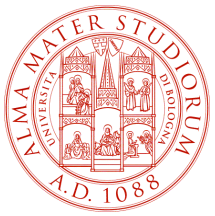


Alma Mater Studiorum - Università di Bologna

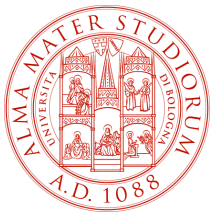
Scuola di Dottorato in Scienze Economiche e Statistiche

Dottorato di Ricerca in
Metodologia Statistica per la Ricerca Scientifica
XXI Ciclo

Gli esiti scolastici nelle scuole di secondo grado di Bologna: un'applicazione della teoria dei modelli a curva latente

Maria Serena Borgia

Dipartimento di Scienze Statistiche "P. Fortunati"
Marzo 2011



Alma Mater Studiorum - Università di Bologna

Scuola di Dottorato in Scienze Economiche e Statistiche

Dottorato di Ricerca in
Metodologia Statistica per la Ricerca Scientifica
XXI Ciclo

Gli esiti scolastici nelle scuole di secondo grado di Bologna: un'applicazione della teoria dei modelli a curva latente

Maria Serena Borgia

Coordinatore:
Prof.ssa Daniela Cocchi

Tutor:
Prof.ssa Stefania Mignani

Settore disciplinare:
SECS-S/05

Dipartimento di Scienze Statistiche "P. Fortunati"
Marzo 2011

A Matilde e Damiano

Indice analitico

Introduzione

Gli amministratori del sistema scuola hanno la necessità di conoscere la realtà in cui operano ed in particolar modo quali sono gli effetti quantitativi dei loro interventi istituzionali. Un elemento di valutazione del sistema scolastico sta nel rendimento degli studenti. La misurazione di tale rendimento risulta essere non intuitiva, data la realtà di riferimento, molto variegata, e la difficoltà di reperimento di dati di base. Al fine di misurare gli esiti del percorso scolastico degli studenti, infatti, proprio nel presente lavoro si dimostra che i dati aggregati possono essere spesso fuorvianti, mentre soltanto dati individuali, sui singoli studenti e per tempi diversi, possono fornire un panorama completo.

L'informazione sulla ricaduta degli interventi istituzionali, ad esempio in quanto a offerta formativa (tipologie di insegnamenti, ore a disposizione, modalità con cui vengono svolte le lezioni, ecc.), è nella maggior parte dei casi carente, in quanto non è possibile disporre di dati completi sul rendimento scolastico, conseguente a tali interventi. Nella consapevolezza, infatti, della molteplicità dei fattori, anche a carattere individuale, che intervengono nel determinare l'esito formativo degli studenti, si ritiene comunque che questo possa essere migliorato anche attraverso interventi tesi a potenziare l'offerta formativa. Ma come si può quantificare il miglioramento apportato senza disporre di una misura dell'esito scolastico? Anche se in questa sede non si è avuta la possibilità di effettuare una "doppia misurazione", quindi prima e dopo un intervento istituzionale, al fine di saggiare gli effetti di quest'ultimo, si è intrapresa un'analisi che vuole essere un punto di partenza per una misurazione accurata e soprattutto quantitativa, piuttosto che qualitativa, dei percorsi scolastici degli studenti. Partendo dalle seppur limitate informazioni disponibili, si è cercata la metodologia più appropriata per lo studio del comportamento nel tempo degli studenti in quanto ad esito scolastico. L'obiettivo è quello, infatti, di cercare un modello che stimi l'esito scolastico a partire da informazioni note sugli studenti e sul contesto in cui essi si trovano. In particolare, si è alla ricerca di un modello che risulti applicabile anche a realtà differenti da quella in questa sede utilizzata e che risulti ampliabile in virtù di maggiori informazioni note per gli individui oggetto di studio.

Nella presente analisi si ha l'eccezionale disponibilità di informazioni di tipo longitudinale per un contingente elevato di individui, grazie all'accordo tra la Facoltà di Scienze Statistiche e la Provincia di Bologna che ha consentito la disponibilità di dati provenienti dal database sull'obbligo formativo (Osservatorio sulla scolarità), detenuto dalla Provincia ai sensi della normativa vigente (legge 144/99).

I dati utilizzati provengono, quindi, da fonte istituzionale e sono da quest'ultima detenuti in base alla normativa: il presente studio ha consentito l'utilizzo di una banca dati già esistente, senza il bisogno di una rilevazione ad hoc, con il doppio vantaggio di trarre informazioni aggiuntive sulla base di dati già disponibili e di evitare i problemi connessi ad una rilevazione diretta (mancate risposte, errori di rilevazione, errori di misurazione, errori di risposta, ecc.).

Al fine di ottenere un modello che potesse al meglio spiegare e riprodurre l'andamento degli esiti scolastici, in termini di promozione da una classe alla successiva e tenendo conto del percorso scolastico nel tempo, si è in questa sede scelto di utilizzare il modello a curva latente condizionato; le variabili esplicative sono state selezionate tra le informazioni disponibili nel dataset, relative al contingente di studenti entrati nella stima del modello (provenienti dall'anagrafe degli studenti in età di obbligo formativo, curata dalla Provincia di Bologna).

In particolare, nel primo capitolo, prima parte, vengono presentati e descritti dati aggregati, provenienti da fonte ministeriale (Ministero dell'Istruzione), dalla cui elaborazione si traggono alcune considerazioni riguardo al percorso scolastico degli studenti, in termini di regolarità scolastica, ripetenza ed abbandono. Nella seconda parte dello stesso capitolo vengono descritti i dati individuali che entreranno nella stima del modello, mostrando la composizione degli studenti ed evidenziando graficamente l'esistenza di relazioni tra l'esito scolastico ed alcune variabili esplicative che poi entreranno nel modello. Nella terza parte del capitolo vengono riportati alcuni risultati di un'analisi su dati campionari precedentemente condotta, al fine di evidenziare la presenza di una relazione stretta tra alcune variabili sulla condizione famigliare degli studenti e la tipologia di scuola frequentata.

Nel secondo capitolo viene descritta la prima fase di stima di un modello a partire dai dati individuali: il modello logistico. Di questo vengono anche elencate le criticità, che hanno indotto alla scelta successiva di un modello dipendente anche dal tempo.

In terzo capitolo viene descritta la metodologia finale utilizzata, cioè la teoria del modello a curva latente condizionato. In particolare vengono descritti gli indicatori poi utilizzati per saggiare la bontà di adattamento dei modelli stimati e per scegliere il modello migliore.

Nel quarto capitolo è delineata la procedura di stima, con il raggiungimento del modello ottimo. Le variabili che risultano significative per spiegare l'andamento delle promozioni individuali nel corso del tempo risultano essere la cittadinanza (gli italiani ottengono mediamente risultati migliori degli stranieri), il sesso (le ragazze hanno un percorso scolastico mediamente migliore dei ragazzi, anche se, mentre la differenza risulta marcata negli istituti tecnici e professionali, al liceo essa risulta maggiormente sfumata), infine la tipologia di scuola frequentata (gli studenti del liceo ottengono risultati migliori, in termini di regolarità scolastica, rispetto ai colleghi che frequentano altri tipi di istituto).

I risultati ottenuti risultano fortemente dipendenti dai dati impiegati, con particolare riguardo al limite territoriale dei dati stessi. Già precedenti analisi evidenziano una forte differenziazione dei risultati scolastici tra studenti del nord e del sud Italia, oltre che tra studenti dei comuni maggiormente popolati e studenti dei comuni di provincia. Sarebbe interessante disporre di dati individuali analoghi a quelli in questa sede utilizzati, ma riferiti all'intero territorio nazionale, oppure comunque ad un zona maggiormente vasta dell'Italia, onde saggiare l'entità dell'influenza della variabile legata al territorio di frequenza scolastica sul percorso scolastico ed in particolare sulla regolarità.

1 Il monitoraggio degli esiti scolastici: un problema aperto

1.1 INTRODUZIONE

La disponibilità di informazioni attendibili e complete sulle principali entità del sistema scolastico (sedi, alunni, docenti, caratteristiche dei percorsi formativi) risente di alcuni fattori di criticità, strettamente legati alle caratteristiche degli strumenti di rilevazione ed alla tempistica di riferimento adottata. Basta pensare, ad esempio, alla mobilità degli studenti nel corso di un medesimo anno scolastico o alle varie procedure di assegnazione e conferma del personale scolastico, per capire quanto sia determinante ai fini del risultato statistico scegliere, ad esempio, un dato riferimento temporale. A ciò si aggiungano le problematiche che derivano dall'utilizzare, per scopi molteplici, dati raccolti con finalità predefinite e a volte molto diverse; ciò può portare, infatti, a diminuire la precisione dei dati in relazione ai fenomeni oggetto d'interesse.

La disponibilità di dati quantitativi sui principali fenomeni del sistema educativo regionale rappresenta una fonte di informazione indispensabile per accompagnare e "governare" lo sviluppo ed il miglioramento delle istituzioni scolastiche. Troppo spesso le decisioni di carattere politico, amministrativo o organizzativo sulla vita della scuola vengono prese sull'onda dell'emergenza, della scadenza burocratica, oppure su valutazioni pregiudiziali, quasi sempre senza un quadro esauriente di dati relativi a tendenze, impatto, contesto.

Gli amministratori del sistema scuola, al fine di porre in opera interventi atti a migliorare l'offerta formativa, infatti, hanno la necessità di conoscere in modo approfondito la struttura della popolazione scolastica, in particolare gli andamenti della scolarità, i ritardi e le regolarità degli allievi, i tassi di successo ed insuccesso, le diverse dinamiche delle iscrizioni, la distribuzione degli esiti, in termini di raggiungimento del titolo di studio, se e in quanto tempo. Per quanto riguarda in particolar modo l'esito scolastico, questo dipende da una serie di fattori individuali, legati alla peculiarità del singolo studente, quindi alla sua condizione personale, ma anche al contesto socio-economico in cui egli vive, inoltre da fattori collettivi quali l'organizzazione dell'istituzione scuola e dal tipo di offerta formativa proposta. Gli interventi istituzionali sono mirati a migliorare gli esiti degli allievi, cercando soprattutto di diminuire l'abbandono precoce (cioè prima di aver conseguito il titolo di studio) e di limitare gli insuccessi, soprattutto in termini di ripetenze. Per fare ciò, vi è il bisogno di disporre di un quadro esatto della situazione attuale, in quanto a rendimento degli studenti ed a fattori che lo determinano, ed anche di un monitoraggio dell'andamento di tale rendimento nel tempo, specie dopo un intervento istituzionale teso ad apportare un miglioramento: solo in tal modo chi amministra il sistema potrà avere informazioni sull'esito del proprio intervento.

La disponibilità di dati per quanto riguarda la scuola, ed in particolare gli esiti scolastici degli studenti, è veramente limitata. In particolare, per quanto riguarda le scuole secondarie di secondo grado (popolazione obiettivo della presente analisi), il Ministero mette a disposizione, anche sul proprio sito Internet, alcuni dati aggregati sugli esiti, in termini di promozione a fine anno scolastico, senza però che vi sia la possibilità di trarre informazioni sull'abbandono, tantomeno sul percorso scolastico dei singoli studenti o anche sulle valutazioni, seppur in forma aggregata. Soltanto le singole istituzioni scolastiche detengono le informazioni sul percorso formativo dettagliato, ciascuna dei propri ragazzi. Gli altri soggetti istituzionali coinvolti nella programmazione e nel monitoraggio dell'offerta formativa (Ministero, Province, Regioni, INVALSI, OCSE, ecc.), che pur sono in

possesso soltanto di dati aggregati, divulgano, tramite pubblicazioni o tramite Internet, elaborazioni dei dati posseduti riguardo alla composizione degli studenti ed anche agli esiti, ma in modo frammentario, non continuativo e soprattutto rendendo molto arduo il compito di svolgere ulteriori ed approfondite analisi in base ai dati pubblicati. Nel seguito del capitolo verranno mostrati alcuni calcoli aggregati sul dataset di tipo anagrafico qui analizzato, proprio al fine di focalizzare l'attenzione sulle compensazioni che derivano dal calcolo di dati aggregati e che possono portare ad una concezione distorta della realtà.

1.2 PRESENTAZIONE DI ALCUNI DATI AGGREGATI

Uno sguardo ad alcuni dati aggregati può dare una prima idea del fenomeno dell'abbandono precoce della scuola ed anche dell'esito scolastico, così si è scelto di riportare in questa sede alcune elaborazioni di dati aggregati, riferiti alla regione Emilia Romagna, provenienti dalla fonte "Rilevazioni integrative" del Ministero dell'Istruzione, che raccoglie dati non individuali direttamente dalle singole scuole. Il dataset del Ministero contiene informazioni, relative a diversi anni scolastici, sulla composizione delle classi delle scuole italiane (i dati vengono raccolti una volta all'anno e si riferiscono alla situazione al 30 settembre di ciascun anno scolastico); vi sono quindi dati sul numero di studenti per ogni anno scolastico, per ogni classe e per ogni classe di età, ma sono presenti anche informazioni sugli insegnanti e sugli edifici scolastici. Le "Rilevazioni integrative" non potevano dirsi fonte completa sui dati italiani nei primi anni di avvio (intorno all'anno 2000), ma già dopo pochi anni si poteva parlare di presenza di dati completi e corretti.
Per dare un'idea del fenomeno, la Tabella 1 contiene alcuni indicatori, riferiti all'Italia, alla regione Emilia Romagna e alla provincia di Bologna.

Tabella 1 – Indicatori globali, anno scolastico 2006/07

	Italia[(a)]	Nord[(a)]	Emilia Romagna[(a)]	provincia di Bologna[(b)]
Studenti (MF)	2.735.135	1.035.010	161.139	30.850
Studenti (F)	1.338.418	513.483	78.574	14.998
Tasso di iscrizione a scuola (MF) [(c)]	92,7	90,0	96,9	93,6
Regolarità (MF) [(d)]	74,5	73,6	74,5	74,4
Regolarità (F) [(d)]	79,5	78,6	79,1	78,3
% Diplomati [(e)]	77,5	71,2	74,5	73,5
% Iscritti al Liceo	41,3	38,5	36,1	44,0
% Iscritti negli istituti artistici	3,8	4,0	4,0	3,3
% Iscritti negli istituti tecnici	34,3	36,0	36,6	32,5
% Iscritti negli istituti professionali	20,6	21,5	23,3	20,3
% Ripetenza (MF)	6,3	6,1	5,4	5,8
% Ripetenza (F)	4,4	4,3	3,9	4,5

(a) Elaborazioni ISTAT a partire dai dati del Ministero dell'Istruzione
(b) Provincia di Bologna - Osservatorio sulla scolarità
(c) Studenti di 14-18 anni iscritti nelle scuole secondarie di secondo grado sulla popolazione residente
(d) Studenti regolari sugli iscritti in tutte le classi
(e) Diplomati sulla popolazione residente di 19 anni di età

Un fattore che può spiegare la tendenza all'uscita da scuola è la ripetenza, quindi la percentuale di ripetenti sugli iscritti è un indicatore della propensione all'abbandono scolastico: in teoria, è più probabile che lascino la scuola quegli studenti che hanno terminato l'anno scolastico senza essere promossi.
Prima di fare delle analisi dettagliate, è utile osservare alcuni dati aggregati sugli studenti dell'Emilia Romagna e di Bologna (dati provenienti dalle "Rilevazioni integrative", Ministero dell'Istruzione), al fine di identificare il

contesto nel quale si sta entrando. Una prima idea di tale contesto è data dal numero di studenti iscritti nelle scuole secondarie di secondo grado nei diversi anni scolastici e per singola classe di appartenenza, confrontati con la popolazione residente in età 14-18 (Tabella 2 e Tabella 3).

Tabella 2 – Studenti iscritti nelle scuole secondarie di secondo grado della regione Emilia Romagna

	I classe	II classe	III classe	IV classe	V classe	TOT	Residenti 14-18 al primo gennaio[1]
a.s. 00/01	32.171	29.054	27.385	24.977	23.166	136.753	153.387
a.s. 01/02	32.796	28.864	28.782	25.505	23.444	139.391	151.727
a.s. 02/03	35.183	29.069	28.669	26.487	24.114	143.522	150.473
a.s. 03/04	35.336	29.956	28.027	25.340	24.193	142.852	150.949
a.s. 04/05	38.079	31.585	30.348	25.703	24.228	149.943	152.584
a.s. 05/06	39.244	33.530	31.468	27.359	23.957	155.558	156.215
a.s. 06/07	39.721	34.391	33.629	27.753	25.645	161.139	161.609
a.s. 07/08	39.700	34.682	34.383	29.096	26.265	164.126	166.313
a.s. 08/09	40.929	34.186	33.701	29.592	27.011	165.419	169.562

Tabella 3 – Studenti iscritti nelle scuole secondarie di secondo grado della provincia di Bologna

	I classe	II classe	III classe	IV classe	V classe	TOT	Residenti 14-18 al primo gennaio[2]
a.s. 00/01	5.575	5.103	4.733	4.359	3.949	23.719	30.851
a.s. 01/02	6.015	5.193	5.186	4.601	4.268	25.263	30.752
a.s. 02/03	6.843	5.721	5.365	5.058	4.605	27.592	30.715
a.s. 03/04	6.859	5.694	5.297	4.635	4.362	26.847	30.810
a.s. 04/05	7.573	6.164	5.715	4.935	4.506	28.893	31.176
a.s. 05/06	7.652	6.536	6.042	5.140	4.626	29.996	31.993
a.s. 06/07	7.896	6.609	6.446	5.188	4.711	30.850	32.976
a.s. 07/08	7.907	6.770	6.475	5.485	5.029	31.666	33.961
a.s. 08/09	8.119	6.713	6.545	5.551	5.193	32.121	34.729

Si può notare l'aumento progressivo di studenti dal primo anno scolastico considerato fino ad arrivare agli ultimi anni, in modo più marcato nell'intera regione (dal 2000 al 2008, gli iscritti sono aumentati del 21%, si parla di oltre 20.000 studenti, e i residenti dell'11%), ma anche nella sola provincia di Bologna (dal 2000 al 2008, gli iscritti sono aumentati del 35%, si parla di circa 8.000 studenti, e i residenti del 13%): vi è stato sicuramente un aumento della popolazione in età di scuola secondaria (sancito anche da un progressivo aumento nell'erogazione delle risorse da parte del Ministero alla Direzione Regionale), ma anche una maggiore scolarizzazione, dovuta probabilmente all'entrata in vigore della legge sull'obbligo formativo, che dal 2000 ha innalzato progressivamente l'obbligo scolastico e formativo (vi è l'obbligo di rimanere all'interno della formazione fino al

[1] Fonte: Regione Emilia Romagna, "statistica self service" (sul sito Internet). Si tratta della somma dei residenti di età 14-18 al primo gennaio di ogni anno (ad esempio, per l'a.s. 2000/01 si tratta dei residenti al primo gennaio 2000).
[2] Fonte: Regione Emilia Romagna, "statistica self service" (sul sito Internet). Si tratta della somma dei residenti di età 14-18 al primo gennaio di ogni anno (ad esempio, per l'a.s. 2000/01 si tratta dei residenti al primo gennaio 2000).

diciottesimo anno di età). Confrontando gli iscritti in quinta classe in ciascun anno scolastico con i residenti diciottenni[3], si ha che la percentuale di ragazzi a scuola passa, dal 72% del 2000 in regione e 63% in provincia, all'80% del 2008 in regione e 77% in provincia. L'aumento di studenti a scuola in un anno scolastico rispetto al precedente, nella regione è stato in media[4] del 2,4% dal 2000/01 al 2008/09 (l'analoga media è dell'1,3% se si considerano i residenti), mentre l'analogo aumento medio nella provincia è stato del 3,9% (l'1,5% se si considerano i residenti). Occorre tener conto del fatto che nei primi anni, in particolar modo nel primo anno scolastico considerato, vi erano ancora alcune problematiche durante la fase di rilevazione dei dati, quindi si può dire che gli aumenti verificatisi negli ultimi anni siano maggiormente rispondenti alla realtà che non gli altri. Nessun dubbio sul fatto che la regione, ed in particolare la provincia, siano collettori di studenti, in quanto vi è in generale un aumento della popolazione scolastica da un anno scolastico al successivo, anche se con un tendenziale decremento del tasso di crescita. In regione, infatti, se nel 2004 è stato toccato il picco del 5% di aumento degli studenti rispetto all'anno precedente, nel 2008 l'analogo aumento è stato dello 0,8%; in provincia, nel 2004 gli studenti sono aumentati del 7%, mentre nel 2008 l'aumento è stato dell'1%. Vi è infine da notare che gli studenti di Bologna rappresentano circa il 20% degli studenti di tutta la regione.

Per dare un'idea della irregolarità nelle scuole superiori, è utile calcolare le differenze percentuali, per ogni anno scolastico, tra il numero di studenti che frequentavano la quinta classe e il numero di studenti che frequentavano la prima classe in ciascun anno scolastico. Come mostrato in Tabella 4 (elaborazione dai dati del Ministero), supponendo che la detta differenza rappresenti, tramite un'analisi per contemporanei, la variazione tra le presenze a scuola all'inizio del percorso ed alla fine dello stesso, si ha che, ad esempio nell'anno scolastico 2008/09, a Bologna, in quinta gli studenti sono calati del 36% rispetto all'entrata in prima. Il dato di un calo di oltre il 30% è abbastanza allarmante.

Tabella 4 – Differenze percentuali tra coloro che in ciascun anno frequentavano la V classe e la I classe

	Emilia Romagna	Bologna
a.s. 00/01	-27,99	-29,17
a.s. 01/02	-28,52	-29,04
a.s. 02/03	-31,46	-32,70
a.s. 03/04	-31,53	-36,40
a.s. 04/05	-36,37	-40,50
a.s. 05/06	-38,95	-39,55
a.s. 06/07	-35,44	-40,34
a.s. 07/08	-33,84	-36,40
a.s. 08/09	-34,00	-36,04

Nelle percentuali di cui sopra occorre però considerare il progressivo aumento degli studenti da un anno all'altro, già menzionato, che di fatto riduce l'entità del problema. Con le osservazioni successive si vedrà però che la percentuale viene sì ridotta, ma non drammaticamente.

Data la disponibilità di dati su più anni scolastici, risulta anche possibile calcolare (supponendo invece un'analisi per generazioni, come quella mostrata in Tabella 5 e proveniente da un'elaborazione dei dati ministeriali) la differenza tra il numero di studenti che frequentavano la prima classe in un dato anno scolastico e il numero di studenti che frequentavano la quinta classe 5 anni dopo (anche se occorre tener presente che non si tratta

[3] Statistica self service della Regione Emilia Romagna.
[4] Media geometrica dei rapporti tra il numero di studenti di ogni anno e quelli del precedente.

esattamente degli stessi studenti, in quanto né la regione né la provincia sono realtà chiuse, ma il contingente è il medesimo).

Tabella 5 – Differenze percentuali tra coloro che frequentavano la I classe e coloro che frequentavano la V classe 5 anni dopo

	Emilia Romagna	Bologna
a.s. 00/01 – I classe	-24,69	-19,17
a.s. 01/02 – I classe	-26,95	-23,09
a.s. 02/03 – I classe	-27,11	-31,16
a.s. 03/04 – I classe	-25,67	-26,68
a.s. 04/05 – I classe	-29,07	-31,43

In Tabella 5 le percentuali sono inferiori a quelle di Tabella 4, ma la situazione è drammatica: circa il 25% degli studenti di Bologna non arriva in 5 anni al diploma, percentuale in aumento negli ultimi anni, quando i dati erano maggiormente attendibili, quindi vi è da pensare che il valore vero si attesti più vicino al 30% che non al 20%.
È anche possibile calcolare (Tabella 6) le differenze assolute tra gli studenti iscritti in una determinata classe in ogni anno scolastico (iscritti[*i,t*], dove *i* è la classe frequentata, e *t* è l'anno scolastico) e gli studenti iscritti alla classe precedente nel precedente anno scolastico (iscritti[*i-1,t-1*]); ciò è stato fatto, però, tenendo anche conto di coloro che non sono usciti dalla scuola, ma che hanno ripetuto la classe, che quindi vengono aggiunti (la differenza assoluta viene aumentata del numero di coloro che hanno ripetuto la classe precedente nello stesso anno scolastico - ripetenti[*i-1,t*] – mentre è diminuita del numero di coloro che hanno ripetuto la stessa classe nello stesso anno scolastico– ripetenti[*i,t*] – e del numero di coloro che hanno ripetuto più di una volta la stessa classe– pluriripetenti[*i-1,t*]). La prima riga della Tabella 6, ad esempio, indica che coloro che erano a scuola in prima nel 2000/01, nell'anno successivo in seconda, sono calati di 2.838 unità, tenuto conto delle ripetenze, ed ancora sono calati di altre 391 unità nel 2002/03 in terza e così via. Quindi la diminuzione di studenti in regione di 7.943 unità (calcolati da Tabella 2), tra la prima nel 2000/01 e la quinta nel 2004/05, deve in realtà essere considerata di 6.780 unità (somma della prima riga in Tabella 6), se si tiene conto delle ripetenze. Questo calcolo è stato eseguito per tentare di identificare gli stessi studenti nei diversi anni scolastici (ovviamente, vi sono anche altri fattori che hanno determinato il movimento degli studenti in regione, ma quella sui ripetenti è l'unica correzione che si può fare sulla base dei dati disponibili).

Tabella 6 – Variazioni assolute degli studenti iscritti nelle scuole superiori della regione Emilia Romagna, da un anno scolastico al successivo, tenendo conto delle ripetenze [5]

	II classe	III classe	IV classe	V classe
Iscritti I classe a.s. 00/01	-2.838	-391	-2.929	-622
Iscritti I classe a.s. 01/02	-2.965	-1.198	-1.816	-1.146
Iscritti I classe a.s. 02/03	-4.407	237	-2.127	-1.031
Iscritti I classe a.s. 03/04	-2.768	-39	-2.896	-863
Iscritti I classe a.s. 04/05	-3.530	196	-3.767	-1.160

[5] La formula utilizzata è la seguente: iscritti(i,t)-ripetenti(i,t)-iscritti(i-1,t-1)+ripetenti(i-1,t)-pluriripetenti (i-1,t) dove i è la classe frequentata, mentre t è l'anno scolastico. La fonte dei dati sulle ripetenze è il Ministero dell'Istruzione, Rilevazioni Integrative.

Tabella 6 – Variazioni assolute degli studenti iscritti nelle scuole superiori della regione Emilia Romagna, da un anno scolastico al successivo, tenendo conto delle ripetenze [5]

Medie	-3.301	-239	-2.707	-964

La Tabella 6 mostra l'andamento degli abbandoni (i ripetenti sono in realtà esclusi, a meno di errori di misurazione del numero di ripetenti) nei diversi anni scolastici. La diminuzione maggiore di studenti (sia in termini assoluti che relativi, si veda anche la Tabella 7, elaborazioni da dati ministeriali) avviene tra le prima e la seconda classe: oltre agli studenti non promossi, più numerosi in prima, molti studenti (per lo più quelli che hanno già frequentato la scuola per 10 anni, quindi sono già stati bocciati in precedenza) iniziano la scuola superiore, ma poi abbandonano il sistema scolastico e (forse) si orientano verso la formazione professionale (ad esempio, i corsi professionali della Provincia); a questo proposito, occorre ricordare che la legge sull'obbligo formativo prevede che uno studente, dopo l'obbligo scolastico possa adempiere quello formativo non solo nella scuola, ma anche nella formazione professionale e nell'apprendistato. Si possono notare anche sensibili cali di studenti nella quarta classe rispetto alla terza: a questo proposito occorre ricordare che in alcuni tipi di scuola, in particolare negli istituti professionali, dopo il terzo anno viene rilasciato un diploma di qualifica, che induce alcuni studenti a terminare a quel punto il percorso scolastico, oltre a determinare un tasso di bocciatura maggiore. È immediatamente evidente che i dati in realtà non sono facili da interpretare: alcuni valori sono frutto di una compensazione, in quanto molti altri fattori vanno ad influire sul comportamento dei dati qui riportati (che sono aggregati). Vi è un continuo movimento di studenti: alcuni di loro iniziano i loro studi in una provincia e poi si trasferiscono in un'altra, sia per motivi famigliari che per motivi strettamente legati allo studio; forse poi ritornano nella precedente provincia; a volte terminano un anno scolastico senza successo e quindi nel successivo frequentano privatamente due anni in uno, per poi ritornare, l'anno successivo, regolarmente nel percorso scolastico pubblico.

Tabella 7 – Variazioni percentuali degli studenti iscritti nella regione Emilia Romagna, da un anno scolastico all'altro, tenendo conto delle ripetenze[6]

	II classe	III classe	IV classe	V classe
Iscritti I classe a.s. 00/01	-8,82	-1,35	-10,22	-2,45
Iscritti I classe a.s. 01/02	-9,04	-4,12	-6,48	-4,46
Iscritti I classe a.s. 02/03	-12,53	0,79	-7,01	-3,77
Iscritti I classe a.s. 03/04	-7,83	-0,12	-9,20	-3,11
Iscritti I classe a.s. 04/05	-9,27	0,62	-12,41	-4,51
Medie[7]	-9,36	-0,81	-8,81	-3,57

Uno sguardo all'analoga Tabella 8 (elaborazioni dai dati aggregati del Ministero), costruita con riferimento ai dati della provincia di Bologna può aiutare a fare altre considerazioni.

[6] La formula usata per il numeratore è la stessa della Tabella 6, mentre il denominatore è iscritti(I-1,t-1) perché è il contingente di riferimento.
[7] Media geometrica dei rapporti

Tabella 8 - Variazioni assolute degli studenti iscritti nelle scuole superiori della provincia di Bologna, da un anno scolastico al successivo, tenendo conto delle ripetenze[8]

	II classe	III classe	IV classe	V classe
Iscritti I classe a.s. 00/01	-249	139	-601	-50
Iscritti I classe a.s. 01/02	-140	-459	-250	-193
Iscritti I classe a.s. 02/03	-1.036	73	-370	-285
Iscritti I classe a.s. 03/04	-487	-139	-690	-58
Iscritti I classe a.s. 04/05	-817	-4	-765	-175
Medie	-546	-78	-535	-152

Le variazioni riportate in Tabella 7, che riguardano gli studenti dell'intera regione, ed in particolare quelle riportate in Tabella 8, sulla provincia di Bologna, comprendono gli studenti stranieri, che in generale appartengono ad una categoria di persone più probabilmente a rischio di abbandono del sistema scolastico e comunque di percorso scolastico accidentato, in quanto provenienti da condizioni famigliari più disagiate rispetto agli altri. Il tasso di abbandono, ed anche il tasso di ripetenza, risentono quindi anche della presenza di questi studenti.

Tabella 9 - Variazioni percentuali degli studenti iscritti nella provincia di Bologna, da un anno scolastico all'altro, tenendo conto delle ripetenze[9]

	II classe	III classe	IV classe	V classe
Iscritti I classe a.s. 00/01	-4,47	2,68	-11,20	-1,08
Iscritti I classe a.s. 01/02	-2,33	-8,02	-4,72	-3,91
Iscritti I classe a.s. 02/03	-15,14	1,28	-6,48	-5,54
Iscritti I classe a.s. 03/04	-7,10	-2,26	-11,42	-1,12
Iscritti I classe a.s. 04/05	-10,79	-0,06	-13,39	-3,55
Medie[10]	-6,55	-1,32	-8,79	-2,47

Come già visto, la dimostrazione che in realtà non tutto è mostrato dai dati aggregati qui presentati (vista l'apparente incongruenza di alcuni dati, non bisogna dimenticare che esiste anche un movimento degli studenti all'interno del territorio nazionale, tra una regione e l'altra) è rappresentata anche dalle variazioni positive che si notano nella classe terza, mentre ci si potrebbero aspettare soltanto variazioni negative o al più nulle. Ciò fa pensare che anche i valori negativi della seconda e quarta classe siano in realtà valori più bassi di quelli registrati

[8] La formula utilizzata è la seguente: iscritti(i,t)-ripetenti(i,t)-iscritti(i-1,t-1)+ripetenti(i-1,t)-pluriripetenti (i-1,t) dove i è la classe frequentata, mentre t è l'anno scolastico.

[9] La formula usata per il numeratore è la stessa della Tabella 6, mentre il denominatore è iscritti(I-1,t-1) perché è il contingente di riferimento.

[10] Media geometrica dei rapporti

(i valori già negativi potrebbero essere di per sé il risultato di un peggiore *drop out* associato ad entrate di altri studenti). Sembra quasi che in regione entrino più studenti da fuori regione che non il contrario. Su questo fatto, si deve considerare che in ogni regione vi sono zone di confine, dove può accadere che uno studente risieda ad una minore distanza da una scuola fuori regione che non all'interno della regione di residenza.
In questa fase descrittiva, si sono ricercate alcune suddivisioni degli studenti in base ad alcune caratteristiche che possano aiutare a risalire alle cause dell'abbandono. Un primo fattore considerato è la regolarità, se cioè un ragazzo frequenta la classe in linea con la propria età (si definisce regolare, ad esempio, un quindicenne che frequenta la seconda classe).

Tabella 10 – Percentuali di studenti non regolari sul totale degli studenti – regione Emilia Romagna[11]

	I classe	II classe	III classe	IV classe	V classe
a.s. 00/01	20,11	21,58	25,06	25,45	25,01
a.s. 01/02	19,85	22,06	25,71	24,71	25,95
a.s. 02/03	20,25	21,74	25,45	25,10	26,77
a.s. 03/04	19,99	21,22	24,46	24,38	26,62
a.s. 04/05	21,17	21,95	24,02	23,82	25,61
a.s. 05/06	21,89	22,88	25,68	23,74	25,91
a.s. 06/07	23,07	23,91	26,11	24,07	25,85
a.s. 07/08	23,76	24,38	27,02	24,57	27,47

Uno sguardo a Tabella 10 e Tabella 11 (elaborazioni dei dati provenienti dalle "Rilevazioni Integrative", rispettivamente riferite alla regione e alla provincia) suggerisce che la regolarità in realtà sembra non influire sull'abbandono precoce del sistema scolastico; infatti mentre la percentuale di studenti che abbandonano differisce nelle diverse classi frequentate e nei diversi anni scolastici, il trend della regolarità pare rimanere per lo più il medesimo in tutti gli anni scolastici, con un leggero aumento dal primo anno scolastico considerato all'ultimo e con invece un aumento pressoché costante da una classe alla successiva (ovviamente il numero di studenti che sono stati almeno una volta bocciati aumenta in modo fisiologico da una classe alla successiva). Le percentuali di studenti non regolari sul totale degli studenti, inoltre, non differiscono in modo rilevante in provincia rispetto all'intera regione.

Tabella 11 - Percentuali di studenti non regolari sul totale degli studenti – provincia di Bologna[12]

	I classe	II classe	III classe	IV classe	V classe
a.s. 00/01	21,18	22,69	27,17	26,56	27,54
a.s. 01/02	21,43	23,17	26,74	26,28	25,97
a.s. 02/03	19,81	22,46	25,36	25,82	27,10
a.s. 03/04	22,16	22,39	25,82	25,48	25,38

[11] Non è stato possibile calcolare il dato per l'anno scolastico 2008/09, per mancanza dei dati suddivisi per età degli studenti.
[12] Non è stato possibile calcolare il dato per l'anno scolastico 2008/09, per mancanza dei dati suddivisi per età degli studenti.

Tabella 11 - Percentuali di studenti non regolari sul totale degli studenti – provincia di Bologna[12]

	I classe	II classe	III classe	IV classe	V classe
a.s. 04/05	22,76	23,26	22,17	22,79	23,38
a.s. 05/06	22,69	24,12	26,64	23,88	26,85
a.s. 06/07	22,14	24,83	27,43	25,17	25,34
a.s. 07/08	23,59	24,68	28,24	25,26	28,73

Dall'osservazione dei dati aggregati sugli esiti scolastici, si può giungere ad altre interessanti considerazioni.
Le singole scuole raccolgono informazioni sugli studenti iscritti e sui loro esiti formativi al termine di ciascun anno scolastico, in seguito le comunicano al Ministero dell'Istruzione. Tali dati, in forma aggregata, vengono resi disponibili anche dalla Regione Emilia Romagna sulla rete Internet. Risulta così agevole confrontare il numero di studenti iscritti all'inizio di ciascun anno scolastico con il numero di studenti valutati alla fine dello stesso anno (Tabella 12, elaborazione da dati del Ministero, Rilevazioni Integrative). Un saldo positivo tra iscritti e scrutinati indica la presenza di abbandono in corso d'anno, mentre un saldo negativo denota un aumento di studenti (nuovi ingressi in regione). Da settembre a giugno è possibile, per ogni studente, cambiare scuola (nel caso di cambiamento, se la scuola di destinazione si trova all'interno della regione, il dato aggregato non cambia, muta invece se lo studente si trasferisce in un'altra regione), ma è anche possibile lasciare il percorso scolastico per orientarsi verso la formazione professionale della Provincia o l'apprendistato. È anche possibile che altri studenti entrino nelle scuole dell'Emilia Romagna da altre regioni o addirittura da stati esteri. Vi sono anche studenti che ritornano a scuola dopo averla lasciata uno o due anni addietro; questi risultano ovviamente avere un ritardo scolastico. Le percentuali di irregolari in generale aumentano nella classe terza: l'obbligo scolastico termina a 18 anni e gli studenti che sono stati bocciati almeno una volta arrivano a compiere (o hanno già compiuto) 18 anni durante la frequenza della classe terza, quindi possono lasciare la scuola (in media nei diversi anni scolastici, la percentuale di studenti al di fuori del percorso di studi regolare in terza è circa del 25%).

Tabella 12 – Differenze percentuali di studenti tra l'inizio e la fine di ciascun anno scolastico (dopo gli esiti finali) - regione Emilia Romagna[13]

	I classe	II classe	III classe	IV classe	V classe
a.s. 00/01	2,99	3,25	-8,26	3,19	NA
a.s. 01/02	2,73	1,46	-13,62	2,09	NA
a.s. 02/03	13,72	13,36	0,58	16,10	NA
a.s. 03/04	4,25	3,19	-8,59	3,67	NA
a.s. 04/05	NA	NA	NA	NA	NA
a.s. 05/06	3,96	1,91	-5,45	4,18	NA
a.s. 06/07	2,99	1,99	3,38	2,93	NA
a.s. 07/08	2,34	2,74	3,64	2,77	2,75

Vengono ora esaminati i dati aggregati sugli studenti che hanno abbandonato la scuola in corso d'anno: hanno comunicato alla scuola il loro abbandono prima del termine dell'anno scolastico (Tabella 13, elaborazioni da dati

[13] I dati non sono disponibili per la classe quinta (escluso il 2007/08) e per l'anno scolastico 2004/2005.

del Ministero). Vi è da notare che questi dati differiscono in modo sensibile dalle percentuali calcolate in base alle differenze tra gli studenti presenti a inizio anno scolastico e gli studenti che arrivano alla fine dell'anno scolastico (Tabella 12, elaborazioni da dati del Ministero). Ciò a dimostrazione dell'entità del movimento di studenti nelle scuole della regione. Per quanto riguarda i dati riferiti all'anno scolastico 2002/03 (Tabella 12), nei dati vi è probabilmente qualche problema, in quanto le percentuali risultano molto diverse rispetto agli altri anni scolastici. Anche per la classe terza, nella medesima tabella, risulta esserci qualche problema nei dati, in quanto gli studenti sembrano aumentare dall'inizio alla fine dell'anno scolastico. Ciò è probabilmente dovuto ad errore di trasmissione dei dati da parte degli istituti professionali, che comunicano separatamente i dati per coloro che superano il test di qualifica e coloro che invece seguono corsi per i quali questo non è previsto, con la conseguenza di una possibile sovrapposizione di informazioni. Occorre anche tener conto del fatto che i dati in Tabella 13 sono in realtà le comunicazioni formali di abbandono fatte alla scuola, quindi mancano probabilmente altri studenti che hanno ugualmente lasciato la scuola. Trattandosi di dati aggregati forniti da enti esterni, non risulta possibile alcuna correzione, ed è anche difficoltoso risalire alle cause che possono aver provocato tali incongruenze. Ci si convince sempre più del fatto che per ottenere risultati attendibili occorre disporre di dati individuali, così da poter seguire il percorso scolastico effettivo di ogni studente.

Tabella 13 – Percentuali di studenti che risultano aver abbandonato la scuola – regione Emilia Romagna[14]

	I classe	II classe	III classe	IV classe	V classe
a.s. 00/01	4,41	3,41	2,63	3,40	1,06
a.s. 01/02	4,51	3,26	3,04	2,96	0,78
a.s. 02/03	4,32	3,67	3,14	2,99	0,75
a.s. 03/04	4,48	3,30	3,48	3,08	0,93
a.s. 04/05	NA	NA	NA	NA	NA
a.s. 05/06	3,20	2,34	2,68	2,76	0,91
a.s. 06/07	5,44	3,52	4,06	3,71	1,31

Tabella 14 – Differenze percentuali di studenti tra l'inizio e la fine di ciascun anno scolastico (dopo gli esiti finali) – provincia di Bologna[15]

	I classe	II classe	III classe	IV classe	V classe
a.s. 00/01	5,97	4,74	-17,45	2,73	NA
a.s. 01/02	-1,30	-3,99	-16,47	-5,11	NA
a.s. 02/03	5,73	3,81	-14,20	5,42	NA
a.s. 03/04	7,68	6,36	0,59	7,92	NA
a.s. 04/05	NA	NA	NA	NA	NA
a.s. 05/06	5,91	2,16	2,50	7,61	NA
a.s. 06/07	4,37	2,80	3,68	2,93	NA
a.s. 07/08	4,51	3,55	3,91	2,66	1,67

[14] I dati non sono disponibili per l'anno scolastico 2004/2005 e a partire dall'anno 2007/08.
[15] I dati non sono disponibili per la classe quinta (escluso il 2007/08) e per l'anno scolastico 2004/2005.

Un fattore che potrebbe spiegare i valori negativi, in quanto a differenza tra numero di studenti a inizio e fine anno scolastico, tanto per la regione quanto per la provincia, è l'ingresso, in corso d'anno, di studenti stranieri. Le percentuali di studenti che hanno comunicato alla scuola l'abbandono non differiscono sostanzialmente tra la provincia e l'intera regione, ciò significa che il fenomeno è pressoché costante in tutto il territorio. Le sostanziali differenze, anche per la provincia di Bologna, tra gli abbandoni comunicati e le differenze sul numero di frequentanti dimostrano la presenza di altri fattori che influenzano questi dati aggregati, fattori a cui è veramente difficoltoso risalire.

Tabella 15 – Percentuali di studenti che risultano aver abbandonato la scuola – provincia di Bologna[16]

	I classe	II classe	III classe	IV classe	V classe
a.s. 00/01	5,29	3,59	2,89	3,12	1,19
a.s. 01/02	4,94	3,43	3,36	3,19	0,68
a.s. 02/03	4,25	3,93	4,03	3,08	1,00
a.s. 03/04	4,83	2,70	3,61	3,02	1,03
a.s. 04/05	NA	NA	NA	NA	NA
a.s. 05/06	3,89	2,05	2,48	3,15	0,69
a.s. 06/07	6,55	3,77	5,01	4,55	1,29

Per analizzare più nel dettaglio il fenomeno del *drop out* e gli esiti scolastici, occorrono certamente dati individuali, in modo da poter seguire ogni singolo studente dal momento della sua entrata nel sistema scolastico fino al termine degli studi, con o senza pieno successo.

Prima di presentare i dati aggregati (Tabelle 16, 17, 18 e 19, elaborazioni da dati ministeriali), che danno una prima idea dell'andamento degli esiti scolastici, occorre specificare che il numero di ragazzi promossi, laddove non espressamente indicato, comprende anche il numero di ragazzi promossi con debito. L'esito scolastico è strettamente legato all'abbandono, in quanto è maggiormente probabile che sia portato ad abbandonare la scuola un ragazzo che consegue i risultati meno brillanti.

Tabella 16 – Percentuali di studenti promossi su quelli presenti a scuola a fine anno scolastico – regione Emilia Romagna[17]

	I classe	II classe	III classe	IV classe	V classe
a.s. 00/01	82,06	87,32	90,88	90,84	NA
a.s. 01/02	81,67	87,06	87,03	90,27	NA
a.s. 02/03	81,86	86,77	86,88	90,32	NA
a.s. 03/04	81,22	86,92	84,63	90,16	NA
a.s. 04/05	NA	NA	NA	NA	96,67
a.s. 05/06	83,52	88,97	84,46	92,17	96,52
a.s. 06/07	81,58	87,74	87,97	90,79	97,65
a.s. 07/08	78,05	84,97	85,46	89,03	97,69

[16] I dati non sono disponibili per l'anno scolastico 2004/2005 e dal 2007/08 in poi.
[17] I dati sulle prime 4 classi non sono disponibili per l'anno scolastico 2004/2005.

Le percentuali di studenti promossi sono pressoché costanti nei diversi anni scolastici, mentre aumentano significativamente nel passaggio da una classe alla successiva, tranne che in terza classe dove, probabilmente in conseguenza degli esami di qualifica (e quindi dell'impossibilità in questo caso di essere promossi con debito), in alcuni anni scolastici si sono abbassate.

Tabella 17 – Percentuali di studenti promossi su quelli presenti a scuola a fine anno scolastico – provincia di Bologna[18]

	I classe	II classe	III classe	IV classe	V classe
a.s. 00/01	78,92	84,32	84,96	88,42	NA
a.s. 01/02	79,16	83,09	83,76	85,69	NA
a.s. 02/03	77,82	82,14	83,11	86,83	NA
a.s. 03/04	79,28	85,54	84,96	91,49	NA
a.s. 04/05	NA	NA	NA	NA	95,72
a.s. 05/06	83,32	87,58	86,50	92,36	98,22
a.s. 06/07	80,93	86,44	87,29	91,32	96,51
a.s. 07/08	78,98	85,27	86,44	90,15	96,62

Le percentuali di studenti promossi relative alla provincia di Bologna sono leggermente più basse di quelle della regione. I valori riscontrati nella classe terza, come detto, includono anche gli studenti che hanno superato l'esame di qualifica negli istituti professionali, e anche in provincia si riscontra, negli stessi anni scolastici, il calo della percentuale di promossi in terza classe rispetto alla seconda.

Tabella 18 – Percentuali di studenti promossi con debito su quelli presenti a scuola a fine anno scolastico - regione Emilia Romagna[19]

	I classe	II classe	III classe	IV classe
a.s. 00/01	40,66	41,19	31,23	37,58
a.s. 01/02	41,10	40,84	31,97	38,16
a.s. 02/03	40,78	42,87	31,45	39,55
a.s. 03/04	42,80	42,63	34,01	38,13
a.s. 04/05	NA	NA	NA	NA
a.s. 05/06	41,83	42,62	37,54	38,28
a.s. 06/07	43,69	42,86	38,55	40,23

Tabella 19 – Percentuali di studenti promossi con debito su quelli presenti a scuola a fine anno scolastico – provincia di Bologna[20]

	I classe	II classe	III classe	IV classe
a.s. 00/01	41,99	43,30	31,88	39,24
a.s. 01/02	42,11	43,68	34,06	39,79
a.s. 02/03	44,42	46,08	33,97	40,71

[18] I dati sulle prime 4 classi non sono disponibili per l'anno scolastico 2004/2005.
[19] I dati non sono disponibili per gli anni scolastici 2004/2005 e 2007/08.
[20] I dati non sono disponibili per gli anni scolastici 2004/2005 e 2007/08.

Tabella 19 – Percentuali di studenti promossi con debito su quelli presenti a scuola a fine anno scolastico – provincia di Bologna[20]

	I classe	II classe	III classe	IV classe
a.s. 03/04	43,49	43,85	36,67	38,05
a.s. 04/05	NA	NA	NA	NA
a.s. 05/06	44,14	43,26	39,24	38,60
a.s. 06/07	44,97	44,79	40,92	40,42

Per quanto riguarda le percentuali di ragazzi promossi con debito formativo, si può notare che nella provincia di Bologna si riscontrano valori superiori rispetto alla media regionale (siccome la percentuale di promossi in provincia è un po' inferiore alla media regionale, si può desumere che a Bologna la percentuale di studenti promossi senza debito sia decisamente inferiore alla media regionale). L'andamento del passaggio da una classe alla successiva è però il medesimo, in provincia e in regione: le prime due classi sono quelle in cui sono percentualmente di più i ragazzi promossi con debito formativo, mentre in terza e in quarta classe la percentuale di promossi con debito diminuisce (in terza si riscontra quella inferiore, anche per la presenza dell'esame di qualifica negli istituti professionali).

Visto che l'analisi principale che verrà qui trattata riguarda una coorte di studenti nati nel 1988 e che hanno frequentato le scuole superiori di Bologna e provincia, vengono nel seguito riportati alcuni dati aggregati su quella coorte, con riferimento anche alla regione Emilia Romagna e rispetto alla popolazione residente, per cercare di capire meglio il contesto in cui ci si trova ad operare.

Un primo confronto, che viene spontaneo, è quello con la popolazione residente: aiuta a capire quanta parte dei ragazzi in età scolare frequentava gli istituti della regione.

In Tabella 20 vengono confrontati gli studenti, nati nel 1988, che frequentavano le scuole della provincia di Bologna e della regione Emilia Romagna (comprese le scuole secondarie di primo grado, in quanto vi sono ragazzi che nel 2002/03 erano già ripetenti, quindi frequentavano al più la terza media), negli anni scolastici dal 2001/02 (nel caso regolare, i ragazzi frequentavano la terza media) al 2006/07 (nel caso regolare, i ragazzi frequentavano la quinta superiore), con i residenti (al primo gennaio di ogni anno) della stessa classe di età nei diversi anni scolastici. I dati relativi ai frequentanti provengono dal database "Rilevazioni Integrative" del Ministero dell'Istruzione, mentre i dati sulla popolazione sono stati ricavati dal sito Internet della Regione Emilia Romagna, "Statistica self-service" (disponibile al momento della presente analisi). I tassi indicati risentono anche della presenza di studenti residenti in regione (o in provincia) che hanno frequentato scuole al di fuori di questa (in particolare, è noto il caso degli studenti che risiedono in provincia di Bologna, ai confini con il territorio ferrarese, e studiano a Ferrara grazie a una maggiore comodità di trasporto), e della presenza di studenti che risiedono in altra regione (o provincia). Quindi è difficile ricavare il tasso di coloro che non frequentano la scuola (non è infatti detto che le entrate e le uscite siano tra loro bilanciate). Bisognerebbe poi anche tener conto di coloro che, pur avendo magari abbandonato il percorso scolastico, sono entrati in un percorso formativo, quale quello dei corsi professionali della Provincia.

Tabella 20 –Tasso di scolarità: ragazzi – nati nel 1988 – frequentanti le scuole secondarie di primo e secondo grado della regione Emilia Romagna e della provincia di Bologna sulla popolazione residente (al primo gennaio di ogni anno) della medesima classe di età

	2001/02	2002/03	2003/04	2004/05	2005/06	2006/07
Emilia Romagna	97,83	100,34	94,63	87,16	85,67	78,64
Bologna	89,05	95,12	86,99	73,62	79,64	73,43

Nella lettura dei dati in Tabella 20, occorre anche considerare la discrepanza dovuta al fatto che vengono confrontati i frequentanti all'inizio dell'anno scolastico (settembre di un certo anno) con i residenti a gennaio dell'anno successivo (l'anno scolastico è il medesimo, ma cambia l'anno solare, seppure siano trascorsi solo alcuni mesi). Occorre anche tener presente il fatto che si sono costruiti dei tassi usando fonti di dati differenti, quindi ciò può portare ad un certo grado di errore (la base dati di riferimento non è la stessa per il numeratore ed il denominatore di uno stesso rapporto).

I tassi di scolarità nella provincia di Bologna sono inferiori rispetto alla media regionale; in generale si ha comunque una diminuzione dei tassi (dopo l'anno scolastico 2002/03) in ogni anno rispetto al precedente, ciò a dimostrazione del fatto che molti studenti iniziano la scuola di secondo grado, ma poi l'abbandonano precocemente. Non bisogna dimenticare la legge sull'obbligo formativo, che prescrive l'obbligo di frequentare la scuola fino ai 15 anni, quindi in generale dopo aver terminato il primo anno di scuola secondaria di secondo grado; in conseguenza di tale legge s vede che la diminuzione del tasso di scolarità inizia proprio nel 2003/04 (per accentuarsi molto negli anni successivi), quando molti ragazzi della coorte esaminata compiono 16 anni.

I tassi di Tabella 20 calcolati per l'anno scolastico 2006/07 sono confrontabili con il tasso di scolarità riportato in Tabella 1, ma occorre tener presente che in Tabella 20 sono presenti i soli ragazzi nati nel 1988, mentre in Tabella 1 sono presenti tutti i ragazzi dai 14 ai 19 anni che erano a scuola nel 2006/07.

Per avere un parametro di confronto anche al di fuori del territorio nazionale, vengono nel seguito riportati i dati di diversi paesi europei (fonte OCSE). Non essendo disponibili dati completi ed affidabili sulla coorte dei nati nel 1988, si sono qui presentati i tassi di scolarità relativi alla coorte del 1989, ritenendo che possano essere comunque una buona base di confronto per i tassi precedentemente calcolati. Questi studenti frequentavano (nel caso regolare) il terzo anno della scuola secondaria di primo grado (o l'anno di studio equivalente negli altri paesi, anche se in effetti il confronto è difficoltoso, vista la diversità del sistema scolastico) nell'anno scolastico 2002/03 (e non 2001/02 come la precedente coorte), mentre frequentavano (nel caso regolare) la quarta superiore (o comunque il quarto anno) nel 2006/07. Non viene riportato il dato relativo al 2007/08 perché nella maggior parte dei paesi europei la scuola termina a 18 anni e non a 19 come in Italia, quindi il tasso risulta incongruente. Il confronto tra la Tabella 20 e la Tabella 21 va quindi effettuato slittando l'anno scolastico: il 2001/02 di Tabella 20 va confrontato con il 2002/03 di Tabella 21 e così via. In Tabella 21 è riportato anche il tasso italiano, calcolato dall'OCSE, per avere un confronto sulla base degli stessi dati (anche se in teoria l'OCSE dovrebbe disporre dei dati forniti dallo stesso Ministero dell'Istruzione italiano).

I tassi della regione Emilia Romagna sono leggermente più bassi della media italiana, ma la differenza è veramente piccola; Bologna si pone in controtendenza, in quanto al quarto anno della scuola superiore il tasso aumenta invece di diminuire, ma ciò potrebbe essere dovuto a un problema contingente nei dati di base.

Tabella 21 – Tasso di scolarità: ragazzi – nati nel 1989 – frequentanti le scuole sulla popolazione residente della stessa classe di età

	2002/03	2003/04	2004/05	2005/06	2006/07
Italia	100,75	100,93	95,15	88,25	83,13
Austria	99,83	99,37	92,42	91,80	77,87
Francia	99,35	99,27	97,94	96,55	89,02
Grecia	94,32	93,74	91,80	100,78	72,66
Netherlands	95,63	98,98	101,47	95,45	85,07
Spagna	105,66	102,91	99,53	94,07	82,43
Svizzera	86,75	98,16	96,91	90,41	86,48
United Kingdom	98,52	99,64	102,24	93,60	70,80

Si passa ora a calcolare (Tabella 22, elaborazione da dati ministeriali) le percentuali di studenti che hanno avuto un corso di studi regolare su tutti gli studenti; gli studenti che sono nel seguito indicati "regolari" sono coloro che non sono mai stati bocciati, quindi frequentano la classe attesa (o al massimo la classe successiva, solo nel caso abbiano iniziato la scuola primaria in anticipo) rispetto alla classe di età. Nel caso della coorte di nati nel 1988, i regolari sono quelli che frequentavano la prima superiore (o al massimo la seconda) nel 2002/03, la seconda superiore (o al massimo la terza) nel 2003/04 e così via. Il denominatore delle percentuali in Tabella 22 è la somma dei ragazzi nati nel 1988 e frequentanti la scuola in ciascun anno scolastico considerato.

Tabella 22 – Percentuali di studenti – nati nel 1988 – con un corso di studi regolare (su tutti i frequentanti), nell'ultimo anno della scuola secondaria di primo grado e nelle scuole secondarie di secondo grado

	2001/02	2002/03	2003/04	2004/05	2005/06	2006/07
Emilia Romagna	92,14	89,95	81,21	77,21	74,21	73,87
Bologna	92,93	90,62	80,22	74,67	73,05	72,00

Le percentuali riferite alla provincia di Bologna sono in generale sotto la media regionale, tranne nei primi due anni scolastici considerati (gli studenti frequentavano l'ultimo anno della scuola di primo grado ed il primo anno della scuola di secondo grado). Occorre considerare anche che la provincia di Bologna è caratterizzata dalla presenza di studenti stranieri, in percentuale maggiore rispetto alle altre province della regione. Il fatto che deve far riflettere è che il 7 - 8% degli studenti risulta già al di fuori del percorso regolare in terza media, per arrivare a un 28% in quinta superiore.

1.3 I DATI DELL'OSSERVATORIO

In questa sede verranno analizzati gli esiti scolastici dei soli studenti delle scuole secondarie di secondo grado, dove esiste il problema dell'abbandono precoce, in quanto i ragazzi hanno già assolto l'obbligo scolastico e, vincolati dalla legge a rimanere nell'ambito formativo anche se non necessariamente in quello scolastico, hanno la facoltà di interrompere il percorso scolastico senza giungere al diploma.

Obiettivo del presente lavoro è stato quello di indagare sul percorso scolastico, cercando di ottenere informazioni sulle variabili che concorrono a determinare l'andamento nel tempo degli esiti, in particolar modo gli scostamenti dalla regolarità, intesa come prosecuzione lineare degli studi. Per fare ciò, era impossibile partire da

dati aggregati, ma occorreva disporre di un contingente di studenti, per poi seguirlo nel tempo. I dati aggregati, come già dimostrato, sono utili a mostrare il panorama e ad evidenziare alcuni aspetti rilevanti, ma non consentono certo di entrare nel dettaglio del problema.

I dati anagrafici che riguardano gli studenti vengono raccolti dalle singole istituzioni scolastiche, ognuna delle quali conserva le informazioni sui propri studenti in un database, con la possibilità di divulgarli soltanto seguendo le norme sulla privacy (Dlg 196/03) e comunque con lunghi procedimenti burocratici. L'idea, perseguita inizialmente, di rivolgersi direttamente ad alcune scuole campione per ottenere dati anagrafici, seppure in forma anonima (era sufficiente avere, per ogni studente, anche solo una parte del codice fiscale, così da non poter risalire direttamente alla persona), è stata presto accantonata, date difficoltà tecniche (ogni scuola ha sistemi di archiviazione differenti e regole di divulgazione diverse) e burocratiche. Grazie alla legge 144/99, che prevede che alcuni enti pubblici, quali i Sevizi per l'impiego e le Province, provvedano annualmente a raccogliere, direttamente dalle singole scuole, alcuni dati individuali sugli studenti in età di obbligo formativo[21], ma soprattutto grazie alla solerzia in particolare della Provincia di Bologna, che ha prontamente attivato la rilevazione (partita nel 2000) ed utilizzato un valido strumento di archiviazione, si è in questa sede potuto disporre di un dataset che potesse rispondere alle esigenze del progetto di studio. I dati sono stati resi disponibili grazie ad un accordo tra la Provincia stessa e la Facoltà di Scienze Statistiche. Ovviamente si sono potute affrontare soltanto alcune problematiche tra quelle inizialmente pensate, basandosi sui dati disponibili, tuttavia si è giunti a conclusioni di sicuro interesse anche per gli amministratori del sistema. I risultati sono comunque, in altra sede, ampliabili qualora si disponesse di informazioni individuali aggiuntive. Data la complessità del sistema scuola, non è certo poca cosa poter disporre di dati anagrafici di tipo longitudinale, quindi di misure ripetute nel tempo, in particolare per un ciclo di studi completo (dalla prima alla quinta classe), per così tanti individui che hanno storie diverse, in termini di diversa provenienza e di diversa scuola frequentata. Esiste una rilevazione anagrafica degli studenti a livello nazionale, condotta dal Ministero dell'Istruzione; tuttavia è molto arduo ottenere la disponibilità anche solo di parti di tale banca dati, che inoltre è partita da solo pochi anni, insufficienti a consentire un'analisi statistica longitudinale (che abbisogna di almeno 3 istanti temporali distinti) e tantomeno che indaghi sull'intero percorso scolastico. Il dataset analizzato in questa sede comprende i soli studenti di Bologna e provincia, quindi ha sicuramente un limite territoriale, tuttavia la numerosità è comunque sufficientemente elevata da garantire la bontà del modello; i risultati potranno essere estesi anche ad altri territori: il modello, se ritenuto valido in questo territorio, potrà essere stimato su altri dati riferiti a un territorio di maggiore ampiezza, ottenendo di giungere a stime maggiormente precise dei parametri, pur mantenendo pressoché invariata la struttura del modello qui ottenuto.

I dati di partenza, come detto, di tipo anagrafico e riferiti agli studenti delle scuole secondarie di secondo grado di Bologna e provincia, sono stati controllati dagli operatori della Provincia, ma è stato necessario fare, preliminarmente allo studio, qualche controllo incrociato aggiuntivo. Infatti, l'obiettivo della Provincia non è prettamente l'analisi dell'intero dataset, quanto piuttosto la ricerca di quegli studenti che abbandonano la scuola prima di compiere i 18 anni, al fine di accertare se abbiano invece intrapreso la strada della formazione professionale, oppure se siano entrati in un percorso di apprendistato o ancora se abbiano del tutto lasciato la

[21] Attualmente, grazie ad un accordo tra la Regione Emilia Romagna, le Province e il Ministero dell'Istruzione, anche la Regione detiene un database con dati anagrafici e sul percorso scolastico egli studenti in età di obbligo formativo.

formazione (nonostante la legge vieti quest'ultima possibilità), cercando di aiutarli a rientrare in un percorso formativo, anche attraverso il contatto diretto (è questo il motivo della presenza di informazioni anagrafiche nel dataset, che le scuole devono, per legge, fornire alle Province). Il database analizzato ("Anagrafe of provinciale") proviene da un'indagine periodica (legge 144 del 17 maggio 1999, che introduce l'obbligo della frequenza di attività formative fino al diciottesimo anno di età o fino all'acquisizione della qualifica professionale o del diploma e che, soprattutto, obbliga i soggetti istituzionali a vigilare sull'adempimento dell'obbligo, costituendo una banca dati di tipo anagrafico degli studenti che all'obbligo sono soggetti; il principio del diritto dovere all'istruzione e alla formazione entro il diciottesimo anno di età, comunque per almeno 12 anni e almeno fino al conseguimento di una qualifica professionale è stato poi rafforzato con il D.lgs 15 aprile 2005, n.76) sugli studenti della provincia di Bologna.

La Provincia conduce l'indagine sull'obbligo formativo, già dall'anno scolastico 1999/2000, su tutti gli studenti (grazie ad un accordo con il Ministero dell'Istruzione vengono rilevati anche i dati sui ragazzi delle scuole secondarie di secondo grado), quindi si può parlare di indagine direttamente sulla popolazione scolastica bolognese (rientrano tutti i ragazzi iscritti nelle scuole di Bologna e provincia, ma occorre tener conto che, nelle scuole della provincia, non compaiono quegli studenti che abitano nel territorio bolognese ma frequentano istituti di altre province, mentre sono presenti studenti che frequentano le scuole bolognesi pur risiedendo in altra provincia). Lo scopo di tale rilevazione è quello di monitorare i giovani in età di obbligo formativo, all'interno del territorio provinciale di Bologna. Le modalità di rilevazione si sono affinate nel tempo, fino a consentire la trasformazione dell'anagrafe in vero e proprio Osservatorio sulla scolarità bolognese. Al fine di costruire l'Osservatorio provinciale coordinando i vari soggetti istituzionali adibiti alla gestione della formazione relativa all'età dell'obbligo formativo, la Provincia di Bologna ha siglato un protocollo d'intesa con i Comuni, le Istituzioni Scolastiche autonome, il Nuovo Circondario di Imola, i soggetti della formazione professionale accreditati per l'obbligo formativo della provincia di Bologna, l'Ufficio Scolastico Regionale per l'Emilia Romagna, l'Ufficio Provinciale di Bologna (dislocazione territoriale del Ministero dell'Istruzione), l'Università di Bologna. L'Osservatorio rappresenta uno strumento utile per la promozione di processi di valutazione ed autovalutazione del sistema scolastico e formativo, per il supporto di processi di *governance* e di programmazione di politiche pubbliche all'interno del sistema scolastico e formativo. I dati dell'Osservatorio provengono da molteplici fonti: dalle Istituzioni Scolastiche, per quanto riguarda alcune informazioni anagrafiche e sulla carriera scolastica; dai Comuni della provincia, per quanto riguarda i dati anagrafici sulla popolazione in età di obbligo formativo; dalle Agenzie adibite al coordinamento della formazione professionale, per quanto riguarda i dati sugli studenti in età di obbligo che hanno scelto di continuare gli studi al di fuori del percorso scolastico. In questa sede verranno analizzati i soli dati provenienti dalle istituzioni scolastiche. L'Osservatorio sulla scolarità si propone anche di attuare azioni di ricerca per realizzare approfondimenti qualitativi su particolari aspetti della scolarità, proprio a partire dalla banca dati anagrafica. La presente ricerca si colloca in un disegno più ampio, teso ad indagare sulle sfaccettature che il fenomeno della scolarità presenta nel territorio provinciale.

Agli inizi dell'indagine, i dati erano incompleti e perciò effettuare analisi accurate risultava molto difficile; ora, dopo anni di lavoro, l'obiettivo di avere un database completo e corretto a partire dai dati forniti alla Provincia dalle singole istituzioni scolastiche si può dire raggiunto. L'indagine presso le Istituzioni Scolastiche viene

condotta dalla Provincia in 3 diversi momenti durante l'anno scolastico: all'inizio dell'anno (ottobre), a metà anno scolastico (febbraio, dopo la fine del primo quadrimestre) e dopo la fine dell'anno scolastico (luglio). La rilevazione comprende tutti gli studenti, dai 13 ai 19 anni di età, ma il dataset analizzato nel presente studio comprende le informazioni sugli studenti che frequentavano le sole scuole secondarie di secondo grado in determinati anni scolastici, come nel seguito specificato. Il dataset, in particolare, contiene informazioni su coloro che sono nati nel 1988 (che quindi avevano 19 anni nel 2007 e che quindi, nel caso di percorso regolare, in quest'ultimo anno si sono diplomati) e che hanno frequentato le scuole della provincia; per ogni individuo, viene tracciata la storia del percorso scolastico, quindi esiste un record per ogni istante temporale considerato (si noti che per storia si intende quella all'interno della provincia, in quanto l'uscita dalla scuola bolognese può significare uscita da scuola in generale, ma anche trasferimento in una scuola di altra provincia; quest'ultima informazione non è contenuta nel dataset). Ogni record contiene informazioni statiche riguardo al luogo e alla data di nascita, al sesso, alla residenza, alla cittadinanza, inoltre informazioni dinamiche sulla scuola frequentata in ciascun anno scolastico, la classe ed anche sulla promozione, la ripetenza, l'eventuale non promozione o l'abbandono della scuola. Poiché l'obiettivo dell'analisi è l'esito scolastico, l'istante temporale che risulta essere di effettivo interesse è il termine di ciascun anno scolastico: sono rientrati nell'analisi soltanto i dati riferiti alla fine di ciascun anno (luglio), senza considerare gli altri due istanti temporali rilevati. Le onde temporali analizzate sono quindi 6:

Luglio 2003	Luglio 2006
Luglio 2004	Luglio 2007
Luglio 2005	Luglio 2008

Uno studente che ha terminato regolarmente il proprio ciclo di studi frequentava la classe quinta a luglio 2007 (anno scolastico 2006/07), mentre aveva terminato il suo primo anno di scuola superiore nel luglio 2003 (anno scolastico 2002/03). La disponibilità di dati ripartiti su così tanti anni scolastici, consente di verificare se anche quegli studenti che hanno un leggero ritardo, nel senso che hanno ripetuto una classe una sola volta (la stima è che siano intorno al 10%), hanno poi raggiunto il diploma mantenendo un solo anno di ritardo.

In questo caso, dato l'obiettivo iniziale di indagare sulle motivazioni, specialmente quelle di carattere collettivo (per risalire alle motivazioni individuali occorrerebbe un indagine ad hoc, con intervista di un campione di studenti), che portano alla deviazione rispetto ad un percorso scolastico regolare, lo si è dovuto adeguare in realtà ai dati disponibili, utilizzandoli al meglio per estrapolare informazioni utili. Nel caso in esame, come già spiegato, i dati sono raccolti e detenuti da un ente pubblico, la cui finalità istituzionale principale non è quella di ricavare analisi statistiche sui comportamenti, quanto invece quella di indagare le singole situazioni critiche. Si sono dovute perciò ricercare le spiegazioni dell'andamento del percorso scolastico soltanto tra le informazioni individuali disponibili, senza poter aggiungere altre variabili, anche se in teoria è possibile che concorrano alla spiegazione del fenomeno. Per questo motivo gli errori che risulteranno dal modello non sono ulteriormente specificabili, cioè è senz'altro possibile che contengano altre componenti che potrebbero essere formalizzate (ad esempio informazioni sul contesto famigliare degli studenti o sulla zona in cui effettivamente vivono), ma in questa sede non è possibile farlo perché non vi è la disponibilità di ulteriori informazioni di base.

1.4 DESCRIZIONE DEL METODO DI ANALISI

Visto che l'obiettivo del presente studio è l'analisi di un andamento nel tempo di esiti scolastici individuali, lo strumento ritenuto, alla fine, maggiormente idoneo allo scopo è il *Latent Curve Model* (LCM). Dopo una preventiva analisi esplorativa del dataset, si è proceduto a concretizzare la variabile oggetto di studio. In un primo tempo si è pensato di studiare il comportamento, nei diversi anni scolastici singolarmente considerati, della variabile d'interesse "Esito scolastico", in termini di promozione o meno: variabile dicotomica con valori 1 (promosso) e 0 (non promosso) (non era disponibile l'informazione riguardo alla valutazione). Data la caratterizzazione di tale variabile, si è scelto di utilizzare, come modello maggiormente idoneo, quello logistico. Con tale approccio, però, non si teneva conto dell'andamento temporale dell'esito nei singoli individui, ma semplicemente si stimava un modello per ogni anno scolastico e classe. Con una complicazione aggiuntiva: raggruppando gli studenti per anno scolastico si otteneva un insieme di frequentanti classi diverse, mentre con il raggruppamento per classe si otteneva un insieme di frequentanti la stessa classe in anni scolastici diversi (uno stesso studente poteva così entrare due volte, se in una classe non era stato promosso). Al fine di cogliere il comportamento prevalente nel tempo e di valutare gli scostamenti da tale andamento, si è scelto un secondo approccio, maggiormente costoso dal punto di vista applicativo, ma sicuramente migliore dal punto di vista dei risultati: il LCM, strumento particolarmente adatto ai dati di tipo longitudinale, come quelli qui disponibili. Nella sua forma più semplice, il modello fornisce una descrizione delle differenze tra gli individui al trascorrere del tempo, attraverso le medie di una funzione lineare deterministica[22]. I parametri del modello sono a loro volta variabili casuali e i loro momenti (la media e la matrice di varianze e covarianze) caratterizzano le differenze individuali nella crescita (dove crescita – *growth* - è intesa nel senso di variazione, positiva o negativa). Se poi vi è la possibilità di misurare anche altre variabili, correlate con quella oggetto d'interesse, si può tentare di risalire alle cause che hanno determinato quell'evoluzione: non solo "Quanto è stato il cambiamento?", ma anche "Cosa ha determinato il cambiamento?".

In generale, data una variabile obiettivo e un insieme di variabili usate come regressori (e di cui si vuole testare il legame con la variabile obiettivo), se sono disponibili dati su una popolazione in un certo istante temporale, è possibile cercare il legame tra la variabile oggetto d'interesse e le altre variabili; questo legame viene misurato in termini di un modello (il modello di regressione) che permette di estendere i risultati ad altri individui, non osservati, provenienti dalla medesima popolazione. Nel caso in cui, invece, siano disponibili dati di tipo longitudinale, sempre provenienti dalla stessa popolazione, risulta possibile esaminare il comportamento dei singoli individui (in termini di medie) durante l'intervallo di tempo considerato e, allo stesso tempo, ricercare le differenze individuali in quel comportamento. Focalizzandosi sulla variabile obiettivo, il fondamento della presente analisi è l'idea che, innanzitutto, esista un modello che mette insieme il fenomeno ed anche il tempo. Vi sono, in particolare, variabili non direttamente osservate (variabili latenti) le cui realizzazioni sono i parametri del modello, e proprio questi vengono stimati sulla base delle osservazioni empiriche.

Una valutazione primaria, che è d'obbligo effettuare, riguarda la forma della relazione tra la variabile obiettivo e il tempo. Se si hanno informazioni a priori, si può costruire il modello sulla base di tali ipotesi; se invece non

[22] Conor V. Dolan et al. "Regime Switching in the Latent Growth Curve Mixture Model"

dispone di tali informazioni, si può procedere a tentativi, sempre però iniziando dalla forma più semplice di legame, che è quello lineare.

Se si sceglie il modello lineare non condizionato, le variabili latenti sono l'intercetta e la pendenza della linea retta, posata sul piano cartesiano che ha i valori della variabile obiettivo (Y) in ordinata e i valori del tempo (t) in ascissa. Le realizzazioni delle variabili latenti sono tante quante il numero di individui (N). L'intercetta rappresenta il valore della variabile Y al tempo 0, all'inizio del processo. La media tra gli individui della variabile latente "intercetta" rappresenta il livello medio della variabile obiettivo nel primo degli istanti di tempo considerati. L'intercetta è il valore iniziale[23] atteso per ogni individuo di quella popolazione. La pendenza rappresenta invece la crescita del fenomeno nel tempo. La pendenza media tra gli individui è la crescita attesa, dove la connotazione di crescita è nel senso di cambiamento, sia positivo che negativo, atteso in un istante temporale rispetto al precedente.

Le variabili casuali "intercetta" e "pendenza" non vengono direttamente osservate, ma vengono stimate dal modello a partire dai dati osservati; in quanto variabili casuali, è possibile ricercare i fattori che hanno determinato il loro comportamento. Il modello può allora essere trasformato in condizionato, aggiungendo alcuni fattori, le covariate o predittori, che spiegano la variabilità di intercetta e pendenza.

L'intercetta e la pendenza sono esse stesse, come già detto, variabili casuali: è plausibile considerare che possa esistere una relazione tra loro, quindi che si possa stimare la loro covarianza (il tasso di variazione dipende in qualche modo dal valore iniziale?).

Un vantaggio del modello qui descritto è che esso rappresenta e descrive non solo il comportamento del gruppo, ma anche la variabilità individuale (la causalità delle differenze tra gli individui))[24].

Nell'ultimo decennio, l'uso del *Latent Curve Model* (LCM) si è particolarmente diffuso[25], grazie soprattutto alla disponibilità di dati longitudinali. Il crescente interesse verso questa tecnica è dovuto al fatto che questa metodologia pone una grande enfasi sulle differenze individuali nel cambiamento.

Si è ritenuto opportuno utilizzare il modello a curva latente, in quanto offriva la possibilità di ricostruire il percorso scolastico degli studenti e di ricercare relazioni di causalità tra l'esito scolastico ed alcune variabili osservate sugli individui.

