Università degli Studi di Firenze

12

VALUTAZIONE DEI PROCESSI FORMATIVI DI TERZO LIVELLO:

CONTRIBUTI METODOLOGICI

a cura di

BRUNO CHIANDOTTO, LEONARDO GRILLI, CARLA RAMPICHINI

VALUTAZIONE E MONITORAGGIO DEI PROCESSI FORMATIVI

Dipartimento di Statistica "*G. Parenti*"

FIRENZE, *DICEMBRE 2005*

campusone

PRESENTAZIONE[1]

L'Ateneo fiorentino da diversi anni dedica particolare attenzione all'attività di valutazione e monitoraggio dei processi formativi grazie anche al notevole impulso dato dal Progetto CampusOne.

La partecipazione dell'Università di Firenze al Progetto CampusOne ha permesso negli ultimi tre anni il perseguimento pressoché completo di alcuni obiettivi ritenuti fondamentali in un'ottica tesa all'innalzamento del livello qualitativo della propria attività formativa. Infatti, in connessione con le attività svolte nel contesto delle ***Azioni di Ateneo*** *relative al* ***Management didattico*** *(Azione 3)****,*** *ai* ***Tirocini e collocamento nel mondo del lavoro*** *(Azione 4) e alla* ***Valutazione della qualità della didattica*** *(Azione 5), sono state programmate e portate a termine indagini il cui obiettivo era quello di acquisire elementi informativi che consentissero di pervenire ad una misura della qualità dei processi formativi offerti dall'Ateneo fiorentino.*

L'attività svolta è consistita nella:

a) raccolta dell'opinione degli studenti frequentanti sulla didattica svolta negli a.a. 2001/02, 2002/03 e 2003/04;
b) valutazione della didattica, delle strutture e dei servizi di supporto alla didattica da parte degli studenti frequentanti e non frequentanti iscritti nell'a.a. 2001/02;
c) misura del carico didattico e valutazione delle modalità di svolgimento delle prove d'esame anche in relazione ai programmi previsti per l'a.a. 2001/02;
d) valutazione ed autovalutazione dei docenti sulla didattica svolta nell'a.a. 2001/02;

[1] Questo lavoro si è avvalso dei contributi previsti nel progetto CAMPUS*ONE* e nel Progetto di Ricerca di Interesse Nazionale (PRIN) "Valutazione del processo di formazione universitaria, sbocchi professionali e pianificazione dei percorsi formativi", anno 2002.

e) *valutazione delle possibili determinanti degli abbandoni e dei tempi di conseguimento del titolo universitario a partire dall'a.a. 1980/81 focalizzando l'attenzione sui laureati dell'anno solare 2000;*

f) *valutazione delle possibili determinanti degli abbandoni degli studi e dei trasferimenti ad altro corso da parte degli studenti immatricolati nell'a.a. 2001/02;*

g) *valutazione della situazione occupazionale dei laureati e diplomati dell'ateneo negli anni solari 1999, 2000, 2001 e 2002;*

h) *indagine sulla condizione occupazionale dei laureati dell'anno solare 1998 a cinque anni dal conseguimento del titolo.*

Il ricco materiale informativo acquisito ha suggerito l'ipotesi di raccogliere il frutto della ricerca del Gruppo VALMON[2] in un'apposita collana al fine di mettere a disposizione di chi opera nel mondo universitario e di chi ne è comunque interessato (giovani, famiglie, mondo del lavoro,…) i risultati di una esperienza che mi auguro possa rivelarsi di qualche utilità.

Questo volume è il n. 12 della Collana VALMON. Essendo la tiratura dei volumi limitata a 500 copie, gli stessi saranno resi disponibili sul sito http://valmon.ds.unifi.it in modo da facilitarne la consultazione da parte di tutti gli interessati.

Bruno Chiandotto

Delegato per la Valutazione della Didattica ed il Monitoraggio dei Processi Formativi
Referente di Ateneo per gli Sbocchi Occupazionali e le Attività di Tirocinio

[2] Il gruppo ***VALMON*** (**Val**utazione e **Mon**itoraggio), coordinato da Bruno Chiandotto e composto da dottorandi e docenti del Dipartimento di Statistica dell'Università degli Studi di Firenze, da diversi anni svolge attività di studio e ricerca nel contesto della valutazione e del monitoraggio dei processi formativi che si svolgono nell'Ateneo fiorentino.

INDICE

PREMESSA

Nel quadro della profonda trasformazione che ha interessato il sistema universitario italiano a seguito della recente riforma dei cicli e degli ordinamenti didattici, oltre che richiesta per legge, è divenuta di primaria importanza la valutazione dei processi formativi. Valutazione che dovrebbe riguardare la capacità delle strutture formative di soddisfare le aspettative dell'utenza sia interna, cioè gli studenti fruitori dei servizi formativi, che esterna, cioè il mondo del lavoro.

La normativa che attualmente regola il Sistema Universitario[3] Italiano attribuisce agli Atenei autonomia finanziaria, manageriale ed organizzativa, riconoscendo la decentralizzazione dei processi decisionali. In particolare, la normativa attribuisce al Ministero dell'Università e della Ricerca (*MIUR*) il compito di definire gli obiettivi principali e le strategie generali di sviluppo del sistema universitario, riconoscendo agli atenei un'ampia autonomia, anche se parte dei finanziamenti accordati è vincolata al soddisfacimento di specifici requisiti.

Ciò ha determinato la necessità della definizione di un sistema di valutazione a struttura piramidale articolato in livelli decisionali distinti: all'apice vi è l'organo di valutazione cosiddetto di I livello rappresentato dal Comitato Nazionale per la Valutazione del Sistema Universitario (*CNVSU*), mentre gli organi di valutazione di II livello si identificano nei Nuclei di Valutazione Interna (*NVI*) dei singoli atenei; ai livelli inferiori si collocano le facoltà ed i corsi di studio.

I principi di decentralizzazione e autonomia inducono in particolare gli organi di I e di II livello, quali responsabili dei risultati ottenuti dalle unità operative loro afferenti, a svolgere intense ed approfondite attività di valutazione e auto-valutazione, in termini di misura di efficienza e di efficacia dell'attività svolta, in un'ottica di qualità. Pertanto, le attività di valutazione si inseriscono nel più ampio contesto della ricerca dell'eccellenza come risposta ad una società in continua evoluzione, dove l'istituzione universitaria non identifica più il solo punto conclusivo di un percorso formativo, ma un riferimento permanente del sapere, del saper fare e del saper essere.

In particolare, la finalità assegnata all'università è quella di creare un vero e proprio incremento del capitale umano, permettendo di sviluppare negli studenti che vi si iscrivono particolari tipologie di competenze che concorrono alla formazione del reddito e, più in generale, allo sviluppo socio-economico del Paese. Tale capitale può venire concettualmente scomposto in una parte manifesta, misurabile tramite le cosiddette abilità specialistiche o professionali (come, ad esempio, la capacità di comprendere e risolvere problemi attraverso opportune strumentazioni tecniche e metodologie acquisite) e in una parte latente, non direttamente misurabile, rappresentata dalle abilità meramente soggettive definite trasversali o personali (come la capacità di organizzazione e progettazione, di collaborare all'interno di gruppi di lavoro, ecc.), oltre che da una parte di competenze di base che contribuiscono a migliorare l'occupabilità e la crescita personale, rendendo più efficaci quelle relativa alla parte manifesta.

La valutazione della qualità della formazione universitaria può essere utilmente collocata nello schema di riferimento proposto da Lockheed e Hanushek nel 1994. Questi due autori suggeriscono una classificazione, riassunta nel prospetto riportato alla pagina seguente, dei criteri su cui impostare analisi valutative circa la qualità della didattica universitaria.

Secondo la concettualizzazione di Lockheed e Hanushek, la stima della performance globale del sistema universitario e, in generale, di un'attività pubblica, può essere scomposta principalmente in tre fasi distinte: la prima relativa al modo in cui le risorse vengono impiegate per ottenere il risultato desiderato (analisi di efficienza); la seconda relativa alla valutazione qualitativa di tale risultato e al grado di raggiungimento degli obiettivi previsti (analisi dell'efficacia); la terza relativa alla percezione soggettiva degli utenti circa il servizio erogato.

[3] DM 509/09 ed anche il DM 270/04, che contiene modifiche al regolamento recante norme concernenti l'autonomia didattica degli atenei ed il DM 49/05 relativo al diploma supplement.

Schema 1 - Ottiche di valutazione della didattica

	INTERNO AL SISTEMA	ESTERNO AL SISTEMA
TERMINI FISICI	o **Efficacia interna** *(effetto dell'ateneo o del corso di laurea sulla capacità di apprendimento dello studente)*	o **Efficacia esterna** *(effetto dell'ateneo o corso di laurea sulla capacità lavorativa del laureato)*
SODDISFAZ.	Soddisfazione dello studente rispetto all'insegnamento	Soddisfazione del laureato rispetto alla condizione lavorativa
TERMINI MONETARI	o **Efficienza interna** *(analisi costi/ricavi aziendali dell'investimento)*	o **Efficienza esterna** *(ritorno economico dovuto al corso di laurea frequentato)*
SODDISFAZ.	Soddisfazione dello studente rispetto alle risorse impiegate	Soddisfazione del laureato rispetto alla condizione economica

Un aspetto fondamentale della qualità misurata in termini di efficienza ed efficacia, è la diversa connotazione che può assumere in relazione al tipo di soggetto interessato. Nel caso della formazione universitaria, in quanto servizio pubblico, i soggetti coinvolti hanno interessi ed aspettative diverse e, talvolta, in conflitto: questi sono gli studenti, il personale docente e non docente, l'ateneo come istituzione nelle sue varie articolazioni (rettorato, facoltà, corsi di studio), il sistema universitario nel suo complesso, e infine la società (le famiglie, il mondo del lavoro e gli enti che a vario titolo promuovono e/o finanziano interventi sulla formazione universitaria).

Il termine "qualità" per uno studente può assumere connotazioni diverse: può, ad esempio, essere collegata al soddisfacimento di bisogni immediati quali il superamento dell'esame, oppure riferito a bisogni futuri quali l'inserimento adeguato nel mondo del lavoro.

Anche per gli erogatori dei servizi formativi il termine "qualità" può assumere connotazioni diverse ma, trattandosi di un servizio pubblico, qualità deve significare, soprattutto, capacità di fornire una risposta complessivamente soddisfacente per la società. Un ateneo efficace e di qualità sarà perciò quella istituzione capace di garantire ai gestori/erogatori (personale docente e non docente) ed ai fruitori dei servizi formativi (gli studenti) e alle parti interessate (mondo del lavoro e, più in generale, all'intera società), certezze riguardo alle proprie capacità di ottenere risultati adeguati agli obiettivi dichiarati e promessi.

I contributi raccolti in questo volume, pur affrontando secondo diverse angolazioni il tema della valutazione della formazione universitaria, trovano nell'approccio statistico l'elemento unificante e si caratterizzano, nella generalità dei casi, per l'elevato contenuto empirico.

Il volume è strutturato in tre parti.

I lavori contenuti nella ***Prima parte*** trattano il ***tema generale della valutazione del sistema universitario*** e illustrano modelli di valutazione e metodologie statistiche utili in questo ambito.

I primi due lavori definiscono l'ambito della valutazione di efficienza e di efficacia del sistema universitario, inquadrando la normativa vigente e descrivendo alcune metodologie e studi di caso. In particolare, il primo lavoro (Bini e Chiandotto, 2003) introduce il problema della valutazione, descrive il nuovo assetto dell'istruzione universitaria in conseguenza della riforma del 2001, e passa in rassegna gli aspetti oggetto di valutazione e gli strumenti metodologici adeguati. Il secondo lavoro (Chiandotto, 2003) propone un modello di valutazione basato sulla teoria delle decisioni e illustra il caso dell'ateneo fiorentino, soffermandosi sulla valutazione della didattica da parte degli studenti frequentanti e sulla valutazione di efficacia esterna in base agli sbocchi occupazionali dei laureati.

Seguono quattro lavori riguardanti metodi statistici utili nell'ambito della valutazione. Il lavoro di Mealli e Chiandotto (2004) illustra il contributo che l'analisi statistica può fornire alla soluzione di problemi decisionali basati su processi di valutazione (ex-ante ed ex-post) facendo specifici riferimenti ad alcune scelte proprie del sistema formativo universitario. Considerando che i dati disponibili ai fini della valutazione del sistema universitario derivano prevalentemente da studi non sperimentali, Mercatanti (2004) affronta il problema dell'inferenza causale in tale contesto, prendendo in esame alcuni aspetti relativi alla gestione dei dati mancanti.

Infine, i due articoli di Riani e Bini (2002) e Bini e Bertaccini (2004) illustrano il metodo della *forward search*, una tecnica robusta di analisi multivariata che consente di identificare unità anomale. Nell'ambito della valutazione tale tecnica permette di individuare istituzioni con livelli di efficienza/efficacia particolarmente alti o bassi, o gruppi di studenti/laureati con caratteristiche peculiari.

Nella ***Seconda parte*** del volume sono raccolti i contributi relativi alla ***valutazione della efficacia interna del sistema universitario***. I primi due lavori applicano metodi robusti di *forward search* per l'individuazione di unità di osservazione anomale. In particolare, il primo lavoro (Bini, 2003) concerne l'individuazione di gruppi di atenei omogenei in base a una serie di indicatori di efficacia interna ed efficienza. Scopo del secondo lavoro (Bini, 2004) è l'analisi dell'abbandono degli studenti dell'Ateneo fiorentino durante il primo anno di studi: la *forward search* consente l'identificazione di gruppi di unità anomale dalla cui composizione strutturale possono essere ricavate informazioni utili all'implementazione di politiche accademiche mirate a ridurre il tasso di abbandono al primo anno di studi.

Anche il lavoro di Chiandotto e Giusti (2004) affronta il problema dell'abbandono universitario, utilizzando un modello multilivello per individuare i fattori esplicativi più rilevanti e i corsi di laurea che presentano tassi di abbandono anomali rispetto a quanto atteso sulla base delle caratteristiche degli iscritti. Sempre ricorrendo ad un modello multilivello, il lavoro di Chiandotto e Variale (2004) descrive le determinanti della durata degli studi universitari; inoltre l'analisi dei residui di secondo livello permette di ottenere interessanti informazioni per quanto riguarda il cosiddetto "effetto corso di laurea" sui tempi di conseguimento del titolo degli studenti.

I due lavori di Grilli e Rampichini (2002, 2003) sviluppano modelli multilivello idonei all'analisi di variabili di risposta di tipo ordinale, quali quelle rilevate nell'ambito della valutazione della didattica da parte degli studenti frequentanti. In particolare, nel primo lavoro si suggerisce l'utilizzo di un modello che tenga esplicitamente conto della possibilità che gruppi di studenti abbiano un diverso metro di giudizio. Nel secondo lavoro il modello viene esteso all'analisi congiunta di più quesiti misurati su scala ordinale.

L'articolo di Rampichini, Grilli e Petrucci (2004) passa in rassegna vari metodi per l'analisi dei giudizi degli studenti e propone un indicatore multi-dimensionale basato sulle stime provenienti da opportuni modelli multilivello.

Infine, il lavoro di Pratesi (2004) propone un metodo basato sul *propensity score matching* per correggere la distorsione da auto-selezione. Tale distorsione si verifica, per esempio, quando si utilizzano dati provenienti da questionari compilati via internet dagli studenti.

La ***Terza parte*** del volume raccoglie i lavori che trattano del problema della ***valutazione dell'efficacia esterna del sistema universitario***. Il primo contributo (Chiandotto, Bini e Bertaccini, 2004) propone un modello strutturale per l'analisi della soddisfazione dei laureati, adattando al caso dell'Università un noto indice utilizzato nell'ambito della *customer satisfaction*.

I lavori di Chiandotto (2003), Chiandotto e Bacci (2004) e Grilli e Mealli (2004) affrontano il problema dell'inserimento dei laureati nel mondo del lavoro. L'articolo di Chiandotto (2003) fa il punto sulle modalità e le finalità della valutazione degli sbocchi occupazionali dei laureati, soffermando l'attenzione sulla situazione dell'Università di Firenze. Il lavoro di Chiandotto e Bacci (2005) si basa su un modello multilivello per individuare le determinanti dell'inserimento occupazionale e misurare il valore aggiunto apportato da ciascun corso di laurea. In un ottica diversa, Grilli e Mealli (2005) propongono l'utilizzo di metodi di inferenza causale al fine di

valutare l'efficacia relativa di due specifici corsi di laurea rispetto all'inserimento occupazionale dei propri laureati.

Il lavoro di Bertaccini (2004) affronta il problema dell'utilizzo delle competenze apprese durante il corso di studi, ricorrendo a tecniche di analisi robusta al fine di individuare gruppi anomali di laureati.

Grilli e Rampichini (2004) soffermano l'attenzione sulla specificazione e stima di un modello multilivello per valutare se e in che misura le competenze utilizzate dai laureati siano state acquisite durante gli studi universitari. Inoltre, Grilli e Rampichini (2003) propongono un modello fattoriale multilivello per lo studio della soddisfazione dei laureati in merito a vari aspetti relativi al lavoro svolto.

Infine, Chiandotto (2004) discute gli aspetti metodologici relativi alla misura della soddisfazione concentrando l'attenzione su una misura indiretta ed ex-post quale il grado di soddisfazione per il lavoro dei neo-laureati.

SOMMARIO

Parte III - Efficacia esterna

1 Chiandotto B., Bini M., Bertaccini B. (2005) Valutazione della qualità della formazione universitaria percepita dai laureati e diplomati dell'Ateneo fiorentino: un'applicazione del modello ECSI, in *Modelli statistici per l'analisi della transizione Università-lavoro*, a cura di C. Crocetta, pp. 87-106, Cleup, Padova.

2 Chiandotto B. (2003) La situazione occupazionale dei laureati: dall'indagine alla pianificazione degli interventi sui processi formativi, in *Transizione Università-lavoro: la definizione delle competenze*, pp. 1-18, a cura di M. Civardi, Cleup, Padova.

3 Chiandotto B., Bacci S. (2005) Un modello multilivello per l'analisi della condizione occupazionale dei laureati, in *Modelli statistici per l'analisi della transizione Università-lavoro*, a cura di C.Crocetta, pp. 211-234, Cleup, Padova.

4 Grilli L., Mealli F. (2005) L'effetto degli studi universitari sull'occupazione: un'applicazione dell'approccio degli "strati principali" all'analisi causale, in *Modelli statistici per l'analisi della transizione Università–lavoro*, a cura di C. Crocetta, pp. 131-154, Cleup, Padova.

5 Bertaccini B. (2004) Valutazione del processo di formazione universitaria: un'analisi robusta dell'efficacia, in *Strategie metodologiche per lo studio della transizione Università-lavoro*, a cura di E. Aureli Cutillo, pp. 37-56, Cleup, Padova.

6 Grilli L., Rampichini C. (2004) A Polytomous Response Multilevel Model with a Non Ignorable Selection Mechanism, *Proceedings of the 19th International Workshop on Statistical Modelling*, pp.194-198, Firenze 4-8 Luglio 2004, Firenze University Press, Firenze.

7 Grilli L., Rampichini C. (2003) A Multilevel Analysis of Graduates' Job Satisfaction, in *Transizione Università-lavoro: la definizione delle competenze*, a cura di M. Civardi, pp. 133-147, Cleup, Padova.

8 Chiandotto B. (2004) Sulla misura della qualità della formazione universitaria, *Note e Studi di Economia*, n. 3, pp. 27-61.

La valutazione del sistema universitario italiano alla luce della riforma dei cicli e degli ordinamenti didattici

Da: Bini M., Chiandotto B. (2003) La valutazione del sistema universitario italiano alla luce della riforma dei cicli e degli ordinamenti didattici, Note e Studi di Economia, n. 2, pp. 29-61.

LA VALUTAZIONE DEL SISTEMA UNIVERSITARIO ITALIANO ALLA LUCE DELLA RIFORMA DEI CICLI E DEGLI ORDINAMENTI DIDATTICI

MATILDE BINI* - BRUNO CHIANDOTTO**

Premessa

La valutazione del sistema universitario ha assunto in questi ultimi anni, sia a livello nazionale che internazionale, importanza fondamentale per le sue vaste implicazioni nei diversi contesti: politico, economico e sociale.

Valutazioni e giudizi con riferimento a persone, imprese e istituzioni, a processi e a risultati, costituiscono un'attività che è sempre stata svolta anche se con modalità spesso informali e non basate su «elementi oggettivi»; l'attività di valutazione formalizzata, cioè basata su approcci sistematici, si è invece molto sviluppata solo negli ultimi decenni – e ciò è spesso dovuto alle molte leggi e normative che la impongono – divenendo un strumento irrinunciabile del management anche delle attività, dei programmi e delle politiche di intervento in campo economico e sociale delle amministrazioni pubbliche, soprattutto laddove si producono servizi alla persona di pubblica utilità (Gori, Vittadini, 1999).

In termini del tutto generali due possibili definizioni di valutazione sono quelle proposte da Ramsden (1988): «un giudizio sistematico del valore o del merito di un qualche oggetto»; e da Feldman (1999): «un'acquisizione sistematica e una valorizzazione delle informazioni atte a fornire un utile feedback circa un determinato oggetto». Entrambe le definizioni descrivono la valutazione come uno sforzo sistematico ed entrambe usano in maniera deliberatamente ambigua il termine «oggetto» che potrebbe fare riferimento ad un programma, ad un'azione, ad una metodologia, così come ad un oggetto in senso stretto, o ad un servizio; la seconda definizione, comunque, rispetto alla prima, enfatizza l'acquisizione e la valutazione finalizzata delle informazioni raccolte piuttosto che limitarsi ad una semplice espressione di valore o di merito. Una definizione più articolata di valutazione è quella propo-

* *Dipartimento di Statistica «G: Parenti». Università degli Studi di Firenze. E-mail: bini@ds.unifi.it.*
** *Dipartimento di Statistica «G: Parenti». Università degli Studi di Firenze. E-mail: chiandot@ds.unifi.it.*

sta da Biggeri (2000): «un'attività di studio e di analisi dei risultati potenziali o effettivi di un programma (o progetto), di una politica d'intervento o di specifiche attività, che si conclude con un atto o documento, più o meno formale, contenente un giudizio di rispondenza o meno dei risultati ad obiettivi o standard determinati a priori e l'indicazione degli eventuali cambiamenti nel programma che si ritengono opportuni».

La valutazione secondo quest'ultima definizione assume, pertanto, la veste di attività strategica a tutti i livelli quale strumento di sostegno, scientifico, dei processi decisionali; per la *verifica*, *ex-ante* delle possibilità, *in itinere* (in corso di realizzazione), conclusiva (a conclusione dei programmi), ed *ex-post* (quando si vedono i frutti delle attività implementate) dell'effettiva realizzazione degli obiettivi programmati. Valutazione intesa quindi non più come semplice attività di ricerca, controllo e giudizio, ma come esercizio finalizzato a supportare decisioni, strategie e eventuali azioni future: «valutare per decidere» (Chiandotto, 2002) un processo dinamico e continuo il cui scopo ultimo è finalizzato al perseguimento di specifici obiettivi.

Ovviamente ogni processo di valutazione presenta connotazioni diverse a seconda del contesto di riferimento ed al momento in cui viene svolto, ma la finalità che si vuol perseguire attraverso il processo è sempre la stessa ed è quella di innescare, anche attraverso strategie d'incentivazione basate sui risultati della valutazione stessa, un sistema di azioni e retroazioni teso al miglioramento della qualità generale delle attività svolte.

Come è noto, la nozione di qualità[1] intesa come il grado con cui un prodotto (o un servizio) soddisfa determinati requisiti, nasce nel mondo della produzione dei beni e in particolare nell'ambito dei sistemi d'organizzazione aziendale, e successivamente si diffonde nell'ambito dei servizi anche di pubblica utilità. In quest'ultimo contesto i criteri

[1] Negli ultimi anni il concetto di qualità si è notevolmente trasformato ed arricchito. Esistono tre principali tipi di approcci (Gori e Vittadini, 1999) al problema della qualità che, nel tempo, si sono modificati. L'approccio *tradizionale*, secondo cui la qualità è concepita come esclusività, prestigio e posizione di vantaggio; proprio a causa di tale concezione era posseduta solo da pochi che potevano accedervi. Successivamente si è assistito ad un impegno da parte degli esperti alla definizione di un approccio al problema di tipo *scientifico-razionalista*, in cui le caratteristiche dei prodotti e dei servizi venivano definite da studiosi, da esperti esterni o professionisti. In tale concezione la qualità era legata al raffronto con degli obiettivi, utilizzando metodi razionali d'analisi dei risultati, senza però impiegare il punto di vista degli utenti. Nell'approccio *manageriale*, la qualità è definita dal grado di soddisfazione dell'utente/cliente, che assume un ruolo centrale nell'analisi, in un ambito competitivo. Per contentare il cliente si rinforzano le posizioni di base con lo scopo di essere più attenti ai bisogni dell'utente. L'utenza può esprimere il grado di soddisfazione rispetto ai servizi erogati, ma non può partecipare attivamente al processo di definizione dei servizi. L'approccio *consumistico* considera la partecipazione attiva da parte dell'utente/cliente nel disegno e nella erogazione dei servizi.

di valutazione cui usualmente si fa riferimento riguardano l'efficacia (grado di raggiungimento degli obiettivi), l'efficienza (grado di ottimizzazione dell'uso delle risorse) e l'equità rispetto all'accesso ai beni e servizi pubblici (Palumbo, 1995).

I termini «efficacia» ed «efficienza», che sono ormai entrati nel linguaggio comune, vengono, purtroppo, spesso scambiati e adattati ai vari contesti senza una definizione precisa, creando sovente, specialmente negli interlocutori meno esperti, confusione e qualche perplessità.

Riguardo l'efficacia, le definizioni sono molteplici e variegate: rapporto tra risultati e obiettivi, in particolare si ha *efficacia interna*, vista come confronto tra risultati e obiettivi relativi a un singolo servizio e alla sua azione; *efficacia esterna*, analizzata tramite la relazione tra obiettivi assegnati e domanda sociale; *efficacia come impatto*, ovvero come confronto tra risultati ottenuti con l'azione e risultati ottenibili senza l'azione; *efficacia come valore aggiunto* derivante del raffronto tra le prestazioni degli agenti, rapporto tra domanda soddisfatta e domanda potenziale, grado di soddisfazione dei bisogni, ecc.

L'efficienza è generalmente definita come «*i*l rapporto tra la somma degli input richiesti per produrre un certo output e quello stesso output» (Hatry, 1986). Le risorse impiegate e l'output possono essere considerate in quantità oppure in valore, dando luogo rispettivamente ai concetti d'efficienza tecnica o produttività ed efficienza economica.

Un altro aspetto fondamentale della qualità misurata in termini di efficienza ed efficacia, è la diversa connotazione che può assumere in relazione al tipo di soggetto interessato. Nel caso della formazione universitaria, in quanto servizio pubblico, i soggetti coinvolti hanno interessi ed aspettative diverse e, talvolta, in conflitto: questi sono gli studenti, il personale docente e non docente, l'ateneo come istituzione nelle sue varie articolazioni (rettorato, facoltà, corsi di studio), il sistema universitario nel suo complesso, e infine la società (le famiglie, il mondo del lavoro e gli enti che a vario titolo promuovono e/o finanziano interventi sulla formazione universitaria).

Il termine «qualità» per uno studente può assumere connotazioni diverse: può, ad esempio, essere collegata al soddisfacimento di bisogni immediati quali il superamento dell'esame, oppure riferito a bisogni futuri quali l'inserimento adeguato nel mondo del lavoro.

Anche per gli erogatori dei servizi formativi il termine «qualità» può assumere connotazioni diverse ma, trattandosi di un servizio pubblico, qualità deve significare, soprattutto, capacità di fornire una risposta complessivamente soddisfacente per la società. Un ateneo efficace e di qualità sarà perciò quell'istituzione capace di garantire ai gestori/erogatori (personale docente e non docente) ed ai fruitori dei servizi formativi (gli studenti) e alle parti interessate (mondo del lavoro e, più in generale, all'intera società), certezze riguardo alle proprie capacità di ottenere risultati adeguati agli obiettivi dichiarati e promessi.

Ma le università sono strutture organizzative complesse, la cui gestione si svolge a diversi livelli; di conseguenza, la qualità del sistema universitario deve essere anche verificata ai diversi livelli decisionali.

1. **Il sistema universitario italiano**

L'università italiana sta vivendo in questi anni un processo di profonda trasformazione, determinato da una ampia autonomia che le è stata riconosciuta coerentemente con il principio costituzionale di cui all'articolo 33 della Costituzione della Repubblica Italiana che così si esprime: «L'arte e la scienza sono libere e libero ne è l'insegnamento» e «le Università hanno il diritto di darsi ordinamenti autonomi nei limiti stabiliti dalle leggi dello Stato».

Il principio costituzionale doveva trovare attuazione tramite l'approvazione di una o più leggi ordinarie in grado di definire il quadro di regole per l'autonomia delle università; in loro assenza, il sistema delle università italiane è rimasto fortemente accentrato fino alla fine degli anni Ottanta. Le università appartenevano all'Amministrazione della Pubblica Istruzione che assegnava annualmente le risorse finanziarie per il funzionamento, vincolate a specifiche destinazioni (stipendi, straordinari, canoni di locazione). I posti di ruolo del personale erano assegnati dal Ministero. I concorsi per l'assunzione del personale docente erano gestiti dal Ministero della Pubblica Istruzione. I corsi degli studi erano compiutamente definiti da decreti ministeriali. Quest'assetto, durato circa quarant'anni, precludeva alle università la possibilità di individuare la propria missione specifica, di definire i propri obiettivi di medio e lungo periodo, di gestire il personale in base a politiche attive delle risorse umane, di progettare autonomamente i percorsi formativi.

Il processo di attuazione del principio costituzionale di autonomia è iniziato nel 1989, con l'emanazione della Legge 168, che oltre ad istituire il Ministero dell'Università e della Ricerca Scientifica (MURST), ha sancito l'autonomia didattica, scientifica, organizzativa, finanziaria e contabile delle università, rinviando a norme successive la definizione dei relativi principi e prevedendo che, ove queste non fossero state emanate, gli atenei avrebbero potuto darsi proprie norme organizzative con gli statuti. In seguito a ciò si è avvertita l'esigenza di istituire un sistema di valutazione della struttura universitaria italiana, per monitorare e controllare le attività didattiche ed amministrative in essa svolte.

L'attenzione mostrata negli ultimi anni dall'opinione pubblica, dalle imprese e dagli operatori sociali nei confronti dell'università, deriva dalla consapevolezza che il buon funzionamento di quest'istituzione in un contesto di completa autonomia è sempre più importante per garantire l'efficienza e la dinamicità dell'intero sistema produttivo. Le «risorse umane» sono considerate una fonte di ricchezza, il livello elevato di istruzione della popolazione è uno dei fattori alla base dello sviluppo economico e un requisito indispensabile per competere sui mer-

cati nazionali ed internazionali; di conseguenza, si può capire come e perché da oltre quindici anni si assista nella maggior parte dei paesi industrializzati ad un ripensamento critico del ruolo e delle funzioni dell'istruzione superiore e, in particolar modo, di quella universitaria. La necessità di incrementare e migliorare la produzione e la circolazione di conoscenza ha focalizzato la discussione intorno al tema dell'efficacia e dell'efficienza del sistema universitario: da più parti si insiste sulla necessità di rendere più chiari ed espliciti gli obiettivi e le finalità della formazione superiore e delle sue istituzioni, in modo da valutare oggettivamente la qualità del sistema formativo in relazione alla spesa sostenuta ed effettuare previsioni strategiche adeguate circa l'offerta e la domanda futura di istruzione.

La necessità della riforma universitaria, e della riforma della scuola a tutti livelli, era già molto sentita alla metà degli anni Ottanta, perché nel sistema educativo italiano erano presenti elementi critici tali da richiedere mutamenti strutturali radicali; fattori critici essenzialmente riconducibili a tre elementi fondamentali: in primo luogo, gli anni di scolarizzazione obbligatoria inferiori agli standard europei, l'età di uscita dalla scuola dell'obbligo di 15 anni (articolo 1 c. 3 Legge 10 febbraio 2000, n. 30) contro la media europea di 16, anche se recentemente «al fine di potenziare la crescita culturale e professionale dei giovani … è progressivamente istituito, a decorrere dall'anno 1999-2000, l'obbligo di frequenza di attività formative fino al compimento del diciottesimo anno di età. Tale obbligo può essere assolto in percorsi anche integrati di istruzione e formazione: a) nel sistema di istruzione scolastica; b) nel sistema della formazione professionale di competenza regionale c) nell'esercizio dell'apprendistato» (articolo 68 c.1 Legge 17 maggio 1999, n. 144), e l'età in cui termina l'obbligo risulta fortemente anticipata rispetto all'età del conseguimento della maturità (per il conseguimento di un diploma di maturità dopo la scuola dell'obbligo occorre completare un ciclo di studi in genere di 5 anni, rispetto ai 3 o 4 negli altri paesi).

In secondo luogo, l'offerta formativa italiana di livello secondario e terziario risultava poco diversificata; l'articolazione dei percorsi, anche per gli istituti tecnici e professionali, era incentrata principalmente su un modello di istruzione con una forte prevalenza di formazione in aula di tipo tradizionale, mentre risultava scarsamente finalizzato alla professionalizzazione delle risorse umane in funzione delle esigenze del mercato del lavoro. In questo modo, il sistema scolastico italiano finiva per non offrire adeguate opportunità formative alternative a coloro che non hanno attitudine agli studi e non garantiva un'adeguata formazione per l'ingresso nel mercato del lavoro; tale inadeguatezza e le difficoltà di assorbimento del mercato del lavoro tendono ad incoraggiare la permanenza nel sistema.

In terzo luogo, il sistema formativo non appariva abbastanza efficiente, non essendo in grado di contenere adeguatamente la dispersione, con la conseguente diffusione dei fenomeni di abbandono prima

del completamento del corso di studi intrapreso e del conseguimento del titolo di studio corrispondente.

Le carenze sopra segnalate danno ragione del grande afflusso all'università in Italia anche se, in proposito, non si può non tenere conto della mancanza di un sistema di istruzione terziario realmente alternativo a quello accademico: una volta conclusa la scuola superiore, le opportunità di formazione aggiuntiva erano quasi esclusivamente di tipo universitario (fatta eccezione per le Accademie e i Corsi di Formazione Professionale di secondo livello), per cui solo una quota molto ridotta dei ragazzi che decidono di proseguire gli studi oltre le secondarie superiori inizia un corso di tipo non universitario.

Il deficit nell'offerta di corsi brevi rispetto a quelli lunghi ha determinato, corrispondentemente ad una scarsa spendibilità del titolo finale sul mercato del lavoro, una ridotta affluenza di studenti ai diplomi universitari. Mentre all'estero i corsi di diploma godono di una lunga ed accreditata tradizione e costituiscono il primo gradino della formazione universitaria, i percorsi didattici dei diplomi universitari e delle lauree italiane erano per lo più paralleli e non sequenziali: l'iscrizione ad un corso di laurea di uno studente che ha già conseguito un diploma universitario non comportava, nella maggior parte dei casi, il riconoscimento integrale del percorso formativo svolto, ma solo di alcuni esami superati.

Dato quanto appena sottolineato, è naturale che fosse bassissima la percentuale di giovani italiani della corrispondente fascia d'età (Tabella 1) che ha concluso un corso di tipo breve (0,3 per cento), mentre tale quota risulta più alta in tutti gli altri paesi considerati e spesso, come nel caso di Belgio (25,4 per cento), Danimarca (23,3 per cento), Finlandia (22,3 per cento) e Irlanda (21 per cento), con scarti di notevole entità. Ovviamente, la situazione appare assai diversa considerando i cicli lunghi (5-6 anni di durata legale): l'Italia, con 15 laureati per 100 giovani di età corrispondente, risulta ai primi posti fra i paesi europei, collocandosi perfino al di sopra degli standard di paesi quali Svezia e Danimarca.

Tuttavia, il tasso di conseguimento del titolo universitario risultava insoddisfacente, se confrontato con la vistosa quota annuale di immatricolazioni; una delle caratteristiche distintive del nostro sistema universitario risiedeva, e risiede tuttora anche se a livelli più attenuati, nell'elevata quota di abbandoni dell'università prima del conseguimento del titolo, quota che nel 1999 sfiorava il 70 per cento degli immatricolati (Tabella 1).

In Italia, la quota di popolazione in età lavorativa 25-64 anni in possesso di un titolo universitario (Tabella 1) è decisamente bassa (9 per cento), di gran lunga inferiore a quella riscontrata in Finlandia (31 per cento), Svezia (29 per cento) e Danimarca (27 per cento), a conferma del fatto che il nostro Paese è fanalino di coda rispetto all'Europa in tema di istruzione universitaria. Esaminando le singole fasce di età,

Tabella 1 – **Indicatori del sistema di istruzione terziaria negli stati dell'Unione Europea (1999) (valori %)**

	Iscritti al 1° anno di un corso universitario per 100 giovani di età corrispondente		*Giovani in possesso di un titolodi studio post-secondario per 100 giovani di età corrispondente*				*Percentuale di studenti che abbandonano gli studi universitari (1996)****	*Percentuale di popolazione (25-64 anni) in possesso di un titolodi studio universitario*				
Nazioni	*Programmi brevi (ISCED 5B)**	*Programmi lunghi (ISCED 5A)***	*Titolo non universitario (ISCED 4)*	*Programmi brevi (ISCED 5B)*	*Programmi lunghi (ISCED 5A)*			*25-34 anni*	*35-44 anni*	*45-54 anni*	*55-64 anni*	*25-64 anni*
					MEDIUM (3-4 anni)	*LONG (5-6 anni)*						
Austria	8	28	-	10,5	0,9	11,1	47	13	12	11	6	12
Belgio	26	30	22,7	25,4	0,9	5,8	37	34	28	22	15	26
Danimarca	34	34	1,2	23,3	-	-	33	29	28	27	19	27
Finlandia	-	-	0,8	22,3	16,4	17,5	25	38	35	29	20	31
Francia	21	35	1,1	17,9	18,5	5,6	45	31	21	18	12	21
Germania	13	28	15,4	11,8	5,2	10,8	28	22	26	24	20	23
Grecia	-	-	13,5	-	-	-	-	26	21	15	9	18
Irlanda	26	28	25,8	21,0	24,8	1,2	23	29	22	16	11	21
Italia	1	40	2,6	0,3	1,1	14,9	66	10	11	10	5	9
Lussemburgo	-	-	4,2	-	-	-	-	21	17	21	12	19
Paesi Bassi	1	54	0,5	0,9	32,3	1,2	30	24	25	21	17	22
Portogallo	-	-	-	-	-	-	51	12	10	9	6	10
Regno Unito	28	45	-	11,4	35,6	1,2	19	27	26	24	19	25
Spagna	11	46	12,7	5,4	12,8	17,5	-	33	23	15	9	21
Svezia	5	65	-	2,7	25,9	1,3	-	32	31	30	22	29
Media OCSE	*15*	*45*	*8,5*	*12,2*	*18,8*	*5,8*	-	*25*	*23*	*20*	*14*	*22*

Fonte: OCSE, *Education at a Giance*, 2001.

* I dati si riferiscono ai corsi «brevi» di primo livello; per l'Italia si tratta dei diplomi universitari.
** I dati si riferiscono ai corsi «lunghi» di primo livello; per l'Italia si tratta delle lauree.
*** 1995 per Danimarca, Francia, Germania e Irlanda; 1993 per il Portogallo.

si rileva che la percentuale di popolazione italiana in età 55-64 anni in possesso di un titolo di studio universitario è il 5 per cento, un valore già distante da quello degli altri paesi europei con i quali dobbiamo confrontarci. Il divario esistente, inoltre, aumenta in maniera preoccupante considerando le fasce di età più giovani, per le quali risulta che in Italia soltanto il 10 per cento della popolazione di età compresa fra i 25 e i 34 anni è in possesso di un titolo universitario, contro la media OCSE del 25 per cento.

Il confronto fra le due classi di età estreme mette in evidenza gli effetti delle politiche adottate dai vari paesi per migliorare gli esiti dell'istruzione universitaria. Mentre l'Italia passa dal 5 per cento al 10 per cento, la Francia passa dal 12 per cento al 31 per cento e la Spagna dal 9 per cento al 33 per cento, indicando così una più lenta e difficoltosa espansione dell'istruzione universitaria nel nostro Paese.

Infine, anche il sistema di istruzione terziaria extra-accademica italiano mostra segnali di debolezza, evidenziati dal fatto che, come già sottolineato, la quota di giovani in possesso di tale titolo risulta una delle più basse tra tutti i paesi considerati: solo il 2,6 per cento dei giovani italiani ha conseguito un titolo di istruzione terziaria non universitaria, a fronte delle quote ben più elevate di Irlanda (25,8 per cento), Belgio (22,7 per cento) e Germania (15,4 per cento).

Riguardo alle disfunzioni sopra richiamate, si deve segnalare che il recente processo di riordino del sistema universitario nazionale (DM 509/99) punta a modificare la situazione descritta. A partire dall'anno accademico 2001/02, infatti, il sistema italiano di studi universitari, in armonia con quelli degli altri paesi dell'Unione Europea, si articola su tre livelli: un primo triennio a cui segue un biennio specialistico, e quindi un successivo triennio che rilascia il titolo di Dottore di Ricerca. A questi livelli potranno affiancarsi altri percorsi di studio mirati all'aggiornamento costante (Master, Corsi di perfezionamento, Corsi di specializzazione) e rivolti sia a studenti che a professionisti.

La riforma prevede, tra l'altro, un'architettura del sistema degli ordinamenti didattici basata su classi di corsi di laurea, che rilasciano una laurea di primo livello dopo l'acquisizione di 180 crediti formativi universitari (CFU), da conseguirsi nell'arco temporale di un triennio. La laurea di secondo livello (*laurea specialistica*) potrà essere conseguita acquisendo, nell'arco temporale di un biennio, ulteriori crediti formativi, fino al raggiungimento di complessivi 300 crediti. Nelle classi di corsi di laurea vengono definiti gli obiettivi qualificanti, le attività formative indispensabili per conseguirli, e tra queste anche le attività di tirocinio da svolgere presso aziende, enti o istituzioni pubbliche o private, il numero minimo di crediti per attività formativa e per ambito disciplinare.

Con l'introduzione del sistema dei crediti formativi universitari, viene definito e misurato il carico di lavoro richiesto allo studente per raggiungere i diversi traguardi formativi; considerando che uno stu-

dente, ogni anno, può dedicare 1.500 ore del proprio tempo allo studio (studio individuale, lezioni, laboratori, stage), 1.500 ore corrispondono a 60 CFU e, conseguentemente, un credito corrisponde a 25 ore di impegno. I crediti non valutano il profitto e rimangono indipendenti dal voto conseguito negli esami o nelle verifiche di altro genere, ma non si acquisiscono se non si superano le prove di accertamento previste, e servono come strumento di tutela del diritto alla mobilità fra percorsi formativi all'interno di un ateneo e dell'intero sistema universitario, italiano e europeo.

Sul fatto che la riforma dei cicli e degli ordinamenti didattici dell'istruzione universitaria, voluta per risolvere i problemi in cui si dibatte il nostro sistema di istruzione superiore, consentirà effettivamente il perseguimento degli obiettivi previsti non si registra l'unanimità dei consensi. La discussione e gli approfondimenti conoscitivi che su questo tema si sono susseguiti negli ultimi tempi, pure se significativi e molto articolati, non hanno prodotto, infatti, conclusioni universalmente condivise: a fronte di convinti sostenitori, che ritengono la riforma potenzialmente in grado di risolvere i problemi in cui si dibatte il Sistema Universitario Italiano, si collocano molti dissenzienti[2] i quali sostengono che dalla riforma non ci si possa attendere altro che effetti perversi e implicazioni sfavorevoli traducibili in costi sociali, economici e culturali destinati ad aggravarsi nel tempo.

2. **L'evoluzione del sistema di valutazione delle università: la recente normativa italiana**

Con la Legge 168 del 1989, richiamata, si consacra il principio di autonomia mediante la sua concreta attuazione, ma già intorno alla metà degli anni Ottanta il dibattito sulla valutazione del sistema universitario aveva iniziato a prendere consistenza. Il primo dispositivo normativo che introduce meccanismi di valutazione è, come già precedentemente menzionato, contenuto nella Legge 168 del 1989, che precisa i confini dell'autonomia finanziaria e contabile delle università, e stabilisce in che modo il nascente Ministero dell'Università e della Ricerca Scientifica e Tecnologica (MURST) attui «forme di controllo interno sull'efficienza e sui risultati di gestione complessiva delle Università».

In conformità a tale normativa molte università istituirono organismi di controllo interno, che però erano generalmente orientati alla verifica dell'attività amministrativa.

Per avere l'istituzione formale di un sistema di valutazione nell'università italiana è necessario attendere il dicembre del 1993, quando la Legge 537 – all'interno di un più generale disegno di revisione della Pubblica Amministrazione – allargò i confini dell'autonomia gestiona-

[2] In proposito si possono consultare, tra gli altri, Monti e Briganti, 2002; Dilorenzo e Stefani, 2003.

le ed organizzativa degli atenei stabilendo l'istituzione dei Nuclei di Valutazione Interna (NVI) nelle singole università e l'Osservatorio Nazionale per la Valutazione del Sistema Universitario (ONVSU).

A livello centrale quindi veniva posto l'Osservatorio (l'attuale Comitato Nazionale per la Valutazione del Sistema Universitario), mentre a livello di ciascun ateneo venivano istituiti i singoli Nuclei di Valutazione. Il sistema non è stato pensato però come una struttura «unica» con al vertice l'Osservatorio; le linee di responsabilità e le interazioni tra i diversi attori coinvolti (Ministero, Comitato, Nuclei, atenei) sono abbastanza complesse e in alcuni aspetti poco o contraddittoriamente definite.

Il Nucleo (Legge 537/93, articolo 5, comma 22) «ha il compito di verificare, mediante analisi comparative dei costi e dei rendimenti, la corretta gestione delle risorse pubbliche, la produttività della ricerca e della didattica, nonché l'imparzialità ed il buon andamento dell'azione amministrativa. I Nuclei determinano i parametri di riferimento del controllo anche su indicazione degli organi generali di direzione, cui riferiscono con apposita relazione almeno annualmente». La relazione deve essere trasmessa (Legge 537/93, articolo 5, comma 22) al MURST, al Consiglio Universitario Nazionale e alla Conferenza permanente dei Rettori, per garantire un doppio controllo, sulla valutazione dei risultati (relativi all'efficienza e alla produttività delle attività di ricerca e di formazione) e sulla verifica dei programmi di sviluppo e di riequilibrio del sistema universitario, anche al fine di provvedere ad una razionale assegnazione delle risorse.

Una nuova legge (Legge 19 ottobre, n. 370 – GU n. 252 del 26.10. 1999) dispone norme volte a disciplinare, più compiutamente, la valutazione del sistema universitario: «Le università adottano un sistema di valutazione interna della gestione amministrativa, delle attività didattiche e di ricerca, degli interventi di sostegno al diritto allo studio …». (articolo 1, comma 1). «Le funzioni di valutazione di cui al comma 1 sono svolte in ciascuna università da un organo collegiale disciplinato dallo statuto dell'università, denominato "nucleo di valutazione di ateneo"»… «Le università assicurano ai nuclei l'autonomia operativa, il diritto di accesso ai dati e alle informazioni necessari, nonché la pubblicità e la diffusione degli atti, nel rispetto della normativa a tutela della riservatezza. I nuclei acquisiscono periodicamente, mantenendo l'anonimato, le opinioni degli studenti frequentanti sulle attività didattiche e trasmettono un'apposita relazione, entro il 30 aprile di ciascun anno, al Ministero dell'università e della ricerca scientifica e tecnologica, e al Comitato per la valutazione del sistema universitario…» (articolo 1, comma 2). «Le Università che non applicano le disposizioni di cui ai commi 1 e 2 entro sei mesi dalla data di entrata in vigore della presente legge sono escluse per un triennio dal riparto dei fondi relativi alla programmazione universitaria, nonché delle quote di riequilibrio …» (articolo 1, comma 3).

La stessa legge, all'articolo 2, prevede l'istituzione del Comitato Nazionale per la Valutazione del Sistema Universitario. «…Il Comitato: a) fissa i criteri generali per la valutazione delle attività delle Università previa consultazione della Conferenza dei Rettori delle Università Italiane (CRUI) del Consiglio Universitario Nazionale (CUN) e del Consiglio Nazionale degli Studenti Universitari (CNSU), ove costituito; b) promuove la sperimentazione, l'applicazione e la diffusione di metodologie e pratiche di valutazione; c) determina ogni triennio la natura delle informazioni e i dati che i nuclei di valutazione degli atenei sono tenuti a comunicare annualmente; d) predispone ed attua, sulla base delle relazioni dei nuclei di valutazione degli atenei e delle altre informazioni acquisite, un programma annuale di valutazioni esterne delle Università o di singole strutture didattiche, approvato dal MURST (ora MIUR), con particolare riferimento alla qualità delle attività universitarie sulla base di standard riconosciuti a livello internazionale nonché della raccomandazione 98/561/CE del Consiglio, del 24 settembre 1998, sulla cooperazione in materia di garanzia della qualità nell'istruzione superiore; e) predispone annualmente una relazione sulle attività di valutazione svolte; f) svolge i compiti assegnati dalla normativa vigente, all'Osservatorio per la valutazione del sistema universitario ...; g) svolge, su richiesta del MURST (MIUR), ulteriori attività consultive, istruttorie, di valutazione, di definizione di standard, di parametri e di normativa tecnica, anche in relazione alle distinte attività delle università, nonché ai progetti e alle proposte presentate dalle medesime».

3. **L'attività di valutazione del sistema universitario: contenuti ed obiettivi**

L'attività di valutazione della formazione universitaria è scandita da un complesso sistema di azioni-decisioni e retroazioni, atte a fornire risposte esaurienti ai seguenti quesiti:

— perché valutare?
— chi deve valutare?
— cosa valutare?
— come valutare?

Non vi è dubbio che alla prima domanda corrisponde l'attività di verifica dell'esistenza o meno di problemi emergenti che interessano la formazione e che sono stati finora discussi.

Con i due interrogativi successivi, invece, si intendono definire i contenuti della valutazione che riflettono, in questo ambito, rispettivamente i soggetti e gli aspetti della formazione; mentre il quarto, identifica gli strumenti (principalmente rappresentati da indicatori) per il processo valutativo, come verrà discusso in seguito.

Se ci si limita a considerare l'attività di formazione, risulta facile procedere all'individuazione dei soggetti interessati al processo di va-

lutazione, nonché delle loro interrelazioni: il *Ministero dell'Istruzione, dell'Università e della Ricerca* (MIUR), gli atenei, le facoltà e i corsi di studio che assieme alla classe dei docenti assumono la duplice veste di soggetti valutati e valutatori; gli studenti, le famiglie e i datori di lavoro, questi ultimi interessati alle capacità acquisite durante il periodo della formazione universitaria.

Riguardo al cosa valutare, bisogna considerare la molteplicità di aspetti che caratterizzano il sistema universitario; ciascuno aspetto dovrebbe costituire oggetto di una specifica attività di valutazione.

a) Valutazione dell'insegnamento e dell'apprendimento

Le università hanno, come ovvio, la missione di insegnare e di fare apprendere. Questo duplice obiettivo corrisponde al fatto che esse devono non soltanto disseminare un elevato livello di conoscenza tra gli studenti, ma soprattutto devono fare apprendere ai giovani strumenti utili per la loro futura occupazione nel mondo del lavoro, come pure fornire un continuo addestramento per coloro che già lavorano.

In Italia solo recentemente (DM 509/99) la disseminazione della conoscenza si è organizzata in lauree di I e di II livello (entrambi orientati verso le professioni), susseguite da masters e dottorati di ricerca (questi ultimi con l'obiettivo di formare i laureati con elevate capacità acquisite). Questa nuova articolazione dovrebbe sostanzialmente produrre una diversificazione della popolazione studentesca secondo l'età, la condizione personale, le possibilità di seguire i corsi (tempo pieno, parziale, a distanza).

L'attività di valutazione di questo nuovo sistema dovrebbe essere effettuata sia da soggetti interni al sistema universitario che valutano i contenuti e l'organizzazione dei corsi di studio (gli studenti e i docenti in primo luogo) allo scopo di migliorarne il livello, sia da soggetti interni non accademici (NVI, MIUR) interessati non solo alla didattica ma anche alle capacità delle strutture e dei servizi di supporto alla didattica offerti. Scopo comune a tutti questi soggetti è una migliore allocazione delle risorse destinate alle attività della didattica: questo obiettivo è «vincolato» alle risorse finanziarie disponibili e, di conseguenza, il contenuto della valutazione riguarda anche l'efficienza oltre che l'efficacia.

Altri problemi che si riscontrano riguardano, ad esempio, la riluttanza del corpo docente nei confronti della valutazione della didattica da parte degli studenti e la necessità , con la connessa difficoltà, di far tacere corsi e/o insegnamenti in presenza di scarsità di risorse e/o di studenti iscritti.

Infine, la valutazione fatta da soggetti esterni quali le famiglie e il mondo del lavoro incide solo marginalmente sui contenuti e l'organizzazione della didattica, a differenza di quella fatta degli studenti che sempre più sono chiamati in prima persona ad esprimere la loro opinione sugli insegnamenti e sui servizi di cui fruiscono.

b) Valutazione della ricerca

Le università svolgono oltre che didattica anche ricerca allo scopo di contribuire ad uno sviluppo economico e sociale: la ricerca di base e la ricerca applicata. La valutazione della ricerca, nonché delle attività, risorse, processi e risultati ad essa connessi, è un'attività cui solo recentemente viene dedicata l'attenzione che merita[3].

In Italia (come già da tempo nella maggior parte dei paesi europei) si sta diffondendo la tendenza a far svolgere attività di valutazione anche a soggetti «esterni» alla ricerca. Il servizio pubblico si pone perciò in maggior relazione con l'università, tenuto conto del fatto che già preesisteva un'attività di valutazione per quanto riguarda la gestione economica e finanziaria delle università.

La valutazione della ricerca interessa sempre più gruppi di unità operative piuttosto che singoli individui, rendendo così più complessi i processi di valutazione stessa. In termini di qualità, intesa principalmente come grado di partecipazione ai progetti e misura dei risultati conseguiti rispetto agli obiettivi prefissati, ma anche come sviluppo di conoscenza per l'organizzazione, reputazione esterna e dotazione di tecnologie (Minelli, Rebora e Turri, 2002), essa richiede inevitabilmente l'impiego di referenti, esperti nel campo di ricerca, e di test di verifica. In entrambi i casi, l'attuale tendenza è l'uso di standard internazionali di qualità.

Gli effetti negativi che si possono riscontrare in questo tipo di attività sono, soprattutto nella ricerca senza rischi, quello di non dare importanza al raggiungimento degli obiettivi di ricerca prefissati, di stimolare la produzione di pubblicazioni facili e in tempi brevi.

In questi ultimi anni la valutazione della ricerca ha assunto un'importanza fondamentale dal punto di vista finanziario: si creano centri di eccellenza per ricevere risorse finanziarie aggiuntive. L'atteggiamento è tuttavia quello di reperire fondi sia in seno all'università che al suo esterno, di concepire nuovi temi di ricerca, ma l'attribuzione dei fondi deve essere controbilanciata dalla presenza di risultati ottenuti.

c) Valutazione dell'efficacia esterna della formazione

Le università devono preparare gli studenti al mondo del lavoro, in particolare alla creazione o aggiornamento di competenze richieste dai cambiamenti nei sistemi di produzione e dal dinamismo del mercato. Nonostante l'importanza di questi due compiti, la valutazione dell'efficacia esterna della formazione rimane un tipo di attività facoltativa, non formalizzata e caratterizzata da una grande varietà di soggetti, contenuti ed obiettivi, utilizzando una varietà di strumenti di valutazione.

---

[3] In proposito si possono utilmente consultare: Breno *et al*., 2003; Vitale e Cerroni (a cura di), 2003.

Gli aspetti possibili da valutare sono diversi: creazione di diplomi orientati verso le professioni, addestramento continuo, inserimento dei laureati/diplomati nel mercato del lavoro, rapporto università-territorio. Lo sviluppo di una valutazione che si occupa dell'inserimento professionale dei laureati/diplomati dipende dalla situazione del mercato del lavoro, dal peso dell'università in un contesto territoriale, dalla specificità dei titoli offerti. Le strutture, le metodologie, le misure sono varie: indagini tempestive, osservatorio all'interno dell'università, osservatorio regionale che collabora con un osservatorio nazionale, ecc.

La ridotta attività di valutazione relativa al rapporto formazione-occupazione-territorio può essere spiegata da una serie di ostacoli dovuti al fatto che, da un lato, i laureati, nella generalità dei casi, mantengono pochi contatti con la loro università rendendo molto problematici i possibili riscontri sull'adeguatezza della formazione ricevuta; dall'altro lato, i datori di lavoro, pur dichiarandosi in linea di principio molto interessati alla collaborazione con le università, in casi molto sporadici danno seguito operativo ai loro propositi.

Le metodologie utilizzate per misurare l'inserimento professionale dei laureati/diplomati dovrebbero tenere conto sempre di un problema centrale: le difficoltà di trovare un lavoro sono dovute alla scarsa qualità degli insegnamenti, propri della laurea conseguita e/o dal deterioramento del mercato del lavoro dovuto ad altri fattori? La valutazione del rapporto università-territorio mette in discussione la diversità dei territori: qual'è lo spazio pertinente per la valutazione? I risultati dell'interrelazione fra il sistema educativo e l'ambiente economico sociale (in particolare il mercato del lavoro), e anche culturale sono particolarmente difficili da cogliere ed interpretare, perché i parametri da considerare sono molteplici.

I dati che consentono un significativo arricchimento conoscitivo sulla qualità e sul livello della preparazione dei laureati e diplomati possono essere acquisiti attraverso rilevazioni che vedono coinvolti i laureati e diplomati stessi, ai quali sono richieste informazioni sulla loro condizione occupazionale, sull'utilizzo delle conoscenze acquisite all'università e sulla qualità dell'eventuale attività lavorativa svolta.

Le indagini sui laureati e diplomati costituiscono uno strumento conoscitivo molto importante della transizione istruzione universitaria-mercato del lavoro, poiché possono combinare e mettere in relazione due scopi convergenti: quello di rilevare la condizione occupazionale dei laureati/diplomati ad un certo intervallo dal conseguimento del titolo e, soprattutto, quello di verificare l'adeguatezza e la qualità del servizio formativo offerto.

Già da molti anni si procede in Italia allo svolgimento di indagini sulla condizione occupazionale dei laureati sia a livello nazionale (ISTAT) che a livello locale da parte di università, facoltà o singoli corsi di studio, anche in collaborazione con enti territoriali.

A livello del tutto generale, un'indicazione sulla utilizzazione dei giovani in possesso di un titolo di studio universitario da parte del sistema economico si può avere sulla base delle informazioni raccolte dall'ISTAT con l'Indagine Trimestrale sulle Forze di Lavoro, in quanto essa fornisce, a livello regionale o di grande ripartizione geografica, i dati sugli occupati, sui disoccupati e su coloro che sono in cerca di lavoro, per titolo di studio conseguito.

Più mirate sono le indagini che l'Istituto Nazionale conduce sui laureati e sui diplomati universitari a tre anni dal conseguimento del titolo: «Indagine sull'inserimento professionale dei laureati...»[4]. Tuttavia, anche questi dati (provenienti da indagini campionarie di notevoli dimensioni e, quindi, tali da consentire conclusioni statisticamente significative a livello nazionale per grandi aree territoriali) non sono in grado di soddisfare l'esigenza fortemente sentita a livello di singolo ateneo di adeguate analisi di efficacia esterna della propria formazione universitaria. Da qui la realizzazione di numerosissime rilevazioni statistiche, sulla transizione università-mercato del lavoro e sull'inserimento professionale di coloro che hanno conseguito un titolo di studio universitario, di interesse di un singolo ateneo o gruppi di atenei.

Allo stato attuale, le rilevazioni più note, anche per il loro più diffuso impatto, riguardo alla transizione università-lavoro sono quelle svolte dall'Osservatorio Statistico dell'Università degli studi di Bologna nell'ambito del progetto ALMALAUREA[5] e, con valenza territorialmente molto più limitata, dalle università lombarde nel contesto del progetto VULCANO[6]. Tra le moltissime iniziative che hanno riguardato questo tema si segnalano, a titolo puramente indicativo l'attività svolta dall'Università degli studi di Pisa nell'ambito del progetto DIOGENE[7], dall'Università degli studi di Padova nell'ambito del proprio

---

[4] Si vedano le pubblicazioni ISTAT, 2003, 1996, 1994, 1994a, 1990.

[5] Il *progetto* ALMALAUREA, la banca dati dei laureati e dei diplomati del sistema universitario italiano per il mondo del lavoro e delle professioni, nasce in via sperimentale nel 1993 per iniziativa dell'Osservatorio Statistico dell'Università di Bologna e comprende attualmente 33 atenei italiani (*Bologna, Bari, Basilicata, Cassino, Catania, Catanzaro, Chieti, Cosenza (Univ. Della Calabria), Ferrara, Firenze, Foggia, Genova, Messina, Milano-IULM, Modena e Reggio Emilia, Molise, Padova, Parma, Perugina, Piemonte Orientale, Reggio Calabria, Roma «LUMSA», Roma Tre, Salerno, Sassari, Siena, Torino, Torino-Politecnico, Trento, Trieste, Udine, Venezia-IUAV, Verona*).
Per ulteriori informazioni, si può consultare il sito Internet www.almalaurea.it.

[6] VULCANO (Vetrina Universitaria Laureati con Curricula per le Aziende Navigabile On-line) è un progetto promosso dalle università lombarde, in collaborazione con il Consorzio Interuniversitario Lombardo per l'Elaborazione Automatica (CILEA). Le università, in ordine di adesione al progetto, sono *Pavia, Brescia, Milano, Bergamo, Milano-Bicocca, Insubria, Cattolica, Bocconi, Politecnico, S.Raffaele*. Per ulteriori informazioni, si possono consultare i siti delle università lombarde.

[7] DIOGENE è un progetto promosso dall'Università degli studi di Pisa per facilitare l'inserimento professionale dei giovani laureati e diplomati dell'Ateneo. Per ulteriori informazioni, si può consultare il sito Internet www.unipi.it/~diogene/.

Osservatorio sul Mercato Locale del Lavoro[8], dall'Università degli studi di Firenze[9] nel contesto della propria attività di valutazione e monitoraggio dei processi formativi.

Le schede di rilevazione utilizzate nelle indagini sopra menzionate sono abbastanza simili; in una sezione, si acquisiscono informazioni sulla soddisfazione della scelta universitaria di formulare un giudizio complessivo sull'esperienza stessa, sui motivi dell'iscrizione, su eventuali attività di qualificazione post-laurea e sull'attuale condizione occupazionale. In una sezione, destinata agli occupati al momento dell'intervista, si acquisiscono informazioni sulla posizione occupazionale, il ramo di attività, la collocazione geografica, le modalità di ottenimento del lavoro, la pertinenza del titolo e più in generale delle competenze acquisite all'università con l'attività svolta, il grado di soddisfazione di alcuni aspetti del lavoro svolto. Si acquisiscono informazioni anche sui non occupati al momento dell'intervista ma che hanno svolto una qualche attività lavorativa dopo il conseguimento del titolo: informazioni sulla posizione occupazionale passata, sulle modalità di ricerca del lavoro, sulla pertinenza dell'attività svolta con le competenze acquisite all'università e sui motivi dell'interruzione.

Una sezione è usualmente riservata ai laureati che non lavorano, in questa sezione si prevede l'acquisizione di informazioni sui motivi della «non ricerca», per coloro che non cercano lavoro, oppure il tipo di lavoro cercato e le azioni compiute in tale direzione, per coloro che cercano lavoro. Una sezione è dedicata all'acquisizione di informazioni sulla famiglia d'origine dell'intervistato.

d) Valutazione del personale docente

Il corpo docente ha come compiti fondamentali l'attività didattica e di ricerca. Le persone sono già valutate al momento del loro reclutamento e nel corso della loro carriera. Questo tipo di valutazione, basata soprattutto sull'attività di ricerca, viene svolta da personale accademico: in particolare da docenti dello stesso settore disciplinare e da docenti del dipartimento di afferenza.

Il reclutamento del corpo docente non si basa tuttavia sul giudizio della persona ma anche sulle politiche di impiego adottate in una determinata università, fortemente condizionate dalle risorse finanziarie disponibili e acquisibili.

Sull'attività didattica si è sviluppato un tipo di valutazione basata sulla raccolta dell'opinione degli studenti e sull'autovalutazione dei docenti.

---

[8] L'Università degli studi di Padova ha avviato dal 2000 un *Osservatorio sul Mercato Locale del Lavoro* che comprende sia rilevazioni dirette presso gli imprenditori e i dirigenti di enti pubblici del Veneto, sia l'osservazione sistematica del mercato del lavoro mediante un'indagine sui laureati e diplomati della stessa università.

[9] Per ulteriori informazioni si può consultare il sito www.unifi.it/aut_dida/indexval.html.

e) Valutazione del personale non docente

Vi sono varie figure di personale non docente che lavorano nell'ambito accademico: managers, ingegneri, tecnici, personale amministrativo... L'attività di valutazione riguarda le singole persone (reclutamento, avanzamento di carriera, mobilità, …) ma anche l'efficienza dell'attività svolta dall'ufficio dove queste lavorano (sia internamente, sia rispetto ai terzi).

Gli obiettivi di valutazione sono molteplici: misurare l'efficacia della gestione, utilizzare il personale nel modo più efficiente ai fini dei compiti tecnico-amministrativi dell'università, creare indicatori di performance e qualità dei servizi offerti, semplificare e razionalizzare le strutture amministrative, trovare un equilibrio fra centralizzazione e decentralizzazione della gestione economico-finanziaria[10]. E ancora, avere una migliore conoscenza del personale, controllare l'adempimento delle regole amministrative, uniformare i carichi di lavoro, rendere più professionale e responsabile il personale, creare nuove funzioni e nuovi posti di lavoro.

Il processo di valutazione del personale non docente è piuttosto lento, soprattutto perché coinvolge un gran numero di persone. Gli aspetti che sono oggetto di valutazione sono soprattutto l'efficienza, la soddisfazione del lavoro, le mansioni, la ripartizione di ruoli, i sistemi di pagamento (per il lavoro e per le prestazioni specifiche).

f) Valutazione delle politiche di governo rettorale

Le università analizzano i bisogni della società e degli utenti? Come si decidono a livello di governo del rettorato gli obiettivi da realizzare? Quali priorità da tenere presenti? Quali risorse stanziate alle priorità? Come si valutano i risultati? Chi sono i soggetti che governano? Questi sono i quesiti su cui attivare una valutazione delle politiche di governo. Il governo dell'università e in particolare il rettore, giocano un ruolo chiave nello sviluppo delle valutazioni potendo addirittura escludere determinati argomenti dalla valutazione e/o rifiutare la valutazione esterna.

Una conseguenza che può verificarsi con la valutazione è il consolidamento di un governo specifico, quello presidenziale o più precisamente quello manager-presidenziale (rettore forte e linea gerarchica amministrativa forte che funziona secondo criteri imprenditoriali). Tuttavia, il perpetuarsi di un tal governo è gestito dalle alleanze o dai compromessi passati con i due governi tradizionali dell'università, quello collegiale (non si può abolire l'influenza dei corpi accademici) e quello burocratico (con una gerarchia amministrativa che controlla l'esecuzione delle regole installate dal servizio pubblico).

---

[10] Per alcuni studi inerenti al tema si vedano, tra gli altri: Minelli, Rebora e Turri, 2002; Catalano, 2002; Azzone, 2003.

g) Valutazione dei mezzi finanziari

In Italia il ruolo dei fondi pubblici nel finanziamento dell'università è un elemento predominante. In una situazione di continua evoluzione dell'istruzione superiore, le autorità pubbliche desiderano il controllo e la razionalizzazione dei mezzi finanziari stanziati alle università.

Valutare i mezzi finanziari significa valutare le risorse e le spese di ogni istituzione e dei relativi componenti. Le risorse e le spese sono presentate sotto la forma di preventivo (risorse e spese per il periodo successivo) e/o di un bilancio (risorse e spese di un periodo precedente). Il processo di preparazione, discussione e voto dal consiglio di amministrazione dell'università, è un tipo di valutazione interna dei mezzi finanziari. Il preventivo ed i bilanci, invece, costituiscono lo strumento principale per le valutazioni effettuate dagli enti esterni alla valutazione.

Il bilancio dell'università è regolato dalla contabilità pubblica, ciò limita l'autonomia finanziaria delle università e di conseguenza, le università hanno talvolta istituito strutture più flessibili e più riservate come ad esempio fondazioni o associazioni, per gestire la completa autonomia finanziaria specialmente per la ricerca o le continue attività di formazione.

In questi ultimi anni si sta affermando la tendenza a effettuare valutazione interna, con nuovi meccanismi di ripartizione e nuove forme di verifica di distribuzione delle risorse fra le diverse strutture (rettorato, facoltà e dipartimenti) presenti nell'università.

h) Valutazione delle strutture

Una delle conseguenze dell'accresciuta dimensione della popolazione studentesca e dell'espandersi dell'offerta formativa in termini qualitativi e quantitativi, è l'aumentata necessità di adeguate strutture di supporto.

Ai fini dell'analisi tre tipi grandi di strutture si possono identificare: strutture accademiche tradizionali (facoltà, corsi di studio, dipartimenti, centri di ricerca...), strutture di sostegno per l'istruzione e ricerca (biblioteche, laboratori, centri di calcolo...), strutture non accademiche (per i servizi amministrativi e tecnici di tutta la gestione universitaria).

La valutazione delle strutture è essenzialmente decisa dalle singole università nell'ambito della propria autonomia, si può trattare di valutazioni interne cui partecipano studenti, docenti personale non docente, oppure di valutazioni esterne ma decise congiuntamente dall'amministrazione centrale dell'università.

4. **Un sistema informativo per la valutazione del sistema universitario**

Sulla base di quanto finora detto riguardo l'intreccio normativo e tutti i soggetti di vario livello preposti, ufficialmente o meno, alla valutazione, emerge in modo evidente la necessità di definizione e strutturazio-

ne di un sistema informativo adeguato alle necessità e nel quale sia presente una chiara articolazione dei diversi livelli e momenti decisionali, delle connessioni tra tali livelli e momenti e, soprattutto, la rilevanza – come punto centrale di riferimento – dell'attività di analisi e formalizzazione dei processi decisionali stessi.

L'esigenza di poter disporre di una base informativa sui settori di possibile intervento deve non solo rendere possibile una conoscenza approfondita sulle varie specificità ma deve, al tempo stesso, consentire l'impostazione di politiche di intervento che siano, al tempo stesso, realizzate in tempi utili e che siano coerenti con i fabbisogni reali delle collettività interessate; infatti, non è infrequente imbattersi in situazioni nelle quali si giustificano decisioni affrettate e non sufficientemente ponderate argomentando sulla carenza nella qualità e/o nella quantità dell'informazione disponibile. Ma una carenza informativa può emergere solo dal confronto con le esigenze conoscitive che l'informazione stessa deve soddisfare, ed è proprio questa necessità di esame comparativo quella che muove la critica, che consente l'evidenziazione delle carenze, che impone un riesame continuo di una qualunque base informativa e che suggerisce la revisione delle procedure di raccolta ed elaborazione dei dati alla luce dei nuovi fatti emergenti.

Spesso, l'operatore pubblico nell'attivare i propri processi decisionali prescinde completamente dalle informazioni anche se qualche volta si basa su dei dati statistici; ma il dato statistico non è necessariamente informazione: il sistema informativo, che è l'unico supporto adeguato di ogni efficace processo decisionale, non si identifica nel sistema statistico.

Un sistema informativo comprende, infatti, l'insieme di tutte le attività che vanno dalla raccolta, alla elaborazione alla trasmissione delle informazioni, sia statistiche che non statistiche, sia formali che informali, finalizzate al perseguimento di specifici obiettivi di governo, di gestione e/o di controllo. Ed è soltanto in tale contesto che il dato (statistico) riesce ad evidenziare tutte le sue potenzialità, ed è soltanto in tale prospettiva che si può parlare di «qualità» del dato e di «valore» dell'informazione.

Per poter consentire lo sfruttamento ottimale del sistema informativo è necessario che il sistema stesso risponda ai requisiti sotto elencati:

— chiara definizione delle necessità degli utilizzatori e delle loro richieste;
— possibilità di un accesso appropriato e continuo ai dati e ai servizi offerti;
— inventario di tutti i dati disponibili e/o acquisibili e delle informazioni pertinenti l'ambiente circostante;
— necessità di costruire insiemi di dati in modo sistematico ed integrato;
— accesso appropriato a connessioni internet che assicurino il raggiungimento in remoto di centri di dati e di sistemi di informazioni.

Tenendo inoltre presente che un sistema informativo che riguarda i processi di formazione e gestione universitaria deve consentire il conseguimento dell'obiettivo generale della formazione stessa che è quello dell'innalzamento degli standard qualitativi dei processi e che tale obiettivo implica l'adozione di politiche di intervento a vari livelli che, a loro volta, comportano una scelta tra un insieme di alternative possibili, ne consegue la necessità di un'attività di valutazione e monitoraggio che dovrebbero essere basate sui criteri sotto elencati:

1. avere una struttura semplice che rilevi i parametri per il controllo;
2. prevedere tutte le dimensioni necessarie per una descrizione esauriente dei processi in itinere e dei loro prevedibili sviluppi;
3. non deve essere una descrizione fine a se stessa, ma un atto che sottintende propositi di intervento tesi all'innalzamento della qualità;
4. deve indurre organizzazione, ma non deve risolversi in puri aspetti organizzativi.

Se si tiene presente il recente avvio della riforma dei cicli e degli ordinamenti didattici e ci si limita a considerare l'attività didattica che si svolge nelle università, si può ragionevolmente ritenere che, o perché richieste dalla normativa vigente o perché, comunque, utili ai vari livelli decisionali (Ministero, Atenei, Facoltà, Corsi di studio, docenti, personale non docente, studenti, famiglie e mondo del lavoro), le informazioni concernenti i processi formativi debbano essere quantomeno tali da consentire un monitoraggio ed una valutazione adeguata:

— delle risorse (finanziarie, strutture, personale docente e non docente) destinate alla didattica;
— delle carriere degli studenti a livello individuale;
— del carico didattico complessivo e dei crediti attribuiti ai vari insegnamenti;
— dei singoli corsi di studio (immatricolazioni, abbandoni, conseguimenti del titolo, ecc.);
— della didattica e dei servizi di supporto alla didattica (segreterie, orientamento, tutorato, biblioteche, ecc.);
— delle attività di orientamento;
— delle attività di tirocinio;
— delle attività didattiche e di tirocinio svolte in collaborazione con università e/o enti, e/o imprese non italiane (progetti speciali: Socrates/Erasmus, Leonardo, ecc.);
— delle attività che sono previste nell'ambito di progetti speciali (Campus, Campus Like, Campus*One*, TRIO, ecc.);
— dell'attività svolta per facilitare l'inserimento dei laureati/diplomati nel mondo del lavoro;
— degli sbocchi occupazionali dei laureati e/o diplomati.

Attività di monitoraggio e valutazione, quelle sopra elencate, certamente utili ai fini di un corretto ed efficiente governo dell'università finalizzato all'impiego ottimale delle risorse e all'innalzamento dei li-

velli qualitativi dei processi che si svolgono nell'ambito delle università[11]. Al riguardo si segnala che una quota non indifferente dei dati cui s'è fatto sopra riferimento sono già disponibili in molte università italiane, o perché costituiscono parte dell'archivio di ateneo, o perché raccolti attraverso apposite indagini; un patrimonio quello disponibile molto ricco ma che, nella generalità dei casi, necessità di un'adeguata risistemazione nel contesto di un sistema informativo adeguato. L'impressione generale è che i sistemi informativi già realizzati in molti atenei italiani siano incompleti e, spesso, configurati in modo non ottimale.

5. **La necessità di un insieme adeguato di indicatori per la valutazione**

Quanto è stato detto finora dovrebbe aver fornito un'idea, seppure di prima approssimazione, su come ci si dovrebbe muovere per avviare a soluzione i problemi in cui si dibatte il sistema universitario italiano.

Tutti questi aspetti (e molti altri neppure menzionati) che risultano fondamentali nel «governo dell'università», hanno un'estrema rilevanza e complessità ed hanno meritato la particolare attenzione che è stata loro dedicata negli ultimi anni; attenzione che si è soffermata, in modo particolare, sulla individuazione e definizione di indicatori adeguati cui fare riferimento nella programmazione e gestione dell'attività didattica e di ricerca che si svolge nelle università.

Molte valutazioni di attività svolte nelle università, in particolare dell'attività didattica, sono già state effettuate, sia in termini di esperienze pilota che come attività di routine, dalla Conferenza dei Rettori e dai Nuclei di Valutazione Interna delle università, dal Comitato Nazionale per la Valutazione del Sistema Universitario, e da diversi gruppi di ricerca o singoli studiosi. Tali esperienze, pur essendo a volte parziali e non estese a tutte le università, hanno indubbiamente fornito e forniscono utili indicazioni, operative e metodologiche.

Tuttavia, molte di esse hanno messo in evidenza che le valutazioni sono sempre «fattibili», ma non sempre altrettanto facilmente accettate dagli «attori» (rettori, direttori amministrativi e personale, in particolare docente) che ne sono coinvolti. Per cui, anche per evitare gli effetti della valutazione e per esigenze «campanilistiche», si criticano i criteri e/o gli indicatori adottati e i metodi di valutazione impiegati, e si cerca di convincere i *policy makers* (il rettore a livello di università e il ministro a livello nazionale) a non mettere in atto le azioni che le valutazioni indicano come necessarie.

---

[11] Sui problemi connessi alla valutazione del Sistema universitario nelle sue varie articolazioni si possono utilmente consultare: Rowley, 1996; Cave, Hanney, Henkel e Kogan, 1997; Modica e Stefani, 1997; Biggeri, 1998 e 1999; Gola, 1998; Université de Toulon, 1998; Bini, 1999; Gori, 2000; Gori e Vittadini, 1999; Catalano, 2002; Chiandotto, 2002; Minelli, Rebora, Turri, 2002; Il progetto Campus*One* al sito Internet www.campusone.it).

Le critiche prevalenti riguardano la non applicabilità della valutazione al caso specifico di interesse (cosa questa che è facilmente dimostrabile essere non vera): gli eccessivi costi e, soprattutto per quanto riguarda gli aspetti statistici, la carenza di dati validi e la non validità degli indicatori impiegati e delle comparazioni tra unità.

È evidente che la valutazione per garantire, ad esempio, la qualità delle attività universitarie e il suo sviluppo richiede costi più o meno elevati (anche in termini comportamentali e psicologici), in relazione alle situazioni strutturali e gestionali di partenza delle singole unità, e può confliggere con gli interessi costituiti, facendo sorgere «resistenze» all'interno del sistema universitario; quando invece si può rilevare, anche dalle esperienze degli altri paesi, che i costi sono relativamente non elevati e che, comunque, alla lunga ripagano ampiamente in termini di risultati conseguiti dalle singole università e dal sistema universitario nel suo complesso.

Per quanto riguarda le critiche all'uso degli indicatori spesso si sostiene che essi: *i)* non sono validi per descrivere il fenomeno di interesse e distorcono la realtà; *ii)* non sono facilmente interpretabili; *iii)* le loro sintesi hanno scarso significato; *iv)* la loro eventuale applicazione per l'attribuzione di incentivi può comportare effetti indesiderati; *v)* nella generalità dei casi non consentono adeguati confronti; *vi)* possono indurre comportamenti impropri con conseguenze negative (Smith, 1995).

Le critiche non sono prive di fondamento: spesso gli indicatori sono inaccurati e/o non validi concettualmente, a volte manca il costrutto e perfino la definizione operativa (Schmitz, 1993); inoltre, le analisi descrittive basate su valori medi non sono sempre di facile interpretazione e non consentono approfondimenti conoscitivi adeguati. Infatti, i dati utilizzati per la costruzione di singoli indicatori di efficacia secondo le modalità di alcune variabili e le differenze riscontrate, non consentono di individuare e valutare correttamente i fattori o caratteristiche (individuali e di contesto) che li influenzano. Altre critiche riguardano l'uso degli indicatori che può risultare inappropriato, o la loro non corretta interpretazione. Si tratta di critiche che dovrebbero stimolare un approfondito lavoro di ricerca teso alla messa a punto di adeguati strumenti di analisi e non giustificano in alcun modo il loro mancato impiego, impiego che in molti casi si è rivelato di estrema utilità. In proposito non si può non sottolineare che il rifiuto dell'uso di indicatori comporta, implicitamente, il rifiuto di qualunque tipologia di valutazione quantitativa.

Relativamente all'adeguatezza degli strumenti statistici di analisi da impiegare risulta opportuno sottolineare che, per il modo in cui è organizzata l'attività di formazione, i dati rilevati si presentano con una struttura gerarchica secondo livelli successivi: si parte dagli individui (gli studenti) di cui si conoscono determinate informazioni (sesso, età, ecc.) che sono e/o possono essere raggruppati nei vari corsi di

studio e, successivamente, per facoltà e/o per sede universitaria. Ne consegue che le analisi descrittive tradizionali, basate su indicatori medi non consentono, ad esempio, di isolare il contributo del corso di laurea e della sede universitaria al risultato conseguito, non consentono, cioè, di verificare quanta parte della variabilità dei risultati osservati sia imputabile al contesto organizzativo dell'università (ai corsi di laurea, alla facoltà e alle sedi) ed alla qualità complessiva del servizio offerto e quanto alle caratteristiche individuali degli studenti; elementi informativi questi che risultano essenziali per una migliore pianificazione e attuazione di eventuali politiche di intervento volte ad incrementare il livello di efficacia. A tal fine, e per misurare gli effetti sui risultati, è necessario che i dati su cui si sviluppa l'analisi siano raccolti a livello individuale e che siano elaborati rispettando la loro struttura gerarchica; finalità questa che può essere perseguita impiegando appropriate metodologie di analisi quali i modelli statistici multilivello (Goldstein, 1995) eventualmente inseriti in adeguati sistemi ad equazioni strutturali (Heck and Thomas, 2000; Rabe-Hesketh *et al.*, 2004).

I risultati conseguibili applicando le metodologie statistiche di analisi appena citate permettono di «misurare» l'effetto del contesto universitario «a parità di condizioni», cioè al netto dell'influenza di altre caratteristiche, che non sono proprie dei corsi di laurea, delle facoltà e/o delle sedi universitarie consentendo cosi a tutte le parti interessate e, in particolare, agli studenti (e alle famiglie) di valutare per quale corso e in quale sede il servizio di formazione offre i migliori esiti.

6. **Gli indicatori proposti dalla Conferenza dei Rettori (CRUI) e dal Comitato Nazionale per la Valutazione del Sistema Universitario (CNVSU)**

Nel 1995 la CRUI ha curato la predisposizione di un documento:«Organizzazione e metodi dei Nuclei di valutazione nelle università italiane. Le proposte della Conferenza dei Rettori» che contiene un elenco molto dettagliato e puntuale di indicatori.

Nel 1998, l'Osservatorio per la Valutazione del Sistema Universitario (l'attuale CNVSU), in attesa di una seconda generazione di indicatori CRUI, propone (DOC 11/98) ai Nuclei di Valutazione Interna di inserire nella relazione annuale un «insieme minimo di 22 indicatori» che, anche se incompleto, sia tale da permettere una visione sufficientemente esplicativa del funzionamento dei vari atenei italiani. In tale ottica si prevede la raccolta di una serie di variabili tali da consentire l'elaborazione di alcuni *indicatori di risultato*, tenendo conto delle risorse disponibili (*indicatori di risorse*), del modo con cui tali risorse sono trasformate in prodotti (*indicatori di processo*) e dell'ambiente in cui l'ateneo si trova ad operare (*indicatori di contesto*).

Se si analizza il documento predisposto dall'Osservatorio appare subito evidente la necessità di un'estensione della base informativa; è

presumibile che l'esigenza fortemente sentita di non sovraccaricare di lavoro i costituendi nuclei di valutazione interna degli atenei abbia indotto a limitare in modo eccessivo il numero degli indicatori rispetto a quelli suggeriti dalla CRUI. Delle carenze informative presenti nell'insieme minimo di indicatori, l'Osservatorio prende coscienza immediatamente; infatti, se si scorre quanto previsto nel documento: «Note tecniche sui dati ed informazioni da trasmettere entro il 2 maggio 2000» si rileva che il numero degli indicatori passa da 22 a 29, numero questo che subisce un ampliamento nell'anno successivo, e per rendersi conto di ciò è sufficiente scorrere il documento: «Note tecniche e dati da trasmettere entro il 30 aprile 2001» predisposto dal CNVSU, ormai subentrato all'Osservatorio. Merita sottolineare che, sia nel documento del 1998 che in quelli successivi, non viene affrontato il problema dell'aggregazione, delle informazioni che si desumono dai vari indicatori, finalizzata all'ottenimento di una misura sintetica dell'attività degli atenei ma si sottolinea, tra l'altro, la necessità di procedere alla definizione di una griglia di pesi da assegnare agli indicatori stessi; problema questo che (per quanto è di conoscenza degli autori di questa nota) è ancora in attesa di una soluzione.

Relativamente all'attività didattica, tenendo anche conto dell'avvio della riforma universitaria, si deve osservare che l'insieme degli indicatori previsti dal CNVSU, nonostante il loro sostanziale incremento rispetto alla proposta iniziale, non sembra coprire tutte le dimensioni d'interesse. A sostegno di questa affermazione si può addurre il documento sull'accreditamento dei corsi di studio (RdR 01/01) predisposto da un apposito gruppo di lavoro nell'ambito delle attività del Comitato. Dall'esame degli elementi da prendere in considerazione ai fini dell'accreditamento di un corso di studi, si rileva come alcune dimensioni non trovino un adeguato corrispettivo nell'elenco degli indicatori previsti dal Comitato.

Tornando alle Note tecniche sopra citate, è apprezzabile la strategia adottata per avanzare le richieste di dati ai nuclei; infatti, negli elenchi delle variabili richieste in tutte le rilevazioni dei Nuclei (da«Nuclei 2000» fino a «Nuclei 2003») non si procede ad un'aggregazione per categoria di indicatori (contesto, risorse, processo e risultato) ma ad un'aggregazione delle variabili per classi di «oggetti»: nella Classe A, vengono inclusi dati relativi agli studenti iscritti, laureati e diplomati; nella Classe B, dati relativi al personale in servizio; nella Classe C, dati finanziari; nella Classe D, altri dati. È da presumere che la procedura utilizzata per richiedere i dati ai nuclei sia stata suggerita da ragioni di opportunità con l'intento di semplificare il lavoro di predisposizione del materiale richiesto. Il materiale raccolto dovrebbe, ovviamente, essere elaborato e reinserito nello schema iniziale che prevede l'aggregazione degli indicatori in quattro categorie distinte secondo la natura degli indicatori stessi. Infatti, è proprio tale strutturazione quella che sembra meglio rispondere alle necessità di individuare gli elementi es-

senziali per la definizione e costruzione di un sistema informativo finalizzato alla risoluzione ottimale dei problemi da affrontare ai vari livelli decisionali.

Nell'ultimo documento predisposto dal Comitato (Note tecniche su dati e informazioni da trasmettere entro il 30 aprile 2003) sono riassunti il dettaglio e le specificazioni sulle informazioni relative all'insieme di variabili da utilizzare per la costruzione di indicatori sull'intero sistema universitario, con particolare riferimento a quelle che dovranno essere raccolte e trasmesse a cura dei Nuclei di valutazione. I dati utilizzati sono stati distinti in 6 sezioni e riguardano: gli studenti, i corsi di studio ed i pareri degli studenti frequentanti; il personale; gli aspetti finanziari; le strutture disponibili; la ricerca scientifica; l'avvio della riforma degli ordinamenti didattici.

Per quanto concerne i dati relativi agli studenti ed ai corsi di studio viene specificato il livello di riferimento: singolo corso di studio, facoltà, ateneo; viene, infine, dedicata particolare attenzione alla raccolta delle opinioni degli studenti frequentanti.

Ovviamente le informazioni richieste dal Comitato sono finalizzate alla costruzione di una batteria di indicatori tali da consentire una valutazione comparativa, in termini di efficienza ed efficacia interna tra atenei; informazioni il cui obiettivo prioritario è la ripartizione tra gli atenei stessi delle risorse disponibili. Non bisogna dimenticare, infatti, che la nuova normativa riconosce al MIUR il compito di definire gli obiettivi principali e le strategie generali di sviluppo del sistema universitario, e di procedere alla sua valutazione, mentre agli atenei viene riconosciuta un'ampia autonomia, anche se parte dei finanziamenti accordati sono vincolati al soddisfacimento di specifici requisiti. Decentralizzazione, autonomia e finanziamenti vincolati implicano che gli organi di secondo livello (gli atenei), che sono i responsabili dei risultati ottenuti dalle unità operative loro afferenti, devono necessariamente svolgere un'intensa ed approfondita attività di valutazione ed autovalutazione in termini di misura dell'efficienza che dell'efficacia cercando, nei limiti delle proprie possibilità, di soddisfare le esigenze conoscitive dei livelli inferiori che operano nel proprio ambito (facoltà, corsi di studio, singoli docenti, studenti) ma anche le esigenze di chi del servizio formativo, o del prodotto del servizio stesso, fruisce o intende fruire (studenti, famiglie e mondo del lavoro).

Sempre in merito a questo tema, deve essere segnalato il recente testo predisposto dalla CRUI nell'ambito del progetto Campus*One* «Guida alla valutazione dei Corsi di Studio (CdS)». L'obiettivo di questo documento riguarda la possibilità di fornire alle unità che offrono il servizio, degli elementi e indicazioni necessari per svolgere un'attività sia di valutazione di efficacia interna, quindi di confronto tra gli obiettivi dichiarati (in seno ai propri programmi) e quelli conseguiti; sia anche di valutazione di efficacia esterna, cioè di analisi tra obiettivi assegnati e domanda sociale (in particolare, esigenze delle famiglie e del

Tabella 2

Dimensione	**Elementi**
Sistema organizzativo	Sistema di gestione Responsabilità Riesame
Esigenze e Obiettivi	Esigenze delle parti interessate Obiettivi generali e politiche Obiettivi di apprendimento
Risorse	Risorse umane Infrastrutture
Processo formativo	Progettazione Erogazione e Apprendimento Servizi di contesto
Risultati, Analisi e Miglioramento	Risultati Analisi e Miglioramento

Fonte: Campus*One*, *Guida alla Valutazione dei Corsi di Studio*, pag. 22.

mondo del lavoro), in seguito alle quali si individuano gli strumenti necessari per modificare i programmi e le prestazioni per una migliore qualità del servizio offerto.

In particolare, il documento propone un modello di valutazione (detto «modello Campus*One* per la valutazione dei CdS») in cui si prevedono le informazioni e i dati che devono essere raccolti, le modalità di raccolta, di elaborazione e di presentazione dei risultati, nonché due tipologie di valutazione: l'autovalutazione e la valutazione esterna. Tale modello si articola in cinque diverse dimensioni, come riportato nella Tabella 2; ciascuna a sua volta è costituita da un insieme di elementi atti a descrivere tutti gli aspetti sia positivi che negativi dell'attività di formazione di un CdS.

Gli strumenti, come appena detto, sono rappresentati dai dati (grezzi) e indicatori e dalle informazioni acquisite tramite quesiti formulati in corrispondenza di ogni elemento.

È auspicabile, come è nell'intenzione della CRUI, che la proposta di questo modello, cui devono attenersi i CdS che partecipano al progetto Campus*One*, possa essere estesa a tutti i CdS del sistema universitario italiano. Ovviamente, all'estensione si dovrà procedere dopo aver apportato al modello tutte le modifiche ed integrazioni che verranno suggerite dalle analisi dei dati raccolti relativamente al triennio di sperimentazione del modello stesso.

