Università degli Studi di Roma Tre

FACOLTÀ DI ECONOMIA
Scuola Dottorale in Economia e Metodi Quantitativi
Sez. Metodi Statistici per l'Economia e l'Impresa

XXV Ciclo

Modelli grafici gerarchici e Item Response Theory: un'applicazione ai dati PISA 2006

Candidato:
Vincenzina Vitale

Supervisori:
Prof.ssa Daniela Marella
Prof.ssa Paola Vicard

Coordinatore:
Prof.ssa Julia Mortera

25 Giugno 2013

Abstract

I modelli dell'Item Response Theory (Lord e Novick 1968; Rasch 1960) sono una particolare classe di modelli matematico - probabilistici la cui diffusione è legata al crescente utilizzo, in particolare in ambito psicometrico e sociale, del questionario come strumento fondamentale per la misurazione di uno o più costrutti latenti. L'idea di base è quella di tradurre le informazioni, ottenute a partire dalle risposte osservate, in misurazioni oggettive del tratto latente, alla stregua di quanto avviene nelle scienze fisiche. Il modello di Rasch, tra i più noti nella classe dei modelli IRT, rispetta i criteri relativi al concetto di misura e presuppone che la probabilità di risposta corretta ad un item, da parte di un soggetto, sia funzione della differenza tra due parametri: l'abilità del soggetto e la difficoltà dell'item.

In particolari contesti applicativi, quale quello dell'Educational Assessment, è frequente che tale classe di modelli venga applicata a matrici di dati incomplete, per le quali è ipotizzabile che il meccanismo generatore del dato mancante sia non ignorabile (Rubin 1976), determinando distorsioni nelle stime dei parametri dei modelli IRT.

A tale proposito, il modello proposto da Holman e Glas (2005) tiene conto del meccanismo generatore del dato mancante considerando, oltre all'abilità, una seconda dimensione latente, la propensione alla risposta, la quale, in caso di dati MNAR (Missing Not At Random), è correlata alla prima dimensione.

Obiettivo principale di questo lavoro di ricerca è stato quello di dimostrare che è possibile, nonché equivalente, definire il modello proposto da Holman e Glas utilizzando il linguaggio e le potenzialità dei modelli grafici (Lauritzen 1996) rendendo evidenti, mediante grafo diretto aciclico, le relazioni di indipendenza condizionata tra le variabili manifeste, e non, del modello. In particolare, si è definito un modello di Rasch bidimensionale, in ottica *between items* (Adams, Wilson e Wang 1997). L'approccio bayesiano ha arricchito, in termini di flessibilità, la rappresentazione grafica consentendo di attribuire, a tutti i nodi stocastici del modello, una distribuzione di probabilità a priori.

La struttura dei dati, un campione italiano dell'indagine PISA 2006, ha suggerito l'introduzione di un ulteriore elemento di complessità: la caratteristica dei dati di essere "annidati" in livelli, o gerarchie, ha spinto la ricerca a considerare l'estensione dei modelli, prima descritti, al caso multilivello (secondo la formulazione che fa riferimento ai modelli GLAMM).

L'analisi dei risultati è caratterizzata dal confronto tra le stime del modello bivariato (che tiene conto della correlazione tra abilità e propensione alla risposta) e quelle ottenute a partire da altri due modelli, l'uno che ha supposto un meccanismo MAR, l'altro che ha utilizzato una matrice completa in cui al posto dei dati mancanti è stato imputato lo zero (la risposta sbagliata).

A corredo dell'analisi, si è aggiunto lo studio delle covariate con l'intento di valutare l'impatto delle stesse su entrambi i processi: l'obiettivo è valutare quali caratteristiche, legate allo studente e al suo background socio - familiare, influenzino le fasi dell'apprendimento e, in questo caso particolare, anche la strategia di risposta.
In ultimo, per completare l'analisi e lo studio delle distorsioni sui parametri d'interesse, si è considerata l'analisi del *Differential Item Functioning* (DIF); lo scopo è, ancora, di comparazione tra i tre modelli al fine di verificare quali ripercussioni si possano avere sull'analisi del DIF qualora il meccanismo generatore del dato mancante non sia ignorabile.
In generale, quindi, gli scopi della ricerca in questione sono stati molteplici e interconnessi; si sono combinate le proprietà dei modelli grafici, in ottica bayesiana, e quelle dei modelli IRT multilivello con l'ulteriore obiettivo di comparazione tra tre diversi approcci al trattamento dei dati mancanti.
L'applicazione al campione dei dati italiani, considerando la stratificazione in gruppi regionali, ha reso più interessante e particolare lo studio confermando, anche in relazione a questa analisi, le forti differenze e sperequazioni, territoriali e sociali, che caratterizzano, da secoli, la nostra penisola.

Ringraziamenti

A conclusione di questo triennio di studi, sento la necessità di ringraziare tutti coloro che hanno contribuito alla mia crescita, non solo professionale ma anche umana.
In primis, ringrazio la prof.ssa Julia Mortera, coordinatrice della scuola dottorale in Economia e Metodi Quantitativi, mentore e guida in questo percorso di studio e di ricerca entusiasmante, dalla quale ho imparato che al rigore scientifico si accompagnano, in egual misura, onestà intellettuale, umiltà ed entusiasmo continuo per il proprio lavoro.
Allo stesso modo, ringrazio la Prof.ssa Paola Vicard e la Prof.ssa Daniela Marella, che hanno supervisionato il mio lavoro di tesi con pazienza e disponibilità. Le ringrazio per il supporto scientifico e morale, per i preziosi consigli che mi hanno elargito, per il bagaglio di conoscenze teoriche che mi hanno trasferito introducendomi allo studio dei Modelli Grafici e dell'Item Response Theory.
Ringrazio, ancora, tutti i ricercatori e i docenti che, attraverso cicli di lezioni e seminari, hanno contribuito ad accrescere il mio bagaglio di conoscenze teoriche e hanno suscitato interesse e curiosità per nuovi spunti di ricerca.
L'esperienza del dottorato è stata molto proficua poiché ho avuto la possibilità di condividere, sin dal primo giorno, un sodalizio non solo professionale ma soprattutto umano con i miei due colleghi di ciclo, Federico e Marco. A loro va la mia stima per i meriti scientifici. Li ringrazio per la pazienza e i bei momenti di condivisione.
Ringrazio tutti gli altri colleghi di dottorato e, in particolare, Flaminia e Francesca, per il loro supporto e per avermi regalato una bellissima amicizia che, sono sicura, durerà lo spazio di una vita. Le ringrazio per le chiacchiere e le risate, per aver condiviso con loro ansie e gioie, fatiche e speranze.
Ringrazio ancora mio marito Pino che, in alcuni momenti, ci ha creduto più di me, che mi ha sostenuta e incentivata, in ogni istante, con infinita pazienza, che ha compreso e condiviso la passione per lo studio e il desiderio di nuove e stimolanti sfide. Ringrazio mia madre e mio fratello per il sostegno morale, mio padre per essermi stato "virtualmente" sempre accanto, i familiari e gli amici.
Concludo questo percorso di studio con la consapevolezza di aver allargato il mio orizzonte di conoscenze, desiderosa di intraprendere e approfondire nuovi sentieri di ricerca, appagata per aver intrecciato rapporti di amicizia nuovi e sinceri e certa che l'esperienza maturata in questo triennio sia stata, tra tutte, la sfida più fruttuosa ed entusiasmante.

Indice

Elenco delle figure

Elenco delle tabelle

Prefazione

Il presente lavoro di tesi si pone come obiettivo primario quello di definire un modello IRT[1] bidimensionale utilizzando le proprietà e il linguaggio dei modelli grafici bayesiani.
La forza della rappresentazione grafica risiede nella sua capacità di rendere evidenti e chiare le relazioni d'indipendenza condizionata tra variabili, sia manifeste che latenti; l'approccio bayesiano arricchisce, in termini di flessibilità, tale rappresentazione grafica poiché consente di attribuire a tutti i nodi stocastici del modello una distribuzione di probabilità *a priori*.
In particolare, si vuole definire un modello di Rasch bidimensionale, in ottica *between items* (Adams, Wilson e Wang 1997), la cui peculiarità consiste nel fatto che ciascun gruppo di items misura una sola delle due dimensioni latenti.
Tale scelta non è casuale, piuttosto è strettamente connessa al secondo obiettivo dell'indagine, che fa riferimento all'analisi e trattamento dei dati mancanti; il fine è valutare eventuali distorsioni nelle stime dei parametri d'interesse (parametri di difficoltà e di abilità) in caso di probabile violazione del principio di ignorabilità del meccanismo generatore del dato mancante, così come definito da Rubin (1976).
A tale scopo vengono proposti tre modelli IRT: un modello univariato che ipotizza un meccanismo di tipo MAR[2] poiché sfrutta la capacità del software di imputare i valori mancanti sulla base della distribuzione predittiva a posteriori, condizionata ai soli dati osservati e ai parametri non noti del modello; un secondo modello univariato che sostituisce al valore mancante lo zero definendo, quindi, la mancata risposta sempre come una risposta sbagliata ed un terzo modello, che fa riferimento alla formulazione bivariata, che

[1] Item Response Theory
[2] Missing At Random

presuppone che i dati mancanti siano MNAR[3] e che, quindi, modellizza il meccanismo generatore del dato mancante.
L'anello di congiunzione tra modelli IRT e *missing data* è rintracciabile nel lavoro di Holman e Glas (2005), i quali hanno proposto di considerare, in caso di probabile violazione del principio di ignorabilità, un modello IRT bidimensionale, che definisce due dimensioni latenti: l'abilità e la propensione alla risposta.
Nel modello MNAR da loro formulato, le due dimensioni latenti sono tra loro correlate; il grado di distorsione delle stime è strettamente connesso alla forza di tale legame e direttamente proporzionale al numero di dati mancanti presente nel dataset.
In questo studio, come già anticipato, si è deciso di considerare tale modello in ottica *between items* (Rose, von Davier e Xu 2010): esso introduce, per la seconda dimensione relativa alla propensione alla risposta, un gruppo di items fittizi, che assumono valore zero in caso di mancata risposta nell'item corrispondente nella matrice dei dati osservati, uno altrimenti.
La scelta del datatset, un campione dei dati PISA 2006, non è casuale ma motivata dalla considerazione secondo la quale, in matrici di dati appartenenti a tale contesto, è abbastanza logico supporre che il pattern della mancata risposta sia tutt'altro che casuale; è lecito, piuttosto, ritenere che a maggiore abilità corrisponda una maggiore propensione alla risposta e, quindi, una correlazione tra queste due dimensioni.
La struttura dei dati ha, poi, suggerito l'introduzione di un ulteriore elemento di complessità nell'indagine; la loro caratteristica di essere "annidati" in livelli, o gerarchie, ha spinto la ricerca a considerare l'estensione dei modelli, prima descritti, al caso multilivello, non potendo ignorare tale caratteristica intrinseca al campione utilizzato. In generale, quindi, gli scopi della ricerca in questione sono molteplici e interconnessi; si vogliono combinare le proprietà dei modelli grafici, in ottica bayesiana, e quelle dei modelli IRT multilivello con l'ulteriore obiettivo di comparazione tra tre diversi approcci al trattamento dei dati mancanti.
A corredo dell'analisi, si aggiunge lo studio delle covariate con l'intento di valutare l'impatto delle stesse su entrambi i processi, sia quello dell'abilità che della propensione alla risposta; l'obiettivo è valutare quali caratteristiche, legate allo studente e al suo background socio - familiare, influenzino le fasi dell'apprendimento e, in questo caso particolare, anche la strategia di risposta.
In ultimo, per completare l'analisi e lo studio delle distorsioni sui parametri d'interesse, si aggiunge l'analisi del *Differential Item Functioning* (DIF); lo

[3]Missing Not At Random

scopo è, ancora, di comparazione tra i tre modelli al fine di verificare quali ripercussioni si possano avere sull'analisi del DIF qualora il meccanismo generatore del dato mancante non sia ignorabile.
Lo sviluppo e la trattazione degli argomenti descritti prevede che il primo capitolo si dedichi interamente alla definizione e trattazione delle proprietà dei modelli grafici, in particolare dei grafi diretti aciclici (DAG), con riguardo alla definizione degli stessi secondo l'approccio bayesiano definito dal software Openbugs.
Il secondo capitolo focalizza l'attenzione, dapprima, sulle proprietà dei modelli IRT, per poi incentrare la trattazione sulle caratteristiche degli stessi in ottica multilivello.
Il terzo capitolo descrive, in breve, i diversi approcci al trattamento dei dati mancanti per poi mostrare, nel dettaglio, la tecnica utilizzata in questo studio.
Il quarto capitolo entra nel dettaglio analitico della definizione dei modelli ed esplicita la rappresentazione grafica corrispondente.
Il quinto e ultimo capitolo conclude la trattazione con l'applicazione dei modelli all'analisi dei dati reali, esaminando e comparando i risultati riferiti ai tre diversi approcci al trattamento dei dati mancanti.

Capitolo 1

Introduzione ai Modelli Grafici

1.1 Genesi e Storia

I modelli grafici hanno origine da applicazioni in diverse aree scientifiche, si spazia dalla Fisica (Gibbs 1902), alla Genetica (Wright 1923, 1934), all'Economia (Wold 1954). In ambito statistico, la loro prima comparsa si accompagna a due tipologie di modelli: *covariance selection models* (Dempster 1972) e *loglinear models* (Haberman 1902). Queste due classi di modelli, unitamente all'utilizzo di alcuni dei concetti fondamentali della *Path Analysis* e dell'indipendenza (condizionata), concorrono a definirne gli aspetti fondamentali.
Segue che un modello grafico si può definire come *un modello statistico probabilistico per vettori aleatori multivariati la cui struttura di indipendenza è caratterizzata da un grafo (di indipendenza condizionata).*
Il contributo scientifico che più ne ha favorito la diffusione è l'articolo scritto da Darroch, Lauritzen e Speed (1980); in esso gli autori dimostrano come un sottoinsieme di modelli log-lineari, i modelli log-lineari grafici appunto, possano essere facilmente interpretati mediante l'applicazione e la verifica delle proprietà markoviane associate al grafo di indipendenza condizionata.
Ad esso si aggiungono altri contributi, non meno importanti, ad opera di Wermuth e Lauritzen (1983, 1989).
Tra la fine degli anni Ottanta e l'inizio degli anni Novanta, i modelli grafici acquistano maggiore visibilità e forza grazie ai contributi scientifici dei seguenti autori: Pearl (1988), Lauritzen e Spiegelhalter (1988), Whittaker (1990), Lauritzen (1996), Cox e Wermuth (1996). Negli anni a seguire fino ad oggi, numerosi sono stati e sono i lavori scientifici che ne diffondono princìpi, proprietà e applicazioni, oltre a sostenerne l'importanza e la validità.