Nel presente lavoro, è stato utilizzato anche il modello logistico: la variabile obiettivo dello studio era così l'esito scolastico in termini di promozione (valore 1) o non promozione (valore 0) e si sono potute inserire le variabili esplicative, al fine di saggiare la relazione tra queste e l'esito. Tuttavia tale applicazione non ha reso possibile la ricostruzione di un percorso scolastico, in quanto, come spiegato nei capitoli successivi, si è reso necessario stimare un modello separatamente per ciascun anno scolastico e classe frequentata.

Nella ricerca di una variabile obiettivo che incorporasse l'informazione sull'esito scolastico di ogni anno, in quale classe, e quella sulla regolarità, con valori dipendenti dall'istante temporale di riferimento, si è giunti alla variabile "*Times Promoted*". Dopo aver costruito una variabile dicotomica che indica se il singolo studente, in ciascun anno scolastico, è stato promosso oppure no, se ne è calcolata una combinazione lineare, in modo da incorporare in una sola informazione sia la classe frequentata che l'esito formativo. La variabile ottenuta è di

[23] Acock "Latent growth curve model: a gentle introduction"
[24] Witta "Latent growth model of cognition in the elderly"
[25] Bollen "On the origins of latent curve model"

tipo numerico (non dicotomica) e indica il numero di volte in cui lo studente è stato promosso, dato l'anno scolastico considerato (così, ad esempio, uno studente che ha frequentato la terza e a fine anno è stato promosso ha il valore 3, mentre uno studente che ha frequentato la quinta e a fine anno non è stato promosso ha il valore 4; occorre ricordare che i tempi considerati sono riferiti a fine anno scolastico). Formalizzando: $TP_t = \sum_{i=1}^{t} 1_i$, dove 1_i è la variabile dicotomica che assume valore 1 se lo studente è stato promosso al tempo *i* e valore 0 se lo studente non è stato promosso al tempo *i*. Quindi uno studente con un corso di studi regolare avrà una sequenza di valori, nei diversi istanti temporali, del tipo: [1,2,3,4,5]. In realtà, quindi, la variabile appena descritta comprende anche l'informazione sulla regolarità, inoltre contiene l'informazione sulla classe frequentata (un valore pari a 3, ad esempio, indica che lo studente ha sicuramente frequentato e positivamente terminato la terza classe). Un limite di tale trasformazione è che uno stesso valore può avere significati diversi a seconda dell'istante temporale in cui si verifica; ad esempio, infatti, il valore 3 nel terzo istante temporale indica una promozione in classe terza, mentre il valore 3 nel quarto istante temporale può significare una bocciatura in quarta. Per questo motivo, anche durante la fase di interpretazione dei risultati, occorre sempre associare i valori della variabile obiettivo con l'istante temporale di riferimento. *Times Promoted* è una combinazione lineare di variabili dicotomiche, che ha il vantaggio di rappresentare la situazione di ogni studente nel tempo, in quanto a promozione, regolarità e ultima classe non frequentata quanto piuttosto terminata con successo.

1.5 STATISTICHE DESCRITTIVE DEL DATASET OGGETTO DI STUDIO

La coorte analizzata, come già premesso, è quella degli studenti nati nel 1988 e presenti nel sistema scolastico della provincia di Bologna (scuole statali e scuole non statali, esclusi i corsi privati di recupero di anni scolastici) in uno o più anni scolastici tra il 2002/03 e il 2007/08. Il numero di studenti presenti nel database è 5.939: 4.385 (74%) di questi sono rimasti nelle scuole bolognesi dal 2002/03 al 2006/07, mentre gli altri sono usciti prima del 2006/07 oppure entrati dopo il 2002/03; 1.008 (17%) studenti sono rimasti nelle scuole bolognesi dal 2003/04 al 2007/08 (sono per lo più coloro che erano già in ritardo scolastico in terza media); 200 studenti (3,4%) sono usciti dalle scuole di Bologna dopo il 2002/03. Non è possibile sapere se questi sono rimasti nel sistema scolastico, trasferendosi in altre province, o se sono entrati in percorsi formativi o ancora se hanno abbandonato il sistema formativo (*drop out*).

Una volta disponibile il dataset contenente i dati individuali (fonte: Provincia di Bologna, Osservatorio sulla scolarità), vi è stato bisogno di un lavoro preliminare di controllo che ha consentito di apportare alcune correzioni nei casi di incongruenze palesi. Essendovi informazioni relative a più di un istante temporale per ciascun anno scolastico (settembre, febbraio e luglio, come in precedenza accennato), si è riusciti a rendere completi i record sull'ultima rilevazione, quella di luglio, che è poi l'unica utilizzata come obiettivo dell'analisi. Il dataset era già nella forma *person period*, cioè un record per ogni individuo ed istante temporale, come richiesto per l'analisi tramite il LCM. Si è reso necessario standardizzare l'informazione riguardo alla tipologia di scuola, ottenendo una notazione sintetica (ad esempio "Liceo") e l'altra analitica (ad esempio "Liceo Classico"). Per ciascun individuo, era disponibile sia l'informazione sulla cittadinanza che sul luogo di nascita; si

è scelto di tener conto della sola cittadinanza, come variabile esplicativa per il modello poi costruito. Ciò in quanto vi sono studenti nati all'estero che probabilmente sono stati adottati, quindi possono considerarsi a tutti gli effetti italiani; soltanto gli studenti che hanno mantenuto la cittadinanza straniera è più probabile che vivano in famiglie non italiane e che quindi si trovino in un contesto che è realmente diverso da quello italiano, fatto che porta più probabilmente a conseguenze anche sul percorso scolastico e formativo.
Una volta controllato e corretto il database, è stato possibile analizzare la struttura dei dati; si è partiti da alcune analisi descrittive. Il numero di record analizzati è 26.292, riferiti agli anni scolastici dal 2002/03 al 2006/07. Il numero di individui analizzati è, come già accennato, 5.939. La distribuzione degli individui nei vari istanti temporali è quella riportata in Tabella 23: vi sono 4.385 studenti (il 73,8% del totale) rimasti nelle scuole bolognesi per 5 anni scolastici, invece 605 vi sono rimasti per 4 anni e così via. Vi sono da aggiungere anche 846 studenti (il 14% del totale) che sono rimasti nelle scuole bolognesi per 6 anni (ma non è disponibile l'informazione su quanti di loro hanno terminato il percorso di studi, giungendo al diploma).

Tabella 23 – Distribuzione degli studenti nei diversi istanti temporali

Numero di istanti temporali	Numero di individui con quel numero di istanti temporali	% di individui con quel numero di istanti temporali
1	315	5,3
2	270	4,6
3	364	6,1
4	605	10,2
5	4.385	73,8

Come già visto, circa il 10% degli individui presenta meno di 3 rilevazioni temporali, cioè il minimo necessario per stimare un modello a curva latente (che è l'obiettivo della presente analisi); nonostante ciò ci si aspetta di riuscire a stimare il modello (senza escludere tali individui) in quanto la percentuale di questi casi è comunque bassa ed il numero assoluto di individui con 3 o più istanti temporali è elevato (oltre 5.000).
L'obiettivo dell'analisi è quello di studiare gli esiti scolastici ed anche, se possibile, la regolarità, cercando di risalire ai fattori che determinano il comportamento degli studenti in termini di insuccesso o addirittura di uscita precoce dal sistema scolastico. Una prima idea per studiare l'abbandono precoce della scuola è stata quella di condurre un'analisi di sopravvivenza: dato il contingente iniziale di studenti nati nel 1988 che frequentavano la prima superiore nell'anno scolastico 2002/03, è seguirli nel tempo e calcolare la probabilità di abbandono della scuola in ciascun anno scolastico; analogamente si può studiare il fenomeno della promozione, considerando come uscite dal contingente i casi di studenti non promossi. Tuttavia vi sono elementi fondamentali di cui occorre tener conto: l'uscita dal sistema scolastico bolognese può non significare uscita dal sistema scolastico (uno studente può trasferirsi in altra provincia e tale informazione non è presente nel database, oltre al fatto che vi sono studenti che escono una volta dal percorso scolastico, ma poi vi rientrano uno o due anni dopo, quindi non possono essere considerati alla stregua di coloro che in effetti abbandonano la scuola); inoltre non è presente nel database l'informazione sugli studenti che nel 2007/08 erano ancora a scuola, cioè non sappiamo se quelli che non si sono diplomati in quell'anno (pur pochi) sono rimasti a scuola oppure no; infine occorre considerare che l'uscita dal sistema scolastico non è da considerarsi come uscita dal sistema formativo, che invece la legge prevede sia un'alternativa perseguibile per i ragazzi che hanno compiuto i 15 anni di età.

Un importante accorgimento durante lo studio di questi dati è quello di evitare l'errore di considerare gli individui del database come provenienti dalla popolazione di ragazzi e ragazze nati (o abitanti) nella provincia di Bologna: essi sono coloro che sono iscritti nelle scuole di Bologna; i nati e coloro che a Bologna risiedono fanno parte, in realtà, di un altro contingente (si intende, per la maggior parte coincidente con quello dei frequentanti, ma non c'è perfetta corrispondenza).

Per quanto riguarda la variabile oggetto di studio, precedentemente descritta, è opportuno in questa sede descrivere il significato dei valori che si possono riscontrare nei diversi individui. Uno studente che ha 0 come valore di *Times Promoted* non è stato promosso nella classe prima in quell'anno scolastico. Uno studente che lascia precocemente la scuola (date le limitazioni prima descritte, soprattutto il fatto che lasciare la scuola bolognese non significa necessariamente uscire dal percorso scolastico) riporta un valore massimo di *Times Promoted* <5. Vi sono casi particolari di studenti che hanno iniziato il loro percorso scolastico fuori dalla provincia di Bologna , ma vi sono entrati in seguito: questi compaiono ad un certo istante temporale successivo al primo considerato con un valore iniziale di *Times Promoted* >1, la loro sequenza di valori non è completa, ma è comunque possibile che abbiano 5 come valore massimo. Vi sono anche pochi casi di studenti che non sono stati promossi in un certo anno scolastico, ma che poi hanno frequentato con successo due anni in uno; a conseguenza di ciò nella sequenza di valori di *Times Promoted* relativa a quell'individuo vi sarà un "gradino". Un caso analogo è quello degli studenti che, specie dopo un esito negativo, si trasferiscono, quindi mancano dal dataset per uno o più anni scolastici, ma poi tornano, anche magari dopo aver completato due classi in uno stesso anno; in questo caso la sequenza di *Times Promoted* sarà ancora più accidentata. Così si possono riscontrare sequenze del tipo [1,2,,,5], indicanti che lo studente si è trasferito in terza, frequentando con successo le altre classi fuori provincia, ma poi vi è rientrato e ha concluso con successo anche la quinta classe. Vi sono anche pochi casi di studenti che hanno cambiato tipologia di scuola da un anno al successivo, quindi pur essendo stati promossi alla fine di un anno scolastico, hanno dovuto ripetere la stessa classe perché il tipo di scuola era diverso (ciò avviene soprattutto se il cambiamento si verifica nelle ultime classi); la sequenza di valori di *Times Promoted*, in questo caso, è del tipo: [1,2,3,4,4,5]. Come già anticipato, vi sono studenti che hanno iniziato la scuola in anticipo di un anno rispetto agli altri, quindi questi in generale frequentavano la classe seconda nel 2002/03 (e non la prima); questi, nel caso di percorso regolare, hanno conseguito il diploma nel 2005/06 e non nel 2006/07 come gli altri. La sequenza di valori di *Times Promoted*, in questo caso, è del tipo: [2,3,4,5,,] (nel caso regolare).

Nel database vi sono molte variabili che possono essere (e verranno) usate come predittori nel modello. La prima è certo la cittadinanza: una variabile dicotomica (chiamata "*Foreign*") che assume valore 0 se la cittadinanza è italiana e 1 se straniera. Si tratta di un predittore invariante nel tempo. Non è stato possibile introdurre nei modelli anche la distinzione tra diverse nazionalità, magari costruendo modelli appositi per i soli cittadini non italiani, dato lo scarso (ai fini della potenziale ulteriore suddivisione in sottogruppi) numero di studenti con cittadinanza non italiana. Un'altra variabile introdotta è il sesso degli studenti. Inoltre si è utilizzata la tipologia di scuola frequentata (che di per sé può variare di anno in anno). Vi sono 20 tipologie diverse di scuola, riassunte in Tabella 24.

Tabella 24 –Tipologie di scuola

Liceo Artistico	ITC – Istituto Tecnico per il Commercio
Liceo Classico	ITCG – Istituto Tecnico Commerciale e per Geometri
IPA – Istituto Professionale per l'Agricoltura	ITG – Istituto Tecnico per Geometri
IPCT – Istituto Professionale per il Commercio e il Turismo	ITI – Istituto Tecnico Industriale
IPIA – Istituto Professionale per l'Industria e l'Artigianato	ITI- Scientifico – Istituto Tecnico Industriale e Liceo Scientifico
IPSAR – Istituto Professionale per i Servizi Alberghieri	Liceo Linguistico
IPSOC – Istituto Professionale per i Servizi Sociali	Liceo Scientifico
Istituto d'arte	Liceo Scienze Sociali
ITA – Istituto Tecnico per l'Agricoltura	Liceo Scienze Sociali e Scientifico
ITAER – Istituto Tecnico Aeronautico	Liceo Scienze Sociali e Linguistico

Le 20 tipologie descritte in Tabella 24 possono essere raggruppate in 4 tipologie principali, che sono poi quelle che verranno utilizzate nei modelli, soprattutto per non rischiare di suddividere gli individui in sottogruppi troppo poco numerosi.

Tabella 25 – Tipologie principali di scuola

Artistici
Licei
Professionali
Tecnici

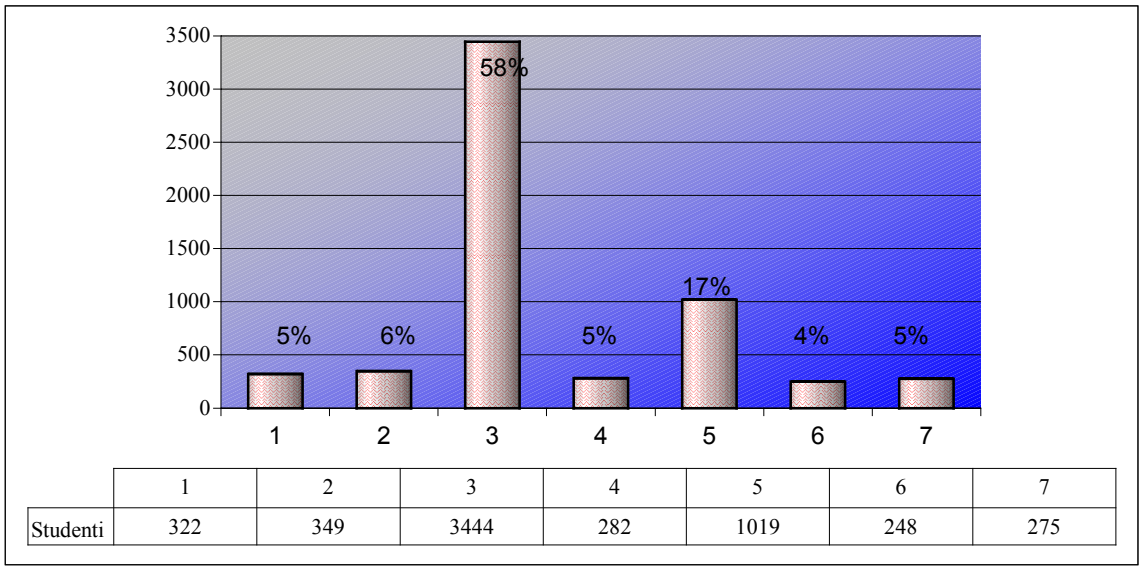
In Tabella 25, la categoria "Licei" comprende il classico, lo scientifico, il liceo di scienze sociali, lo scientifico scienze sociali, il linguistico scienze sociali e l'ITI Scientifico; "Artistici" comprende il liceo artistico e l'istituto d'arte; "Professionali" e "Tecnici" comprendono il primo le tipologie indicate come "Istituto Professionale" ed il secondo le tipologie indicate come "Istituto Tecnico". Il caso dell'ITI Scientifico (che è un indirizzo del Tecnico che deve essere considerato molto più vicino al liceo che all'istituto tecnico) è particolare e viene appunto annoverato tra i Licei.

Un'altra informazione utile per spiegare la variabile obiettivo è se lo studente abbia cambiato scuola durante il percorso scolastico oppure se sia rimasto nella stessa scuola (variabile dicotomica invariante nel tempo). In generale, si suppone sia più probabile che chi cambia scuola ottenga risultati peggiori rispetto a chi rimane, anche perché il cambiamento è rivolto a istituti di solito meno impegnativi. Tale variabile verrà di fatto utilizzata soltanto a fini descrittivi, in quanto non è risultata significativa per spiegare la variabile *Times Promoted*.

Tabella 26 – Ambiti territoriali secondo la ripartizione della Provincia di Bologna

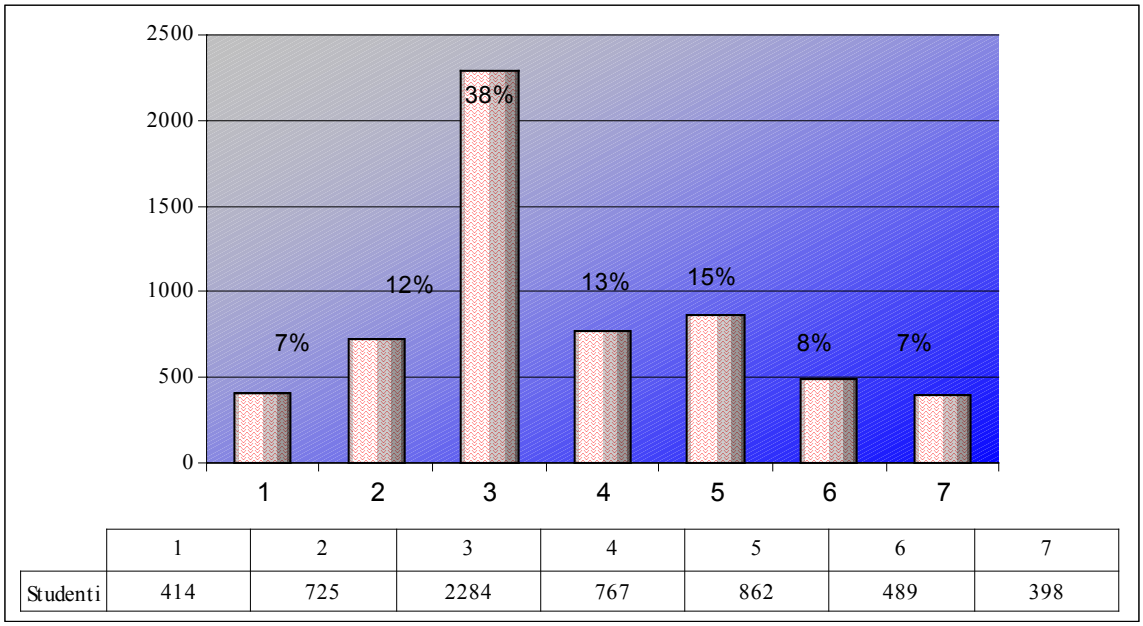
Ambito	Comuni compresi
1	San Giovanni in Persiceto, Crevalcore, Sala Bolognese, Calderara, Angola, Sant'Agata Bolognese.
2	Zola Predosa, Bazzano, Casalecchio di Reno, Castello di Serravalle, Crespellano, Monte San Pietro, Monteveglio, Sasso Marconi, Savigno
3	Bologna città
4	Argelato, Baricella, Benivoglio, Budrio, Castello d'Argile, Castel Maggiore, Castenaso, Galliera, Granarolo, Malalbergo, Minerbio, Molinella, Pieve di Cento, San Giorgio di Piano, San Pietro in Casale
5	Borgo Tossignano, Casalfiumanese, Castel del Rio, Castel Guelfo, Castel San Pietro, Dozza, Fontanelice, Imola, Medicina, Mordano
6	San Lazzaro di Savena, Ozzano Emilia, Pianoro, Monterenzio, Monghidoro, Loiano
7	Porretta Terme, Vergato, Granaglione, Castel di Casio, Grizzana Morandi, Lizzano in Belvedere, Castiglione dei Pepoli, Camugnano, Castel d'Aiano, Gaggio Montano, Monzuno, Marzabotto, San Benedetto Val di Sambro.

Per una descrizione del contesto, è utile vedere la distribuzione degli studenti nei vari comuni della provincia di Bologna, tenendo conto del fatto che nel comune di Bologna vi sono ovviamente molti più studenti che in ogni altro comune e sono presenti quasi tutte le tipologie di scuola, mentre negli altri comuni le scelte sono più ridotte. La Provincia di Bologna distingue il territorio in ambiti territoriali, in particolare in Tabella 26 è riportata tale ripartizione. In Grafico 1 è mostrata la distribuzione degli studenti nelle scuole dei diverse Comuni (si considera l'ultima scuola frequentata), raggruppate a seconda dell'ambito territoriale. In Grafico 2 è mostrata la distribuzione degli studenti nei diversi ambiti territoriali, a seconda della residenza anagrafica (si considera l'ultimo dato sulla residenza).



	1	2	3	4	5	6	7
Studenti	322	349	3444	282	1019	248	275

Grafico 1 – Ripartizione percentuale degli studenti per ambito di appartenenza dell'ultima scuola frequentata



	1	2	3	4	5	6	7
Studenti	414	725	2284	767	862	489	398

Grafico 2 – Ripartizione percentuale degli studenti per ambito di appartenenza in quanto a ultima residenza anagrafica

Come dato di contesto, si può vedere la distribuzione delle percentuali di alunni con corso di studi regolare nei diversi comuni della provincia, sempre raggruppati per ambiti, ma occorre sempre tener conto del fatto che le

tipologie di scuola presenti nel territorio sono diverse da un comune all'altro. Comunque anche questa è una conferma del fatto che la tipologia di scuola frequentata influisce sull'esito, in particolare sulla regolarità.
È interessante notare che, considerando i 647 studenti che hanno terminato con successo la quinta classe nel 2006/07, rappresentanti l'11% del contingente studiato, questi si distribuiscono differentemente rispetto agli altri nei diversi ambiti della provincia. Considerando la scuola frequentata, infatti, il 61% di essi frequenta una scuola di Bologna città (contro il 58% dell'intero contingente), il 7% frequenta una scuola dell'ambito 1 e solo il 3% frequenta una scuola dell'ambito 7; le percentuali di studenti con successo variano inoltre a seconda dell'ambito di appartenenza della scuola (come si nota anche in Grafico 3): queste sono rispettivamente, nei 7 ambiti, pari a [15%; 8%; 11%; 8%; 10%; 13%; 7%]. Le analoghe percentuali calcolate suddividendo gli studenti per ambito di residenza mostrano un andamento maggiormente regolare: [12%; 11%; 10%; 12%; 10%; 14%; 9%].

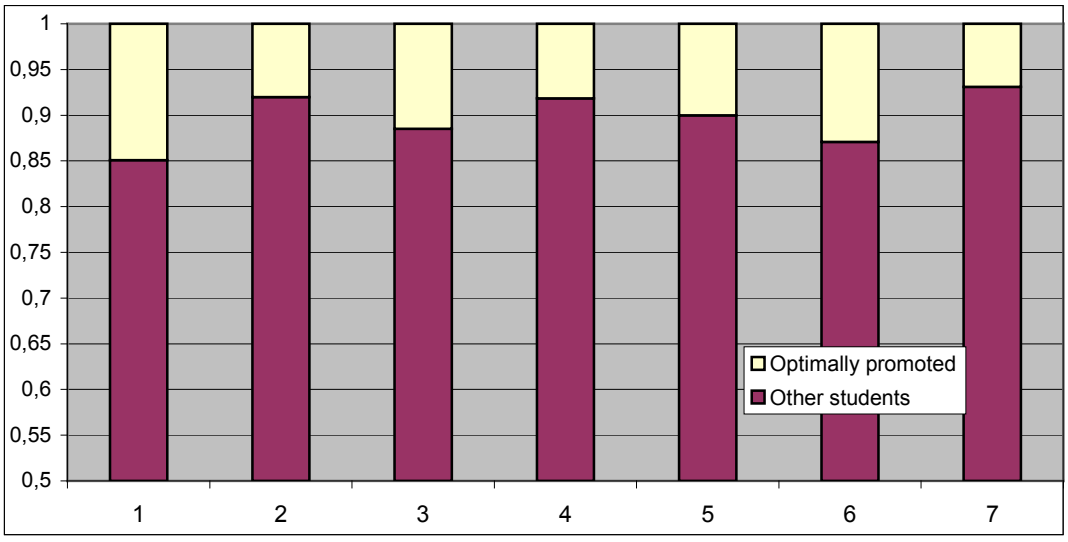


Grafico 3 – Composizione degli studenti frequentanti le scuole nei vari ambiti territoriali per ottimamente promossi (diplomati in cinque anni) e non.

Un esempio evidente della differenza, in quanto a probabilità di avere un percorso di studi irregolare tra gli studenti che frequentano i diversi tipi di scuola, è dato dal comune di Castel San Pietro Terme: qui vi sono soltanto istituti tecnici e professionali e la percentuale di alunni regolari è la più bassa della provincia. Ciò fa immediatamente pensare che in questi tipi di istituti sia meno probabile trovarsi nella condizione di regolarità, ma esaminando le percentuali relative anche agli altri comuni si vede che, anche laddove le probabilità di avere un percorso regolare sono simili (ad esempio confrontando il comune di Bologna con il comune di San Lazzaro di Savena), le percentuali di studenti che frequentano gli istituti professionali e tecnici sono diverse. È chiaro che sulla diversità tra i gruppi di studenti che frequentano le scuole dei vari comuni influisce sia il fattore tipo di scuola che il fattore legato al comune.

La distribuzione dei 5.939 studenti per classi ed anni scolastici è quella descritta in Tabella 27.
In Tabella 27 si può notare un aumento consistente di studenti tra il secondo ed il primo anno scolastico considerato: questo si spiega con le nuove entrate in prima classe di studenti provenienti dalla terza media (erano già stati bocciati alle medie, quindi sono entrati nella scuola superiore con un anno di ritardo rispetto ai regolari).

Tabella 27 – Distribuzione degli studenti nati nel 1988, per diverse classi e anni scolastici

	2002/03	2003/04	2004/05	2005/06	2006/07	promossi	2007/08	promossi
1	5.277	916	256	51	5	0	3	0
2	89	4.548	954	283	47	17	7	2
3		95	4.144	1.000	337	218	66	34
4			87	3.722	893	671	304	174
5				83	3.505	3.377	772	646
Totali	5.366	5.559	5.441	5.139	4.787	4.283	1.152	856

4.534 dei 5.277 studenti (86%) che frequentavano la classe prima nell'anno scolastico 2002/03 sono stati promossi, quindi almeno 14 (4.548 – 4.534) studenti sono entrati nelle scuole superiori di Bologna nell'anno scolastico 2003/04 (ma potrebbero anche esserci altri studenti che sono usciti e quindi altri che sono entrati, come vi sono studenti che nel 2002/03 frequentavano la classe seconda e non sono stati promossi). La percentuale di studenti promossi, considerando soltanto coloro che frequentavano la classe seconda nel 2002/03, è del 93% (mentre la percentuale di studenti promossi in seconda classe è in generale attorno all'86%): sono quegli studenti che hanno iniziato in anticipo il ciclo di studi (entrando un anno prima degli altri in prima elementare) ed evidentemente hanno raggiunto risultati in generale migliori degli altri. Calcolando le stesse percentuali negli anni scolastici successivi, le percentuali di promozione risultano del 94% (in classe terza), del 96% (in classe quarta) e del 99% (in classe quinta), mentre le stesse percentuali riferite agli studenti che hanno iniziato non anticipatamente il ciclo di studi sono del 90% (in classe terza), del 94% (in classe quarta) e del 96% (in classe quinta). Si può notare che la percentuale di studenti promossi è più bassa in prima, ma poi aumenta nelle classi successive. Lo stesso aumento si verifica sia per quegli studenti che hanno iniziato regolarmente il ciclo di studi che per coloro che lo hanno iniziato in anticipo. Per quanto riguarda i diplomati, si ha che la percentuale di studenti che hanno conseguito il diploma nel 2006/07 (quindi al termine di un ciclo di studi regolare), su quelli che frequentavano la classe prima nel 2002/03, è del 64% (3.377 studenti su 5.277); occorre però tener presente che questa percentuale non considera coloro che si sono poi diplomati negli anni successivi, quindi in ritardo rispetto al ciclo regolare. Considerando poi tutti i 5.366 studenti che erano a scuola nel 2002/03, si può dire che il 76% di essi è giunto al diploma al più in sei anni. Rimane un 24% di non diplomati (con al più il ritardo di un anno), che è in linea con la percentuale di dispersi dalla prima classe alla quinta, già trovata nei dati regionali e provinciali. Già da queste poche considerazioni sui dati aggregati, si evince la difficoltà a trarre considerazioni precise, visto che tali dati celano informazioni in realtà dettagliate e sono frutto di compensazioni. In questo caso, però, risultano disponibili le relative informazioni disaggregate sui singoli studenti, quindi si possono fare analisi maggiormente dettagliate, fino a tracciare il percorso scolastico di ogni singolo studente. Nell'anno scolastico 2002/03 vi erano 5.366 studenti nelle scuole secondarie di secondo grado della provincia di Bologna: 4.385 (82%) di essi erano a scuola nell'anno scolastico 2006/07; 4.011 (75%) di essi sono stati promossi nell'anno scolastico 2006/07; 3.406 (63%) di essi si sono diplomati nell'anno scolastico 2005/06 o nel

2006/07; 846 (16%) di essi erano ancora a scuola nell'anno scolastico 2007/08 (erano quindi in ritardo scolastico); 525 (10%) di essi si sono diplomati nell'anno scolastico 2007/08 (con un anno di ritardo rispetto alla situazione regolare).

Considerato che uno studente ha seguito un corso di studi regolare se, durante il periodo considerato, è stato promosso per 5 volte, terminando il proprio corso di studi in 5 anni (giungendo quindi al diploma), la distribuzione degli studenti qui chiamati regolari è quella nel seguito illustrata. È importante notare che vi sono alcuni studenti che hanno iniziato la scuola in ritardo, frequentando la classe prima nell'anno scolastico 2004/05; questi, anche se sono stati sempre promossi, comunque frequentavano la terza classe nell'anno scolastico 2006/07. Tali studenti non rientrano tra quelli considerati regolari, anche perché non vi sono sufficienti informazioni riguardo al completamento del loro percorso di studi.

Come mostrato in Grafico 4, gli studenti che hanno conseguito un corso regolare degli studi, detti appunto "regolari", sono oltre la metà, cioè 4.023 (il 68% del totale), mentre gli altri sono 1.916.

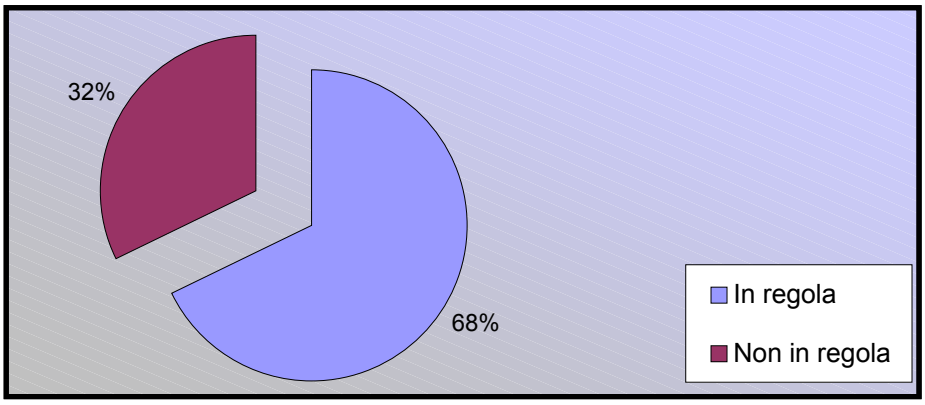


Grafico 4 – Ripartizione percentuale degli studenti tra chi ha avuto un percorso di studi regolare e chi no

Riportati in Grafico 5, gli studenti con cittadinanza non italiana sono 361 (6%), mentre coloro che hanno la cittadinanza italiana sono 5.578 (94%). Risulta evidente in grafico che un percorso di studi regolare caratterizza maggiormente gli studenti italiani rispetto ai cittadini stranieri.

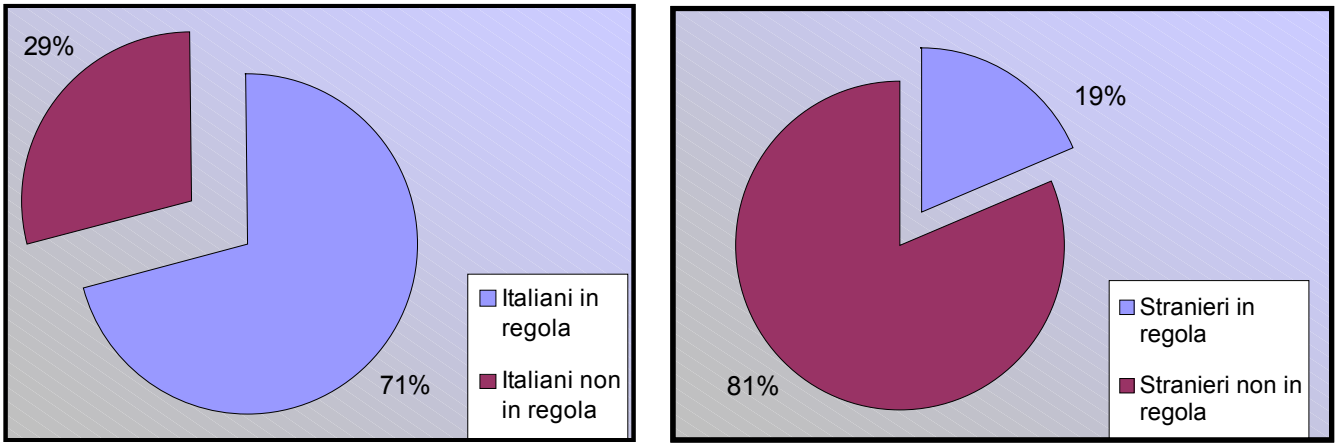


Grafico 5 – Ripartizione percentuale degli studenti tra chi ha avuto un percorso regolare e chi no: italiani e stranieri

Le studentesse sono 2.893 (48,7%), mentre i ragazzi sono 3.046 (51,3%). La loro distribuzione in base alla regolarità scolastica, riportata in Grafico 6, mostra una netta propensione delle ragazze alla regolarità rispetto ai colleghi maschi.

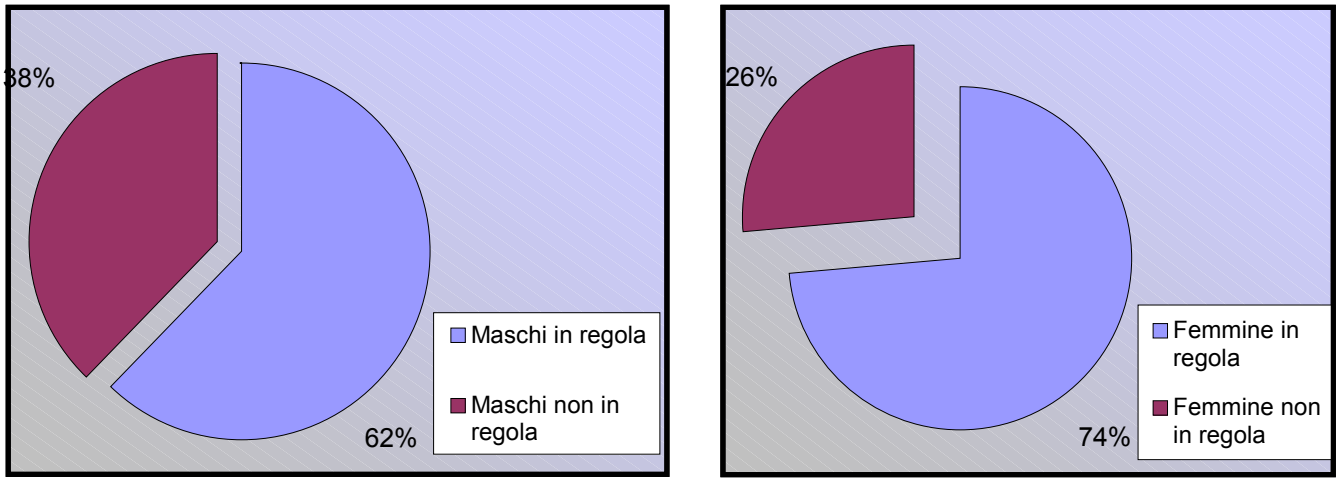


Grafico 6 – Ripartizione percentuale degli studenti tra coloro che hanno avuto un percorso regolare e gli altri: maschi e femmine

Si noti che gli studenti con cittadinanza straniera sono così composti: 187 (52%, il 6,5% di tutte le ragazze) femmine e 174 (48%, il 5,7% di tutti i ragazzi) maschi; le ragazze con un percorso regolare sono il 21% di tutte le ragazze con cittadinanza non italiana, mentre i ragazzi con un percorso regolare sono il 16% di tutti i ragazzi con cittadinanza non italiana. Ciò a ulteriore riprova della propensione delle ragazze ad avere esiti scolastici migliori dei ragazzi, non solo per i cittadini italiani.

Gli studenti che almeno una volta sono stati bocciati sono 1.986 (33%), mentre gli studenti che sono sempre stati promossi sono 3.953 (67%); in Grafico 7 risulta evidente la stretta relazione tra regolarità e promozione.

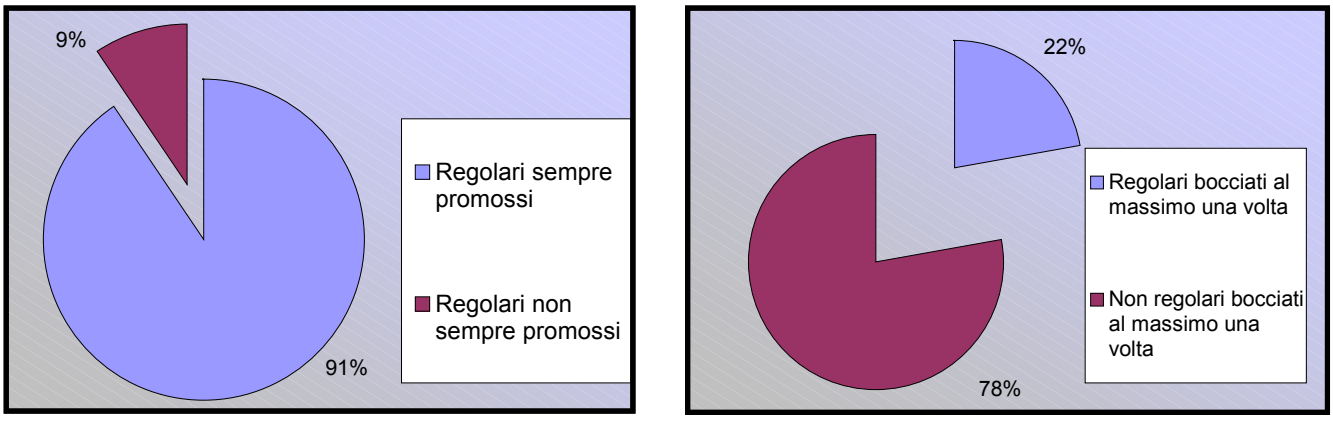


Grafico 7 – Ripartizione percentuale degli studenti tra coloro che hanno avuto un percorso regolare e gli altri: quelli sempre promossi e gli altri

La distinzione tra regolarità e promozione è in realtà sottile: gli studenti che sono sempre stati promossi e che però non hanno avuto un percorso regolare sono quelli che hanno lasciato la scuola (meglio, in questo caso la scuola bolognese, perché non si tratta di una realtà chiusa) prima di terminare la classe quinta. Quegli studenti che sono stati bocciati una sola volta e che però si può dire abbiano ugualmente un percorso regolare sono quei

pochi che hanno iniziato la scuola in anticipo di un anno, quindi frequentavano la classe seconda nel 2002/03, ma poi sono stati bocciati, terminando ugualmente la classe quinta nell'anno scolastico 2006/07.
Gli studenti che, durante il periodo analizzato, hanno cambiato istituto almeno una volta sono 899 (15%), mentre gli altri sono 5.040 (85%). Il Grafico 8 mostra che il cambiamento di istituto non agevola la regolarità scolastica, anche in quanto spesso tale scelta è conseguente ad un esito scolastico negativo.

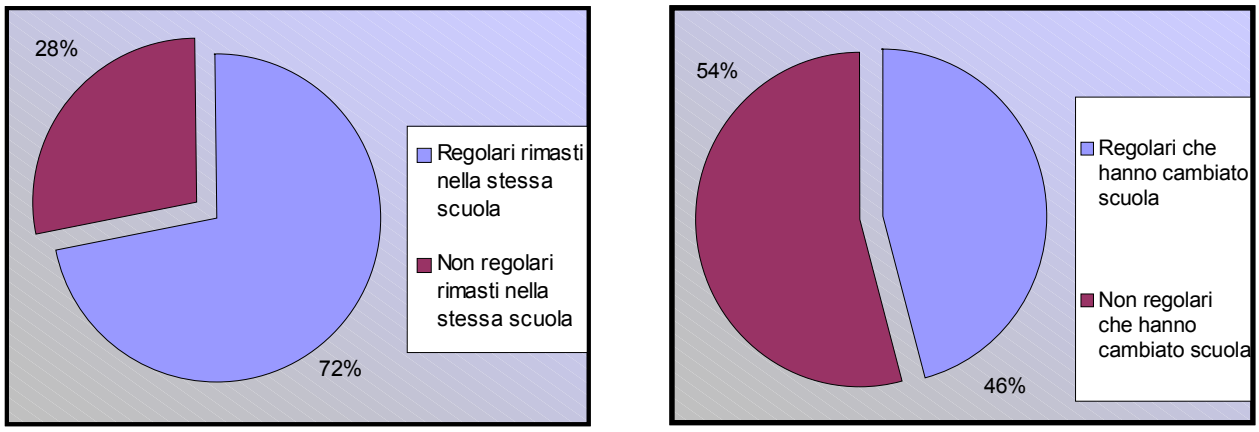
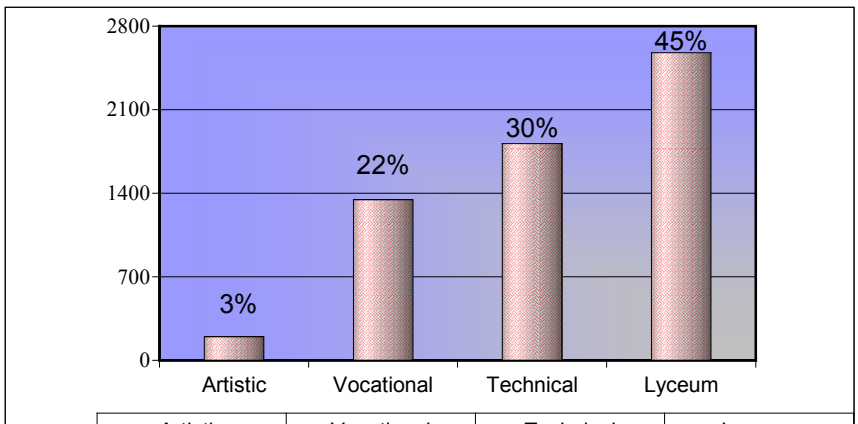


Grafico 8 – Ripartizione percentuale degli studenti tra coloro che hanno avuto un percorso regolare e gli altri: fedeli allo stesso istituto e no

Le differenze tra le diverse categorie di individui, in quanto a percentuali di regolarità, suggeriscono che le variabili usate per distinguere i diversi gruppi spieghino, o comunque concorrano a spiegare, la regolarità del percorso scolastico. Ciò incoraggia a costruire un modello che possa anche quantificare le relazioni.
Vi sono 899 studenti che hanno cambiato almeno una volta istituto: tenendo conto del fatto che essi rappresentano circa il 15% di tutti gli studenti, è possibile, per avere un'idea dell'effetto della tipologia di scuola frequentata sulla regolarità scolastica, considerare soltanto l'ultima scuola frequentata e trattare, per un attimo, tale variabile come invariante nel tempo. Il cambiamento di scuola frequentata può essere anche considerato come misura del successo scolastico. Come si vedrà in seguito, è possibile analizzare la mobilità degli studenti da una tipologia di scuola all'altra.
La distribuzione degli studenti del dataset per differenti tipologie di scuola (tenendo conto dell'ultima scuola frequentata, nei casi di cambiamento di tipologia di scuola da parte del singolo ragazzo) è quella presentata nel Grafico 9.

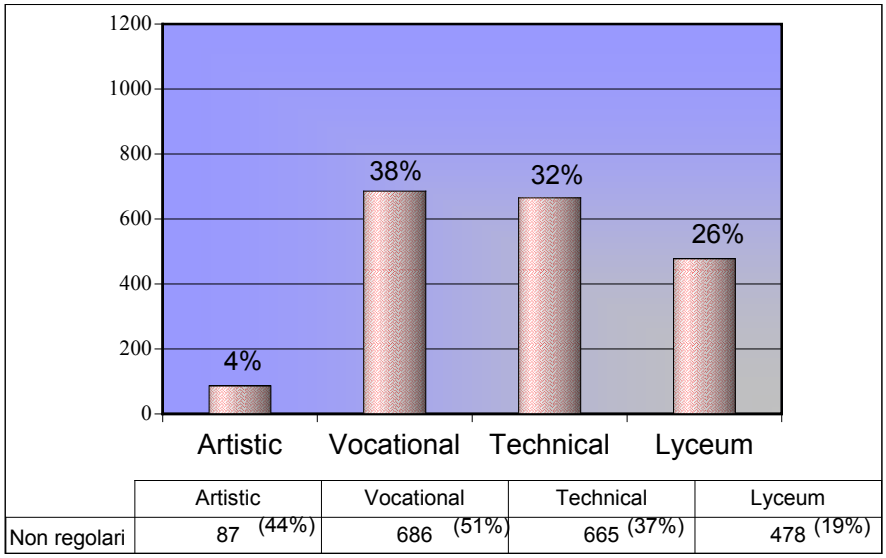


	Artistic	Vocational	Technical	Lyceum
Studenti	199	1347	1816	2577

Grafico 9 – Ripartizione percentuale degli studenti nelle diverse tipologie di scuola

Ancora, in Grafico 9, appare evidente la maggiore propensione, da parte dei ragazzi, a rivolgersi verso i licei che non alle altre tipologie di scuola. Il maggiore afflusso ai licei può in parte essere spiegato anche dall'incremento di offerta formativa apportata dalla trasformazione degli ex istituti magistrali, ora divenuti liceo pedagogico, liceo scienze sociali, ecc. che si sono rivelati essere alla portata anche di categorie di studenti che in passato si rivolgevano maggiormente ad istituti tecnici o professionali.

Il Grafico 10 mostra invece le percentuali di studenti con percorso di studi non regolare, distinte per tipologia di scuola frequentata.



	Artistic	Vocational	Technical	Lyceum
Non regolari	87 (44%)	686 (51%)	665 (37%)	478 (19%)

Grafico 10 – Ripartizione percentuale degli studenti con un percorso di studi non regolare nelle diverse tipologie di scuola

Dal Grafico 10 si evince che le percentuali di irregolarità maggiori si riscontrano negli istituti professionali e poi negli istituti tecnici. Si ha inoltre che negli istituti professionali il 51% dei ragazzi frequentanti risulta avere un percorso irregolare, negli istituti artistici il 44% dei ragazzi non ha un percorso regolare, mentre la stessa percentuale risulta del 37% negli istituti tecnici e del 19% nei licei: è ancora una volta confermato il maggiore successo degli studenti che frequentano quest'ultima tipologia di scuola rispetto a tutti gli altri.

Ritornando alla variabile oggetto di studio, *Times Promoted*, in Grafico 11 si possono vedere alcuni suoi andamenti tipici in alcune tipologie di studenti, attraverso il *trellis plot* costruito per taluni individui.

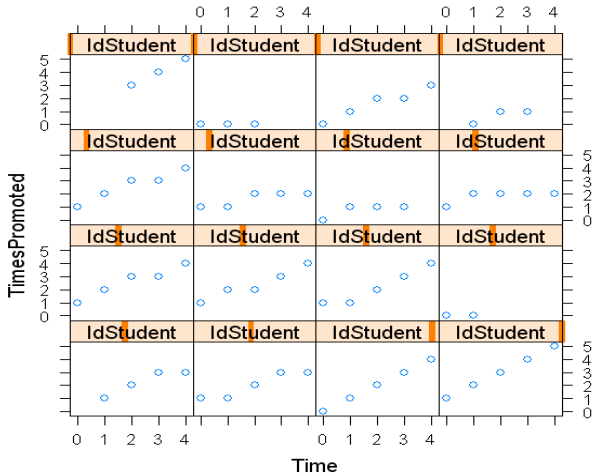


Grafico 11 – *Trellis plot* raffiguranti gli andamenti della variabile *Times Promoted* per diversi casi studiati

Vi sono 16 diversi casi rappresentati in Grafico 11; sono, nell'ordine:

1. studenti, con un corso di studi regolare, che hanno iniziato la scuola al di fuori della provincia di Bologna, ma poi si sono trasferiti e l'hanno terminata (promossi in quinta) in questa provincia;
2. studenti che non sono mai stati promossi e che hanno lasciato la scuola (o la scuola bolognese) dopo aver ripetuto per tre volte la classe prima;
3. studenti che hanno ripetuto la prima (valore 0 di *Times Promoted*) e la terza classe;
4. studenti che non sono stati promossi in classe prima nell'anno scolastico 2003/04, ma sono stati promossi in prima nel 2004/05 e hanno ripetuto la seconda l'anno successivo, per poi lasciare la scuola bolognese;
5. studenti che hanno ripetuto la classe quarta e sono passati in quinta nell'anno scolastico 2007/08;
6. studenti che sono stati bocciati in seconda classe una volta e poi anche in terza per due volte, quindi nel 2007/08 ripetevano la terza;
7. studenti che hanno ripetuto la prima classe, poi non hanno superato nemmeno la seconda, quindi l'hanno ripetuta per due volte e poi hanno lasciato la scuola (bolognese);
8. studenti che hanno ripetuto la seconda classe per tre volte, però nel 2006/07 risultavano ancora a scuola;
9. studenti che hanno ripetuto una volta la classe quarta e che nel 2006/07 hanno superato tale classe per passare in quinta l'anno successivo;
10. studenti che hanno ripetuto una volta la terza classe e che si trovavano in quinta nel 2007/08;
11. studenti che si trovavano in quinta nel 2007/08, ma essendo stati bocciati una volta in seconda classe;
12. studenti che hanno ripetuto due volte la prima classe, senza essere promossi, e poi hanno lasciato la scuola (bolognese);
13. studenti che sono arrivati alla scuola superiore con un anno di ritardo rispetto alla situazione regolare e che sono stati bocciati in classe quarta, quindi nel 2007/08 si trovavano a ripetere tale classe;
14. studenti che hanno ripetuto la seconda classe ed anche la quarta, quindi nel 2007/08 ripetevano la quarta;
15. studenti che hanno ripetuto la classe prima e che nel 2007/08 frequentavano la quinta classe;
16. studenti regolari, che hanno terminato il ciclo di studi nel 2006/07.

L'approssimazione lineare degli stessi casi è riportata in Grafico 12.

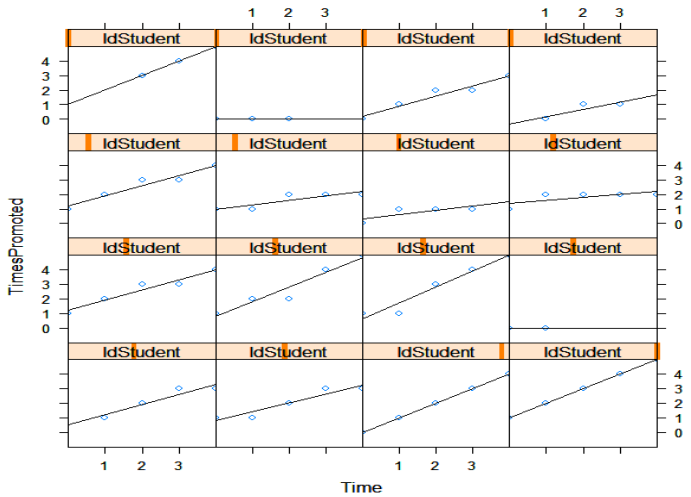


Grafico 12 - Andamenti della variabile *Times Promoted* per diversi casi studiati, approssimazione lineare

Data la situazione variegata dei comportamenti appena illustrati, è utile vedere le proporzioni di studenti che seguono i trend espressi. Il 56% degli studenti presenta la sequenza ottima di *Times Promoted* ([1,2,3,4,5]); il 26% degli studenti ha una sequenza incompleta (hanno lasciato la scuola prima del 2006/07 oppure vi sono entrati dopo il 2002/03); l'11% degli studenti ha una sequenza completa della variabile obiettivo (sono rimasti nelle scuole bolognese dal 2002/03 al 2006/07) ma si trovava in classe quarta (e non in quinta) nel 2006/07, presentava dunque un ritardo nel percorso scolastico.

Si passa ora ad analizzare anche graficamente gli esiti scolastici, in termini di promozione, degli studenti presenti nel dataset. Si possono così distinguere coloro che sono stati promossi da coloro che non lo sono stati, ma anche da coloro che hanno ottenuto una promozione con debito formativo (vi è da notare che, nel periodo considerato, non c'era la possibilità di essere rimandati, come lo è stato in passato e come è tuttora; il debito formativo può tuttavia essere paragonato all'attuale materia da recuperare a settembre, con la sostanziale differenza, però, che il debito non mette in discussione l'anno scolastico come invece l'esame a settembre). Lasciando per un attimo da parte il problema dell'abbandono precoce, si passa ora all'esame dell'esito scolastico. Nei dati aggregati nel seguito presentati, vengono considerati gli studenti che frequentavano una data classe in ciascun anno scolastico: per esempio, le percentuali di studenti promossi in prima classe sono calcolate tenendo conto soltanto di quegli studenti (nati nel 1988) che frequentavano la prima classe nell'anno scolastico 2002/03 e in modo analogo per le classi successive. In Grafico 13, sono stati considerati tutti gli studenti e tutti gli istanti temporali, quindi in questa percentuale "globale" entrano due volte nella stessa classe i ragazzi che non sono stati promossi, una volta come promossi ed una volta come no. Si ha comunque un'idea del fenomeno, che viene maggiormente precisata in Tabella 28.

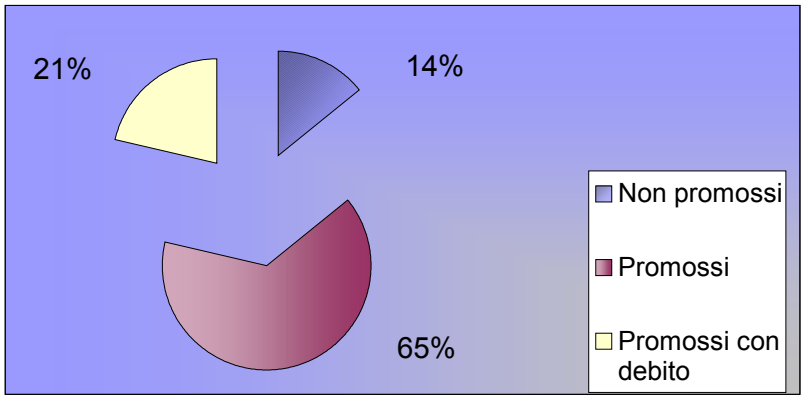


Grafico 13 – Distribuzione degli studenti per esito di fine anno

Tabella 28 – Percentuali di studenti promossi

Classe	% Promossi	% Promossi con debito[26]	% Non promossi
1	83,21	3,58	13,21
2	56,01	35,29	8,70
3	64,54	26,91	8,55
4	66,45	28,82	4,73
5	96,90	-	3,10

I risultati così ottenuti differiscono da quelli presentati nei dati aggregati in premessa. È importante considerare che questi dati sono relativi ai soli studenti nati nel 1988 e che frequentavano le suddette classi in uno degli anni scolastici considerati, quindi sono solo parzialmente confrontabili con i dati aggregati relativi agli studenti, nati in qualunque anno, che frequentavano ciascuna classe in ogni anno scolastico considerato. Per esempio, nella classe prima di Tabella 28, sono compresi tutti gli studenti nati nel 1988 che frequentavano tale classe nelle scuole bolognesi in uno degli anni scolastici presenti nel dataset, mentre i dati aggregati sono relativi a tutti gli studenti iscritti in prima nell'anno scolastico 2002/03, non solo i nati del 1988: si tratta di contingenti diversi. Il

[26] Rapporto tra numero di ragazzi promossi con debito e totale dei frequentanti.

confronto è reso in parte accettabile dal fatto che i dati aggregati si riferiscono a quegli studenti che erano a scuola alla fine dell'anno scolastico, e analogamente i dati individuali si riferiscono a quegli studenti che hanno terminato l'anno scolastico a scuola, sono cioè esclusi coloro che hanno lasciato la scuola prima del termine dell'anno scolastico. Si nota anche però che le percentuali di ragazzi promossi nelle diverse classi risultano sempre maggiori nel contingente dei nati nel 1988 rispetto ai valori che si presentano nei dati aggregati. Ciò può anche essere dovuto a qualche errore presente nel database (nonostante i controlli incrociati effettuati, non è possibile escluderlo), però occorre considerare un elemento fondamentale: le percentuali calcolate sul contingente 1988 per le diverse classi comprendono in realtà un gran numero di ragazzi regolari (trattandosi dei soli nati nel 1988, è chiaro che all'aumentare della progressione delle classi aumenta il numero di regolari). Alcuni confronti: in prima classe la percentuale di studenti promossi (anche con debito) è dell'87% (Tabella 28) per gli studenti nati nel 1988, mentre risulta del 78% (Tabella 17) per i frequentanti nel 2002/03; in classe seconda, per il contingente dei nati nel 1988 la percentuale di promozione è del 91,3% (Tabella 28), mentre è dell'85,5% (Tabella 17) per i frequentanti nel 2003/04; in terza, la percentuale di promossi calcolata per il contingente del 1988 è del 91,4% (Tabella 28), mentre quella calcolata su dati aggregati non è disponibile; in quarta, la percentuale di promossi per i nati nel 1988 è del 95,3% (Tabella 28), mentre è del 92,4% (Tabella 17) per i frequentanti nel 2005/06.

Passando ad osservare le percentuali di ragazzi promossi con debito sul totale dei promossi, si nota immediatamente la presenza di un errore nel dataset dei dati individuali (soltanto 187 ragazzi frequentanti la prima classe risultano aver avuto un debito contro i 4.534 promossi, con una percentuale intorno al 4%). Per le classi successive, le percentuali (ragazzi promossi con debito sul totale dei promossi) sui dati individuali (Tabella 28) sono analoghe a quelle calcolate sui dati aggregati (Tabella 19). In classe seconda, infatti, tale percentuale risulta del 39% sui dati individuali e del 44% sui dati aggregati; in terza si ha un 29% nei dati individuali e un 30% nei dati aggregati; in quarta la percentuale è del 30% sui dati individuali e del 38% sui dati aggregati.

Sempre con riferimento all'esito scolastico, in termini di promozione alla classe successiva, nel seguito vengono presentati alcuni grafici che mostrano le diverse percentuali di promozione riscontrate in diversi gruppi di studenti (vengono come sempre considerati a denominatore soltanto gli studenti che concludono l'anno scolastico a scuola, escludendo quindi coloro che abbandonano prima del termine).

La prima distinzione è quella in base al sesso, rappresentata in Grafico 14 e descritta in Tabella 29 e Tabella 30.

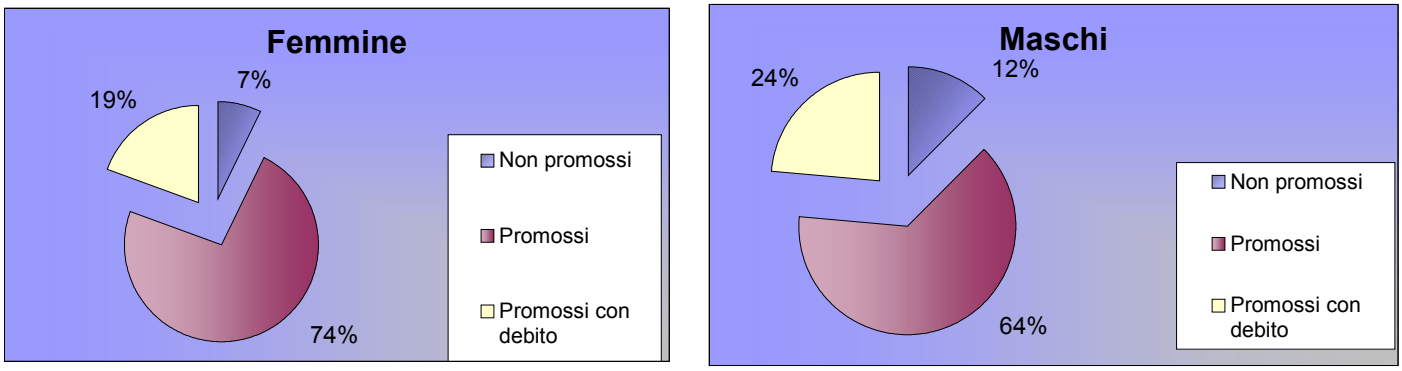


Grafico 14 - Distribuzione degli studenti per esito di fine anno: femmine e maschi

Tabella 29 – Percentuali di studenti promossi – Femmine

Classe	% Promossi	% Promossi con debito[27]	% Non promossi
1	78,93	10,16	10,91
2	58,75	33,75	7,50
3	68,05	24,47	7,48
4	68,61	26,18	5,21
5	96,46	-	3,54

Tabella 30 – Percentuali di studenti promossi - Maschi

Classe	% Promossi	% Promossi con debito[28]	% Non promossi
1	72,31	10,89	16,80
2	43,32	43,79	12,89
3	54,43	32,49	13,08
4	52,61	37,87	9,52
5	94,68	-	5,32

Dal Grafico 14 risultano evidenti le differenze tra maschi e femmine (i ragazzi sono maggiormente a rischio di non promozione che non le femmine, in tutte le classi), ma per entrambi si ha un trend analogo: il rischio di non promozione più alto si ha in classe prima, poi c'è una diminuzione, con una lieve ripresa poi in classe terza (più accentuata nei maschi) e seguita da un calo nei due anni successivi. Anche per quanto riguarda la probabilità di promozione con debito, i maschi sono maggiormente a rischio rispetto alle femmine: in prima, il 13% dei ragazzi promossi e l'11% delle ragazze promosse ha un debito; in seconda le stesse percentuali sono del 50% per i maschi e del 36% per le femmine; in terza si riscontra un 37% per i ragazzi contro un 26% delle ragazze; in quarta, infine, il rischio di essere promossi con debito è del 42% per i maschi contro il 28% delle femmine.
Un'altra distinzione tra gruppi diversi di studenti è quella in base alla variabile cittadinanza, rappresentata in Grafico 15: studenti stranieri (con cittadinanza non italiana) e italiani (come in precedenza, i dati tengono conto dei soli presenti a fine anno scolastico).

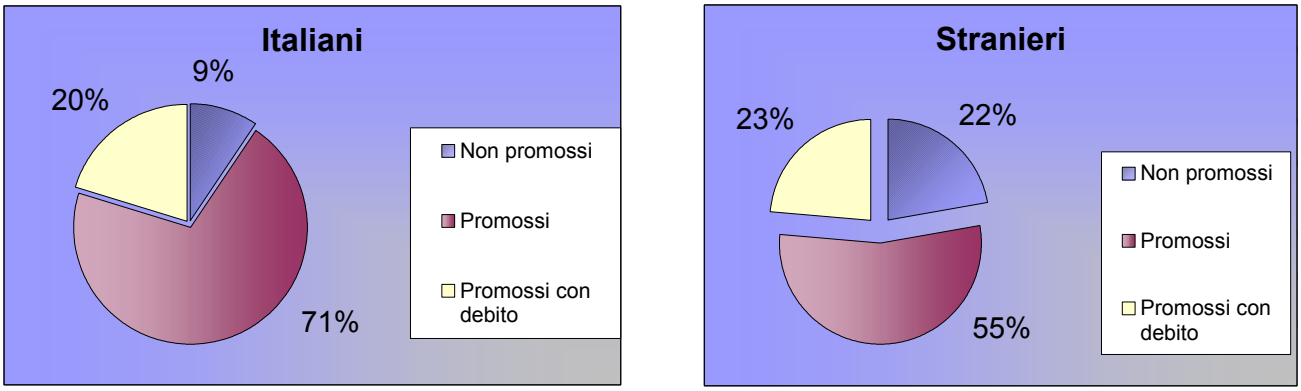


Grafico 15 – Distribuzione degli studenti per esito di fine anno: italiani e non italiani

Tabella 31 – Percentuali di studenti promossi - Italiani

Classe	% Promossi	% Promossi con debito[29]	% Non promossi
1	77,61	9,15	13,24
2	52,01	38,27	9,72
3	61,23	28,71	10,06
4	61,27	31,80	6,93
5	95,73	-	4,27

Tabella 32 – Per Percentuali di studenti promossi - Stranieri

Classe	% Promossi	% Promossi con debito[30]	% Non promossi
1	38,15	35,08	26,77
2	32,93	45,78	21,29
3	61,23	23,35	15,42
4	49,13	33,53	17,34
5	90,82	-	9,18

[27] Rapporto tra numero di ragazzi promossi con debito e totale dei frequentanti.
[28] Rapporto tra numero di ragazzi promossi con debito e totale dei frequentanti.
[29] Rapporto tra numero di ragazzi promossi con debito e totale dei frequentanti.
[30] Rapporto tra numero di ragazzi promossi con debito e totale dei frequentanti.

Come evidente anche dal Grafico 15, introducendo la distinzione fra studenti con cittadinanza italiana e non, si notano forti differenze tra i due gruppi così individuati. Già dal rischio di non promozione, descritto in Tabella 31 e Tabella 32, si può vedere il forte effetto della variabile cittadinanza sull'esito scolastico: un 27% di ragazzi stranieri non promossi in classe prima, contro il 13% degli italiani; in seconda si riscontra un 21% contro il 10% degli italiani; in terza la differenza si attenua (il 15% per gli stranieri contro il 10% per gli italiani); il trend prosegue, in quarta, per entrambi in calo, ma il rischio rimane più alto per gli stranieri, con un 9% contro il 4%. In terza classe si può notare una diminuzione del rischio di promozione con debito per gli stranieri (il 28% dei promossi ha un debito) rispetto al quella degli italiani (il 32% ha un debito, tuttavia la percentuale di non promozione rimane più alta per gli stranieri): ciò è probabilmente dovuto al fatto che gli studenti con cittadinanza non italiana frequentano più spesso istituti professionali, dove in terza classe viene effettuato il test di qualifica, non è quindi possibile ottenere la promozione con debito. Occorre anche tener conto del fatto che gli stranieri diminuiscono procedendo nelle diverse classi, proprio perché aumenta la probabilità di regolarità (sono sempre i nati nel 1988), inoltre aumenta anche il loro tasso di regolarità e quindi anche la probabilità di essere promossi.

Un'altra variabile che influisce sull'esito scolastico in termini di promozione è la tipologia di scuola frequentata, rappresentata in Grafico 16.

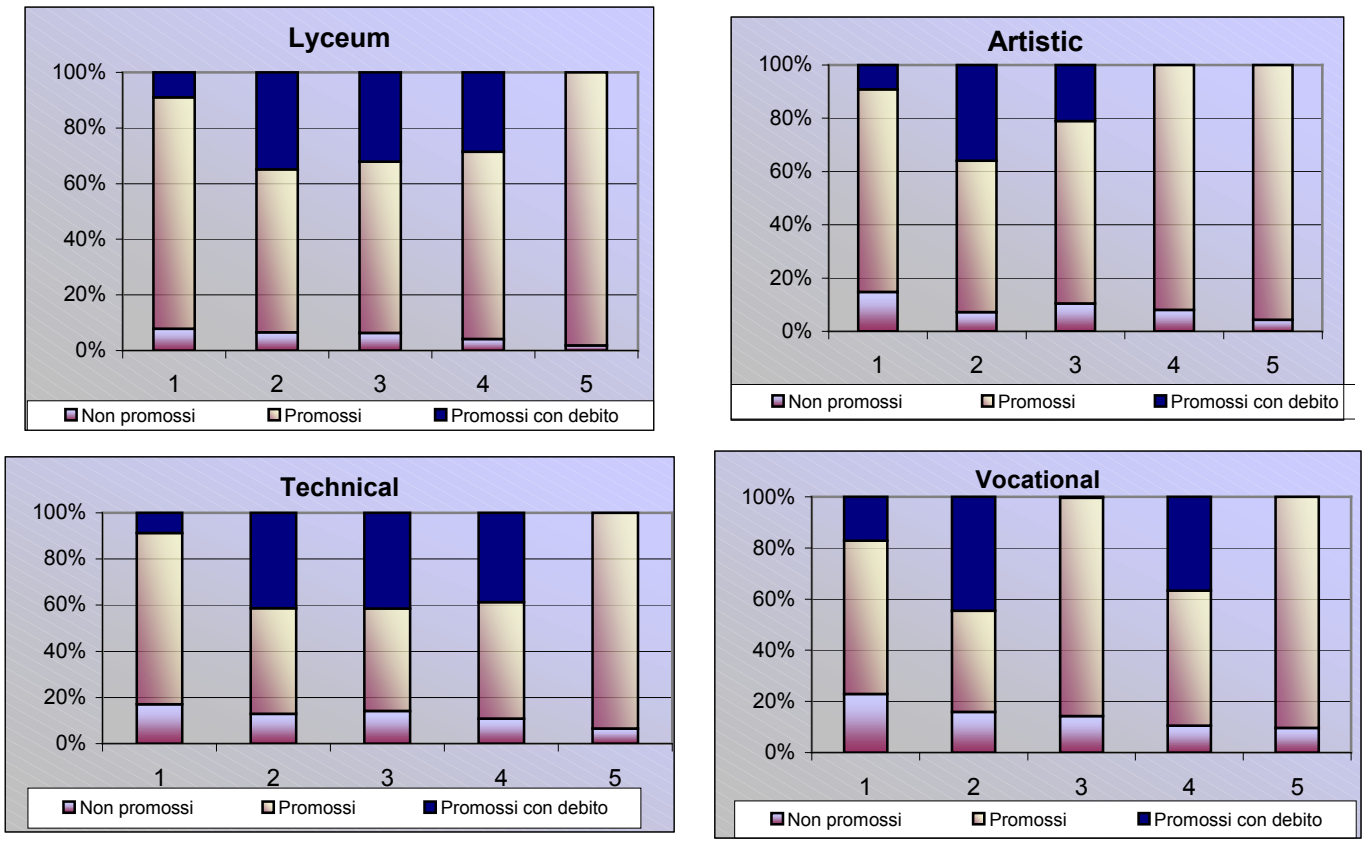


Grafico 16 - Distribuzione degli studenti per esito di fine anno: distinzione per tipo di scuola

È immediatamente visibile, in Grafico 16, la distinzione tra gli istituti professionali e le altre tipologie di scuola per quanto riguarda il terzo anno: negli istituti professionali in terza si sostiene l'esame di qualifica, quindi non è possibile essere promossi con debito, ma soltanto essere promossi o non promossi. La presenza di tale esame è anche il motivo per cui alcuni studenti lasciano la scuola professionale dopo aver terminato la terza classe. Fortunatamente si tratta di non molti casi, ma tale fenomeno si può vedere osservando le percentuali, calcolate

sul dataset dei dati individuali, di coloro che abbandonano la scuola dopo la classe terza nelle diverse tipologie di istituto: tale percentuale risulta del 18% negli istituti professionali, mentre del 10% nei tecnici e del 6% nei licei. Un'altra situazione particolare è quella degli istituti artistici: ve ne sono di due tipi, il liceo e l'istituto d'arte. Al liceo artistico è possibile essere promossi con debito in classe quarta, ma nel caso in esame nessuno degli studenti ha conseguito un debito; all'istituto d'arte, invece, non è proprio possibile essere promossi con debito in quarta perché è l'ultima classe dell'istituto.

Con particolare riferimento alla classe prima, si può vedere la variabilità del rischio di non promozione da una tipologia di scuola all'altra: la percentuale di studenti non promossi è dell'8% nei licei, del 15% negli istituti artistici, del 17% nei tecnici e del 23% nei professionali. Ma anche nelle classi successive si hanno, sebbene in minor misura, evidenti differenze. In classe seconda, il rischio di non promozione non arriva al 7% nei licei, è poco più del 7% negli istituti artistici, mentre cala passando al 13% nei tecnici e al 16% nei professionali (un raddoppio). In classe terza, diminuisce passando al 6% nei licei, mentre aumenta passando al 10% negli istituti artistici, mentre passa al 14% nei tecnici e nei professionali. In classe quarta, il rischio di non promozione diminuisce, ponendosi intorno al 4% nei licei e intorno all'8% negli istituti artistici, mentre intorno all'11% nei tecnici e nei professionali. In classe quinta, il rischio di non promozione cala ulteriormente in tutte le tipologie di scuola, portandosi al 2% nei licei, al 4% negli istituti artistici, al 6% nei tecnici ed al 9% nei professionali. Si notano anche differenze nelle percentuali di studenti promossi con debito sul totale dei promossi: un range del 30 – 37% (nelle diverse classi frequentate) caratterizza i licei, mentre negli istituti artistici lo stesso intervallo è del 24 – 39%, nei tecnici è del 43 – 49% e nei professionali risulta del 41 – 53%. Ancora una volta si nota il maggior rischio degli istituti tecnici e professionali, dove non solo è meno probabile essere promossi, ma anche essere promossi senza debito.

Procedendo con le analisi descrittive del dataset oggetto di studio e grazie alla disponibilità di dati individuali nel tempo, si può considerare l'esito scolastico dei soli 4.385 studenti che erano presenti nelle scuole di Bologna e provincia in tutti gli anni scolastici considerati (2002/03, 2003/2004, 2004/05, 2005/06 e 2006/07). Ciò permette di analizzare il comportamento di uno stesso contingente di individui nell'arco di tutto l'intervallo temporale considerato. Occorre tener conto che in questo modo vengono esclusi coloro che erano presenti solo in alcuni degli anni scolastici considerati, quindi più probabilmente studenti con un percorso irregolare; gli esiti dei 4.385 sono perciò tendenzialmente migliori di quelli riportati dall'intero contingente analizzato.

Una descrizione del fenomeno è quella riportata in Grafico 17, dove i 4.385 studenti sono suddivisi per promozione: coloro i quali sono stati sempre promossi (si tratta quindi dei regolari) vengono distinti a seconda del numero di volte in cui hanno conseguito la promozione con debito.