Una stessa metodologia di valutazione renderebbe possibili analisi comparative significative tra CdS diversi di uno stesso ateneo e tra lo stesso CdS di atenei diversi.

7. **Considerazioni conclusive**

A conclusione di questa nota, può risultare di qualche utilità un breve commento sul significato e sugli esiti dell'attività di valutazione che ha interessato, negli ultimi anni, il sistema universitario italiano per quanto attiene agli aspetti che sono stati, e sono tutt'ora, presi in maggiore considerazione: la misura dell'efficacia interna ed esterna del servizio formativo offerto.

Per ciò che concerne la misura dell'efficacia interna, tra le fonti informative più significative e rilevanti, anche perché richiesta dalla normativa vigente, si colloca la valutazione della didattica da parte degli studenti frequentanti.

Attraverso la raccolta delle opinioni degli studenti frequentanti, l'obiettivo primario che s'intende perseguire è quello della individuazione dei fattori che facilitano o che ostacolano l'apprendimento da parte degli studenti stessi sia in termini di svolgimento dell'attività didattica sia riguardo alle condizioni logistiche in cui la stessa si realizza che ai servizi di supporto. Ovviamente, i valori assunti dagli indicatori provenienti dall'elaborazione dell'opinione degli studenti sono soltanto uno degli elementi, di rilevanza certamente non marginale, su cui basare la valutazione del processo formativo ed è fondamentale che questi indicatori non vengano utilizzati per meccanismi automatici di premio/sanzione, ma che invece passino, insieme alle altre informazioni, attraverso il filtro di un giudizio competente e coerente con una corretta politica di assicurazione della qualità. In proposito si deve sottolineare che dimensioni della didattica quali aule, dotazioni e attrezzature, carico di lavoro, fanno parte delle esperienze dirette sulle quali lo studente ha titolo per rispondere e che è utile raccogliere notizie su tali fattori sia ai fini della «assicurazione della qualità» (per una funzione di auto-controllo sugli effetti delle scelte operate, in materia di didattica, da parte del CdS) sia ai fini di «audit» da parte degli organismi di livello più elevato (Emerson, Mosteler e Youtz, 2000).

Fino ad un recente passato, però, la scarsa disponibilità di strumenti per la verifica dell'efficacia formativa di ciascuna università e per il confronto dei risultati ottenuti in corsi identici presso sedi diverse, la scarsa attenzione e l'insufficiente approfondimento dei principali aspetti strutturali dell'università nel suo complesso, hanno favorito l'accumulo di risultati negativi per l'intero sistema formativo universitario e determinato valutazioni distorte sul sistema di istruzione superiore italiano, anche a livello di comunità internazionale.

Gli approfondimenti conoscitivi su alcuni fenomeni quali l'abbandono degli studi universitari (Bulgarelli, 2002), i tempi di conseguimento del titolo e l'attività di formazione post laurea (Bertaccini, 2000; Bulgarelli, 2002; Chiandotto, 2001; Chiandotto e Bertaccini, 2003) effettuati negli ultimi anni hanno consentito non solo una soddisfacente quantificazione della loro dimensione, ma hanno anche indot-

to riflessioni di portata più generale il cui esito si è risolto in una migliore comprensione dei punti di maggiore criticità presenti nell'organizzazione complessiva di servizi formativi offerti. Infatti, è ragionevole presumere, ad esempio, che l'elevato tasso di abbandono (circa il 30 per cento degli immatricolati non si iscrive al II anno) degli studi universitari e l'eccessivo prolungamento della durata degli studi emersi dalle indagini (in alcuni casi più che doppia rispetto alla durata legale), oltre che misura dell'efficacia interna, sia anche misura dell'efficienza in quanto dipendente dall'uso non ottimale delle risorse: dei docenti, delle strutture didattiche e degli studenti stessi le cui capacità non vengono certamente valorizzate da corsi il cui contenuto risulti eccessivamente pesante o di livello non adeguato (a causa di carenze pregresse dovute alla formazione pre-universitaria o al mancato coordinamento dei livelli e dei contenuti dei diversi corsi universitari). Si tratta di un'allocazione non ottimale di risorse o, comunque, di un'allocazione non in linea con processi formativi in grado di soddisfare l'esigenza, comune alla maggior parte dei laureati, di un rapido inserimento nel mercato del lavoro. Inoltre, riguardo alla formazione post-laurea si deve osservare che, se per un verso il desiderio di raggiungere ulteriori livelli di qualificazione deve essere considerato positivamente, per altro verso, il fenomeno potrebbe sottintendere l'incapacità di una parte dei percorsi didattici offerti dal sistema universitario italiano nel fornire ai laureati un bagaglio di conoscenze tale da poter essere immediatamente speso nel mondo del lavoro; e ciò anche a prescindere dal fatto che l'ulteriore formazione non fa che procrastinare ulteriormente l'ingresso nel mondo del lavoro.

Per quanto attiene agli sbocchi occupazionali dei laureati è ormai diffusa la consapevolezza di quanto l'ingresso dei giovani nel mondo del lavoro rappresenti un momento particolarmente critico rispetto al quale i giovani stessi e l'intera società sono chiamati a confrontarsi quotidianamente; la transizione dalla scuola al lavoro per i possessori di una preparazione a livello universitario, seppure meno problematica e più rapida di coloro che non possiedono tale preparazione, presenta delle connotazioni negative che, se analizzate in modo adeguato, potrebbero essere, se non completamente eliminate, quantomeno, sostanzialmente ridotte.

Dall'esame dei dati sui livelli occupazionali (ALMALAUREA, Bertaccini, 2000; Chiandotto, 2001; Chiandotto e Bertaccini, 2003) si può ragionevolmente concludere che la disoccupazione giovanile non è fenomeno diffuso tra coloro che possiedono un titolo di studio universitario anche se, in proposito, si deve sottolineare che una quota non irrilevante di intervistati prosegue un'attività iniziata prima del conseguimento del titolo.

Gli elevati tassi di abbandono, l'eccessiva durata degli studi e lo scarso potere professionalizzante dei titoli conseguiti sono stati forse gli elementi che hanno inciso in modo più significativo sulla struttura-

zione dei nuovi cicli di studio: un primo triennio, nel quale devono essere previsti anche insegnamenti professionalizzanti, seguito da un biennio di specializzazione.

Ci si può domandare se tali problemi hanno trovato la giusta soluzione con la riforma dei cicli e degli ordinamenti didattici dell'università. In proposito quale dubbio più che ragionevole può essere avanzato. Infatti, riguardo al problema degli abbandoni e a quello della durata degli studi, anche se è prematuro esprimere una valutazione compiuta, i primi segnali non appaiono davvero confortanti: i tassi di abbandono al primo anno nella generalità dei casi non si sono ridotti, i crediti acquisiti dagli studenti nel corso del primo anno di studi (forse a causa dell'eccessiva frammentazione degli insegnamenti che ha comportato un notevole incremento nel numero delle prove d'esame) si collocano mediamente sulla quota del 50 per cento di quelli previsti, fatto che se si ripete anche nei due anni successivi comporta necessariamente un raddoppio della durata degli studi.

Per quanto concerne, infine, la «professionalizzazione» prevista nel primo triennio, non si può non tenere presente che: *a)* in alcuni percorsi di studio può essere fornita soltanto una base professionalizzante e non competenze professionali direttamente spendibili sul mercato del lavoro; *b)* in altri precorsi, quelli che per loro natura non consentono l'acquisizione ne di professionalità ne di base professionale a basso livello, la «voglia» di professionalizzare a tutti i costi può indurre allo snaturamento dei percorsi stessi impedendo, di fatto, un'articolazione adeguata e direttamente finalizzata al perseguimento di professionalità di livello elevato. Relativamente a questo aspetto, molto più significativo appare, invece, l'inserimento nei curricula delle attività di tirocinio presso aziende o enti pubblici e/o privati; infatti, queste attività risultano praticamente obbligatorie in molti percorsi di studio e, oltre a consentire una effettiva acquisizione di professionalità, rappresentano anche un canale privilegiato per l'assunzione. In proposito, però, non si può non prendere atto che per molti corsi di studio, anche in dipendenza del contesto ambientale nel quale sono collocati i vari atenei, le attività di tirocinio non potranno che essere svolte in misura limitata a causa della mancanza di un numero sufficiente di interlocutori del mondo del lavoro.

Gli autori di questa nota, se per un verso condividono interamente i principi che hanno ispirato la riforma, vale a dire la necessità di migliorare l'efficienza e l'efficacia di un organismo che assorbe risorse pubbliche, in un periodo di crescente contenimento della spesa dello Stato, e il desiderio di elevare gli standard non solo della formazione ma anche della ricerca in un contesto di crescente competizione internazionale, per altro verso, sono fermamente convinti che il peso crescente, rispetto all'amministrazione centrale dello Stato, dell'autonomia degli atenei, deve essere bilanciato ed equilibrato attraverso la verifica continua (*valutazione e monitoraggio*) delle attività svolte e dei

risultati conseguiti. Attività di valutazione e monitoraggio del tutto sterile se non è accompagnata da fattivi interventi (*misurare per migliorare*) da parte degli organi preposti al governo e alla gestione dei processi formativi: il MIUR, il Senato Accademico, il Consiglio di Amministrazione, i Consigli di Facoltà e i Consigli di corso di studi, le Commissione per la didattica, i singoli docenti. Soltanto percorrendo questa via è possibile, se non risolvere, quanto meno avviare a soluzione i problemi che hanno afflitto, e ancora oggi affliggono, il sistema universitario italiano.

BIBLIOGRAFIA

Azzone G. (a cura di) (2003), *Analisi dell'efficacia e dell'efficienza delle attività amministrative delle università: il progetto good practice II*. www.cnvsu.it/indagini/programmi_ricerca/view.asp?ID_PDR=8.

Biggeri L. (2000), *Valutazione: idee, esperienze, problemi. Una sfida per gli Statistici* in *Atti della XL Riunione Scientifica della Società italiana di Statistica*, Firenze, 26-28 aprile 2000.

Biggeri L. (1999), *Autonomia e valutazione dell'insegnamento nel sistema universitario italiano, Giornata di studio su «L'insegnamento universitario in Italia». Accademia dei Lincei*, Roma, 21 gennaio 1999.

Biggeri L. (1998), *Programmazione e valutazione dello sviluppo del sistema universitario*, reprint dell'articolo pubblicato su: *La programmazione del sistema universitario*, Università Ricerca n. 2, 1998, Reprint 1/98.

Bini M. (1999), *Valutazione della Efficacia dell'Istruzione Universitaria rispetto al Mercato del Lavoro*. RdR 03/99, Osservatorio per la Valutazione del Sistema Universitario - MURST – Roma, consultabile anche sul sito www.cnvsu.it/publidoc/comitato/default.asp.

Bertaccini B. (a cura di) (2000), *I laureati dell'Ateneo Fiorentino dell'anno 1998 – Profilo e sbocchi occupazionali*, Università degli Studi di Firenze, consultabile anche sul sito www.unifi.it/aut_dida/indexval.html.

Breno E., *et al.* (2002), *La ricerca scientifica nelle università italiane. Una prima analisi delle citazioni della banca dati ISI*, Pubblicazioni CRUI, Roma, www.crui.it/data/allegati/links/902/ISI_imp.pdf.

Bulgarelli G. (2002), *Esito degli studi degli immatricolati dell'Ateneo Fiorentino dal 1980/81 al 1997/98,* Università degli Studi di Firenze, consultabile anche sul sito www.unifi.it/aut_dida/indexval.html.

Bulgarelli G. (a cura di) (2001), *I laureati dell'Ateneo Fiorentino dell'anno 1997 – Profilo e sbocchi occupazionali,* Università degli Studi di Firenze, www.unifi.it/aut_dida/indexval.html.

Dilorenzo P. - Stefani E. (2003), *La riforma universitaria. Una indagine sui docenti: dall'estraneità al coinvolgimento*, Pubblicazioni CRUI, Roma, consultabile anche sul sito www.fondazionecrui.it/?Arg=208.

Catalano G. (a cura di) (2002), *La valutazione delle attività amministrative delle università: il Progetto «Good practices»*, Il Mulino.

Cave M., *et al*. (1997), *The Use of Performance Indicators in Higher Education, The Challenge of the Quality Movement*, ed. Jessica Kingsley.

Chiandotto B. (2002), *Valutazione dei processi formativi: cosa, come e perché, in Valutazione della Didattica e dei Servizi nel Sistema Università. Atti della giornata di Studio*, Fisciano, 31 maggio 2002. CUSL, Salerno 2002.

Chiandotto B. (2001), *Profilo e condizione occupazionale dei laureati dell'Ateneo Fiorentino ad uno, due e tre anni dal conseguimento del titolo*, Università degli Studi di Firenze, www.unifi.it/aut_dida/indexval.html.

Chiandotto B. - Bertaccini B. (2003), *I laureati dell'Ateneo Fiorentino dell'anno 1999 – Profilo e sbocchi occupazionali,* Università degli Studi di Firenze, consultabile anche sul sito www.unifi.it/aut_dida/indexval.html.

Comitato Nazionale per la Valutazione del Sistema Universitario (2001), *Rapporto finale del Gruppo di lavoro «Accreditamento dei Corsi di*

Studio». RdR 01/01, MIUR - CNVSU, Roma, consultabile www.cnvsu.it/publidoc/comitato/default.asp.
Comitato Nazionale per la Valutazione del Sistema Universitario (2003), *Note tecniche sui dati e informazioni da trasmettere entro il 30 Aprile 2003*, MURST – Roma, www.cnvsu.it/publidoc/comitato/default.asp.
Comitato Nazionale per la Valutazione del Sistema Universitario (2002), *Note tecniche sui dati e informazioni da trasmettere entro il 30 Aprile 2002*, MURST – Roma, www.cnvsu.it/dati/nuclei/default.asp.
Comitato Nazionale per la Valutazione del Sistema Universitario (2001), *Note tecniche sui dati e informazioni da trasmettere entro il 30 Aprile 2001*, MURST – Roma, www.cnvsu.it/dati/nuclei/default.asp.
Comitato Nazionale per la Valutazione del Sistema Universitario (2000), *Note tecniche sui dati e informazioni da trasmettere entro il 30 Aprile 2000*, MURST – Roma, www.cnvsu.it/publidoc/comitato/default.asp.
Emerson J.D. - Mosteller F. - Youtz C. (2000), *Students Can Help Improve College Teaching: A Review and an Agenda for the Statistics Profession*. Statitics for thr 21st Century – Methodologies for Applications of the Future. Rao C.R. and Székely G.J. editors. Marcel Dekker, Inc. New York.
Ewell P.T. (1999), *Linking Performance measures to resource allocation: exploring unmapped terrain*, *Quality in Higher Education*, vol.5, 3.
Feldman S. (1999), *Only connect - Professors and teachers with a common mission*. *Academe-Bulletin of the AAUP,* vol. 85.
Fondazione CRUI (2003), *Guida alla Valutazione dei Corsi di Studio*, Pubblicazioni CRUI, Roma, consultabile anche sul sito www.crui.it//data/allegati/links/902/GuiValimp.pdf.
Gola M. (1998), *La didattica universitaria e la sua valutazione,* Comitato paritetico per la didattica, Corso strumenti e metodologie per il formatore, Politecnico di Torino.
Goldstein H. (1995), *Multilevel Statistical Models*, London, Edward Arnold.
Gori E. (2000), *La valutazione del sistema di istruzione: il caso dell'università*, www.unisi/didatticavalut/atti/8-appendice/crisp/crisp.htm.
Gori E. - Vittadini G. (a cura di) (1999), *Qualità e valutazione nei servizi di pubblica utilità*, Milano, Etas Libri.
Hatry H. P. (1986), *Efficiency Measurement for local Governement services*. The Urban Institut, Washington DC.
Heck H. H. - Thomas S. L. (2000), *An Introduction to Multilevel Modeling Techniques*, Lawrence Erlbaum Associates, London, Publishers.
ISTAT (2003), *Inserimento professionale dei laureati. Indagine 2001*, *Informazioni n.28*, Roma, ISTAT.
ISTAT (1996), *Inserimento professionale dei laureati. Indagine 1995, Informazioni n.10*, Roma, ISTAT.
ISTAT (1994), *Indagine longitudinale sugli sbocchi professionali dei laureati. Anni 1989-1991*, *Informazioni n. 25*, Roma, ISTAT.
ISTAT (1994a), *Indagine 1991 sugli sbocchi professionali dei laureati. Informazioni n.1*, Roma, ISTAT.
ISTAT (1990), *Indagine 1989 sugli sbocchi professionali dei laureati*, *Informazioni n.17*, Roma, ISTAT.
Minelli E. - Rebora G. - Turri M. (2002), *Il valore dell'università. La valutazione della didattica, della ricerca, dei servizi negli atenei,* Milano, Guerini e Associati.

Modica L. - Stefani E. (a cura di) (1997), *Valutazione delle attività didattiche. Le esperienze condotte dalla CRUI*, Documenti, vol. 5, Roma, CRUI, dicembre 1997, consultabile anche sul sito www.quipo.it/netpaper/cruival.doc.
Monti A. - Briganti A. (2002), *Rapporto sull'istruzione universitaria in Italia – Costi e rischi della riforma* (a cura di), Milano, Franco Angeli. Il volume raccoglie i contributi presentati nella Giornata di studio su *La riforma del sistema universitario nel quadro dell'integrazione europea*, tenutasi a Roma il 26 settembre 2001.
Osservatorio per la Valutazione del Sistema Universitario (1998), *Informazioni per la preparazione delle relazioni dei Nuclei di valutazione interna e un insieme minimo di indicatori*. DOC 11/98, MURST - OVSU, Roma, www.cnvsu.it/publidoc/comitato/default.asp.
OCSE (2001), *Education at a Glance 2001. Organization for Economic Co-operation and Development Indicators*, OECD.
Palumbo M. (1995), *Indicatori e valutazione di efficacia delle policies* in *Sociologia e ricerca sociale*, n. 47-48.
Rabe-Hesketh S. - Skrondal A. - Pickles A. (2004), *Generalized multilevel structural equation modelling* in *Psychometrika*, in press.
Ramsden P. (1988), *Improving Learning: New Perspectives* in London, Kogan Page.
Rebora G. (1999), *La valutazione dei risultati nelle amministrazioni pubbliche*, Milano, Guerini e Associati.
Rowley J. (1996), *Measuring Quality in Higher Education* in *Quality in Higher Education*, vol. 2.
Schmitz C.C. (1993), *Assessing the validity of higher education indicators* in *Journal of Higher Education*, vol.64, 5.
Smith P. (1995), *On the unintended consequences of publishing performance indicators in the public sector* in *International Journal of Public Administration*, 18.
Université de Toulon (1998), *Qualité totale et enseignement superieur* in Atti del Convegno tenutosi a Toulon, 3-4 settembre 1998, organizzato in collaborazione con l'Università di Verona.
Yorke M. (1998), *Performance Indicators Relating to Student Development: can they be trusted?*, *Quality in Higher Education*, vol. 4, n. 1.
Viale R. - Cerroni A. (a cura di) (2003), *Valutare la scienza*, Rubettino Ed.

Valutazione dei processi formativi: cosa, come e perché

Da: Chiandotto B. (2003) Valutazione dei processi formativi: cosa, come e perché, in Valutazione della Didattica e dei Servizi nel Sistema Università, a cura di D'Esposito M.R. , pp. 35-86, CUSL, Salerno.

Valutazione dei processi formativi: cosa, come e perché

Bruno Chiandotto
Dipartimento di Statistica "G. Parenti"
Viale Morgagni, 59 Firenze
chiandot@ds.unifi.it

Premessa

I problemi (abbandoni, tempi di conseguimento del titolo, frattura tra preparazione universitaria e mondo del lavoro, ecc.) in cui si dibatte il Sistema Universitario nel nostro Paese, sono numerosi e di difficile soluzione.

La discussione e gli approfondimenti conoscitivi che sul tema si sono susseguiti negli ultimi anni sembrano avviati nella giusta direzione, ed in questo contesto appare particolarmente efficace l'attività che hanno svolto e stanno svolgendo l'Osservatorio per la Valutazione del Sistema Universitario (OVSU), il **Comitato Nazionale per la Valutazione del Sistema Universitario** (**CNVSU**) - subentrato all'Osservatorio - e la **Conferenza dei Rettori delle Università Italiane** (**CRUI**).

Prima di entrare nel merito dell'argomento oggetto della nota risulta utile richiamare l'attenzione su un fatto talmente ovvio che, proprio per la sua ovvietà, viene spesso trascurato: "**per risolvere i problemi bisogna conoscerli**", e la conoscenza non può che derivare da un'attenta ed approfondita valutazione del materiale informativo disponibile.

L'esigenza di poter disporre di una solida base informativa sui settori di possibile intervento deve non solo permettere l'acquisizione di una conoscenza approfondita sulle varie specificità ma deve, al tempo stesso, consentire l'impostazione di politiche di intervento coerenti (**decisioni razionali**) con i fabbisogni reali delle collettività interessate. Non è infrequente imbattersi in situazioni nelle quali si giustificano decisioni affrettate e non sufficientemente ponderate argomentando sulla carenza nella qualità e/o nella quantità dell'informazione disponibile, ma come è possibile parlare di carenza della base informativa se la stessa non viene confrontata con le esigenze conoscitive che l'informazione stessa deve soddisfare, ed è proprio questa necessità di esame comparativo quella che muove la critica, che consente l'evidenziazione delle carenze, che impone un riesame continuo di una qualunque base informativa e che suggerisce la revisione delle procedure di raccolta ed elaborazione dei dati alla luce dei nuovi fatti emergenti.

Spesso, l'operatore pubblico nell'attivare i propri processi decisionali prescinde completamente dalle informazioni anche se qualche volta si basa su dei dati statistici; ma il dato statistico non è necessariamente informazione: il **sistema informativo**, che è l'unico supporto adeguato di ogni efficace processo decisionale, non si identifica nel **sistema statistico**.

Un sistema informativo comprende, infatti, l'insieme di tutte le attività che vanno dalla raccolta, alla elaborazione alla trasmissione delle informazioni, sia statistiche che non statistiche, sia formali che informali, finalizzate al perseguimento di specifici obiettivi di governo, di gestione e/o di controllo. Ed è soltanto in tale contesto che il dato (statistico) riesce ad evidenziare tutte le sue potenzialità, ed è soltanto in tale prospettiva che si può parlare di "**qualità**" del dato e di "**valore**" dell'**informazione**.

Per poter consentire lo sfruttamento ottimale del Sistema Informativo è necessario che il Sistema stesso risponda ai requisiti sotto elencati:

a) chiara definizione delle necessità degli utilizzatori e delle loro richieste;
b) possibilità di un accesso appropriato e continuo ai dati e ai servizi offerti;

c) inventario di tutti i dati disponibili e delle informazioni pertinenti l'ambiente circostante;
d) necessità di costruire insiemi di dati in modo sistematico ed integrato;
e) accesso appropriato a telecomunicazioni e connessioni internet che assicurino il raggiungimento in remoto di centri di dati e di sistemi di informazioni;

Infine, occorre sottolineare l'importanza di costruire un sistema di indicatori tali da consentire un'effettiva valutazione sia delle attività che producono effetti nel medio-lungo periodo, sia delle attività correnti in termini di processo ed in termini di risultati; da qui la necessità di fornire risposte esaurienti ai quesiti:

1. **cosa valutare?**
2. **come valutare ?**
3. **perché valutare ?**

Il tema della valutazione interna nelle Università è stato introdotto, dal punto di vista normativo, dalle leggi 168/89 e 537/93. La prima istituisce il Ministero dell'Università e della Ricerca Scientifica e Tecnologica (MURST) e prevede l'attuazione di forme di controllo interno sull'efficienza e sui risultati della gestione nelle Università; la seconda prevede l'istituzione nelle Università dei Nuclei di valutazione interna.

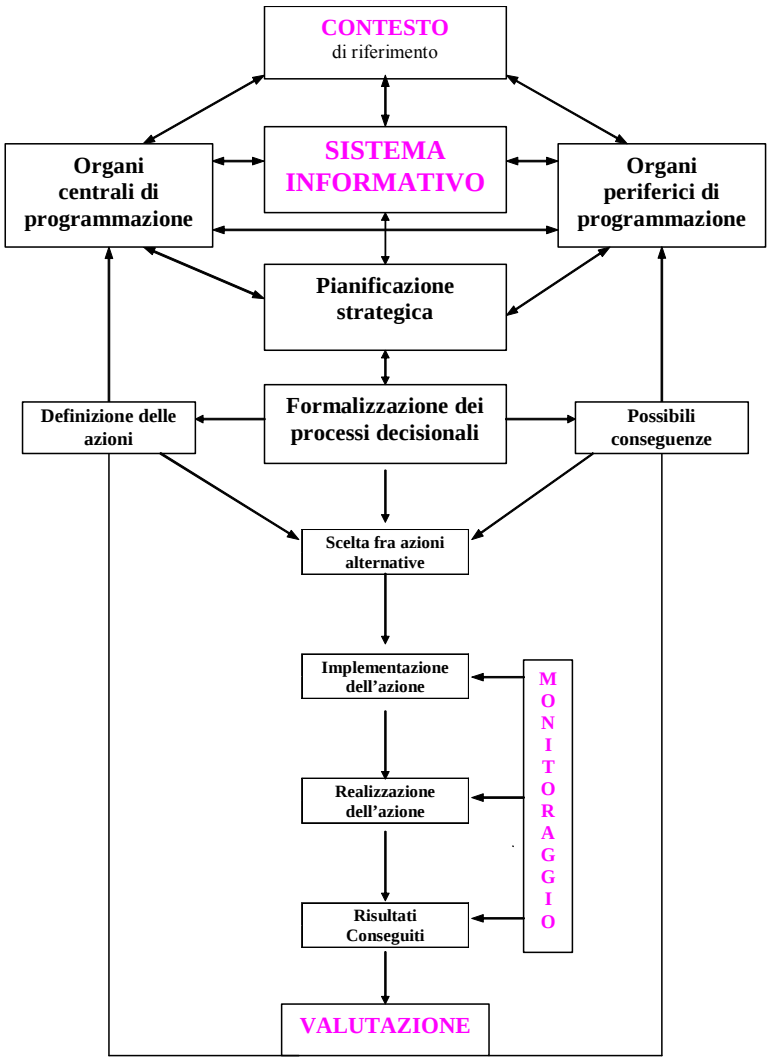
Il Nucleo (L. 537/93, articolo 5, comma 22) "*ha il compito di verificare, mediante analisi comparative dei costi e dei rendimenti, la corretta gestione delle risorse pubbliche, la produttività della ricerca e della didattica, nonché l'imparzialità ed il buon andamento dell'azione amministrativa. I Nuclei determinano i parametri di riferimento del controllo anche su indicazione degli organi generali di direzione, cui riferiscono con apposita relazione almeno annualmente*". La relazione deve essere trasmessa (L. 537/93, articolo 5, comma 22) al MURST, al Consiglio Universitario Nazionale e alla Conferenza permanente dei Rettori, per garantire un doppio controllo, sulla valutazione dei risultati (relativi all'efficienza e alla produttività delle attività di ricerca e di formazione) e sulla verifica dei programmi di sviluppo e di riequilibrio del sistema universitario, anche al fine di provvedere ad una razionale assegnazione delle risorse.

Una nuova legge (L. 19 ottobre, n. 370 – G.U. n. 252 del 26.10. 1999) dispone norme volte a disciplinare, più compiutamente, la valutazione del sistema universitario: "*Le università adottano un sistema di valutazione interna della gestione amministrativa, delle attività didattiche e di ricerca, degli interventi di sostegno al diritto allo studio …".* (art. 1, comma 1). "*Le funzioni di valutazione di cui al comma 1 sono svolte in ciascuna università da un organo collegiale disciplinato dallo statuto dell'università, denominato "nucleo di valutazione di ateneo*",……. "*Le università assicurano ai nuclei l'autonomia operativa, il diritto di accesso ai dati e alle informazioni necessari, nonché la pubblicità e la diffusione degli atti, nel rispetto della normativa a tutela della riservatezza. I nuclei acquisiscono periodicamente, mantenendo l'anonimato, le opinioni degli studenti frequentanti sulle attività didattiche e trasmettono un'apposita relazione, entro il 30 aprile di ciascun anno, al Ministero dell'università e della ricerca scientifica e tecnologica, e al Comitato per la valutazione del sistema universitario.*…"(art. 1, comma 2). "*Le Università che non applicano le disposizioni di cui ai commi 1 e 2 entro sei mesi dalla data di entrata in vigore della presente legge sono escluse per un triennio dal riparto dei fondi relativi alla programmazione universitaria, nonché delle quote di riequilibrio …..…...*" (articolo 1, comma 3).

La stessa legge, all'articolo 2, prevede l'istituzione del Comitato Nazionale per la Valutazione del Sistema Universitario. "……*Il Comitato: a) fissa i criteri generali per la valutazione delle attività delle Università previa consultazione della Conferenza dei Rettori delle Università Italiane (CRUI) del Consiglio Universitario Nazionale (CUN) e del Consiglio Nazionale degli Studenti Universitari (CNSU), ove costituito; b) promuove la sperimentazione, l'applicazione e la diffusione di metodologie e pratiche di valutazione; c) determina ogni triennio la natura delle informazioni e i dati che i nuclei di valutazione degli atenei sono tenuti a comunicare annualmente; d) predispone ed attua, sulla base delle relazioni dei nuclei di valutazione degli atenei e delle altre informazioni*

acquisite, un programma annuale di valutazioni esterne delle Università o di singole strutture didattiche, approvato dal MURST (ora MIUR), con particolare riferimento alla qualità delle attività universitarie sulla base di standard riconosciuti a livello internazionale nonché della raccomandazione 98/561/CE del Consiglio, del 24 settembre 1998, sulla cooperazione in materia di garanzia della qualità nell'istruzione superiore; e) predispone annualmente una relazione sulle attività di valutazione svolte; f) svolge i compiti assegnati dalla normativa vigente, all'Osservatorio per la valutazione del sistema universitario ….. .; g) svolge, su richiesta del MURST (MIUR), ulteriori attività consultive, istruttorie, di valutazione, di definizione di standard, di parametri e di normativa tecnica, anche in relazione alle distinte attività delle università, nonché ai progetti e alle proposte presentate dalle medesime.".

Da quanto sopra richiamato emerge in modo evidente la necessità di definizione e strutturazione di un sistema informativo adeguato alle necessità e nel quale sia presente una chiara articolazione dei diversi livelli e momenti decisionali, delle connessioni tra tali livelli e momenti e, soprattutto, la rilevanza - come punto centrale di riferimento - dell'attività di analisi e formalizzazione dei processi decisionali stessi. Al riguardo si può proporre una semplice schematizzazione come quella sotto riportata.



Ovviamente, lo schema proposto va'interpretato in funzione del livello decisionale in cui ci si colloca. Se, ad esempio, si fa riferimento al sistema universitario nel suo complesso e si individua

come organo centrale di programmazione il Ministero, gli organi periferici sono gli atenei; se si individua, invece, l'Ateneo come organo centrale di programmazione, gli organi periferici sono le facoltà, i corsi di studio, i dipartimenti e i singoli docenti.

Stabilito il livello decisionale d'interesse si potrà procedere alla specificazione dettagliata delle esigenze conoscitive da soddisfare per rendere possibili politiche d'intervento finalizzate al perseguimento di livelli elevati di qualità, di efficienza ed efficacia dell'attività didattica, dell'attività di ricerca e della gestione che si svolge presso le diverse sedi universitarie.

1. Valutazione e decisioni razionali

Qualunque problema decisionale da risolvere, dal più banale al più complesso, richiede la chiara definizione del problema stesso e l'individuazione delle relazioni che connettono i vari elementi che lo caratterizzano.

Il quadro logico di riferimento e le informazioni sono gli ingredienti essenziali di ogni processo decisionale, la teoria delle decisioni, la teoria statistica (Savage, 1972) ed i metodi e i modelli sviluppati in questi ambiti disciplinari sono gli strumenti necessari per lo svolgimento ottimale di ogni processo decisionale, decisioni che devono essere, nella generalità dei casi, prese in situazioni di conoscenza parziale della realtà in cui si opera. Conoscenza parziale perché risulta impossibile o non conveniente acquisire tutte le informazioni relative agli aspetti che interessano pur essendo, almeno teoricamente, possibile una loro acquisizione totale; altro caso è quello della conoscenza parziale perché le informazioni non sono neanche potenzialmente disponibili.

La teoria delle decisioni fissa principi razionali di comportamento che consentono la derivazione di regole di scelta ottimale (Allais, 1953; Von Neumann e Morgenstern, 1953; Fishburn, 1982, 1988°, 1988b e 1989; French, 1986). Gli sviluppi più recenti di tale teoria consentono anche di valutare e correggere eventuali incoerenze e contraddizioni nel comportamento dei decisori (Tversky e Kahneman, 1986 e 1992; Quiggin, 1993; Camerer e Ho, 1994).

Pertanto, oggetto di studio della teoria delle decisioni è il processo decisionale. Attraverso l'analisi del comportamento degli attori (individui o gruppi) coinvolti nel processo si procede, cioè, all'esame di come i decisori prendono o dovrebbero prendere delle decisioni.

Le applicazioni della teoria spaziano dalle speculazioni astratte, relative ad agenti idealmente razionali, ai suggerimenti pratici per la risoluzione di specifici problemi decisionali. I teorici della decisione indagano sulle conseguenze logiche di differenti regole decisionali o esplorano gli aspetti logico-matematici di diverse descrizioni di comportamento razionale (Allais, 1953; Ellsberg, 1961; Fishburn, 1988b); gli applicati sono invece interessati all'esame dei processi decisionali così come gli stessi si svolgono nella realtà.

In questa ottica si è soliti distinguere la teoria delle decisioni in due filoni principali: **teoria normativa** e **teoria descrittiva**. Chi si occupa di teoria descrittiva cerca di scoprire come le decisioni vengono prese nei diversi contesti operativi; chi si occupa di teoria normativa analizza il modo con cui le decisioni dovrebbero essere prese facendo riferimento ad agenti idealmente razionali. Questa distinzione è utile ma alquanto artificiale, essendo l'informazione sul modo effettivo di prendere decisioni, certamente rilevante ai fini della fissazione di regole su come le decisioni devono essere prese; d'altro lato nessuno studio sul comportamento effettivo di agenti può consentire il conseguimento di risultati soddisfacenti se lo stesso non viene, in qualche modo, posto a confronto con una sorta di comportamento ideale.

La teoria descrittiva delle decisioni non interessa in questa sede essendo oggetto di discipline specifiche quali la psicologia, la sociologia e, per alcuni aspetti, l'economia; qui ci si occuperà di come le decisioni dovrebbero essere prese per massimizzare il benessere e non di come le decisioni sono effettivamente prese. Ma, il riferimento alla teoria normativa non può essere assoluto, si deve, infatti, tenere conto di tutta una serie di vincoli e di condizionamenti che emergono dall'analisi dei

processi reali affinché le **regole di comportamento razionale** possono tradursi in comportamenti effettivi.

Per caratterizzare e distinguere questo specifico sviluppo della teoria normativa delle decisioni alcuni autori (Tversky e Kahneman, 1986 e 1992) hanno suggerito la dizione **teoria prescrittiva** che si caratterizza, appunto per il fatto che le regole ideali di comportamento razionale analizzate devono poter essere tradotte in comportamenti reali.

Un'altra importante distinzione operata all'interno della teoria delle decisioni è quella tra decisioni individuali e decisioni di gruppo (French, 1986). Da sottolineare che ai fini di questa distinzione una decisione individuale non deve necessariamente riferirsi ad un singolo individuo, anche le imprese, le associazioni, i partiti, le nazioni, le regioni, le università, ecc., quando mirano al conseguimento di un obiettivo comune della organizzazione, prendono decisioni individuali. Si parla, invece, di decisioni di gruppo quando gli individui che appartengono alla stessa organizzazione manifestano opinioni diverse rispetto ai fini o alle priorità del gruppo.

La parte più rilevante della ricerca relativa alla teoria delle decisioni di gruppo è stata rivolta allo sviluppo di strategie comuni per governare i vari componenti del gruppo e alla distribuzione delle risorse all'interno del gruppo stesso ed in questo ambito assumono, spesso, grande rilevanza aspetti etici e morali. All'opposto, nella teoria delle decisioni individuali ci si concentra sul problema di come gli individui possono favorire i propri interessi, qualunque sia la loro natura, non riconoscendo alcuna rilevanza ad aspetti etici e/o morali; potrebbe essere pertanto possibile per un agente idealmente razionale trovarsi in condizioni migliori violando la strategia comune del gruppo di appartenenza.

Qualunque decisione, sia essa individuale o di gruppo, comporta una scelta tra più alternative, o azioni, o atti, ciascuna delle quali produrrà una tra più conseguenze che dipenderà dalle condizioni del contesto, **stato di natura**, nel quale il processo decisionale si svolge. Le decisioni, sono, pertanto, costituite da **azioni**, **stati** e **conseguenze**, con le ultime che dipendono dall'azione e dallo stato in cui l'azione si verifica.

Quando si analizza un problema di decisione, l'analista, che può essere lo stesso soggetto che prende la decisione, deve individuare l'insieme rilevante delle azioni, degli stati e delle conseguenze per caratterizzare in modo adeguato il problema stesso. Attraverso l'individuazione di azioni, stati e conseguenze e costruendo, eventualmente, una **tavola** o un **albero di decisione**, si procede alla specificazione del problema decisionale.

Un ulteriore ed interessante aspetto connesso alla specificazione del problema decisionale è quello relativo alla distinzione tra **decisione giusta** e **decisione razionale**. La decisione di chi agisce è giusta se si risolve in esiti ottimali, se si disponesse di una conoscenza completa del futuro basterebbe fare riferimento al solo principio: **"prendi la decisione giusta"**, purtroppo, la maggior parte delle decisioni è basata sul ciò che si ritiene possa accadere e non su quello che accadrà realmente. Nella quasi totalità dei casi risulta quindi impossibile prendere una decisione giusta, si dovrà allora prendere una decisione razionale, valutando al meglio l'insieme parziale di informazioni a disposizione riguardo al vero stato del mondo, e non è affatto scontata l'equivalenza: **decisione razionale = decisione giusta**.

Se l'agente, il decisore, conoscesse le conseguenze di ciascuna azione (**decisioni in situazione di certezza**) il problema di scelta si ridurrebbe al confronto tra le conseguenze e la scelta razionale equivarrebbe alla scelta giusta, sempre che il decisore sia in grado di esprimere, in modo razionale, le sue preferenze riguardo alle conseguenze stesse. Il comportamento razionale consente, in altre parole, l'individuazione dell'alternativa ottimale che comporta il conseguimento del massimo beneficio.

Se le conseguenze in corrispondenza di ciascuna azione sono più di una e si dispone di una misura della probabilità sulle possibilità della loro realizzazione, si parla di **decisioni in situazioni di rischio o incertezza**.

Le innumerevoli teorie (normative e prescrittive) delle decisioni che sono state proposte e si sono successivamente sviluppate sono, nella generalità dei casi, riferite alla **teoria dell'utilità**

attesa (EU - Expected Utility Theory) proposta da **von Neumann** e **Morgenstern** (1953), la quale, pur rappresentando la teoria normativa per eccellenza (tale viene considerata dalla maggioranza degli studiosi), si è, tutto sommato, anch'essa sviluppate in un'ottica prescrittiva come risposta alle carenze riscontrate nella prima formulazione della teoria normativa delle decisioni che prevedeva la massimizzazione del **valore monetario atteso (EMV –Expected Monetary Value)**.

I problemi decisionali nei quali in corrispondenza di ciascuna azione sono possibili conseguenze diverse, e nei quali sono note le probabilità (oggettive o soggettive) ad esse associate, possono essere risolti in modo del tutto soddisfacente poiché si dimostra che (*teorema*) se un decisore agisce conformandosi ad un certo insieme di postulati di comportamento razionale allora esiste una funzione a valori reali definita sull'insieme delle conseguenze e se il decisore sceglie l'azione cui corrisponde il massimo dell'utilità attesa egli agisce in modo conforme al proprio schema di preferenze massimizzando il proprio beneficio (French, 1986).

Dal teorema ne consegue che il **criterio ottimale di scelta** in situazioni di rischio o incertezza è quello della **massimizzazione dell'utilità attesa**.

E' noto, e ne sono esempio i numerosi paradossi presenti in letteratura, come i comportamenti degli individui non siano spesso in accordo con i principi di razionalità sui quali si basa il modello (classico) dell'utilità attesa (Ellsberg, 1961). Come già sottolineato, questo aspetto ha indotto molti autori a considerare il modello di von Neumann e Morgenstern inadeguato come strumento operativo; in particolare, il divario che spesso si osserva fra il comportamento ideale ipotizzato in un modello normativo e il comportamento effettivo degli individui è stato il motivo principale di rivisitazioni e critiche, nonché la base per lo sviluppo di teorie delle decisioni che si discostano da quella classica. I modelli decisionali normativi, infatti, pur traendo origine da comportamenti reali, si discostano dagli stessi comportamenti proprio per la loro idealizzazione e astrazione dalle situazioni reali. Tuttavia, ciò non deve necessariamente indurre al rifiuto dei modelli normativi e all'accettazione di quelli descrittivi, il cui scopo è quello della identificazione della natura e struttura delle preferenze degli individui dai quali trarre modelli che permettano di configurare preferenze e decisioni non ancora manifestate

La semplice descrizione dei comportamenti individuali, infatti, risulta in alcuni contesti altrettanto insoddisfacente, in quanto, se posti di fronte alle proprie incoerenze, molte individui cercano di ovviare alle incoerenze stesse proprio attraverso una rivisitazione e sistemazione delle proprie scelte in accordo con quanto previsto dai metodi normativi. A questo proposito, alcuni autori hanno evidenziato il fatto che l'analisi delle decisioni dovrebbe indirizzarsi sempre più verso una risposta alla domanda: **è possibile per gli individui operare in modo tale da non contraddire il proprio schema di preferenze?** Dovrebbe, cioè, suggerire comportamenti ottimali, senza però fare troppa violenza sulle attitudini più profonde del decisore (Keller, 1992; Wakker e Deneffe, 1996; Herstein e Milnor, 1998). In quest'ottica si colloca l'approccio prescrittivo alla teoria delle decisioni: un'analisi prescrittiva dovrebbe sviluppare procedure volte ad eliminare o ridurre violazioni dei principi cardine delle scelte razionali (Tversky e Kahneman, 1986 e 1992).

I modelli prescrittivi sono dunque orientati ad avvicinare i comportamenti degli individui a schemi decisionali razionalmente coerenti; tali modelli contemplano solitamente assiomi più deboli rispetto a quelli classici o, addirittura, possono anche non trovare inizialmente una giustificazione su base assiomatica.

Tornando al tema della relazione tra informazioni e decisione, si deve sottolineare che la mancanza di razionalità nell'azione pubblica, mancanza intesa nel senso di assenza di strutturazione adeguata dei processi decisionali posti in essere, ha fatto si che in passato molta produzione d'informazione sia stata stimolata, non tanto da reali necessità degli utilizzatori, quanto dalle convinzioni, non necessariamente corrette, che i produttori di informazioni si sono create sulla domanda potenziale, in una funzione di supplenza, di chi era chiamato (il decisore) istituzionalmente a farlo.

2. Sistema di indicatori per le decisioni

In apertura di questa nota sono state ricordati i punti critici più rilevanti del sistema universitario italiano per quanto attiene all'attività di formazione: abbandoni, tempi di conseguimento del titolo, scarso collegamento tra formazione universitaria e mondo del lavoro; sempre in apertura è stata sottolineata l'attività della CRUI e del CNVSU che si sono preoccupati e si preoccupano di stabilire regole e/o di fornire suggerimenti che consentano di superare i punti critici richiamati. L'attenzione si è soffermata, in particolare, sulla individuazione e definizione di indicatori adeguati cui fare riferimento nella programmazione e gestione dell'attività didattica e di ricerca che si svolge nelle università.

Nel 1995 la CRUI ha curato la predisposizione di un documento:"*Organizzazione e metodi dei Nuclei di valutazione nelle università italiane. Le proposte della Conferenza dei Rettori*" che contiene un elenco molto dettagliato e puntuale di indicatori.

Nel 1998, l'Osservatorio per la Valutazione del Sistema Universitario, in attesa di una seconda generazione di indicatori CRUI, propone (DOC 11/98) ai nuclei di valutazione di inserire nella relazione annuale un "**insieme minimo di 22 indicatori**" che, anche se incompleto, sia tale da permettere una visione sufficientemente esplicativa del funzionamento dei vari atenei italiani. In tale ottica si prevede la raccolta di una serie di variabili tali da consentire l'elaborazione di alcuni **indicatori di risultato**, tenendo conto delle risorse disponibili (**indicatori di risorse**), del modo con cui tali risorse sono trasformate in prodotti (**indicatori di processo**) e dell'ambiente in cui l'ateneo si trova ad operare (**indicatori di contesto**).

Se si analizza il documento predisposto dall'Osservatorio appare subito evidente la necessità di una estensione della base informativa; è presumibile che l'esigenza fortemente sentita di non sovraccaricare di lavoro i costituendi nuclei di valutazione interna degli atenei abbia indotto a limitare in modo eccessivo il numero degli indicatori rispetto a quelli suggeriti dalla CRUI. Delle carenze informative presenti nell'insieme minimo di indicatori, l'Osservatorio prende coscienza immediatamente; infatti, se si scorre quanto previsto nel documento: "*Note tecniche sui dati ed informazioni da trasmettere entro il 2 maggio 2000*" si rileva che il numero degli indicatori passa da 22 a 29, numero questo che subisce un ulteriore ampliamento nell'anno successivo, e per rendersi conto di ciò e sufficiente scorrere il documento: "*Note tecniche e dati da trasmettere entro il 30 aprile 2001*" predisposto dal Comitato Nazionale per la Valutazione del Sistema Universitario (CNVSU) ormai subentrato all'Osservatorio.

Vale la pena sottolineare che, sia nel documento del 1998 che in quelli successivi, non viene affrontato il problema dell'aggregazione, delle informazioni che si desumono dai vari indicatori, finalizzata all'ottenimento di una misura sintetica dell'attività degli atenei ma si sottolinea, tra l'altro, la necessità di procedere alla definizione di una griglia di pesi da assegnare agli indicatori stessi; problema questo che (per quanto è di conoscenza dello scrivente) è ancora in attesa di soluzione.

Relativamente all'attività didattica, tenendo anche conto dell'avvio della riforma universitaria, si deve osservare che l'insieme degli indicatori previsti dal CNVSU, nonostante il loro sostanziale incremento rispetto alla proposta iniziale, non sembra coprire tutte le dimensioni d'interesse. A sostegno di questa affermazione si può addurre l'ottimo documento sull'accreditamento dei corsi di studio (RdR/01) predisposto da un apposito gruppo di lavoro nell'ambito delle attività del Comitato. Dall'esame degli elementi da prendere in considerazione ai fini dell'accreditamento di un corso di studi, si rileva come alcune dimensioni non trovino un adeguato corrispettivo nell'elenco degli indicatori previsti dal Comitato.

Tornando alle Note tecniche sopra citate, è apprezzabile la strategia adottata per avanzare le richieste di dati ai nuclei; infatti, negli elenchi delle variabili richieste nelle rilevazioni "*Nuclei 2000*" e "*Nuclei 2001*" non si procede ad una aggregazione per categoria di indicatori (contesto, risorse, processo e risultato) ma ad una aggregazione delle variabili per classi di "oggetti": nella Classe A, vengono inclusi dati relativi agli studenti iscritti, laureati e diplomati; nella Classe B, dati

relativi al personale in servizio; nella Classe C, dati finanziari; nella Classe D, altri dati. E' da presumere che la procedura utilizzata per richiedere i dati ai nuclei sia stata suggerita da ragioni di opportunità con l'intento di semplificare il lavoro di predisposizione del materiale richiesto. Ovviamente il materiale raccolto deve essere elaborato e reinserito nello schema iniziale che prevede l'aggregazione degli indicatori in quattro categorie distinte secondo la natura degli indicatori stessi. Infatti, è proprio tale strutturazione quella che sembra meglio rispondere alle necessità di individuazione degli elementi essenziali per la definizione e costruzione di un sistema informativo finalizzato alla risoluzione ottimale dei problemi da affrontare ai vari **livelli decisionali**.

Se ci si limita a considerare l'attività di formazione delle università risulta facile procedere all'elencazione dei livelli decisionali e, di conseguenza, dei soggetti utilizzatori del sistema informativo; livelli che sono:

- **Ministero**
- **Atenei**
- **Facoltà**
- **Corsi di studio**
- **Docenti**
- **Studenti**
- **Famiglie**
- **Mondo del lavoro.**

Per quanto attiene gli utilizzatori docenti, studenti, famiglie e mondo del lavoro, appare scontato che gli stessi possono contribuire alla costruzione del sistema informativo solo attraverso l'esplicitazione delle loro esigenze conoscitive; il compito di definizione e realizzazione del sistema informativo deve essere svolto dai soggetti erogatori del servizio formativo ai livelli più elevati, cioè, dal ministero, dagli atenei, dalle facoltà e dai corsi di studio. In realtà, se è certo che alla definizione del sistema informativo devono contribuire, ed anche in modo molto consistente, le facoltà, i corsi di studio ed i docenti, risulta più che ragionevole ipotizzare una responsabilità piena del Ministero e degli Atenei per quanto attiene l'effettiva realizzazione dei sistemi informativi stessi, ovviamente, le facoltà, i corsi di studio ed i docenti possono procedere autonomamente alle integrazioni del sistema informativo che ritengono opportune in funzione del soddisfacimento di specifiche esigenze conoscitive.

Dalle considerazioni sopra svolte si possono, anche se in modo parziale e molto sommario, illustrare gli elementi su cui deve essere basato un sistema informativo e le dimensioni cui il sistema stesso deve fare riferimento.

Tenendo presente che un **sistema informativo** che riguarda i processi di formazione universitaria deve consentire il conseguimento dell'obiettivo generale della formazione stessa che è quello dell'**innalzamento degli standard qualitativi** dei processi e che tale obiettivo implica l'adozione di politiche di intervento a vari livelli che, a loro volta, comportano una scelta tra un'insieme di alternative possibili, ne consegue la necessità di una attività di valutazione e monitoraggio.

Da quanto sopra sottolineato ne consegue che un **Sistema Informativo** per la gestione, valutazione e controllo dei processi formativi che si svolgono nelle Università dovrebbe essere basata sui criteri sotto elencati:

1. avere una struttura semplice che metta in evidenza i parametri per il controllo;
2. prevedere tutte le dimensioni necessarie per una descrizione esauriente dei processi in itinere e dei loro prevedibili sviluppi;
3. non deve essere una descrizione fine a se stessa, ma un atto che sottintende propositi di intervento tesi all'innalzamento della qualità
4. deve indurre organizzazione, ma non deve risolversi in puri aspetti organizzativi;

Inoltre, limitando il riferimento all'attività didattica che si svolge negli atenei, un sistema informativo soddisfacente dovrebbe quantomeno comprendere le seguenti cinque dimensioni:

1. obiettivi;
2. sistema organizzativo;
3. risorse;
4. processo formativo;
5. risultati.

Dimensioni che, a loro volta, si articolano in diversi elementi, che permettono di descrivere le potenzialità del processo ai fini del conseguimento degli obiettivi prefissati. Gli elementi caratterizzanti ciascuna dimensione sono:

- gli obiettivi generali e le esigenze delle parti interessate;
- le responsabilità ed il sistema di gestione;
- le risorse umane e le infrastrutture;
- la progettazione e l'erogazione dei servizi (anche di supporto);
- la valutazione dei risultati in funzione degli obiettivi prefissati;
- gli strumenti e metodi per il monitoraggio;
- gli strumenti e metodi per l'attuazione degli interventi.

Le considerazioni sopra svolte dovrebbero aver fornito un'idea, seppure molto grossolana e di prima approssimazione, su come ci si dovrebbe muovere per avviare a soluzione i problemi in cui si dibatte il sistema universitario italiano. In particolare si è sottolineata la necessità d'inquadramento di tutta la problematica in una logica decisionale. Il riferimento a tale logica consente di concludere che se si procede all'effettuazione delle scelte ispirandosi a criteri di comportamento razionale, esiste ed è possibile individuare una decisione ottimale (decisione razionale) compatibile con lo schema di preferenze del decisore.

Ovviamente quanto detto risulta estremamente schematico ed oltremodo semplicistico; infatti, anche se non è stato esplicitamente dichiarato, si è fatto riferimento a decisioni semplici anziché decisioni multiple e sequenziali, a decisioni ad un solo criterio anziché decisioni multicriterio (Keeney e Raiffa, 1976), è stata considerata l'unidimensionalità delle conseguenze anziché la multidimensionalità, non sono stati considerati i condizionamenti sui decisori che operano ai vari livelli provenienti dal contesto operativo di riferimento (ambiente, decisori di livello superiore, decisori di livello inferiore, decisioni pregresse, ecc.). Tutti questi aspetti, e tanti altri neppure menzionati, caratteristici dell'analisi decisionale risultano fondamentali nel "**governo dell'università**"; ma si tratta di aspetti, non ancora trattati a livello teorico, nella specificità che qui interessa, ma la cui estrema rilevanza e complessità meriterebbero particolare attenzione; in proposito si potrebbe pensare alla costituzione di un apposito gruppo di lavoro a carattere interdisciplinare il cui obiettivo è quello di pervenire alla definizione di metodi, protocolli, standard qualitativi e quantitativi condivisi per valutare, in un'ottica decisionale, i processi e i risultati relativi all'efficienza, all'efficacia e qualità delle attività di formazione, di ricerca e di gestione.

Tornando al tema del **"cosa valutare"** per poter decidere razionalmente, non si può non fare riferimento alle specificazioni previste nei documenti della CRUI e del CNVSU segnalati in precedenza; in proposito conviene, inoltre, considerare le: "*Note tecniche su dati e informazioni da trasmettere entro il 30 aprile 2002*" che offre un quadro aggiornato della visione del Comitato in merito al problema della valutazione del sistema universitario.

Nel documento predisposto dal Comitato vengono riassunti il dettaglio e le specificazioni sulle informazioni relative all'insieme di variabili da utilizzare per la costruzione di indicatori sull'intero sistema universitario, con particolare riferimento a quelle che dovranno essere raccolte e trasmesse a cura dei Nuclei di valutazione. I dati utilizzati sono stati distinti in 6 sezioni e riguardano: gli studenti, i corsi di studio ed i pareri degli studenti frequentanti; il personale; gli

aspetti finanziari; le strutture disponibili; la ricerca scientifica; l'avvio della riforma degli ordinamenti didattici.

Per quanto concerne i dati relativi agli studenti ed ai corsi di studio viene specificato il livello di riferimento: singolo corso di studio, facoltà, ateneo; viene, infine, dedicata particolare attenzione alla raccolta delle opinioni degli studenti frequentanti.

Ovviamente le informazioni richieste dal Comitato sono finalizzate alla costruzioni di una batteria di indicatori tali da consentire una valutazione comparativa, in termini di efficienza ed efficacia, tra atenei; informazioni il cui obiettivo prioritario è la ripartizione tra gli atenei stessi delle risorse disponibili. Proprio per la loro natura queste informazioni, anche se estremamente utili, non sono in grado di garantire il perseguimento degli obiettivi di ateneo richiamati in precedenza, ed in particolare non consentono un'adeguata e completa valutazione di tutti gli aspetti d'interesse nell'ottica di intervento finalizzata all'innalzamento del livello qualitativo dei processi formativi.

Tenendo presente anche l'avvio della riforma degli ordinamenti didattici, quali ulteriori informazioni concernenti l'attività didattica, oltre a quelle previste nel documento del Comitato, dovrebbero essere raccolte ed elaborate adeguatamente? Si può ragionevolmente ritenere che, o perché richieste dalla normativa vigente o perché, comunque, utili al governo dell'università, le informazioni concernenti l'attività didattica devono essere quantomeno tali da consentire:

1. la valutazione della didattica da parte degli studenti frequentanti;
2. la valutazione dei servizi di supporto alla didattica (segreterie, orientamento, tutorato, biblioteche, ecc.) da parte degli studenti (frequentanti e non frequentanti);
3. la misura del carico didattico (monitoraggio e valutazione dei crediti attribuiti ai vari insegnamenti);
4. il monitoraggio e la valutazione delle carriere degli studenti a livello individuale;
5. la valutazione/autovalutazione della didattica e dei servizi di supporto alla didattica da parte del personale docente e non docente;
6. la valutazione relativa all'impiego delle risorse (finanziarie, strutture, personale docente e non docente) destinate alla didattica;
7. la valutazione a livello individuale dell'attività didattica svolta dal personale docente;
8. il monitoraggio e la valutazione a livello individuale degli abbandoni e dei trasferimenti degli studenti;
9. il monitoraggio, la valutazione e l'autovalutazione dei singoli corsi di studio (immatricolazioni, abbandoni, conseguimenti del titolo, attività di orientamento, tutorato e di supporto attivate, ecc.);
10. il monitoraggio e la valutazione delle attività relative alla formazione permanente;
11. il monitoraggio e la valutazione delle attività di tirocinio;
12. il monitoraggio e la valutazione delle attività di orientamento;
13. il monitoraggio e la valutazione delle attività che sono ricompresse nell'ambito di progetti speciali (Campus, Campus Like, Campus One, TRIO, ecc…. .;
14. il monitoraggio e la valutazione delle attività didattiche e di tirocinio svolte in collaborazione con università e/o enti, e/o imprese non italiane (progetti Socrates/ Erasmus, Leonardo, ecc.);
15. il monitoraggio e la valutazione degli sbocchi occupazionali dei laureati/diplomati;
16. il monitoraggio e valutazione dell'attività svolta per facilitare l'inserimento dei laureati/diplomati nel mondo del lavoro.

Attività di monitoraggio e valutazione, quelle sopra elencate, certamente utili ai fini di un corretto ed efficiente governo dell'università finalizzato all'impiego ottimale delle risorse e all'innalzamento dei livelli qualitativi dei processi che si svolgono nell'ambito delle università[1]. Al

[1] Sui problemi connessi alla valutazione del Sistema universitario nelle sue varie articolazioni si possono utilmente consultare: Rowley, 1996; Cave, Hanney, Henkel e Kogan, 1997; Modica e Stefani, 1997; Biggeri, 1998 e 1999; Gola, 1998; Università de Toulon, 1998; Bini, 1999; Gori e Vittadini, 1999; Gola, Squarzoni, Stefani, Tosi e Tronci, 2002).

riguardo si segnala che una quota non indifferente dei dati cui s'è fatto sopra riferimento sono già disponibili in molte università italiane, o perché costituiscono parte dell'archivio di Ateneo, o perché raccolti attraverso apposite indagini; un patrimonio quello disponibile molto ricco ma che, nella generalità dei casi, necessità di una adeguata risistemazione nel contesto di sistema informativo adeguato. L'impressione dell'estensore di questa nota e che i sistemi informativi già realizzati in molti atenei italiani siano incompleti e, quasi sempre, configurati in modo non ottimale.

3. Un caso di studio: l'attività di valutazione svolta dall'Ateneo Fiorentino

Riguardo a quanto sopra richiamato, l'Ateneo Fiorentino si è impegnato negli ultimi anni in modo consistente conseguendo dei risultati che possono essere considerati complessivamente, se non ottimali, abbastanza soddisfacenti. Qui di seguito vengono richiamate alcune delle attività svolte; in particolare, si farà riferimento alla valutazione della didattica da parte degli studenti frequentanti ed agli sbocchi occupazionali dei laureati/diplomati.

Relativamente alla valutazione della didattica, si sottolinea che attraverso la raccolta delle opinioni degli studenti frequentanti, l'obiettivo primario che s'intende perseguire è quello della individuazione dei fattori che facilitano o che ostacolano l'apprendimento da parte degli studenti stessi sia in termini di svolgimento dell'attività didattica sia riguardo alle condizioni logistiche in cui la stessa si realizza.

Non bisogna, inoltre, dimenticare che i valori assunti dali indicatori provenienti dall'elaborazione dell'opinione degli studenti sono soltanto uno degli elementi, di rilevanza certamente non marginale, su cui basare la valutazione del processo formativo ed è fondamentale che questi indicatori non vengano utilizzati per meccanismi automatici di premio/sanzione, ma che invece passino, insieme alle altre informazioni, attraverso il filtro di un giudizio competente e coerente con una corretta politica di assicurazione della qualità.

Considerazioni analoghe valgono per tutto quanto attiene alle condizioni in cui si svolge la didattica (aule, dotazioni e attrezzature) potendo sostenere che questi sono, in effetti, fattori affidati alla gestione e al controllo del Corso di Studio. E ancora, per la valutazione dei carichi di studio, sia del singolo insegnamento sia degli insegnamenti da seguire in contemporanea: anche questi fattori dovrebbero appartenere a una corretta progettazione del processo formativo, e dovrebbero trovare una corretta espressione tramite i 'crediti'.

In proposito si deve sottolineare che dimensioni della didattica quali aule, dotazioni e attrezzature, carico di lavoro, fanno parte delle esperienze dirette sulle quali lo studente ha titolo per rispondere e che è utile raccogliere notizie su tali fattori sia ai fini della "*assicurazione della qualità*" (per una funzione di auto-controllo sugli effetti delle scelte operate, in materia di didattica, da parte del Corso di Studi) sia ai fini di "*audit*" da parte degli organismi di livello più elevato.

Si deve, infine, ribadire che i dati raccolti e le elaborazioni effettuate possono costituire una fonte informativa molto articolata e densa di implicazioni operative e che spetta poi agli organi preposti al governo e gestione dei processi formativi (Senato Accademico, Consiglio di Amministrazione, Consigli di Facoltà, Consigli di corso di studi, Commissioni per la didattica e singoli docenti) trarne profitto: interventi finalizzati al perseguimento di più elevati standard qualitativi.

Per quanto attiene agli sbocchi occupazionali dei laureati è ormai diffusa la consapevolezza di quanto l'ingresso dei giovani nel mondo del lavoro rappresenti un momento particolarmente critico rispetto al quale i giovani stessi e l'intera società sono chiamati a confrontarsi quotidianamente; la transizione dalla scuola al lavoro per i possessori di una preparazione a livello universitario, seppure meno problematica e più rapida di coloro che non possiedono tale preparazione, presenta delle connotazioni negative che, se analizzate in modo adeguato, potrebbero essere, se non completamente eliminate, quantomeno, sostanzialmente ridotte.

Tra gli aspetti negativi della formazione universitaria che meritano particolare attenzione, e sui quali si può certamente intervenire in modo fattivo, si devono annoverare la già segnalata eccessiva durata degli studi e l'inserimento professionale molto spesso inadeguato al percorso di studi concluso.

Fino ad un recente passato, però, la scarsa disponibilità di strumenti per la verifica dell'efficacia formativa di ciascuna Università e per il confronto dei risultati ottenuti in corsi identici presso sedi diverse, la scarsa attenzione e l'insufficiente approfondimento dei principali aspetti strutturali dell'Università nel suo complesso, hanno favorito l'accumulo di risultati negativi per l'intero sistema formativo universitario e determinato valutazioni distorte sul sistema di istruzione superiore italiano, anche a livello di comunità internazionale.

In questo contesto - sia per rispondere a quesiti di corretto impiego delle risorse, sia per verificare la validità dei percorsi didattici intrapresi, ma anche in una più ampia ottica di analisi della competitività sul mercato della formazione superiore - hanno preso avvio in questi ultimi anni, rivestendo subito importanza fondamentale, i progetti di valutazione della qualità misurata in termini d'efficienza e d'efficacia (interna e d esterna) dei processi formativi.

Anche l'Ateneo Fiorentino ha di recente dedicato particolare attenzione alla misura dell'efficienza e dell'efficacia esterna, progettando e realizzando indagini sulla valutazione degli sbocchi occupazionali dei propri laureati/diplomati.

Le indagini si collocano nell'ambito della misura dell'efficacia esterna, ma per certi versi anche dell'efficienza del sistema di formazione; è ragionevole, infatti, presumere che, ad esempio, l'eccessivo prolungamento della durata degli studi emerso dalle indagini (in alcuni casi più che doppia rispetto alla durata legale) sia da attribuire all'uso non ottimale delle risorse: dei docenti, delle strutture didattiche e degli studenti stessi le cui capacità non vengono certamente valorizzate da corsi il cui contenuto risulti eccessivamente pesante o di livello non adeguato (a causa di carenze pregresse dovute alla formazione pre-universitaria o al mancato coordinamento dei livelli e dei contenuti dei diversi corsi universitari). Si tratta di un'allocazione non ottimale di risorse o, comunque, di un'allocazione non in linea con processi formativi in grado di soddisfare l'esigenza, comune alla maggior parte dei laureati, di un adeguato e rapido inserimento nel mercato del lavoro.

3.1 Valutazione della didattica

Alla valutazione della didattica da parte degli studenti, l'Ateneo Fiorentino ha, negli ultimi anni dedicato particolare attenzione; tra il materiale prodotto in conseguenza dell'attività svolta si segnalano in particolare i rapporti sotto elencati (i documenti sono consultabili sul sito http://www.unifi.it/aut_dida/indexval.html).

1. Valutazione della didattica da parte degli studenti (VDS) - relazione sul triennio 1996/1999.
2. Valutazione della didattica da parte degli studenti (VDS) attraverso un questionario unico - relazione sulla sperimentazione a.a. 1999/2000.
3. Valutazione della didattica da parte degli studenti (VDS) - a.a. 2000-2001 - Relazione tecnica.
4. Valutazione della didattica da parte degli studenti (VDS) - a.a. 2000-2001 - Dati aggregati per Facoltà e Corso di Laurea/Diploma.
5. Valutazione della didattica da parte degli studenti frequentanti - a.a. 2001-2002, I° semestre.

Di seguito si ripropongono, in forma estremamente sintetica, alcuni commenti, grafici e tabelle contenuti: a) nella Relazione tecnica "*La valutazione della didattica da parte degli studenti attraverso un questionario unico: Relazione sulla rilevazione dell'Ateneo Fiorentino nell'A.A.*

2000/2001"; b) nel Rapporto "*Valutazione della didattica da parte degli studenti (VDS) - a.a. 2000-2001 - Dati aggregati per Facoltà e Corso di Laurea/Diploma*"; si presenteranno, inoltre, i risultati di alcune elaborazioni sui dati relativi agli studenti che hanno frequentato nel I° semestre dell'anno accademico corrente.

Il questionario[2] impiegato nella rilevazione è composto da una scheda fronte-retro identica per tutto l'Ateneo, ad eccezione dei quesiti Q17-Q21, che sono stati formulati dalle singole Commissioni di Facoltà o Corso di Laurea/Diploma per soddisfare specifiche esigenze conoscitive relative, in particolare, all'attività di supporto didattico.

La facciata anteriore del questionario risulta composta dalle seguenti parti:

1. Intestazione e codici;
2. Informazioni sullo studente;
3. Quesiti di valutazione, raggruppati nelle seguenti sezioni:
   a. **Aule e attrezzature** (Q1-Q2);
   b. **Carico di lavoro e organizzazione della didattica** (Q3-Q5);
   c. **Lezioni** (Q6-Q16), ulteriormente distinguibile in una parte relativa agli aspetti formali, quali rispetto del programma, dell'orario ecc. (Q6-Q12) e in una parte relativa alla qualità dell'insegnamento (Q13-Q16);
   d. **Aspetti specifici** del corso di studi (Q17-Q21);
   e. **Informazioni aggiuntive** (Q22-Q26);
4. **Suggerimenti**.

La facciata posteriore contiene le istruzioni per la compilazione e la consegna del questionario, una nota sul carattere anonimo delle informazioni fornite e sul loro utilizzo e uno spazio per le osservazioni personali.

L'elenco completo dei quesiti comuni a tutti i corsi di studio è quello sotto riportato nella pagina seguente.

[2] Il questionario deriva da una rielaborazione di quello proposto in B. Chiandotto e M. Gola (1999).

	Decisamente NO	Più NO che si	Più SI che no	Decisamente SI
Aule ed attrezzature				
1. Le aule dove si svolgono le lezioni sono adeguate *(si vede, si sente, si trova posto)***?**	1 ❒	2 ❒	3 ❒	4 ❒
2. I locali e le attrezzature per *eventuali* esperienze pratiche *(esercitazioni, laboratori, ecc.)* **sono adeguate?** *Esperienze pratiche non previste* ❒	1 ❒	2 ❒	3 ❒	4 ❒
Carico di lavoro e organizzazione della didattica				
3. Il carico di lavoro richiesto dall'insegnamento è accettabile?	1 ❒	2 ❒	3 ❒	4 ❒
4. Il carico di lavoro complessivo richiesto per gli insegnamenti previsti nel piano di studio ufficiale per ilo periodo di riferimento *(trimestre, semestre, anno)* è accettabile? *Piano di studi non previsto* ❒	1 ❒	2 ❒	3 ❒	4 ❒
5. L'organizzazione *(orario, calendario, esami, ecc.)* **degli insegnamenti previsti nel piano di studi per il periodo di riferimento** *(trimestre, semestre, anno)* **è adeguata?** *Piano di studi non previsto* ❒	1 ❒	2 ❒	3 ❒	4 ❒
Lezioni (Se le lezioni sono tenute da più docenti esprimere una valutazione media)	1 ❒	2 ❒	3 ❒	4 ❒
6. Le lezioni sono aderenti al programma? *Programma non previsto* ❒	1 ❒	2 ❒	3 ❒	4 ❒
7. Le lezioni vengono tenute rispettando il calendario ufficiale? *Calendario non previsto* ❒	1 ❒	2 ❒	3 ❒	4 ❒
8. Il docente rispetta l'orario concordato per le lezioni *(puntualità del docente)***?**	1 ❒	2 ❒	3 ❒	4 ❒
9. Il docente è reperibile durante l'orario di ricevimento?	1 ❒	2 ❒	3 ❒	4 ❒
10. Sono state dichiarate le modalità e le regole di esame?	1 ❒	2 ❒	3 ❒	4 ❒
11. Il materiale didattico consigliato *(libri, dispense)* è sufficiente per la comprensione della materia?	1 ❒	2 ❒	3 ❒	4 ❒
12. Il materiale didattico consigliato è facilmente reperibile?	1 ❒	2 ❒	3 ❒	4 ❒
13. Il docente espone gli argomenti in modo chiaro?	1 ❒	2 ❒	3 ❒	4 ❒
14. Gli argomenti affrontati nelle lezioni sono trattati in modo esauriente?	1 ❒	2 ❒	3 ❒	4 ❒
15. Il docente stimola/motiva l'interesse verso gli argomenti?	1 ❒	2 ❒	3 ❒	4 ❒
16. Il docente risponde esaurientemente alle richieste di chiarimento?	1 ❒	2 ❒	3 ❒	4 ❒
Aspetti specifici del corso di studi non previsto	1 ❒	2 ❒	3 ❒	4 ❒
17. ❒	1 ❒	2 ❒	3 ❒	4 ❒
18. ❒	1 ❒	2 ❒	3 ❒	4 ❒
19. ❒	1 ❒	2 ❒	3 ❒	4 ❒
20. ❒	1 ❒	2 ❒	3 ❒	4 ❒
21. ❒	1 ❒	2 ❒	3 ❒	4 ❒
Informazioni aggiuntive	1 ❒	2 ❒	3 ❒	4 ❒
22. La frequenza alle lezioni e/o esercitazioni è accompagnata da una regolare attività di studio?	1 ❒	2 ❒	3 ❒	4 ❒
23. Gli argomenti trattati sono risultati nuovi rispetto a quelli affrontati in insegnamenti precedenti?	1 ❒	2 ❒	3 ❒	4 ❒
24. Le conoscenze preliminari possedute sono risultate sufficienti per affrontare l'insegnamento?	1 ❒	2 ❒	3 ❒	4 ❒
25. Indipendentemente da come è stato svolto l'insegnamento, sei interessato a questa disciplina?	1 ❒	2 ❒	3 ❒	4 ❒
26. Sei globalmente soddisfatto di questo insegnante?	1 ❒	2 ❒	3 ❒	4 ❒

Suggerimenti *(sul retro della scheda si possono formulare anche suggerimenti scritti)*

Alleggerire il carico didattico complessivo ❒ - Aumentare l'attività di supporto didattico ❒ - Fornire più conoscenze di base ❒ - Eliminare dal programma argomenti già trattati in altri corsi ❒ - Migliorare iòl coordinamento con altri corsi e/o moduli ❒ - Migliorare la qualità del materiale didattico ❒ - Fornire in anticipo il materiale didattico ❒ - Inserire prove d'esame intermedie ❒

VALUTAZIONE DELLA DIDATTICA A.A. 2000/01

I questionari raccolti sono stati 59330, per un totale di 2415 corsi monitorati (almeno una scheda) e 2059 corsi valutabili (almeno sei schede). I dati disaggregati per le due fasi sono riportati nel prospetto che segue:

	Questionari	Corsi monitorati	Corsi valutabili	Tasso di risp.
Prima fase	31609	1172	1005	72.02%
Seconda fase	27721	1243	1054	56.08%
Totale	59330	2415	2059	62.64%

La stima del tasso di risposta, riportata nell'ultima colonna, è data dal rapporto fra la somma del numero di schede raccolte e la somma dei corrispondenti numeri medi di frequentanti. A livello di Ateneo, la stima del tasso di risposta è pari a 62.64%, un risultato che lascia intravedere un ampio margine di miglioramento (la stima - basata sul numero degli studenti iscritti in corso, sul numero medio di corsi frequentati in un anno e sulla % di studenti frequentanti - del numero massimo di schede ottenibili è di circa 92000).

Se si osservano i tassi di partecipazione a livello di singola facoltà o corso di laurea/diploma si rileva una situazione estremamente diversificata. A livello di facoltà si passa da un valore minimo di 20.63% per Giurisprudenza ad un valore massimo di 72.95% per Economia; a livello di corso di laurea/diploma, il divario tra i tassi di partecipazione è ancora più marcato, il valore minimo ed il valore massimo si registrano nell'ambito della Facoltà di Medicina e Chirurgia e riguardano, rispettivamente, il Corso di laurea In Scienze Motorie (19,14%) ed il Corso di diploma in Tecnico Sanitario di Laboratorio (96,43%). In proposito si deve, comunque, sottolineare l'elevato numero di corsi di laurea/diploma con tassi di partecipazione superiori all'80% e che alcuni corsi non sono stati monitorati, ovviamente, per questi ultimi il tasso di partecipazione è pari a zero.

Relativamente alla Facoltà di Giurisprudenza si deve, inoltre, rilevare che le schede compilate sono in numero estremamente ridotto 722, su un totale di 6401 studenti iscritti di cui 2728 in corso; per contro nella facoltà di Scienze Politiche le schede compilate sono state 4976, su un totale di 4090 studenti iscritti di cui 1969 in corso. Il numero estremamente esiguo delle valutazioni espresse dagli studenti della Facoltà di Giurisprudenza ingenera forti perplessità sul loro grado di rappresentatività, pertanto, nella lettura ed interpretazione delle tabelle e dei grafici si deve tener presente che i valori che riguardano questa Facoltà hanno una valenza molto limitata

Un'idea complessiva del "grado di rappresentatività" dei dati raccolti si desume osservando i valori riportati nella Fig. 1. Il divario tra la percentuale di studenti iscritti e la percentuale di questionari compilati misurato, ad esempio, dal rapporto tra le due percentuali, anche se risente pesantemente della diversa frequenza alle lezioni degli studenti iscritti nei vari corsi di laurea/diploma, fornisce un'indicazione utile, anche a fini operativi, sul diverso grado di "coinvolgimento" degli studenti. Le facoltà con il più elevato coinvolgimento degli studenti sono: Farmacia (2,63), Scienze M.F.N. (1,83), Agraria (1,74) e Ingegneria (1,73); mentre le facoltà in cui si registrano i livelli più bassi di coinvolgimento sono: Giurisprudenza (0,11), Scienze della Formazione (0,66) e Architettura (0,69).

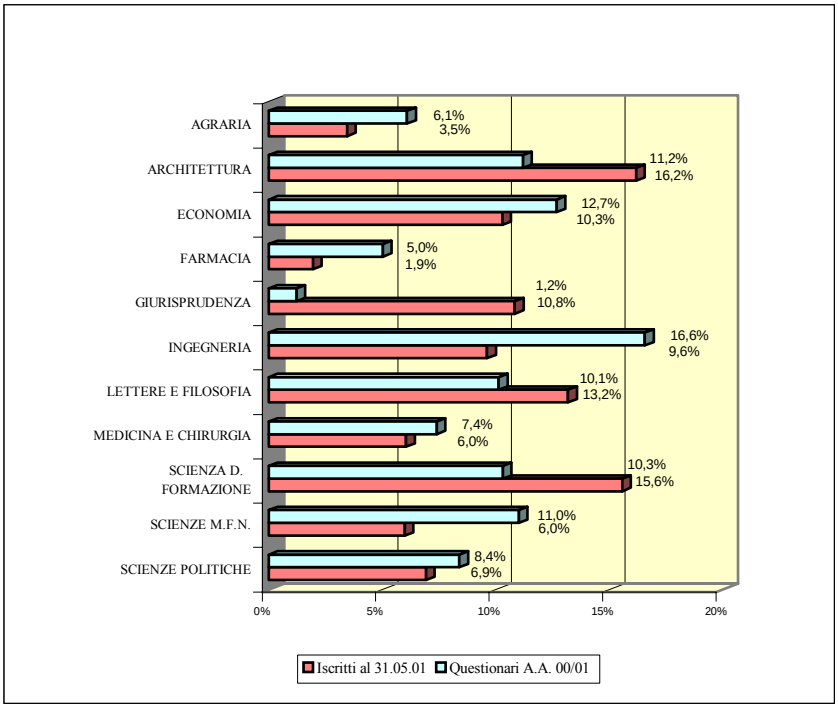


Fig. 1 – Università di Firenze: studenti iscritti nell'A.A. 2000/2001 e "coinvolgimento" all'indagine per Facoltà (valori %)

Nota: I valori riportati nel grafico non tengono conto dei questionari compilati via Web.

Per ogni Corso di Laurea/Diploma, sono stati predisposti e consegnate alle presidenze di facoltà o di corso di studi le seguenti informazioni:

1. archivio integrale dei dati della rilevazione;
2. tabella riassuntiva degli insegnamenti monitorati (qualunque sia il numero di schede raccolte), con indicazione del numero di schede raccolte e del numero medio di frequentanti;
3. tabelle e grafici che sintetizzano la valutazione a livello di:
   a. Facoltà,
   b. Corso di Laurea/Diploma,
   c. Singoli insegnamenti.

Alle presidenze sono, inoltre state consegnate anche le schede cartacee che dovrebbero essere state messe a disposizione dei singoli docenti interessati in modo da consentire ai docenti stessi di prendere visione delle osservazioni personali espresse dagli studenti e che sono riportate sul retro della scheda di rilevazione.

Le elaborazioni hanno interessato solo gli insegnamenti con almeno 6 schede valide raccolte.

La tabella riassuntiva riporta, in ordine di codice, gli insegnamenti monitorati, con indicazione del numero di schede raccolte, del numero medio di frequentanti e del codice di CdL/DU degli studenti che hanno compilato le schede. Per gli insegnamenti tenuti da più docenti sono indicati anche il codice e nome del docente, secondo l'attribuzione effettuata dai rilevatori.

I risultati della valutazione sono sintetizzati, a livello di Facoltà, di Corso di Laurea/Diploma o di singolo insegnamento, da tabelle e da grafici con la stessa struttura di quelli riportati (a titolo esemplificativo) nelle pagine successive e che sono relativi al Corso di Laurea in Scienze Politiche.

La tabella indica, per ogni quesito, il numero di risposte e le percentuali di risposta 1, risposta 2, risposta 3 e risposta 4; la tabella riporta, inoltre, la mediana, la media aritmetica e lo scarto quadratico medio che si ottengono attribuendo alle risposte i seguenti punteggi:

risposta 1 (decisamente no) → punti 2
risposta 2 (più no che sì) → punti 5
risposta 3 (più sì che no) → punti 7
risposta 4 (decisamente sì) → punti 10

L'attribuzione dei punteggi, che facilita la lettura dei risultati, è ovviamente opinabile; tuttavia alcune sperimentazioni effettuate mostrano che l'attribuzione di punteggi diversi (purché ragionevoli) non altera nella sostanza i risultati della valutazione. La tabella riporta inoltre, per ogni suggerimento (seguendo l'ordine della scheda), il numero di studenti che ha formulato il suggerimento e la percentuale sul numero di schede valutative del corso.

Il grafico è costituito da una coppia di diagrammi cartesiani; il primo diagramma fornisce una immagine visiva delle valutazioni: ognuno dei 26 quesiti viene rappresentato assumendo come ordinata la media aritmetica e come ascissa lo scarto quadratico medio; il secondo diagramma, utile a fini comparativi, mette a confronto le medie aritmetiche dei 26 quesiti dell'insegnamento[3] con quelle del CdL/DU di appartenenza (per il grafico del CdL/DU il confronto è con la Facoltà di appartenenza). A questo proposito, vogliamo sottolineare che quello con la media del CdL è solo uno dei possibili confronti e va interpretato con cautela: ad esempio, se nella rilevazione effettuata la maggior parte delle valutazioni raccolte si riferisce ad insegnamenti facoltativi che notoriamente ottengono buoni giudizi, la media del CdL è sovrastimata. Inoltre, è bene tener sempre presente il valore assoluto dei giudizi, poiché è possibile che la valutazione di un insegnamento sia bassa se confrontata con la media del proprio CdL, ma comunque buona in termini assoluti (o rispetto a insegnamenti analoghi impartiti di altri CdL).

[3] Nel grafico non vengono rappresentati i quesiti che hanno ricevuto meno di 6 valutazioni - tipicamente quelli relativi al supporto didattico nei corsi in cui questo non è previsto.

Facoltà: "11 - SCIENZE POLITICHE"
CdL: "25 - SCIENZE POLITICHE"

Numero di schede valide raccolte: 1822

Quesito	n. r.	% no	% no/si	% si/no	% si	IMI	IMedia	SQMI
Q1	1827	10.1	23.5	37.5	29.2	7	6.90	2.47
Q2	821	20.8	34.0	32.3	12.9	5	5.67	2.44
Q3	1813	3.7	13.4	47.9	35.0	7	7.09	2.06
Q4	1605	6.0	25.5	52.9	15.5	7	6.65	1.95
Q5	1637	8.2	25.5	40.8	16.5	7	6.57	2.09
Q6	1774	1.4	5.9	31.2	61.5	10	8.66	1.84
Q7	1759	0.6	2.6	26.9	69.8	10	9.01	1.58
Q8	1809	1.1	3.2	20.1	75.6	10	9.15	1.60
Q9	1583	1.2	3.0	33.0	62.7	10	8.76	1.72
Q10	1797	4.9	8.8	21.1	65.2	10	8.54	2.26
Q11	1805	3.8	11.7	35.9	48.6	7	8.02	2.17
Q12	1806	3.7	10.1	34.4	51.8	10	8.17	2.15
Q13	1811	2.8	10.0	33.2	53.9	10	8.28	2.08
Q14	1809	2.2	12.6	41.6	43.7	7	7.95	2.02
Q15	1806	4.7	13.6	37.4	44.4	7	7.83	2.24
Q16	1811	1.6	6.5	33.5	58.5	10	8.54	1.88
Q17	672	6.2	15.9	45.7	32.1	7	7.32	2.23
Q18	739	2.9	10.3	51.7	35.1	7	7.70	1.95
Q19	1555	38.0	33.0	17.9	11.1	5	4.77	2.62
Q20	1654	2.4	7.5	36.2	54.0	10	8.35	1.98
Q21	1446	16.2	36.5	24.5	22.8	5	6.15	2.61
Q22	1806	4.8	23.8	47.8	23.5	7	6.99	2.07
Q23	1794	12.4	26.3	36.1	25.2	7	6.61	2.50
Q24	1795	7.5	19.4	46.9	26.2	7	7.02	2.24
Q25	1813	3.1	9.1	35.5	52.2	10	8.33	2.08
Q26	1817	4.3	10.8	40.0	44.9	7	7.92	2.17

Suppl.	n. r.	%
S1	373	20.4
S2	144	7.9
S3	253	13.8
S4	130	7.1
S5	208	11.4
S6	214	11.7
S7	230	12.6
S8	391	21.3

Legenda:
n. r. = numero di risposte
% no = % risposte decisamente no (peso: 1)
% no/si = % risposte più no che sì (peso: 3)
% si/no = % risposte più sì che no (peso: 7)
% si = % risposte decisamente sì (peso: 10)
IMI = Idoneità
SQMI = Scarto Quadratico Medio

Tab. 1 – Università di Firenze: valutazione della didattica da parte degli studenti frequentanti nell'A.A. 2000/01.

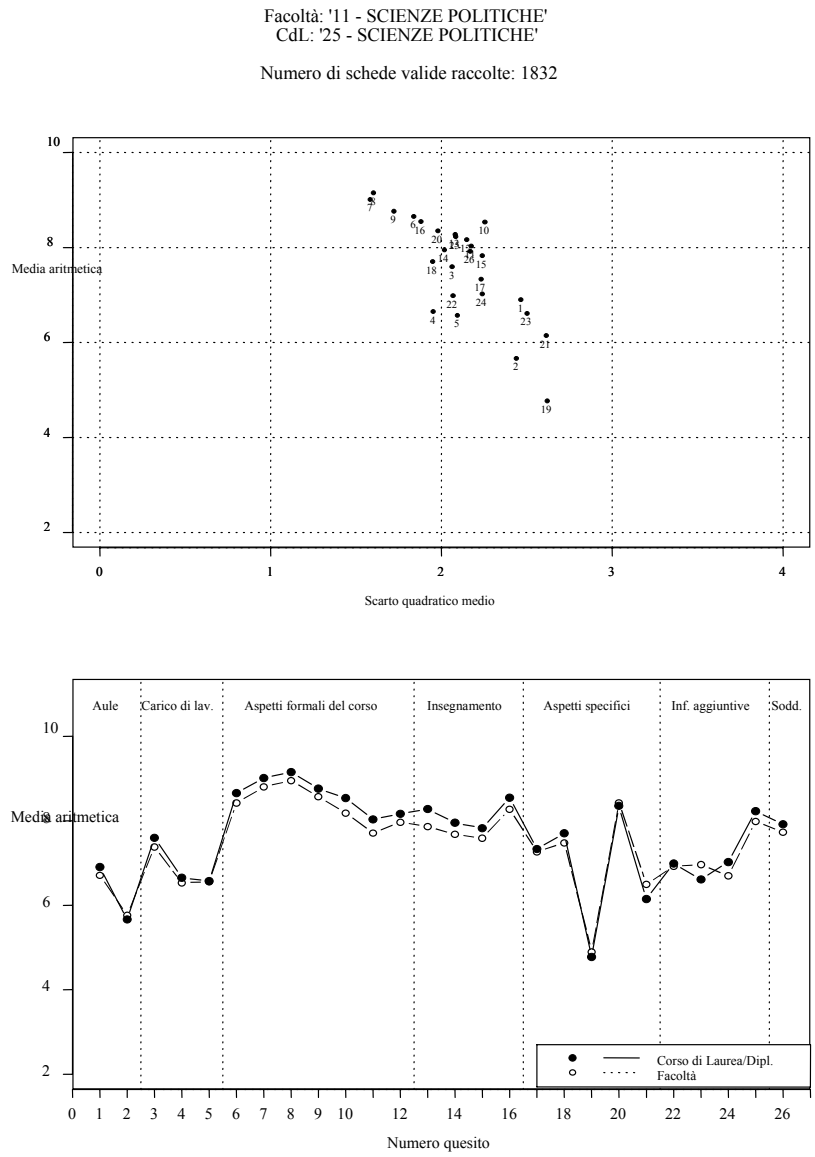


Fig. 2 – Università di Firenze: valutazione della didattica da parte degli studenti frequentanti nell'A.A. 2000/01.

Sui dati rilevati si è proceduto all'effettuazione di ulteriori elaborazioni a livello aggregato ed alla costruzione di graduatorie a livello di facoltà e di corso di studi.

Dall'esame della Fig. 2 si rileva come la valutazione (media e relativa all'intero ateneo) espressa dagli studenti in merito ai vari punti toccati dal questionario non sia omogenea. Per alcuni aspetti, emerge un giudizio decisamente positivo (quesiti Q6-Q9) o molto soddisfacente (quesiti Q10-Q16), per altri aspetti il giudizio si colloca a livelli che sfiorano la sufficienza[4] (quesiti Q4 e Q5); mentre, la valutazione relativa alla disponibilità di locali e attrezzature per attività di supporto (quesito Q2), il giudizio risulta decisamente negativo. Non raggiungono la sufficienza nemmeno le valutazioni medie relative alla regolarità degli studi (quesito Q22) e quelle relative alle conoscenze preliminari (quesito Q24).

[4] A ragione della scala utilizzata si può, ragionevolmente, collocare il livello di **sufficienza** sul valore numerico **7**.

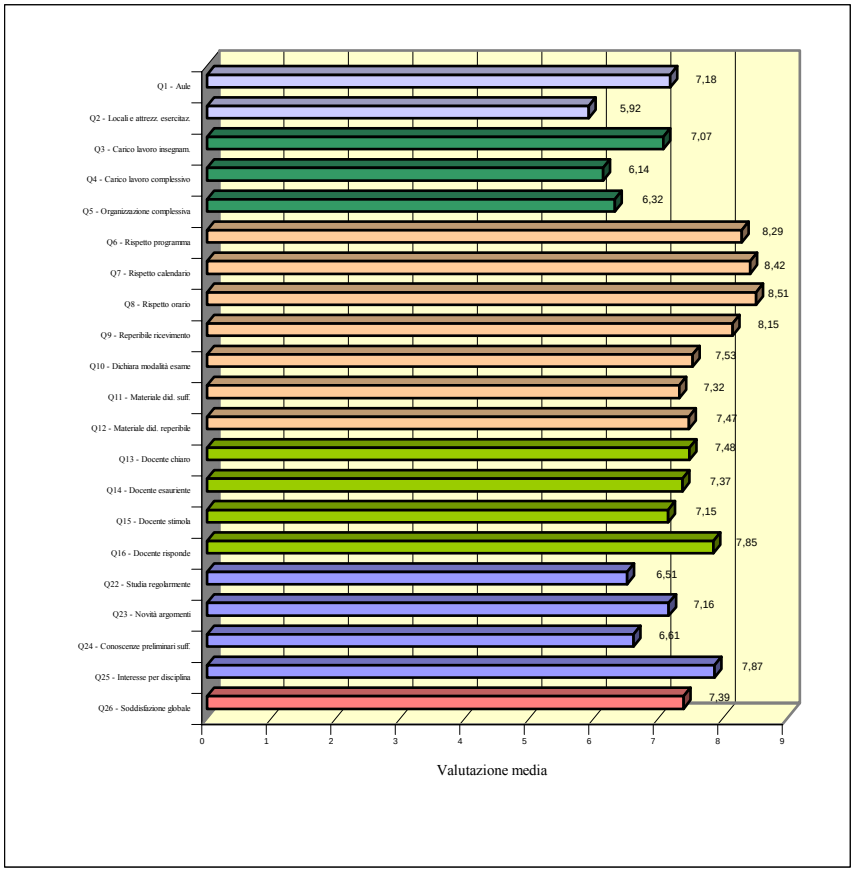


Fig. 2 – Università di Firenze: valutazioni medie della didattica per l'intero Ateneo A.A. 2000/2001

In estrema sintesi, si può affermare che gli studenti:

i. "**promuovono**" i docenti, sia per quanto attiene gli aspetti formali connessi alla didattica (quesiti Q6-Q12), sia per quanto concerne gli aspetti sostanziali (quesiti Q13-Q16, Q25 e Q26);

ii. "**bocciano**" il corpo docente riguardo alla organizzazione della didattica (quesiti Q4, Q5 e Q24);

iii. "**bocciano**" l'Ateneo (il Ministero?) relativamente alla disponibilità di locali ed attrezzature (quesito Q2);

iv. si "**autobocciano**", non accompagnando con una regolare attività di studio la frequenza alle lezioni e alle esercitazioni (quesito Q22).