1.1.1 Usi, Potenzialità e Applicazioni

"*The powerful feature of graphical models are partly related to the fact that graphs and graphs structures are easy to communicate to computers, but not least their visual rappresentation*"[1]. La frase appena citata descrive con chiarezza quali siano le potenzialità dei modelli grafici e, di conseguenza, i suoi campi applicativi.
Il grafo è un potente mezzo di rappresentazione visiva, tale aspetto non è da sottovalutare in quanto favorisce, in molti casi, una migliore comunicazione tra statistici e ricercatori. La struttura del grafo ha, inoltre, il grande vantaggio di poter essere "appresa" dai moderni calcolatori.
Ultimo aspetto, ma non meno importante, i modelli grafici si caratterizzano per la loro modularità per cui diventa possibile gestire problemi complessi decomponendoli in sottostrutture più semplici.
Al fine di comprendere proprietà e potenzialità di tale classe di modelli, i paragrafi 1.2 e 1.3 introdurranno ai concetti fondamentali relativi alla teoria dei grafi; i paragrafi 1.4, 1.5 e 1.6 approfondiranno il concetto di indipendenza condizionata, o meglio la relazione tra esso e il concetto di separazione nel grafo, nonché le proprietà di Markov per grafi indiretti e per grafi diretti aciclici. L'ultimo paragrafo si concentrerà sulla definizione dei modelli grafici in ottica bayesiana e sul legame tra essi e l'algoritmo del Gibbs Sampling.
Sulla base di quanto esposto, diventa facile intuire quali siano le destinazioni d'uso dei modelli grafici: essi trovano, in particolare, larga diffusione nell'ambito del Data Mining, come Sistemi Esperti Probabilistici, come strumenti per eseguire Inferenza Causale e come supporto per la Teoria delle Decisioni.

1.2 Cenni di Teoria dei Grafi

Un modello grafico è un modello statistico multivariato per un vettore aleatorio che verifica vincoli di indipendenza (condizionata) rappresentati mediante grafo. È chiaro che, per meglio comprendere il significato dei modelli in questione, diventa essenziale definire cos'è un grafo.
Un *grafo* G è formato da due insiemi V ed E, possiamo quindi scrivere:

$$\mathbf{G} = (\mathbf{V}, \mathbf{E}) \tag{1.1}$$

dove $V = \{1, ...k\}$ rappresenta l'insieme finito dei vertici (o nodi), mentre $E \subseteq V \times V$ rappresenta l'insieme finito degli archi.
I vertici, o nodi, rappresentano le nostre variabili casuali, gli archi le relazioni di dipendenza (condizionata).

[1] Inciso tratto dal materiale didattico riferito al ciclo di lezioni del Prof. S. L. Lauritzen presso Università di Roma Tre, Novembre 2012

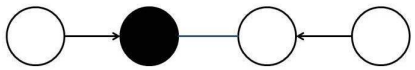


Figura 1.1: Grafo con archi sia diretti che indiretti, con un solo vertice discreto e i restanti continui.

L'arco $(\alpha, \beta) \in E$ si definisce *diretto* se l'insieme E contiene la coppia ordinata (α, β): il vertice α è il nodo *genitore* del vertice β, mentre quest'ultimo è il nodo *figlio* del vertice α.
L'arco $(\alpha, \beta) \in E$ si definisce *indiretto* se l'insieme E contiene entrambe le coppie (α, β) e (β, α).
Se tutti gli archi sono non orientati il grafo stesso si definisce *indiretto*.
Definiamo l'insieme dei vertici V come:

$$V = \Delta \cup \Gamma \; con \; \Delta \cap \Gamma = \emptyset$$

Δ rappresenta l'insieme delle variabili qualitative, Γ quello delle variabili quantitative. I vertici in Δ si dicono *discreti*, quelli in Γ *continui*.
Se un grafo è composto da una sola delle due tipologie si dice *puro*.
Per la rappresentazione grafica dei vertici si utilizzano i cerchi neri ($\bullet$) per le variabili discrete, quelli bianchi ($\circ$) per le variabili continue.
Gli archi, invece, sono rappresentati da linee o da frecce a seconda se si debba rappresentare, rispettivamente, un arco indiretto o diretto (vedi Fig.1.1).
Si indica con $\alpha \rightarrow \beta$ l'arco diretto, con $\alpha \sim \beta$ quello indiretto, così come proposto in Lauritzen (1996).
Se l'obiettivo è studiare relazioni di dipendenza che non abbiano caratteristiche di direzionalità, è possibile ottenere un grafo indiretto, a partire dal grafo rappresentato in Figura 1.1, sostituendo alle frecce orientate archi non orientati.
Sia $A \subseteq V$,

$$G_A = (A, E_A)$$

si definisce sottografo indotto da A, dove $E_A = E \cap (A \times A)$ lo si ottiene a partire da G mantenendo i soli archi con entrambi gli estremi in A.
Un grafo è *completo* se tutti i vertici sono collegati da un arco.
Una *clique* è un sottografo massimale completo, per il quale l'introduzione di un altro vertice lo renderebbe incompleto.
Come già evidenziato, se $\alpha \rightarrow \beta$, α è nodo genitore di β e β nodo figlio di α. L'insieme dei genitori del nodo β si indica con $pa(\beta)$ e l'insieme dei nodi figli con $ch(\alpha)$.

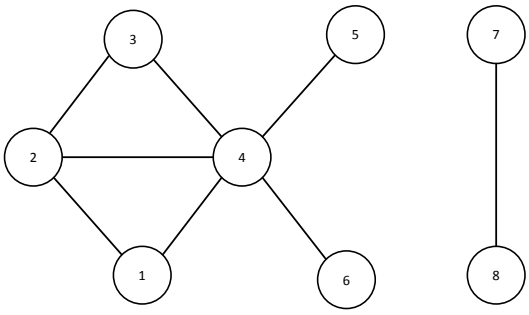


Figura 1.2: Esempio di Grafo non orientato e non connesso.

Due nodi si dicono *adiacenti*(o *vicini*) se sono uniti da un arco non orientato. L'insieme dei *vicini* di α si indica con $ne(\alpha)$.
La *frontiera* di un nodo α è l'insieme dei vertici in $V \setminus \alpha$ che sono genitori o vicini di α. Si indica con il simbolo $bd(\alpha)$.
Con il simbolo $cl(\alpha)$ si indica la chiusura di α, ossia $cl(\alpha) = \alpha \cup bd(\alpha)$.
Si definisce *cammino* di lunghezza n tra α e β la sequenza $\alpha = \alpha_0, ..., \alpha_n = \beta$ di vertici distinti tale che l'arco $(\alpha_{i-1}, \alpha_i) \in E$ per $i = 1, ...n$.
Si definisce, invece, *ciclo* di lunghezza n un *cammino* in cui $\alpha = \beta$. Il ciclo si dice *diretto* se contiene almeno un arco orientato.
Due vertici α e β si dicono *connessi* se esiste un cammino da α a β e da β a α.
Un grafo si dice *connesso* se tutti i nodi sono a due a due connessi.
Dato un grafo G non connesso, si definiscono *componenti connesse* i suoi sottografi massimali connessi. Ad esempio, in Figura 1.2, si ha un grafo non connesso in cui le componenti connesse sono $\{1, 2, 3, 4, 5, 6\}$ e $\{7, 8\}$; $\{4, 5\}$ induce un sottografo connesso ma non è una componente connessa in quanto non è il sottografo massimale connesso.
Dato un grafo indiretto ed un ciclo di lunghezza maggiore di 3, si definisce *corda* l'arco che unisce due nodi non consecutivi del ciclo. Un grafo è, inoltre, *triangolato* se non contiene cicli di 4 o più vertici senza una corda.

Separazione Un sottinsieme $C \subseteq V$ separa due vertici, α e β, se tutti i cammini da α a β intersecano C. Quest'ultimo separa due sottoinsiemi, $A \subseteq V$ e $B \subseteq V$, se separa ogni coppia di vertici $\alpha \in A$ e $\beta \in B$.
C è detto *separatore* di A e B.

Decomponiblità Consideriamo un grafo indiretto G ed una terna (A, B, C), partizione di V; la terna costituisce una *decomposizione forte* di G se:

- C separa A da B;
- C è un sottinsieme completo di V;
- $C \subseteq \Delta \vee B \subseteq \Gamma$.

La partizione $V = A \cup B \cup C$ decompone G in $G_{A \cup C}$ e $G_{B \cup C}$.
Si ha una *decomposizione debole* se non vale la terza proprietà.
Se $A \neq \emptyset$ e $B \neq \emptyset$, la decomposizione è detta *propria*.
Un grafo si dice *primo* se non ammette una decomposizione *propria*.
Si può dire che un grafo è *decomponibile* quando lo si può scomporre nelle sue cliques. In particolare, è *decomponibile* se:

- È completo;
- Esiste una decomposizione propria (A, B, C) in sottografi decomponibili $G_{A \cup C}$ e $G_{B \cup C}$;
- È triangolato.

Un grafo è *decomponibile in senso debole* se può essere decomposto nelle sue cliques mediante una decomposizione debole. Un grafo decomponibile lo è anche in senso debole, non vale il viceversa.
In Figura 1.3, è riportato un esempio di grafo decomponibile. Se il medesimo grafo non avesse l'arco (2,4), non sarebbe decomponibile.

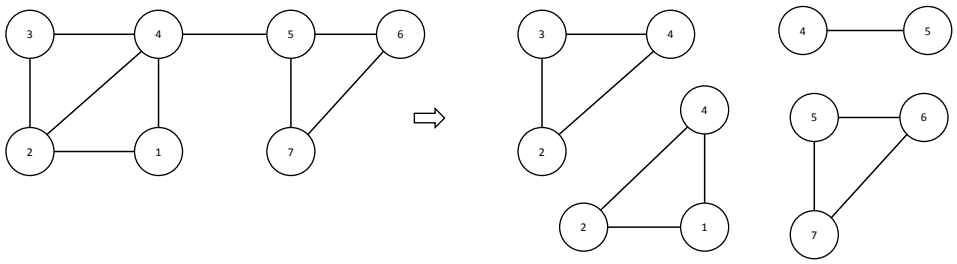


Figura 1.3: Esempio di grafo decomponibile

1.3 Tipologie di Grafi

Escludendo la trattazione dei grafi con archi bidirezionali, i restanti si distinguono in tre tipi:

- Grafi indiretti (UG), caratterizzati da archi non orientati;
- Grafi diretti aciclici (DAG), caratterizzati dal fatto che non possono contenere cicli orientati e dalla presenza di soli archi orientati;
- Grafi a Catena (CG), caratterizzati da entrambe le tipologie di archi. In particolare, sia ammessa su V una partizione ordinata di blocchi non vuoti $B_1, ..., B_k$, con $k \geq 1$, allora si ha che:

  $(\alpha, \beta) = \alpha \sim \beta$ se $\alpha, \beta \in B_i$

  $(\alpha, \beta) = \alpha \rightarrow \beta$ se $\alpha \in B_i$ e $\beta \in B_j$, con $j > i$

In Figura 1.4 sono riportati tre esempi di grafo, uno per ciascuna tipologia descritta.

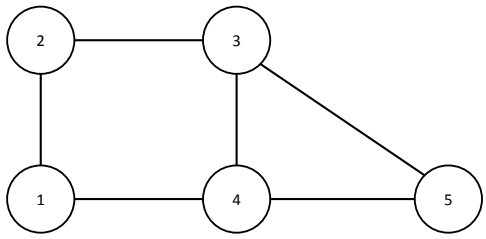
Avendo definito i grafi diretti aciclici, è necessaria una piccola digressione relativa alle seguenti definizioni proprie della teoria dei grafi: dati due nodi α e β ed un cammino discendente $\alpha \rightarrow ... \rightarrow \beta$ tra di essi, α si definisce *antenato* di β e β *discendente* di α.
Si indica con $an(\alpha)$ l'insieme degli *antenati* di α, con $de(\alpha)$ l'insieme dei *discendenti* e con $nd(\alpha)$ l'insieme dei *non discendenti* di α.
Sia $A \subseteq V$, si definisce *insieme ancestrale* di A, e lo si indica con $An(A)$, l'insieme di tutti gli antenati di A. Se A non ha nodi genitori, è esso stesso ancestrale.

1.4 Indipendenza Condizionata

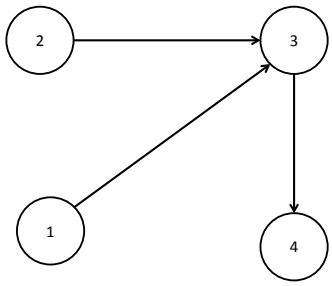
Data la terna di variabili casuali X, Y, Z con distribuzione congiunta P o densità f, diremo che X è ***indipendente*** da Y ***condizionatamente*** a Z, in simboli $X \perp\!\!\!\perp Y \mid Z$, se e solo se

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z) \qquad (1.2)$$

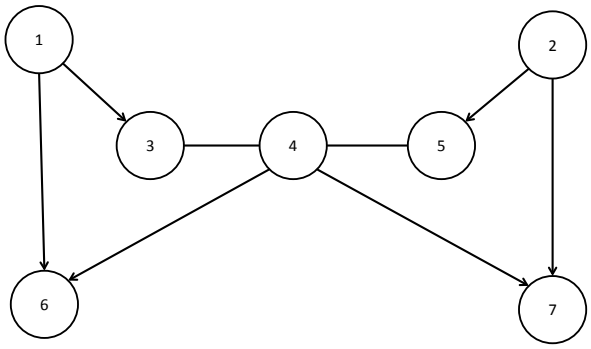
per tutte le z con $P(Z = z) > 0$.



(a)



(b)



(c)

Figura 1.4: Tipologie di grafi: indiretto (a), diretto aciclico (b), a catena (c)

Oppure se

$$f_{XY|Z}(x, y \mid z) = f_{X|Z}(x \mid z) f_{Y|Z}(y \mid z) \tag{1.3}$$

per tutte le z con $f_z(z) > 0$.

È chiaro, sulla base di quanto esposto, che dire $X \perp\!\!\!\perp Y \mid Z$ significa affermare che la presenza di Z rende Y irrilevante per spiegare X.
Se, invece, X è *indipendente* da Y marginalmente, in simboli, scriveremo: $X \perp\!\!\!\perp Y$.

Formulazioni alternative ed equivalenti di indipendenza condizionata sono le seguenti:

$$f_{XYZ}(x, y, z) f_Z(z) = f_{XZ}(x, z) f_{YZ}(y, z) \tag{1.4}$$

Vale per ogni z con densità continue.

$$\exists a, b : f_{XYZ}(x, y, z) = a(x, z) b(y, z) \tag{1.5}$$

Essa è il criterio di fattorizzazione per l'indipendenza condizionata. a e b sono funzioni che non coincidono necessariamente con le densità marginali.

$$f_{X|YZ}(x, y, z) = f_{X|Z}(x, z) \tag{1.6}$$

1.4.1 Proprietà

Date le variabili X, Y, Z, W e una funzione misurabile arbitraria t sullo spazio campionario di X, ne seguono le seguenti proprietà:

A.1 Se $X \perp\!\!\!\perp Y \mid Z$, vale che $Y \perp\!\!\!\perp X \mid Z$;

A.2 Se $X \perp\!\!\!\perp Y \mid Z$ e $H = t(X)$, vale che $H \perp\!\!\!\perp Y \mid Z$;

A.3 Se $X \perp\!\!\!\perp Y \mid Z$ e $H = t(X)$, vale che $X \perp\!\!\!\perp Y \mid (Z, H)$;

A.4 Se $X \perp\!\!\!\perp Y \mid Z$ e $X \perp\!\!\!\perp W \mid (Y, Z)$, vale che $X \perp\!\!\!\perp (W, Y) \mid Z$;

A.5 Se $X \perp\!\!\!\perp Y \mid Z$ e $X \perp\!\!\!\perp Z \mid Y$, vale che $X \perp\!\!\!\perp (Y, Z)$.