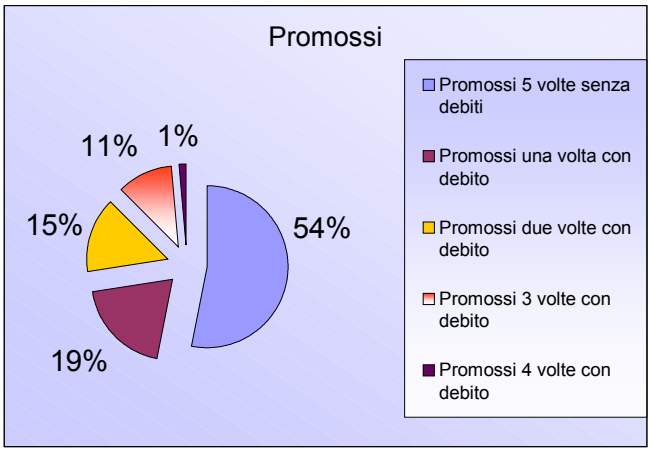


Grafico 17 – Distribuzione degli studenti promossi in base ai debiti riportati

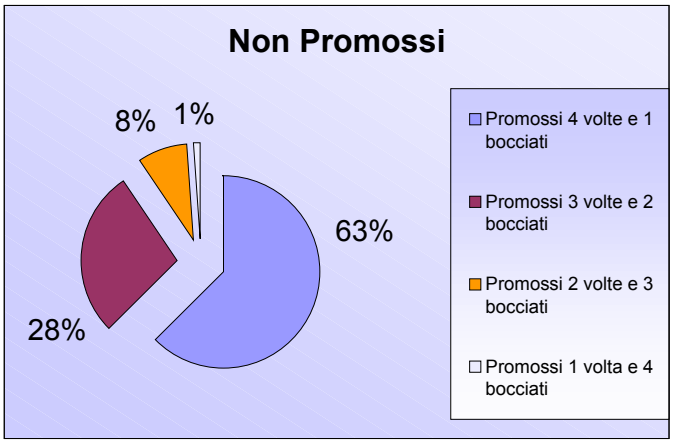


Grafico 18 – Distribuzione degli studenti non promossi in base al numero di ripetenze

Come si nota nel Grafico 17, soltanto l'1% degli studenti sempre promossi ha ottenuto sempre la promozione con debito, mentre poco meno della metà degli studenti sempre promossi (il 46%) è stato promosso con debito almeno una volta. Fortunatamente, la maggior parte degli studenti promossi (il 54%) non ha mai avuto debiti formativi. Come mostrato in Grafico 18, più della metà (il 63%) degli studenti bocciati almeno una volta, lo è stato di fatto una volta sola. In totale, il 76% dei 4.385 studenti che erano nelle scuole bolognesi negli anni scolastici dal 2002/03 al 2006/07 è sempre stato promosso.

Considerando ora soltanto quei 3.313 studenti che sono stati sempre promossi, si analizzano alcune delle loro caratteristiche, sempre a fini di indagine sui fattori che determinano il successo scolastico. Se si considerano i 1.756 (40% del totale dei ragazzi) che sono stati sempre promossi senza debito (qui denominati ottimamente promossi), si nota che si tratta per lo più di ragazze (per il 60%) e da italiani (soltanto l'1% ha cittadinanza straniera). In Grafico 19 è mostrata, differentemente per ragazzi e ragazze, la ripartizione tra studenti ottimamente promossi e gli altri; analogamente, in Grafico 20 è rappresentata la medesima ripartizione, distintamente per italiani e stranieri. Per quanto riguarda la tipologia di scuola frequentata (considerando l'ultima scuola frequentata in caso di cambiamento), si tratta in prevalenza di studenti del liceo (il 61% frequenta un liceo, il 3% un istituto artistico, il 25% un istituto tecnico e solo l'11% un istituto professionale). Il Grafico 21 mostra la differenza, tra gli studenti frequentanti i diversi tipi di scuola, in termini di percentuale di studenti ottimamente promossi. Si ha infine che il 91% degli studenti sempre promossi senza debito non ha mai cambiato istituto durante il percorso scolastico, percentuale maggiore rispetto a quella calcolata sull'intero contingente, pari all'85%. In Grafico 22 si nota la differenza in termini di promozione tra chi ha cambiato istituto e chi no.

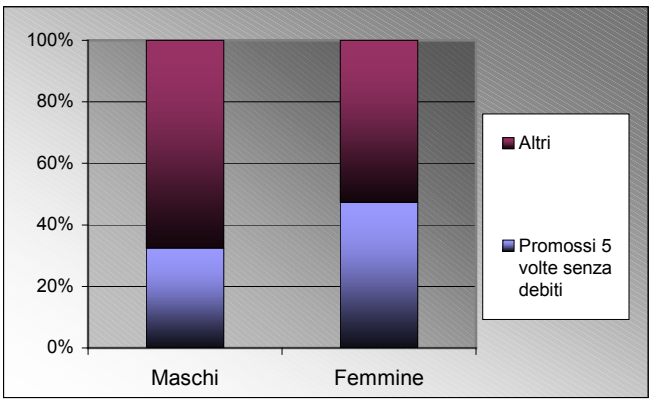


Grafico 19 – Ripartizione degli studenti sempre promossi in base al sesso

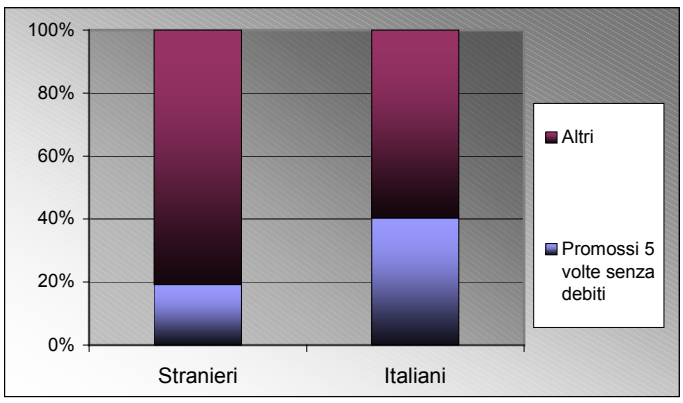


Grafico 20 – Ripartizione degli studenti sempre promossi in base alla nazionalità

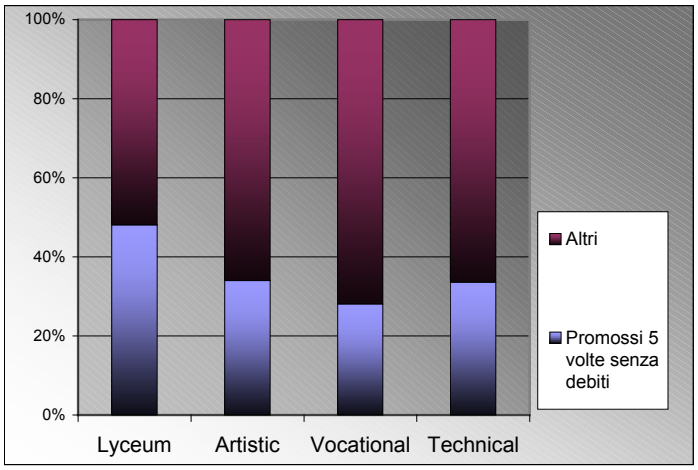


Grafico 21 - Ripartizione degli studenti sempre promossi in base al tipo di scuola

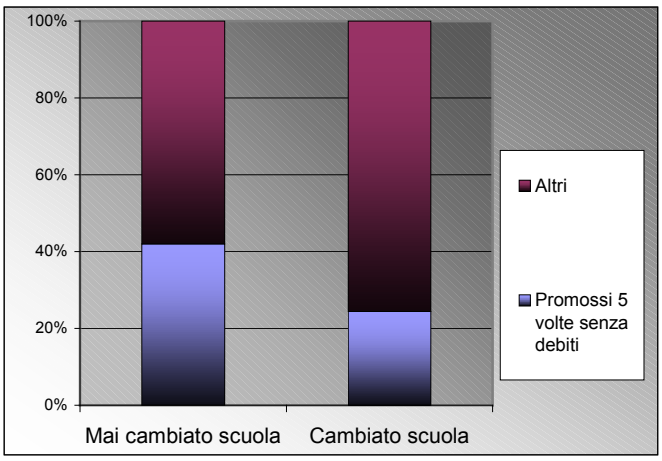


Grafico 22 – Studenti sempre promossi e fedeltà all'istituto

Per quanto riguarda invece coloro i quali sono sempre stati promossi, indipendentemente dal debito, si ha che anch'essi sono per la maggior parte ragazze (per il 55%) e studenti con cittadinanza italiana (anche in questo caso solo l'1% è formato da studenti con cittadinanza straniera). La percentuale di studenti ottimamente promossi, quindi senza mai conseguire un debito, su quelli sempre promossi è diversa a seconda dei vari gruppi: il 59% per le ragazze contro il 46% per i ragazzi; il 53% per gli studenti con cittadinanza italiana contro il 37% per quelli che hanno cittadinanza straniera; il 56% per gli studenti del liceo, contro il 53% per i ragazzi degli istituti artistici, il 51% per i ragazzi degli istituti professionali ed il 48% per quelli dei tecnici. La composizione per diverse tipologie di scuola (considerando l'ultima scuola frequentata in caso di cambiamento) dei ragazzi che sono sempre stati promossi, indipendentemente dai debiti, è decisamente a favore dei licei: il 57% di questi studenti frequentava il liceo, il 3% un istituto artistico, il 28% un istituto tecnico ed il 12% un professionale. Se si osservano le analoghe percentuali, ma calcolate tenendo conto della prima scuola frequentata, si hanno leggere variazioni, che denotano però movimenti degli studenti: i licei passano dal 57% al 59% (data la più alta percentuale ottenuta considerando la prima scuola frequentata, si evince che probabilmente vi è stato un calo di ragazzi nel tempo) e gli istituti tecnici dal 28% al 26% (data invece la più bassa percentuale considerando la prima scuola frequentata, vi è stato probabilmente un aumento di ragazzi nel tempo).

Un altro interessante punto di vista, per spiegare il successo scolastico, è quello che tiene conto dei cambiamenti nella tipologia di scuola frequentata durante il percorso scolastico. Infatti, si può dire che uno studente completi con successo il percorso scolastico non solo con riferimento alla promozione, ma anche alla prosecuzione della scelta sulla tipologia di scuola: possiamo considerare uno studente di successo come quello che rimane nella medesima tipologia di scuola inizialmente scelta, essendo sempre promosso.

Viene ora considerato il gruppo di studenti che frequentavano la classe prima o seconda nel primo anno scolastico considerato, il 2002/03, e che sono poi rimasti nelle scuole bolognesi fino al 2006/07 (o 2005/06 per quelli che hanno iniziato in anticipo). Sono i 4.385 studenti già analizzati (per il 51% ragazze e per l'1% cittadini non italiani), 659 dei quali (il 15%) ha cambiato tipologia di istituto, almeno una volta, durante il proprio percorso scolastico: questi sono composti per il 47% da ragazze e per il 4% da studenti con cittadinanza non italiana. Guardando gli esiti scolastici di tali studenti, si nota che esiste una differenza tra coloro che hanno cambiato tipologia di scuola e coloro che invece sono rimasti nella stessa scuola inizialmente scelta. In particolare, la differenza è più consistente nelle prime due classi, mentre diminuisce nelle classi successive. In

Tabella 33 sono illustrate le percentuali di composizione degli studenti promossi, ma occorre tener presente che sono percentuali riferite agli studenti che frequentavano la prima classe nel 2002/03, la seconda nel 2003/04 e così via (si tratta dei soli studenti regolari).

Tabella 33 – Percentuali di studenti promossi, per ciascuna classe frequentata

	Studenti che non hanno cambiato tipologia di scuola		Studenti che hanno cambiato tipologia di scuola	
Classe	Promossi senza debiti	Promossi con debito	Promossi senza debiti	Promossi con debito
1	93,1	3,4	71,1	4,9
2	61,3	34,7	45,9	37,3
3	65,6	28,1	63,6	20,3
4	66,9	28,4	65,3	32,2
5	97,0	-	94,6	-
Medie	76,5	19,2	66,9	18,3

Le differenze, in termini di rischio di non promozione, tra il gruppo degli studenti che ha proseguito la propria scelta e quello degli altri, sono notevoli nelle prime 3 classi: coloro che hanno cambiato tipologia di scuola presentano una percentuale di promozione inferiore di 20 punti percentuale in classe prima e di circa 10 punti percentuale nelle classi seconda e terza, mentre il rischio di non promozione è pressoché analogo tra i due gruppi nelle ultime due classi (quando coloro che hanno cambiato scuola l'avevano già cambiata, ottenendo quindi migliori risultati).

Guardando la tipologia di scuola frequentata da questi studenti, presentata in Tabella 34, si possono fare altre interessanti considerazioni, in particolare riguardo ai movimenti degli studenti.

Tabella 34 – Percentuali di studenti che frequentavano le diverse tipologie di scuola

	a.s. 2002/03	a.s. 2003/04	a.s. 2004/05	a.s. 2005/06	a.s. 2006/07
Licei	54,8	53,3	52,4	51,8	50,4
Istituti artistici	2,8	3,1	3,2	3,2	3,2
Istituti tecnici	29,6	29,7	29,9	29,6	30,3
Istituti professionali	12,8	13,9	14,5	15,4	16,1

Dalla Tabella 34 emerge che le percentuali di studenti che frequentavano gli istituti tecnici e quelli artistici nei diversi anni scolastici considerati rimangono pressoché costanti nel tempo, mentre le stesse percentuali riferite agli istituti professionali aumentano nel tempo (passando da un 13% a un 16%) e quelle riferite ai licei diminuiscono nel tempo (passando da un 55% a un 50%). Ciò vuol dire che, a partire da tali dati aggregati, pare che gli istituti professionali siano in generale collettori di studenti (gli studenti che cambiano tipologia di scuola si rivolgono verso questo tipo di istituti), mentre i licei possano dirsi, per taluni studenti, punti di partenza che poi si rivelano essere scelte sbagliate.

Avendo a disposizione dati individuali, però, a partire dalla classe frequentata è stato possibile ricostruire il percorso di alcuni individui tipo, anche per quanto riguarda la mobilità. In questo modo si è pervenuti a risultati più analitici, in parte anche diversi da quelli in Tabella 34, che sono il frutto anche di compensazioni.

Soltanto il 4,8% degli studenti che, nell'anno scolastico 2002/03, frequentavano la classe prima in un istituto artistico ha cambiato tipologia di scuola, passando per lo più ad istituti professionali. La percentuale di studenti

che ha cambiato scuola partendo in prima, nel 2002/03, da un liceo è dell'8,8%; il 61% di essi si è rivolto ad istituti tecnici, mentre il 32% ad istituti professionali. Anche alcuni studenti (l'8,8%) che frequentavano, nel 2002/03, la prima classe di un istituto tecnico hanno cambiato tipologia di scuola, scegliendo per l'82% un istituto professionale, ma anche un istituto artistico o, un po' più spesso, un liceo. La percentuale di studenti che ha cambiato scuola partendo in prima dal professionale è più bassa che negli altri casi (il 3,6%) ed il 70% di essi ha scelto un istituto tecnico.

Il 37% degli studenti che ha cambiato almeno una volta tipologia di scuola lo ha fatto nell'anno scolastico 2003/04, dopo aver frequentato la classe prima.

Si può anche dare uno sguardo alle percentuali di arrivo, dopo aver analizzato quelle di partenza. Il 18% degli studenti che nel 2006/07 frequentavano gli istituti artistici provenivano da altre tipologie di scuola (la maggior parte di essi da un liceo). Soltanto lo 0,8% degli studenti che nel 2006/07 frequentavano un liceo proveniva da un'altra tipologia di scuola (di solito, infatti, il liceo perde studenti e non ne acquista). L'11% degli studenti che frequentavano un istituto tecnico nel 2006/07 proviene da altre tipologie di scuola: per il 93% provenivano da un liceo. Il 23% degli studenti che frequentavano un istituto professionale nel 2006/07 veniva da un'altra tipologia di scuola, per il 41% da un liceo. L'alta percentuale degli istituti professionali conferma la loro caratteristica di collettori di studenti.

È interessante indagare anche sul comportamento di alcuni singoli individui che hanno cambiato tipologia di scuola, per dimostrare ancora una volta che il fenomeno oggetto di studio è una realtà variegata, quindi ogni sintesi nasconde un trend comune con scostamenti individuali anche rilevanti.

Tra gli studenti che frequentavano gli istituti tecnici nel primo anno scolastico considerato, è utile segnalare 6 casi particolari di alcuni studenti che hanno cambiato tipologia di scuola più di una volta. 4 di questi studenti hanno effettuato il cambiamento per fini migliorativi, quindi contro la tendenza generale di lasciare una scuola che presenta maggiori difficoltà: sono passati da un istituto tecnico ad un liceo, non certo per rimediare ad una situazione negativa ma piuttosto per cercare di migliorare il proprio percorso scolastico; alla fine però sono ritornati indietro, anche se in realtà sono sempre stati promossi, o al massimo bocciati una sola volta. Gli altri 2 dei 6 studenti possono essere considerati casi particolari in quanto sono passati da un istituto professionale ad un tecnico, senza però essere stati bocciati; forse anche in questo caso si può parlare di scelta per fini migliorativi.

Come è facile attendersi, la maggior parte degli studenti che ha effettuato il cambiamento è stato bocciato almeno una volta, e ciò è accaduto nell'anno scolastico precedente il cambiamento. Considerando gli studenti che frequentavano la prima classe degli istituti tecnici nel 2002/03 e che hanno poi cambiato tipologia di scuola, il 75% di coloro che hanno effettuato il cambiamento nel 2003/04 è stato bocciato l'anno precedente, il 38% di coloro che hanno cambiato istituto nel 2004/05 non era stato promosso l'anno precedente, il 69% di quelli che si sono rivolti ad altra tipologia di scuola nel 2005/06 non aveva conseguito la promozione l'anno precedente e l'analoga percentuale è del 67% per il 2006/07.

Tra gli studenti che nel 2002/03 frequentavano un liceo e che poi hanno cambiato tipologia di scuola, vi sono 3 casi particolari da evidenziare; si tratta ancora una volta di studenti che hanno effettuato il cambiamento per più di una volta. Anche in questi casi si può parlare di cambiamento per migliorare la propria condizione, non soltanto come rimedio ad una situazione negativa. Questi studenti, dopo una bocciatura, si sono rivolti ad un

istituto tecnico, dove sono stati promossi; tuttavia dopo un anno sono ritornati al liceo, ottenendo ancora una volta la promozione.

Per quanto riguarda gli studenti che all'inizio frequentavano un liceo, differentemente da coloro che invece frequentavano un istituto tecnico, soltanto alcuni di coloro che hanno effettuato un cambiamento di scuola lo hanno fatto a seguito di un fallimento (bocciatura). In particolare si ha che il 55% di coloro che hanno cambiato tipologia di scuola nel 2003/04 non ha ottenuto la promozione l'anno precedente; il 44% di quelli che si sono rivolti ad altra scuola nel 2004/05 lo ha fatto in seguito a bocciatura; le analoghe percentuali sono del 46% per il 2005/06 e del 20% per il 2006/07 (quindi decisamente inferiore a quelle riferite agli istituti tecnici).

A partire dai dati individuali, si può costruire un'interessante tavola di sopravvivenza, presentata in Tabella 35: i 5.366 studenti, che frequentavano le scuole di Bologna e provincia nell'anno scolastico 2002/03 (in classe prima o al più seconda), sono stati seguiti durante il loro percorso scolastico. Bisogna però sempre considerare che non è possibile sapere se coloro che sono usciti dalla scuola bolognese sono effettivamente usciti dal percorso scolastico o si sono semplicemente trasferiti in altra provincia.

Tabella 35 – Tavola di sopravvivenza: studenti che sono rimasti nelle scuole di Bologna, su 1000 studenti che frequentavano tali scuole nell'a.s. 2002/03

	Classe I	Classe II	Classe III	Classe IV	Classe V	TOTALI	TOTALI Valori assoluti
2007/08		1	4	36	117	158	846
2006/07		2	37	135	643	817	4.385
2005/06	2	31	152	684	14	883	4.738
2004/05	15	136	762	14		927	4.976
2003/04	104	843	16			963	5.166
2002/03	983	17				1.000	5.366

Dalla Tabella 35 emerge che, su 1.000 studenti che erano a scuola nell'anno scolastico 2002/03 (in prima o in seconda), soltanto 657 (643+14) hanno poi frequentato, con percorso regolare, l'ultimo anno della scuola secondaria di secondo grado (su 5.366 studenti, sono 3.518 regolari). Il 16% degli studenti iniziali (846 studenti) si trovava ancora a scuola nel 2007/08, quindi questi studenti sono stati bocciati almeno una volta durante il loro percorso scolastico. Dai dati (non visibile in tabella) emerge anche che 183 studenti su 1.000 risultano usciti dalla scuola bolognese prima di ultimare la quinta classe.

Un altro punto di vista della valutazione del successo scolastico, come precedentemente accennato, è quello di considerare la prosecuzione della scuola inizialmente scelta: uno studente completa con successo il proprio percorso scolastico se rimane nella tipologia di scuola scelta il primo anno (la sua scelta iniziale si è rivelata quella giusta). Si possono a tal fine analizzare i 4.385 studenti rimasti all'interno della scuola bolognese (escludendo quindi le uscite) per i 5 anni scolastici considerati. È utile costruire anche in questo caso una tavola di sopravvivenza (Tabella 36), ma in questo caso distinta per tipologia di scuola frequentata ed aggiungendo l'informazione aggiuntiva relativa al conseguimento del diploma, così da evidenziare il completamento del percorso scolastico.

Tabella 36 – Percentuali di "sopravvivenza" per gli studenti che frequentavano diverse tipologie di scuola, distinzione per classe

	Classe II	Classe III	Classe IV	ClasseV	Diplomati
Percentuali di studenti regolari[31] sui 2.403 studenti che frequentavano la classe prima al liceo	93,0	88,4	83,9	80,7	79,4
Percentuali di studenti regolari sui 124 studenti che frequentavano la classe prima all'istituto artistico	93,6	89,5	82,3	79,0	75,0
Percentuali di studenti regolari sui 1.299 studenti che frequentavano la classe prima all'istituto tecnico	91,2	83,9	75,4	69,3	66,4
Percentuali di studenti regolari sui 559 studenti che frequentavano la classe prima all'istituto professionale	91,4	84,6	75,9	71,0	66,4

In Tabella 36 si può vedere che le percentuali più basse si riscontrano negli istituti tecnici e in quelli professionali. In particolare, il 7% di quegli studenti che frequentavano la classe prima in un liceo non era presente in seconda nell'anno successivo; l'analoga percentuale per gli istituti tecnici e per quelli professionali è del 9%. Per quanto invece riguarda il passaggio dalla seconda alla terza classe, la percentuale di coloro che mancavano in terza rispetto ai presenti in seconda è del 5% nei licei, mentre dell'8% negli istituti tecnici e del 7% negli istituti professionali: l'incidenza di coloro che abbandonano diminuisce rispetto alla classe precedente, ma il divario tra tipologie di scuola aumenta leggermente. La percentuale di coloro che hanno lasciato la scuola bolognese nel passaggio tra la terza e la quarta classe rimane del 5% nei licei, mentre aumenta, passando al 10%, sia negli istituti tecnici che nei professionali. La percentuale di coloro che invece abbandonano nel passaggio alla quinta classe è in netto calo (probabilmente in questa fase si può veramente parlare di trasferimento, più che di abbandono, anche se qualche caso di abbandono ancora rimane), passando al 4% nei licei, all'8% negli istituti tecnici e al 6% negli istituti professionali.

Un'altra differenza si riscontra nella percentuale dei diplomati, maggiore al liceo e decisamente inferiore negli istituti tecnici e professionali.

Ritornando all'analisi della mobilità degli studenti da una tipologia di scuola all'altra, senza tener conto della promozione da una classe alla successiva, si può costruire un'ulteriore tavola di sopravvivenza, che non tiene conto della classe frequentata ma solo della presenza dello studente in ciascun anno scolastico considerato. Dalla Tabella 37 si evince che, su 1.000 studenti che hanno iniziato il liceo oppure l'istituto tecnico nel 2002/03, 912 erano ancora al liceo o al tecnico nel 2006/07, indipendentemente dalla classe frequentata in quell'anno, quindi al netto del successo in termini di promozione. Su 1.000 studenti che hanno invece iniziato l'istituto professionale nel 202/03, 964 sono rimasti al professionale fino al 2006/07. Ciò dimostra ancora una volta la maggiore capacità degli istituti professionali di trattenere gli studenti, rispetto ai licei e agli istituti tecnici.

[31] Studenti con un percorso di studi regolare

Tabella 37 – Percentuali di "sopravvivenza" per gli studenti che frequentavano diverse tipologie di scuola, distinzione per presenza

	2003/04	2004/05	2005/06	2006/07
Percentuali di studenti ancora a scuola sui 2.403 che frequentavano la classe prima al liceo	96,8	94,8	93,7	91,2
Percentuali di studenti ancora a scuola sui 124 che frequentavano la classe prima all'istituto artistico	98,4	96,8	95,2	95,2
Percentuali di studenti ancora a scuola sui 1.299 che frequentavano la classe prima all'istituto tecnico	96,6	94,6	92,6	91,2
Percentuali di studenti ancora a scuola sui 559 che frequentavano la classe prima all'istituto professionale	98,4	97,5	96,6	96,4

Dalla Tabella 37, si evince anche che la maggior parte degli studenti che hanno lasciato la tipologia di scuola inizialmente scelta lo ha fatto nel secondo anno: al liceo, coloro che hanno cambiato tipologia di scuola nel secondo anno sono il 40% di tutti coloro che lo hanno fatto; l'analoga percentuale è del 45% per gli istituti professionali e del 46% per gli istituti tecnici.

Oltre al cambiamento di tipologia di scuola, seguendo la distinzione tra licei, istituti artistici, tecnici e professionali, gli studenti hanno effettuato, nel periodo considerato, anche cambiamenti in termini di indirizzo di studio, inteso all'interno della medesima tipologia di scuola (ad esempio, uno studente che passa dal liceo classico allo scientifico).

Considerando soltanto alcune tipologie di indirizzo tra le più comuni, si è costruita una tavola di sopravvivenza (Tabella 38) che riporta le percentuali di studenti di successo, intendendo come tale la prosecuzione della scelta inizialmente fatta in termini di indirizzo di studio unitamente alla promozione.

Tabella 38 – Percentuali di "sopravvivenza" per gli studenti che hanno proseguito con esito positivo la scelta inizialmente fatta in termini di indirizzo di studi, distinzione per classe frequentata

	Classe II	Classe III	Classe IV	ClasseV	Diplomati
Percentuali di studenti regolari sui 314 studenti che frequentavano la classe prima nel 2002/03 al liceo classico	89,2	84,7	78,3	77,4	76,4
Percentuali di studenti regolari sui 1.543 studenti che frequentavano la classe prima nel 2002/03 al liceo scientifico	89,4	84,6	80,5	76,3	75,1
Percentuali di studenti regolari sui 205 studenti che frequentavano la classe prima nel 2002/03 al liceo linguistico o scienze sociali	84,0	79,4	75,1	73,2	72,3
Percentuali di studenti regolari sui 505 studenti che frequentavano la classe prima nel 2002/03 all'istituto tecnico industriale	89,5	80,8	69,5	63,8	60,2
Percentuali di studenti regolari sui 512 studenti che frequentavano la classe prima nel 2002/03 all'istituto tecnico commerciale	92,2	87,9	80,9	68,6	66,7

Tabella 38 – Percentuali di "sopravvivenza" per gli studenti che hanno proseguito con esito positivo la scelta inizialmente fatta in termini di indirizzo di studi, distinzione per classe frequentata

	Classe II	Classe III	Classe IV	ClasseV	Diplomati
Percentuali di studenti regolari sui 163 studenti che frequentavano la classe prima nel 2002/03 all'istituto professionale industriale	91,4	81,6	70,6	57,7	54,0
Percentuali di studenti regolari sui 171 studenti che frequentavano la classe prima nel 2002/03 all'istituto professionale per commercio o turismo	89,5	84,8	77,2	74,3	69,0

In Tabella 38 si può vedere che le percentuali sono differenti anche all'interno dei diversi indirizzi di studio: il liceo classico è l'indirizzo con la maggior probabilità di successo (sia in termini di prosecuzione della scelta che in termini di promozione), mentre l'istituto professionale industriale è l'indirizzo con la probabilità di successo inferiore. Si può anche notare che le probabilità di successo calcolate tenendo conto dell'indirizzo di studi sono più basse rispetto a quelle calcolate tenendo soltanto conto del cambiamento in termini di tipologia di scuola: vi sono studenti che cambiano scuola, pur rimanendo all'interno di una stessa tipologia nel senso ampio del termine. Vi è molta differenza, ad esempio, tra l'indirizzo professionale per l'industria e quello professionale per il commercio: in quest'ultimo la probabilità di successo è in generale maggiore, anche se per la sola classe prima risulta inferiore. Vi sono differenze anche nelle percentuali dei diplomati, da un indirizzo di studi all'altro: la percentuale più alta di diplomati si riscontra al liceo classico (76%), mentre quella più bassa è quella dell'istituto professionale industriale (54%). Vi sono anche tipologie di indirizzi che si dimostrano attrarre gli studenti piuttosto che perderli. Se si considerano quegli studenti che frequentavano le diverse tipologie di indirizzi nell'anno scolastico 2006/07, si può vedere quanti di essi vi sono rimasti a seguito di una scelta iniziale e quanti invece vi sono arrivati in un momento successivo al primo anno. Si ha così che il 44% degli studenti che frequentavano un istituto professionale per il commercio nel 2006/07 proveniva da altro tipo di scuola (mentre soltanto il 6% di coloro che inizialmente frequentavano tale tipo di scuola l'hanno poi abbandonato); il 27% degli studenti che nel 2006/07 frequentavano un istituto professionale industriale proveniva da altro tipo di scuola (mentre il 19% degli studenti che inizialmente frequentavano tale tipo di scuola l'hanno poi lasciato); le analoghe percentuali per i diversi indirizzi degli istituti tecnici rimangono attorno al 13%, quindi stanno molto al di sotto di quelle relative ai professionali (anche la percentuale di studenti che ha abbandonato si pone sullo stesso livello, il bilancio in questo caso è in pareggio).

Su 1.000 studenti che hanno iniziato l'istituto professionale per il commercio nel 2002/03, 1.513 si trovavano a frequentare il medesimo tipo di scuola nel 2006/07 (ciò denota un sensibile aumento); su 1.000 studenti che hanno invece iniziato l'istituto professionale industriale, ve ne erano 1.364 nel 2006/07. Il conteggio analogo effettuato sugli istituti tecnici produce un amento di circa 100 studenti. In Grafico 23, sono illustrate le probabilità di successo: queste sono calcolate come probabilità che uno studente rimanga nella stessa tipologia di scuola, pur proseguendo la classe frequentata nei diversi anni scolastici considerati (successo in termini di prosecuzione della scelta e di regolarità).

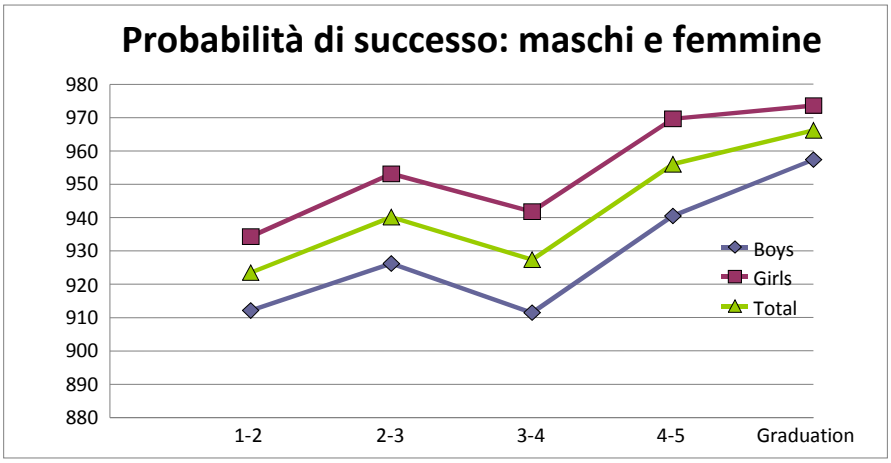


Grafico 23 – Probabilità di passaggio da ogni classe alla successiva, per tutti gli studenti e distintamente per sesso

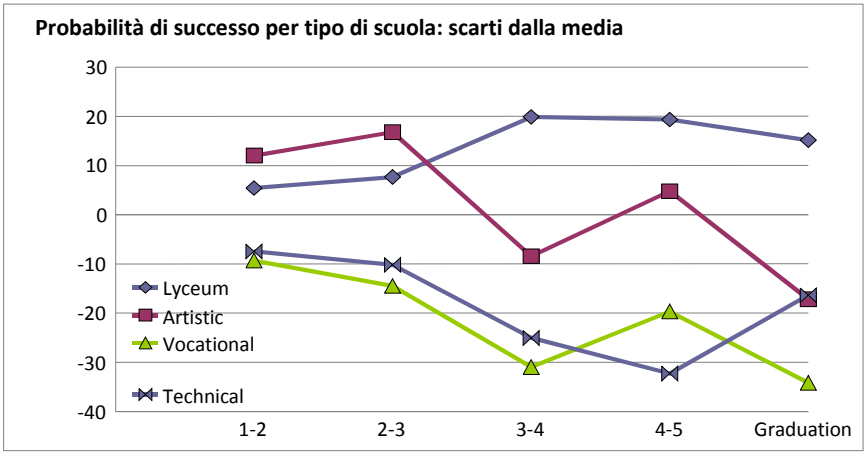


Grafico 24 – Probabilità di passaggio da ogni classe alla successiva (in termini di scarto dalla media) distintamente per tipo di scuola

Nel Grafico 23 si può vedere che è più probabile per una ragazza rimanere nello stesso tipo di istituto ed avere un risultato positivo che non per un ragazzo; la probabilità di cambiare istituto (e di non essere promossi) ha un incremento tra la classe terza e la quarta (la probabilità di successo diminuisce), così tanto che la probabilità di cambiare tipologia di istituto (e di non essere promossi) risulta pressoché la medesima dopo la classe prima e dopo la classe terza. Come mostrato in Grafico 24, gli studenti che hanno cambiato tipologia di scuola e che non sono stati promossi è più probabile che si trovassero in un istituto professionale che non in un altro tipo di scuola, mentre le percentuali più basse di risultati negativi si registrano nei licei. Per quanto riguarda il Grafico 24, riportante la distinzione per tipologia di scuola, occorre tener presente che sono rappresentati gli scarti dalla media: in realtà le differenze sono molto basse (il range delle probabilità di successo è [900; 980]). È anche interessante analizzare il comportamento degli studenti (Tabella 39), ponendo uguale a 1.000 il numero di studenti nelle diverse tipologie di scuola nel primo anno scolastico considerato e in classe prima, in modo tale da poter confrontare in termini relativi la prosecuzione della scelta fatta ed allo stesso tempo la regolarità.

Tabella 39 – Studenti rimasti nella scuola inizialmente scelta (dati 1.000 studenti che frequentavano la classe prima di ogni tipologia di istituto nel 2002/03)

	Licei	Istituti artistici	Istituti tecnici	Istituti professionali
Classe I	1.000	1.000	1.000	1.000
Classe II	968	984	966	984
Classe III	948	968	946	975

Tabella 39 – Studenti rimasti nella scuola inizialmente scelta (dati 1.000 studenti che frequentavano la classe prima di ogni tipologia di istituto nel 2002/03)

	Licei	Istituti artistici	Istituti tecnici	Istituti professionali
Classe IV	937	952	926	966
Classe V	912	952	912	964
Diplomati	875	903	816	810

Dalla Tabella 39 si evince che gli istituti professionali sembrano le scuole da cui si esce di più; tuttavia, anche alla luce dei risultati precedentemente ottenuti, occorre tener conto del fatto che in questo caso il dato risente fortemente del fattore promozione, quindi è a maggior ragione vero che negli istituti professionali si viene promossi di meno, ma non che da essi si esce di più. A questo proposito è utile calcolare le percentuali di studenti che provenivano dalla stessa tipologia di scuola su tutti quelli che erano presenti a scuola nel 2006/07: dato il numero assoluto di studenti che frequentavano ciascuna tipologia di scuola nel 2006/07 (2.210 i licei, 144 gli istituti artistici, 1.330 gli istituti tecnici, 701 gli istituti professionali), si è calcolato il numero di studenti che si trovavano nel medesimo tipo di scuola anche nel 2002/03. Al fine di agevolare i confronti, il dato si è poi standardizzato, ponendo uguale a 1.000 gli studenti che frequentavano ciascuna tipologia di scuola nell'anno scolastico 2002/03; in tal modo si è ottenuto che 1.301 studenti frequentavano gli istituti professionali nel 2006/07, 1.122 i tecnici e solo 1.008 i licei. Ecco un'ulteriore dimostrazione del fatto che gli istituti professionali più che perdere studenti in realtà ne sono collettori.

Nel seguente Grafico 25, viene mostrato questo fatto, considerando gli spostamenti degli studenti da una tipologia di scuola all'altra senza invece considerare la promozione da una classe alla successiva. Si può così evidenziare la probabilità di rimanere nella stessa tipologia di scuola dall'inizio alla fine del percorso scolastico, al netto della promozione. Il Grafico 25 riporta, in termini di scarti dalla media, le probabilità che uno studente, che frequentava un dato tipo di scuola nel 2002/03, vi sia rimasto fino al 2006/07. I licei sono quelle scuole da cui è più probabile spostarsi.

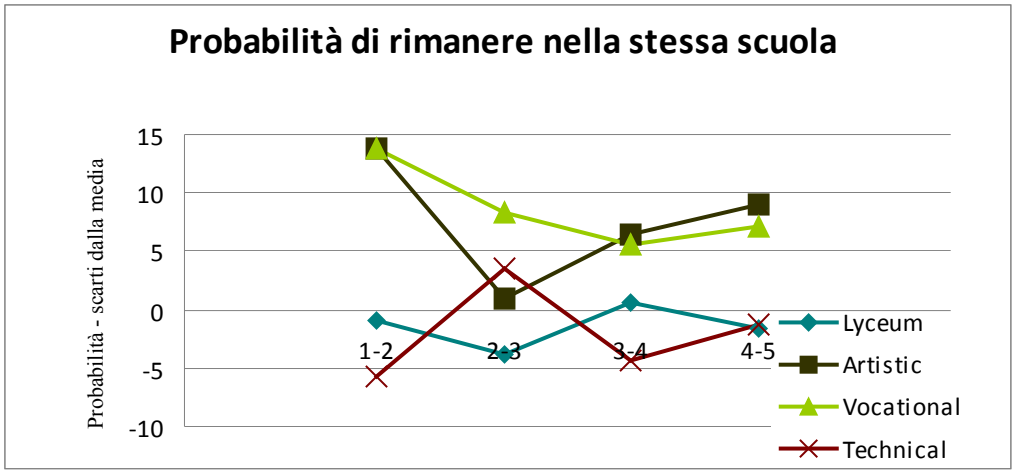


Grafico 25 – Probabilità di permanenza nella stessa scuola nel passaggio da una classe alla successiva: distinzione per tipo di scuola

Si è scelto di introdurre nel modello anche un'ulteriore variabile, che rappresenta la necessità del singolo studente di spostarsi rispetto alle vicinanze della propria abitazione per raggiungere la scuola frequentata. A tal fine, si è tenuto conto di una suddivisione, operata dalla Provincia di Bologna, della provincia stessa in 7 diversi

ambiti territoriali. A partire dall'ambito di appartenenza della scuola frequentata da ogni singolo studente e dall'ambito di appartenenza della sua residenza, si è costruita una variabile dicotomica indicante (valore 0) se lo studente rimane nello stesso ambito, oppure se (valore 1) frequenta un istituto che si trova in ambito diverso. Occorre soffermarsi sull'interpretazione di tale variabile, costruita con l'intento di distinguere gli studenti che, per scelta o per necessità, devono ogni giorno effettuare un lungo percorso per raggiungere la scuola da coloro che invece vi abitano vicino. Non è dato conoscere a priori se tale differenza abbia ripercussioni positive o negative sull'esito scolastico: ci si può benissimo aspettare che la lontananza dal luogo di studio abbia un'influenza negativa laddove si tratta di scelta obbligata, mentre che non abbia alcuna influenza laddove si tratta di scelta libera da parte del singolo studente. Una limitazione importante di questa scelta è la presenza di zone di confine, per cui può accadere che uno studente frequenti una scuola in ambito diverso, ma in realtà più vicina alla propria abitazione rispetto ad una scuola dello stesso ambito. Un altro fattore da tener presente è che non è detto che il dato sulla residenza indicato nel database sia quello di domicilio effettivo. Per questi motivi occorre attendersi risultati complessi e di difficile interpretazione, dall'introduzione di questa variabile nel modello. Analizzando i dati sul contingente studiato, sembra che gli studenti che dimorano nello stesso ambito in cui si trova la scuola abbiano risultati medi, in termini di promozione, leggermente migliori degli altri, seppure con una differenza che non è detto sia significativa, come mostrato dal Grafico 26.

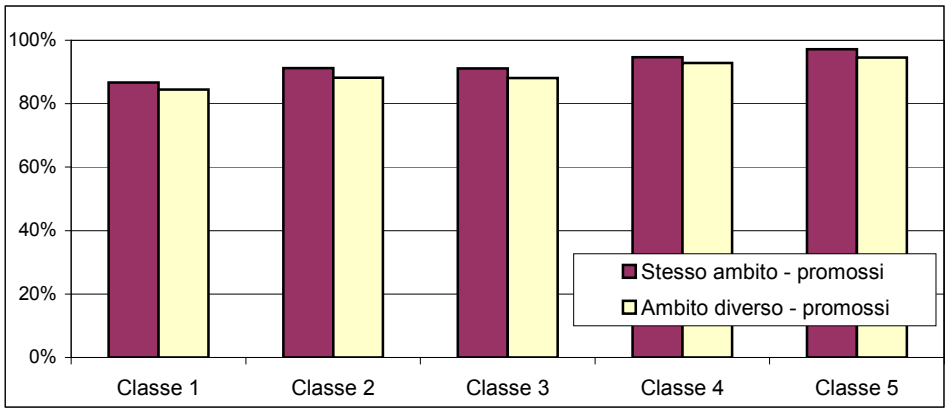


Grafico 26 – Probabilità di promozione degli studenti per ambito territoriale, di residenza e della scuola, e classe frequentata

1.6 IL FATTORE "TIPOLOGIA DI SCUOLA"

All'inizio dell'anno scolastico 2007/08, agli studenti che frequentavano il primo anno di alcune scuole secondarie di secondo grado della provincia di Bologna è stato somministrato, a cura della Provincia, un questionario finalizzato a rilevare alcune informazioni riguardanti il background familiare degli studenti stessi. Il campione di scuole, inizialmente scelto in modo tale da rappresentare la realtà scolastica bolognese, era alla fine formato dai dati provenienti da quegli istituti che effettivamente avevano fornito una risposta; si era così conseguito uno sbilanciamento del numero di ragazzi in favore dei frequentanti i licei. I dati raccolti sono poi stati elaborati e si è riusciti ad ottenere un indicatore di contesto familiare che contribuisce a spiegare il successo o l'insuccesso scolastico del singolo alunno[32].

In particolare, per giungere a tale indicatore sintetico, si sono utilizzate alcune informazioni sui componenti del nucleo familiare degli studenti. La prima informazione entrata nell'indicatore è lo status occupazionale dei genitori, con la suddivisione dei diversi rami di attività in 6 gruppi principali: nel primo gruppo entrano le attività imprenditoriali (anche di tipo agricolo), di dirigente e di libero professionista; il secondo gruppo comprende artigiani e commercianti; fanno invece parte del terzo gruppo insegnanti, impiegati, militari e quadri intermedi; operai comuni e specializzati entrano nel quarto gruppo; il quinto è costituito da lavoratori a domicilio e collaboratori domestici, mentre nel sesto gruppo entrano tutte le altre tipologie di attività.

Importante indicatore del contesto familiare è stato ritenuto anche lo stato civile dei genitori, combinato con le informazioni sulla convivenza; si sono così considerati diversi gruppi di studenti: quelli che vivono con genitori sposati o comunque conviventi, quelli con genitori separati o divorziati in casa, infine coloro che vivono soltanto con la madre oppure con il padre.

Altra informazione considerata nell'analisi è stato il titolo di studio dei genitori, utilizzando una suddivisione in 3 macro categorie: titolo di studio uguale o inferiore alla terza media, qualifica professionale o diploma, laurea o post laurea.

Si è infine introdotta nell'indicatore la variabile Classe sociale, costruita a partire da due informazioni: quella relativa all'attività lavorativa dei genitori e quella relativa al loro titolo di studio. Si sono costruiti così 4 gruppi di studenti: quelli con genitori appartenenti alla cosiddetta Borghesia, che sono quindi dirigenti, imprenditori, liberi professionisti, quadri intermedi oppure militari con almeno la laurea; quelli con genitori appartenenti alla Classe media impiegatizia, che cioè sono impiegati, insegnanti, militari senza laurea o comunque lavoratori dipendenti con almeno il diploma; quelli ancora con genitori appartenenti alla Piccola borghesia urbana e rurale, ossia artigiani, commercianti o lavoratori autonomi; quelli infine con genitori che possono considerarsi appartenenti alla Classe operaia, cioè operai, comuni o specializzati, lavoratori domiciliari, militari senza diploma o altri lavoratori dipendenti senza diploma.

Partendo da tali informazioni, si è arrivati ad un indicatore sintetico della condizione familiare dello studente, riuscendo ad utilizzare poi tale informazione come variabile esplicativa in un modello che spiega il rendimento scolastico, in termini di promozione alla classe successiva. Si è così dimostrato che un elemento di fondamentale importanza per il rendimento scolastico è appunto la condizione familiare di partenza: gli studenti con un valore

[32] Tesi di laurea di Irene Martelli – a.a. 2008/09 Facoltà di Scienze Statistiche

basso dell'indicatore, cioè quegli studenti che vivono in un contesto famigliare caratterizzato da disunione dei componenti, da basso livello culturale e probabilmente da un basso livello di reddito, dato dalla tipologia di attività lavorativa, conseguono in media risultati peggiori dei loro coetanei che invece vivono in un contesto famigliare più agiato.

Utilizzando gli stessi dati campionari, in questa sede, si è condotta un'analisi di tipo descrittivo, atta a sondare se esista una relazione tra il contesto famigliare (quindi le variabili che sono entrate a far parte dell'indicatore sintetico costruito) e la tipologia di scuola frequentata. I dati di base per il modello a curva latente, infatti, non comprendono alcuna informazione riguardo al contesto famigliare degli studenti; tuttavia sarebbe interessante scoprire se l'informazione riguardo alla tipologia di scuola frequentata possa essere utilizzata in qualche modo anche come indicatore del contesto famigliare (una *proxy*). Se infatti dai dati emergesse che gli studenti appartenenti a famiglie benestanti, unite e acculturate sono orientati alla scelta del liceo, mentre coloro che appartengono a famiglie a basso reddito, con genitori che hanno un titolo di studio non elevato e magari nemmeno conviventi, frequentano per lo più gli istituti professionali o al massimo i tecnici, si potrebbero interpretare i risultati raggiunti con il modello LCM anche alla luce di questo. Vista la differenza significativa, in termini di rendimento scolastico, tra gli studenti frequentanti le diverse tipologie di scuola, questa non sarà da attribuirsi soltanto alla diversità intrinseca tra una scuola e l'altra, ma anche senz'altro alla diversa composizione della popolazione scolastica. Quindi i possibili interventi da parte di un soggetto pubblico dovranno essere orientati, in special modo per alcune tipologie di scuola, non soltanto a migliorare l'offerta formativa, ma anche a mettere in atto strategie per sanare le singole situazioni di difficoltà derivanti dal contesto famigliare, così da permettere a tutti gli studenti, indipendentemente dal contesto di partenza, il raggiungimento di buoni risultati.

Al fine di mostrare che la relazione tra contesto famigliare e tipologia di scuola frequentata esiste, viene nel seguito presentata un'analisi descrittiva della composizione degli studenti nel campione sul quale è stata condotta l'indagine si cui sopra esposto.

Considerando il tipo di attività lavorativa dei genitori, la distribuzione per tipologia di scuola risulta quella in Tabella 40. In questa e nelle seguenti non vengono considerati gli studenti che presentano valori mancanti (quindi il totale degli studenti considerati nelle diverse tabelle è variabile e mai pari al totale degli studenti del campione, cioè 2.029).

Tabella 40 - Distribuzione degli studenti per stato occupazionale del padre

Tipologia di scuola	Stato occupazionale del padre	Studenti	% composizione	% Promossi
Liceo	Operaio comune o specializzato; collaboratore domiciliare	140	12,94%	80,71%
	Altro	942	87,06%	91,72%
Istituto Tecnico	Operaio comune o specializzato; collaboratore domiciliare	114	24,95%	76,32%
	Altro	343	75,05%	80,76%
Istituto Professionale	Operaio comune o specializzato; collaboratore domiciliare	124	43,21%	75,00%
	Altro	163	56,79%	71,17%

Per una corretta interpretazione dei dati presentati, occorre sottolineare che le percentuali di promozione sono influenzate anche dalla numerosità campionaria dei singoli gruppi di unità statistiche, quindi alcuni gruppi a scarsa numerosità offrono stime della probabilità di promozione scarsamente precise. Invece, quindi, di soffermarsi sul valore della stima, è da focalizzare l'attenzione sulla differenza tra le probabilità di promozione dei diversi gruppi, ma soprattutto sulle percentuali di composizione degli studenti frequentanti una stessa tipologia di scuola.

Tabella 41 - Distribuzione degli studenti per stato occupazionale della madre

Tipologia di scuola	Stato occupazionale della madre	Studenti	% composizione	% Promossi
Liceo	Operaio comune o specializzato; collaboratore domiciliare	83	8,33%	72,29%
	Altro	913	91,67%	91,57%
Istituto Tecnico	Operaio comune o specializzato; collaboratore domiciliare	95	23,28%	66,32%
	Altro	313	76,72%	82,75%
Istituto Professionale	Operaio comune o specializzato; collaboratore domiciliare	88	34,38%	68,18%
	Altro	168	65,62%	76,19%

Nelle Tabelle 40 e 41 si è evidenziata la differenziazione tra gli studenti con genitori che svolgono le mansioni di operaio e quelli i cui genitori sono imprenditori, liberi professionisti, dirigenti, impiegati, insegnanti, ecc. Già operando tale semplice suddivisione si vede la notevole differenza nella probabilità di essere promossi: in tutte le tipologie di scuola (tranne che nell'istituto professionale, nel caso di suddivisione per professione del padre, ma può esserci un problema legato anche alla non elevata numerosità campionaria) i figli degli operai ottengono un minore successo scolastico. Tuttavia tale differenza è senz'altro accentuata nei licei, mentre è decisamente più bassa negli istituti professionali; ciò si rileva anche confrontando le percentuali calcolate con la probabilità di promozione totale, senza la suddivisione per professione dei genitori: in generale si ha che la probabilità di promozione al liceo è attorno al 90% (per i figli di operai scende all'80%), mentre al tecnico rimane intorno al 78% (scende al 66% per i figli di madri operaie), fino a scendere fino al 70% nell'istituto professionale.

Ma da questi dati emerge un'altra importante realtà: i figli di genitori benestanti è più probabile che frequentino il liceo (si nota che in questa tipologia di scuola soltanto il 10% degli studenti ha genitori operai), mentre quegli studenti che hanno genitori operai hanno maggiore propensione alla frequenza degli istituti professionali (ben il 60% degli studenti che frequentano questa tipologia di scuola ha genitori operai). Se si considerano gli studenti che hanno il padre operaio, questi sono ugualmente distribuiti nelle tre tipologie di scuola, leggermente a favore dei licei (il 37% di essi frequenta il liceo, il 30% l'istituto tecnico ed il 33% il professionale); considerando invece tutti gli altri studenti, si ha che ben il 65% di essi frequenta il liceo, mentre solo l'11% frequenta l'istituto

professionale. Se si esaminano i dati dal punto di vista dell'occupazione della madre, la situazione risulta del tutto analoga, ma la percentuale dei figli di madri non operaie frequentanti il liceo sale al 68%.
Anche dall'analisi della distribuzione degli studenti per tipologia famigliare, distinguendo quelli provenienti da un ambiente unito da coloro che hanno a che fare con situazioni di divisione dei componenti del nucleo famigliare, emerge che la proporzione di coloro che ottengono la promozione è superiore per il primo gruppo (Tabella 42), indistintamente in tutte le tipologie di scuola (al liceo la probabilità di promozione scende dal 90% media generale all'82%, al tecnico dal 78% al 67%, al professionale dal 70% al 62%).

Tabella 42 - Distribuzione degli studenti per tipologia famigliare

Tipologia di scuola	Tipologia famigliare	Studenti	% composizione	% Promossi
Liceo	Studente che vive con genitori sposati o conviventi	971	83,42%	90,73%
	Altra situazione	193	16,58%	82,38%
Istituto Tecnico	Studente che vive con genitori sposati o conviventi	305	79,63%	79,67%
	Altra situazione	78	20,37%	66,67%
Istituto Professionale	Studente che vive con genitori sposati o conviventi	168	63,64%	77,38%
	Altra situazione	96	36,36%	62,50%

La distribuzione degli studenti per tipologia famigliare mostra anche la maggiore propensione degli studenti che hanno problemi famigliari verso la scelta dell'istituto professionale, a discapito del liceo: oltre l'80% dei frequentanti il liceo sono ragazzi provenienti da famiglie solide, mentre la medesima percentuale per gli istituti tecnici scende a poco meno dell'80% e per gli istituti professionali cala drasticamente al 64%. Se si dà uno sguardo alla composizione degli studenti provenienti da famiglie con situazione stabile, si nota che il 67% di essi frequenta il liceo, mentre solo l'11% di essi frequenta l'istituto professionale.
Se si considera il titolo di studio dei genitori, in particolare la variabile che riassume il titolo di studio più elevato tra i due, si ha una ulteriore conferma del fatto che gli studenti provenienti da famiglie meno agiate, quindi con anche più bassi livelli culturali, non solo conseguono risultati peggiori, ma si concentrano negli istituti professionali a scapito dei licei (Tabella 43). Si può notare che risulta che negli istituti professionali la percentuale di promozione sia più elevata per chi ha genitori con titolo di studio inferiore, tuttavia occorre considerare la scarsa numerosità dell'altro gruppo (soltanto 29 studenti), che fa in modo che la stima della percentuale in popolazione non sia attendibile. Importante da segnalare è invece la proporzione di studenti con almeno uno dei genitori laureato, che negli istituti professionali è solo del 9%, contro il 43% dei licei.

Tabella 43 - Distribuzione degli studenti per titolo di studio massimo dei due genitori

Tipologia di scuola	Titolo di studio massimo	Studenti	% composizione	% Promossi
Liceo	Inferiore al diploma	625	56,77%	87,52%

Tabella 43 - Distribuzione degli studenti per titolo di studio massimo dei due genitori

Tipologia di scuola	Titolo di studio massimo	Studenti	% composizione	% Promossi
	Laurea o post laurea	476	43,23%	92,86%
Istituto Tecnico	Inferiore al diploma	425	89,29%	77,41%
	Laurea o post laurea	51	10,71%	86,27%
Istituto Professionale	Inferiore al diploma	282	90,68%	73,76%
	Laurea o post laurea	29	9,32%	65,52%

Considerando gli studenti che hanno almeno uno dei due genitori laureato, è importante sapere che l'86% di essi frequenta il liceo, mentre solo il 9% frequenta l'istituto tecnico e il 5% frequenta l'istituto professionale.

Sempre considerando le variabili incluse nell'indicatore di contesto familiare, si passa ora a considerare la Classe sociale, variabile costruita[33] tenendo conto dell'attività lavorativa e del titolo di studio dei genitori, come già descritto.

La distribuzione degli studenti per classe sociale del padre e tipologia di scuola (Tabella 44) mostra ancora una volta che i figli di genitori benestanti conseguono risultati scolastici migliori degli altri: al liceo la probabilità di promozione sale oltre il 90% medio generale per i figli di genitori appartenenti almeno alla classe media, mentre rimane al di sotto dell'80% per i figli degli operai; analogo andamento si rileva per le altre tipologie di scuola. Analizzando la composizione degli studenti, si evince che la maggior parte degli iscritti al liceo ha il padre appartenente alla borghesia o alla classe media impiegatizia (circa il 74%), mentre più del 50% degli iscritti all'istituto professionale ha un padre operaio; il tecnico si pone a metà strada, con un 51% di iscritti aventi il padre appartenente alla borghesia o alla classe media impiegatizia. È anche interessante notare che il 70% degli studenti con padre appartenente alla borghesia o alla classe media impiegatizia frequenta il liceo, il 20% l'istituto tecnico e solo il 10% frequenta l'istituto professionale. Invece gli studenti con padre operaio sono ugualmente suddivisi nelle tre tipologie di scuola (36% al liceo e 32% rispettivamente al tecnico e al professionale).

Tabella 44 - Distribuzione degli studenti per classe sociale del padre

Tipologia di scuola	Classe sociale del padre	Studenti	% composizione	% Promossi
Liceo	Classe operaia urbana e rurale	164	15,16%	79,88%
	Piccola borghesia urbana e rurale	121	11,18%	91,74%
	Classe media impiegatizia	373	34,47%	93,30%
	Borghesia	424	39,19%	91,27%
Istituto Tecnico	Classe operaia urbana e rurale	143	31,29%	76,22%
	Piccola borghesia urbana e rurale	81	17,72%	70,37%

[33] Tesi di laurea di Irene Martelli – a.a. 2008/09 Facoltà di Scienze Statistiche

Tabella 44 - Distribuzione degli studenti per classe sociale del padre

Tipologia di scuola	Classe sociale del padre	Studenti	% composizione	% Promossi
	Classe media impiegatizia	148	32,39%	87,16%
	Borghesia	85	18,60%	81,18%
Istituto Professionale	Classe operaia urbana e rurale	145	50,52%	71,72%
	Piccola borghesia urbana e rurale	37	12,89%	70,27%
	Classe media impiegatizia	62	21,60%	74,19%
	Borghesia	43	14,98%	76,74%

Anche la distribuzione degli studenti per classe sociale della madre (Tabella 45) mostra un generale andamento migliore di coloro i quali hanno la madre appartenente ad una classe sociale alta. La percentuale relativa ai figli di madre appartenente alla borghesia che frequentano il professionale non è da considerarsi stima precisa, data la bassa numerosità campionaria (24 unità statistiche). Rispetto alla classe sociale del padre, la situazione cambia leggermente, anche se l'andamento complessivo delle percentuali di composizione non è del tutto diverso. La percentuale di ragazzi con madre appartenente alla borghesia o alla classe media impiegatizia rispetto ai frequentanti il liceo è più alta che nel caso precedente, raggiungendo l'83%, ma con una percentuale maggiore di madri appartenenti alla classe media. Ciò è spiegabile dal fatto che è ancora vero che meno donne che non uomini raggiungono posizioni di responsabilità. Questo è confermato dalle basse (inferiori rispetto ai padri) percentuali di studenti con madre che appartiene alla borghesia in tutte le tipologie di scuola.

Analogamente alla distribuzione per classe sociale del padre, è bassa la proporzione di studenti con madre operaia che frequenta il liceo: solo il 10% dei frequentanti il liceo ha madre operaia, mentre l'analoga percentuale per il tecnico è il 29% e per il professionale è il 40%. Come accade per la classe sociale del padre, però, il contingente di studenti con madre operaia è ugualmente distribuito tra le diverse tipologie di scuola (il 37% frequenta il tecnico, il 31% il liceo e il 32% il professionale). Il contingente, invece, dei figli di madre appartenente alla borghesia o al ceto medio è diversamente ripartito tra le diverse tipologie di scuola: il 68% frequenta il liceo, il 21% frequenta il tecnico e solo l'11% il professionale (senza sostanziale differenziazione rispetto alla classe sociale del padre). Ciò a sottolineare che sono proprio i figli di genitori appartenenti alle classi più alte ad orientarsi verso i licei, mentre i figli degli operai non si può dire abbiano una propensione decisa verso un tipo di scuola.

Tabella 45 - Distribuzione degli studenti per classe sociale della madre

Tipologia di scuola	Classe sociale della madre	Studenti	% composizione	% Promossi
Liceo	Classe operaia urbana e rurale	101	10,14%	73,27%
	Piccola borghesia urbana e rurale	64	6,43%	87,50%

Tabella 45 - Distribuzione degli studenti per classe sociale della madre

Tipologia di scuola	Classe sociale della madre	Studenti	% composizione	% Promossi
	Classe media impiegatizia	630	63,25%	92,06%
	Borghesia	201	20,18%	92,54%
Istituto Tecnico	Classe operaia urbana e rurale	117	28,68%	66,67%
	Piccola borghesia urbana e rurale	38	9,31%	73,68%
	Classe media impiegatizia	222	54,41%	86,04%
	Borghesia	31	7,60%	80,65%
Istituto Professionale	Classe operaia urbana e rurale	102	39,84%	69,61%
	Piccola borghesia urbana e rurale	20	7,81%	70,00%
	Classe media impiegatizia	110	42,97%	79,09%
	Borghesia	24	9,38%	66,67%

Considerando la classe sociale massima dei due genitori (Tabella 46), non emerge nulla di nuovo. Viene confermata la propensione dei figli di genitori appartenenti alla borghesia o alla classe media impiegatizia a scegliere il liceo (il 67% dei figli di tali genitori frequenta il liceo, il 21% frequenta il tecnico e solo l'11% frequenta il professionale); invece per quanto riguarda i figli degli operai non vi è una scelta netta, anche se, analizzando i dati ripartiti per classe sociale massima, emerge che in realtà gli studenti che hanno entrambi i genitori operai (quelli che appunto hanno tale classe come massima) hanno una propensione per il tecnico e il professionale rispetto al liceo (tra tutti coloro che hanno entrambi i genitori operai, il 35% sceglie il tecnico, il 37% sceglie il professionale e solo il 28% sceglie il liceo).

Tabella 46 - Distribuzione degli studenti per classe sociale massima dei due genitori

Tipologia di scuola	Classe sociale massima	Studenti	% composizione	% Promossi
Liceo	Classe operaia urbana e rurale	52	5,38%	61,54%
	Piccola borghesia urbana e rurale	43	4,45%	81,40%
	Classe media impiegatizia	453	46,85%	93,60%
	Borghesia	419	43,33%	91,89%
Istituto Tecnico	Classe operaia urbana e rurale	64	16,80%	68,75%
	Piccola borghesia urbana e rurale	39	10,24%	64,10%
	Classe media impiegatizia	195	51,18%	85,64%
	Borghesia	83	21,78%	81,93%

Tabella 46 - Distribuzione degli studenti per classe sociale massima dei due genitori

Tipologia di scuola	Classe sociale massima	Studenti	% composizione	% Promossi
Istituto Professionale	Classe operaia urbana e rurale	67	28,76%	70,15%
	Piccola borghesia urbana e rurale	23	9,87%	60,87%
	Classe media impiegatizia	97	41,63%	80,41%
	Borghesia	46	19,74%	76,09%

Dall'analisi descrittiva condotta, emerge che effettivamente esiste un legame stretto tra il background famigliare dello studente e la sua scelta nel tipo di scuola, anche se rimane vero che esistono casi di studenti che vivono in famiglie sulla carta disagiate e che poi frequentano licei ottenendo buoni risultati. Si può però dire che, in media, gli studenti che hanno genitori con alto titolo di studio, con attività lavorativa che richiede una maggiore responsabilità e con più probabilmente un livello di reddito medio-alto, hanno una propensione maggiore a frequentare i licei, a scapito soprattutto degli istituti professionali, oltre a conseguire risultati in media migliori in tutte le tipologie di scuola. In questo senso, l'analisi dei dati sui rendimenti scolastici può ragionevolmente utilizzare la variabile "Tipologia di scuola" come riferita anche in realtà allo stato sociale dei ragazzi, specie quando si analizzano le differenze, sempre significative, tra il liceo e gli altri tipi di scuola. Viceversa non è del tutto corretto dire che i ragazzi appartenenti a famiglie di operai e caratterizzate da livelli culturali non elevati si dirigono preferibilmente verso alcuni tipi di scuola piuttosto che in altri, anche se i dati confermano che quei ragazzi cresciuti in famiglie di soli operai più difficilmente degli altri orientano le proprie scelte verso il liceo. Anche il "Rapporto sulla scuola in Italia 2010" (Fondazione Giovanni Agnelli) evidenzia che, in generale, gli studenti provenienti da famiglie con genitori che hanno un titolo di studio elevato, e quindi di classe sociale più alta, è più probabile che frequentino i licei. Fatte queste premesse, si ritiene in questa sede sufficientemente corretto basarsi sui risultati ottenuti dall'analisi del modello a curva latente, in particolare sulle differenze tra studenti che frequentano i diversi tipi di istituto, per poter affermare che tali differenze non derivano soltanto dal fatto intrinseco di frequentare un diverso tipo di scuola, quanto anche dal diverso contesto famigliare in cui vivono tali studenti, in particolare dal livello socio-economico della loro famiglia. Discorso diverso, infine, si deve fare per il contesto famigliare in termini di convivenza: è solo in parte vero che gli studenti che provengono da famiglie con entrambi i genitori conviventi sono maggiormente propensi a frequentare il liceo, dato che la percentuale dei frequentanti il liceo tra i figli di genitori conviventi è del 67%, mentre tra i figli di genitori non conviventi l'analoga percentuale è del 53%, più bassa ma non così tanto inferiore come accade per le altre variabili sopra considerate. Quindi la frequenza di una diversa tipologia di scuola non è buon indicatore della tipologia del contesto famigliare in termini di convivenza. Nel seguito (Grafico 27 e Grafico 28) è rappresentata graficamente la situazione qui descritta.

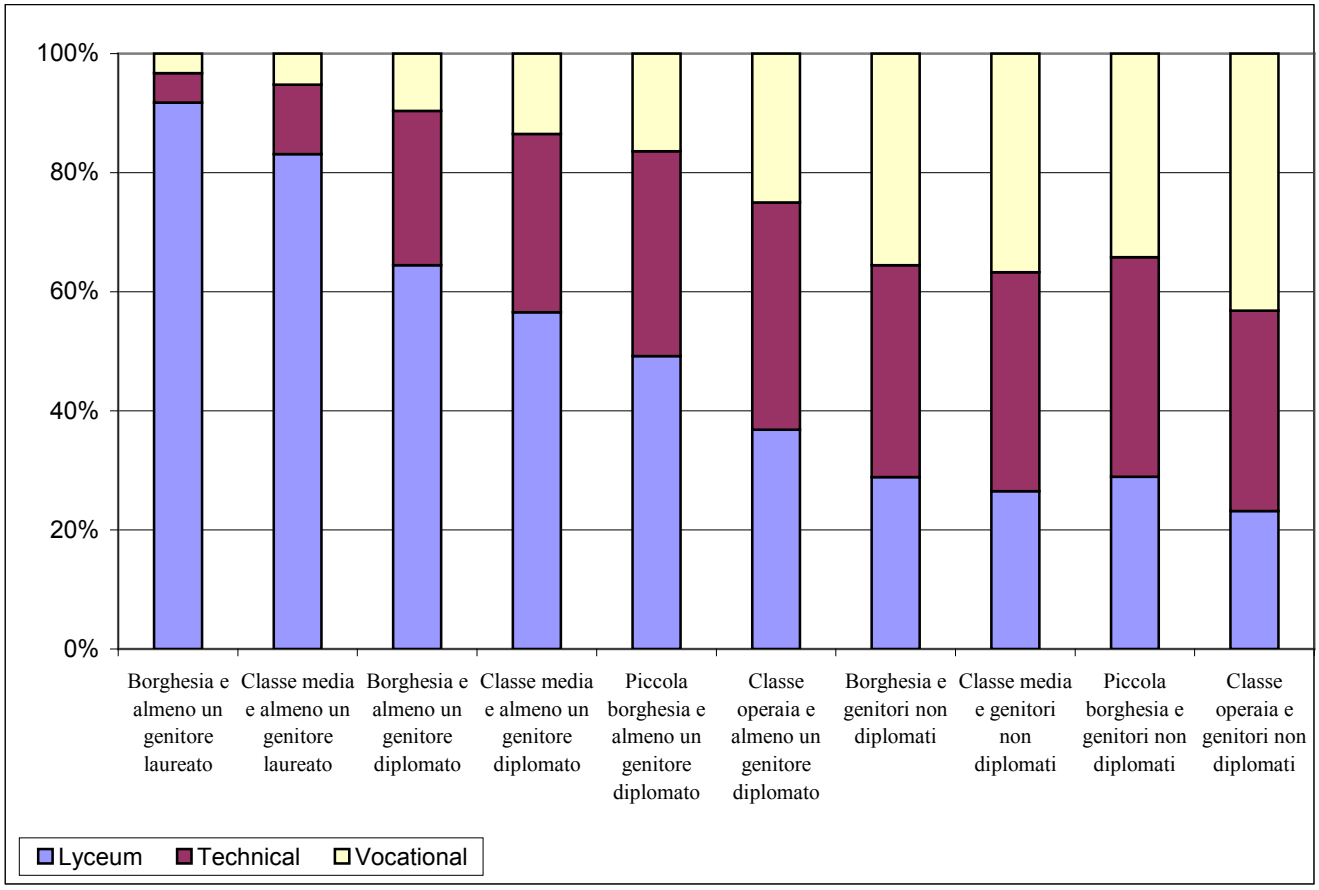


Grafico 27 – Composizione percentuale degli studenti nelle varie tipologie di scuola per status socio economico dei genitori

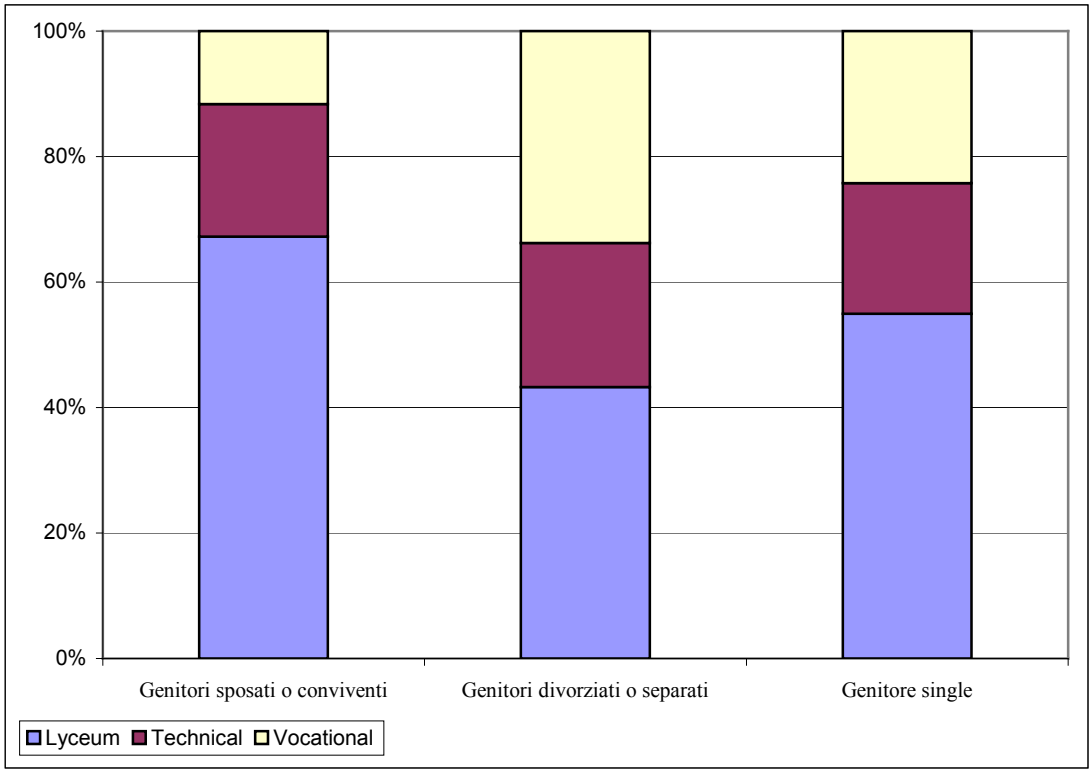


Grafico 28 – Composizione percentuale nelle varie tipologie di scuola per stato di convivenza dei genitori

2 Stima del modello logistico

Perseguendo l'obiettivo di studiare gli esiti scolastici, si è partiti dall'applicazione di un modello di regressione logistica, ponendo come variabile oggetto d'interesse, dicotomica, l'esito di fine anno scolastico, in termini di promozione o non promozione. In prima istanza si è infatti scelto di non introdurre l'elemento temporale, ma di trattare ogni singolo anno scolastico come a se stante. La variabile dicotomica, che assume il valore 1 in caso di esito positivo (promozione) ed il valore 0 in caso di esito negativo (non promozione), è stata usata come variabile dipendente nel modello di regressione in cui Sesso, Cittadinanza e Tipologia di scuola sono i regressori. Si è applicato il modello di regressione logistica, particolarmente adatto per variabili dicotomiche, quali l'obiettivo della presente analisi:

$\log it(p) = \ln\left(\frac{p}{1-p}\right) = X\beta = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$, dove p è la probabilità di promozione, con $p \in (0,1)$, mentre le X_j sono i regressori.

Le probabilità di successo (esito positivo), con regressori dicotomici, vengono in tal modo definite:

$$p(1) = P(Y = 1 \mid X_j = 1, X_{i \neq j} = 0) = \frac{e^{\beta_0 + \beta_j}}{1 + e^{\beta_0 + \beta_j}}$$

$$p(0) = P(Y = 1 \mid X_j = 0, X_{i \neq j} = 0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}.$$

Le variabili Sesso e Cittadinanza (se italiana o straniera) sono di per sé variabili dicotomiche (Sesso=0 sono le femmine, mentre Cittadinanza=0 sono gli studenti con cittadinanza italiana), invece la variabile Tipologia di scuola è categorica; nel caso in esame assume 4 valori: "liceo", "istituto artistico", "istituto tecnico" e "istituto professionale". Per tale variabile, si è dovuto procedere alla scomposizione in 4 variabili dicotomiche, una per ogni tipologia di scuola (la prima è stata poi usata come riferimento, quindi non introdotta nel modello). Tenendo conto anche della variabile dicotomica legata all'ambito territoriale, ecco infine il modello logistico ottenuto:

$\log it(p) = \beta_0 + \beta_1 Sex + \beta_2 Foreign + \beta_3 Artistic + \beta_4 Vocational + \beta_5 Technical + \beta_6 DifferentAmbit + \varepsilon$,

dove la probabilità obiettivo è quella di essere promossi (indipendentemente dal debito), tenendo conto dell'anno scolastico considerato e della classe frequentata. A questo punto si è dovuta prendere una ulteriore decisione, riguardo ai contingenti di individui da utilizzare per la stima dei parametri. Con la suddivisione, infatti, del dataset soltanto in base all'anno scolastico, stimando un modello per ogni gruppo di individui che frequentava le scuole bolognesi appunto in ciascun anno scolastico, sarebbero stati utilizzati per lo più gli studenti regolari, tuttavia sarebbero entrati studenti che frequentavano classi diverse, con le conseguenze di rendere maggiormente complicata l'interpretazione dei risultati, e inoltre di rendere impossibile trarre informazioni sulle determinanti gli esiti scolastici per ciascuna classe. Separando invece i diversi modelli soltanto sulla base della classe frequentata, indipendentemente dall'anno scolastico, in ciascuno di tali modelli sarebbero entrati più di una volta gli stessi individui, rilevati in anni scolastici diversi, ma frequentanti la medesima classe in quanto non promossi. Date tali considerazioni, si è infine scelto di costruire diversi modelli, uno per ciascun anno scolastico e classe frequentata, pur consapevoli della limitazione data dall'introduzione nei modelli soltanto degli individui con corso di studi regolare. Ci si aspetta, dunque, di ottenere in realtà sovrastime della probabilità effettiva di

successo. L'interpretazione della probabilità di successo deve infatti essere legata agli studenti che sono arrivati in ciascuna classe nell'anno scolastico regolare rispetto all'età.