Ovviamente, il dato medio di Ateneo è la risultante di una situazione disomogenea a livello di facoltà e di corso di laurea/diploma.

Dall'esame dei dati (relativamente a tutti i quesiti ed a tutte le sezioni) emergono indicazioni di estremo interesse; nelle righe che seguono, si richiama l'attenzione su alcuni aspetti particolarmente interessanti; le considerazioni svolte riguardano nella generalità dei casi le facoltà ma, ovviamente, le maggiori potenzialità informative del materiale prodotto si collocano a livello di corso di laurea/diploma. I dati completi a livello di facoltà e di singolo corso di studi con le relative graduatorie sono riportati nel già citato "Valutazione della didattica da parte degli studenti (VDS) - a.a. 2000-2001 - Dati aggregati per Facoltà e Corso di Laurea/Diploma" consultabile sul sito http://www.unifi.it/aut_dida/indexval.html

Nella lettura e nella interpretazione dei dati si deve tener conto dei limiti che hanno gli indicatori sintetici, soprattutto se si tratta di indicatori semplici (non ponderati) come quelli qui

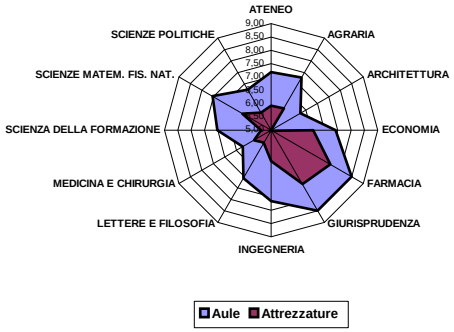
proposti, "costruiti" sulla base di valutazioni soggettive che, per loro natura, non possono tener conto di tutte le sfaccettature e le implicazioni, oggettive e soggettive, che determinano le valutazioni stesse[5].

Di seguito vengono proposti, corredati da brevi commenti, alcuni grafici riguardanti aspetti della didattica particolarmente interessanti[6].

Relativamente alla **dotazione di aule** (quesito Q1), a fronte di un dato medio di ateneo che supera la sufficienza (7.18), si collocano dati medi di facoltà che evidenziano chiaramente situazioni di **disagio**; si tratta, in particolare, delle facoltà di **Scienze Politiche** (6.76), **Architettura** (6.28), **Medicina e Chirurgia** (6.25). Se si passa all'esame della situazione al livello più specifico si osservano situazioni ancora più problematiche e che interessano anche corsi di laurea/diploma di facoltà diverse da quelle sopra segnalate. Ed è questo il caso del corso di laurea in **Scienze della formazione primaria** (6.21), del corso di laurea in **Economia aziendale** (5.94) e del corso di diploma in **Servizio Sociale** (5.83).

La situazione concernente la disponibilità di **locali ed attrezzature** (quesiti Q1 e Q2, Fig. 3) per esercitazioni, laboratori, ecc., che risulta decisamente **insufficiente** a livello medio di ateneo, si rivela "disastrosa" in molte facoltà: **Agraria** (5.95), **Scienze Politiche** (5.75), **Medicina e Chirurgia** (5.76), **Lettere e Filosofia** (5.54), **Scienza della Formazione** (5.37) ed **Architettura** (4.98).

Fig. 3 - Università di Firenze: VALUTAZIONE DELLA DIDATTICA A.A. 2000/2001

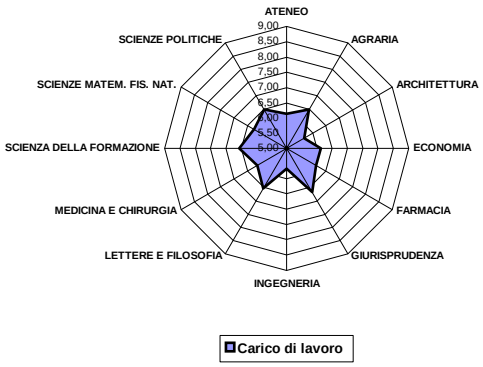


Riguardo al **carico di lavoro complessivo** (quesito Q4, Fig. 4), dai dati emerge in modo netto una situazione di **disagio generalizzato**: a fronte di un dato medio di ateneo pari 6.14, si registra un valore massimo (comunque inferiore alla sufficienza) nella facoltà di Giurisprudenza (6.65) ed un valore minimo nella facoltà di Ingegneria (5.67) e Architettura (5.67).

[5] Alcune considerazioni sui problemi di ponderazione e sulle possibilità di "depurare" le valutazioni dagli effetti di fattori strutturali eventualmente presenti sono riportate in: *B. Chiandotto e M. Gola, cit.*

[6] Le figure riportate consentono un'immediata percezione delle diverse valutazioni espresse dagli studenti frequentanti, sia in termini di livelli (dimensione delle figure), sia in termini di squilibri (regolarità delle figure) presenti a livello di facoltà.

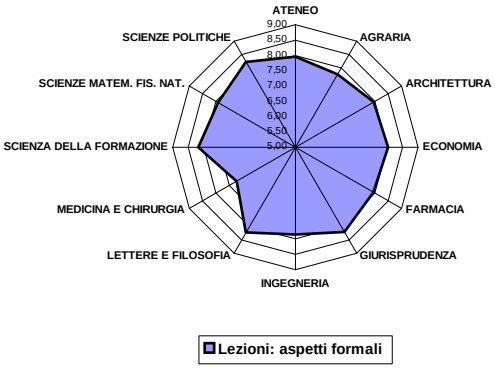
Fig. 4 - Università di Firenze: VALUTAZIONE DELLA DIDATTICA A.A. 2000/2001



All**'organizzazione della didattica** (quesito Q5) viene attribuita una valutazione **insufficiente** in tutte le facoltà dell'**ateneo**, particolarmente negativo è il giudizio per le facoltà di Lettere e Filosofia (5.99) e Architettura (5.98).

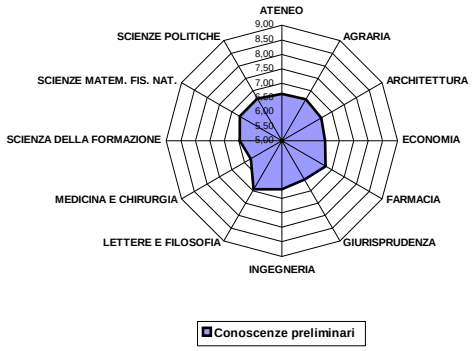
Per ciò che concerne i quesiti che interessano più direttamente i docenti, le opinioni degli studenti risultano del tutto gratificanti; infatti, le valutazioni si collocano, nella generalità dei casi, a livelli superiori a 7. I punteggi relativi agli aspetti più strettamente formali sono riportati nella Fig. 5.

Fig. 5 - Università di Firenze: VALUTAZIONE DELLA DIDATTICA A.A. 2000/2001



Per quanto concerne le **conoscenze preliminari** (quesito Q24, Fig. 6), la valutazione media espressa dagli studenti **non** raggiunge la **sufficienza** in nessuna realtà. Ma, se a livello di facoltà non si registrano diversificazioni eccessive (si passa da un minimo di 6.26 per Medicina e Chirurgia ad un massimo di 6.94 per Lettere e Filosofia), quando si scende a livello di corso di laurea/diploma emergono differenze più accentuate; si passa, infatti, da un valore minimo per Scienze e Tecnologie Alimentari pari a 5.98, ad un valore massimo di 7.22 per Produzioni Vegetali.

Fig. 6 - Università di Firenze: VALUTAZIONE DELLA DIDATTICA A.A. 2000/2001



Il quadro complessivo, articolato per Corsi di laurea/diploma e per Facoltà, "disegnato" dagli studenti che hanno frequentato le lezioni offerte dall'Ateneo Fiorentino nell'a.a. 2000/2001 è riportato nel Rapporto sopra segnalato.

VALUTAZIONE DELLA DIDATTICA A.A. 2001/02 - I° SEMESTRE[7]

Anche se l'avvio della riforma dei corsi di studio rende difficile il confronto con i risultati dello scorso A.A., si riportano nel prospetto che segue i dati, distinti per fase (semestre), delle ultime tre rilevazioni effettuate.

	Questionari	Corsi monitorati	Corsi valutabili	Tasso di risp.
A.A. 2000/2001				
Prima fase	31609	1172	1005	72.02%
Seconda fase	27721	1243	1054	56.08%
Totale	59330	2415	2059	62.64%
A.A. 2001/2002				
Prima fase	42422	1560	1317	61.69%

Si segnala che i dati si riferiscono al materiale pervenuto al 31 gennaio dell'anno corrente. Il quadro completo della rilevazione è riportato nel Rapporto: "*Valutazione della didattica da parte degli studenti frequentanti - a.a. 2001-2002, I° semestre*" consultabile nel sito dell'Università di Firenze più volte richiamato.

Anche per quanto concerne l'opinione degli studenti che hanno frequentato i corsi nel **I° semestre dell'a.a. 2001/02**, per ogni Corso di Laurea/Diploma, sono state effettuate le usuali elaborazioni, e cioè si è proceduto alla predisposizione:

[7] In alcune facoltà si tratta non del I° semestre ma del primo (quadrimestre) o dei primi due periodi (trimestri) di lezione.

1. della tabella riassuntiva degli insegnamenti monitorati (qualunque sia il numero di schede raccolte), con indicazione del numero di schede raccolte e del numero medio di frequentanti;
2. delle tabelle e grafici che sintetizzano la valutazione a livello di:
   - facoltà,
   - corso di studi,
   - singolo insegnamento.

Tali elaborati, l'archivio integrale dei dati della rilevazione e le schede cartacee sono state consegnate alle presidenze di facoltà o di corso di studi. Mancano le elaborazioni a livello aggregato che verranno effettuate al completamento della rilevazione per l'intera attività didattica svolta nel corrente a.a..

Le opinioni espresse dagli studenti che hanno frequentato i corsi nel primo semestre dell'anno accademico corrente hanno un contenuto informativo che può risultare particolarmente significativo, e ciò vale soprattutto per gli insegnamenti che hanno subito variazioni consistenti rispetto al passato. Al riguardo è auspicabile una prima ed attenta riflessione, sia a livello individuale sia a livello di corso di studi, sull'impatto che l'avvio dei nuovi ordinamenti ha avuto nella percezione degli studenti. Per valutare la potenzialità informativa e l'utilità del materiale raccolto sono state effettuate a titolo sperimentale ulteriori elaborazioni rispetto a quelle sopra illustrate; in particolare, si è proceduto all'analisi dell'opinione espressa dagli studenti della Facoltà di Ingegneria distinguendo gli studenti stessi in due gruppi: studenti iscritti al I° anno e studenti iscritti agli anni successivi e si è proceduto ad effettuare alcune elaborazioni di cui si riporta qui di seguito una breve sintesi. Le stesse elaborazioni sono state effettuate sulle opinioni espresse dagli studenti che hanno frequentato i corsi nel I° semestre dell'a.a. 2000/2001 e si è proceduto ad un confronto dei risultati conseguiti.

Nella Fig. 7 vengono riportati i profili medi relativi alle opinioni espresse dagli studenti iscritti al I° anno e quelle espresse dagli studenti iscritti agli anni successivi. Osservando la figura non emergono diversità molto rilevanti fra le opinioni espresse dai due gruppi di studenti.

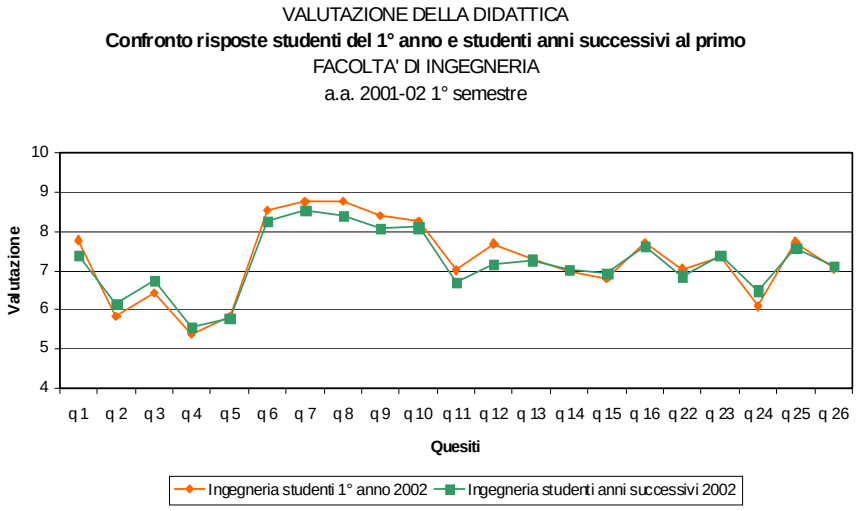


Fig. 7 - Università di Firenze: valutazione della didattica A.A. 2001/02

Si rileva comunque una maggiore attenzione, da parte dei titolari i corsi di insegnamento del primo anno, al rispetto degli orari delle lezioni e di ricevimento e nella scelta ed indicazione del

materiale didattico; inoltre, gli studenti di I° anno avvertono un maggior peso del carico didattico ed una più accentuata carenza riguardo alle conoscenze preliminari (quesito 24).

La diversità tra le opinioni dei due gruppi di studenti emerge invece in modo evidente se si considerano i profili medi a livello di singolo corso di studi; le opinioni espresse dagli studenti degli anni successivi disegnano dei profili medi che presentano una certa omogeneità (Fig. 8) mentre risultano decisamente diversi i profili relativi alle valutazioni espresse dagli studenti iscritti al I° anno (Fig. 8).

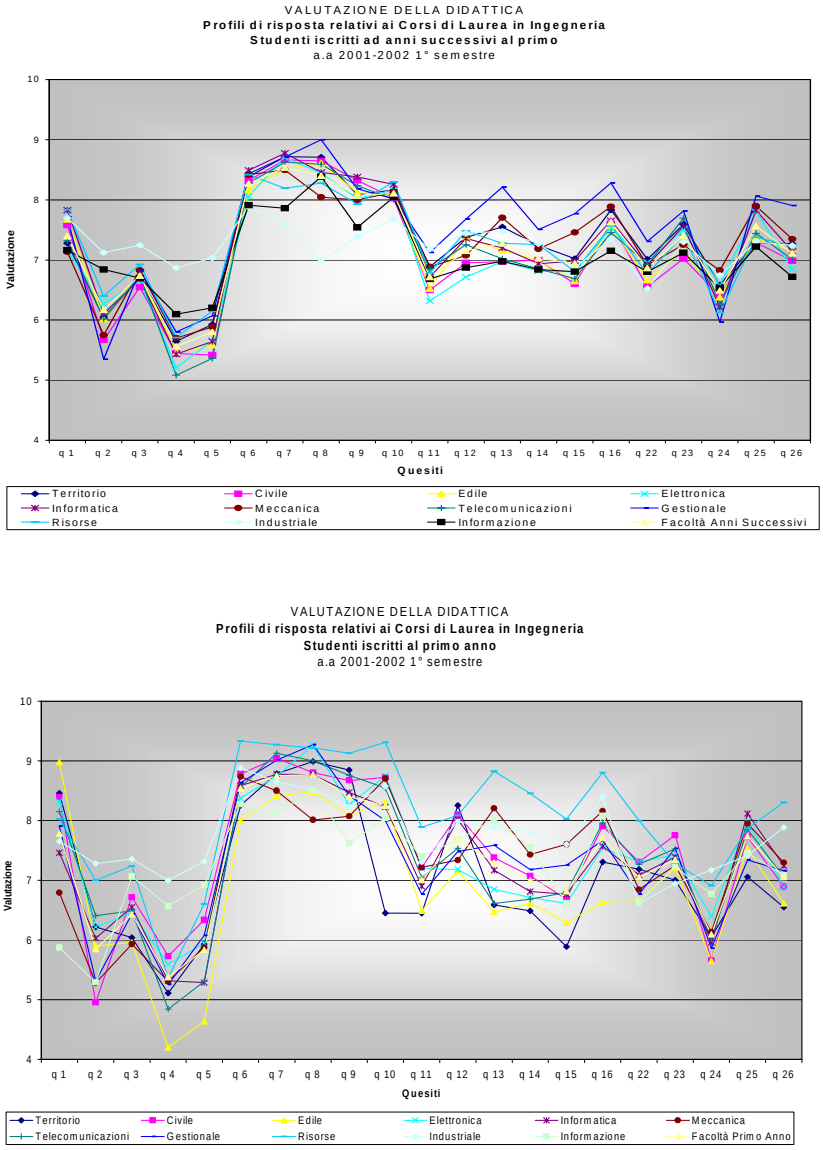


Fig. 8 - Università di Firenze: valutazione della didattica A.A. 2001/02

La maggiore variabilità nei profili di risposta per CdL da parte degli studenti del I° anno rispetto agli studenti degli anni successivi si rileva anche dalla analisi per gruppi. Ad un taglio dell'albero (Fig. 9) alla distanza di legame 1.3 per gli studenti degli anni successivi al primo, dall'unione dei gruppi rimangono fuori soltanto i CdL in Ing. Gestionale, Ing. Industriale ed Ing. Dell'Informazione.

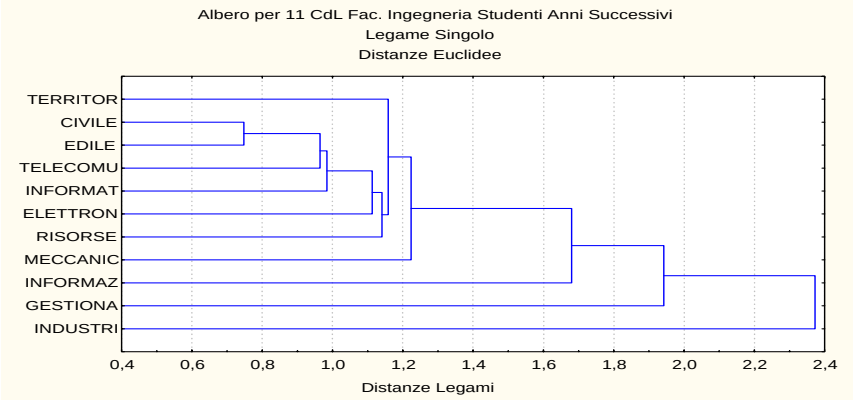


Fig. 9 - Università di Firenze: valutazione della didattica A.A. 2001/02

La stessa analisi dei gruppi applicata alle risposte fornite dagli studenti del I° anno fornisce un albero (Fig. 10) che se tagliato alla stessa distanza (1,3) non individua alcun gruppo. Il primo che viene a formarsi è costituito dai CdL in Telecomunicazioni ed Informatica ai quali si aggrega a distanza molto ravvicinata il CdL in Ing. Elettronica. I restanti CdL sono invece distanti fra di loro a conferma dell'alta variabilità e diversità di risposta dovuta molto verosimilmente alla:

1. eterogeneità della popolazione studentesca del 1° anno;
2. effetto della riforma universitaria;
3. diversità di risposta nell'offerta didattica e nella organizzazione degli insegnamenti previsti al I° anno dei vari Corsi di Laurea.

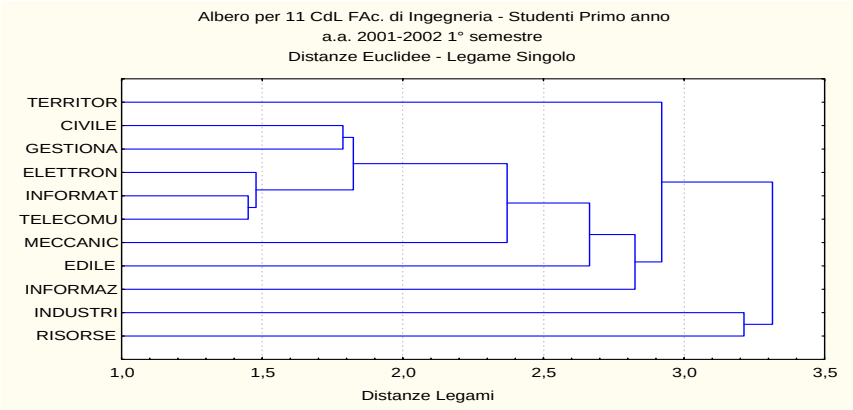


Fig. 10 - Università di Firenze: valutazione della didattica A.A. 2001/02

Ulteriori elementi informativi per valutare l'impatto della riforma sulla valutazione degli studenti si possono ragionevolmente desumere dal confronto (Fig. 11) tra le valutazioni espresse dagli studenti che hanno frequentato i corsi nel I° semestre dell'a.a. 2000/01 con quelle espresse dagli studenti che frequentato i corsi nel I° semestre dell'a.a. 2001/02.

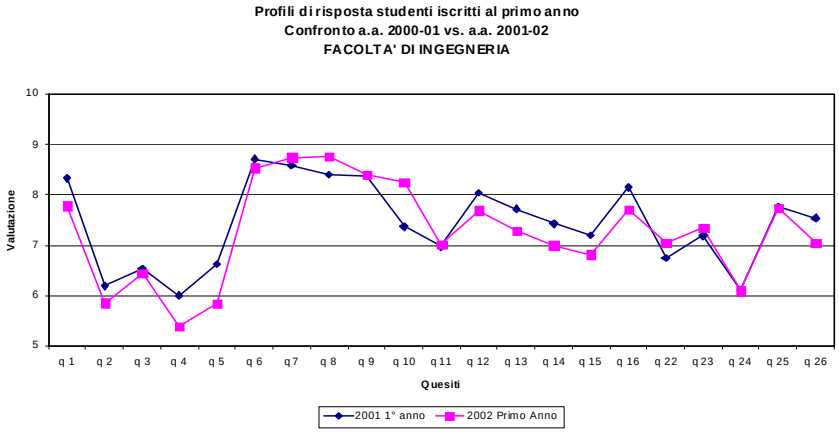


Fig. 11 - Università di Firenze: valutazione della didattica A.A. 2001/02

Osservando il grafico si rileva che le opinioni espresse dagli studenti che hanno frequentato nell'anno accademico corrente esprimono valutazioni più elevate, rispetto all'anno precedente, relativamente ad alcuni aspetti formali della didattica, per contro, risulta decisamente più negativa la valutazione relativa al carico di lavoro ed alla "capacita" didattica dei docenti. Verosimilmente**,** le modifiche apportate ai programmi e soprattutto il completamento degli stessi in un lasso di tempo più breve e la maggiore vicinanza fra le lezioni, può aver generato maggiori difficoltà e fornito minor tempo allo studente nel recepire le nozioni impartite. Lo stesso grado di soddisfazione globale espresso sull'insegnamento risente della valutazione inferiore data ai docenti e dell'accresciuto carico didattico: si ha una diminuzione della media di circa 0.5, valore abbastanza elevato se si considera il fatto che i profili di risposta della Facoltà di Ingegneria presentano storicamente caratteristiche di scarsa variabilità.

Scarsa variabilità che si riscontra nel diagramma che riporta i profili di risposta degli studenti degli anni successivi al primo (Fig. 12), praticamente sovrapposti con le le uniche eccezioni rappresentate dal quesito 5 riguardante l'organizzazione della didattica e dal quesito 24 sulle conoscenze preliminari, due quesiti correlati con la variazione dei programmi attuata. Ciò confermerebbe che una parte rilevante delle differenze riscontrate nel primo anno è da attribuire all'introduzione dei nuovi ordinamenti didattici.

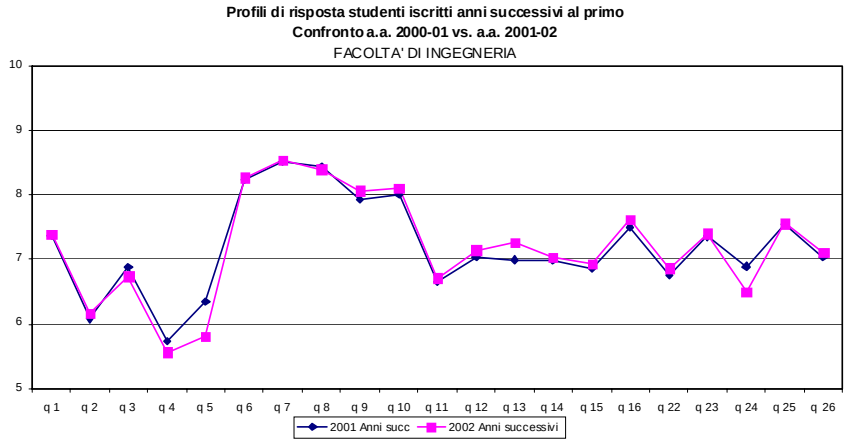


Fig. 12 - Università di Firenze: valutazione della didattica A.A. 2001/02

3.2 Sbocchi occupazionali

Le indagini sugli sbocchi occupazionali condotte dal Dipartimento di Statistica "G: Parenti", rientrano nell'ambito del più ampio progetto ALMALAUREA[8], ed hanno interessato tutti i giovani cha hanno conseguito il titolo negli anni solari 1998 e 1999[9].

Com'è noto, le rilevazione effettuate direttamente da ALMALAUREA coinvolgono i laureati della sola sessione estiva ad uno, due e tre anni dal conseguimento del titolo; l'estensione al collettivo dei laureati/diplomati di tutte le sessioni dell'intero anno solare è stata suggerita dell'esigenza di:

a) ottenere una comprensione globale e più puntuale di quella che è *la qualità del prodotto finito* dell'Ateneo fiorentino;
b) raggiungere un maggiore grado di attendibilità dei dati a livello di singolo corso di studi.

Obiettivi che specificano ulteriormente quelli propri dell'indagine ALMALAUREA che vengono sommariamente sotto elencati:

1. monitorare l'evoluzione nel tempo dei percorsi occupazionali dei laureati o diplomati;
2. approfondire la conoscenza dell'inserimento professionale dei laureati o diplomati indagati, ricorrendo anche alla documentazione di origine amministrativa in possesso dei vari Atenei e dello stesso Osservatorio Statistico;

---

[8] Il ***progetto ALMALAUREA***, la banca dati dei laureati e dei diplomati del sistema universitario italiano per il mondo del lavoro e delle professioni, nasce in via sperimentale nel 1993 per iniziativa dell'Osservatorio Statistico dell'Università di Bologna e comprende attualmente 28 atenei italiani.

a) Lo scopo principale di ALMALAUREA è quello di colmare la separazione esistente tra offerta di personale qualificato e domanda del medesimo da parte del mondo del lavoro e delle professioni, problema questo comune a gran parte delle Università italiane.
b) In particolare, ALMALAUREA si propone di:
c) analizzare l'efficacia interna delle strutture formative universitarie attraverso apposite indagini;
d) agevolare l'inserimento nel mondo del lavoro dei giovani laureati o diplomati del sistema universitario italiano, tramite la gestione di un'apposita banca dati denominata ALMALAUREA, costituita sia da informazioni di carattere amministrativo, sia dai dati raccolti tramite un apposito questionario compilato dai laureandi e diplomandi;
e) analizzare l'efficacia esterna delle proposte formative degli Atenei, attraverso una successiva indagine sugli sbocchi occupazionali dei laureati o diplomati della sola sessione estiva, ad uno, due e tre anni circa dal conseguimento del titolo;
f) aggiornare progressivamente i curricula formativo-professionali dei laureati o diplomati già presenti nella banca dati ALMALAUREA.

Per ulteriori informazioni, si può consultare il sito Internet www.almalaurea.it .

[9] Chi fosse interessato ad un approfondimento dei risultati delle analisi svolte può consultare i seguenti rapporti:

a. ***I laureati dell'Ateneo Fiorentino dell'anno 1997 – Profilo e sbocchi occupazionali.***

Il Rapporto, curato da **Giacomo Bulgarelli**, riguarda il profilo di tutti i laureati dell'anno solare 1997 e la condizione occupazionale dei laureati della sola sessione estiva.

b. ***I laureati dell'Ateneo Fiorentino dell'anno 1998 – Profilo e sbocchi occupazionali***.
Il Rapporto, curato da **Bruno Bertaccini**, consultabile sul sito internet: www.unifi.it/aut_dida/indexval.html, riguarda il profilo e la condizione occupazionale di tutti i laureati dell'anno solare 1998. Copia cartacea è disponibile a richiesta.

c. *Profilo e condizione occupazionale dei laureati dell'Ateneo Fiorentino ad uno, due e tre anni dal conseguimento del titolo,* curato da **Bruno Chiandotto,** consultabile sul sito internet: www.unifi.it/aut_dida/indexval.html, riguarda il profilo dei laureati nell'anno solare 2000 e la condizione occupazionale dei laureati nelle sessioni estive 1997, 1998 e 1999, ad uno due e tre anni dal conseguimento del titolo.

3. ottenere informazioni che siano comparabili con quelle raccolte da altre ricerche, svolte in ambiti territoriali più estesi;
4. conciliare la propensione a richiedere quante più notizie possibile con i vincoli di tempo e di costo della rilevazione.

Le sezioni in cui si articola il questionario utilizzato nella rilevazione sono:

a) la prima, diretta ad accertare se l'intervistato è soddisfatto della scelta universitaria fatta, i motivi che lo hanno indotto ad iscriversi, eventuali attività di qualificazione post-laurea e l'attuale posizione occupazionale;
b) la seconda, diretta ai laureati occupati, rilevando informazioni sulla posizione occupazionale, il ramo di attività, la collocazione geografica, le modalità di ottenimento del lavoro, la pertinenza del titolo e più in generale delle competenze acquisite all'università con l'attività svolta, il grado di soddisfazione di alcuni aspetti del lavoro;
c) la terza, rivolta a quei laureati che hanno lavorato dopo il conseguimento del titolo ma che al momento dell'intervista non sono occupati. Si raccolgono informazioni sulla passata posizione occupazionale, le modalità di ricerca del lavoro, la pertinenza dell'attività svolta con le competenze acquisite e il motivo d'interruzione;
d) la quarta, idonea a rilevare alcune informazioni relative ai giovani che non lavorano o che, comunque, cercano un nuovo lavoro. Si rilevano i motivi della "non ricerca", oppure il tipo di lavoro cercato e le azioni compiute in tale direzione;
e) l'ultima, dedicata a descrivere un breve quadro sulla famiglia d'origine dell'intervistato.

La comprensione e valutazione di questi aspetti non può prescindere dall'esame di quelle che sono le peculiarità dei laureati (caratteristiche socio-economiche, tipo di maturità, conoscenze linguistiche e informatiche, eventuali esperienze lavorative e/o di studi all'estero ecc. Occorre, inoltre, sottolineare che l'esperienza universitaria del contingente esaminato è sì caratterizzata da uno stesso punto di arrivo (anno di laurea), ma da diversi punti di partenza (anno di immatricolazione), definendo durate in cui si è certamente evoluta l'offerta formativa dell'Ateneo, ma si è anche trasformata la capacità d'assorbimento da parte del mercato del lavoro.

Queste considerazioni rendono articolato l'insieme delle fonti di acquisizione delle informazioni, che risulta complessivamente composto da:

a) l'archivio amministrativo d'Ateneo per una valutazione globale dei percorsi di studio (superiori e universitari) intrapresi da ogni laureato, in termini di tipo di disciplina , giudizio conseguito e durata;
b) il questionario ALMALAUREA compilato dai laureandi/diplomandi, fondamentale al fine di rilevarne le caratteristiche socio-economiche, il livello di alfabetizzazione informatica, le conoscenze linguistiche, le eventuali esperienze lavorative, ecc.;
c) il *Questionario sugli Sbocchi Occupazionali*, sviluppato dall'Osservatorio Statistico dell'Università di Bologna nell'ambito del progetto ALMALAUREA
d) Al questionario di base ALMALAUREA, il Dipartimento di Statistica ha provveduto ad aggiungere alcuni quesiti volti sia a migliorare la comprensione di tematiche inerenti la qualità del lavoro svolto da coloro che sono stati impegnati in almeno un'attività lavorativa dopo la laurea, sia ad indagare sulla soddisfazione complessiva di coloro che sono stati gli utenti dei vari percorsi didattici proposti dall'Ateneo.

L'assorbimento dei laureati da parte del mercato del lavoro è certamente la misura più significativa di efficacia esterna da non disgiungere, comunque, dall'utilizzo nell'attività lavorativa delle competenze acquisite all'università (altro "misuratore" di efficacia esterna estremamente utile nel processo di monitoraggio continuo dei processi formativi).

Sul primo aspetto le indagini svolte offrono elementi confortanti (Figg. 13,14 e 15); le percentuali di disoccupati, cioè di laureati che non lavorano e sono alla ricerca di un lavoro risultano molto contenute; in proposito si deve, inoltre, sottolineare che alcuni laureati non occupati

in cerca di lavoro non possono essere definiti disoccupati in senso stretto, poiché una parte non irrilevante di essi risulta impegnata in servizio di leva o in attività di formazione/qualificazione post-laurea/diploma. E questo è, ad esempio, certamente il caso di Giurisprudenza per quei laureati che intendono svolgere la professione di avvocato e per i quali è previsto un praticantato di non breve durata e per Medicina e Chirurgia.

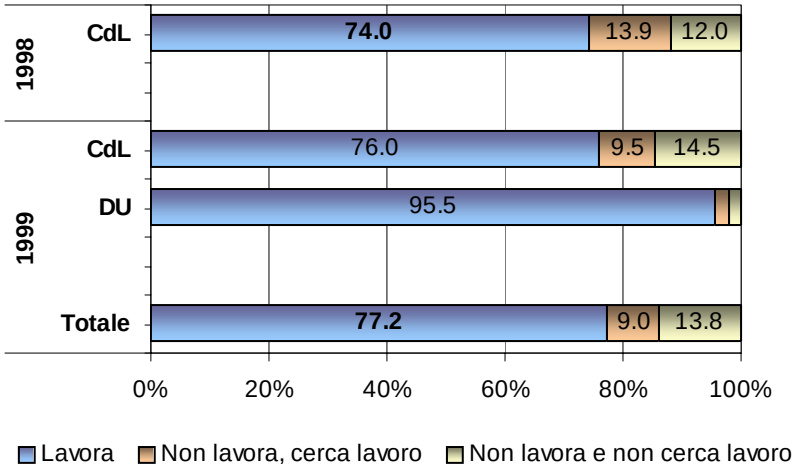


Fig. 13 - Università di Firenze: condizione occupazionale dei laureati/diplomati

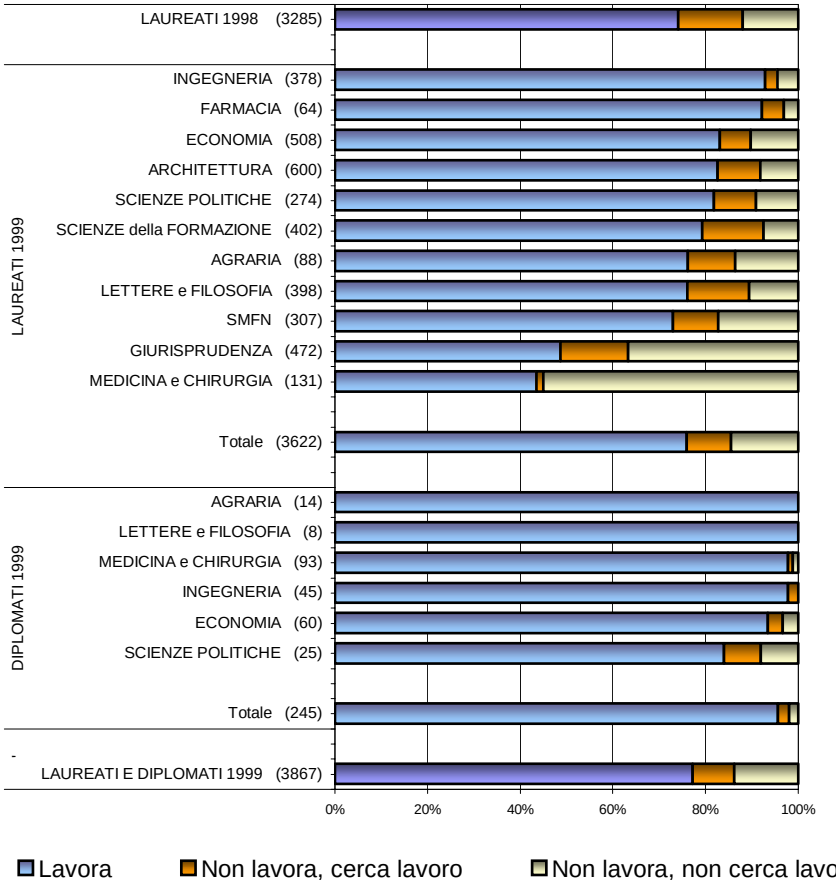


Fig. 14- Università di Firenze: condizione occupazionale dei laureati/diplomati

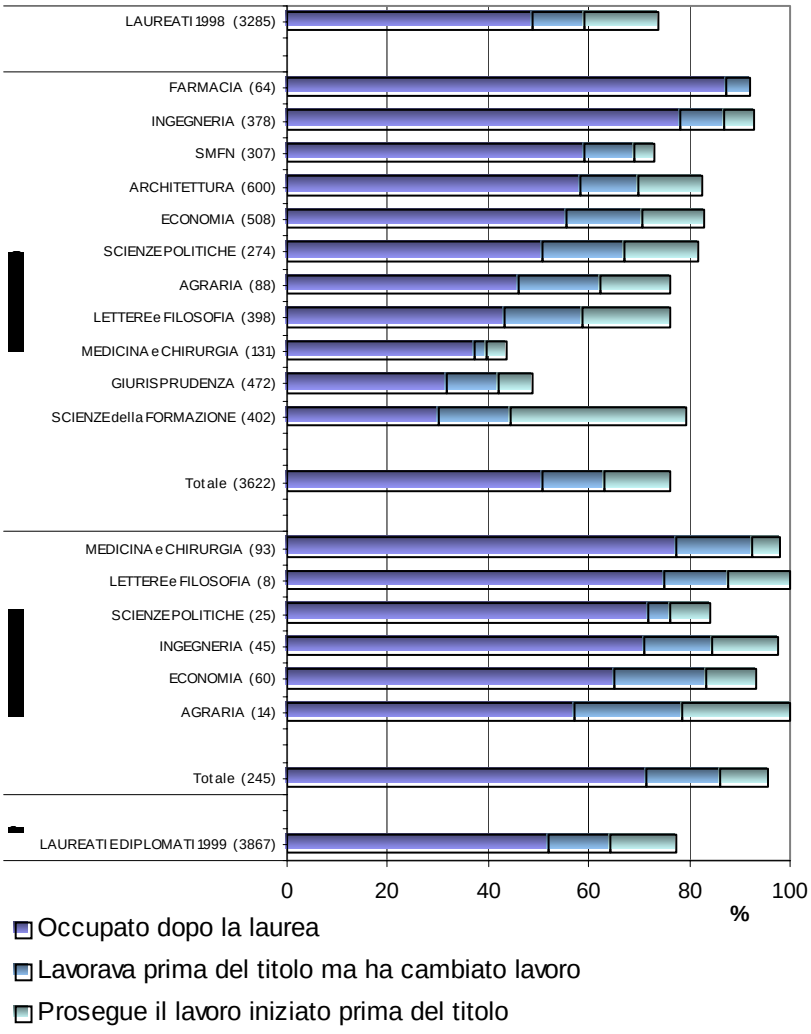


Fig. 15 - Università di Firenze: condizione occupazionale dei laureati/diplomati

Relativamente meno confortante è la risposta fornita al quesito relativo all'utilizzo delle competenze acquisite nei corsi universitari (Fig. 16). Infatti, quasi un terzo degli intervistati dichiara di utilizzare poco o per niente le competenze acquisite, con un massimo di utilizzo a Medicina e Chirurgia ed un minimo a Scienze Politiche. Ma anche in questo caso si deve tenere presente il tipo di formazione offerta dai due corsi di laurea: Medicina e Chirurgia offre una formazione molto professionalizzante il cui riscontro sul posto di lavoro è facile ed immediato; molto più problematica è l'interpretazione per il laureato in Scienze Politiche la cui attività lavorativa richiede spesso uno spettro molto ampio di competenze, complessivamente acquisite all'università ma, difficilmente attribuibili a singoli corsi seguiti.

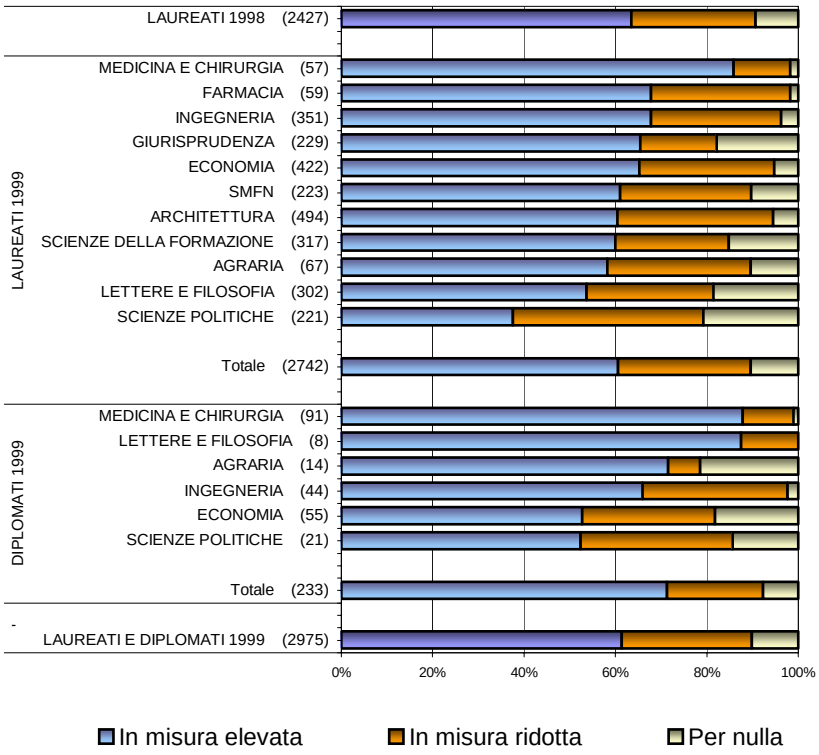


Fig. 16 - Università di Firenze: laureati/diplomati per grado di utilizzo delle competenze acquisite

Ad una conclusione non molto diversa si perviene se si fa riferimento all'efficacia[10] del titolo di studio conseguito (Fig. 17).

[10] Le cinque classi di "efficacia nel lavoro svolto" dai laureati/diplomati occupati sono:

1) *Molto efficace* – per gli occupati la cui laurea/diploma è richiesta per legge, e che utilizzano in maniera elevata le competenze universitarie acquisite;

2) *Efficace* – per gli occupati la cui laurea/diploma non è richiesta per legge, ma è di fatto necessaria o comunque utile, e che utilizzano in maniera elevata le competenze universitarie acquisite;

3) *Abbastanza efficace* – per gli occupati che utilizzano in maniera ridotta o non utilizzano per niente, le competenze universitarie acquisite, ma la cui laurea è richiesta per legge o, di fatto, necessaria;

4) *Poco efficace* – per gli occupati la cui laurea/diploma è utile in qualche senso, ma che non utilizzano per niente, o utilizzano in maniera ridotta le competenze universitarie acquisite;

5) *Per niente efficace* – per gli occupati la cui laurea/diploma non è utile né necessaria in alcun senso, e che non utilizzano per niente, o utilizzano in maniera ridotta, le competenze universitarie acquisite.

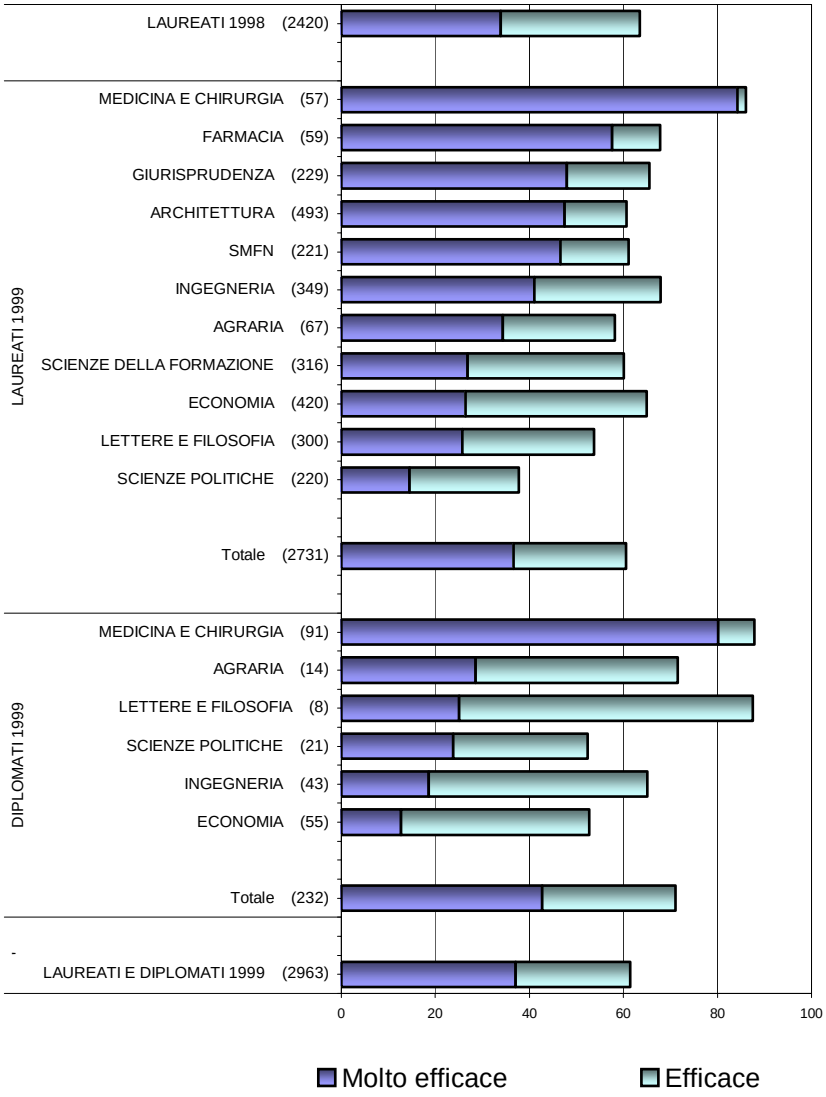


Fig. 17 - Università di Firenze: laureati/diplomati per efficacia del titolo conseguito

Un ulteriore elemento poco confortante è l'estrema difformità, nell'ambito dei diversi corsi di studio, della votazione agli esami e della votazione di laurea (Figg. 18 e 19). La grande variabilità dei criteri di valutazione adottati nei diversi contesti determina certamente sperequazioni ingiustificate nell'accesso al mondo del lavoro, soprattutto nelle situazioni (es. concorsi pubblici) in cui la votazione riportata si traduce in punteggi nella formazione delle graduatorie.

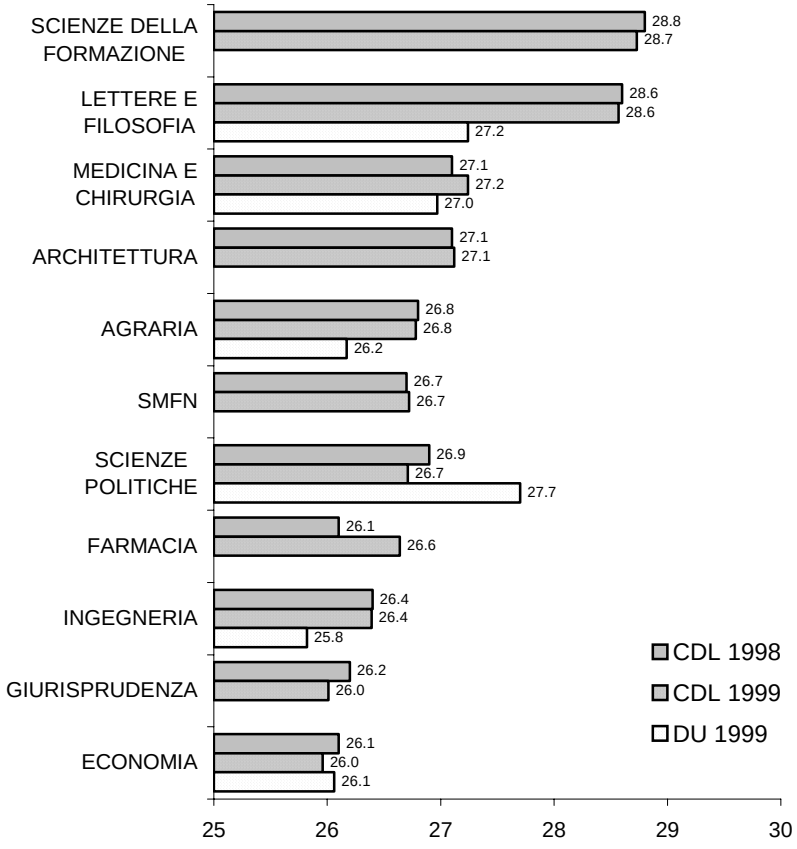


Fig. 18 - Università di Firenze: laureati/diplomati per voto medio agli esami

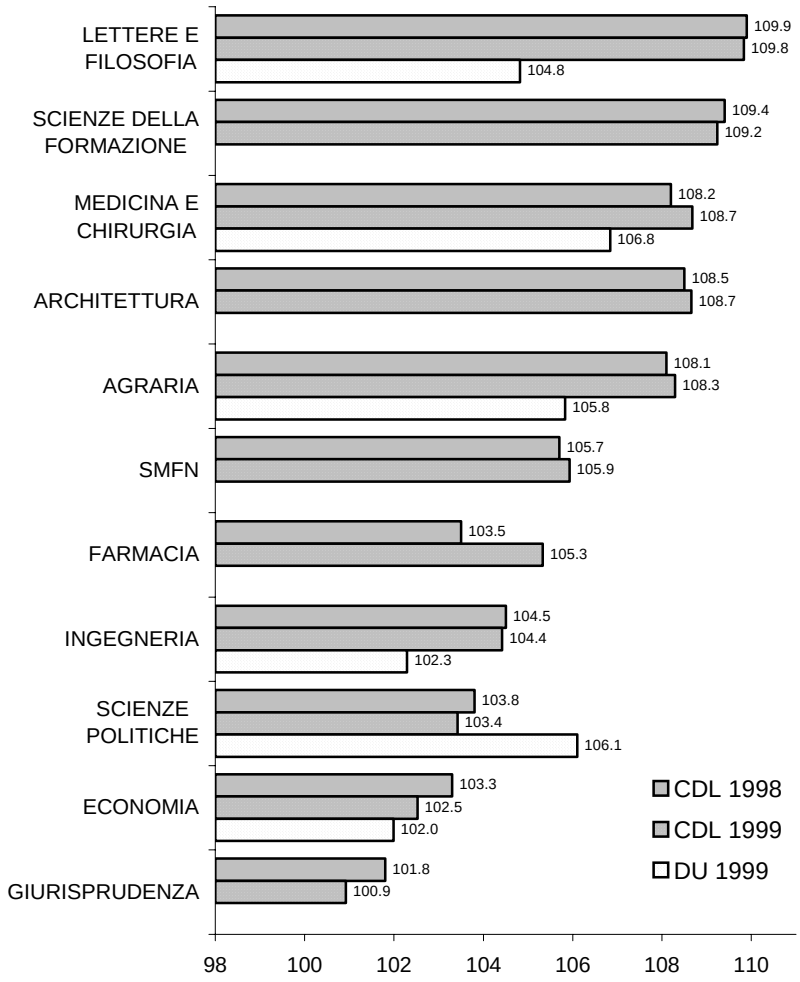


Fig. 19 - Università di Firenze: laureati/diplomati per voto medio alla laurea/diploma

Decisamente negativa si presenta la situazione sul fronte degli abbandoni[11] dei tempi di conseguimento del titolo, a sostegno di tale affermazione basta scorrere le figure riportate nelle pagine seguenti (Figg. 20 e 21); la gravità della situazione risulta ancora più accentuata se si considera la situazione presente negli altri Paesi dell'Unione Europea.

Maggiori informazioni, sulla condizione occupazionale dei laureati/diplomati dell'Ateneo Fiorentino nell'anno solare 1998 sono riportati nel Rapporto: "**Sbocchi occupazionali dei laureati dell'Ateneo fiorentino nell'anno solare 1998**"; ulteriori approfondimenti sono riportati nel Rapporto: "**Profilo e condizione occupazionale dei laureati dell'Ateneo fiorentino ad uno, due e tre anni dal conseguimento del titolo**". Entrambi i documenti sono consultabili sul sito http://www.unifi.it/aut_dida/indexval.html; entro breve termine, sempre sullo stesso sito, sarà consultabile anche il Rapporto sulla condizione occupazionale dei laureati/diplomati dell'anno solare 1999.

[11] Nell'ambito delle attività di valutazione e monitoraggio dei processi formativi è stato predisposto il Rapporto: "**Immatricolati dell'Ateneo Fiorentino dal 1980/81 al 1997/98: *esito degli studi***", che sarà entro breve termine consultabile sul sito dell'Università di Firenze.

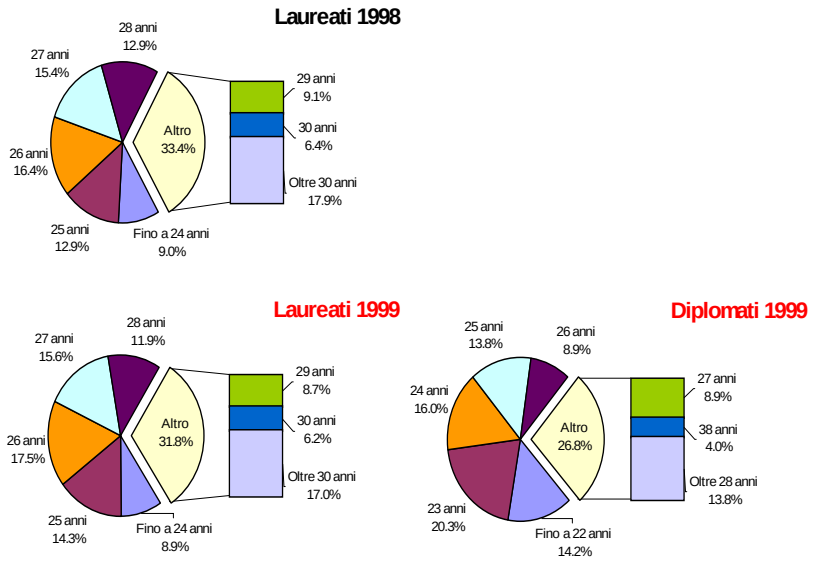
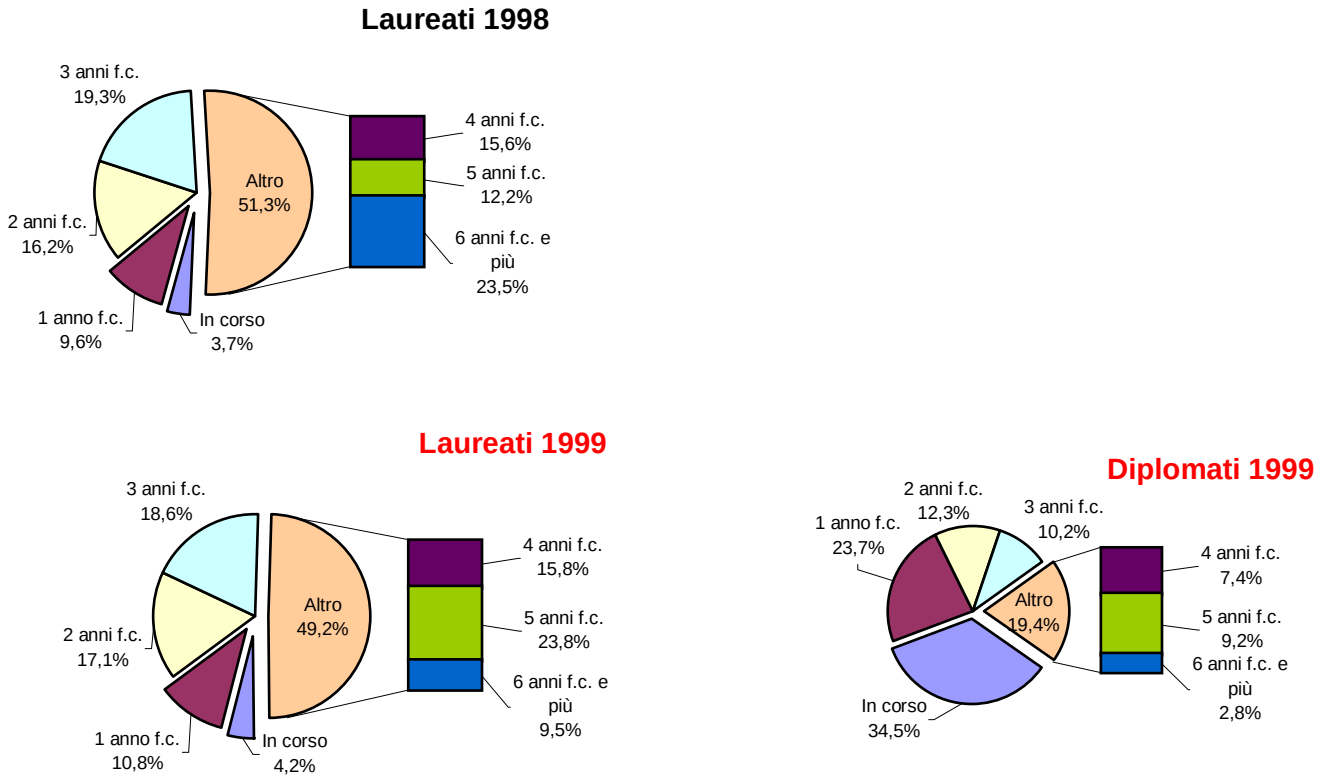


Fig. 19 – Università di Firenze: laureati/diplomati per tempi di conseguimento del titolo



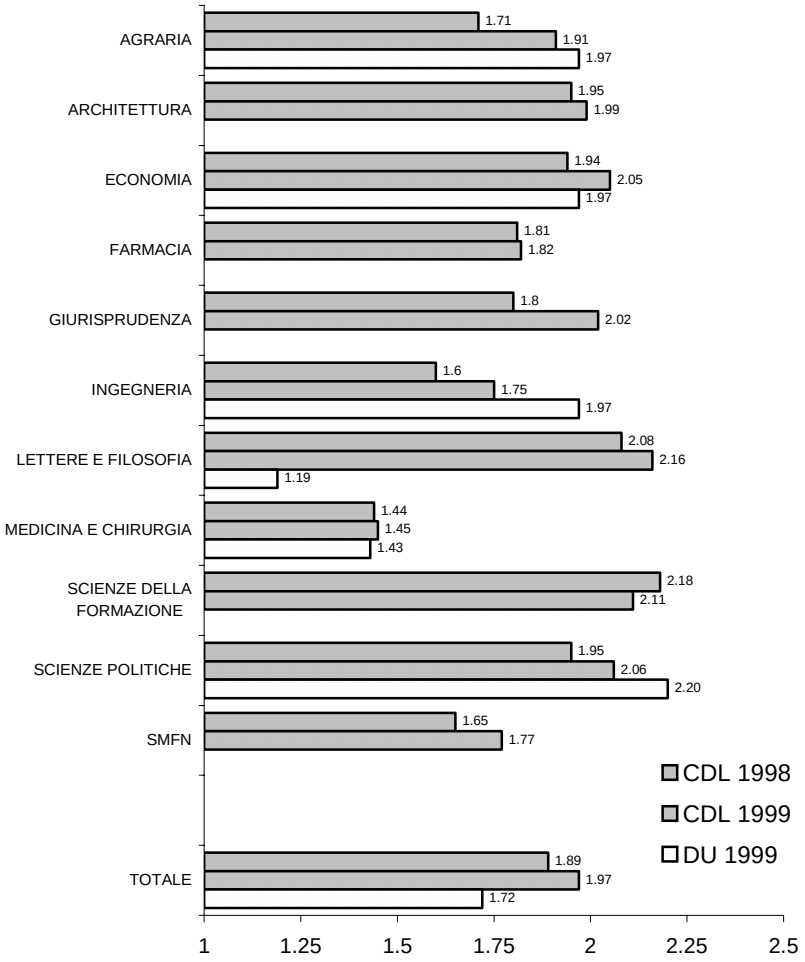


Fig. 21 - Università di Firenze: laureati/diplomati per indice di durata[12]

Bibliografia

Allais M. (1953) - "Le comportament de l'homme rationel devant le risque: critique des axiom et postulates de l'ecole americane" in "*Econometrica*", Vol. 21, n° 4, pagg. 503 – 546, 1953

Biggeri L. (1998) - Programmazione e valutazione dello sviluppo del sistema universitario, *Università Ricerca*, n. 2.

Biggeri L. (1999) - Autonomia e valutazione dell'insegnamento nel sistema universitario italiano,in Atti della giornata di studio "*L'insegnamento universitario in Italia*", Accademia dei Lincei, Roma.

Bini M. (1999) - *Valutazione dell'efficacia dell'istruzione universitaria nei riguardi del mercato del lavoro*, Osservatorio per la valutazione del Sistema Universitario, Roma.

Camerer C. F., Ho T.(1994) - "Violations of the betweenness axiom and nonlinearity in probability" in "*Journal of Risk and Uncertainty*", Vol. 8.

[12]L'indice di durata degli studi è definito dal rapporto tra la durata effettiva degli studi e quella legale del rispettivo corso di laurea/diploma.

Cave M., **Hanney S.**, **Henkel M.**, **Kogan M.** (1997) - *The use of Performance Indicators in Higher Education. The Challenge of Quality Movement*, Higher Education Policy Series 34, London.

Chiandotto B., **Gola M.** (1999) - *Questionario di base da utilizzare per l'attuazione di un programma per la valutazione della didattica da parte degli studenti*. Relazione finale del gruppo di ricerca. Osservatorio per la valutazione del Sistema Universitario. MURST, Roma.

Ellsberg D. (1961) - "Risk, ambiguity, and the Savage axioms" in "*Quarterly Journal of Economics*", Vol. 75.

Fishburn P. (1982) - "*The foundations of expected utility*", D. Reidel Publishing Company, Dordrecht, Holland.

Fishburn P. (1988 a) - "Normative theories of decision making under risk and under uncertainty" in Bell D.E., Raiffa H., Tversky A.- "*Decision Making*", Cambridge University Press, Cambridge.

Fishburn P. (1988 b) - "Expected utility: an anniversary and a new era" in "*Journal of Risk and* U*ncertainty*", n.1.

Fishburn P. (1989) - "Retrospective on the utility theory of Von Neumann and Morgenstern" in "*Journal of Risk and Uncertainty*", n. 2.

French S. (1986) - *Decision Theory: an Introduction to the Mathematics and Rationality, Ellis Horwood Ltd., Chichster.*

Gola M. (1998) - *La didattica universitaria e la sua valutazione,* Comitato paritetico per la didattica, Corso strumenti e metodologie per il formatore, Politecnico di Torino.

Gola M., **Squarzoni A.**, **Stefani E.**, **Tosi P.**, **Tronci M.** (2002) - Campus*one* - *La metodologia di valutazione della qualità dei processi e dei prodotti delle attività universitarie: Bozza di lavoro* -, CRUI, Roma.

Gori E., **Vittadini G.** (1999) - Qualità e valutazione dei servizi di pubblica utilità, Etas, Milano.

Herstein I.N., Milnor J. (1998) - "An axiomatic approach to measurable utility" in "*Econometrica*", Vol. 21.

Keeney R.L., Raiffa H. (1976) -"Decisions with multiple objectives: preferences and value trade-offs", Wiley & Sons, New York.

Keller L.R. (1992) - "Properties of utility theories and related empirical phenomena" in Edwards W., "*Utility theories: measurements and applications*", Kluwer Academic publishers, Boston, 1992

Modica L. e **Stefani E.** (1997) - *Valutazione delle attività didattiche. Le esperienze condotte dalla CRUI*, Pubblicazioni CRUI.

OECD (1997) - *Education at a Glance. Indicators. Centre for Educational Research and Innovation*.

Quiggin J. (1993) - "*Generalized expected utility (the rank-dependent model)*", Kluwer Academic Publishers, Dordrecht.

Rowley J. (1996) - Measuring Quality in Higher Education, *Quality in Higher Education*, vol. 2.

Université de Toulon (1998) - *Qualité totale et enseignement superieur*, Atti del Covegno tenutosi a Toulon il 3-4 settembre, 1998 organizzato in collaborazione con l'Università di Verona.

Savage L.J. (1972) - "*The foundations of statistics*", Dover Publications inc., New York.

Tversky A., Kahneman D. (1986) - "Rational choice and the framing of decisions" in "*Journal of Business*", Vol. 59.

Tversky A., **Kahneman D.** (1992) - "Advances in prospect theory: cumulative representation of uncertainty" in "*Journal of Risk and Uncertainty*", n. 5.

Von Neumann J., Morgenstern O. (1953) - "*Theory of games and economic behavior*", Princeton, Princeton University Press.

Wakker P., Deneffe D. (1996) - "Eliciting Von Neumann-Morgenstern utilities when probabilities are distorted or unknown" in "*Management Science*", Vol. 42, n. 8.

Nota – I documenti citati nel testo prodotti nell'ambito dell'attività dell' Osservatorio per la Valutazione del Sistema Universitario e del Comitato Nazionale per la Valutazione del Sistema Universitario sono consultabili sul sito: http://www.cnvsu.it; i documenti prodotti nell'ambito della CRUI sono consultabili sul sito: http://www.crui.it.

Decision oriented evaluation

Da: Mealli F., Chiandotto B.(2004) Decision oriented evaluation, Atti della XLII Riunione Scientifica della Società Italiana di Statistica, Bari 9-11 giugno 2004, pp. 209-220, Cleup, Padova.

Decision-oriented statistical evaluation

Valutazione orientata alle decisioni statistiche

Bruno Chiandotto, Fabrizia Mealli
Dipartimento di Statistica "G. Parenti"
Viale Morgagni, 59 Firenze
chiandot@ds.unifi.it, mealli@ds.unifi.it

Riassunto: Il lavoro tratta dei possibili diversi obiettivi di un'analisi valutativa, evidenziando i momenti in cui è necessario l'esplicito coinvolgimento del decisore (pubblico) con la propria struttura di preferenze. In tale contesto, si illustra il contributo che l'analisi statistica può fornire nella risoluzione di problemi decisionali basati su processi di valutazione (ex-ante ed ex-post) facendo specifici riferimenti ad alcune scelte proprie del sistema formativo universitario.

Keywords: decision analysis, program evaluation, utility theory, effectiveness.

1. Introduction

Over the last decades Evaluation has been the focus of many papers in various fields, ranging from Economics and Econometrics (Heckman et al., 1999), Statistics (Holland, 1986), to Psychology, Political Science (Meny and Thoenig, 1991), and Law. In a very broad sense, Evaluation is seen as a multidisciplinary and integrated approach to produce evidence on systems that can serve to a) understand how the systems work; b) to assess whether certain objectives have been met; c) to promote and support evidence-based policy choices; d) to draw "scenarios" to help choosing among alternative new policies.
So, a fundamental objective of evaluation is to provide policy makers with useful information for choosing policies. In this paper we are going to restrict the attention to the so-called impact Evaluation of policies from a statistical point of view, bringing to the discussion Evaluation examples regarding Educational systems.
We will focus on the role of the decision maker in guiding the ex-post analysis by defining the subject of the evaluation as well as in analyzing the result of the evaluation. Formal decision analysis is also, and especially, needed in the ex-ante evaluation of alternative interventions, because evaluation has to take into account the policymaker's attitude towards inequality and uncertainty in an appropriate way.
We will show that a proper decision-oriented statistical evaluation can be achieved not only if appropriate statistical tools are implemented for the production of evidence from available data, but also if those are used together with tools that enable to elicit individuals and policy makers preferences, in an interdisciplinary approach.
In particular, in the paper we would like to focus on specific questions which characterize impact evaluation analysis. Subject to impact evaluation are usually specific policies or interventions aimed at modifying in a certain desired way the condition or behavior of participants.
Within Educational system there are a variety of aspects that can be subject to evaluation in a very wide sense. Examples include the evaluation of the quality of

teaching, the evaluation of Educational systems' (internal and external) efficiency and effectiveness (Bini and Chiandotto, 2003). Some typical examples of more specific policies subject to impact evaluation are student aid policies, such as University grants, whose main objective is to guarantee higher education to motivated students irrespective of their income; the comparison of different teaching methods, whose main objective is to identify the method that is most effective in helping the students perform well. Another problem that can be cast in this framework is that of the evaluation of student orientation policies, whose objective is to provide students with information that can help them choose the career that mostly matches their preferences and attitudes, thus reducing drop out rates and duration of the studies and so improving the efficiency of the whole system.

All these problems of empirical research can be described as problems of causal inference, because they seek to estimate the effects of certain policies/interventions. It is usually a very difficult and often controversial analysis (Holland and Rubin, 1983).

Problems of causal inference involve "what if " statements and thus counterfactual outcomes and are usually motivated by policy concerns; they can be "translated" into a treatment-control situation typical of the experimental framework.

The fact the treatment is not assigned at random[1] prevents the possibility of comparing "treated" and "untreated" individuals, as such comparison is very unlikely to have a causal interpretation, because the two groups are different irrespective of their treatment status.

A growing list of papers in the statistical and economic literature (e.g., Rosenbaum and Rubin, 1983; Manski, 1990; Angrist et al., 1996; Heckman and Smith, 1998; Pearl, 2000) have tried to identify causal effects of interventions from observational, i.e. non experimental, studies, although using the conceptual framework of randomised experiments and the so-called *potential outcomes* approach, that allows to translate causal problems into a statistical model. Some identification strategies for causal effects can be found even in non experimental settings: as we will see, data alone do not suffice to identify treatment effects, but suitable assumptions, possibly containing prior information available to the researchers, are always needed.

The potential outcomes approach to causal inference, that we will use in the sequel, is rooted in the statistical work on randomized experiments by Fisher and Neyman, as extended by Rubin (1974, 1990) and subsequently by others. This perspective was called *Rubin's Causal Model* (RCM) by Holland (1986) because it views causal inference as a problem of missing data with explicit mathematical modeling of the assignment mechanism as a process for revealing the observed data[2].

The starting essential feature of this approach is to define a causal effect as the comparison of the potential outcomes on the same unit measured at the same time.

---

[1] This is the usual case in the social sciences were only observational data are available. Social experiments are still relatively rare, although they have been designed and implemented especially in North America: despite the ethical problems they may arise, they could be easily used for example to assess the effectiveness of different orientation policies, by selecting at random small group of students.

[2] Another approach to causal modeling is based on the theory of causal graphs. Recent development in this field, which has many links with structural equation modeling, has provided a formal theory for evaluating causal effects (Pearl, 2000). The use of graphs provides a tool for integrating statistical and subject-matter information: it facilitates to explicate the assumptions underlying the causal model and to decide whether the assumptions are sufficient for obtaining estimates of causal effects from observed quantities. The two formalizations have many links (see for example Pearl (2000) for a discussion on this).

Consider a binary treatment: $Y(0)$ is the value of the outcome variable Y if the unit is exposed to treatment $T = 0$, and $Y(1)$ is the value of Y if exposed to treatment $T = 1$; this same notation can be extended to the case treatment levels are more than two. Only one of these potential outcomes can be observed, the one corresponding to the treatment the unit received, yet causal effects are defined by their comparison, e.g., $Y(1) - Y(0)$ or $Y(1)/Y(0)$. Thus, causal inference becomes a problem of inference with missing data, because only one of the two potential outcomes is observed, the other one becoming a counterfactual. Also, causal effects defined as comparisons of potential outcomes force to think of a causal effect of a treatment only if the treatment status can be thought of as having been manipulated in some way. The focus of impact analysis is usually that of estimating a feature of the distribution of $Y(1)$-$Y(0)$, e.g., the average treatment effect $ATE = E(Y(1) - Y(0))$ or the average treatment effect for subpopulations of individuals defined by the value of some variable, most notably the subpopulation of the treated individuals: $ATT = E(Y(1) - Y(0) \mid T = 1)$. Alternatively, the focus can involve only comparisons of the two marginal distributions of $Y(0)$ and $Y(1)$.
In the following sections we will show how an Evaluation process can be thought of as a decision problem, where the results of a causal analysis are used to judge the effects of some interventions. We will distinguish ex-post (section 2) from ex-ante evaluation of policy interventions that have not yet been implemented (section 3).

2. Evaluation as a decision problem

Even in the preliminary step of the analysis, formal decision making tools are needed to guide the Evaluation process. First of all, the decision maker should inform the analyst on which outcome variable to choose, i.e., he should clearly identify what is the condition or behavior that the intervention aimed at modifying in a desired direction. This is usually not an easy task; the analyst usually has to cope with laws, where the policy objectives are only vaguely specified. More than formal methods for preferences elicitation, such as those for eliciting utility functions, there is the need of questions and explicit answers by decision makers. With this respect, policy makers can be aided, as usually done for the elicitation (often hierarchical) of criteria to be included in multi-attribute utility functions and multi-criteria decision making (Bana e Costa, 2001, Keeney et al., 1990), where an informative discussion between the analyst and the decision maker, can help both find the most appropriate outcome variables to focus on.
Secondly, restricting the analysis to average treatment effects (ATE), even if sometimes considered also for subsets of individuals defined on observed pre-treatment characteristics, does not usually reflect what the decision maker would like to know.
In order to illustrate this point, let us consider the following Higher Education related example.
The agency that administrates students' aids for the University of Florence (Azienda per il Diritto allo Studio) offers every year some scholarships to eligible freshmen, where eligibility is based on merit and economic needs. The main objective of this intervention is to give equal opportunity to achieve higher education to motivated students irrespective of their income. As grants' award is usually constrained by a fixed budget, not all eligible students who apply for a grant can receive one: so the applicants are ranked on the basis of an economic indicator S, that depends in a deterministic way on family income, real and personal assets and family structure, and only a varying

percentage of high ranking applicants with S lower then s_0 receive a grant. In agreement with the objective of the intervention, the decision maker should first be interested in knowing whether those receiving a grant can reach the same education level as students that are equally motivated but belong to richer families. Motivation is usually a variable that is difficult to measure, so the analysis can be diverted to explore whether the grant is an effective tool to prevent good students from lower-income families from dropping out of higher education and to achieve a degree in a reasonable time period: indeed, low income students are characterized by a high drop out rate, as well as a long duration of their studies, because of their usual need to work while studying.

The analyst must first answer the following question: under which conditions are the distributions of $Y(1)$ and $Y(0)$ identified? In this case $Y(0)$ and $Y(1)$ are either the binary variable taking value 1 if the student drops out and 0 otherwise or the continuous variable representing time to degree achievement, T is the treatment indicator being 1 if the student receives the grant and 0 otherwise. In order to answer the question one must make assumptions on the assignment mechanism. The assignment mechanism is a stochastic rule for assigning treatments to units and thereby for revealing $Y(0)$ or $Y(1)$ for each unit: it can depend on other measurements, i.e. $P(T = 1|Y(0), Y(1), X)$; if these other measurements are observed values, then the assignment mechanism is ignorable; if, given observed values, it involves missing values, possibly even missing Y's, then it is non ignorable. Unconfoundedness is a special case of ignorable missing mechanism and holds when $P(T = 1|Y(0), Y(1), X) = P(T = 1| X)$, where X is a vector of observed pre-treatment variables; basically this assumption says that exposure to treatment is random within cells defined by the variables X; matching methods rely on this assumption. Although very strong, the plausibility of this assumptions heavily relies on the amount and on the quality of the information on the individuals contained in X.

In the case of the evaluation of grants' effect, the study is a so-called sharp regression-discontinuity design (Thistelthwaite and Campbell, 1960), where T is known to depend in a deterministic way on some observable variable S (the economic indicator in our case) $T= f(S)$, where S takes on a continuum of values and the points s_0 where the function f is discontinuous are assumed to be known. In this case the following identifying condition can be exploited: $P(T = 1|Y(0), Y(1), X, S = s_0) = P(T = 1| X, S = s_0)$. The intuition is that, in the neighborhood of s_0, the grant status T and the potential outcomes are independent, so that the sample of individuals within a very small interval around each cutoff point, who have essentially the same S value, are very similar to a randomized experiment at the cutoff value (Mealli and Rampichini, 2001). This condition, if accepted to be plausible, allows to identify the distribution of the two potential outcomes for students having an economic indicator value equal to the cutoff point s_0 (Hahn et al., 2000). Note that, whatever model we apply to the data, evidence on the causal effects is contained only around the cut-off: if we extend the evidence for other subgroups of people (for example having different values of S), we make an improper use of extrapolation, and the reliability of such causal effects' estimates cannot be assessed[3]. Usually, only the ATE at the cut-off point is estimated, i.e., in this case, $E(Y(1)-Y(0)|X, S= s_0)$; as pointed out by Dehejia (2004) and others, focusing the attention on ATE does not always solve the decision problem of assessing the two treatments.

[3] The same argument can be applied when regression models for the estimation of causal effects under unconfoundedness are used, without previously checking if the distribution of the covariates in the two groups of treatments overlap to a large extent or not.

An appropriate alternative is to compare the distribution of the potential outcomes by formal decision theory methods. The distributions in the outcome variables space should contain all the uncertainty from the model as well as from the parameter estimation: according to the Bayesian approach, a distribution with these characteristics is the posterior predictive distribution of $Y(0)$ and $Y(1)$, i.e., $f(Y(0)|\text{data})$ and $f(Y(1)|\text{data})$. In order to derive the posterior predictive distribution one has to specify a statistical model for the outcome variables (see for a very simple example Gelman et al., 1995, ch. 7) $f(Y(t) \mid \theta)$, and elicit a priori distributions of the parameters characterizing the model, $\pi(\theta)$, so that $f(Y(t) \mid data) = \int f(Y(t) \mid \theta)\pi(\theta \mid data)d\theta$. The model must be consistent with the assumptions on the assignment mechanism that allow identification of causal parameters, because the model will be able to capture uncertainty correctly only if correctly specified[4]. Y can be multivariate (e.g., it may include drop-out and academic performance) the analyst must be very careful in specifying the model, as some variables in Y might be "intermediate" with respect to others so that a simple multivariate model might not be appropriated, as will be more clear in subsequent examples.

In order to judge whether treatment is effective for individuals with certain characteristics X, one should first elicit decision makers preferences with respect to Y, i.e., specify a utility function, that expresses policy makers risk attitudes and preferences over uncertain values of Y. Recent developments in methods for utility elicitation can be used and applied (Fennema and Van Assen, 1999). When Y is multivariate, multi-attribute utility functions should be elicit, with explicit consideration of the trade-offs between attributes (see French, 1988, for a deep discussion on these issues, and on the implicit assumptions of commonly used additive utility functions).

Once a utility function is elicited, treatments should then be judged comparing $E(U(Y(1)))$ to $E(U(Y(0)))$. As it is well known, and shown in numerous examples, in cases where ATE cannot "distinguish" between the two treatments, expected utility comparison can lead to clearcut decisions.

Another more informative alternative to ATE is looking at other feature of the distribution of potential outcome differences, for example, the decision maker might be interested in knowing the proportion of individuals who have benefited from the intervention, i.e., $P(Y(1)\text{-}Y(0)>0)$; equal values of ATE can be generated by very different values of this proportion. The estimation of this proportion requires the knowledge of the *joint* distribution of potential outcomes, that depends on the correlation structure of potential outcomes on which we do not have direct evidence.

The data contain only limited information on the joint distribution but, under suitable assumptions, bounds on this quantity can be derived, as explained with the next example.

Consider again the problem of evaluating the impact of university grants on time to graduation; this variable is not observed on everybody. Indeed duration is only observed and defined for students who do not drop-out: thus a properly defined causal effect on duration is the effect of the grant on the subset of students who would not drop-out irrespective of receiving a grant or not. The analysis should first assess the effect on

[4] Recent advances in Bayesian model choice and model checking have also been applied in the specific context of causal inference; see, among others, Hirano et al. (2000) and Mealli et al. (2004). Posterior predictive distributions can be easily approximated by simulations if parameters' posterior distributions are obtained via MCMC methods.

drop-out and then, once the group of students who do not drop-out irrespective of receiving the grant is *identified*, the effect on duration can be assessed. In order to simultaneously consider the two phenomena, we can adopt the principal strata framework (Frangakis and Rubin, 1999), that allows to properly account for intermediate outcomes, such as drop-out in this case, and unobservables.
The *intermediate variable* $D_i(T_i)$ is 1 or 0 if student i drops out or does not drop out , when receiving treatment T_i .The *response variable* $Y_i(T_i)$ the duration of the studies of student i when receiving treatment T_i. Since, again, for each individual the treatment variable assumes a single value, for every post-treatment variable only one of the two potential versions can be observed; because both the treatment variable and the intermediate variable are dichotomous, four *principal strata* can be defined in the following way through the latent variable L_i: L_i = 'DD' (Dropout, Dropout) if $D_i(1)=1$ and $D_i(0)=1$; L_i = 'DN' (Dropout, No dropout) if $D_i(1)=1$ and $D_i(0)=0$; L_i = 'ND' (No dropout, Dropout) if $D_i(1)=0$ and $D_i(0)=1$; L_i = 'NN' (No dropout, No dropout) if $D_i(1)=0$ and $D_i(0)=0$. The principal strata are latent classes and the data allow to put bounds or even estimate, under certain assumptions, the probability that a given student belongs to a certain latent class. If the purpose is to evaluate the effect of the grant on drop-out, note that the probability of belonging to a principal strata answers the question about the proportion of individuals who have benefited from the intervention, that is $P(L_i = \text{'ND'}) = \pi_{ND:i}$. Also note that the probabilities of the principal strata, $\pi_{ND:i}$, $\pi_{DD:i}$, $\pi_{DN:i}$, $\pi_{NN:i}$,are interesting per se, as they throw light on how the average effect has been achieved. In fact, the average causal effect on dropout is

$$P(D_i(1)=1) - P(D_i(0)=1) = (\pi_{DD:i} + \pi_{DN:i}) - (\pi_{ND:i} + \pi_{DD:i}) = \pi_{DN:i} - \pi_{ND:i}$$

Therefore, similar values of ATE can be produced by very different proportions of students who benefit or are "harmed" by the interventions. Decision maker preferences on these proportions can be elicited in order to judge an intervention.
We can put bounds on the probabilities for the principal strata following the procedure used in Zhang and Rubin (2004). The bounds can be sharpened by making some suitable assumptions (see also Grilli and Mealli, 2004).
If the purpose is to evaluate the impact of the grant on the duration of the studies, the response variable Y is defined only for the non dropouts, so the causal effect $Y_i(1)$-$Y_i(0)$ is properly defined only for the NN stratum, i.e. the students who graduate with or without the grant. Again, bounds on this causal effect can be derived; more informative results, can be obtained by specifying suitable parametric models that include some identifying conditions, to be fitted with likelihood or Bayesian methods.
Note that these issues just presented are usually neglected in the literature of measuring the amount of human capital, by means of multilevel models. Also, the same approach can be applied to jointly analyze internal and external effectiveness (with respect to the labor market) of different degree programs as in Grilli and Mealli (2004).

3. Decision-oriented (ex-ante) evaluation

The usual goal of ex-ante evaluation is to perform an evaluations of policy interventions that have not yet been implemented. Such policies may well be variations of previously adopted interventions. For example, as far as the problem of distributing university

grants, the policy maker might be interested in changing the criteria that identify the eligible students, or change the amount of the grants, or change the tool of student aid policies, for example offering *goods* (accommodation, meals, books) instead of a money transfer.
Interventions should be evaluated using social welfare functions reflecting the individual preferences, embedded in their utility functions, as well as the policy maker preferences, e.g. his degree of inequality aversion.
Such issues have been pointed out and analyzed by economists who raised the issue of making informed public choices, but are usually neglected in the evaluation literature. Exceptions include Manski (1990) who develops nonparametric bounds for the expected welfare from different post-evaluation assignment rules; Mealli and Pudney (1994) that analyze the cost-effectiveness of youth training funding for a wide range of the social evaluation of successful employment, also deriving bounds on optimal policy welfare; and Heckman and Smith (1998) who consider the evaluation of various welfare functions, some of which requiring information on the joint distribution of potential outcomes, as seen in the previous section.
Consider the following decision problem faced by a policy maker, which is a relatively standard model of welfare Economics: let the potential outcome(s) in the presence of treatment (policy) k for individual i be Y_{ki}, and let the preference of person i for outcome Y be denoted by $U_i(Y)$. It is postulated that a social welfare function exists[5] that is defined over utilities of the N individuals of a society or a target population (e.g. all the students enrolled in a given University):

$$W(k) = W(U_1(Y_{k1}), U_2(Y_{k2}), \ldots, U_N(Y_{kN})) \quad (1)$$

A choice based on such a function would choose the treatment k with the highest value of W or, in case of uncertainty, the highest expected welfare[6].
There are several issues that need to be solved before the above criterion can become operational; because of the difficulty and subjectivity of some of them, social welfare functions usually do not govern public decision making. Nevertheless, we believe that in more restricted problems (such as that of the distribution of grants among students) they can help making informed decision. These issues are the topic of next subsections.

3.1 How do we elicit preferences?

In principle the same procedures and tools, quoted in the previous section, could be applied to elicit individuals preferences U_i. These include also recent advances in Multi Criteria Decision Making such as in Bana e Costa and Chagas (2004), with interactive tools as aid for preferences elicitations, that would very well match the needs of some program participants, such as students. Despite that, we believe that in most cases it

[5] Here we do not discuss the issue related to Arrow's impossibility theorem, as well as to the problem of making interpersonal comparison. For a formal treatment of the issue, where assumptions on the decision maker preferences are made explicit, see Deaton and Muellbauer (1980) and French (1988).
[6] As in the ex-post analysis, we could consider here also criteria that involve the knowledge of the joint distribution of potential outcome; for example policy k could be preferred to policy j if the proportion of people benefiting from k is greater than 0.5. We limit our attention on criteria that requires, ex-ante, only the marginal distributions of outcomes under different policies (Dehejia, 2004).

would be almost impossible to elicit every individual *i*'s preferences, because of timing and costs. We can still think at possible solutions to this problem: for example we could elicit the preference of the modal, median or average individual, and use only his utility function. Alternatively we could assume preferences depend on observed characteristics, and use the estimation of discrete choice models from revealed preferences and *translate* them into a multiattribute utility function $U_i(Y)$.
The choice of a functional form for W and the specification of its arguments is usually rather complicated. Let us first assume (1) is indeed describing the policy maker decision problem; then, once the individual utility functions have been quantified, the choice of how to aggregate them depends on the decision-maker attitudes toward inequality. The following simple functional form can be used, that encompasses, depending on the value of ε, the inequality-neutral utilitarian preferences (ε = 0), different degrees of inequality aversion (ε > 0) to the leximin or Rawlsian preferences ($\varepsilon = \infty$) that embodies an extreme concerns with the worst-off individual in the group (Deaton and Muelbauer, 1980):

$$W(k) = W(U_1(Y_{k1}), U_2(Y_{k2}), \ldots, U_N(Y_{kN})) = \frac{\sum_h \left(U_h(Y_{kh})\right)^{1-\varepsilon}}{1-\varepsilon}.$$

Yet, policy makers preferences do not always match those of the individuals: decision makers may want to exclude (include) some variables in the Y vector that might be (not) relevant for the students. Also, even if the Y may be the same, the overall objective of the decision maker might not involve individual preferences and risk attitudes. If individual tastes are not explicitly taken into account but only the opportunity sets faced by different individuals with whom we are concerned, then to call W a social welfare function is misleading, since it is only an aggregate weighting functions:

$$W(k) = W(Y_{k1}, Y_{k2}, \ldots, Y_{kN})$$

What to include in Y and how to aggregate the Y's depends on the explicit objective of the decision maker. In the University grant example, the objective can be that of maximizing the number of students from low income families who get a degree, or maximizing the number of students who get a degree (L) *and* reducing the time to get a degree (D); in this second case W could be specified as follows $Y_{kh} = \{D_{kh}, L_{kh}\}$:

$$W(k) = \sum_h \left(\alpha D_{kh} + L_{kh}\right)$$

where α represents the trade-off between the two attributes.
In the presence of limited resources, the aim might be, not only that of ex-ante evaluating different interventions, but that of finding the "best" way of distributing resources. In this case the decision maker might be interested in exploiting treatment effect heterogeneity, i.e., one kind of intervention may be effective for some people but not for others. This approach is usually called *profiling*, and is not always accepted, in particular it is not obvious to what extent observed characteristics can be used to chose

between alternatives: the analyst may be precluded from using certain covariates (say gender, previous merit). Note that maximization of $W(k)$ with respect to alternative distribution of resources is a problem that would require appropriate operational research tools. In Compagnino and Gori (1992) a similar approach was used in order to find an *optimal* allocation of university grants.

3.2 How can we derive predictions of the Y_{ki} ?

The estimation of marginal distributions of the different Y_{ki} is a rather difficult task, especially if some of the alternative interventions under consideration have never been implemented before. A structural behavioral model must be specified and estimated from available data; the model should be able to predict the modifications of behavior or conditions induced by the different interventions or the different ways of making the intervention available in the future. For example, in the university grant problem, suppose grants have never been assigned to students before; thus we would need a model where students behavior (for example their academic performance as well as their decision to continue their study and achieve a university degree) are linked to some target variables, in this case their income, or the income of their family, because the grant can indeed be thought of as a money transfer that is changing their endowment.
The model can be specified at different degree of complexity, also depending on the availability of data. A simple model would be one for the probability of getting a degree, or the duration of the studies, where this depends on the student's characteristics, in particular merit and income (or some equivalent scale for it), and student choices, for example the faculty of enrolment.
A more complex structural model would be a dynamic behavioral model (Wolpin, 1996) in which each student maximizes his/her lifetime earnings and takes schooling decision, given an outside wage and some intertemporal discounting factor, where the disutility of loss income today is higher for the poor than for the rich. This model would be a lot richer than the previous one, but would also require more assumptions and more information (for example those on wages and on the returns to schooling) that are not always readily available.
If grants have already been assigned in the past, one can use the results of the ex-post analysis as a benchmark for the ex-ante analysis. If the structural model is able to reproduce the ex-post results up to a reasonable approximation, there is hope it to be appropriate for simulating the effects of alternative programs: Attanasio and Meghir (2001), Borguignon et al. (2002), Todd and Wolpin (2003) are examples on how this comparison can be made.
An implicit assumption that is usually made here is that results obtained for one population can be extrapolate to other populations and/or to other time period; recently this issue has been discussed in Hotz et al. (2004). Also general equilibrium effect are usually neglected, i.e., we can define the potential outcomes of an individual irrespective of the (level of the) treatments the others receive. This assumption in usually called SUTVA (Stable Unit Treatment Value Assignment) in the statistical literature.
Note that the output of such models, can be used also to inform individual decisions. For example, we could provide students with predictions on their academic performance and labor market performance, given their observed characteristics, with different

degree programs, to help them choose the one they mostly prefer. This information can form the basis for a formal multi-criteria career choice problem (Bana e Costa and Chagas, 2004). This issue was discussed in Goracci (2000) where an optimal allocation of students among the various degree programs was found, optimal in the sense of maximizing a welfare function that depends on the estimated duration of the studies and on the labor market opportunities of each student given his/her characteristics.
As said in section 2, we believe that taking a Bayesian perspective to decision analysis, rather than frequentist one (as in Manski, 2003) allows to take account of all the uncertainty in a simple way by using the predictive distribution, now of the single Y_{ki}:'s. If $W(k)$includes individual preference U_i , then we can first substitute the U_i with their certain equivalents, and apply the chosen $W(k)$ in order to derive the expected welfare from policy k. If instead $W(k)$ does not include individual preferences, then we can compute the posterior predictive distribution for W, by drawing from each individual predictive distribution of Y_{ki}, and then summarize it by its expected value.

4. Concluding remarks

The main objective of Evaluation is to provide policy makers with useful information for choosing among policies. In this paper, attention was restricted to the so-called impact Evaluation of policies. We have shown the role of the decision maker in conducting both ex-post and ex-ante analysis. Formal decision analysis is needed to properly account for individuals' and policymakers' attitude towards uncertainty and inequality. Despite the fact that a rigorous Evaluation can only be implemented if appropriate statistical tools are used, the paper shows that Evaluation will benefit also from tools made available from other disciplines, such as Economics and Operational Research, that allow to elicit individuals and policy makers preferences and judge interventions.