La condizione A.5 vale solo sotto opportune condizioni, ossia se $f(x, y, z, w) > 0$.

Si considerino, ora, le medesime proprietà definite relativamente ai vertici di un grafo.

Grafoidi e Semigrafoidi

Se consideriamo un insieme finito di V e i sottinsiemi disgiunti (A, B, C, D), un *modello di indipendenza* $\perp_\sigma$ su V è un *semigrafoide* se, per tutti i sottinsiemi citati, valgono:

B.1 *Simmetria*
Se $A \perp_\sigma B \mid C$, vale che $B \perp_\sigma A \mid C$;

B.2 *Decomposizione*
Se $A \perp_\sigma (B \cup D) \mid C$, vale che $A \perp_\sigma B \mid C$ e $A \perp_\sigma D \mid C$;

B.3 *Unione Debole*
Se $A \perp_\sigma (B \cup D) \mid C$, vale che $A \perp_\sigma B \mid (C \cup D)$;

B.4 *Contrazione*
Se $A \perp_\sigma B \mid C$ e $A \perp_\sigma D \mid (B \cup C)$, vale che $A \perp_\sigma (B \cup D) \mid C$.

Se il semigrafoide soddisfa anche la proprietà

B.5 *Intersezione*
Se $A \perp_\sigma B \mid (C \cup D)$ e $A \perp_\sigma C \mid (B \cup D)$, vale che $A \perp_\sigma (B \cup C) \mid D$;

esso è detto *Grafoide*.

Possiamo definire un modello di indipendenza $\perp\!\!\!\perp$ per un sistema di V variabili casuali X_v, con $v \in V$ e distribuzione P, attraverso la seguente relazione:

$$A \perp\!\!\!\perp B \mid C \iff X_A \perp\!\!\!\perp X_B \mid X_C$$

dove $X_A = (X_v, v \in A)$.

Separazione e Indipendenza condizionata. È evidente che esiste una relazione importante tra i concetti di *separazione* nel grafo e di *indipendenza condizionata* tra variabili causali.
Il grafo di indipendenza condizionata di $(X_\alpha)_{\alpha \in V}$ è rappresentato da un grafo indiretto $G = (V, E)$, con archi mancanti tra le coppie di vertici corrispondenti alle coppie di variabili casuali che sono indipendenti condizionatamente alle restanti variabili. Avremo, cioè che l'arco $(\alpha, \beta) \notin E \iff X_\alpha \perp\!\!\!\perp X_\beta \mid X_{V \setminus \{\alpha, \beta\}}$.

Il grafo G, quindi, fornisce informazioni riguardo le interdipendenze tra variabili, ossia se i nodi che le rappresentano sono adiacenti o separati.
È bene, a tale riguardo, enunciare le tre proprietà di Markov per grafi indiretti per poi soffermare l'attenzione sulle medesime proprietà relativamente ai grafi diretti aciclici.

1.5 Proprietà di Markov per Grafi Indiretti

Dato un grafo indiretto $G = (V, E)$ ed un vettore aleatorio $(X_\alpha)_{\alpha \in V}$, la distribuzione di probabilità congiunta P verifica la proprietà di Markov:

(P) **a coppie** se, per ogni coppia (α, β) di nodi non adiacenti

$$X_\alpha \perp\!\!\!\perp X_\beta \mid X_{V \setminus \{\alpha,\beta\}};$$

(L) **locale** se, per ogni vertice $\alpha \in V$

$$X_\alpha \perp\!\!\!\perp X_{V \setminus cl(\alpha)} \mid X_{bd(\alpha)};$$

(G) **globale** se, per ogni terna (A, B, S) di sottinsiemi disgiunti di V tali che S separa A da B in G

$$X_A \perp\!\!\!\perp X_B \mid X_S.$$

Le proprietà appena descritte verificano la seguente relazione:

$$(G) \Rightarrow (L) \Rightarrow (P)$$

e dipendono dalle proprietà A.1 - A.4 enunuciate per l'indipendenza condizionata. Se si verifica anche che $f(x) > 0$, quindi se risulta valida anche la proprietà A.5, allora

$$(P) \Rightarrow (G)$$

e quindi

$$(G) \Leftrightarrow (L) \Leftrightarrow (P)$$

(Pearl e Paz 1987).

Come il concetto di indipendenza condizionata è legato alla fattorizzazione, così lo sono anche le proprietà di Markov.

Definiamo la proprietà (F) di **fattorizzazione** nel seguente modo:

Per ogni sottinsieme completo $a \subseteq V$, esiste una funzione non negativa ψ_a che dipende da x solo tramite x_a; la distribuzione di probabilità di X_V fattorizza secondo G, o soddisfa (F), se

$$f(x) = \prod_{a\, completo} \psi_a(x)$$

Essendo le cliques sottografi massimali completi, possiamo equivalentemente scrivere:

$$f(x) = \prod_{c \in C} \psi_c(x)$$

indicando con C l'insieme delle cliques di G.

Siccome $(F) \Rightarrow (G)$, se vale la proprietà B.5 o se $f > 0$, abbiamo che $(P) \Rightarrow (F)$ e quindi:

$$(F) \Leftrightarrow (G) \Leftrightarrow (L) \Leftrightarrow (P)$$

Vale che, se la distribuzione P soddisfa la proprietà di fattorizzazione (F) rispetto a G e (A,B,S) è una decomposizione di G, anche $P_{A \cup S}$ e $P_{B \cup S}$ soddisfano (F) rispetto a $G_{A \cup S}$ e $G_{B \cup S}$; possiamo, quindi, scrivere che

$$f(x) = \frac{f_{A \cup S}(x) f_{B \cup S}(x)}{f_S(x)}$$

in termini di cliques, per un grafo decomponibile, si ha:

$$f(x) = \frac{\prod_{c \in C} f_c(x)}{\prod_{s \in S} f_s(x)}$$

indicando con S l'insieme dei separatori delle cliques del grafo G.

1.6 Proprietà di Markov per Grafi Diretti Aciclici

Senza entrare nel merito e nel dibattito relativo al concetto di *causalità*, è però possibile affermare che, in molti casi, il ruolo assunto dalle variabili in un modello è tutt'altro che simmetrico; piuttosto, esiste un criterio di ordine

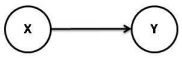


Figura 1.5: Esempio di relazione diretta

parziale o, comunque, è possibile stabilire che una variabile X sia causa di un'altra variabile Y (ma non il viceversa). I grafi diretti aciclici consentono di rappresentare mediante diagramma tale relazione, come mostra la Figura 1.5. Come anticipato nei paragrafi precedenti, nei DAG non è possibile avere cicli diretti, ma è possibile stabilire che i vertici del grafo siano tutti ordinati. Questo significa che ogni arco orientato nel grafo può avere una sola direzione e che i diretti antecedenti del nodo α costituiscono i $pa(\alpha)$. L'ordinamento *a priori* dei vertici consente, a ciascun nodo, di avere ben definito il proprio *passato*, *presente* e *futuro*.
La definizione di DAG è la medesima di quella dei grafi indiretti, l'unica differenza è che, mentre nei secondi, due variabili sono indipendenti condizionatamente al *resto* delle altre variabili del modello, nei primi si condiziona alle variabili che rappresentano il *passato*.
Diremo che una distribuzione di probabilità P su X_V **fattorizza** rispetto ad un DAG D se la sua funzione di probabilità, o densità f, assume la forma

$$f(x) = \prod_{v \in V} f_v(x_v \mid x_{pa(v)})$$

Si pensi, ad esempio, alla catena di Markov rappresentata in Figura 1.6, con $X_{i+1} \perp\!\!\!\perp (X_1, \ldots, X_{i-1}) \mid X_i$, per $i = 3, \ldots, n$

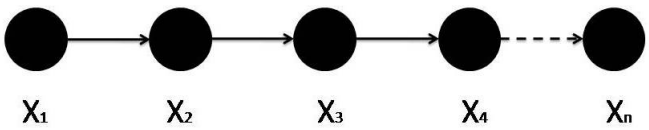


Figura 1.6: Esempio di catena di Markov

Consideriamo, ad esempio, il grafo rappresentato in Figura 1.7; esso ha come fattorizzazione:

$$f(x) = f(x_1)f(x_2)f(x_3 \mid x_2)f(x_4 \mid x_1, x_2, x_3)f(x_5 \mid x_4)f(x_6 \mid x_3, x_5)$$

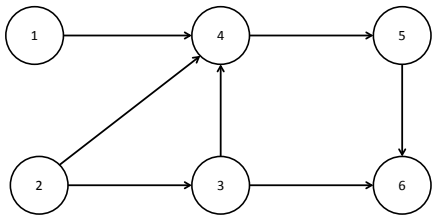


Figura 1.7: Esempio di fattorizzazione

Diremo, invece, che una distribuzione P soddisfa la proprietà **locale** di Markov rispetto ad un DAG D se, $\forall\ \alpha \in V$, vale che

$$X_\alpha \perp\!\!\!\perp X_{nd(\alpha)\setminus pa(\alpha)} \mid X_{pa(\alpha)}$$

Diremo, in ultimo, che una distribuzione P soddisfa la proprietà **globale** di Markov rispetto ad un DAG D se:

$$A \perp_D B \mid S \Rightarrow\ X_A \perp\!\!\!\perp X_B \mid X_S$$

Il simbolo $\perp_D$ indica la ***d-separazione***, uno dei due criteri per verificare l'indipendenza condizionata, introdotto da Pearl (1986a,b).
È sempre vero che $(F) \Leftrightarrow (G) \Leftrightarrow (L)$.

Criterio di *d - separazione*

Un cammino π tra i vertici α e β è *bloccato* da S se contiene un vertice $\gamma \in \pi$ tale che:

- $\gamma \in S$ e non ci sono configurazioni a V, ossia $\rightarrow \gamma \leftarrow$;
- Nè γ nè i suoi discendenti sono in S e ci sono configurazioni a V su γ.

In tutti gli altri casi i cammini si dicono *attivi*. Quindi, due sottinsiemi di vertici A e B sono *d-separati* da S se tutti i cammini da A a B sono bloccati da S.
In termini di variabili causali avremo che X_α è indipendente da X_β condizionatamente ad un insieme di variabili X_S, con $(\alpha, \beta \notin S)$, se tutti i cammini

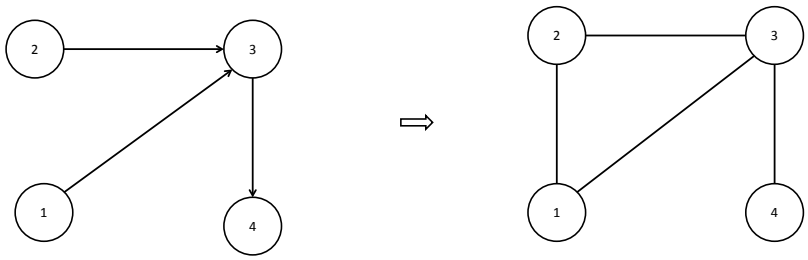


Figura 1.8: Esempio di moralizzazione di un DAG

tra X_α e X_β sono bloccati da X_S. Le due variabili sono, cioè, *d-separate* condizionatamente a X_S.

Criterio alternativo alla *d - separazione*

Il secondo criterio, equivalente alla *d - separazione*, è stato introdotto da Lauritzen *et al.*(1990) e utilizza il grafo morale D^m di un DAG D.
D^m si costruisce, a partire da D, unendo con un arco (sposando) le coppie di nodi che hanno figli in comune e, successivamente, sostituendo le frecce con archi non orientati, come rappresentato in Figura 1.8. Si chiama grafo morale poiché "sposa" i nodi genitori; tale terminologia è dovuta a Lauritzen e Spiegelhalter (1988).
Inoltre, se la distribuzione P fattorizza rispetto a D, fattorizza anche rispetto al suo grafo morale; vale anche che, se P soddisfa tutte le proprietà di Markov, le soddisfa anche D^m. Partendo da questi presupposti, come dimostrato da Lauritzen (1996), è possibile definire un criterio alternativo per lo studio dell'indipendenza condizionata sulla base dei seguenti passi:

- A partire da D, otteniamo il sottografo $D_{An(A \cup B \cup S)}$, ossia quello indotto dall'insieme ancestrale di $A \cup B \cup S$;
- Otteniamo il suo grafo morale, ossia $(D_{An(A \cup B \cup S)})^m$.

Diremo che $A \perp_D B \mid S \Leftrightarrow S$ separa A da B nel grafo indiretto (moralizzato), ottenuto dalla procedura appena indicata.

Equivalenze di Markov. Due DAG sono Markov equivalenti quando verificano le medesime relazioni di indipendenza condizionata. Questo accade quando hanno le stesse configurazioni a V e, eliminate le direzioni alle frecce, essi hanno gli stessi archi (si dice hanno lo stesso scheletro). Se vogliamo verificare le equivalenze di Markov tra un grafo diretto ed uno indiretto vale che il grafo indiretto G_u è Markov equivalente al DAG G_D se quest'ultimo non contiene configurazioni a V (quindi non presuppone l'aggiunta di alcun arco); è ancora equivalente a qualche DAG G_d se e solo se è triangolato. Si ha, invece, che un DAG G_D è equivalente ad un qualche grafo (morale) indiretto G_u se e solo se non contiene configurazioni a V.

1.7 Modelli Grafici Bayesiani

I princìpi dell'Inferenza Bayesiana. La Statistica Bayesiana distingue, in maniera abbastanza netta, tra quantità osservate (i dati) e quantità non osservate (parametri, dati mancanti, variabili latenti). I parametri vengono trattati, a differenza che nell'approccio frequentista, come variabili casuali per cui diventa necessario attribuire loro una distribuzione di probabilità, detta *a priori*. Indichiamo con $X = x$ i dati osservati e con θ il parametro del modello, si ha che la verosimiglianza

$$L(\theta, x) \propto p(x \mid \theta).$$

La distribuzione *a priori* su θ è $p(\theta)$ e riflette il nostro grado di incertezza sul parametro incognito, prima di osservare i dati. Per fare inferenza su θ, possiamo sfruttare il fatto che, essendo x noto, è possibile ottenere la distribuzione *a posteriori*, condizionando ai dati osservati, mediante il teorema di Bayes:

$$p(\theta \mid x) = \frac{p(\theta)p(x \mid \theta)}{\int p(\theta)p(x \mid \theta)d\theta} \propto p(\theta)p(x \mid \theta)$$

ossia:
Probabilità a posteriori $\propto$ Probabilità a priori $\times$ verosimiglianza.
La distribuzione *a posteriori* esprime il nostro grado di incertezza su θ, dopo aver osservato i dati.
In moltissimi casi, $P(\theta \mid x)$ risulta molto complessa e non in forma chiusa, quindi non derivabile analiticamente. Siccome uno degli scopi dell'inferenza bayesiana è ottenere la distribuzione marginale *a posteriori* di θ_i, ossia:

$$p(\theta_i \mid x) = \int\int \dots \int p(\theta \mid x) d\theta_{\setminus i} \qquad con \quad \theta_{\setminus i} = (\theta_1, \dots, \theta_{i-1}, \dots, \theta_{i+1}, \dots \theta_p)$$

l'integrazione numerica, e quindi i metodi **Markov Chain Monte Carlo** (Gilks, Richardson e Spiegelhalter 1996), diventano uno strumento essenziale. Gli integrali vengono valutati mediante una simulazione Monte Carlo da una catena di Markov costruita in maniera tale che la sua distribuzione stazionaria sia l'*a posteriori* cercata. Tra gli algoritmi sviluppati a tale fine, il **Gibbs sampling** (Gelfand e Smith 1990; Geman e Geman 1984) risulta particolarmente utile quando si vogliono esplorare le relazioni di indipendenza condizionata. Prima di approfondire il legame tra Gibbs Sampler e DAG, è opportuno definire alcune caratteristiche dei modelli grafici bayesiani, sviluppati mediante il software Openbugs[2](o Winbugs\Bugs).