Prima di iniziare con la stima dei diversi modelli riferiti ai diversi istanti temporali, in Tabella 47 viene riportata la distribuzione degli studenti nei diversi anni scolastici e classi considerati.

Tabella 47 – Distribuzione degli studenti considerati nei modelli di regressione logistica

Anno scolastico e classe frequentata	Numero di studenti	Femmine	Cittadinanza straniera	Ambito diverso	Studenti che frequentano il liceo
2002/03 – classe I	5.277	2.608 (49,4%)	89 (1,7%)	1.797 (34,1%)	2.615 (49,6%)
2003/04 – classe II	4.548	2.327 (51,2%)	66 (1,5%)	1.530 (33,6%)	2.384 (52,4%)
2004/05 – classe III	4.144	2.162 (52,2%)	52 (1,3%)	1.392 (33,6%)	2.247 (54,2%)
2005/06 – classe IV	3.722	1.986 (53,4%)	39 (1,0%)	1.220 (32,8%)	2.099 (56,4%)
2006/07 – classe V	3.505	1.891 (54,0%)	35 (1,0%)	1.138 (32,5%)	1.984 (56,6%)

Già dalla Tabella 47 si possono intuire i legami tra l'appartenenza degli studenti ai vari gruppi individuati dalle variabili dicotomiche e l'esito scolastico: mentre il numero assoluto degli studenti diminuisce drasticamente dalla classe prima alla quinta (si stanno analizzando i soli regolari), la percentuale delle ragazze aumenta, come anche quella degli studenti del liceo, mentre diminuisce la percentuale di cittadini non italiani ed anche quella di coloro che frequentano una scuola in ambito territoriale diverso. Si può già capire che la frequenza al liceo influisce positivamente sull'esito scolastico, come anche l'appartenenza al sesso femminile, mentre le altre due variabili influiscono negativamente sulla propensione alla promozione.

Sono state stimate 5 regressioni logistiche, dapprima introducendo tutte le variabili ed anche le interazioni tra di esse, poi scegliendo le variabili significative, attraverso una procedura di tipo *stepwise backward*, sulla base dell'indice di Akaike (AIC). Si è scelto il valore 0,1 come soglia di significatività (si sono mantenute nel modello le variabili con coefficienti significativi oltre la soglia del 90% di significatività).

Occorre considerare che il numero di studenti con cittadinanza non italiana è di molto inferiore a quello degli italiani, motivo per cui occorre essere molto cauti sull'interpretazione dei risultati relativi agli stranieri. Per esempio, i modelli stimati per gli ultimi anni scolastici considerati fanno emergere che non vi sono differenze significative tra italiani e stranieri, ma ciò è vero soltanto perché nelle ultime classi gli stranieri sono davvero pochi e quei pochi sicuramente presentano caratteristiche (dal punto di vista sociale ed economico) che li accomunano più agli studenti italiani che non a quelli stranieri. Ancora, a causa della scarsa numerosità, occorre prestare attenzione ai risultati ottenuti sugli istituti artistici, frequentati da relativamente pochi studenti.

Considerando quindi il Tempo = 0 (anno scolastico 2002/03) e la classe prima, il modello diventa (AIC = 3.931):

$$\log it(p) = 2{,}959 - 0{,}416\,Sex - 0{,}510\,Foreign - 1{,}488\,Art - 1{,}852\,Vocational - 0{,}941\,Technical + \varepsilon$$

dove l'intercetta indica il livello stimato del logaritmo della variabile obiettivo per le ragazze italiane che frequentano un liceo nello stesso ambito di residenza (in particolare, anche che nel 2002/03 frequentavano la classe prima). Come già spiegato, la variabile categorica Tipo di scuola entra nel modello nella forma di più variabili dicotomiche; in questo caso una modalità (in particolare, la modalità Liceo) viene utilizzata come riferimento, quindi non entra nel modello in forma di variabile distinta. Da questo modello emerge che la probabilità di essere promossi in classe prima, in generale, non differisce significativamente tra gli studenti che frequentano una scuola nello stesso ambito di residenza e coloro che invece debbono spostarsi in altro ambito.
Per meglio interpretare i risultati della stima del modello, viene ora calcolato l'*odds ratio* per ciascun regressore. Dati gli *odds* per ciascun possibile valore dei regressori:

$$odds\ (x_j = 0) = \frac{P(Y = 1 \mid X = x_1,..., x_j = 0,,..., x_k)}{P(Y = 0 \mid X = x_1,..., x_j = 0,,..., x_k)} = \frac{P(Y = 1 \mid X = x_1,..., x_j = 0,,..., x_k)}{1 - P(Y = 1 \mid X = x_1,..., x_j = 0,,..., x_k)} =$$
$$= \exp(\beta_0 + \beta_1 x_1 + ... + \beta_{j-1} x_{j-1} + \beta_{j+1} x_{j+1} + ... + \beta_k x_k)$$

$$odds\ (x_j = 1) = \frac{P(Y = 1 \mid X = x_1,..., x_j = 1,..., x_k)}{P(Y = 0 \mid X = x_1,..., x_j = 1,..., x_k)} = \frac{P(Y = 1 \mid X = x_1,..., x_j = 1,..., x_k)}{1 - P(Y = 1 \mid X = x_1,..., x_j = 1,..., x_k)} =$$
$$= \exp(\beta_0 + \beta_1 x_1 + ... + \beta_{j-1} x_{j-1} + \beta_j + \beta_{j+1} x_{j+1} + ... + \beta_k x_k)$$

L'*odds ratio*, per la *j*-esima variabile esplicativa risulta:

$$oddsratio(x_j) = \frac{odds(x_j = 1)}{odds(x_j = 0)} = \exp(\beta_j).$$

Tale rapporto misura l'effetto del *j*-esimo regressore sulla propensione della variabile dipendente Y ad assumere il valore 1. Gli *odds ratio* calcolati per ogni regressore sono indipendenti dagli altri.
In Tabella 48 sono mostrati gli *odds ratio* (exp{β}, dove β è il generico coefficiente di un regressore risultato significativo nella spiegazione della probabilità di successo) per tutti i regressori. Tali rapporti sono calcolati sulla base del modello stimato: rappresentano gli effetti netti di ogni singolo regressore sulla propensione alla promozione.

Tabella 48 – Odds ratio per ogni variabile esplicativa – classe I	
Maschi	0,659
Istituto artistico	0,226
Istituto tecnico	0,390
Istituto professionale	0,157
Cittadinanza straniera	0,601

Osservando gli *odds ratio,* si nota che sono tutti inferiori a 1, quindi, ad esempio, i maschi hanno una propensione ad essere promossi in classe prima inferiore rispetto alle femmine.
Le stime delle probabilità di promozione per gli studenti che frequentano la classe prima nell'anno scolastico 2002/03 (assunte come stime della probabilità di promozione genericamente in classe prima) sono rappresentate dalla seguente espressione: $\hat{p} = \frac{e^{X\hat{\beta}}}{1 + e^{X\hat{\beta}}}$.

In base a tale relazione, si possono calcolare le stime delle probabilità per alcuni individui tipici, caratterizzati da diverse combinazioni dei valori delle variabili esplicative, come mostrato in Tabella 49.

Tabella 49 – Stime delle probabilità di promozione ed intervalli di confidenza – anno scolastico 2002/03 – classe prima (valori osservati)

Studenti	$\hat{p}$ studenti Italiani	Numero di studenti	$\hat{p}$ studenti con cittadinanza straniera	Numero di studenti
Ragazze che frequentano il liceo	0,951 [0,946; 0,955] (0,950)	1.409	0,921 [0,892; 0,942] (0,900)	10
Ragazze che frequentano l'istituto artistico	0,813 [0,762; 0,855] (0,789)	128	0,724 [0,601; 0,820] (NA)	0
Ragazze che frequentano l'istituto tecnico	0,883 [0,859; 0,903] (0,888)	543	0,819 [0,741; 0,877] (0,875)	8
Ragazze che frequentano l'istituto professionale	0,752 [0,711; 0,788] (0,755)	481	0,645 [0,536; 0,741] (0,621)	29
Ragazzi che frequentano il liceo	0,927 [0,914; 0,938] (0,928)	978	0,884 [0,833; 0,921] (0,750)	4
Ragazzi che frequentano l'istituto artistico	0,742 [0,660; 0,810] (0,814)	43	0,633 [0,477; 0,766] (NA)	0
Ragazzi che frequentano l'istituto tecnico	0,832 [0,787; 0,870] (0,830)	1071	0,749 [0,634; 0,837] (0,714)	21
Ragazzi che frequentano l'istituto professionale	0,666 [0,598; 0,728] (0,662)	535	0,545 [0,411; 0,673] (0,647)	17

Quelli presentati in Tabella 49 sono i valori stimati per la variabile obiettivo, soltanto relativamente a quegli studenti, nati nel 1988, che frequentavano la classe prima nel 2002/03 (quindi gli studenti che hanno iniziato in anticipo la scuola sono esclusi).

Si ha che le probabilità di essere promossi per i soli studenti che frequentano la classe prima sono leggermente più basse delle probabilità degli studenti che frequentano la scuola (quindi la classe prima o la seconda) nell'anno scolastico 2002/03: questo conferma, come precedentemente visto, che gli studenti che hanno iniziato la scuola in anticipo hanno in generale risultati migliori degli altri.

Il modello riproduce adeguatamente il trend generale: le ragazze hanno risultati migliori dei ragazzi, gli studenti italiani hanno maggior successo degli studenti con cittadinanza straniera, gli studenti che frequentano gli istituti professionali conseguono i peggiori risultati, mentre gli studenti che frequentano il liceo ottengono in generale i risultati migliori. Si può anche notare che il modello riproduce la differenza positiva, della probabilità di essere promossi, tra gli studenti che frequentano gli istituti professionali che si trovano in ambito diverso rispetto a quello di residenza e gli studenti degli stessi istituti professionali, che però frequentano un istituto collocato nello stesso ambito di residenza (ciò vale tanto per le ragazze quanto per i ragazzi). I valori stimati sono meno lontani dai valori osservati per gli studenti italiani che non per quelli stranieri; lo stesso vale per gli studenti dei licei, degli istituti tecnici e professionali rispetto a quelli degli istituti artistici. Tuttavia, come già spiegato, non bisogna dimenticare che gli studenti con cittadinanza non italiana sono in numero molto inferiore agli altri (come anche mostrato in Tabella 42), e lo stesso dicasi per gli studenti che frequentano gli istituti artistici.

La differenza tra maschi e femmine risulta significativa (gli intervalli di confidenza sono nettamente separati) per i ragazzi dei licei e quelli dei professionali (studenti italiani, gli stranieri hanno numerosità eccessivamente ridotte per poter testare la significatività delle differenze), mentre si verifica che gli intervalli di confidenza

stimati per i ragazzi e le ragazze dei tecnici hanno alcuni punti in comune. Le differenze tra gli studenti che frequentano le diverse tipologie di scuola sono nette e significative per quanto riguarda gli studenti del liceo, i cui risultati sono migliori degli altri, e gli studenti del professionale, i cui risultati sono quelli nettamente peggiori. Gli studenti degli artistici non risultano avere risultati significativamente differenti da quelli dei tecnici o dei professionali, ma occorre tener conto della ridotta numerosità, che riduce la precisione delle stime. Gli studenti dei tecnici si collocano tra i colleghi dei licei ed i colleghi dei professionali, con risultati ancora significativamente diversi. Risultano simili in realtà i risultati conseguiti dalle ragazze dei professionali e dai ragazzi dei tecnici; ciò è visibile dalla leggera sovrapposizione dei rispettivi intervalli di confidenza.
Considerando ora il tempo = 1 (anno scolastico 2003/04) e la classe seconda, le stime dei coefficienti risultano essere tutte significative al livello di probabilità del 99%. Il modello diventa (AIC = 2.748):

$$\log it(p) = 3{,}115 - 0{,}368Sex - 1{,}082Foreign - 0{,}837Art - 1{,}524Vocational - 0{,}818Technical + \varepsilon$$

dove l'intercetta indica il livello stimato del logaritmo della variabile obiettivo per le ragazze italiane che frequentano un liceo nello stesso ambito di residenza (in particolare che nel 2003/04 frequentavano la classe seconda). Questo modello conferma che gli studenti con cittadinanza non italiana hanno ottenuto risultati significativamente peggiori degli italiani, inoltre mostra ancora una volta che le ragazze ottengono in generale risultati migliori dei ragazzi. La differenza tra studenti che frequentano la scuola in ambito diverso da quello di residenza non risultano ottenere risultati significativamente differenti rispetto agli altri. La Tabella 50 mostra gli *odds ratio* calcolati sulla base del modello stimato.

Tabella 50 – Odds ratio per ogni variabile esplicativa – classe II

Maschi	0,692
Istituto artistico	0,433
Istituto tecnico	0,441
Istituto professionale	0,218
Cittadinanza straniera	0,339

Guardando gli *odds ratio,* si nota che sono anche qui tutti inferiori a 1, quindi, ad esempio, gli studenti dei professionali hanno una propensione ad essere promossi in classe prima inferiore rispetto agli altri.
La Tabella 51 mostra alcune probabilità, stimate dal modello, di essere promossi per gli studenti che frequentano la classe seconda nell'anno scolastico 2003/04.

Tabella 51 – Stime delle probabilità di promozione – anno scolastico 2003/04 – classe seconda (valori osservati)

Studenti	$\hat{p}$ studenti Italiani	Numero di studenti	$\hat{p}$ studenti con cittadinanza straniera	Numero di studenti
Ragazze che frequentano il liceo	0,957 [0,952; 0,962] (0,949)	1343	0,884 [0,838; 0,919] (0,600)	10
Ragazze che frequentano l'istituto artistico	0,907 [0,866; 0,936] (0,902)	102	0,768 [0,623; 0,868] NA	0

Tabella 51 – Stime delle probabilità di promozione – anno scolastico 2003/04 – classe seconda (valori osservati)

Studenti	$\hat{p}$ studenti Italiani	Numero di studenti	$\hat{p}$ studenti con cittadinanza straniera	Numero di studenti
Ragazze che frequentano l'istituto professionale	0,831 [0,794; 0,862] (0,840)	362	0,625 [0,496; 0,738] (0,842)	19
Ragazze che frequentano l'istituto tecnico	0,909 [0,887; 927] (0,924)	484	0,771 [0,666; 0,850] (0,714)	7
Ragazzi che frequentano il liceo	0,940 [0,926; 0,951] (0,954)	912	0,841 [0,763; 0,897] (1,000)	5
Ragazzi che frequentano l'istituto artistico	0,871 [0,801; 0,919] (0,886)	35	0,696 [0,508; 0,836] NA	0
Ragazzi che frequentano l'istituto professionale	0,773 [0,706; 0,828] (0,757)	354	0,535 [0,380; 0,684] (0,364)	11
Ragazzi che frequentano l'istituto tecnico	0,873 [0,830; 0,907] (0,865)	890	0,700 [0,554; 0,814] (0,714)	14

Anche per la seconda classe valgono le medesime considerazioni già riportate. Il modello stima significativa la differenza tra maschi e femmine al liceo e al professionale, mentre al tecnico si ha una sovrapposizione degli intervalli di confidenza. In particolare, aumenta, rispetto al modello costruito per la prima classe, la sovrapposizione degli intervalli di confidenza tra le ragazze dei professionali ed i ragazzi dei tecnici. La sovrapposizione si verifica anche all'istituto artistico, tuttavia in questo caso si deve tener conto della scarsa precisione delle stime, dovuta alla scarsa numerosità. La significatività delle differenze in base al sesso va testata basandosi sui risultati dei soli studenti italiani; gli studenti con cittadinanza non italiana sono troppo poco numerosi. Le differenze in quanto a probabilità di promozione sono inoltre stimate essere significative tra studenti che frequentano i diversi tipi di scuola: licei (qui gli studenti hanno la più alta probabilità di essere promossi) e istituti tecnici e professionali (dove gli studenti hanno la probabilità di promozione inferiore).
Passando ora a considerare il tempo = 2 (anno scolastico 2004/05) e la classe terza, le stime dei coefficienti per sesso e tipologia di scuola risultano tutte significative al livello di probabilità del 99%, mentre il coefficiente relativo alla cittadinanza risulta significativo, ma al livello di probabilità del 95%. Il modello fa quindi emergere che la differenza tra italiani e non italiani diminuisce procedendo con l'anno scolastico e la classe frequentata. Occorre sempre tenere in considerazione il fatto che gli stranieri sono in numero molto inferiore rispetto agli italiani, inoltre che a maggior ragione gli stranieri con un percorso regolare sono ancora meno rispetto agli italiani (si stanno considerando solo coloro che frequentavano la classe terza nel 2004/05); inoltre gli stranieri regolari sono in realtà, via via procedendo con la classe frequentata, sempre più simili in realtà agli italiani. La differenza tra studenti che frequentano la scuola nello stesso ambito di residenza e gli altri non è stimata essere significativa. Il modello diventa (AIC = 2.569):

$$\log it(p) = 2{,}991 - 0{,}426 Sex - 0{,}691 Foreign - 0{,}884 Art - 1{,}270 Vocational - 0{,}773 Technical + \varepsilon$$

dove l'intercetta ha significato analogo al modello costruito per gli anni precedenti. La Tabella 52 mostra gli *odds ratio* calcolati sulla base del modello stimato.

Tabella 52 – Odds ratio per ogni variabile esplicativa – classe III	
Maschi	0,653
Istituti artistici	0,413
Istituti tecnici	0,462
Istituti professionali	0,281
Cittadinanza straniera	0,501

Dall'osservazione degli *odds ratio,* emerge che sono anche in questo caso tutti inferiori a 1, quindi, ad esempio, gli studenti con cittadinanza straniera hanno una propensione ad essere promossi in classe prima inferiore rispetto agli altri.

La Tabella 53 mostra alcune stime delle probabilità per gruppi diversi di studenti che frequentano la classe terza nell'anno scolastico 2004/05.

Tabella 53 – Stime delle probabilità di promozione – anno scolastico 2004/05 – classe terza (valori osservati)

Studenti	$\hat{p}$ studenti Italiani	Numero di studenti	$\hat{p}$ studenti con cittadinanza straniera	Numero di studenti
Ragazze che frequentano il liceo	0,952 [0,947; 0,956] (0,953)	1285	0,909 [0,863; 0,940] (0,875)	8
Ragazze che frequentano l'istituto artistico	0,892 [0,848; 0,925] (0,901)	91	0,805 [0,661; 0,897] NA	0
Ragazze che frequentano l'istituto professionale	0,848 [0,814; 0,878] (0,831)	307	0,737 [0,606; 0,836] (0,867)	15
Ragazze che frequentano l'istituto tecnico	0,902 [0,879; 0,921] (0,906)	447	0,822 [0,720; 0,892] (0,778)	9
Ragazzi che frequentano il liceo	0,929 [0,913; 0,942] (0,930)	883	0,867 [0,787; 0,920] (0,500)	6
Ragazzi che frequentano l'istituto artistico	0,843 [0,764; 0,899] (0,839)	31	0,729 [0,533; 0,864] (0,000)	1
Ragazzi che frequentano l'istituto professionale	0,785 [0,719; 0,839] (0,793)	275	0,751 [0,600; 0,858] (1,000)	9
Ragazzi che frequentano l'istituto tecnico	0,857 [0,810; 0,894] (0,855)	773	0,647 [0,474; 0,788] (1,000)	4

Le stime delle probabilità, illustrate in Tabella 53 e calcolate per gli studenti frequentanti la classe terza, sono tutte più alte delle analoghe probabilità calcolate per tutti gli studenti che frequentavano le scuole bolognesi nell'anno scolastico 2004/05. Ciò vuol dire che è in generale più probabile che uno studente sia promosso se la sua carriera pregressa è regolare piuttosto che nel caso opposto. Ciò è sempre più vero procedendo da una classe alla successiva. Le stime delle probabilità sono più vicine a quelle osservate che nei precedenti modelli, stimati per le classi prima e seconda.

Si può notare, in questo modello come in quelli stimati per le altre classi ed anni scolastici, che gli intervalli di confidenza sono tanto meno ampi quanto più numeroso è il contingente di riferimento: è una dimostrazione dell'aumento, in quanto a precisione delle stime, all'aumentare della numerosità (se la numerosità è bassa, aumenta la stima dell'errore standard, che è a sua volta funzione inversa della numerosità campionaria, e di conseguenza gli intervalli di confidenza, i cui estremi sono funzioni dirette dell'errore standard, si ampliano).

Il modello evidenzia che gli studenti stranieri hanno risultati non tanto peggiori rispetto agli italiani, come era sembrato nei modelli precedenti. Tuttavia rimane il problema della scarsa numerosità degli studenti con cittadinanza non italiana, che rende difficile estendere tale risultato ad una popolazione più ampia. Non è da escludersi, più probabilmente, che tale risultato sia vero per quegli stranieri che arrivano regolarmente alla classe terza, che però sicuramente si isolano rispetto alla condizione della maggior parte degli stranieri, che invece si trova in posizione non regolare rispetto al corso di studi.

Una costante di tutti questi modelli è la differenza tra le varie tipologie di scuola, in particolare tra il liceo e gli altri istituti, e tra maschi e femmine (non confermata negli istituti artistici, ma al momento non verificabile data la scarsa numerosità).

Si nota la sovrapposizione degli intervalli di confidenza tra ragazze dei professionali e ragazzi dei tecnici: con la progressione del tempo e in particolare della classe frequentata la differenza tra tali due categorie si affievolisce (gli intervalli di confidenza praticamente coincidono).

Le ragazze dei tecnici iniziano ad eguagliare i loro risultati con i ragazzi dei licei: si verifica una lieve sovrapposizione degli intervalli di confidenza.

Considerando ora il tempo = 3 (anno scolastico 2005/06) e la classe quarta, le stime dei coefficienti risultano essere tutte significative al livello di probabilità del 99%, eccetto quello relativo all'istituto artistico, che è significativo al livello di probabilità del 90%. La cittadinanza risulta non essere più significativa (e ciò è anche ovvio: gli studenti stranieri che giungono regolarmente alla classe quarta, oltre a essere in numero molto esiguo, hanno anche molto probabilmente caratteristiche più simili agli italiani che agli altri stranieri, che in generale non riescono a mantenere la regolarità scolastica). Nemmeno la diversificazione tra scuole collocate in ambito diverso o uguale a quello di residenza è più significativa. Il modello diventa (AIC = 1.616):

$$\log it(p) = 3{,}606 - 0{,}525 Sex - 0{,}787 Art - 1{,}199 Vocational - 0{,}919 Technical + \varepsilon$$

dove l'intercetta ha significato analogo a quello dei modelli precedenti, con la differenza che non esiste differenziazione tra studenti italiani e stranieri. La Tabella 54 mostra gli *odds ratio* calcolati sulla base del modello stimato.

Tabella 54 – Odds ratio per ogni variabile esplicativa – classe IV	
Maschi	0,592
Istituti artistici	0,455
Istituti tecnici	0,399
Istituti professionali	0,301

Gli *odds ratio*, tutti inferiori a 1 come anche nei modelli precedenti, rivelano che vi sono ancora differenze tra le varie tipologie di studenti, tuttavia non vi sono più differenze, in termini di propensione alla promozione, tra studenti con cittadinanza italiana e straniera. Ciò è principalmente dovuto al fatto che molto probabilmente quei pochi stranieri che in quarta sono in posizione regolare sono del tutto simili agli studenti italiani. Inoltre si ha che in classe quarta l'effetto mobilità per lo studente non influisce sui suoi risultati.

Le stime della probabilità di essere promossi per quegli studenti che frequentavano la classe quarta nell'anno scolastico 2005/06 sono visualizzate in Tabella 55, per i vari gruppi distinti da valori diversi delle variabili esplicative (le probabilità stimate sono uguali per diversa cittadinanza e diversa corrispondenza tra gli ambiti, ma quelle osservate sono diverse, per questo si è mantenuta la precedente distinzione in gruppi).

Tabella 55 – Stime delle probabilità di promozione – anno scolastico 2005/06 – classe quarta (valori osservati)

Studenti	$\hat{p}$ studenti Italiani	Numero di studenti	$\hat{p}$ studenti con cittadinanza straniera	Numero di studenti
Ragazze che frequentano il liceo	0,974 [0,970; 0,977] (0,977)	1226	0,974 [0,970; 0,977] (1,000)	6
Ragazze che frequentano l'istituto artistico	0,944 [0,906; 0,967] (0,963)	82	0,944 [0,906; 0,967] NA	0
Ragazze che frequentano l'istituto professionale	0,917 [0,888; 0,940] (0,907)	248	0,917 [0,888; 0,940] (1,000)	13
Ragazze che frequentano l'istituto tecnico	0,936 [0,915; 0,952] (0,926)	404	0,936 [0,915; 0,952] (0,857)	7
Ragazzi che frequentano il liceo	0,956 [0,942; 0,967] (0,950)	826	0,956 [0,942; 0,967] (1,000)	3
Ragazzi che frequentano l'istituto artistico	0,908 [0,831; 0,952] (0,846)	26	0,909 [0,831; 0,952] NA	0
Ragazzi che frequentano l'istituto professionale	0,868 [0,802; 0,914] (0,872)	211	0,868 [0,802; 0,914] (1,000)	4
Ragazzi che frequentano l'istituto tecnico	0,897 [0,846; 0,932] (0,908)	660	0,897 [0,846; 0,932] (0,500)	6

Occorre tener conto del fatto che nelle classi quarte degli istituti artistici non vi erano studenti stranieri. Come suggerisce il modello, anche nei valori osservati non sembra esservi in realtà differenza tra gli esiti degli studenti che frequentano una scuola situata nello stesso ambito di residenza e gli altri. Non è invece possibile confrontare le probabilità di successo degli studenti italiani e stranieri osservate, data la troppo scarsa numerosità degli studenti con cittadinanza straniera ed in particolare considerato che quelli presenti sono sicuramente autoselezionati rispetto ai cittadini stranieri e forse anche maggiormente motivati.

Anche per il modello relativo alla classe quarta, le stime delle probabilità di essere promossi per quegli studenti che frequentano appunto la classe quarta sono maggiori delle probabilità osservate per tutti gli studenti che erano a scuola nell'anno scolastico 2005/06 (compresi quindi quelli che erano in ritardo scolastico). Il modello conferma, inoltre, le differenze positive, in termini di esito scolastico, tra femmine e maschi. Inoltre viene ulteriormente confermato che gli studenti dei licei (in particolare le ragazze) conseguono risultati migliori di tutti gli altri studenti, mentre gli istituti professionali risultano essere quelli dove gli studenti (in particolare i ragazzi) ottengono i risultati peggiori.

Le ragazze dei professionali, tuttavia, in quarta risulta che abbiano risultati migliori dei colleghi maschi dei tecnici: gli intervalli di confidenza sono per larga parte sovrapposti, ma l'estremo superiore per le ragazze è un valore più alto di quello dei ragazzi. Per chi arriva con regolarità in quarta, pare vero che si possano ottenere risultati migliori ai professionali che ai tecnici.

Considerando ora il tempo = 4 (anno scolastico 2006/07) e la classe quinta, le stime dei coefficienti risultano essere tutte significative al livello di probabilità del 99%, o al massimo del 95% (per il coefficiente relativo ai

tecnici). La differenza tra i risultati conseguiti dagli italiani e quelli degli stranieri non risulta significativa. Nemmeno la differenza tra studenti che frequentano scuole dello stesso ambito e di ambito diverso rispetto a quello di residenza risulta significativa, ad eccezione degli studenti che frequentano gli istituti tecnici. Il modello diventa (AIC = 1.039):

$$\log it(p) = 4{,}279 - 0{,}557 Sex - 1{,}186 Art - 0{,}553 Technical - 1{,}631 Vocational$$
$$- 0{,}934 Technical * DifferentAmbit + \varepsilon$$

dove l'intercetta indica, come anche in precedenza, il valore dl logaritmo della probabilità per le ragazze che frequentano il liceo. La Tabella 56 mostra gli *odds ratio* calcolati sulla base del modello stimato.

Tabella 56 – Odds ratio per ogni variabile esplicativa – classe V

Maschi	0,573
Istituti artistici	0,305
Istituti tecnici	0,575
Istituti professionali	0,196
Diverso ambito* Ist. tecnici	0,393

Dagli *odds ratio*, ancora inferiori a 1, emerge che ancora esiste una differenza, in termini di propensione alla promozione, tra le tipologie di studenti: ad esempio, i maschi hanno una propensione alla promozione inferiore rispetto alle femmine. Analogamente agli altri modelli, gli studenti che conseguono i risultati peggiori sono quelli che frequentano gli istituti professionali. Inoltre, mentre in generale, in classe quinta, non vi è differenza tra gli studenti che frequentano una scuola nello stesso ambito di residenza e coloro frequentano una scuola in ambito diverso, in questo caso sembra permanere la lieve differenza, in quanto ad ambito, per i soli istituti tecnici, dove gli studenti conseguono risultati migliori se frequentano una scuola nello stesso ambito di residenza.

Le stime delle probabilità di essere promossi per gli studenti che frequentavano la classe quinta nel 2006/07 sono mostrate in Tabella 57.

Tabella 57 – Stime delle probabilità di promozione – anno scolastico 2006/07 – classe quinta (valori osservati)

Studenti	$\hat{p}$ studenti Italiani	Numero di studenti	$\hat{p}$ studenti con cittadinanza straniera	Numero di studenti
Ragazze che frequentano il liceo	0,986 [0,983; 0,989] (0,990)	1187	0,986 [0,983; 0,989] (0,833)	6
Ragazze che frequentano l'istituto artistico	0,957 [0,917; 0,978] (0,949)	78	0,957 [0,917; 0,978] NA	0
Ragazze che frequentano l'istituto professionale	0,934 [0,901; 0,956] (0,916)	226	0,934 [0,901; 0,956] (1,000)	13
Ragazze che frequentano l'istituto tecnico nello stesso ambito di residenza	0,976 [0,962; 0,985] (0,975)	244	0,976 [0,962; 0,985] (1,000)	4
Ragazze che frequentano l'istituto tecnico in ambito diverso rispetto a quello di residenza	0,942 [0,883; 0,974] (0,947)	131	0,942 [0,883; 0,974] (1,000)	2

Tabella 57 – Stime delle probabilità di promozione – anno scolastico 2006/07 – classe quinta (valori osservati)

Studenti	$\hat{p}$ studenti Italiani	Numero di studenti	$\hat{p}$ studenti con cittadinanza straniera	Numero di studenti
Ragazzi che frequentano il liceo	0,976 [0,966; 0,984] (0,972)	789	0,976 [0,966; 0,984] (1,000)	3
Ragazzi che frequentano l'istituto artistico	0,927 [0,839; 0,968] (0,954)	22	0,927 [0,839; 0,968] NA	0
Ragazzi che frequentano l'istituto professionale	0,890 [0,811; 0,938] (0,910)	189	0,890 [0,811; 0,938] (0,750)	4
Ragazzi che frequentano l'istituto tecnico nello stesso ambito di residenza	0,960 [0,924; 0,979] (0,960)	374	0,960 [0,924; 0,979] (1,000)	2
Ragazzi che frequentano l'istituto tecnico in ambito diverso rispetto a quello di residenza	0,903 [0,781; 0,961] (0,900)	230	0,903 [0,781; 0,961] (1,000)	1

Come nei casi precedenti, le probabilità di successo per gli studenti che frequentano la classe quinta sono più alte delle analoghe probabilità calcolate per tutti gli studenti a scuola nel 2006/07: è infatti più probabile essere promossi in classe quinta che non nelle altre classi, specialmente se si è nella condizione di studente regolare.

Il modello riproduce adeguatamente la, seppur lieve, differenza tra gli studenti che frequentano gli istituti tecnici nello stesso ambito di residenza e gli studenti che frequentano gli istituti tecnici in un ambito diverso; ben riproduce anche le differenze, decisamente inferiori rispetto ai modelli precedenti, tra gli esiti di quegli studenti che frequentano le diverse tipologie di scuola e di maschi e femmine separatamente. Non si possono ricavare informazioni riguardo alle differenze tra italiani e stranieri, per l'esiguo numero di studenti con cittadinanza non italiana che si trova in classe quinta.

Per quanto riguarda la variabile ambito che rappresenta la mobilità, gli intervalli di confidenza stimati hanno in realtà punti in comune, quindi è ancora incerto se si possa attribuire una differenza significativa. Vi è inoltre da notare la maggiore ampiezza dell'intervallo calcolato in caso di ambito diverso rispetto al caso di stesso ambito. La minore precisione non pare dovuta a scarsa numerosità, quanto forse ad altri fattori non noti che in realtà differenziano gli studenti del gruppo.

Confrontando gli intervalli di confidenza relativi alle ragazze dei tecnici ed ai ragazzi dei licei, si nota una sovrapposizione nel caso di stesso ambito (per i tecnici): in quinta classe non risulta significativa la differenza tra tali due categorie di studenti, già messa in dubbio nei due anni precedenti.

E' pur vero che, in generale, in quinta, le differenze si affievoliscono: chi è arrivato qui con un percorso regolare è assai difficile che "inciampi sull'ultimo gradino". Persino gli intervalli di confidenza stimati per i ragazzi dei professionali e le ragazze dei licei hanno punti in comune!!!

Si può notare, infatti, che in quinta classe si ha una sovrapposizione degli intervalli di confidenza non riscontrata nei modelli precedenti.

La Tabella 58 mostra le probabilità di essere promossi, stimate ed osservate, per alcuni individui tipici nei diversi anni scolastici e classi considerati.

Tabella 58 – Stime delle probabilità di promozione per gli studenti italiani (valori osservati)

Studenti	2003/04 classe I	2004/05 classe II	2005/06 classe III	2006/07 classe IV	2007/08 classe V
Ragazze che frequentano il liceo	0,951 (0,950)	0,957 (0,949)	0,952 (0,953)	0,974 (0,977)	0,986 (0,990)
Ragazze che frequentano l'istituto artistico	0,813 (0,789)	0,907 (0,902)	0,892 (0,901)	0,944 (0,963)	0,957 (0,949)
Ragazze che frequentano l'istituto professionale	0,752 (0,755)	0,831 (0,840)	0,848 (0,831)	0,917 (0,907)	0,934 (0,916)
Ragazze che frequentano l'istituto tecnico	0,883 (0,888)	0,909 (0,924)	0,902 (0,906)	0,936 (0,926)	0,976 (0,965)
Ragazzi che frequentano il liceo	0,927 (0,928)	0,940 (0,954)	0,929 (0,930)	0,956 (0,950)	0,976 (0,972)
Ragazzi che frequentano l'istituto artistico	0,742 (0,814)	0,871 (0,886)	0,843 (0,839)	0,908 (0,846)	0,927 (0,955)
Ragazzi che frequentano l'istituto professionale	0,666 (0,662)	0,773 (0,757)	0,785 (0,793)	0,868 (0,872)	0,890 (0,910)
Ragazzi che frequentano l'istituto tecnico	0,832 (0,830)	0,873 (0,865)	0,857 (0,855)	0,897 (0,908)	0,903 (0,937)

Si può notare che i modelli rappresentano con un buon livello di approssimazione i dati osservati quando si ha un numero sufficiente di osservazioni, per il singolo gruppo per cui si stima la probabilità; per esempio, occorre sempre usare cautela nell'interpretazione delle stime relative agli studenti stranieri, proprio dato il loro numero il più delle volte esiguo a garantire la precisione delle stime. La differenza tra studenti che frequentano una scuola nello stesso ambito di residenza e gli altri è invece dimostrata dal modello non essere significativa nella spiegazione della probabilità di successo. Calcolando le probabilità medie di essere promossi, in ciascun anno e classe e distintamente per maschi e femmine, si può notare che il modello riesce a riprodurre adeguatamente le osservazioni. In Tabella 59 vengono riportate le probabilità medie stimate ed osservate, inoltre la differenza in termini percentuali tra maschi e femmine, che si attesta attorno al 4%. La differenza riscontrata in quinta risulta in realtà di dubbia significatività.

Tabella 59 – Stime delle probabilità di promozione medie per gli studenti italiani (osservazioni empiriche)

	2003/04 classe I	2004/05 classe II	2005/06 classe III	2006/07 classe IV	2007/08 classe V
Femmine	0,889 (0,889)	0,925 (0,924)	0,924 (0,923)	0,958 (0,957)	0,976 (0,974)
Maschi	0,798 (0,798)	0,885 (0,885)	0,879 (0,880)	0,922 (0,923)	0,938 (0,951)
Differenza dei maschi rispetto alle femmine	10% (10%)	4% (4%)	5% (5%)	4% (4%)	4% (4%)

Il modello ben riproduce le probabilità osservate anche per quanto riguarda la distinzione per tipologia di scuola frequentata. In particolare, si ha che le differenze percentuali nei vari anni in termini di probabilità di promozione tra tecnici e licei sono stimate essere del [9%; 7%; 7%; 6%; 5%] (le analoghe percentuali osservate sono [9%; 7%; 7%; 5%; 4%]) mentre la differenza percentuale tra professionali e licei è stimata essere [26%; 15%; 13%; 7%; 7%] (le analoghe percentuali osservate sono [26%; 16%; 14%; 8%; 7%]).

I modelli non sempre rappresentano il trend temporale della probabilità di essere promossi, perché sono calcolati separatamente per ogni anno scolastico. È invece possibile, ottenendo certo migliori risultati, introdurre nel modello anche la variabile Tempo, in modo da tener conto dell'andamento temporale.

Rimane sempre vero, in tutti gli istanti temporali (tranne che in classe quinta, dove le differenze si assottigliano notevolmente), che la probabilità significativamente più alta di essere promossi è quella degli studenti che frequentano il liceo, mentre quella significativamente più bassa è quella degli studenti, in particolare maschi, che frequentano gli istituti professionali (per entrambi, maschi e femmine e per studenti con cittadinanza italiana e straniera, inoltre per studenti frequentanti una scuola dello stesso ambito di residenza e gli altri). Questa è una conferma quantitativa delle valutazioni fatte sulla sola base dei dati descrittivi sulla regolarità scolastica.

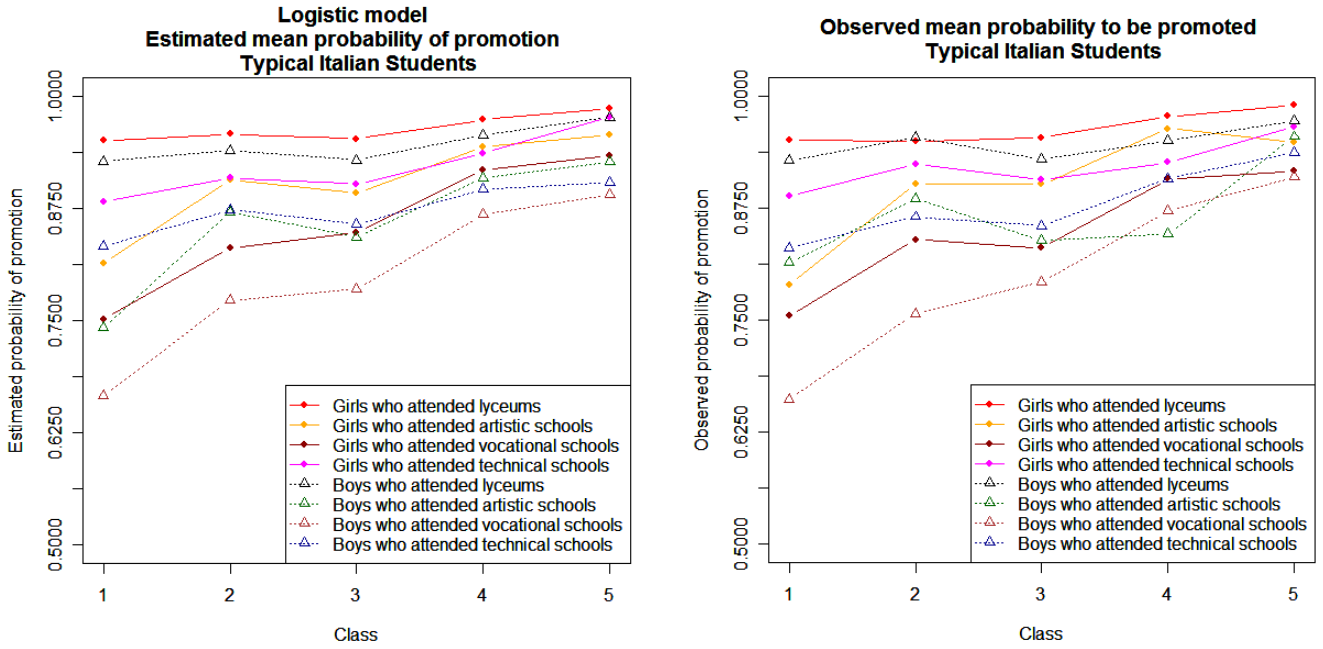


Grafico 29 – Probabilità di promozione degli studenti con cittadinanza italiana: medie stimate dal modello logistico e medie osservate

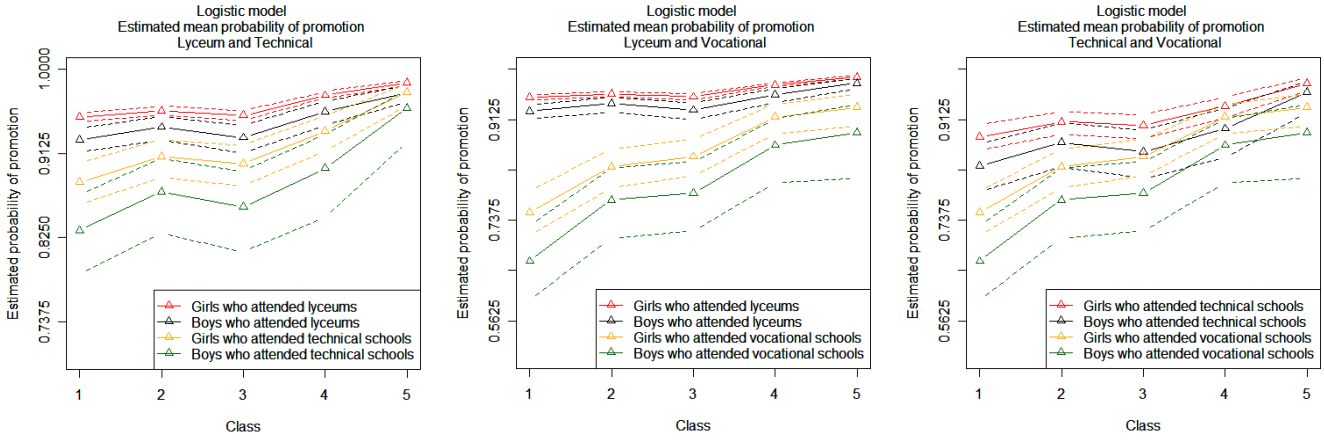


Grafico 30 – Probabilità di promozione degli studenti con cittadinanza italiana: confronto delle stime, con intervalli di confidenza, tra tipologie di scuola diverse

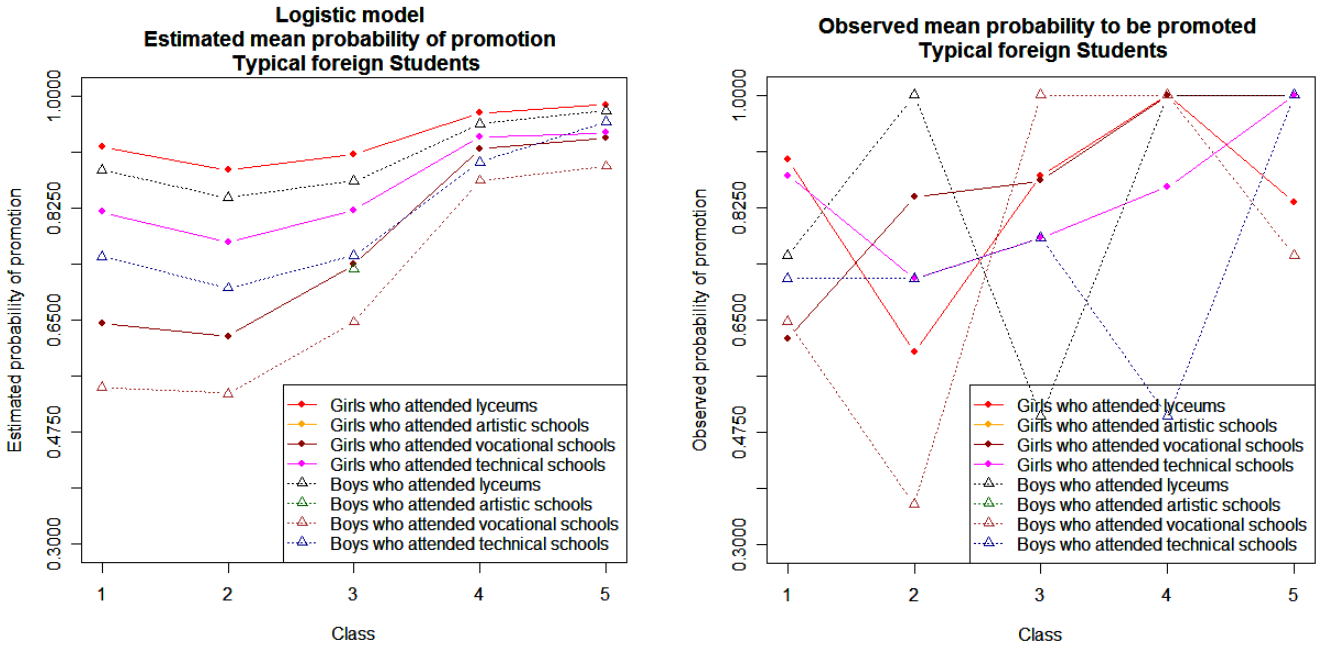


Grafico 31 – Probabilità di promozione degli studenti con cittadinanza non italiana: medie stimate dal modello logistico e medie osservate

Si possono notare le diverse ampiezze degli intervalli di confidenza (Grafico 30) stimati per gli studenti delle diverse tipologie di scuola: le stime relative ai licei sono maggiormente precise in quanto hanno errori standard decisamente inferiori rispetto alle altre. Sempre osservando gli intervalli di confidenza stimati, si ha che in realtà la differenza significativa è tra gli studenti dei licei (che conseguono risultati migliori) e tutti gli altri. Gli studenti dei tecnici e dei professionali conseguono in realtà risultati non sempre significativamente diversi, in particolare la differenza risulta significativa soltanto nelle prime classi frequentate.

Tabella 60 – *Odds ratio* dei diversi modelli per ciascun anno scolastico considerato

	Classe I	Classe II	Classe III	Classe IV	Classe V
Genere					
F (rif.)	1	1	1	1	1
M	0,659	0,692	0,653	0,592	0,573
Tipologia di scuola					
Liceo (rif.)	1	1	1	1	1
Istituto artistico	0,226	0,433	0,413	0,455	0,305
Istituto tecnico	0,390	0,441	0,462	0,399	0,575
Istituto professionale	0,157	0,218	0,281	0,301	0,196
Cittadinanza					
Italiani (rif.)	1	1	1	1	1
Stranieri	0,601	0,339	0,501	n.s.	n.s.
Ambito					
Stesso ambito (rif.)	1	1	1	1	1
Diverso ambito	n.s.	n.s.	n.s.	n.s.	n.s.
Interazioni					
Ist. Tec.* Diverso Ambito	-	-	-	-	0,393

In Tabella 60 si può notare che la differenza negativa nella probabilità di essere promossi tra ragazzi e ragazze aumenta lievemente nelle classi quarta e quinta. Inoltre, la differenza negativa tra gli studenti con cittadinanza non italiana e gli italiani è stimata significativa nelle prime tre classi, mentre non lo risulta più nelle ultime due; ciò è dato dal fatto, già più volte ripetuto, che sono veramente pochi gli studenti stranieri che arrivano in quarta o quinta classe con percorso regolare ed quei pochi sono in realtà simili agli italiani. Per gli studenti dell'istituto tecnico, la differenza nella probabilità di promozione rispetto ai colleghi del liceo è piuttosto alta e costante nel tempo; in quinta, la differenza diminuisce notevolmente. Per gli studenti dell'istituto professionale, la differenza rispetto a quelli del liceo è sempre la più alta, con una diminuzione progressiva nelle prime quattro classi, pur rimanendo sempre al di sotto della differenza presentata dai ragazzi degli istituti tecnici; in quinta classe, a differenza dei colleghi degli istituti tecnici, gli studenti degli istituti professionali presentano un peggioramento della differenza rispetto a chi frequenta il liceo.

3 Il modello a curva latente

3.1 LCM - ANALISI DESCRITTIVA

Prima di delineare la struttura del modello, è utile condurre un'analisi esplorativa del dataset. A tal fine, si può avere una visione generale del comportamento della variabile obiettivo nel dataset disponibile, attraverso un'analisi grafica di un gruppo casuale di individui condotta con i grafici di crescita (*growth plots*): essi mostrano l'andamento della variabile oggetto d'interesse nei periodi considerati per alcuni individui scelti in modo casuale (è opportuno utilizzare dimensioni degli assi identiche nei diversi grafici, per un adeguato confronto).

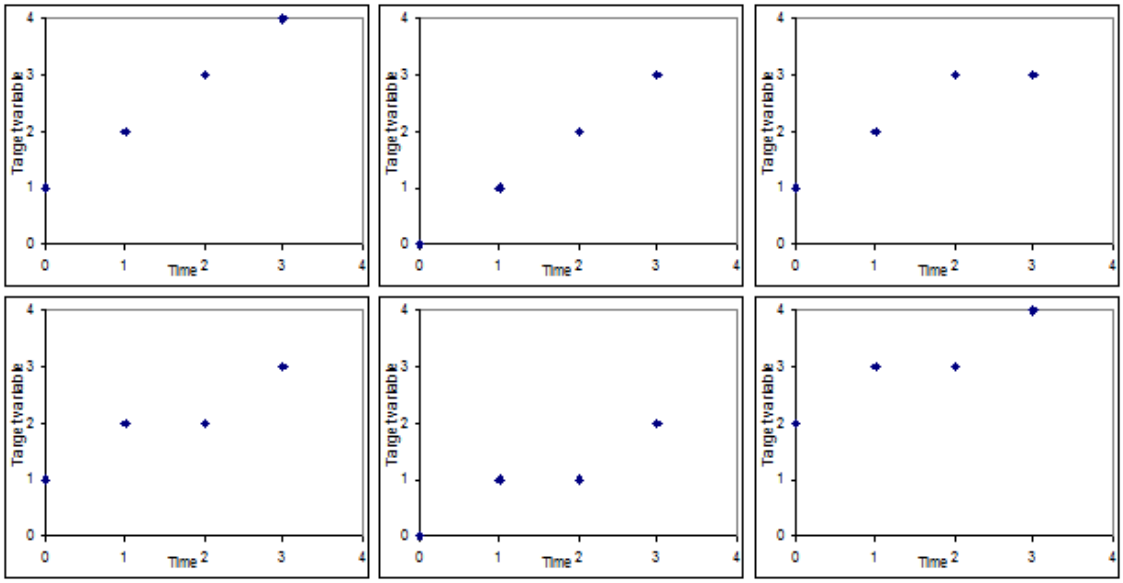


Grafico 32 – Esempio di *growth plots*

È possibile modellare le traiettorie di crescita empiriche con un metodo non parametrico, solo per fini esplorativi (si possono vedere la posizione, la forma ed eventuali punti di svolta). Si possono anche inserire in un singolo grafico tutte le traiettorie modellate, o anche soltanto alcune di esse, per confrontarle tra di loro. In questo modo si può osservare un insieme di traiettorie, al fine di individuare i comportamenti estremi (*outliers*) ed anche di vedere se le traiettorie sono tra loro vicine oppure si discostano molto le une dalle altre. Come vedremo meglio in seguito, vi è la possibilità di stimare un modello lineare per ciascun individuo e di vedere graficamente il comportamento di ognuno in termini di intercetta (valore iniziale) e di pendenza (tasso di variazione). Anche le rette stimate possono essere inserite in un grafico (tutte o un campione casuale di esse), permettendo considerazioni analoghe a quelle sopra descritte.

3.2 LCM - IL MODELLO

Nel caso univariato, data la variabile obiettivo Y, per la quale sono disponibili misure su N individui (i = 1,….N), ripetute in T istanti temporali (t = 1,….T), è possibile stimare un LCM (*Latent Curve Model*). Il modello, in generale, è il seguente[34]:

[34] Bollen "On the origins of latent curve model"

$$Y_{it} = g_1(t)\eta_{i1} + g_2(t)\eta_{i2} + ... + g_K(t)\eta_{iK} + \varepsilon_{it}$$

dove $g_k(t)$ è una funzione del tempo relativa al k-esimo fattore, *k*=1,..,K; η_{ik} è il coefficiente casuale, o fattore o variabile latente, per l'individuo *i*-esimo e il *k*-esimo fattore, mentre ε_{it} è la componente di errore dell'individuo *i*-esimo al tempo *t* (con la consueta assunzione $E(\varepsilon_{it}) = 0$, per tutte le osservazioni e per tutti i tempi, inoltre, $\text{cov}(\varepsilon_{it+s}, \eta_{ik}) = 0$ per tutti gli *i, s* e *k*, cioè mancanza di correlazione tra i termini di errore e i fattori, per ogni individuo ed ogni tempo, infine $E(\varepsilon_{it}, \varepsilon_{i+j,t+s}) = 0$ per tutti i *i ,t, j, s* e per *j, s* >0, cioè termini di errore incorrelati per tempi e individui differenti). La varianza del termine di errore è $E(\varepsilon_{it}^2) = \text{var}(\varepsilon_{it}) = \vartheta_{\varepsilon_{it}}$. Un'assunzione che frequentemente viene fatta è che questa varianza sia uguale tra gli individui: $\text{var}(\varepsilon_{it}) = \vartheta_{\varepsilon_t}$. I fattori η_{ik} possono essere tra loro correlati (*oblique factors*) oppure incorrelati (*orthogonal factors*), a seconda del modello impiegato. Ciascuno di questi fattori è il peso individuale su ogni curva $g_k(t)$. La grandezza del fattore indica la traiettoria dominante di ogni individuo: nella maggior parte dei casi soltanto un fattore è dominante (al massimo due) e gli altri sono prossimi a 0.

La tipologia di LCM che al momento viene dai più utilizzata parte dall'ipotesi lineare (la dipendenza di Y dal tempo è rappresentata come legame lineare); vengono usati due termini di curva, quindi K=2, (il primo termine è una costante pari a 1, quindi $g_1(t) = 1$, con $\eta_{i1} = \alpha_i$, e il secondo termine è lineare, $g_2(t) = \lambda_t$, con una pendenza lineare casuale $\eta_{i2} = \beta_i$), attraverso il modello individuale[35]:

$$Y_{it} = \alpha_i + \beta_i\lambda_t + \varepsilon_{it}.$$

In esso, λ_t mostra il valore della variabile tempo, mentre α_i e β_i sono le realizzazioni, per l'individuo *i*-esimo delle variabili latenti intercetta $\eta_1 = \alpha$ e pendenza $\eta_2 = \beta$; ε_{it} riassume il contenuto residuo, non spiegato dal modello, per l'individuo *i*-esimo al tempo *t*.

Il termine di errore comprende anche: l'errore di misurazione nelle misure ripetute della variabile obiettivo; gli errori di approssimazione utilizzando la forma funzionale; gli errori dovuti alla possibile omissione di variabili esplicative; gli errori dovuti a una componente stocastica. Come di consueto, accade che la traiettoria reale abbia una forma più complessa di quella stimata.

Il modello di cui abbiamo parlato è chiamato "modello di primo livello" (analogo al modello gerarchico, unidimensionale che vedremo in seguito): esso contiene alcuni elementi, detti coefficienti, che hanno bisogno di una ulteriore specificazione. Siccome questi coefficienti sono a loro volta variabili aleatorie, ognuno di essi può essere scomposto in una parte fissa (il valore atteso) ed in una parte variabile (la differenza individuale dal valore atteso), al fine di ottenere il modello di secondo livello , che diventa (seguendo la notazione del modello gerarchico):

[35] Curran "Comparing Three Modern Approaches to Longitudinal Data Analysis: An Examination of a Single Developmental Sample"

$$\alpha_i = \mu_\alpha + \zeta_{\alpha_i}$$

$$\beta_i = \mu_\beta + \zeta_{\beta_i}$$

Quello descritto è il cosiddetto modello di crescita non condizionato, siccome non presuppone che alcuna variabile latente dipenda da una o più variabili esplicative note.

Un primo modello più semplice è quello chiamato delle medie non condizionate[36], senza variabili esplicative, quindi con l'esclusione anche della variabile tempo:

$$Y_{it} = \alpha_i + \varepsilon_{it}$$

$$\alpha_i = \mu_\alpha + \zeta_{\alpha_i}$$

In questo caso, la traiettoria di cambiamento individuale è completamente piatta; l'intercetta è α_i (la vera media temporale di Y per l'individuo *i*), e l'intercetta media è μ_α (la media vera di Y tra gli individui in tutti gli intervalli temporali). I termini di errore vengono interpretati come scostamenti per ogni individuo nei diversi istanti temporali (ε_{it}) e come scostamenti tra tutti gli individui mediamente in tutti i tempi (ζ_{α_i}).

In generale, nel modello di primo livello, l'attenzione si posa sulla caratterizzazione del modello di cambiamento individuale, cioè sulla traiettoria di crescita individuale; l'obiettivo di questa analisi è quello di descrivere la forma della traiettoria individuale[37]. Nel modello di secondo livello, l'attenzione si posa sulle differenze interindividuali nel cambiamento; l'obiettivo di tale analisi è quello di determinare le relazioni tra i predittori e le traiettorie individuali.

Unendo le equazioni del modello di crescita non condizionato, si ottiene (notazione *composite* o a struttura mista o compatta):

$$Y_{it} = \left(\mu_\alpha + \mu_\beta \lambda_t\right) + \left(\zeta_{\alpha_i} + \zeta_{\beta_i} \lambda_t + \varepsilon_{it}\right).$$

In tale espressione si può notare che la variabile obiettivo dipende da una componente fissa per tutti gli individui (l'intercetta media, la pendenza media e il tempo) e da una componente complessa di disturbo, diversa tra gli individui e nei vari istanti temporali, la cui varianza dipende da λ_t e quindi varia al trascorrere del tempo[38].

Utilizzando questa scomposizione, quando si valuta il modello, si controlla sia la componente fissa (l'intercetta e la pendenza media) che quella aleatoria (le variazioni individuali attorno alla media). È anche possibile valutare lo scostamento della singola variabile dal livello medio della variabile obiettivo, come anche le differenze nel tasso di cambiamento[39].

Le assunzioni sono le stesse del modello lineare: media degli errori nulla ($E(\varepsilon_{it}) = 0 \ \forall i,t \quad E(\zeta_{\alpha_i}) = 0$ $E(\zeta_{\beta_i}) = 0 \quad \forall i$) ed errori tra loro incorrelati (ε_{it} incorrelati, ζ_{α_i} incorrelati, ζ_{β_i} incorrelati). Inoltre

[36] Singer, Willett "Applied longitudinal data analysis – modelling change and event occurrence"
[37] Singer, Willett "Applied longitudinal data analysis – modelling change and event occurrence"
[38] Curran "Latent Curve Models - A structural equation perspective"
[39] Witta "Latent growth model of cognition in the elderly"

$COV\left(\varepsilon_{it},\zeta_{\alpha_i}\right)=0$, $COV\left(\varepsilon_{it},\zeta_{\beta_i}\right)=0$, $COV\left(\zeta_{\alpha_i},\zeta_{\beta_j}\right)=0$ per $i \neq j$ (si noti che la covarianza tra gli errori dei due fattori latenti per individui diversi è nulla, per la legge dei grandi numeri in presenza di un dataset comprendente molti individui, mentre la covarianza tra gli errori relativi allo stesso individuo deve essere valutata dal modello!).

Prima di stimare i parametri del modello, si ha bisogno di valutarne l'identificazione[40]: vi sono valori unici dei parametri del modello, determinati dalla struttura del modello stesso? In altri termini, è necessario considerare se i parametri noti siano in numero sufficiente per stimare i parametri non noti. Un modello identificato è un modello in cui il valore di uno specifico parametro identifica in modo univoco il modello, e non può esservi un'altra equivalente formulazione fornita dal valore di un altro parametro. Nel modello *unconditional*, i parametri noti sono i momenti del primo e del second'ordine (qui le considerazioni vengono ristrette a questi due momenti) della variabile Y ($E(y_{it})$, $\text{var}(y_{it})$, $\text{cov}(y_{it}, y_{i,t-s})$ s>0), mentre i parametri incogniti sono le medie delle variabili latenti e le varianze e covarianze dei loro errori (μ_α, μ_β, $\text{var}(\zeta_{\alpha_i})=\psi_{\alpha\alpha}$, $\text{var}(\zeta_{\beta_i})=\psi_{\beta\beta}$, $\text{cov}(\zeta_{\alpha_i},\zeta_{\beta_i})=\psi_{\alpha\beta}$), la varianza degli errori (var (ε_{it})) e le determinazioni del tempo, λ_t.

Un'identificazione esatta del modello richiede tanti parametri noti quanti incogniti. Si noti che il numero delle medie, delle varianze e delle covarianze della variabile obiettivo dipende dalle unità temporali considerate (in generale, sono $\frac{1}{2}T(T+3)$), mentre il numero dei parametri incogniti è NT (le varianze degli errori) + T (i valori λ_t) + $\frac{1}{2}K(K+3)$ (i parametri legati alle variabili latenti, dove K è il numero di tali variabili, che nel caso lineare è 2).

Al fine della stima, occorre in qualche modo restringere il numero dei parametri, facendo ipotesi a priori. Generalmente, è meglio presumere di avere informazioni sul tempo, quindi si può porre $\lambda_t = t-1$, e quindi non si ha più bisogno di stimare questi parametri, ma si assume che siano già noti (così, nel modello lineare, si hanno NT+5 parametri da stimare, contro $\frac{1}{2}T(T+3)$ parametri già noti). Per l'identificazione del modello, si ha bisogno di ulteriori assunzioni, la più comune delle quali, in presenza di un numero sufficiente di individui, è di porre la varianza degli errori uniforme per tutti gli individui, ma diversa nei differenti istanti temporali: var (ε_{it}) = var (ε_t). I parametri da stimare diventano così T+5. Se si considerano 2 istanti di tempo (avendo così 7 parametri da stimare e 5 già noti) il modello è sottoidentificato; mentre se si considerano 3 istanti di tempo (avendo così 8 parametri da stimare e 9 già noti) il modello è identificato: per ottenere una valutazione del modello, quindi, si ha bisogno di raccogliere informazioni su almeno 3 unità di tempo distinte.

Il modello a curva latente può essere visto come separato su due livelli.

Al primo livello, il modello descrive come gli individui cambiano nel tempo (modello di crescita individuale), mentre al secondo livello descrive come sono diversi i cambiamenti tra gli individui. Per formulare una prima ipotesi sulla tipologia di modello di primo livello che ha probabilmente generato il campione osservato, è utile guardare i grafici di crescita empirici. Comunque, specialmente quando non sono disponibili molte onde

[40] Curran "Latent Curve Models - A structural equation perspective"

temporali, l'ipotesi lineare è quella maggiormente utilizzata. Data una variabile endogena (Y) e una funzione del tempo λ_t, il modello di primo livello, per *i*=1,…,N e *t*=1,…,T è:

$Y_{it} = \alpha_i + \beta_i \lambda_t + \varepsilon_{it}$ Questa equazione assume che una linea retta rappresenti in modo adeguato il vero cambiamento nel tempo di ogni individuo, mentre che gli scostamenti osservati dal modello siano dovuti ad errori di misurazione casuali. Si noti che lo stesso modello viene utilizzato per dati *time structured* (quando tutti gli individui hanno la stessa distribuzione delle onde temporali) e per dataset in cui la distribuzione dei tempi e della loro spaziatura è diversa per i differenti individui.

La parte stocastica del modello, ε_{it}, rappresenta la parte dell'*i*-esimo valore della variabile endogena non spiegata dalla variabile esogena. Si può anche introdurre un'altra variabile esogena con l'obiettivo di ridurre l'errore: è possibile che ε_{it} del modello senza covariate sia non solo dovuto ad errore di misurazione. Per adattare il modello ai dati osservati, è necessario fare alcune assunzioni su questi errori. Con la regressione OLS, l'assunzione è $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$, dove σ_ε^2 rappresenta la variabilità degli scostamenti individuali dalla traiettoria vera. Il problema di questa assunzione è quello di non essere realistica: infatti, visto che la misurazione riguarda la stessa persona in diverse occasioni temporali, questo probabilmente porta ad autocorrelazione degli errori; inoltre, la precisione della misurazione può variare nel tempo, come anche le varianze possono essere diverse nel tempo. Non ci sono soluzioni certe per questo problema. La notazione compatta del modello, ottenuta sostituendo i parametri di primo livello con le loro espressioni di secondo livello, può tuttavia aiutare a risolvere la situazione. Il modello di secondo livello si concentra sulle differenze interindividuali nelle traiettorie di cambiamento; come tutti i modelli statistici, descrive il processo ipotizzato nella popolazione, della quale il campione è una realizzazione empirica. I risultati del modello di secondo livello sono i parametri del modello di primo livello, in particolare vi è una espressione specifica per ognuno dei parametri di primo livello e questa espressione mostra la relazione tra il parametro (il valore vero del parametro) e una variabile esogena (Z), più una parte aleatoria che specifica la variazione individuale. Nel caso lineare, le equazioni (una per l'intercetta ed una per la pendenza) sono:

$\alpha_i = \gamma_{00} + \gamma_{01} Z_i + \zeta_{0i}$ e $\beta_i = \gamma_{10} + \gamma_{11} Z_i + \zeta_{1i}$.

Ogni componente ed ogni individuo ha il proprio termine di errore, così un individuo può differire in modo stocastico dagli altri. γ_{00}, γ_{01}, γ_{10}, e γ_{11} sono chiamati effetti fissi e rappresentano le differenze nella traiettoria di cambiamento basata sui valori del predittore (dati alcuni valori di Z è possibile determinare l'intercetta e la pendenza attese sulla base dei coefficienti stimati dal modello di secondo livello). La parte aleatoria del modello, ζ_{0i} e ζ_{1i}, indica quella porzione dei valori dei parametri non spiegati dal modello. L'interesse sta nelle varianze degli errori (σ_0^2, σ_1^2 e σ_{01}) e non nei loro valori specifici stessi. Queste varianze sono chiamate condizionate, perché rappresentano quella parte tralasciata dei parametri del modello, dopo aver considerato gli effetti dei regressori del modello stesso (il termine inglese indica appunto condizionatamente alla presenza dei regressori). La covarianza rappresenta l'associazione tra lo status iniziale e il tasso di variazione. Per adattare il modello ai dati osservati, è necessario fare alcune assunzioni. Ritornando al modello più semplice, quello delle medie non condizionate, le assunzioni riguardano la distribuzione degli errori: $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$ e $\zeta_{0i} \sim N(0, \sigma_0^2)$.

Qui σ_ε^2 è la varianza entro gli individui, mentre σ_0^2 è la varianza tra gli individui. È importante testare la significatività di queste varianze al fine di continuare ad approfondire il modello: se una componente della varianza è 0, non c'è bisogno di spiegare tale varianza perché è troppo piccola. Una volta che si è condotto il test d'ipotesi sul singolo parametro (la statistica test è la consueta, cioè la stima del parametro divisa per il suo errore standard asintotico), con il rifiuto dell'ipotesi nulla di valore 0, è possibile misurare la grandezza relativa delle varianze: è così utile il coefficiente di correlazione in popolazione, *intraclass*, $\rho = \frac{\sigma_0^2}{\sigma_\varepsilon^2 + \sigma_0^2}$. Questo coefficiente di popolazione può essere stimato sostituendo le varianze in popolazione con quelle stimate. Esso descrive la parte della varianza totale dovuta alla variabilità tra gli individui; è anche il coefficiente di correlazione degli errori, perché σ_0^2 è il coefficiente degli errori di secondo livello, che non dipende dal tempo.

Per quanto riguarda il modello non condizionato, dove $Y_{it} = (\mu_\alpha + \mu_\beta \lambda_t) + (\zeta_{\alpha_i} + \zeta_{\beta_i} \lambda_t + \varepsilon_{it})$ è la forma a struttura mista, le assunzioni diventano: $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$ e $\begin{pmatrix} \zeta_{\alpha_i} \\ \zeta_{\beta_i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{pmatrix} \right)$. Qui la covarianza σ_{10} quantifica la relazione tra lo status iniziale in popolazione e i cambiamenti effettivi. Ogni individuo ha T (t=1,…,T) valori degli errori della notazione compatta, $(\zeta_{\alpha_i} + \zeta_{\beta_i} \lambda_t + \varepsilon_{it})$, componenti che possono essere eteroschedastiche ed autocorrelate. È possibile ora ottenere la varianza in popolazione dei termini di errore della forma compatta al tempo t: $\sigma^2_{residual_t} = \sigma_0^2 + \sigma_1^2 \lambda_t^2 + 2\sigma_{01}\lambda_t + \sigma_\varepsilon^2$ e testare l'omoschedasticità.

Inoltre, si può ottenere il coefficiente di correlazione tra i diversi istanti temporali:

$$\rho_{residual_t residual_s} = \frac{\sigma_0^2 + \sigma_{01}(\lambda_t + \lambda_s) + \sigma_1^2 \lambda_t \lambda_s}{\sqrt{\sigma^2_{residual_t} \sigma^2_{residual_s}}}$$ e testare l'autocorrelazione.

La varianza degli errori di primo livello e la matrice di varianze e covarianze di primo livello, insieme considerate, sono note come componenti di varianza.

Nel caso di modello condizionato, i modelli iniziali possono confluire in una singola specificazione:

$Y_{it} = \gamma_{00} + \gamma_{01} Z_i + \zeta_{0i} + (\gamma_{10} + \gamma_{11} Z_i + \zeta_{1i}) X_{it} + \varepsilon_{it}$ che diventa

$Y_{it} = (\gamma_{00} + \gamma_{01} Z_i + \gamma_{10} X_{it} + \gamma_{11} Z_i X_{it}) + (\zeta_{1i} X_{it} + \zeta_{0i} + \varepsilon_{it})$ dove la componente nella prima parentesi è quella strutturale, mentre la componente nella seconda parentesi è quella stocastica. La prima contiene i predittori e gli effetti fissi, mentre l'altra contiene gli effetti casuali (le variazioni degli individui dal valore atteso e le variazioni di ogni individuo nei diversi istanti temporali; in particolare, l'errore della forma compatta rappresenta la differenza tra il valore osservato ed il valore atteso dell'individuo i al tempo t). Questo termine di errore può essere autocorrelato ed eteroschedastico entro gli individui. È invece probabile che la parte non spiegata del risultato di ogni individuo abbia una diversa varianza, considerando i diversi istanti temporali (ζ_{1i} è moltiplicato per X_{it}, che è il valore del tempo). La causa più comune di eteroschedasticità è l'omissione

di qualche predittore. La presenza di autocorrelazione significa che parti non spiegate del risultato di ogni individuo sono correlate con tutte le altri parti, considerando i diversi tempi. Ciò può essere causato anche dall'omissione di qualche predittore.

Usare un modello non condizionato (*unconditional*) aiuta a valutare se vi sia una variabilità del risultato che sia prevedibile. In particolare, il modello a medie non condizionate (*unconditional means*) misura la variabilità della distribuzione, ma non la colloca negli individui o nel tempo, mentre il modello non condizionato di crescita (*unconditional growth*) permette la distinzione della variabilità tra gli individui e gli istanti temporali, senza però spiegarne le cause (senza quindi variabili esogene).

Per quanto riguarda la valutazione della bontà del modello, nell'analisi di regressione classica, il valore di R^2 misura quanta parte di varianza è spiegata dal modello stesso. Nei dati longitudinali, la varianza totale è scomposta in diverse componenti: σ_ε^2, σ_0^2 e σ_1^2. Quindi si ha la necessità di calcolare uno pseudo R^2 che indichi quanta parte di varianza sia spiegata dal modello. Come primo passo, è necessario calcolare un risultato predetto dal modello (per ogni individuo e per ciascun tempo), poi occorre calcolare il quadrato della correlazione tra i valori osservati e quelli stimati. La statistica ottenuta stima la porzione della variabilità totale spiegata dal modello.

Con la misurazione della variabilità residua, cioè la porzione di varianza non spiegata dal modello, è possibile valutare l'eventuale necessità di aggiungere predittori al modello. Aggiungere variabili esplicative, infatti, porta, nel caso siano significative, ad una diminuzione della variabilità e questo decremento rappresenta il miglioramento nel *fitting*. Lo pseudo R^2 viene costruito in termini di confronto di modelli diversi: per esempio, $pseudoR^2_\varepsilon = \frac{\hat{\sigma}^2_\varepsilon(umm) - \hat{\sigma}^2_\varepsilon(ugm)}{\hat{\sigma}^2_\varepsilon(umm)}$, dove *umm* è il modello a medie non condizionate, mentre *ugm* è il modello non condizionato di crescita, opera un confronto tra i due modelli. In particolare, tale indice misura la proporzione della variabilità entro gli individui spiegata dal predittore "tempo". Dopo aver stimato un modello aggiungendo un predittore, è possibile testare il miglioramento: $pseudoR^2_\varepsilon = \frac{\hat{\sigma}^2_\varepsilon(ugm) - \hat{\sigma}^2_\varepsilon(cm)}{\hat{\sigma}^2_\varepsilon(ugm)}$, dove *ugm* è il modello non condizionato di crescita, mentre *cm* è il modello condizionato di crescita.

3.3 LCM - STIMA DEL MODELLO

Tra i vari metodi per stimare il modello[41] (si considera il solo modello lineare), si esamina ora il metodo chiamato "*case by case*" (Wishart, 1938). Questa metodologia modifica una delle assunzioni di cui si è parlato nel precedente paragrafo: invece di assumere la varianza degli errori come costante tra gli individui e variabile nel tempo, qui tale varianza viene considerata costante nel tempo e diversa per ogni individuo: $\text{var}(\varepsilon_{it}) = \text{var}(\varepsilon_i)$. Inoltre, se si assume che non vi siano dati mancanti, è possibile ottenere uno stimatore BLUE (Best Linear Unbiased Estimator) per l'intercetta e la pendenza, per ogni individuo

[41] Curran "Latent Curve Models - A structural equation perspective"

($\hat{\beta}_i = \frac{\sum_{t=1}^{T} (\lambda_t - \bar{\lambda})(y_{it} - \bar{y}_i)}{\sum_{t=1}^{T} (\lambda_t - \bar{\lambda})^2}$ e $\hat{\alpha}_i = \bar{y}_i - \hat{\beta}_i \bar{\lambda}$). L'idea di base del metodo *case by case* è in effetti quella di stimare tanti modelli quanti sono gli individui, arrivando così ad N stime delle variabili latenti (coefficienti, intercetta e pendenza); tali stime portano alla costruzione di alcuni indici sintetici (tipicamente medie, $\hat{\mu}_\alpha = \frac{1}{N}\sum_{i=1}^{N} \hat{\alpha}_i$, $\hat{\mu}_\beta = \frac{1}{N}\sum_{i=1}^{N} \hat{\beta}_i$, ed errori standard, $s.e.(\hat{\mu}_\alpha) = \sqrt{\frac{1}{N}\text{var}(\hat{\alpha})}$ $= \sqrt{\frac{1}{N(N-1)}\sum_{i=1}^{N}(\hat{\alpha}_i - \hat{\mu}_\alpha)^2}$, $s.e.(\hat{\mu}_\beta) = \sqrt{\frac{1}{N(N-1)}\sum_{i=1}^{N}(\hat{\beta}_i - \hat{\mu}_\beta)^2}$), di intervalli di confidenza e di test di significatività.