References

Angrist J., Imbens G., Rubin D.B (1996) Identification of causal effects using instrumental variables, *Journal of the American Statistical Association*, 91, 444-472.

Attanasio O. , Meghir C. & Santiago A. (2001) Education Choices in Mexico: Using a structural Model and a Randomized Experiment to Evacuate PROGRESA, mimeo.

Bana e Costa C.A. (2001) The use of multi-criteria decision analysis to support the search for less conflicting policy options in a multi-actor context: case study, *Journal of Multi-Criteria Decision Analysis*, 10, 2, 111-125.

Bana e Costa C.A. & Chagas M.P. (2004) A career choice problem: an example of how to use MACBETH to build a quantitative value model based on qualitative value judgements, *European Journal of Operational Research*, 153, 2, 323-331.

Bini M. & Chiandotto B. (2003) La valutazione del sistema universitario italiano alla luce della riforma dei cicli e degli ordinamenti didattici, *Studi e Note di Economia*, 2, 29-61.

Bourguignon F., Ferriera F.H. & Leite P. (2002) Ex-ante evaluation of conditional cash transfer programs: the case of Bolsa Escola, World Bank Policy Research Working Paper n. 2916, The World Bank, Washington D.C.

Bleichrodt H. Pinto J.L. & Wakker P. (2001) Making descriptive use of prospect theory to improve the prescriptive use of expected utility in www.fee.uva.nl/creed/wakker/pcf/corrsg.pdf.

Chiandotto B. (2002) Valutazione dei processi formativi: cosa, come e perché, in Valutazione della didattica e dei servizi nel sistema Università, *Atti della giornata di Studio, Fisciano*, CUSL, Salerno.

Compagnino A. & Gori E. (1992) *Il controllo di gestione degli enti per il diritto allo studio universitario. Efficienza ed efficacia*, Franco Angeli, Milano.

Deaton A. & Muelbauer J. (1980) *Economics and consumer behavior*, Cambridge University Press, Cambridge.

Dehejia R.H. & Wahba S. (1999) Causal effects in nonexperimental studies: reevaluating the evaluation of training programs, *Journal of the American Statistical Association*, 94, 1053-1062.

Dehejia R.H., (2004) Program Evaluation as a Decision Problem, forthcoming, *Journal of Econometrics*.

Fennema H. & Van Assen M. (1999) Measuring the utility of losses by means of the tradeoff method in *Journal of Risk and Uncertainty*, Vol. 17.

Frankgakis C. E. & Rubin, D. B. (2002) Principal Stratification in Causal Inference, *Biometrics,* 58, 21-29.

French S. (1988) *Decision Theory*, Ellis Horwood Limited, UK.

Gelman A., Carlin J.B., Stern H.S. & Rubin D.B. (1995) *Bayesian Data Analysis*, Chapman & Hall, London, UK.

Goracci L. (2000) *Scelte educative e sbocchi occupazionali degli studenti universitari: un'applicazione ai dati di Almalaurea*, Tesi di laurea, Corso di Laurea in Scienze Statistiche ed Economiche, Università di Firenze.

Grilli L. & Mealli F. (2004) Analysis of the Effectiveness of Degree Programmes by means of Principal Stratification, mimeo.

Hahn J., Todd P. & Van der Klaauw W. (2001) Identification and estimation of treatment effects with a regression-discontinuity design, *Econometrica*, 69, 1, 201-209.

Heckman J.J. (1989) Causal Inference and nonrandom samples, *Journal of Educational Statistics*, 14, 159-168.

Heckman J.J & Hotz V.J. (1989) Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training, *Journal of the American Statistical Association*, 84, 408, 862-874.

Heckman J. & Smith J. (1998) Evaluating the Welfare State. In Strom, S. (Ed.), *Econometrics in the 20th Century: The Ragnar Frisch Centenary*, Cambridge University Press, Econometric Society Monograph Series, Cambridge.

Heckman J.J., Lalonde R. & Smith, J. (1999) The economics and econometrics of active labour market programs, in Ashenfelter O. e Card D. (eds), *Handbook of Labor Economics*, vol. 3a, North Holland, Amsterdam.

Hirano K., Imbens G.W., Rubin D.B. & Zhou X. (2000) Assessing the effect of an influenza vaccine in an encouragement design, *Biostatistics*, 1, 69-88.

Holland P. (1986) Statistics and Causal Inference, *Journal of the American Statistical Association*, 81, 945-970.

Holland P.W. & Rubin D.B. (1983) On Lord's paradox. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement*, Hillsdale, NJ: Erlbaum.

Hotz J., Imbens G.W. & Mortimer J. (2004) Predicting the Efficacy of Future Training Programs Using Past Experiences, *Journal of Econometrics*, forthcoming.
Manski C.F. (1990) Nonparametric Bounds on Treatment Effects, *American Economic Review Papers and Proceedings*, 80, 319-323.
Manski C.F. (2000) Identification Problems and Decisions under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Choice, *Journal of Econometrics*, 95, 415-442.
Manski C.F. (2004) Statistical Treatment Rules for Heterogeneous Populations: with Application to Randomized Experiments, *Econometrica*, forthcoming.
Mealli F., Pudney S. & Thomas J. (1994) Youth employment and the optimal structure of Youth Training: an econometric analysis, Discussion Papers in Economics n. 94/14, University of Leicester.
Mealli F. & Rampichini C. (2001) Impact evaluation of University grants, *JSM, New York, Proceedings.*
Mealli F. & Rubin D.B. (2003) Commentary: Assumptions allowing the estimation of direct causal effects, *Journal of Econometrics*, 112, 79-87.
Mealli F., Imbens G.W., Ferro S.& Biggeri A. (2004) Analyzing a randomized experiments for breast self-examination with noncompliance and missing outcomes, *Biostatistics*, forthcoming.
Meny Y. & Thoenig J.C. (1991), *Le politiche pubbliche*, Il Mulino, Bologna.
Keefer D.L., Kirkwood C.W. & Corner J.L. (2004) Perspective on Decision Analysis, Applications, 1990-2001, *Decision Analysis*, 1, 5-24.
Keeney R.L., von Winterfeldt D. & Eppel T. (1990), Eliciting values for complex policy decisions, *Management Science*, 36, 9, 1011-1030.
Pearl J. (2000), *Causality*, Cambridge University Press.
Rosenbaum P. & Rubin D.B. (1983), The central role of the propensity score in observational studies for causal effects, *Biometrika,* 70, 41-55.
Rubin D.B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology*, 66, 668-701.
Rubin D.B. (1990) Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies. *Statistical Science* 5, 472-480.
Thistelthwaite D. & Campbell D. (1960) Regression-discontinuity analysis: an alternative to the ex-post facto experiment, *Journal of Educational Psychology*, 51, 309-317.
Todd P. & Wolpin K.I. (2003) Using a social experiment to validate a dynamic behavioral model of child schooling and fertility: assessing the impact of school subsidy in Mexico, mimeo.
Tsoukias A. (2003) From decision theory to decision aiding methodology, DIMACS Technical report 2003-21.
Wolpin K.I. (1996) Public-Policy Uses of Discrete-Choice Dynamic Programming Models, *American Economic Review*, Papers and Proceedings, 86,2, 427-432.
Zhang J. & Rubin D. B. (2004) Censoring due to death via principal stratification, *Journal of Educational and Behavioral Statistics,* forthcoming.
Zhang J., Rubin D. B. & Mealli F. (2004) Evaluating the Effects of Training Programs on Wages with Experimental Data, mimeo.

La gestione dei dati mancanti nei modelli di inferenza causale: il caso degli esperimenti naturali

Da: Mercatanti A. (2004) La gestione dei dati mancanti nei modelli di inferenza causale: il caso degli esperimenti naturali, in Strategie metodologiche per lo studio della transizione Università- lavoro, a cura di E. Aureli Cutillo, pp. 271-280, Cleup, Padova.

La gestione dei dati mancanti nei modelli di inferenza causale: il caso degli esperimenti naturali

Andrea Mercatanti[1]
Dip. di Statistica e Matematica Applicata all'Economia, Università di Pisa

Riassunto. Il lavoro prende in esame alcuni aspetti relativi alla gestione dei dati mancanti nei contesti di inferenza causale usualmente denominati "esperimenti naturali". Dopo un'introduzione alla problematica e l'illustrazione di una funzione di verosimiglianza adeguata al trattamento dei dati mancanti sotto condizioni di ignorabilità, verrà proposto un esempio basato su simulazioni. L'esempio è teso a mettere in evidenza l'importanza di un'attenta considerazione del meccanismo generatore dei dati mancanti in un'analisi di massima verosimiglianza. A tal proposito sarà presa in considerazione una popolazione ipotetica le cui caratteristiche rispecchiano quelle del dataset che Angrist e Krueger (1991) hanno utilizzato ai fini della valutazione dell'effetto della scolarizzazione sul reddito.
Parole chiave: Inferenza causale, variabili strumentali, dati mancanti, rendimento della scolarizzazione.

1. Introduzione alla problematica

I metodi statistici per l'inferenza causale sono stati oggetto negli ultimi anni, e lo sono tuttora, di continuo studio e di approfondita ricerca sia metodologica che applicativa. Il fine di questi metodi sta nella quantificazione delle relazioni di causa ad effetto, argomento di preminenza centrale non solo nel dibattito teorico statistico ma in un più ampio contesto filosofico scientifico. Dal punto di vista statistico, una delle giustificazioni ad un siffatto interesse è da ricercarsi nelle vaste possibilità di applicazione che spaziano dall'epidemiologia, alla microeconomia, alle problematiche della valutazione di efficacia, toccando anche i problemi dell'istruzione. In particolare, un settore applicativo relativamente recente che riguarda sia aspetti di tipo microecono-

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

mico che aspetti connessi all'istruzione è rappresentato dalla valutazione del cosiddetto *return to scholling*, inteso come il guadagno in termini monetari causato dall'aver conseguito un certo grado di istruzione. In questo campo fanno scuola alcuni lavori tra i quali Card (1995), Angrist e Krueger (1991), Ichino e Winter-Ebmer (1999), i quali al fine di risolvere il problema dell'autoselezione al trattamento tipico degli studi osservazionali utilizzano variabili strumentali aventi sostanziale significato sia dal punto di vista statistico causale, cioè aventi un'effetto significativo sul trattamento, sia da quello microeconomico, in quanto possono essere considerate come variabili operanti dal lato dei costi e quindi facilmente implementabili nell'ambito di una teoria di ottimo marginalista. Un semplice riferimento ad alcuni importanti lavori di microeconometria è già stato sufficiente per toccare alcune questioni tipiche e controverse nella valutazione dei nessi di causalità in ambito non sperimentale, quali l'autoselezione al trattamento e l'uso di variabili strumentali. Vediamo di darne una breve ma il più possibile esauriente spiegazione e di circoscriverne l'ambito all'oggetto del presente lavoro.

Al netto di alcuni recenti contributi, Dawid (2002), Pearl (2000), il modello statistico metodologico che probabilmente è stato maggiormente utilizzato nell'analisi di microdati cross-section è il cosiddetto "modello causale di Rubin", Holland (1986), il quale si basa sull'idea di controfattualità di Fisheriana memoria. Il concetto di controfattualità impone che l'effetto causale di un trattamento su di una variabile di risposta venga definito come confronto (tipicamente una differenza) tra i valori assunti dalla variabile di risposta in corrispondenza dei possibili valori assumibili dal trattamento. Ad esempio, l'effetto del conseguimento di un diploma di Laurea sul reddito di una certo laureato osservato all'età di 50 anni, secondo questa impostazione viene definito come differenza tra il reddito osservato all'età di 50 anni e il reddito che si sarebbe osservato per la stessa persona, sempre a 50 anni, ma nell'ipotesi che la stessa non avesse conseguito il diploma di Laurea. Ovviamente non sarà mai possibile una valutazione del genere tant'è che il dilemma viene risolto facendo riferimento ad una definizione di effetto causale come media degli effetti individuali in una popolazione di riferimento. In questo modo gli effetti causali possono essere facilmente stimati dai risultati di processi sperimentali, i quali vengono posti in essere mediante l'assegnazione casuale (la cosiddetta randomizzazione) di unità a due distinti gruppi differenziati per il fatto che tutti gli appartenenti ad un gruppo (casi) sono forzatamente sottoposti al trattamento mentre a tutti gli appartenenti all'altro gruppo (controlli) viene impedita la somministrazione del trattamento oppure viene forzatamente somministrato un trattamento alternativo. Siffatti procedimenti sperimentali, sebbene costituiscano una situazione ottimale dal punto di vista inferenziale, rappresentano però un'eccezione, soprattutto nelle scienze economiche e sociali, in quanto motivi di ordine etico sovente ne creano impedimento. Conseguentemente, in campo economico e sociale chi vuole eseguire analisi di causalità si trova

il più delle volte ad operare su dati provenienti da fonti non-sperimentali. E' ad esempio il caso dei già citati contributi riguardanti la valutazione del *return to schooling*. In queste situazioni di lavoro i risultati inferenziali, rispetto al caso sperimentale, sono però più difficili da raggiungere a causa dell'autoselezione al trattamento, cioè dell'assegnazione al trattamento non derivante da una processo di randomizzazione. Si ricorre allora a modelli che possono essere parametrici o semi-parametrici, oppure all'uso di variabili strumentali. In particolare la metodologia basata sulle variabili strumentali consente di porre in essere quello che viene comunemente denominato "esperimento naturale", Ichino (2001). Si tratta, in sintesi, di identificare una variabile che abbia le caratteristiche per poter essere considerata un'assegnazione casuale al trattamento ma rispetto alla quale, a differenza degli esperimenti veri e propri, non esistono vincoli coercitivi. Ad esempio; Angrist e Krueger (1991) nell'intento di valutare l'effetto della scolarizzazione sul reddito (negli Stati Uniti) utilizzano una variabile strumentale binaria che discrimina se l'individuo è nato o meno nell'ultimo trimestre dell'anno; tale scelta deriva dalla constatazione che la legislazione statunitense, da un lato impedisce l'abbandono scolastico prima del compimento del sedicesimo anno di età, ma dall'altro fa coincidere l'inizio della frequenza della scuola con l'anno solare. In questo modo ai ragazzi nati nell'ultimo trimestre, essendo costretti dalla legislazione a restare a scuola più a lungo, corrisponde una minore propensione ad abbandonarla prima del conseguimento del diploma di Scuola Superiore "High School". La data di nascita viene quindi utilizzata come variabile strumentale, può infatti essere considerata un'assegnazione casuale al trattamento che gli individui non sono però obbligati a rispettare. Nell'ambito di un esperimento naturale è inoltre possibile classificare le unità statistiche in base a come reagiscono all'assegnazione al trattamento, si parla infatti di *compliers* per gli individui che adottano il trattamento in conseguenza dell'assegnazione e di *noncompliers* per i restanti. Nell'esempio, i *compliers* sono i ragazzi che rimangono un'anno in più a scuola poiché obbligati dal fatto di essere nati nell'ultimo trimestre, ma che avrebbero abbandonato se nati negli altri trimestri.

Una problematica rilevante e frequente nelle analisi statistiche di dataset di tipo osservazionale è costituita dalla presenza di dati mancanti. In particolare, il problema appare particolarmente evidente quando la fonte dei dati è costituita da risposte a questionari e in particolare di fronte a domande delicate quali ad esempio quelle riguardo al reddito percepito. Il presente lavoro vuole prendere in considerazione la gestione dei dati mancanti nell'analisi di esperimenti naturali, imponendo condizioni di ignorabilità per il meccanismo generatore degli stessi.

2. La funzione di verosimiglianza per un esperimento naturale sotto condizioni di ignorabilità per il meccanismo generatore dei dati mancanti

Volendo affrontare la questione dei dati mancanti in dataset da utilizzare per la stima di effetti causali mediante l'uso di variabili strumentali, ossia nel contesto dei cosiddetti esperimenti naturali, occorre tener presente fin da subito che nelle analisi di causalità basate sul concetto di eventi controfattuali i dati mancanti propriamente detti (come ad esempio le non risposte a domande di questionari) non sono l'unica fonte di informazione mancante. Abbiamo appena accennato nel paragrafo precedente al fatto che a livello individuale soltanto uno dei due eventi controfattuali viene osservato, l'altro costituisce un'evento ipotetico ma comunque necessario alla costruzione di un modello metodologico che abbia significato dal punto di vista della filosofia della causalità. Occorre quindi tenere in considerazione questa duplice tipologia di informazioni mancanti nell'esplicitazione della funzione di verosimiglianza da utilizzare.

Una soluzione semplice al problema dei dati mancanti è costituita, non solo nel contesto degli esperimenti naturali, dall'eliminazione delle unità statistiche che presentano almeno un valore mancante per le variabili. Questo è possibile soltanto se il meccanismo che ha generato i dati mancanti soddisfa la condizione usualmente detta *Missing Completely At Random* (MCAR), Little e Rubin (1987), la quale impone che la probabilità di avere valori mancanti sia la stessa per ogni unità statistica. La condizione MCAR appare però in molti casi eccessivamente restrittiva. Nei casi in cui non si vogliano affrontare le difficoltà insite nella specificazione di un meccanismo generatore dei dati mancanti ma, allo stesso tempo, si cerchi di soddisfare assunzioni più blande, si fa usualmente ricorso a condizioni di cosiddetta ignorabilità del meccanismo generatore dei dati mancanti, Rubin (1976). Queste condizioni permettono di far riferimento alla funzione di verosimiglianza ottenibile dall'integrazione della funzione di verosimiglianza completa (cioè quella che si avrebbe nell'ipotesi di assenza di dati mancanti) rispetto alle quantità non osservate. In tal senso è allora possibile "ignorare" il modello probabilistico che genera i valori mancanti. Per una esauriente illustrazione delle condizioni di ignorabilità si può far riferimento a Rubin (1976). In estrema sintesi, le condizioni di ignorabilità vengono soddisfatte quando le probabilità di non osservare i dati dipendono soltanto dalle quantità osservate. Nel caso di un'esperimento naturale per il quale il meccanismo generatore dei dati mancanti soddisfa le condizioni di ignorabilità, occorre però tener presente che l'integrazione della funzione di verosimiglianza completa rispetto ai dati mancanti va estesa alle quantità non osservate di tipo controfattuale. In termini formali, facciamo riferimento ad un'esperimento naturale per il quale indichiamo con Y la variabile di risposta, con D il trattamento e con Z la variabile strumentale (da in-

tendersi come assegnazione casuale al trattamento). Ipotiziamo altresì che vengano soddisfatte le ipotesi, usualmente adottate per l'identificazione di effetti causali per mezzo di variabili strumentali, che consentono di valutare l'effetto di un trattamento per i soli soggetti che agiscono in accordo all'assegnazione al trattamento (*compliers*), Angrist et al. (1996). Tutto ciò dato, la funzione di verosimiglianza sulla quale operare può essere così formalizzata, Mercatanti (2003):

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} \int \ldots \int f(\underline{d_i}, \underline{y_i}; \boldsymbol{\theta})\, d\, \underline{d_{mis,i}}\, d\, \underline{y_{mis,i}}\; dz_{mis,i}\, dd_{mis,i}\, dy_{mis,i} \qquad (1)$$

dove si è indicato: con $\underline{d_i}$ la coppia dei valori assunti dal trattamento in corrispondenza delle due alternative assegnazioni al trattamento; con $\underline{y_i}$ la coppia dei valori assunti dalla variabile di risposta in corrispondenza delle due alternative assegnazioni al trattamento; con $\underline{d_{mis,i}}$ il valore del trattamento non osservato nella coppia $\underline{d_i}$; con $\underline{y_{mis,i}}$ il valore della variabile di risposta non osservata nella coppia $\underline{y_i}$; e con $z_{mis,i}$, $d_{mis,i}$, *e* $y_{mis,i}$ gli eventuali valori mancanti delle variabili *z, d,* e *y* rispettivamente. Il vettore parametrico $\boldsymbol{\theta}$ è dato da:

$$(\pi_z, \omega_a, \omega_n, \omega_c, \boldsymbol{\eta}_a, \boldsymbol{\eta}_n, \boldsymbol{\eta}_{c0}, \boldsymbol{\eta}_{c1})$$

dove π_z è la probabilità di assegnazione al trattamento; ω_a, ω_n, e ω_c sono le probabilità di appartenenza ad un certo *compliance status* cioè rispettivamente ad uno dei gruppi denominati *always-takers*, *never-takers*[2] e *compliers*; $\boldsymbol{\eta}_a$ e $\boldsymbol{\eta}_n$ sono rispettivamente i vettori dei parametri delle distribuzioni di probabilità per gli *always-takers* e per i *never-takers*; $\boldsymbol{\eta}_{c0}$ e $\boldsymbol{\eta}_{c1}$ sono i vettori dei parametri delle distribuzioni di probabilità per i *compliers*, rispettivamente non assegnati e assegnati al trattamento.

3. Un esempio basato su simulazioni

In questo paragrafo, mediante il ricorso a simulazioni, è proposto un'esempio riguardante l'analisi di massima verosimiglianza per un'esperimento naturale il cui dataset sia perturbato da dati mancanti. Si ipotizza che il meccanismo generatore degli stessi soddisfi le condizioni di ignorabilità. Verranno creati 1000 dataset artificiali, ognuno di 1000 unità statistiche, estratti tutti dalla medesima ipotetica popolazione. Su ognuno di questi dataset viene eseguita una stima di massima verosimiglianza utilizzando la funzione (1). A fini comparativi i risultati verranno confrontati con quelli ottenibili

[2] Con il termine *always-takers* si intendono gli individui che assumono sempre il trattamento (indipendentemente dall'assegnazione); con il termine *never-takers* si intendono gli individui che non assumono mai il trattamento (anch'essi indipendentemente dall'assegnazione).

sugli stessi dataset artificiali ma attuando una più semplicistica eliminazione delle unità statistiche aventi almeno un valore mancante (procedura che presupporrebbe il soddisfacimento della condizione MCAR ma che spesso viene negligentemente adottata nelle ricerche applicate).

Data l'importanza assunta nel tempo dal lavoro di Angrist e Krueger (1991) (A&K) riguardante la valutazione dell'effetto della scolarizzazione sul reddito, si è ritenuto interessante attribuire ai parametri della popolazione artificiale valori attinenti a quelli della popolazione utilizzata dai suddetti autori. A questo fine si è fatto riferimento ad un successivo contributo, Imbens e Rubin (1997) (I&R), nel quale i parametri della popolazione sono stati stimati col metodo della massima verosimiglianza e ipotizzando una distribuzione normale per la variabile di risposta. I&R hanno assunto come variabile di risposta il logaritmo del reddito settimanale osservato nel 1980, come trattamento una variabile binaria che vale 0 se l'individuo ha frequentato le scuole per meno di dodici anni e 1 altrimenti, e come variabile strumentale una variabile binaria che vale 0 se l'individuo è nato nel primo trimestre dell'anno e 1 se è nato nell'ultimo. Il dataset utilizzato da I&R è costituito dalle osservazioni su 162515 individui nati negli Stati Uniti tra il 1/1/1930 e il 31/12/1939 durante il primo o l'ultimo trimestre.

I valori del vettore parametrico usato per le simulazione sono quindi i seguenti:

$$(\pi_z = 0.5, \omega_a = 0.762, \omega_n = 0.218, \omega_c = 0.02, \mu_a = 5.99, \sigma_a = 0.64,$$
$$\mu_n = 5.6, \sigma_n = 0.714, \mu_{c0} = 5.53, \sigma_{c0} = 0.714, \mu_{c1} = 6.03, \sigma_{c1} = 0.64)$$

e come detto corrispondono alle stime di massima verosimiglianza calcolate da I&R sul dataset originale utilizzato da A&K. Si noti come avendo assunto una distribuzione normale per la variabile di risposta, i parametri della distribuzione relativa ad ogni *compliance status* siano la media e lo scarto quadratico medio. Per ogni dataset artificiale, costituito mediante estrazione dalla popolazione di riferimento, vengono poi generati i dati mancanti in base ad un meccanismo soddisfacente le condizioni di ignorabilità, e che viene riportato in Tabella 1. Si può osservare come nel caso considerato la variabile trattamento sia sempre osservata, a differenza della variabile di risposta la cui osservabilità dipende dal valore assunto dal trattamento. Tenendo presente che si fa riferimento ad una ipotetica situazione per la quale la variabile di risposta è il logaritmo del reddito e il trattamento è un indice di scolarizzazione, si sta allora ipotizzando che la probabilità di osservare il reddito dipenda dal grado di istruzione, in particolare che sia più elevata (p=0.9 contro p=0.7) in corrispondenza di un grado di scolarizzazione più elevato.

La Tabella 2 mostra i risultati delle simulazioni per gli elementi più interessanti del vettore parametrico. Si può osservare come l'analisi di massima verosimiglianza condotta eliminando erroneamente gli individui che presentano dati mancanti (cioè presupponendo l'ipotesi MCAR) produca nel complesso dei 1000 dataset artifi-

ciali una distorsione media ed una radice quadrata dell'errore quadratico medio sempre superiori a quelli ottenuti dalla massimizzazione della funzione di verosimiglianza (1), che invece tiene in considerazione le condizioni di ignorabilità. Ciò appare particolarmente evidente riguardo ai parametri relativi alla distribuzione della variabile di risposta per i *compliers*. Si noti in particolare come per i parametri μ_{c0} e σ_{c1} i valori della distorsione media sotto l'ipotesi MCAR siano rispettivamente di –0.2425 e di –0.3482 contro 0.0370 e –0.0109 ottenuti sotto l'ipotesi di ignorabilità. Tenendo presente che il maggior interesse nelle analisi degli esperimenti naturali sta prevalentemente nello studio del comportamento degli individui che obbediscono all'assegnazione al trattamento, i risultati della simulazione sottolineano l'importanza di una calibrata e attenta considerazione del meccanismo generatore dei dati mancanti.

Tabella 1. *Definizione del modello probabilistico generatore dei dati mancanti.*

P(Y non è osservato|D=1)=	0.1
P(Y non è osservato|D=0)=	0.3
P(Z non è osservato|D=1)=	0.1
P(Z non è osservato|D=0)=	0.1

Tabella 2. *Distorsione media e radice quad. dell'errore quad. medio di stime MLE condotte su 1000 dataset artificiali di numerosità 1000 dalla popolazione ipotetica.*

	MLE sotto ipotesi di:	Distorsione Media	Rad. Quad. Errore Quad. Medio
π_z	Ignorabilità	8.86 x 10-5	0.0063
	MCAR	0.0022	0.0101
ω_a	Ignorabilità	-0.0145	0.0225
	MCAR	0.0504	0.0533
ω_n	Ignorabilità	-0.0139	0.0214
	MCAR	-0.0658	0.0674
ω_c	Ignorabilità	0.0154	0.0233
	MCAR	0.0285	0.0365
μ_{c0}	Ignorabilità	0.0370	0.2720
	MCAR	-0.2425	0.6696
μ_{c1}	Ignorabilità	-0.0587	0.3109
	MCAR	0.1446	0.9082
σ_{c0}	Ignorabilità	-0.2073	0.3949
	MCAR	0.3870	0.4019
σ_{c1}	Ignorabilità	-0.0109	0.2022
	MCAR	-0.3482	0.4530

Riferimenti bibliografici

ANGRIST J.D., A.B. KRUEGER (1991) Does compulsory school attendance affect schooling and earnings?, *Quaterly Journal of Economics,* **61**: 979-1014.

ANGRIST J.D., G.W. IMBENS, D.B. RUBIN (1996) Identification of causal effects using instrumental variables, *J.A.S.A.,* **91**: 444-455.

CARD D. (1995) Earnings, schooling, and ability revisited, *Research in labor economics,* **14**: 23-48.

DAWID P. (2002) Influence diagrams for causal modelling and inference, *International Statistical Review,* **70**, 161-189.

HOLLAND P.W. (1986) Statistics and causal inference, *J.A.S.A.,* **81**, 945-970.

ICHINO A. (2001) Il problema della causalità. Una introduzione generale e un esempio. In: BRUCCHI LUCHINO, *Manuale di economia del lavoro,* Il Mulino, Bologna, 459-483.

ICHINO A., R. WINTER-EBMER (1999) Lower and upper bounds of returns to schooling: an exercise in IV estimation with different instruments, *European economic review,* **43**, 889-901.

IMBENS G.W., D.R. RUBIN (1997) Bayesian inference for causal effects in randomized experiments with noncompliance, *The annals of statistics,* **25**, 305-327.

IMBENS G.W., D.B. RUBIN (1997) Estimating outcome distributions for compliers in instrumental variables models, *Review of economic studies,* **64**, 555-574.

LITTLE R.J.A., D.B. RUBIN (1987) *Statistical analysis with missing data*, J.Wiley and Sons, New York.

MERCATANTI A. (2003) Analyzing a ramdomized experiment with imperfect compliance and ignorable conditions for missing data, *Computational Statistics and Data Analysis,* in corso di stampa.

PEARL J. (2000) *Causality: models, reasoning, and inference,* Cambridge University Press, Cambridge, UK.

RUBIN D.B. (1976) Inference and missing data, *Biometrika,* **63**, 581-59.

Robust and Efficient Dimension Reduction

Da: Riani M., Bini M. (2002) Robust and Efficient Dimension Reduction, Relazione invitata alla XLI Riunione Scientifica della Società Italiana di Statistica, Milano 5-7 giugno 2002, pp. 295-306, Cleup, Padova.

Robust and Efficient Dimension Reduction

Metodi robusti ed efficienti per la riduzione delle dimensioni

Marco Riani
Dipartimento di Economia
Università di Parma
mriani@unipr.it

Matilde Bini[1]
Dipartimento di Statistica G. Parenti
Università di Firenze
bini@ds.unifi.it

Riassunto: L'obiettivo di questo lavoro è quello di estendere la procedura di forward search (Atkinson and Riani (2000)) alle componenti principali. L'approccio suggerito parte da un *data set* ridotto privo di outliers ed include sequenzialmente le rimanenti osservazioni in base ad una misura via via crescente di anomalia delle stesse. La procedura proposta consente di ottenere un approccio unificato all'analisi delle trasformazioni multivariate ed all'individuazione degli outliers e delle osservazioni influenti nelle componenti principali. Per illustrare il nuovo approccio si utilizza un data set che si riferisce alla valutazione della qualità delle università italiane.

Keywords: Principal components, robust methods, multivariate transformations, forward search, masked outliers, robustness.

1. Introduction

Principal component analysis (PCA) essentially uses eigenvalues and eigenvectors of covariance or correlation matrices, whose statistical properties are difficult to study. Most results are obtained under normal assumptions (e.g. Mardia *et al.* (1979); p. 229). This implies that at present few methods are available to assess the stability of the results of PCA. Often the assumption of multivariate normality is approximately true only after that the data have been appropriately transformed. Unfortunately, in multivariate data is very difficult to test and validate a particular transformation due to the well known masking and swamping problems (Velilla (1995)). The lack of a proper choice of the most appropriate transformation may lead to overestimate (underestimate) the importance of particular variables (Riani and Atkinson (2001)), and to a wrong interpretation of the extracted components. The difficulties and intricacies of the choice of the best multivariate transformation usually lead the analyst to apply PCA to untransformed data. Finally, as concerns outlier detection in PCA, the traditional approach is based on robust estimates of the covariance and correlation matrix or on projection pursuit (Hubert *et al.* (2002)). These estimators, however, are seldom used in practice because they are difficult to compute and unacceptably inefficient.

In this paper we suggest a unified robust and efficient approach to outlier detection, multivariate transformations, robust estimation of covariance matrices and dimension reduction. The suggested approach is an extension to PCA of the forward search (FS) methodology as described by Atkinson and Riani (2000) for regression data.

The paper is structured as follows: in section 2 we introduce notation and recall the likelihood ratio test for multivariate transformations. In section 3 we describe the general

[1]Acknowledgments: the dataset used in this work has kindly been made available by the National University Evaluation Committee-Ministry of Education, Universities and Research (MIUR).

principles of our robust diagnostic method. In section 4 we show how the FS, free from masking and swamping problems, can be used to find and test a set of transformation parameters, to reduce dimensionality and to monitor the stability of the extracted principal components. In order to show the power of our procedure, in section 5 we introduce and use a data set coming from the National University Evaluation Committee (NUEC).

2. Tests for transformation in multivariate analysis

Let Y be a sample data matrix of dimension $n \times p$ and let y_{ij} be the ith observation on response j. In the extension of the Box and Cox (1964) family to multivariate responses the normalized transformation of y_{ij} is

$$\begin{array}{rcll} z_{ij}(\lambda_j) & = & (y_{ij}^{\lambda_j} - 1)/\lambda_j \dot{y}_j^{\lambda_j - 1} & (\lambda \neq 0) \\ & = & \dot{y}_j \log y_{ij} & (\lambda = 0), \end{array} \tag{1}$$

where $\dot{y}_j$ is the geometric mean of the jth response. The value $\lambda_j = 1$ $(j = 1, \ldots, p)$ corresponds to no transformation of any of the responses.

The likelihood ratio test to validate the hypothesis $\lambda = \lambda_0$, can be expressed as the ratio of the determinants of two covariance matrices (Atkinson (2002)):

$$T_{LR} = n \log\{|\hat{\Sigma}(\lambda_0)|/|\hat{\Sigma}(\hat{\lambda})|\} \tag{2}$$

where $\hat{\lambda}$ is the vector of p parameter estimates found by numerical search maximizing the transformed normal log likelihood. The value of T_{LR} must be compared with the χ^2 distribution on p degrees of freedom. In order to find the most appropriate transformation we use both the maximum likelihood estimates (MLE) and the values of the test statistic. However, as we see in the next sections, in a confirmatory stage it is better to look at the value of the test rather than parameter estimates, because if the likelihood is flat, the estimates can vary widely without conveying any useful information about the transformation.

Remark: From equation (2) it is clear that if we want to use standard deletion methods to validate a particular set of transformations, we have to delete each observation in turn and every time remaximize the likelihood. Because of the necessity for remaximization, this approach becomes infeasible if multiple deletions are required. In addition, if there is a group of influential observations, the backwards approach may fail due to masking.

3. The Forward Search for PCA

Principal components analysis looks for a few linear combinations which can be used to summarize the data, losing in the process as little information as possible. Let $\tilde{Y}$ be the sample centered data matrix $\tilde{Y} = Y - 1\bar{y}'$, and v a standardized vector ($v'v = 1$). The linear combination of the columns of $\tilde{Y}$, $z = \tilde{Y}v$ which has the largest variance is obtained when v is the standardized eigenvector corresponding to the largest eigenvalue of $\hat{\Sigma} = V\Gamma V'$: V is the matrix which contains the eigenvectors ($V = (v_1, \ldots, v_p)$) and Γ is the diagonal matrix containing the eigenvalues $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_p)$ in non decreasing order. The j-th principal component z_j is defined as $z_j = \tilde{Y}v_j$, $j = 1, \ldots, p$. As is well known, the sum of the first k eigenvalues divided by the sum of all the eigenvalues

of the matrix $\hat{\Sigma}$ represents the proportion of variation explained by the first k principal components.

The small sample distribution of the eigenvalues and of the eigenvectors is very complicated even if all parent correlations vanish. The reason comes from the fact that the eigenvalues are non rational functions of the elements of $\hat{\Sigma}$. The large sample results stem from the theorem (Anderson (1963)) which states that for *normal data*, when the eigenvalues of Σ are distinct, the sample principal components and the eigenvalues are the maximum likelihood estimators of the corresponding population parameters. This implies that is very important to try to transform the data to reach approximate normality and to remove outliers. If the values of the parameters of the model were known ($\mu(\lambda)$ and $\Sigma(\lambda)$), there would be no difficulty in detecting the outliers, which would have large Mahalanobis distances. The difficulty arises because the outliers are included in the data used to estimate the parameters, which can then be badly biased.

Most methods for outlier detection therefore seek to divide the data into two parts, a larger "clean" part and the outliers. The clean data are then used for parameter estimation. The simplest example of this division of the data into two parts is in the use of single deletion diagnostics, where the division is into one potential outlier and the rest of the data. It is clear, however, that if multiple deletions are required there is a combinatorial explosion of the number of cases that have to be considered by such backwards working. Finally, the situation is complicated by the fact that outliers in one transformed scale may not be outliers in another scale.

In the FS such larger subsamples of outlier free observations are found by starting from small subsets and incrementing them with observations which have small Mahalanobis distances, and so are unlikely to be outliers. More precisely, the FS is made up of three steps: choice of the initial subset, progressing in the search and monitoring.

Step 1: Choice of the Initial Subset

We find an initial subset of moderate size by robust analysis of the matrix of bivariate scatter plots. The initial subset of r observations consists of those observations which are not outlying on any scatter plot, found as the intersection of all points lying within a robust contour containing a specified portion of the data and inside the univariate boxplot. There are two versions of the robust bivariate contour. The first uses convex hull peeling and B-spline smoothing (Zani *et al.* (1998)). The second, more simple but less robust, is based on robust ellipses (Riani and Zani (1997)). An important property of this method is that the size of the subset can easily be increased or decreased by changing the level of the contour. In the examples which follow we always use the quicker version, because we simply need to find an initial subset of a certain size which is not contaminated by outliers. Finally, since it is the extreme observations which provide the evidence for transformations, such a robust subset will provide a good start to the search for many values of λ.

Step 2: Adding Observations during the Forward Search

In every step of the forward search, given a subset $S_*^{(m)}$ of size m $(m = r, \ldots, n-1)$, we move to a subset of size $(m+1)$ by selecting the $(m+1)$ units with the $(m+1)$ smallest Mahalanobis distances.

More specifically, let $\hat{\mu}_{i,S_*^{(n)}} = \hat{\mu}_i$ be the estimated response using all the observations. The squared Mahalanobis distance for observation i is

$$d_i^2 = (y_i - \hat{\mu}_i)^T \hat{\Sigma}^{-1} (y_i - \hat{\mu}_i) = e_i^T \hat{\Sigma}^{-1} e_i, \tag{3}$$

where $\hat{\mu}_i$ is simply the arithmetic mean of the $y_1, \ldots, y_n$ or may come from regression and $\hat{\Sigma} = \hat{\Sigma}_{S_*^{(n)}}$ is the sample covariance matrix with

$$\hat{\Sigma}_{jk} = \sum_i (y_{ij} - \hat{\mu}_{ij})(y_{ik} - \hat{\mu}_{ik})/(n-r) = \sum_i e_{ij} e_{ik}/(n-q), \tag{4}$$

and q is the dimension of the vector of regression parameters.

Now suppose that a subset $S_*^{(m)}$ of m observations is used to estimate the regression and covariances. Let the estimates be $\hat{\mu}_{i,S_*^{(m)}}$ and $\hat{\Sigma}_{S_*^{(m)}}$, yielding the set of squared Mahalanobis distances

$$d^2_{i,S_*^{(m)}} = \left(y_i - \hat{\mu}_{i,S_*^{(m)}}\right)^T \hat{\Sigma}^{-1}_{S_*^{(m)}} \left(y_i - \hat{\mu}_{i,S_*^{(m)}}\right) \qquad i = 1, \ldots, n \tag{5}$$

At step m we move to step $m+1$ by selecting the $m+1$ units with the smallest $m+1$ $d^2_{i,S_*^{(m)}}$. Usually just one new unit joins the subset. It may also happen that two or more units join $S_*^{(m)}$ as one or more leave. However, our experience is that such an event is quite unusual, only occurring when the search includes one unit which belongs to a cluster of outliers. At the next step the remaining outliers in the cluster seem less outlying and so several may be included at once. Of course, several other units then have to leave the subset.

Step 3 Monitoring the search

In multivariate analysis we first try to find a set of transformation parameters to reach approximate normality. In each step of the search (as m goes from r to n (where $p << r < n$)) we initially monitor the evolution of MLE and of the likelihood ratio test for transformations and of the Mahalanobis distances using the procedure described in section 4. The changes which occur, will be associated with the introduction of particular observations into the subset m used for fitting. Once a reasonable set of values of transformation parameters λ is found, if our purpose is dimension reduction (PCA), in order to find out if the results are stable, in each step of the search we monitor the elements of the eigenvalues:

$$\Gamma_{S_*^{(r)}}, \ldots, \Gamma_{S_*^{(n)}} \tag{6}$$

and the eigenvectors

$$V_{S_*^{(r)}}, \ldots, V_{S_*^{(n)}} \tag{7}$$

of the transformed variables. In general, the forward search estimator $\hat{\theta}_{FS}$ is defined as the collection of maximum likelihood estimators in each step of the forward search, that is:

$$\hat{\theta}_{FS} = \left(\hat{\theta}_r, \ldots, \hat{\theta}_n\right). \tag{8}$$

We use a robust method for the estimation of $\hat{\theta}_r$, but maximum likelihood (that is fully efficient) estimators during the remainder of the search. The zero breakdown point of

maximum likelihood estimators, in the context of the forward search, has advantages. The introduction of atypical (influential) observations is signalled by sharp changes in the curves which monitor the percentage of variance explained by the first principal components, the elements of the eigenvectors or other statistics at every step. In this context, the robustness of the method does not derive from the choice of a particular estimator with a high breakdown point, but from the progressive inclusion of the units into a subset which, in the first steps, is outlier free.

The search which we use avoids, in the first steps, the inclusion of outliers and provides a natural ordering of the data according to the specified null model. It is therefore possible to know how many observations are compatible with a particular specification. Furthermore, the suggested approach enables us to analyze the inferential effect of the atypical units (outliers) on the results of statistical analyses (Cerioli and Riani (1999)).

4. Finding a transformation and reducing the dimension with the forward search

With just one variable for transformation it is comparatively easy to use our forward search to find satisfactory transformations, if such exist, and the observations that are influential in their choice. Atkinson and Riani (2000) perform a forward search using five standard values of λ ($-1, -0,5,$ 0, 0.5 and 1) and monitor the likelihood ratio statistic for transformations for each search. However, with p variables for transformation there would be 5^p combinations of the standard values. Whether or not the calculations are time consuming, trying to absorb and sort the information would be difficult (Riani and Atkinson (2001)). We therefore suggest three steps to help structure the search for a transformation:

1. Run a forward search through the data, ordering the observations at each m by Mahalanobis distances calculated from untransformed observations. Estimate λ at each value of m. Use the results to select a set of transformation parameters.
2. Rerun the forward search using distances calculated with the parameters selected in the first step, again estimating λ for each m. If some change is suggested in λ, repeat this step until a reasonable set of transformations has been found. Let this be λ_R.
3. Test the suggested transformation. We expand each transformation parameter in turn around the five common values of λ $(-1, -0.5, 0, 0.5, 1)$, using the values of the vector λ_R for transforming the other variables. In this way we turn a multivariate problem into a series of univariate ones. In each search we can test the transformation by comparing the likelihood ratio test with χ^2 on 1 degree of freedom. But we prefer to use the signed square root of the likelihood ratio in order to learn whether lower or higher values of λ are indicated.

If the final purpose of the analysis is dimension reduction, once a reasonable set of transformation parameters has been found and tested, we rerun the FS using the transformed variables and monitor the Mahalanobis distances, the percentage of variance explained by the different components and the stability of the eigenvectors. In the next sections we exemplify this procedure for a real data set concerning the evaluation activity of the Italian Universities.

5. The evaluation of university activity through the forward search

The university service can be modelled as a production process where inputs such as capital, labour and organisational factors, are used to produce teaching and research outputs and outcomes. Many stakeholders such as university policy-makers (i.e. Ministry of the Education - MIUR), teachers, researchers, technical and administrative staff and students, interact among them inside the process and are interested to measure the performance of the process. In the literature, in order to measure the productivity of the research and teaching, and to carry out comparative cost-efficiency analyses, it has been suggested to collect 29 indicators which can be grouped in the following four classes (Ewell (1999); Report n.11/98 in http://www.murst.it/osservatorio/pubblic.htm): **Resources indicators**, i.e. indicators of resources as available funds, staff, etc.; **Contextual indicators**, i.e. indicators of the context where the university is working, that indicate the degree of the importance of the university, size, the socio-economic environment, etc.; **Process indicators**, that should inform about the organization, facilities and results of the teaching and research processes; **Outcome indicators**, that should inform about the final results and the degree of quality of the teaching and research activities (Biggeri and Bini (2000); http://www.murst.it/valutazionecomitato/activnuc.htm).

In Italy starting from March 2000, the National University Evaluation Committee (NUEC) yearly conducts census surveys on all the 73 Italian universities to monitor the situation of the university system and to organize the statistical information system in order to carry out the requested evaluations. The information collected concerns the following aspects:

- the student and academic staff characteristics;
- the teaching service and the studying conditions of students (quality of teaching, classroom availability, laboratories, libraries, location of the offered services);
- the facilities (funds included) devoted to teaching and research activities;
- the managing and administrative organization as well as the working conditions of the technical and administrative staff.

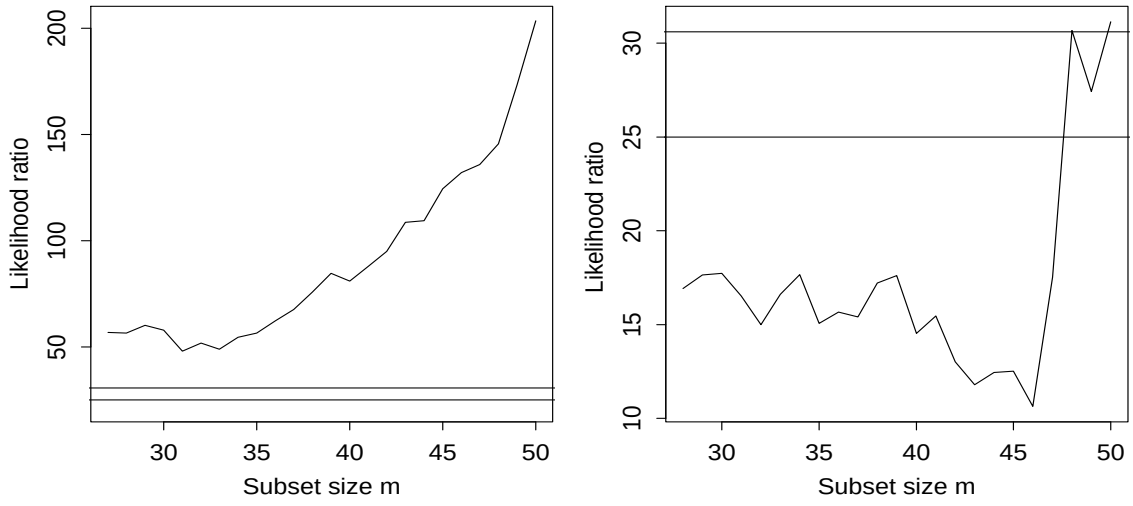
In our study we constructed 15 indicators $(p = 15)$ for 50 public Universities $(n = 50)$[2]. More precisely the variables which have been considered are the following:

- **Outcome**: graduation rate within institutional time (y_1); retention rate = 1- drop out rate (y_2).
- **Resources**: faculty/students (y_3); university current expenditure per student (y_4); research funds (y_5); administrative staff per faculty (y_6).
- **Contextual**: university size (y_7); enrollment rate of the best freshmen (y_8); enrollment rate of freshmen from classic and scientific high schools (y_9); extra-regional attraction index (y_{10}).
- **Process**: non institutional time students ratio (y_{11}); private research grants per single member of faculty (y_{12}); financial self-sufficiency (y_{13}); course completing ratio (y_{14}); expenditure for technical staff per ordinary funds (y_{15}).

[2]We have only included public Universities since the private Universities have a different organization of their structures and are evaluated using different criteria of judgement from the public institutions, for receiving funds and incentives from MIUR.

The scatter plot matrix of these data (not given here for lack of space) shows that the distribution of some variables is highly skewed and that maybe some outliers are present in the data. The MLE of the transformation parameters using all the observations are $\hat{\lambda} =$ (0.19, 0.02, 0.65, 1.22, 0.14, -1.24, -0.16, -0.01, 0.78, 0.49, 1.70, 0.46, 0.37, 1.71, 2.01). The null hypothesis of no transformation, that is all $\lambda_j = 1$, yields a value of $T_{LR} = 203.37$. Given that this value must be compared with a χ^2_{15}, it is clear that the p-value is virtually zero. The question is the following: is this large value of T_{LR} due to outliers or is the need of transformation spread throughout the data? In addition: what is the effect of each unit on the MLE of the transformation parameter for each variable? The left hand panel of Figure 1, which reports a FS for the hypothesis of no transformation enables us to state that the need for transformation is spread throughout the data, because the hypothesis of no transformation is always rejected in each step of the forward search. The monitoring of the maximum likelihood estimates of the transformation parameters as

Figure 1: Forward search for the Likelihood ratio test. $H_0 : \lambda = 1$ for all variables. The evidence of transformation is spread throughout the data (left). $H_0 : \lambda = (0.5, 0, 0.5, 0.5, 0.5, -1, 0, 0, 0.5, 0.5, 1, 0.5, 0.5, 1, 1)$. The transformation is acceptable up to step $m = 47$ (right). The two horizontal lines are respectively the 5% and 1% points of the associated χ^2 distribution



described in step 1 of the previous section (not given here for lack of space), suggests that a set of reasonable values for λ is $\lambda_R =$ (0.5,0,0.5,0.5,0.5,-1,0,0,0.5,0.5,1,0.5,0.5,1,1). The FS using these new values is shown in the 3 panels of Figure 2. This plot enables us to state what are the variables whose MLE are stable, what is the effect of the introduction of each unit on the MLE of the transformation parameters for each variable, and what is the impact of the outliers on the suggested transformation. For example, as concerns the MLE of λ for variable 4, up to $m = 46$ the curve is stable and always lies in the interval 0-0.5. At step $m = 47$ the curve shows a sudden upward jump and the final MLE becomes greater than 1. The same thing, even if less marked, happens for variable 3 (y_3). Finally, the curves for y_{11}, y_{14} and y_{15}, given in the right hand panel of Figure 2, vary widely and show that perhaps no transformation is necessary for these variables. The interpretation of the large changes in the values of the estimated parameters in Figure 2 is aided by considering the likelihood surface. Figure 3 shows how the loglikelihood changes when $m = 46$ and the other 14 values of λ are at their overall value of $\hat{\lambda}$. For example, the

Figure 2: Maximum likelihood estimates of the transformation parameters. The number by the curves refer to the variables

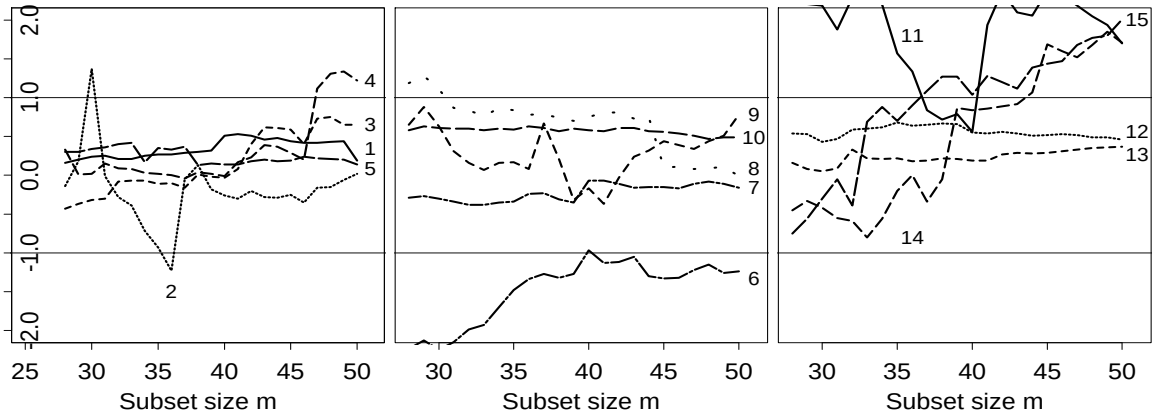
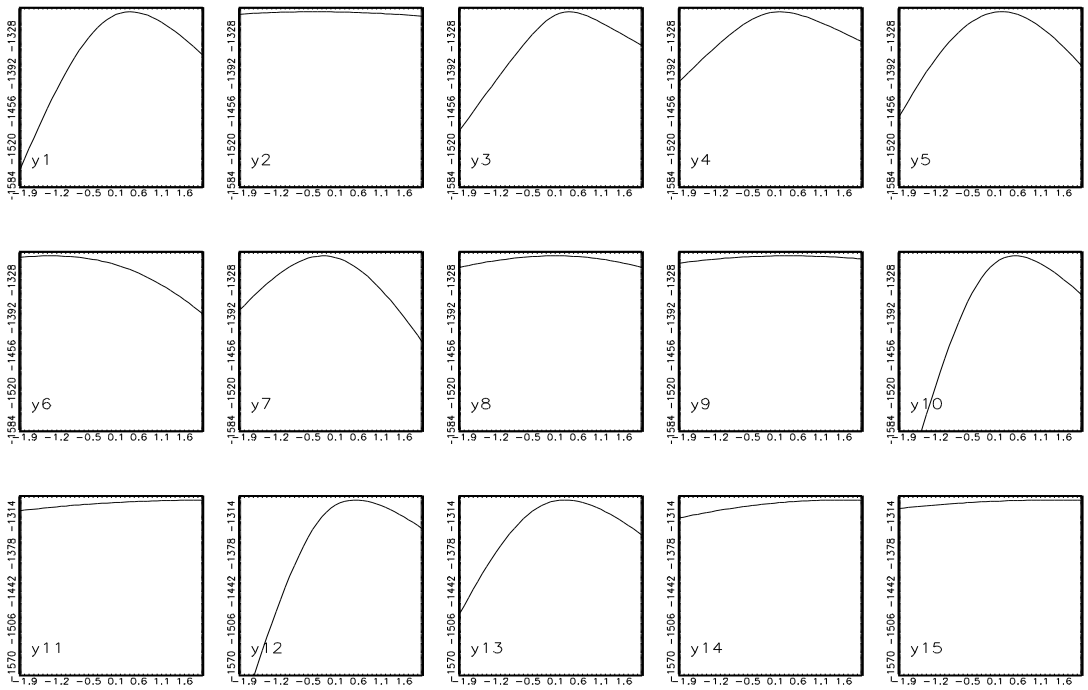


Figure 3: Likelihoods as function of the 15 components of λ when $m = 46$ and $H_0 : \lambda = \lambda_R$. Each panel refers to a variable



surfaces for y_{11}, y_{14} and y_{15} are relatively flat so that a small change in the shape by the addition of one unit can cause a large change in the position of the maximum and so of the value of $\hat{\lambda}$. In Figure 3 we have chosen to show step $m = 46$, because the monitoring of maximum and minimum Mahalanobis distances for the units not in the subset (see Atkinson and Riani (2000) or Atkinson (2002) for a detailed description of these plots) clearly shows that the last 4 units which enter the search (12, 23, 30 and 37), must be considered as atypical. The effect of these 4 units is also clear in the right hand panel of Figure 1 which monitors the likelihood ratio test for the final suggested transformation. Notice that working backwards it would have been impossible to discover the suggested combination of values of λ. In fact, using standard deletion diagnostics, the deletion of 1

or 2 observations does not enable us to see at all the tremendous effect these 4 units have on the likelihood ratio test.

Figure 4: Transformed data using λ_R. Forward plots of (left) maximum Mahalanobis distance (MD) inside subset and (right) minimum Mahalanobis distance not in the subset used to compute the centroid and the sample covariance matrix

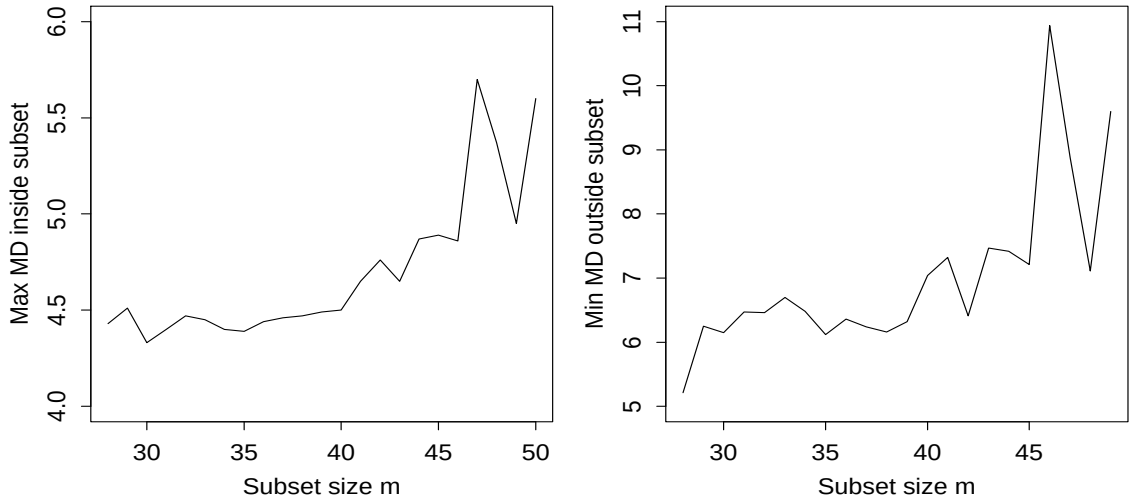
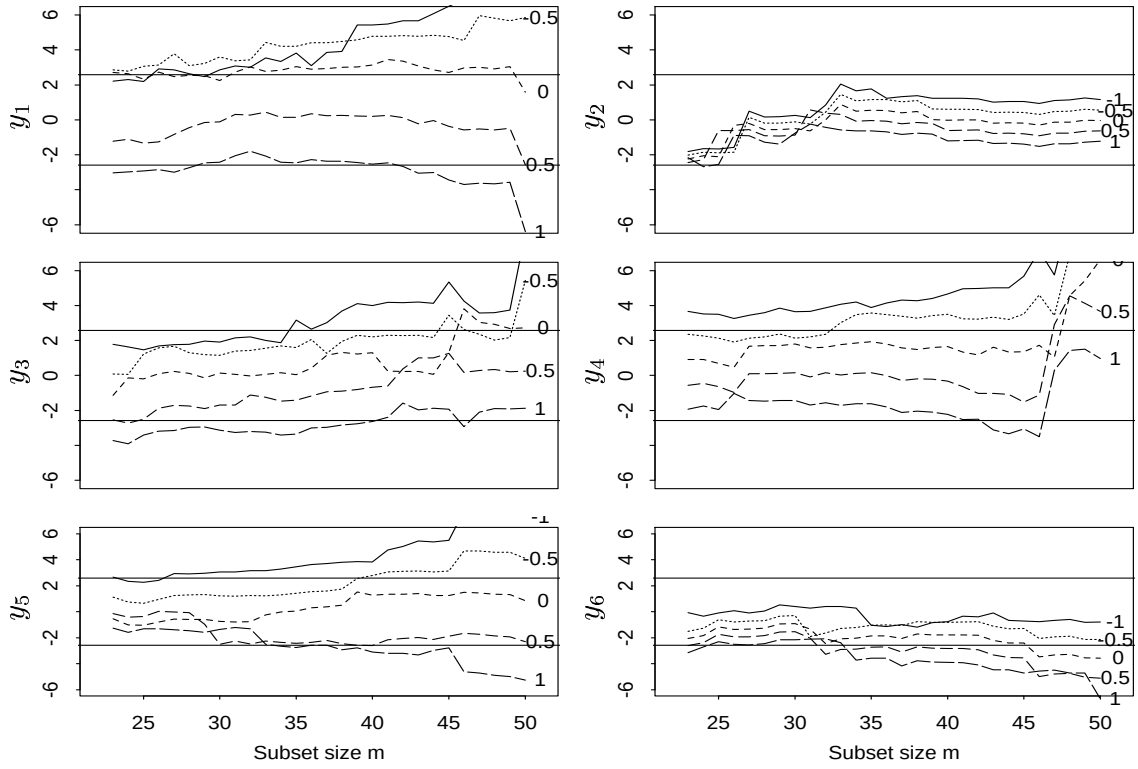


Figure 5: Signed square root of the likelihood ratio test for transformation, confirming $\lambda_R = (0.5, 0, 0.5, 0.5, 0.5, -1, 0, 0, 0.5, 0.5, 1, 0.5, 0.5, 1, 1)$. The numbers by the curves are the values of λ. Due to lack of space we only show the results for the first 6 variables

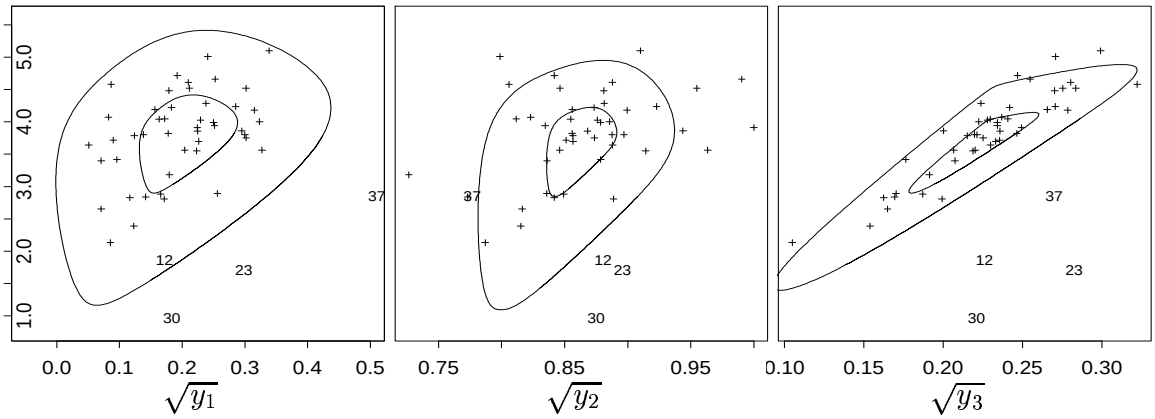


After finding a reasonable set of transformation parameters, we can use a series of forward searches to determine which observations are influential in determining the differ-

ences between the parameter estimates. For each search with each of the 5 most relevant values of the vector $\lambda_k = (-1, -0.5, 0.0.5, 1)$, we use the other 14 values from vector λ_R. We have thus replaced the 5^{15} factorial search by a one variable at a time search with five factors (the λ's), each at five levels. We test the transformation by use of the likelihood ratio for each value of λ_k, which is on one degree of freedom. Calculation of each value of the statistic in the forward search requires a numerical maximization. The fan plots of the 5 forward searches for the first 6 variables are given in Figure 5 (those for the remaining variables are not given for lack of space). Within each panel we give a plot of the signed square root of the likelihood ratio statistic for the usual five values of λ. Use of the signed square root gives plots which cogently illustrate whether lower or higher values of λ are preferred. For example, the top left panel of Figure 5 shows that the value of 0.5 is the only one which is acceptable for y_1. Notice at the end the effect of the outliers which make $\lambda_1 = 0$ acceptable and $\lambda_1 = 0.5$ unacceptable. As concerns y_2, as the flat surface of the likelihood had already suggested (Figure 3), even if the log is the best transformation we cannot reject any value of λ in the interval $(-1, 1)$. Notice that in our suggested approach we not only suggest a transformation which does not suffer from masking and swamping problems, but we are also able to measure the impact of each unit and to illustrate what are the other values of λ which are acceptable for each variable. When multiple values of λ are acceptable we can both use a priori information and choose the value of λ which is in accordance with the other variables belonging to the same group. For example, given that $\lambda = 0.5$ is the only reasonable value of λ for y_1, we choose this value also for y_2 in order to have the same transformation parameter for the two variables referred to *Outcome*.

Finally, our method provides a link between the evidence for a transformation and the scatter plots of the data. For example, Figure 6 shows the scatter diagrams of $\sqrt{y}_4$ versus $\sqrt{y}_1$, $\sqrt{y}_2$ and $\sqrt{y}_3$ with superimposed robust bivariate boxplots (Zani *et al.* (1998)). This

Figure 6: Scatter plots $\sqrt{y}_4 = \sqrt{\text{university current expenditure per student}}$ versus $\sqrt{y}_1 = \sqrt{\text{graduation rate within institutional time}}$, $\sqrt{y}_2 = \sqrt{\text{retention rate}}$ and $\sqrt{y}_3 = \sqrt{\text{faculty/students}}$ with superimposed robust bivariate boxplots. The numbers refer to the 4 outliers



plot clearly shows that the 4 outliers (12 Catania, 23 Modena, 30 Palermo and 37 Roma Tor Vergata) are the Universities which have a combination of values of the first 4 indicators which do not follow the general shape. For example, they have current expenditures per student (y_4) which are much smaller than their corresponding ratio faculty/students

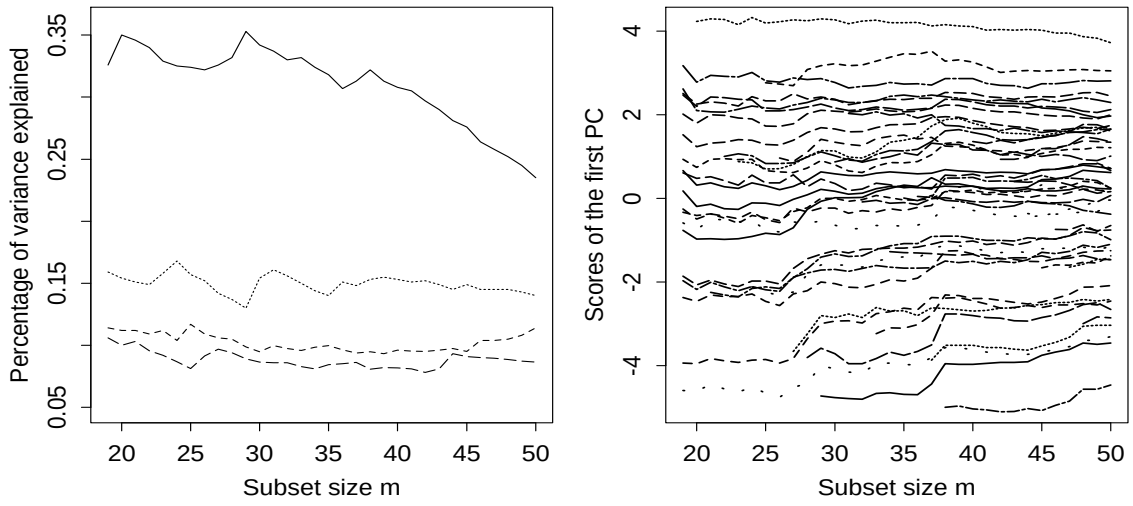
Table 1: Correlation matrix among the first 4 variables. Upper part: original untransformed variables including outliers. Lower part: transformed variables using $\lambda = (0.5, 0.5, 0.5, 0.5)$ and deleting the outliers

	y_1	y_2	y_3	y_4
y_1	1	0.09	0.35	0.11
y_2	0.39	1	0.26	0.27
y_3	0.36	0.35	1	0.65
y_4	0.43	0.36	0.91	1

(y_3). The sources of these observations should then be checked for anomalies and transcription errors. The 3 superimposed boxplots show that the spread of the data in the top right corner of the plots is slightly larger than that in the lower left part. The square root transformation has reduced the larger variability in correspondence of the high values. Finally, in order to show the effects of the outliers on the correlations, in Table 1 we show the $4\times$ 4 block of the correlation matrix referred to the first 4 variables before and after the transformation. The upper part shows the original correlations considering all the 50 units and without transforming the variables. The lower part shows the same correlations computed on the transformed variables after deleting the outliers. All the correlations increase and the average increase is 0.18. This implies that the variance of the extracted principal components will be much larger than that which uses untransformed variables.

Once a reasonable set of transformations has been found, if our purpose is dimension reduction, we can monitor the eigenvalues and the eigenvectors of the covariance and correlation matrices to check the stability of the extracted principal components. Figure 7(left) gives the 4 curves referred to the percentage of variance explained by the first 4 principal components. The trajectories are pretty stable, even if there is a slight de-

Figure 7: Percentage of variance explained by the first 4 principal components using transformed and standardized variables (left). Scores of the first principal component (right)



creasing pattern of the curve referred to the first eigenvalue in the last part of the search. However, this does not affect the ordering of the scores of the first principal component as Figure 7(right) clearly shows. This reflects the strong stability of the elements of the first

eigenvector. As concerns the interpretation of the 4 principal components, the monitoring of the eigenvectors (not given here for lack of space) suggests that the first principal component is positively correlated with all our transformed variables. This means that it can be interpreted as a global performance indicator. As concerns the second principal component, the Universities with high scores in this dimension are those which have high rates of graduation, retention and completion (y_1, y_2 and y_{14}), high faculty per administrative staff ($1/y_6$) and extra regional attraction index (y_{10}), but low dimensions (y_7), research funds and grants (y_5 and y_{12}), enrollment rates of the freshmen (y_8 and y_9), and expenditures for technical staff (y_{15}). High values for the third principal component still identify universities with high values of y_1, y_2 and y_{14}, but differently from before, high ratios of contextual indicators ($y_6 - y_9$) except y_{10}, and completion rate (y_{14}), and low values of resources indicators, grants and self sufficiency ($y_3 - y_6$, y_{12}, y_{13}) and extra regional attraction index (y_{10}). Finally, the fourth principal component is mainly determined (with positive sign) by research funds (y_6) and enrollment rate of the best freshmen (y_8) (with negative sign).

References

Anderson T.W. (1963) *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.

Atkinson (2002) Robust diagnostics for multivariate analysis and classification problems, in: *Proceedings of the XLI Riunione Scientifica della Società Italiana di Statistica*, Milano.

Atkinson A.C. and Riani M. (2000) *Robust Diagnostic Regression Analysis*, Springer, New York.

Biggeri L. and Bini M. (2000) The variability of the italian university cost per student: a multilevel cost function model, in: *ASA 2000, Proceedings of the Section on Statistical Education*, Washington, D.C.: The Association, Indianapolis.

Box G.E.P. and Cox D.R. (1964) An analysis of transformations (with discussion), *Journal of the Royal Statistical Society, Series B*, 26, 211–246.

Cerioli A. and Riani M. (1999) The ordering of spatial data and the detection of multiple outliers, *Journal of Computational and Graphical Statistics*, 8, 239–258.

Ewell P.T. (1999) Linking performance measures to resource allocation: exploring unmapped terrain, *Quality in Higher Education*, 5, 191–208.

Hubert M., Rousseeuw P. and Verboven S. (2002) A fast method for robust principal components with applications to chemometrics, *Chemometrics and Intelligent Laboratory Systems*, 60, 101–111.

Mardia K.V., Kent J.T. and Bibby J.M. (1979) *Multivariate Analysis*, Academic Press, London.

Riani M. and Atkinson A.C. (2001) A unified approach to outliers, influence and transformations in discriminant analysis, *Journal of Computational and Graphical Statistics*, 10, 513–544.

Riani M. and Zani S. (1997) An iterative method for the detection of multivariate outliers, *Metron*, 55, 101–117.

Velilla S. (1995) Diagnostics and robust estimation in multivariate data transformations, *Journal of the American Statistical Association*, 90, 945–951.

Zani S., Riani M. and Corbellini A. (1998) Robust bivariate boxplots and multiple outlier detection, *Computational Statistics and Data Analysis*, 28, 257–270.

Forward search nell'analisi di regressione

Da: Bini M., Bertaccini B. (2004) Forward search nell'analisi di regressione, in Strategie metodologiche per lo studio della transizione Università-lavoro, a cura di E. Aureli Cutillo, pp. 19-36, Cleup, Padova

Forward Search nell'analisi di regressione[1]

Matilde Bini **Bruno Bertaccini**
Dipartimento di Statistica "G. Parenti"
Università degli Studi di Firenze

Riassunto. Le strumentazioni proprie dell'analisi descrittiva o le rappresentazioni mediante i modelli classici di regressione costituiscono le tecniche statistiche cui generalmente si ricorre per studiare varie problematiche in qualunque ambito disciplinare, sebbene spesso non si tenga conto dell'eventuale presenza, tra i dati rilevati, di situazioni atipiche multivariate. Per ovviare a tale inconveniente, che si rivela particolarmente gravoso quando i valori anomali sono numerosi e si "mascherano" a vicenda, è stato recentemente proposto un approccio di analisi innovativo chiamato ***forward search***, robusto ed applicabile con successo anche in situazioni complesse, che parte da un dataset ridotto privo di outliers ed include sequenzialmente le rimanenti osservazioni in base ad una misura sempre crescente di "anomalia" delle stesse (Atkinson e Riani, 2000). L'impiego della *forward search* applicata ai modelli di regressione, consente di ordinare le osservazioni in base al loro grado di aderenza ad un determinato modello e di cogliere l'effetto inferenziale di ciascuna unità sui risultati ottenuti.

Parole chiave: dati anomali, forward search, minimi quadrati mediani, robustezza.

1. Introduzione

La valutazione del sistema formativo superiore - e più in generale di quello universitario - e la sua misura in termini statistici, hanno assunto in questi ultimi anni importanza fondamentale per le vaste implicazioni che questa comporta in ogni ambito politico, economico e sociale. Le analisi svolte in tale direzione utilizzando i metodi e

[1] Il presente lavoro è stato finanziato nell'ambito del PRIN 2002, cofinanziato dal MIUR "Transizioni Università-lavoro e valorizzazione delle competenze professionali dei laureati: modelli e metodi di analisi multidimensionali delle determinanti". Coordinatore nazionale è L. Fabbris, coordinatore del gruppo di Firenze è B.Chiandotto (titolo del progetto dell'unità di ricerca locale "Valutazione del processo formativo universitario, sbocchi professionali e pianificazione dei percorsi formativi: modelli e metodi").

le tecniche proprie della statistica descrittiva, pur fornendo soltanto un quadro interpretativo e di riferimento essenziale, hanno mostrato tutta la loro utilità nel porre in luce gli elementi di criticità del sistema universitario (Bertaccini, 2000; Bini, 1999; Chiandotto e Bertaccini, 2003).

Infatti, com'è noto, i dati medi su singoli indicatori non riescono ad evidenziare gli effetti "netti" dei fattori nella determinazione dei valori assunti dagli indicatori stessi, a causa delle interazioni con le altre variabili. Da qui la necessità di approfondimenti conoscitivi adeguati sul complesso sistema di relazioni che incidono sul fenomeno in esame, ricorrendo all'introduzione di modelli analitici capaci di rappresentare in modo opportuno la realtà oggetto di studio.

Il modello statistico di regressione, come ampiamente dimostrato in letteratura, fornisce una risposta più che soddisfacente a queste necessità. Tuttavia, nelle applicazioni descritte negli articoli successivi (Bertaccini, 2004; Bini, 2004), esso costituirà solo la base di partenza di un'analisi più complessa e completa, svolta ricorrendo all'algoritmo di ***forward search*** proposto da Atkinson e Riani (2000) che, per la prima volta, viene impiegato quale supporto ad eventuali politiche correttive, sia in fase di distribuzione delle risorse che in fase di programmazione dei corsi di studio (numero, tipologia e articolazione), dato che ne è stata riscontrata la validità nell'accertare particolari condizioni di estrema efficacia del titolo conseguito all'interno di più ampi contesti caratterizzati da livelli d'efficacia mediocri, oppure nell'individuare contingenze tali da giustificare livelli particolarmente bassi di abbandono degli studi in ambiti in cui il fenomeno si presenta particolarmente gravoso.

Il presente lavoro intende fornire un approfondimento metodologico del suddetto algoritmo. La procedura parte dall'adattamento del modello ad un gruppo minimale d'osservazioni necessarie a stimarne i parametri, per poi passare a valutare adattamenti a sottoinsiemi sempre più ampi. Il risultato è l'ordinamento delle osservazioni rispetto al loro grado di vicinanza nei confronti del modello supposto. Se il modello concorda con i dati, l'adattamento robusto e quello dei minimi quadrati produrranno risultati simili, sia nelle stime dei parametri sia in quelle degli errori. Ma spesso le stime ed i residui del modello adattato cambiano sensibilmente durante tutto il processo di ricerca. Il monitoraggio di questi cambiamenti e di varie statistiche, usualmente impiegate nell'inferenza dei modelli di regressione, consente la collezione di un insieme d'informazioni in grado non solo di individuare gli *outlier* ma, aspetto ancor più importante, di comprendere il peso che ciascuna ha sull'inferenza del modello.

Il lavoro si articola in cinque paragrafi di cui il secondo introduce le problematiche di stima causate dalla presenza di dati anomali, il terzo illustra le proprietà dei minimi quadrati mediani quale approccio di stima robusto al modello di regressione, il quarto ed il quinto sono dedicati alla presentazione dell'algoritmo *forward search*

rispettivamente nel caso dei modelli lineari classici e alla sua estensione ai modelli lineari generalizzati; infine nel sesto sono riportate alcune considerazioni conclusive.

2. Il problema degli outlier

Il modello di regressione è forse il più importante strumento statistico adoperato nelle applicazioni della statistica alle varie discipline scientifiche. Tra i possibili metodi di stima, il metodo dei minimi quadrati (*Ordinary Least Squares* OLS *regression analysis*) può essere considerato una delle pietre miliari della statistica classica.

Le proprietà possedute dagli stimatori OLS ne legittimano la popolarità, ma non ne giustificano l'abuso che talvolta ne viene fatto, prestando scarsa attenzione sia alla verifica delle ipotesi di specificazione che all'eventuale presenza, tra i dati rilevati, di osservazioni anomale che non sembrano essere generate dal modello ipotizzato. Per inquadrare, allo stesso tempo, semplificare i termini del problema basti ricordare che, nei modelli di regressione, le stime dei p parametri dipendono da p statistiche calcolate su tutte le informazioni rilevate, e se alcune di queste si differenziano in qualche misura dal corpo dei dati, il processo d'adattamento può mascherare tali differenze o, all'opposto, esserne fortemente influenzato.

Convenzionalmente denotati con il termine inglese "*outliers*", i dati anomali possono verificarsi per errori commessi in fase di registrazione, a causa della misura di fenomeni eccezionali, o possono identificare unità appartenenti a popolazioni differenti entrate accidentalmente nel campione. Non è solo la variabile risposta ad essere soggetta ad eventuali irregolarità; in un certo senso, è più probabile che eventuali *outlier* siano presenti in una delle p variabili esplicative, poiché si hanno più occasioni di rilevare dati anomali.

Se i parametri del modello fossero noti, non si avrebbero difficoltà nell'individuazione degli *outlier*, unità che evidenzierebbero i termini d'errore più elevati. Le difficoltà invece emergono dal momento in cui i parametri del modello devono essere stimati tramite un insieme d'osservazioni in cui possono essere presenti unità anomale. Per ovviare alla loro presenza, sono stati recentemente proposti nuovi metodi di stima definiti *robusti* o *resistenti*, così qualificati in letteratura per la loro proprietà di produrre stime non facilmente condizionabili da dati contaminati. Questi metodi concordano nell'identificare come *outlier* le unità che evidenziano i residui più elevati. Se invece il modello venisse adattato ricorrendo alla strumentazione classica, tale tecnica d'identificazione non risulterebbe altrettanto efficace in quanto le anomalie possono influenzare pesantemente le stime e quindi alterare i valori residui.

Un approccio alternativo al problema è quello di avvalersi delle cosiddette *analisi diagnostiche*, che prevedono il calcolo di statistiche in grado di individuare le anomalie e tra queste quelle influenti. Queste possono essere così esaminate, e successivamente eliminate o corrette, in modo da consentire il riadattamento del modello mediante le tecniche classiche (Cook e Weisberg, 1982; Atkinson, 1985). Il grosso limite di queste procedure è però costituito, all'aumentare del numero delle potenziali anomalie, dall'esplosione combinatoria dei possibili sottoinsiemi di volta in volta da esaminare[2].

Altri metodi per l'investigazione contemporanea di più *outlier* si avvalgono di tecniche robuste per l'ordinamento delle osservazioni in base ai valori dei residui. Tra questi una menzione particolare spetta all'algoritmo di *forward search*, che costituisce motivo d'approfondimento di questo lavoro.

3. Un approccio robusto al modello di regressione

Com'è noto, con la tecnica dei minimi quadrati ordinari si perviene ad una stima dei coefficienti $\boldsymbol{\beta}$ minimizzando la somma dei quadrati dei residui. In questo caso la misura di distanza a $(y_i - \hat{y}_i)^2$ adottata attribuisce alle unità con i residui più grandi un peso relativamente maggiore. Pertanto, se si osservano relativamente poche osservazioni con termini di errore ε_i eccezionalmente elevati rispetto al corpo dei dati, si possono avere pesanti ripercussioni sulle stime prodotte, soprattutto in corrispondenza dei cosiddetti *high leverage points*. Questa estrema sensibilità dei minimi quadrati ordinari ha condotto ad identificare con il termine "*robusto*" (Box e Andersen, 1955) quei metodi di stima che continuano a possedere proprietà desiderabili nonostante parte dei dati sia in qualche misura contaminata.

Per formalizzare questo aspetto Donoho e Huber (1983) introducono il concetto di *breakdown point* (in Rousseeuw, 1987). Sia $Z = (X, y)$ di dimensione $n \times (p+1)$ la matrice dei dati e sia T uno stimatore di un parametro θ del modello di regressione; ciò significa che $T(Z) = \hat{\theta}$.

Si considerino ora tutti i possibili campioni contaminati Z' ottenuti rimpiazzando un qualsiasi numero *m* delle osservazioni originarie con valori arbitrari. Si denoti con

[2] Si osservi che gli approcci robusti applicati alla regressione, nonostante si pongano lo stesso obiettivo della diagnostica, procedono secondo un'ottica totalmente opposta, adattando in prima analisi il modello secondo tecniche che rendano giustizia al corpo dei dati, per poi passare all'esame delle unità che si differenziano maggiormente dai valori predetti. Tuttavia, molto spesso gli approcci robusti e le analisi diagnostiche conducono agli stessi risultati.

$bias(m;T,Z)$ la massima distorsione che può essere causata da tale contaminazione, cioè

$$bias(m;T,Z) = \sup_{Z'} \left\| T(Z') - T(Z) \right\|.$$

Se $bias(m;T,Z) \to \infty$ allora significa che *m outlier* possono avere un effetto arbitrariamente grande su *T*, provocando la perdita di funzionalità dello stimatore. Pertanto, il *breakdown point* dello stimatore *T* nel campione Z è definito come

$$\varepsilon_n^*(T,Z) = \min\left\{ \frac{m}{n}; bias(m;T,Z) \to \infty \right\}.$$

In altre parole, il *breakdown point* non è altro che la più piccola frazione di contaminazione che può portare lo stimatore $T(Z')$ ad assumere valori arbitrariamente lontani da *T*(*Z*). Si osservi che in questa definizione non viene fatto alcun riferimento alla distribuzione di probabilità dei dati osservati.

Tra i vari approcci robusti proposti in letteratura, che generalmente concordano nell'identificare come *outlier* le unità che evidenziano i residui più elevati, una menzione particolare spetta agli studi effettuati da Huber e al metodo di stima dei *minimi quadrati mediani* (*Least Median of Squares* LMS - Rousseeuw, 1984), tecnica quest'ultima che verrà ora esaminata in dettaglio per il fondamentale ruolo svolto nella procedura proposta da Atkinson e Riani e presentata nel prosieguo di questo lavoro.

Per il modello di regressione lineare $E(Y) = X\beta$, con *X* di rango pieno *p*, sia *b* una qualunque stima di β. Con *n* osservazioni, i residui calcolati secondo questa stima sono $e_i(b) = y_i - x_i^T b$.

La stima dei *minimi quadrati mediani* $\hat{\beta}_p^*$ è il valore di *b* che minimizza la mediana dei residui quadrati $e_i^2(b)$. Quindi $\hat{\beta}_p^*$ minimizza il parametro di scala

$$\phi(b) = e^2_{[med]}(b) \qquad (1)$$

dove $e^2_{[k]}(b)$ è il *k*-esimo residuo quadrato ordinato e *med* è la parte intera di $(n+p+1)/2$ [3].

[3] Rispetto alla nota formula della mediana *med* = *int* [(n+1)/2], questa formula tiene conto del fatto che i primi p residui ordinati sono zero in quanto il risultato di un modello di regressione con *p* parametri e *p* osservazioni è un'iperpiano passante esattamente per i *p* punti.

La definizione di $\hat{\beta}_p^*$ data in (1) non fornisce alcuna indicazione su come ottenere l'effettiva stima dei parametri. Poiché la superficie da minimizzare ha molti minimi locali, si deve ricorre ad una qualche approssimazione. Rousseeuw (1984) ha propone un'approssimazione di $\hat{\beta}_p^*$ ottenuta tramite una ricerca su insiemi di *p* osservazioni estratti casualmente dalle informazioni a disposizione, la cui formalizzazione sarà illustrata all'inizio del prossimo capitolo. Osservare che non è assolutamente detto che l'iperpiano dei minimi quadrati mediani passi esattamente per *p* punti; in altre parole, il metodo proposto da Rousseeuw costituisce un'approssimazione del reale $\hat{\beta}_p^*$ in quanto il minimo assoluto della (1) può non trovarsi tra le sole *p*-tuple ma in sottoinsiemi di osservazioni più grandi.

È interessante a questo punto osservare che, lo stimatore che soddisfa la (1) ha un *breakdown point* del 50%. In altre parole, per avere ripercussioni sulle stime è necessario che almeno la metà dei dati siano *outlier*; altrimenti, nel caso in cui la percentuale di contaminazione sia inferiore, il metodo LMS sarà comunque in grado, quando *n* è sufficientemente grande, di produrre una stima *non distorta* dell'iperpiano di regressione. Questo è il massimo *breakdown point* tollerabile da un modello di regressione.

Il comportamento molto robusto dello stimatore LMS è in netto contrasto con quello dei minimi quadrati ordinari $\hat{\beta}_{OLS}$ che minimizza la

$$S(b) = \sum_{i=1}^{n} e_i^2(b) .$$

In tal caso, basta che un solo *outlier* assuma valori arbitrariamente molto elevati per causare una variazione sensibile nel valore di $\hat{\beta}_{OLS}$: in questo caso, poiché la frazione 1/*n* tende a zero al crescere dei casi campionati, il *breakdown point* è zero.

4. Forward Search per Modelli Lineari Classici

Il grosso limite delle tecniche diagnostiche per accertare la presenza di *outlier* sta nella loro intrinseca impossibilità di individuare più di 3 o 4 *outlier* contemporaneamente, data l'esplosione combinatoria del numero di sottoinsiemi di volta in volta da considerare. L'algoritmo "*forward search*" proposto da Atkinson e Riani esprime tutta la sua efficacia proprio sopperendo a tali limitazioni, coniugando la capacità propria delle analisi diagnostiche di identificare congiuntamente gruppi di *outlier* alle

proprietà espresse dai metodi di stima robusti, in particolare dai *minimi quadrati mediani*.

I tre passi fondamentali in cui si articola sono:

1. scelta del miglior sottoinsieme iniziale (*starting set*) libero da *outlier*;
2. aggiunta di osservazioni durante la *forward search*;
3. monitoraggio delle statistiche idonee ad individuare gli *outlier* durante l'avanzamento dell'analisi;

che saranno descritti in dettaglio nei prossimi sottoparagrafi.

Nel paragrafo successivo verrà illustrata l'estensione dell'algoritmo al caso dei modelli lineari generalizzati.

3.1 Scelta dell'insieme iniziale

Il miglior sottoinsieme iniziale d'analisi viene individuato ricorrendo all'approssimazione proposta da Rousseuw ai minimi quadrati mediani, che garantisce un insieme di partenza (*starting set*) libero da *outlier*.

Si cercherà ora di dare una definizione formale di questo fondamentale passo dell'algoritmo.

Sia $Z = (X, y)$ di dimensione $n \times (p+1)$. Se *n* non è troppo grande e $p << n$ si è detto che la scelta del sottoinsieme iniziale avviene valutando esaustivamente tutte le $\binom{n}{p}$ distinte *p*-tuple $S^{(p)}_{i_1,\ldots,i_p} \equiv \{z_{i_1},\ldots,z_{i_p}\}$, dove $z^T_{i_j}$ è la i_j-esima riga di Z, con $j = 1,\ldots,p$ e $1 \le i_j \ne i_{j*} \le n$. In particolare, sia $\boldsymbol{\iota}^T = [i_1,...,i_p]$ e sia $e_{i,S^{(p)}_{\boldsymbol{\iota}}}$ il residuo dei minimi quadrati per l'*i*-esima unità dato che il modello di regressione è stato adattato con le sole osservazioni in $S^{(p)}_{\boldsymbol{\iota}}$. Lo *starting set* è dunque individuato dalla $S^{(p)}_{*}$ cha soddisfa la

$$e^2_{[med],S^{(p)}_{*}} = \min_{\boldsymbol{\iota}} \left[e^2_{[med],S^{(p)}_{\boldsymbol{\iota}}} \right] \qquad (2)$$

dove $e^2_{[k],S^{(p)}_{\iota}}$ è il *k*-esimo residuo quadrato ordinato tra gli $e^2_{i,S^{(p)}_{\iota}}$, con $i = 1,\ldots,n$, e *med* è la parte intera di $(n+p+1)/2$. Se $\binom{n}{p}$ è troppo grande, la scelta del sottoinsieme iniziale avviene sempre con lo stesso criterio della (2), ma ricorrendo in alternativa all'estrazione di 3000 *p*-tuple dalla matrice dei dati *Z*.

Quindi, in totale accordo con la proposta di Rousseeuw, la *forward search*, a seconda della dimensione del problema e tenuto conto delle capacità degli attuali elaboratori, ricerca il minimo della (2) su un massimo di 3000 *p*-tuple tra tutte le possibili estraibili da un campione di *n* osservazioni. Qualora il numero di tutte le possibili *p*-tuple fosse inferiore a 3000 la *forward search* effettuerà una valutazione esaustiva di tali sottoinsiemi.

3.2 Aggiunta di osservazioni durante la forward search

Dato un sottoinsieme $S^{(m)}_{*}$ di dimensione $m \geq p$, la *forward search* muove verso il sottoinsieme $S^{(m+1)}_{*}$ selezionando le unità cui corrispondono i primi $m+1$ residui ordinati $e^2_{[k],S^{(m)}_{*}}$. Tale procedura termina nel momento in cui tutte le osservazioni sono state incluse nel sottoinsieme, cioè quando $S^{(m)}_{*} = S^{(n)}$. Per facilitare la comprensione dell'algoritmo, in **Figura 1** sono illustrati i risultati dei vari passaggi della procedura su un insieme esemplificativo di dati.

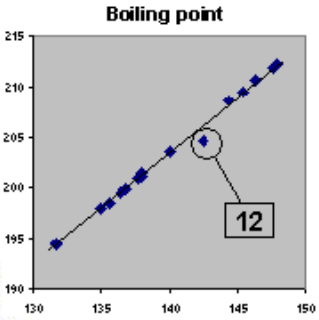
Lo stimatore di *forward search* $\hat{\beta}_{FS}$ è quindi definito dall'insieme degli $n-p+1$ stimatori dei minimi quadrati ordinari ottenuti ad ogni passo della procedura; cioè

$$\hat{\beta}_{FS} = \left(\hat{\beta}^*_p, \hat{\beta}^*_{p+1}, \ldots, \hat{\beta}^*_{n-1}, \hat{\beta}^*_n \equiv \hat{\beta}_n\right).$$

Come già osservato nelle note introduttive di questo capitolo, il passaggio dalla dimensione *m* a quella $m+1$ comporta, nella maggior parte dei casi, l'ingresso di una nuova unità in aggiunta a quelle che già compongono il sottoinsieme costituito al passo precedente. Ma può anche accadere che due o più unità entrino in $S^{(m+1)}_{*}$ mentre una o più escono.