1.7.1 I DAG in Openbugs: caratteristiche della Rappresentazione Grafica

Come anticipato nel paragrafo precedente, l'analisi bayesiana presuppone che sia le variabili osservate che i parametri siano variabili casuali, per cui un modello grafico bayesiano utilizza i nodi sia per la rappresentazione dei dati osservati che per i parametri del modello.
Il *doodle* è lo strumento di Openbugs che permette la rappresentazione grafica del modello.
In particolare, utilizza i seguenti tipi di nodi:

- **Costanti** per rappresentare valori fissati o assegnati nei dati. Non possono avere nodi genitori e sono rappresentati mediante un rettangolo;
- **Stocastici** per rappresentare variabili casuali, sia osservate che non osservate, a cui vanno assegnate distribuzioni di probabilità. Sono rappresentati tramite ovali;
- **Deterministici** per rappresentare funzioni logiche di altri vertici e quindi, a seconda della specificazione, possono essere costanti, osservati o non osservati. Anch'essi sono rappresentati tramite ovali.

Le relazioni di dipendenza sono rappresentate mediante due tipologie di archi orientati:

- **freccia singola** ($\rightarrow$) per rappresentare dipendenze stocastiche;
- **freccia doppia** ($\Rightarrow$) per rappresentare dipendenze funzionali (logiche).

[2]OpenBUGS è la versione Open Source di WinBUGS (Bayesian inference Using Gibbs Sampling)

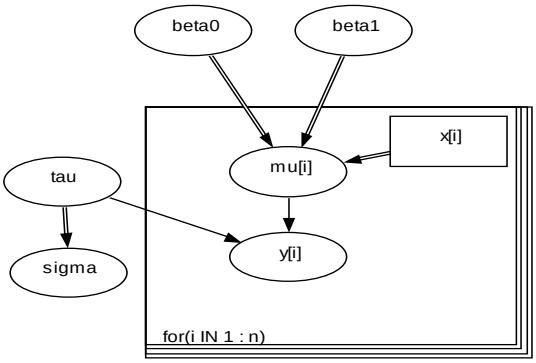


Figura 1.9: DAG per il modello di regressione $y_i \sim N(\mu_i, \tau)$

Se la struttura di alcuni o tutti i vertici è ripetuta, per esempio se si hanno dei cicli del tipo $i = 1, \ldots n$, si racchiudono i nodi in questione nei cosiddetti *plates*, indicizzati per i, come in Figura 1.9. In essa si rappresenta graficamente un modello di regressione lineare del tipo:

$$y_i \sim N(\mu_i, \tau)$$

con $i = 1, \ldots, n$ e $\mu_i = \beta_0 + \beta_1 x_i$.
τ è il parametro di precisione, ossia l'inverso della varianza, per cui $\tau = 1/\sigma^2$.
Sia i nodi che i *plates* possono essere annidati se si hanno strutture dei dati di tipo gerarchico.
Essendo interessati principalmente alle relazioni di tipo stocastico tra i parametri incogniti del modello e i dati, si ha che, una volta individuato tale legame, le dipendenze logiche vengono collassate e i nodi costanti ignorati.
I termini *nodo genitore* e *nodo figlio* fanno riferimento, quindi, alle quantità stocastiche.
A partire dal DAG è possibile ottenere il corrispondente codice del modello nel linguaggio Bugs. I vertici del grafo devono apparire sempre alla sinistra dell'equazione che ne specifica la relazione con i propri nodi genitori.
Se si tratta di una dipendenza stocastica, verrà riprodotto il simbolo "$\sim$", se si tratta di una dipendenza deterministica si avrà "$\leftarrow$"[3].
I *plates*, rappresentati graficamente, vengono riprodotti nel codice attraverso i cicli.

[3]Tali simboli non vanno confusi con quelli analoghi, descritti nel paragrafo 1.2, utilizzati per indicare la tipologia di arco, indiretto o diretto.

1.7.2 DAG e Gibbs Sampling

È utile, a questo punto, mostrare il legame esistente tra il DAG e l'algoritmo del Gibbs Sampling; ossia, conviene mostrare come il modello grafico in questione possa essere utilizzato per promuovere computazioni efficienti per l'inferenza.
Come già evidenziato, l'utilizzo dei metodi MCMC consente di generare un campione da una distribuzione che comprende tutte le variabili aleatorie non osservate del modello, senza conoscere la forma analitica della loro distribuzione congiunta o marginale. Usando tali metodi non è necessario dover specificare le distribuzioni *a posteriori*, ma solo le distribuzioni a *priori* e la verosimiglianza.
I metodi MCMC non fanno altro che estendere i metodi di integrazione Monte Carlo consentendo di estrarre campioni *dipendenti* da una catena di Markov, la cui distribuzione stazionaria è quella desiderata, l'*a posteriori* appunto.
Il Gibbs sampling è un esempio di algoritmo MCMC.
Per $\theta \mid x$, esso si articola nei seguenti passi:

- Inizializza la catena con valori (non troppo a caso) $\theta_1^0, \theta_2^0, \ldots, \theta_J^0$
- Ad ogni iterazione:

  campiona θ_1^{t+1} da $p(\theta_1^{t+1} \mid \theta_2^t, \ldots \theta_J^t)$
  campiona θ_2^{t+1} da $p(\theta_2^{t+1} \mid \theta_1^{t+1}, \theta_3^t, \ldots \theta_J^t)$
  $\vdots$
  campiona θ_J^{t+1} da $p(\theta_J^{t+1} \mid \theta_1^{t+1}, \theta_2^{t+1}, \ldots \theta_{J-1}^{t+1})$
- Il secondo step va ripetuto N volte fino a convergenza, ottenendo un campione per ogni θ
- $\forall\theta$, per $N \to \infty$, $\dfrac{1}{N}\Sigma_t h(\theta^{(t)}) \approx E(h(\theta \mid x))$

Le distribuzioni condizionate, definite nel secondo step del Gibbs sampling, sono dette *full conditionals* poiché condizionano rispetto a tutti gli altri parametri.
È proprio in relazione a ciò che è possibile stabilire un legame tra i modelli grafici e l'algoritmo in questione; in particolare, la rappresentazione grafica consente l'individuazione delle variabili da considerare per la specificazione della distribuzione *full conditional*.
Sulla base della proprietà locale di Markov per i DAG, è possibile esprimere

la distribuzione *full conditional* per un grafo diretto aciclico come:

$$p(v \mid V \setminus v) \propto p(v \mid pa(v)) \prod_{w \in ch(v)} p(w \mid pa(w)) \tag{1.7}$$

$$= p(v \mid bl(v)) \tag{1.8}$$

dove $bl(v) = pa(v) \cup ch(v) \cup \{\cup_{w \in ch(v)} pa(w) \setminus \{v\}\}$ è il *Markov blanket* del nodo v.
Va ricordato che Openbugs utilizza, oltre al Gibbs sampling, anche gli algoritmi Slice sampling (Neal 1997) e Metropolis-Hastings (Hastings 1970; Metropolis et al. 1953). In particolare il Gibbs sampling è un caso speciale del M-H. Quest'ultimo è appropriato per distribuzioni *full conditionals* complesse, ossia quando esse non hanno forma nota.
Il software è un *sistema esperto* per cui gestisce autonomamente la scelta dello schema di campionamento più appropriato, a seconda se la distribuzione *full conditional* del nodo è, o non è, nota in forma chiusa. Per lo studio e l'analisi della convergenza della serie ottenute si rimanda al capitolo 5.

Capitolo 2

Item Response Theory in ottica multilivello

2.1 Introduzione

I modelli IRT trovano larga applicazione in ambito psicometrico, la loro diffusione è direttamente proporzionale al crescente utilizzo, in particolare nelle scienze sociali, del questionario come strumento principale per la misurazione di uno o più costrutti latenti.
I problemi legati alla costruzione dei test, alla valutazione ed interpretazione dei risultati ottenuti, hanno incoraggiato lo sviluppo dell'Item Response Theory.
A partire dalla seconda metà del '900, questo approccio viene ampiamente utilizzato allo scopo di produrre misurazioni valide di stati soggettivi, quindi non direttamente osservabili, detti anche *costrutti latenti*; con essi si fa riferimento a caratteristiche quali l'intelligenza, o più in generale l'abilità, la soddisfazione del consumatore, la qualità della vita, solo per citare alcuni esempi.
I lavori di Lord e Novick (1968) e del matematico danese Rasch (1960) vengono citati, in letteratura, come punti di riferimento per questo genere di modelli; seguono poi altri importanti contributi ad opera di Andrich (1978), Masters e Wright (1997).
Per quanto concerne l'approccio bayesiano, i primi sviluppi risalgono agli anni Ottanta; a riguardo sono da citare i lavori di Mislevy (1986), Rigdon e Tsutakawa (1983), Swaminathan e Gifford (1982, 1985).
L'avvento di nuovi software ha favorito la diffusione delle tecniche MCMC per l'Item Response Theory, come dimostrato dai numerosi lavori prodotti negli ultimi decenni (Beguin e Glas (2001); Bolt e Lall (2003); Bradlow, Wai-

ner e Wang (1999); Fox e Glas (2001); Johnson e Sinharay (2005); Patz e Junker (1999a,b); Fox (2010)).

2.2 Modelli IRT

Come anticipato nell'introduzione a questo capitolo, i modelli IRT sono modelli probabilistici che fondano la loro idea sulla possibilità di misurare costrutti latenti, i quali non possono essere direttamente osservati ma possono essere misurati, indirettamente, attraverso variabili manifeste.
La soddisfazione rispetto ad un determinato servizio, ad esempio, può essere misurata solo indirettamente, tramite le risposte fornite alle domande sui diversi aspetti che la definiscono; gli items di un questionario, o meglio le risposte ad essi associate, possono essere visti come indicatori parziali della variabile latente.
I questionari rappresentano uno strumento efficace di valutazione del fenomeno solo in relazione ai suoi aspetti qualitativi, dal momento che la somma dei punteggi grezzi non ha alcun significato quantitativo. I valori numerici assegnati alle risposte costituiscono una scala qualitativa ordinale mentre l'obiettivo è quello di collocare i punteggi osservati su di una scala quantitativa.
A seguito di ciò diventa necessario individuare dei metodi di misura capaci di tradurre l'informazione, ottenuta a partire dalle risposte osservate, in una ***misura*** sintetica del tratto latente.
A differenza del conteggio, la misura è pur sempre un costrutto astratto per cui uno dei principali problemi è quello di riuscire a renderla *oggettiva*, così come accade per le misure sviluppate nelle scienze fisiche; essa deve essere tale da non dipendere dal contesto applicativo, invariante rispetto al campione dei rispondenti e all'insieme degli items.
Sulla base di quanto appena detto, affinché una misura possa definirsi tale è necessario che risponda ai princìpi fondamentali dell'**unidimensionalità** e della **specifica oggettività**. Il primo requisito impone che gli items siano indicatori del medesimo ed unico costrutto latente; il secondo afferma che, così come avviene nelle scienze naturali, i metodi utilizzati devono consentire di misurare caratteristiche specifiche del soggetto senza che il processo di misurazione risulti alterato da caratteristiche dell'individuo diverse da quella di interesse, o da altri individui o, ancora, da caratteristiche relative allo strumento (questionario) utilizzato.

2.2.1 Il modello di Rasch

Il modello introdotto da Rasch (1960), tra i più noti nella classe dei modelli IRT, rispetta i criteri appena citati in relazione al concetto di misura.
Diventa possibile tradurre i punteggi osservati in manifestazioni di un *continuum* latente: è possibile associare, ad ogni item e ad ogni soggetto, un numero reale riferito ad una stessa dimensione e secondo una scala comune.
Propriamente, il modello di Rasch si basa sulla trasformazione logistica delle probabilità associate alle possibili risposte fornite dai soggetti ai diversi items.
Presuppone che la probabilità di risposta corretta ad un item, da parte di un soggetto, sia funzione della differenza tra due parametri: θ_j $(j = 1, \ldots, J)$, l'*abilità* del soggetto, e β_i $(i = 1, \ldots, I)$, la *difficoltà* dell'item.
Facendo riferimento ai soli items dicotomici, il concetto di ordinamento in questa analisi va inteso nel senso di riportare lungo un *continuum* gli items, ordinati dal più facile al più difficile, e i soggetti, dal meno abile al più abile.

Supponendo di avere a disposizione la matrice $\mathbf{Y}$ contenente le risposte dei J soggetti agli I items, con elementi y_{ij} uguali a 0 o 1, il modello matematico che governa la probabilità di risposta soddisfa le seguenti assunzioni:

- **Unidimensionalità**.
  Tutti gli I items misurano il medesimo tratto latente $\boldsymbol{\theta}$;

- **Monotonicità**.
  Per ogni item i, la probabilità di risposta è una funzione del tratto latente e prende il nome di ICC (*Item Characteristic Curve*); essa si caratterizza per essere una funzione continua e monotòna, crescente per la modalità di risposta uguale ad 1.
  Ne segue che, ad ogni livello di abilità, corrisponde una certa probabilità che l'individuo dia una risposta corretta all'item; nel caso di un item tipico, tale probabilità sarà piccola per i soggetti con bassi livelli di abilità, crescerà man mano che si considerino individui con livelli di abilità elevati.

- **Indipendenza locale degli items**.
  Dato il livello di abilità θ_j per l'individuo $j-esimo$, le risposte $Y_{ij} = y_{ij}$ agli items sono tra loro indipendenti. Considerato il vettore di valori osservati $\mathbf{y_j} = (y_{1j}, y_{2j}, \ldots y_{Ij})$, l'indipendenza locale implica che:

$$P(\mathbf{y_j}|\theta_j) = P(y_{1j}|\theta_j)P(y_{2j}|\theta_j)\ldots P(y_{Ij}|\theta_j) = \prod_{i=1}^{I} P(y_{ij}|\theta_j) \qquad (2.1)$$

Risulta abbastanza chiaro che essa è equivalente alla definizione di **indipendenza condizionata**.

- **Assenza di Guessing**.
  Essa presuppone che quanto più il livello di abilità del soggetto tende a valori piccoli, tanto più la probabilità di risposta corretta all'item $i - esimo$ tende a zero; al contrario, quanto più il livello di abilità assume valori elevati, tanto più la probabilità di risposta corretta si avvicina a 1.

- **Sufficienza dei punteggi grezzi**.
  Le statistiche dei punteggi grezzi $R_j = \sum_{i=1}^{I} Y_{ij}$ e $S_i = \sum_{j=1}^{J} Y_{ij}$ sono statistiche sufficienti, rispettivamente, per θ_j e β_i.