Gli errori standard delle stime di intercetta e pendenza non possono essere direttamente usati per stimare le varianze dei coefficienti, ma necessitano di una correzione che dipende dal modello degli errori. Data la varianza $\text{var}(\varepsilon) = \frac{1}{N}\sum_{i=1}^{N} \text{var}(\varepsilon_i)$, si ha che:

$$\hat{\psi}_{\alpha\alpha} = \text{var}(\hat{\alpha}) - \frac{\text{var}(\varepsilon)\sum_{t=1}^{T} \lambda_t^2}{T\sum_{t=1}^{T} (\lambda_t - \bar{\lambda})^2} \text{ e}$$

$$\hat{\psi}_{\beta\beta} = \text{var}(\hat{\beta}) - \frac{\text{var}(\varepsilon)}{\sum_{t=1}^{T} (\lambda_t - \bar{\lambda})^2},$$

potendo in tal modo ottenere stimatori corretti delle varianze di intercette e pendenza.

Il metodo descritto è intuitivo e porta facilmente a risultati comprensibili, ma non fornisce una stima globale del modello. Inoltre esso impone alcune forti limitazioni sulla struttura degli errori: al fine di ottenere stimatori OLS, è necessario supporre che la varianza degli errori sia costante nel tempo per ogni individuo ($\text{var}(\varepsilon_{it}) = \text{var}(\varepsilon_i)$). Comunque, quando si calcola $\text{var}(\varepsilon) = \frac{1}{N}\sum_{i=1}^{N} \text{var}(\varepsilon_i)$, per correggere le stime delle varianze delle variabili latenti o quando si confrontano le varianze degli errori, si assume che le varianze osservate siano realizzazioni empiriche (e quindi affette da errori) di una variabile aleatoria, la cui media è $\text{var}(\varepsilon)$, quindi si presume che tutti gli individui abbiano in realtà la stessa varianza degli errori. Questa è un'assunzione molto restrittiva, che non considera la presenza di una struttura degli errori più complessa, che potrebbe invece esistere.

Siccome i parametri del modello sono a loro volta variabili aleatorie, è interessante valutare in che modo dipendano da altre variabili osservate, fisse o anche variabili nel tempo. Con il metodo *case by case*, questa struttura diventa abbastanza difficile da rappresentare e stimare.

L'approccio *case by case* sceglie per prima cosa una forma funzionale esplicita per catturare la relazione tra le misure ripetute e la misurazione del tempo[42]. In seguito adatta questa funzione ad ogni individuo, per poi mettere insieme le stime individuali dei parametri. Il problema rimane quello di trovare una funzione appropriata per ogni individuo. In generale, è utile adottare trasformazioni delle misure ripetute o della scala temporale, oppure anche provare diverse forme funzionali. Ma potrebbe anche essere possibile avere una diversa forma funzionale ottimale per ciascun individuo, o gruppo di individui.

Baker (1954) per primo utilizzò tecniche di analisi fattoriale per stimare il *Latent Curve Model*[43]: utilizzò uno stimatore di sistema al fine di stimare simultaneamente tutti i parametri; inoltre stimò anche le curve, analizzando una matrice di correlazione (sebbene ciò comportasse la rimozione delle differenze nelle varianze delle misure ripetute), senza utilizzare una forma funzionale specifica e senza stimare i valori dei fattori latenti.

Rao e Tucker (1958) cercarono di affinare tale tecnica. Rao utilizzò un modello di analisi fattoriale basato sulle differenze temporali per singolo individuo, invece Tucker considerò la presenza di una relazione funzionale tra la variabile obiettivo in un determinato istante temporale e la misura del tempo. Il vantaggio di questi modelli stava nel fatto che non vi fosse la necessità di specificare l'esatta natura della funzione del tempo o della funzione dei parametri ϑ_i. Era possibile trovare le traiettorie e stimare la loro importanza per ogni caso, attraverso la previsione dei valori dei fattori latenti. In questi approcci vi era però un grosso problema: non era possibile ottenere una soluzione unica perché il modello era sottoidentificato.

L'analisi fattoriale fornisce i pesi per ogni individuo su ogni fattore. Questi pesi vengono stimati facilmente dall'approccio *case by case*. Quindi, data la forma funzionale tra le misure ripetute ed il tempo, è più semplice usare l'approccio *case by case* piuttosto che gli altri citati. Recentemente, i ricercatori hanno perso l'interesse sugli individui; non tendono a stimare un modello per ogni caso in quanto non ritengono di fondamentale interesse i parametri individuali. Focalizzano invece l'attenzione sulla matrice delle covarianze e sulle medie delle misure ripetute. Quindi hanno anche perduto interesse alla stima del modello individuale. Al momento, la forma funzionale che viene maggiormente utilizzata è il modello lineare, con qualche occasionale scelta di forma quadratica.

L'approccio strutturale è il punto di partenza più frequentemente usato per la stima del LCM, in particolare la metodologia più comune è la stima di massima verosimiglianza.

L'idea di base è che esista una traiettoria latente, non osservata, che produce tutte le traiettorie osservate (l'andamento nel tempo di ogni individuo). La traiettoria latente è quella definita dai parametri che vengono stimati dal modello: intercetta e pendenza, nel modello lineare.

Il punto di partenza è la rappresentazione matriciale del modello:

$y = \Lambda\ \eta + \varepsilon$ $\quad y$ (Tx1) Λ (Txm) η (mx1) ε (Tx1) m è il numero dei fattori latenti (2 nel caso lineare).

La rappresentazione completa sarebbe $y = \tau_y + \Lambda\ \eta + \varepsilon$ dove τ_y sono le intercette delle misure ripetute (un vettore Tx1), ma sono poste =0, in modo da assegnare ai parametri il significato di punto di partenza (α) e di tasso di variazione per unità temporale (β).

Nel caso di traiettoria lineare, il modello diventa:

[42] Bollen "On the origins of latent curve model"
[43] Bollen "On the origins of latent curve model"

$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ . \\ . \\ y_{iT} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ . & . \\ . & . \\ 1 & T-1 \end{pmatrix} \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ . \\ . \\ \varepsilon_{iT} \end{pmatrix} \quad i = 1,\ldots,N \text{ equivalente a } y = \Lambda \eta + \varepsilon$$

La matrice Λ contiene i valori fissi convenzionali e rappresenta i valori del tempo. La colonna unitaria, collegata al fattore intercette, fa in modo che il primo fattore latente influenzi tutte le misure ripetute lungo tutti gli istanti temporali[44]. Per quanto riguarda la seconda colonna, questa è solo una delle possibilità (i valori di *t* dovrebbero comunque riflettere la spaziatura tra i diversi istanti temporali)[45]; in questo caso, le informazioni sono disponibili ad intervalli di tempo equidistanti e il valore iniziale 0 indica che l'intercetta rappresenta il valore atteso al primo istante di tempo considerato. Una peculiarità del LCM è che i valori della matrice Λ sono fissati a priori; altre analisi, infatti, mirano a stimare anche questi valori (o almeno alcuni di essi). L'obiettivo è ora quello di stimare i parametri, ma anche le loro varianze e covarianze, oltre alle varianze dei termini di errore.

In un modo simile a quello già visto, si può esprimere il vettore delle variabili latenti come composto da una parte fissa (le medie) e da una parte variabile (gli errori):

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix} + \begin{pmatrix} \zeta_{\alpha_i} \\ \zeta_{\beta_i} \end{pmatrix} \quad \text{equivalente a} \quad \eta = \mu_\eta + \zeta$$

Si ottiene così l'equazione strutturale:

$$y_i = \Lambda (\mu_\eta + \zeta_i) + \varepsilon_i$$

La media diventa: $E(y) = \Lambda \mu_\eta$

La varianza diventa: $\Sigma = \Lambda \Psi \Lambda' + \Theta_\varepsilon$

dove Σ (TxT) è la matrice di varianze e covarianze di Y; Θ_ε (TxT) è la matrice di varianze e covarianze del termine di errore, che a sua volta è la matrice diagonale di var (ε_t) *t*=1,…,T; Λ (Tx2) è la matrice dei *factor loadings*; Ψ (2x2) è la matrice di varianze e covarianze degli errori dei fattori latenti: $\begin{pmatrix} \psi_{\alpha\alpha} & \psi_{\alpha\beta} \\ \psi_{\alpha\beta} & \psi_{\beta\beta} \end{pmatrix}$.

Nel modello più semplice, senza variabili esplicative per i fattori latenti: $\Sigma_{\eta\eta} = \Psi$ cioè la varianza dei fattori latenti *s* (η) è la varianza degli errori stessa (ζ)[46].

La descrizione grafica di un modello lineare non condizionato, con 3 "onde" di dati (sono cioè considerati 3 istanti temporali), è quella in Grafico 33.

In questo diagramma, i quadrati rappresentano le variabili osservate, qualcosa di misurabile direttamente; i cerchi sono le variabili latenti. Una freccia unidirezionale indica l'ipotesi di causalità della variabile dalla quale parte la freccia rispetto alla variabile di arrivo. I numeri indicati in rosso indicano l'entità della relazione: [1, 1, 1] indica che l'influenza della variabile latente alpha è costante nei diversi istanti temporali; [0, 1, 2] indica che

[44] Curran "Latent Curve Models - A structural equation perspective"
[45] Fuzhong Li et al. "Modeling Interaction Effects in Latent Growth Curve Models"
[46] Curran "Latent Curve Models - A structural equation perspective"

la relazione con beta dipende dal tempo (vi sono 3 istanti temporali tra loro equidistanti). Una freccia bidirezionale indica una covarianza (ma se la freccia arriva nello stesso posto da cui parte, indica una varianza).

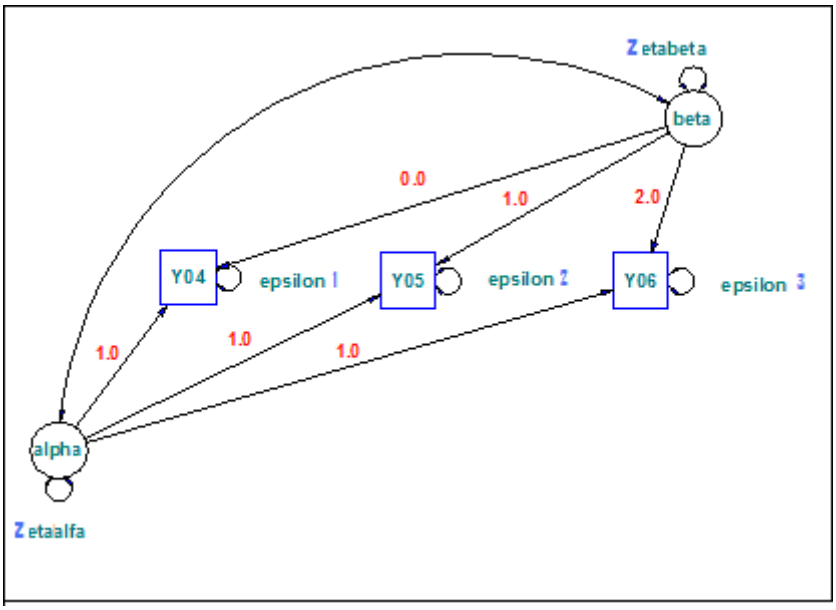


Grafico 33 – Rappresentazione di un modello lineare non condizionato

Al fine di stimare il modello descritto, lo stimatore maggiormente utilizzato è lo stimatore di massima verosimiglianza (ML). Questo, in effetti, gode di molte proprietà e proprio per questo è usato così di frequente in questo tipo di analisi. Prima di procedere con la stima, è però necessario porre alcune ipotesi sugli errori: questi si assumono omoschedastici tra gli individui nello stesso tempo ($\text{var}(\varepsilon_{it}) = \text{var}(\varepsilon_t)$ per tutti gli *i*); inoltre si pone che abbiano varianze e covarianze non note; ciascuno dei termini di errore ($\zeta_{\alpha_i}, \zeta_{\beta_i}, \varepsilon_{it}$) si assume distribuito normalmente con media 0; gli errori del modello di primo livello (ε_{it}) sono posti indipendenti dagli errori di secondo livello ($\zeta_{\alpha_i}, \zeta_{\beta_i}$); infine tutti i termini di errore sono ipotizzati indipendenti dai predittori del modello.

Definendo $\vartheta = (\mu_\alpha, \mu_\beta, \lambda_t, \text{var}(\varepsilon_t), \psi_{\alpha\alpha}, \psi_{\beta\beta}, \psi_{\alpha\beta})$ come il vettore dei parametri del modello,

Si ha $\mu = \mu(\vartheta)$, cioè $\begin{pmatrix} \mu_{y_1} \\ \mu_{y_2} \\ . \\ . \\ \mu_{y_T} \end{pmatrix} = \begin{pmatrix} \mu_\alpha + \lambda_1 \mu_\beta \\ \mu_\alpha + \lambda_2 \mu_\beta \\ . \\ . \\ \mu_\alpha + \lambda_T \mu_\beta \end{pmatrix}$, l'equazione strutturale delle medie

e $\Sigma = \Sigma(\vartheta)$ cioè

$$\begin{pmatrix} \text{var}(y_1) & \text{cov}(y_1, y_2) & . & . & \text{cov}(y_1, y_T) \\ \text{cov}(y_2, y_1) & \text{var}(y_2) & . & . & \text{cov}(y_2, y_T) \\ . & . & . & . & . \\ . & . & . & . & . \\ \text{cov}(y_T, y_1) & \text{cov}(y_T, y_2) & . & . & \text{var}(y_T) \end{pmatrix} = \begin{pmatrix} \psi_{\alpha\alpha} + \lambda_1^2 \psi_{\beta\beta} + 2\lambda_1 \psi_{\alpha\beta} + \text{var}(\varepsilon_1) & . & . & . & \psi_{\alpha\alpha} + \lambda_1 \lambda_T \psi_{\beta\beta} + (\lambda_1 + \lambda_T) \psi_{\alpha\beta} \\ \psi_{\alpha\alpha} + \lambda_2 \lambda_1 \psi_{\beta\beta} + (\lambda_{21} + \lambda_1) \psi_{\alpha\beta} & . & . & . & \psi_{\alpha\alpha} + \lambda_2 \lambda_T \psi_{\beta\beta} + (\lambda_2 + \lambda_T) \psi_{\alpha\beta} \\ . & . & . & . & . \\ . & . & . & . & . \\ \psi_{\alpha\alpha} + \lambda_T \lambda_1 \psi_{\beta\beta} + (\lambda_T + \lambda_1) \psi_{\alpha\beta} & . & . & . & \psi_{\alpha\alpha} + \lambda_T^2 \psi_{\beta\beta} + 2\lambda_T \psi_{\alpha\beta} + \text{var}(\varepsilon_T) \end{pmatrix},$$

la struttura delle varianze e covarianze.

Le medie, le varianze e le covarianze delle variabili osservate sono, nel modello, funzioni dei parametri. Sostituendo i valori in popolazione (μ e Σ) con i valori osservati (campione) ($\bar{y}$ e S), è possibile ottenere le stime di massima verosimiglianza, a partire dalla funzione di verosimiglianza (espressa in modo che debba essere minimizzata):

$$\ln\left|\Sigma(\vartheta)\right| - \ln\left|S\right| + tr\left|\Sigma(\vartheta)^{-1}S\right| - p - (\bar{y} - \mu(\vartheta))'\Sigma(\vartheta)^{-1}(\bar{y} - \mu(\vartheta))$$

dove *p* è il numero di variabili osservate nel modello.

Lo stimatore, ottenuto minimizzando la funzione considerata, è consistente, asintoticamente corretto, asintoticamente distribuito normalmente e asintoticamente efficiente[47]. Inoltre, una qualsiasi funzione delle stime è ancora una stima ML, quindi le traiettorie di crescita predette dal modello sono a loro volta stime ML delle traiettorie di crescita vere. Non vi è una numerosità campionaria minima richiesta, in generale i ricercatori raccomandano almeno N=100, ma una numerosità adeguata per garantire le proprietà asintotiche sarebbe intorno ai 500 individui. In realtà, la numerosità adeguata varia a seconda del contesto in cui si opera.

Vi è da notare che ogni individuo contribuisce alla funzione ML tante volte quanti istanti temporali vengono considerati. $\Sigma(\vartheta)$ è la matrice di varianze e covarianze derivante dal modello (*model implied*); essa contribuirà a dire al ricercatore quanto bene il modello si adatta ai dati osservati.

È possibile distinguere tra due tipi di stime ML[48]: la tipologia *full* e quella *restricted*. Nella *full* (FML) vengono stimati congiuntamente tutti i parametri incogniti (effetti fissi e componenti di varianza), in tal modo le stime delle varianze contengono le stime degli effetti fissi senza tener conto della loro incertezza, quindi producono sottostime, specialmente nei piccoli campioni. Nella classica regressione *cross section* è possibile stimare la varianza residua $\hat{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{N}\hat{\varepsilon}_i^2$, dove $\hat{\varepsilon}_i^2$ sono i quadrati degli errori della regressione stimata; le stime dei parametri ($\hat{\beta}$) sono stimatori FML e i gradi di libertà sono assunti rimanere *n*, mentre, a causa delle *p+1* stime dei parametri, sono in realtà di meno. Lo stimatore corretto della varianza residua è perciò $\hat{\sigma}_\varepsilon^2 = \frac{1}{n-(p+1)}\sum_{i=1}^{N}\hat{\varepsilon}_i^2$. Analogamente è possibile ottenere lo stimatore corretto per dati longitudinali, con il metodo RML. Viene utilizzato un procedimento iterativo. Inizialmente si fa una stima degli effetti fissi con altri metodi (OLS o GLS); poi le stime dei γ entrano nella stima dell'errore per ogni individuo in ciascun istante temporale. È possibile scrivere la funzione di massima verosimiglianza (e il suo logaritmo) in modo che come dati vi siano gli errori e che l'obiettivo sia quello di stimare le loro componenti di varianza, ponendo i γ come noti.

[47] Curran "Latent Curve Models - A structural equation perspective"
[48] Singer, Willett "Applied longitudinal data analysis – modelling change and event occurrence"

3.4 LCM - DIPENDENZA DEI FATTORI LATENTI DA VARIABILI ESOGENE: IL MODELLO CONDIZIONATO

Al fine di risalire ai motivi del cambiamento, si può aggiungere al modello un insieme di variabili che contribuiscano a determinare i fattori latenti.

$\eta_i = \mu_\eta + \zeta_i$, in generale, diventa $\eta_i = \mu_\eta + \Gamma X_i + \zeta_i$, dove μ_η è un vettore *m*x1 (*m* è il numero dei fattori, nel caso lineare *m*=2) che contiene le intercette dei fattori della curva latente; Γ è la matrice dei pesi dei predittori esogeni; η_i è il vettore dei coefficienti, nel caso lineare α_i e β_i.

Una premessa chiave che occorre fare in questo tipo di analisi è che i predittori esogeni siano variabili invarianti nel tempo (che rimangano cioè le medesime per ogni individuo lungo tutto l'intervallo temporale considerato); se sono, per loro natura, variabili che assumono diversi valori nel tempo, l'ipotesi che bisogna fare è che esse siano misurate soltanto una volta, di solito nel primo istante temporale considerato, ma potrebbe scegliersi anche un altro momento, e che quel valore sia assunto essere costante nel tempo.

Come primo strumento esplorativo[49] per vedere graficamente l'impatto di alcune variabili esplicative (quelle ritenute maggiormente descrittive del fenomeno esaminato), si può utilizzare un grafico delle traiettorie individuali livellate (ad esempio attraverso una procedura OLS), visualizzate in modo separato per i diversi gruppi di individui caratterizzati dai valori ritenuti principali (o per i diversi intervalli temporali) delle variabili esogene. In questo modo, si può vedere la differenza tra i diversi tipi di individui in termini di intercette rispetto anche all'intercetta media (il punto iniziale medio) ed in termini di pendenze, rispetto anche alla pendenza media (il tasso di variazione medio).

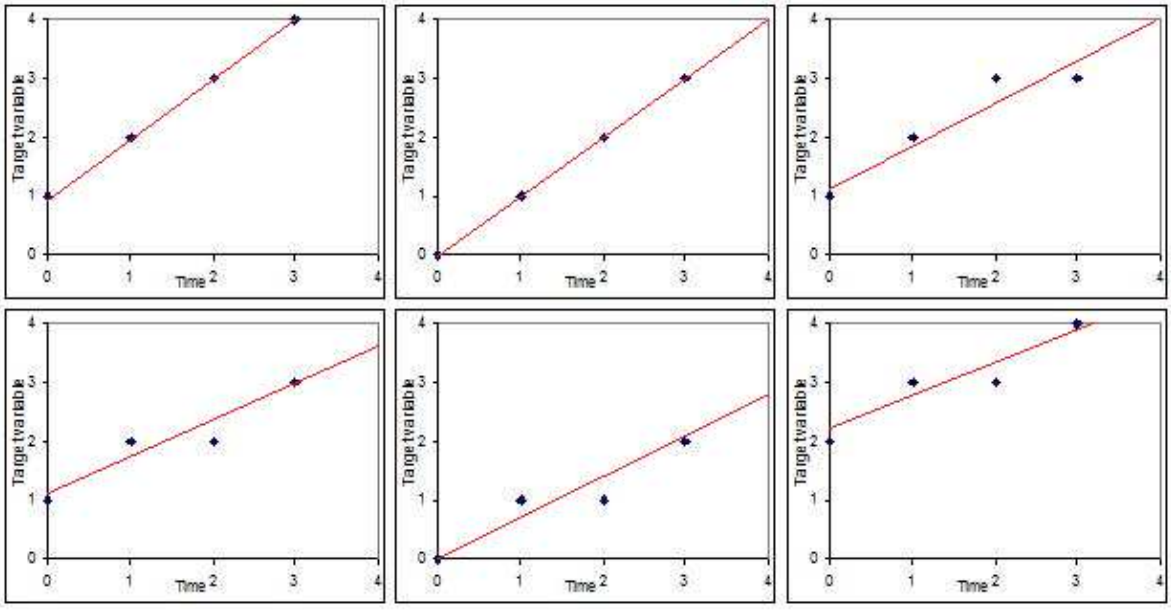


Grafico 34 – Esempio di *growth plots* con traiettorie livellate

Un'altra modalità di tipo grafico, utile per una iniziale esplorazione delle relazioni tra i predittori e le traiettorie stimate, è quella di considerare le intercette e le pendenze stimate come risultati di modelli in cui queste sono le variabili dipendenti, mentre le variabili esogene sono i predittori. È possibile vedere graficamente questa relazione, inserendo in grafico le intercette o le pendenze stimate sui valori dei predittori (un grafico per ogni

[49] Singer, Willett "Applied longitudinal data analysis – modelling change and event occurrence"

predittore). Ovviamente, queste analisi sono meramente esplorative e sono utili soltanto per dare una prima idea dei comportamenti delle variabili. Per un'analisi accurata c'è bisogno di misure quantitative.

Modello *combined*: $y_i = \Lambda\left(\mu_\eta + \Gamma X_i\right) + \Lambda\,\zeta_i + \varepsilon_i$.

La media[50] diventa: $\mu\left(\vartheta\right) = \Lambda\left(\mu_\eta + \Gamma\,\mu_x\right)$.

La struttura delle varianze e covarianze diventa: $\Sigma\left(\vartheta\right) = \Lambda\left(\Gamma\Omega\,\Gamma' + \Psi\right)\Lambda' + \Theta_\varepsilon$,

con μ_x vettore delle medie e Ω matrice di varianze e covarianze (*p*x*p*) dei *p* predittori esogeni.

Lo stimatore più diffuso si ottiene con il metodo *Direct Maximum Likelihood*:

$$\ln L\left(\vartheta\right) = \sum_{i=1}^{N} \ln L_i\left(\vartheta\right) \quad \ln L_i\left(\vartheta\right) = K_i - \frac{1}{2}\ln\left|\Sigma_i\left(\vartheta\right)\right| - \frac{1}{2}\left(z_i - \mu_i\left(\vartheta\right)\right)'\Sigma_i\left(\vartheta\right)\left(z_i - \mu_i\left(\vartheta\right)\right),$$

dove z_i è il vettore delle variabili osservate per l'individuo *i*-esimo (la variabile obiettivo Y e le variabili esogene X); K_i è una costante incorrelata a ϑ, $\Sigma_i\left(\vartheta\right) = E\left[\begin{pmatrix} y_i - \mu_y \\ x_i - \mu_x \end{pmatrix}\begin{pmatrix} y_i - \mu_y \\ x_i - \mu_x \end{pmatrix}'\right]$ è la matrice di varianze e covarianze della variabile obiettivo e delle variabili esogene, allo stesso modo $\mu_i\left(\vartheta\right) = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}$.

Introdurre variabili esogene nel modello porta necessariamente a dover interpretare i loro effetti: gli effetti diretti sui fattori latenti e quelli indiretti sulle misure ripetute. Gli effetti diretti vengono interpretati come nel modello lineare classico: il coefficiente del regressore indica di quanto cambia un fattore latente per ogni unità di variazione del regressore stesso. Quello seguente è il modello con un solo regressore X:

$$\alpha_i = \mu_\alpha + \gamma_{11}x_{1i} + \zeta_{\alpha_i} \quad \text{e}$$

$$\beta_i = \mu_\beta + \gamma_{21}x_{1i} + \zeta_{\beta_i}.$$

Se X è, per esempio, una variabile dummy, μ_α e μ_β sono l'intercetta media e la pendenza media della traiettoria latente per gli individui che hanno x_1=0, mentre γ_{11} e γ_{21} rappresentano la differenza in termini, rispettivamente, di intercetta e di pendenza, degli individui che hanno x_1=1 rispetto all'altro gruppo. Se x_1 è continua, μ_α e μ_β continuano a rappresentare l'intercetta e la pendenza medie per coloro che hanno x_1=0, mentre γ_{11} e γ_{21} rappresentano il cambiamento atteso, rispettivamente, in termini di intercetta e di pendenza, per un cambiamento unitario di x_1.

In modo analogo è possibile interpretare i coefficienti di una regressione multipla. In questo caso, vi è la complicazione che potrebbero esservi delle interazioni tra i regressori: non è possibile interpretare l'influenza di una variabile esplicativa senza considerare tutte le altre. Lasciando per un momento da parte le interazioni, i coefficienti in questo caso rappresentano l'effetto di ciascuna variabile esogena sui fattori latenti, senza considerare le altre. Gli effetti indiretti delle esogene del fattore intercetta sulle misure ripetute non sono nulla di più che gli effetti diretti stessi, visto che i *factor loadings* sono =1. Per quanto invece riguarda le esogene del

[50] Coffman, Millsap "Evaluating Latent Growth Curve Models Using Individual Fit Statistics"

fattore pendenza, vi è una difficoltà aggiuntiva: l'influenza di esse sulle misure ripetute varia in funzione del tempo (λ_t). Inoltre, l'effetto delle variabili esplicative sulla variabile obiettivo, mediato dai coefficienti di pendenza, subisce una componente di interazione tra tali variabili ed il tempo (λ_t). Questo diviene ancora più chiaro se si analizza il modello in forma compatta, nel caso di una singola variabile esogena:

$$y_{it} = \left(\mu_\alpha + \mu_\beta \lambda_t + \gamma_{11} x_{1i} + \gamma_{21} \lambda_t x_{1i}\right) + \left(\zeta_{\alpha_i} + \zeta_{\beta_i} \lambda_t + \varepsilon_{it}\right).$$

L'attenzione del ricercatore si focalizza anche sulle interazioni tra le variabili esplicative[51]. Questi vorrebbe conoscere quali variabili sono in grado di prevedere il perché alcuni individui hanno un tasso di variazione alto, mentre alcuni altri ne hanno uno inferiore o addirittura mantengono lo stesso valore nel tempo; d'altronde è anche interessato a sapere se l'influenza di una variabile sullo stato iniziale della variabile obiettivo e sul suo tasso di variazione dipende dal livello della variabile stessa. L'interazione può entrare nel modello aggiungendo un termine moltiplicativo nella regressione delle variabili latenti. Il modello con due variabili esogene per intercetta e pendenza è il seguente:

$$\alpha_i = \mu_\alpha + \gamma_{11} x_1 + \gamma_{12} x_2 + \gamma_{13} x_1 x_2 + \zeta_{\alpha_i} \text{ intercetta,}$$

$$\beta_i = \mu_\beta + \gamma_{21} x_1 + \gamma_{22} x_2 + \gamma_{23} x_1 x_2 + \zeta_{\beta_i} \text{ pendenza.}$$

I coefficienti dell'intercetta possono essere interpretati come gli effetti diretti dei regressori sullo stato iniziale della variabile obiettivo; se γ_{13} è significativo, allora l'effetto di x_1 sul livello iniziale di *Y* dipende dal livello di x_2 , indicando l'effetto di interazione. I parametri γ_{21} e γ_{22} indicano quanto le variabili esogene influiscono sulla relazione tra *Y* e il tempo, cioè il tasso di variazione nel tempo della variabile obiettivo (se entrambi i parametri sono positivi, quegli individui che hanno alti valori di x_1 e di x_2 avranno un tasso di variazione di *Y* più alto degli altri individui). Il coefficiente γ_{23} misura quanto l'effetto della prima variabile esogena sul tasso di variazione dipende dall'altra variabile (un valore positivo indica che quegli individui che hanno alti livelli di x_2 hanno una forte influenza di x_1 sul tasso di variazione della variabile obiettivo).

Se il ricercatore considera l'ipotesi di aggiungere un regressore i cui valori cambiano nel tempo (*time varying predictor*)[52], il modo più semplice è partire dalla specificazione composite del modello non condizionato:

$$Y_{it} = \left(\mu_\alpha + \mu_\beta \lambda_{it}\right) + \left(\zeta_{\alpha_i} + \zeta_{\beta_i} \lambda_{it} + \varepsilon_{it}\right).$$

È ora possibile aggiungere l'effetto della covariata con valori che variano nel tempo nella parte fissa del modello:

$$Y_{it} = \left(\mu_\alpha + \mu_\beta \lambda_{it} + \mu_\gamma X_{it}\right) + \left(\zeta_{\alpha_i} + \zeta_{\beta_i} \lambda_{it} + \varepsilon_{it}\right).$$

In questo modello, il valore della variabile obiettivo dipende anche dal contemporaneo valore del regressore X. Il parametro μ_β indica il tasso di variazione medio in popolazione per unità di tempo, mentre μ_γ indica la differenza media in popolazione nel tempo tra individui con diversi livelli di X; l'intercetta μ_α indica il valore

[51] Fuzhong Li et al. "Modelling Interaction Effects in Latent Growth Curve Models**"**
[52] Singer, Willett "Applied longitudinal data analysis – modelling change and event occurrence"

di Y al tempo =0 e al valore 0 del predittore X. La differenza tra gli individui è data, come di consueto, dagli elementi stocastici ζ_{α_i}, per l'intercetta, e ζ_{β_i} per la pendenza.

Lo stesso modello può essere scomposto in due sotto modelli di primo e secondo livello:

$Y_{it} = \alpha_i + \beta_i \lambda_{it} + \gamma_i X_{it} + \varepsilon_{it}$ primo livello,

$$\left.\begin{aligned} \alpha_i &= \mu_\alpha + \zeta_{\alpha_i} \\ \beta_i &= \mu_\beta + \zeta_{\beta_i} \\ \gamma_i &= \mu_\gamma \end{aligned}\right\} \text{ secondo livello.}$$

Si possono anche aggiungere variabili esogene invarianti nel tempo, al secondo livello.

L'ultima equazione suppone che l'effetto della variabile con valori non costanti nel tempo sia costante tra gli individui (in popolazione); questa assunzione (γ_{11} e γ_{21} sono costanti nel tempo) è richiesta per una variabile costante nel tempo (*time invariant*), ma può essere rivisitata nel caso di variabilità nel tempo, a seconda delle assunzioni fatte nel modello (tale ipotesi può essere considerata nel caso ve ne sia realmente la necessità e soltanto con una disponibilità di molte osservazioni); se si vuole considerare la variazione casuale tra gli individui in popolazione, la stessa equazione diventa: $\gamma_i = \mu_\gamma + \zeta_{\gamma_i}$. In questo caso, le nuove assunzioni del modello sono:

$$\varepsilon_{it} \sim N\left(0, \sigma_\varepsilon^2\right) \text{ e } \begin{pmatrix} \zeta_{\alpha_i} \\ \zeta_{\beta_i} \\ \zeta_{\gamma_i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} & \sigma_{\alpha\gamma} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 & \sigma_{\beta\gamma} \\ \sigma_{\alpha\gamma} & \sigma_{\beta\gamma} & \sigma_\gamma^2 \end{pmatrix} \right);$$

l'aggiunta di un termine di errore comporta 3 varianze in più. È sempre meglio iniziare da un modello più semplice prima di stimare altri modelli come quello descritto.

Una ulteriore differenza tra le esogene invarianti nel tempo e quelle che invece variano nel tempo sta negli effetti sulle varianze. L'aggiunta di un regressore costante nel tempo non influisce sulla componente di varianza al primo livello σ_ε^2, perché quel regressore non spiega direttamente la variazione entro gli individui; però lo stesso può spiegare alcune parti delle varianze di secondo livello σ_α^2 e σ_β^2, in quanto i suoi effetti ricadono sulla variazione tra gli individui di intercetta e pendenza. Un regressore che varia nel tempo, invece, può influire sia sulle varianze di primo livello che su quelle di secondo livello, infatti può spiegare la variazione diretta tra gli individui ed anche quella tra le intercette e pendenze individuali.

Una cosa importante da notare è che non ha senso confrontare le varianze in modelli diversi. La ragione sta nel fatto che, aggiungendo un regressore che varia nel tempo, cambia il significato dei parametri. Infatti, mentre nel modello non condizionato l'intercetta α_i è il valore individuale della variabile obiettivo al tempo 0, nel modello condizionato con un regressore che varia nel tempo, lo stesso parametro indica il valore individuale al tempo 0, ma con anche il valore di X =0; allo stesso modo, per la pendenza, nel modello condizionato β_i è un tasso di variazione condizionato, dato l'effetto del regressore.

La rappresentazione grafica di un modello con regressori con valori che cambiano nel tempo è quella riportata in Grafico 35 (3 istanti temporali ed una sola variabile esogena W).

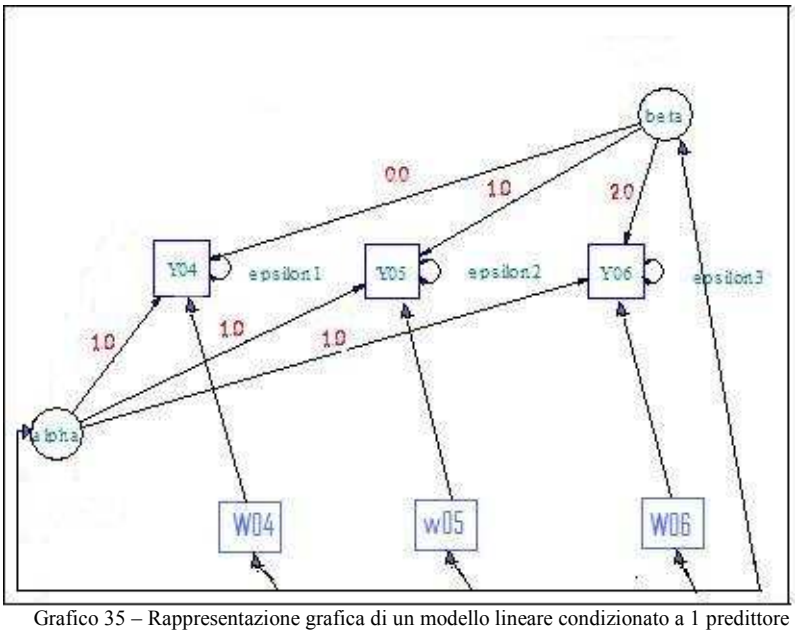


Grafico 35 – Rappresentazione grafica di un modello lineare condizionato a 1 predittore

La variabile obiettivo viene così espressa in funzione non solo dei coefficienti casuali, ma anche del regressore variabile nel tempo. Il regressore influisce sui valori osservati e sul tempo. Risulta possibile stimare l'influenza del regressore e poi esaminare il processo di crescita senza considerare gli effetti di quest'ultimo.

Il modello specificato è non condizionato, nel senso che i fattori latenti non rientrano in un modello di regressione con variabili esogene costanti nel tempo (come nel modello condizionato senza regressori che variano nel tempo).

È ancora possibile aggiungere, in questo senso, una ulteriore specificazione dei parametri del modello.

L'equazione di primo livello rimane: $y_{it} = \alpha_i + \beta_i \lambda_t + \gamma_t w_{it} + \varepsilon_{it}$, dove w è la variabile esogena che varia nel tempo (TVC). La corrispondente forma matriciale è: $y_i = \Lambda_{y\eta}\eta_i + \Gamma_{yw} w_i + \varepsilon_i$. Le equazioni di secondo livello, dato un insieme di Q regressori esogeni che non variano nel tempo, diventano:

$$\alpha_i = \mu_\alpha + \sum_{q=1}^{Q} \gamma_{\alpha q} x_{iq} + \zeta_{\alpha_i} \text{ e } \beta_i = \mu_\beta + \sum_{q=1}^{Q} \gamma_{\beta q} x_{iq} + \zeta_{\beta_i};$$

la corrispondente espressione matriciale è $\eta_i = \mu_\eta + \Gamma_{\eta x} x_i + \zeta_i$. La forma compatta ottenuta combinando le espressioni sopra è: $y_i = \Lambda_{y\eta}\mu_\eta + \Lambda_{y\eta}\Gamma_{\eta x} x_i + \Gamma_{yw} w_i + \left(\Lambda_{y\eta}\zeta_i + \varepsilon_i\right)$.

Si noti che gli effetti fissi dei fattori latenti di crescita $\Lambda_{y\eta}\mu_\eta$ e i loro effetti casuali $\left(\Lambda_{y\eta}\zeta_i + \varepsilon_i\right)$ rispecchiano la stabilità e il cambiamento sulla variabile obiettivo, al netto degli effetti del regressore che varia nel tempo.

La rappresentazione grafica (3 istanti temporali, una variabile esogena con valori dipendenti dal tempo, W, e 2 regressori costanti nel tempo, X1 e X2) è quella in Grafico 36.

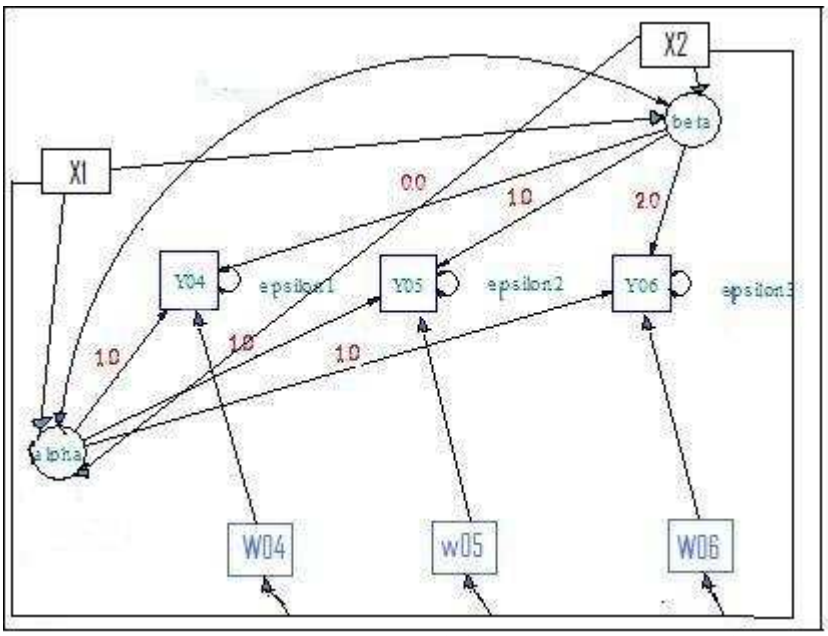


Grafico 36 – Rappresentazione grafica di un modello lineare condizionato con 1 predittore che vaia nel tempo e 2 predittori costanti nel tempo

Il modello appena analizzato comprende soltanto gli effetti principali del regressore e del tempo sulla variabile obiettivo, quindi non viene considerato alcun effetto di interazione tra questi fattori. Graficamente, le traiettorie per valori diversi del predittore (X) sono parallele. Introducendo un effetto di interazione tra il regressore ed il tempo, il modello diventa: $Y_{it} = \left(\mu_\alpha + \mu_\beta \lambda_{it} + \mu_\gamma X_{it} + \mu_{\beta\gamma} \lambda_{it} X_{it}\right) + \left(\zeta_{\alpha_i} + \zeta_{\beta_i} \lambda_{it} + \varepsilon_{it}\right)$.

Il coefficiente di interazione $\mu_{\beta\gamma}$ indica di quanto l'effetto del regressore sulla variabile obiettivo varia nel tempo, ma indica anche quanto il tasso di variazione della variabile obiettivo nel tempo differisce dal valore della variabile esogena X. Se la stima dei parametri del modello suggerisce che il coefficiente direttamente legato al tempo è molto piccolo, è possibile tentare di stimare un modello diverso, dove l'effetto principale del tempo è nullo, ma dove l'interazione tra il regressore e il tempo rimane:

$$Y_{it} = \left(\mu_\alpha + \mu_\gamma X_{it} + \mu_{\beta\gamma} \lambda_{it} X_{it}\right) + \left(\zeta_{\alpha_i} + \zeta_{\alpha\beta} \lambda_{it} X_{it} + \varepsilon_{it}\right).$$

In questo modello, accade che gli effetti fissi e gli effetti casuali si allineano, nel senso che entrambi i tipi di effetti del tempo vengono rimossi, ma viene aggiunto un effetto casuale dell'interazione. Se il ricercatore ritiene che l'effetto della variabile esogena non sia costante per tutti gli individui, può aggiungere un termine casuale per il coefficiente di tale variabile, ottenendo in tal modo:

$$Y_{it} = \left(\mu_\alpha + \mu_\gamma X_{it} + \mu_{\beta\gamma} \lambda_{it} X_{it}\right) + \left(\zeta_{\alpha_i} + \zeta_\gamma X_{it} + \zeta_{\alpha\beta} \lambda_{it} X_{it} + \varepsilon_{it}\right).$$

Così ognuno degli effetti fissi ha il proprio corrispondente effetto casuale.

3.5 LCM - VALUTAZIONE DELLA BONTÀ DI ADATTAMENTO DEL MODELLO

Una volta che il ricercatore ha specificato il modello e che i parametri sono stati stimati, il passo più importante è quello di testare la bontà di adattamento del modello, cioè è importante sapere se il modello stimato possa ben rappresentare i dati empirici. È possibile testare l'ipotesi di buon modello, sia con un approccio individuale che tramite un approccio di tipo strutturale.

3.5.1 APPROCCIO INDIVIDUALE

Nella regressione per singolo caso, l'osservazione delle intercette e delle pendenze individuali può aiutare a valutare la bontà del modello. Un primo approccio è di tipo grafico; si possono inserire in grafici distinti le distribuzioni di frequenza degli α_i e dei β_i, come mostrato in Grafico 37.

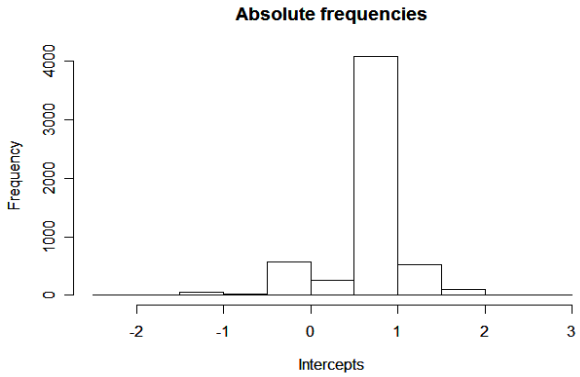


Grafico 37 – Esempio di distribuzione di frequenza, descritta con istogrammi, delle intercette di un LCM

L'istogramma, di cui un esempio relativo alle sole intercette in Grafico 37, mostra con chiarezza la concentrazione delle stime attorno all'intercetta e alla pendenza media, il *range* dei valori e la variabilità delle stime. Se quindi l'istogramma mostra valori dell'intercetta tutti positivi (o negativi), vuol dire che tutti gli individui in popolazione hanno positivi (o negativi) valori iniziali della variabile obiettivo; similmente per la pendenza, il ricercatore può ragionevolmente pensare che tutti gli individui abbiano il tasso di variazione della variabile obiettivo nel tempo che va nella medesima direzione. In questi grafici è anche semplice identificare gli *outliers* nella distribuzione delle intercette e delle pendenze (valori troppo alti o troppo bassi caratterizzati da una bassa densità).

Un secondo strumento per valutare il modello *case by case* è l'osservazione dei valori di R^2 per ogni regressione: questo fornisce informazioni sulla vicinanza della traiettoria lineare ai dati osservati. Ogni R_i^2 dovrebbe essere quanto più vicino possibile ai valori estremi del range di questo indice: 0 e 1. Inserendo gli R_i^2 in un grafico, si può vedere quali valori sono buoni e quali invece sono troppo vicini a 0; il grafico inoltre mostra la dispersione dei valori attorno alla loro media ed anche la eventuale presenza di *outliers*.

Un altro modo per vedere la corrispondenza tra il modello ed i dati osservati è quello di inserire, insieme nello stesso grafico, la traiettoria delle medie temporali della variabile obiettivo e della linea ottenuta dalla media delle intercette e pendenze stimate. Al migliorare del modello, diminuisce la differenza tra le due linee.

Per l'approccio descritto, un limite è che non risulta possibile calcolare un indice di bontà di adattamento dell'intero modello, preso nella sua globalità.

In termini di approccio SEM, è possibile condurre un test d'ipotesi su ogni effetto fisso, γ_{00}, γ_{01}, γ_{10}, e γ_{11}, utilizzando un test su ogni singolo parametro[53]. L'ipotesi nulla più comune su ciascun parametro è $H_0 : \gamma = 0$, ma si può scegliere di testare un altro valore d'interesse; la statistica test è $z = \frac{\hat{\gamma}}{ase(\hat{\gamma})}$ dove il denominatore è l'errore standard asintotico della stima del parametro. Se l'ipotesi nulla non viene rifiutata, ciò suggerisce che il parametro ha un ruolo nella spiegazione del comportamento della variabile obiettivo. In particolare, se non vengono rifiutate entrambe le ipotesi nulle sui parametri γ_{01} e γ_{11}, ciò vuol dire che le differenze, in termini di stato iniziale e di tasso di variazione, tra gli individui con valori diversi della variabile esogena X sono statisticamente significative.

Un altro tipo di test riguarda le stime delle componenti di varianza. Queste componenti riassumono la variabilità del risultato (quello di ogni individuo attorno alla sua traiettoria vera e quello tra gli individui) in un modello corretto. È perciò possibile testare se sia rimasta qualche variazione residua del risultato che potrebbe essere magari spiegata da eventuali altri predittori. Un primo metodo, anche se al momento non molto raccomandato, è il test sul singolo parametro (su σ_ε^2 per il primo livello e su σ_0^2, σ_1^2 e σ_{10} per il secondo livello), che fornisce alcune informazioni sulla significatività dei termini di errore (per esempio, scoprire che σ_{10} non è significativa vuol dire che lo stato iniziale ed il tasso di variazione non sono realmente correlati, mentre se σ_ε^2 è significativa, il ricercatore deve pensare che via sia una qualche variabilità nel risultato non spiegata dal modello).

3.5.2 APPROCCIO STRUTTURALE

Data la funzione di massima verosimiglianza F_{ML}, un indice di bontà di adattamento molto usato è la statistica test calcolata come rapporto di verosimiglianze $T_{ML} = (N-1)F_{ML}$, basata sul metodo ML di stima dei parametri (si noti che una differenza di logaritmi è uguale al logaritmo di un rapporto). T_{ML} viene calcolata con le stime ML; così la funzione di massima verosimiglianza diventa:

$\hat{F}_{ML} = \ln\left|\Sigma\left(\hat{\vartheta}\right)\right| - \ln|S| + tr\left|\Sigma\left(\hat{\vartheta}\right)^{-1} S\right| - p - \left(\bar{y} - \mu\left(\hat{\vartheta}\right)\right)' \Sigma\left(\hat{\vartheta}\right)^{-1}\left(\bar{y} - \mu\left(\hat{\vartheta}\right)\right)$ dove $\hat{\vartheta}$ è la stima ML di ϑ, quindi $\Sigma\left(\hat{\vartheta}\right)$ e $\mu\left(\hat{\vartheta}\right)$ sono la matrice di varianze e covarianze *model implied* e il vettore delle medie *model implied,* mentre *p* è il numero di variabili osservate. Questa funzione permette di testare l'ipotesi nulla $H_0 : \mu = \mu(\vartheta); \Sigma = \Sigma(\vartheta)$, su entrambe, la media e la matrice di varianze e covarianze in popolazione. La caratteristica chiave di questa analisi è il test di questa ipotesi, dove μ è la media in popolazione e $\mu(\vartheta)$ è la media *model implied*, mentre Σ è la matrice di varianze e covarianze in popolazione e $\Sigma(\vartheta)$ è la matrice di varianze e covarianze *model implied* (ϑ è il vettore dei parametri del modello). Ma l'ipotesi più importante che deve essere testata è quella sulla matrice di varianze e covarianze: si deve cioè testare se la matrice di varianze e covarianze in popolazione delle variabili osservate sia una funzione dei parametri liberi incogniti oppure no.

[53] Singer, Willett "Applied longitudinal data analysis – modelling change and event occurrence"

Se le assunzioni sul metodo di stima sono verificate (o se comunque la distribuzione della variabile obiettivo non ha una eccessiva curtosi, come si vedrà spiegato nel seguito) e se l'ipotesi nulla non è falsificata, T_{ML} è distribuito come un χ^2 centrale con $\frac{1}{2}T(T+3)-u$ gradi di libertà, dove T è il numero di istanti temporali e *u* è il numero di parametri liberi del modello. Se il modello è esattamente identificato, la statistica test è =0, quindi il test non risulta di particolare utilità. Tale test di adattamento esatto è di solito irrealistico, poiché difficilmente un modello che utilizza dati reali è esente da errori; in realtà, siccome i modelli empirici sono spesso specificati con errori, un test formale di adattamento esatto non è il miglior metodo per valutare il modello; un banale errore di specificazione può infatti portare al rifiuto di un modello anche se questo potrebbe adeguatamente riprodurre la matrice di varianze e covarianze in popolazione[54].

Prima di costruire un test basato sulla massima verosimiglianza, comunque, è necessario controllare le assunzioni di quel tipo di stima. L'ipotesi basilare è la distribuzione normale della variabile obiettivo. Non è facile avere distribuzioni normali, tuttavia si può a ragion veduta tener conto del lavoro di Browne (1984), il quale provò che è sufficiente avere distribuzioni che, anche se non normali, hanno la stessa curtosi di una normale (o che in generale non hanno un eccesso di curtosi), per poter utilizzare lo stimatore ML senza perdere troppe delle sue proprietà.

Se la variabile obiettivo non è distribuita normalmente, i test che si basano sulla ML sono comunque asintoticamente validi, sempre che i fattori di disturbo ε_{it} siano indipendenti da α e β. Questa assunzione è molto più forte della precedente. Gli ε_{it} potrebbero essere incorrelati con α e con β anche se la $\text{var}(\varepsilon_{it})$ non fosse costante tra i diversi casi, ma α e β sarebbero correlati. Questa correlazione viola la condizione di indipendenza.

Sotto l'ipotesi di indipendenza, le statistiche test sono asintoticamente corrette.

Il problema è che non è facile determinare se gli errori e i parametri siano indipendenti. Per testare questa condizione, è disponibile una statistica chi-quadro corretto, che permette di identificare un'eventuale eccesso di curtosi. Altrimenti è possibile ricampionare (bootstrap) per formare *standard error* simulati in modo empirico.

Un altro approccio[55] è quello di usare una funzione di adattamento WLS:

$F_{WLS} = (s-\sigma(\vartheta))' W^{-1} (s-\sigma(\vartheta))$ dove *s* è il vettore delle varianze e delle covarianze delle variabili osservate, $\sigma(\vartheta)$ è il corrispondente vettore della varianze e covarianze stimate dal modello ($\sigma(\vartheta)$ e *s* sono i vettori contenenti gli elementi non ridondanti delle matrici $\Sigma(\vartheta)$ e *S*), e W è una matrice dei pesi. Asintoticamente $T_{WLS} = (N-1)F_{WLS}$ segue una distribuzione chi-quadrato, quando la scelta di W è uno stimatore consistente della matrice di varianze e covarianze di *s*.

Un altro problema del test T_{ML} è dovuto alla distribuzione χ^2 stessa: se la distribuzione della variabile osservata presenta un eccesso di curtosi, ciò porta la statistica test ad avere valori troppo alti o troppo bassi,

[54] Chen et al. "An empirical evaluation of the use of fixed cut-off points in RMSEA test statistic in SEM"
[55] Hipp, Bollen "Model fit in SEM with censored ordinal dichotomous variables"

indipendentemente dalla bontà del modello. La distribuzione χ^2 della statistica test, inoltre, è asintotica, quindi il campione deve essere grande abbastanza (N>100). Un campione piccolo porta a valori alti della statistica test, anche con un buon modello. Con grandi campioni, occorre comunque valutare la potenza del test χ^2. La potenza del test corrisponde alla capacità della statistica test di rifiutare un'ipotesi nulla falsa. Grazie ad un campione grande, infatti, è possibile trovare anche piccoli errori di specificazione del modello. Per misurare la potenza, è necessario esaminare la distribuzione della statistica test quando l'ipotesi nulla non è vera. In questo caso, se l'errore di specificazione non è molto grande, la distribuzione è un χ^2 non centrale; quindi non basta definire i gradi di libertà per caratterizzare la distribuzione, ma si ha anche bisogno di un parametro di non centralità (che nel caso di distribuzione centrale sarebbe =0). Mentre nel caso di χ^2 centrale (H_0 non è falsa), la statistica test non dipende dal numero di osservazioni (N), se H_0 è falsa, il parametro di non centralità e la statistica test sono positivamente correlate con N. Quindi, all'aumentare di N, anche la statistica test aumenta; in grandi campioni, infatti, si possono reperire più facilmente rispetto a campioni piccoli possibili errori di specificazione.

Un altro modo più semplice per gestire la tendenza del χ^2 ad essere maggiore con grandi campioni è di confrontare i risultati in termini di χ^2 con quelli che provengono da altri indici.

L'indice di Tucker Lewis è il primo, qui trattato, di una serie di indici basati sul confronto del modello ipotizzato con un modello base. Quest'ultimo è un modello più restrittivo. Di solito, il modello base ha le varianze delle variabili osservate come parametri liberi mentre le covarianze sono nulle. Le medie del modello base possono essere parametri liberi.

Quindi $TLI = \hat{\rho}_2 = \left(\frac{T_b}{df_b} - \frac{T_h}{df_h} \right) \frac{1}{\frac{T_b}{df_b} - 1}$ dove T_b e df_b sono, rispettivamente, la statistica test e i gradi di libertà del modello base, mentre T_h e df_h sono quelli del modello ipotizzato. Ogni modello entra nel confronto tenendo conto dei propri gradi di libertà. In termini di valori delle funzioni di adattamento:

$$TLI = \left(\frac{F_b}{df_b} - \frac{F_h}{df_h} \right) \frac{1}{\frac{F_b}{df_b} - \left(\frac{1}{N-1} \right)}.$$

Il range del *TLI* è [0,1], ma è possibile riscontrare valori al di fuori di esso, un buon modello ha un *TLI*=1.

Nell'universo dei campioni, la distribuzione del *TLI* ha per lo più la stessa media per numerosità campionarie diverse. Ma la sua varianza cresce parecchio di più di quella degli altri indici. Infatti, vi sono due modi in cui la numerosità campionaria può influire un indice di *fitting*. Il primo è il caso in cui N entra direttamente nel calcolo dell'indice; il secondo è quello in cui la media della distribuzione campionaria dell'indice è correlata ad N[56]. Dato un modello, supponendo di disegnare molti campioni di uguale numerosità N, indipendenti e casuali, dei valori delle variabili, è possibile pensare ad una stima del modello per ogni campione. Quindi si può formare una

[56] Bollen "A new incremental fit index for GSEM"

distribuzione degli indici di *fitting*. Questo passo può essere ripetuto per diverse numerosità campionarie. In tal modo, può risultare disponibile la distribuzione delle variazioni dell'indice rispetto a N, divenendo così possibile analizzare la relazione tra l'indice di *fitting* e la numerosità campionaria. È chiaro che ciò non si può realizzare in tutti i casi studiati, però in generale alcune simulazioni condotte con il metodo di Montecarlo hanno mostrato le relazioni tra gli indici qui presi in esame e la numerosità campionaria.

Bollen (1989) ha implementato l'*Incremental Fit Index*: $IFI = \frac{T_b - T_h}{T_b - df_h}$ che ha lo stesso range del *TLI* e il cui valore ideale, per il caso di un buon modello, è 1. Sono state condotte diverse simulazioni che hanno mostrato che il valore di questo indice rimane abbastanza stabile anche per diverse numerosità campionarie.

Bentler (1990) e Mc Donald e Mash (1990) hanno per primi calcolato il *Relative Non Centrality Index* $RNI = \frac{(T_b - df_b) - (T_h - df_h)}{T_b - df_b}$ ed anche il simile *Comparative Fit Index* (*CFI*); questi indici mostrano valori per lo più simili a quelli dell'*IFI*.

Si è in precedenza descritto il metodo per saggiare la bontà del modello rispetto ai dati osservati, basato sul test d'ipotesi sugli effetti fissi e sui componenti di varianza (approccio basato sul singolo parametro). Tale metodo può aiutare a determinare se vi sia la necessità di rendere più complesso un modello semplice. Tuttavia questo metodo non si è rivelato molto efficiente; la maggior parte degli statistici preferisce utilizzare un'altra metodologia, basata sulla *deviance statistic*[57]. Il punto di partenza è la funzione di verosimiglianza: le stime ML sono quei valori che massimizzano la funzione di log verosimiglianza, cioè il logaritmo della verosimiglianza congiunta di tutti i dati osservati. Date le stime ML e i dati osservati, è possibile determinare la grandezza della funzione di verosimiglianza: la statistica di log verosimiglianza campionaria, chiamata LL. Questa rappresenta anche un metodo per confrontare modelli diversi: quanto più il valore della statistica LL è grande, tanto migliore è l'adattamento del modello ai dati (se la statistica test è negativa, il valore migliore è quello più vicino a 0). La cosiddetta *deviance statistic* mette a confronto le log verosimiglianze di due modelli, quello ipotizzato ed il modello base. L'indice $Deviance = -2\left[LL_{hypothesized} - LL_{baseline}\right]$ quantifica il peggioramento del modello ipotizzato nei confronti del modello base: quanto minore è il valore dell'indice, tanto migliore è il modello ipotizzato. Il modello base maggiormente utilizzato è quello che porta a un adattamento perfetto, quindi il massimo della sua funzione di verosimiglianza è 1 (è la probabilità che esso riproduca perfettamente i dati osservati), e il logaritmo corrispondente è perciò 0. Quindi la *deviance statistic* diventa –2LL o –2logL. Per confrontare due modelli attraverso questo test statistico, devono essere verificati alcuni requisiti specifici: ogni modello deve essere stato stimato sugli stessi dati (se i dati non sono completi per uno dei due modelli, occorre eliminare gli stessi dati mancanti per calcolare l'altro modello); uno dei due modelli deve essere annidato nell'altro (uno dei due modelli deve essere stato specificato ponendo dei vincoli sui parametri liberi dell'altro modello); si deve inoltre controllare che i due modelli siano stati stimati con lo stesso metodo (per esempio, entrambi con la RML o con a FML). Se è stata applicata la stima FML, la *deviance statistic* permette di testare ipotesi su una qualsiasi combinazione dei parametri, degli effetti fissi e delle componenti di varianza; se invece è stata utilizzata la stima RML, è possibili soltanto testare ipotesi sui componenti di varianza. Sotto l'ipotesi nulla

[57] Singer, Willett "Applied longitudinal data analysis – modelling change and event occurrence"

che i vincoli specificati siano veri, la differenza tra la *deviance statistic* di un modello completo e quella di un modello ridotto è distribuita asintoticamente come un χ^2 con gradi di libertà uguali al numero dei vincoli indipendenti imposti. Questi test risultano particolarmente utili quando è necessario confrontare modelli con un predittore in più al secondo livello.

Bentler e Bonett hanno implementato il *normed fit index*[58]:

$\Delta_1 = \frac{\chi_b^2 - \chi_h^2}{\chi_b^2} = \frac{F_b - F_h}{F_b}$ dove si confrontano i valori dello stimatore χ^2, quello del modello base (*b*) e quello del modello ipotizzato (*h*); analogamente per i valori delle funzioni di fitting (*F*). Tale indice misura il miglioramento, in proporzione, in quanto a bontà di adattamento, del passare dal modello base al modello ipotizzato. Quanto più Δ_1 è vicino a 1, tanto migliore è l'adattamento del modello ai dati.

L'appena descritto *normed fit index* presenta il problema che la media della sua distribuzione campionaria è positivamente correlata con la numerosità campionaria (quindi l'indice tende a fornire un'immagine troppo pessimistica di adattamento nel caso di piccoli campioni, anche se in realtà l'adattamento è buono). Comunque non esiste ancora nessun aggiustamento per i gradi di libertà che possa risolvere questo problema.

Per avere un modello valido con variabili osservate che non presentino un'eccessiva curtosi, lo stimatore chi-quadro di χ_h^2 segue una distribuzione asintotica appunto chi-quadro. La media di un chi-quadro sono i suoi gradi di libertà, quindi per grandi campioni la media di χ_h^2 è approssimativamente df_h. Per un modello correttamente stimato, il numeratore di Δ_1 è in media $\left(\overline{\chi}_b^2 - df_h\right)$, dove $\overline{\chi}_b^2$ è il valore medio del chi-quadrato relativo al modello base. Se in media $\left(\overline{\chi}_b^2 - df_h\right)$ è quanto ci si aspetta per un modello corretto, questo potrebbe essere il denominatore con cui confrontare $\left(\chi_b^2 - \chi_h^2\right)$, ottenendo:

$\Delta_2 = \frac{\chi_b^2 - \chi_h^2}{\chi_b^2 - df_h}$, dove χ_b^2 è l'unico stimatore di $\overline{\chi}_b^2$ disponibile. In questo modo, i modelli con un minor numero di parametri hanno valori di Δ_2 più alti, mantenendo le altre quantità invariate; inoltre la numerosità campionaria viene presa in considerazione, infatti riscrivendo l'indice in termini dei valori delle funzioni di *fitting* si ha che:

$\Delta_2 = \frac{F_b - F_h}{F_b - \left(df_h / (N-1)\right)}$. Al diminuire di N aumenta Δ_2. Non esiste un range fisso per questo indice. Valori molto più bassi di 1 indicano un adattamento scadente, mentre valori di molto superiori a 1 potrebbero indicare, anche se non con certezza, un *overfitting*. La media della distribuzione campionaria di Δ_2 pare essere non correlata ad N.

Tutte le misure di cui sopra sono pesantemente dipendenti dal modello base utilizzato. In più, si è riscontrato che esse sono suscettibili anche all'influenza dei metodi di stima[59].

[58] Bollen "A new incremental fit index for GSEM"

Nel caso si desideri confrontare modelli non annidati e che comprendano insiemi diversi di predittori, bisogna ricorrere ad altri indici che non sono basati su un confronto tra modelli, come quelli nel seguito riportati.
Jöreskog e Sörbom (1986) hanno introdotto altre due misure di bontà di adattamento[60]:

$$GFI = 1 - tr\left[\left(\hat{\Sigma}^{-1}S - I\right)^2\right] / tr\left[\left(\hat{\Sigma}^{-1}S\right)^2\right]$$

$$AGFI = 1 - \left[q(q+1)/(2df_h)\right](1 - GFI)$$

dove S è la matrice campionaria di varianze e covarianze delle (q) variabili osservate, $\hat{\Sigma}$ è la matrice di varianze e covarianze *sample implied*, mentre I è la matrice identità. GFI ha un massimo normato di 1, ma può assumere valori negativi; anche AGFI ha un massimo normato di 1, ma tiene anche conto dei gradi di libertà. Date S e $\hat{\Sigma}$, N non ha influenza su questi indici, in termini calcolatori; tuttavia alcune simulazioni hanno mostrato che le medie delle distribuzioni campionarie di ambo gli indici sono positivamente correlate con N.
Hoelther (1983) ha implementato il *Critical N*:

$CN = \left(crit\chi^2\right) \Big/ F_h + 1$ dove $\left(crit\chi^2\right)$ è il valore critico di chi-quadro con df_h gradi di libertà e ad un livello alpha prefissato. Non ha un range determinato, come anche non subisce l'influenza di N, ma le medie delle distribuzioni campionarie di CN sembrano positivamente correlate con la numerosità campionaria N.
Radice dell'errore medio di approssimazione (*root mean square error of approximation*):

$$RMSEA = \sqrt{\frac{T_h - df_h}{(N-1)df_h}}.$$

Questa statistica test è asintoticamente distribuita come un chi-quadrato non centrale (con parametro di non centralità λ, valore atteso *df*+ λ, varianza *2df+4* λ)[61], quando l'ipotesi nulla non è rifiutata, ma quando anche l'errore non è troppo grande; il parametro di non centralità λ è una misura del grado di errore di specificazione del modello ipotizzato. Sotto le assunzioni di numerosità campionaria elevata, assenza di eccesso di curtosi e non impropria specificazione del modello, T_h segue una distribuzione chi-quadrato centrale, con *df* come valore atteso e *2df* come varianza. Il numeratore sotto radice è asintoticamente uno stimatore corretto del parametro di non centralità per il χ^2 non centrale. I suoi valori sono >0, ma se l'adattamento è buono, il valore è molto vicino a 0. È possibile calcolare il suo intervallo di confidenza: $CI = \left(\sqrt{\frac{\hat{\lambda}_L}{df\ (N-1)}}; \sqrt{\frac{\hat{\lambda}_U}{df\ (N-1)}}\right)$ dove $\hat{\lambda}_L$ e $\hat{\lambda}_U$ sono i limiti inferiore e superiore del chi-quadrato non centrale considerato. La distribuzione asintotica è verificata data la non eccessiva curtosi della distribuzione della variabile obiettivo, la numerosità campionaria sufficientemente elevata e l'errore di approssimazione non eccessivamente alto rispetto all'errore di stima. Utilizzando sia la stima dell'indice che il suo intervallo di confidenza, è possibile effettuare un qualsiasi test d'ipotesi classico. Il primo è il test di adattamento perfetto: se l'ipotesi nulla è $H_0 : \varepsilon = 0$, dove ε è il valore

[59] Chen et al. "An empirical evaluation of the use of fixed cut-off points in RMSEA test statistic in SEM"
[60] Bollen "Overall fit in covariance structure models: two types of sample size effects"
[61] Curran, Bollen et al. "Finite sampling properties of RMSEA"

di RMSEA in popolazione, l'ipotesi nulla è rifiutata se l'estremo inferiore dell'intervallo di confidenza è >0. Si tratta in effetti di un test χ^2. Il secondo test è quello di adattamento stretto: l'ipotesi nulla è $H_0 : \varepsilon \leq c$, dove *c* è una costante arbitraria. L'ipotesi nulla è rifiutata se la statistica test è maggiore di un valore di *cut-off c* che definisce un'area nella coda superiore della distribuzione chi-quadrato. L'accettazione dell'ipotesi nulla comporta il mantenimento del modello proposto. Questo test è maggiormente realistico del precedente. La scelta del punto di *cut-off* è fondamentale per la valutazione del modello attraverso l'indice RMSEA. Browne e Cudeck (1993) hanno raccomandato il valore ottimale dell'indice RMSEA di 0,05 o inferiore per avere un adattamento stretto del modello; mentre un valore di 0,08 o inferiore rappresenta un errore di approssimazione ragionevole; se il valore di RMSEA è superiore a 0,1, è ragionevole pensare che il modello non sia in grado di rappresentare i dati osservati. Il punto di *cut-off* uguale a 0,05 è stato adottato come "*gold standard*".

Browne e Cudeck (1993) hanno effettuato la distinzione tra due tipi di errori[62]: gli errori di stima e gli errori di approssimazione. Data Σ_0 come matrice di varianze e covarianze in popolazione, $\hat{\Sigma}_0$ come matrice di varianze e covarianze in popolazione stimata dal modello ritenuto più adatto e data $\hat{\Sigma}$ come matrice di varianze e covarianze nel campione (S) stimata dal modello ritenuto più adatto, l'errore di approssimazione è il grado di non adattamento tra la matrice di varianze e covarianze in popolazione Σ_0 e la matrice di varianze e covarianze in popolazione stimata dal modello (*model implied*) $\hat{\Sigma}_0$, mentre l'errore di stima è il grado di non adattamento tra la matrice di varianze e covarianze campionaria stimata dal modello $\hat{\Sigma}$ e la matrice di varianze e covarianze in popolazione stimata dal modello $\hat{\Sigma}_0$. Il grado del primo errore di adattamento viene stimato come funzione della discrepanza dovuta all'approssimazione: $F_0 = \left(\Sigma_0, \hat{\Sigma}_0\right)$. Il grado del secondo errore di adattamento viene stimato come funzione: $F_0 = \left(\hat{\Sigma}_0, \hat{\Sigma}\right)$. L'errore di approssimazione è d'interesse chiave nella valutazione del modello: così ritennero Steiger e Lind (1980) e Browne e Cudeck (1993), che definirono l'indice RMSEA in popolazione come $\varepsilon = \sqrt{\frac{F_0}{df}}$. Qui F_0 è una somma adeguatamente pesata di scarti quadratici tra la matrice di varianze e covarianze in popolazione Σ_0 e la matrice di varianze e covarianze stimata con il modello ottimo in termini di adattamento; l'aggiunta di ulteriori gradi di libertà viene operata al fine di tener conto della complessità del modello. Tuttavia questo valore di RMSEA non è praticamente calcolabile in quanto F_0 è un valore in popolazione: deve perciò essere stimato con il campione. La stima $\hat{F}$, la funzione di discrepanza, è uno stimatore distorto di F_0, quindi lo stimatore corretto dell'errore di approssimazione è $\hat{F}_0 = \hat{F} - \frac{df}{N-1}$.

Questo va a correggere anche l'indice RMSEA= $\sqrt{\frac{\hat{F}_0}{df}} = \sqrt{\frac{\hat{F}(N-1)-df}{df(N-1)}} = \sqrt{\frac{T-df}{df(N-1)}} = \sqrt{\frac{\hat{\lambda}}{df(N-1)}}$, stima campionaria di ε.

Alcuni studi empirici hanno mostrato che, in media, i valori di RMSEA vengono sovrastimati con numerosità campionarie non elevate e per modelli specificati propriamente, mentre la distorsione diminuisce all'aumentare

[62] Curran, Bollen et al. "Finite sampling properties of RMSEA"

della numerosità campionaria, anche se si verifica un aumento nell'errore di specificazione del modello. Inoltre, il suddetto indice è meno sensibile a un errore di specificazione inferiore mentre è più sensibile a errori di specificazione del modello importanti. Ancora, è stato mostrato che i valori di RMSEA siano gonfiati in presenza di scarsa numerosità campionaria, mentre che non mantengano tale comportamento con numerosità campionarie elevate[63].

Altri indici che non richiedono il confronto con un modello base, basati sempre sulla funzione di log verosimiglianza con qualche penalizzazione (a seconda del numero di parametri liberi, che aumentano la funzione LL, e del numero di istanti temporali):

$$AIC = T_h - 2u \quad \text{(Akaike)}$$

$$BIC = T_h - u\ln(T) \quad \text{(Schwarz)}$$

dove *u* è il numero di parametri liberi del modello e T è il numero di istanti temporali considerati. Più sono bassi i valori, migliore è il modello, ma per piccole differenze nei valori degli indici per modelli diversi, la decisione non è scontata. Raftery mostrò che differenze dell'ordine di 0-2 si possono considerare piccole, mentre differenze dell'ordine di 6-10 sono da ritenersi elevate.

Se si ha la necessità di confrontare modelli che non sono annidati l'un l'altro, per esempio modelli con insiemi di predittori diversi, meglio usare gli indici AIC e BIC.

3.5.3 ADATTAMENTO PER SINGOLA COMPONENTE

Esiste un'intera classe di test statistici che si pongono l'obiettivo di valutare parti del modello invece di questo stesso nella sua globalità. Si tratta di un altro modo per stimare la corrispondenza del modello ai dati osservati. Questi test[64] sono disponibili sia utilizzando l'approccio per singolo caso che uno stimatore ML globale.

In questo caso l'obiettivo è quello di esaminare le stime dei parametri e i loro errori, in particolare è quello di ricercare possibili soluzioni improprie, ad esempio i casi di varianza degli errori negativa. Un primo modo per trattare quest'ultimo caso è quello di costruire un test di significatività (*z*-test) per determinare se il valore negativo è soltanto frutto di una fluttuazione da 0 dovuta all'errore di campionamento. La statistica test, in questo caso, si costruisce come rapporto tra la stima del singolo parametro ed il suo errore standard asintotico; la distribuzione della statistica test è la normale standardizzata.

Un altro test di significatività è quello di Wald: il rapporto tra la stima del parametro e il suo *standard error* asintotico viene elevato al quadrato e la statistica test risultante si confronta con la distribuzione chi-quadrato con 1 grado di libertà. Il problema di ambedue tali test è che l'errore standard asintotico potrebbe non essere una buona stima per l'errore standard delle stime. Il loro vantaggio è invece la possibilità di testare anche alcune ipotesi composte (per esempio, $H_0 : \gamma_{00} = 0$ e $\gamma_{10} = 0$) su effetti multipli, indipendentemente dal metodo di

[63] Curran, Bollen et al. "Finite sampling properties of RMSEA"
[64] Curran "Latent Curve Models - A structural equation perspective"

stima utilizzato. Si noti che l'ipotesi nell'esempio potrebbe anche essere formulata come segue: $H_0 : 1\gamma_{00} + 0\gamma_{11} + 0\gamma_{10} = 0$ e $0\gamma_{00} + 0\gamma_{11} + 1\gamma_{10} = 0$, con $C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ e $\gamma' = \begin{pmatrix} \gamma_{00} \\ \gamma_{11} \\ \gamma_{10} \end{pmatrix}$. Quindi $H_0 : C\gamma' = 0$.

Un vantaggio del test di Wald è che può essere generalizzato nel caso di test simultanei di diverse varianze degli errori. Utilizzando tale test, diventa inoltre possibile determinare se varianze dell'errore negative siano dovute alle fluttuazioni campionarie. La statistica di Wald mette a confronto il quadrato di una combinazione lineare pesata dei parametri con la stima della sua varianza. Sotto l'ipotesi nulla, la statistica di Wald ha una distribuzione χ^2 con gradi di libertà uguali al numero di righe della matrice C (che è il numero di vincoli indipendenti).

Il problema delle soluzioni improprie si verifica maggiormente nei piccoli campioni che non nei grandi[65]; ciò è dovuto al fatto che nei piccoli campioni vi sono molte più fluttuazioni che non nei grandi campioni. La relazione tra l'errata specificazione del modello e la presenza di soluzioni improprie non è diretta; la combinazione dell'errata specificazione dei parametri di popolazione, della deviazione standard e della distribuzione della varianza degli errori determina la probabilità di trovare varianze degli errori negativa. Così il numero di soluzioni improprie non è un buon indicatore dell'errore di specificazione del modello. Inoltre, la presenza di soluzioni improprie tende a significare che la distorsione delle stime dei parametri del modello sarà maggiore che non in campioni che presentano solo soluzioni proprie.