Figura 1. *Indici delle unità che entrano nel sottocampione durante le varie fasi dell'algoritmo di ricerca, applicato ai dati di Forbes (17 osservazioni) sul punto d'ebollizione dell'acqua in relazione a differenti livelli di pressione atmosferica (fonte: Atkinson – Riani, 2000). Le analisi effettuate segnalano chiaramente la forte anomalia della 12° osservazione rispetto al resto dei dati.*

```
> fwdlm(formula = log.Pressure ~ ., data = forbes)
2   4 16
3   4 16 17          unit(s) entered: 17
4   4 17 16 5        unit(s) entered: 5
5   16 17 4 5 7      unit(s) entered: 7
6   16 5 17 4 7 8    unit(s) entered: 8
7   16 17 5 7 8 4 13         unit(s) entered: 13
8   5 16 7 8 17 4 13 6       unit(s) entered: 6
9   5 4 16 17 7 8 13 6 10    unit(s) entered: 10
10  4 16 5 17 7 13 8 6 10 3          unit(s) entered: 3
11  4 16 17 5 6 13 7 8 10 3 2        unit(s) entered: 2
12  6 16 4 17 13 3 5 10 2 7 8 11     unit(s) entered: 11
13  6 16 17 3 4 10 13 2 5 7 8 11 9           unit(s) entered: 9
14  6 16 17 3 10 2 4 13 5 7 8 11 9 15        unit(s) entered: 15
15  6 17 10 16 3 4 2 5 11 13 7 8 9 15 14     unit(s) entered: 14
16  6 10 3 4 17 16 11 2 5 9 7 15 8 13 14 1           unit(s) entered: 1
17  3 10 2 6 17 16 11 9 4 5 13 15 7 8 1 14 12        unit(s) entered: 12
```



L'esperienza empirica ha indotto gli autori che hanno proposto la procedura ad affermare che tale evento, di per sé abbastanza inusuale, tende a verificarsi quando la ricerca include un'osservazione che appartiene ad un gruppo di *outlier*; infatti, al passo successivo, i rimanenti *outlier* del gruppo evidenziano un comportamento meno anomalo e alcuni di loro possono entrare allo stesso momento nel sottoinsieme di analisi. Le simulazioni effettuate dagli stessi autori hanno anche evidenziato la lieve perdita di stabilità del metodo di ricerca se al posto dei residui ordinari vengono impiegati i residui studentizzati.

Da sottolineare infine che l'approccio proposto abbina l'estrema robustezza dei minimi quadrati mediani LMS all'efficienza degli stimatori dei minimi quadrati ordinari. La robustezza dell'algoritmo però non deriva tanto dalla scelta di un particolare stimatore con un alto *breakdown point*, ma dalla progressiva inclusione delle osservazioni in un insieme che, nella fase iniziale, è ritenuto libero da *outlier*. In altre parole, questo metodo non è tanto sensibile alla tecnica impiegata per la selezione dell'insieme iniziale, quanto al fatto che questo sia costituito da osservazioni non anomale o eventualmente contenga *outlier* mascherati che saranno immediatamente rimossi nelle prime fasi d'avanzamento della procedura.

3.3 Monitoraggio del processo

La stima di σ^2 non rimane costante durante le fasi del processo, in quanto ad ogni passo entrano a far parte del sottoinsieme d'analisi le *m* osservazioni con i più piccoli residui, con $m: p+1,\ldots,n$. Quindi, anche in assenza di *outlier*, si ha che $s^2_{S_*^{(m)}} \leq s^2_{S^{(n)}} = s^2$ per $m < n$; generalmente, la curva tracciata da $s^2_{S_*^{(m)}}$ esibisce una prima fase di leggera crescita tipica dell'ingresso di osservazioni che concordano con il modello di regressione ipotizzato, ed una seconda fase di crescita più ripida diretta conseguenza dell'eventuale presenza di *outlier* e del loro peso rispetto al corpo dei dati.

Una rappresentazione grafica molto importante è quella che consente di verificare il comportamento di tutti gli *n* residui ad ogni passo dell'analisi. Valori elevati dei residui in corrispondenza di osservazioni che non appartengono all'insieme d'analisi sono un chiaro segnale della presenza di *outlier*. A causa della forte dipendenza di $s^2_{S_*^{(m)}}$ da *m*, tutti i residui vengono standardizzati rispetto alla media finale dei residui quadrati s^2.

Il grafico dei valori di *leverage* si rivela un ulteriore utile strumento di diagnosi delle osservazioni anomale. Dal momento in cui ogni unità entra a far parte del sottoinsieme $S_*^{(m)}$, ne vengono tracciati i valori di *leverage* $h_{i,S_*^{(m)}}$ assunti nei passi successivi dell'analisi, dove

$$h_{i,S_*^{(m)}} = x_i^T \left(X^T_{S_*^{(m)}} X_{S_*^{(m)}} \right)^{-1} x_i \qquad \text{con} \quad i \in S_*^{(m)} \quad \text{e} \quad m = p,\ldots,n .$$

All'inizio della ricerca il sottoinsieme $S_*^{(p)}$ è formato soltanto da *p* osservazioni ognuna delle quali ha *leverage* pari ad uno. Dopodiché i *leverage* decrescono. Gli *outlier*, che entrano a far parte del sottoinsieme $S_*^{(m)}$ nei passi finali dell'analisi, possono mostrare punti di *leverage* più alti rispetto al resto delle osservazioni, anche se non sono infrequenti situazioni in cui le unità che compongono l'insieme iniziale mostrano i valori più elevati per tutto il tempo dell'analisi.

L'importanza del monitoraggio delle varie fasi della ricerca può essere più facilmente compresa osservando le **Figure 2** e **3**, anch'esse diretto risultato di analisi effettuate su dati esemplificativi.

Figura 2. *Forward plot della stima di σ^2 e dell'indice di determinazione R^2 durante le varie fasi dell'algoritmo di ricerca applicato ai dati di Forbes (cfr. Figura 1), dai quali risulta chiara l'individuazione di un'outlier all'ultimo passo della procedura.*

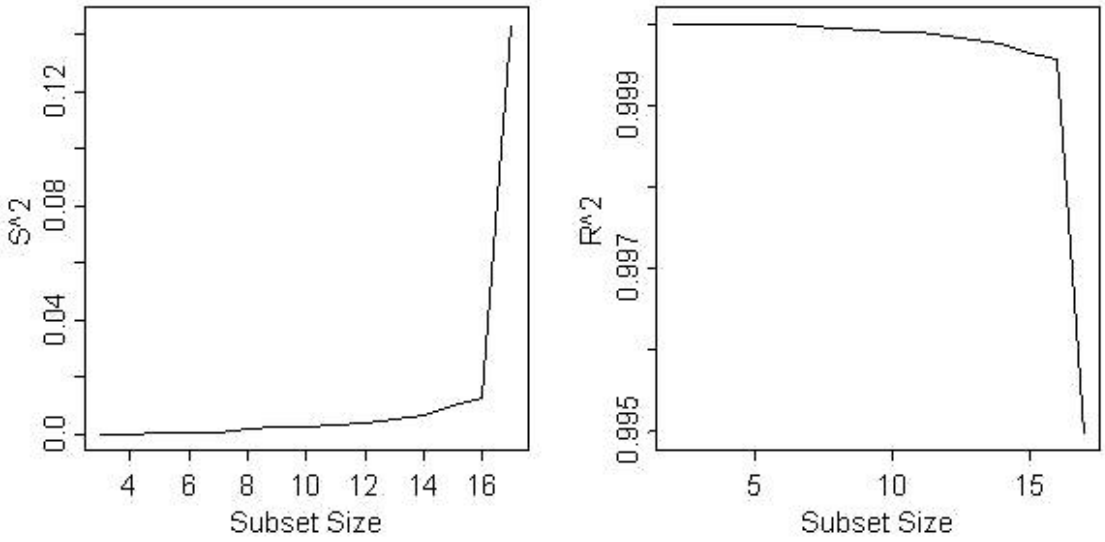
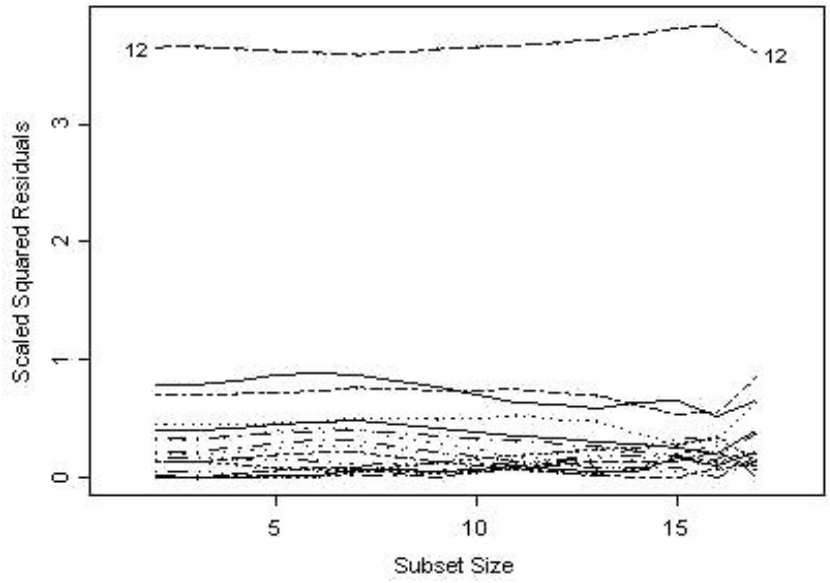


Figura 3. *Forward plot dei residui quadrati scalati durante le varie fasi dell'algoritmo di ricerca applicato ai dati di Forbes (cfr. Figura 1); l'unità outlier 12 mostra, ad ogni passo dell'analisi, un residuo quadrato molto alto rispetto a tutte le altre osservazioni.*



5. Estensione ai Modelli Lineari Generalizzati

L'algoritmo *forward search* può essere esteso al caso dei modelli lineari generalizzati (*Generalized Linear Models* GLMs) che, come sappiamo, sono un'estensione dei modelli lineari classici a variabili di risposta non normalmente distribuite, il cui valore atteso è modellato secondo una qualche funzione (Agresti, 2002).

In particolare, in questo capitolo saranno illustrate le necessarie modifiche relative al caso dei modelli per dati binomiale in virtù delle applicazioni presentate negli articoli successivi di Bertaccini e Bini.

5.1 Richiami ai modelli lineari generalizzati per dati binomiali

Si supponga quindi di disporre di un campione di dati individuali, e che per ogni unità osservata si abbiano due possibili modalità, convenzionalmente indicate genericamente con i termini *successo* e *insuccesso*. Quando N è molto grande si tende generalmente a raggruppare i dati in classi di variabili esplicative. Tale operazione, alquanto agevole quando le covariate sono discrete ed assumono un numero abbastanza limitato di modalità, è in realtà piuttosto delicata quando le variabili assumono un numero finito piuttosto elevato o un'infinità numerabile o non di valori poiché, in tali casi, la suddivisione in classi è generalmente arbitraria e può incidere notevolmente sui risultati del modello. Nel caso di dati raggruppati le risposte assumono la forma

$$R_1/m_1,\ldots,R_n/m_n$$

in cui R_i è la variabile casuale del numero di successi nelle m_i prove indipendenti dell'i-esimo gruppo con $i=1,\ldots,n$. Si ha pertanto che $R_i \sim B(\pi_i, m_i)$, dove le R_i sono indipendenti. Gli interi positivi m_i sono i cosiddetti denominatori binomiali.

L'importanza della classificazione è principalmente dovuta al fatto che le approssimazioni normali risultano appropriate per dati raggruppati; infatti, la teoria asintotica è basata sulla condizione che $m_i \to \infty$, per ogni i e non su $N \to \infty$.

È immediato comprendere che i modelli di regressione classici non sono applicabili a risposte così strutturate in quanto:

1. R_i è Binomiale con indice m_i eventualmente uguale ad uno nel caso individuale, pertanto deve essere abbandonata l'assunzione di normalità;
2. le medie $E(R_i)=m_i\pi_i$ sono funzione delle probabilità di successo. Ricorrendo alla funzione legame identità si rischia che per certi valori dei parametri si ottengano delle probabilità stimate non comprese tra 0 e 1;
3. le varianze $Var(R_i)=m_i\pi_i(1-\pi_i)$ non sono costanti, pertanto deve essere abbandonata l'assunzione di omoschedasticità.

La distribuzione può essere convenientemente riscritta in termini relativi; si ha allora $Y_i = R_i / m_i$, con $E(Y_i) = \pi_i$ e $Var(Y_i) = \pi_i(1-\pi_i)/m_i$. I link più appropriati sono in questo caso rappresentati da funzioni monotone con forma ad S, dominio la retta reale e codominio una misura di probabilità. È infatti plausibile pensare che una variazione nei livelli di una delle covariate abbia un impatto inferiore quando π è vicino a 0 o 1 di quando e' vicino a 0.5.

In questo lavoro, particolare attenzione sarà prestata al *logit link* $g(\mu) = \log\left[\frac{\mu}{1-\mu}\right]$ proprio dei *modelli logistici*, la cui formulazione generica, nel caso di dati raggruppati con covariate qualitative (dette anche fattori), diviene

$$\log\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = \beta_1 + d_2^2\beta_2^2 + \ldots + d_{l_2}^2\beta_{l_2}^2 + d_2^3\beta_2^3 + \ldots + d_{l_3}^3\beta_{l_3}^3 + \\ + d_2^p\beta_2^p \ldots + d_{l_p}^p\beta_{l_p}^p$$

dove

1. x_i indica l'*i*-esima delle *n* possibili combinazioni osservate[4] dei livelli dei fattori $X_1, \ldots, X_p$, con. $i = 1, \ldots, n$;
2. β_h^j è il coefficiente relativo al *h*-esimo livello del *j*-esimo fattore ad l_j livelli, con $j = 2, \ldots, p$ indice di covariata e $h = 2, \ldots, l_j$ indice di livello;
3. d_h^j è la variabile *dummy* originata dalla parametrizzazione dei fattori che assume valore 1 se X_j assume l'*h*° livello e zero altrimenti.

Poiché i GLM richiedono generalmente metodi iterativi di stima dei parametri, l'algebra di *deletion diagnostic* in questo contesto si complica notevolmente. Molte quantità sono ora generate ricorrendo all'algoritmo di adattamento dei minimi quadrati ponderati. Per esempio la *hat matrix H*, nel caso classico definita dalla $H = X(X^TX)^{-1}X^T$, diventa ora

$$H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$$

dove *W* è la matrice diagonale

$$W = diag\{m_1\hat{\pi}_1(1-\hat{\pi}_1), \ldots, m_n\hat{\pi}_n(1-\hat{\pi}_n)\}$$

[4] Notare che non è assolutamente detto, che nelle indagini osservazionali, si disponga di osservazioni in corrispondenza a tutte le possibili combinazioni dei livelli dei fattori, come invece avviene in molti disegni sperimentali, dove però gli esperimenti sono controllati.

contenente le varianze binomiali valutate nel punto di massima log-verosimiglianza.

A differenza dei modelli di regressione classici, nei GLM sono definibili differenti tipi di residui, tra cui i più utilizzati sono:

1. *Residui di Pearson*

   La semplice definizione del residuo dei minimi quadrati ordinari conduce, nel caso dei modelli per dati binomiali dove la varianza di *Y* dipende dalla media, ai residui di Pearson

$$r_{prs} = \frac{(r_i - m_i\hat{\pi}_i)}{\sqrt{m_i\hat{\pi}_i(1-\hat{\pi}_i)}}.$$

   La sommatoria dei quadrati di questi residui conduce alla statistica

$$\mathrm{X}^2 = \sum_{i=1}^{n} \frac{(r_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)}$$

   detta *chi-quadro di Pearson*, frequentemente impiegata nei test sulla bontà d'adattamento del modello. Nell'ipotesi che il modello adattato sia il vero modello generatore dei dati e se $m_i \to \infty$, con $i = 1,\ldots,n$, allora

$$\mathrm{X}^2 \overset{A}{\sim} \chi^2_{n-p}.$$

2. *Residui di devianza*

   La devianza, che generalizza la somma dei quadrati dei residui della regressione ordinaria $S(\hat{\beta}) = \sum_{i=1}^{n} e_i^2$, può essere vista come la somma di *n* quantità

$$D(\underline{\hat{\beta}}) = \sum_{i=1}^{n} d_i^2,$$

   anche se le componenti d_i^2 non sono quadrati di semplici quantità, ma risultano genericamente definiti come

$$r_{d_i} = sign(r_i - m_i\hat{\pi}_i)d_i$$

   dove, dalla definizione della statistica G^2,

$$di = \sqrt{2}\left(r_i \log \frac{r_i}{m_i\hat{\pi}_i} + (m_i - r_i)\log\left(\frac{m_i - r_i}{m_i - m_i\hat{\pi}_i}\right)\right)^{1/2}. \qquad (3)$$

Se gli m_i sono sufficientemente grandi, generalmente i residui di devianza mostrano una distribuzione molto più simile ad una normale di quella dei residui di Pearson.

Tutte le considerazioni finora svolte sono basate sull'ipotesi che i dati osservati siano stati generati da un modello con funzione legame di tipo logistico, che potrebbe però rivelarsi non adeguata allo specifico caso in esame. Può pertanto essere consigliabile verificare la bontà del legame prescelto attraverso un opportuno test.

Si supponga che il legame utilizzato nell'adattamento ai dati sia $g(\mu)$ quando la vera funzione è in realtà $g^*(\mu) = \eta$, dove η è il predittore lineare. Si definisca $h(\eta) = g\left\{g^{*-1}(\eta)\right\}$. Si può allora scrivere che

$$g(\mu) = g\left\{g^{*-1}(\eta)\right\} = h(\eta).$$

Se il legame adottato è corretto, si ha che $h(\eta) = \eta$. Altrimenti $h(\eta)$ è una funzione non lineare in η. Si deve perciò verificare che $g(\mu)$ sia una funzione lineare di η. Lo sviluppo in serie di Taylor di tale funzione in un intorno di zero è data da

$$\begin{aligned} g(\mu) = h(\eta) &= h(0) + h'(0)\eta + h''(0)*\eta^2/2 + \ldots \\ &= \alpha + \delta x^T \beta + \gamma\eta^2 \end{aligned}$$

dove α, δ e γ sono scalari. Poiché β deve essere stimato, la precedente espressione si semplifica in

$$g(\mu) = x^{*T}\beta^* + \gamma\eta^2,$$

a condizione che il modello adattato contenga l'intercetta. Il test di bontà d'adattamento della funzione legame si riduce allora a verificare l'ipotesi nulla $H_0 : \gamma = 0$. Per un approfondimento metodologico di questa procedura di test si rimanda al testo di McCullagh e Nelder (1989).

5.2 Forward Search e modelli per dati binomiali

L'algoritmo della *forward search* nel caso dei GLM per dati binomiali è simile a quello per la regressione classica, ad eccezione del fatto che al posto dei residui dei minimi quadrati vengono impiegati i residui di devianza quadratici d_i^2 definiti nella (3). In questo caso, quindi, la procedura inizia selezionando in modo casuale sottoinsiemi di dimensione *p*, scegliendo quello per il quale il valore della componente mediana della devianza risulta essere il più piccolo.

Quanto detto ha validità generale, tranne nel particolare caso dei modelli per dati binari; infatti, dal momento che, per la variabile risposta, il numero di zero non è uguale al numero di uno, il metodo di stima dei minimi quadrati mediani introdurrà nell'insieme minimo iniziale solo osservazioni con la modalità di risposta più frequente. Perciò occorre modificare sostanzialmente l'algoritmo di ricerca, per mantenere un bilanciamento di entrambi i tipi di risposta durante le varie fasi d'avanzamento della procedura.

6. Conclusioni

Come si è visto, l'algoritmo "*forward search*" esprime tutta la sua efficacia coniugando la capacità propria delle analisi diagnostiche di identificare congiuntamente gruppi di *outlier* alle proprietà espresse dai metodi di stima robusti.

Negli studi relativi alla valutazione dell'efficacia del titolo universitario nei confronti dell'inserimento professionale (Bertaccini, 2004) e all'abbandono degli studi universitari (Bini, 2004), l'impiego di tale tecnica, riuscendo in maniera iterativa ad accertare la presenza di eventuali situazioni anomale o difformità tra le osservazioni che incidono sulla "capacità" rappresentativa del modello, costituisce un ulteriore elemento informativo per intervenire in maniera efficiente ed efficace in sede di attivazione di adeguate politiche di intervento. In altri termini, l'obiettivo principale dell'impiego dell'algoritmo non è stato tanto quello di individuare e successivamente eliminare unità anomale al fine di ottenere un modello più stabile e confacente alla realtà, quanto piuttosto quello di utilizzare le informazioni provenienti dall'esame della composizione strutturale di tali anomalie per condurre analisi approfondite sul fenomeno oggetto di studio, capaci di fornire informazioni utili all'innalzamento della qualità dei processi formativi. Nei due articoli sopra citati è infatti stata riscontrata la validità della metodologia nell'accertare particolari condizioni di estrema efficacia del titolo conseguito all'interno di più ampi contesti caratterizzati da livelli d'efficacia mediocri, oppure nell'individuare contingenze tali da giustificare livelli particolarmente bassi di abbandono degli studi in ambiti in cui il fenomeno si presenta particolarmente gravoso.

Riferimenti bibliografici

Agresti A. (2002). *Categorical Data Analysis* - second edition. Wiley, New York.

Atkinson A. C. (1985). *Plots, Transformations and Regression*. Oxford University Press, Oxford.

Atkinson A. C., Riani M. (2000). *Robust Diagnostic Regression Analysis*. Springer, New York.

Bertaccini B. (2000). *Misure di efficacia esterna dell'istruzione universitaria: indicatori statistici e analisi robusta.* (Tesi di laurea). Università degli Studi di Firenze.

Bertaccini B. (2004). Valutazione del processo di formazione universitaria: un'analisi robusta dell'efficacia. Pubblicato in questo volume.

Bini M. (1999). Valutazione della Efficacia dell'Istruzione Universitaria rispetto al Mercato del Lavoro. Rdr 03/99. Osservatorio per la Valutazione del Sistema Universitario - Ministero dell'Università e della Ricerca Scientifica e Tecnologica.

Bini M. (2004). Valutazione del processo di formazione universitaria: un'analisi robusta degli abbandoni. Pubblicato in questo volume.

Chatterjee S., Hadi A. S. (1988). *Sensitivity Analysis in Linear Regression*. Wiley, New York.

Chiandotto B., Bertaccini B. (2003). "*Profilo e Sbocchi occupazionali dei laureati e diplomati dell'Ateneo fiorentino nell'anno 1999*". Università degli Studi di Firenze.

Cook R. D., Weisberg S. (1982). *Residual and Influence in Regression*. Chapman and Hall, Londra.

Donoho D. L., Huber P. J. (1983). The notation of breakdown point. In *A festschrift for Erich Lehmann*, edited by Bickel P., Doksum K., Hodges J. L. Wadsworth.

McCullagh P., Nelder J.A. (1989). *Generalized Linear Models*. Chapman and Hall, Londra.

Rousseeuw, P. J. (1984). Least Median of Square Regression. *Journal of the American Statistical Association 85*, pp. 633-639

Rousseeuw, P. J. e Leroy A. M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley

Robust multivariate methods for the analysis of the university performance

Da: Bini M. (2003) Robust multivariate methods for the analysis of the university performance, Series Studies in Classification, Data Analysis, and Knowledge Organization, pp. 285-292, Springer-Verlag.

Robust Multivariate Methods for the Analysis of the University Performance

Matilde Bini

Dipartimento di Statistica "G. Parenti"
Università di Firenze, Italy
bini@ds.unifi.it

Abstract. One of the most important problems among the methodological issues discussed in cluster analysis is the identification of the correct number of clusters and the correct allocation of units to their natural clusters. In this paper we use the forward search algorithm, recently proposed by Atkinson, Riani and Cerioli (2004) to scrutinize in a robust and efficient way the output of k-means clustering algorithm. The method is applied to a data set containing efficiency and effectiveness indicators, collected by the National University Evaluation Committee (NUEC), used to evaluate the performance of Italian universities.

1 Introduction

The forward search is a powerful general method for detecting multiple masked outliers and for determining their effect on models fitted to data. The aim of this study is to show how the forward search algorithm can be used to validate the output of a cluster analysis algorithm. The suggested approach enables us to scrutinize in a robust and efficient way the degree of belonging of each unit to its appropriate cluster and the degree of overlapping among the different groups which have been found. As an illustration of the suggested approach, we tackle the problem of the performance university measurement. The data set considered in the analysis includes indicators, which derive from data of the past census survey conducted by NUEC in 2000, concerning 50 public universities of the academic year 1998-99. The variables have been actually defined using the information that each university usually has to collect for the National Statistical System and for the Ministry of Education and Research, and they have been proposed in 1998 by National University Evaluation Committee (NUEC) (ex Observatory for the Evaluation of University System until 1998) as a minimum set of indicators to perform efficiency evaluation of the universities activities (see details in Biggeri and Bini, 2001). Among the large number of the proposed indicators (29), in order to show how the new method works, we have selected only few of them. After the presentation, in section 2, of the data set used and the results of the classical cluster analyses, here conventionally named *"traditional"*, the application of the forward search and the comments of the results obtained are illustrated in section 3. Finally, the last section is devoted to some concluding remarks.

2 The data set and the traditional cluster analysis

Starting from March 2000, the NUEC yearly conducts census surveys with the co-operation of all the 73 Italian universities, to implement a statistical information system useful to monitor the university organization and carry out the requested evaluations. The information collected concern many aspects of the educational service (for the detailed list of variables and the data set, see the web site www.cnvsu.it). On the basis of this information, it is possible to compute a set of indicators (29) for the measurement and the evaluation of the performance of single units which produce this service. The set can be arranged in four classes (Ewell, 1999): **Outcome (output) indicators**, that should inform about the final results and the degree of quality of the teaching and research activities; **Resources indicators**, i.e. indicators of resources as funds, staff, etc. available; **Contextual indicators**, i.e. indicators of the context where the university is working, of the socio-economic environment; **Process indicators**, that should inform about the organization, facilities and results of the teaching and research processes.

To implement the present study on clustering the Italian universities in homogeneous groups of units, the data of the survey conducted in 2000 are used. They include a set of 50 public universities obtained by the exclusion of the private universities, since they did not received the most part of the ordinary resources from the MIUR, and of the universities established less than 6 years ago, because they do not have information available. Considering the number of elementary units and the objective of this study, only the following indicators have been considered: graduation rate within institutional time (X_1) (outcome indicator); faculty/students ratio (X_2), research funds (X_3) and administrative staff per faculty (X_4) (resources indicators); private research grants per single member of faculty (X_5) and expenditure for technical staff per "traditional funds" (X_6) (process indicators). The *"traditional"* clustering techniques, usually, do not assure that units are allocated to the appropriate clusters, and this can lead to the problem of the incorrect assignment of policies to units belonging to "wrong" groups that causes with no doubt side effects and iniquity among interventions. The forward search applied to this algorithm can be used to solve this problem. Before to perform the cluster analysis, the inspection of the scatter plot matrix of the data shows that the distribution of some variables is highly skewed and that maybe some outliers are present in the data. We therefore proceed to estimate the values of the Box-Cox (1964) transformation parameters using the robust procedure described in Riani and Atkinson (2001) (for details concerning the transformation parameters which have been found, see Riani and Bini, 2002). Given that the data appropriately transformed satisfy the multivariate normality assumption, the cluster analysis have been performed using the transformed data. We started according to classical approach using first the hierarchical and then the non-hierarchical methods. Several hierarchical cluster analyses (Krzanowski and Marriott, 1995) are performed on

Groups	X_1	X_2	X_3	X_4	X_5	X_6
G_1	0,3500	0,1309	0,5938	0,6304	0,2787	−0,3780
G_2	−0,8829	−0,8775	−0,2352	−0,4380	−0,5938	−0,3867
G_3	0,5806	0,8375	−0,4459	−0,2600	0,3386	0,8919

Table 1. Centroids of groups from k-means algorithm

data set, using different distances (Euclidean and Mahalanobis) and different linkages (single, complete and average). A reasonable clustering in terms of number of groups and degree of homogeneity among units in each group could be the one having 3 groups of units obtained with Euclidean distances and average linkages. Starting from this result, the study continues with the non-hierarchical k-means algorithm, using Euclidean distances and a starting number of groups equal to three. The method yields three clusters (G1, G2, G3), each one having respectively size of 18, 17, 15 units, and with specific characteristics, as it is outlined by the centroids of groups reported in Table 1: G1 contains universities with high resources performance; into G2 there are universities with low resources and process performance; universities included in G3 have high global performance. Some graphical representations given in Figure 1, that plot units against the first three principal components as axes, allow us a further useful identification of the characteristics of groups, and also enable us to identify the degree of possible overlapping among groups. The components correspond to the 68% of the total variance. In particular, their proportions are respectively equal to 30%, 19.6% and 18.4%.

The degree of overlapping of the different clusters can be highlighted by plotting robust bivariate contours (Zani, Riani and Corbellini, 1998; Rousseeuw, Ruts and Tukey, 1999), containing the 50% of the data for each group which have been found (see Figure 1). They clearly show that the central parts for the 3 groups are separated in the space of the first two principal components, but overlap considerably in the space of the second and third principal components.

As concerns the interpretation of these plots, the correlation histograms suggest that the first component is positively correlated with all variables. This means that it can be interpreted as a global performance indicator. Regarding the second principal component, the universities with high scores in this dimension are those which have high rates of research funds, administrative personnel, expenditure for technical staff, and a lower rate of graduation, but bad performance in terms of private grants (X_5 having negative sign). High values for the third principal component identify universities with very bad profile as concerns the research and private funds (X_3 and X_5 have negative sign) and quite high values of graduation rate and faculty/student ratio.

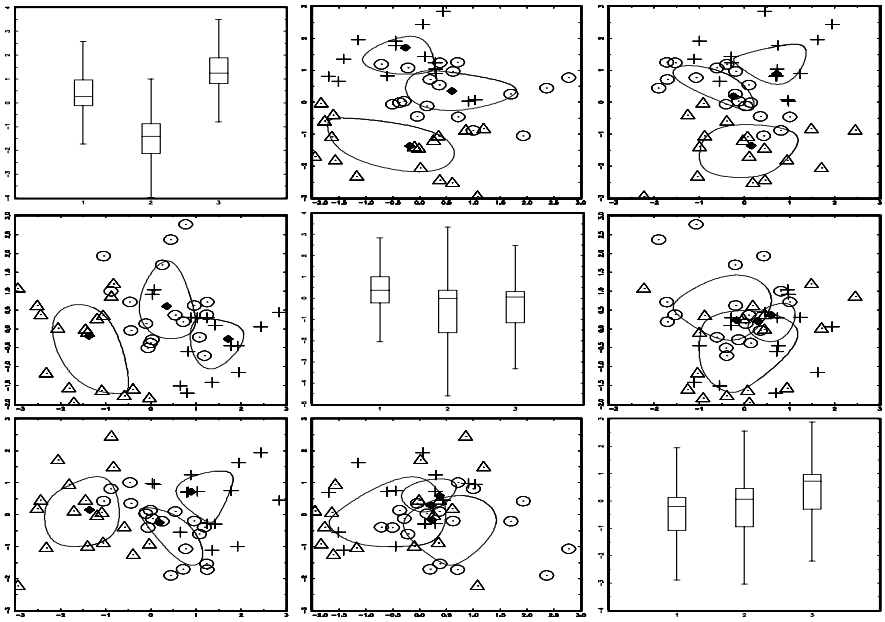


Fig. 1. Position of the units in the space of the first 3 principal components with univariate boxplots (main diagonal) and bivariate contours containing 50% of the data for each group. The black diamond denotes the position of the robust centroid for each group

3 Robust validation of cluster analysis output through the forward search

In most statistical analyses, it occurs that single or groups of observations may affect inappropriately the results obtained using statistical methods. Robust procedures clearly reveal this problem and they solve it by down-weighting or even discarding the influential units from the bulk of data. Very recently, a powerful general procedure based on the *forward search* through the data, as alternative approach to the traditional ones used to detect outliers, has been proposed by Atkinson and Riani (2000) for generalized linear models, and by Atkinson, Riani and Cerioli (2004) for multivariate methods. It is able to identify observations, referred as *outliers*, which are different to the majority of the data, and to determine their effects on inference made about the model or on results from statistical methods. They may be a few units or they may well form a large part of the data and indicate unsuspected structure which is often impossible to be detected from a method applied to all the data. The feature of this new approach is that at each stage of the

forward search it is fundamental to use information such as parameters and plots of Mahalanobis distances to guide to a suitable model.

In the present paper we apply this algorithm (fs) to cluster analysis, but performing in the preliminary analysis the k-means algorithm as alternative method to identify possible clusters, named tentative clusters, to the one adopted by the mentioned authors, and that we briefly summarize as follows: "In the preliminary analysis the data can be explored using scatter plots combined with forward plots of Mahalanobis distances of the units in order to find some tentative clusters. Groups of units are tentatively detected by looking at the behaviour of Mahalanobis distances at seemingly interesting points in the forward search. These often correspond to apparent separations in forward plots of distances, or of peaks in plots such as that of maximum distances of units within the subset..." (see details in chapters 2 and 7 of Atkinson, Riani and Cerioli book).

Hence, our starting point is the output which comes from a k-means cluster analysis using Euclidean distances. As a result we obtained three clusters of sizes 18, 17 and 15. We numbered the units arbitrarily within the groups. Questions of interest include whether the clusters are well separated and whether the units are correctly clustered. In the confirmatory stage of a cluster analysis we used the forward search with a few units in one of the tentative clusters. Let there be m units in the subset. We take as our next subset the units with the m+1 smallest distances. The process continues until all n units are included. During this process we monitor the distances of units of interest. If the units are correctly and strongly clustered, all units in a cluster will have distances that follow similar trajectories to each other. These trajectories will be markedly different from those of units in other clusters. Figure 2 is a plot of the distances for the 17 units of Group 2 from a search that starts by fitting some of the units in Group 2.

We see that unit 20, which is the last to join the group, is far from the other units in Group 2 until it affects the estimated Mahalanobis distances by being used to estimate the parameters. Unit 28 steadily diverges from the group as the search progresses and units from other clusters are included in the subset used for estimation. We can also look at forward plots of distances for units in all groups. Figure 3 plots distances for all units for the search shown in Figure 2. The central three panels, for units in Group 2, separate out, by unit number, the 17 traces of distances we saw one on top of another in Figure 2. In the first panel of the row unit 20 stands out as different, as does, to a lesser extent, unit 28 in the second panel. Otherwise the units seem to have similar traces of their distances, which are generally rather different from those in the top row for units in Group 1. All units in Group 1 seem to have a peak around m = 16 and decline thereafter. Unit 20 looks as if it might belong to this group, although its trace increases at the end. The traces in the last row, for units in Group 3, are again somewhat different, particularly in the second panel, where they decline steadily. We can repeat these plots

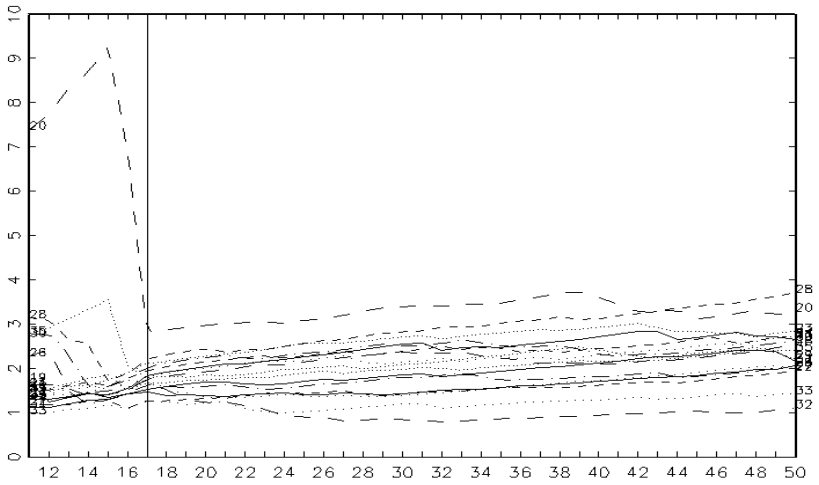


Fig. 2. Monitoring of Mahalanobis distances for the 17 units classified in group 2 by the k-means algorithm

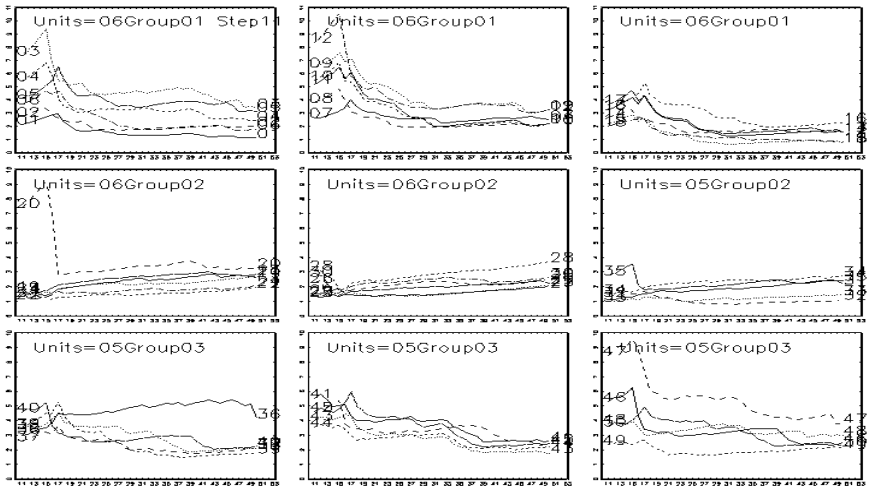


Fig. 3. Forward plot of Mahalanobis distances divided into the 3 groups produced by the k-means algorithm. Each row refers to a group

for searches starting in Group 1 and in Group 3 and so get a clearer idea of which units have distances that vary together and so form a natural cluster. We would indeed present these plots, but instead, we close with Figure 4

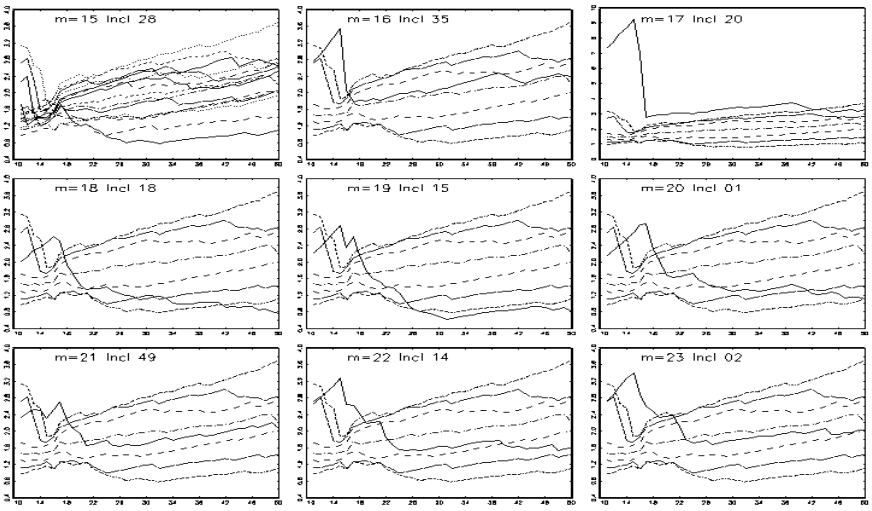


Fig. 4. Mahalanobis distances plots from m=15 for individual units from a search starting in group 2. The plotted percentage points are at 2.5%, 12.5%, 25%, 50% and the symmetrical upper points of the empirical distribution

which shows the trajectories for individual units during the search, starting with units in Group 2. The first panel of Figure 4 shows the distances for the first 15 units to join. These are used to form a reference distribution. In the second panel of the first row we show the distance for unit 35 which joins when m = 16 against a background of the reference distribution. It has a sharp peak at m = 15, just before it joins, which however is as nothing compared to the peak for unit 20, which we have already discussed. The three units in the central row of panels all come from our tentative cluster 1 and all do behave similarly; they have broad peaks around m = 15 and then have distances that decrease markedly. The bottom row shows units 49, 14 and 2, which join as m increases from 21 to 23. These traces are very similar to those in the second row, except that the distances do not decrease to such an extent later in the search. This analysis shows that Group 2 seems to be a coherent cluster, apart from units 20 and 28. However, Figure 4 confirms the impression from some panels of Figure 3 that the current separation between Groups 1 and 3 is not satisfactory.

4 Concluding remarks

Regrettably, we showed only few plots of our example, and no more results about the new clustering obtained, but the purpose of this paper is not to

answer substantive questions about the clustering of units, due to this applied study. Instead it is two fold: 1) to propose the use of k-means algorithm to find some tentative clusters rather than scatterplots combined with the forward plots of Mahalanobis distances of the units, as suggested by Atkinson, Riani and Cerioli (2004); 2) to show how the forward search enables us to explore the characteristics of individual units and so move towards an improved clustering.

References

ATKINSON, A.C., RIANI, M. (2000): *Robust Diagnostic Regression Analysis.* Springer, New York.

ATKINSON, A.C., RIANI, M. and CERIOLI A. (2004): *Exploring Multivariate Data with the Forward Search.*Springer, New York.

BIGGERI, L., BINI, M. (2001): Evaluation at university and state leveling Italy: need for a system of evaluation and indicators. *Tertiary Education and Management, 7, 149–162.*

BOX, G.E.P., COX, D.R. (1964): An analysis of transformations (with discussion).*Journal of the Royal Statistical Society, Series B, 26, 211–246.*

EWELL, P.T. (1999): Linking performance measures to resource allocation: exploring unmapped terrain. *Quality in Higher Education, 5(3), 191–208.*

KRZANOWSKI, W.J. and MARRIOTT, F.H.C. (1995): *Kendall's Library of Statistics 2: Multivariate Analysis, Part 2.* London:Edward Arnold.

RIANI, M., ATKINSON, A.C. (2001): A Unified Approach to Outliers, Influence, and Transformations in Discriminant Analysis. *Journal of Computational and Graphical Statistics, Vol. 10, 513–544.*

RIANI, M., BINI, M. (2002): Robust and Efficient Dimension Reduction. *Atti della XLI Riunione Scientifica della Società Italiana di Statistica,* 295–306. Milano, 5-7 Giugno 2002.

ROUSSEEUW, P.J., RUTS, I. and TUKEY, J.W. (1999): The bagplot: A Bivariate Boxplot. *The American Statistician, Volume 53, Number 4, 382–387.*

YORKE, M. (1998): Performance Indicators Relating to Student Development: Can They be Trusted?. *Quality in Higher Education, 4(1), 45–61.*

ZANI, S., RIANI, M. and CORBELLINI, A. (1998): Robust Bivariate Boxplots and Multiple Outlier Detection, *Computational Statistics and Data Analysis, Vol. 28, 257–270.*

Valutazione del processo di formazione universitaria: un'analisi robusta degli abbandoni

Da: Bini M. (2004) Valutazione del processo di formazione universitaria: un'analisi robusta degli abbandoni, in Strategie metodologiche per lo studio della transizione Università-lavoro, a cura di E. Aureli Cutillo, pp. 57-72, Cleup, Padova

Valutazione del processo di formazione universitaria: un'analisi robusta degli abbandoni

Matilde Bini[1]
Dipartimento di Statistica "G. Parenti"
Università degli Studi di Firenze

Riassunto. L'impiego di procedure robuste nella stima dei modelli di regressione può produrre, quale sottoprodotto dell'analisi, sottoinsiemi di dati anomali, la cui individuazione può fornire informazioni utili alla comprensione di fenomeni peculiari specifici sottosegmenti della popolazione oggetto di studio.

Scopo di questo lavoro è analizzare il problema dell'abbandono degli studenti universitari dell'Ateneo fiorentino alla fine del primo anno di studi. La base dati utilizzata a tal fine è costituita da fonti di tipo amministrativo corrette mediante un'apposita indagine sulle cause d'abbandono degli immatricolati nell'*a.a.* 2001/02. Una volta individuati i fattori che influiscono sullo stato d'iscrizione mediante l'impiego dei modelli lineari generalizzati classici, l'introduzione di una particolare tecnica robusta ha consentito l'identificazione di gruppi di unità anomale dalla cui composizione strutturale possono essere ricavate informazioni utili all'implementazione di politiche accademiche mirate, volte a ridurre il tasso di abbandono al primo anno di studi.

Parole chiave: abbandono degli studi universitari, outliers, forward search.

1. Introduzione

Valutazioni e giudizi con riferimento a persone e istituzioni, processi e risultati costituiscono un'attività che è sempre stata, e viene sempre, svolta in qualsiasi società anche se con modalità informali e a volte molto soggettive. L'attività di valutazione

[1] Il presente lavoro è stato finanziato nell'ambito del PRIN 2002, cofinanziato dal MIUR "Transizioni Università-lavoro e valorizzazione delle competenze professionali dei laureati: modelli e metodi di analisi multidimensionali delle determinanti". Coordinatore nazionale è L. Fabbris, coordinatore del gruppo di Firenze è B.Chiandotto (titolo del progetto dell'unità di ricerca locale "Valutazione del processo formativo universitario, sbocchi professionali e pianificazione dei percorsi formativi: modelli e metodi").

formalizzata, cioè basata su approcci sistematici, si è invece molto sviluppata solo negli ultimi decenni - e ciò è spesso dovuto alle molte leggi e normative che la impongono - divenendo uno strumento irrinunciabile del management dei programmi e delle politiche di intervento in campo economico e sociale e delle attività in genere delle amministrazioni pubbliche, soprattutto laddove si producono *Servizi alla Persona di Pubblica Utilità* (Gori e Vittadini, 1999).

Come sappiamo, gli ultimi anni hanno visto tutta l'amministrazione pubblica italiana evolversi in maniera significativa; in particolare, la nuova normativa riguardante il Sistema Universitario Italiano riconosce al MIUR (Ministero dell'Istruzione, dell'Università e della Ricerca) il compito di definire gli obiettivi principali e le strategie generali di sviluppo del sistema stesso, e di procedere alla sua valutazione, mentre agli Atenei viene riconosciuta ampia autonomia[2], anche se parte dei finanziamenti accordati sono vincolati al soddisfacimento di specifici requisiti. Decentralizzazione, autonomia e finanziamenti vincolati implicano che gli Atenei, che sono i responsabili dei risultati ottenuti dalle unità operative loro afferenti, devono necessariamente svolgere un'intensa ed approfondita attività di autovalutazione sia in termini di misura dell'*efficienza* che dell'*efficacia*.

Ai fini di una loro adeguata misurazione, il *Comitato Nazionale di Valutazione del Sistema Universitario* ha proposto un set minimo di indicatori e di relative variabili necessarie alla loro computazione; tra questi uno dei più importanti risulta il tasso di abbandono, che continua ad assumere valori superiori al 50% in molte realtà universitarie[3] anche dopo la recente riforma dei cicli, con pesanti ripercussioni nella pianificazione delle politiche d'Ateneo ma anche, a livello più ampio, nella società nel suo complesso.

Come la logica farebbe ritenere, studi passati indicano le caratteristiche individuali degli studenti al momento dell'immatricolazione come uno dei più importanti fattori che incidono sulla probabilità di abbandono, anche se le possibili modifiche delle stesse intervenute durante i primi anni di studio, così come la qualità dell'attività didattica, rappresentano aspetti tutt'altro che irrilevanti ai fini dell'analisi del problema. A tale proposito è necessario precisare che la definizione di abbandono assume connotati diversi a seconda che si consideri quale soggetto principale l'Ateneo nel suo complesso (definizione comunemente adottata dagli organi accademici e dai principali istituti statistici) o lo specifico corso di studi di prima immatricolazione. In quest'ultimo caso vengono considerati come abbandoni non solo le interruzioni effettive della carriera ed i trasferimenti verso altri Atenei, ma anche i passaggi tra corsi di studi diversi, effettuati all'interno dello stesso Ateneo. Nella piena consapevolezza

---

[2] L'autonomia finanziaria, manageriale ed organizzativa di ogni Ateneo è stata introdotta dalle leggi 168/89, 537/1993, 59/1997 e 127/1997.

[3] Addirittura, più del 25% degli studenti abbandonano l'università dopo un solo anno di corso.

che anche un semplice passaggio di corso possa di fatto costituire un importante punto di rottura nella carriera di uno studente - dal momento che può comportare una perdita di tempo e di risorse sia per lo studente stesso che, indirettamente, per il corso di studi da lui scelto -, è anche vero che la rilevante quota di coloro che ogni anno effettuano un passaggio interno è verosimilmente da attribuire alla sopravvenuta presa di coscienza dell'effettiva compatibilità tra le proprie attitudini e la reale consistenza dei contenuti didattici offerti dal corso di prima immatricolazione.

Al fine di procedere all'individuazione degli effetti netti delle possibili determinanti del fenomeno in questione si rende necessaria l'introduzione di una modellistica adeguata, "rappresentativa" della realtà che si intende analizzare. Il *modello statistico di regressione*, come ampiamente dimostrato in letteratura, fornisce una risposta più che soddisfacente a queste necessità qualora siano soddisfatte le ipotesi di specificazione su cui è basato. In realtà la scarsa attenzione prestata sia alla verifica di tali ipotesi che all'eventuale presenza, tra i dati rilevati, di osservazioni anomale che non sembrano essere generate dal modello ipotizzato può comportare un certo livello di distorsione dei risultati ottenuti (per una discussione approfondita su aspetti si veda Bini e Bertaccini, 2004 in questo volume). In questo lavoro, tale modello costituirà solo il primo gradino di un moderno approccio all'analisi di regressione denominato *forward search* (Atkinson e Riani, 2000) che, per la prima volta viene impiegato quale supporto ad eventuali politiche correttive, sia in fase di distribuzione delle risorse che in fase di programmazione dei corsi di studio (numero, tipologia e articolazione).

Ricorrendo quindi alle informazioni derivanti dagli archivi amministrativi dell'università, corrette mediante un'apposita indagine svolta dall'Ateno fiorentino sugli immatricolati nell'*a.a.* 2001/02 non più risultanti iscritti allo stesso corso nell'*a.a.* successivo, scopo di questo lavoro è verificare l'idoneità della metodologia sopra citata, i cui aspetti tecnici sono illustrati in Bertaccini e Bini (2004), nel fornire nuove chiavi di lettura nei confronti dello scottante problema degli abbandoni.

Il ricorso a tale approccio, in grado di individuare eventuali anomalie o difformità tra le osservazioni che incidono sulla "capacità" rappresentativa del modello, è in questo caso legittimato dal fatto che potrebbero verificarsi contingenze tali da giustificare livelli di abbandono particolarmente elevati all'interno di più ampi contesti in cui il fenomeno è generalmente contenuto. L'identificazione di queste specifiche situazioni può suggerire l'implementazione di politiche accademiche mirate, volte a contenere il tasso di abbandono dopo il primo anno di studi.

2. La base dati utilizzata

Dalle fonti amministrative dell'Ateneo fiorentino risulta che nell'*a.a.* 2002/03 si sono verificati 2908 casi d'abbandono dell'iniziale corso scelto che costituiscono quasi il 30% delle 10053 immatricolazioni registrate nell'anno accademico precedente.

Date le pesanti conseguenze che questo fenomeno comporta nella pianificazione dell'attività didattica ai vari livelli d'ateneo, l'Università degli Studi di Firenze ha condotto nel giugno del 2003 una rilevazione sulle possibili cause dell'abbandono[4] su tutti gli studenti che si sono immatricolati nell'*a.a.* 2001/02 e che, nel corso dei controlli effettuati durante tutto il primo anno di studi, sono risultati nelle seguenti condizioni:

- *Passati* (P) ad un altro corso di laurea presso l'Università di Firenze;
- *Trasferiti* (T) ad un'altra università;
- *Rinunciatari* (R), ovvero coloro che hanno presentato domanda di rinuncia agli studi;
- *Impliciti* (I), ovvero coloro che non risultano iscritti all'*a.a.* 2002/03 e non rientrano nelle precedenti categorie;
- *Sospesi* (S), per es. per svolgere il servizio di leva; tali studenti possono essere considerati nella categoria degli impliciti;

Le interviste, della durata massima di 5 minuti, sono state condotte ricorrendo alle tecniche *C.A.T.I.* (*Computer Aided Telephone Interviewing*) che, com'è noto, consentono una riduzione sensibile dei tempi d'indagine ed il raggiungimento di elevati tassi di risposta rispetto ad altre modalità di intervista (dei 1839 individui effettivamente contattati data la veridicità delle informazioni anagrafiche, la partecipazione all'indagine è stata del 97,7%).

Le informazioni rilevate tramite il questionario sugli abbandoni hanno consentito la rettifica (per errori o ritardi di registrazione da parte delle segreterie studenti) di parte degli stati d'iscrizione presenti negli archivi amministrativi d'Ateneo. Pertanto, la base dati completa include informazioni sul profilo anagrafico (sesso, età, residenza), sugli studi secondari superiori (tipo e voto di maturità), sulla facoltà e corso di studi di prima immatricolazione, sullo status occupazionale e la posizione nei confronti degli obblighi di leva al momento dell'iscrizione, nonché lo stato d'iscrizione corretto e le eventuali motivazioni che hanno causato l'abbandono[5].

---

[4] L'indagine è stata condotta grazie ad i finanziamenti provenienti dai progetti *CAMPUS*ONE e *OUTCOMES.*

[5] Mediante l'indagine è stato possibile collezionare informazioni sulle cause dell'abbandono per motivi imputabili all'organizzazione didattica o per motivazioni personali ovvero sulle cause del trasferimento ad altro corso e/o facoltà dell'Ateneo fiorentino o altro Ateneo.

3. Modelli per il tasso d'abbandono e *forward search*

Dal questionario sulle cause dell'abbandono risultano evidenti le due prospettive secondo le quali può essere condotta l'analisi: considerare l'abbandono come fenomeno esclusivamente legato all'organizzazione dei corsi di studi o, in un'ottica più ampia, determinato anche dalle politiche organizzative e dalla capacità attrattiva dell'Ateneo. A causa della rilevante quota di coloro che, ogni anno, effettuano un passaggio interno, si è scelto di utilizzare quale variabile di risposta un indicatore dello stato d'iscrizione che assume i valori:

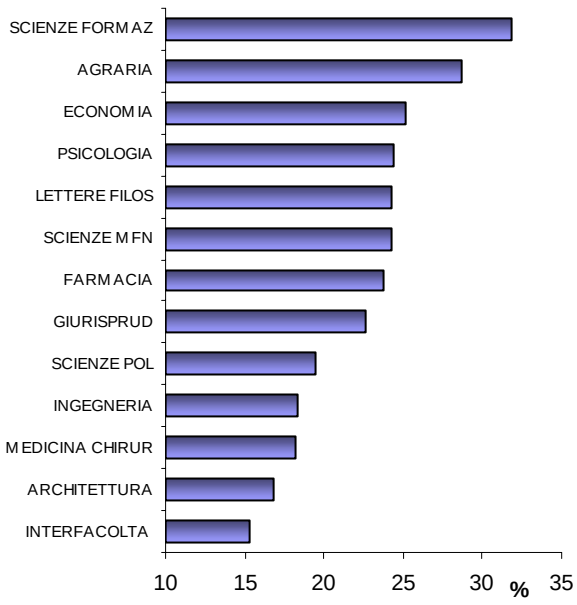
$Y = 1$, se l'individuo ha abbandonato il corso d'immatricolazione e non risulta passato, nell'anno accademico successivo, a nessun altro corso attivo presso l'Ateneo fiorentino (gli stati di riferimento contemplati dal questionario sono "*trasferito*", "*rinunciatario*", "*implicito*" e "*sospeso*");

$Y = 0$, se l'individuo risulta nell'anno accademico successivo ancora iscritto al corso iniziale, è passato o ha fatto domanda di passaggio ad altro corso dell'Ateneo fiorentino.

La **Tavola 1** illustra la distribuzione dei valori medi della variabile risposta sul contingente d'analisi per facoltà.

Tavola 1. *Distribuzione del tasso di abbandono nel contingente d'analisi, per facoltà*

Facoltà	***Response Y***	
	Immatr.	**% (*Y*=1)**
AGRARIA	*358*	28,7
ARCHITETTURA	*1065*	16,8
ECONOMIA	*1029*	25,2
FARMACIA	*197*	23,8
GIURISPRUDENZA	*777*	22,6
INGEGNERIA	*889*	18,3
LETTERE E FILOSOFIA	*1518*	24,3
MEDICINA E CHIRURGIA	*717*	18,3
PSICOLOGIA	*1429*	24,4
SCIENZE DELLA FORMAZIONE	*462*	31,8
SCIENZE MFN	*649*	24,2
SCIENZE POLITICHE	*450*	19,5
INTERFACOLTA	*513*	15,3
Totale	***10053***	22,3



Una volta definita la variabile di risposta ed individuati i fattori da utilizzare nel ruolo di variabili esplicative, è possibile ricorrere all'algoritmo di *forward search* per l'analisi dei casi anomali seguendo due diversi approcci:

1. si può condurre l'analisi su dati aggregati[6] secondo i livelli dei fattori individuati, in modo da costruire un modello logistico per dati binomiali, ed in tal caso i dati anomali individuati saranno costituiti da gruppi di soggetti con certe caratteristiche comuni ed eventualmente appartenenti alla stessa facoltà. L'analisi della composizione strutturale di questi gruppi potrebbe ad esempio evidenziare contingenze tali da giustificare tassi d'abbandono modesti in più ampi contesti in cui il fenomeno dell'abbandono risulta particolarmente gravoso;
2. si può altresì condurre l'analisi su dati individuali, e allora i possibili dati anomali saranno costituiti da individui con particolari caratteristiche che giustificheranno le differenze con i valori predetti dal modello. Questo approccio però non garantisce l'identificazione di gruppi di soggetti con caratteristiche comuni prefissate (ad esempio, soggetti appartenenti ad una stesso corso di studi), anche se è possibile effettuare l'analisi su tutto il gruppo di unità il cui ingresso nel modello non produce incrementi significativi nel valore della devianza spiegata.

Per valutare l'impatto dell'organizzazione didattica delle varie facoltà quale possibile causa dell'abbandono, si è scelto di condurre le analisi secondo il primo degli approcci sopra descritti.

Preliminari analisi di tipo descrittivo (Giusti, 2004) hanno consentito di individuare le covariate, tra tutte quelle presenti nel dataset, che indicano una certo livello d'associazione con la variabile risposta. Di queste, solo la facoltà ed il corso di studi risultano strettamente relate alla qualità dell'organizzazione dei percorsi formativi, mentre le altre sono pertinenti al profilo degli individui (cfr. **Tavola 2**).

Le omissioni presenti nell'archivio amministrativo in corrispondenza delle variabili correlate al fenomeno oggetto di studio hanno comportato una riduzione del contingente d'analisi a 9007 studenti, senza peraltro alterare la distribuzione dei tassi d'abbandono illustrata in Tavola 1.

[6] Si parla di dati aggregati quando si procede al raggruppamento di tutte quelle osservazioni che manifestano gli stessi livelli dei fattori d'interesse.

Tavola 2. *Elenco delle variabili esplicative che si sono rivelate correlate con la variabile risposta*

Fattore	Descrizione	Livelli
DEGREE	Facoltà	1 = 'AGRARIA' 2 = 'ARCHITETTURA' 3 = 'ECONOMIA' 6 = 'FARMACIA' 7 = 'GIURISPRUDENZA' 8 = 'INGEGNERIA' 9 = 'LETTERE E FILOSOFIA' 10 = 'MEDICINA E CHIRURGIA' 11 = 'SCIENZE DELLA FORMAZIONE' 12 = 'SCIENZE POLITICHE' 13 = 'PSICOLOGIA' 14 = 'SCIENZE MFN' 15 = 'INTERFACOLTÀ'
COURSE	Corso di studi	104 livelli
SEX	Sesso	1 ='Maschio'; 2 ='Femmina'
COUNTY	Regione di residenza	1 = 'Firenze - Hinterland' 2 = 'Altri comuni delle province di Firenze e Prato' 3 = 'Altre province della Toscana' 4 = 'Altre regioni del Centro Nord' 5 = 'Sud e Isole'
HSCHOOL	Tipo maturità	1 = 'Classica'; 2 = 'Scientifica' 3 = 'Tecnica'; 4 = 'Altra maturità
HSSCORE	Voto alla maturità	1 = '60 - 56' 2 = '55 - 51' 3 = '50 - 46' 4 = '45 - 41' 5 = '40 - 36'
AGEENROLL	Età d'immatricolazione	1 = 'Meno di 20' 2 = '20' 3 = '21 - 25' 4 = 'Oltre 25'
WORK	Stato occupazionale all'immatric.	1 = 'Occupato' 2 = 'Non occupato'

Tutte le variabili rilevate che si sono mostrate correlate con il fenomeno oggetto di studio sono state inserite come possibili predittive[7] in un modello logistico per dati binari, automatizzato con l'opzione *backward elimination*. Si osservi che la funzione

[7] Dall'elenco delle possibili variabili esplicative, sono state eliminate a priori quelle variabili che dimostravano una forte associazione con la facoltà di provenienza. Inoltre, sono state ripartite in classi le variabili di tipo numerico come il voto medio agli esami e alla laurea. La perdita d'informazione collegata a questa trasformazione, se per un verso provoca conseguenze negative del tutto irrilevanti ai fini dell'analisi che si vuol condurre, per l'altro comporta notevoli vantaggi sia sul versante computazionale, sia sul piano dell'interpretabilità dei risultati.

di questo modello è puramente esplorativa, poiché vi si ricorre per riuscire ad individuare il sottoinsieme di variabili esplicative dal più alto potere discriminante, e che tutte le covariate inserite sono di tipo categorico. La **Tavola 3** ne illustra i risultati principali.

Tavola 3. *Principali risultati del modello logistico per dati binari*

The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	784.3454	22	<.0001
Score	811.3674	22	<.0001
Wald	729.1014	22	<.0001

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
AGEENROLL	3	352.3030	<.0001
COUNTY	4	78.6473	<.0001
DEGREE	12	82.7623	<.0001
HSSCORE	3	101.8830	<.0001

Sulla base dei livelli dei fattori individuati in questa fase preliminare, che risultano essere la *facoltà*, la *residenza* l'*età all'immatricolazione* ed il *voto alla* maturità, si procede al raggruppamento degli individui prestando attenzione alle numerosità dei gruppi generati. Com'è noto, se i denominatori binomiali non sono sufficientemente grandi, le usuali statistiche sulla bontà d'adattamento del modello non convergono ad una distribuzione nota. Inoltre è assai probabile che gruppi molto piccoli, peraltro dallo scarso potere informativo, comportino un effetto di disturbo nelle varie fasi dell'algoritmo di ricerca[8]. Pertanto si è imposto il vincolo che i gruppi costituiti contengano più di cinque individui.

Il modello per dati binomiali adattato al nuovo dataset produce i risultati illustrati in **Tavola 4**. I gruppi costituiti sono 454.

[8] Per esempio, in un gruppo di 2 individui i possibili valori della variabile di risposta sono 0: nessun successo; 0,5: un successo; 1: due successi, ed è pertanto probabile che questi valori vengano identificati come anomalie rispetto al corpo dei valori predetti dal modello. Analogo discorso può essere fatto per gruppi di 3, 4 o 5 individui.

Tavola 4. *Principali risultati del modello logistico per dati raggruppati in base ai livelli dei fattori inizialmente individuati*

```
Summary of
glm(formula = y ~ County + Degree + AgeEnroll + HSscore,
    family = binomial(link = "logit"), data = abnd03, weights = enr)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.8730178  -0.8601809   0.0091846   0.7433140   2.9480225

Coefficients:
              Estimate Std. Error z value  Pr(>|z|)
(Intercept)   -1.298283   0.155187 -8.3659 < 2.2e-16 ***
County2        0.256007   0.081815  3.1291 0.0017534 **
County3        0.342038   0.073441  4.6573 3.203e-06 ***
County4        0.654752   0.108170  6.0530 1.422e-09 ***
County5        0.708523   0.098155  7.2184 5.260e-13 ***
Degree2       -0.593890   0.167855 -3.5381 0.0004030 ***
Degree4       -0.189403   0.275305 -0.6880 0.4914699
Degree5       -0.115360   0.171284 -0.6735 0.5006290
Degree6       -0.164708   0.170176 -0.9679 0.3331091
Degree7       -0.177111   0.155681 -1.1376 0.2552667
Degree9       -0.603253   0.179289 -3.3647 0.0007663 ***
Degree10       0.143121   0.174118  0.8220 0.4110916
Degree11      -0.342498   0.200154 -1.7112 0.0870500 .
Degree49      -0.295448   0.156798 -1.8843 0.0595299 .
Degree52      -0.045730   0.163929 -0.2790 0.7802749
Degree56      -0.274399   0.187682 -1.4620 0.1437303
Degree58      -0.606601   0.204628 -2.9644 0.0030326 **
AgeEnroll2     0.552379   0.078484  7.0381 1.949e-12 ***
AgeEnroll3     0.958635   0.079691 12.0293 < 2.2e-16 ***
AgeEnroll4     1.262348   0.114303 11.0438 < 2.2e-16 ***
HSscore2      -0.374010   0.063411 -5.8982 3.676e-09 ***
HSscore3      -0.710940   0.097891 -7.2626 3.797e-13 ***
HSscore4      -0.983368   0.110800 -8.8752 < 2.2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Number of Fisher Scoring iterations: 3

Null deviance: 1126.89  on 453  degrees of freedom
Residual deviance:  495.73  on 431  degrees of freedom
AIC: 1688.75

Analysis of Deviance Table
Terms added sequentially (first to last)

             Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                           453   1126.889
County        4   61.712       449   1065.178 1.2666e-12
Degree       12   81.135       437    984.042 2.5057e-12
AgeEnroll     3  383.222       434    600.820 9.5323e-83
HSscore       3  105.091       431    495.730 1.2492e-22
```

La tavola di analisi della devianza di questo modello evidenzia nel complesso un adattamento più che soddisfacente (la devianza residua assume un valore di poco superiore ai gradi di libertà residui); evidenti sono anche gli effetti dei fattori nonché dei singoli livelli, se si eccettuano quelli relativi ad alcune facoltà che non mostrano effetti significativi sulla risposta a causa dei valori elevati assunti dagli errori standard.

Una volta individuato il modello su cui lavorare si procede all'applicazione dell'algoritmo della *forward search*, un particolare approccio alla regressione la cui idea base è quella di ordinare le osservazioni campionarie in base ad una misura sempre crescente di "anomalia" delle stesse (la metodologia è presentata in Bini e Bertaccini, 2004).

Tra i risultati più interessanti di quest'analisi assume particolare rilevanza il grafico della statistica test sulla bontà di adattamento della funzione legame (cfr. **Figura 1**). I valori della statistica test evidenziano un andamento decrescente nella parte finale dell'algoritmo di ricerca al di fuori dei limiti di significatività[9], causa la presenza di gruppi di individui che differiscono sensibilmente dal corpo dei dati. Pertanto già da quest'analisi si evince che gli ultimi 98 gruppi sarebbero meritevoli di approfondimenti conoscitivi riguardo alla loro composizione.

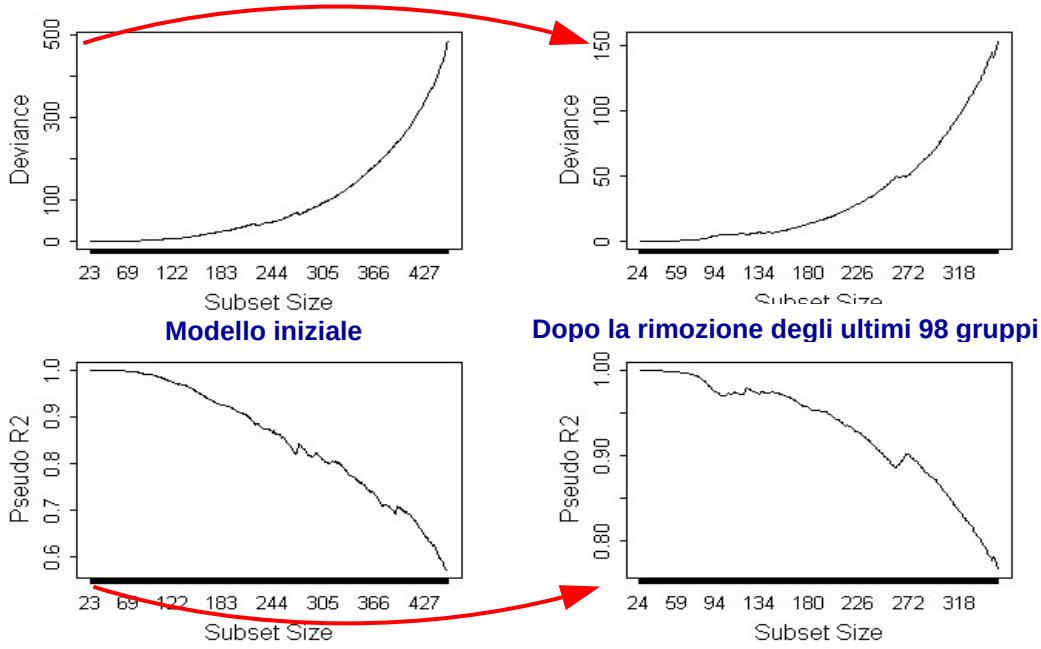
Figura 1. *Forward Search: test sulla bontà della funzione legame*

[9] Il test è strutturato in maniera tale che valori statisticamente uguali a zero ($1-\alpha = 95\%$), propendono per la bontà della funzione scelta.

La presenza di osservazioni con comportamento anomalo emerge anche dai grafici relativi ai valori assunti dalla devianza spiegata e dall'indice di determinazione R^2. Gli ultimi 98 gruppi causano un incremento esponenziale della devianza residua (da 150 a 500) ed un forte decremento dei valori assunti dalla statistica Pseudo-R^2 (da 0.80 a 0.60).

Sebbene non rientri tra gli obiettivi di questo lavoro, può essere interessante comprendere come si sarebbero modificati i risultati una volta rimossi dall'analisi i gruppi che entrano nella fasi finali della ricerca (cfr. **Figure 2** e **3**).

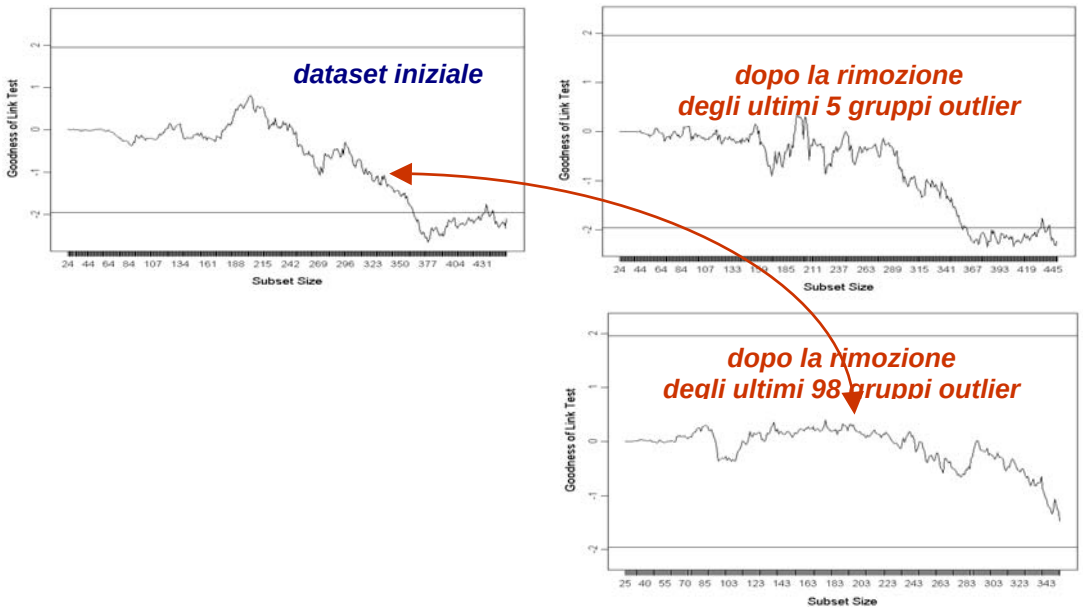
Figura 2. *Forward Search: devianza del modello prima e dopo l'esclusione dei gruppi anomali*



Il contributo in termini di adattamento e di stabilità della statistica test sulla bontà della funzione legame che questa operazione d'esclusione ha comportato è del tutto evidente.

Una volta identificati i gruppi anomali è, come detto, necessario passare all'analisi della loro composizione. Dato lo scopo dimostrativo di questo lavoro, ci limiteremo ad esaminare due particolari circostanze che illustrano situazioni antitetiche del fenomeno degli abbandoni.

Figura 3. *Forward Search: test sulla bontà della funzione legame prima e dopo l'esclusione dei gruppi anomali*



L'ultimo gruppo che entra nell'analisi, etichettato con il numero 270, è composto 12 individui di Medicina e Chirurgia di cui 11 iscritti al DU in Infermeria. I tassi di abbandono sono pari all'83,3% per quanto riguarda il gruppo (10 studenti su 12), e all'81,8% per quanto riguarda il corso (9 su 11), valore quest'ultimo molto elevato se confrontato con quello medio calcolato sul totale immatricolati nel corso di diploma, pari al 27,4% (cfr. **Tavola 5**).

Il fenomeno trova verosimile giustificazione nelle caratteristiche che delineano il profilo del gruppo: trattasi di particolari individui con un'età all'immatricolazione nettamente superiore ed un giudizio alla maturità più basso rispetto ai valori medi generali espressi dal corso di diploma in questione. Inoltre, il gruppo risulta composto da individui tutti residenti in Toscana, ma al dì fuori delle province di Firenze e Prato mentre, per l'intero corso, la percentuale di studenti con tale profilo di residenza è pari al 27,4%. Le analisi condotte fanno perciò ritenere che il fenomeno dell'abbandono sia imputabile più alle caratteristiche individuali degli studenti che all'organizzazione didattica del corso. Comunque, le informazioni supplementari a disposizione dall'indagine sulle cause di abbandono hanno consentito di riscontrare la veridicità di tali conclusioni.

Tavola 5. *Analisi della composizione del cluster 270*

Composizione del gruppo (12 individui / 10 abbandoni)

Corso	Obs	Variabili	valori medi
DU Infermeria (92% del gruppo)	11	Età immatricolazione	32.54
		Voto alla maturità *(36 - 60)*	37.90
		Y (tasso abbandono)	***83.3%** - gruppo* ***81.8%** - corso*

Nota (*residenza*):

tutti immatricolati residenti in Toscana ma fuori delle province di Firenze e Prato

Caratteristiche del corso

Corso	Obs	Variabili	valori medi
DU Infermeria	197	Età immatricolazione	21.73
		Voto alla maturità *(36 - 60)*	42.13
		Y (tasso abbandono)	***27.4%***

Nota (*residenza*):

Immatr residenti in Toscana ma fuori delle province di Firenze e Prato: 29,4%

Il terzultimo gruppo che entra nell'analisi, etichettato con il numero 62, mostra una situazione diametralmente opposta a quella rilevata nel caso precedente. Il gruppo risulta composto da 47 individui immatricolati in vari corsi di Scienze MFN, tutti residenti nei comuni dell'hinterland fiorentino ed iscritti all'università immediatamente dopo la maturità conseguita peraltro con un giudizio medio di 58,9; di questi, nessuno abbandono il corso di prima immatricolazione (cfr. **Tavola 6**).

Per contro, l'analisi del profilo medio degli iscritti nella facoltà di Scienze MFN, svela un età media d'immatricolazione pari a 21,1 ed un voto medio alla maturità pari a 45,3. Pertanto, anche alla luce del tasso generale di abbandono riportato da questa facoltà che, dopo le limitazioni imposte al contingente d'analisi precedentemente discusse, risulta pari al 19,3%, si può senza ombra di dubbio concludere che i caratteri distintivi dei soggetti di questo gruppo contribuiscono certamente al proseguimento dell'esperienza universitaria.

Tavola 6. *Analisi della composizione del cluster 62*

Composizione del gruppo (47 individui / 0 abbandoni)

Corso	Obs	Variabili	valori medi
Immatricolati In vari corsi della facoltà di SMFN	47	Età immatricolazione	18.9
		Voto alla maturità *(36 - 60)*	58.9
		Y (tasso abbandono)	***0.0%***

Nota:
Immatricolati tutti provenienti dai comuni dell'hinterland fiorentino

Caratteristiche della facoltà

	Obs	Variable	Mean
Scienze MFN	109	Età immatricolazione	21.09
		Voto alla maturità *(36 - 60)*	45.27
		Y (tasso abbandono)	***19.3%***

Nota:
Immatricolati tutti provenienti dai comuni dell'hinterland fiorentino: (32% del totale facoltà)

4. Note conclusive

Le analisi svolte in questo contesto e la constatazione di un determinato livello di performance della formazione universitaria, costituiscono un utile supporto per la pianificazione di interventi ed azioni sull'organizzazione delle strutture ma soprattutto dell'attività didattica.

Tuttavia, per effettuare un'analisi più approfondita sul complesso sistema di relazioni e fattori che influenzano il tasso di abbandono degli studi universitari, è necessario ricorrere all'impiego di appropriati modelli analitici.

Le diagnostiche robuste applicate all'analisi della regressione, sono capaci non soltanto di fornire una risposta a questa necessità, come del resto anche i modelli di regressione basati su metodi di stima classici, ma anche di identificare unità o gruppi di unità (osservazioni anomale) aventi particolari caratteristiche. Queste ultime rappresentano una fonte di informazione utile per la definizione di nuovi programmi di didattica, allo scopo di migliorarne la qualità, e conseguentemente, di ridurre il tasso di abbandono degli studi.

Riferimenti bibliografici

Atkinson A. C., Riani M. (2000). *Robust Diagnostic Regression Analysis*. Springer, New York.

Bertaccini B. (2000). *Misure di efficacia esterna dell'istruzione universitaria: indicatori statistici e analisi robusta.* (Tesi di laurea). Università degli Studi di Firenze.

Bertaccini B., Bini M. (2004). Forward search nell'analisi di regressione. Pubblicato in questo volume.

Biggeri L. (2000). Valutazione: idee, esperienze, problemi. Una sfida per gli statistici. *Atti della XL Riunione Scientifica della Società Italiana di Statistica*.

Biggeri L., Bini M. (2001). Evaluation at university and state level in Italy: need for a system of evaluation and indicators. *Tertiary Education and Menagement 7*, 149-162

Giusti C. (2004). *L'abbandono degli studi nell'Ateneo fiorentino: evoluzione nel periodo 1980-2000 e applicazione di un modello gerarchico non lineare agli immatricolati nell'anno accademico 2001/02.* (Tesi di laurea). Università degli Studi di Firenze.

Gori E., Vittadini G. (1999). La valutazione dell'efficienza ed efficacia dei servizi alla persona: impostazione e metodi. In Gori e Vittadini, a cura di (1999), "*Qualità e Valutazione nei Servizi di Pubblica Utilità*". Etas, Milano

L’abbandono degli studi universitari

Da: Chiandotto B., Giusti C. (2005) L’abbandono degli studi universitari, in Modelli statistici per l’analisi della transizione Università-lavoro, a cura di C. Crocetta, pp. 1-22, Cleup, Padova.

L'abbandono degli studi universitari[1]

Bruno Chiandotto **Caterina Giusti**
Dipartimento di Statistica "G. Parenti" - Università degli Studi di Firenze

Riassunto. Nel lavoro si analizza il fenomeno dell'abbandono degli studi, una delle maggiori criticità del sistema universitario italiano. Per cercare di individuare le possibili determinanti del fenomeno è stata svolta un'analisi dei dati individuali relativi agli studenti immatricolati presso l'Ateneo fiorentino nel ventennio 1980-2000 e nell'a.a. 2001/02. Su questi ultimi dati è stato applicato un *modello di regressione logistica con intercetta casuale a due livelli* per valutare l'effetto "netto" esercitato sia dai fattori individuali che da quelli istituzionali (variabili relative ai corsi di studio). Tale modello tiene conto del fatto che gli studenti (unità di primo livello) risultano naturalmente aggregati in Corsi di Laurea (unità di secondo livello). Nelle analisi sono stati considerati abbandoni non solo le interruzioni effettive della carriera universitaria ed i trasferimenti verso altri Atenei, ma anche i passaggi tra Corsi di Laurea.
Parole chiave: Abbandoni universitari, Analisi per coorti, Modelli multilivello, Regressione logistica multilivello.

1. Introduzione

Negli ultimi decenni il sistema universitario italiano si è caratterizzato, all'interno del panorama internazionale dell'istruzione terziaria, per la presenza ed il progressivo aggravamento di una serie di situazioni di particolare criticità; infatti, "qualunque sia la misura presa a riferimento, a partire dalle risorse finanziarie fino al numero dei laureati, il sistema universitario italiano appare debole ed arretrato, in sostanza non ancora *europeo*"(Associazione TreeLLLe, 2003).

[1] Il presente lavoro è stato finanziato nell'ambito del PRIN 2002, cofinanziato dal MIUR "Transizioni Università-lavoro e valorizzazione delle competenze professionali dei laureati: modelli e metodi di analisi multidimensionali delle determinanti". Coordinatore nazionale è Luigi Fabbris, coordinatore del gruppo di Firenze è Bruno Chiandotto (titolo del progetto dell'unità di ricerca locale "Valutazione del processo formativo universitario, sbocchi professionali e pianificazione dei percorsi formativi: modelli e metodi").
L'idea iniziale, la struttura e l'impostazione del lavoro sono dovuti al contributo di entrambi gli autori, mentre le elaborazioni e l'implementazione del modello vanno attribuite a C. Giusti.

Un tale stato di cose non dipende tanto da una minore quota, rispetto agli altri Paesi europei, di giovani diplomati che decidono di intraprendere gli studi universitari, quanto piuttosto dal fenomeno degli abbandoni: mediamente negli ultimi anni più del 25% degli studenti ha lasciato l'Università in Italia dopo un solo anno di corso, percentuale che s'incrementa notevolmente, come si avrà modo di verificare in seguito, negli anni successivi al primo (MURST, 1998).

Il fenomeno degli abbandoni, tipico del sistema universitario italiano, appare ancora più accentuato se si analizza la situazione dell'Università di Firenze; ciò induce a presumere che le conclusioni di un approfondimento conoscitivo su questo fenomeno utilizzando i dati fiorentini possano essere ragionevolmente estese anche a gran parte degli altri Atenei italiani.

Riguardo ai dati utilizzati si deve precisare che l'unità statistica di riferimento considerata non è l'intero Ateneo ma il singolo corso di studi; pertanto, sono stati considerati abbandoni non solo le interruzioni effettive della carriera universitaria ed i trasferimenti verso altri Atenei, ma anche i passaggi tra Corsi di Laurea effettuati all'interno dell'Università di Firenze. La motivazione alla base di tale scelta è che anche un semplice passaggio può costituire, di fatto, un importante punto di rottura della carriera universitaria di uno studente, comportando spesso una perdita di tempo e di risorse, sia per lo studente stesso sia per il corso di studi da lui scelto, del tutto simile a quella caratterizzante l'abbandono degli studi universitari.

L'individuazione delle possibili determinanti del fenomeno degli abbandoni dovrebbe suggerire interventi finalizzati alla eliminazione di una tale criticità[2].

Il secondo paragrafo di questa nota è dedicato ad una sintetica illustrazione dei risultati dell'analisi finalizzata all'individuazione dell'eventuale influenza esercitata sull'esito degli studi universitari sia dal Corso di Laurea che da caratteristiche individuali, quali genere, residenza, diploma di scuola superiore, ecc.[3], relativamente agli immatricolati presso l'Università di Firenze negli anni accademici dal 1980/81 al 2000/01.

Nel terzo paragrafo vengono riassunti, altrettanto sinteticamente, i risultati della medesima analisi condotta relativamente ai 10053 studenti immatricolati presso l'Università degli Studi di Firenze nell'a.a. 2001/02[4], anno in cui è entrata in vigore la riforma dei cicli e degli ordinamenti didattici dell'Università italiana.

---

[2] Sul problema della valutazione dei processi formativi finalizzata all'eliminazione di eventuali criticità presenti nel sistema si veda Chiandotto B. (2002).

[3] Una trattazione più dettagliata si trova in Giusti C. (2004); un altro significativo contributo sull'argomento è stato fornito da Bulgarelli G. (2002).

[4] Anche in questo caso si tratta di un'esposizione estremamente sintetica; maggiori dettagli si trovano in Giusti C. (2004). Conviene in ogni caso precisare che i dati considerati per le analisi dei primi due paragrafi provengono dall'archivio amministrativo dell'Università degli Studi di Firenze e sono stati messi a disposizione dall'Ufficio Servizi Statistici e Controllo di Gestione dell'Ateneo. Attraverso tali informazioni si è proceduto a classificare come "abbandoni" gli studenti che nei periodi di tempo

Per pervenire alla misura dell'effetto "netto" eventualmente esercitato da possibili determinanti (sia individuali che istituzionali) del fenomeno degli abbandoni si è fatto ricorso, facendo sempre riferimento agli immatricolati dell'a.a. 2001/02, ai modelli gerarchici o di regressione multilivello che, com'è noto, hanno la principale caratteristica di tenere in considerazione la struttura gerarchica dei dati oggetto di studio. I risultati delle analisi condotte sono riportati nel quarto paragrafo; alcune sintetiche conclusioni completano la nota.

2. Esito degli studi universitari degli immatricolati nell'Ateneo fiorentino nel periodo 1980-2000

In questo paragrafo viene offerto un quadro descrittivo[5] dell'esito degli studi degli immatricolati presso l'Ateneo di Firenze negli anni accademici tra il 1980/81 ed il 2000/01, rivolgendo particolare attenzione al fenomeno degli abbandoni.

Per analizzare tutte le informazioni disponibili si sarebbero potuti seguire due principali approcci: l'analisi "per contemporanei" o "trasversale" e l'analisi "per coorti" o "longitudinale". In questa sede è stato adottato l'approccio longitudinale; scegliendo come evento di comune origine l'immatricolazione presso l'Università di Firenze in un determinato anno accademico, sono state individuate all'interno della popolazione oggetto di studio 21 distinte coorti.

A tale proposito bisogna osservare che l'analisi degli esiti delle carriere mette in evidenza il principale difetto dell'approccio per coorti, ovvero la possibilità di valutare solamente i dati meno recenti, cioè quelli che si ottengono dopo aver osservato ogni coorte per un certo numero di anni, in modo che ciascun individuo abbia avuto il tempo di "sperimentare" il suo esito finale. In realtà il fenomeno degli abbandoni "colpisce soprattutto gli iscritti ai primi due anni che, insieme, raccolgono più della metà delle mancate reiscrizioni complessive" (Istat, 2003); infatti, il Ministero dell'Istruzione, dell'Università e della Ricerca, nell'indagine sull'abbandono universitario condotta nel 2001, ha scelto di quantificare tale

---

considerati risultavano aver effettuato un passaggio di corso, un trasferimento ad altro Ateneo, aver presentato domanda di rinuncia agli studi o non aver rinnovato l'iscrizione nello stesso Corso di Laurea ("abbandoni impliciti").

[5] Gli studenti immatricolati, ovvero "iscritti per la prima volta al primo anno di un Corso di Laurea o di Diploma Universitario" secondo la definizione dell'Istat, sono stati classificati in base al Corso di Laurea di prima iscrizione; per tali studenti si dispone di informazioni classificabili in "variabili d'ingresso" (principalmente dati anagrafici e relativi agli studi preuniversitari), "di soggiorno" (per esempio informazioni su eventuali passaggi di corso, rinunce) e "d'uscita" (esito finale degli studi). Le variabili d'ingresso e quelle "in itinere" rappresentano i fattori individuali, o variabili esplicative, che si suppone possano influenzare l'esito e la durata degli studi. Tali informazioni risultano aggiornate, per ciascuna delle unità di analisi, al 31 luglio 2003.

fenomeno proprio attraverso il numero di abbandoni tra il primo ed il secondo anno di corso. L'applicazione di tale metodologia d'analisi ha consentito, pertanto, di valutare il fenomeno dell'abbandono per tutte le 21 coorti considerate senza alcuna censura, dal momento che il tempo minimo di osservazione risultava pari a tre anni (coorte 2000/01).

Se si considerano i soli 116841 studenti che si sono immatricolati dall'a.a. 1980/81 al 1993/94, ovvero le generazioni per le quali si dispone di almeno dieci anni di osservazione, risulta immediatamente evidente come il fenomeno dell'abbandono del Corso di Laurea di prima immatricolazione assuma nell'Ateneo fiorentino dimensioni alquanto preoccupanti; la percentuale media di studenti che abbandonano il proprio CdL durante il primo anno è infatti pari al 27.8% (cfr. Figura 1), cioè a più di un quarto del totale degli iscritti. Se si considerano gli abbandoni nei primi due anni, la quota di studenti che lasciano il proprio corso sale al 39.3%, mentre la percentuale degli abbandoni nei primi tre anni risulta pari al 45.2%.

A dieci anni di distanza dall'immatricolazione, si osserva una percentuale media di studenti laureati nel Corso di Laurea di immatricolazione pari ad appena il 30.5% del contingente iniziale, mentre la quota complessiva di abbandoni è pari al 56.8% delle matricole; una frazione non trascurabile di studenti (12.7%) risulta infine ancora iscritta allo stesso CdL dopo dieci anni di carriera universitaria.

Analizzando le percentuali di abbandono in ogni singolo anno (cfr. Figura 2), si ottiene conferma del fatto che il fenomeno della mancata reiscrizione nel Corso di Laurea di prima immatricolazione riguardi prevalentemente il primo anno ed il secondo anno di corso.

Nella Figura 3 sono riportate le percentuali medie di abbandono, calcolate su tutto il ventennio, nei primi due anni di corso. Rispetto ad un valore medio di Ateneo pari al 39.7%, si osservano valori molto elevati per le Facoltà di Economia (46.3%), Scienze Politiche (46.2%), SMFN e Scienze della Formazione (entrambe con valori attorno al 45.5%).

All'opposto i valori più bassi si osservano per Architettura (28.7%) e Medicina e Chirurgia (24.5%). La netta caratterizzazione di queste due Facoltà fa pensare che le limitazioni al numero massimo di immatricolazioni, esistenti a Firenze proprio per questi due indirizzi di studio (anche se non nell'intero ventennio), possano influenzare la probabilità di abbandono; infatti, poiché è necessario superare una prova di ammissione prima di potersi iscrivere, si può supporre che gli studenti che vi riescono siano più motivati rispetto a quelli che si immatricolano ad un Corso di Laurea ad accesso libero.