Sulla base delle assunzioni appena enunciate, il modello di Rasch esprime la probabilità di risposta come:

$$P(Y_{ij} = y_{ij}|\theta_j, \beta_i) = \frac{exp[y_{ij}(\theta_j - \beta_i)]}{1 + exp(\theta_j - \beta_i)} \tag{2.2}$$

È agevole verificare che la probabilità di risposta corretta $y_{ij} = 1$ è pari a $1/2$ solo se $\theta_j = \beta_i$, cioè quando l'abilità del soggetto è uguale alla difficoltà della domanda.
Come già anticipato negli assunti del modello, al variare di θ, fissato β_i, si ottiene la curva logistica detta *Item Characteristic Curve* (ICC); essa è una funzione monotòna crescente rispetto a θ: all'aumentare del livello di abilità θ, aumenta la probabilità di rispondere correttamente all'item. Tale funzione tende a zero se $\theta \to -\infty$, tende a 1 se $\theta \to +\infty$. Maggiore è il valore di β_i, maggiore deve essere l'abilità del soggetto per riuscire a rispondere correttamente all'item.
Ogni item ha la sua ICC e, nel caso del modello di Rasch, esse sono tutte parallele tra loro poiché hanno tutte la stessa pendenza (si ipotizza che gli items abbiano la stessa capacità discriminatoria rispetto alla variabile latente), come mostrato in Figura 2.1.
L'equazione 2.2 implica che il logaritmo degli odd-ratios sia pari, proprio, alla differenza tra i parametri di abilità e di difficoltà:

$$\log \left[\frac{Pr(Y_{ij} = 1)}{Pr(Y_{ij} = 0)} \right] = \theta_j - \beta_i \tag{2.3}$$

La trasformazione in logit fa sì che i parametri del modello θ e β assumano valori reali.
Come già detto, le ipotesi alla base del modello garantiscono che le stime dei

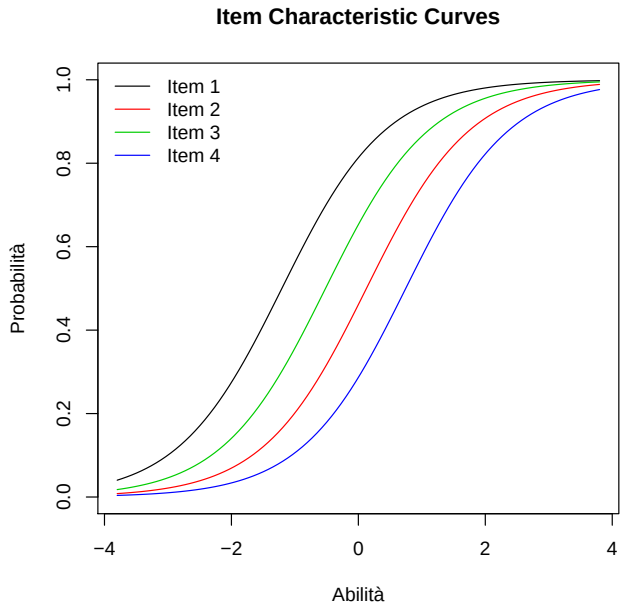


Figura 2.1: Item Characteristic Curves per il modello di Rasch con riferimento a quattro items

parametri rispettino i princìpi di unidimensionalità e specifica oggettività. Ciò che contraddistingue il modello di Rasch dagli altri modelli IRT è che esso è l'unico ad ammettere statistiche sufficienti: il punteggio totale delle risposte corrette da parte di un individuo è una statistica sufficiente per la stima della sua abilità latente, analogamente la somma delle risposte corrette di tutti gli individui su ciascun item è una statistica sufficiente per la stima del suo livello di difficoltà.
Ciò implica che, noto il punteggio complessivo del soggetto, che per definizione non dipende dalla difficoltà degli items, nessun'altra informazione sull'abilità dei soggetti è contenuta nei vettori delle risposte: dato il punteggio complessivo di ciascun individuo, la probabilità di ciascun pattern di risposta dipende solo dai parametri di difficoltà (Baker e Kim 2004; Wright e Masters 1982).
Di conseguenza, è vero pure che la probabilità, condizionata al punteggio complessivo di ciascun item, dipende dai soli parametri di abilità.
Tale proprietà è nota come *separabilità dei parametri* ed è grazie ad essa che può essere garantita la specifica oggettività (Gori, Sanarico e Plazzi 2005): la misura dei soggetti non dipende dal particolare insieme di items somministrati, quindi dallo strumento di misura impiegato, né la misura di ciascun

item dipende, a sua volta, dal campione di individui in esame. Come già detto, con il modello di Rasch, diventa possibile ordinare gli items sulla base del loro livello di difficoltà e gli individui sulla base delle loro abilità; tali graduatorie possono esse ovviamente confrontate tra loro.
Si può non solo stabilire se un item sia più o meno facile di un altro, oppure se un soggetto sia più o meno abile di un altro, ma è anche possibile prevedere la probabilità di risposta corretta ponendo a confronto abilità dell'individuo e difficoltà dell'item.
Se, per esempio, il logit della difficoltà dell'item è uguale al logit dell'abilità del soggetto, si può affermare che tale individuo ha il 50% di probabilità di rispondere correttamente al quesito.
Il modello di Rasch è anche detto **modello logistico ad 1 parametro** poiché prevede un solo parametro degli item.

2.2.2 I modelli IRT a due e tre parametri

Una seconda classe di modelli IRT, nota come **modello logistico a 2 parametri**, introduce, accanto al parametro di difficoltà, un parametro di discriminazione. Ciò implica che item diversi possano discriminare in maniera differente rispetto alla variabile latente di interesse. In caso di risposta corretta, la probabilità diventa:

$$P(Y_{ij} = 1|\theta_j, a_i, \beta_i) = \frac{exp(a_i\theta_j - \beta_i)}{1 + exp(a_i\theta_j - \beta_i)} \tag{2.4}$$

Come conseguenza si ha che ogni ICC ha un coefficiente angolare pari ad a_i, quindi, a differenza che nel modello di Rasch, le ICC possono intersecarsi come mostrato in Figura 2.2.
Items a cui corrispondono parametri a_i più alti, quindi, discriminano maggiormente gli individui con abilità bassa da quelli con abilità alta.
Una terza classe di modelli IRT, nota come **modello logistico a 3 parametri**, introduce il cosiddetto parametro di *guessing*, prevedendo la possibilità che un individuo, che ignori la risposta corretta, possa comunque rispondere esattamente affidandosi al caso. La probabilità di risposta corretta diventa:

$$P(Y_{ij} = 1|\theta_j, a_i, \beta_i, c_i) = c_i + (1 - c_i)\frac{exp(a_i\theta_j - \beta_i)}{1 + exp(a_i\theta_j - \beta_i)} \tag{2.5}$$

il parametro c_i indica, quindi, la probabilità di risposta corretta all'item i da parte di individui con un livello molto basso di abilità.

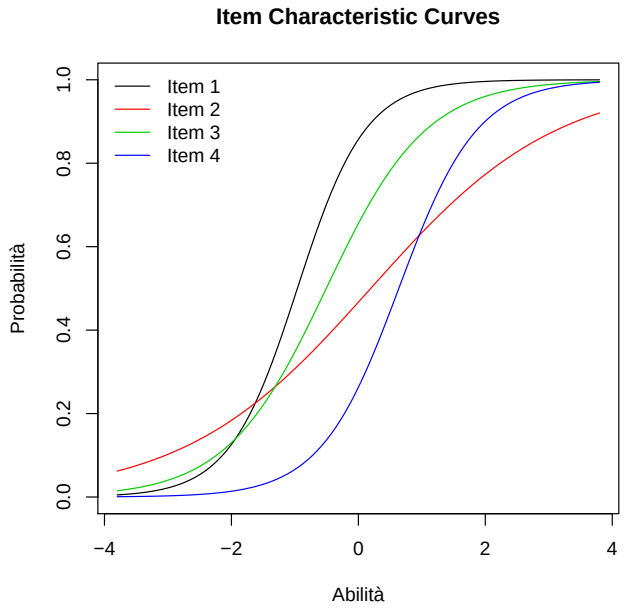


Figura 2.2: Item Characteristic Curves per il modello a 2 parametri con riferimento a quattro items

Considerare tale parametro comporta l'introduzione di un asintoto orizzontale diverso da zero per le ICC, particolare che lo rende diverso dai precedenti due modelli, in cui gli individui con bassissima abilità hanno probabilità nulla di rispondere correttamente all'item.

Va sottolineato che se da un lato l'introduzione di ulteriori parametri nel modello apporta un miglioramento in termini di flessibilità, dall'altro determina la rinuncia a caratteristiche quali la sufficienza dei punteggi grezzi e, quindi, il venir meno del princìpio fondamentale della specifica oggettività della misura.

2.2.3 I modelli per items politomici

I modelli sin qui trattati fanno riferimento ad items dicotomici per cui, in caso di più di due categorie di risposta su scala ordinale, è necessario ricorrere ad un'estensione politomica del modello.
Le due possibili estensioni per il modello di Rasch sono il ***Rating Scale Model*** (RSM) (Andrich 1978) e il ***Partial Credit Model*** (PCM) (Masters 1982).

I modelli in questione considerano, nella loro funzione di probabilità, i cosiddetti *parametri di soglia*, che rappresentano una misura della difficoltà di accordare la preferenza alle diverse possibili categorie di risposta associate a ciascun item.
Il primo modello può essere applicato quando, per tutti gli items, il numero delle categorie di risposta è uguale e le "soglie" sono vincolate ad essere identiche; il secondo modello ne rappresenta una generalizzazione poiché i vincoli citati possono non sussistere. Le soglie possono non essere equidistanti, favorendo la possibilità che il passaggio da una categoria di risposta a quella adiacente non sia costante. L'ordinamento delle soglie è invece importante: è opportuno che la difficoltà di ciascuna soglia sia maggiore di quelle che la precedono, altrimenti si avrebbe qualche categoria ridondante.
In riferimento alla formulazione più generale, e quindi al *Partial Credit Model*, se si organizzano le categorie in maniera tale che $0 < 1 < 2 < \ldots < m_i$, la probabilità di ottenere il punteggio y, anziché $y - 1$, cresce in maniera monotòna all'aumentare del livello di abilità e il PCM si esprime come:

$$\frac{P_{ijy}}{P_{ijy-1} + P_{ijy}} = \frac{exp(\theta_j - \beta_{iy})}{1 + exp(\theta_j - \beta_{iy})} \tag{2.6}$$

con $y = 1, 2, \ldots, m_i$.
In particolare:

1. P_{ijy} rappresenta la probabilità che il soggetto j risponda y all'item i;
2. P_{ijy-1} rappresenta la probabilità che il soggetto j risponda $y-1$ all'item i;
3. θ_j è il parametro di abilità del soggetto j;
4. β_{iy} è il parametro di difficoltà dell'item i.

θ_j e β_{iy} governano la funzione di probabilità di rispondere y, anziché $y - 1$, per l'item i.
Riscrivendo il modello come la probabilità non condizionata di ogni possibile outcome $0, 1, \ldots, m_i$ del soggetto j rispetto all'item i, il PCM diventa:

$$P(Y_{ij} = y|\theta_j, \beta_{ik}) = \frac{exp[\sum_{k=0}^{y}(\theta_j - \beta_{ik})]}{\sum_{h=0}^{m_i} exp \sum_{k=0}^{h}(\theta_j - \beta_{ik})} \tag{2.7}$$

con $y = 0, 1, \ldots, h, \ldots, m_i$.

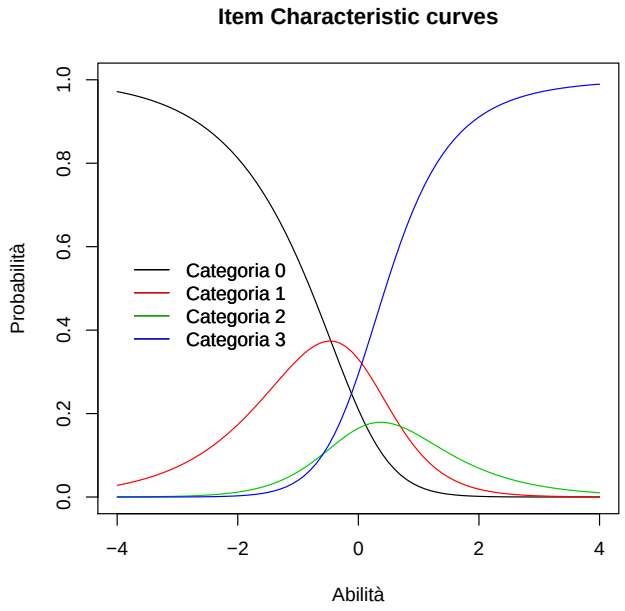


Figura 2.3: Item Characteristic Curves per il modello PCM con riferimento alle categorie di un singolo item

Per convenzione:

$\sum_{k=0}^{0}(\theta_j - \beta_{ik}) \equiv 0$ e $\sum_{k=0}^{h}(\theta_j - \beta_{ik}) \equiv \sum_{k=1}^{h}(\theta_j - \beta_{ik})$.

Un esempio di ICC, tipica per il PCM, è rappresentato in Figura 2.3. Per items con sole due categorie di risposta è agevole verificare che il PCM diventa il modello di Rasch.

2.2.4 Stima dei parametri: approccio bayesiano

Come anticipato nel paragrafo 1.7 del precedente capitolo, nell'approccio bayesiano, i parametri del modello sono variabili casuali a cui è associata una distribuzione di probabilità *a priori*, la quale riflette il grado di incertezza sul valore vero del parametro, prima di osservare i dati.
Nei modelli IRT, in particolare, l'obiettivo è la stima dei parametri di difficoltà e di abilità, per cui oggetto dell'inferenza diventa la loro distribuzione *a posteriori*.
Volendo costruire un modello bayesiano psicometrico, sia $\boldsymbol{\theta}$ il parametro latente dell'abilità con probabilità *a priori* $p(\boldsymbol{\theta})$ e sia $\boldsymbol{\omega}$ l'insieme dei parametri

che definiscono gli items con probabilità *a priori* $p(\boldsymbol{\omega})$, sulla base del teorema di Bayes, la densità congiunta *a posteriori* è definita come:

$$p(\boldsymbol{\theta}, \boldsymbol{\omega}|\mathbf{y}) = \frac{p(\mathbf{y}|(\boldsymbol{\theta}, \boldsymbol{\omega})p(\boldsymbol{\theta}, \boldsymbol{\omega})}{p(\mathbf{y})} \tag{2.8}$$

Dal momento che le densità *a priori* sono considerate indipendenti tra loro, possiamo scrivere:

$$p(\boldsymbol{\theta}, \boldsymbol{\omega}|\mathbf{y}) = \frac{p(\mathbf{y}|(\boldsymbol{\theta}, \boldsymbol{\omega})p(\boldsymbol{\theta})p(\boldsymbol{\omega})}{p(\mathbf{y})} \tag{2.9}$$

Ne segue che, se siamo interessati alla stima di $\boldsymbol{\theta}$, i parametri di riferimento degli items diventano parametri di disturbo e quindi:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \int p(\boldsymbol{\theta}, \boldsymbol{\omega}|\mathbf{y}) d\boldsymbol{\omega} = \int \frac{p(\mathbf{y}|(\boldsymbol{\theta}, \boldsymbol{\omega})p(\boldsymbol{\theta})p(\boldsymbol{\omega})}{p(\mathbf{y})} d\boldsymbol{\omega} \tag{2.10}$$

Allo stesso modo, se siamo interessati alla stima dei parametri riferiti agli items, integreremo rispetto ai parametri di abilità.
L'integrale, come vedremo nei paragrafi successivi, si complica nel momento in cui si considerano modelli bayesiani gerarchici.
Come è stato già messo in evidenza nel capitolo precedente, la risoluzione di questi integrali avviene attraverso l'utilizzo delle metodologie MCMC.
Una differenza sostanziale tra i metodi MCMC e altri metodi di stima noti in letteratura (Massima Verosimiglianza, ad esempio) consiste nel fatto che essi stimano una distribuzione di probabilità per il parametro di interesse, piuttosto che un valore puntuale.
Trattandosi di modelli bayesiani, è bene fare alcune precisazioni in relazione alla scelta delle distribuzioni *a priori* da utilizzare per i parametri del modello. Si suole distinguere tra distribuzioni *a priori coniugate* e *non coniugate*.
Una distribuzione *a priori coniugata* ha la caratteristica di restituire distribuzioni *a posteriori* della stessa famiglia delle distribuzioni *a priori*; ciò rende più agevole la fase di campionamento dell'MCMC in quanto diventa nota la forma funzionale della distribuzione *a posteriori*.
Ovviamente è da precisare che l'esistenza di distribuzioni *a priori coniugate* dipende dal tipo di modello scelto.
Per i motivi appena citati, è preferibile selezionare, quindi, distribuzioni *a priori coniugate*.
L'influenza dell'*a priori* può essere controllata attraverso la scelta dei parametri che la specificano e che prendono il nome di *hyperparameters*. Essi

non devono necessariamente assumere un valore fissato, si può attribuire loro una distribuzione *a priori* che prende il nome di *hyperprior*. Ciò conduce, ovviamente, a definire modelli di tipo gerarchico.
Infine, per fare in modo che siano i dati a fornire la massima informazione possibile, è necessario far ricorso a distribuzioni *a priori* cosiddette *non informative*, caratterizzate da varianza elevata.
Si precisa che tali criteri di scelta delle distribuzioni sono gli stessi utilizzati per la definizione dei modelli oggetto del presente studio.