La presenza di varianze negative porta a stime ML improprie. Una volta esaminati i valori inferiori a zero, è importante valutare la significatività statistica delle varianze e delle covarianze. Le varianze sono quelle di intercetta ($\hat{\psi}_{\alpha\alpha}$) e pendenza ($\hat{\psi}_{\beta\beta}$). Costruendo i loro intervalli di confidenza è possibile svelare se i loro *range* includano il valore zero. Se questo è incluso nell'intervallo, ciò vuol dire che gli individui non hanno differenze significative in quanto a intercetta e/o a pendenza rispetto alla traiettoria media. Anche per quanto riguarda la singola varianza dell'errore si può costruire l'intervallo di confidenza. Dopo aver identificato i valori impropri (negativi), è interessante vedere quali sono le implicazioni, in termini di $R^2_{y_t}$ della $\text{var}(\varepsilon_t)$,con *t=1,..T*.

$R^2_{y_t}$ stesso, invece, misura quanto le variazioni della variabile obiettivo possano essere spiegate dal modello: valori alti indicano un buon adattamento del modello. Nel modello globale, i valori di R^2 differiscono nei diversi istanti temporali, ma rimangono gli stessi per tutti gli individui in un dato tempo; nel modello per singolo caso i valori di R^2 sono differenti a seconda degli individui, ma non nei diversi istanti temporali con riferimento ad ogni singolo individuo.

Il metodo dei momenti residui, altro utile strumento per valutare la bontà di adattamento per singola componente (*component fit*), è basato sulla stima ML delle medie, delle varianze e covarianze delle variabili osservate. Data $\bar{y}$, la media campionaria osservata, e $\mu(\vartheta)$, la media del modello (dipende dalle medie stimate delle variabili latenti), il vettore della media residua è: $\bar{y} - \mu(\vartheta)$. I valori che si trovano in questo vettore nei diversi istanti

[65] Bollen, Chen et al. "Improper solutions in structural equation models"

temporali aiutano a capire se il modello ha sovrastimato oppure sottostimato le variabili osservate. La matrice delle varianze e covarianze residue $S - \Sigma(\vartheta)$ mostra se il modello sovrastima o sottostima i momenti del second'ordine. Un utile grafico, che permette di vedere quanto il modello sia in grado di stimare le varianze, è quello che contiene i valori delle varianze osservate (ordinati in senso crescente secondo il tempo) plottati sui valori delle varianze *model implied* (anch'essi nello stesso ordine). Un altro strumento è il grafico delle covarianze stimate su quelle osservate: una retta con l'inclinazione di 45° indica una corrispondenza perfetta tra le due covarianze.

Un'interessante e intuitiva domanda sui dati è se le varianze siano realmente differenti nei diversi istanti temporali. Si può testare, con una statistica test di tipo chi-quadrato, se le differenze osservate tra le varianze nel tempo siano significative. Per fare ciò, occorre stimare un modello dove tutte le varianze sono vincolate ad un medesimo valore (il modello T_r con df_r gradi di libertà) ed anche un altro modello dove invece le varianze sono stimate in modo libero (il modello T_u con df_u gradi di libertà, dove la differenza tra *r* e *u* è il numero di varianze); quindi si può costruire la statistica test con la differenza tra i due modelli, che è distribuita come un chi-quadrato con *u-r* gradi di libertà.

3.5.4 VALUTAZIONE GRAFICA DELLE ASSUNZIONI DEL MODELLO

Una volta che il modello è stato specificato, il ricercatore pone alcune assunzioni sulla popolazione. Tuttavia tutto ciò che egli può testare è in realtà il comportamento del campione. Una semplice modalità per esaminare la forma funzionale è quella di osservare, graficamente, il comportamento della variabile obiettivo su quello dei regressori. Al primo livello, è utile usare gli *empirical growth plots* individuali con le traiettorie individuali stimate con il metodo OLS. Al secondo livello, è invece utile plottare le stime OLS dei parametri di crescita individuali su ogni variabile esogena di secondo livello. Per capire se la normalità distributiva può essere ipotizzata, le stime degli errori, $\hat{\varepsilon}_{ij}$ $\hat{\zeta}_{0i}$ e $\hat{\zeta}_{1i}$ chiamati errori di riga (*raw residuals*), possono essere visualizzati in un grafico sui relativi punteggi nel caso normale (*normal probability plot*), come esemplificato in Grafico 38; se la distribuzione è una linea, si può dare per accertata la normalità distributiva.

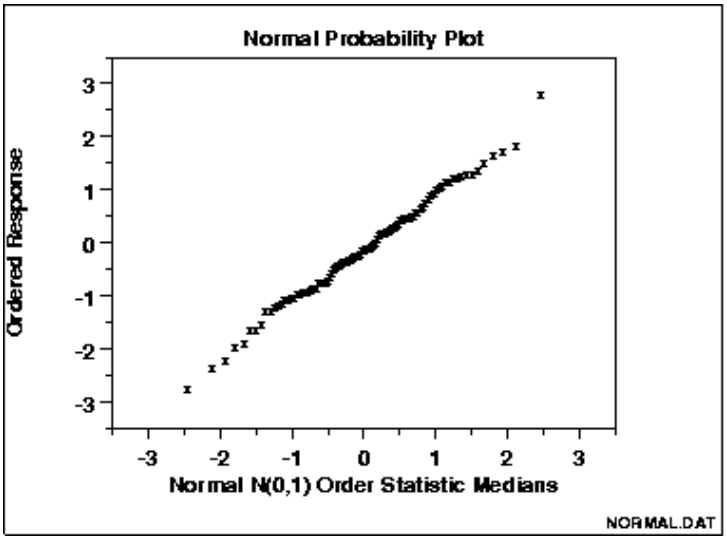


Grafico 38 – Esempio di *normal probability plot*

Al medesimo scopo, si possono inserire in un grafico anche gli errori di riga standardizzati: se approssimativamente il 95% di essi rimane all'interno di ± 2 volte la deviazione standard del loro centro, si può ritenere corretta l'assunzione di normalità distributiva. Gli stessi errori standardizzati, ordinati semplicemente per identificativo, possono anche entrare in un altro grafico che aiuta a mettere in evidenza eventuali *outliers*. Il grafico dei *raw residuals* sui regressori (ogni errore di riga su ciascun regressore di primo o di secondo livello) permette di valutare la presenza di omoschedasticità: se tale assunzione è verificata, la variabilità degli errori sarà approssimativamente uguale per diversi valori di ciascun regressore.

3.6 LCM - LA STRUTTURA DELLE COVARIANZE DEGLI ERRORI

Può essere utile, in alcuni casi, focalizzare l'attenzione sugli effetti casuali (*random effects*) del modello[66] per descrivere la struttura delle covarianze degli errori. Partendo dal modello di primo livello standard $y_{it} = \alpha_i + \beta_i \lambda_t + \varepsilon_{it}$, dove i dati sono *time structured* (tutti gli individui hanno la stessa struttura degli istanti temporali) e dove i valori del tempo sono indicati con 0,1,2,..,T-1, mentre gli effetti casuali, i.i.d. (indipendenti lungo l'asse dei tempi ed anche per i diversi individui ed anche identicamente distribuiti), provengono da una distribuzione normale univariata $\varepsilon_{it} \sim N\left(0, \sigma_\varepsilon^2\right)$. Per ogni parametro viene specificato un diverso modello di secondo livello, che tiene conto di una variabile esogena indipendente dal tempo. Così:

$\alpha_i = \gamma_{00} + \gamma_{01} X_i + \zeta_{0i}$

$\beta_i = \gamma_{10} + \gamma_{11} X_i + \zeta_{1i}$

dove $\begin{pmatrix} \zeta_{0i} \\ \zeta_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{pmatrix} \right)$ e $\begin{pmatrix} \zeta_{0i} \\ \zeta_{1i} \end{pmatrix}$ sono i.i.d.

Nel modello di secondo livello, gli effetti fissi catturano gli effetti della variabile esogena sulle traiettorie medie, mentre gli effetti casuali rappresentano quella parte di valori dei parametri che è spiegata dal modello. La differenza tra considerare separatamente gli effetti casuali di secondo livello (come se ognuno di essi provenisse da una distribuzione univariata) ed invece supporre la loro distribuzione bivariata sta nel fatto che, con quest'ultima modalità, è possibile tener conto del legame tra l'intercetta vera e la vera pendenza. Focalizzando l'attenzione sugli effetti casuali, un metodo utile per la stima è RML.

La specificazione compatta del modello è:

$y_{it} = \gamma_{00} + \gamma_{01} X_i + \zeta_{0i} + (\gamma_{10} + \gamma_{11} X_i + \zeta_{1i})\lambda_t + \varepsilon_{it}$ che diventa

$y_{it} = \gamma_{00} + \gamma_{10} \lambda_t + \gamma_{01} X_i + \gamma_{11} X_i \lambda_t + (\zeta_{0i} + \zeta_{1i} \lambda_t + \varepsilon_{it})$.

Il termine dentro le parentesi è la parte stocastica, chiamata r_{it}; il resto del polinomio rappresenta l'ipotesi fatta sulla dipendenza della variabile obiettivo dal tempo e dalla variabile esogena. r_{it} può essere vista come combinazione lineare degli effetti casuali. Il modello nella notazione compatta, una volta sostituita la parte stocastica con r_{it}, può essere riguardato come un modello di regressione standard, con due variabili esogene e con un termine di interazione. È possibile stimare il modello con un'analisi di regressione di tipo GLS, facendo assunzioni specifiche sulla distribuzione degli errori. Per quanto riguarda l'analisi OLS standard, l'ipotesi

[66] Singer, Willett "Applied longitudinal data analysis – modelling change and event occurrence"

dovrebbe essere che tutti i termini di errore (TxN elementi) siano distribuiti come una distribuzione normale multivariata, con un vettore delle medie nullo e una matrice di varianze e covarianze diagonale (la matrice delle covarianze degli errori) di σ_r^2. Così ogni r_{it} è una variabile casuale normale e le r_{it} sono tra loro indipendenti. Questa ipotesi è troppo restrittiva per dati di tipo longitudinale: in effetti, se l'ipotesi di indipendenza tra un individuo e l'altro è ragionevole, l'ipotesi di indipendenza e omoschedasticità nei diversi istanti di tempo non lo è proprio. Una matrice di varianze e covarianze più accettabile è quella di tipo diagonale a blocchi, dove ogni blocco corrisponde ad un individuo nei diversi istanti temporali considerati. Il generico blocco della matrice è il seguente: $\Sigma_r = \begin{pmatrix} \sigma_{r_1}^2 & . & \sigma_{r_T r_1} \\ . & . & . \\ \sigma_{r_1 r_T} & . & \sigma_{r_T}^2 \end{pmatrix}$, una matrice TxT delle varianze e covarianze nel tempo, uguale per tutti gli individui. Ogni individuo avrà errori indipendentemente dagli altri individui. Andare ad indagare sugli effetti casuali significa che i *composite residuals* hanno una distribuzione multivariata come quella già specificata e l'analisi sta anche nella stima dei valori di questa matrice. La distribuzione degli r_{it} è dovuta alla specificazione degli r_{it} stessi: siccome sono combinazioni lineari di variabili casuali normali, anch'essi sono variabili casuali normali; quindi le loro medie sono vettori nulli grazie alla struttura della combinazione lineare stessa. Gli elementi di Σ_r sono dipendenti dal tempo.

Il generico elemento della sua diagonale è: $\sigma_{r_t}^2 = \text{var}(\zeta_{0i} + \zeta_{1i}\lambda_t + \varepsilon_{it}) = \sigma_0^2 + \sigma_1^2\lambda_t^2 + 2\sigma_{01}\lambda_t + \sigma_\varepsilon^2$.

La sua stima può essere ottenuta sostituendo le stime delle varianze degli errori e i valori del tempo nella stessa espressione. Le stime possono mostrare se l'ipotesi di eteroschedasticità sia vera (attraverso un confronto statistico). La generica covarianza diventa: $\sigma_{r_t r_{t'}} = \sigma_0^2 + \sigma_{01}(\lambda_t + \lambda_{t'}) + \sigma_1^2\lambda_t\lambda_{t'}$ e la sua stima è ottenuta sostituendo gli elementi con le stime delle varianze e delle covarianze e con i valori del tempo. Si noti che la grandezza della covarianza dipende anche dai valori del tempo, visto che c'è un termine che dipende dal prodotto di due tempi diversi. Se tutte le componenti di errore fossero prossime a zero, anche le covarianze sarebbero zero, quindi la matrice Σ_r sarebbe diagonale e gli errori omoschedastici. Se soltanto la varianza di secondo livello σ_1^2 e le covarianze fossero molto piccole, tutti gli elementi di covarianza in Σ_r sarebbero uguali, di valore σ_0^2, e l'omoschedasticità sarebbe ancora verificata, con tutti gli elementi di varianza uguali a $\sigma_0^2 + \sigma_\varepsilon^2$.

Le ipotesi sulla struttura degli errori delle covarianze conduce a costruire modelli diversi. Disponendo di un dataset relativo al fenomeno oggetto di studio, è necessario ipotizzare diversi modelli e poi testare la loro bontà di adattamento ai dati, per poi scegliere il modello migliore. Per testare la performance dei diversi modelli si possono utilizzare diversi metodi, come verrà nel seguito descritto. Risulta possibile specificare modelli con effetti fissi identici, ma diversa struttura degli errori delle covarianze. Certamente in questa fase aiuta avere a disposizione informazioni a priori sui dati, così da definire un insieme iniziale di modelli per cominciare la procedura di selezione del modello ottimo. Durante la fase di confronto, è forse meglio usare il metodo di stima *restricted*, in modo che la statistica di bontà di adattamento ottenuta rifletta soltanto la parte stocastica del modello.

Una prima scelta per la struttura della matrice degli errori delle covarianze è quella con una matrice **non strutturata** (*unstructured*): Σ_r ha valori tutti diversi, quindi ci sono T parametri incogniti, uno per ciascuna varianza, più le varianze incognite ((T-1)+(T-2)+…+(1)). In questo tipo di matrice non vi sono vincoli; la *deviance statistic*, dati gli stessi effetti fissi, sarà sempre la più piccola. Se T è grande, il numero di parametri cresce troppo. Se T non è molto grande, questa ipotesi può diventare il punto di partenza per l'analisi esplorativa. Un'altra possibilità è la scelta della matrice Σ_r **simmetrica composta** (*compound symmetric*): tutte le varianze sono uguali a $\sigma^2 + \sigma_1^2$, quindi c'è omoschedasticità tra un istante temporale e l'altro, mentre le covarianze sono costanti, σ_1^2. In questo caso, la variazione residua della vera intercetta delle traiettorie di crescita è prossima a zero, e così anche la covarianza residua; quindi tale scelta risulta buona quando i dati sembrano mostrare pendenze con una piccola o anche nulla varianza residua. La matrice **simmetrica eterogenea** (*heterogeneous compound symmetric*) porta invece a eteroschedasticità nella varianza degli errori, quindi le varianze, cioè gli elementi sulla sua diagonale principale, sono σ_1^2, σ_2^2, …, σ_T^2; le covarianze sono diverse per ogni coppia di errori, ma tutti sono il prodotto delle corrispondenti deviazioni standard ($\sigma_{r_i r_j} = \sigma_{r_i}\sigma_{r_j}\rho$ dove ρ è una parametro di autocorrelazione costante, di solito posto ≤ 1). Il numero dei parametri da stimare è uguale al numero degli istanti temporali considerati più 1. Un altro tipo di matrice Σ_r ipotizzabile è quella **autoregressiva del prim'ordine** (*first order autoregressive*). L'assunzione, in questo caso, è di omoschedasticità degli errori, quindi tutti gli elementi sulla diagonale principale sono uguali a σ^2; inoltre, coppie di errori diverse hanno uguali covarianze, se la loro covarianza si trova sulla stessa banda parallela alla diagonale principale della matrice e il loro valore di covarianza è il prodotto di σ^2 e di un coefficiente ρ (≤ 1) (quindi sono una frazione degli elementi sulla diagonale principale), moltiplicato per se stesso un numero di volte uguale alla sua posizione rispetto alla diagonale principale (così gli elementi sulla prima banda immediatamente sotto la diagonale principale sono moltiplicati per ρ, gli elementi sulla diagonale ancora sotto sono moltiplicati per ρ^2 e così via). Questo fatto porta a valori decrescenti man mano che ci si allontana dalla diagonale principale. In questo tipo di modello vi sono pochi parametri da stimare. Una matrice delle covarianze degli errori del tutto simile a quella appena descritta, ma con meno vincoli, è la matrice **autoregressiva eterogenea** (*heterogeneous autoregressive*), dove i valori sulla diagonale principale sono diversi tra loro. Al di fuori della diagonale principale, gli elementi sono tutti moltiplicati per il coefficiente ρ, sempre ≤ 1 (ancora moltiplicato per se stesso un numero di volte uguale alla sua posizione rispetto alla diagonale principale, in modo che la proprietà di valori decrescenti via via che ci si allontana dalla diagonale principale sia preservata), ma vengono aggiunti più elementi di varianza; l'elemento generico al di fuori della diagonale principale è $\sigma_{r_i r_j} = \sigma_{r_i}\sigma_{r_j}\rho^{i+j-2}$. Questa assunzione è più costosa in termini di gradi di libertà, ma è anche maggiormente flessibile. Rispetto alla matrice autoregressiva prima descritta, in questo caso la *deviance statistic* è in generale migliore. La matrice delle covarianze degli errori chiamata **Toeplitz** è invece simile alla matrice autoregressiva del prim'ordine, in quanto le bande parallele alla diagonale principale hanno tutte gli stessi valori, tuttavia questi ultimi non sono una funzione degli elementi

della diagonale principale. Quindi tutte le varianze sono uguali a σ^2, mentre le covarianze sono uguali a σ_1 nella banda immediatamente vicina alla diagonale principale, σ_2 nella diagonale successive e così via, fino a σ_{T-1}. Il numero di parametri da stimare è T. Questo tipo di assunzione è più costosa e più flessibile di quella autoregressiva, ma è più parsimoniosa e meno flessibile di quella autoregressiva eterogenea.

La cosa più importante di cui tener conto ai fini del confronto di modelli differenti, in termini di matrice di errori delle covarianze, è l'utilizzo di più di un test di bontà di adattamento. È veramente importante guardare la *deviance statistic* come anche gli indici come AIC e BIC. L'indicatore di bontà di adattamento ottenuto utilizzando la matrice non strutturata è ovviamente quello migliore, in quanto negli altri casi vengono posti vincoli, ma potrebbe rivelarsi una scelta migliore quella a favore di un modello che si adatti in modo un po' peggiore ai dati ma che abbia un numero inferiore di parametri liberi. Specie quando si ha a che fare con lunghe serie di dati, è meglio utilizzare un modello più parsimonioso.

L'analisi della struttura delle covarianze può aiutare anche a saggiare la bontà di adattamento del LCM, quindi nella proposta iniziale nonché nella scelta del modello migliore. La traiettoria di cambiamento individuale di primo livello viene considerata come il modello *Y-measurement* (quel modello che specifica le relazioni tra le variabili osservate e i fattori latenti); le differenze individuali di secondo livello vengono trattate come nel modello strutturale (dove esistono relazioni tra le variabili latenti); l'aggiunta di variabili esogene costanti nel tempo è il corrispondente del modello *X-measurement* (dove le variabili esogene sono spiegate in termini di media in popolazione e di scarti dalla media stessa). Per prima cosa si ha bisogno della matrice di varianze e covarianze e del vettore delle medie del campione, dove gli input sono le misure ripetute della variabile risultato e delle variabili esogene. Dato il modello lineare standard, $Y_{it} = \alpha_i + \beta_i \lambda_t + \varepsilon_{it}$, questo può essere espresso in forma matriciale come $\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ . \\ Y_{iT} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ . \\ 0 \end{pmatrix} + \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ . & . \\ 1 & t_T \end{pmatrix} \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ . \\ \varepsilon_{iT} \end{pmatrix}$. Questa forma è molto simile al modello *Y-measurement*, che è descritto dall'equazione: $\overline{Y} = \overline{\tau}_y + \Lambda_y \overline{\eta} + \overline{\varepsilon}$, dove $\overline{Y}$ è la variabile risultato, $\overline{\eta}$ è la variabile latente, Λ_y è il fattore di scala, $\overline{\tau}_y$ è la media in popolazione della variabile risultato, mentre $\overline{\varepsilon}$ è l'errore di misurazione. Questo modello ipotizza che il vettore degli errori $\overline{\varepsilon}$ provenga da una distribuzione normale multivariata (con un vettore delle medie nullo e una matrice di varianze e covarianze $\Theta_\varepsilon = \text{cov}(\varepsilon_{i1},...,\varepsilon_{iT})$). In Θ_ε i parametri di varianza della diagonale principale stimano la variabilità di ciascun errore, mentre i parametri di covarianza stimano il grado di associazione tra gli errori. La forma di questa matrice riflette le assunzioni del modello: omoschedasticità ed autocorrelazione. L'equazione strutturale è $\overline{\eta} = \overline{\tau}_\eta + \Gamma \overline{\xi} + B\overline{\eta} + \overline{\zeta}$, quindi le variabili latenti sono spiegate come scostamenti dalla media in popolazione $\overline{\tau}_\eta$, spiegati da variabili esogene e da un elemento di errore; il termine $\Gamma \overline{\xi}$ esplica invece la relazione tra i parametri latenti e i predittori, mentre $B\overline{\eta}$ rappresenta le relazioni tra gli stessi parametri (si noti

che gli elementi sulla diagonale principale di B sono ovviamente uguali a zero, visto che è impossibile la previsione di un costrutto endogeno da parte dello stesso). La scelta dei valori interni alle matrici Γ e B è alla base della struttura del modello. Il vettore degli errori è posto provenire da una distribuzione normale multivariata, con vettore delle medie nullo e matrice di varianze e covarianze $\Psi = \text{cov}(\zeta_{1i,}, \zeta_{2i})$; il numero dei valori di errore è uguale al numero dei coefficienti del modello (2 nel caso lineare). Gli elementi sulla diagonale principale di Ψ sono le varianze degli errori (variabilità entro), mentre gli elementi al di fuori della diagonale principale sono le covarianze tra gli errori (variabilità tra). I metodi dell'analisi basata sulla struttura delle covarianze possono essere usati per stimare i parametri di crescita e le matrici delle covarianze degli errori. Se un primo sguardo alle stime suggerisce che vi sia una qualche eterogeneità nelle traiettorie di crescita spiegabile da qualche variabile esogena, è possibile incorporare tale variabile nel modello di secondo livello. Il regressore entra nel modello *X-measurement*: $X = \tau_x + \Lambda_x \xi + \delta$. Se il regressore è soltanto uno, X deve contenere un singolo elemento, l'errore di misurazione δ deve contenere come singolo elemento zero (il predittore è misurato senza errore), anche τ_x deve contenere come singolo elemento zero, mentre Λ_x deve contenere come singolo elemento 1. Si noti che $E\{X_i\} = E\{\tau_x + \lambda^x \xi_i + \delta_i\}$ quindi $\mu_X = \tau_x + \lambda^x \mu_\xi$; di solito la media della variabile esogena è posta uguale ad uno solo di quei due elementi. Se la media è forzata essere uguale a τ_x, l'elemento μ_ξ, media della variabile latente, è zero e il regressore sarà centrato sul suo valore medio; se invece la media è forzata uguale a $\lambda^x \mu_\xi$, il regressore non sarà centrato. Una volta specificato il modello *X-measurement*, è possibile usare l'analisi della struttura delle covarianze per specificare le relazioni tra i parametri di crescita e il regressore.

Un'interessante analisi della relazione tra la variabile obiettivo e una variabile esogena che varia nel tempo è la ricerca dell'associazione tra i cambiamenti della prima e quelli del secondo. È necessario in questo caso specificare il modello di crescita individuale per la variabile obiettivo e per il regressore. Dato il consueto modello $Y_{it} = \alpha_i + \beta_i \lambda_t + \varepsilon_{it}$ (ipotesi lineare), deve anche essere specificato l'ulteriore modello (sotto un'altra ipotesi lineare) $X_{it} = \alpha_i' + \beta_i' \lambda_t + \delta_{it}$, dove l'interpretazione dei parametri è analoga al precedente: α_i' è il vero punto di partenza individuale per il regressore, mentre β_i' è il suo vero tasso di cambiamento individuale. La specificazione di primo livello è: $\begin{pmatrix} X_{i1} \\ . \\ . \\ X_{iT} \end{pmatrix} = \begin{pmatrix} 0 \\ . \\ . \\ 0 \end{pmatrix} + \begin{pmatrix} 1 & t_1 \\ . & t_2 \\ . & . \\ 1 & t_T \end{pmatrix} \begin{pmatrix} \alpha_i' \\ \beta_i' \end{pmatrix} + \begin{pmatrix} \delta_{i1} \\ . \\ . \\ \delta_{iT} \end{pmatrix}$ oppure $\overline{X} = \overline{\tau}_x + \Lambda_x \overline{\xi} + \overline{\delta}$, dove $\overline{\xi}$ è il vettore dei parametri latenti del regressore esogeno. La media in popolazione è $k = \begin{pmatrix} \mu_{\alpha_i'} \\ \mu_{\beta_i'} \end{pmatrix}$ mentre la matrice di varianze e covarianze in popolazione è

$\Phi = \begin{pmatrix} \sigma^2_{\alpha'_i} & \sigma_{\alpha'_i\ \beta'_i} \\ \sigma_{\alpha'_i\ \beta'_i} & \sigma^2_{\beta'_i} \end{pmatrix}$. È necessario specificare anche la struttura degli errori delle covarianze. Per il regressore, tale matrice sarà nella forma: $\Phi_\delta = \mathrm{cov}(\delta)$ (le sue dimensioni sono TxT). Risulta possibile fare alcune assunzioni su questa matrice, per esempio, se l'ipotesi è di eteroschedasticità e di indipendenza degli errori, la matrice sarà diagonale, con gli elementi sulla diagonale principale diversi tra loro, del tipo $\sigma^2_{\delta_t}$. Occorre stabilire anche la relazione tra gli errori della variabile risultato e gli errori della variabile esogena: $\Phi_{\delta\varepsilon} = \mathrm{cov}(\delta\varepsilon)$; dove gli elementi sulla diagonale principale rappresentano la correlazione tra un istante temporale e l'altro (entro quindi la variabile tempo), mentre gli altri elementi indicano la correlazione tra i diversi istanti temporali. La relazione tra i parametri latenti delle variabili endogene ed esogene è espressa dalla:

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \tau_0 \\ \tau_1 \end{pmatrix} + \begin{pmatrix} \gamma_{\alpha\alpha'} & \gamma_{\alpha'\beta} \\ \gamma_{\alpha\beta'} & \gamma_{\beta\beta'} \end{pmatrix} \begin{pmatrix} X_{\alpha_i} \\ X_{\beta_i} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} + \begin{pmatrix} \zeta_{\alpha_i} \\ \zeta_{\beta_i} \end{pmatrix},$$

che può essere vista come il modello strutturale stesso $\overline{\eta} = \overline{\tau}_\eta + \Gamma\overline{\xi} + B\overline{\eta} + \overline{\zeta}$. La matrice Γ contiene i parametri della regressione di secondo livello, che catturano la relazione diretta tra le variazioni nella variabile risultato e le variazioni nella variabile esogena.

La rappresentazione grafica di un esempio di questo modello è quella in Grafico 39 (Y è la variabile obiettivo, misurata per 3 diversi istanti temporali, mentre X è un predittore con valori che cambiano nel tempo).

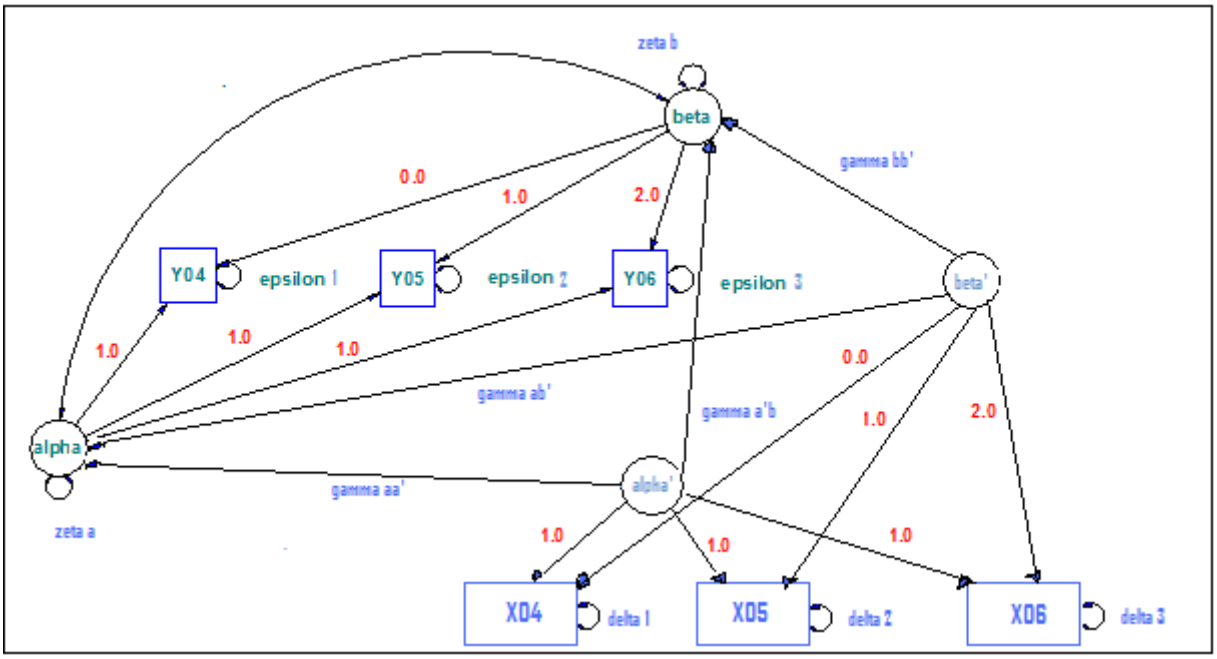


Grafico 39- Rappresentazione grafica di un modello lineare condizionato con 1 predittore di tipo *time varying*

4 Stima del modello a curva latente

4.1 ANALISI PER SINGOLO CASO

Dopo una prima analisi dell'esito scolastico in termini di variabile dicotomica e senza tener conto dell'elemento temporale, si è focalizzata l'attenzione sull'andamento nel tempo e sui vari gruppi di individui, formati dai diversi valori delle medesime variabili esplicative usate nei precedenti modelli, per quanto riguarda la variabile obiettivo *Times promoted*.

Il primo passo è stato quello di stimare una regressione del tipo "caso per caso" (basandosi sugli anni scolastici dal 2002/03 al 2006/07, escludendo il 2007/08, quando gli studenti con percorso regolare erano già usciti da scuola). Vi sono 315 studenti presenti a scuola in uno solo degli anni scolastici considerati e per questi non è stato possibile stimare una regressione. Potrebbe sembrare che tutti questi studenti abbiano abbandonato precocemente la scuola, ma non è esattamente così, o almeno non tutti lo hanno fatto. Infatti il 58% di essi ha frequentato la classe prima nel 2002/03 (di questi, il 91% non è stato promosso), quindi molti di essi hanno probabilmente lasciato la scuola, ma potrebbero esservi anche alcuni che hanno in realtà continuato gli studi al di fuori della provincia di Bologna. Il 17% di essi ha frequentato la classe prima nel 2003/04 (con un anno di ritardo rispetto alla situazione regolare e di questi il 95% non è stato promosso) ed il 6% di essi ha frequentato la classe prima nel 2004/05 (il 75% di essi ha cittadinanza non italiana). È quindi probabile che questo 23% degli studenti presenti in un solo anno scolastico tra quelli considerati, trovandosi in posizione non regolare rispetto al percorso di studi, abbia scelto di frequentare un solo anno di scuola, anche a seguito dell'obbligo fino ai 15 anni, e poi di uscire dal percorso scolastico, forse anche per seguire un corso di formazione professionale. Gli altri studenti presenti in un solo anno scolastico (il 19%) frequentavano una classe successiva alla seconda in un anno scolastico che indicava la regolarità del loro percorso scolastico, è quindi probabile che siano rimasti a studiare a Bologna per un solo anno scolastico, ma che abbiano poi completato il percorso di studi in altra provincia (il 53% di questi ha ottenuto la promozione, inoltre si tratta di studenti con cittadinanza non italiana per il 47% e ciò potrebbe in parte spiegare la loro esigenza di mobilità). Guardando la composizione di questo gruppo di studenti, la maggior parte dei quali (ma non tutti) ha molto probabilmente abbandonato precocemente il percorso scolastico, si notano alcuni elementi di rischio (tenendo presente le conclusioni tratte dai modelli precedenti): sono per la maggior parte (il 65%) maschi, con cittadinanza non italiana (il 25%), inoltre più della metà di essi ha frequentato un istituto professionale (il 51%, mentre la percentuale media di studenti che frequentano tale tipologia di istituto è del 20%).

Considerando ora soltanto gli individui rilevati per due o più istanti temporali, si sono stimate 5.624 regressioni, ottenendo una distribuzione di intercette e di pendenze. È importante notare che la percentuale degli studenti che si sono diplomati è del 61%, considerando soltanto questi 5.624 individui, mentre la percentuale di diplomati scende al 58% se si considerano tutti i 5.939 studenti.

La statistica R-quadro, calcolata per ognuna delle regressioni, ha la distribuzione mostrata nel Grafico 40.

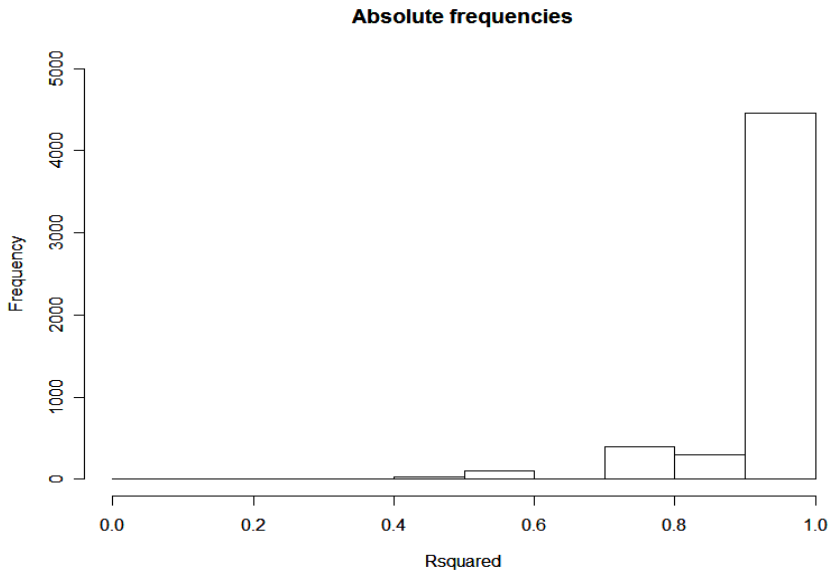


Grafico 40 – Distribuzione di frequenza degli indici R^2 calcolati per ognuna delle regressioni sui singoli individui del dataset

Il valore di R-quadro medio è = 0.958. I valori di R-quadro sono confortanti: il 96% della variabilità è in media spiegata dal modello lineare. Occorre tener conto del fatto che il calcolo di R-quadro, per quegli individui che hanno solo valori pari a 0 della variabile obiettivo (non sono mai stati promossi, anche se hanno frequentato per più di un anno scolastico la stessa classe) porta ad un valore Null. Il valore 1 di tale indice è invece collegato a quegli studenti che hanno frequentato tutte le cinque classi e che non sono mai stati bocciati. L'indice R-quadro ha un valore per ogni studente introdotto nel modello, ma tale valore rimane il medesimo per lo stesso individuo nel tempo. Per quanto invece riguarda i singoli coefficienti (intercetta e pendenza), occorre prestare particolare attenzione al loro significato. I grafici 41 e 42 mostrano le loro distribuzioni di frequenza.

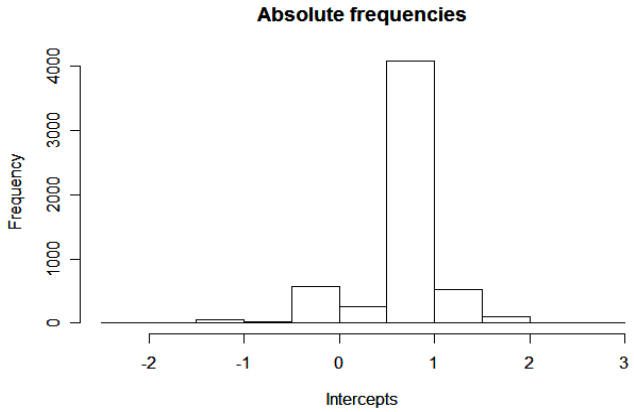


Grafico 41 – Distribuzione di frequenza delle intercette ottenute dalle regressioni calcolate sui singoli individui del dataset

L'intercetta indica, in generale, se lo studente è stato promosso oppure no alla classe frequentata al tempo 0. Nel caso di corso di studi regolare, l'intercetta deve essere pari a 1 (lo studente è stato promosso in classe prima al tempo 0); se uno studente ha iniziato il proprio corso di studi al di fuori della provincia di Bologna e compare al tempo 3 o al tempo 4 con un corso di studi regolare, ha anch'egli un'intercetta pari a 1. Si osserva che il 97% degli studenti che si è diplomato entro il 2007 ha un'intercetta uguale a 1. Tuttavia la sola intercetta non è certo un buon indicatore della buona prosecuzione degli studi; esso sintetizza, infatti, soltanto la regolarità nella classe prima (infatti, anche lo studente che ha ripetuto più di una volta una classe successiva alla prima e poi ha

abbandonato la scuola ha un'intercetta pari a 1). L'intercetta media è 0,860, un valore non molto diverso da 1, che consideriamo il caso regolare; la deviazione standard è 0,451. Il valore estremo di intercetta pari a 3 è quello dei pochissimi individui che hanno frequentato (anche più di una volta senza successo) soltanto la classe terza a Bologna, oppure che hanno frequentato la classe terza con una promozione e la quarta con una bocciatura, e che dopo l'ennesima bocciatura hanno lasciato la scuola (oppure la scuola bolognese, specie i molti riscontrati che vivono in realtà fuori Bologna). Le intercette estreme di valore compreso in]2 ;3[sono quelle relative a quegli studenti che sono rimasti a Bologna per meno di 5 anni scolastici, che hanno frequentato la prima classe al di fuori della provincia e che sono stati almeno una volta bocciati (per esempio, uno studente che ha frequentato 3 classi fuori Bologna, iniziando il corso di studi in anticipo di un anno, e che poi ha frequentato la quarta all'interno della provincia ed ha ottenuto esito negativo; inoltre che in seguito ha ripetuto la quarta con successo ed è stato poi promosso anche in classe quinta nel 2006/07); tra questi studenti, riguardo a coloro che non risultano aver completato il percorso scolastico giungendo al diploma, non è dato sapere se hanno lasciato definitivamente la scuola o se si sono trasferiti in altra provincia (si è notato che si tratta di studenti nati tutti al di fuori della regione Emilia Romagna). Il valore pari a 2 caratterizza gli studenti che hanno iniziato la scuola in anticipo (un anno prima degli altri) e che hanno poi avuto un corso di studi regolare, anche se vi sono due casi di studenti che hanno invece lasciato la scuola bolognese. Il 2% degli studenti diplomati entro il 2006/07 ha un'intercetta uguale a 2. Il valore dell'intercetta <1 indica un individuo con un ritardo rispetto alla regolarità, che quindi ha cominciato la scuola uno, due o più anni dopo rispetto al caso regolare, oppure anche un individuo che è stato ripetutamente bocciato. Da ciò possiamo desumere che i valori dell'intercetta che denotano almeno una buona partenza sono in generale 1 e 2. I valori d'intercetta compresi nell'intervallo]1, 2[sono quelli relativi agli individui promossi in classe prima, ma poi bocciati uno o più anni dopo. Se uno studente, per esempio, è stato promosso nelle prime tre classi e poi ha lasciato la scuola, ha comunque intercetta pari a 1. Ecco come quindi i soli valori dell'intercetta nulla dicono sull'abbandono precoce della scuola. È anche vero, come già accennato, che non è completamente corretto analizzare l'abbandono scolastico attraverso questi dati riferiti alla sola popolazione di Bologna, in quanto non vi è distinzione tra coloro che si spostano dalla provincia e coloro che possono essere considerati come *drop out*.

Se l'intercetta media calcolata sui 5.624 individui è 0,860, l'intervallo di confidenza, determinato dall'espressione

$$s.e.(\hat{\mu}_\alpha) = \sqrt{\frac{\sum (\hat{\alpha}_i - \hat{\mu}_\alpha)^2 / (N-1)}{N}}$$ è CI(95%)=[0,860±1,96*0,006]=[0,848;0,872].

I valori poco inferiori a 0 indicano una o più bocciature in classe prima ed eventualmente anche altre bocciature in seguito (per esempio, un'intercetta di 0,1 è quella di un individuo che è stato bocciato una volta in classe prima e una volta in seconda; un'intercetta di 0,2 è quella di uno studente che ha ripetuto la prima e la terza classe). Molti dei ragazzi caratterizzati da una tale intercetta scompaiono dal dataset prima del quinto anno rilevato, quindi probabilmente questi hanno effettivamente poi abbandonato la scuola. Un valore dell'intercetta uguale a –1 caratterizza quegli studenti che hanno iniziato la scuola superiore in ritardo, quindi frequentavano la classe prima nell'anno scolastico 2004/05 e sono stati promossi oppure frequentavano la prima nel 2003/04 e sono stati bocciati. I rarissimi valori di intercetta pari a –1,5 sono collegabili a quegli studenti che sono stati

bocciati in prima, ma che poi hanno concluso con successo due classi in un solo anno scolastico, presentandosi regolarmente alla classe successiva. Il valore pari a –2 è relativo agli studenti (per la maggior parte con cittadinanza non italiana) che hanno iniziato la scuola superiore con molto ritardo rispetto alla situazione regolare, che quindi frequentavano la prima nel 2005/06; questi studenti erano ancora a scuola nel 2007/08.
La pendenza indica il tasso di variazione della variabile obiettivo *Times Promoted*. Non è possibile utilizzare la sola pendenza come indicatore dell'abbandono scolastico.

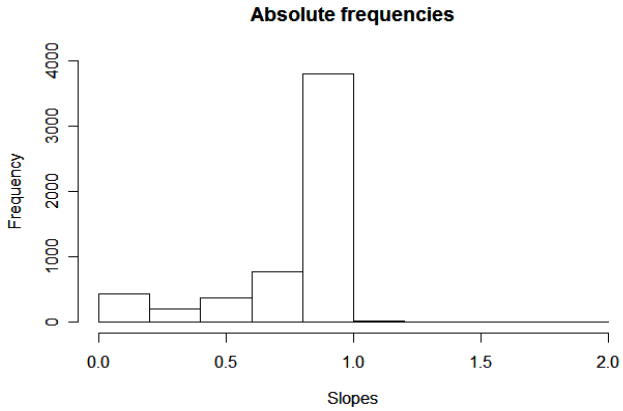


Grafico 42 – Distribuzione di frequenza delle pendenze ottenute dalle regressioni calcolate sui singoli individui del dataset

Il valore minimo della pendenza, cioè 0, indica quegli studenti che hanno frequentato soltanto una classe, per più di un anno scolastico (più volte bocciati), ma che poi hanno abbandonato la scuola (o hanno lasciato Bologna). Il valore 0,5 rappresenta quegli studenti che hanno frequentato due volte la classe seconda (e che poi hanno abbandonato la scuola) oppure anche due volte la seconda e la terza (e poi anch'essi hanno abbandonato la scuola). La pendenza pari a 0,9 è propria degli individui che hanno ripetuto la classe terza oppure la quarta e che poi hanno concluso con successo due classi in un solo anno scolastico; lo 0,5% degli studenti che si sono diplomati nel 2006/07 (quindi regolarmente) hanno un tale valore della pendenza. Il valore 1 denota una progressione lineare nella variabile *Times promoted*: [0,1,2] oppure [1,2,3], ecc.. Gli studenti con tale valore sono stati promossi in tutti gli anni scolastici in cui erano presenti, o al più sono stati bocciati una sola volta in classe prima; potrebbero però aver lasciato la scuola (o la scuola bolognese) prima del diploma. Il 98% degli studenti che si sono diplomati nel 2006/07 è caratterizzato dalla pendenza pari a 1. Un livello pari a 1,1 denota una bocciatura in seconda, ma lo stesso studente ha poi concluso con successo due classi (seconda e terza) in un solo anno, fino a frequentare con successo finale la classe quarta nel 2006/07. La pendenza pari a 1,17 è quella degli studenti che hanno prima frequentato per alcuni anni la scuola fuori Bologna (concludendo due classi in un solo anno), ma che poi sono stati bocciati, e alla fine si sono diplomati nel 2006/07, quindi regolarmente. In particolare, l'1,5% dei diplomati entro il 2006/07 ha tale livello di pendenza. Il valore 1,2 rappresenta coloro i quali sono stati bocciati in classe prima, ma che poi hanno concluso con successo due classi in un anno (la prima e la seconda) e si sono ugualmente diplomati nel 2006/07. Il valore 1,26 è invece proprio di uno studente che ha frequentato per alcuni anni la scuola fuori Bologna (anche due classi in un anno), ma che ha comunque frequentato la prima, con una bocciatura, e poi le ultime classi, essendo promosso, a Bologna, diplomandosi nel 2006/07. Il livello di pendenza pari a 1,3 denota una bocciatura nelle prime classi, ma lo studente ha poi frequentato due classi in un anno, quindi si è diplomato, o al più ha terminato con successo la quarta, nel

2006/07. Rarissimo è il caso del valore 1,4, proprio di uno studente che è stato bocciato, ma che poi ha frequentato addirittura 3 classi in un anno, diplomandosi nel 2006/07. Pendenza uguale a 1,5 è relativa a coloro i quali hanno frequentato la prima e la seconda classe in un solo anno ma che poi hanno abbandonato la scuola (o la scuola bolognese) al termine della terza o della quarta classe. Gli studenti che si sono diplomati regolarmente nel 2006/07 hanno diversi valori della pendenza, quindi questa non è un buon indicatore nemmeno della performance scolastica. Soltanto una studentessa straniera presenta una pendenza pari a 2: ha frequentato la prima ed è stata bocciata, poi ha concluso prima e seconda classe in un solo anno, ma poi ha lasciato la scuola (o Bologna). In generale si può però dire che il valore della pendenza che denota una buona performance scolastica (anche se non in tutti i casi, come visto) è il valore quanto più vicino a 1.

La pendenza media calcolata sui 5.624 studenti è 0,832, mentre l'intervallo di confidenza, calcolato in base all'espressione:

$$s.e.(\hat{\mu}_\beta) = \sqrt{\frac{\sum (\hat{\beta}_i - \hat{\mu}_\beta)^2 / (N-1)}{N}}$$ è CI(95%)=[0,832±1,96*0,004]=[0,824;0,840].

La bontà di adattamento del modello sembra alta, in ragione dei valori alti di R-quadro: per la maggior parte degli studenti la traiettoria lineare è quella che meglio approssima il trend della variabile obiettivo *Times promoted*. La traiettoria stimata con la procedura OLS appare riflettere adeguatamente le medie osservate, come si vede dal Grafico 43 (le medie osservate sono rappresentate da triangoli, mentre quelle stimate da rombi).

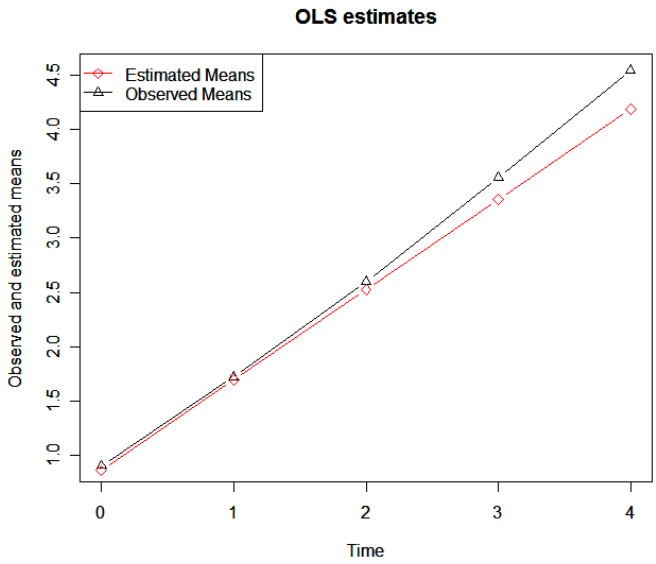


Grafico 43 – Rappresentazione grafica della variabile obiettivo: medie temporali osservate e stimate dal modello

Utilizzando il metodo di misura della bontà di adattamento per singola componente (*components of fit*[67]), si possono esaminare, in quanto a bontà di adattamento, alcune parti del modello spiegato, invece di una statistica sintetica per tutto il modello. Una possibilità è quella di valutare le differenze tra le medie stimate dal modello e le medie osservate nei diversi istanti temporali: [0,043; 0,034; 0,075; 0,201; 0,363]; il modello sottostima i valori medi della variabile obiettivo (come messo in evidenza anche dal Grafico 43).

[67] Si rimanda al capitolo sulla trattazione della teoria.

Si possono poi stimare le varianze dei singoli coefficienti al fine di valutare la precisione delle stime. Uno stimatore corretto delle varianze degli errori è $\text{var}(\varepsilon_i)=\frac{\sum_{t=1}^{T} e_{it}^2}{T-2}$, dove e_{it}^2 è il quadrato dell'errore per lo studente *i* al tempo *t*. In questo modo, si ottengono N (5.624) stime delle varianze. Sotto l'ipotesi che le varianze degli errori siano uguali per tutti gli individui, queste stime possono tutte essere considerate stime in realtà di un'unica varianza, che viene a questo punto stimata come:

$\text{var}(\varepsilon)=\frac{\sum_{i=1}^{N}\text{var}(\varepsilon_i)}{N}$=0,034, errore medio compiuto stimando le traiettorie individuali con il metodo OLS.

Si possono quindi ottenere le stime corrette delle varianze dei coefficienti:

$$\hat{\psi}_{\alpha\alpha} = \text{var}(\hat{\alpha}) - \frac{\text{var}(\varepsilon)\sum_{t=1}^{T}\lambda_t^2}{T\left(\sum_{t=1}^{T}(\lambda_t-\bar{\lambda})^2\right)}=0{,}181$$

$$\hat{\psi}_{\beta\beta} = \text{var}(\hat{\beta}) - \frac{\text{var}(\varepsilon)}{\sum_{t=1}^{T}(\lambda_t-\bar{\lambda})^2}=0{,}09$$

Queste stime entrano nella matrice di varianze e covarianze stimata; in tal modo si può confrontare quest'ultima con la matrice di varianze e covarianze osservata ed esaminare la matrice di varianze e covarianze degli errori:

$S-\Sigma(\hat{\theta})$ che diventa nel caso in esame $\begin{pmatrix} -0{,}08 & -0{,}07 & -0{,}11 & -0{,}14 & -0{,}2 \\ -0{,}07 & 0 & -0{,}10 & -0{,}26 & -0{,}43 \\ -0{,}11 & -0{,}10 & -0{,}11 & -0{,}36 & -0{,}66 \\ -0{,}14 & -0{,}26 & -0{,}36 & -0{,}46 & -0{,}94 \\ -0{,}2 & -0{,}43 & -0{,}66 & -0{,}94 & -1{,}13 \end{pmatrix}$, mostrando una forte sottostima della matrice di varianze e covarianze da parte del modello (5 valori sono sottostimati di circa il 60%, 4 valori sono sottostimati di circa il 45% e solo 1 valore è perfettamente stimato, mentre due valori sono sottostimati di meno del 20%). Il grado di sottostima cresce nel tempo (come anche la sottostima delle medie).

La differenza tra le due matrici è ben visibile nel Grafico 44.

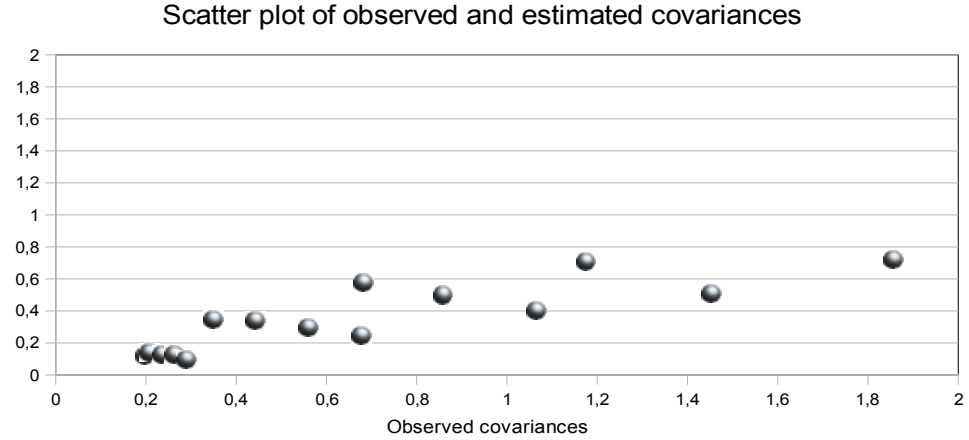


Grafico 44 – Rappresentazione grafica della relazione tra varianze e covarianze osservate (in ascissa) e stimate dal modello *case by case* (in ordinata)

Nel Grafico 44, un andamento sulla linea bisettrice degli assi mostrerebbe un modello che stima perfettamente le covarianze, ma è evidente che non è questo il caso. Un elemento che non si vede ma che va segnalato è che il numero di istanti temporali non è costante per tutti gli individui, quindi le covarianze osservate sono state calcolate soltanto per quegli individui che avevano entrambi i relativi istanti temporali (ogni covarianza ha pertanto un riferimento su un numero diverso di individui).

Uno svantaggio del modello appena descritto è quello di non poter introdurre variabili esplicative oltre al tempo. Vi sarebbe la possibilità di stimare un modello diverso per ogni sottogruppo di studenti identificato da valori diversi delle variabili esplicative (ad esempio un modello per i maschi ed uno per le femmine, uno per i frequentanti ogni diversa tipologia di scuola). Si è però scelto di non percorrere questa strada per l'assenza di un indicatore globale della bontà del modello: se infatti, come di fatto accade, i diversi modelli si adattano in modo diverso ai dati, i rispettivi risultati non possono essere tra loro confrontati.

4.2 STIMA DEL MODELLO A CURVA LATENTE (LCM)

4.2.1 INTRODUZIONE ALLA PROCEDURA DI STIMA

Il passo successivo dell'analisi è stato quello di stimare un LCM sull'intero dataset (5.939 studenti).
In questo caso, differentemente dalla stima caso per caso, è stato possibile stimare un modello che comprendesse anche quegli studenti per i quali era presente la rilevazione di un solo istante temporale, in quanto la procedura prende in carico tutti i dati, non essendo basata sulla singola regressione. L'unica condizione, qui rispettata, è che il numero di individui con una sola rilevazione nel tempo sia in percentuale sufficientemente contenuta.
Il software utilizzato per stimare il modello è R, in particolare si è utilizzata la procedura *lme*, della libreria *nlme*. Questa procedura stima un modello lineare ad effetti misti. Il modello di crescita può infatti essere visto come un caso particolare del modello lineare ad effetti misti (fissi e casuali). È presentata nel seguito la specificazione del modello.

$$y_{it} = \beta_1 x_{1it} + ... + \beta_p x_{pit} + b_{t1} z_{1it} + ... + b_{tq} z_{qit} + \varepsilon_{it}$$

con $b_{tk} \sim N\left(0, \psi_k^2\right)$, $\quad \text{cov}\left(b_k, b_{k'}\right) = \psi_{kk'}$

con $\varepsilon_{it} \sim N\left(0, \sigma^2 \lambda_{tii}\right)$ $\quad \text{cov}\left(\varepsilon_{ti}, \varepsilon_{ti'}\right) = \sigma^2 \lambda_{tii'}$

dove y_{it} è la variabile obiettivo;

$\beta_1, ..., \beta_p$ sono i coefficienti degli effetti fissi, cioè i coefficienti delle variabili esplicative, costanti per tutte le unità di tempo (l'unità di tempo è trattata come sottogruppo);

$x_1, ..., x_p$ sono le variabili esplicative – effetti fissi, con un valore per ogni individuo e per ogni istante temporale;

$b_{t1}, ..., b_{tq}$ sono i coefficienti degli effetti casuali, sono diversi per ogni unità di tempo e la loro distribuzione ipotizzata è una Normale Multivariata; ogni b_{tk} è visto come variabile casuale, quindi è in realtà diverso da un coefficiente;

$z_1, ..., z_q$ sono i regressori – effetti casuali, con un valore per ogni individuo e per ogni istante temporale;

ψ_{kk} sono le varianze e covarianze degli effetti casuali, sono costanti per tutte le unità di tempo, ma sono diverse per ogni regressore;

ε_{it} sono gli errori per ogni individuo e unità di tempo; la loro distribuzione ipotizzata è la Normale Multivariata;

$\sigma^2 \lambda_{tii'}$ è la covarianza tra gli errori al tempo *t*.

La notazione matriciale del modello è la seguente:

$$y_t = x_t \beta + Z_t b_i + \varepsilon_i \quad \text{con} \quad b_t \sim N_q\left(0, \Psi\right) \text{ e } \varepsilon_t \sim N_{n_t}\left(0, \sigma^2 \Lambda_t\right)$$

dove $\sigma^2\Lambda_t$ (n_txn_t) è la matrice di varianze e covarianze degli errori, mentre Ψ (qxq) è la matrice di varianze e covarianze degli effetti casuali.

La prima equazione specificata risulta assai simile alla specificazione compatta del modello a curva latente; ciò risulta ancor più evidente effettuando alcune sostituzioni nella notazione: $\beta_1, ..., \beta_p$ e $x_1, ..., x_p$ sono

rispettivamente i coefficienti e le variabili che non cambiano nel tempo, quindi sono i coefficienti e le variabili delle equazioni di secondo livello (regressori di intercetta e pendenza), mentre $b_{t1},...,b_{tq}$ e $z_1,....,z_q$ sono i coefficienti e le variabili caratterizzati da valori che cambiano nel tempo, quindi sono gli elementi che fanno parte dell'equazione di primo livello; possono essere regressori che variano nel tempo.

4.2.2 RICERCA DEL MODELLO OTTIMALE PER L'INTERO DATASET

Il primo modello stimato sull'intero dataset è un modello base: delle **medie non condizionate** (*Unconditional Means*) che non prevede l'introduzione nemmeno della variabile Tempo. Attraverso la stima di questo modello è possibile valutare la variabilità entro gli individui, mentre nei modelli più complessi tale variabilità verrà scomposta in due parti, la prima facente riferimento allo stato iniziale (intercetta) e la seconda riferita invece al tasso di cambiamento (pendenza). La variabilità entro gli individui, quindi tra i diversi istanti temporali, è espressa dalla varianza di primo livello, mentre la variabilità tra gli individui emerge dal secondo livello.
Si è usato il metodo di stima FML[68].
Così, il **modello A**, delle medie non condizionate, risulta essere il seguente:

$y_{it} = \alpha_i + \varepsilon_{it}$ con ipotesi $\varepsilon_{it} \sim N\left(0, \sigma_\varepsilon^2\right)$

$\alpha_i = \gamma_{00} + \zeta_{0i}$ con ipotesi $\zeta_{0i} \sim N\left(0, \sigma_0^2\right)$

La stima di γ_{00} è 2,596, significativa (p-value=0), intercetta media degli individui. Questo modello assume che il trend della variabile obiettivo sia una linea piatta; ogni linea individuale si trova a distanza α_i dall'origine, mentre la posizione media è γ_{00}; α_i è la media specifica di ogni persona (per diversi valori del tempo), mentre γ_{00} è la media generale. In questo caso, le differenze tra gli studenti sono molto vicine a 0. La stima 2,596 è più bassa del valore 3, valore atteso per la progressione regolare nel tempo della variabile obiettivo *Times promoted*: [1,2,3,4,5]. Questo fatto indica che i valori della variabile obiettivo tendono ad essere più bassi di quello che si riscontrerebbe nel caso di una progressione regolare, quindi gli studenti hanno in media un trend che sta al di sotto di quello del caso di perfetta regolarità.
L'intervallo di confidenza stimato per il parametro γ_{00} è [2,587; 2,605], non comprendente il valore 3 del caso di perfetta regolarità.
Nel seguito verranno introdotti alcuni predittori; per ora ci si sofferma sui valori medi della variabile *Times promoted* distintamente per le diverse categorie di individui: le ragazze italiane hanno un valor medio di 2,59, i ragazzi italiani hanno un valor medio di 2,35, mentre per le ragazze con cittadinanza straniera lo stesso valor medio è di 1,81 e di 1,46 per i ragazzi stranieri. Anche queste differenze suggeriscono che alcune caratteristiche degli studenti potrebbero effettivamente distinguerli in base agli esiti scolastici.

La stima di σ_ε^2 è 2,122 (il corrispondente errore standard è 1,457): rappresenta la somma della varianza entro gli individui e della varianza entro gli istanti temporali. Nella progressione regolare, l'errore standard è 1,414, quindi il valore 1,457 indica una variabilità maggiore della variabile obiettivo rispetto alla situazione regolare.

[68] Si veda il capitolo sulla teoria alla base dei modelli.

La stima di σ_0^2 è 0,0001: rappresenta la varianza tra gli individui (la differenza tra le medie individuali e la media generale). Non risulta essere un valore significativo.

Il secondo modello qui stimato è quello che tiene conto della sola variabile temporale: il modello a curva latente (chiamato anche di crescita) **non condizionato** (*Unconditional Growth*).

Si è utilizzato il metodo di stima FML. Quindi, il **modello B** diventa:

$$y_{it} = \alpha_i + \beta_i Time_{it} + \varepsilon_{it} \quad \text{con ipotesi } \varepsilon_{it} \sim N\left(0, \sigma_\varepsilon^2\right)$$

$$\begin{aligned} \alpha_i &= \gamma_{00} + \zeta_{0i} \\ \beta_i &= \gamma_{10} + \zeta_{1i} \end{aligned} \text{ con ipotesi } \begin{pmatrix} \zeta_{0i} \\ \zeta_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{pmatrix}\right)$$

La variabile tempo è stata costruita a partire da ciascun anno di riferimento: i valori attribuiti sono [0,1,2,3,4] (anni scolastici, rispettivamente, 2002/03, 2003/04, 2004/05, 2005/06, 2006/07). I periodi sono costanti, in altre parole gli istanti temporali sono tra loro equidistanti.

In questo modello, non vi sono regressori, ma i valori della variabile obiettivo dipendono soltanto dal tempo.

Tabella 61 – Stima dei parametri: modello B

	Parametro	Errore standard	P-value
γ_{00} intercetta media	0,819	0,006	0
γ_{10} pendenza media	0,833	0,004	0
σ_ε^2	0,0377		
σ_0^2	0,174		
σ_1^2	0,067		

La stima dell' intercetta media degli individui è significativa (p-value=0); si noti che l'intercetta media nella regressione caso per caso era 0,860, ma occorre tener conto del fatto che mancavano i 315 individui con un solo valore della variabile obiettivo *Times promoted*. Il valore stimato è inferiore a 1 (valore ottimo per l'intercetta, cioè intercetta degli studenti con percorso regolare), così indicando la presenza di bocciature già in classe prima.

La stima della pendenza media degli individui è significativa (p-value=0); la pendenza media nella regressione *case by case* era 0,832. Nella progressione regolare, la pendenza dovrebbe essere 1; il valore stimato conferma la presenza di esiti negativi nelle classi successive alla prima.

La stima di σ_ε^2 (il corrispondente errore standard è =0,194) rappresenta la varianza entro gli individui e tra i diversi istanti temporali. Se tutti gli individui fossero stati in situazione di regolarità scolastica, l'errore standard sarebbe stato molto vicino a 0; il valore 0,194 indica pertanto la presenza di individui con situazioni diverse dalla condizione regolare. Ovviamente, questa varianza è molto inferiore a quella del modello A, infatti il tempo spiega gran parte della variabilità entro gli individui.

La stima di σ_0^2 (il corrispondente errore standard è =0,417) rappresenta la varianza tra le intercette individuali. Data l'intercetta media, 0,819, la differenza che intercorre tra uno studente e l'altro è data dalla differenza dalla media generale, che è in media ±0,417.

La stima di σ_1^2 (il corrispondente errore standard è =0,258) rappresenta la varianza tra le pendenze individuali. Gli studenti differiscono in media ±0,258 dalla pendenza media 0,833.

R stima il coefficiente di correlazione tra intercetta e pendenza: $\hat{\rho}_{10}$ =0,507.

Così $\hat{\sigma}_{10}$=0,507*0,258*0,417=0,055 stima la covarianza tra i due coefficienti. Ciò vuol dire che vi è una certa correlazione tra l'esito scolastico degli studenti nel primo anno di scuola ed il loro esito scolastico negli anni successivi; circa il 5% del trend degli esiti scolastici dipende dal risultato del primo anno di scuola.

Prima di procedere con l'analisi dei risultati, viene nel seguito illustrata la fase di test delle ipotesi del modello.

L'omoschedasticità dei termini di errore è una ipotesi del modello. Data la specificazione compatta di questo:

$y_{it} = \gamma_{00} + \gamma_{10} Time_{it} + \zeta_{0i} + \zeta_{1i} Time_{it} + \varepsilon_{it}$.

È possibile calcolare le stime delle varianze dei termini di errore per ogni istante temporale considerato, [0,1,2,3,4]: [0,01550;0,02956;0,03129;0,01927;0,01811].

Tra questi valori si può vedere la presenza di eteroschedasticità.

Gli errori si assumono non essere autocorrelati; nel seguito è illustrato il test di tale ipotesi.

Si può ottenere un vettore delle correlazioni tra gli errori nei diversi istanti temporali:

[cor(0,1);cor(1,2);cor(2,3);cor(3,4)] = [-0,280;-0,484;-0,079;0,326].

Non vi è un'alta correlazione degli errori (il coefficiente di correlazione ha il proprio massimo in |1|, quindi questi valori non risultano essere molto alti). L'autocorrelazione è particolarmente bassa tra gli errori degli istanti 0 e 1 e degli istanti 2 e 3.

Un'altra ipotesi del modello è la distribuzione normale degli errori ε_{it} .

La distribuzione degli errori (il valore massimo è 1,195, il valore minimo è –1,436, mentre i valori più probabili sono [-0,0015; -0,00618; 0,00785; 0,0125] ciascuno con una probabilità del 13%; il 26% dei valori è compreso nell'intervallo [-0,005; 0,005]) è mostrata nel Grafico 45.

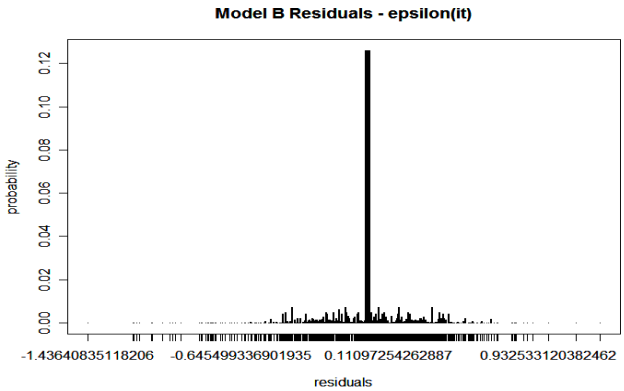


Grafico 45 – Distribuzione di probabilità degli errori del modello B

È anche possibile valutare graficamente l'ipotesi di normalità distributiva, attraverso il *qq-plot*, rappresentato nel Grafico 46, oppure osservando il Grafico 47, in cui i valori degli errori standardizzati sono raggruppati nei

diversi istanti temporali: in caso di normalità distributiva, la concentrazione degli errori dovrebbe essere non troppo differente nei diversi istanti considerati.

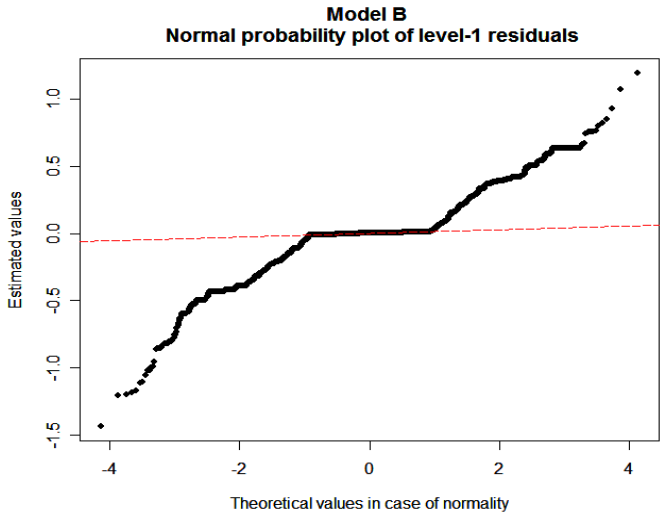


Grafico 46 – Valori degli errori del modello B stimati e valori teorici in caso di normalità distributiva

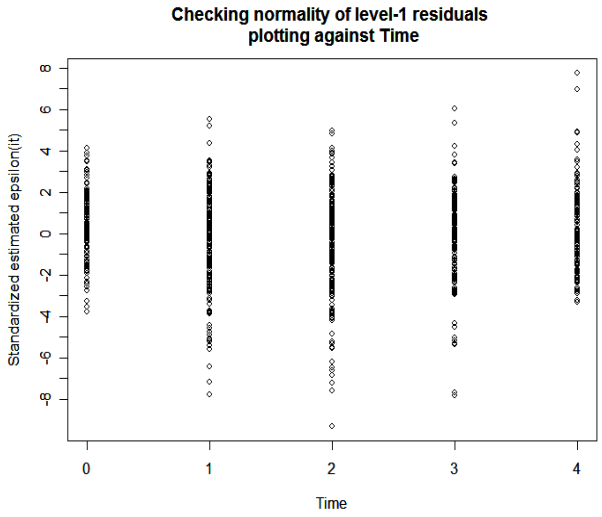


Grafico 47 – Distribuzione degli errori standardizzati del modello B nei diversi istanti temporali

Nel caso di distribuzione normale, in Grafico 46 dovrebbe comparire una linea retta. In questo caso, l'andamento è storto nel centro. Si può vedere che la maggior parte dei valori standardizzati rimangono nell'intervallo [-2; 2] (2.419 dei 26.292 valori rimangono fuori da questo intervallo; questi rappresentano circa il 9%, mentre nel caso di normalità distributiva la stessa percentuale dovrebbe rimanere attorno al 5%). Il test di Kolmogorov Smirnov mostra che in effetti l'ipotesi di normalità distributiva non è verificata. Anche il calcolo dell'indice di curtosi risulta pari a –2,99, indicante un eccesso di curtosi.

Nei grafici 48 e 49 sono presentate le distribuzione di probabilità degli errori di secondo livello.

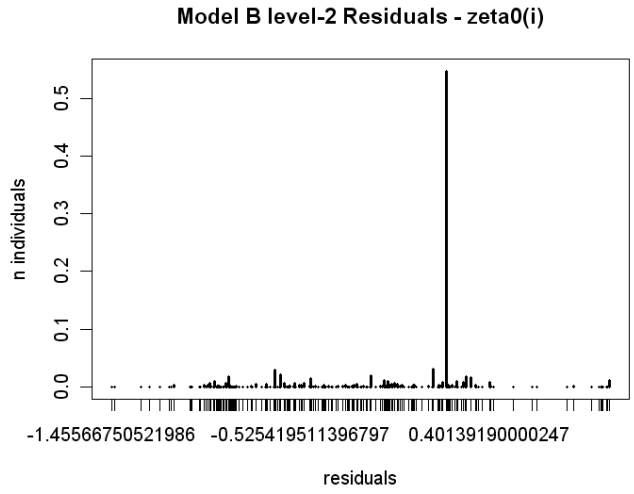


Grafico 48 – Distribuzione di probabilità degli errori di secondo livello del modello B relativi all'intercetta

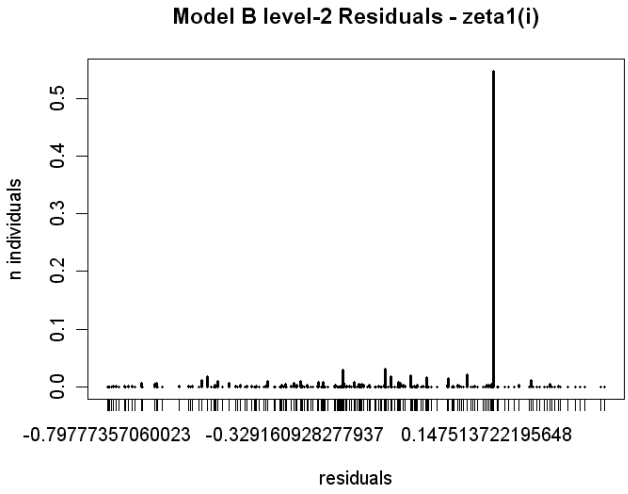


Grafico 49 – Distribuzione di probabilità degli errori di secondo livello del modello B relativi alla pendenza

I *qq-plots* relativi agli errori di secondo livello sono rappresentati nei Grafici 50 e 51.

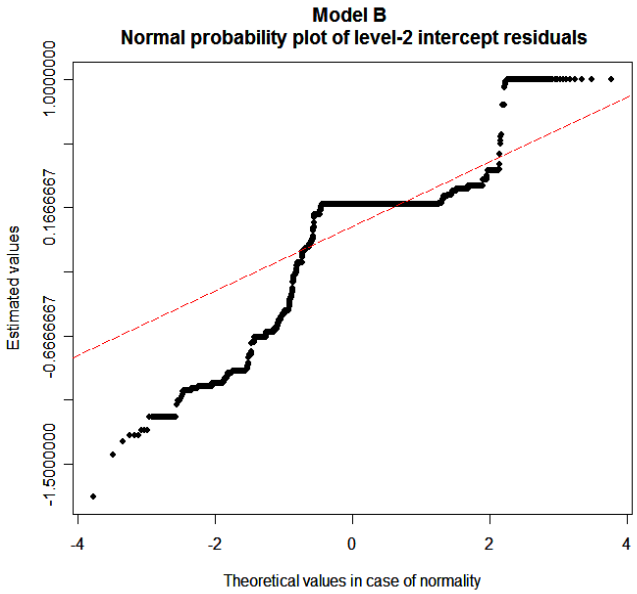


Grafico 50 – Valori degli errori di secondo livello, relativi all'intercetta, del modello B stimati e valori teorici in caso di normalità distributiva

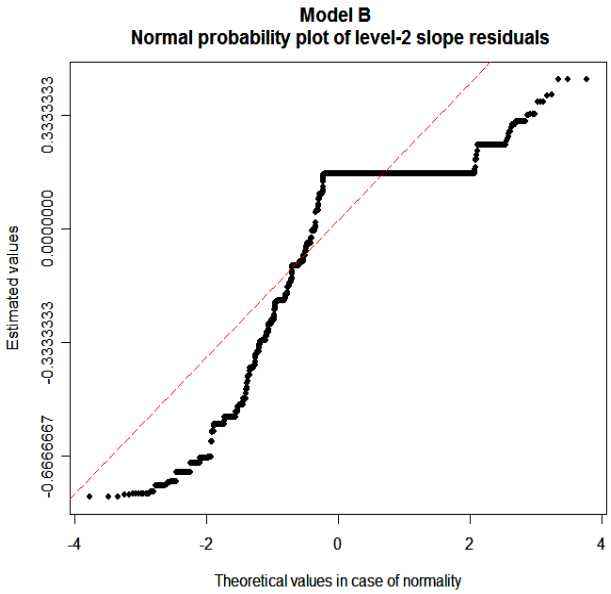


Grafico 51 – Valori degli errori di secondo livello, relativi alla pendenza, del modello B stimati e valori teorici in caso di normalità distributiva

La normalità distributiva non può essere assunta. Il test di Kolmogorov Smirnov fornisce una risposta negativa per entrambi i tipi di errore; inoltre circa l'8% degli errori standardizzati rimane al di fuori dell'intervallo [-2; 2]. Lo stimatore ML mantiene le sue proprietà asintotiche desiderabili (e qui la numerosità è 26.292, quindi si può dire elevata) in presenza di distribuzioni, seppure non normali, ma con una curtosi non eccessiva[69]. In questo caso, la curtosi degli errori di primo e di secondo livello è circa –2,99, quindi comunque un po' alta.