Figura 1. *Esito degli studi per i dieci a.a. successivi a quello di immatricolazione (valori %).*

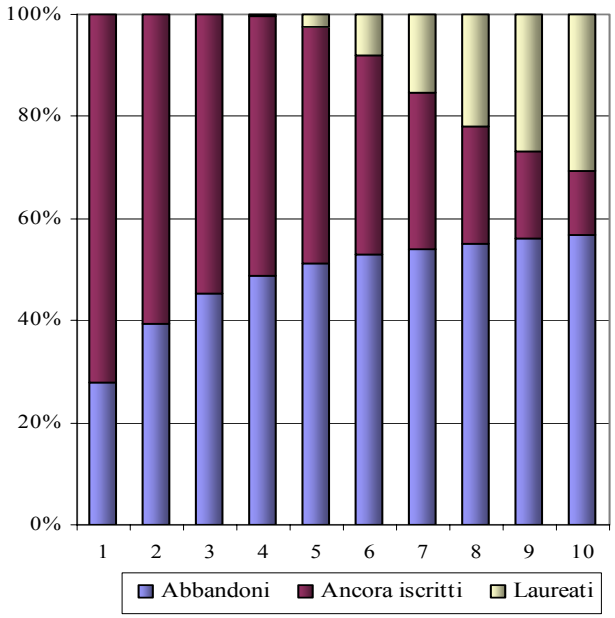
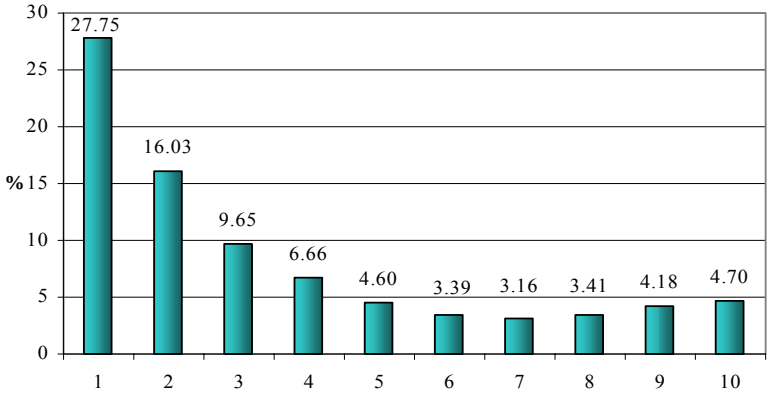


Figura 2. *Abbandoni nel periodo 1980-2000, per anno di corso (% medie).*

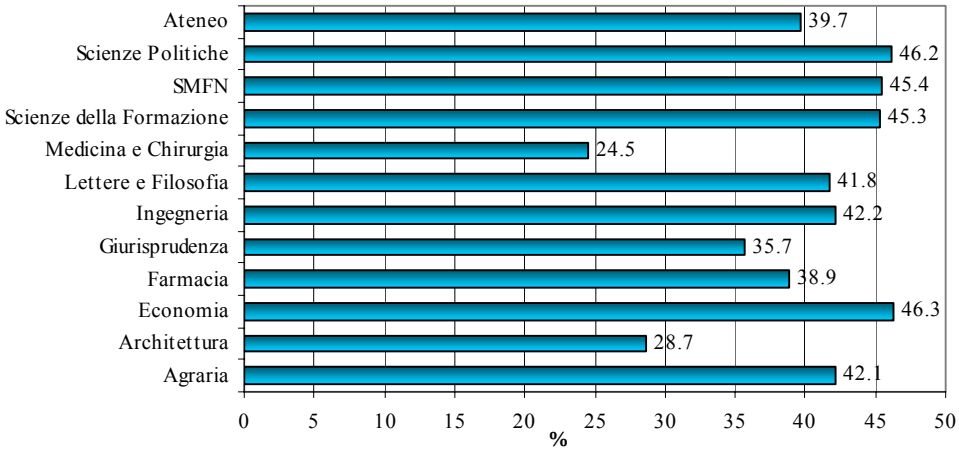


Studiando il fenomeno dell'abbandono più dettagliatamente[6], si osservano percentuali relative a Corsi di Laurea appartenenti alla stessa Facoltà anche molto diverse tra loro; l'analisi condotta a livello di Facoltà risulta perciò, in un certo senso, "distorta", in quanto i valori di Facoltà rappresentano una media tra tutti i CdL e non consentono di cogliere le differenze esistenti fra i vari indirizzi di studio. Stato di

[6] Per i dati relativi ai singoli Corsi di Laurea si rimanda a Giusti C. (2004).

fatto questo che suggerisce, naturalmente, un'analisi del fenomeno degli abbandoni a livello di Corso di Laurea.

Figura 3. *Tasso medio di abbandono dopo due anni nel periodo 1980-2000, per Facoltà.*



Nel processo teso all'individuazione delle possibili determinanti del fenomeno degli abbandoni risulta di una certa utilità misurare il grado di associazione tra la proporzione di studenti che hanno abbandonato gli studi nei primi due anni di corso e altri caratteri ritenuti rilevanti ai fini dell'analisi condotta. Si è pertanto proceduto al computo degli indici *V* di Cramer e *Chi-quadro* di Pearson; i risultati delle elaborazioni effettuate sono riportati nella Tabella 1.

Dall'esame dei dati riportati nella tabella si rileva un discreto livello di associazione fra l'esito degli studi entro 2 anni dall'immatricolazione e la Facoltà di appartenenza dello studente; ancor più significativo risulta però il dato relativo alla relazione fra esito e Corso di Laurea, il che sottolinea ancora una volta come condurre un'analisi al solo livello di Facoltà comporti necessariamente una perdita d'informazioni. Non viene, invece, evidenziata alcuna relazione tra il sesso e l'esito degli studi.

Il tipo di studi preuniversitari svolti ed il voto conseguito risultano fortemente connessi al tasso di abbandono; infatti, la percentuale media di studenti che decidono di abbandonare entro due anni gli studi nel Corso di Laurea di immatricolazione è pari al 29.4% tra i liceali, mentre sale notevolmente tra i diplomati presso istituti tecnici e professionali, per i quali tale quota è pari rispettivamente al 52.7% e 60.2%. Tali valori risultano pressoché stabili per tutte le coorti considerate.

Per quanto riguarda il voto alla maturità, si rileva che passando da una classe di voto a quella superiore il tasso medio di abbandono diminuisce di quasi due punti percentuali.

Tabella 1. *Statistiche d'associazione: periodo 1980 - 2000.*

Caratteri: **esito dopo 2 anni** *vs*	*Statistica*	*Valore*	*GdL*	*Prob*
Anno d'immatricolazione	Chi-quadro	481.1453	20	<.0001
	V di Cramer	0.0526		
Facoltà	Chi-quadro	3514.2521	10	<.0001
	V di Cramer	0.1421		
Corso di Laurea	Chi-quadro	4635.6176	41	<.0001
	V di Cramer	0.1632		
Sesso	Chi-quadro	274.1894	1	<.0001
	V di Cramer	-0.0397		
Voto di Maturità	Chi-quadro	4033.9811	12	<.0001
	V di Cramer	0.1542		
Tipo di Maturità	Chi-quadro	9210.3959	3	<.0001
	V di Cramer	0.2324		
Residenza	Chi-quadro	246.1852	5	<.0001
	V di Cramer	0.0376		
Regolarità studi superiori	Chi-quadro	10507.1918	5	<.0001
	V di Cramer	0.2457		
Ritardo immatricolazione	Chi-quadro	6377.0199	7	<.0001
	V di Cramer	0.1933		

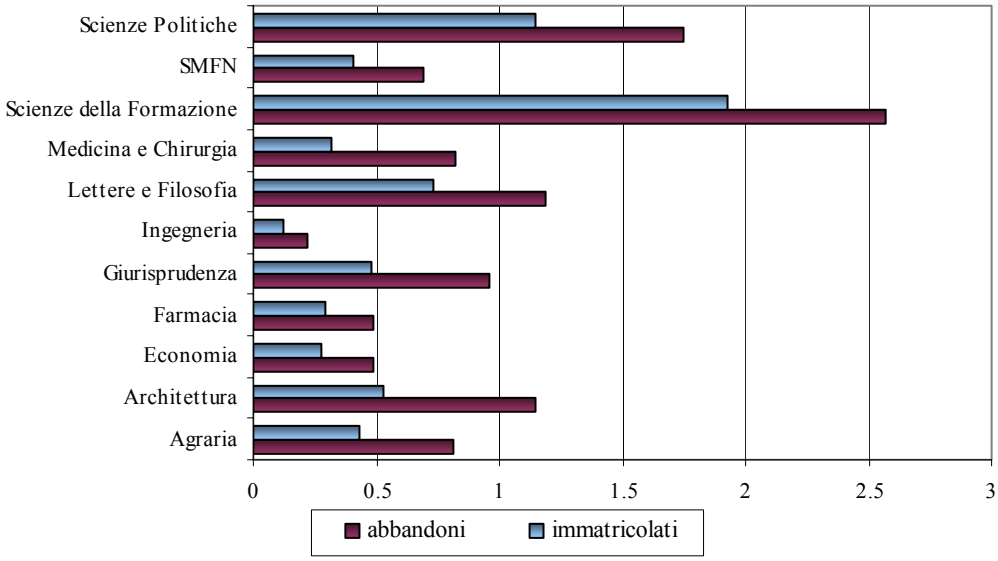
Meno significativa risulta invece l'analisi del tasso di abbandono condotta distinguendo gli studenti delle varie coorti in base alla propria residenza, come mostrato anche dai valori delle due statistiche calcolate. L'unica rilevante differenza che si osserva è quella tra il dato relativo ai residenti fuori regione ed i toscani: mentre per i primi la percentuale media di abbandoni è circa del 37%, per i residenti a Firenze o in una qualsiasi delle altre province toscane è leggermente superiore, attorno al 40.2%.

Per quanto riguarda l'influenza di eventuali "irregolarità" sperimentate nel percorso scolastico sull'esito degli studi universitari, emerge una netta distinzione tra il tasso di abbandono entro due anni degli studenti "regolari" e non: tra coloro che hanno conseguito la maturità a più di 19 anni si osserva una percentuale media di abbandono superiore di quasi il 20% rispetto ai diplomati entro i 19 anni, e tale differenza è pressoché costante per tutte le coorti analizzate.

Un'ulteriore conferma all'ipotesi che iniziare l'Università ad un'età più avanzata rispetto a quella normalmente prevista possa costituire un ostacolo al

proseguimento degli studi deriva infine dall'analisi del tempo medio di attesa tra il conseguimento del diploma e l'immatricolazione all'Università (cfr. Figura 4).

Figura 4. *Tempo medio di attesa prima dell'immatricolazione, per Facoltà e esito.*



Osservando la Figura 4 risulta evidente che coloro che abbandonano il proprio CdL hanno atteso prima dell'immatricolazione un tempo medio maggiore di coloro che invece sono ancora iscritti dopo 2 anni. Tale fenomeno appare differenziato a livello delle singole Facoltà, come evidenziato in figura.

I valori del *Chi-quadro* di Pearson e dell'indice *V* di Cramer (cfr. Tabella 1) confermano il significativo livello di associazione che lega l'esito degli studi dopo due anni sia con l'età al conseguimento del diploma di maturità che con il tempo di attesa prima dell'immatricolazione, reso discreto attraverso la suddivisione in 8 classi distinte.

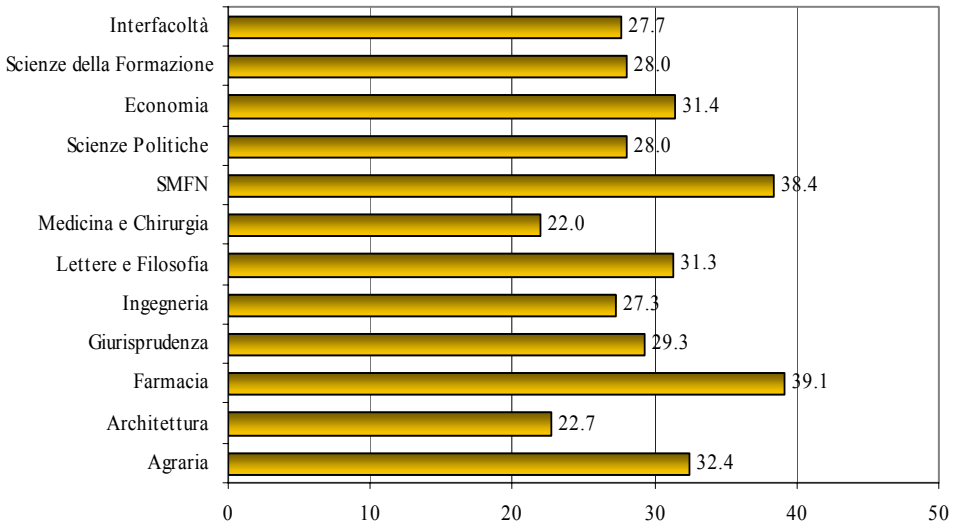
3. Gli abbandoni degli immatricolati nell'a.a. 2001/02

Le matricole del 2001/02 che alla fine del giugno 2003 rientravano nella categoria degli abbandoni dopo il primo anno di Università erano 2908 su 10053, ovvero il 28.9% del totale degli iscritti; tale valore è inferiore sia al 32.3% della coorte 2000/01 che al 29.9% di quella immediatamente precedente.

Come si è avuto modo di osservare nel corso del primo paragrafo, maggiori indicazioni dovrebbero derivare dall'analisi del fenomeno condotta a livello delle

Facoltà e, soprattutto, dei singoli corso di studi. I dati a livello di Facoltà sono riportati nella Figura 5.

Figura 5. *Abbandoni dopo un anno degli immatricolati nell'a.a. 2001/02, per Facoltà (valori %).*



A livello di Facoltà le quote più elevate di mancate iscrizioni al secondo anno sono state registrate a Farmacia (39.1%) ed a SMFN (38.4%); seguono Agraria, Economia e Lettere e Filosofia, tutte attorno al 32% di abbandoni, mentre per Giurisprudenza, Scienze della Formazione, Scienze Politiche, Ingegneria ed i corsi Interfacoltà tale valore scende a circa il 28%. Le quote più basse si osservano infine per Architettura e Medicina e Chirurgia, rispettivamente con il 22.7% ed il 22%.

Per quanto riguarda i dati relativi ai singoli Corsi di Laurea[7] si vede che le percentuali di abbandono più elevate appartengono a Facoltà di indirizzo prevalentemente scientifico, anche se diversi CdL appartengono alla Facoltà di Lettere. Si osserva inoltre che molti di questi corsi di studio sono tra quelli di nuova istituzione; si può allora ipotizzare che tali corsi siano riusciti ad "attirare" un numero piuttosto consistente di studenti, che però si sono successivamente resi conto di non essere veramente interessati a quei percorsi di studio. Infine, l'estrema variabilità riscontrata tra Corsi di Laurea della stessa Facoltà porta a concludere che anche per la coorte dell'a.a. 2001/02 emerge la necessità di analizzare il fenomeno degli abbandoni dopo un anno proprio a tale livello di osservazione.

Anche per questi dati si è proceduto alla misura del grado di associazione tra proporzione di studenti che hanno abbandonato gli studi nel primo anno di corso e i

[7] Per i dati relativi ai singoli Corsi di Laurea si rimanda a Giusti C. (2004).

principali caratteri individuali; i risultati dell'elaborazioni effettuate sono riportati nella Tabella 2.

Relativamente al sesso degli studenti, si osserva un valore significativo per l'associazione di tale variabile con l'esito degli studi sia tra il totale degli studenti che tra i soli abbandoni[8]; il valore della *V* di Cramer evidenzia però, in entrambi i casi, un'intensità di legame piuttosto bassa, indicando quindi che l'esito degli studi dopo un solo anno non sembra eccessivamente associato al sesso degli studenti.

La Facoltà d'immatricolazione dello studente mostra, attraverso il calcolo delle statistiche di associazione, un legame leggermente più forte con l'esito degli studi rispetto a quanto osservato relativamente al sesso, e ciò risulta vero, in particolare, andando a distinguere tra i diversi tipi di abbandono.

Passando all'esame delle altre possibili determinanti degli abbandoni, si rileva che la residenza, così come era emerso anche dall'analisi relativa alle coorti di immatricolati dal 1980/81 al 2000/01, non sembra esercitare una particolare influenza sull'esito degli studi.

Il tipo di maturità conseguita e la relativa votazione, come si è già avuto modo di rilevare, mostrano invece un'influenza molto significativa sull'esito degli studi dopo un anno: la percentuale di studenti che hanno conseguito la maturità liceale, classica o scientifica, è particolarmente bassa (23.7%) tra gli abbandoni impliciti, ed è inferiore alla media osservata per tutti gli immatricolati anche tra coloro che hanno presentato domanda di rinuncia (34.5%).

Ancora più interessante risulta l'analisi del voto conseguito alla maturità: gli studenti che hanno abbandonato il Corso di Laurea presentano una distribuzione della votazione al diploma spostata verso le classi più basse di voto; infatti, la percentuale registrata per gli abbandoni è sempre superiore a quella generale nelle classi da 60 a 75 centesimi, con una differenza massima nella classe di votazione più bassa, 60-62; mentre la quota di studenti presenti nelle classi da 75 a 100 centesimi è invece sempre inferiore per gli abbandoni e, anche in questo caso, la differenza massima si osserva per la classe di voto estrema, quella dei 100/100.

[8] Relativamente agli immatricolati nell'a.a. 2001/02, a differenza dell'elaborazione illustrata nel paragrafo precedente, è risultato possibile condurre le analisi distinguendo quattro diverse tipologie di abbandono: il passaggio di Corso di Laurea, il trasferimento ad altro Ateneo, la rinuncia formale agli studi e la mancata reiscrizione al secondo anno di corso, categoria quest'ultima dei cosiddetti "abbandoni impliciti".

Tabella 2. *Statistiche d'associazione: anno accademico 2001/02.*

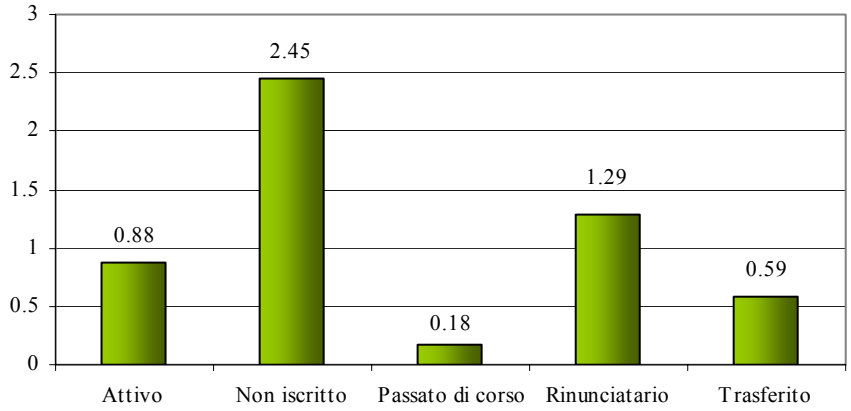
Caratteri: **esito dopo 1 anno** *vs*	*Statistica*	*Valore*	*GdL*	*Prob*
Genere (Tutti gli studenti)	Chi-quadro	49.1868	4	<.0001
	V di Cramer	0.0699		
Genere (Solo abbandoni)	Chi-quadro	6.3306	3	<.0001
	V di Cramer	0.0467		
Facoltà (Tutti gli studenti)	Chi-quadro	86.3551	11	<.0001
	V di Cramer	0.0927		
Facoltà (Solo abbandoni)	Chi-quadro	262.3537	33	<.0001
	V di Cramer	0.1734		
Residenza (Tutti gli studenti)	Chi-quadro	320.3917	16	<.0001
	V di Cramer	0.0927		
Residenza (Solo abbandoni)	Chi-quadro	273.5687	12	<.0001
	V di Cramer	0.1771		
Maturità (Tutti gli studenti)	Chi-quadro	414.5744	12	<.0001
	V di Cramer	0.1172		
Maturità (Solo abbandoni)	Chi-quadro	258.3219	9	<.0001
	V di Cramer	0.1721		
Voto Maturità (Tutti gli studenti)	Chi-quadro	328.5039	32	<.0001
	V di Cramer	0.0917		
Voto Maturità (Solo abbandoni)	Chi-quadro	119.1988	24	<.0001
	V di Cramer	0.1189		
Maturità (Tutti gli studenti)	Chi-quadro	414.5744	12	<.0001
	V di Cramer	0.1172		
Maturità (Solo abbandoni)	Chi-quadro	258.3219	9	<.0001
	V di Cramer	0.1721		
Voto Maturità (Tutti gli studenti)	Chi-quadro	328.5039	32	<.0001
	V di Cramer	0.0917		
Voto Maturità (Solo abbandoni)	Chi-quadro	119.1988	24	<.0001
	V di Cramer	0.1189		
Regolarità studi preuniv. (Tutti gli studenti)	Chi-quadro	423.8752	4	<.0001
	V di Cramer	0.2054		
Ritardo immatricolazione (Tutti gli studenti)	Chi-quadro	188.8819	7	<.0001
	V di Cramer	0.1371		
Ritardo immatricolazione (Solo abbandoni)	Chi-quadro	282.3551	21	<.0001
	V di Cramer	0.1799		

L'utilità di distinguere gli studenti secondo il tipo di abbandono è particolarmente evidente proprio nell'analisi del voto di conseguimento del diploma: il voto medio ottenuto è infatti superiore alla media generale, calcolata su tutti gli immatricolati, sia per i trasferiti che per i passati ad altro Corso di Laurea; inferiore alla media è invece il dato relativo ai rinunciatari ed agli impliciti.

La forte associazione della regolarità della carriera preuniversitaria con l'esito degli studi alla fine del primo anno è confermata anche dal valore della *V* di Cramer, piuttosto elevato (cfr. Tabella 2) sia fra tutti gli studenti che tra i soli abbandoni.

Un ultimo interessante aspetto da considerare riguarda il tempo di attesa tra il conseguimento del diploma e l'iscrizione all'Università. È presumibile, infatti, che gli studenti che si immatricolano nello stesso anno in cui conseguono la maturità abbiano un maggiore interesse verso la loro carriera universitaria, non avendo probabilmente nemmeno provato a cercare lavoro o ad intraprendere un altro percorso formativo. I dati (cfr. Figura 6) sembrano confermare questa ipotesi, dal momento che il tempo medio di attesa passa dallo 0.88 degli ancora attivi dopo un anno all'1.13 osservato in media tra coloro che abbandonano. Ancor più interessante è il dato relativo alle diverse tipologie di abbandono: coloro che lasciano il Corso di Laurea di immatricolazione effettuando un passaggio hanno un'attesa media molto bassa, addirittura inferiore a quella degli ancora attivi. Opposto è invece il discorso riguardante coloro che sospendono semplicemente gli studi, mentre un risultato intermedio si osserva infine per gli studenti che abbandonano in seguito ad un trasferimento o alla presentazione di una rinuncia formale agli studi.

Figura 6. *Anni medi di attesa prima di immatricolarsi, per stato di carriera dopo un anno.*



Occorre tuttavia ricordare che è necessario essere molto cauti nel trarre conclusioni, poiché in questa analisi si classificano gli studenti in base al loro stato di carriera dopo un solo anno di corso; la distribuzione che ne deriva ha quindi

caratteristiche del tutto peculiari, e differisce sicuramente da quella che sarà possibile calcolare non appena saranno disponibili i dati relativi agli anni accademici successivi.

4. Il modello gerarchico a due livelli

Come anticipato nell'introduzione, per procedere ad una più soddisfacente individuazione delle possibili determinanti degli abbandoni universitari si è fatto ricorso ai modelli di regressione multilivello[9].

Nei due paragrafi precedenti è stato analizzato il fenomeno degli abbandoni degli studi universitari nell'Ateneo fiorentino sia relativamente al periodo 1980-2000 che all'a.a. 2001/02, utilizzando prevalentemente tecniche statistiche di tipo descrittivo; tali tecniche non consentono di pervenire ad una misura dell'effetto netto delle determinanti (fattori esplicativi) del fenomeno di interesse, che nel caso specifico è rappresentato dalla probabilità individuale di abbandonare il corso di studi di prima immatricolazione. Tale possibilità è invece offerta dai modelli di regressione e, in particolare, dai modelli di regressione di tipo multilivello.

In proposito vale la pena ricordare che il ricorso ad un modello di regressione multilivello è consigliabile ogni volta che le unità su cui si rileva il fenomeno oggetto di studio (dette unità di primo livello) risultano naturalmente aggregate in gruppi differenti (unità di secondo livello), che a loro volta possono essere aggregate in unità di terzo livello e così via. In tali casi è ragionevole ritenere che la variabilità del fenomeno dipenda non solo da variabili esplicative individuali (o di primo livello), ma altresì dal fatto che un certo individuo appartenga ad un determinato gruppo avente caratteristiche peculiari che lo contraddistinguono dagli altri gruppi; nel caso specifico trattato, gli studenti (unità di primo livello) risultano naturalmente aggregati in Corsi di Laurea (unità di secondo livello).

La finalità che s'intende perseguire attraverso il ricorso ai modelli di regressione multilivello è, dunque, l'individuazione delle variabili maggiormente esplicative dell'abbandono degli studi misurando anche, nel contempo, l'incidenza del fenomeno stesso. I dati cui si farà riferimento sono quelli relativi agli immatricolati dell'a.a. 2001/02, il che consente un approfondimento conoscitivo del fenomeno limitatamente agli abbandoni che si registrano ad un solo anno dall'immatricolazione.

La variabile risposta presa in considerazione è la situazione al 30 giugno 2003 degli immatricolati dell'Ateneo fiorentino nell'a.a. 2001/02; le modalità che essa può

[9] Al riguardo si possono consultare, tra gli altri, i volumi di Goldstein H. (2003) e Snijders T., Bosker R. (1999).

assumere sono due: lo studente non si è reiscritto allo stesso corso di studi oppure lo studente si è reiscritto allo stesso corso. Trattandosi di una variabile binaria, il modello prescelto è stato il modello di tipo logistico a due livelli, preferito ad un modello di tipo probit per la maggiore facilità di interpretazione dei risultati attraverso il calcolo degli odds[10].

Per quanto riguarda le unità di primo livello, ovvero tutti gli studenti immatricolati per la prima volta all'Università di Firenze nell'a.a. 2001/02, si disponeva di un totale di 10053 individui, ma per alcuni di questi studenti non si possedeva l'informazione su due interessanti variabili di analisi, il titolo di studio di scuola superiore ed il relativo voto conseguito. Alla fine si è potuto disporre, quindi, di 9770 unità di primo livello, dal momento che gli individui con dati mancanti sono stati esclusi automaticamente dalla procedura.

Le unità di secondo livello sono i Corsi di Laurea attivati presso l'Ateneo fiorentino nell'a.a. 2001/02. Su un totale di 99 Corsi di Laurea sono state però considerate solo 83 unità, dal momento che alcune di esse possedevano un numero di studenti molto esiguo, addirittura inferiore a 10; si è allora deciso, per includere comunque nell'analisi tali studenti, di aggregare tra loro alcuni dei Corsi di Laurea.

Coerentemente a quanto suggerito dalla teoria, i passi della procedura di stima impiegata sono stati:

- stima del ***modello nullo***, al fine di verificare, attraverso la stima della varianza dei residui di secondo livello, la ragionevolezza del ricorso ad un'analisi multilivello;
- stima del ***modello a intercetta casuale*** comprendente tutte le variabili di primo livello (comprese le interazioni) risultate significative[11];
- stima del ***modello a intercetta casuale finale***[12] (due livelli) costituito dalle sole variabili di primo e secondo livello (comprese le interazioni) risultate significative.

Di seguito verranno riportati e commentati in forma estesa soltanto i risultati relativi alla stima del *modello a intercetta casuale finale*, cioè del modello gerarchico di regressione logistica a due livelli ad intercetta casuale:

$$y_{ij} = \text{logit}\,(P_{ij}) = \gamma_0 + \sum_{h=1}^{r} \gamma_h X_{hij} + U_{0j}$$

$$U_{0j} \sim \text{N}(0, \tau_0^2)$$

[10] Gli odds sono dati dal rapporto tra la probabilità che uno studente abbandoni il corso di studi di immatricolazione rispetto alla probabilità che lo stesso studente si reiscriva allo stesso corso.

[11] Si segnala che, nel caso trattato, tutte le variabili inserite nel modello sono risultate significative.

[12] Si puntualizza che sono stati testati anche modelli con coefficienti casuali; in questo caso però l'algoritmo di massimizzazione della funzione di verosimiglianza non ha mai raggiunto la convergenza.

La variabile risposta utilizzata è misurata al primo livello, il livello individuale, ed è dicotomica, ovvero

$$y_{ij} = \begin{cases} 1 & \text{se lo studente } i \text{ del CdL } j \text{ abbandona} \\ 0 & \text{se non abbandona} \end{cases}$$

mentre le variabili esplicative X_h sono sia di primo che di secondo livello (ovvero relative ai CdL) e sono sia continue che categoriche.

P_{ij} indica la probabilità di abbandono dello studente *i* del CdL *j*, e quindi si ha:

$$y_{ij} \mid U_{0j} \sim \text{Bernoulli}\,(P_{ij})$$

Gli U_{0j}, che costituiscono l'elemento distintivo del modello gerarchico, rappresentano gli effetti casuali relativi alle unità di secondo livello. Tali entità, dunque, esprimono l'effetto residuo esercitato da ciascun CdL nei riguardi della variabile risposta, una volta controllato l'effetto delle covariate X_h. Relativamente a tali componenti casuali si ipotizza una distribuzione normale, con media nulla e varianza costante τ_0^2.

La procedura utilizzata per stimare il modello di regressione logistica a due livelli è la PROC NLMIXED[13] del software SAS-STAT.

[13]Tale procedura prevede la specificazione del predittore lineare, come funzione delle variabili esplicative, e della funzione *link* utilizzata. Quando la convergenza viene raggiunta con successo, l'*output* di questa procedura fornisce la stima dei parametri, del loro errore standard e include inoltre, relativamente a ciascun singolo parametro, la verifica della sua significatività attraverso il test *t* di Wald. Una migliore procedura di convergenza può essere assicurata dalla specificazione del valore iniziale dei parametri da stimare, valori altrimenti posti uguali ad uno di *default*.
Per quanto riguarda la procedura di stima, è importante sottolineare che la PROC NLMIXED massimizza un'approssimazione numerica dell'esatta verosimiglianza marginale del modello non lineare, attraverso il *metodo di quadratura di Gauss-Hermite*. Questo fa sì che la misura della Devianza fornita nell'*output* possa essere utilizzata per confrontare modelli diversi attraverso il test del Rapporto di Verosimiglianza; bisogna rilevare, tuttavia, che la procedura non prevede la possibilità di calcolare tale tipo di test, che deve quindi essere calcolato "manualmente". Una caratteristica interessante del metodo di quadratura utilizzato dalla PROC NLMIXED è che questo risolve l'integrale della verosimiglianza marginale utilizzando la cosiddetta versione *adattiva* del *metodo di Gauss-Hermite*. Tale versione fornisce un'approssimazione dell'integrale generalmente più accurata di quella che si otterrebbe con la *quadratura di Gauss-Hermite* "standard" che utilizza il medesimo numero di punti di quadratura (SAS INSTITUTE INC., 1999).

4.1 Il modello stimato ed i risultati ottenuti

Come già segnalato, il primo passo della procedura ha previsto la stima di un modello di regressione a due livelli senza variabili esplicative, ovvero del cosiddetto *modello nullo*:

$$\text{logit}(P_j) = \gamma_0 + U_{0j}$$

Attraverso tale modello è stato possibile valutare la significatività del parametro τ_0^2, che esprime la varianza dei residui di secondo livello U_{0j}; si è provveduto infatti a confrontare la Devianza (che corrisponde a meno due volte il logaritmo naturale della *verosimiglianza*) del modello precedente con quella ottenuta per lo stesso modello ma senza le componenti U_{0j}, svolgendo il test del Rapporto di Verosimiglianza.

In particolare, con la stima del modello nullo si è ottenuta una Devianza pari a 11563; per il modello nullo di regressione logistica ad un solo livello si è ottenuta invece una Devianza pari a 11724: anche tale informazione è stata calcolata attraverso la PROC NLMIXED, omettendo lo statement RANDOM relativo alle componenti casuali di secondo livello. Il relativo test del Rapporto di Verosimiglianza è risultato altamente significativo, indicando dunque che effettivamente il CdL di appartenenza dello studente ha un effetto significativo nel determinare la probabilità di abbandono.

Una volta verificata l'effettiva esistenza di un'organizzazione dei dati su due livelli di analisi, lo studio è proseguito con la stima del modello multilivello completo, ovvero contenente le variabili esplicative sia di primo che di secondo livello.

Relativamente alle variabili categoriche introdotte nel modello, si è reso necessario individuare per ciascuna di esse una modalità base o di riferimento nei confronti della quale valutare l'effetto di tutti gli altri livelli esistenti, come in una normale regressione logistica. Relativamente a tale aspetto si è scelto di considerare come modalità base di ciascuna covariata la caratteristica più diffusa nella popolazione di studio, secondo i risultati ottenuti attraverso l'analisi descrittiva preliminare; l'individuo che possiede tutte queste caratteristiche verrà denominato *individuo base*[14].

[14] Relativamente all'unica variabile continua disponibile, il voto di conseguimento del diploma di scuola media superiore, si è scelto di calcolare per ciascuno studente lo scarto tra il voto da lui conseguito ed il valore medio calcolato all'interno del suo CdL. Si è scelto dunque di utilizzare l'approccio *group mean centering*, che consente di considerare il cosiddetto "frog-pond effect" (Hox J.J., 2002). In termini relativi all'istruzione questa teoria si riferisce al fatto che uno studente dotato di media intelligenza può essere considerato molto intelligente se si trova in una classe in cui gli altri studenti sono molto scarsi, oppure poco intelligente se i suoi compagni sono tutti molto capaci.

Le variabili esplicative prese in considerazione per la stima del modello completo sono state scelte sulla base delle analisi descrittive svolte, di cui molto sommariamente riferito nei paragrafi precedenti, e sulla base della conoscenza del fenomeno che è basata anche su un'indagine telefonica rivolta agli immatricolati dell'a.a. 2001/02 che non risultavano iscritti allo stesso corso di studi al 30 giugno 2003[15].

In particolare, le caratteristiche individuali e le relative covariate di primo livello considerate nel modello sono state:

- ✓ il **genere**: la variabile considerata (*sesso*) è dicotomica ed assume valore 0 se lo studente è femmina, 1 se maschio;
- ✓ il **tipo di maturità**: tale variabile è stata introdotta nel modello attraverso la creazione di tre *dummy*; avendo scelto come base la maturità di tipo liceale, le due variabili presenti nel modello risultano essere quella relativa alla maturità tecnica o professionale (*prof_tecnica*) e alla maturità di altro tipo (*altra_mat*).
- ✓ la **residenza**: anche in questo caso sono state create tre *dummy*, relative rispettivamente alla residenza a Firenze (variabile non presente nel modello), a Arezzo, Pistoia o Prato (*pendolari*) e a qualsiasi altra residenza (*altra_res*).
- ✓ **regolarità degli studi preuniversitari**: tale caratteristica è indicata dalla variabile dicotomica *eta_mat*, che assume valore 1 nel caso in cui lo studente abbia conseguito la maturità ad età maggiore di 19 anni, 0 altrimenti;
- ✓ **ritardo nell'immatricolazione**: anche in questo caso è una variabile dicotomica (*ritardo_imm*) ad indicare la presenza dell'"irregolarità" di carriera, costituita da un tempo di attesa tra il diploma e l'immatricolazione maggiore ad un anno;
- ✓ **regolarità carriera preuniversitaria**: la variabile *eta_per_ritardo* esprime l'interazione tra le due precedenti, ed è dunque una variabile dicotomica che assume valore 1 se *eta_mat*=1 e *ritardo_imm*=1, 0 altrimenti;
- ✓ **voto di maturità**: per la variabile continua data dal voto riportato alla maturità espresso in centesimi (*voto_mat*) si è effettuata la centratura rispetto alla media del CdL cui appartiene lo studente.

[15] Per comprendere i motivi all'origine della scelta di abbandonare dopo un solo anno il Corso di Laurea di immatricolazione da parte degli studenti iscrittisi nell'a.a. 2001/2002, nel luglio 2003 è stata effettuata un'apposita indagine telefonica, la prima realizzata nell'Ateneo fiorentino relativamente a tale argomento. La speranza era quella di trarre indicazioni che potessero aiutare a comprendere meglio tale fenomeno e a capire quali politiche adottare, a livello di Ateneo ma soprattutto dei singoli Corsi di Laurea. Una trattazione estesa dei risultati dell'indagine, che è stata di tipo censuario e che ha coinvolto 2715 studenti, si trova in Giusti C. (2004).

Date tali variabili, si ha che l'***individuo base***, ovvero colui che possiede tutte le modalità di riferimento scelte per le variabili di analisi, è:

- **femmina;**
- **possiede maturità liceale** (scientifica o classica)**;**
- **risiede a Firenze;**
- **ha conseguito la maturità ad un'età minore o uguale a 19 anni** (ovvero si può supporre che non abbia mai sperimentato episodi di ripetente)**;**
- **si è immatricolato all'Università nello stesso anno in cui ha conseguito il diploma di maturità;**
- **ha ottenuto un voto di maturità pari al voto medio del CdL in cui si è immatricolato**.

Una delle caratteristiche più interessanti dei modelli multilivello è che gli stessi permettono di considerare anche variabili esplicative relative al livello superiore di analisi. In questo modo si può cercare di ridurre la correlazione presente all'interno delle unità di secondo livello, tentando quindi di "spiegare" almeno in parte la variabilità degli effetti casuali U_{0j}. Il passo successivo dell'analisi è consistito dunque nel cercare di individuare variabili esplicative di secondo livello che risultassero esercitare un effetto significativo sul logit delle probabilità di abbandono.

Attraverso il test di Wald al livello di significatività del 5% si sono individuate due variabili esplicative significative misurate al livello dei CdL: la variabile dicotomica indicante la presenza del *numero chiuso* delle immatricolazioni, e la variabile continua esprimente la percentuale di studenti con carriera preuniversitaria "irregolare" (maturità conseguita a più di 19 anni e/o attesa tra il diploma e l'immatricolazione pari ad almeno un anno).

Seguendo la medesima procedura utilizzata relativamente alle variabili di primo livello, per la variabile indicante la presenza del numero chiuso si è scelto come modalità di riferimento l'assenza dello stesso, mentre per quanto riguarda l'altra variabile di secondo livello, essendo questa continua, è stata centrata attorno alla media generale, calcolata fra i vari CdL.

Le variabili esplicative di secondo livello considerate sono state quindi:

- ✓ **numero chiuso**: la variabile dicotomica *num_chiuso* assume valore pari ad 1 se il CdL possiede limitazioni al numero di immatricolazioni, 0 altrimenti;
- ✓ **regolarità degli studi preuniversitari**: la variabile continua esprimente la percentuale di studenti con irregolarità di carriera (*eta_mat*=1 e/o *ritardo_imm*=1) iscritti al CdL è centrata rispetto alla media generale.

Tabella 3. *Parametri stimati con il modello ad intercetta casuale "completo".*

Parametro fisso	**Stima**	**Standard error**	***p-value***
intercetta	-1.609	0.12	<.0001
sesso	0.1114	0.05279	0,0379
prof_tecnica	0.5619	0.05482	<.0001
altra_mat	0.4265	0.06972	<.0001
pendolari	-0.1724	0.06275	0,0074
altra_res	0.1512	0.05603	0,0085
eta_mat	0.4389	0.06426	<.0001
ritardo_imm	0.4214	0.08443	<.0001
voto_mat	-0.0199	0.002093	<.0001
eta_per_ritardo	-0.3106	0.1184	0,0104
irreg_medio	0.0058	0.2658	0,0318
num_chiuso	-0.5749	0.1721	0,0013
Parametro casuale	**Stima**	**Standard error**	***p-value***
Varianza τ_0^2	0.1254	0.03009	<.0001

Introducendo le variabili esplicative di secondo livello la varianza degli U_{0j} si riduce, passando dal valore di 0.1869, ottenuto per il modello con le sole covariate di primo livello, a τ_0^2=0.1254 (cfr. Tabella 3). Di conseguenza si osserva anche una riduzione della correlazione infragruppo, che risulta adesso pari a 0.037, contro lo 0.045 precedentemente ottenuto. Ancor più significativa risulta la riduzione della variabilità di secondo livello rispetto al modello nullo; il che induce a concludere che le variabili relative ai CdL introdotte nel modello finale sono riuscite a spiegare il 33% circa della variabilità degli U_{0j}.

Per interpretare i risultati ottenuti per i parametri fissi, le stime sono state trasformate in probabilità di abbandono attraverso l'impiego della funzione logistica; per esempio, la stima ottenuta per l'intercetta del modello implica una probabilità di abbandono per l'*individuo base* iscritto ad un *CdL base* (ovvero senza numero chiuso e con una percentuale di studenti con "irregolarità" di carriera uguale alla media generale) pari al 16.7%:

$$\hat{\pi}_0 = \frac{\exp(-1.609)}{1+\exp(-1.609)} = 0.167$$

Utilizzando tale risultato è possibile interpretare le stime restanti andando a vedere in che modo le varie caratteristiche modificano la probabilità di abbandono dell'*individuo base*. I risultati delle trasformazioni sono riportati nella Tabella 4.

Si nota subito che essere maschio aumenta, seppur in modo limitato, la probabilità di abbandono: questo conferma quanto ottenuto in sede di analisi descrittiva, dal momento che per le femmine si era individuata una quota di abbandoni alla fine del primo anno di corso inferiore a quella dei maschi.

Per quanto riguarda il diploma di scuola superiore, possedere una maturità tecnica o professionale piuttosto che liceale aumenta in modo considerevole la probabilità di abbandono; il valore ottenuto per l'effetto "maturità professionale o tecnica" è infatti il più alto in valore assoluto tra quelli di livello individuale. Anche possedere un qualsiasi altro tipo di diploma aumenta la probabilità di abbandonare, seppur in misura inferiore.

Risiedere nelle province di Arezzo, Pistoia o Prato, ovvero rientrare nella categoria dei cosiddetti "pendolari", riduce la probabilità di abbandono, dal momento che questa risulta inferiore rispetto a quella dell'*individuo base*, che risiede a Firenze. L'effetto di una qualsiasi altra residenza agisce invece nel senso opposto, facendo cioè aumentare la probabilità di interrompere gli studi nel CdL di prima immatricolazione.

Aver sperimentato una qualche irregolarità nella carriera scolastica preuniversitaria, fatto misurato in modo indiretto dall'età dello studente al conseguimento della maturità, ha un effetto negativo e piuttosto consistente sulla probabilità di abbandono. Tale effetto risulta poi del tutto simile ad un'altra possibile "irregolarità", ovvero l'aver atteso un anno o più tra il superamento dell'esame di maturità e l'immatricolazione all'Università. Il termine d'interazione di queste due ultime variabili indica poi che se uno studente ha sperimentato entrambi gli episodi di "irregolarità", l'effetto negativo sulla sua probabilità di abbandono risulterà "mitigato" rispetto a quello che si avrebbe sommando semplicemente i due singoli effetti. Sempre relativamente alle variabili misurate al livello individuale, si ha infine che l'incremento di un'unità del voto di maturità rispetto alla media di CdL ha l'effetto di ridurre la probabilità di abbandono individuale.

Per quanto riguarda le variabili di secondo livello, si osserva che la presenza del numero chiuso comporta una variazione positiva assai consistente della probabilità di abbandono individuale: in termini percentuali, infatti, questa variabile risulta esercitare l'effetto maggiore tra quelli stimati. Si può quindi affermare che, a parità di tutti gli altri fattori, dover superare una prova di accesso per potersi immatricolare ad un dato CdL si tradurrà presumibilmente in una maggiore motivazione ed interesse a portare aventi il percorso di studi intrapreso.

Infine, un aumento dell'1% rispetto alla media generale della percentuale di studenti del CdL che hanno sperimentato almeno una delle due "irregolarità" di carriera più volte citate avrà l'effetto di aumentare, seppur lievemente, le probabilità di abbandono individuali degli studenti di quel CdL.

Tabella 4. *Interpretazione delle stime ottenute con il modello completo.*

Probabilità di abbandono *individuo base* = 16,7%				
Parametro fisso	**Significato**	**Stima**	**Probabilità di abbandono** (%)	**Variazione % rispetto all'*individuo base***
sesso	maschio	0.1114	18,31	+9.63%
prof_tecnica	maturità professionale o tecnica	0.5619	26,02	+55.81%
altra_mat	maturità di altro tipo	0.4265	23,5	+40.71%
pendolari	Arezzo, Pistoia o Prato	-0.1724	14,44	-13.55%
altra_res	altra residenza	0.1512	18,91	+13.24%
eta_mat	maturità conseguita ad età >19	0.4389	23,72	+42.05%
ritardo_imm	immatricolazione almeno un anno dopo la maturità	0.4214	23,41	+40.16%
eta_per_ritardo	interazione tra le due variabili precedenti	-0.3106	12,81	-23.28%
voto_mat	maggiore di un'unità rispetto alla media di CdL	-0.0199	16,42	-1.65%
irreg_medio	maggiore dell'1% rispetto alla media generale	0.0058	16,78	+0.48%
num_chiuso	presente	-0.5749	10,13	-39.3%

Ovviamente non bisogna dimenticare l'effetto esercitato sulle probabilità di abbandono dalle componenti casuali di secondo livello U_{0j}. Le stime di tali residui, dette *stime di Bayes*, possono essere impiegate, inoltre, sia per valutare il particolare effetto esercitato da ciascuno dei CdL sulla probabilità di abbandono P_{ij}, che per verificare l'ipotesi di normalità relativa alla distribuzione degli U_{0j} stessi.

Se per esempio si indica con τ_0 la radice quadrata della varianza di secondo livello, possiamo calcolare le variazioni di probabilità rispetto al valore base di 0.167 dovute ad alcune realizzazioni dell'effetto casuale U_{0j} (cfr. Tabella 5). Risulta così evidente che, a parità di caratteristiche sia di primo che di secondo livello, frequentare un CdL piuttosto che un altro può modificare notevolmente la probabilità di abbandono individuale.

L'impiego più interessante delle *stime di Bayes* consiste però nell'utilizzare tali valori per confrontare tra loro i vari CdL, dal momento che il residuo U_{0j} rappresenterà l'effetto esercitato sulle probabilità di abbandono individuali dal *j-esimo* CdL, una volta controllato per l'effetto di tutte variabili esplicative.

Tabella 5. *Effetto dei parametri casuali*

Ipotetico valore dell'effetto casuale	**Probabilità dell'*individuo base*** (%)	**Variazione percentuale della probabilità dell'*individuo base***
$-2\hat{\tau}_0 = -0.708$	8,98	-46,2%
$-\hat{\tau}_0 = -0.354$	12,33	-26,15%
$\hat{\tau}_0 = 0.354$	22,22	+33.07%
$+2\hat{\tau}_0 = 0.708$	28,94	+73,27%

Il corso nei confronti del quale si è ottenuta la *stima di Bayes* più elevata è quello in Scienze Biologiche, seguito da Informatica, CdL che si distinguono quindi per l'effetto particolarmente negativo che esercitano nei confronti dell'abbandono. È interessante notare, inoltre, che se non si tiene conto delle covariate di secondo livello, facendo sì che la variabilità degli U_{0j} non sia "controllata" in nessun modo, risultano esercitare un effetto significativamente positivo nel ridurre la probabilità di abbandono, contrariamente a quanto succede con il modello completo, i CdL in Medicina e Chirurgia, Odontoiatria, Progettazione della Moda e Architettura. Questi sono quattro dei sette CdL per i quali è presente una qualche forma di numero chiuso: evidentemente, quindi, nel modello nullo questi CdL godono del fatto che le limitazioni all'accesso contribuiscono a ridurre la probabilità di abbandono, mentre controllando l'effetto di tale variabile (modello completo), gli U_{0j} corrispondenti ne risultano "penalizzati" ed il loro valore aumenta.

5. Conclusioni

Nella prima parte del presente lavoro si è proceduto all'analisi descrittiva, accompagnata dal calcolo di alcune statistiche di associazione, relativamente alle principali caratteristiche individuali degli immatricolati presso l'Ateneo fiorentino tra l'a.a. 1980/81 ed il 2001/02, rivolgendo particolare attenzione al fenomeno degli abbandoni. Tali analisi sono risultate strumentali ad una prima comprensione del fenomeno stesso ed alla successiva stima del modello multilivello.

Il modello gerarchico a due livelli ad intercetta casuale presentato ha consentito la valutazione del fenomeno dell'abbandono universitario secondo una nuova prospettiva di analisi. È risultato possibile, infatti, valutare l'effetto netto esercitato sulla probabilità di abbandono individuale degli studenti non solo dalle

loro caratteristiche personali, ma anche da alcune variabili misurate a livello dei Corsi di Laurea dell'Ateneo.

La PROC NLMIXED del software SAS ha inoltre consentito la stima delle componenti casuali di secondo livello, permettendo di ottenere una sorta di "graduatoria" di efficacia relativa dei vari CdL nei confronti del fenomeno analizzato. Un risultato di questo tipo dovrebbe aiutare gli organi di governo di Ateneo a capire in quale direzione concentrare maggiore attenzione e risorse per ridurre il fenomeno dell'abbandono.

I risultati delle analisi svolte, molto sommariamente richiamati in questa nota, giustificano ampiamente, a nostro parere, il ricorso ai modelli multilivello quando si procede all'analisi di dati che riguardano gli studenti universitari[16]; infatti, è del tutto evidente la natura gerarchica dei dati: le unità di primo livello sono gli studenti o i laureati/diplomati, mentre le unità di secondo livello sono i corsi di studio. Ovviamente la gerarchizzazione può essere estesa ad un numero di livelli più elevato: ad esempio le Facoltà possono rappresentare il terzo livello e gli Atenei il quarto livello.

Riferimenti bibliografici

ASSOCIAZIONE TREELLLE (2003) *Università italiana, università europea? Dati, proposte e questioni aperte*, Quaderno n.3, Genova.

BULGARELLI, G. (2002) *Esito degli studi degli immatricolati dell'Ateneo Fiorentino dal 1980/81 al 1997/98,* Università degli Studi di Firenze, consultabile anche sul sito www.unifi.it/aut_dida/indexval.html.

CHIANDOTTO B. (2002) *Valutazione dei processi formativi: cosa, come e perché, in Valutazione della Didattica e dei Servizi nel Sistema Università*. Atti della giornata di Studio, Fisciano, 31 maggio 2002. CUSL, Salerno 2002**.**

GIUSTI C. (2004) *L'abbandono degli studi nell'Ateneo fiorentino: evoluzione nel periodo 1980 - 2000 e applicazione di un modello gerarchico non lineare agli immatricolati nell'a.a. 2001/02***.** Tesi di laurea, Università degli Studi di Firenze.

[16] In tale direzione si sta movendo da tempo il gruppo ***VALMON*** (**Val**utazione e **Mon**itoraggio). Il gruppo, coordinato da B. Chiandotto e costituito da laureandi, dottorandi e docenti del Dipartimento di Statistica dell'Università degli Studi di Firenze, da diversi anni svolge attività di studio e ricerca nel contesto della valutazione e del monitoraggio dei processi formativi che si svolgono nell'Ateneo fiorentino. Tale interesse è testimoniato, tra l'altro, da altri due lavori presentati in questa sede: "*Un modello multilivello per l'analisi della condizione occupazionale dei laureati"* (Chiandotto B. e Bacci S.); "*Un modello multilivello per l'analisi della durata degli studi universitari"* (Chiandotto B. e Varriale R.).

GOLDSTEIN H. (2003) *Multilevel Statistical Models***,** Edward Arnold, London.

HOX J.J. (2002) *Multilevel Analysis: Techniques and Applications***,** LAWRENCE ERLBAUM ASSOCIATES, Mahwah (New Jersey), London.

ISTAT (2003) *Università e lavoro 2003*, consultabile sul sito internet http://www.istat.it/Societ-/Istruzione (al 06/11/2003).

MURST (1998) *L'evoluzione della domanda di formazione universitaria: studenti, laureati e studenti equivalenti*, consultabile sul sito internet: http://www.murst.it/valutazionecomitato/attivnuc.htm (al 11/12/2003).

OCSE (2002) *Education at a Glance – OECD Indicators 2002*, consultabile sul sito internet: http://www.oecd.org/ (al 03/11/2003)

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

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

Un modello multilivello per l'analisi della durata degli studi universitari

Da: Chiandotto B., Varriale R. (2005) Un modello multilivello per l'analisi della durata degli studi universitari, in Modelli statistici per l'analisi della transizione Università-lavoro, pp. 63-86, a cura di C. Crocetta, Cleup, Padova

Un modello multilivello per l'analisi della durata degli studi universitari[1]

Bruno Chiandotto **Roberta Varriale**
Dipartimento di Statistica "G. Parenti" - Università degli Studi di Firenze

Riassunto. Nel lavoro si analizza il fenomeno dei tempi di conseguimento della laurea, una delle maggiori criticità del sistema universitario italiano. Per cercare di individuarne le possibili determinanti è stata svolta un'analisi sia sui dati di archivio relativi agli studenti immatricolatisi presso l'Ateneo fiorentino nel ventennio 1980-2000, sia sui dati (di archivio e raccolti nell'ambito del progetto AlmaLaurea) relativi ai laureati nell'anno solare 2000. Su questi ultimi dati, avendo come finalità la misura dell'effetto "netto" esercitato dai fattori individuali e da fattori istituzionali (variabili specifiche dei corsi di studio) sui tempi di conseguimento del titolo, è stato introdotto un *modello lineare gerarchico a due livelli*; tale modello tiene conto del fatto che gli studenti (unità di primo livello) risultano naturalmente aggregati nei Corsi di Laurea (unità di secondo livello).
Parole chiave: Tempi di conseguimento del titolo, Modelli multilivello, Regressione lineare gerarchica.

1. Introduzione

Tra gli aspetti negativi che hanno caratterizzato e caratterizzano ancora oggi il sistema universitario italiano assumono particolare rilevanza gli abbandoni e la durata delle carriere: la percentuale di studenti che abbandonano gli studi in Italia è eccessiva e, per coloro che invece riescono a conseguire il titolo universitario, il tempo impiegato per concludere il percorso di studi è troppo elevato.

[1] Il presente lavoro è stato finanziato nell'ambito del progetto "Transizioni Università-lavoro e valorizzazione delle competenze professionali dei laureati: modelli e metodi di analisi multidimensionali delle determinanti", cofinanziato dal MIUR; coordinatore nazionale è Luigi Fabbris, coordinatore del gruppo di Firenze è Bruno Chiandotto (titolo del progetto dell'unità di ricerca locale "Valutazione del processo formativo universitario, sbocchi professionali e pianificazione dei percorsi formativi: modelli e metodi"). L'idea iniziale, la struttura e l'impostazione del lavoro sono dovuti al contributo di entrambi gli autori, mentre le elaborazioni e l'implementazione del modello vanno attribuite a R. Varriale.

Il problema dell'eccessiva durata delle carriere universitarie, tipico del sistema universitario italiano, appare ancora più accentuato se si analizza la situazione dell'Ateneo fiorentino (Chiandotto B. e Bertaccini B., 2003), il che induce a presumere che, su questo fenomeno, le conclusioni di un approfondimento conoscitivo utilizzando i dati fiorentini possano essere ragionevolmente estese anche a gran parte degli altri Atenei italiani. L'individuazione delle possibili determinanti del fenomeno dei tempi di conseguimento del titolo eccessivamente lunghi dovrebbe suggerire interventi finalizzati alla eliminazione di una tale criticità[2].

Il secondo paragrafo di questa nota è dedicato ad una sintetica illustrazione dei risultati dell'analisi svolta sugli immatricolati presso l'Università di Firenze negli anni accademici dal 1980/81 al 2000/01, finalizzata all'individuazione dell'eventuale influenza esercitata sulla durata degli studi sia dal Corso di Laurea sia da caratteristiche individuali (quali genere, residenza, diploma di scuola superiore, ecc.)[3].

Nel terzo paragrafo vengono riassunti, altrettanto sinteticamente, i risultati dell'analisi relativa agli studenti che, essendosi immatricolati presso l'Università di Firenze dall'anno accademico 1980/81 all'anno accademico 2000/01, e non avendo mai effettuato un passaggio di corso di studi, hanno conseguito la laurea presso l'Università di Firenze durante l'anno solare 2000[4].

Successivamente, facendo sempre riferimento ai laureati dell'anno 2000, per pervenire alla misura dell'effetto "netto" eventualmente esercitato da possibili determinanti (sia individuali che istituzionali) sulla durata degli studi, sono stati introdotti i modelli gerarchici o di regressione multilivello; il ricorso a tali modelli è stato suggerito dalla struttura dei dati che è di tipo gerarchico a due livelli: le unità di primo livello sono gli studenti, quelle di secondo livello sono i Corsi di Laurea. I risultati delle analisi condotte sono riportati nel quarto paragrafo; alcune conclusioni completano la nota.

[2] Sul problema della valutazione dei processi formativi finalizzata alla eliminazione di eventuali criticità presenti nel sistema universitario si veda Chiandotto B. (2002).

[3] Una trattazione più dettagliata si trova in Varriale R. (2004), un altro significativo contributo sull'argomento è stato fornito da Bulgarelli G. (2002).

[4] Anche in questo caso si tratta di un'esposizione estremamente sintetica, maggiori dettagli si trovano in Varriale R. (2004), sullo stesso argomento si può utilmente consultare Chiandotto B., Bacci S. e Bertaccini B. (2004).

2. Esito degli studi universitari degli immatricolati nell'Ateneo fiorentino nel periodo 1980-2000

Secondo la definizione utilizzata dall'Istat[5], sono stati considerati immatricolati gli studenti "iscritti per la prima volta al primo anno di un Corso di Laurea o di Diploma Universitario"; in particolare, sono stati esaminati gli immatricolati ai soli Corsi di Laurea.

Ai fini dell'analisi, è stato adottato l'approccio longitudinale, scegliendo come evento di comune origine l'immatricolazione presso l'Università di Firenze in un determinato anno accademico; all'interno della popolazione oggetto di studio sono state, pertanto, individuate 21 coorti. Ogni coorte è stata osservata per 10 anni[6], trascorso tale periodo lo studente può: aver abbandonato gli studi (*abbandono*), essersi già laureato (*laureato*), essere ancora iscritto (*iscritto*). Per valutare il fenomeno dei tempi di conseguimento del titolo sono state, pertanto, considerate 13 coorti.

I laureati presso l'Ateneo fiorentino tra il 1980 e il 31 luglio 2003, immatricolatisi nel periodo intercorso tra l'a.a. 1980/81 e l'a.a. 1992/93, sono stati 32636; a livello di Ateneo il tasso medio di laurea è del 30.4%; mentre i tassi registrati per le diverse Facoltà variano da un minimo di 22.9% per Scienze della formazione ad un massimo di 40.6% per Medicina e Chirurgia.

Solo il 3.2% del totale dei laureati dell'Ateneo completa il ciclo di studi in corso, mentre più dell'80% lo fa con almeno 2 anni di ritardo. Il tempo che gli studenti impiegano per concludere gli studi universitari ha un ovvio riflesso sull'età che i laureati stessi hanno al conseguimento del titolo: l'età media di Ateneo è di 26.8 anni, leggermente più elevata per i maschi (27 anni) rispetto alle femmine (26.7

[5] Gli studenti rientranti nella suddetta definizione di immatricolati sono stati classificati in base al Corso di Laurea di prima iscrizione; per tali studenti si dispone di informazioni classificabili in "variabili d'ingresso" (principalmente dati anagrafici e relativi agli studi pre-universitari), "di soggiorno" (per esempio informazioni su eventuali passaggi di corso, rinunce) e "d'uscita" (esito finale degli studi). Le variabili d'ingresso e quelle "in itinere" rappresentano i fattori individuali, o variabili esplicative, che si suppone possano influenzare l'esito e la durata degli studi. Tali informazioni risultano aggiornate, per ciascuna delle unità di analisi, al 31 luglio 2003.

[6] Nella scelta di tale periodo si è tenuto conto che la durata media degli studi risulta pari a 7-8 anni e che entro 8 anni dall'immatricolazione si registra più del 70% del totale delle lauree osservate per ogni generazione. Inoltre, è stato rilevato che il tasso marginale di crescita del tasso di laurea tende nel tempo ad attestarsi su valori abbastanza costanti e che scegliendo un periodo di osservazione pari a 15 anni gli studenti che avrebbero fatto parte del collettivo di riferimento per le successive analisi sarebbero stati solo 69560 (anziché 107267) sui 174072 studenti immatricolatisi tra l'a.a. 1980/81 e l'a.a. 2000/01.

anni), abbastanza simile per tutte le Facoltà, tranne che per la Facoltà di Scienze della Formazione nella quale i laureati hanno un'età media di quasi 28 anni.

Ovviamente, anche l'analisi basata sull'indice di durata[7] porta a conclusioni non confortanti. A livello di Ateneo, infatti, l'indice di durata medio assume il valore di 1.7: questo vuol dire che gli studenti impiegano più di una volta e mezzo del tempo ritenuto necessario per legge a terminare gli studi universitari.

La Facoltà che presenta l'indice di durata più basso è Medicina e Chirurgia (1.24), mentre la Facoltà con l'indice più alto è Economia (1.9), seguita da Lettere e Filosofia (1.86) e Giurisprudenza (1.83). Come prevedibile, sono gli studenti con un diploma di tipo liceale e coloro che hanno riportato votazioni più elevate all'esame di maturità a presentare valori più bassi dell'indice di durata.

3. Tempi di conseguimento del titolo dei laureati nell'anno solare 2000

In questa parte del lavoro vengono analizzati i dati forniti dal Consorzio Interuniversitario AlmaLaurea[8] relativi agli studenti che, appartenendo alla popolazione esaminata nel paragrafo precedente, hanno conseguito la laurea presso l'Università degli studi di Firenze durante l'anno solare 2000; i dati utilizzati sono stati forniti dal Consorzio Interuniversitario AlmaLaurea e derivano sia da indagini predisposte nell'ambito del Progetto AlmaLaurea sia da fonti di tipo amministrativo. Il tipo di strumento utilizzato per le prime è il questionario strutturato compilato dai laureandi (tasso di risposta pari al 91%) al conseguimento del titolo, composto da domande chiuse a risposta unica e suddiviso in 6 aree tematiche che riguardano: notizie anagrafiche, curriculum scolastico e formativo, notizie sull'esperienza

[7] L'indice di durata (I_d) delle singole Facoltà e CdL è costruito rapportando la durata effettiva degli studi alla durata legale del rispettivo corso; tale indice rende possibile il confronto tra laureati delle diverse Facoltà e diversi CdL, cresce al crescere del ritardo e assume valori maggiori o uguali a 1.

[8] Il consorzio interuniversitario *ALMALAUREA* nasce nel 1994 per iniziativa dell'Osservatorio Statistico dell'Università di Bologna ed attualmente è gestito dalle Università aderenti con il sostegno del Ministero dell'Istruzione, dell'Università e della Ricerca.
I principali obiettivi dei servizi offerti da *ALMALAUREA* sono, da una parte, quelli di assicurare agli organi di governo degli Atenei appartenenti al consorzio, ai Nuclei di Valutazione, alle commissioni impegnate nella didattica e nell'orientamento, attendibili e tempestive basi documentarie e di verifica, volte a favorire i processi decisionali e la programmazione delle attività; dall'altra di creare una sempre più stretta collaborazione tra università e mondo produttivo, facilitando, attraverso la propria banca dati, l'accesso dei giovani al mercato del lavoro italiano ed internazionale.
Per ulteriori informazioni, si può consultare il sito Internet: www.almalaurea.it

universitaria appena conclusa, situazione lavorativa, notizie sulla famiglia, intenzioni e prospettive future.

I dati AlmaLaurea sono stati successivamente integrati con quelli forniti dall'Ufficio Servizi Statistici e Controllo di Gestione dell'Università di Firenze utilizzati per le analisi a cui si è fatto riferimento nel paragrafo precedente, in modo da consentire il confronto dei risultati conseguiti nelle due diverse analisi. Nel procedere all'integrazione tra i due insiemi di dati, però, alcuni records relativi a studenti laureati presso l'Ateneo fiorentino nell'anno solare 2000 non sono stati utilizzati[9]; per tale motivo il collettivo di riferimento oggetto delle successive analisi è risultato composto da 4382 studenti. Inoltre, sono stati esclusi dalle indagini quei laureati che non hanno compiuto l'intero ciclo di studi nell'Ateneo fiorentino e nello stesso Corso di Laurea, ottenendo una popolazione di riferimento ***3978*** unità.

Come già sottolineato, uno degli aspetti più negativi che caratterizza la figura del laureato "tipo" è l'età molto elevata al conseguimento del titolo, età elevata che è il diretto riflesso dell'eccessiva durata degli studi; infatti, solo l'11% dei laureati consegue il titolo ad un'età inferiore ai 24 anni, mentre quasi il 30% termina gli studi universitari ad un'età superiore ai 28 anni. A livello di Facoltà, i laureati in Architettura presentato un'età media al conseguimento del titolo più elevata (29.6), mentre i laureati in Scienze Matematiche Fisiche e Naturali sono i più giovani con un'età media di 26.9 anni.

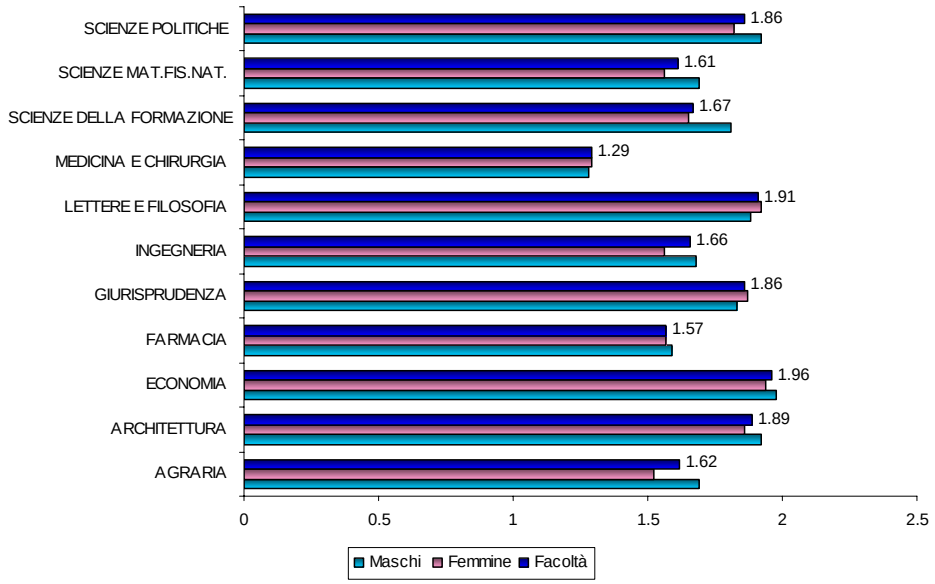
Naturalmente, la variabile ora analizzata serve solo a fornire una indicazione, seppure interessante, della durata degli studi universitari; infatti, molteplici sono i fattori che possono determinare l'innalzamento dell'età al raggiungimento del termine degli studi, e non tutti possono essere imputabili alla durata del piano di studi del corso prescelto. Si è preferito analizzare, pertanto, la durata degli studi universitari sia attraverso un approccio diretto, ovvero attraverso la sua misurazione in anni, sia indirettamente facendo ricorso all'indice di durata.

Dall'analisi della variabile *durata degli studi* per il contingente dei laureati dell'anno solare 2000, il risultato è tutt'altro che confortante. A livello di Ateneo, confrontando le durate legali delle varie Facoltà con quelle medie effettivamente impiegate dagli studenti, si può notare come gli studenti impiegano mediamente da 1.5 a 4.4 anni in più rispetto al tempo considerato necessario per legge al raggiungimento del termine degli studi. Inoltre, mentre solo il 5% degli studenti analizzati si laurea in corso, ben il 50% di questi si laurea dopo il quarto anno fuori corso.

[9] Degli originari 4846 laureati: 195 studenti si sono immatricolati in altri Atenei, 53 studenti si sono immatricolati prima dell'a.a. 1980/81, 41 studenti erano già in possesso di altro titolo, 4 studenti si sono immatricolati ad un diploma, 140 studenti si sono già immatricolati in precedenza senza arrivare al conseguimento del titolo, 25 studenti si sono immatricolati ad anni successivo al primo, 6 studenti si sono immatricolati oltre il 31 Luglio di ogni anno.

Facendo riferimento all'indice di durata, all'interno dell'intero Ateneo si registra un valore medio pari a 1.8: questo significa che gli studenti impiegano quasi il doppio del tempo ritenuto necessario per legge a completare gli studi universitari. I valori assunti da questo indice a livello di Facoltà sono riportati nella Figura 1.

Figura 1. *Indice di durata per Facoltà e sesso*



Successivamente, si è cercato di individuare le possibili relazioni esistenti tra i caratteri di interesse e la durata degli studi e, a tal fine, oltre ad effettuare un'analisi di tipo descrittivo, si è proceduto al computo di due statistiche di associazione, il Chi-quadro di Pearson e la V di Cramer[10].

[10] Questi indici sono calcolati attraverso il confronto tra le frequenze osservate e le frequenze teoriche nell'ipotesi di indipendenza stocastica tra i caratteri considerati e, mentre il primo è espresso in termini assoluti, l'indice di Cramer varia tra 0 e 1. A ragione del contesto di analisi, valori superiori a 0.10 di tale indice inducono a concludere a favore della presenza di un livello di dipendenza abbastanza elevato tra i caratteri analizzati.
La statistica *Chi-quadro* è solitamente utilizzata per l'analisi di distribuzioni discrete, ma può essere calcolata anche per distribuzioni continue raggruppando i dati in classi di modalità; considerazioni analoghe valgono per l'indice *V* di Cramer. Per questo motivo si è dovuto procedere alla suddivisione in classi dell'indice di durata e, dato che questa variabile misura il tempo impiegato dallo studente per conseguire il titolo e quindi cresce all'aumentare degli anni di iscrizione "fuori corso", nel farlo si è cercato di creare una corrispondenza tra classe dell'indice e il numero degli anni fuori corso; le classi così ottenute sono 6. Classe 1 – (0-1.14) – 0; Classe 2 – (1.1401-1.37) – 1; Classe 3 – (1.3701-1.62) – 2 ; Classe 4 – (1.6201-1.87) – 3; Classe 5 – (1.8701-2.12) – 4; Classe 6 – (2.1201-7) – Più di 4 anni, dove, rispettivamente, si riporta (in parentesi) l'intervallo dell'indice di durata ed il numero di anni fuori corso.

Se si procede al confronto tra il valore assunto dall'indice di durata per i due sessi si riscontra una lievissima differenza a livello complessivo (0.04) a favore del genere femminile, differenza questa che non presenta grande variabilità anche a livello di singole Facoltà. Se si fa, invece, riferimento alle statistiche di associazione, si rileva la presenza di un legame tra i due caratteri essendo abbastanza elevato il valore assunto dalla V di Cramer (0.10).

Un risultato inatteso è quello concernente la relazione esistente fra *residenza* degli studenti e *durata* degli studi universitari; infatti, la residenza degli studenti non sembra incidere in maniera così netta sul valore dell'indice di durata: tale valore rimane identico per gli studenti provenienti da Firenze e dalle province di Prato, Pistoia e Arezzo e lievemente più alto per coloro con residenza nelle altre province della Toscana; un valore più alto dell'indice, invece, si rileva per i giovani con residenza fuori dalla Toscana. Anche l'analisi delle statistiche di associazione (che rileva come la V di Cramer assume un valore pari a 0.08) sembra confermare la mancanza di un forte legame tra la residenza dello studente e la durata dei suoi studi universitari, ma è da ricordare che in realtà la variabile d'interesse sarebbe il *domicilio* degli studenti e non la loro residenza.

Una certa incidenza sulla *durata* degli studi ha, invece, il *titolo di studio dei genitori* dei laureati. Tale fatto è confermato sia dal valore della V di Cramer (0.12) sia dalla semplice lettura dei dati: da una situazione in cui entrambi i genitori sono laureati e l'indice assume un valore medio di 1.58 si passa a situazioni in cui in famiglia vi è al più una licenza elementare dove si registra un indice medio di 1.97. Piuttosto basso è invece il valore assunto dall'indice V (0.07) quando si considera la *classe sociale*[11] della famiglia di appartenenza.

Esaminando l'indice di durata in funzione della carriera preuniversitaria si osservano risultati del tutto prevedibili: i giovani con una carriera preuniversitaria regolare presentano un valore dell'indice di durata minore rispetto a chi ha affrontato l'esame di maturità con uno o più anni di ritardo; si registra una relazione inversa tra *voto alla maturità* e indice di durata (al crescere della votazione al diploma del laureato diminuisce il valore assunto dall'indice di durata); gli studenti che hanno concluso in tempi più contenuti la carriera universitaria sono quelli provenienti dal liceo scientifico e classico, mentre quelli che vi hanno impiegato più tempo provengono da altri tipi di maturità e scuole secondarie di tipo tecnico.

[11] Per la classificazione della variabile *classe sociale* si è adottato lo schema proposto da A. Cobalti e A. Schizzerotto, *La mobilità sociale in Italia*, Bologna, Il Mulino, 1994, adottato anche da ALMALAUREA. La posizione socio-economica può assumere le modalità borghesia, classe media impiegatizia, piccola borghesia e classe operaia. In proposito si segnala che gli imprenditori, i libero professionisti e i dirigenti, appartengono alla *borghesia* indipendentemente dal titolo; gli impiegati o intermedi con laurea sono nella *classe media impiegatizia*; i lavoratori in proprio, i soci di cooperative e i coadiuvanti appartengono alla *piccola borghesia*; gli impiegati con un titolo di studio della scuola dell'obbligo, gli operai ed i lavoratori a domicilio sono nella classe operaia

Le statistiche di associazione mostrano come vi sia un forte legame tra carriera preuniversitaria dello studente e tempo impiegato per conseguire la laurea: la significatività delle statistiche Chi-quadro di Pearson è sempre molto elevata e la V di Cramer è sempre superiore a 0.10. Da notare, è che la relazione più intensa si ha tra voto alla maturità e indice di durata (la V di Cramer è quasi pari a 0.13) a conferma dell'influenza esercitata da questo fattore sull'esito della carriera universitaria.

Per quanto riguarda le variabili relative alla carriera universitaria dello studente, è possibile innanzitutto osservare come, al crescere del *ritardo dell'immatricolazione* all'Università, cresca anche il valore dell'indice di durata. Ancora, è possibile osservare che coloro che hanno ottenuto risultati migliori sia a livello di *voto medio* agli esami che alla laurea impiegano meno tempo per conseguire il titolo. Il fatto che esista una forte relazione tra i risultati ottenuti dagli studenti per quanto riguarda la votazione conseguita e il tempo di conseguimento del titolo è confermato, inoltre, dal calcolo delle statistiche di associazione: la significatività del Chi-quadro è sempre molto elevata e la V di Cramer supera in entrambi i casi il valore di 0.14.

Riguardo il modo di vivere l'esperienza universitaria, l'elevato valore della V di Cramer (0.21) mostra come vi sia un legame molto forte tra tempi di conseguimento del titolo e la *frequenza* alle lezioni; infatti, chi frequenta con regolarità tutti o quasi tutti i corsi impiega meno tempo ($I_d = 1.75$) rispetto a chi frequenta saltuariamente, al più, alcuni corsi ($I_d = 2$).

Il fenomeno della frequenza alle lezioni è sicuramente collegato a quello delle *esperienze lavorative* durante la carriera universitaria: chi non lavora conclude gli studi nettamente prima (il valore dell'indice di durata è di 1.64 contro 1.86 per coloro che lavorano) e tra chi lavora ha più difficoltà a mantenere un buon ritmo di studi chi ha un contratto di lavoro di tipo stabile rispetto a chi ha rapporti di lavoro di tipo occasionale.

Interessanti sono i risultati che si ottengono sia dal calcolo dell'indice di durata in relazione alla necessità o meno di svolgere attività di *stage o tirocinio* per il completamento degli studi sia dal calcolo delle statistiche di associazione tra queste variabili: queste attività sembrano incidere positivamente sulla durata della carriera universitaria (il valore dell'indice è di 1.54 per coloro che sono stati coinvolti in tali attività e 1.84 per gli altri) ed il loro legame con i tempi di conseguimento del titolo sembra molto forte (il valore della V di Cramer è ben 0.26).

In relazione al rapporto che può esistere tra tempi di laurea e posizione nei confronti degli *obblighi di leva*, si registra un valore molto alto dell'indice di durata per gli studenti che hanno già svolto il servizio militare o civile ($I_d = 2.02$), mentre una situazione migliore si rileva per chi si trova nella condizione di non dover

svolgere il servizio militare ($I_d = 1.83$). Valori molto bassi dell'indice si osservano per gli studenti che stanno adempiendo agli obblighi di leva, o per coloro che stanno aspettando di farlo, probabilmente dovuto al fatto che, consapevoli dei propri obblighi, tali studenti hanno organizzato in maniera migliore i propri piani di studio (la forte relazione tra questa variabile e i tempi di durata è rilevata dalla V di Cramer pari a 0.37).

Per quanto riguarda la *soddisfazione* sull'esperienza universitaria appena conclusa, coloro che ne danno un giudizio ottimo sono coloro che hanno impiegato meno tempo a concludere gli studi ($I_d = 1.57$), mentre coloro che impiegano più tempo non sono coloro che ne danno un giudizio pessimo, ma mediocre. Anche in questo caso, comunque, la V di Cramer assume un valore abbastanza alto (0.14).

Relativamente all'ipotesi di *reiscrizione* all'Università, inoltre, sono coloro che vorrebbero reiscriversi allo stesso Corso di Laurea ad aver impiegato meno tempo alla conclusione degli studi ($I_d = 1.76$), mentre gli studenti che non vorrebbero reiscriversi all'Università sono coloro per cui si osserva il valore dell'indice di durata più elevato (1.96).

Da segnalare, infine, il forte grado di associazione tra indice di durata e *Facoltà* (V di Cramer pari a 0.26) che risulta ancora più elevato quando si misura l'associazione tra durata e *Corso di studi* (V di Cramer pari a 0.36).

4. Il modello gerarchico a due livelli

Come anticipato nell'introduzione, per procedere ad una più soddisfacente individuazione delle possibili determinanti dei tempi di conseguimento del titolo si è fatto ricorso ai modelli di regressione multilivello[12].

Nei due paragrafi precedenti è stato analizzato il fenomeno della durata degli studi nell'Ateneo fiorentino, sia relativamente al periodo 1980-2000 che ai laureati dell'anno solare 2000, facendo ricorso a tecniche statistiche sostanzialmente di tipo descrittivo, tecniche che non consentono di pervenire ad una misura dell'effetto netto delle determinanti (fattori esplicativi) del fenomeno di interesse che nel caso specifico è rappresentato dall'indice di durata. Tale possibilità è, invece, offerta, dai modelli di regressione e, in particolare, dai modelli di regressione di tipo multilivello. In proposito, vale la pena ricordare che il ricorso ad un modello di regressione multilivello è consigliabile ogni volta che le unità (dette unità di primo livello) su cui

[12] Al riguardo si possono consultare, tra gli altri, i volumi di Goldstein H. (2003) e Snijders A.B., Bosker R. J. (1999).

si rileva il fenomeno oggetto di studio risultano naturalmente aggregate in gruppi differenti (le unità di secondo livello), che a loro volta possono essere aggregate in unità di terzo livello e così via: in tali casi è ragionevole ritenere che la variabilità del fenomeno dipenda non solo da variabili esplicative individuali (o di primo livello), ma altresì dal fatto che un certo individuo appartenga ad un determinato gruppo avente caratteristiche peculiari che lo contraddistinguono dagli altri gruppi; nel caso specifico trattato, gli studenti (unità di primo livello) risultano naturalmente aggregati in Corsi di Laurea (unità di secondo livello).

Come già sottolineato, la variabile risposta considerata è l'indice di durata. Il "vantaggio" principale che proviene dall'utilizzo di questo indicatore è che attraverso di esso è possibile effettuare un confronto tra tempi di laurea osservati in diverse Facoltà e diversi CdL; unico "svantaggio" nell'utilizzo della variabile continua *ind_durata* è che la sua distribuzione ha un andamento di tipo normale, ma troncato a sinistra (il valore minimo osservato è 1)[13].

I dati cui si farà riferimento sono quelli relativi ai laureati dell'anno solare 2000 (unità di primo livello), che si sono immatricolati per la prima volta nell'Ateneo fiorentino a partire dall'a.a. 1980/81 e che non hanno cambiato corso di studi. Come precedentemente illustrato, sono stati esclusi dall'analisi quei laureati che non hanno compiuto l'intero ciclo di studi nell'Ateneo fiorentino e nello stesso Corso di Laurea, ottenendo così una popolazione di riferimento di *3978* unità; l'insieme di dati così ottenuto è risultato, però, composto da numerosi records con dati mancanti relativi ad alcune variabili e, dato che il software utilizzato per l'applicazione del modello di regressione multilivello esclude questi records dall'analisi e volendo evitare di ricorrere a complicate tecniche di imputazione, si è proceduto a cancellare tali records, ottenendo così una popolazione di riferimento composta da ***1896*** osservazioni.

Come unità di secondo livello sono stati scelti i Corsi di Laurea in cui gli studenti hanno conseguito il titolo di studio e non le rispettive Facoltà di appartenenza in quanto si è ritenuto che solo dall'analisi di questi si potessero ottenere informazioni utili alla spiegazione del fenomeno dei tempi di laurea all'interno dell'Ateneo fiorentino. Infatti, è possibile osservare come spesso le Facoltà sono suddivise al loro interno in CdL con caratteristiche sostanzialmente differenti tra loro. Inoltre, come prevedibile, anche dall'analisi dei tempi medi di laurea è possibile osservare come i diversi CdL all'interno delle stesse Facoltà differiscono notevolmente tra loro; solo per fare un esempio relativo alla Facoltà di

[13] Il fatto che la variabile *ind_durata* sia troncata a sinistra è una delle possibili cause della non normalità dei residui di regressione. Quest'ultimo aspetto ha comunque delle conseguenze soprattutto sul valore degli errori standard delle stime dei coefficienti di regressione e delle componenti di varianza (e di conseguenza sulla validità dei test utilizzati), e non su quello delle stime puntuali di tali parametri.

Economia, il CdL in Scienze Statistiche ed Attuariali ha un indice di durata pari a 1.60 mentre per Economia e Commercio si osserva un valore dell'indice addirittura pari a 1.97 (quasi il doppio della durata legale!). Un'ultima osservazione che conferma quanto appena descritto riguarda il già segnalato (cfr. paragrafo 2) alto grado di associazione riscontrato sia tra i tempi di laurea degli studenti e le Facoltà frequentate sia tra i tempi di laurea degli studenti ed i loro CdL: associazione molto forte nel primo caso, ma notevolmente più forte nel secondo.

Coerentemente a quanto suggerito dalla teoria i passi della procedura di stima impiegata sono stati:

- stima del ***modello nullo***, al fine di verificare, attraverso la stima della varianza dei residui di secondo livello, la ragionevolezza del ricorso ad un'analisi multilivello e di scomporre la variabilità del fenomeno tra variabilità *entro* i gruppi e variabilità *tra* gruppi.
- stima del ***modello a intercetta casuale*** comprendente tutte le variabili di primo livello (compresi termini quadratici e interazioni) risultate significative.
- stima del ***modello a intercetta casuale finale***[14] (due livelli) costituito dalle variabili di primo e secondo livello (compresi termini quadratici e interazioni) risultate significative.

Di seguito verranno riportati e commentati soltanto i risultati relativi alla stima dei due Modelli a intercetta casuale comprendenti, rispettivamente, le sole variabili di primo livello e quelle di primo e secondo livello relative alla stima finale.

Il modello utilizzato è un modello multilivello ad intercetta casuale del tipo:

$$Y_{ij} = \left(\gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j\right) + \left(u_{oj} + r_{ij}\right)$$

dove:

$$r_{ij} \sim iid\ \mathrm{N}(0,\sigma^2)$$
$$u_{0j} \sim iid\ \mathrm{N}(0,\tau_{00})$$

i residui del modello, r_{ij} e $u_{0j,}$ sono indipendenti tra loro, j è l'indice utilizzato per descrivere i gruppi (Corsi di Laurea - $j = 1,2,...37$), mentre i è l'indice che descrive le unità (laureati all'interno di ogni gruppo - $i = 1,2,...n_j$); la variabile risposta Y_{ij} (misurata al livello individuale) è rappresentata dall'indice di durata degli studi ed ha distribuzione di tipo normale troncata nella coda di sinistra (il valore minimo osservato è 1); X_{ij} indicano le variabili esplicative, di primo livello mentre Z_j rappresentano le variabili esplicative di secondo livello. I termini r_{ij} e u_{0j} rappresentano gli errori residui del modello (ovvero quella parte di variabilità di Y_{ij} che non è catturata dalle variabili esplicative) rispettivamente a livello individuale ed a livello di gruppo

[14] Si puntualizza che sono stati testati anche modelli con coefficienti casuali; in questo caso, però, l'algoritmo utilizzato non ha mai raggiunto la convergenza.

Per l'applicazione del modello lineare gerarchico è stata utilizzata la PROC MIXED del software SAS[15].

4.1 Il modello stimato ed i risultati ottenuti

Come già segnalato, il primo passo della procedura prevede la stima di un modello di regressione a due livelli senza variabili esplicative, ovvero del cosiddetto modello nullo.

In particolare, attraverso il modello nullo è stato possibile esplicitare il coefficiente di correlazione intra-classe ρ, che misura il grado di omogeneità tra osservazioni appartenenti allo stesso gruppo: in questo caso, quasi il 40% della varianza totale dell'indice di durata è dovuta all'effetto del Corso di Laurea. Inoltre, è stato calcolato il valore della devianza - ovvero il grado di "non adattabilità" del modello (Hox J.J., 2002) - che è stato successivamente utilizzato come benchmark per il confronto di differenti modelli.

Una volta verificata l'effettiva esistenza di un'organizzazione dei dati su due livelli di analisi, lo studio è proseguito con la stima del modello multilivello (completo) contenente le variabili esplicative.

Relativamente alle variabili categoriche introdotte nel modello, si è reso necessario individuare per ciascuna di esse una modalità base o di riferimento nei confronti della quale valutare l'effetto di tutti gli altri livelli esistenti[16], come in una normale regressione. Relativamente a tale aspetto si è scelto di considerare come modalità base di ciascuna covariata la caratteristica più diffusa nella popolazione di studio, secondo i risultati ottenuti attraverso l'analisi descrittiva preliminare; l'individuo che possiede tutte queste caratteristiche verrà denominato *individuo-base*.

Per semplificare l'interpretazione dei risultati, si è scelto di centrare le variabili continue di primo livello rispetto alla loro media di gruppo e quelle di

---

[15] La procedura PROC MIXED permette di scegliere i metodi di stima dei parametri. Nel contesto dei modelli multilivello gli stimatori maggiormente impiegati sono quelli di *Massima Verosimiglianza* (Maximum Likelihood, ML) e quelli della *Massima Verosimiglianza residua* (Residual Maximum Likelihood, REML). Questi due metodi danno risultati molto simili per quanto riguarda la stima dei coefficienti di regressione mentre differiscono maggiormente nella stima delle componenti della varianza; inoltre, la devianza calcolata attraverso il metodo REML può essere utilizzata nei *test del rapporto di verosimiglianza* solo se i due modelli comparati sono composti dalla stessa parte fissa e differente parte casuale. Per tale motivo in questa applicazione è stato scelto di calcolare le stime dei coefficienti di regressione attraverso il metodo della *Massima Verosimiglianza* (specificando l'argomento METHOD=ML nella procedura PROC MIXED).

[16] Dovendo utilizzare queste variabili ai fini dell'applicazione ai dati del modello multilivello, si è scelto talvolta di adottare classificazioni delle variabili meno dettagliate rispetto a quelle presentate durante l'analisi descrittiva. Naturalmente, un'analisi che tenga presente classificazioni più dettagliate delle variabili potrà essere spunto per approfondimenti successivi.

secondo livello rispetto alla loro media totale[17]. Si è scelto dunque di utilizzare l'approccio *group mean centering*, che consente di considerare il cosiddetto "frog-pond effect" (Hox J.J., 2002). Nel caso qui considerato, ad esempio, scegliendo come variabile esplicativa il voto al diploma riportato dagli studenti (*voto*), studiare questo effetto vuol dire analizzare come varia la relazione tra tempo di conseguimento del titolo e la variabile voto, in dipendenza dal voto medio al diploma osservabile nello specifico Corso di Laurea.

Le variabili esplicative utilizzate per la selezione del modello sono state scelte in base ai suggerimenti forniti dalla conoscenza del fenomeno e alle conclusioni risultanti dall'analisi descrittiva svolta.

Le variabili di primo livello, riportate in Tabella 1, possono essere classificate in tre gruppi, ognuno dei quali riguarda un differente aspetto della vita dello studente:

- **variabili legate ai caratteri strutturali** (variabili 1-4)
- **variabili legate alla preparazione preuniversitaria** (variabili 5-7)
- **variabili legate all'esperienza universitaria** (variabili 8-14).

Tabella 1. *Variabili esplicative di primo livello*

n.	*Nome variabile*	Descrizione	Modalità di risposta	Categoria di riferimento
1	*sesso*	Sesso	1= maschi 2= femmine	Femmine
2	*residenza*	Residenza	1= fuori Toscana 2= altra provincia Toscana 3= Po - Pt - Ar 4= Firenze	Firenze
3	*tit_gen*	Titolo di studio dei genitori[18]	1= al più un diploma inferiore 2= almeno un genitore con diploma superiore 3= almeno un genitore con laurea	Almeno un genitore con laurea
4	*cl_sociale*	Classe sociale della famiglia di origine	1= borghesia 2= classe operaia 3= classe media impiegatizia o piccola borghesia	Classe media impiegatizia o piccola borghesia

[17] Il calcolo della media di gruppo e della media totale delle medie di gruppo, come suggerito da Snijders e Bosker (1999), è stato effettuato su tutti i valori individuali a disposizione per la determinata variabile analizzata prima della procedura di cancellazione dei records aventi dati mancanti relativi ad altre variabili.

[18] Per questa variabile è stata scelta come categoria di riferimento non la modalità più frequente (che sarebbe stata "almeno un genitore con diploma superiore"), ma quella ritenuta più significativa per studiare il contributo del livello di istruzione presente in famiglia sui tempi di laurea dello studente.

n.	*Nome variabile*	Descrizione	Modalità di risposta	Categoria di riferimento
5	*diploma*	Tipo di diploma di scuola superiore	1= altro 2= tecnico 3= liceale	Liceale
6	*voto_dipl*	Voto al diploma di scuola superiore (in sessantesimi)		Variabile centrata rispetto alla media di gruppo
7	*eta_dipl*	Età al diploma di scuola superiore	1= maggiore di 19 anni (percorso di studi non regolare) 2= minore o uguale a 19 anni (percorso di studi regolare)	Percorso di studi regolare
8	*rit_iscr*	Tempo tra l'esame di maturità e l'iscrizione all'Università	1= maggiore o uguale a 1 anno 2= 0 anni	0 anni
9	*frequenz*	Frequenza alle lezioni	1= non regolare 2= regolare ad almeno alcuni corsi	Frequenza regolare
10	*esp_lav*	Tipo di esperienze lavorative durante gli studi universitari[19]	1= stabile 2= non stabile	Non stabile
11	*tiroc*	Tirocinio o stage svolto per il completamento degli studi	1= sì 2= no	No
12	*voto_30*	Votazione media riportata agli esami (in trentesimi)		Variabile centrata rispetto alla media di gruppo
13	*tempo_tesi*	Tempo impiegato per la stesura della tesi (in mesi)		Variabile centrata rispetto alla media di gruppo
14	*militare*	Servizio militare o civile svolto durante gli studi	1= svolto durante gli studi universitari 2= non svolto durante gli studi universitari.	Non svolto durante gli studi universitari

[19] La scelta di inserire come variabile di primo livello il tipo di esperienze lavorative avute durante gli studi universitari e non il fatto di aver avuto o meno tali esperienze è dovuta al fatto che tutti i laureati appartenenti alla popolazione analizzata hanno indicato di aver avuto almeno un'esperienza di tipo lavorativo durante gli studi.

Nella tabella, per ogni variabile sono stati indicati: un nome convenzionale, una breve descrizione del suo significato, le modalità di risposta ricodificate e la categoria di riferimento (categoria riferita all'individuo-base).

Da quanto indicato nella Tabella 1 emerge il profilo dell'individuo-base che risulta essere:

- **femmina**
- **residente a Firenze**
- **almeno un genitore con laurea**
- **appartenenza alla classe media impiegatizia o alla piccola borghesia**
- **diploma di tipo liceale**
- **voto al diploma medio all'interno del proprio CdL**
- **percorso di studi pre-universitari regolare**
- **iscritto subito all'Università**
- **frequenza regolare alle lezioni universitarie**
- **esperienza di lavoro non stabile**
- **nessuna attività di tirocinio o stage per il completamento degli studi**
- **votazione media agli esami pari alla media nel proprio CdL**
- **tempo medio all'interno del proprio CdL per la stesura della tesi.**

Le variabili di secondo livello (Tabella 2) utilizzate sono di tipo contestuale, ovvero variabili che si riferiscono a caratteristiche proprie di ogni Corso di Laurea, espresse attraverso il valore della media di gruppo delle variabili di primo livello[20].