2.3 Modelli multilivello (o gerarchici)

Nel contesto educativo, è assai frequente che i dati a disposizione abbiano una struttura gerarchica, ciò significa che gli individui non sono campionati in maniera indipendente e casuale dalla popolazione di interesse. Piuttosto, in molte situazioni, i soggetti sono annidati in gruppi e la popolazione di interesse è formata, quindi, da sottopopolazioni.
Con riferimento all'Item Response Theory è frequente supporre che gli items siano annidati (*nested*) negli individui che, a loro volta, possono essere annidati in livelli gerarchici superiori (scuole, nazioni etc..): la diretta conseguenza è che osservazioni appartenenti al medesimo *cluster* sono, quindi, correlate tra loro.
L'esistenza di tali livelli di gerarchia porta a complesse strutture di dipendenza, con diverse fonti di variabilità che fanno riferimento a ciascun livello di aggregazione.
Questo implica che il modello statistico di riferimento non può prescindere da tale caratteristica intrinseca ai dati e che l'inferenza deve essere, essa stessa, condotta a differenti livelli di aggregazione.
Le risposte agli items, che fanno quindi riferimento al livello 1 della gerarchia, il livello *within respondent*, sono modellate attraverso la verosimiglianza condizionata, assumendo l'ipotesi di indipendenza condizionata, fissato il parametro di abilità del soggetto. Ai livelli più alti, il modello *between respondent* modella l'eterogeneità tra i soggetti.
Come si avrà modo di vedere in seguito, per dati gerarchici, l'approccio bayesiano è particolarmente appropriato; la sua flessibilità e la sua capacità di gestire problemi anche molto complessi, con strutture di dipendenza come quelle appena descritte, lo rendono particolarmente attraente.

2.3.1 Modelli IRT multilivello

Da quanto descritto nell'introduzione ai modelli gerarchici, risulta chiaro come sia stato necessario, soprattutto in ambito psicometrico, combinare le proprietà e gli assunti di base dei modelli IRT con quelli propri dei modelli multilivello, dal momento che ciò comporta una serie di vantaggi: primo fra tutti, consente di analizzare i dati in "cluster" tenendo conto di entrambe le fonti di variabilità, sia quella *within* che *between*.
Inoltre, consente di ottenere le stime dei tratti latenti con riferimento ai diversi livelli di aggregazione, oltre ad offrire la possibilità di inserire nel modello covariate ed effetti di interazione.
Come sarà mostrato in seguito, il modello IRT multilivello ha il vantaggio di restituire stime più accurate della relazione esistente tra predittori e tratti latenti poiché effettua, simultaneamente, non solo la stima dei parametri del modello IRT, ma anche le stime relative ai coefficienti associati ai predittori.
E' da sottolineare che l'utilizzo dei modelli IRT in ottica multilivello è piuttosto recente; in seno a tale filone, i ricercatori hanno verificato che i modelli IRT possono essere interpretati nell'ottica dei **modelli lineari generalizzati gerarchici** (Raudenbush e Bryk 2002) o come **modelli misti generalizzati**, lineari o non lineari (Rijmen et al. 2003).
Uno dei contributi più significativi, in relazione ai modelli analizzati, è da attribuire a Kamata (2001, 2002); nei suoi lavori ha dimostrato come il modello di Rasch possa essere rappresentato mediante un modello lineare generalizzato a due livelli estendendo, successivamente, il risultato al caso di tre livelli gerarchici.
Particolarmente interessante per la comprensione dei modelli multilivello risulta essere l'articolo di Kamata e Cheong (2007); in esso gli autori effettuano una chiara disamina delle diverse prospettive di interpretazione del modello di Rasch in ottica multilivello.
La prima di esse fa riferimento al fatto che le abilità dei soggetti possono essere viste come effetti *random* (Adams, Wilson e Wu 1997; Hedeker e Gibbons 1993; Spiegelhalter et al. 1996.); uno degli obiettivi legati a questo approccio è quello di facilitare la stima di massima verosimiglianza marginale dei parametri di difficoltà, evitando il cosiddetto "*Neyman - Scott problem*", ossia l'inconsistenza degli stimatori quando sia i parametri degli items che quelli di abilità vengono stimati simultaneamente.
Tale approccio conduce in maniera naturale a considerare i modelli IRT come modelli a effetti misti: i parametri di difficoltà sono gli effetti fissi mentre i parametri di abilità rappresentano gli effetti casuali.
La seconda prospettiva fa riferimento alla possibilità di rappresentare l'effetto random come combinazione lineare di effetti fissi e di effetti casuali. A soste-

gno di ciò, Adams e Wilson (1996) hanno proposto un modello più generale la cui sigla **RCMLM** è l'acronimo di *Random Coefficient Multinomial Logit Model* e la cui generalizzazione al caso multidimensionale (**MRCMLM**) è contenuta nell'articolo dell'anno successivo (Adams, Wilson e Wang 1997). L'importanza di questo modello consiste nel fatto che, non solo si considerano i parametri delle abilità dei soggetti come effetti random, ma si rende possibile includere nel modello, in qualità di predittori, anche caratteristiche proprie dei soggetti.
Questa modellizzazione è, però, applicabile solo a dati con due livelli gerarchici e può includere solo effetti random al secondo livello.
Una terza prospettiva è rappresentata dai modelli IRT multi-gruppo, in cui si assume che gli individui siano raggruppati sulla base di caratteristiche comuni quali, per esempio, il gruppo etnico o la scuola di appartenenza (Bock e Zimowski 1997.).
Gli autori definiscono anche una quarta prospettiva che fa riferimento alla possibilità di scomporre gli *item parameters* in diverse componenti, come avviene nel *Linear Logistic Test Model* (Fischer 1995).
A conclusione di questa breve disamina, merita di essere menzionato il contributo di Cheong e Raudenbush (2000); gli autori valutano il modello di Rasch, non solo in ottica multilivello, ma anche nella sua versione estesa al caso multidimensionale.
In ottica bayesiana merita particolare cenno il contributo di Fox (2004) (vedi anche Fox e Glas (2001, 2003)), il quale definisce, come vedremo in seguito, un modello IRT la cui variabile latente è una variabile dipendente in un modello di regressione multilivello.

2.3.2 Esempi di modellizzazione IRT gerarchica

Essendo l'approccio proposto da Kamata (2001) tra i più noti e tra i più applicati nel contesto di riferimento, è bene passare ad una sua descrizione dettagliata.
Prendendo a riferimento il modello di Rasch dell'equazione 2.3, si dimostra che esso è interpretabile come un modello a tre livelli con intercetta casuale. Tale modello considera come unità di primo livello le risposte agli items che, a loro volta, sono annidate negli individui (unità di secondo livello) i quali sono poi annidati in gruppi (unità di terzo livello), come rappresentato in Figura 2.4.
Il primo livello del modello *unconditional* assume che il logit della probabilità che l'individuo $j - esimo$, con $j = 1, \ldots, J$, appartenente al gruppo g, con $g = 1, \ldots, G$, risponda correttamente all'item i, con $i = 1, \ldots, I - 1$, sia

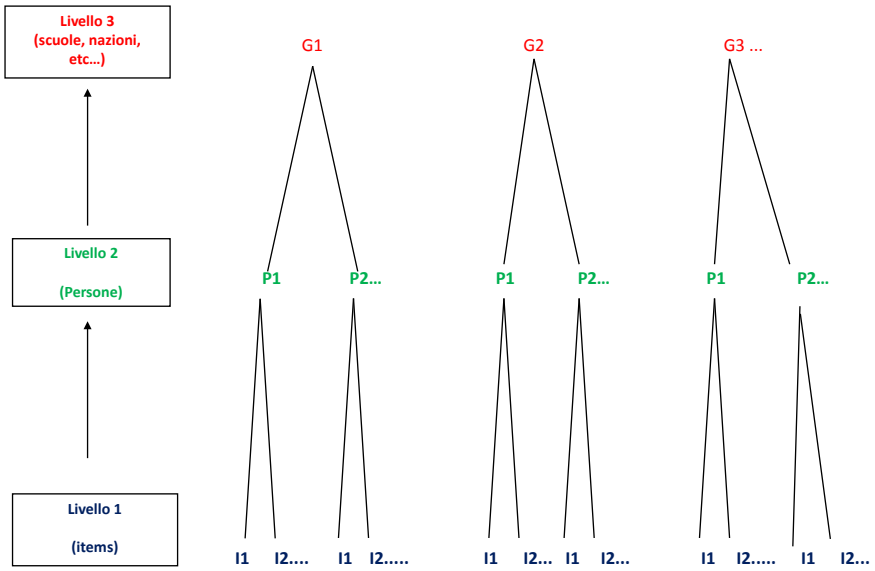


Figura 2.4: Esempio di struttura gerarchica a tre livelli.

esprimibile come:

$$log\left(\frac{p_{ijg}}{1-p_{ijg}}\right) = \beta_{0jg} + \beta_{1jg}X_{1jg} + \beta_{2jg}X_{2jg} + \ldots + \beta_{(I-1)jg}X_{(I-1)jg} \quad (2.11)$$

$$= \beta_{0jg} + \sum_{q=1}^{I-1} \beta_{qjg}X_{qjg} \quad (2.12)$$

X_{qjg} $(q = 1, \ldots, I-1)$ indica la $q - esima$ variabile dummy per lo studente $j - esimo$, appartenente al gruppo g, tale che:

$$X_{qjg} = \begin{cases} 1 & \text{se } q = i, \\ 0 & \text{altrimenti.} \end{cases}$$

β_{0jg} rappresenta il termine d'intercetta, ossia l'effetto dell'item di riferimento, l'$I - esimo$ quindi, mentre β_{qjg} rappresenta l'effetto del $q - esimo$ item rispetto all'item di riferimento.
In genere si sceglie come item di riferimento quello che, dall'analisi IRT classica, risulta essere l'item più semplice. E' da notare che viene esclusa in tale modello la dummy riferita all'$I - esimo$ item, a causa della presenza dell'intercetta; è comunque possibile riparametrizzare il modello non includendo il

termine di intercetta e inserendo le dummies per ogni item, come mostrano Rijmen et al. (2003).
Il modello, al secondo livello, considera la variabilità degli individui all'interno dei gruppi, ragion per cui i coefficienti $\beta_{1jg}, \ldots, \beta_{(I-1)jg}$ diventano effetti fissi ottenendo che:

$$\begin{cases} \beta_{0jg} = \gamma_{00g} + u_{0jg} \\ \beta_{1jg} = \gamma_{10g} \\ \vdots \\ \beta_{(k-1)jg} = \gamma_{(I-1)0g} \end{cases}$$

u_{0jg} rappresenta la componente random dell'intercetta, si distribuisce come una $N(0, \tau_\beta)$ e indica propriamente la deviazione dell'individuo j dalla media del gruppo g.
Ovviamente, in questo caso, l'effetto random è una variabile latente.
Il modello al terzo livello, tenendo in considerazione la variabilità tra i gruppi, assume la forma:

$$\begin{cases} \gamma_{00g} = \pi_{000} + r_{00g} \\ \gamma_{10g} = \pi_{100} \\ \vdots \\ \gamma_{(I-1)0g} = \pi_{(I-1)00} \end{cases}$$

In questo caso l'effetto random $r_{00g} \sim N(0, \tau_\gamma)$.
π_{000} può essere visto come l'effetto dell'item di riferimento nel campione.
E' abbastanza chiaro che, anche in relazione al terzo livello, gli item sono considerati effetti fissi sia tra i gruppi che tra gli individui cosi da ottenere, per tutti i gruppi e tutti gli individui, le stesse stime dei parametri di difficoltà.
E' agevole verificare che, combinando i livelli, il modello che ne risulta è analogo al modello di Rasch; in particolare per uno specifico individuo j appartenente ad uno specifico gruppo g, con riferimento all'item i (dove $i = q$), si ha:

$$log\left(\frac{p_{ijg}}{1 - p_{ijg}}\right) = \pi_{000} + \pi_{q00} + u_{0jg} + r_{00g} \tag{2.13}$$

l'equivalenza con il modello di Rasch la si ottiene se poniamo $u_{0jg} + r_{00g} = \theta_j$ e $\beta_i = -(\pi_{q00} + \pi_{000})$, per $i = q$.
Il modello può essere esteso includendo predittori al secondo e al terzo livello della gerarchia e, come proposto in Kamata e Cheong (2007)(*vedi anche* Cheong e Raudenbush (2000)), può essere esteso al caso multidimensionale

in ottica *between*; ciò significa che gruppi di items diversi misurano, ciascuno, un tratto latente differente o, per meglio dire, un item può riferirsi ad una soltanto delle variabili latenti presenti nel modello.