Nel seguito si intende invece misurare la bontà di adattamento del modello rispetto ai dati osservati. La statistica qui calcolata parte dalla correlazione tra i valori osservati e quelli stimati dal modello; si noti che il modello stima un valore per ogni istante temporale, ma la sequenza è uguale per tutti gli individui. I valori stimati sono perciò i seguenti: [0,819; 1,652; 2,486; 3,319; 4,152], per ognuno degli istanti temporali [0,1,2,3,4].

Si può notare che il primo valore stimato è di poco inferiore a 1, che è il valore rilevato nel caso di percorso regolare: è chiaro che vi sono molti studenti con un corso di studi irregolare, per questo motivo la prima stima è inferiore a 1.

La statistica R-quadro per il modello B è 0,769: indica un buon adattamento del modello ai dati osservati (il 77% della variabilità è spiegata dal modello). È anche possibile valutare la bontà di adattamento di ogni modello costruito rispetto gli altri modelli via via applicati, tramite la valutazione del miglioramento nella spiegazione della variabilità dell'esito scolastico.

Nell'analisi multivariata, viene usata la statistica R-quadro corretto per calcolare quanta parte di varianza viene spiegata aggiungendo ulteriori variabili esogene. Nell'analisi di dati longitudinali, in modo analogo, si utilizza un R-quadro corretto per calcolare la varianza spiegata, nella variabile obiettivo, grazie all'introduzione nel modello di un ulteriore predittore.

Il miglioramento verificatosi nel modello B rispetto al modello A, diventa così $pseudoR^2_{\varepsilon} = 0{,}982$: denota un forte miglioramento, in quanto il 98% della varianza entro gli individui viene spiegata dal predittore tempo.

[69] Curran "Latent Curve Models - A structural equation perspective"

Un modo per valutare la bontà di adattamento ai dati da parte del modello è quello di confrontare le medie temporali stimate con quelle osservate (Grafico 52).

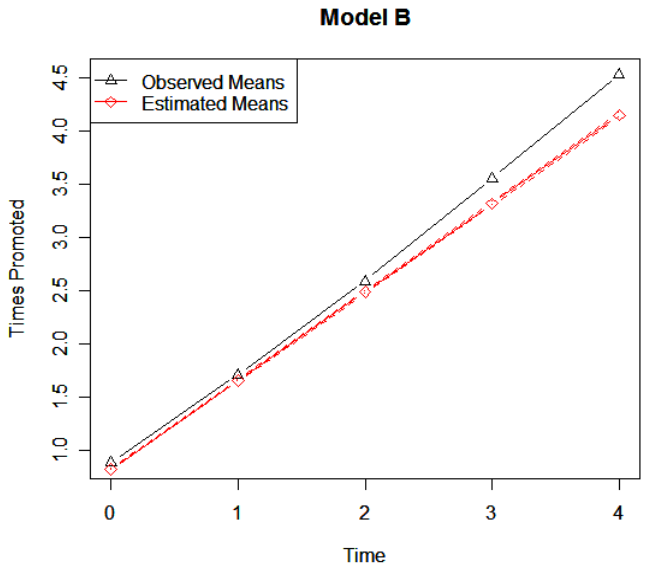


Grafico 52 – Linea delle medie stimate dal modello B (con relativo intervallo di confidenza) ed osservate

La successione dei valori stimati della variabile obiettivo, nei diversi istanti temporali, è [0,819; 1,652; 2,486; 3,319; 4,152], mentre la successione dei valori osservati è [0,877; 1,707; 2,588; 3,556; 4,534]; in Grafico 52,si può notare che l'intervallo di confidenza, oltre a non comprendere i valori osservati, è molto stretto: le stime hanno una bassa variabilità.

È chiaro che il modello sottostima i valori medi, come accadeva anche nel metodo per singolo caso; in quest'ultimo modello, inoltre, accade che le differenze siano addirittura inferiori rispetto al modello B (l'incremento delle differenze nei diversi istanti temporali, registrato nel modello B rispetto alla regressione per singolo caso, va da 0,01 a 0,03), che quindi sottostima maggiormente i valori osservati che non il precedente modello.

Come già visto in precedenza, è anche qui possibile valutare le differenze anche tra le varianze osservate e quelle stimate dal modello.

$S - \Sigma(\hat{\theta})$ diventa $\begin{pmatrix} -0{,}02 & -0{,}07 & -0{,}14 & -0{,}21 & -0{,}30 \\ -0{,}07 & 0 & -0{,}12 & -0{,}29 & -0{,}46 \\ -0{,}14 & -0{,}12 & -0{,}09 & -0{,}33 & -0{,}62 \\ -0{,}21 & -0{,}29 & -0{,}33 & -0{,}39 & -0{,}82 \\ -0{,}30 & -0{,}46 & -0{,}62 & -0{,}82 & -0{,}93 \end{pmatrix}$, che mostra una generale sottostima della matrice di varianze e covarianze, tuttavia si registra un miglioramento rispetto al modello per singolo caso*:* nel modello B, infatti, in generale le differenze risultano inferiori rispetto al precedente modello, inoltre qui la differenza massima è 0,93, mentre nell'altro modello era 1,13; 3 valori sono poi sottostimati di meno del 20%, mentre nel modello precedente se ne trovavano soltanto 2.

Anche in questo caso occorre notare che il numero di istanti temporali sui quali sono state calcolate le varianze non è costante per tutti gli individui, quindi le covarianze osservate sono state calcolate soltanto sulla base degli studenti presenti in ambo gli istanti temporali via via considerati.

Al fine di migliorare le stime, si è proceduto alla costruzione di modelli che hanno tenuto conto di alcune variabili esplicative, in particolare inserite nel modello di secondo livello (quindi con valori fissi nel tempo). La prima variabile esplicativa considerata è stata la cittadinanza. Si è consapevoli che tale scelta può determinare un errore di attribuzione, nel senso che, intendendo come stranieri coloro i quali conseguono un esito scolastico differente dagli altri per motivi sociali ed economici dati dalla loro condizione, non è detto che la distinzione per cittadinanza riesca a cogliere tale diversità. Tuttavia tale distinzione è stata reputata la migliore, avendo a disposizione informazioni sul luogo di nascita, sul luogo di residenza e sulla cittadinanza. Si è quindi introdotto nel modello la variabile esogena *Foreign*, indicando con tale termine la variabile relativa alla cittadinanza.

Si è pertanto stimato il modello **condizionato in base alla cittadinanza** (*Foreign*), utilizzando una variabile dicotomica che indica se uno studente ha cittadinanza italiana (valore =0) oppure straniera (valore =1). Il metodo di stima utilizzato è FML.

Ecco quindi come è stato costruito il **modello C**:

$$y_{it} = \alpha_i + \beta_i Time_{it} + \varepsilon_{it} \quad \text{con ipotesi } \varepsilon_{it} \sim N\left(0, \sigma_\varepsilon^2\right)$$

$$\begin{aligned} \alpha_i &= \gamma_{00} + \gamma_{01} Foreign_i + \zeta_{0i} \\ \beta_i &= \gamma_{10} + \gamma_{11} Foreign_i + \zeta_{1i} \end{aligned} \quad \text{con ipotesi } \begin{pmatrix} \zeta_{0i} \\ \zeta_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{pmatrix} \right)$$

Foreign è stato usato come regressore per i valori di pendenza e intercetta. In Tabella 62 sono mostrate le stime dei parametri ed i relativi errori standard.

Tabella 62 – Stima dei parametri: modello C

	Parametro	Errore standard	P-value
γ_{00} intercetta media italiani	0,854	0,006	0
γ_{10} pendenza media italiani	0,850	0,004	0
γ_{01} scostamento intercetta stranieri	-0,665	0,027	0
γ_{11} scostamento pendenza stranieri	-0,221	0,015	0
σ_ε^2	0,0377		
σ_0^2	0,156		
σ_1^2	0,062		

La stima di σ_ε^2 rimane invariata rispetto al modello precedente: questa varianza non cambia aggiungendo predittori di secondo livello.

La correlazione tra intercette e pendenze è $\hat{\rho}_{10}$ =0,461 (inferiore rispetto al modello B).

Per testare il miglioramento ottenuto grazie al modello, vengono calcolate le statistiche pseudo R-quadro per ognuno dei fattori latenti (nel caso lineare, l'intercetta e la pendenza). In questo caso, i miglioramento ottenuto con il modello C, rispetto al modello B, è misurato dagli indici:

$pseudoR^2_0$ =0,1023

$pseudoR^2_1$ =0,0681

Non si registra un forte miglioramento (del 10% per l'intercetta e del 7% per la pendenza).

Si può anche vedere se via sia una migliore rappresentazione dei dati osservati rispetto al modello precedente. I valori stimati, calcolati separatamente per gli studenti italiani e per gli stranieri, sono mostrati in Tabella 63.

Tabella 63 – Medie stimate, con rispettivi intervalli di confidenza, dal modello C (medie osservate) della variabile obiettivo *Times Promoted*

	2002/2003	2003/2004	2004/2005	2005/2006	2006/2007
Studenti italiani	0,854 [0,848-0,860] (0,881)	1,704 [1,695-1,713] (1,742)	2,554 [2,541-2,566] (2,654)	3,403 [3,387-3,420] (3,618)	4,253 [4,233-4,273] (4,605)
Studenti stranieri	0,189 [0,157-0,222] (0,745)	0,818 [0,767-0,870] (1,095)	1,447 [1,377-1,518] (1,620)	2,077 [1,987-2,166] (2,542)	2,707 [2,597-2,814] (3,404)

La statistica R-quadro è 0,794 (il 79% della variabilità è spiegata dal modello, registrando un leggero miglioramento rispetto al modello B, dove il 77% della variabilità era spiegata dal modello). Confrontando i valori stimati con quelli osservati, è evidente la differenza tra studenti italiani e studenti stranieri: in media gli italiani sono arrivati al quinto anno concludendo con successo la classe quarta, mentre gli stranieri sono arrivati al quinto anno concludendo con successo la classe seconda. Risulta anche evidente la differenza nella bontà di adattamento del modello ai dati osservati tra italiani e stranieri. A questo proposito è importante notare che gli italiani sono in maggior numero, inoltre che gli stranieri hanno progressioni della variabile obiettivo con una più alta variabilità da un individuo all'altro rispetto agli italiani. Per questi motivi le stime risultano migliori per gli studenti italiani.

Perché il modello rappresentasse con maggiore efficacia la situazione osservata, si è deciso di introdurre la variabile esplicativa indicante il sesso; si è passati poi ad evidenziare anche le differenze tra le progressioni della variabile obiettivo distintamente di maschi e femmine. Si è costruito il modello **condizionato in base a cittadinanza** (*Foreign*) **e sesso** (*Sex*), utilizzando il metodo di stima FML.

Quindi il **modello D** diventa:

$$y_{it} = \alpha_i + \beta_i Time_{it} + \varepsilon_{it} \text{ con ipotesi } \varepsilon_{it} \sim N\left(0, \sigma_\varepsilon^2\right)$$

$$\begin{aligned} \alpha_i &= \gamma_{00} + \gamma_{01} Foreign_i + \gamma_{02} Sex_i + \zeta_{0i} \\ \beta_i &= \gamma_{10} + \gamma_{11} Foreign_i + \gamma_{12} Sex_i + \zeta_{1i} \end{aligned} \text{ con ipotesi } \begin{pmatrix} \zeta_{0i} \\ \zeta_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{pmatrix}\right)$$

Foreign e *Sex* sono stati usati come regressori dicotomici per i valori di intercetta e pendenza.

Tabella 64 – Stima dei parametri: modello D			
	Parametro	Errore standard	P-value
γ_{00} intercetta media ragazze italiane	0,894	0,008	0
γ_{10} pendenza media ragazze italiane	0,894	0,005	0
γ_{01} scostamento intercetta stranieri	-0,667	0,027	0

Tabella 64 – Stima dei parametri: modello D			
	Parametro	Errore standard	P-value
γ_{11} scostamento pendenza stranieri	-0,224	0,015	0
γ_{02} scostamento intercetta maschi	-0,078	0,011	0
γ_{12} scostamento pendenza maschi	-0,086	0,007	0
σ_{ε}^2	0,0377		
σ_0^2	0,155		
σ_1^2	0,060		

In questo modello, il confronto tra maschi e femmine è difficoltoso, in quanto vi sono due variabili esogene; la differenza netta tra italiani e stranieri è ben visibile nel modello C, mentre la differenza netta tra maschi e femmine è visibile nel modello che utilizza soltanto il sesso come regressore. In particolare, il modello in cui viene introdotto soltanto il regressore sesso stima una differenza di intercetta tra maschi e femmine pari a –0,073 (i ragazzi è più probabile che ripetano la classe prima rispetto alle ragazze). Dallo stesso modello, si ottiene una differenza di pendenza tra maschi e femmine di –0,085 (i ragazzi è più probabile che abbiano un corso di studi non regolare, quindi con una progressione più lontana da quella regolare pari a 1, rispetto alle ragazze).

Il confronto dei risultati dei modelli condizionati mostra che vi sono differenze significative tra ragazzi e ragazze e tra studenti italiani e stranieri, ma anche sicuramente che gli studenti con cittadinanza non italiana hanno un corso di studi molto più irregolare che non quelli con cittadinanza italiana e tale differenza è molto maggiore rispetto a quella riscontrata tra maschi e femmine.

Non vengono qui presentati i risultati completi del modello che utilizza il sesso come unico regressore, in quanto ha comportato un miglioramento, rispetto al modello non condizionato, inferiore rispetto al modello invece presentato. Si è scelto quindi di inserire il sesso come secondo regressore di secondo livello.

La correlazione tra le intercette e le pendenze è $\hat{\rho}_{10}$ =0,454 (inferiore rispetto al modello C).

Per testare il miglioramento del modello, sono state calcolate le statistiche pseudo R-quadro di confronto tra il modello D ed il modello base B, che risultano:

$pseudoR_0^2$ =0,1108

$pseudoR_1^2$ =0,0983.

Occorre, come già visto, calcolare una di queste statistiche per ciascuno dei fattori latenti, in questo caso 2. Si è calcolato il miglioramento rispetto al modello B (senza regressori di secondo livello), ma analogamente si può confrontare il miglioramento ottenuto con il modello D rispetto a quello ottenuto con il modello C; in tal caso il miglioramento non è stato molto accentuato, ma maggiore rispetto al modello C: l'11% per l'intercetta, contro il 10% del modello C, e il 10% per la pendenza, contro il 7% del modello C.

I valori stimati (e quelli osservati) vengono (Tabella 65) calcolati separatamente per ragazzi e ragazze italiani e stranieri, così da permettere i confronti direttamente sulle medie stimate.

Tabella 65 – Medie stimate, con relativi intervalli di confidenza, dal modello D (medie osservate) della variabile obiettivo *Times Promoted*

	2002/2003	2003/2004	2004/2005	2005/2006	2006/2007
Ragazze italiane	0,894 [0,886-0,902] (0,913)	1,788 [1,775-1,801] (1,815)	2,682 [2,664-2,700] (2,752)	3,576 [3,554-3,599] (3,739)	4,470 [4,443-4,498] (4,718)
Ragazzi italiani	0,816 [0,797-0,836] (0,847)	1,625 [1,594-1,655] (1,673)	2,433 [2,390-2,475] (2,559)	3,241 [3,187-3,295] (3,499)	4,049 [3,984-4,115] (4,491)
Ragazze straniere	0,227 [0,193-0,262] (0,723)	0,897 [0,843-0,952] (0,969)	1,568 [1,493-1,643] (1,561)	2,238 [2,143-2,333] (2,496)	2,908 [2,793-3,023] (3,301)
Ragazzi stranieri	0,150 [0,104-1,195] (0,690)	0,734 [0,661-0,807] (0,852)	1,318 [1,219-1,418] (1,286)	1,903 [1,776-2,029] (2,115)	2,487 [2,334-2,641] (3,038)

Si può notare che le ragazze hanno, in media, un percorso scolastico migliore dei ragazzi in ambo i casi, sia che si tratti di cittadini italiani che stranieri, e il modello ben riproduce queste differenze. Né i ragazzi con cittadinanza non italiana né le ragazze con cittadinanza non italiana terminano con successo la classe quarta dopo i cinque anni; il modello stima che soltanto le ragazze straniere concludano con successo la terza classe dopo cinque anni; in particolare, il modello stima, rispecchiando le osservazioni, che le ragazze non italiane arrivino almeno al termine della terza classe, mentre sottostima il risultato dei ragazzi stranieri, per i quali le osservazioni mostrano che invece arrivino mediamente alla conclusione della terza classe, in cinque anni.

Gli intervalli di confidenza sono tra loro separati, quindi la differenza delle stime tra i vari gruppi di studenti risulta significativa. Si può notare che tali intervalli stimati risultano poco ampi, a conseguenza della non elevata variabilità delle stime, misurata dal loro errore standard.

La statistica R-quadro risulta pari a 0,796 (il 79,6% della variabilità è spiegata dal modello, con un leggero miglioramento rispetto al modello C, dove il 79,4% della variabilità era spiegata dal modello). La differenza tra le medie stimate e quelle osservate è maggiore per gli studenti stranieri (occorre sempre considerare che gli italiani sono in numero molto maggiore rispetto agli stranieri, che rappresentano il solo 7% della popolazione); risulta inoltre in aumento nel tempo, quindi le stime relative ai primi anni si avvicinano maggiormente ai valori osservati rispetto alle stime relative agli ultimi anni. A questo proposito, non bisogna dimenticare che il modello stima un trend lineare nel tempo, quindi può verificarsi in realtà che negli ultimi anni vi sia una maggiore deviazione dalla linearità rispetto agli anni precedenti: anche per quegli studenti per i quali ci si aspettava un insuccesso, dati i precedenti, si può verificare in realtà un miglioramento.

Si può notare che è possibile la stima di modelli diversi per diversi gruppi di individui (per esempio, maschi e femmine) invece di calcolare un unico modello per tutti gli individui, aggiungendo regressori. Dalla stima di modelli diversi per maschi e femmine (due diversi modelli C, con la variabile *Foreign* usata come regressore di secondo livello), si ottengono in realtà risultati simili (soltanto con valori leggermente più bassi), in termini di intercetta e di pendenza, a quelli ottenuti dal modello completo. Tuttavia un vantaggio del modello completo è che in questo modo si possono ottenere le statistiche che saggiano la bontà di adattamento per il modello riferito all'intero dataset.

Un altro passo compiuto per migliorare il modello è stato quello di considerare il confronto tra l'ambito di appartenenza della scuola frequentata da ogni studente e l'ambito di residenza dello studente stesso. Come già spiegato nelle premesse, tale confronto è stato effettuato per ciascun istante temporale, ottenendo una variabile

dicotomica che indica se in quell'anno scolastico lo studente ha frequentato un istituto situato nello stesso ambito di residenza oppure se in ambito diverso. Questa variabile rappresenta il grado di mobilità di ogni studente per raggiungere la scuola. Nelle statistiche descrittive si aveva che gli studenti che frequentano un istituto in ambito diverso da quello di residenza hanno in generale risultati peggiori degli altri, ad eccezione degli studenti che frequentano gli istituti professionali. Nel seguito sono presentati i relativi risultati del LCM.
Prima di tutto, in Tabella 66 viene presentata la distribuzione delle numerosità degli studenti nei vari gruppi individuati dai valori delle variabili dicotomiche a questo punto introdotte nel modello.

Tabella 66 – Distribuzione nei diversi anni scolastici del numero di studenti e degli istanti temporali nei diversi gruppi, identificati da valori diversi delle variabili esplicative

Cittadinanza	Sesso	Ambito	N. Istanti temporali	N. Studenti 2002/03	N. Studenti 2003/04	N. Studenti 2004/05	N. Studenti 2005/06	N. Studenti 2006/07
Italiana	F	Stesso ambito	8.198	1.726	1.725	1.666	1.591	1.490
Italiana	F	Ambito diverso	4.297	888	888	879	848	794
Italiana	M	Stesso ambito	8.250	1.762	1.777	1.682	1.579	1.450
Italiana	M	Ambito diverso	4.432	901	932	917	869	813
Totale			25.177	5.277	5.322	5.144	4.887	4.547
Straniera	F	Stesso ambito	414	33	83	106	96	96
Straniera	F	Ambito diverso	194	14	46	51	43	40
Straniera	M	Stesso ambito	321	23	70	89	71	68
Straniera	M	Ambito diverso	186	19	38	51	42	36
Totale			1.115	89	237	297	252	240
Totale			26.292	5.366	5.559	5.441	5.139	4.787

I gruppi degli studenti con cittadinanza straniera sono quelli meno numerosi, tuttavia le numerosità sono sufficienti per la stima del **modello E**, che risulta la scelta migliore tra modelli simili che tenevano conto anche delle interazioni tra i regressori.

$$y_{it} = \alpha_i + \beta_i Time_{it} + \varepsilon_{it} \text{ con ipotesi } \varepsilon_{it} \sim N\left(0, \sigma_\varepsilon^2\right)$$

$$\alpha_i = \gamma_{00} + \gamma_{01} Foreign_i + \gamma_{02} Sex_i + \gamma_{03} Foreign_i * Ambit_i + \zeta_{0i}$$

$$\beta_i = \gamma_{10} + \gamma_{11} Foreign_i + \gamma_{12} Sex_i + \gamma_{13} Foreign_i * Ambit_i + \gamma_{14} Foreign_i * Ambit_i * Sex_i + \zeta_{1i}$$

$$\text{con ipotesi } \begin{pmatrix} \zeta_{0i} \\ \zeta_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{pmatrix} \right)$$

La variabile esplicativa *Ambit* è una variabile dicotomica che assume valore 0 se la scuola frequentata si trova nello stesso ambito di residenza dello studente, mentre assume valore 1 se gli ambiti della scuola e di residenza sono diversi. Le stime dei parametri sono riportate in Tabella 67.

Tabella 67 – Stima dei parametri: modello E

	Parametro	Errore standard	P-value
γ_{00} intercetta media ragazze italiane stesso ambito	0,894	0,008	0
γ_{10} pendenza media ragazze italiane stesso ambito	0,896	0,005	0
γ_{01} scostamento intercetta stranieri	-0,586	0,028	0

Tabella 67 – Stima dei parametri: modello E

	Parametro	Errore standard	P-value
γ_{11} scostamento pendenza stranieri	-0,200	0,016	0
γ_{02} scostamento intercetta maschi	-0,080	0,011	0
γ_{12} scostamento pendenza maschi	-0,089	0,007	0
γ_{03} scostamento intercetta stranieri ambito diverso	0,179	0,040	0
γ_{13} scostamento pendenza stranieri ambito diverso	-0,094	0,022	0
γ_{14} scostamento pendenza stranieri maschi ambito diverso	0,104	0,030	0,00006
σ_{ε}^2	0,0377		
σ_0^2	0,158		
σ_1^2	0,060		

Il modello stima che non vi sia differenza, in quanto a valore iniziale (al tempo 0) della variabile obiettivo, tra gli studenti italiani che frequentano una scuola dello stesso ambito e quelli che si spostano in ambito diverso, mentre che la stessa differenza sia significativa per i soli studenti stranieri. Il valore positivo del parametro indica che gli studenti stranieri che frequentano una scuola in un ambito diverso rispetto a quello di residenza hanno valori iniziali della variabile *Times promoted* più alti rispetto agli studenti stranieri che frequentano la scuola nello stesso ambito di residenza.

Il modello stima inoltre che non vi sia differenza, in media, tra gli studenti italiani che frequentano la scuola nello stesso ambito di residenza e gli stessi studenti che frequentano la scuola in ambito diverso per quanto riguarda il tasso di variazione della variabile obiettivo *Times promoted*. Il valore negativo del parametro γ_{13} indica che gli studenti stranieri che frequentano la scuola in ambito diverso rispetto a quello di residenza hanno un tasso di variazione di *Times promoted* in media peggiore (più basso) rispetto agli altri studenti stranieri, anche se il loro valore iniziale risultava migliore (si veda il parametro γ_{03}).

Non si può parlare di miglioramento nella bontà di adattamento in questo modello rispetto al precedente modello D (si può quindi concludere che la variabile *Ambit* non sia in realtà una variabile significativa); la statistica pseudo Rquadro conferma infatti che non vi è un forte miglioramento rispetto sempre al modello base B.

$pseudoR^2_0$ =0,088

$pseudoR^2_1$ =0,095

Un vantaggio di questi tipi di modelli è che si possono ottenere i valori stimati per alcuni individui tipici, così da poter confrontare gli individui con diversi valori dei regressori. Considerando il **modello D**, ritenuto quello ottimo per l'intero dataset, nei Grafici 53 e 54 sono rappresentate le medie stimate (con i rispettivi intervalli di confidenza) e quelle osservate per alcuni studenti tipo.

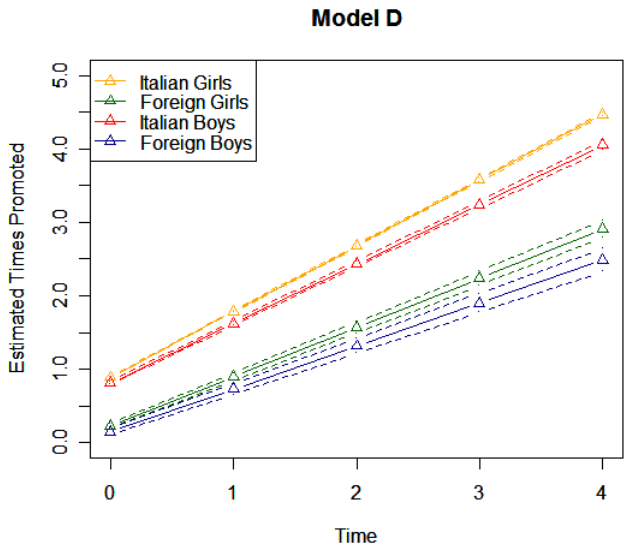


Grafico 53 – Distribuzione delle medie nel tempo, e dei relativi intervalli di confidenza, di individui tipo stimate dal modello D

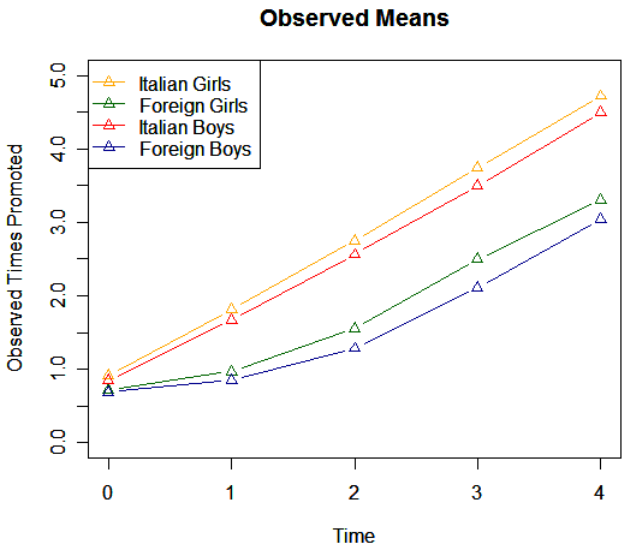


Grafico 54 – Distribuzione delle medie nel tempo di individui tipo osservate

Il **modello D** porta a stime migliori dei valori medi relativi agli studenti italiani rispetto alle stime dei valori medi degli studenti stranieri. Bisogna sempre considerare che gli studenti con cittadinanza non italiana sono in numero molto inferiore rispetto agli italiani, quindi la stima è meno precisa. Il modello stima al meglio le medie relative alle ragazze italiane; tuttavia riproduce anche la differenza negativa tra maschi e femmine, anche se, specie per i ragazzi, sottostima il valore iniziale ed anche quello relativo all'ultimo istante temporale considerato.

Il modello si adatta meglio ai dati relativi agli italiani, che rappresentano di gran lunga la maggior parte degli studenti del dataset e che mostrano avere un andamento nel tempo più vicino alla linearità; per quanto riguarda gli stranieri, il modello sottostima fortemente il valore iniziale e quello finale, mentre ben riproduce l'andamento intermedio. Vi è da notare che il modello ben riproduce due circostanze: innanzitutto che mentre gli italiani terminano mediamente la seconda classe dopo 2 anni di scuola, gli altri arrivano mediamente alla fine della prima classe nel medesimo tempo; poi che i cittadini non italiani arrivano in media, dopo 5 anni, a terminare la terza classe, mentre gli italiani giungono mediamente al termine almeno della quarta.

Come si evince dal Grafico 55, il modello sottostima le medie temporali della variabile obiettivo (i valori medi osservati stanno al di sopra dell'intervallo di confidenza stimato).

Guardando i valori stimati, si ha che relativamente a 1.033 valori (il 3,9% di tutti i valori) vi è una differenza tra stima e valore osservato di *Times promoted* maggiore di ±1,5 (se invece si considerano i soli studenti italiani, la percentuale si abbassa al 3,4%). Inoltre il 9% dei valori presenta l'analoga differenza pari a ±1 (considerando i soli studenti italiani, la stessa percentuale diventa del 7,9%).

Si può notare che gli intervalli di confidenza delle stime sono tra loro separati, a dimostrazione del fatto che la differenza tra i gruppi di individui identificati dai valori delle variabili esplicative è significativa.

I valori osservati si trovano leggermente al di sopra degli intervalli di confidenza stimati. Il modello si può dire che riproduca le differenze tra studenti appartenenti ai diversi gruppi sociali.

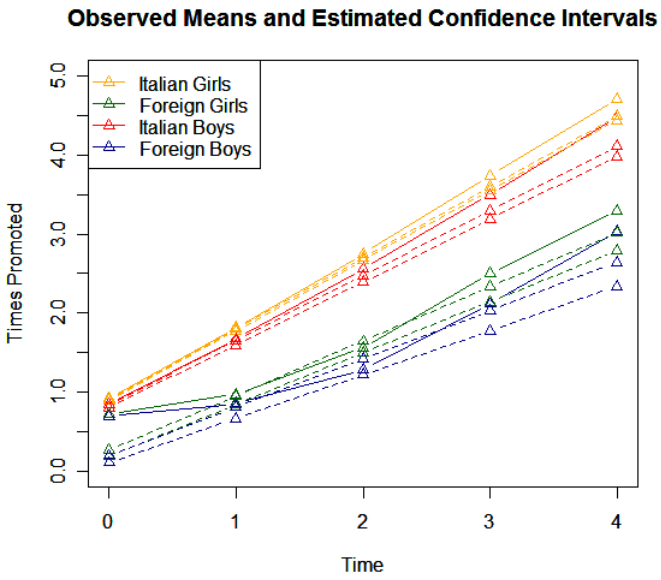


Grafico 55 – Medie osservate su gruppi di individui ed intervalli di confidenza stimati dal modello D

4.2.3 RICERCA DEL MODELLO OTTIMALE PER I SOLI STUDENTI ITALIANI

Un'altra estensione del modello è quella che tiene conto della tipologia di scuola frequentata dagli studenti. A causa del numero esiguo di studenti stranieri, per stimare il modello che tenesse conto della tipologia di scuola, è stato necessario considerare soltanto i 5.513 studenti con cittadinanza italiana.

In Tabella 68 è rappresentata la distribuzione di tali studenti nei diversi anni scolastici e tipi di scuola. Vi è da notare il contingente di studenti di ogni anno scolastico non è esattamente il medesimo rispetto a quello dell'anno precedente o del successivo, fatto dovuto alla mobilità degli studenti. Si nota in particolare l'aumento di studenti nel secondo anno considerato, dovuto all'ingresso nella scuola secondaria di secondo grado di coloro che sono stati bocciati al primo grado (soprattutto maschi che entrano in scuole diverse dal liceo).

Tabella 68 – Distribuzione negli anni scolastici del numero di studenti e degli istanti temporali nei diversi gruppi, identificati da valori diversi della variabile tipo di scuola

Cittadinanza	Sesso	Tipo di scuola	N. Studenti 2002/03	N. Studenti 2003/04	N. Studenti 2004/05	N. Studenti 2005/06	N. Studenti 2006/07
Italiana	F	Licei	1.455	1.453	1.444	1.416	1.326
Italiana	M	Licei	1.010	1.004	999	988	926
Italiana	F	Artistici	129	133	126	123	117
Italiana	M	Artistici	43	50	46	39	41
Italiana	F	Tecnici	546	546	527	497	477
Italiana	M	Tecnici	1.074	1.091	1.036	970	889
Italiana	F	Professionali	484	481	448	403	364
Italiana	M	Professionali	536	564	518	451	407
		Totale	5.277	5.322	5.144	4.887	4.547

Il **modello F**, ottenuto dopo una procedura di ottimizzazione che teneva in considerazione anche le interazioni tra i diversi regressori, è risultato essere quello nel seguito specificato.

$y_{it} = \alpha_i + \beta_i Time_{it} + \varepsilon_{it}$ con la consueta ipotesi $\varepsilon_{it} \sim N\left(0, \sigma_\varepsilon^2\right)$

$$\alpha_i = \gamma_{00} + \gamma_{04} Art_i + \gamma_{05} Voc_i + \gamma_{06} Tech_i + \gamma_{07} Sex_i Voc_i + \gamma_{08} Sex_i Tech_i + \zeta_{0i}$$
$$\beta_i = \gamma_{10} + \gamma_{11} Art_i + \gamma_{13} Voc_i + \gamma_{15} Tech_i + \gamma_{17} Sex_i Voc_i + \gamma_{18} Sex_i Tech_i + \zeta_{1i}$$

con la consueta ipotesi $\begin{pmatrix} \zeta_{0i} \\ \zeta_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{pmatrix} \right)$

Gli indicatori pseudo R-quadro, che confrontano il miglioramento ottenuto con il modello F, rispetto ad un modello base che non tiene conto della distinzione per tipologia di scuola, risultano, rispettivamente per intercetta e pendenza:

$pseudoR\ _0^2$ =0,1013

$pseudoR\ _1^2$ =0,1203

Vi è da notare che, nell'espressione precedente, ci si riferisce a un modello analogo al modello B già presentato, ma che tiene in considerazione i soli studenti italiani.

Le variabili esogene *Art*, *Voc* e *Tech* sono tutte dicotomiche ed assumono valore 1 solo se la scuola frequentata è, rispettivamente, un istituto artistico, un professionale oppure un tecnico (il riferimento è il liceo).

Con questo modello, la statistica R-quadro è risultata pari a 0,820 (il valore più alto dei modelli finora considerati). Ciò significa che la tipologia di scuola frequentata spiega molta parte della variabilità di *Times promoted*; occorre però considerare che in questo modello non vi sono gli studenti stranieri, che comunque risultavano essere la parte della popolazione che i modelli precedenti riuscivano a stimare in modo peggiore rispetto agli italiani. Nel modello F, i coefficienti dei regressori risultano tutti significativi (Tabella 69), con un p-value <0,001.

Tabella 69 – Stima dei parametri: modello F

	Parametro	Errore standard	P-value
γ_{00} intercetta media studenti dei licei	0,969	0,008	0
γ_{10} pendenza media studenti dei licei	0,938	0,005	0
γ_{04} scostamento intercetta studenti degli artistici	-0,243	0,030	0
γ_{11} scostamento pendenza studenti degli artistici	-0,091	0,018	0
γ_{05} scostamento intercetta studenti dei professionali	-0,246	0,019	0
γ_{13} scostamento pendenza studenti dei professionali	-0,178	0,012	0
γ_{06} scostamento intercetta studenti dei tecnici	-0,083	0,018	0
γ_{15} scostamento pendenza studenti dei tecnici	-0,062	0,011	0
γ_{07} scostamento intercetta maschi dei professionali	-0,139	0,024	0
γ_{17} scostamento pendenza maschi dei professionali	-0,077	0,015	0
γ_{08} scostamento intercetta maschi dei tecnici	-0,071	0,020	0

Tabella 69 – Stima dei parametri: modello F			
	Parametro	Errore standard	P-value
γ_{18} scostamento pendenza maschi dei tecnici	-0,101	0,012	0
σ_{ε}^2	0,0361		
σ_0^2	0,133		
σ_1^2	0,051		

Si può notare che non vi è una differenza significativa, tanto nel valore iniziale quanto nel tasso di variazione, tra maschi e femmine che frequentano i licei o gli istituti artistici, mentre l'analoga differenza risulta significativa soltanto negli istituti tecnici e professionali; ciò è anche stato confermato dalla stima del modello che ha come riferimento gli studenti che frequentano l'istituto artistico e nel quale il coefficiente dell'interazione tra la frequenza del liceo e il sesso non è risultato significativo.

Le medie stimate (con i relativi intervalli di confidenza stimati) dal modello F (ed osservate) per studenti tipo sono descritte in Tabella 70 e rappresentate in Grafico 56 e Grafico 57.

Tabella 70 – Medie stimate, con relativi intervalli di confidenza, dal modello F (medie osservate) della variabile obiettivo *Times Promoted*					
	2002/2003	2003/2004	2004/2005	2005/2006	2006/2007
Studenti italiani che frequentano il liceo	0,969 [0,961-0,977] (M 0,958 F 0,981)	1,907 [1,895-1,920] (M 1,915 F 1,932)	2,845 [2,828-2,863] (M 2,848 F 2,886)	3,784 [3,762-3,806] (M 3,781 F 3,881)	4,722 [4,695-4,749] (M 4,762 F 4,852)
Studenti italiani che frequentano l'istituto artistico	0,727 [0,689-0,765] (M 0,814 F 0,798)	1,574 [1,513-1,635] (M 1,500 F 1,624)	2,422 [2,338-2,505] (M 2,326 F 2,571)	3,269 [3,162-3,376] (M 3,436 F 3,528)	4,117 [3,987-4,246] (M 4,219 F 4,504)
Ragazze italiane che frequentano l'istituto tecnico	0,887 [0,787-0,862] (0,899)	1,762 [1,720-1,805] (1,788)	2,638 [2,654-2,747] (2,727)	3,514 [3,587-3,690] (3,704)	4,389 [4,520-4,633] (4,696)
Ragazzi italiani che frequentano l'istituto tecnico	0,815 [0,695-0,811] (0,837)	1,590 [1,515-1,665] (1,640)	2,365 [2,335-2,519] (2,502)	3,140 [3,155-3,374] (3,416)	3,914 [3,975-4,228] (4,451)
Ragazze italiane che frequentano l'istituto professionale	0,723 [0,696-0,750] (0,756)	1,484 [1,440-1,527] (1,547)	2,244 [2,184-2,304] (2,397)	3,005 [2,928-3,081] (3,347)	3,765 [3,671-3,858] (4,324)
Ragazzi italiani che frequentano l'istituto professionale	0,584 [0,533-0,635] (0,660)	1,268 [1,186-1,350] (1,319)	1,951 [1,837-2,065] (2,139)	2,635 [2,489-2,780] (3,064)	3,318 [3,141-3,495] (3,988)

Confrontando le stime ed i relativi intervalli di confidenza, si nota che (Grafici 58, 59, 60, 61) esiste una differenza significativa (e gli intervalli di confidenza stimati non hanno punti in comune) tra maschi e femmine negli istituti tecnici e professionali (nei tecnici, soltanto al primo anno, si ha una leggera sovrapposizione). Esiste inoltre una differenza significativa tra gli studenti dei licei e quelli degli altri istituti. Gli intervalli di confidenza delle stime sono inoltre nettamente separati tra maschi di istituti tecnici e professionali ed anche tra femmine degli istituti tecnici e professionali. Si verifica invece una sovrapposizione degli intervalli di confidenza tra studenti degli istituti artistici e studenti degli istituti tecnici. Occorre a tale proposito tenere conto del fatto che proprio le stime che riguardano gli istituti artistici non sono da ritenersi affidabili data l'esigua numerosità.

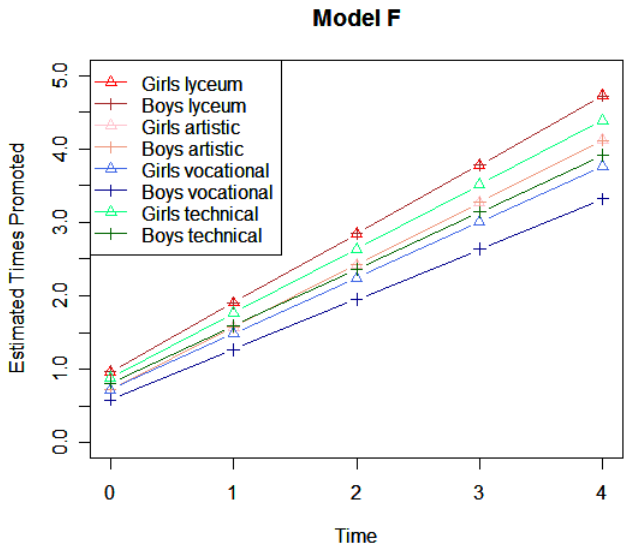


Grafico 56 - Distribuzione delle medie nel tempo di individui tipo stimate dal modello F

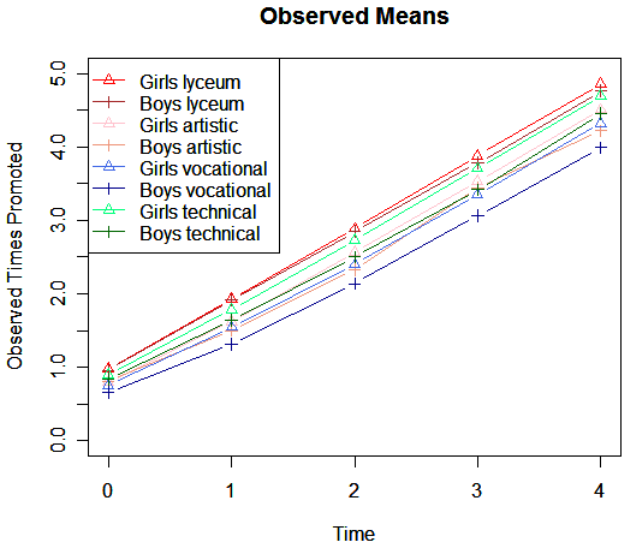


Grafico 57 - Distribuzione delle medie nel tempo di individui tipo osservate

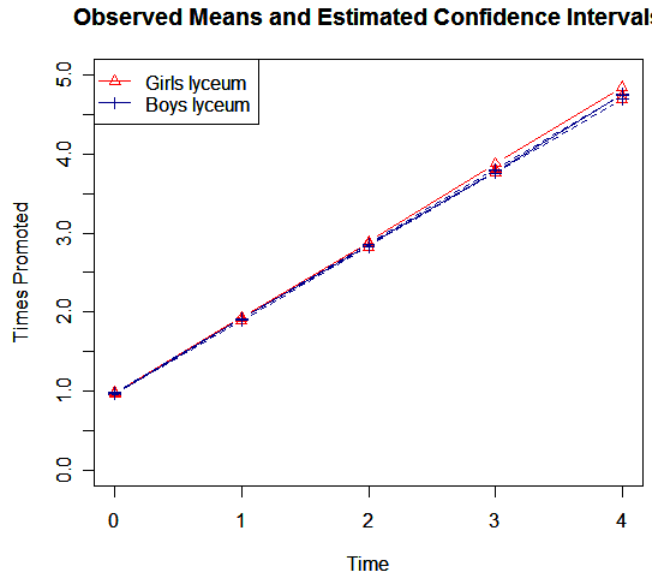


Grafico 58 – Valori osservati e intervalli di confidenza stimati - licei

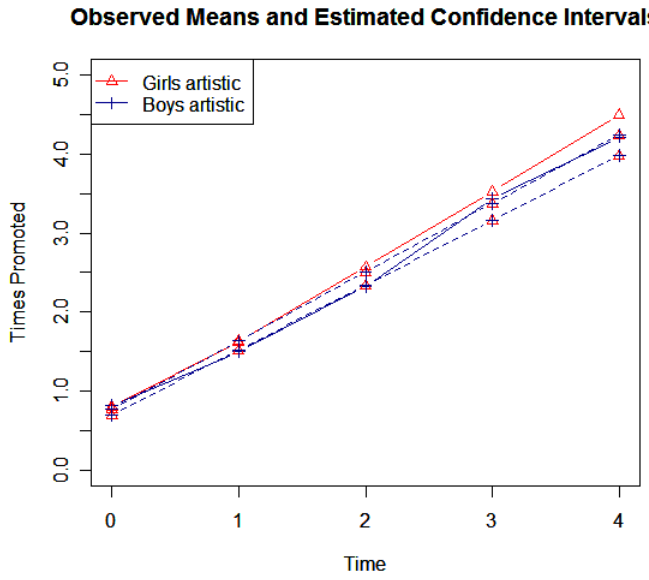


Grafico 59 – Valori osservati e intervalli di confidenza stimati - artistici

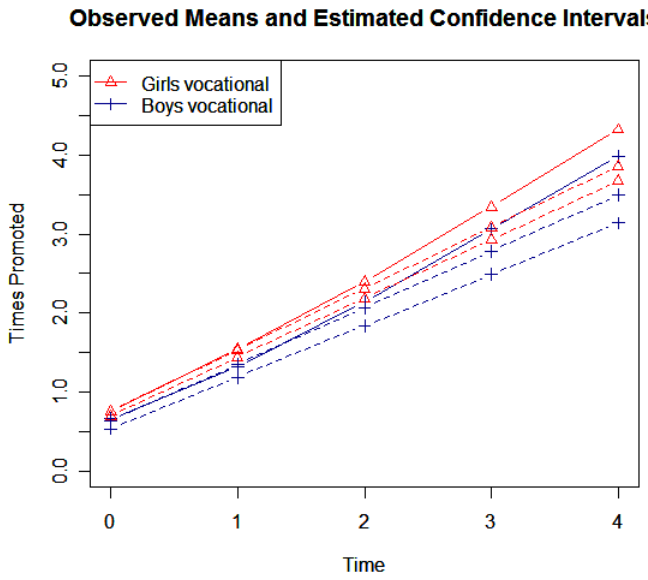


Grafico 60 – Valori osservati e intervalli di confidenza stimati - professionali

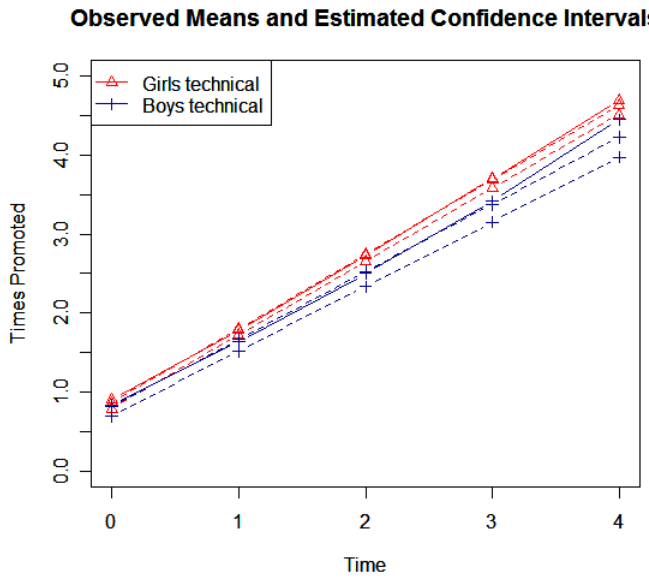


Grafico 61 – Valori osservati e intervalli di confidenza stimati - tecnici

Dal confronto grafico, si può notare che il modello tende a sovrastimare le differenze, tuttavia riproduce fedelmente alcune evidenti relazioni: il percorso dei frequentanti i licei è in assoluto il migliore, mentre quello dei ragazzi dei professionali risulta in assoluto il peggiore; le ragazze dei professionali hanno un percorso di poco peggiore rispetto ai ragazzi dei tecnici e le ragazze dei tecnici hanno un percorso di poco peggiore rispetto ai frequentanti il liceo. Il modello tende inoltre a sottostimare i valori della variabile obiettivo: si può notare che i valori osservati rimangono lievemente al di sopra dell'estremo superiore dell'intervallo di confidenza stimato. Le stime migliori, da questo punto di vista) risultano quelle relative ai licei e ai tecnici.

Per quanto riguarda la verifica delle assunzioni del modello, purtroppo la situazione non si è modificata di molto, rispetto ai modelli precedenti. Le rappresentazioni grafiche degli errori di primo livello sono riportate nei Grafici 62, 63 e 64.

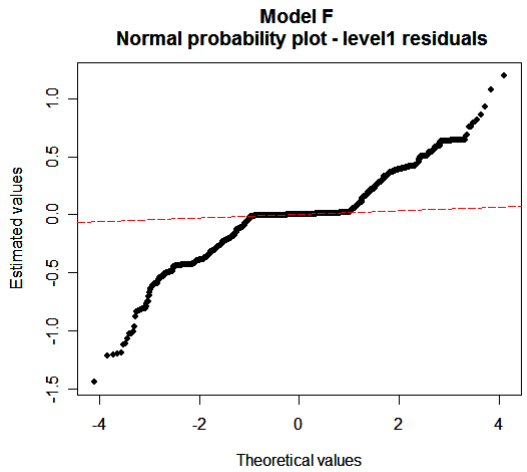


Grafico 62 – *Normal probability plot* delle stime degli errori di primo livello del modello F

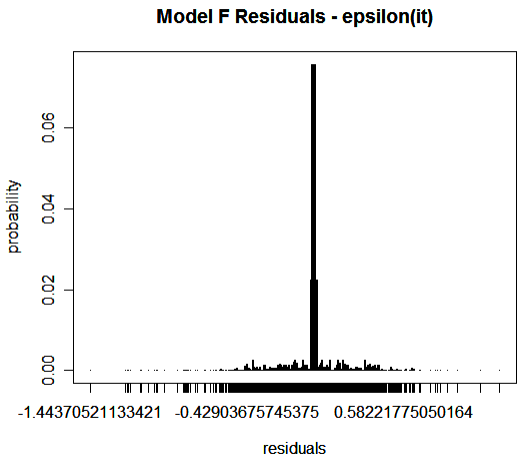


Grafico 63 – Distribuzione di probabilità delle stime degli errori di primo livello del modello F

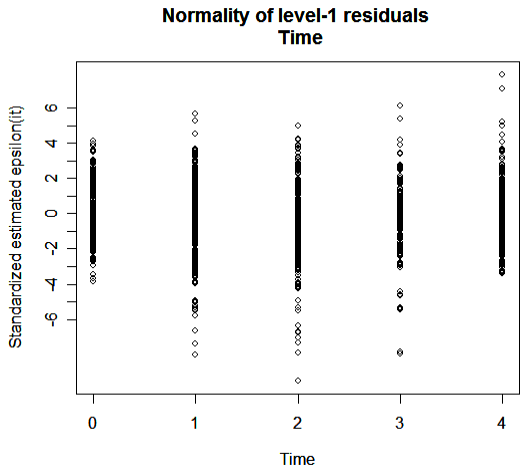


Grafico 64 – Distribuzione nei diversi intervalli temporali delle stime degli errori standardizzati

Il 9% degli errori standardizzati presenta valori <-2 oppure >2. La statistica test di Kolmogorov Smirnov è 0,35.

La normalità distributiva non può essere ipotizzata.

Gli errori di secondo livello sono rappresentati nei Grafici 65 e 66.

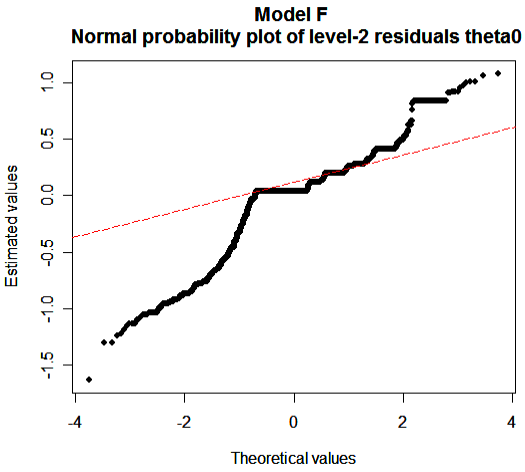


Grafico 65 – *Normal probability plot* degli errori di secondo livello relativi all'intercetta, modello F

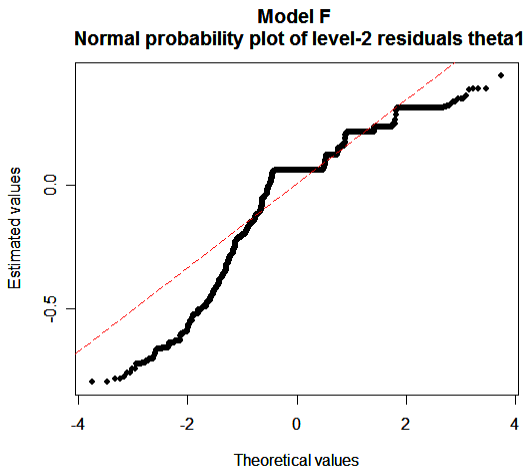


Grafico 66 – *Normal probability plot* degli errori di secondo livello relativi alla pendenza, modello F

Il 9% degli errori standardizzati (per l'intercetta) ha valori <-2 oppure >2. La statistica test di Kolmogorov Smirnov risulta pari a 0,31. Il 7% degli errori standardizzati (per la pendenza) risulta pari a >2 oppure <-2. La statistica test di Kolmogorov Smirnov risulta pari a 0,37. La normalità distributiva non può assumersi.

Se fosse verificata l'ipotesi di omoschedasticità, le differenze tra le varianze nei diversi istanti temporali non dovrebbe essere significativa. Le varianze degli errori nei diversi anni scolastici considerati sono: [0,01551; 0,02966; 0,03134; 0,01934; 0,01812]. L'omoschedasticità degli errori in questo caso può dirsi ipotesi realistica, anche se vi è una leggera eteroschedasticità tra i tempi 1 e 2. Si noti che i valori delle varianze sono leggermente maggiori in questo modello rispetto al modello B, senza regressori.

Un'altra ipotesi del modello è l'assenza di correlazione tra gli errori nei diversi istanti temporali considerati. I valori di autocorrelazione sono: [cor(0,1); cor(1,2); cor(2,3); cor(3,4)] = [-0,47; -0,03; -0,12; -0,62]. Considerando che si parla di alta correlazione se il valore della correlazione stessa è vicina ad 1, tali valori non sono molto alti, specialmente alcuni di questi (le correlazioni tra gli errori degli istanti 1 e 2 e degli istanti 2 e 3). I valori della correlazione sono leggermente più bassi di quelli riscontrati nel modello B.

In seguito, si è proceduto con la stima di un modello che tenesse conto anche del confronto tra ambito di residenza dello studente e ambito in cui è collocata la scuola frequentata. Tale modello mostra che la variabile dicotomica che esprime questo confronto non è significativa, ad eccezione della sua influenza sul tasso di variazione della variabile obiettivo per quegli studenti che frequentano gli istituti professionali: gli studenti di questi istituti frequentanti una scuola in ambito diverso rispetto a quello di residenza presentano un tasso di variazione della variabile obiettivo leggermente più alto degli stessi studenti frequentanti una scuola nello stesso ambito di residenza.

Si è pertanto giunti alla stima del **modello G**, che tiene conto dell'effetto, seppur limitato, della variabile esogena rappresentante l'ambito diverso tra scuola frequentata e residenza (*Ambit*).

$$y_{it} = \alpha_i + \beta_i Time_{it} + \varepsilon_{it} \quad \text{con ipotesi } \varepsilon_{it} \sim N\left(0, \sigma_\varepsilon^2\right)$$

$$\alpha_i = \gamma_{00} + \gamma_{04} Art_i + \gamma_{05} Voc_i + \gamma_{06} Tech_i + \gamma_{07} Sex_i Voc_i + \gamma_{08} Sex_i Tech_i + \zeta_{0i}$$

$$\beta_i = \gamma_{10} + \gamma_{13} Voc_i + \gamma_{15} Tech_i + \gamma_{17} Sex_i Voc_i + \gamma_{18} Sex_i Tech_i + \gamma_{19} Ambit_i Voc_i + \zeta_{1i}$$

$$\text{con ipotesi } \begin{pmatrix} \zeta_{0i} \\ \zeta_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{pmatrix}\right).$$

Nel modello G sono stati inclusi soltanto i regressori con coefficienti significativi (p-value <0,001), quindi la variabile *Ambit*, che esprime il diverso ambito della residenza dello studente rispetto a quello della scuola frequentata, è stata introdotta soltanto nell'interazione tra tempo e istituto professionale (gli studenti che frequentano gli istituti professionali nello stesso ambito di residenza hanno un tasso di variazione di *Times promoted* diverso rispetto agli studenti che frequentano istituti professionali situati in ambito diverso da quello di residenza). La variabile indicante il sesso è invece risultata significativa soltanto per quegli studenti che frequentano gli istituti tecnici e professionali, tanto per spiegare l'intercetta che la pendenza.

Tabella 71 – Stima dei parametri: modello G

	Parametro	Errore standard	P-value
γ_{00} intercetta media studenti dei licei	0,966	0,008	0
γ_{10} pendenza media studenti dei licei	0,932	0,005	0
γ_{04} scostamento intercetta studenti degli artistici	-0,203	0,029	0
γ_{05} scostamento intercetta studenti dei professionali	-0,242	0,019	0
γ_{13} scostamento pendenza studenti dei professionali	-0,198	0,013	0
γ_{06} scostamento intercetta studenti dei tecnici	-0,080	0,018	0
γ_{15} scostamento pendenza studenti dei tecnici	-0,056	0,011	0
γ_{07} scostamento intercetta maschi dei professionali	-0,139	0,024	0
γ_{17} scostamento pendenza maschi dei professionali	-0,081	0,015	0
γ_{08} scostamento intercetta maschi dei tecnici	-0,071	0,020	0

Tabella 71 – Stima dei parametri: modello G			
	Parametro	Errore standard	P-value
γ_{18} scostamento pendenza maschi dei tecnici	-0,101	0,012	0
γ_{19} scostamento pendenza studenti dei professionali ambito diverso	0,056	0,009	0
σ_{ε}^2	0,0361		
σ_0^2	0,133		
σ_1^2	0,051		

Il modello G porta ad un leggerissimo miglioramento rispetto al modello F, se entrambi confrontati con il modello base senza distinzione per tipologia di scuola.

$pseudoR_0^2$=0,1004

$pseudoR_1^2$=0,1171

I valori dello pseudo R-quadro sono, in questo caso, leggermente inferiori, tuttavia l'introduzione della ulteriore variabile non si può dire abbia apportato un miglioramento significativo. Si può quindi dire che il modello ottimale sia in realtà il **modello F**.

Considerando tale modello F, contenente le interazioni tra le variabili esogene, l'interpretazione data ai coefficienti suggerisce che risulta in realtà difficoltoso attribuire un significato preciso a ciascun coefficiente, quindi spiegare l'effetto di ogni esogena sulla variabile obiettivo. Per capire meglio il significato dei singoli coefficienti, tuttavia, si può ricorrere ad un escamotage, come nel seguito mostrato.

Considerando allora il modello F, è possibile determinare gli istanti temporali nei quali le variabili esplicative influiscono in modo significativo sui valori della variabile obiettivo, mediate dai fattori latenti.

Si può così scrivere la relazione tra *Times promoted* e i regressori come condizionata dai valori del tempo.

$$\hat{y}_{t|\lambda_t} = \gamma_{00} + \gamma_{04} Art_i + \gamma_{05} Voc_i + \gamma_{06} Tech_i +$$
$$+ \gamma_{10}\lambda_t + \gamma_{11} Art_i\lambda_t + \gamma_{13} Voc_i\lambda_t + \gamma_{15} Tech_i\lambda_t + \left(\gamma_{07} Voc_i + \gamma_{08} Tech_i + \gamma_{17} Voc_i\lambda_t + \gamma_{18} Tech_i\lambda_t\right) Sex_i$$

Data, per esempio, la variabile *Sex* come regressore obiettivo, $\left(\gamma_{07} Voc_i + \gamma_{08} Tech_i + \gamma_{17} Voc_i\lambda_t + \gamma_{18} Tech_i\lambda_t\right)$ può essere visto come la pendenza della regressione della variabile Y sulla variabile *Sex*, dato un certo valore del tempo λ_t, mentre tutti gli altri termini dell'equazione possono essere visti come intercetta di tale regressione. Quindi si può dire che, dati valori fissi degli altri predittori, quella descritta è la relazione lineare tra *Times promoted* e *Sex*. È anche possibile ora testare la significatività della relazione tra la Y e *Sex* per ogni valore del tempo.

Si ha così che, posti:

$$\hat{\omega}_0 = \gamma_{00} + \gamma_{04} Art_i + \gamma_{05} Voc_i + \gamma_{06} Tech_i + \gamma_{10}\lambda_t + \gamma_{11} Art_i\lambda_t + \gamma_{13} Voc_i\lambda_t + \gamma_{15} Tech_i\lambda_t$$

$$\hat{\omega}_1 = \left(\gamma_{07} Voc_i + \gamma_{08} Tech_i + \gamma_{17} Voc_i\lambda_t + \gamma_{18} Tech_i\lambda_t\right),$$

queste espressioni rappresentano le stime di intercetta e pendenza, per ogni istante temporale, della variabile *Times promoted* regressa sulla variabile *Sex* (siccome questa è dicotomica, l'intercetta è il valore medio nei diversi istanti temporali, relativo alle femmine, della variabile *Times Promoted* nel tempo, mentre la pendenza è la differenza media, nei diversi istanti temporali, tra maschi e femmine).

Si possono anche calcolare le stime degli errori standard, per ognuna delle combinazioni dei valori dei regressori.

Si è riscontrato che tutte le combinazioni dei valori portano a intercetta e pendenza significative, quindi si può dire che la relazione tra *Times promoted* e il sesso per gli studenti sia significativa in tutti gli istanti temporali. Il risultato prima mostrato dai grafici (cioè che i ragazzi conseguono risultati in generale peggiori rispetto alle ragazze) è ora confermato dal modello. Tuttavia, risulta che la stima per gli studenti degli istituti artistici e dei licei ha il valore $\hat{\omega}_1 = 0$ in tutti i tempi, ciò vuol dire che in tali istituti in realtà la differenza tra maschi e femmine non esiste. Occorre però tener conto anche della bassa numerosità degli studenti degli artistici se confrontata con quelli dei licei. Al termine di tali valutazioni, è corretto dire che la relazione tra il sesso di appartenenza e la variabile obiettivo è significativa soltanto negli istituti tecnici (dove il valore di $\hat{\omega}_1$ è, nei diversi istanti temporali, pari a [-0,071; -0,172; -0,341; -0,442; -0,543]) e nei professionali (dove il valore di $\hat{\omega}_1$ è, nei diversi istanti temporali, pari a [-0,139; -0,216; -0,293; -0,37; -0,447], quindi qui la differenza tra maschi e femmine è maggiormente accentuata rispetto ai tecnici).

La differenza tra maschi e femmine dei tecnici nei diversi istanti temporali è stimata essere mediamente[70] attorno al 9%. La differenza tra maschi e femmine dei professionali nei diversi istanti temporali è stimata essere mediamente del 12%.

Ripercorrendo i valori della Tabella 70, si può notare che in classe prima i ragazzi sembrano avere un andamento della variabile obiettivo simile a quello delle ragazze, ma poi hanno un peggioramento nel tempo. La differenza media negativa, inoltre, tra ragazzi e ragazze è maggiormente accentuata negli istituti professionali che non nei tecnici e inoltre aumenta nel tempo; il modello, come si può anche vedere dal confronto di Grafico 56 e Grafico 57, sovrastima la differenza tra maschi e femmine che frequentano gli istituti professionali, mentre si adatta maggiormente a rappresentare la differenza tra maschi e femmine che frequentano gli istituti tecnici.

Passando ora ad analizzare la matrice delle covarianze nell'espressione compatta[71], si ha che gli errori del modello risultano essere:

$r_{it} = \varepsilon_{it} + \zeta_{0i} + \zeta_{1i} Time_t$.

Se si sostituisce r_{it} nel modello, si ottiene un generico modello di regressione multipla, con una variabile esplicativa "particolare", che è il Tempo, e con le interazioni tra le variabili esplicative (il tempo e le altre variabili). L'assunzione sulla distribuzione di questi errori è quella normale, ma la matrice di varianze e covarianze non è diagonale, bensì diagonale a blocchi.

[70] Media geometrica delle differenze nei diversi istanti temporali.

[71] Singer, Willett "Applied longitudinal data analysis – modelling change and event occurrence"

Questo è perché si assume che gli errori siano eteroschedastici e che siano correlati nel tempo entro ogni individuo, ma si assume anche che le covarianze e le varianze siano costanti per tutti gli individui per ogni istante temporale, mentre che le covarianze tra gli errori dei diversi individui siano =0.
Si può quindi ricavare l'espressione nel seguito descritta.

$$r_{it} \sim N\left(0, \begin{bmatrix} \Sigma_r & 0 & 0 & 0 \\ 0 & \Sigma_r & 0 & 0 \\ 0 & 0 & . & 0 \\ 0 & 0 & 0 & \Sigma_r \end{bmatrix}\right) \quad \text{con} \quad \Sigma_r = \begin{bmatrix} \sigma^2_{r_1} & \sigma_{r_1 r_2} & . & \sigma_{r_1 r_T} \\ \sigma_{r_1 r_2} & \sigma^2_{r_2} & . & \sigma_{r_2 r_T} \\ . & . & . & . \\ \sigma_{r_T r_1} & \sigma_{r_T r_2} & . & \sigma^2_{r_T} \end{bmatrix}.$$

Le varianze stimate degli errori *composite* nei diversi istanti temporali sono le seguenti:

$$\sigma^2_{r_t} = \text{var}\left(\varepsilon_{it} + \zeta_{0i} + \zeta_{1t} Time_t\right) = \sigma^2_\varepsilon + \sigma^2_0 + 2\sigma_{01} Time_t + \sigma^2_1 Time^2_t$$

In particolare, nel modello F, queste stime risultano:
[0,169; 0,290; 0,513; 0,839; 1,266].
È ora possibile vedere la medesima equazione da un altro punto di vista:

$$\sigma^2_{r_t} = \left(\sigma^2_\varepsilon + \frac{\sigma^2_0 \sigma^2_1 - \sigma^2_{01}}{\sigma^2_1}\right) + \sigma^2_1 \left(Time_t + \frac{\sigma_{01}}{\sigma^2_1}\right)^2;$$

questa espressione mostra che il LCM assume una relazione quadratica tra le varianze degli errori e il tempo. Ciò non è sempre verificato nella pratica: questo è un primo problema del modello. Si ha che nel caso in cui tutti i componenti di varianza di secondo livello sono prossimi a 0 (cioè quando i regressori di secondo livello spiegano molta parte della variabilità tra gli individui in quanto a intercetta e pendenza), la matrice di varianze e covarianze degli errori è prossima ad essere omoschedastica, con tutte le varianze vicine a σ^2_ε. Nel caso in cui σ^2_1 e σ_{01} siano entrambi prossimi a 0 (cioè quando la pendenza non differisce molto tra i diversi individui), la matrice di varianze e covarianze degli errori è ancora prossima ad essere omoschedastica, con tutte le varianze vicine a $\left(\sigma^2_\varepsilon + \sigma^2_0\right)$. In questi due casi, la dipendenza temporale delle varianze è molto debole. Quello in esame non fa parte di alcuno di tali due casi. Nel presente studio, come spesso accade, la differenza relativa tra la varianza riferita alla pendenza e la covarianza porta le varianze stimate a raggiungere il loro minimo al di fuori dell'intervallo di tempo considerato (0,169 non è il minimo).
Le stime delle covarianze degli errori *composite,* nei diversi istanti temporali, sono:

$$\sigma^2_{r_t r_s} = \sigma^2_0 + \sigma_{01}\left(Time_t + Time_s\right) + \sigma^2_1 Time_t Time_s$$

mentre $\rho_{r_t r_s} = \sigma_{r_t r_s} \Big/ \sqrt{\sigma^2_{r_t} \sigma^2_{r_s}}$ rappresenta la correlazione.

Siccome non è ovvio che gli errori della forma compatta specificati siano appropriati in modo uniforme ai dati osservati, è possibile specificare una struttura delle covarianze alternativa.
Dopo aver ipotizzato strutture alternative, si può valutare la loro bontà di adattamento attraverso i metodi standard come AIC e BIC. Ogni nuovo modello avrà identici effetti fissi, ma una diversa struttura delle covarianze degli errori.

Uno sguardo alle stime dei valori per la variabile *Times promoted* da parte del modello F mostra che circa l'11% dei valori assoluti delle differenze tra i valori osservati e quelli stimati sono superiori a 1, mentre circa lo 0,7% sono superiori a 2. In particolare, l'1,6% delle differenze assolute relative al tempo 0, il 7,1% delle differenze assolute relative al tempo 1, il 12% delle differenze assolute relative al tempo 2, il 13,1% delle differenze assolute relative al tempo 3 e il 25,5% delle differenze assolute relative al tempo 4 sono superiori a 1. Inoltre, il 4,6% delle differenze assolute relative agli studenti del liceo, il 9,1% delle differenze assolute relative agli studenti degli artistici, il 25,1% delle differenze assolute relative agli studenti dei professionali ed il 14,2% delle differenze assolute relative agli studenti dei tecnici risultano superiori a 1.

È possibile ipotizzare altre strutture della matrice di varianze e covarianze degli errori; nel seguito viene modificato il modello F sulla base delle diverse assunzioni su tale matrice.

La matrice di varianze e covarianze **non strutturata** è quella che non pone alcuna condizione sulla struttura, quindi gli elementi assumono i valori che i dati osservati richiedono. Un suo vantaggio è quello di avere la *deviance statistic* sempre inferiore rispetto ad ogni altro modello basato su struttura diversa (mantenendo lo stesso insieme di effetti fissi), proprio perché non vengono imposte restrizioni, quindi il suo adattamento ai dati risulta sempre il migliore. Un suo svantaggio è invece quello di avere troppi parametri incogniti, specialmente se gli istanti temporali non sono pochi. Nel caso in esame, gli istanti temporali considerati sono 5, perciò i parametri incogniti sono 15 (5 varianze e 10 covarianze). Quindi si ottiene:

$$\Sigma_r unstr = \begin{bmatrix} 0{,}138 & 0{,}171 & 0{,}197 & 0{,}227 & 0{,}263 \\ 0{,}171 & 0{,}317 & 0{,}368 & 0{,}427 & 0{,}495 \\ 0{,}197 & 0{,}368 & 0{,}545 & 0{,}634 & 0{,}733 \\ 0{,}227 & 0{,}427 & 0{,}634 & 0{,}833 & 0{,}967 \\ 0{,}263 & 0{,}495 & 0{,}733 & 0{,}967 & 1{,}193 \end{bmatrix} \text{ e } \mathrm{P}_r unstr = \begin{bmatrix} 1{,}000 & 0{,}815 & 0{,}719 & 0{,}669 & 0{,}647 \\ 0{,}815 & 1{,}000 & 0{,}886 & 0{,}830 & 0{,}804 \\ 0{,}719 & 0{,}886 & 1{,}000 & 0{,}941 & 0{,}910 \\ 0{,}669 & 0{,}830 & 0{,}941 & 1{,}000 & 0{,}970 \\ 0{,}647 & 0{,}804 & 0{,}910 & 0{,}970 & 1{,}000 \end{bmatrix}.$$

La matrice di varianze e covarianze degli errori **simmetrica composta** ipotizza, invece, l'omoschedasticità degli elementi sulla diagonale principale (le varianze sono tutte uguali a $\left(\sigma^2 + \sigma_1^2\right)$) ed una covarianza costante per tutte le coppie di errori. Questo è un caso particolare del modello standard, dove vi è una bassa variazione residua nell'intercetta vera delle traiettorie. In questo caso, vi sono soltanto due parametri incogniti, σ^2 e σ_1^2. Utilizzando come base il modello F, le matrici delle covarianze e delle correlazioni diventano:

$$\Sigma_r csym = \begin{bmatrix} 0{,}415 & 0{,}305 & 0{,}305 & 0{,}305 & 0{,}305 \\ 0{,}305 & 0{,}415 & 0{,}305 & 0{,}305 & 0{,}305 \\ 0{,}305 & 0{,}305 & 0{,}415 & 0{,}305 & 0{,}305 \\ 0{,}305 & 0{,}305 & 0{,}305 & 0{,}415 & 0{,}305 \\ 0{,}305 & 0{,}305 & 0{,}305 & 0{,}305 & 0{,}415 \end{bmatrix} \text{ e } \mathrm{P}_r csym = \begin{bmatrix} 1{,}000 & 0{,}733 & 0{,}733 & 0{,}733 & 0{,}733 \\ 0{,}733 & 1{,}000 & 0{,}733 & 0{,}733 & 0{,}733 \\ 0{,}733 & 0{,}733 & 1{,}000 & 0{,}733 & 0{,}733 \\ 0{,}733 & 0{,}733 & 0{,}733 & 1{,}000 & 0{,}733 \\ 0{,}733 & 0{,}733 & 0{,}733 & 0{,}733 & 1{,}000 \end{bmatrix}.$$

La matrice di varianze e covarianze degli errori **simmetrica eterogenea** è una variazione del tipo precedente di matrice, ma qui gli elementi sulla diagonale principale sono diversi (eteroschedasticità) ed anche le covarianze sono poste diverse tra loro; in particolare, le covarianze sono il prodotto delle deviazioni standard degli errori e di un parametro di correlazione costante ρ, con valore sempre compreso nell'intervallo [0, 1] (le covarianze sono della forma $\sigma_t \sigma_s \rho$). I parametri incogniti risultano quindi 6, uno per ciascun tempo (varianze) più uno per

il parametro di correlazione ρ. Al di fuori della diagonale principale della matrice di correlazione, tutti i valori sono uguali al parametro di correlazione (0,819). Si hanno quindi:

$$\Sigma_r hcsym = \begin{bmatrix} 0{,}159 & 0{,}181 & 0{,}228 & 0{,}272 & 0{,}308 \\ 0{,}181 & 0{,}309 & 0{,}318 & 0{,}379 & 0{,}430 \\ 0{,}228 & 0{,}318 & 0{,}488 & 0{,}477 & 0{,}541 \\ 0{,}272 & 0{,}379 & 0{,}477 & 0{,}694 & 0{,}645 \\ 0{,}308 & 0{,}430 & 0{,}541 & 0{,}645 & 0{,}893 \end{bmatrix} \text{ e } \mathrm{P}_r hcsym = \begin{bmatrix} 1{,}000 & 0{,}819 & 0{,}819 & 0{,}819 & 0{,}819 \\ 0{,}819 & 1{,}000 & 0{,}819 & 0{,}819 & 0{,}819 \\ 0{,}819 & 0{,}819 & 1{,}000 & 0{,}819 & 0{,}819 \\ 0{,}819 & 0{,}819 & 0{,}819 & 1{,}000 & 0{,}819 \\ 0{,}819 & 0{,}819 & 0{,}819 & 0{,}819 & 1{,}000 \end{bmatrix}.$$

La matrice di varianze e covarianze degli errori **autoregressiva del prim'ordine**, invece, ha una struttura autoregressiva del prim'ordine, con omoschedasticità degli elementi sulla diagonale principale; ha inoltre covarianze identiche nelle bande parallele alla diagonale principale. Le covarianze sono poste come il prodotto della varianza residua e di un parametro di correlazione degli errori ρ, sempre compreso nell'intervallo [0, 1]. Questo parametro ha la potenza 1 nella prima banda parallela alla diagonale principale, la potenza 2 nella seconda banda e così via. Allontanandosi dalla diagonale principale, le covarianze diminuiscono. Al di fuori della diagonale principale della matrice di correlazione, infatti, i valori sono uguali al parametro di correlazione (nella prima banda parallela alla diagonale principale), al suo quadrato (nella seconda banda) e così via. I parametri incogniti sono 2: la varianza costante (0,447) e il parametro di correlazione (0,879). Quindi, nel caso in esame:

$$\Sigma_r AR1 = \begin{bmatrix} 0{,}447 & 0{,}393 & 0{,}346 & 0{,}304 & 0{,}267 \\ 0{,}393 & 0{,}447 & 0{,}393 & 0{,}346 & 0{,}304 \\ 0{,}346 & 0{,}393 & 0{,}447 & 0{,}393 & 0{,}346 \\ 0{,}304 & 0{,}346 & 0{,}393 & 0{,}447 & 0{,}393 \\ 0{,}267 & 0{,}304 & 0{,}346 & 0{,}393 & 0{,}447 \end{bmatrix} \text{ e } \mathrm{P}_r AR1 = \begin{bmatrix} 1{,}000 & 0{,}879 & 0{,}773 & 0{,}680 & 0{,}597 \\ 0{,}879 & 1{,}000 & 0{,}879 & 0{,}773 & 0{,}680 \\ 0{,}773 & 0{,}879 & 1{,}000 & 0{,}879 & 0{,}773 \\ 0{,}680 & 0{,}773 & 0{,}879 & 1{,}000 & 0{,}879 \\ 0{,}597 & 0{,}680 & 0{,}773 & 0{,}879 & 1{,}000 \end{bmatrix}.$$

La matrice di varianze e covarianze degli errori **autoregressiva eterogenea** ha elementi diversi (eteroschedasticità) sulla diagonale principale e covarianze diverse anche all'interno della stessa banda parallela alla diagonale principale. Il parametro di correlazione è moltiplicato per il prodotto delle deviazioni standard degli errori. I parametri incogniti, nel caso in esame, sono 6: le 5 varianze e il parametro di correlazione. La struttura delle bande è visibile nella matrice di correlazione. Tale matrice utilizza ulteriori gradi di libertà, tuttavia è maggiormente flessibile rispetto alle precedenti. Le matrici delle varianze e covarianze e delle correlazioni, sempre tenendo come riferimento il modello F, diventano:

$$\Sigma_r hAR1 = \begin{bmatrix} 0{,}175 & 0{,}230 & 0{,}243 & 0{,}237 & 0{,}233 \\ 0{,}230 & 0{,}374 & 0{,}395 & 0{,}386 & 0{,}380 \\ 0{,}243 & 0{,}395 & 0{,}519 & 0{,}506 & 0{,}499 \\ 0{,}237 & 0{,}386 & 0{,}506 & 0{,}615 & 0{,}606 \\ 0{,}233 & 0{,}380 & 0{,}499 & 0{,}606 & 0{,}741 \end{bmatrix} \text{ e } \mathrm{P}_r hAR1 = \begin{bmatrix} 1{,}000 & 0{,}897 & 0{,}804 & 0{,}721 & 0{,}647 \\ 0{,}897 & 1{,}000 & 0{,}897 & 0{,}804 & 0{,}721 \\ 0{,}804 & 0{,}897 & 1{,}000 & 0{,}897 & 0{,}804 \\ 0{,}721 & 0{,}804 & 0{,}897 & 1{,}000 & 0{,}897 \\ 0{,}647 & 0{,}721 & 0{,}804 & 0{,}897 & 1{,}000 \end{bmatrix}.$$

Risulta ora possibile confrontare la bontà di adattamento dei modelli così specificati attraverso gli indicatori già utilizzati, AIC e BIC (Tabella 72).