Seguendo la strategia di selezione[21] sopra descritta si è ottenuto il modello:

$$
\begin{aligned}
& ind\_dur_{ij} = \gamma_{00} + \gamma_{10j} sesso_{ij} + \gamma_{20j} tit\_gen_{ij} + \gamma_{30j} diploma_{ij} + \gamma_{40j} voto\_dipl_{ij} + \\
& + \gamma_{50j} frequenz_{ij} + \gamma_{60j} esp\_lav_{ij} + \gamma_{70j} tiroc_{ij} + \gamma_{80j} voto\_30_{ij} + \gamma_{90j} tempo\_tesi_{ij} + \\
& + \gamma_{10,0j} militare_{ij} + \gamma_{11,0j} sesso * tempo\_tesi_{ij} + \gamma_{12,0j} sesso * voto\_30_{ij} + \\
& + \gamma_{13,0j} diploma * tit\_gen_{ij} + \gamma_{14,0j} voto\_dipl * tempo\_tesi_{ij} + \left(u_{0j} + r_{ij}\right)
\end{aligned}
$$

[20] Nel caso in cui le variabili di primo di livello sono categoriche, le rispettive variabili di secondo livello sono espresse dalla percentuale di studenti in ogni CdL per cui si osserva una modalità di risposta diversa da quella base.

[21] Si è provveduto a migliorare di volta in volta il modello inserendovi differenti variabili esplicative e le loro interazioni e togliendo quelle covariate non risultate significative attraverso l'analisi del *test di Wald* al livello di significatività del 5%. Per un confronto tra modelli ottenuti attraverso l'inserimento di parametri aggiuntivi ci si è basati sul *test della devianza* al livello di significatività del 5%; nonostante il software proceda in automatico al calcolo della statistica della *devianza*, il test ad esso relativo è stato calcolato manualmente. Inoltre, per confrontare modelli con differenti parametri, è stato utilizzato l'indice di adattamento *AIC* (Akaike's Information Criterion).

Tabella 2. *Variabili esplicative di secondo livello*

n.	*Nome variabile*	Descrizione	Modalità di risposta	Valore di riferimento
1	*maschi_CdL*	Percentuale di maschi nel CdL		
2	*noliceali_CdL*	Percentuale di studenti con un diploma superiore diverso dal diploma liceale		
3	*voto_dipl_CdL*	Voto medio nel CdL riportato dagli studenti al diploma di scuola superiore (in sessantesimi)		Variabile centrata rispetto alla media totale
4	*frequenz_CdL*	Percentuale di studenti che non frequentano regolarmente tutte o quasi tutte lezioni		
5	*tiroc_CdL*	Percentuale di studenti che hanno svolto attività di tirocinio o stage per il completamento degli studi		
6	*voto_30_CdL*	Votazione media nel CdL riportata agli esami (in trentesimi)		Variabile centrata rispetto alla media totale
7	*tempo_tesi_CdL*	Tempo medio nel CdL impiegato per la stesura della tesi (in mesi)		Variabile centrata rispetto alla media totale
8	*numchiuso*	Presenza nel CdL di limitazioni all'accesso	1= sì 2= no	No

I risultati ottenuti sono riportati nella Tabella 3; i coefficienti riportati in quarta colonna indicano quanto cambia il tempo di laurea di uno studente iscritto ad un Corso di Laurea di durata quadriennale al variare di un'unità delle variabili esplicative corrispondenti.

L'intercetta $\gamma_{00} = 1.65$ indica il valore dell'indice di durata (che per un Corso di Laurea di durata quadriennale corrisponde a circa 6 anni e 7 mesi) quando le variabili esplicative di primo livello assumono un valore pari a 0 e gli errori di primo e di secondo livello sono nulli, ovvero quando è osservato il tempo di conseguimento del titolo del cosiddetto individuo-base all'interno di un CdL-base.

Tra le variabili di primo livello che non sono risultate significative (*residenza, cl_sociale, eta_dipl, rit_iscr*) vi è la residenza dello studente. Tale risultato conferma le osservazioni svolte in precedenza: la residenza degli studenti non sembra incidere in maniera così netta sul valore dell'indice di durata anche se, come già segnalato, la variabile da considerare dovrebbe essere il domicilio e non la residenza degli studenti.

Tabella 3. *Effetti fissi: coefficienti di regressione*

Effetto	Categorie	Stima	STIMA (espressa in mesi)	Errore standard	Valore t	Pr > |t|
Intercetta		1.647	79.08	0.0567	29.05	<.0001
sesso	1	-0.102	-4.891	0.0233	-4.38	0.0001
sesso	2	0	0	.	.	.
tit_gen	1	0.188	9.010	0.0294	6.39	<.0001
tit_gen	2	0.109	5.251	0.0281	3.9	0.0002
tit_gen	3	0	0	.	.	.
diploma	1	0.0243	1.164	0.0577	0.42	0.6759
diploma	2	0.180	8.664	0.0571	3.16	0.0027
diploma	3	0	0	.	.	.
voto_dipl		-0.011	-0.513	0.0014	-7.65	<.0001
frequenz	1	0.083	4.00	0.0352	2.37	0.0319
frequenz	2	0	0	.	.	.
esp_lav	1	0.132	6.350	0.0224	5.91	<.0001
esp_lav	2	0	0	.	.	.
tiroc	1	-0.140	-6.730	0.0323	-4.34	0.0002
tiroc	2	0	0	.	.	.
voto_30		-0.011	-0.544	0.0095	-1.19	0.2338
tempo_tesi		0.013	0.647	0.0022	5.98	<.0001
militare	1	0.228	10.944	0.0265	8.6	<.0001
militare	2	0	0	.	.	.
sesso*tempo_tesi	1	0.009	0.419	0.0033	2.65	0.0082
sesso*tempo_tesi	2	0	0	.	.	.
sesso*voto_30	1	-0.032	-1.561	0.0128	-2.53	0.0114
sesso*voto_30	2	0	0	.	.	.
diploma*tit_gen	1*1	0.056	2.683	0.0662	0.84	0.4026
diploma*tit_gen	1*2	0.057	2.720	0.0705	0.8	0.4257
diploma*tit_gen	1*3	0	0	.	.	.
diploma*tit_gen	2*1	-0.181	-8.688	0.0638	-2.84	0.0068
diploma*tit_gen	2*2	-0.190	-9.106	0.0666	-2.85	0.0065
diploma*tit_gen	2*3	0	0	.	.	.
diploma*tit_gen	3*1	0	0	.	.	.
diploma*tit_gen	3*2	0	0	.	.	.
diploma*tit_gen	3*3	0	0	.	.	.
voto_dipl*tempo_tesi		0.001	0.030	0.0003	2.37	0.0177

Le stime dei coefficienti di regressione che hanno un valore positivo sono quelle riguardanti la relazione tra indice di durata e le variabili *tit_gen, diploma,*

frequenz, esp_lav, tempo_tesi, militare: questo significa che "allontanandosi" dal profilo base si ha un incremento dell'indice di durata e quindi un aumento dei tempi di laurea.

La variabile *diploma* è significativa se presa nel suo insieme ma, dall'analisi dei risultati riportati in Tabella 3, il passaggio da una situazione in cui lo studente ha una preparazione preuniversitaria di tipo liceale ad una situazione in cui il tipo di scuola superiore frequentata ricade nella categoria "altro" (né liceale, né tecnica) non sembra avere influenza sui tempi di laurea degli studenti, mentre significativo è avere una preparazione tecnica rispetto a quella liceale[22].

Molto interessante, inoltre, è osservare che, assunto come casuale l'effetto del Corso di Laurea, essere maschio ha un effetto positivo sui tempi di laurea. Qualora le altre variabili esplicative hanno valore pari a 0, si avrà:

$$ind\_dur_j(maschio) = 1{,}65 - 0{,}10(maschio = 1)_{ij} + (u_{0j} + r_{ij})$$

e

$$ind\_dur_j(femm) = 1{,}65 - 0{,}10(femm = 0)_{ij} + (u_{0j} + r_{ij})$$

quindi, un valore dell'indice di durata di circa 1.55 (che per un Corso di Laurea di durata quadriennale corrisponde a circa 6 anni e 2 mesi) per i maschi e 1.65 per le femmine (corrispondente a circa 6 anni e 7 mesi).

Una volta inserite nel modello le variabili esplicative di primo livello e le loro interazioni, sono state aggiunte anche le variabili esplicative di secondo livello. Il modello di regressione specificato separatamente per i singoli gruppi risulta, pertanto, espresso dall'equazione:

$$Y_{ij} = \left(\beta_{0j} + \sum_p \beta_{pj} X_{pij} \right) + r_{ij}$$

dove:

$$\beta_{0j} = \gamma_{00} + \sum_q \gamma_{0q} Z_{qj} + u_{0j}$$

Seguendo la strategia di selezione del modello sopra descritta, per il coefficiente β_{0j} si è ottenuta l'equazione:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} voto\_dipl\_CdL + \gamma_{02} numchiuso + u_{0j}$$

Quindi, il modello completo assume la forma:

[22] Si segnala che sono state utilizzate anche altre classificazioni interne della variabile diploma, ma nessuna di queste ha apportato dei miglioramenti complessivi al modello.

$$ind\_dur_{ij} = \gamma_{00} + \gamma_{10j} sesso_{ij} + \gamma_{20j} tit\_gen_{ij} + \gamma_{30j} diploma_{ij} + \gamma_{40j} voto\_dipl_{ij} +$$
$$+ \gamma_{50j} frequenz_{ij} + \gamma_{60j} esp\_lav_{ij} + \gamma_{70j} tiroc_{ij} + \gamma_{80j} voto\_30_{ij} + \gamma_{90j} tempo\_tesi_{ij} +$$
$$+ \gamma_{10,0j} militare_{ij} + \gamma_{11,0j} sesso * tempo\_tesi_{ij} + \gamma_{12,0j} sesso * voto\_30_{ij} +$$
$$+ \gamma_{13,0j} diploma * tit\_gen_{ij} + \gamma_{14,0j} voto\_dipl * tempo\_tesi_{ij} +$$
$$+ \gamma_{01} voto\_dipl\_CdL + \gamma_{02} numchiuso + (u_{0j} + r_{ij})$$

Nella Tabella 4 sono riportati i risultati ottenuti.

L'intercetta $\gamma_{00} = 1.69$ indica il valore dell'indice di durata (che per un Corso di Laurea di durata quadriennale corrisponde a circa 6 anni e 9 mesi) quando tutte le variabili esplicative hanno un valore pari a 0 e gli errori di primo e di secondo livello sono nulli, ovvero quando si osserva il tempo di conseguimento del titolo del cosiddetto individuo-base all'interno di un CdL-base.

I coefficienti di regressione delle covariate *voto_dipl_CdL* e *numchiuso* esprimono l'effetto di queste due variabili di secondo livello sulla media tra gruppi dell'indice di durata. Il fatto che il coefficiente γ_{01} sia negativo indica che all'aumentare del voto medio che gli studenti hanno ottenuto al diploma migliorano i tempi di laurea medi all'interno del gruppo. Ancora più interessante, inoltre, è analizzare il coefficiente di regressione della variabile *numchiuso*. Quando $u_{0j} = 0$, si ha:

$$\beta_{0j} = 1.69 - 0.37(numchiuso)$$

quindi:

$$\beta_{0j}(numchiuso = 1) = 1.32$$
$$\beta_{0j}(numchiuso = 0) = 1.69$$

Questo vuol dire che il valore medio dell'indice di durata risulta inferiore di 0.37 (corrispondente a circa 1 anno e quasi 6 mesi per un Corso di Laurea di durata quadriennale) per quei CdL in cui vi è il cosiddetto numero chiuso rispetto a quelli in cui non esiste nessuna limitazione all'accesso per le immatricolazioni.

Le variabili di secondo livello che singolarmente sono risultate significative, ma che successivamente, attraverso l'applicazione del test della devianza e il calcolo dell'indice di adattamento AIC sono state escluse dal modello, sono state: *noliceali_CdL*, *frequenz_CdL*, *tempo_tesi_CdL*. Questo indica che vi è una relazione tra il tipo di preparazione preuniversitaria degli studenti che si iscrivono in un determinato CdL, la loro frequenza media e il tempo medio richiesto per la stesura della tesi e i tempi medi di laurea osservati all'interno dello specifico CdL, ma che le variabili *voto_dipl_CdL* e *numchiuso* spiegano una maggior variabilità del fenomeno risposta.

Tabella 4. *Effetti fissi: coefficienti di regressione*

Effetto	Categorie	Stima	STIMA (espressa in mesi)	Errore standard	Valore T	Pr > |t|
Intercetta		1.689	81.058	0.0500	33.78	<.0001
sesso	1	-0.096	-4.612	0.0229	-4.19	0.0002
sesso	2	0	0	.	.	.
tit_gen	1	0.168	8.050	0.0291	5.77	<.0001
tit_gen	2	0.089	4.291	0.0278	3.21	0.002
tit_gen	3	0	0	.	.	.
diploma	1	-0.002	-0.109	0.0570	-0.04	0.9685
diploma	2	0.170	8.150	0.0563	3.01	0.004
diploma	3	0	0	.	.	.
voto_dipl		-0.011	-0.513	0.0014	-7.75	<.0001
frequenz	1	0.085	4.068	0.0347	2.44	0.0275
frequenz	2	0	0	.	.	.
esp_lav	1	0.127	6.086	0.0221	5.75	<.0001
esp_lav	2	0	0	.	.	.
tiroc	1	-0.149	-7.162	0.0317	-4.7	<.0001
tiroc	2	0	0	.	.	.
voto_30		-0.010	-0.479	0.0094	-1.06	0.288
tempo_tesi		0.013	0.625	0.0022	5.85	<.0001
militare	1	0.209	10.042	0.0262	7.97	<.0001
militare	2	0	0	.	.	.
sesso*tempo_tesi	1	-0.035	-1.674	0.0127	-2.76	0.0059
sesso*tempo_tesi	2	0	0	.	.	.
sesso*voto_30	1	0.009	0.442	0.0032	2.83	0.0047
sesso*voto_30	2	0	0	.	.	.
diploma*tit_gen	1*1	0.070	3.342	0.0652	1.07	0.2915
diploma*tit_gen	1*2	0.073	3.527	0.0695	1.06	0.2961
diploma*tit_gen	1*3	0	0	.	.	.
diploma*tit_gen	2*1	-0.174	-8.338	0.0629	-2.76	0.0083
diploma*tit_gen	2*2	-0.176	-8.462	0.0657	-2.68	0.0101
diploma*tit_gen	2*3	0	0	.	.	.
diploma*tit_gen	3*1	0	0	.	.	.
diploma*tit_gen	3*2	0	0	.	.	.
diploma*tit_gen	3*3	0	0	.	.	.
voto_dipl*tempo_tesi		0.001	0.029	0.0003	2.29	0.0221
voto_dipl_CdL		-0.050	-2.418	0.0155	-3.26	0.0025
numchiuso	1	-0.378	-18.154	0.0521	-7.26	.
numchiuso	2	0	0	.	.	.

Passando all'analisi dei coefficienti di regressione delle variabili di primo livello e delle loro interazioni si rileva come questi sostanzialmente non siano

cambiati rispetto al modello precedentemente presentato. Inoltre, ancora una volta, le stime dei coefficienti di regressione che hanno un valore positivo sono quelle riguardanti la relazione tra indice di durata e le variabili *tit_gen*, *diploma*, *frequenz*, *esp_lav*, *tempo_tesi*, *militare*.

Infine, è stata svolta un'analisi dei residui sia di primo che di secondo livello per verificare alcune ipotesi poste alla base del modello.

In particolare, è risultato che il valore della varianza di entrambe le componenti residue sono inferiori rispetto a quelli stimati attraverso il modello nullo: parte della variabilità della variabile risposta dovuta sia all'effetto individuale che a quello di gruppo è stata spiegata attraverso l'inserimento delle variabili esplicative. Inoltre, è stato calcolato il coefficiente residuo di correlazione intra-classe; comparando tale valore con quello calcolato attraverso il modello vuoto, si osserva come attraverso l'inserimento delle variabili esplicative vi sia stata una diminuzione della percentuale della varianza totale dell'indice di durata dovuta all'effetto del Corso di Laurea.

L'analisi dei residui di secondo livello relativi al modello nullo ed al modello finale ha permesso, tra le altre cose, di ottenere interessanti informazioni per quanto riguarda il cosiddetto "effetto Corso di Laurea" sui tempi di conseguimento del titolo degli studenti; infatti, è stato possibile costruire una sorta di "graduatoria" dei Corsi di Laurea in termini di efficacia dovuta esclusivamente alle loro peculiarità. In particolare, è stato rilevato che parte della variabilità dei tempi di laurea degli studenti che attraverso una prima analisi di natura descrittiva sembrava dovuta all'effetto proprio dei diversi Corsi di Laurea è in realtà spiegabile altrimenti. Solo per far un esempio, presupponendo come casuale l'effetto proprio del gruppo, i Corsi di Laurea in Medicina e Chirurgia ed in Odontoiatria e protesi dentaria sono risultati molto efficienti in termini di tempi di conseguimento del titolo degli studenti mentre dopo l'inserimento nel modello delle variabili esplicative tale effetto positivo è in qualche modo "scomparso". Questo può essere spiegato dal fatto che i due CdL considerati godono degli effetti positivi nei confronti dei tempi di laurea esercitati dalla presenza delle limitazioni all'accesso delle immatricolazioni, effetti che sono stati controllati inserendo nel modello la variabile esplicativa relativa alla presenza del numero chiuso all'interno del corso di studi. Inoltre, si rileva che sia presupponendo la totale casualità dell'effetto del Corso di Laurea sia cercando di spiegarlo attraverso l'utilizzo delle variabili esplicative, gli effetti "estremamente" positivi e quelli "estremamente" negativi dei CdL rimangono gli stessi: solo per fare un esempio, i CdL ad avere un effetto migliore sono Psicologia, Scienze Forestali e Scienze dell'Educazione, mentre quelli ad avere effetti peggiori sono Materie Letterarie e Lingue e Letterature Straniere Moderne.

5. Conclusioni

In questa nota sono stati descritti molto sinteticamente i risultati di un'analisi dei tempi di conseguimento del titolo dei laureati dell'Ateneo fiorentino soffermando l'attenzione sui laureati dell'anno solare 2000; su tali dati si è anche proceduto alla stima di un modello multilivello. Obiettivo dell'applicazione è stato quello di esaminare i comportamenti individuali degli studenti universitari per quanto riguarda i tempi di laurea in funzione di variabili esplicative di primo e secondo livello, dove le unità di primo livello sono gli studenti e quelle di secondo livello sono i Corsi di Laurea. I risultati dell'analisi possono essere utili ai fini di una miglior comprensione di un fenomeno considerato unanimemente una criticità assoluta del sistema universitario italiano.

Il modello a cui si è giunti è un modello lineare gerarchico ad intercetta casuale, in cui si suppone un effetto costante tra gruppi delle variabili esplicative sulla variabile risposta (Y_{ij}).

Le covariate di primo livello che sono risultate significative nello spiegare i tempi di laurea degli studenti sono sia variabili legate ai loro caratteri strutturali (genere, titolo di studio dei genitori), sia variabili legate alla loro preparazione preuniversitaria (il tipo di scuola superiore frequentata, la votazione riportata all'esame di maturità), sia variabili legate alla loro carriera universitaria (frequenza alle lezioni, il tipo di esperienze lavorative avute durante gli studi, lo svolgimento o meno di un tirocinio, la votazione riportata agli esami, il tempo impiegato per la stesura della tesi, lo svolgimento o meno del servizio militare o civile durante gli studi); inoltre, è stato possibile rilevare come il fenomeno "tempi di laurea" è spiegato anche da alcune interazioni tra le variabili di primo livello. A livello di Corso di Laurea, i tempi impiegati dagli studenti per il conseguimento del titolo dipendono sia dal voto medio riportato dagli stessi all'esame di maturità, sia dalla presenza o meno del cosiddetto "numero chiuso" all'interno del CdL.

Naturalmente, le politiche universitarie d'intervento che dovrebbero essere messe in atto ai fini della risoluzione del problema dell'eccessiva durata degli studi potranno riguardare solo variabili legate alla vita universitaria degli studenti che in qualche modo "agiscono" sui tempi di laurea degli stessi. Dall'analisi di tali variabili è possibile rilevare come queste si trovano in relazione con l'organizzazione interna dei corsi di studi: il tipo di frequenza richiesta, il fatto di lasciar tempo o meno per diversi tipi di esperienze lavorative (stabili o non stabili), il tempo richiesto per la stesura tesi, l'obbligatorietà o meno di svolgere attività di tirocinio o stage, possono essere considerati indici di una "buona o cattiva" organizzazione interna della corso di studi. Anche il fatto che dall'applicazione del modello sia risultato significativo il cosiddetto "numero chiuso" nello spiegare le differenze tra CdL rileva come una miglior organizzazione del corso incida sui tempi di laurea degli studenti. Infatti,

generalmente, la limitazione all'accesso delle immatricolazioni, convogliando all'interno dei Corsi di Laurea solo un circoscritto numero di studenti solitamente molto motivati, ha degli effetti positivi sull'organizzazione della didattica, sulla gestione dei servizi agli studenti, sul numero dei docenti per studente, ecc. Anche l'inserimento nei curricula di attività di tirocinio o stage può avere degli effetti positivi sui tempi di laurea degli studenti agendo positivamente sull'organizzazione interna dei piani di studi dei CdL.

I risultati delle analisi svolte, molto sommariamente richiamati in questa nota, giustificano ampiamente, a nostro parere, il ricorso ai modelli multilivello quando si procede all'analisi di dati che riguardano gli studenti universitari[23]; infatti, è del tutto evidente la natura gerarchica dei dati: le unità di primo livello sono gli studenti o i laureati/diplomati, mentre le unità di secondo livello sono i corsi di studio. Ovviamente la gerarchizzazione può essere estesa ad un numero di livelli più elevato: ad esempio le Facoltà possono rappresentare il terzo livello e gli Atenei il quarto livello.

Riferimenti bibliografici

BULGARELLI G. (2002) *Esito degli studi degli immatricolati dell'Ateneo Fiorentino dal 1980/81 al 1997/9,* Università degli Studi di Firenze, consultabile anche sul sito www.unifi.it/aut_dida/indexval.html.

CHIANDOTTO B. (2002) *Valutazione dei processi formativi: cosa, come e perché, in Valutazione della Didattica e dei Servizi nel Sistema Università*.In D'ESPOSITO M.R. (a cura di) *Valutazione della Didattica e dei Servizi nel Sistema Università*. CUSL, Salerno 2002.

CHIANDOTTO B., BACCI S., BERTACCINI B. (2004) *I laureati e diplomati dell'Ateneo Fiorentino dell'anno 2000: profilo e sbocchi professionali*, Università degli Studi di Firenze.

CHIANDOTTO B., BERTACCINI B. (2003) *I laureati e diplomati dell'Ateneo Fiorentino dell'anno 1999: profilo e sbocchi professionali*, Università degli Studi di Firenze.

[23] In tale direzione si sta movendo da tempo il gruppo ***VALMON*** (**Val**utazione e **Mon**itoraggio). Il gruppo, coordinato da B.Chiandotto e costituito da laureandi, dottorandi e docenti del Dipartimento di Statistica dell'Università degli Studi di Firenze, da diversi anni svolge attività di studio e ricerca nel contesto della valutazione e del monitoraggio dei processi formativi che si svolgono nell'Ateneo fiorentino. Tale interesse è testimoniato, tra l'altro, da altri due lavori presentati in questa sede: "*Un modello multilivello per l'analisi della condizione occupazionale dei laureati"* (Chiandotto B. e Bacci S.); "*L'abbandono degli studi universitari"* (Chiandotto B. e Giusti C.).

COBALTI A., SCHIZZEROTTO A. (1994) *La mobilità sociale in Italia*, Il Mulino, Bologna.

GOLDSTEIN H. (2003) *Multilevel Statistical Models*, Edward Arnold, London.

HOX J.J. (2002) *Multilevel Analysis: Techniques and Applications*, LAWRENCE ERLBAUM ASSOCIATES, Mahwah (New Jersey), London.

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

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

VARRIALE R. (2004) *Tempi di conseguimento del titolo nell'Università degli Studi di Firenze nel periodo 1980-2000 e applicazione di un modello lineare gerarchico ai laureati nell'anno solare 2000,* Tesi di laurea, Università degli Studi di Firenze.

Specification issues in stratified variance component ordinal response models

Da: Grilli, L., Rampichini C. (2002) Specification issues in stratified variance component ordinal response models, Statistical Modelling, Vol. 2, pp. 251-264.

Specification issues in stratified variance component ordinal response models

Leonardo Grilli and Carla Rampichini
Department of Statistics 'G. Parenti', University of Florence, Italy

Abstract: The paper presents some criteria for the specification of ordinal variance component models when the units are grouped in a limited number of strata. The base model is specified using a latent variable approach, allowing the first level variance, the second level variance, and the thresholds to vary according to the strata. However this model is not identifiable. The paper discusses some alternative assumptions that overcome the identification problem and illustrates a general strategy for model selection. The proposed methodology is applied to the analysis of course programme evaluations based on student ratings, referring to three different schools of the University of Florence. The adopted model takes into account both the ordinal scale of the ratings and the hierarchical nature of the phenomenon. In this framework, the specification of the latent variable distributions is crucial, since a different first level variance among the schools would substantially change the interpretation of model parameters, as confirmed by the limited simulation study presented in the paper.

Key words: ordinal response models; variance component models; varying thresholds; course programme evaluation

Data and software link available from: http://stat.uibk.ac.at/SMIJ
Received June 2001; **revised** October 2001, August 2002; **accepted** September 2002

1 Introduction

The interest in multilevel techniques for ordinal response variables has increased considerably in the last few years (Fielding, 1999). Such techniques can be useful, for example, in course programme evaluations based on student ratings (Emerson *et al.*, 2000). In this framework, it is usual to suppose that the observed ordinal measurement comes from a latent continuous variable. In order to make that model identifiable, it is necessary to make assumptions on the latent variable distribution, for example, normal with unit first level variance. However, when the units are grouped into strata, and the model parameters are allowed to vary with such strata, a new identification problem arises, requiring an additional assumption. The correct specification of the latent variable distribution is especially important in the case of comparisons among the strata: in particular, a different first level variance across the strata would change substantially the interpretation of the estimated model parameters.

Address for correspondence: L Grilli, Department of Statistics 'G. Parenti', University of Florence Viale Morgagni, 59, 50134 Florence, Italy. E-mail: grilli@ds.unifi.it

© Arnold 2002
10.1191/1471082x02st041oa

As an example, we consider the student course evaluation ratings collected at regular intervals by the University of Florence. The data have a hierarchical structure with ratings nested in courses, which are nested in schools; however, the schools are too few to constitute a model level, so they should be considered as strata. When building a two-level ordinal model for the student ratings one should allow the parameters to be different across the schools, including the first level standard deviations, which act in the ordinal model as scaling factors for the other parameters. Failure to adjust for different scaling factors invalidates the comparisons among schools.

The paper is organized as follows. The second section presents the standard variance component ordinal response model, illustrating the classical identification problem arising from the latent variable formulation. The third section introduces a generalization of this model that allows for comparison among strata, presenting three alternative assumptions that overcome the identifiability problem, and proposing a general model selection strategy. In Section 4 the proposed methodology is applied to the student evaluation data; in the same section we present the results of a simulation study designed to show the effects of incorrectly assuming equal first level variances and to evaluate the power of the test for such equality. Section 5 concludes with some remarks.

2 The standard model

Suppose that an observed ordinal response variable Y, with $k = 1, 2, \ldots, K$ levels, comes, through a set of thresholds, from a latent continuous variable $\tilde{Y}$ following a variance component model (Hedeker and Gibbons, 1994):

$$\tilde{Y}_{ij} = \alpha + \beta x_{ij} + \tau u_j + \varepsilon_{ij}, \tag{2.1}$$

with $i = 1, 2, \ldots, n_j$ elementary units for the jth cluster the ($j = 1, 2, \ldots, J$). In equation (2.1) α is the intercept, x_{ij} is a covariate and β the corresponding slope, the random variables ε_{ij} and u_j are the disturbances at the first (elementary units) and second (cluster) levels respectively, and τ^2 is the second level variance.

For the disturbances of model (2.1) the usual assumptions are:

(i) The ε_{ij}'s are iid with zero mean and $\mathrm{Var}(\varepsilon_{ij}) = \sigma^2$ (first level variance);
(ii) $u_j \overset{\text{iid}}{\sim} N(0, 1)$;
(iii) The ε_{ij}'s and u_j's are mutually independent.

The observed ordinal variable Y is linked to the latent one $\tilde{Y}$ through the following relationship:

$$\{Y_{ij} = k\} \Leftrightarrow \{\gamma_{k-1} < \tilde{Y}_{ij} \leq \gamma_k\},$$

where the thresholds satisfy $-\infty = \gamma_0 \leq \gamma_1 \leq \cdots \leq \gamma_{K-1} \leq \gamma_K = +\infty$. For identifiability reasons, γ_1 is fixed to zero, so there are $K - 2$ estimable thresholds.

Conditional on u_j, the model probability for unit i of cluster j is:

$$P(Y_{ij} = k \mid u_j) = P(\gamma_{k-1} < \tilde{Y}_{ij} \leq \gamma_k \mid u_j)$$
$$= P(\tilde{Y}_{ij} \leq \gamma_k \mid u_j) - P(\tilde{Y}_{ij} \leq \gamma_{k-1} \mid u_j),$$

with

$$\begin{aligned} P(\tilde{Y}_{ij} \leq \gamma_k \mid u_j) &= P(\varepsilon_{ij} \leq \gamma_k - [\alpha + \beta x_{ij} + \tau u_j] \mid u_j) \\ &= P\left[\frac{\varepsilon_{ij}}{(\sigma/c)} \leq \frac{\gamma_k}{(\sigma/c)} - \left(\frac{\alpha}{(\sigma/c)} + \frac{\beta}{(\sigma/c)} x_{ij} + \frac{\tau}{(\sigma/c)} u_j\right) \mid u_j\right] \\ &= F\left[\frac{\gamma_k}{(\sigma/c)} - \left(\frac{\alpha}{(\sigma/c)} + \frac{\beta}{(\sigma/c)} x_{ij} + \frac{\tau}{(\sigma/c)} u_j\right)\right] \\ &= F[\gamma_{\sigma,k} - (\alpha_\sigma + \beta_\sigma x_{ij} + \tau_\sigma u_j)], \end{aligned} \tag{2.2}$$

where $F(\cdot)$ is the distribution function of the 'standardized' first level error term $\varepsilon_{ij}(c/\sigma)$, which has variance c^2. Usually the value of the constant c is chosen in such a way that $F(\cdot)$ has a convenient standard form: typical choices are $c = 1$ for the normal distribution, $c = \sqrt{\pi^2/3}$ for the logistic distribution and $c = \sqrt{\pi^2/6}$ for the complementary log–log distribution.

Note that all the model parameters are defined in terms of σ/c, the standard deviation of the 'standardized' first level error term, which depends on the unknown σ (this fact is denoted by the presence of the symbol σ in the subscript of the parameters). Thus, only the ratios of the model parameters to the standard deviation of the first level error term are identifiable; one popular way to overcome this identifiability problem is to fix σ, usually to 1.

The overall marginal likelihood for the defined model is:

$$L(\psi) = \prod_{j=1}^{J} \int_{-\infty}^{\infty} \prod_{i=1}^{n_j} \left[\prod_{k=1}^{K} P(Y_{ij} = k|u)^{d_{ijk}}\right] \phi(u) \mathrm{d}u,$$

where $\psi = \{\alpha, \beta, \tau, \gamma_2, \gamma_3, \ldots, \gamma_{K-1}\}$, d_{ijk} is the indicator variable of the event $\{Y_{ij} = k\}$ and ϕ is the standard normal density.

3 Model specification in the presence of strata

Suppose now that the elementary units can be grouped into a limited number of strata $h = 1, \ldots, H$, across which the model parameters are free to vary. The clusters can be cross-classified with the strata or nested within them. Note that in the last case the strata define a new third level; however, if the number of strata is small (say, below ten), the inclusion of the corresponding random component is not advisable and this new third level is most appropriately modelled by stratum-dependent parameters.

Denoting with the superscript (h) the stratum to which the quantities are referred, model (2.1) for stratum h can be written in the following way:

$$\tilde{Y}_{ij}^{(h)} = \alpha^{(h)} + \beta^{(h)} x_{ij} + \tau^{(h)} u_j + \varepsilon_{ij}^{(h)}, \tag{3.1}$$

with the following assumptions:

(i) The $\varepsilon_{ij}^{(h)}$'s are independent with zero mean and stratum-dependent variance $\mathrm{Var}(\varepsilon_{ij}^{(h)}) = \sigma^2(1+\theta^{(h)})^2$;
(ii) $u_j \overset{\text{iid}}{\sim} N(0,1)$;
(iii) The $\varepsilon_{ij}^{(h)}$'s and u_j's are mutually independent.

Therefore the parameters of model (3.1) are

$$\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(H)}, \beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(H)}, \tau^{(1)}, \tau^{(2)}, \ldots, \tau^{(H)}, \sigma, \theta^{(2)}, \ldots, \theta^{(H)}.$$

Note that the $\theta^{(h)}$'s are intended to measure the difference in the first level variance from the reference stratum $h = 1$, for which $\theta^{(1)} \equiv 0$.

Consequently, the model probabilities corresponding to (2.2) are stratum-dependent:

$$\begin{aligned} P\left(\tilde{Y}_{ij}^{(h)} \le \gamma_k^{(h)} | u_j\right) &= P\left(\frac{\varepsilon_{ij}^{(h)}}{(\sigma/c)(1+\theta^{(h)})} \le \frac{\gamma_k^{(h)}}{(\sigma/c)(1+\theta^{(h)})} \right. \\ &\qquad \left. - \left[\frac{\alpha^{(h)}}{(\sigma/c)(1+\theta^{(h)})} + \frac{\beta^{(h)}}{(\sigma/c)(1+\theta^{(h)})} x_{ij} + \frac{\tau^{(h)}}{(\sigma/c)(1+\theta^{(h)})} u_j \right] \middle| u_j \right) \\ &= F\left(\frac{\gamma_{\sigma,k}^{(h)}}{(1+\theta^{(h)})} - \left[\frac{\alpha_\sigma^{(h)}}{(1+\theta^{(h)})} + \frac{\beta_\sigma^{(h)}}{(1+\theta^{(h)})} x_{ij} + \frac{\tau_\sigma^{(h)}}{(1+\theta^{(h)})} u_j \right] \right) \\ &= F\left(\gamma_{\sigma,k}^{(h)*} - [\alpha_\sigma^{(h)*} + \beta_\sigma^{(h)*} x_{ij} + \tau_\sigma^{(h)*} u_j] \right), \end{aligned} \tag{3.2}$$

where the superscript $(h)^*$ indicates that the parameter is divided by $(1+\theta^{(h)})$. In the following, the expression 'h-parameter' will denote an original parameter divided by (σ/c), for example, $\alpha_\sigma^{(h)} = \alpha^{(h)}/(\sigma/c)$, while the expression '$h^*$-parameter' will refer to an h-parameter divided by $(1+\theta^{(h)})$. Note that the number of estimable thresholds is now $H \times (K-2)$.

The conditional likelihood for the jth group is now:

$$L_j(\psi \mid u_j) = \prod_{i=1}^{n_j} \prod_{h=1}^{H} \prod_{k=1}^{K} P\left(Y_{ij}^{(h)} = k \mid u_j\right)^{d_{ijhk}}, \tag{3.3}$$

where ψ is the set of estimable parameters, while d_{ijhk} is an indicator variable that assumes the value one if the ith individual of the jth cluster belongs to the hth stratum and has an observed value $Y_{ij}^{(h)} = k$.

The overall marginal likelihood is:

$$L(\psi) = \prod_{j=1}^{J} \int_{-\infty}^{\infty} L_j(\psi|u)\phi(u)\mathrm{d}u.$$

The h^*-parameters can be easily estimated by allowing the intercept, the second level variance and the thresholds to vary across the H strata. However some additional assumptions must be made in order to estimate the h-parameters, which are the parameters of interest. For example, if a certain stratum has a different h^*-intercept, it can be that the true h-intercept is different, but it can also be that its level one variance is different, or a mixture of the two cases.

Three possible assumptions that overcome this identifiability problem are the following:

1. $\theta^{(2)} = \theta^{(3)} = \cdots = \theta^{(H)} = 0$ (common first level variance): in this case the h-parameters are the same as the h^*-parameters.
2. $\tau_\sigma^{(1)} = \tau_\sigma^{(2)} = \cdots = \tau_\sigma^{(H)}$ (common second level variance): in this case the parameters $\theta^{(h)}$ can be estimated from the following identity:

$$\frac{\tau_\sigma^{(1)*}}{\tau_\sigma^{(h)*}} = \frac{\tau_\sigma^{(1)}}{[\tau_\sigma^{(h)}/(1+\theta^{(h)})]} = \frac{\tau_\sigma^{(1)}}{\tau_\sigma^{(h)}}(1+\theta^{(h)}) = 1+\theta^{(h)}.$$

   The original intercepts and thresholds are then easily calculated by multiplying the h^*-parameters by $(1+\theta^{(h)})$. For example,

$$\alpha_\sigma^{(h)} = \alpha_\sigma^{(h)*}(1+\theta^{(h)}). \tag{3.4}$$

3. $\beta_\sigma^{(1)} = \beta_\sigma^{(2)} = \cdots = \beta_\sigma^{(H)}$ (common regression coefficient): we can proceed as in case 2, using the following identity:

$$\frac{\beta_\sigma^{(1)*}}{\beta_\sigma^{(h)*}} = \frac{\beta_\sigma^{(1)}}{[\beta_\sigma^{(h)}/(1+\theta^{(h)})]} = \frac{\beta_\sigma^{(1)}}{\beta_\sigma^{(h)}}(1+\theta^{(h)}) = 1+\theta^{(h)}. \tag{3.5}$$

Often the researcher acts as in case 1, tacitly assuming identical first level variances. It should be stressed that such an assumption, which is crucial for the interpretation of the results, is not testable and its validity in the data at hand is difficult to assess. However, there are other ways to proceed. The second choice (identical second level variances) is similar to the first one, since it simply shifts the assumption from the first to the second level variance. But the third choice (identical regression slopes) is somewhat different, since such an assumption concerns not a variance parameter, but an association parameter and is consequently easier to justify. In fact, it is more common to have some *a priori* knowledge of the regression coefficients than of the variances. Moreover,

the validity of such an assumption in the data at hand can be investigated through some technique which can help to explore the association among the latent variable and the covariate of interest. For example, one might assign a set of scores to the levels of the ordinal variable, fit a separate linear regression model for each of the H strata and then compare the slopes; when the covariate is also ordinal, a closely related technique is to compare the H Spearman correlations between the response variable and the covariate. This strategy has a theoretical justification in the well-known fact that association parameter estimators are usually more robust to model misspecifications than variance parameter estimators. In particular, Fielding (1999) found that, for the fixed regression coefficients, the ordinal variance component model leads essentially to the same conclusions as the linear variance component model on the scores of the ordinal variable (using various scoring systems); on the other hand, the conclusions for the variance components are significantly different.

A possible general strategy is the following:

1. Choose the covariate for which the assumption $\beta_\sigma^{(1)} = \cdots = \beta_\sigma^{(H)}$ is more reasonable (using *a priori* information, separate linear regression models, Spearman correlations or the like).
2. Fit the model including the covariate chosen in step 1, allowing all the parameters (the h^*-parameters in our notation) to vary among the H strata.
3. Assuming $\beta_\sigma^{(1)} = \cdots = \beta_\sigma^{(H)}$, use the estimated $\beta_\sigma^{(h)*}$'s to obtain an estimate of the $\theta^{(h)}$'s.
4. Test the hypothesis that the $\theta^{(h)}$'s ($h = 2, 3, \ldots, H$) are jointly null by carrying out a test for the equality of the $\beta_\sigma^{(h)*}$'s; if such a hypothesis is rejected, perform a sequence of tests to identify the subset of $\theta^{(h)*}$'s that are significantly different from zero.
5. If the hypothesis that all the $\theta^{(h)}$'s are null is not rejected, then go on with model selection in the usual manner (the interpretation of the results is straightforward, since in this case the h-parameters equal the h^*-parameters).
6. Otherwise, for the strata whose corresponding $\theta^{(h)}$ is significantly different from zero, it is necessary to correct the h^*-estimates with an estimate of the factor $(1 + \theta^{(h)})$, as in (3.4). In this case the model selection should be modified to take into account the restrictions on the parameters. For example, if $\theta^{(2)}$ is different from zero, testing the hypothesis $\alpha_\sigma^{(1)} = \alpha_\sigma^{(2)}$ amounts to testing the hypothesis

$$\frac{\alpha_\sigma^{(1)*}}{\alpha_\sigma^{(2)*}} = \frac{\beta_\sigma^{(1)*}}{\beta_\sigma^{(2)*}} \tag{3.6}$$

One way of testing (3.6) is to carry out a Wald test with the aid of the delta method. However, this technique may be inadequate in such a complex testing problem (Godfrey, 1991), so it is advisable to fit a restricted model that satisfies the nonlinear constraint (3.6) and carry out a deviance test (the required restricted estimates can be easily obtained with the estimation procedure used for the application described in Section 4.1).

The strategy just outlined is to be recommended when the assumption of a common regression slope is more justifiable than the corresponding assumptions on

the variances. Obviously, if no suitable regressor is found, one is forced to impose some restrictions on the first level variances or, alternatively, on the second level variances.

A crucial step in the proposed strategy is the test on the equality of the first level variances, that is, $\theta^{(h)} = 0$ for $h = 2, 3, \ldots, H$. In fact, incorrectly failing to reject this hypothesis causes an erroneous interpretation of all model parameters, as is clear from (3.2) and from the simulation results presented in Section 4.2. In step 4 above we suggested performing this test in the context of the most general model, imposing no further restrictions beyond those required for the $\beta^{(h)}$'s. In this way one is protected against potentially serious biases due to incorrect restrictions. The limited simulation experiment of Section 4.2 indicates that the deviance test comparing the general model with the restricted model ($\beta_\sigma^{(h)*}$'s constant among the strata) has enough power; further investigation is needed to study the power of the test under different levels of complexity of the model and for more combinations of first level variance values.

4 Application: course evaluation

The method outlined in Section 3 is applied to the data gathered in a survey on course evaluations carried out by the University of Florence, in all the schools of the university, for classes in the second semester of the 1999–2000 academic year. Specifically, we will refer to the results from the schools of Engineering, Science and Letters. The data have a hierarchical structure: ratings are nested in courses that are nested in schools. The total number of clusters represented by the courses is 370, while the number of strata represented by the schools is three. Table 1 reports, for each school, the number of respondents, the number of courses evaluated and the minimum, median and maximum numbers of respondents per course.

In the present application we focus on the item relative to overall satisfaction, which required a response on a four-level ordinal scale: (1) decidedly no; (2) more no than yes; (3) more yes than no; (4) decidedly yes. The aim of the analysis is to establish if the different evaluations expressed by the students in the three schools might, to some extent, be attributed to a different 'measurement scale', that is, to a different way of

Table 1 Number of respondents, number of courses evaluated, and minimum, median, and maximum number of respondents per course for the University of Florence, academic year 1999–2000, second semester

			Respondents per course		
School	No. of respondents	No. of courses	Min	Median	Max
Engineering	3165	150	4	16	71
Science	1633	103	4	13	52
Letters	1932	117	3	10	118
Total	6730	370	3	13	118

interpreting the levels of the ordinal scale (obviously, the evaluations expressed by each student are influenced also by their characteristics and expectations).

4.1 Empirical results

The analysis begins by fitting the most general model presented in Section 3, that is, the model allowing the intercept, the second and first level variances and the thresholds to vary among the schools, with the first threshold for all the schools fixed to 0 and Engineering as the reference stratum. In the present application we always assume that the first level disturbances have a Gaussian distribution, leading to a probit model specification.

In order to recover the h-parameters from the estimable h^*-parameters, we introduce into the model a covariate with an assumed common slope. In this case the covariate is the answer (on a four-level ordinal scale) to the question whether the student will take the exam at the first examination session (covariate *exam*). The choice of this covariate is motivated by our knowledge of the phenomenon, since we have no reason to suppose a differential effect of the covariate on the latent evaluation among the schools. Moreover, this covariate shows very similar values of the Spearman correlation coefficient with the overall satisfaction among the three schools (0.32, 0.34, and 0.36, respectively, for Engineering, Science, and Letters).

The estimation was carried out with the NLMIXED procedure of the SAS software (SAS Institute, 1999). In order to fit the model described in Section 3, we exploited the capability of the procedure that allows us to write down the conditional likelihood (3.3). PROC NLMIXED fits nonlinear multilevel models by maximizing a numerical approximation to the marginal likelihood. Different integral approximations are available: for the present paper we used a 10-point Gaussian quadrature (Hedeker and Gibbons, 1994). A variety of alternative optimization techniques are available to carry out the maximization; the default, used in this paper, is a dual quasi-Newton algorithm, where dual means that at each iteration the upgrading concerns the Cholesky factor of an approximate Hessian instead of an approximation of the inverse Hessian (SAS Institute, 1999).

The fitted models are described in Table 2 in terms of assumptions on the h-parameters. Note that Model 1 does not imply any restriction on the estimable h^*-parameters, while in Models 2 to 7 the assumption of equal first level variances ($\theta^{(S)} = \theta^{(L)} = 0$) implies the equivalence between h- and h^*-parameters. Table 3 presents the results relative to the fitted models in terms of the estimable h^*-parameters.

Comparing Models 1 and 2 in terms of deviance, the hypothesis of equal first level variances among the schools is not to be rejected: the deviance difference is 2 (14651 − 14649), not significant in a chi-squared distribution with d.f. = 2. This fact can be appreciated also by looking at the estimates of the $\theta^{(h)}$'s obtained through formula (3.5) applied to Model 1, which are −0.0801 (s.e. = 0.08441) for Science and −0.1035 (s.e. = 0.07382) for Letters.

Model selection can now proceed in the usual way. First of all, from Model 2 it seems that Engineering and Science have the same thresholds and second level variance. This impression is confirmed by the very small increase of the deviance when fitting Model 3 ($\chi^2 = 14653 - 14651 = 2$, d.f. = 3).

Table 2 Assumptions on *h*-parameters for models of Table 3

Model	Assumptions
1	$\beta_\sigma^{(E)} = \beta_\sigma^{(S)} = \beta_\sigma^{(L)}$
2	$\beta_\sigma^{(E)} = \beta_\sigma^{(S)} = \beta_\sigma^{(L)}; \theta^{(S)} = \theta^{(L)} = 0$
3	$\beta_\sigma^{(E)} = \beta_\sigma^{(S)} = \beta_\sigma^{(L)}; \theta^{(S)} = \theta^{(L)} = 0; \tau_\sigma^{(E)} = \tau_\sigma^{(S)}; \gamma_{\sigma,k}^{(E)} = \gamma_{\sigma,k}^{(S)}$ $(k = 2, 3)$
4	$\beta_\sigma^{(E)} = \beta_\sigma^{(S)} = \beta_\sigma^{(L)}; \theta^{(S)} = \theta^{(L)} = 0; \tau_\sigma^{(E)} = \tau_\sigma^{(S)} = \tau_\sigma^{(L)}; \gamma_{\sigma,k}^{(E)} = \gamma_{\sigma,k}^{(S)}$ $(k = 2, 3)$
5	$\beta_\sigma^{(E)} = \beta_\sigma^{(S)} = \beta_\sigma^{(L)}; \theta^{(S)} = \theta^{(L)} = 0; \gamma_{\sigma,k}^{(E)} = \gamma_{\sigma,k}^{(S)} = \gamma_{\sigma,k}^{(L)}$ $(k = 2, 3)$
6	$\theta^{(S)} = \sigma^{(L)} = 0$
7	$\theta^{(S)} = \theta^{(L)} = 0; \tau_\sigma^{(E)} = \tau_\sigma^{(S)}; \gamma_{\sigma,k}^{(E)} = \gamma_{\sigma,k}^{(S)}$ $(k = 2, 3)$

The assumptions refer to the general model described in section 3, with $K = 4$ and $H = 3$.
The superscript denotes the school: E = Engineering (baseline stratum), S = Science, L = Letters.

Further simplifications of the model are not supported by the data: for example, Table 3 reports the estimates for the model with constant second level variance and varying thresholds for Letters (Model 4) and for the model with varying second level variance but fixed thresholds (Model 5).

Note that imposing fixed thresholds (Model 5) causes an important loss of fit with respect to Model 2 ($\chi^2 = 14717 - 14651 = 66$, d.f. $= 4$), so the data strongly support the hypothesis of a different 'measurement scale' for the students of Letters. The consequences of this fact can be appreciated by comparing Models 3 and 5: in Model 5 the higher ratings obtained by the courses in the school of Letters as compared to those in Engineering are totally attributed to higher latent evaluations ($\hat{\alpha}_\sigma^{(L)*} - \hat{\alpha}_\sigma^{(E)*} = 0.6143$ with s.e. 0.0899), while in Model 3 are attributed partly to higher latent evaluations ($\hat{\alpha}_\sigma^{(L)*} - \hat{\alpha}_\sigma^{(E)*} = 0.2444$ with s.e. 0.1001) and partly to a more favorable 'measurement scale' (the lower values of the thresholds for Letters imply that, for the same latent evaluation, the expressed rating on the ordinal scale is greater or equal).

Another interesting feature of Model 3 is that, although the first level variance is constant among the schools, the second level variance is not, with Letters having a significantly lower value: 0.5006 versus 0.7208. The intraclass correlation coefficient, $\rho = \tau^2/(1 + \tau^2)$, is 0.2004 for Letters and 0.3464 for the other schools: this means that in the school of Letters the proportion of variance attributable to the courses is substantially lower and, consequently, the student ratings have a lower discriminant power.

Note that the conclusions just outlined rely heavily on the hypothesis of constant first level variance among the schools, highlighting the practical importance of devising a procedure to assess the validity of such a hypothesis.

Since, in the present application, the inclusion of the covariate *exam* is instrumental and not of direct interest, after testing for equal first level variances we also performed the model selection without the covariate. Table 3 reports the fitting of Models 6 and 7, which are the no-covariate counterparts of Models 2 and 3, respectively. As expected, the omission of the covariate *exam* leads to the same substantive conclusions, since its effect is approximately constant among the strata. Figure 1 represents, for each school,

Table 3 Results of model selection

	Model						
	With covariate					Without cov.	
	1	2	3	4	5	6	7
Intercept							
$\alpha_\sigma^{(E)*}$	0.3160	0.2539	0.2549	0.2479	0.1636	1.4704	1.4745
$\alpha_\sigma^{(S)*} - \alpha_\sigma^{(E)*}$	*0.1305*	0.2301	0.2274	0.2244	0.2141	0.2578	0.2425
$\alpha_\sigma^{(L)*} - \alpha_\sigma^{(E)*}$	*0.1149*	0.2450	0.2444	0.2909	0.6143	0.2410	0.2370
Exam							
$\beta_\sigma^{(E)*}$	0.4206	0.4436	0.4434	0.4428	0.4459		
$\beta_\sigma^{(S)*}$	0.4572	"	"	"	"		
$\beta_\sigma^{(L)*}$	0.4692	"	"	"	"		
Rand. par.							
$\tau_\sigma^{(E)*}$	0.7505	0.7525	0.7208	0.6692	0.7096	0.7620	0.7596
$\tau_\sigma^{(S)*}$	0.6902	0.6907	"	"	0.6587	0.7554	"
$\tau_\sigma^{(L)*}$	0.5000	0.5008	0.5006	"	0.5788	0.5600	0.5600
Thresholds							
$\gamma_{\sigma,2}^{(E)*}$	0.9984	1.0022	1.0131	1.0081	0.9603	0.9360	0.9488
$\gamma_{\sigma,2}^{(S)*}$	1.0408	1.0388	"	"	"	0.9787	"
$\gamma_{\sigma,2}^{(L)*}$	0.8032	0.7985	0.7984	0.8070	"	0.7226	0.7226
$\gamma_{\sigma,3}^{(E)*}$	2.5334	2.5438	2.5368	2.5236	2.3856	2.3751	2.3736
$\gamma_{\sigma,3}^{(S)*}$	2.5333	2.5278	"	"	"	2.3756	"
$\gamma_{\sigma,3}^{(L)*}$	2.0132	2.0022	2.0021	2.0279	"	1.8299	1.8299
N. par.	15	13	10	9	9	12	9
$-2 \log L$	14649	14651	14653	14663	14717	15445	15446

The assumptions on h-parameters for each model are defined in Table 2.
The estimates in italics are not significant at the 5% level; the symbol " indicates that the value is constrained to be equal to the value in the above cell; the superscript denotes the school: E = Engineering (baseline stratum), S = Science, L = Letters.

the marginal distribution and the mean of the latent variable and the corresponding thresholds estimated from Model 7 of Table 3. It is worth noting that, with respect to the distribution of Engineering, the distribution of Science is simply shifted to the right, while the distribution of Letters is shifted to the right and has a lower variance. Moreover, the thresholds of Letters are shifted to the left, with the last threshold being almost equal to the mean, so that the area under the density function on the right of the last threshold (i.e., the model estimated proportion of very satisfied students) is about 0.5.

4.2 Simulation results

In order to illustrate the potential bias in the estimators due to ignoring different first level variances, we conducted a brief Monte Carlo simulation experiment, generating

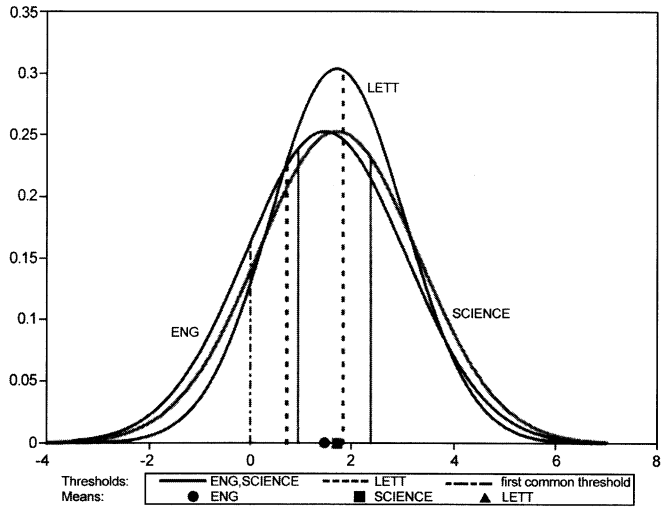


Figure 1 Estimated latent variable distributions, thresholds and means for the schools of Engineering (ENG), Letters (LETT) and SCIENCE from Model 7 of Table 3

Edward Arnold (Publishers) Limited

data with the same hierarchical structure (first and second level number of observations) and the same covariate values as the observed data; the parameter values of the intercepts, covariate, second level variances and thresholds were fixed approximately at the estimated values from Model 2 of Table 3. We simulated a data set for each of four different combinations of first level variance values. The same simulation experiment allows for the evaluation of the power of the deviance test of fixed 5% size for the hypothesis of equal first level variances, mentioned in point 4 of the general strategy presented in Section 3.

The simulation results are reported in Tables 4 and 5. Note that the values in Table 4 refers to the h-parameters, not to the h^*-parameters. The distinction is irrelevant for Model 2, since it assumes $\theta^{(S)} = \theta^{(L)} = 0$, but is crucial for Model 1, for which h^*-estimates have been converted into h-estimates, using expressions (3.4) and (3.5).

Table 4 shows the consequences of a model misspecification in two opposite situations. When $\theta^{(S)} = \theta^{(L)} = 0$, Model 1 is the wrong model because it fails to impose the restriction on the first level variances: in this case, however, the misspecified model leads to approximately unbiased estimators and the only loss is in terms of efficiency.

On the contrary, when the $\theta^{(h)}$'s are different the wrong model is Model 2, which imposes the incorrect restriction $\theta^{(S)} = \theta^{(L)} = 0$: in this case the estimates of the intercepts, which are used to compare the schools, are completely misleading, while the estimates of the other parameters are considerably attenuated, except for Engineering.

The test for equal first level variances is therefore a crucial step in model selection. Table 5 shows, for four combinations of first level variance values, the Monte Carlo means of the estimated $\theta^{(h)}$'s, the deviance test statistic G and the corresponding power.

Table 4 True values and Monte Carlo means of estimated *h*-parameters for Model 1 (M1) and Model 2 (M2) of Table 2 (Monte Carlo standard errors in brackets) based on 500 replications

		$\theta^{(S)}=\theta^{(L)}=0$		$\theta^{(S)}=0.25,\ \theta^{(L)}=0.50$	
Parameter	True	M1	M2	M1	M2
Intercept					
$\alpha^{(E)}$	0.25	0.2602	0.2650	0.2509	0.4113
		(0.1282)	(0.1209)	(0.1299)	(0.1221)
$\alpha^{(S)}-\alpha^{(E)}$	0.25	0.2525	0.2388	0.2081	−0.0609
		(0.1958)	(0.1586)	(0.2353)	(0.1350)
$\alpha^{(L)}-\alpha^{(E)}$	0.35	0.3499	0.3362	0.2446	−0.2422
		(0.1948)	(0.1376)	(0.2690)	(0.1279)
Exam					
$\beta^{(E)}=\beta^{(S)}=\beta^{(L)}$	0.45	0.4493	0.4495	0.4492	0.3805
		(0.0233)	(0.0160)	(0.0232)	(0.0158)
Random parameter					
$\tau^{(E)}$	0.75	0.6905	0.6887	0.6901	0.6864
		(0.0558)	(0.0556)	(0.0558)	(0.0556)
$\tau^{(S)}$	0.69	0.6786	0.6771	0.6846	0.5477
		(0.0931)	(0.0697)	(0.1005)	(0.0542)
$\tau^{(L)}$	0.50	0.4927	0.4892	0.4930	0.3360
		(0.0663)	(0.0538)	(0.0804)	(0.0428)
Thresholds					
$\gamma_2^{(E)}$	1.00	0.9975	0.9974	0.9976	0.9844
		(0.0349)	(0.0346)	(0.0350)	(0.0341)
$\gamma_3^{(E)}-\gamma_2^{(E)}$	1.55	1.5456	1.5455	1.5456	1.5251
		(0.0340)	(0.0335)	(0.0340)	(0.0327)
$\gamma_2^{(S)}$	1.00	0.9994	0.9964	1.0010	0.7998
		(0.1064)	(0.0551)	(0.1224)	(0.0449)
$\gamma_3^{(S)}-\gamma_2^{(S)}$	1.53	1.5362	1.5315	1.5401	1.2308
		(0.1481)	(0.0507)	(0.1693)	(0.0395)
$\gamma_2^{(L)}$	0.80	0.8057	0.8023	0.8060	0.5384
		(0.0873)	(0.0519)	(0.0968)	(0.0318)
$\gamma_3^{(L)}-\gamma_2^{(L)}$	1.20	1.2025	1.1977	1.2083	0.8068
		(0.1079)	(0.0429)	(0.1397)	(0.0328)

Table 5 True values and Monte Carlo means of estimated $\theta^{(h)}$'s, deviance test statistic G and power of the test (Monte Carlo standard errors in brackets)

True		Estimates		Deviance test H_0: $\theta^{(S)}=\theta^{(L)}=0$	
$\theta^{(S)}$	$\theta^{(L)}$	$\theta^{(S)}$	$\theta^{(L)}$	G	$Pr\{G>\chi^2_{\alpha=0.05}\|H_0\}$
0	0	0.0029	0.0041	2.0898	0.064
		(0.0952)	(0.0875)		
0	0.25	0.0026	0.2592	8.8798	0.636
		(0.0952)	(0.1242)		
0	0.50	0.0025	0.5155	21.8070	0.982
		(0.0950)	(0.1671)		
0.25	0.50	0.2554	0.5153	19.8334	0.970
		(0.1386)	(0.1671)		

$\chi^2_{\alpha=0.05}=5.9915$, df = 2. Replications = 500.

The actual size of the test is close to the nominal one, while the behavior of the power seems to be satisfactory. Note that, given the different consequences of Type I and Type II errors, it may be worthwhile to increase the size of the test in order to obtain a higher power.

5 Concluding remarks

The paper has discussed the issues that arise in the specification of ordinal variance component models in the presence of strata, with an application to the analysis of student evaluations.

We suggest an approach to overcoming the identification problem due to the stratum-dependent first level variance by introducing a covariate with a common slope among the strata. In a specific application, this choice can be justified on the grounds of prior knowledge and on the basis of techniques that attempt to describe the behavior of the latent variable. The reliability of these techniques in various practical situations needs to be assessed through a large simulation study.

It should be noted that the same identification problems affect the standard one level ordinal model; the role of the second level variance is to lead to a more realistic model for the phenomenon under study, also allowing a broader and more interesting discussion.

Since in our application the use of the covariate *exam* was merely instrumental, Table 3 also presents the results from the model with no covariate. Indeed, the exclusion of the covariate used to test for equal first level variances is not expected to modify the conclusions of such a hypothesis; on the contrary, the inclusion of further covariates is potentially harmful, especially if the new covariates have varying effects among the strata. To avoid such problems the hypothesis of equal first level variances should be tested in the more general model (i.e., the model including all the relevant covariates), imposing the appropriate restrictions on the slopes of the covariates which are assumed to have a constant effect among the strata.

As for the distributional form of the disturbances, the trials that we made suggest that this issue is not crucial in the present application. However, it would be very useful to carry out a sensitivity analysis and to develop a formal procedure for the selection of the distributional form; such a selection is particularly relevant for the first level disturbances, whose distribution determines the link function of the model.

Acknowledgements

We wish to thank the Editor and two anonymous referees for their suggestions, which contributed to substantially improving the paper.

Financial support from MURST research projects 'Evaluation of quality and effectiveness in human services, with particular attention to health and education' and 'Statistical education and evaluation: instruments, methods and new technologies' is gratefully acknowledged.

References

Emerson JD, Mosteller F, Youtz C (2000) Students can help improve college teaching: a review and an agenda for the statistics profession. In Rao CR, Szèekely GJ eds. *Statistics for the 21st Century: Methodologies for Applications of the Future.* New York: Marcel Dekker.

Fielding A (1999) Why use arbitrary point scores?: ordered categories in models of educational progress. *JRSS A*, **162**, 303–28.

Godfrey LG (1991) *Misspecification Test in Econometrics: The Lagrange Multiplier Principle and Other Approaches.* New York: Cambridge University Press.

Hedeker D, Gibbons RD (1994) A random-effects ordinal regression model for multilevel analysis. *Biometrics*, **50**, 933–44.

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

Alternative specifications of multivariate multilevel probit ordinal response models

Da: Grilli L., Rampichini C. (2003) Alternative specifications of multivariate multilevel probit ordinal response models, Journal of Educational and Behavioral Statistics, Vol. 28, pp. 31-44.

Alternative specifications of multivariate multilevel probit ordinal response models

Leonardo Grilli
Carla Rampichini
University of Florence

Multivariate multilevel models for ordinal variables are quite complex with respect to both interpretation and estimation. The specification in terms of a multivariate latent distribution and a set of thresholds helps in the interpretation of the variance-covariance parameters. However most existing estimation algorithms for multilevel models can be used only if the model is reparametrized as a univariate model with an additional dummy bottom level. Moreover the univariate formulation allows the model to be cast in the framework of Generalised Linear Latent and Mixed Models (Rabe-Hesketh, Pickles & Skrondal, 2001a), a rather general class which includes, as special cases, structural equations and factor models. The paper outlines the multivariate latent distribution specification and the corresponding interpretation issues; then it shows the univariate formulation, along with some alternative parametrizations which are useful in the estimation phase. An application to student ratings data illustrates the interpretation of the parameters and the estimation procedures, with a discussion of some computational issues.

Keywords: *multilevel models, ordinal variables, latent variables, multivariate models, numerical integration*

In the last years several methods for the analysis of ordinal multivariate multilevel data have been proposed (Rabe-Hesketh et al., 2001a, Rabe-Hesketh, Skrondal & Pickles, 2002; Mazzolli, 2001; Lillard & Panis, 2000). One popular way to define an ordinal response model relies on the concepts of latent variables and thresholds. When the model is at the same time

multivariate and multilevel, the formulation based on a multivariate latent distribution (discussed in detail in the following) is particularly appealing, since it helps one to understand the complex correlation structure implied by the model.

However this formulation is not suitable for standard algorithms used with multilevel models. In fact, the existence of such algorithms has motivated researchers to look for an alternative specification in terms of univariate models (Rabe-Hesketh et al., 2001a). In the paper, considering the two-level case for the sake of simplicity, we show how a two-level multivariate model for ordinal responses can be reparametrized as a three-level univariate model with a dummy bottom level. In the paper attention will be restricted to estimation methods based on numerical integration, which produce accurate results even in very complex models, though they may require long computational times. We also discuss some alternative parametrizations of the univariate specification which can increase the computational efficiency.

An application to the evaluation of university courses by means of student ratings illustrates the modelling and computing issues, making some broad comparison among existing software.

Basic specification of the model

Let $Y_{ij}^{(h)}$ be the h-th observed ordinal variable ($h = 1, 2, \cdots, H$) for the i-th subject ($i = 1, 2, \cdots n_j$) of the j-th cluster ($j = 1, 2, \cdots J$). In the application presented in the paper the *clusters* are the courses, the *subjects* are the questionnaires and the *ordinal variables* are the ratings on two items of the questionnaire (i.e., $H = 2$).

Now assume that each of the observed responses $Y_{ij}^{(h)}$, which takes values in $\{1, 2, \ldots, C\}$ (for the sake of simplicity, with the same C for all h), is generated by a latent variable $\tilde{Y}_{ij}^{(h)}$ through the following relationship:

$$\left\{Y_{ij}^{(h)} = c^{(h)}\right\} \quad \Leftrightarrow \quad \left\{\gamma_{c^{(h)}-1}^{(h)} < \tilde{Y}_{ij}^{(h)} \leq \gamma_{c^{(h)}}^{(h)}\right\}, \tag{1}$$

where the thresholds satisfy $-\infty = \gamma_0^{(h)} \leq \gamma_1^{(h)} \leq \ldots \leq \gamma_{C-1}^{(h)} \leq \gamma_C^{(h)} = +\infty$.

Now let us consider the following multivariate two-level null model for the latent variables:

$$\tilde{Y}_{ij}^{(h)} = \alpha^{(h)} + u_j^{(h)} + \varepsilon_{ij}^{(h)}, \qquad h = 1, \cdots, H \tag{2}$$

where, for each h, $\alpha^{(h)}$ is the mean, $u_j^{(h)}$ is the cluster's random effect (level two error) and $\varepsilon_{ij}^{(h)}$ is the subject's disturbance (level one error). The errors are assumed to be distributed as

$$\left(\varepsilon_{ij}^{(1)}, \cdots, \varepsilon_{ij}^{(H)}\right)' \overset{iid}{\sim} N(\mathbf{0}, \mathbf{\Sigma}_\varepsilon) \qquad \left(u_j^{(1)}, \cdots, u_j^{(H)}\right)' \overset{iid}{\sim} N(\mathbf{0}, \mathbf{\Omega}). \tag{3}$$

For example, for H=2 the covariance matrices are

$$\mathbf{\Sigma}_\varepsilon = \begin{pmatrix} \sigma^2_{\varepsilon_1} & \\ \sigma_{\varepsilon_1\varepsilon_2} & \sigma^2_{\varepsilon_2} \end{pmatrix}, \quad \mathbf{\Omega} = \begin{pmatrix} \tau_1^2 & \\ \tau_{12} & \tau_2^2 \end{pmatrix}. \tag{4}$$

Moreover the first and second level errors are assumed to be independent, so $\mathrm{Cov}(\varepsilon_{ij}^{(h)}, u_j^{(k)}) = 0, \forall i, j, h, k$. These standard distributional assumptions on the errors allow the model to be both clearly interpretable and estimable with conventional statistical packages; note also that the Gaussian hypothesis is not as restrictive as it may seem, since it refers to latent variables.

The previous model specification implies the following conditional covariance structure for any couple of latent variables $(\tilde{Y}_{ij}^{(h)}, \tilde{Y}_{ij}^{(k)})$:

$$\begin{aligned} \mathrm{Cov}(\tilde{Y}_{ij}^{(h)}, \tilde{Y}_{i'j'}^{(k)} \mid u_j^{(h)}, u_j^{(k)}) &= \mathrm{E}(\varepsilon_{ij}^{(h)} \varepsilon_{i'j'}^{(k)}) \\ &= \begin{cases} \sigma^2_{\varepsilon_h} & \text{if} \quad k = h, j = j', i = i' \\ \sigma_{\varepsilon_h\varepsilon_k} & \text{if} \quad k \neq h, j = j', i = i' \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{5}$$

Then the ratio $\sigma_{\varepsilon_h\varepsilon_k}/\sigma_{\varepsilon_h}\sigma_{\varepsilon_k}$ can be interpreted as the conditional polychoric correlation (Drasgow, 1986) between the ordinal variables $(Y_{ij}^{(h)}, Y_{ij}^{(k)})$.

The unconditional covariance structure is:

$$\mathrm{Cov}(\tilde{Y}_{ij}^{(h)}, \tilde{Y}_{i'j'}^{(k)}) = \mathrm{E}(\varepsilon_{ij}^{(h)} \varepsilon_{i'j'}^{(k)}) + \mathrm{E}(u_j^{(h)} u_{j'}^{(k)}), \tag{6}$$

with $\mathrm{Cov}(\tilde{Y}_{ij}^{(h)}, \tilde{Y}_{i'j'}^{(k)}) = 0$ if $j \neq j'$, while the expression of $\mathrm{Cov}(\tilde{Y}_{ij}^{(h)}, \tilde{Y}_{i'j'}^{(k)})$ for $j = j'$ is reported in Table 1.

From the expressions reported in Table 1, three types of correlation can be defined:

- the correlation between the same variable for two distinct subjects of the same cluster, that is the intraclass correlation coefficient, ICC, representing also the proportion of variance explained by the clusters:

$$\mathrm{Corr}(\tilde{Y}_{ij}^{(h)}, \tilde{Y}_{i'j}^{(h)}) = \tau_h^2/(\sigma^2_{\varepsilon_h} + \tau_h^2) \qquad h = 1, \cdots, H;$$

Table 1: $\mathrm{Cov}(\tilde{y}_{ij}^{(h)}, \tilde{y}_{i'j'}^{(k)})$ for $j = j'$ (same cluster).

	$i = i'$	$i \neq i'$
$h = k$	$\sigma^2_{\varepsilon_h} + \tau^2_h$	τ^2_h
$h \neq k$	$\sigma_{\varepsilon_h \varepsilon_k} + \tau_{hk}$	τ_{hk}

- the correlation between two variables for the same subject (marginal polychoric correlation):

$$\mathrm{Corr}(\tilde{Y}_{ij}^{(h)}, \tilde{Y}_{ij}^{(k)}) = (\sigma_{\varepsilon_h \varepsilon_k} + \tau_{hk}) / \sqrt{(\sigma^2_{\varepsilon_h} + \tau^2_h)(\sigma^2_{\varepsilon_k} + \tau^2_k)}, \qquad h \neq k;$$

- the correlation between two variables for two distinct subjects of the same cluster:

$$\mathrm{Corr}(\tilde{Y}_{ij}^{(h)}, \tilde{Y}_{i'j}^{(k)}) = \tau_{hk} / \sqrt{(\sigma^2_{\varepsilon_h} + \tau^2_h)(\sigma^2_{\varepsilon_k} + \tau^2_k)}, \qquad h \neq k.$$

The cluster random effects $u_j^{(h)}$ may be viewed as second level factors, so the model described so far may be interpreted as a particular H-factor model. The one-factor version is obtained by specifying:

$$u_j^{(h)} = \lambda_h w_j, \qquad h = 1, 2, \cdots, H$$

where $w_j \overset{iid}{\sim} N(0,1)$ and the λ_h's are parameters. In this case, for any couple $u_j^{(h)}$ and $u_j^{(k)}$ the variances are distinct, but the factors are perfectly correlated. The unconditional covariances (6) are easily derived posing $\tau_h = \lambda_h$ and $\tau_{hk} = \lambda_h \lambda_k$.

To make the ordinal model identifiable, it is necessary to impose some constraints: in the following we assume $\gamma_h^{(1)} = 0$ and $\sigma_{\varepsilon_h} = 1$ for each $h, h = 1, \cdots, H$. Note that the model has H^2 estimable variance-covariance parameters, $H(H+1)/2$ at cluster level and $H(H-1)/2$ at subject level. As an example, for $H = 2$, the model has four estimable variance-covariance parameters, three at cluster level (τ_1^2, τ_2^2, τ_{12}) and one at subject level ($\sigma_{\varepsilon_1 \varepsilon_2}$). In the following we denote the set of all estimable parameters with $\boldsymbol{\theta}$.

The full model likelihood can be derived in the following steps. First, the conditional likelihood for subject i of cluster j is

$$L_{ij}(\boldsymbol{\theta} \,|\, \mathbf{u}) = \prod_{\mathbf{c} \in \mathcal{C}} \left[P\left(\bigcap_{h=1}^{H} \left\{ Y_{ij}^{(h)} = c^{(h)} \right\} \;\middle|\; \mathbf{u} \right) \right]^{d_{ij\mathbf{c}}}, \tag{7}$$

where $\mathcal{C}$ is the set of all admissible values of the vector $\mathbf{c} = (c^{(1)}, \cdots, c^{(H)})'$, $d_{ij\mathbf{c}}$ is the indicator function of the event $\left\{\bigcap_{h=1}^{H}\{Y_{ij}^{(h)} = c^{(h)}\}\right\}$, and $\mathbf{u} = (u^{(1)}, \cdots, u^{(H)})'$. Note that the relationship between the observed and latent variables (1) and the hypotheses on the latent model (2) imply that

$$\begin{aligned} & P\left(\bigcap_{h=1}^{H}\left\{Y_{ij}^{(h)} = c^{(h)}\right\} \;\middle|\; \mathbf{u}\right) \qquad (8) \\ = \; & P\left(\bigcap_{h=1}^{H}\left\{\gamma_{c^{(h)}-1}^{(h)} < \tilde{Y}_{ij}^{(h)} \leq \gamma_{c^{(h)}}^{(h)}\right\} \;\middle|\; \mathbf{u}\right) \\ = \; & E_{\boldsymbol{\varepsilon}}\left[\prod_{h=1}^{H} I\left\{\gamma_{c^{(h)}-1}^{(h)} - \alpha^{(h)} - u^{(h)} < \varepsilon^{(h)} \leq \gamma_{c^{(h)}}^{(h)} - \alpha^{(h)} - u^{(h)}\right\} \;\middle|\; \mathbf{u}\right], \end{aligned}$$

where $\boldsymbol{\varepsilon} = (\varepsilon^{(h)}, \cdots, \varepsilon^{(H)})'$; therefore, computation of the $L_{ij}(\boldsymbol{\theta} \,|\, \mathbf{u})$ involves an integral with respect to a multivariate Gaussian density.

Then the marginal likelihood for cluster j is

$$L_j(\boldsymbol{\theta}) = E_{\mathbf{u}}\left[\prod_{i=1}^{n_j} L_{ij}(\boldsymbol{\theta} \,|\, \mathbf{u})\right], \qquad (9)$$

involving another integral with respect to a multivariate Gaussian density. Finally, the overall marginal likelihood is

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{J} L_j(\boldsymbol{\theta}). \qquad (10)$$

Maximization of the marginal likelihood (10) requires the solution of the multiple integrals at subject and cluster levels.

In the bivariate case ($H = 2$) a little programming allows the maximization to be accomplished by the NLMIXED procedure of the SAS system (SAS Institute, 1999), which allows to specify an arbitrary conditional likelihood programmable with SAS statements. In fact the probabilities (8) can be written using the bivariate Gaussian distribution function and are calculated through finite differences, while the integration with respect to the random effects $u^{(1)}, u^{(2)}$ is performed through (possibly adaptive) Gaussian quadrature. The approximated marginal likelihood is maximized using a dual quasi-Newton algorithm.

For $H > 2$, the approach is still feasible in principle, but the likelihood involves the multivariate Gaussian density that is not directly available in

SAS. Actually, to our knowledge, no software is currently available to perform this kind of maximization. However, as shown in the next section, the model can be written in a form suitable for the routines implemented in standard multilevel analysis software.

Alternative specification of the model

In the case of continuous response variables it is customary to fit multivariate multilevel models by means of the estimation routines designed for univariate multilevel models (Snijders and Bosker, 1999). The procedure relies on the fact that the responses of the subjects may be viewed as an additional dummy bottom level. The same trick can be used for non linear multivariate multilevel models, as is done in packages such as GLLAMM (Rabe-Hesketh, Pickles & Skrondal, 2001b) and aML (Lillard & Panis, 2000). However, defining a complex multivariate model in a univariate framework may obscure the real meaning of the model. Therefore it is useful to highlight the relationship between the familiar two-level specification in terms of a multivariate latent distribution and the three-level univariate specification used by the specialized software.

First note that the multivariate two-level latent model (2) can be viewed as a univariate three-level model where the H responses form the new bottom level. In formal terms, this amounts to setting up a reparametrization based on the following decomposition of the first level error:

$$\varepsilon_{ij}^{(h)} = v_{ij}^{(h)} + \xi_{ij}^{(h)}, \tag{11}$$

where the v and ξ errors are independent, with:

$$\begin{aligned} \left(v_{ij}^{(1)}, \cdots, v_{ij}^{(H)}\right)' &\overset{iid}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma_v}) \\ \left(\xi_{ij}^{(1)}, \cdots, \xi_{ij}^{(H)}\right)' &\overset{iid}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{I}) \end{aligned} \tag{12}$$

It follows that the variance-covariance matrix of the ε errors is decomposed as $\boldsymbol{\Sigma_\varepsilon} = \boldsymbol{\Sigma_v} + \boldsymbol{I}$. For example, when $H = 2$:

$$\begin{pmatrix} \sigma^2_{\varepsilon_1} & \\ \sigma_{\varepsilon_1 v_2} & \sigma^2_{\varepsilon_2} \end{pmatrix} = \begin{pmatrix} \sigma^2_{v_1} & \\ \sigma_{v_1 v_2} & \sigma^2_{v_2} \end{pmatrix} + \begin{pmatrix} 1 & \\ 0 & 1 \end{pmatrix}.$$

The constraints on the parameters in the basic specification of the model, described in the previous section, must be reported also in the alternative

specification, imposing H constraints on the $H(H+1)/2$ parameters of $\mathbf{\Sigma}_v$, for example $\sigma^2_{v_1} = \cdots = \sigma^2_{v_H} = 1$.

Under decomposition (11), the latent model (2) becomes:

$$\tilde{Y}^{(h)}_{ij} = \alpha^{(h)} + u^{(h)}_j + v^{(h)}_{ij} + \xi^{(h)}_{ij}, \quad h = 1, \cdots, H. \tag{13}$$

Note that the responses are mutually independent conditionally on the random effects at subject level $(v^{(1)}_{ij}, \cdots, v^{(H)}_{ij})$ and cluster level $(u^{(1)}_j, \cdots, u^{(H)}_j)$. The conditional covariance structure for the latent variables is as (5), with $\sigma^2_{\varepsilon_h} = 1 + \sigma^2_{v_h}$ and $\sigma_{\varepsilon_h \varepsilon_k} = \sigma_{v_h v_k}$. Note that the conditional polychoric correlation between $Y^{(h)}_{ij}$ and $Y^{(k)}_{ij}$ is now $\sigma_{v_h v_k}/\sqrt{(\sigma^2_{v_h}+1)(\sigma^2_{v_k}+1)}$, which is different from the correlation between the subject's random effects $v^{(h)}_{ij}$ and $v^{(k)}_{ij}$.

It is important to note that the basic specification of the model, based on equations (2), and the alternative specification, based on equations (13), differ in the scaling factor used to ensure identifiability. In fact, in the alternative specification it is natural to fix to one the standard deviations of $\xi^{(h)}, \cdots, \xi^{(H)}$, while in the basic specification the corresponding constraint is imposed on the standard deviations of $\varepsilon^{(1)}, \cdots, \varepsilon^{(H)}$. In terms of the alternative specification (13), the constraints used in the basic specification amount to scaling the parameters of the h-th latent equation by a factor $\sqrt{1+\sigma^2_{v_h}}$, so the estimable parameters are smaller in magnitude, specifically they are $1/\sqrt{1+\sigma^2_{v_h}}$ times the corresponding parameters of the alternative specification.

The error decomposition (11) allows the likelihood to be written in a different way. In particular, the probability (8) is now

$$\begin{aligned}
& P\left(\bigcap_{h=1}^{H} \left\{Y^{(h)}_{ij} = c^{(h)}\right\} \;\middle|\; \mathbf{u}\right) \qquad (14)\\
= \;& P\left(\bigcap_{h=1}^{H} \left\{\gamma^{(h)}_{c^{(h)}-1} < \tilde{Y}^{(h)}_{ij} \leq \gamma^{(h)}_{c^{(h)}}\right\} \;\middle|\; \mathbf{u}\right)\\
= \;& E_{\mathbf{v}}\left[P\left(\bigcap_{h=1}^{H} \left\{\gamma^{(h)}_{c^{(h)}-1} < \tilde{Y}^{(h)}_{ij} \leq \gamma^{(h)}_{c^{(h)}}\right\} \;\middle|\; \mathbf{u}, \mathbf{v}\right)\right]\\
= \;& E_{\mathbf{v}}\left[\prod_{h=1}^{H} P\left(\left\{\gamma^{(h)}_{c^{(h)}-1} < \tilde{Y}^{(h)}_{ij} \leq \gamma^{(h)}_{c^{(h)}}\right\} \;\middle|\; u^{(h)}, v^{(h)}\right)\right]
\end{aligned}$$

$$= \quad E_{\mathbf{v}}\left[\prod_{h=1}^{H}\left(F_{\xi^{(h)}}(\gamma^{(h)}_{c^{(h)}} - \eta^{(h)}) - F_{\xi^{(h)}}(\gamma^{(h)}_{c^{(h)}-1} - \eta^{(h)})\right)\right],$$

where $\eta^{(h)} = \alpha^{(h)} + u^{(h)} + v^{(h)}$, $\mathbf{v} = (v^{(1)}, \cdots, v^{(H)})'$ and $F_{\xi^{(h)}}$ is the distribution function of the response-specific error $\xi^{(h)}$. This formulation allows the likelihood to be maximized with the standard algorithms for three-level models, making the joint analysis of multiple responses straightforward.

The three-level version of the model involves an additional step of integration, so it is likely to be computationally more heavy. However, the computational burden can be reduced by eliminating one of the v random effects (errors at subject level). In fact, the identifiability constraints imply that only $H(H-1)/2$ variance-covariance parameters are estimated at subject level, so instead of using H random effects with $H(H+1)/2$ parameters and H constraints, one can insert $H-1$ random effects. However, in this way some restriction is imposed on the correlations. For example, when $H = 2$, the two-level latent model (13) becomes:

$$\begin{aligned} \tilde{Y}^{(1)}_{ij} &= \alpha^{(1)} + u^{(1)}_j + v_{ij} + \xi^{(1)}_{ij} \\ \tilde{Y}^{(2)}_{ij} &= \alpha^{(2)} + u^{(2)}_j + v_{ij} + \xi^{(2)}_{ij} \end{aligned} \tag{15}$$

with $v_{ij} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2_v)$.

The conditional polychoric correlation is now $\sigma^2_v/(1+\sigma^2_v)$, which is restricted to be positive. A negative correlation is obtained multiplying v_{ij} by -1 in one of the two equations. In the model selection process one should try both versions to discover the sign of the correlation.

An alternative parametrization, which avoids the restriction on the correlations is the following:

$$\tilde{Y}^{(h)}_{ij} = \alpha^{(h)} + u^{(h)}_j + \lambda_h v_{ij} + \xi^{(h)}_{ij}, \quad h = 1, \cdots, H \tag{16}$$

with $v_{ij} \overset{iid}{\sim} \mathrm{N}(0, 1)$, and λ_1 fixed to one. The parameters λ_h determine the conditional variances of $\tilde{Y}^{(h)}_{ij}$, $\lambda^2_h + 1$, and the conditional polychoric correlation of the couples $(Y^{(h)}_{ij}, Y^{(k)}_{ij})$: $\lambda_h\lambda_k/\sqrt{(\lambda^2_h+1)(\lambda^2_k+1)}$.

An advantage of the three-level univariate formulation is that the inclusion of subjects with some missing responses is straightforward, as in any multilevel formulation of multivariate models (Snijders & Bosker, 1999, p.

200). This strategy gives unbiased results under the usual MAR (missing at random) assumption.

It is worthwhile to note that the three-level univariate formulation allows the model to be cast in the framework of Generalised Linear Latent and Mixed Models (Rabe-Hesketh et al., 2001a), a rather general class which includes, as special cases, multilevel structural equation and factor models.

Finally, it should be noted that the three-level univariate specifications shown in this section are equivalent to the two-level multivariate specification of the previous section only because of the assumption of Gaussian disturbances at response and subject levels. Without such an assumption the model based on the error decomposition (11) in general does not correspond to a model with a well-known multivariate distribution for the subjects' responses. However the multivariate Gaussian distribution may still be a good approximation of the true distribution: for example, this should be the case when using the logistic distribution at the lower level (i.e., replacing the probit link with the logit link), since, at least for binary responses, probit and logit models are empirically virtually indistinguishable (Rabe-Hesketh, Skrondal & Pickles, 2001).

Application

The data

The model presented in the previous sections has been used to analyze some of the data gathered in the survey of course quality carried out by the University of Florence in the 2000-2001 academic year.

Specifically, we considered the ratings relative to the courses held in the School of Pharmacy, excluding the courses for which there were less than five respondents. In order to compare the different estimation algorithms, we also excluded the questionnaires (3% of the total) with a missing response in either of the two jointly analyzed items, i.e., course workload (Q3: *Is the course workload acceptable?*) and clarity of the teacher (Q13). Altogether, 2888 questionnaires have been considered, corresponding to 87 courses. The number of ratings per course varies from 6 to 136, with a median of 32.

Table 2 reports the sample bivariate distribution of the jointly analyzed items, Q3 and Q13, measured on an ordinal scale: 1. decidedly no; 2. more no than yes; 3. more yes than no; 4. decidedly yes.

A standard measure of association among the two items is the polychoric correlation (Drasgow, 1986), which is the correlation coefficient of the un-

derlying bivariate normal distribution. Ignoring the hierarchical structure of the data, the polychoric correlation can be estimated by means of a null one-level bivariate probit model, which is fitted by the *plcorr* option of the SAS FREQ procedure applied to the bivariate sample distribution of Table 2. The resulting estimate is 0.448 (s.e. 0.019, see model (a) of Table 3).

Table 2: Ratings by course workload (Q3) and teacher clarity (Q13). The University of Florence, School of Pharmacy, academic year 2000-2001.

	Teacher clarity				*Total*	
Workload	1	2	3	4	N	%
1	44	24	30	6	104	3.6
2	64	138	186	76	464	16.1
3	116	268	718	486	1588	55.0
4	26	56	198	452	732	25.3
Total	250	486	1132	1020	2888	
%	8.7	16.8	39.2	35.3		100.0

In general, the student rating to a given item for a certain course may depend on characteristics measured at the following hierarchical levels: a) the student (background, expectations, etc.); b) the course (subject-matter, organization, professor, readings); c) the course of study or the school or department (halls, laboratories, students guidance, etc.); d) the university. Therefore a full analysis would require a complex multivariate multilevel model with several covariates (Snijders and Bosker, 1999). However since the application has a chiefly illustrative purpose, we limit the analysis to a single school, thereby eliminating the need for school and university levels; moreover we omit the covariates.

Model selection and estimation

Now let us consider the bivariate version of the multivariate two-level model defined in the first section, which can be used to jointly analyze the items Q3 and Q13. The bivariate model includes the correlation structure between the two items, which is interesting in itself and might also influence the other model parameters.

In order to identify the model, for each item we fix the first threshold $\gamma_1^{(h)}$ to zero and the first level standard deviation σ_{ε_h} to one, i.e., at the first level only the correlation is estimated.

Interpreting the random effects as course-level factors, two alternative models can be set up:

1. *one-factor model*: there is a single random effect at course level, entering the two linear predictors with different factor loadings;

2. *two-factor model*: there are two random effects at course level (one for each item), whose variances and covariance can be estimated.

The one-factor model is a special case of the two-factor model in which the factors are perfectly correlated. Having a single factor is useful in that the courses can be easily ranked on the basis of the predicted values of that factor; however such a model should be used only if supported by the data.

Table 3 reports the results for three types of bivariate model: (a) single-level (no factor) model, (b) two-level one-factor model and (c) two-level two-factor model. The estimates were obtained with the NLMIXED procedure of SAS, using for models (b) and (c) non-adaptive Gaussian quadrature with 21 points.

In terms of deviances ($-2logL$) the two-factor model is clearly preferable to the one-factor model. Even if the two models lead to a similar estimate of the marginal polychoric correlation (0.48 versus 0.45), in the two-factor model the correlation between the factors (0.62) is farther from unity, which is the value assumed by the one-factor model. The consequence of incorrectly assuming the existence of a single course-level factor seems to concern mainly item Q3: in the one-factor model the ICC goes down to 0.06, causing an attenuation in the intercept and thresholds.

Therefore the ranking of the courses on the basis of the considered items cannot rely upon a single measure. Instead, the predicted values of both factors (second level residuals) should be computed and plotted, as in Figure 1, where the best courses lie in the first quadrant, while the worst courses lie in the third quadrant. Extreme cases should be selected for further investigation.

Alternative estimation algorithms

In principle every software designed for three-level ordinal models is suitable to fit the model in the alternative specification described in the paper, in which the ratings of the subjects are considered as an additional dummy

Table 3: Bivariate null models fitted with SAS NLMIXED.

Parameter	(a) single level no factor		(b) two levels one factor		(c) two levels two factors	
	Estim	*s.e.*	*Estim*	*s.e.*	*Estim*	*s.e.*
fixed						
$\alpha^{(Q3)}$	1.80	0.044	1.81	0.046	1.99	0.063
$\gamma_2^{(Q3)}$	0.95	0.041	0.97	0.042	1.08	0.047
$\gamma_3^{(Q3)}$	2.46	0.048	2.53	0.050	2.79	0.056
$\alpha^{(Q13)}$	1.36	0.033	1.68	0.058	1.57	0.053
$\gamma_2^{(Q13)}$	0.69	0.029	0.94	0.038	0.94	0.038
$\gamma_3^{(Q13)}$	1.73	0.036	2.33	0.049	2.33	0.050
covariance						
$\rho_{\tilde{y}^{(Q3)}\tilde{y}^{(Q13)}\|u^{(Q3)},u^{(Q33)}}$	.	.	0.40	0.023	0.42	0.021
$\rho_{\tilde{y}^{(Q3)}\tilde{y}^{(Q13)}}$	0.45	0.019	0.45	0.019	0.48	0.025
$\mathrm{ICC}^{(Q3)}$	0	.	0.06	0.011	0.25	0.029
$\mathrm{ICC}^{(Q13)}$	0	.	0.43	0.023	0.50	0.025
$\rho_{u^{(Q3)}u^{(Q13)}}$	.	.	1	.	0.62	0.050
-2log$L$	13058		12029		11738	
n. of param.	7		9		10	

Note. For models (b) and (c) we use non-adaptive Gaussian quadrature with 21 points.