L'altro esempio di formulazione alternativa dei modelli IRT in ottica multilivello fa riferimento alla classe dei modelli **GLLAMM** (Generalized Linear Latent and Mixed Model).
In relazione a tali modelli si è soliti definire, per prima cosa, la forma dell'aspettativa condizionata delle risposte y (i pedici per le unità di osservazione vengono omessi per motivi di semplificazione notazionale).
L'aspettativa condizionata della risposta, dati due insiemi di variabili esplicative $\mathbf{x}$ e $\mathbf{z}$ ed un vettore di variabili latenti $\boldsymbol{\eta}$, è specificata tramite una funzione di link $g(.)$ ed un predittore lineare ν come segue:

$$g(E[y|\mathbf{x}, \boldsymbol{\eta}, \mathbf{z}]) = \nu \tag{2.14}$$

Il link può essere uno tra quelli utilizzati nei modelli lineari misti generalizzati.
Per un modello con L livelli ed $M_l (m = 1, \ldots, M_l)$ variabili latenti per ciascun livello $l (l = 2, \ldots, L)$, il predittore lineare ha la forma:

$$\boldsymbol{\nu} = \mathbf{x}'\boldsymbol{\beta} + \sum_{l=2}^{L} \sum_{m=1}^{M_l} \eta_m^{(l)} \mathbf{z}_m^{(l)'} \boldsymbol{\lambda}_m^{(l)} \tag{2.15}$$

con il primo elemento di $\boldsymbol{\lambda}_m^{(l)}$ posto uguale a 1.
Gli elementi di $\mathbf{x}$ sono le covariate associate agli effetti fissi $\boldsymbol{\beta}$, $\eta_m^{(l)}$ è l'$m-esima$ variabile latente al livello l ed $\boldsymbol{\eta}$ è il vettore delle variabili latenti tale che:

$$\boldsymbol{\eta} = (\eta_1^{(2)}, \eta_2^{(2)}, \ldots, \eta_{M_2}^{(2)}, \eta_1^{(3)}, \eta_2^{(3)}, \ldots, \eta_{M_3}^{(3)}, \ldots, \eta_1^{(L)}, \eta_2^{(L)}, \ldots, \eta_{M_L}^{(L)}) \tag{2.16}$$

Le variabili latenti appartenenti allo stesso livello sono generalmente correlate fra loro, mentre si assume l'indipendenza tra variabili latenti che fanno riferimento a livelli differenti.
Ciascuna variabile latente è moltiplicata per una combinazione lineare delle variabili $\mathbf{z}_m^{(l)} \boldsymbol{\lambda}_m^{(l)}$. L'apice di $\mathbf{z}_m^{(l)}$ indica che la variabile latente a cui si riferisce varia al livello l, mentre $\mathbf{z}_m^{(l)}$ in genere varia al livello immediatamente inferiore.
Come caso particolare dei modelli GLAMM è possibile definire il modello lineare misto generalizzato (GLMM); è sufficiente usare una sola variabile

esplicativa $z_{m1}^{(l)}$ per ogni variabile latente presente nel modello (con $\boldsymbol{\lambda}_{m1}^{(l)} = 1$). Si ha quindi la semplificazione:

$$\boldsymbol{\nu} = \mathbf{x}'\boldsymbol{\beta} + \sum_{l=2}^{L} \sum_{m=1}^{M_l} \eta_m^{(l)} z_{m1}^{(l)} \tag{2.17}$$

dove in genere $z_{11}^{(l)} = 1$.
Nel caso di un modello con tre livelli gerarchici, utilizzando al solito i pedici (i, j, g) per indicare le unità di primo, secondo e terzo livello, considerando le intercette random al secondo e terzo livello ed un coefficiente random al secondo livello, si ha:

$$\nu_{ijg} = \mathbf{x}'_{ijg}\boldsymbol{\beta} + \eta_{1jg}^{(2)} + \eta_{2jg}^{(2)} z_{2ijg}^{(2)} + \eta_{1g}^{(3)} \tag{2.18}$$

Al fine di mostrare la relazione tra i modelli GLMM e i modelli IRT, supponiamo di voler definire un modello di regressione per dati in cluster (con i unità di primo livello e j unità di secondo livello), e estendiamo l'ipotesi che due variabili latenti, $\boldsymbol{\eta}_j$, inducano dipendenza *within cluster* per le y_{ij}, date le covariate; il modello GLMM può essere scritto come:

$$g(E[\mathbf{y}_j|\mathbf{X}_j, \boldsymbol{\eta}_j, \mathbf{Z}_j]) = \boldsymbol{\nu}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\boldsymbol{\eta}_j \tag{2.19}$$

g è la funzione link, $\boldsymbol{\nu}_j$ è il predittore lineare mentre $\mathbf{X}_j$ sono le covariate associate agli effetti fissi $\boldsymbol{\beta}$ e $\mathbf{Z}_j$ rappresentano le covariate per gli effetti random $\boldsymbol{\eta}_j$.
Per essi si assume che $\boldsymbol{\eta}_j \sim N(0, \boldsymbol{\Psi})$.
È agevole verificare che esiste una equivalenza tra il modello lineare misto generalizzato(GLMM) e i modelli IRT, in quanto questi ultimi possono essere definiti come:

$$\boldsymbol{\nu}_j = \mathbf{I}\boldsymbol{\beta} + \boldsymbol{\Lambda}\boldsymbol{\theta}_j \tag{2.20}$$

Il modello di misura IRT può essere letto come un modello lineare misto generalizzato assumendo che:

- Gli items corrispondono alle unità di primo livello mentre gli individui rappresentano le unità di secondo livello;
- La matrice identità $\mathbf{I}$ sostituisce la matrice delle covariate $\mathbf{X}_j$;
- $\boldsymbol{\Lambda}$ sostituisce $\mathbf{Z}_j$ con la differenza che, però, $\boldsymbol{\Lambda}$ non varia tra i soggetti.

Anche in questo caso, i parametri di difficoltà sono gli effetti fissi del modello mentre i parametri delle abilità rappresentano gli effetti random.
Considerando il modello logistico ad 1 parametro e assumendo tre livelli gerarchici, l'equivalenza tra i due modelli implica che è possibile definire il modello di Rasch a tre livelli come:

$$log\left[\frac{p_{ijg}}{(1-p_{ijg})}\right] = \nu_{ijg} = \beta_i + \theta_{jg}^{(2)} + \theta_g^{(3)} \tag{2.21}$$

In questa formulazione $\theta_{jg}^{(2)} + \theta_g^{(3)}$ rappresenta, quindi, l'abilità del soggetto $j - esimo$ appartenente al gruppo g; risulta chiaro che, nel modello a tre livelli, i parametri di abilità sono somma di due componenti:

- $\theta_{jg}^{(2)}$ - l'abilità specifica associata al soggetto $j - esimo$ del gruppo g (indica di quanto devia l'abilità del soggetto j in relazione all'abilità media degli individui del gruppo g).
- $\theta_g^{(3)}$ - l'abilità media degli individui appartenenti al cluster g.

Come ulteriore esempio di approccio ai modelli multilivello merita dettaglio l'interpretazione fornita da Fox e Glas (2001); gli autori citati considerano una struttura gerarchica in cui gli studenti sono le unità di primo livello e le scuole le unità di secondo livello.
In particolare, l'abilità degli studenti è vista come variabile risposta in un modello di regressione multilivello; ovviamente, la variabile risposta è, però, latente.
Supponendo di avere le scuole indicizzate con $g = 1, \ldots, G$, si ha che gli studenti sono annidati in esse e indicizzati con $j = 1, \ldots, n_g$; indicando con $\mathbf{x_{jg}} = (x_{0jg}, x_{1jg}, \ldots, x_{Qjg})^t$ le covariate a livello studente, con in genere $x_{0jg} = 1$, il modello al primo livello viene scritto come:

$$\theta_{jg} = b_{0g} + \ldots + b_{qg}x_{qjg} + \ldots + b_{Qg}x_{Qjg} + e_{jg} \;\; con\; e_{jg} \sim N(0, \sigma^2) \tag{2.22}$$

I parametri di regressione possono variare tra le scuole per cui, denotando con $\mathbf{w_{qg}} = (w_{0qg}, w_{1qg}, \ldots, x_{Sqg})^t$ le covariate a livello delle scuole, con in genere $w_{0qg} = 1$, al secondo livello si ha:

$$b_{qg} = \gamma_{q0} + \ldots + \gamma_{qs}w_{sqg} + \ldots + \gamma_{qS}w_{Sqg} + u_{qg} \;\; per\; q = 0, \ldots, Q \tag{2.23}$$

$\mathbf{u_g}$, in questo caso, diventa un vettore normale multivariato con media 0 e matrice di varianza e covarianza $\mathbf{T}$, il cui generico elemento è $\tau_{qq'}^2$ per $q, q' = 0, \ldots, Q$.

Vi sono casi in cui è preferibile che non tutti i regressori del primo livello siano effetti casuali, costringendo in tal modo alcuni dei regressori ad essere effetti fissi.
E' agevole dimostrare che tale formulazione può essere vista come un modello ad effetti misti lineare. Se, per esempio, si considera il modello senza regressori, abbiamo che:

$$\begin{aligned} \theta_{jg} &= b_{0g} + e_{jg} \\ b_{0g} &= \gamma_{00} + u_{0g} \end{aligned} \tag{2.24}$$

θ_{jg} è modellato dall'intercetta b_{0g}, che è specifica per ogni scuola, e dal termine d'errore e_{jg} che si distribuisce come una normale con media zero e varianza σ^2_θ.
A sua volta l'intercetta casuale è definita da γ_{00}, ossia la media generale, e dal termine d'errore u_{0g} che si distribuisce normalmente con media zero e varianza τ^2_{00}.
Si assume, inoltre, che due abilità appartenenti al medesimo gruppo siano correlate mentre due abilità che fanno riferimento a due scuole differenti non lo siano, per cui:

$$Cov(\theta_{jg}, \theta_{j'g'}) = Cov(e_{jg}, e_{j'g'}) + Cov(u_{0g}, u_{0g'}) \tag{2.25}$$

$$= \begin{cases} \sigma^2_\theta + \tau^2_{00} & per\ j = j^{'}, g = g^{'} \\ \tau^2_{00} & per\ j \neq j^{'}, g = g^{'} \\ 0 & per\ g \neq g^{'} \end{cases} \tag{2.26}$$

È possibile notare come il modello multilivello senza covariate sia un modello a effetti misti lineare dal momento che:

$$\theta_{jg} = \underbrace{\gamma_{00}}_{parte\ fissa} + \underbrace{u_{0g} + e_{jg}}_{parte\ casuale} \tag{2.27}$$

I tre approcci descritti fino a questo momento costituiscono, spesso, la base di partenza per numerose estensioni ed interpretazioni; è da precisare che non solo quest'ultimo approccio, ma anche i precedenti, trovano applicazione in ottica bayesiana. Ne sono esempi, a riguardo, i lavori di Cho (2007), Kim (2007), Chaimongkol, Huffer e Kamata (2007), Jiao et al. (2012), Park e Bolt (2008).
Come si avrà modo di approfondire nel quarto capitolo, il presente lavoro di tesi fa riferimento al secondo approccio per la definizione del modello, considerandone una formulazione in ottica bayesiana.

2.4 Algoritmi di stima

In generale, diversi sono gli algoritmi e i programmi sviluppati per la stima dei modelli IRT multilivello.
In alcuni casi viene proposto un approccio a due step: esso consiste nello stimare i parametri di abilità con un modello IRT, per ciascun soggetto, al primo step, ma gli errori standard associati a tali stime non sono poi modellati nel secondo step ottenendo, necessariamente, stime distorte dei parametri. La distorsione è tanto maggiore quanto più la dimensione totale del campione è piccola.
Adams, Wilson e Wu (1997), Cheong e Raudenbush (2000) e Mislevy (1987) procedono, invece, alla stima dei modelli gerarchici utilizzando, per esempio, approssimazioni per grandi campioni o stime empiriche bayesiane, facendo ricorso, quindi, alla teoria della distribuzione normale. Ciò comporta, per esempio, l'introduzione di vincoli sull'ampiezza campionaria minima.
L'approccio bayesiano, invece, è libero dal vincolo di approssimazione normale ed è in grado di gestire anche problemi di integrazione complessi. Le distribuzioni a posteriori, sufficientemente complicate nei modelli gerarchici, vengono rappresentate come una collezione di distribuzioni di probabilità condizionate, di forma nota, attraverso l'algoritmo del Gibbs Sampler (*vedi ad es.* Maier (2001)); nel caso in cui le distribuzioni full conditional non siano di forma nota si fa ricorso all'algoritmo Metropolis-Hastings.
Patz e Junker (1999a,b) forniscono una descrizione dettagliata dell'utilizzo di tali algoritmi nel contesto dei modelli IRT.
In generale vale che, in un modello di risposta gerarchico, la densità a posteriori di interesse è costruita sulla base del modello di risposta e sulle distribuzioni *a priori* dei parametri.
In un contesto gerarchico, le distribuzioni *a priori* fanno riferimento e ai parametri del modello e ai rispettivi *hyperparameter*.
Supponendo di indicare con $\boldsymbol{\theta}_P = (\mu_\theta, \sigma^2_\theta)$ gli *hyperparameter* associati ai parametri di abilità e con $\boldsymbol{\beta}_P = (\mu_\beta, \sigma^2_\beta)$ gli *hyperparameter* associati ai parametri degli items, la densità *a posteriori* si scrive come:

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}) &\propto \int\int p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\beta}) p(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{\theta}_P, \boldsymbol{\beta}_P) p(\boldsymbol{\theta}_P, \boldsymbol{\beta}_P) d\boldsymbol{\beta}_P d\boldsymbol{\theta}_P \\ &\propto \int\int p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\beta}) p(\boldsymbol{\theta}|\boldsymbol{\theta}_P)(\boldsymbol{\beta}|\boldsymbol{\beta}_P) p(\boldsymbol{\theta}_P) p(\boldsymbol{\beta}_P) d\boldsymbol{\beta}_P d\boldsymbol{\theta}_P \\ &\propto \int\int \prod_{i,j} [p(y_{ij}|\theta_j, \beta_i) p(\theta_j|\boldsymbol{\theta}_P)(\beta_i|\boldsymbol{\beta}_P)] p(\boldsymbol{\theta}_P) p(\boldsymbol{\beta}_P) d\boldsymbol{\beta}_P d\boldsymbol{\theta}_P \quad (2.28) \end{aligned}$$

La connotazione gerarchica è ben evidente nella prima equazione, dove i parametri delle distribuzioni *a priori* sono, a loro volta, definiti tramite di-

stribuzioni di probabilità. I parametri associati agli item e quelli associati ai soggetti sono, poi, posti indipendenti l'uno dall'altro, così come gli *hyperparameter*, inducendo la fattorizzazione presente nella seconda equazione.
L'ultima fattorizzazione rende ancor più chiaro l'approccio gerarchico. Le osservazioni sono modellate, al primo livello, come indipendenti condizionatamente ai parametri di difficoltà e di abilità: $p(y_{ij}|\theta_j, \beta_i)$ rappresenta la parte relativa alla verosimiglianza del modello.
Al secondo livello, sono definite le distribuzioni *a priori* per i parametri del primo livello: $p(\theta_j|\boldsymbol{\theta}_P)$ descrive l'eterogeneità tra gli individui mentre $p(\beta_i|\boldsymbol{\beta}_P)$ quella tra gli items.
Al terzo livello vengono definite le cosiddette *hyperpriors* per i parametri delle *a priori* definite al livello precedente.
L'inferenza sui parametri di interesse deriva dalla combinazione dell'informazione derivante dai dati e da quella relativa alle distribuzioni *a priori*. Nei modelli gerarchici, quindi, l'informazione *a priori* si combina con la verosimiglianza, includendo una fonte di incertezza aggiuntiva dovuta alla presenza degli *hyperparameter*.
I problemi relativi all'integrazione aumentano in maniera consistente e proporzionale alla complessità dei parametri coinvolti. Se, per esempio, siamo interessati al calcolo della densità *a posteriori* dei parametri di abilità, il calcolo sarà il seguente:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto \int\int\int p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\beta})p(\boldsymbol{\theta}|\boldsymbol{\theta}_P)p(\boldsymbol{\theta}_P)p(\boldsymbol{\beta}|\boldsymbol{\beta}_P)p(\boldsymbol{\beta}_P)d\boldsymbol{\theta}_P d\boldsymbol{\beta} d\boldsymbol{\beta}_P \tag{2.29}$$

Se si assumono distribuzioni *a priori* normali, tali integrazioni possono avvenire utilizzando la quadratura Gauss Hermite, sebbene ottenere un'approssimazione soddisfacente di tale integrale, attraverso metodi di integrazione numerica, sia abbastanza complesso.
I metodi MCMC, come più volte detto, riescono a stimare tutti i parametri del modello simultaneamente e si rivelano particolarmente utili in questi casi, quando si considerano modelli gerarchici con elevati livelli di complessità.
Chiudendo questa parentesi di tipo computazionale merita cenno, a solo titolo di esempio, ancora il contributo di Fox e Glas (2001); nello stesso articolo, gli autori generalizzano al caso multilivello il risultato ottenuto da Albert (1992), il quale ha implementato il Gibbs sampler per un *normal ogive model* a due parametri. Partendo dal modello che considera $i(i = 1, \ldots, I)$ items, si ha:

$$P(Y_{ijg}) = \Phi(a_i\theta_{jg} - \beta_i) \tag{2.30}$$

Φ si indica la funzione cumulata della distribuzione normale standard, a_i e β_i rappresentano i parametri di discriminazione e di difficoltà, che sono indicati con $\omega_i = (a_i, \beta_i)^t$.