Tabella 72 – Confronto dei diversi modelli costruiti in base ad una diversa struttura della matrice di varianze e covarianze degli errori.

Matrice delle covarianze degli errori del modello	Gradi di libertà	AIC	BIC	Log Likelihood
Standard		17.203	17.333	-8.585
Non strutturata	27	15.421	15.640	-7.683
Simmetrica composta	14	30.434	30.548	-15.203
Simmetrica eterogenea	18	23.472	23.618	-11.718
Autoregressiva del prim'ordine	14	22.226	22.340	-11.099
Autoregressiva eterogenea	18	18.036	18.183	-9.000

Si può vedere che il modello migliore, secondo gli indici, è quello con una matrice di varianze e covarianze degli errori non strutturata, come già ci si aspettava. Tenendo anche conto del numero di parametri incogniti, cercando quindi di ridurlo, si ha che un buon modello è quello che utilizza la matrice di varianze e covarianze degli errori autoregressiva eterogenea del prim'ordine (18 gradi di libertà, contro i 27 della matrice non strutturata). In Tabella 73 sono mostrate le stime dei parametri derivanti dai modelli reputati migliori.

Tabella 73 – Stime dei parametri derivanti dai modelli che utilizzano diverse strutture per la matrice di varianze e covarianze degli errori, avendo come riferimento il modello F.

Parametro	Standard	Non strutturata	Autoregressiva eterogenea del prim'ordine	Autoregressiva del prim'ordine
Intercetta	0,9692	0,9668	0,9701	0,9632
Ist. Artistico	-0,2427	-0,2161	-0,1886	-0,2329
Ist. Professionale	-0,2460	-0,2440	-0,2265	-0,2714
Ist. Tecnico	-0,0826	-0,0853	-0,0765	-0,0957
Tempo	0,9381	0,9398	0,9446	0,9503
Tempo: Ist. Artistico	-0,0906	-0,0906	-0,0801	-0,0684
Tempo: Ist. Professionale	-0,1777	-0,1741	-0,1486	-0,1197
Tempo: Ist. Tecnico	-0,0624	-0,0582	-0,0472	-0,0345
Sesso: Ist. Professionale	-0,1389	-0,1266	-0,1216	-0,1506
Sesso: Ist. Tecnico	-0,0714	-0,0782	-0,0821	-0,0898
Tempo:Sesso: Ist. Professionale	-0,0770	-0,0857	-0,0764	-0,0647
Tempo:Sesso: Ist. Tecnico	-0,1009	-0,1004	-0,0881	-0,0794
AIC	17.203	15.421	18.036	22.226
BIC	17.333	15.640	18.183	22.340
Deviance statistic	17.171	15.367	18.000	22.198

Non tutti i modelli attribuiscono uguale significatività ai parametri. Il modello in cui è ipotizzata la matrice autoregressiva del prim'ordine attribuisce un p-value di 0,0022 al parametro di frequenza dell'Ist. Tecnico e un p-value di 0,009 al parametro relativo all'interazione tra la frequenza dell'Ist. Tecnico e il Sesso. Guardando i valori dei parametri di queste due ultime variabili, si può desumere che la differenza per quanto riguarda lo stato

iniziale (quindi la probabilità di promozione in classe prima) tra gli studenti che frequentano il liceo e quelli che frequentano invece l'istituto tecnico non è molto alta, mentre l'analoga differenza riferita al tasso di variazione (quindi le probabilità di promozione nelle classi successive e la prosecuzione della regolarità) è maggiore, come dimostra la significatività dell'interazione tra la variabile rappresentante la frequenza dell'Ist. Tecnico e il Tempo.

Si può anche vedere che le direzioni delle relazioni sono perfettamente riprodotte da tutti i modelli. Esaminando graficamente gli errori di primo livello, però, non si può dire che l'ipotesi di normalità distributiva sia verificata. I *normality plot* dei due modelli migliori sono rappresentati in Grafico 67 e Grafico 68, mentre la distribuzione degli errori standardizzati per diversi valori del tempo sono rappresentati in Grafico 69 e Grafico 70.

Dal confronto dei grafici di normalità distributiva relativi al modello F con matrice delle covarianze standard (Grafico 62) e a quello con matrice delle covarianze di tipo autoregressivo eterogeneo (Grafico 68), emerge la spiccata maggiore linearità nel secondo caso, motivo per cui è preferibile utilizzare una tale matrice delle covarianze rispetto al caso standard.

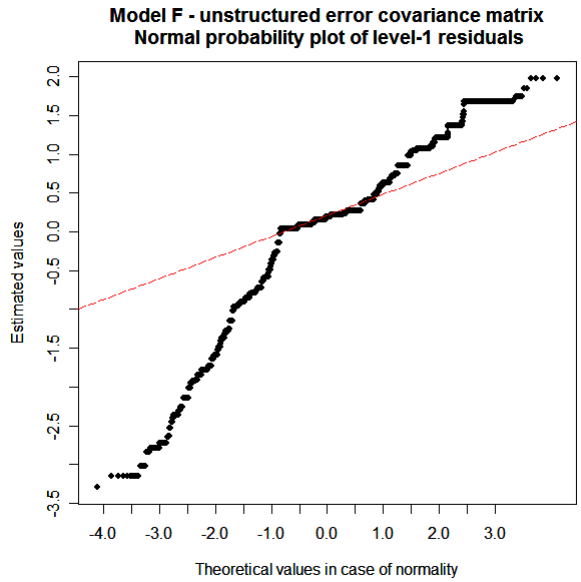


Grafico 67 - *Normal probability plot* degli errori di primo livello, modello F con ipotesi di matrice di varianze e covarianze degli errori non strutturata

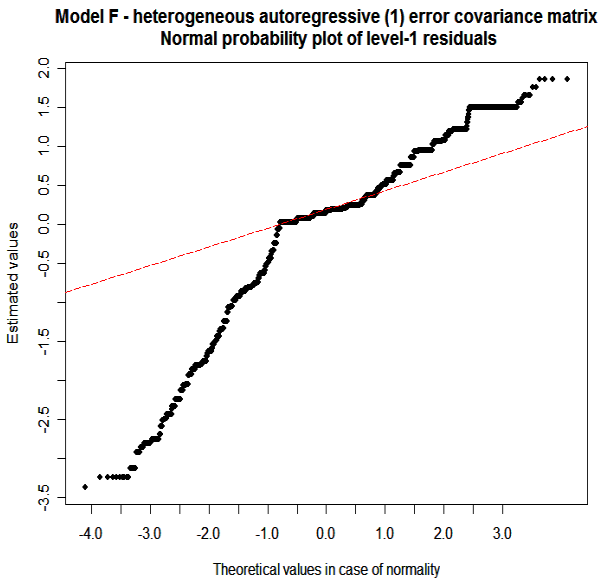


Grafico 68 - *Normal probability plot* degli errori di primo livello, modello F con ipotesi di matrice di varianze e covarianze degli errori autoregressiva eterogenea

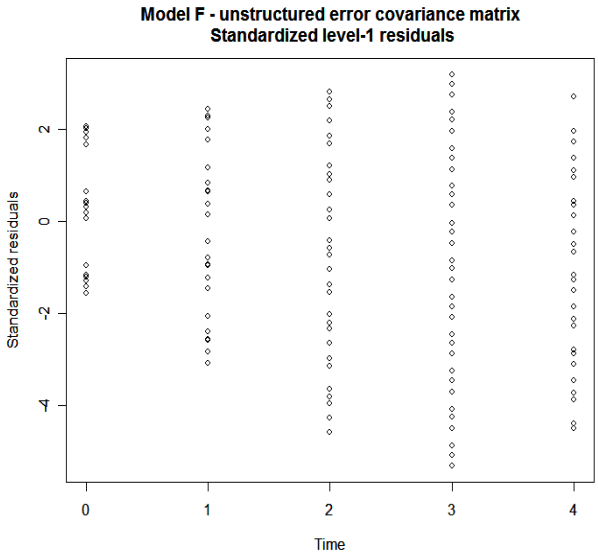


Grafico 69 – Distribuzione delle stime degli errori standardizzate nei diversi intervalli temporali, modello F con ipotesi di matrice di varianze e covarianze degli errori non strutturata

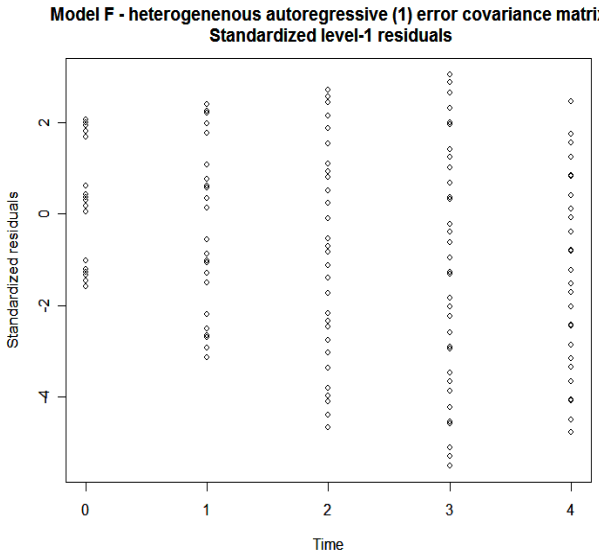


Grafico 70 – Distribuzione delle stime degli errori standardizzate nei diversi intervalli temporali, modello F con ipotesi di matrice di varianze e covarianze degli errori autoregressiva eterogenea

I grafici ed anche la statistica test di Kolmogorov Smirnov confermano che la distribuzione degli errori non è molto vicina alla normale; ma se si confrontano le percentuali di errori standardizzati al di fuori dell'intervallo [-2; 2], si ha che risultano del 9% nel modello F con la matrice di varianze e covarianze degli errori standard, invece del 5,4% nello stesso modello F, ma in cui è ipotizzata una matrice di varianze e covarianze degli errori autoregressiva eterogenea (questo è un altro sintomo di maggiore bontà di adattamento del modello, visto che con una tale percentuale si è maggiormente vicini ad una distribuzione regolare degli errori). Alla fine, il modello F in cui è ipotizzata una matrice di varianze e covarianze degli errori del tipo autoregressiva (del prim'ordine) eterogenea è preferibile a tutti gli altri modelli, in quanto si adatta in modo migliore ai dati, considerato anche il numero limitato di gradi di libertà.

Nel seguito è illustrato (Grafici 71, 72 e 73) un confronto grafico dei valori osservati e di quelli stimati dai diversi modelli della variabile *Times promoted* (è chiaro che i modelli non stimano perfettamente i valori interi, che invece sono gli unici possibili, e in generale si ha una sottostima).

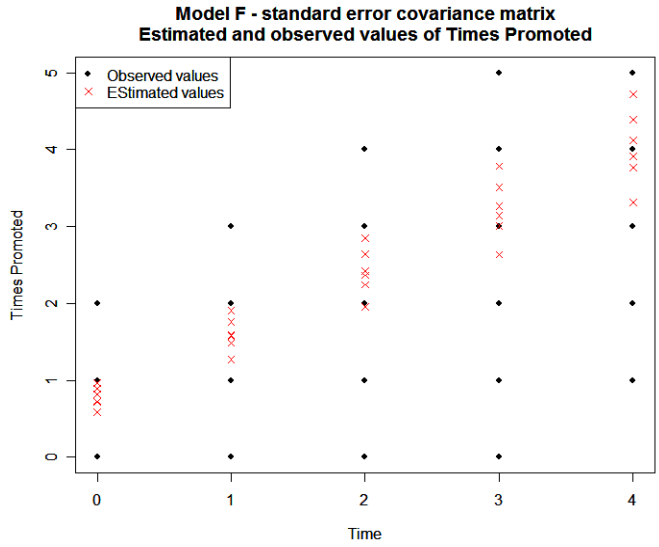


Grafico 71 – Distribuzione dei valori osservati e stimati della variabile obiettivo nel tempo, modello F con ipotesi di matrice di varianze e covarianze degli errori standard

Tempo 0 – l' 85% dei valori osservati è =1;l'80% dei valori stimati rimane all'interno dell'intervallo [0,73; 0,97].
Tempo 1 – 76% dei valori osservati è =2; l'80% dei valori stimati rimane all'interno dell'intervallo [1,57; 1,91].
Tempo 2 – il 72% dei valori osservati è =3; il 60% dei valori stimati rimane all'interno dell'intervallo [2,42; 2,85].
Tempo 3 – il 71% dei valori osservati è =4; il 59% dei valori stimati rimane all'interno dell'intervallo [3,51; 3,78].
Tempo 4 – il 74% dei valori osservati è =5; il 60% dei valori stimati rimane all'interno dell'intervallo [4,39; 4,72].

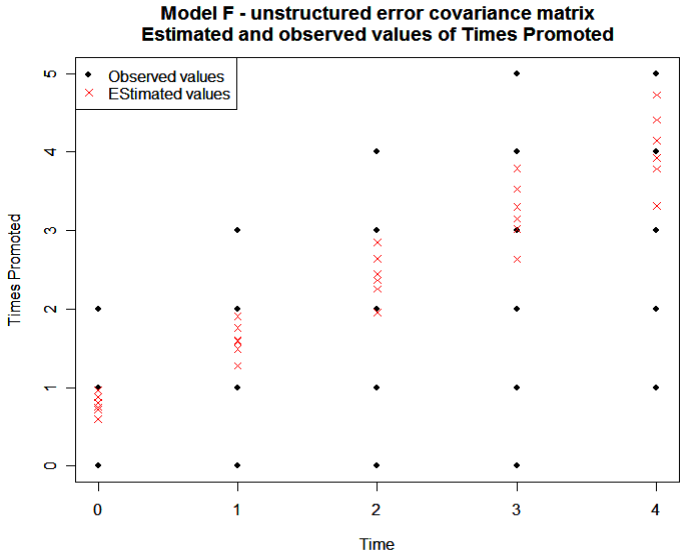


Grafico 72 – Distribuzione dei valori osservati e stimati della variabile obiettivo nel tempo, modello F con ipotesi di matrice di varianze e covarianze degli errori non strutturata

Tempo 0 – l'85% dei valori osservati è =1; l'81% dei valori stimati rimane all'interno dell'intervallo [0,75; 0,97].
Tempo 1 – il 76% dei valori osservati è =2; l'80% dei valori stimati rimane all'interno dell'intervallo [1,58; 1,91].
Tempo 2 – il 72% dei valori osservati è =3; il 60% dei valori stimati rimane all'interno dell'intervallo [2,45; 2,85].
Tempo 3 – il 71% dei valori osservati è =4; il 62% dei valori stimati rimane all'interno dell'intervallo [3,30; 3,79].
Tempo 4 – il 74% dei valori osservati è =5; il 63% dei valori stimati rimane all'interno dell'intervallo [4,15; 4,73].

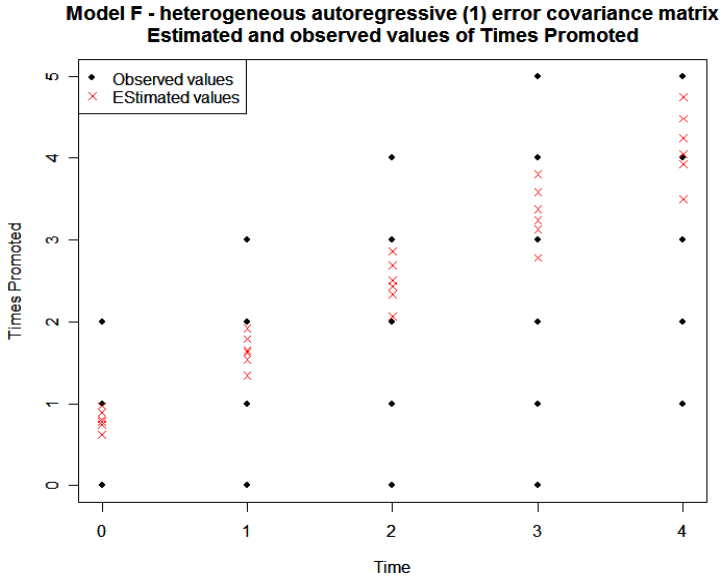


Grafico 73 – Distribuzione dei valori osservati e stimati della variabile obiettivo nel tempo, modello F con ipotesi di matrice di varianze e covarianze degli errori autoregressiva eterogenea

Tempo 0 – l'85% dei valori osservati è =1; l'81% dei valori stimati rimane all'interno dell'intervallo [0,78; 0,97].
Tempo 1 – il 76% dei valori osservati è =2; l'80% dei valori stimati rimane all'interno dell'intervallo [1,62; 1,92].
Tempo 2 – il 72% dei valori osservati è =3; il 61% dei valori stimati rimane all'interno dell'intervallo [2,51; 2,86].
Tempo 3 – il 71% dei valori osservati è =4; il 62% dei valori stimati rimane all'interno dell'intervallo [3,37; 3,80].
Tempo 4 – il 74% dei valori osservati è =5; il 64% dei valori stimati rimane all'interno dell'intervallo [4,24; 4,75].

Il modello F con ipotesi di matrice di varianze e covarianze degli errori autoregressiva eterogenea è quello che si adatta al meglio ai dati osservati.

In Tabella 74 sono riportate le medie stimate della variabile obiettivo, in base al modello ritenuto ottimale.

Tabella 74 – Medie stimate, con relativi intervalli di confidenza, dal modello F, con matrice delle covarianze dei termini di errore autoregressiva eterogenea, (medie osservate) della variabile obiettivo *Times Promoted*

	2002/2003	2003/2004	2004/2005	2005/2006	2006/2007
Studenti italiani che frequentano il liceo	0,970 [0,962-0,978] (M 0,958 F 0,981)	1,915 [1,903-1,926] (M 1,915 F 1,932)	2,859 [2,845-2,874] (M 2,848 F 2,886)	3,804 [3,786-3,822] (M 3,781 F 3,881)	4,748 [4,727-4,769] (M 4,762 F 4,852)
Studenti italiani che frequentano l'istituto artistico	0,781 [0,743-0,819] (M 0,814 F 0,798)	1,646 [1,591-1,700] (M 1,500 F 1,624)	2,510 [2,440-2,581] (M 2,326 F 2,571)	3,375 [3,288-3,462] (M 3,436 F 3,528)	4,239 [4,136-4,343] (M 4,219 F 4,504)
Ragazze italiane che frequentano l'istituto tecnico	0,894 [0,867-0,920] (0,899)	1,791 [1,753-1,829] (1,788)	2,688 [2,639-2,737] (2,727)	3,586 [3,525-3,646] (3,704)	4,483 [4,411-4,555] (4,696)
Ragazzi italiani che frequentano l'istituto tecnico	0,811 [0,765-0,858] (0,837)	1,621 [1,554-1,688] (1,640)	2,430 [2,343-2,518] (2,502)	3,239 [3,131-3,347] (3,416)	4,049 [3,920-4,177] (4,451)
Ragazze italiane che frequentano l'istituto professionale	0,744 [0,716-0,771] (0,756)	1,540 [1,500-1,579] (1,547)	2,336 [2,284-2,387] (2,397)	3,132 [3,068-3,195] (3,347)	3,928 [3,852-4,003] (4,324)
Ragazzi italiani che frequentano l'istituto professionale	0,622 [0,571-0,673] (0,660)	1,342 [1,268-1,415] (1,319)	2,061 [1,964-2,158] (2,139)	2,781 [2,661-2,901] (3,064)	3,501 [3,358-3,643] (3,988)

Si nota che, rispetto al modello F standard, le stime hanno in generale valori più alti e si avvicinano maggiormente alle osservazioni, come dimostrano anche i test sulla bontà di adattamento. Dal confronto degli intervalli di confidenza ottenuti dai due modelli analoghi, emerge inoltre che la loro ampiezza è in generale inferiore se si fa riferimento al modello che utilizza l'ipotesi autoregressiva eterogenea per la matrice delle covarianze dei termini di errore. Sempre con riferimento agli intervalli di confidenza, si può in generale notare che sono di minore ampiezza se si considerano le stime riguardanti i ragazzi dei licei, mentre tale ampiezza aumenta se si considerano i ragazzi degli istituti tecnici e professionali, con conseguente diminuzione di precisione per le stime.

4.2.4 RICERCA DEL MODELLO OTTIMALE PER I SOLI STUDENTI ITALIANI PRESENTI A SCUOLA PER TUTTO L'INTERVALLO TEMPORALE CONSIDERATO

Utilizzando il modello che comprende soltanto i 4.385 studenti che erano nelle scuole bolognesi in tutti gli anni scolastici considerati (anni scolastici 2002/03, 2003/04, 2004/05, 2005/06 e 2006/07), si può osservare il trend della variabile *Times promoted* soltanto per questo gruppo ristretto e cercare un modello che lo riproduca. Per tener conto della variabile indicante la tipologia di scuola frequentata, variabile risultata significativa nella spiegazione del trend, si è scelto di utilizzare, per la stima del modello, soltanto i dati sugli studenti con cittadinanza italiana.

Si è così stimato il **modello F2** (che tiene conto soltanto dei 4.333 studenti italiani):

$$y_{it} = \alpha_i + \beta_i Time_{it} + \varepsilon_{it} \quad \text{con ipotesi } \varepsilon_{it} \sim N\left(0, \sigma_\varepsilon^2\right)$$

$$\alpha_i = \gamma_{00} + \gamma_{02} Sex + \gamma_{05} Voc_i + \zeta_{0i}$$
$$\beta_i = \gamma_{10} + \gamma_{12} Sex + \gamma_{13} Voc_i + \gamma_{14} Art_i + \gamma_{15} Tech_i + \zeta_{1i}$$

con la consueta ipotesi $\begin{pmatrix} \zeta_{0i} \\ \zeta_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{pmatrix}\right)$.

Inoltre si è proceduto a stimare il corrispondente modello con matrice di varianze e covarianze degli errori autoregressiva eterogenea, ottenendo risultati migliori. Si può notare che, in questo caso, il valore iniziale di *Times promoted* è stimato come funzione del Sesso e della tipologia di scuola frequentata, ma soltanto se si tratta di un professionale: la probabilità di essere promossi in classe prima è stimata differire tra maschi e femmine e inoltre tra studenti che frequentano istituti professionali e gli altri, mentre la medesima probabilità è identica per studenti del liceo e studenti degli istituti artistici e tecnici. Il tasso di variazione di *Times promoted* è invece stimato differire anche tra gli studenti che frequentano il liceo e quelli dell'istituto artistico e dell'istituto tecnico. Secondo gli indici AIC (pari a 4.275 nel modello standard e a 3.852 nel modello autoregressivo eterogeneo) e BIC (pari a 4.379 nel modello standard e a 3.964 nel modello autoregressivo eterogeneo) ed anche guardando il valore della *deviance statistic* (pari a -2124 nel modello standard e a -1912 nel modello autoregressivo eterogeneo), il modello migliore è quello che ipotizza una matrice autoregressiva eterogenea.

La stima (ipotesi autoregressiva eterogenea) del valore iniziale di *Times promoted* per le ragazze del liceo è 0,948 (la corrispondente probabilità osservata di essere promossi in classe prima è 0,935), mentre per i ragazzi del liceo è 0,917 (la corrispondente probabilità osservata di essere promossi in classe prima è 0,920). La stima (sempre nell'ipotesi autoregressiva eterogenea) del tasso di variazione della variabile obiettivo per le ragazze che frequentano il liceo è 0,964; l'analoga stima per i ragazzi che frequentano il liceo è 0,935. La stima dei parametri del modello risulta:

$$y_{it} = \alpha_i + \beta_i Time_{it} + \varepsilon_{it}$$

$$\alpha_i = 0{,}948 - 0{,}031 Sex + 0{,}015 Voc_i + \zeta_{0i}$$
$$\beta_i = 0{,}964 - 0{,}031 Sex - 0{,}045 Voc_i - 0{,}024 Art_i - 0{,}035 Tech_i + \zeta_{1i}$$

La differenza stimata in quanto a valore iniziale tra ragazzi e ragazze è –0,031; la differenza stimata in quanto a tasso di variazione tra ragazzi e ragazze è mediamente nel tempo –0,031; invece, la differenza in quanto a valore della variabile al tempo 0 tra gli studenti che frequentano l'istituto professionale rispetto agli altri è 0,015 (gli studenti dei professionali hanno mediamente un valore più alto degli altri il primo anno, nonostante nel seguito i risultati peggiorino più velocemente rispetto agli altri studenti). La progressione media nel tempo della variabile obiettivo è peggiore, rispetto agli studenti del liceo, per gli studenti degli istituti artistici (-0,024), degli istituti tecnici (-0,035) e degli istituti professionali (-0,045).

La differenza media, risultante dalla media geometrica dei rapporti tra stime e osservazioni, relativa alle ragazze del liceo è pari all'1% (il modello sottostima i valori osservati); l'analoga differenza media per le ragazze dell'artistico risulta pari al 3% (il modello sovrastima le osservazioni). La differenza media per i ragazzi del liceo è del 2% (il modello sottostima le osservazioni); per quanto riguarda invece i ragazzi degli artistici, risulta dello 0,8% (occorre però considerare che il modello ora sovrastima e ora sottostima i dati osservati). Per quanto riguarda gli istituti tecnici, la differenza media relativa alle ragazze è dello 0,5%, mentre per i ragazzi si attesta intorno allo 0,2%. Negli istituti professionali, la differenza tra stime ed osservazioni risulta pari al 4% per le ragazze, mentre si porta al 5% per i ragazzi (il modello sovrastima i dati).

Per quanto riguarda le ipotesi del modello, la distribuzione degli errori non è molto vicina a quella normale, come mostrato in Grafico 74.

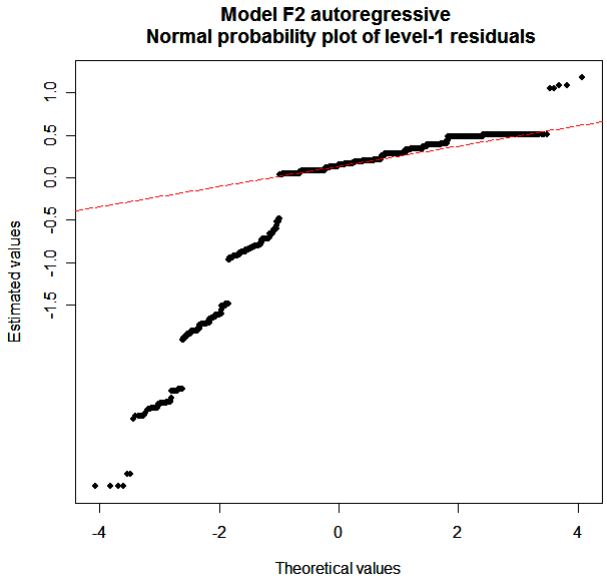


Grafico 74 – *Normal probability plot* degli errori di primo livello nel modello F2, con matrice di varianze e covarianze degli errori autoregressiva

Per quanto riguarda ancora la distribuzione degli errori, si ha che la normalità è forzata per il modello con matrice di varianze e covarianze degli errori standard, e ciò è dimostrato anche dal controllo sui singoli valori degli errori stessi: circa il 9% degli errori standardizzati di secondo livello nel caso dell'intercetta e circa il 6% degli stessi errori nel caso della pendenza rimangono al di fuori dell'intervallo [-2; 2], che sono percentuali sempre superiori al 5% richiesto per il caso di possibile normalità distributiva. Invece la situazione cambia nel caso di matrice autoregressiva eterogenea: in questo caso non è possibile separare gli errori di primo livello per intercetta e pendenza, tuttavia considerandoli tutti insieme, la percentuale di errori standardizzati che rimane al di

fuori dell'intervallo [-2; 2] è del 3,4%, quindi inferiore al 5% richiesto per la normalità. Si può dire che si è più vicini alla normalità distributiva che non in tutti i modelli precedenti.

Passando al confronto dei valori stimati per la variabile *Times promoted* con i valori osservati, si ha che il 97% dei valori stimati sono compresi nell'intervallo]valore vero –1; valore vero +1[. In particolare, il 93% delle differenze tra valori osservati e stimati relativi al Tempo 0 rimangono nell'intervallo [-0,3; 0,3]; l'88% delle differenze tra valori osservati e stimati relativi al Tempo 1 rimangono nell'intervallo [-0,3; 0,3]; mentre le analoghe percentuali risultano dell'82% per il Tempo 2, del 61% per il Tempo 3 e del 28% per il Tempo 4.

Osservando il grafico che riporta i valori stimati in ordinata e quelli osservati in ascissa (Grafico 75), si può notare che la maggior parte dei valori si trova sulla diagonale principale. Inoltre, il modello tende a sovrastimare la variabile obiettivo per valori osservati più bassi, mentre tende a sottostimarla per valori osservati più alti: il modello appiattisce proprio in virtù dell'ipotesi lineare.

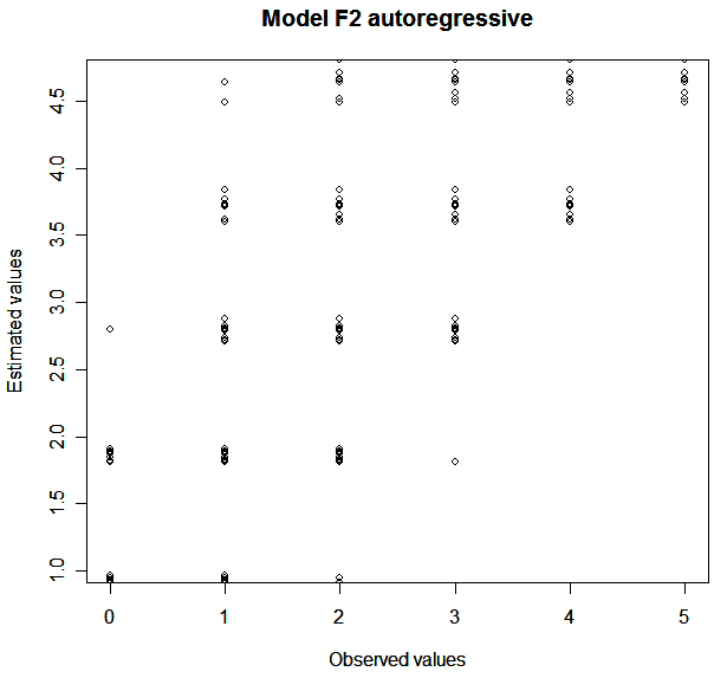


Grafico 75 – Distribuzione nel tempo dei valori della variabile obiettivo osservati e stimati dal modello F2, con ipotesi di matrice di varianze e covarianze degli errori autoregressiva

I valori veri della variabile *Times promoted* =0 (che rappresenta l'1,4% dei suoi valori) sono per la maggior parte sovrastimati: l'86% di essi è stimato essere inferiore a 0,95.
L'87% dei valori veri di *Times promoted* =1 è stimato essere compreso tra 0,9 e 1.
L'80% dei valori veri di *Times promoted* =2 è stimato essere compreso tra 0,8 e 2.
Il 78% dei valori veri di *Times promoted* =3 è stimato essere compreso tra 2,7 e 3.
Il 65% dei valori veri di *Times promoted* =4 è stimato essere compreso tra 3,7 e 4.
Il 78% dei valori veri di *Times promoted* =5 è stimato essere compreso tra 4,6 e 5.

Il Grafico 76 mostra che in realtà le stime fornite dal modello per i diversi gruppi di individui sono tra loro non distanti. Occorre meglio indagare sull'ampiezza degli intervalli di confidenza stimati, al fine di saggiare se effettivamente non vi siano estremi in comune almeno tra alcuni gruppi di individui. Dal confronto tra il Grafico 76 ed il Grafico 77 si evince che il modello comunque ben riproduce le medie di gruppo.

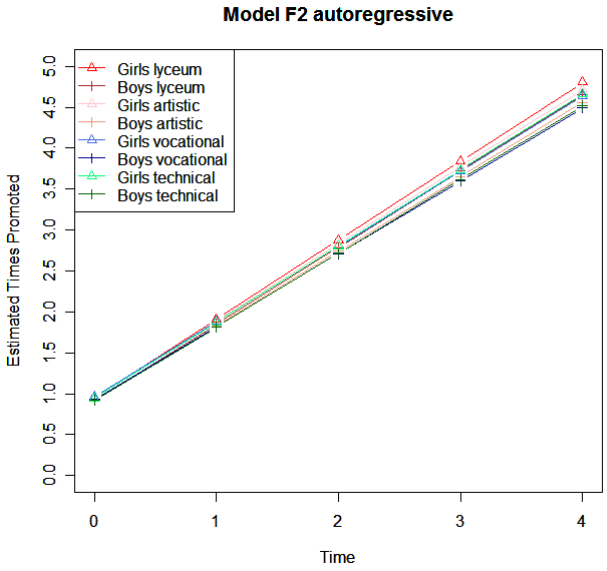


Grafico 76 – Distribuzione delle medie nel tempo di individui tipo stimate dal modello F2

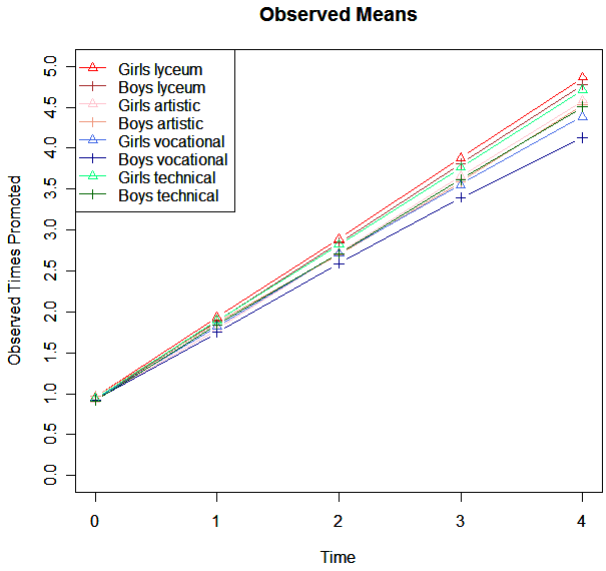


Grafico 77 - Distribuzione delle medie nel tempo di individui tipo osservate

I grafici 78 e 79 mostrano gli intervalli di confidenza stimati (linee tratteggiate) e le medie osservate (linee continue) per alcuni individui tipo. Si nota che gli intervalli di confidenza sono davvero molto vicini tra loro. La separazione degli intervalli di confidenza si ha tra maschi e femmine, in ciascun tipo di scuola ed in tutti gli intervalli temporali, come si evince dalla Tabella 68, dove sono specificati gli estremi degli intervalli di confidenza stimati. Se anche il coefficiente che rappresenta la differenza tra il valore iniziale della variabile obiettivo per gli studenti dei professionali e gli altri studenti risulta significativo, dall'esame degli intervalli di confidenza si nota che tali intervalli, per individui appartenenti a gruppi diversi, hanno in realtà punti in comune: la differenza non si può dire significativa. Per quanto riguarda le stime relative agli istituti artistici, si verifica che gli intervalli di confidenza abbiano punti in comune con quelli relativi agli studenti degli altri tipi di scuola: si può dire che vi sia una differenza negativa significativa tra gli studenti dell'artistico e quelli, di medesimo sesso, del liceo (negli anni successivi al primo); non si può invece dire che vi sia una differenza significativa tra gli studenti dell'artistico e quelli, di medesimo sesso, dell'istituto tecnico e professionale; si può inoltre affermare che esista una differenza positiva significativa tra ragazze dell'artistico e ragazzi del tecnico o del professionale, come anche che sia significativa la differenza negativa tra ragazzi dell'artistico e ragazze del tecnico o del professionale. Le considerazioni che riguardano gli studenti degli artistici devono però sempre tener conto della scarsa numerosità. Dall'esame della Tabella 75, emerge che le ragazze del liceo ottengono la migliore performance rispetto a tutti gli altri, in tutti gli anni considerati: gli intervalli di confidenza sono nettamente separati da quelli stimati per gli altri gruppi di studenti. Sono invece i ragazzi dei tecnici e dei professionali a conseguire la peggiore successione di valori di *Times Promoted*, anche se il divario negativo rispetto ai ragazzi dei licei, che pure è inesistente nel primo anno, cresce andando avanti. Si può dire che non esista differenza significativa nell'andamento nel tempo della variabile obiettivo tra le ragazze che frequentano i tecnici e quelle dei professionali; tale andamento non è nemmeno significativamente differente da quello dei ragazzi del liceo. Le ragazze che frequentano istituti tecnici o professionali presentano una differenza positiva significativa, in tutti gli anni considerati, rispetto ai ragazzi che frequentano gli stessi istituti.

Tabella 75 – Medie stimate, con relativi intervalli di confidenza, dal modello F2 (medie osservate) della variabile obiettivo *Times Promoted*					
	2002/2003	2003/2004	2004/2005	2005/2006	2006/2007
Ragazze che frequentano il liceo	0,948 [0,942-0,954] (0,958)	1,912 [1,903-1,921] (1,935)	2,876 [2,864-2,888] (2,890)	3,840 [3,825-3,855] (3,881)	4,804 [4,786-4,822] (4,864)
Ragazzi che frequentano il liceo	0,916 [0,902-0,931] (0,926)	1,851 [1,830-1,873] (1,889)	2,786 [2,758-2,814] (2,845)	3,721 [3,686-3,756] (3,808)	4,656 [4,614-4,698] (4,777)
Ragazze che frequentano l'istituto artistico	0,948 [0,942-0,954] (0,925)	1,888 [1,869-1,907] (1,788)	2,829 [2,796-2,861] (2,717)	3,769 [3,724-3,814] (3,657)	4,709 [4,651-4,768] (4,582)
Ragazzi che frequentano l'istituto artistico	0.916 [0,902-0,931] (0,968)	1,828 [1,796-1,859] (1,875)	2,739 [2,690-2,787] (2,697)	3,650 [3,585-3,715] (3,588)	4,561 [4,479-4,643] (4,529)
Ragazze che frequentano l'istituto tecnico	0,948 [0,942-0,954] (0,938)	1,877 [1,864-1,890] (1,894)	2,806 [2,786-2,826] (2,821)	3,735 [3,708-3,762] (3,763)	4,664 [4,630-4,698] (4,715)
Ragazzi che frequentano l'istituto tecnico	0,917 [0,902-0,931] (0,915)	1,816 [1,791-1,842] (1,845)	2,716 [2,680-2,752] (2,713)	3,616 [3,569-3,663] (3,616)	4,516 [4,458-4,573] (4,504)
Ragazze che frequentano l'istituto professionale	0,962 [0,946-0,979] (0,939)	1,881 [1,857-1,906] (1,821)	2,800 [2,767-2,833] (2,711)	3,719 [3,678-3,759] (3,554)	4,637 [4,588-4,686] (4,384)
Ragazzi che frequentano l'istituto professionale	0,931 [0,906-0,956] (0,925)	1,821 [1,784-1,858] (1,749)	2,710 [2,661-2,759] (2,591)	3,600 [3,539-3,660] (3,399)	4,489 [4,416-4,562] (4,127)

Confrontando i risultati ottenuti con il modello F2, che considera i soli studenti rimasti a scuola per tutti i cinque anni considerati, e quelli ottenuti con il modello F, che include anche coloro che hanno lasciato la scuola bolognese o che sono entrati dopo il primo anno considerato, si nota che nel primo caso le differenze, stimate significative dal secondo modello, si assottigliano maggiormente, fino a scomparire. Questo vuol forse dire che la vera e significativa differenza in quanto ad andamento degli esiti scolastici nel tempo è dovuta soprattutto alla diversità tra gli studenti che rimangono nel sistema scolastico e coloro che invece lo abbandonano precocemente o che comunque si spostano una o più volte da un istituto all'altro.

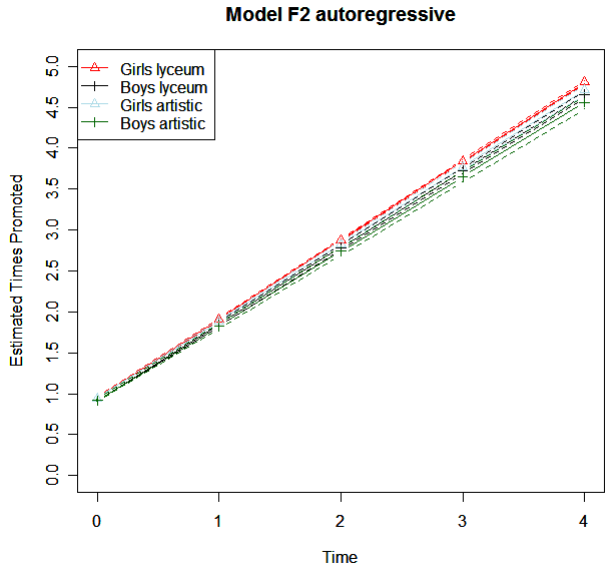


Grafico 78 – Valori osservati e intervalli di confidenza stimati – licei e artistici

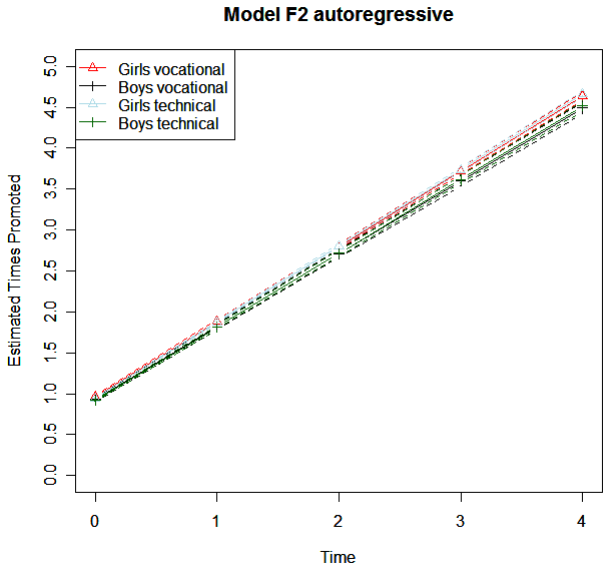


Grafico 79 – Valori osservati e intervalli di confidenza stimati – tecnici e professionali

Al fine di cercare di migliorare ulteriormente le stime, si è intentata la strada delle stime basate sul metodo empirico di Bayes[72]. Le traiettorie medie in popolazione, in questo caso, sono state ottenute a partire dai valori

[72] Singer, Willett "Applied longitudinal data analysis – modelling change and event occurrence"

individuali dei predittori, ma aggiungendo poi una informazione specifica, con l'utilizzo degli errori di secondo livello.

A partire dal modello F2 con matrice di varianze e covarianze degli errori di tipo standard, le stime dei parametri di primo livello sono le seguenti:

$$\hat{\alpha}_i = 0{,}945 - 0{,}021 Sex + 0{,}013 Voc_i$$

$$\hat{\beta}_i = 0{,}958 - 0{,}033 Sex - 0{,}019 Voc_i - 0{,}032 Tech_i - 0{,}007\, Art_i$$

Tali stime portano a traiettorie identiche per tutti gli individui che hanno la stessa combinazione dei valori dei predittori. È possibile correggere le traiettorie individuali tenendo conto degli errori di secondo livello: considerando l'individuo *i*, i parametri del modello F2 sono quelli prima specificati, ma possono essere corretti con gli errori di secondo livello individuali. Calcolando le differenze tra i valori di *Times promoted* osservati e quelli stimati con il metodo di Bayes, si ha che il 95% delle differenze relative al tempo 0, l'89% delle differenze relative al tempo 1 e al tempo 2, il 92% delle differenze relative al tempo 3 ed il 93% delle differenze relative al tempo 4 sono comprese nell'intervallo]-0,3; 0,3[.

Il Grafico 80, costruito in modo analogo al precedente, in questo caso diventa quello nel seguito riportato.

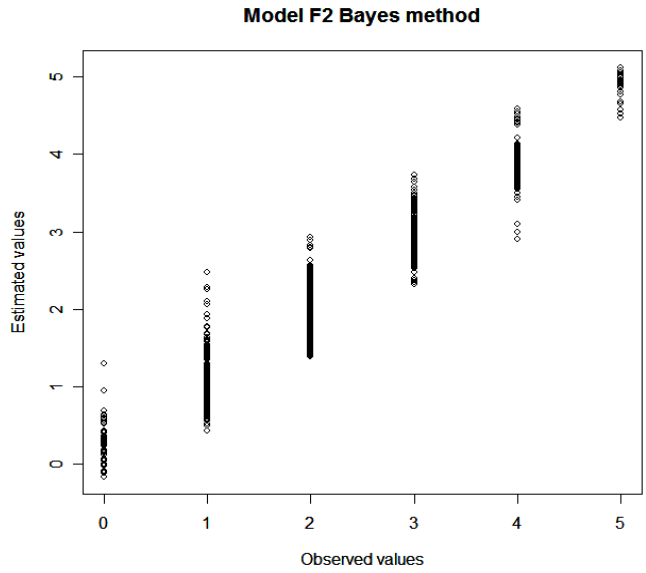


Grafico 80 –Valori della variabile obiettivo osservati nel tempo e stimati dal modello F2, con ipotesi di matrice di varianze e covarianze degli errori standard, ma con la correzione apportata con il metodo di Bayes

Il 92% dei valori veri di *Times promoted* =0 è stimato essere circa inferiore a 0,45.
Il 99% dei valori veri di *Times promoted* =1 è stimato rimanere nell'intervallo]0,5; 1,5[.
Il 98% dei valori veri di *Times promoted* =2 è stimato rimanere nell'intervallo]1,5; 2,5[.
Il 98% dei valori veri di *Times promoted* =3 è stimato rimanere nell'intervallo]2,5; 3,5[.
Il 99,6% dei valori veri di *Times promoted* =4 è stimato rimanere nell'intervallo]3,5; 4,5[.
Il 99,9% dei valori veri di *Times promoted* =5 è stimato rimanere nell'intervallo]4,5; 5,15[.

La stima è molto migliore rispetto al modello F2 con matrice di varianze e covarianze degli errori di tipo standard ed è anche migliore rispetto al modello simile con matrice di tipo autoregressivo eterogeneo. Ma occorre considerare, quando si utilizzano questi diversi tipi di stime, che il modello basato sul metodo della massima verosimiglianza porta a stimatori corretti, ma questi possono avere una perdita di efficienza; invece le stime di tipo empirico sono distorte, tuttavia maggiormente precise, quindi la loro variabilità è inferiore. Quest'ultimo metodo di stima è maggiormente appropriato per spiegare il fenomeno, come può essere nel caso in esame, che non per previsioni. Tuttavia, nel caso in esame, si ritiene maggiormente appropriato l'utilizzo del modello F2 con matrice di varianze e covarianze autoregressiva eterogenea, che quindi viene scelto come modello ottimale per il dataset specificato all'inizio del capitolo.

Tabella 76 – Risultati dei modelli a curva latente – Variabile obiettivo: *Times Promoted*

		Parametro	Modello A Medie non condizionate	Modello B Non condizionato	Modello C Condizionato cittadinanza	Modello D Condizionato cittadinanza e sesso	Modello F(1) Condizionato sesso e tipo scuola	Modello G(1) Condizionato sesso, tipo scuola, ambito
Stato iniziale	Intercetta s.e.	γ_{00}	2,596*** (0,009)	0,819*** (0,006)	0,854*** (0,006)	0,894*** (0,008)	0,969*** (0,008)	0,966*** (0,008)
	Cittadinanza straniera s.e.	γ_{01}			-0,665*** (0,027)	-0,667*** (0,027)	Non presenti	Non presenti
	Sesso (Maschi) s.e.	γ_{02}				-0,078*** (0,011)	-0,139(P) –0,071(T) *** (0,024) (0,020)	-0,139(P) –0,071(T) *** (0,024) (0,020)
	Ist. Professionale (P) s.e.	γ_{05}					-0,246*** (0,019)	-0,242*** (0,019)
	Ist. Tecnico (T) s.e.	γ_{06}					-0,083*** (0,018)	-0,080*** (0,018)
	Ist. Artistico s.e.	γ_{04}					-0,243*** (0,030)	-0,203*** (0,029)
Tasso di variazione	Pendenza s.e.	γ_{10}		0,833*** (0,004)	0,850*** (0,004)	0,894*** (0,005)	0,938*** (0,005)	0,932*** (0,005)
	Cittadinanza straniera s.e.	γ_{11}			-0,221*** (0,015)	-0,224*** (0,015)	Non presenti	Non presenti
	Sesso (Maschi) s.e.	γ_{12}				-0,086*** (0,007)	-0,077(P) –0,101(T) *** (0,015) (0,012)	-0,081(P) –0,101(T) *** (0,015) (0,012)
	Ist. Professionale (P) s.e.	γ_{13}					-0,178*** (0,012)	-0,198*** (0,013)
	Ist. Tecnico (T) s.e.	γ_{15}					-0,062*** (0,011)	-0,056*** (0,011)
	Ist. Artistico s.e.	γ_{11}					-0,091*** (0,018)	Non significativo
I livello	Varianza entro gli individui	σ^2_ε	2,122	0,0377	0,0377	0,0377	0,0361(2)	0,0361(2)
II livello	Varianza – stato iniziale	σ^2_0	0,000000011	0,174	0,156	0,155	0,133	0,133
	Varianza – tasso di variazione	σ^2_1		0,067	0,062	0,060	0,051	0,051
	Covarianza	σ_{10}		0,055	0,045	0,044	0,035	0,035
	R-quadro	R^2		0,769	0,794	0,796	0,820	0,820
	Pseudo R-quadro	$pseudoR^2_\varepsilon$		0,982	0,982	0,982	0,983	0,983
	Pseudo R-quadro	$pseudoR^2_0$			0,1023	0,1108	0,1013	0,1004
	Pseudo R-quadro	$pseudoR^2_1$			0,0681	0,0983	0,1203	0,1171
AIC			94.406	21.229	20.504	20.339	17.203	17.191
BIC			94.430	21.278	20.569	20.421	17.333	17.322
Deviance Statistic			94.400	21.217	20.488	20.319	17.171	17.159

*** p-value<0,001; ** p-value<0,01; * p-value<0,1 - (1) I valori sono riferiti agli studenti italiani che frequentano il liceo (2) il modello di riferimento è un modello B1 con soli studenti italiani

5 Conclusioni

La maggiore propensione delle ragazze ad ottenere buoni risultati scolastici rispetto ai colleghi maschi è un dato ormai assodato e risultante da tutte le analisi di dati sulla scolarità. In questa sede si è cercato di quantificare tale differenza, in termini di esito scolastico nei 5 anni delle scuole superiori. Il modello logistico stima direttamente la differenza, in termini di probabilità di essere promossi, tra maschi e femmine: nelle diverse classi ed anni scolastici considerati, la differenza media[67] percentuale di probabilità di promozione, in meno dei ragazzi rispetto alle ragazze, è stimata essere attorno al 6%. Il modello logistico ha anche il limite dovuto al contingente di riferimento: questo comprende i soli regolari, con la conseguenza che la probabilità di promozione nelle ultime classi risulta altamente sovrastimata; in particolare in quinta risulta non significativa la differenza tra maschi e femmine, proprio in virtù del fatto che la stima della probabilità di promozione risulta molto bassa.

Il modello a curva latente fornisce un risultato che necessita di una interpretazione preliminare. Considerando il modello F (con matrice delle covarianze autoregressiva), reputato quello ottimo e che si basa sui dati dei soli studenti italiani, emerge che la stima della differenza negativa tra maschi e femmine, in termini di differenza nella variabile obiettivo *Times Promoted*[68], risulta pari al 7% (la differenza percentuale calcolata direttamente sui dati risulta del 6,6%); in particolare, la detta differenza è significativa soltanto negli istituti tecnici e professionali, mentre al liceo mediamente i ragazzi e le ragazze conseguono risultati stimati dal modello non significativamente diversi. La citata differenza è stimata essere maggiormente accentuata negli istituti professionali rispetto ai tecnici: mentre nei tecnici si attesta mediamente negli anni intorno al 9,5% (l'analoga percentuale osservata è del 7,2%), ai professionali è stimato un valore medio negli anni di circa il 12% (l'analoga percentuale osservata è dell'11%). Occorre precisare che tali percentuali si riferiscono alla differenza in termini di valori della variabile obiettivo *Times Promoted*, mentre le percentuali calcolate per il modello logistico si riferiscono alla differenza in quanto a probabilità di promozione in ciascun anno e ciascuna classe. A questo proposito, occorre anche dire che l'informazione sulla classe non è potuta entrare in modo diretto nella stima, utilizzando il modello a curva latente invece del modello logistico, tuttavia si è potuto in più risalire al percorso scolastico. La valutazione dei valori medi della variabile obiettivo, stimati dal modello a curva latente, deve anche tener conto del fatto che al tempo 0 mancavano ancora quegli studenti che ancora frequentavano l'ultimo anno delle scuole secondarie di primo grado perché già in ritardo scolastico, inoltre delle bocciature successive (nel modello logistico, invece, entrano soltanto gli studenti in regola con gli studi).

Considerato che il valor medio della variabile obiettivo negli anni, per i soli studenti con percorso scolastico regolare è pari a 3, la stima (modello F) di tale valor medio risulta di 2,72 per le ragazze (2,79 osservato) e di 2,53 per i ragazzi (2,61 osservato). La progressione media stimata per le ragazze è [0,90; 1,81; 2,71; 3,63; 4,54] (la progressione osservata è [0,91; 1,82; 2,75; 3,74; 4,72]), mentre quella stimata per i ragazzi risulta [0,83; 1,67; 2,52; 3,38; 4,24] (la progressione osservata è [0,85; 1,67; 2,56; 3,50; 4,49]).

[67] Differenza di medie geometriche.

[68] Una volta calcolate le stime della variabile obiettivo per gruppi di studenti, in base alle stime dei parametri, si sono calcolate le differenze percentuali tra tali stime riferite a gruppi diversi e a ciascun istante temporale, per poi fare la media geometrica, nei vari tempi, di tali differenze.

Tabella 77 – Medie ponderate della variabile obiettivo *Times Promoted* nel tempo – stime da modello F e osservazioni	
Ragazze che frequentano il liceo	2,86 (2,91)
Ragazzi che frequentano il liceo	2,86 (2,85)
Ragazze che frequentano l'istituto artistico	2,51 (2,61)
Ragazzi che frequentano l'istituto artistico	2,51 (2,46)
Ragazze che frequentano l'istituto tecnico	2,69 (2,76)
Ragazzi che frequentano l'istituto tecnico	2,43 (2,57)
Ragazze che frequentano l'istituto professionale	2,34 (2,47)
Ragazzi che frequentano l'istituto professionale	2,06 (2,23)

Le medie osservate nel tempo della variabile obiettivo, riportate in Tabella 77, sono state calcolate tenendo conto della numerosità, per questo sono medie ponderate.

Le differenze percentuali medie tra maschi e femmine sono di poco sovrastimate dal modello per i professionali (il valor medio dei ragazzi è stimato essere del 12% inferiore a quello delle ragazze contro una differenza percentuale osservata dell'11%) e per i tecnici (il valor medio dei ragazzi è stimato essere del 9,5% inferiore a quello delle ragazze contro una differenza percentuale osservata del 7,2%). Per quanto riguarda il 2% di differenza percentuale media osservata tra maschi e femmine del liceo, il modello la stima non significativa. Per ciò che concerne invece gli istituti artistici, il modello stima non significativa la differenza osservata del 6%; la valutazione dell'adattamento deve però considerare la bassa numerosità degli studenti degli artistici, che ha come conseguenza principale l'inferiore attendibilità delle stime.

Per quanto riguarda le differenze tra studenti che frequentano le diverse tipologie di istituto, non si è riusciti ad individuare le differenze tra i frequentanti tutte le tipologie di scuola, data la scarsa numerosità degli studenti presenti nel dataset analizzato (se divisi secondo tutte le micro tipologie di cui è formato il nostro sistema scolastico), tuttavia si è riusciti a trarre informazioni riguardo agli esiti scolastici separatamente degli studenti che frequentano le diverse macro tipologie di istituto (liceo, artistico, tecnico e professionale). Dal modello logistico, emerge che la stima della differenza percentuale tra la probabilità di promozione, mediamente in ciascun anno e classe, degli studenti dei professionali rispetto a quelli dei licei risulta inferiore del 14%, mentre l'analoga stima relativa agli studenti dei tecnici rispetto a quelli dei licei è mediamente inferiore del 6%. Il modello a curva latente ritenuto ottimo stima che la differenza media percentuale negli anni degli studenti dei professionali sia, per quanto riguarda il valore della variabile *Times Promoted*, del 24% (il 20% osservato) inferiore rispetto a quelli dei licei, mentre l'analoga percentuale per gli studenti dei tecnici è stimata essere del 12% (il 9% osservato) inferiore rispetto a quelli dei licei; infine, i ragazzi che frequentano gli istituti artistici sono stimati avere in media valori della variabile obiettivo inferiori del 13% (il 12% osservato) rispetto ai ragazzi dei licei.

Dal confronto tra stime ed osservazioni, si evince che il modello ben si adatta ai dati osservati. Si ha che soltanto il 3% delle differenze tra valori stimati e valori osservati (773 differenze su 25177) sono uguali o superiori a 1,5; solo l'1% di tali differenze sono uguali o superiori a 2.

I Grafici 81, 82, 83 e 84 mostrano la separazione degli intervalli di confidenza tra le diverse tipologie di scuola, in particolare vengono confrontati tra loro i risultati relativi alle diverse tipologie di scuola, separatamente considerate. Nei grafici sono ben visibili anche le differenze tra studenti di sesso opposto (per i soli istituti tecnici e professionali) e la separazione dei relativi intervalli di confidenza stimati.

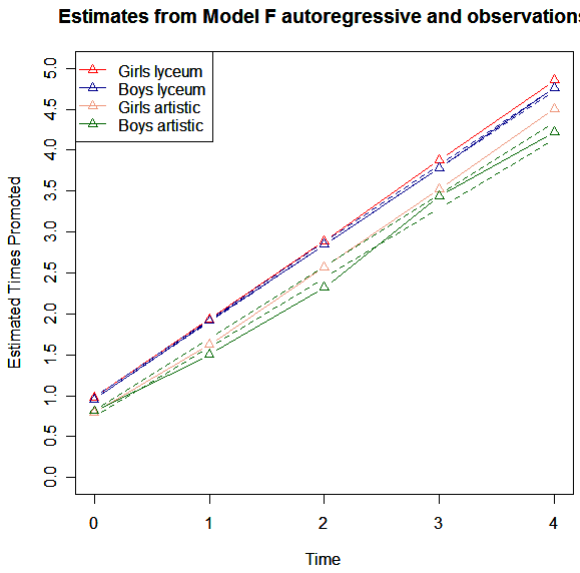


Grafico 81 – Valori osservati (linee continue) ed intervalli di confidenza stimati dal modello F autoregressivo eterogeneo (linee tratteggiate) – studenti dei licei e degli artistici

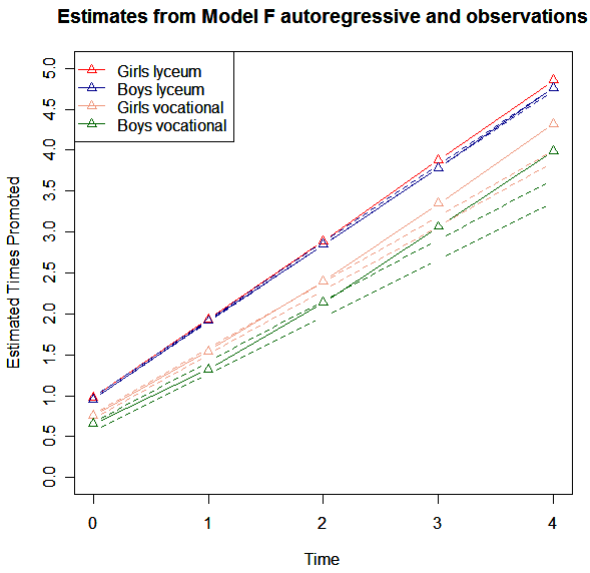


Grafico 82 – Valori osservati (linee continue) ed intervalli di confidenza stimati dal modello F autoregressivo eterogeneo (linee tratteggiate) – studenti dei licei e dei professionali

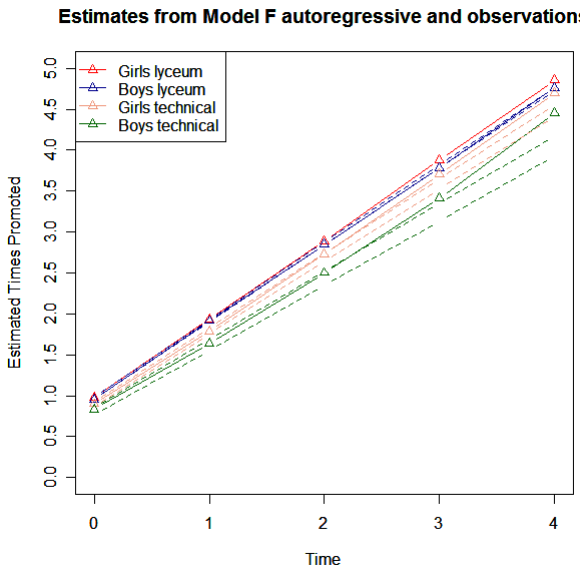


Grafico 83 – Valori osservati (linee continue) ed intervalli di confidenza stimati dal modello F autoregressivo eterogeneo (linee tratteggiate) – studenti dei licei e dei professionali

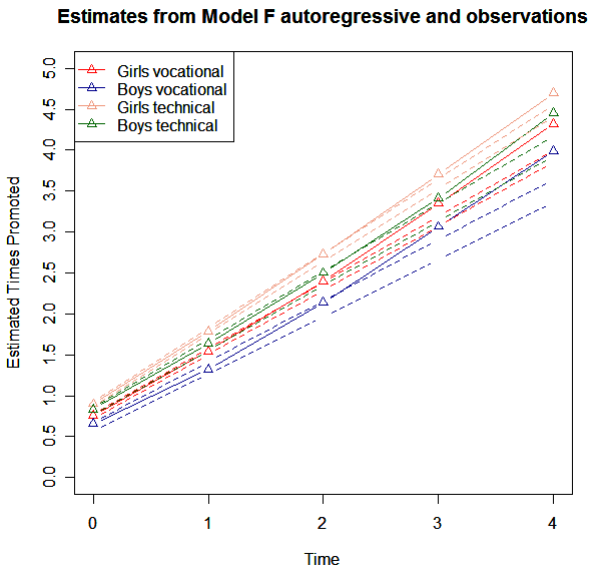


Grafico 84 – Valori osservati (linee continue) ed intervalli di confidenza stimati dal modello F autoregressivo eterogeneo (linee tratteggiate) – studenti dei tecnici e dei professionali

Vi è da notare il fatto che l'unica sovrapposizione tra intervalli di confidenza si verifica tra i risultati delle ragazze che frequentano i professionali e i ragazzi che frequentano invece i tecnici, maggiormente marcata,

specie nei primi due anni considerati. La conseguenza di tale sovrapposizione è il dubbio sulla significatività della differenza tra tali categorie di studenti, almeno nei primi anni di scuola.

Si può quindi rivedere l'asserto sulla generica differenza in quanto ad esiti scolastici tra studenti che frequentano gli istituti tecnici e professionali, affermando che in realtà la differenza è significativa soltanto tra ragazzi delle due tipologie di scuola e tra ragazze delle due tipologie di scuola, mentre la differenza tra le ragazze che frequentano i professionali ed i colleghi maschi che frequentano i tecnici è sicuramente inferiore, tanto da risultare di dubbia significatività, quantomeno nei primi anni di scuola. Il fatto che i risultati conseguiti dalle ragazze dei professionali e dai ragazzi dei tecnici siano simili nei primi anni, per poi differenziarsi negli anni successivi, è probabilmente sintomo del fatto che è soprattutto la tipologia di scuola in cui si è inseriti a determinare il risultato: nei tecnici si tende a conseguire risultati migliori che non nei professionali, anche partendo da condizioni omogenee. La differenza si accentua permanendo nella stessa tipologia di scuola.

La differenza comunque sostanziale è tra gli studenti dei licei e gli studenti che frequentano le altre tipologie di scuola, siano maschi o femmine. La permanenza al liceo favorisce esiti ulteriormente positivi.

Tali risultati non tengono tuttavia conto, come precedentemente accennato, della variegata realtà scolastica in termini di indirizzi di studio. La scuola secondaria di secondo grado, infatti, è caratterizzata da una molteplicità di indirizzi di studio e sperimentazioni, specie negli istituti tecnici e professionali, ma a volte anche nei licei, che si articolano in sperimentazioni linguistiche piuttosto che di tipo scientifico. Sarebbe interessante, disponendo di un numero maggiore di individui, studiare separatamente l'andamento degli esiti scolastici dei frequentanti i singoli indirizzi, o anche per raggruppamenti comunque ad un livello maggiormente disaggregato rispetto alla suddivisione per generica tipologia. In questa sede non è stato possibile effettuare una ulteriore suddivisione, proprio per l'eccessiva frammentazione del dataset che ne sarebbe derivata, con la conseguente perdita di precisione delle stime. In particolare, occorre tener conto del fatto che alcuni tipi di indirizzo attivati presso gli istituti tecnici sono di fatto molto simili al percorso di alcuni licei (scientifico e linguistico); ciò porta in realtà ad una omogeneizzazione dei percorsi scolastici anche tra studenti che frequentano tipologie di scuola diverse. Con la recente riforma della scuola secondaria di secondo grado, è stato notevolmente ridotto il numero di indirizzi di studio. Inoltre si è cercato di superare l'intersezione che si era creata negli anni, grazie ai diversi tipi di sperimentazione, tra corsi di studio di scuole anche di tipologia diversa. Se in futuro si potrà quindi disporre di dati individuali relativi a studenti che hanno frequentato le scuole del dopo riforma, sarà probabilmente possibile differenziare i percorsi scolastici in modo più approfondito rispetto al presente studio.

Un limite di fatto incontrato in questa sede è certo l'ampiezza del contingente di riferimento. In virtù della legge sull'obbligo formativo, anche alcune Regioni si sono attivate con lo scopo di raccogliere le informazioni individuali sugli studenti; le relative banche dati non sono al momento disponibili. Il Ministero dell'Istruzione ha avviato, già da alcuni anni, l'anagrafe degli studenti, raccogliendo i dati individuali direttamente dalle istituzioni scolastiche, in diversi momenti dell'anno. Dall'analisi statistica di tali banche dati potranno forse emergere informazioni aggiuntive anche sui fattori che concorrono alla spiegazione dell'andamento degli esiti scolastici.

Dalla stima del citato modello F2, che utilizza i soli studenti che risultano presenti in tutti gli anni considerati, e dal suo confronto con il modello che invece si basa sull'intero dataset, emerge un'altra importante realtà.

Se si escludono dalla procedura di stima coloro che partono con uno svantaggio, che quindi hanno già subito almeno una bocciatura alla scuola di primo grado, e coloro che escono prematuramente dal sistema scolastico, le differenze tra le categorie esaminate si assottigliano fino a perdere di significatività. Chi non parte svantaggiato e chi rimane all'interno del sistema ottiene risultati pressoché analoghi, a prescindere dalla tipologia di scuola frequentata e dal sesso. Le differenze, invece, tra le diverse categorie di studenti acquistano significatività se si analizzano le frange. Nel *drop-out* e nel disagio si differenziano realmente tra loro le categorie.

Non solo. Il sistema non pare in grado di invertire le sorti: chi parte con uno svantaggio è più probabile che si rivolga ad un istituto professionale e che poi lì rimanga, conseguendo magari ulteriori insuccessi; chi parte già in regola con il percorso di studi, è più probabile che si rivolga ad un liceo e che ottenga ulteriori successi.

BIBLIOGRAFIA

Alan C. Acock **"*Latent Growth Curve Analysis: A Gentle Introduction*" -**Department of Human Development and Family Sciences - Oregon State University - Oregon Research Institute 1999

Bengt Muthen and Tihomir Asparouhov "***Longitudinal Data Analysis: Handbooks of Modern Statistical Methods***" Chapman & Hall/CRC, 2008

Bollen, K.A. 2007 "***On the Origins of Latent Curve Models***" Pages 79-98 in Robert Cudeck and Robert MacCallum (eds) Factor Analysis at 100. Mahwah, NJ:Lawrence Erlbaum Associates.

Conor V. Dolan, Verena D. Schmittmann, Gitta H. Lubke, Michael C. Neale "***Regime Switching in the Latent Growth Curve Mixture Model***" STRUCTURAL EQUATION MODELING, *12*(1), 94–119, 2005

Donna L. Coffman, Roger E. Millsap ***"Evaluating Latent Growth Curve Models Using Individual Fit Statistics"*** STRUCTURAL EQUATION MODELING, *13*(1), 1–27 - 2006

Feiniang Chen, Kenneth A. Bollen et al. "***Impoper solutions in structural equation models***" , Sociological methods and research Vol.29 n.4 2001.

Feiniang Chen, Kenneth A. Bollen, Patrick J. Curran et al. "***An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models***", Sociological methods and research Vol.36 n.4 2008.

Fondazione Giovanni Agnelli "***Rapporto sulla scuola in Italia 2010***" Editori Laterza

Fuzhong Li, Terry E. Duncan, Alan Acock "***Modeling Interaction Effects in Latent Growth Curve Models***", STRUCTURAL EQUATION MODELING, *7*(4), 497–533, 2000

Gerhard Arminger, Ronald J. Schoenberg "***Pseudo maximum likelihood estimation and a test for misspecification in mean and covariance structure models"*** Psychometrika, 1989

Irene Martelli, Tesi di Laurea "***L'analisi della relazione tra rendimento scolastico e contesto familiare: uno studio relativo alle scuola secondarie di II grado della provincia di Bologna***" a.a. 2008/09 Facoltà di Scienze Statistiche

John R. Hipp, Kenneth A. Bollen "***Model fit in SEM with censored, ordinal and dichotomous variables: testing vanishing tetrads***", paper presented at 2002 Psychometric Society meeting.

Judith D. Singer, John B. Willett "***Applied longitudinal data analysis – modelling change and event occurrence***" Oxford 2003

Kenneth A. Bollen "***A new incremental fit index for general structural equation models***" Sociological methods and research Vol.17 n.3 1989

Kenneth A. Bollen "***Overall fit in covariance structure models: two types of sample size effects***" Psychological Bullettin vol.107 n.2 1990.

Lea E. Witta, Stephen A. Sivo "***Latent growth model of cognition in the elderly***" Paper presented at the annual meeting of the American Educational Research Association, 2003

Patrick J. Curran ***"Comparing Three Modern Approaches to Longitudinal Data Analysis: An Examination of a Single Developmental Sample"*** A Symposium to be Conducted at the 1997 Biennial Meeting of the Society for Research in Child development Washington, D.C.

Patrick J. Curran, Daniel J. Bauer, and Michael T. Willoughby "***Testing Main Effects and Interactions in Latent Curve Analysis***" Psychological Methods 2004, Vol. 9, No. 2, 220–237

Patrick J. Curran, Kenneth A. Bollen "***A Hybrid Latent Trajectory Model of Stability and Change: Applications in Developmental Psychopathology***" - Paper Presented at the 1999 Biennial meeting of the Society for Research on Child Development, Albuquerque, New Mexico.

Patrick J. Curran, Kenneth A. Bollen "***Latent Curve Models – A structural equation perspective*"** Wiley 2006

Patrick J. Curran, Kenneth A. Bollen, Feinian Chen, Pamela Paxton, James Kirby "***Finite sampling properties of the point estimates and confidence intervals of RMSEA***" Sociological methods and research Vol.32, No.2, 2003.