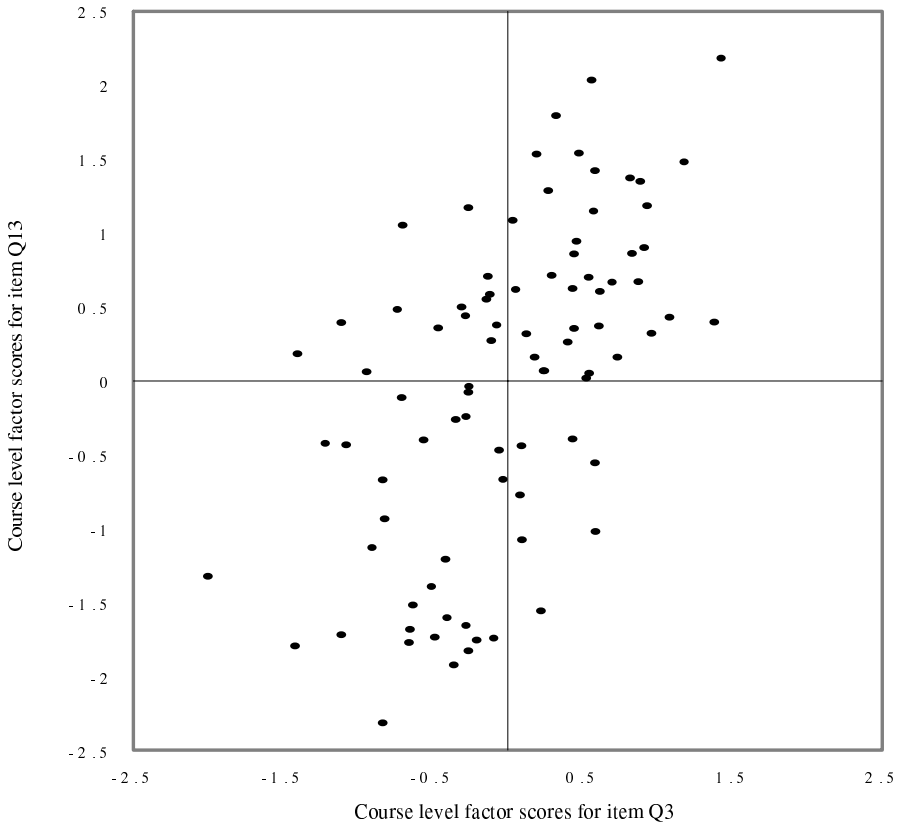
bottom level. This is not the case for SAS NLMIXED, which is currently limited to two-level structures.

In this application we used the GLLAMM procedure of Stata (Rabe-Hesketh et al., 2001b) and the aML software (Lillard & Panis, 2000).

The algorithms implemented in these packages rely on Gaussian quadrature to solve the integrals that appear in the marginal likelihood and admit, in principle, an arbitrary number of levels.

GLLAMM uses Stata's maximum likelihood functions to maximize the marginal likelihood by means of a modified Newton Raphson algorithm based on numerical first and second derivatives of the marginal log-likelihood; the predicted second level residuals can be obtained with the *gllapred* command of Stata (Fig. 1). In order to apply the algorithm, it is necessary to expand the data set, generating a number of records equal to the number of valid item responses.

aML is a statistical software for estimating multilevel multiprocess mod-


2 . 5
2
1 . 5
1
0 . 5
0
- 0 . 5
- 1
- 1 . 5
- 2
- 2 . 5
- 2 . 5
- 1 . 5
- 0 . 5
0 . 5
1 . 5
2 . 5
Course level factor scores for item Q13
Course level factor scores for item Q3

els. It handles a wide variety of models, among which random coefficients ordered probit and logit models. aML calculates maximum likelihood estimates using the BHHH optimization algorithm, i.e., the search direction is equal to minus the inverse of an approximation to the matrix of second derivatives times the gradient, where the approximation to the Hessian is equal to minus the sum over observations of the outer product of the gradient. At present the software does not compute the second level residuals.

We have also fitted the bivariate null model (c) of Table 3 with GLLAMM and aML. In both cases the estimates have been scaled to allow a correct comparison with NLMIXED results. The scaled estimates are reported in Table 4.

Using an adequate number of quadrature points, GLLAMM and aML yield substantially equal estimates, while NLMIXED estimates are slightly different. More investigation is needed to understand the source of this difference.

Computational remarks

The choice of the number of quadrature points at the higher level is crucial in the case of two or more random effects, as in model (c) of Table 3. As an example consider the results reported in Table 4.

First, let us consider the basic specification version of the model defined in (1)-(3) and fitted with SAS NLMIXED. In this case there is only one level which need numerical integration and we found that 21 points were adequate. Note that with 5 points the variance-covariance estimates are totally misleading and change in a substantive manner if the order of the equations is inverted! This phenomenon was noted also in univariate models with two random effects, in which the estimates obtained with few quadrature points are sensitive to the order in which the random effects enter the equation.

The situation is more complex for the univariate three-level version of the model defined in equations (15), which requires numerical integration at two levels of the hierarchy. However, with both GLLAMM and aML, it is clear that the course level, which has two random effects, is more demanding, in terms of quadrature points, than the subject level, which has a single random effect. For example, the estimates obtained with 21 quadrature points at both levels are the same as the estimates obtained with 21 points at the second level and 10 points at the first level.

In general, the computational time rapidly increases with the number of

Table 4: Bivariate null model (c) of Table 3: SAS NLMIXED estimates with different quadrature points number, GLLAMM and aML estimates.

Parameter	SAS NLMIXED			aML	GLLAMM
	5 qp	*10 qp*	*21 qp*	*10;21 qp*	*10;21 qp*
Fixed					
$\alpha^{(Q3)}$	1.934	1.854	1.988	2.055	2.055
$\gamma_2^{(Q3)}$	1.072	1.082	1.081	1.081	1.081
$\gamma_3^{(Q3)}$	2.764	2.782	2.785	2.785	2.786
$\alpha^{(Q13)}$	1.547	1.342	1.571	1.877	1.879
$\gamma_2^{(Q13)}$	0.924	0.928	0.940	0.936	0.935
$\gamma_3^{(Q13)}$	2.295	2.314	2.332	2.325	2.325
Covariance					
$\rho_{\tilde{y}^{(Q3)}\tilde{y}^{(Q13)}\|u^{(Q3)},u^{(Q13)}}$	0.425	0.422	0.421	0.424	0.423
$ICC^{(Q3)}$	0.194	0.219	0.245	0.228	0.228
$ICC^{(Q13)}$	0.302	0.522	0.501	0.401	0.400
$\rho_{u^{(Q3)}u^{(Q13)}}$	0.289	0.677	0.624	0.552	0.550
-2log$L$	11781	11767	11738	11732	11732

quadrature points.

Finally, note that both SAS NLMIXED and GLLAMM allow adaptive quadrature, which should yield the same level of precision of the estimates with a lesser number of quadrature points, thus reducing computational time.

Conclusions

The paper discussed the issues of specification and estimation in multivariate multilevel ordinal models, showing that they are strictly connected. In principle, model specification and interpretation are quite straightforward. However, existing estimation routines can be used only if some special parametriziation of the model is used. Moreover, since the estimation requires considerable computational effort, the model should be specified so that the estimation algorithm works in the most efficient manner.

Fitting a multilevel generalized linear model requires the solution of multiple integrals. Numerical integration with Gaussian quadrature (Anderson & Aitkin, 1985; Hedeker & Gibbons, 1994) allows for marginal maximum likelihood estimation, but becomes very cumbersome with multiple level struc-

tures and random effect covariates.

Various alternative solutions, often less precise but more efficient in terms of computational times, have been proposed, such as MQL and PQL (Goldstein & Rasbash, 1996), EM algorithm (Stiratelli, Laird & Ware, 1984), Gibbs sampling (Zeger & Karim, 1991), and Laplace approximations (Raudenbush & Yang, 1998). These techniques are implemented in specialized packages, like Mlwin (Goldstein et al., 1998) and HLM (Raudenbush, Bryk, Cheong, & Congdon, 2000).

Simulation-based estimators are an interesting alternative, particularly Simulated Maximum Likelihood (Gouriéroux and Monfort, 1991; Mazzolli, 2001; Calzolari, Mealli & Rampichini, 2001). However, this method is not yet implemented in standard software and its efficiency needs to be assessed.

Finally, note that, apart from computational considerations, the three-level univariate specification of the model is more appealing in that it casts the multivariate multilevel model in the Generalized Linear Latent and Mixed Models framework, which includes more complex models, such as multilevel factor and structural equations models.

References

Anderson, D. A., & Aitkin M. (1985). Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society B*, *47*, 203–210.

Calzolari G., Mealli F. , & Rampichini C. (2001). Alternative simulation-based estimators of logit models with random effects. *Quaderni del Dipartimento di Statistica 'G. Parenti'*, *48*, Firenze.

Drasgow F. (1986) Polychoric and polyserial correlations. In: N. L. Johnson, & S. Kotz, (Eds.) *Encyclopedia of statistical sciences*, *7*. New York: Wiley.

Goldstein H., Rasbash J., Plewis I., Draper D., Browne W., Yang M., Woodhouse G., & Healy M. J. R. (1998). *A User's Guide to MLwin*. London: Institute of Education.

Goldstein H., & Rasbasch J. (1996). Improved approximations for multilevel models with binary responses. *Working Paper, Multilevel Models Project* (University of London).

Gouriéroux C., & Monfort A. (1991). Simulation Based Econometrics in Models with Heterogeneity. *Annales d'Economie et de Statistique*, *20*, 69-107.

Hedeker D., & Gibbons R. D. (1994). A random effects ordinal regression model for multilevel analysis. *Biometrics*, *50*, 933-944.

Lillard L. A., & Panis S. (2000). *aML Multilevel Multiprocess Statistical Software, Release 1.0*. Los Angeles, California: EconWare.

Mazzolli B. (2001). A multilevel structural equation model with polytomous and dichotomous data, *mimeo* presented at the Third International Conference on Multilevel Analysis, Amsterdam, April 9-10, 2001.

Calzolari G., Mealli F. , & Rampichini C. (2001). Alternative simulation-based estimators of logit models with random effects, *Quaderni del Dipartimento di Statistica 'G. Parenti'*, *48*, Firenze: Dipartimento di Statistica 'G. Parenti'.

Rabe-Hesketh S., Pickles A., & Skrondal A. (2001a). GLLAMM: A general class of multilevel models and Stata program. *Multilevel Modelling Newsletter*, *13*(1), 17-23.

Rabe-Hesketh S., Pickles A., & Skrondal A. (2001b). GLLAMM Manual, *Technical report 2001/01*, London: Department of Biostatistics and Computing, Institute of Psychiatry, King's College, University of London.

Rabe-Hesketh S., Skrondal A., & Pickles A. (2001). Generalized Multilevel Parameterization of Multivariate Random Effects Models for Categorical Data. *Biometrics*, *57*, 1256-1264.

Rabe-Hesketh S., Skrondal A., & Pickles A. (2002). Generalized multilevel structural equation modelling, conditionally accepted in *Psychometrika*.

Raudenbush S., Bryk A., Cheong Y.F., & Congdon (2000). *HLM 5. Hierarchical linear and non linear models*, Lincolnwood, IL: Scientific Software International, Inc.

Raudenbush S., & Yang M. (1988). Maximum Likelihood for Hierarchical Models via High-order Multivariate Laplace Approximation, *mimeo*, University of Michigan.

Snijders T. A. B., & Bosker R. J. (1999) *Multilevel Analysis. An introduction to basic and advanced multilevel modelling*. London: Sage.

Stiratelli R., Laird N., & Ware J. H. (1984). Random-effects models for serial observations with binary response, *Biometrics 40*, 961 - 971.

Zeger S. L., & Karim R. M. (1991). Generalized linear models with random effects: a Gibbs sampler approach, *Journal of the American Statistical Association*, *86*, 79-86.

Authors

LEONARDO GRILLI is Assistant Professor of Statistics at the Department of Statistics "G. Parenti", University of Florence: grilli@ds.unifi.it. His main research interest is multilevel modelling in social sciences.

CARLA RAMPICHINI is Professor of Statistics at the Department of Statistics "G. Parenti", University of Florence: carla@ds.unifi.it. Her research interests include transition models, program evaluation, multilevel models and educational statistics.

Acknowledgements

We are grateful to Sophia Rabe-Hesketh for her constructive and detailed comments on a preliminary version of the paper and GLLAMM use, to Tom Snijders for his valuable comments and to Stan Panis for his useful suggestions on aML use and model specification.

Analysis of university course evaluations: from descriptive measures to multilevel models

Da: Rampichini C., Grilli L., Petrucci A. (2004) Analysis of university course evaluations: from descriptive measures to multilevel models, Statistical Methods and Applications, Vol. 13, n. 3, pp. 357-373.

Statistical Methods & Applications (2004) 13: 357–373
DOI: 10.1007/s10260-004-0087-1

SMA
© Springer-Verlag 2004

Analysis of university course evaluations: from descriptive measures to multilevel models

Carla Rampichini, Leonardo Grilli, Alessandra Petrucci

Dipartimento di Statistica "G. Parenti", Viale Morgagni 59, 50134 Firenze, Italy
(e-mail: {carla,grilli,alex}@ds.unifi.it)

Abstract. In the paper we present a comprehensive methodology for the analysis of student ratings of university courses. First, simple descriptive measures, which take into account the ordinal nature of the ratings, are discussed. Then net measures, which adjust for the characteristics of the students, are obtained through multilevel modelling. Finally, the measures relative to the various aspects of the course are synthesized through a weighted mean, building gross or net multidimensional indicators of course quality. The different indicators are then contrasted with respect to the rankings of courses they induce.

Key words: Course evaluation, student ratings, ordinal variables, multilevel models

1. Introduction

Student ratings are an old and widely recognized instrument to evaluate university courses (Emerson et al., 2000). Also in Italy there is a growing interest on this issue (D'Esposito, 2002; Grilli and Rampichini, 2002a and 2003; Pagani and Seghieri, 2002).

The statistical analysis of student ratings need special techniques which take into account the ordinal nature of the ratings, the multivariate nature of the data (the questionnaire includes several items) and the hierarchical structure of the phenomenon (ratings are nested in courses which are nested in schools).

Moreover, if one wishes to use the students' satisfaction as a measure of course quality, it should be recognised that the satisfaction of a student, as expressed

The current study has been financed with the MURST (40%) grant "Production and experimentation of computer-assisted systems to survey the quality of teaching at the university level and the success of university graduates in the job market" and with the MURST (40%) grant "Evaluation of quality and effectiveness in human services, with particular attention to health and education".

by the ratings, depends not only on the course characteristics of interest (lecture hall, clarity of the teacher, readings and so on), but also on the student's traits and expectations. Therefore a fair comparison among courses requires the calculation of net measures that adjust for individual characteristics. Such measures can be obtained, among others, by means of multilevel models (Goldstein, 2003; Snijders and Bosker, 1999).

The structure of the paper is as follows. In Section 2 the data at hand for the illustration of the proposed methods are described, while in Section 3 descriptive measures that take into account the ordinal nature of the ratings are proposed and discussed. In Section 4 net measures, useful for comparative evaluations, are obtained via multilevel models. Section 5 is devoted to the development of gross and net multidimensional indicators of course quality, which are compared with respect to the induced rankings of courses. Section 6 concludes.

2. The data

The methods presented in the paper will be illustrated with data gathered in the survey of course evaluation carried out by the University of Florence, in all schools of the university, for classes in the second semester of the 1999–2000 academic year.

The survey form was based on the proposal of a unique questionnaire for course evaluation by students formulated by the National Committee for the Evaluation of the University System (Chiandotto and Gola, 1999). The present application refers to required courses taken during the first year of the degree or diploma in the School of Engineering. Table 1 reports the number of enrolled students, the number of courses evaluated and the proportion of respondents per type of degree or diploma (to insure privacy, codes of courses for the degree or diploma have been substituted with labels). To avoid comparisons based on too few ratings, we considered only the courses with at least five respondents.

A total of 767 questionnaires have been considered, corresponding to 30 courses in the first year of the school of Engineering. For each course, the number of respondents varied from a minimum of 5 to a maximum of 60.

The average proportion of respondents is about 40% (with a minimum of 10.5% and a maximum of 76.2%). The low proportion found in some courses may be due to: (a) the substantial drop-out and transfer rates that are typical of the first year of enrolment; (b) the timing of the evaluations toward the end of the course, bringing with it a reduced class attendance level. Each respondent filled an individual data form including information about the student's academic history and then as many course evaluation forms as there were courses to be evaluated.

The individual data form and the evaluation forms are linked by an unambiguous code. The evaluation form is divided into a preliminary section, containing information about the course (the identification code, the name of the professor) and the attendance rate of the student, and a main section with 26 items concerning the evaluation of various aspects of the course and some student's characteristics. All the items in the main section require the same type of ordinal response:

1. decidedly *no*; 2. more *no* than *yes*; 3. more *yes* than *no*; 4. decidedly *yes*.

Table 1. Number of courses evaluated and the percentage of respondents relative to students enrolled. The University of Florence, School of Engineering, first year, academic year 1999–2000, second semester

Degree/ diploma	Enrolment	Courses evaluated	% Respondents Average	Min	Max
A	127	1	38.6	38.6	38.6
B	120	2	49.6	46.7	52.5
C	157	3	28.5	12.7	49.0
D	105	4	50.7	38.1	76.2
E	104	3	60.9	44.2	70.2
F	81	2	28.4	25.9	30.9
G	44	3	11.4	11.4	11.4
H	52	5	50.8	46.2	55.8
I	134	3	18.2	10.5	23.9
L	23	4	42.4	39.1	43.5
Tot		**30**	**39.3**	**10.5**	**76.2**

3. Descriptive measures

To simplify the analysis and dissemination of the results, student ratings are usually converted into scores and treated as quantitative variables. However, the arbitrariness implicit in the choice in the scoring system may be felt unacceptable by some of the subjects interested in the evaluation process, so it is advisable to develop simple descriptive measures directly based on the ordinal variables.

For example, the ratings' distribution of any course could be summarized by a location index and a dispersion index. For an ordinal variable the usual location index is the median, while for the dispersion index the choice is less obvious. Here we suggest using the following index (Leti, 1983, pp. 290-297; Piccarreta, 2001; Grilli and Rampichini, 2002b):

$$D = \sum_{k=1}^{K-1} F_k\,(1 - F_k), \quad 0 \leq D \leq \frac{K-1}{4}, \tag{1}$$

where $k = 1, 2, ..., K$ are the categories of the item of interest Y, and F_k is the relative cumulative distribution function of the Y variable. The supremum for the index D, reported in (1), actually holds only when the number of observations N is uneven (Leti, 1983). Nonetheless, for values of N that are sufficiently high, the supremum approximately holds also when N is even.

The dispersion index is often reported in standardized terms:

$$d = \frac{D}{(K-1)/4}, \quad 0 \leq d \leq 1. \tag{2}$$

Other dispersion indexes, also known as indexes of qualitative variation, may be used (Blair and Lacy, 2000, Yager, 2001); here the D measure was selected because of the decomposition property shown in Sect. 3.2.

Table 2. Median and standardized dispersion index for two courses in the first year. The University of Florence, School of Engineering, first year, academic year 1999-2000, second semester

Item number and definition		Course n. 8		Course n. 25	
		Me	d	Me	d
General aspects of the course					
1.	Lecture hall	4.0	0.35	3.0	0.32
4.	Course workload	3.0	0.41	3.0	0.21
11.	Readings	3.0	0.66	2.5	0.33
9.	On schedule with program	3.0	0.42	3.0	0.50
10.	Keeps scheduled hours	3.0	0.55	3.5	0.33
12.	Clear exam rules	4.0	0.44	3.0	0.53
17.	Availability of professor (outside class)	3.0	0.32	3.0	0.75
Teaching ability					
13.	Depth	3.0	0.57	3.0	0.50
14.	Clarity	3.0	0.64	1.0	0.32
15.	Ability to motivate	3.0	0.69	1.0	0.43
16.	Ability to answer questions	3.0	0.56	2.0	0.53
Student characteristics					
23.	Will take the exam in the first session	3.0	0.77	3.0	0.53
24.	Student's previous knowledge of the topic	2.0	0.72	3.0	0.53
25.	Student's interest in topic	2.0	0.77	4.0	0.53
26.	Overall satisfaction	2.0	0.69	2.0	0.50

3.1. Summarizing the ratings for a single course

The proposed indexes have been computed for some of the items of the questionnaire used in the survey described in Sect. 2 (a few items have been discarded due to problems in the survey and/or in the data coding). For example, Table 2 presents the values of the median and the standardized dispersion index (2) for two courses.

The problem remains how to summarize this information. In this regard, it is useful to provide the reader with a graphical representation of the indexes, in order to highlight particular features of each course. For any course, we suggest to represent the median values for each item on the y-axis and the standardized dispersion index d on the x-axis. Items that show lower dispersion are considered more reliable in their representation of course quality. Figures 1 and 2 are examples corresponding to the data reported in Table 2. It can be seen that for course n. 8 the median value of overall satisfaction is low (at least 50% of respondents are not satisfied with the course), but there is considerable variability in the individual responses.

This dissatisfaction seems to be tied to the lack of basic knowledge and interest for a good portion of the students. Still, both the professor and the readings, as well as the organization of the course, receive a favorable overall evaluation. The professor of this course could gather, from these results, the need to provide students with

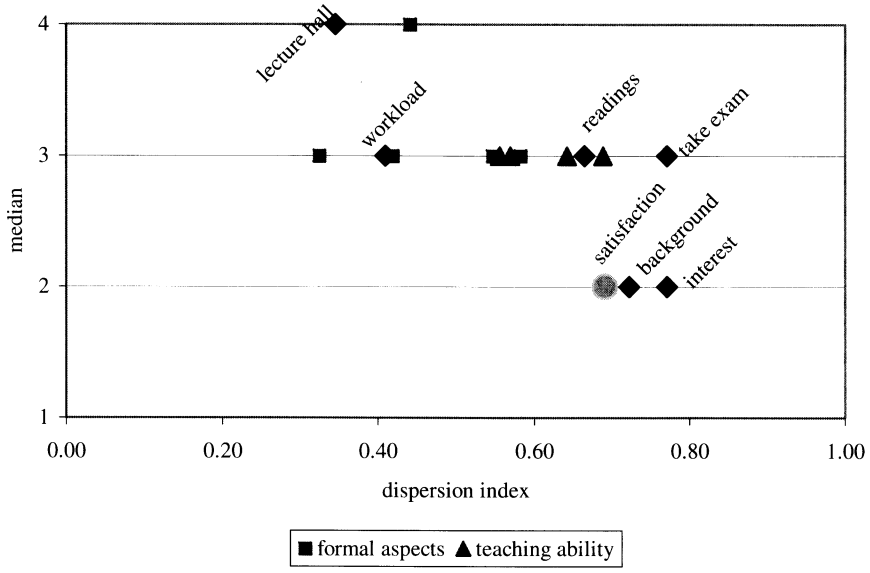


Fig. 1. Evaluation of course n. 8. Median and standardised dispersion index of selected items

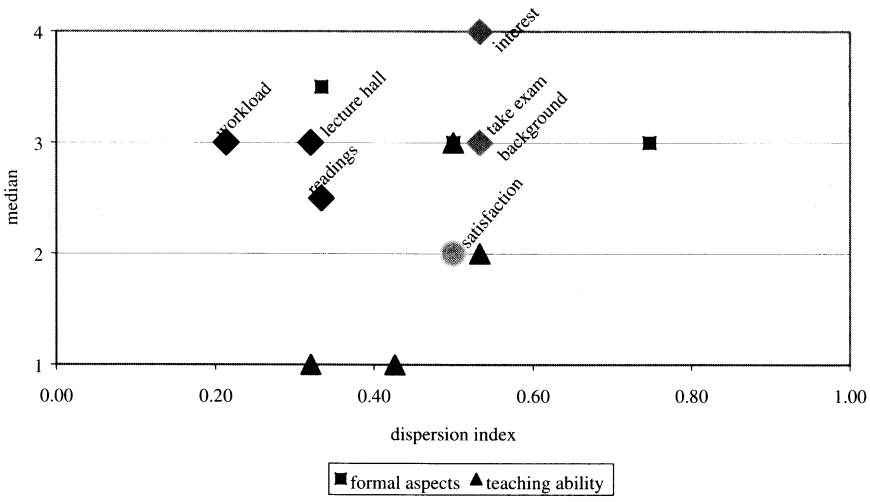


Fig. 2. Evaluation of course n. 25. Median and standardised dispersion index of selected items

a more extensive introduction to the material and to attempt to make the material itself more engaging for those who are less interested.

Course n. 25 as the same level of satisfaction than course n. 8, but for different reasons. In fact, the students show a strong interest in the subject and an adequate base of knowledge. In this case the dissatisfaction clearly comes from a strong negative evaluation of the teaching capacity of the professor (rendering possible improvements more difficult).

In this example the response scale has only four categories, so the median is a somewhat rough location index. Alternatively, one may use the proportion of favourable evaluations (in this case categories three and four).

3.2. Decomposition of the total dispersion: dispersion "between" and "within" courses

It is possible to decompose the overall dispersion of each item in two parts, "between" D_B and "within" D_W the courses (Grilli and Rampichini, 2002b):

$$D_W = \sum_{j=1}^{J} \pi_j D_j,\ D_B = \sum_{j=1}^{J} \pi_j Z_{2j}^2\,, \tag{3}$$

$$D_j = \sum_{k=1}^{K-1} F_{jk}\,(1 - F_{jk})\,, \tag{4}$$

$$Z_{2j}^2 = \sum_{k=1}^{K-1} (F_{jk} - F_k)^2,\quad F_k = \sum_{j=1}^{J} \pi_j F_{jk}\ ,$$

where D_j is the dispersion index calculated for j-th course, $j = 1, 2, ..., J$, Z_{2j} is the quadratic dissimilarity index (Leti, 1983, p. 153) between the j-th conditional distribution (that is relative to the j-th course) and the marginal distribution of the ordinal variable of interest, F_k represents the relative marginal distribution function, obtained as a mixture of the conditional relative distributions F_{jk}, and $\pi_j = N_j/N$ is the proportion of evaluation forms relative to the j-th course.

The total dispersion D is, therefore, obtainable as the sum of D_B and D_W. The proportion of dispersion among courses (Piccarreta, 2001)

$$\delta = D_B/D \tag{5}$$

provides an indication of how to interpret the different evaluations formulated for an item. In fact, the higher is the value of δ, the more is the influence of the course's characteristics on the expressed ratings. Therefore the items with the higher value of δ are the most reliable indicators of course quality.

As an example, Table 3 shows the percentage conditional distributions by course and the marginal distribution relative to the *overall satisfaction* for the course, the proportion of observations per course $\pi_j (j = 1, 2, \ldots, 30)$ and the values of the dispersion indexes D_j and D. Note that in this application $(K-1)/4 = 0.75$, thus these indexes, on the basis of (1), assume values between 0 and 0.75.

The index D_j is highest for course n. 26 (0.534) and lowest for course n. 28 (0.188), for which the evaluations are quite concentrated in the fourth category. A synthetic measure of the dispersion "within" the courses is given by the D_W index, which is equal to 0.433, while the dispersion "between" courses is obtained by subtracting from the total value $D = 0.509$ the value D_W, such that $D_B = 0.075$. The proportion of dispersion linked to course δ is, therefore, equal to 0.148. This fairly low value means that, in explaining the differences in judgments formulated

Table 3. Overall satisfaction with the course. Conditional percentage distributions, percentage of observations by course and dispersion indexes. The University of Florence, School of Engineering, first year, academic year 1999–2000, second semester

Course	Evaluation				$100\pi_j$	D_j
	1	2	3	4		
1	18.60	41.86	25.58	13.95	4.93	0.511
2	10.61	30.30	43.94	15.15	7.56	0.465
3	17.24	51.72	24.14	6.90	3.32	0.421
4	10.00	26.00	34.00	30.00	5.73	0.530
5	17.65	52.94	17.65	11.76	1.95	0.457
6	11.43	14.29	40.00	34.29	4.01	0.518
7	5.88	11.76	47.06	35.29	1.95	0.429
8	20.93	30.23	37.21	11.63	4.93	0.518
9	3.03	12.12	59.09	25.76	7.56	0.349
10	0	0	50.00	50.00	0.46	0.250
11	0	0	60.00	40.00	0.57	0.240
12	4.17	12.50	45.83	37.50	2.75	0.413
13	13.64	50.00	27.27	9.09	2.52	0.432
14	12.50	33.33	45.83	8.33	2.75	0.434
15	4.55	18.18	54.55	22.73	2.52	0.395
16	2.94	19.12	45.58	32.35	7.79	0.419
17	20.00	31.11	35.56	13.33	5.15	0.525
18	0	8.11	29.72	62.16	4.24	0.310
19	25.58	6.98	58.14	9.30	4.93	0.494
20	15.71	50.00	30.00	4.29	8.02	0.399
21	0	25.00	68.75	6.25	1.83	0.246
22	10.00	50.00	40.00	0	1.15	0.330
23	15.38	34.62	38.46	11.54	2.98	0.482
24	0	7.41	55.56	37.04	3.09	0.302
25	25.00	50.00	25.00	0	0.46	0.375
26	50.00	19.23	23.08	7.69	2.98	0.534
27	22.22	33.33	44.44	0	1.03	0.420
28	0	0	25.00	75.00	0.92	0.188
29	0	50.00	37.50	12.50	0.92	0.359
30	0	44.44	33.33	22.22	1.03	0.420
Tot.	**12.14**	**26.92**	**40.32**	**20.62**	**100.00**	**0.509**

about the course, individual variability plays a prominent role with respect to that tied to the characteristics of the course.

Should one wish to compare courses on the basis of students' evaluations, these results provide a clear indication toward the construction of net indexes that take into account the individual heterogeneity, as we will show in the next section.

4. Comparative evaluations

The descriptive indicators proposed in the previous section are useful mainly for the analysis of a single course. However, when it comes to comparisons among

courses, they are not appropriate since the evaluations expressed by the students are influenced by individual expectations and traits of the students themselves. Any comparison among courses that is based on descriptive indicators, therefore, could be misleading.

In general, the rating of a student to a given item for a certain course may depend on any of the following:

- Characteristics of the student (background, expectations, etc.);
- Characteristics of the course (subject-matter, organization, professor, readings);
- Characteristics of the course of study or of the school or department (halls, laboratories, students guidance, etc.);
- Characteristics of the University.

To make fair comparisons among courses it is necessary to isolate the influence of each of these components on the expressed evaluations. In particular, it is important to net out the effects related to the individual characteristics of students in order to obtain a fair ranking of courses, schools, universities etc., thereby bringing to light possibly critical situations to than further pursue.

In the present application we consider a single school, so the comparisons among the courses do not require adjusting for the characteristics of the school or university. The adjustment for *observed* individual characteristics could be carried out by calculating, for each item, net indicators derived from multilevel models (Goldstein, 2003; Snijders and Bosker, 1999), that allow for the estimation of the effect of the course net of the influence of observed individual variables.

4.1. The model

The model used in the analysis is defined as follows. Let assume that the observed ordinal response variable Y, with $k = 1, \ldots, K$ levels, is generated by a latent continuous variable Y^*, which represents the intensity of the concept expressed, following a two-level random intercept model:

$$Y_{ij}^* = \alpha + \boldsymbol{\beta}' \mathbf{x}_{ij} + u_j + \varepsilon_{ij} \tag{6}$$

with $i = 1, 2, \ldots, n_j$ ratings for the j-th course ($j = 1, 2, \ldots, J$). In (6) $\mathbf{x}_{ij}$ is the vector of dimension $p \times 1$ of observed variables, measured at the individual or course level; α is the intercept, $\boldsymbol{\beta}$ is the vector $p \times 1$ of the fixed coefficients; the random variables ε_{ij} and u_j are the disturbances, respectively at the first (individual) and second (course) level, for which we assume:

$$\begin{aligned} & u_j \overset{iid}{\sim} N(0, \sigma_u^2),\ \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)\,, \\ & Cov(u_j, \varepsilon_{ij}) = 0\,,\ \forall i, j \end{aligned} \tag{7}$$

where σ_ε^2 and σ_u^2 are the variance components, respectively at the first and second level.

The observed ordinal variable Y is linked to the latent Y^* through the following relationship:

$$\begin{aligned} Y_{ij} = k \quad &\text{if and only if} \quad \gamma_{k-1} < Y_{ij}^* \leq \gamma_k, \\ -\infty = \gamma_0 \leq \gamma_1 \leq \ldots \leq \gamma_{K-1} \leq \gamma_K &= +\infty \end{aligned} \tag{8}$$

Therefore, denoting with $\eta_{ij} = \alpha + \boldsymbol{\beta}' \mathbf{x}_{ij} + u_j$ the linear predictor, the conditional probability of observing a response equal to k is

$$P(Y_{ij} = k|u_j) = P(\gamma_{k-1} < Y_{ij}^* \leq \gamma_k|u_j) = \Phi(\gamma_k - \eta_{ij}) - \Phi(\gamma_{k-1} - \eta_{ij}),$$

where Φ is the standard Gaussian distribution function. Equations (6)–(8) define a two-level random intercept ordinal probit model. In order to identify the model, the following standard assumptions are made: (i) $\sigma_\varepsilon^2 = 1$; and (ii) $\gamma_1 = 0$.

Note that the assumptions on the disturbances (7) imply that, conditional on the covariates, the evaluations of different courses are independent, even if they are expressed by the same student (we will discuss this point later). On the contrary, the ratings of a single course are correlated, since:

$$Corr(Y_{ij}^*, Y_{lm}^*) = \rho = \begin{cases} 0 & \text{if } j \neq m \\ \dfrac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2} & \text{if } j = m \end{cases}$$

The intraclass correlation coefficient ρ also provides a measure of the amount of variance among courses relative to the overall variance.

4.2. Results

In order to make an initial assessment of the proportion of variance in the evaluations linked to the courses, it is useful to estimate, for each item of the questionnaire, the random intercept probit model (6)–(8) without covariates (*null model*). The model parameters were estimated by means of the software MIXOR (Hedecker and Gibbons, 1996), which performs marginal maximum likelihood with Gaussian quadrature.

Table 4 reports the estimation of the intraclass correlation coefficient ρ obtained from the null model for each item of the questionnaire, together with the descriptive indexes defined in the Sect. 3.

As already noted, the estimate of ρ derived from the null model represents the proportion of variance of the latent evaluation Y^* attributable to the courses. The higher the value of ρ, the more the various evaluations can be explained by characteristics of the course. Consistently, the proportion of variability among courses is particularly high for the items directly related to characteristics of the course (particularly teacher clarity, lecture hall and course workload), while the items most linked to the student (previous knowledge of the subject, interest in the subject and the intention of taking the exam in the first session) show a low value. In no case, however, do these coefficients result as being not significant. As expected, the

Table 4. Median, dispersion and proportion of dispersion among courses δ for the observed ratings Y, and intraclass correlation coefficient ρ estimated for the latent ratings Y^* by the null model for each item

	Item	Median	d	δ	ρ
1.	Lecture hall	3	0.62	0.24	0.29
4.	Course workload	3	0.59	0.13	0.24
9.	On schedule with program	4	0.52	0.10	0.11
10.	Keeps scheduled hours	3	0.62	0.08	0.09
11.	Readings	3	0.70	0.15	0.23
12.	Clear exam rules	3	0.67	0.13	0.18
13.	Depth	3	0.66	0.17	0.24
14.	Clarity	3	0.76	0.25	0.36
15.	Ability to motivate	2	0.72	0.14	0.20
16.	Ability to answer questions	3	0.64	0.12	0.16
17.	Availability of professor (outside class)	3	0.58	0.11	0.14
23.	Will take the exam	3	0.67	0.08	0.09
24.	Student's previous knowledge of the subject	2	0.71	0.10	0.16
25.	Student's interest in the subject	3	0.70	0.09	0.11
26.	Overall satisfaction	3	0.68	0.15	0.21

indexes ρ and δ provide analogous information, though the values of ρ are always higher.

The next step is to explain the variability among courses, conditionally on observed individual characteristics. In the context of our application, the variability of the responses can be attributed to the following factors:

- Observed individual characteristics: sex, age group, high school record and type of high school attended, year of first enrolment, full-time or part-time student;
- Self-assessed individual characteristics: previous familiarity with topic, interest in the subject, attendance with the intention of taking the exam in the first examination session;
- Characteristics of the course: suitability of classroom space, course workload, readings, teaching ability of professor, number of attendants. Unfortunately, the number of attendants is not available for this application.

As an example, we briefly discuss the results of the analysis on the responses to the item relative to the overall satisfaction (Table 5). It is worth to note that the observed individual characteristics (sex, age group, etc.) are not significant, with the only exception of full-time status, and so they are not included in the models reported in Table 5.

The null model provides an estimate of the intraclass correlation coefficient ρ equal to 0.21: this means that about one fifth of the total variability found in the evaluations is attributable to characteristics of the course. In general, the inclusion of covariates at course level necessarily reduce ρ, while an individual level covariate has an unpredictable effect, since it acts on both σ^2_ε and σ^2_u (σ^2_u may even increase, see for example Snijders and Bosker, 1999). In our application none of the indi-

Table 5. Random intercept probit models for overall satisfaction (item n. 26)

Model for overall satisfaction		n. param.	$-2LogL$	ρ
1.	Null	2	1900.4	0.21
2.	Full-time student (FT)	3	1897.6	0.21
3.	FT, exam (item n. 23)	6	1794.3	0.20
4.	FT, previous knowledge (item n. 24)	6	1769.6	0.18
5.	FT, interest (item n. 25)	6	1630.7	0.26
6.	FT, previous knowledge, exam, interest	12	1539.3	0.24

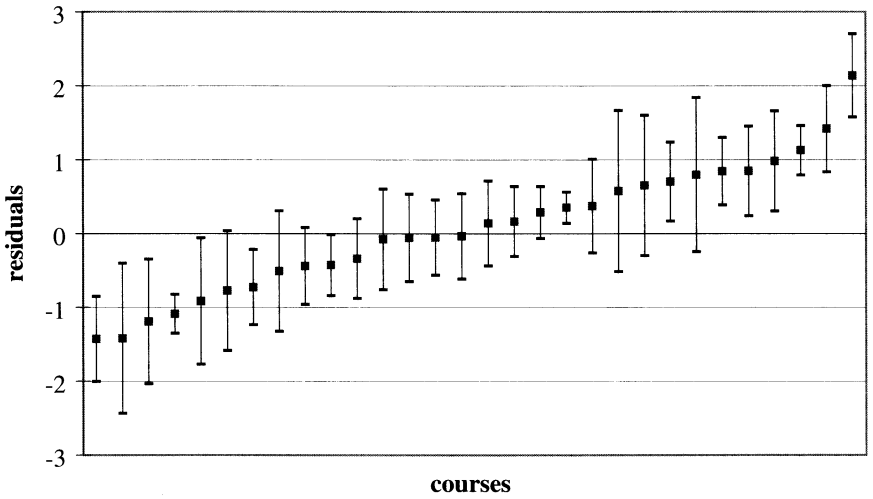


Fig. 3. Model 6 of Table 5: point estimates and 95% pariwise confidence intervals for the second level residuals

vidual characteristics (age, sex, etc.), with the exception of the status of full-time versus part-time student, is statistically significant in explaining the ratings. On the other hand, self-assessed individual characteristics have a strong effect: in fact, the deviance ($-2logL$) reduces from 1897.6 to 1539.3 after the introduction of the following variables: (i) whether the student will take the exam at the first examination session (*exam*); (ii) whether previous knowledge of the subject is judged by the student as sufficient for meeting the challenge of the course (*previous knowledge*); and (iii) whether the student is interested in the subject, independently from how the course is taught (*interest*). Each of these variables, by their nature ordinal with four categories, has been inserted in the model as a set of three dichotomous variables. Nevertheless, these covariates do nothing to reduce the variability among courses. On the contrary, they allow for a distinction among students with different expectations and background, putting into clearer focus the differences between courses, as indicated by the increase in the value of ρ.

With the aim of pointing out the 'anomalous' courses, both positive and negative, we have estimated the second level residuals u_j from Model 6 in Table 5 (Fig. 3).

These residuals can be interpreted as a measure of the effect of the course on the evaluations expressed by students, conditional on the variables inserted into the model; therefore they can be considered as 'net' indicators of the course quality with respect to the given item (Goldstein and Spigelhalter,1996; Spencer and Fielding, 1998). The 95% pairwise confidence intervals around the residuals are constructed so that two residuals are considered significantly different from one another if and only if the respective intervals are disjoint (Goldstein and Healy, 1995).

Figure 3 highlights both the extremely negative and extremely positive situations. Further analysis 'on the field' will provide elements for evaluation of these courses and indications for improvement of the worst situations.

4.3. Students' unobserved heterogeneity

Each student filled several course evaluation forms (see Sect. 2) linked by an identification code, so the ratings can be cross-classified by course and by student. The inclusion in the model of a student random effect allows for individual unobserved characteristics, which may cause a correlation among the ratings expressed by the same student. The presence of random effects for both the course and the student gives rise to a cross-classified multilevel model (Rasbash and Goldstein, 1994; Browne et al., 2001), a conceptually simple extension that, however, requires a considerable effort to be fitted with an ordinal response variable. Therefore to evaluate the role of the students' unobserved heterogeneity, we first fitted a model with a single random effect at student level and the same covariates as in Model 6 of Table 5, obtaining an estimated intraclass correlation coefficient ρ equal to 0.024, which is non significant by the likelihood ratio test and considerably smaller than the corresponding value previously found for the course-level component. Moreover, to evaluate the effect of ignoring students' unobserved heterogeneity, we transformed the ordinal observed ratings in continuous variables through a conditional mean scoring (Fielding, 1997) and fitted a cross-classified model by means of SAS PROC MIXED (SAS Institute, 1999). The estimated student variance component is quite low and, comparing with the corresponding model without unobserved heterogeneity, there is no relevant change in the value of the other fixed and random parameters, nor in the course residuals.

For these reasons, we did not try to fit the cross-classified model for the ordinal ratings and simply ignored the students' unobserved heterogeneity component.

5. Multidimensional indicators of course quality

So far, we have considered only one item at a time, trying to single out the individual and course characteristics that explain the variability of the responses provided by students and, on this basis, to bring into focus any anomalous situations.

Nonetheless, the quality of a course (or the set of courses for a given course of study, school or university) is a multidimensional concept: multiple factors contribute to form the final teaching result of a degree program. This aspect can be taken into account by devising a multidimensional indicator of course quality, based on a

Table 6. Principal Components Analysis on polychoric correlation matrix

	ITEM	Rotated (varimax) Factor Pattern with 4 Principal Components				Total Communality
		PC1	**PC2**	**PC3**	**PC4**	
1.	Lecture hall	0.24	0.17	**0.92**	0.11	0.94
4.	Course workload	0.34	0.41	0.22	**0.67**	0.79
9.	On schedule with program	0.31	**0.84**	0.10	0.03	0.81
10.	Keeps scheduled hours	0.22	**0.80**	0.12	0.19	0.74
11.	Readings	0.41	**0.68**	0.12	0.20	0.67
12.	Clear exam rules	0.34	**0.54**	0.39	0.07	0.57
13.	Depth	**0.74**	0.37	0.12	0.18	0.73
14.	Clarity	**0.81**	0.34	0.14	0.09	0.79
15.	Ability to motivate	**0.78**	0.25	0.19	0.10	0.71
16.	Ability to answer questions	**0.77**	0.30	0.24	0.05	0.75
17.	Availability of professor	**0.54**	0.49	0.19	0.41	0.75
26.	Overall satisfaction	**0.74**	0.20	0.15	0.35	0.74
% Total variance		0.55	0.08	0.06	0.06	0.75

battery of items that characterize the quality of a course, such that the information provided is not redundant and is sufficiently explanatory of the general level of satisfaction in the class. In any case the choice of the set of items is a crucial step that necessarily involves some arbitrariness.

5.1. Selection of the items

A general strategy to select the items could be based on a preliminary principal component analysis on the estimated correlation matrix. Given the ordinal nature of the ratings, it is advisable to use the polychoric correlation (Drasgow, 1986), which is the correlation coefficient of the underlying bivariate normal distribution. Ignoring the hierarchical structure of the data, the polychoric correlation can be estimated by means of a null bivariate probit model, which is fitted by the *plcorr* option of the SAS FREQ procedure (SAS Institute, 1999). The inspection of the estimated polychoric correlation matrix on the data at hand by means of a principal component analysis (Table 6) clearly shows the presence of at least 4 dimensions of the satisfaction: "Teaching ability", "Organizational aspects", "Lecture hall adequacy" and "Course workload". This result confirms the multidimensional nature of the satisfaction and the first principal component cannot be used as a synthetic measure of course satisfaction. Note that the overall satisfaction is strongly related to teaching ability, so it cannot be used as a single indicator of the course quality.

In order to define the elements to be included in the synthetic indicator, two strategies could be followed: to compute the scores on the first four principal components, or to select a set of items that represents the four dimensions. Since the results are to be disseminated to a heterogeneous audience, we prefer to follow the second strategy, which leads to a simpler interpretation of the results. For the third

and fourth dimensions, we select the only item with a relevant correlation with the corresponding principal component. For the first and second dimensions, the chosen item is the one with the higher value of δ (Table 4). Thus, the selected items are: Lecture hall (item n. 1), Course workload (item n. 4), Readings (item n. 11), Teacher clarity (item n. 14).

5.2. Definition of the multidimensional indicators

For a given course j, a multidimensional indicator of course quality I_j could be defined as a weighted average across the chosen items ($q = 1, \ldots, Q$) of the item-specific indicators I_{qj}, where the weights w_q represents the reliability of I_{qj}. In symbols:

$$I_j = \frac{1}{\sum_{q=1}^{Q} w_q} \sum_{q=1}^{Q} I_{qj} w_q. \tag{9}$$

Depending on the nature of the item-specific indicators I_{qj}, the multidimensional indicators can be of several types. Two useful classifications are the following:

- gross versus net indicators: gross indicators do not adjust for student characteristics, while net indicators do;
- simple versus model-based: simple indicators use only descriptive measures, while model-based indicators rely upon a statistical model.

In the following we use three indicators: I^{SG} (simple/gross), I^{MG} (model-based/gross) and I^{MN} (model-based/net). They all have the structure of formula (9), with I_{qj} and w_q defined in Table 7. Note that w_q, the weight of the indicator for the q-th item, is a measure of dispersion, or variance, "between" courses. This choice is motivated by the fact that the higher is the proportion of dispersion, or variance, due to the courses, the more are the considered item ratings reliable as an indicator of course quality.

5.3. Application

On the basis of the indicators defined in Sect. 5.2, it is possible to build a general ranking of courses through which it is possible to isolate extreme cases (Table 8, where rank 1 corresponds to the worst course and 30 to the better course).

The correlation among the three rankings of Table 8, reported in Table 9, is very high: it is 0.93 for the two gross indicators and 0.91 for the two model-based indicators. Particularly, the extremes of the three rankings are the same: the worst courses and the best courses are, in this particular example, identified by all the proposed rankings. However, there are some courses that have a clearly different position in the three rankings. Course n. 5, for example, "gains" nine positions when the net instead of the gross indicator is considered, that is, due to the adjustment on student characteristics (different levels of interest in the discipline, knowledge

Table 7. Definition of multidimensional indicators of course quality based upon formula (9)

I_j	I_{qj}	w_q
I_j^{SG} (simple/gross)	Me_{qj} (median)	$D_{B,q}$ (dispersion between)
I_j^{MG} (model-based/ gross)	$\hat{u}_{qj}^{NULL}$ (2^{nd} level residual from the null model)	$\hat{\sigma}_{u,q}^{2\ NULL}$ (2^{nd} level variance from the null model)
I_j^{MN} (model-based/ net)	$\hat{u}_{qj}$ (2^{nd} level residual from the model with student's covariates)	$\hat{\sigma}_{u,q}^2$ (2^{nd} level variance from the model with student's covariates)

Table 8. Evaluation of courses: multidimensional indicators of course quality (rankings). The University of Florence, School of Engineering, first year, academic year 1999–2000, second semester

Course	I^{SG}	I^{MG}	I^{MN}
26	1	1	1
1	4	3	2
22	3	2	3
25	2	8	4
17	7	6	5
20	11	7	6
13	5	4	7
15	**21**	**17**	**8**
23	12	5	9
2	6	9	10
14	9.5	12	11
4	9.5	11	12
11	18	19	13
27	8	13	14
3	13	10	15
19	14	16	16
21	18	15	17
10	18	22	18
6	18	20	19
7	22	23	20
16	24.5	21	21
29	23	24	22
5	**15**	**14**	**23**
12	18	25	24
28	29	28	25
8	24.5	18	26
30	26	26	27
9	27.5	27	28
18	30	30	29
24	27.5	29	30

Table 9. Correlations between rankings obtained with different indicators. The University of Florence, School of Engineering, first year, academic year 1999–2000, second semester

	I^{SG}	I^{MG}	I^{MN}
I^{SG}	1.00	–	–
I^{MG}	0.93	1.00	–
I^{MN}	0.90	0.91	1.00

bases, and intention of taking the exam in the first session). Course n. 15, instead, "loses" positions, going from seventeenth to the eighth position in the ranking.

We wish to stress that the ranking obtained from the net model-based multidimensional indicator should be interpreted with caution, due to the uncertainty of the residuals, as is clear from Figure 3. The main use of this ranking is to characterize groups of courses (e.g., it may be that all the courses of a certain subject appear in the bottom of the ranking) and to find out extreme cases.

6. Conclusions

In the paper we proposed the construction of some multidimensional indicators of course quality, as perceived by students, giving to each of the chosen items of the evaluation questionnaire a weight proportional to its 'between course' variability. The proposed indicators, which take explicitly into account the ordinal nature of the ratings, can be distinguished in: simple versus model-based indicators, and gross versus net indicators. Model-based net indicators are more appropriate in principle, since the analysis shows that student ratings are strongly influenced by individual traits and expectations. A measure of the course component, adjusting for observed student characteristics, can be obtained by means of appropriate statistical models, such as ordinal random intercept models.

The proposed indicators have been used to derive rankings of the courses, which are useful instruments for planning purposes.

In the present application, the gross indicator based on descriptive measures leads essentially to the same conclusions as the model-based net indicator. However, this result may be due to the homogeneity of the first year student body in Engineering and cannot be generalized to the case of courses with more heterogeneous student bodies (different enrollment years, different schools, etc.), where the phenomenon of self-selection may play a major role.

Note that the multidimensional indicators defined in Sect. 5 are obtained from a two-step process: (i) an item-by-item analysis and (ii) synthesis. This synthesis may be based on other methods, for example multi-criteria analysis (Keeney and Raiffa, 1976). Alternatively, the two-step process could be replaced by a joint analysis of the items by means of a multivariate multilevel model (Rabe-Hesketh et al., 2003; Grilli and Rampichini, 2003). In particular, for the construction of a unidimensional scale a standard methodology is the Rasch model (Bond and Fox, 2001; Beltyukova and Fox, 2002), extended to the multilevel case (Fox and Glas, 2001).

References

Beltyukova SA, Fox CM (2002) Student satisfaction as a measure of student development: towards a universal metric. Journal of College Student Development 43(2): 1–12

Blair J, Lacy MG (2000) Statistics of ordinal variation. Sociological Methods & Research 28: 251–280

Bond TG, Fox C M (2001) Applying the Rasch model: fundamental measurement in the human sciences. Erlbaum, Mahwah, NJ

Browne WJ, Goldstein H, Rasbash J (2001) Multiple Membership multiple classification (MMMC) models. Statistical Modelling 1: 103–124

Chiandotto B, Gola MM (1999) Questionario di base da utilizzare per l'attuazione di un programma per la valutazione della didattica da parte degli studenti. Osservatorio Nazionale per la Valutazione del Sistema Universitario, Roma

D'Esposito MR (ed) (2002) Valutazione della didattica e dei servizi nel sistema università. Atti della Giornata di Studio, Fisciano, 31 maggio 2002

Drasgow F (1986) Polychoric and polyserial correlations. In: Johnson NL, Kotz S (eds) Encyclopaedia of statistical sciences, vol 7. Wiley, New York

Emerson JD, Mosteller F, Youtz C (2000) Students can help improve college teaching: a review and an agenda for the statistics profession. In: Rao CR, Székely GJ (eds) Statistics for the 21st century: methodologies for applications of the future. Marcel Dekker, New York

Fielding A (1997) On scoring ordered classifications. British Journal of Mathematical and Statistical Psychology 50: 285–307

Fox J-P, Glas CAW (2001) Bayesian estimation of a multilevel IRT model using Gibbs sampling. Psychometrika 66: 269–286

Goldstein H (2003) Multilevel Statistical Models, 3rd. edition. Edward Arnold, London

Goldstein H, Healy MJR (1995) The graphical presentation of a collection of means. Journal of the Royal Statistical Society A, 158: 175–177

Goldstein H, Spiegelhalter DJ (1996) League tables and their limitations: statistical issues in comparisons of institutional performance. Journal of the Royal Statistical Society A, 159: 385–443

Grilli L, Rampichini C (2002a) Specification issues in stratified variance component ordinal response models. Statistical Modelling 2: 251–264

Grilli L, Rampichini C (2002b) Scomposizione della dispersione per variabili statistiche ordinali. Statistica anno LXII, n 1: 111–116

Grilli L, Rampichini C (2003) Alternative specifications of multivariate multilevel probit ordinal response models. Journal of Educational and Behavioural Statistics 28(1): 31–44

Hedeker D, Gibbons RD (1996) MIXOR: A computer program for mixed-effects ordinal regression analysis. Computer Methods and Programs in Biomedicine 49: 157–176

Keeney RL, Raiffa H (1976) Decision with multiple objectives: preferences and value tradeoffs. Wiley, New York

Leti G (1983) Statistica Descrittiva. Il Mulino, Bologna

Muthén BO (1994) Multilevel Covariance Structure Analysis. Sociological Methods & Research 22: 376–398

Pagani L, Seghieri C (2002) A Statistical Analysis of Teaching Effectiveness from Students' Point of View. In: Mrvar A, Ferligoj A (eds) Developments in Statistics . Metodološki zvezki 17, Ljubljana, FDV, pp 197–208

Piccarreta R (2001) A new measure of nominal-ordinal association. Journal of Applied Statistics 28: 107–120

Rabe-Hesketh S, Skrondal A, Pickles A (2004) Generalized Multilevel structural equation modelling. Psychometrika (in press)

Rasbash J, Goldstein H (1994) Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. J. Educational and Behavioural Statistics 19: 337–350

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

Snijders TAB, Bosker RJ (1999) Multilevel Analysis. An introduction to basic and advanced multilevel modelling. Sage, Londra

Spencer A, Fielding A (1998) Comparison of modelling strategies for value added analyses of educational data. In: Marx B, Friedl H (eds) Statistical Modelling: Proceeding of the 13th International Workshop on Statistical Modelling. Louisiana State University, Baton Rouge

Yager RR (2001) Dissonance. A measure of variability for ordinal random variables. International Journal of Uncertainty, Fuzziness and Knowledge based Systems 9: 39–53

Indagini via Internet sugli studenti: propensity score matching e stime da campioni non probabilistici

Da: Pratesi M. (2004) Indagini via Internet sugli studenti: propensity score matching e stime da campioni non probabilistici, in Strategie metodologiche per lo studio della transizione Università-lavoro, a cura di E. Aureli Cutillo, pp. 297-312, Cleup, Padova

Indagini via Internet sugli studenti universitari: propensity score matching e stime da campioni non probabilistici

Monica Pratesi[1]
Dipartimento di Statistica e di Matematica Applicata all'Economia
Facoltà di Economia, Università degli Studi di Pisa
Via Cosimo Ridolfi, 10 - I-56124 Pisa
E-mail: m.pratesi@ec.unipi.it

Riassunto. In molti casi le indagini via Internet sugli studenti si avvalgono delle risposte ottenute su un campione auto-selezionato di rispondenti che, spontaneamente, compilano il questionario incontrato durante la navigazione in rete. Anche in presenza di tassi di risposta soddisfacenti i risultati sono estendibili alla popolazione da cui proviene il campione soltanto a prezzo di ipotesi riduttive e spesso non verificabili sul comportamento di chi non ha risposto. In questo lavoro si propone di correggere la distorsione da auto-selezione tramite la tecnica di *propensity score matching*. Viene presentato uno studio di simulazione nel quale si stimano alcuni parametri descrittivi (totale, media e scarto quadratico medio) della popolazione obiettivo associando alle stime una valutazione del loro errore quadratico medio empirico. Si mostra anche come la distorsione da auto-selezione possa essere corretta condizionatamente alla specificazione di un opportuno modello di stima per i *propensity scores*.
Parole chiave: Indagini via Internet, distorsione da auto-selezione, propensity score matching

1. Introduzione

La rilevazione di dati utili per la valutazione del processo formativo universitario è tradizionalmente condotta su campioni probabilistici di studenti e/o laureati

[1] Il presente lavoro è stato finanziato nell'ambito del PRIN 2002, cofinanziato dal MIUR "Transizioni Università-lavoro e valorizzazione delle competenze professionali dei laureati: modelli e metodi di analisi multidimensionali delle determinanti". Coordinatore nazionale è L. Fabbris, coordinatore del gruppo di Firenze è il prof. B.Chiandotto (titolo del progetto dell'unità di ricerca locale "Valutazione del processo formativo universitario, sbocchi professionali e pianificazione dei percorsi formativi: modelli e metodi").

intervistati per telefono, oppure durante una visita dell'intervistatore (intervista faccia a faccia) o anche lasciati liberi di compilare autonomamente il questionario ricevuto per posta.

Recentemente, si sono sperimentate anche nuove metodologie che si avvalgono di Internet per la somministrazione del questionario e l'auto-intervista. In particolare, l'Ateneo di Pisa ha promosso negli ultimi mesi la rilevazione via Internet sulla qualità della vita degli studenti universitari a Pisa (http://www.studenti.unipi.it) e già da tempo, nello stesso Ateneo, è attivo un servizio di raccolta informazioni sul curriculum dei laureati nell'Ateneo che faciliti il loro inserimento nel mondo del lavoro (http://www.diogenet.net).

Risultati significativi sulle potenzialità ed i limiti di Internet per le rilevazioni statistiche sugli studenti e/o laureati erano già stati segnalati per l'Ateneo di Padova e Firenze su temi riguardanti la qualità della didattica universitaria e la rilevazione sull'inserimento lavorativo e professionale dei laureati e dei diplomati (Fabbris e Giusti, 2001) [2].

Nel caso di rilevazione via Internet sulla qualità della didattica universitaria è emerso che il principale problema da affrontare è la volontarietà della partecipazione. E cioè si è notato che, mentre è relativamente facile svolgere la rilevazione in aula con questionario cartaceo, risulta problematico imporre e regolamentare l'accesso ai computer per la rilevazione attraverso il questionario su *web*. Tra l'altro, l'uso di postazioni Internet non strutturate, in genere Personal Computer disponibili a domicilio allo studente, può risultare non sempre facile per studenti di facoltà caratterizzate da una minore diffusione dell'informatica nell'offerta didattica e nelle competenze di base degli iscritti.

Per quanto riguarda le rilevazioni sull'inserimento lavorativo e professionale dei laureati e dei diplomati, si è notato invece che l'uso di Internet avrebbe consentirebbe un considerevole risparmio economico rispetto alla rilevazione telefonica con sistema CATI (Computer Assisted Telephone Interviewing) e avrebbe dato la possibilità di far aggiornare direttamente al laureato le modifiche della propria situazione professionale. Tuttavia, la conclusione dello studio citato per Padova e Firenze era che la rilevazione telefonica risultava essere il canale preferibile anche per questo tipo di rilevazioni per la sistematicità e la ricchezza delle informazioni ottenibili e per la maggiore concentrazione temporale dei periodi di rilevazione, soprattutto per la possibilità di svolgere la rilevazione sui laureati su base campionaria. In sostanza anche in questo caso si segnalava come principale problema la volontarietà della parte-

[2] Gli autori si riferiscono alle rilevazioni via Internet indicandole come rilevazioni effettuate con la tecnica *web*-CASI. Infatti, nel caso di rilevazione via Internet, il questionario elettronico è composto da più pagine *web* e la rilevazione è condotta con un sistema CASI, cioè di tipo Computer Assisted Self Interviewing.

cipazione all'indagine via Internet e la conseguente distorsione da auto-selezione da cui sarebbero affetti i risultati.

L'obiettivo di questo contributo è proprio quello di analizzare le modalità con cui la distorsione da auto selezione può essere affrontata in questo tipo di indagini. Negli ultimi anni vari approcci di natura non sperimentale sono stati sviluppati per correggere la distorsione da auto-selezione. Una di queste tecniche, conosciuta come 'Propensity Score Matching' e nata nell'ambito degli studi medici di tipo caso-controllo, è dovuta a Rosenbaum e Rubin (1983) ed è stata applicata recentemente anche a dati rilevati con indagini via Internet (Couper 2000; Terhanian et al. 2001; Varedian e Forsman 2002; Schonlau et al., 2002, Biffignandi e Pratesi, 2003).

L'argomento di questo contributo è proprio l'ulteriore esplorazione delle potenzialità di questo approccio nel giungere ad inferenze dai dati raccolti con campioni non probabilistici affetti da auto-selezione. L'attenzione è diretta sia al caso in cui la popolazione obiettivo dell'indagine sia costituita esclusivamente da studenti/laureati utenti Internet, sia al caso in cui essa comprenda anche individui senza accesso alla rete.

Nell'economia del lavoro non si considera la possibilità, per altro ancora remota in questo contesto applicativo, di poter contare su liste di utenti Internet dalle quali estrarre un campione probabilistico, ma ci si limita a considerare il caso, assai più comune, nel quale il questionario *web* sia incontrato durante la navigazione e compilato per scelta volontaria dallo studente/laureato.

Diamo anche per risolto il problema della identificazione del rispondente come membro effettivo della popolazione obiettivo d'interesse: riteniamo infatti che tale problema, nel caso di popolazioni di iscritti all'Università o laureati sia tecnicamente risolvibile tramite l'inserimento della matricola e la garanzia successiva dell'anonimato delle risposte fornite. A titolo di esempio si può far riferimento al caso del questionario sulla condizione studentesca sul sito citato http://www.studenti.unipi.it. Esso è rigorosamente anonimo. Utilizza infatti un codice di accesso con matricola criptata per fare in modo che venga compilato esclusivamente dagli studenti e che ciascuno lo compili una sola volta.

Nel seguito, indicheremo le indagini via Internet condotte su un campione non probabilistico di studenti ed affetto da auto-selezione con la dizione indagini via Web, o, più semplicemente, indagini Web. Il sottoinsieme di risposte dei partecipanti all'indagine viene denominato campione Web. La struttura secondo la quale sono presentati i risultati del lavoro è la seguente. Nel paragrafo 2 si esamina il contesto originario nel quale è nata la metodologia nota come *propensity matching*, le assunzioni sottostanti la sua validità e le circostanze nelle quali, nel contesto dell'indagine Web, quelle ipotesi possono essere accettate. Nel paragrafo si indicano anche le possibili popolazioni obiettivo dell'inferenza. Il metodo proposto per la correzione della

distorsione da auto-selezione è descritto nel paragrafo 3, nel quale si discute il processo di generazione del campione Web e gli stimatori di alcuni parametri descrittivi della popolazione obiettivo (media e deviazione standard). Le proprietà statistiche degli stimatori proposti sono studiate attraverso un esperimento Monte Carlo. Alla descrizione delle procedure di simulazione e alla discussione dei risultati sono dedicati i paragrafi 4 e 5.

2. L'auto-selezione ed il matching tramite propensity scores

Le tecniche di *propensity scores matching*, nate come abbiamo detto nell'ambito di studi medici di tipo caso-controllo, hanno come obiettivo la trasformazione di risultati ottenuti su casi auto-selezionati in risultati che godono delle stesse proprietà di quelli provenienti da studi caso-controllo di tipo randomizzato.

Si dimostra infatti che è possibile simulare l'assegnazione casuale di casi e controlli, combinando i casi a disposizione con controlli *post hoc* individuati tramite un singolo indice (chiamato *propensity*), che esprima la probabilità di essere esposti al trattamento. Tale probabilità dovrebbe tenere conto simultaneamente di tutte le variabili che si ritiene possano influenzare il paragone tra casi e controlli (dette variabili pre-intervento). Una volta effettuato il *matching* tra casi e controlli, è possibile stimare correttamente l'esito del trattamento sui casi considerati.

Il metodo di *matching* (appaiamento) è fondato su alcune ipotesi restrittive (Rosenbaum e Rubin, 1983, Rubin e Zanutto, 2002) che non hanno però impedito la sua vasta applicazione anche in campi lontani da quello di origine. Infatti, il suo uso è attualmente esteso alle politiche di valutazione del mercato del lavoro. Da citare sono le applicazioni negli anni'90 (Dehija e Wahba, 1999), alcuni studi molto critici sull'applicabilità del metodo al contesto economico (Smith e Todd, 2000; Heckman et al. 1998) e recenti contributi metodologici che estendono la teoria inizialmente proposta all'individuazione di opportuni controlli anche nel caso di trattamento multiplo (Imbens, 2000).

Negli ultimi tre anni, tali tecniche hanno attratto l'attenzione anche di studiosi dediti all'impostazione delle indagini statistiche ed all'analisi dei dati con queste raccolti. In particolare il *matching* tramite *propensity* ha fornito una possibile soluzione all'analisi dei dati ottenuti su campioni autoselezionati di utenti di Internet (Schonlau et al., 2002, Biffignandi e Pratesi, 2003).

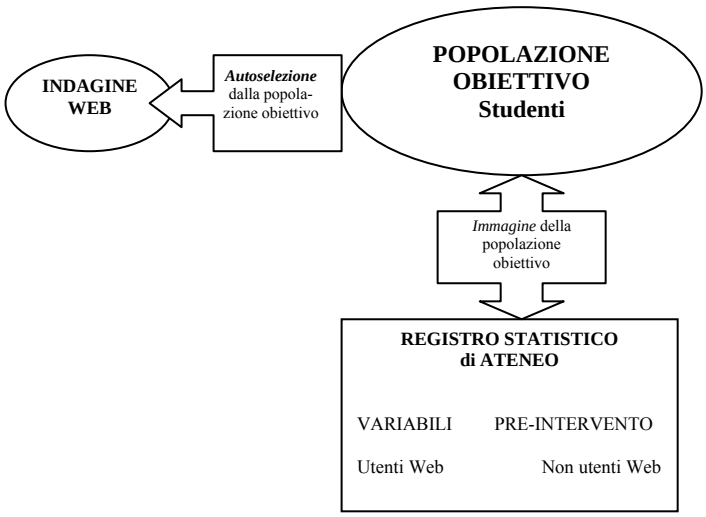
In altre parole, la tecnica è stata usata per costruire un gruppo di controllo per i rispondenti all'indagine (casi). L'assegnazione dei casi al trattamento (l'indagine via Web) non è infatti frutto di un assegnazione casuale controllata dal ricercatore: il campione è auto-selezionato e le risposte sono verosimilmente affette da distorsione.

Tale distorsione può essere corretta ponderando le stime ottenute dall'indagine tramite gli stessi *propensity scores* usati per il *matching*.

Ma andiamo per ordine. I rispondenti sono appaiati con opportuni controlli scelti tra i non rispondenti. I controlli *post hoc* sono individuati a parità di *propensity score*. Vale a dire ciascun rispondente è appaiato con il non rispondente che ha lo stesso propensity score (*propensity scores matching*). I propensity scores esprimono la probabilità di rispondere all'indagine Web. Tale probabilità è stimata in funzione di tutte le variabili osservabili che si ritiene possano influenzare il paragone tra chi ha risposto all'indagine (i casi) e coloro che invece non hanno partecipato (i controlli). I casi appaiati (che possono essere anche in numero minore rispetto ai rispondenti originari) sono poi ponderati affinchè le risposte da essi fornite ai quesiti dell'indagine siano estese anche a coloro che fanno parte del gruppo dei controlli.

Il contesto applicativo a cui ci riferiamo può essere descritto semplicemente facendo ricorso allo schema presentato in Figura 1. Nella figura è rappresentata una situazione comune a molti contesti applicativi di tipo socio-economico e in particolare molto frequente nel caso di indagini Web su studenti universitari e/o laureati. In questo caso, infatti, i singoli Atenei sono obbligati per finalità certificative a mantenere registri statistici informatizzati riferiti alla popolazione obiettivo.

Figura 1: *Rapporto tra popolazione obiettivo, registro statistico ed indagine Web*



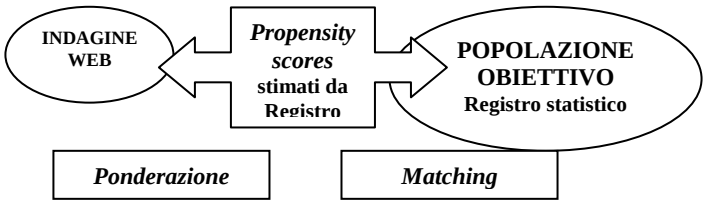
In buona sostanza, non si possiede una lista completa ed aggiornata degli studenti utenti Internet, ma si è in grado di definire concettualmente la popolazione obiettivo cui essi appartengono.

Essa inoltre è spesso rappresentata da registri statistici attivi presso gli uffici statistici di Ateneo che ne curano il livello qualitativo e dai quali si possono avere molte informazioni sulle cosiddette variabili pre-intervento e sulla possibilità di accedere ad Internet da parte dei membri della popolazione[3].

L'esistenza del registro statistico permette la definizione e la stima dei propensity scores per la popolazione da cui proviene il gruppo dei casi, cioè l'insieme dei partecipanti all'indagine Web diretta alla popolazione obiettivo rappresentata dal Registro.

La figura 2 schematizza le funzioni assolte dai propensity scores come strumento prima di matching e poi di ponderazione dei dati raccolti. Una volta effettuato il matching, i propensity sono di nuovo usati per la ponderazione dei risultati raccolti con l'indagine secondo la metodologia proposta nel paragrafo 3 al fine di stimare alcuni parametri descrittivi della popolazione obiettivo.

Figura 2: *Funzione dei propensity scores stimati da Registro*



3. Il metodo di stima proposto

Nella nostra prospettiva, il campione Web è il risultato di un processo di selezione spontanea dalla popolazione di dimensione N che può essere descritto a livello individuale dall'indicatore R_i ($R_i = 1$ se l'individuo partecipa all'indagine, $R_i = 0$ altrimenti, $i = 1,\ldots,N$). Il valore atteso dell'indicatore è la probabilità di autoselezione dell'individuo $E(R_i) = p_i$. La dimensione del campione è una variabile casuale $n = \sum_i^N R_i$.

[3] Citiamo come caso noto quello dell'archivio dei laureati dell'Ateneo di Pisa aderenti a Diogenet che è corredato anche dell'indirizzo di posta elettronica dei laureati.

Se p_i fosse nota, il problema inferenziale potrebbe essere risolto usando gli strumenti tradizionali di stima da campioni casuali per i quali la probabilità di inclusione nel campione è nota per ogni elemento della popolazione obiettivo (Särndal et al., 1992). Tale probabilità non è però nota e deve essere stimata. Una possibile stima di p_i è costituita dai propensity score, che sono nel nostro caso buone proxy per la probabilità di partecipare all'indagine Web.

L'idea è quella di stimare la probabilità individuale p_i (*propensity score*) usando un modello di regressione logistica, nel quale la variabile risposta sia R_i e le variabili pre-intervento siano i regressori. I parametri stimati sono poi usati per stimare i *propensity scores* del campione Web, usando le stesse variabili pre-intervento rilevate questa volta sul campione Web. Dopo aver appaiato il gruppo dei casi con i controlli della popolazione tramite il *matching* che ha come chiave i *propensity scores* stimati, i dati del campione Web sono opportunamente ponderati per stimare alcuni parametri descrittivi (media e deviazione standard) della variabile di studio nella popolazione.

I dati necessari

Il metodo di stima dei propensity necessita di dati sulle variabili pre-intervento riferiti sia al campione Web sia ai membri della popolazione obiettivo che non hanno partecipato all'indagine.

Tali dati possono essere ottenuti da una precedente indagine campionaria rappresentativa della popolazione oppure da un registro statistico. Ci sono esempi in cui i propensity scores sono stimati a partire da dati ottenuti con un'indagine campionaria telefonica sulla popolazione (Varedian and Forsman, 2002). In questo lavoro immaginiamo di stimare il modello per i propensity scores con il supporto di un registro statistico riferito all'intera popolazione piuttosto che sui dati di un' indagine campionaria. I motivi sono legati alla natura stessa dei dati campionari. A causa delle mancate risposte, o di sovracampionamento in particolari sottogruppi, i dati dell'indagine possono non essere direttamente rappresentativi dell'intera popolazione. E' pratica comune ponderare per eliminare la distorsione da non-risposta e gli effetti del sovracampionamento. Non è ancora chiaro quale debba essere il ruolo di questi pesi nelle operazioni precedenti il matching, cioè nella stima del modello per i propensity, e nelle operazioni successive al matching. Alcuni sostengono che i pesi debbano essere usati solo nell'adattamento del modello per i propensity, altri ritengono che una volta effettuato l'appaiamento debbano essere ponderati sia i casi sia i controlli (Bryson A., 2001; Green et al., 2001). La discussione è ancora aperta, per questo motivo preferiamo evitare i problemi precedenti, ed agire come se avessimo a disposizione un registro statistico aggiornato della popolazione target. Questo presupposto può essere accettato in molti campi di applicazione (Biffignandi e Pratesi, 2002, Pratesi et al, 2004).

La stima dei parametri descrittivi della popolazione

Per completezza proponiamo due diversi approcci per la stima dei parametri descrittivi. Entrambi fanno uso di pesi basati sui propensity: 1) il primo è detto di "Horvitz e Thompson", perchè produce uno stimatore che richiama lo stimatore tradizionale di Horvitz e Thompson, 2) il secondo è presentato come approccio di "Pseudo stratificazione a posteriori" per distinguerlo dalla post-stratificazione tradizionale. Parliamo di pseudo stratificazione a posteriori perché, nel nostro caso, la dimensione degli strati nella popolazione e nel campione sono il risultato di un metodo di stima e non di un conteggio come nel caso tradizionale.

1. L'approccio di tipo "Horvitz-Thompson". La probabilità di partecipare all'indagine è assimilata ai propensity scores stimati sulla popolazione, $\hat{p}_i = P(R_i = 1 \mid X_i)$, dato il vettore X_i di variabili pre-intervento. Gli scores stimati $\hat{p}_i$ sono usati per trovare le unità di controllo nella popolazione appaiabili allo stesso livello di probabilità: i controlli per il gruppo dei casi sono quelle unità per le quali $1 - \hat{p}_i = P(R_i = 0 \mid X_i) = \hat{p}_i = P(R_i = 1 \mid X_i)$. Dopo il matching, le unità del campione Web che trovano un adeguato il controllo sono ponderate con pesi pari direttamente al reciproco dei propensity scores, i pesi sono cioè $w_i = 1/\hat{p}_i$. Il matching sui propensity scores riduce gruppo dei casi all'insieme per il quale sono stati trovati controlli adeguati nella popolazione. Questo assicura che per il gruppo dei casi usato nella stima, date le variabili pre-intervento (i regressori nel modello per i propensity scores) sia verificato il presupposto di indipendenza condizionale. Gli stimatori proposti per il totale, la media e la deviazione standard della popolazione sono riassunti in Tabella 1. La performance di questi stimatori è valutata attraverso lo studio di simulazione descritto nel paragrafo 4.

2. L'approccio di Pseudo Post-stratificazione. Alla base di questo approccio c'è l'idea di suddividere in strati popolazione e campione secondo la distribuzione di frequenza dei propensity scores stimati. La popolazione originale è stratificata, non distinguendo i casi dai controlli, in base alla distribuzione di frequenza dei propensity scores stimati. Indichiamo con N_h la dimensione di uno strato nella popolazione. La dimensione della popolazione è ottenuta come $N = \sum_h N_h$. Dopo il matching, possiamo stratificare anche il campione. Indichiamo con n_h la dimensione dell'analogo strato nel campione. Come nella post-stratificazione tradizionale, l'appartenenza delle unità del campione agli strati è determinata dopo la selezione del campione stesso. Gli stimatori proposti per i parametri descrittivi della popolazione sono riassunti in Tabella 1. Gli stimatori sono calco-

lati per ciascuno dei campioni generati nello studio di simulazione descritto nel paragrafo 4.

Tabella 1. *Gli stimatori proposti*

	Stimatore
Parametro	*Approccio tipo Horvitz-Thompson*
Media	$\bar{y}_p = \sum_i y_i w_i / \sum_i w_i$
Standard deviation	$\sigma(y)_p = \sqrt{\sum_i \left(y_i - \bar{y}_p\right)^2 w_i / \sum_i w_i}$
Parametro	*Approccio di Pseudo post-stratificazione*
Media	$\bar{y}_{ps} = \sum_h N_h / N \sum_i y_{ih} / n_h$
Standard deviation	$\sigma(y)_{ps} = \sqrt{\sum_h N_h / n_h \sum_i \left(y_{ih} - \bar{y}_h\right)^2 / N}$

4. Lo studio di simulazione

La performance relativa dei metodi di aggiustamento per ponderazione esaminati è stata valutata tramite tre esperimenti di tipo Monte Carlo su tre diverse popolazioni simulate di studenti e/o laureati. La popolazione indicata come *popolazione originaria* rappresenta una situazione dove la relazione tra la variabile di studio e le variabili di pre-intervento è di tipo lineare e la distribuzione delle variabili di pre-intervento è la stessa (anche se con momenti diversi) per gli utenti e i non utenti Web. Dalla popolazione originaria sono state ottenute popolazioni di tipo diverso, come descritto qui di seguito:

1. *Popolazione originaria*: la relazione tra la variabile di studio e le variabili di pre-intervento è di tipo lineare e la distribuzione delle variabili di pre-intervento è la stessa (anche se con momenti diversi) per gli utenti e i non utenti Web.
   Rispondenti Web:

$x_1 = N(6,4), x_2 = Gamma(5),\ x_3 = \exp(-t) + 3$

<u>*Non Rispondenti Web*</u>:

$x_1 = N(10,5), x_2 = Gamma(3),\ x_3 = \exp(-t)$

la variable di studio è una funzione lineare delle variabili osservate: $y = 0.2x_1 + 0.3x_2 + 0.5x_3 + 10u$ dove u ha una distribuzione uniforme nell'intervallo [0,1].

2. *Popolazione 1:* Rimuoviamo il presupposto di relazione lineare, considerando la variabile di studio come funzione non lineare delle variabili osservate, distribuite ancora una volta come descritto nel paragrafo 4.1: $y = 0.2/x_1 + 0.3x_2^2 + 0.5x_3 + 10u$ dove u ha una distribuzione uniforme nell'intervallo [0,1].
3. *Popolazione 2:* Le variabili pre-intervento si distribuiscono in modo diverso per utenti e non utenti Web:

   <u>*Rispondenti Web*</u>:

   $x_1 = N(6,4), x_2 = Gamma(5),\ x_3 = \exp(-t) + 3$

   <u>*Non Rispondenti Web*</u>:

   $x_1 = \log(N(10,5)), x_2 = Gamma(3),\ x_3 = \exp(-t)$

   la variable di studio è ancora una funzione lineare delle variabili osservate:

   $y = 0.2x_1 + 0.3x_2 + 0.5x_3 + 10u$

Il processo di autoselezione che ha generato il gruppo dei casi è stato simulato selezionando da ogni popolazione Web un campione di navigatori con probabilità di autoselezione individuale uguale al propensity score stimato sulla popolazione.

Il processo di generazione del campione Web è stato ripetuto 100 volte, ottenendo di volta in volta gruppi di casi di dimensione n=20. La stima dei propensity scores è stata ottenuta sulla popolazione in base ad una regressione logistica che avesse come regressori tutte le variabili pre-intervento. La ricerca di controlli adeguati è stata effettuata basandosi su propensity con 3 cifre decimali uguali, poi 2 cifre ed, infine, 1 sola cifra decimale uguale a quella dei casi (Parsons, 2001).

La completezza del matching, e la qualità delle coppie appaiate sono descritte in Tabella 2, riassumendo i risultati dei 100 campioni (n=20) per la popolazione originaria. In tabella si mostra solo la media di x_1 per i casi ed i controlli appaiati.

Tabella 2. *Sintesi del matching*

Algoritmo	% casi appaiati	Casi x_1 (Media±ds)	Controlli x_1 (Media±ds)	Propensity Score delle coppie: Differenza media assoluta
Matching su 3 cifre	40%	5.54± 2.70	6.95 ± 3.05	.000813
Matching su 2 cifre	80%	5.56± 1.53	7.34 ± 1.48	.0014
Matching su 1 cifra	90%	5.22 ± 1.45	7.42 ± 1.28	.0232
Popolazione		**5.87 ± 3.95**	**9.97 ± 5.00**	

Dalla tabella si vede che imponendo un criterio di matching restrittivo la percentuale di coppie appaiate diminuisce ma la qualità delle coppie appaiate migliora. Tale qualità si nota dalla diminuzione della differenza tra la distribuzione della variabile x_1 nei casi e nei controlli appaiati. Ovviamente, imponendo il matching su un numero maggiore di cifre, la differenza media assoluta nei propensity score delle coppie appaiate diminuisce.

I risultati presentati nel paragrafo 5 sono basati tutti su unità appaiate sulla prima cifra decimale: il livello di completezza del matching è alto mentre il bilanciamento tra casi e controlli ottenuto sulla base del matching non è il migliore possibile anche se la differenza media nei propensity delle coppie è, a nostro avviso, del tutto accettabile (0.0232)[4].

Le figure 3 e 4 mostrano la distribuzione di propensity score stimati sui casi e sui controlli per la popolazione originaria. Sul grafico sono evidenziate le classi di valori per i quali il matching tra casi e controlli è possibile.

[4] Dalla Tabella 2 si nota che le deviazioni standard sono decisamente più elevate nel matching a 3 cifre rispetto a quello a 2 e a 1 cifra: al crescere del numero delle cifre utili per il matching le medie dei casi e dei controlli si avvicinano, come del resto era atteso, ma il numero di coppie formate diminuisce e la variabilità nel gruppo ottenuto aumenta. In questa situazione, appare preferibile lavorare su un gruppo di casi e controlli meno vicini in media, ma di dimensione maggiore e con osservazioni che si addensano maggiormente attorno ai valori medi di gruppo. Anche per questi motivi si è deciso di effettuare la simulazione su unità appaiate soltanto sulla prima cifra decimale.

Figura 3: *Casi: Propensity scores stimati sulla popolazione originaria*

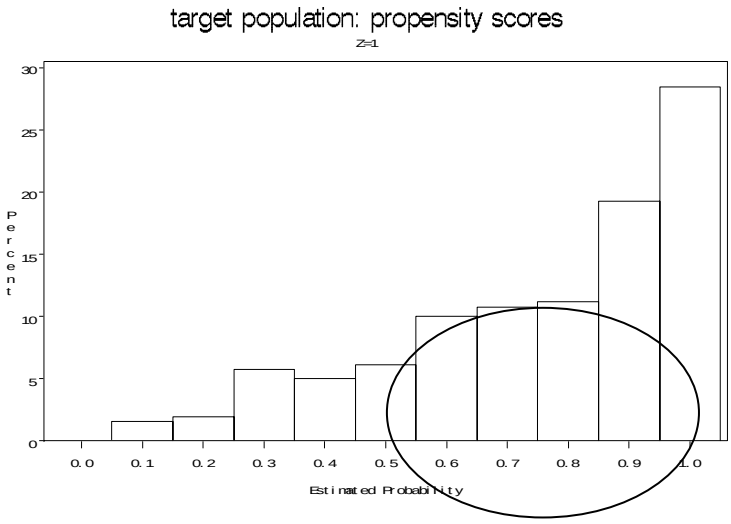
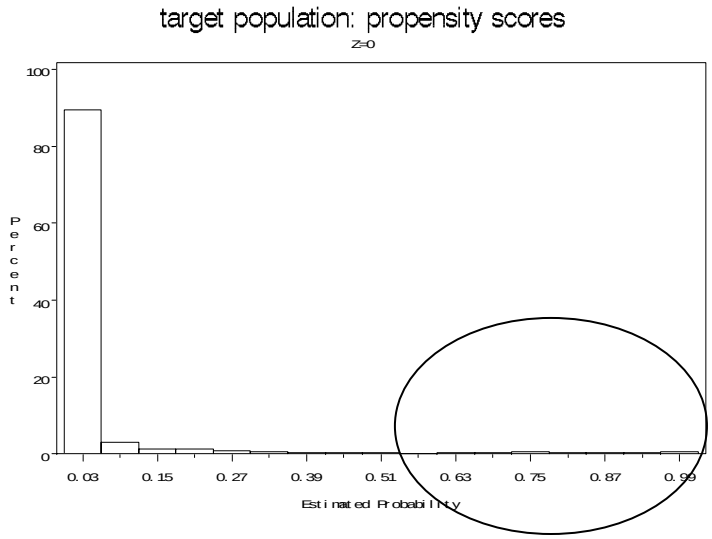


Figura 4: *Controlli: Propensity scores stimati sulla popolazione originaria*



5. Commento dei risultati e alcune considerazioni conclusive

I risultati presentati nelle Tabelle 3 e 4 considerano come target l'intera popolazione. L'attenzione è sulla stima della media e della deviazione standard della variabile di studio effettuata usando gli stimatori presentati in Tabella 1: stimatori del tipo "Horvitz-Thompson" e stimatori di "Pseudo stratificazione a posteriori".
I risultati sono discussi sulla base della distorsione relativa empirica (D.R.) degli stimatori ottenuta sui 100 campioni Web generati come descritto nel paragrafo 4: in formula $D.R. = (T - \theta) \Big/ \sqrt{EQM(T)}$, dove $EQM(T)$ è l'errore medio quadrato empirico dello stimatore T e θ è il parametro di interesse.

Per tutti i modelli di popolazione i risultati migliori sono quelli dello stimatore pseudo post-stratificato: la distorsione relativa di tale stimatore è sempre più bassa di quella dello stimatore H.T. per tutti i parametri descrittivi considerati (vedi tabelle 3 e 4).

Tabella 3*: Target: Popolazione complessiva: Distorsione Relativa percentuale degli stimatori*

Stimatori	**Popolazione originaria**		**Popolazione 1**	
	H. T.	*Pseudo post*	*H. T.*	*Pseudo post*
$\bar{y}_{ps}$	40,34	14,54	38,46	22,83
$\sigma(y)_{ps}$	-79,31	-62,24	-26,32	-21,28

Tabella 4*: Target: Popolazione complessiva: Distorsione Relativa percentuale degli stimatori*

Stimatori	**Populazione originaria**		**Popolazione 2**	
	H. T.	*Pseudo post*	*H. T.*	*Pseudo post*
$\bar{y}_{ps}$	40,34	14,54	14,44	13,97
$\sigma(y)_{ps}$	-79,31	-62,24	21,17	22,75

La performance della stimatore Pseudo post stratificato non rimane la stessa quando, nel modello lineare, le variabili pre-intervento hanno distribuzione diversa nel gruppo dei rispondenti Web ed in quello dei non rispondenti Web (Popolazione 2). In

questo caso le differenze tra stimatore Pseudo post stratificato e stimatore H. T. non sono così rilevanti (vedi Tabella 4). Il comportamento dei metodi di aggiustamento studiati sembra quindi risentire delle caratteristiche del modello che genera la popolazione obiettivo. Comunque, anche in situazioni dove i casi ed i controlli sono ben diversi in base alle variabili pre-intervento (Popolazione 2, lo stimatore di Pseudo post stratificazione fornisce risultati accettabili in confronto a quello di H.T..

In altre parole, i risultati ottenuti suggeriscono che, nel caso di propensity scores stimati con regressione logistica, è preferibile usare classi di propensity (o post strati sui propensity) piuttosto che ponderare direttamente con i propensity stimati, come si fa quando si usano stimatori dei tipo Horvitz-Thompson.

Dai risultati presentati, infatti, pare che la ponderazione del tipo H. T. non sia sufficiente a compensare la distorsione da autoselezione principalmente perchè il livello medio dei propensity sul campione è alto: questo produce una bassa inflazione dei totali campionari ed una maggiore distorsione relativa dello stimatore. Inoltre, può insorgere un altro problema con conseguenze opposte: in mancanza di ulteriori aggiustamenti le probabilità stimate possono essere molto piccole ed instabili ed il loro reciproco (il peso) può essere molto alto. Ciò avviene proprio a causa del metodo di stima usato per il modello logistico (minimi quadrati ponderati iterativi); la classificazione dei propensity in post strati crea invece categorie discrete di pesi che generano aggiustamenti più stabili, non affetti dalla piccola dimensione di alcuni propensity scores stimati.

In conclusione i nostri risultati mostrano come il matching tramite propensity combinato con la strategia di stima che noi chiamiamo di Pseudo post stratificazione sia una soluzione promettente per l'aggiustamento della distorsione da auto selezione che affligge in molti casi le indagini Web. Si mostra infatti come la distorsione da auto-selezione possa essere limitata condizionatamente alla specificazione di un opportuno modello di stima per i *propensity scores*.

Riferimenti bibliografici

BIFFIGNANDI S., PRATESI M. (2002), Internet surveys: the role of time in italian firms response behaviour, *Journal of Research in Official Statistics*, Volume 5, number **2**: pp.53-66.

BIFFIGNANDI S., PRATESI M. (2003), Potentiality of propensity scores Methods in Weighting for Web Surveys: a simulation study., Quaderni del DMSIA, Università degli Studi di Bergamo.

BRYSON A. (2001), *The Union Membership Wage Premium: An Analysis Using Propensity Score Matching*, Centre for Economic Performance, Working Paper n. 1160, London School of Economics.

COUPER M. P. (2000), Web Surveys, A review of Issues and Approaches, *Public Opinion Quarterly*, vol. 64, pp. 464-494.
DEHEJIA R. H., WAHBA S. (1999), Causal Effects in Nonexperimental Studies: re-evaluating the evaluation of training programs, *Journal of the American Statistical Association,* vol. 94, n.448, pp.1053-1062
FABBRIS L., GIUSTI A., (2001), Il progetto EXPERTUM: Sperimentazione di sistemi *computer assisted* per la rilevazione della valutazione della didattica universitaria da parte degli studenti e dell'inserimento lavorativo e professionale dei laureati e dei diplomati, *Atti del convegno intermedio della S.I.S 2001* "Processi e metodi statistici di valutazione", Roma 4-6 /6/2001.
GREEN H., CONNOLY H., MARSH A., BRYSON A. (2001), *The Longer-term Effects of Voluntary Participation in ONE*, Department of Work and Pensions, Research report Number 149.
HECKMAN J., ICHIMURA H., SMITH J., TODD P. (1998), Characterizing selection bias using experimental data, *Econometrica*, 66 (**5**), pp. 1017-1098
IMBENS G. (2000), The Role of Propensity Score in Estimating Dose-Response Functions, *Biometrika,* vol. 87, pp.706-710
PARSONS LS. (2001), Reducing Bias in a Propensity Score Matched-Pair Sample Using Greedy Matching Techniques, Proceedings of the Twenty-Sixth Annual SAS Users Group International Conference, Cary, NC: SAS Institute Inc.
PRATESI M., LOZAR MANFREDA K., BIFFIGNANDI S., VEHOVAR V. (2004), List based Web Surveys: qualità, timeliness and non-response in the steps of the participation flow, *Journal of Official Statistics*, *forthcoming*.
ROSENBAUM, P.R, AND D.B. RUBIN, (1983), "The Central Role of the Propensity Score in Observational studies for Causal Effect", *Biometrik*a, vol 70, pp. 41-55.
RUBIN, D. AND E. ZANUTTO (2001), Using matched substitutes to adjust for nonignorable nonresponse through multiple imputation, in *Survey Nonresponse* (R.GROVES, D. DILLMAN, J. ELTINGE AND R. LITTLE, eds). New York John Wiley, Chapter 26, pp.389-402.
SÄRNDAL C. E., SWENSSON B., WRETMAN J. (1992), *Model assisted survey sampling*, Springer Verlag
SMITH J., TODD P. (2000), *Does matching overcome LaLonde critique of nonexperimental estimators?,* mimeo, downloadable from http://www.bsos.umd.edu/econ/jsmith/Papers.html.
SCHONLAU, M., FRICKER R., ELLIOTT, M. (2002), Evaluation of Web survey Methodology, RAND, Santa Monica, CA. 2002.
TAYLOR H. (2000), Does Internet research work? Comparing Online survey Result with telephone Survey, *International Journal of Market Research*, 42 (**1**) 58-63
TERHANIAN G., R. SMITH, J. BREMER, R.K. THOMAS (2001), Exploiting Analytical Advances: Minimizing the biases Associated with Internet–Based Surveys of Non–Random Samples, ARF/ESOMAR: *Worldwide Online Measurement*, ESOMAR Publication Services, vol. 248, pp. 247-272.
VAREDIAN M., FORSMAN G. (2002), *Comparing propensity score weighting with other weighting methods: A case study on Web data*, unpublished report.