θ_{jg} è modellato sulla base della (2.22) e della (2.23). A questo punto è chiaro che la distribuzione a posteriori congiunta di tutti i parametri coinvolti nel modello multilivello è molto complessa, ragion per cui gli autori hanno sviluppato una procedura MCMC ad hoc.
Per implementare il Gibbs sampler hanno utilizzato la tecnica detta di *data augmentation*; essa consiste nell'introdurre le variabili indipendenti Z_{ijg}, caratterizzate dal fatto che hanno distribuzioni normali, la cui media è $a_i\theta_{jg}-\beta_i$ e la cui varianza è pari ad 1.
L'assunzione è che

$$Y_{ijg} = \begin{cases} 1 & \text{se } Z_{ijg} > 0 \\ 0 & \text{altrimenti} \end{cases} \tag{2.31}$$

Attraverso l'MCMC da loro implementato, è possibile ottenere le distribuzioni *full conditional* a partire dalla distribuzione congiunta a posteriori definita come:

$$\begin{aligned} p(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\omega}, \mathbf{b}, \sigma_\theta^2, \boldsymbol{\gamma}, \mathbf{T} | \mathbf{Y}, \mathbf{X}, \mathbf{W}) \propto \prod_{g=1}^{G} \prod_{j=1}^{n_g} \left(\left(\prod_{i=1}^{I} p(Z_{ijg} | \theta_{jg}, \omega_i, y_{ijg}) \right) p(\theta_{jg} | \mathbf{b}_g, \sigma_\theta^2, \mathbf{X}_g) \right) \\ p(\mathbf{b}_g | \boldsymbol{\gamma}, \mathbf{T}, \mathbf{W}_g) p(\boldsymbol{\gamma} | \mathbf{T}) \\ p(\boldsymbol{\omega}) p(\sigma_\theta^2) p(\mathbf{T}) \end{aligned} \tag{2.32}$$

A fine del presente capitolo è doveroso precisare che la trattazione relativa ai modelli IRT, in ottica multidimensionale, è stata demandata al capitolo successivo al solo fine di rendere più evidente il legame tra i modelli in questione e la letteratura relativa alla modellizzazione del meccanismo generatore del dato mancante.

Capitolo 3

Approccio all'analisi dei Dati Mancanti mediante l'Item Response Theory

3.1 Introduzione

Il presente capitolo ha come scopo quello di evidenziare il ruolo dell'Item Response Theory nell'analisi e trattamento dei dati mancanti. Vasta è la letteratura in merito, sia in ottica frequentista che bayesiana; il crescente interesse da parte dei ricercatori è dettato dalla necessità di gestire il meccanismo che governa i *missing values*, al fine di ridurre la distorsione delle stime dei parametri di interesse in matrici di dati incomplete.
Prima di proseguire è bene far notare che, in generale, si distinguono tre tipologie principali di dati mancanti, riconducibili a: mancata copertura (non-coverage), mancata risposta totale (unit or total nonresponse) e mancata risposta parziale (item nonresponse).
La mancata copertura dell'indagine è dovuta alla non inclusione di alcune unità della popolazione obiettivo nella lista di campionamento determinando, per esse, una probabilità nulla di entrare a far parte del campione.
La mancata risposta totale si verifica quando, per l'unità selezionata, non si dispone di alcun dato; la causa è, in genere, il rifiuto di collaborare da parte dell'intervistato o l'impossibilità di contattarlo.
Si è, invece, in presenza di mancata risposta parziale quando solo una parte delle informazioni relative ad una data unità statistica non risulta disponibile; ciò si verifica, ad esempio, nei casi in cui l'intervistatore dimentica di porre una domanda o di registrare una risposta oppure l'intervistato non è in grado, o non ha intenzione, di fornire una determinata informazione.

In ambito psicometrico, in particolare nelle indagini su larga scala quali **PISA** (Programme of International Student Assessment), **TIMSS** (Trends in International Mathematics and Science Studies) o **NEAP** (National Assessment of Educational Progress) per esempio, è frequente dover gestire matrici di dati con un numero consistente di valori mancanti.
In questo genere di *surveys*, gli items sono organizzati nei cosiddetti *booklets*, i quali, a loro volta, sono assegnati agli studenti in maniera casuale. Questo fa sì che gli items somministrati a ciascuno studente siano differenti a seconda del *booklet* assegnato.
La presenza di dati mancanti in relazione agli items non somministrati sono noti, in letteratura, con il termine di *missing by design*, ossia valori mancanti dovuti al disegno campionario.
I dati mancanti che non riguardano tale tipologia possono, invece, avere origini e cause differenti: scarsa preparazione, mancanza di tempo, strategie di autoselezione dei quesiti da parte dello studente, mancanza di motivazione, insieme a numerose altre ragioni.
Il target di questo studio è proprio il trattamento della seconda tipologia di *missing* poiché, come vedremo nel dettaglio nei prossimi paragrafi, essi possono determinare, in certe condizioni, stime distorte e inefficienti dei parametri del modello.

3.2 Il meccanismo generatore dei dati mancanti

Sulla base di quanto illustrato da Rubin (Little e Rubin 2002; Rubin 1976), è possibile distinguere tre diverse tipologie di *missing* in base ai diversi meccanismi generatori del dato mancante:

- **Missing completely at random** (MCAR);
- **Missing at random** (MAR);
- **Missing not at random** (MNAR).

Per entrare nel dettaglio di quanto appena esposto, si definisca la matrice dei dati osservati $\mathbf{Y}$ partizionata come $\mathbf{Y} \equiv (\mathbf{Y}_{obs}, \mathbf{Y}_{miss})$, si indichi con $\mathbf{Z}$ la matrice delle covariate e con ξ il parametro che governa la funzione di *missing*.
Rubin introduce il concetto di variabile indicatrice della presenza\assenza di *missing*, intesa come variabile casuale a cui è associata una distribuzione di probabilità.
A tale proposito, si definisca la matrice indicatrice $\mathbf{D}$ il cui generico elemento

è tale che:

$$d_i = \begin{cases} 1 & \text{se } Y_i \text{ è osservato} \\ 0 & \text{altrimenti} \end{cases} \tag{3.1}$$

Se la distribuzione dei dati mancanti è indipendente dai dati osservati e non osservati di $\mathbf{Y}$ e dalla matrice delle covariate $\mathbf{Z}$, cioè se vale:

$$Pr(\mathbf{D}|\mathbf{Y}_{obs}, \mathbf{Y}_{miss}, \mathbf{Z}, \xi) = Pr(\mathbf{D}|\xi) \tag{3.2}$$

allora il missing è *completely at random.*

Se la distribuzione dei dati mancanti, invece, dipende dai soli dati osservati $\mathbf{Y}_{obs}$ e dalla matrice delle covariate $\mathbf{Z}$, quindi se vale:

$$Pr(\mathbf{D}|\mathbf{Y}_{obs}, \mathbf{Y}_{miss}, \mathbf{Z}, \xi) = Pr(\mathbf{D}|\mathbf{Y}_{obs}, \mathbf{Z}, \xi) \tag{3.3}$$

il missing è *at random.*

Se, in ultimo, si verifica che la distribuzione condizionata dei dati mancanti dipende proprio dai dati non osservati $\mathbf{Y}_{miss}$, ottenendo:

$$Pr(\mathbf{D}|\mathbf{Y}_{obs}, \mathbf{Y}_{miss}, \mathbf{Z}, \xi) \neq Pr(\mathbf{D}|\mathbf{Y}_{obs}, \mathbf{Z}, \xi) \tag{3.4}$$

il missing è *not at random.*

A questo punto, è necessario distinguere quale di questi meccanismi possa definirsi ***ignorabile*** e quale ***non ignorabile***.
Si indichi con $\boldsymbol{\theta}$ il vettore dei parametri riferiti alla distribuzione di probabilità di $\mathbf{Y}$; se il processo generatore dei dati mancanti è MAR (o MCAR) e il vettore dei parametri $\boldsymbol{\theta}$ è distinto dal vettore dei parametri $\boldsymbol{\xi}$, allora il meccanismo è ignorabile. Dal punto di vista bayesiano, ciò si traduce nella necessità che sia verificata, oltre all'ipotesi MAR, anche la condizione che le distribuzioni a priori su $\boldsymbol{\theta}$ e $\boldsymbol{\xi}$ siano indipendenti.
Sulla base di quanto appena esposto, nell'ipotesi di meccanismo ignorabile del dato mancante, è possibile stimare il vettore dei parametri $\boldsymbol{\theta}$ senza dover specificare la distribuzione del meccanismo generatore del dato mancante.

Per il trattamento della mancata risposta, le strategie proposte sono numerose; in letteratura si distingue tra metodi *ad hoc* e metodi *model based.* In questo lavoro l'attenzione sarà rivolta, in particolar modo, alla seconda categoria.

3.2.1 Metodi ad hoc

La Complete Case Analysis (CCA) procede con l'analisi delle sole unità per le quali si hanno a disposizione informazioni complete. Tale metodologia è implementata in molti software come opzione di default e se, da un lato, ha il vantaggio di essere molto semplice in termini di applicabilità, dall'altro, produce serie distorsioni e inefficienze nelle stime qualora il meccanismo generatore del dato mancante non sia MCAR.
Vi è da aggiungere che, anche nei casi in cui sia ipotizzabile una distribuzione dei dati mancanti completamente casuale, la perdita di informazione legata alla riduzione delle unità analizzate determina, al pari, una riduzione della precisione delle stime.
Un altro approccio, noto come Available Case Analysis (ACA), consiste nell'effettuare l'analisi dei soli casi disponibili, ossia considerare tutti i casi per i quali la variabile di interesse è presente; ciò ha lo svantaggio ovvio di ottenere, per ogni variabile, indicatori la cui base campionaria è differente.
Tale metodo produce stime consistenti solo sotto ipotesi MAR.
E' chiaro che i metodi citati, sebbene di facile applicazione, si basano su un'ipotesi molto stringente, ossia l'assunzione di meccanismo generatore del dato mancante casuale.
Un terzo approccio fa riferimento ai metodi di imputazione singola, che hanno lo scopo di assegnare un valore sostitutivo al dato mancante ottenendo, in tal modo, una matrice di dati completa. Tale pratica è molto diffusa poiché lavorare sul dataset, come se questo fosse completo, facilita l'analisi dei dati e rende più semplice la presentazione dei risultati; tale metodologia, inoltre, consente di effettuare analisi che, molto spesso, conducono a risultati consistenti, a differenza di quanto avviene nel caso di dataset incompleti.
Per contro, non garantisce di ottenere stime meno distorte e ciò dipende, in particolar modo, dalla tipologia di *missing*, dalla procedura utilizzata e dal tipo di stima applicato. Sicuramente l'aspetto meno convincente è connesso al fatto che i valori imputati vengono trattati come se fossero stati effettivamente osservati, ignorando due fonti di variabilità: quella dovuta al meccanismo di mancata risposta e quella dovuta alle imputazioni. Il rischio è di ottenere sottostime degli errori standard.
Si è soliti distinguere i metodi di imputazione in parametrici e non parametrici, a seconda che essi facciano uso o meno di un modello a priori esplicitato (in genere si utilizza il modello stesso applicato per descrivere i dati).
In ultimo vi sono dei casi in cui, piuttosto che utilizzare tali metodi, si preferisce considerare il *missing* come una categoria di risposta aggiuntiva.

3.2.2 Metodi model-based

I metodi appartenenti a tale categoria, in contrasto con quelli accennati in precedenza, combinano l'informazione disponibile contenuta nei dati osservati con quella derivante da un'esplicita assunzione riguardante il meccanismo generatore del dato mancante. Fanno parte di tale categoria i metodi basati sull'algoritmo EM, i metodi basati sui pesi campionari, l'imputazione multipla e i modelli bayesiani *full probability*.
Il metodo basato sull'algoritmo EM (Dempster, Laird e Rubin 1977) consente, attraverso una procedura iterativa, di ottenere le stime di massima verosimiglianza dei parametri; partendo da un valore iniziale, ad ogni iterazione, applica due step o passi: *Expectation step* e *Maximisation step*.
Il passo E calcola i valori attesi condizionati dei *missing*, dati i valori osservati e le stime correnti dei parametri, sostituendo ai dati mancanti i valori attesi.
In realtà, è corretto affermare che l'algoritmo EM non sostituisce direttamente i *missing* con i valori attesi, piuttosto le funzioni di $\mathbf{Y}_{miss}$ che compaiono nella log-verosimiglianza dei dati completi.
Il passo M calcola le stime di massima verosimiglianza sui dati "completati", come se non ci fossero dati mancanti.
Questo algoritmo può essere applicato anche per meccanismi non ignorabili.
Per quanto riguarda i metodi di riponderazione, essi si applicano in casi di unità completamente non osservate. Modificano i pesi campionari (in genere, in presenza di dati completi, si associa ad ogni caso un peso che è inversamente proporzionale alla probabilità di selezione) in maniera tale da redistribuire il peso complessivo dei non rispondenti sui rispondenti.
Quando il dataset contiene informazioni utilizzabili a fini predittivi, una pratica assai diffusa è rappresentata dall'imputazione multipla; l'idea di base del metodo è quella di generare più di un valore ($m > 2$) da imputare per ogni dato mancante, campionando da un'opportuna distribuzione ritenuta plausibile per $y_{i,miss}$. Sostituendo in tutti gli items del dataset i valori generati alla prima estrazione, poi quelli alla seconda fino all'$m-esima$, si ottengono m matrici complete. Su ognuna di esse, separatamente, vengono effettuate le analisi statistiche usate per matrici complete.
Successivamente, gli m valori delle stime dei parametri e delle relative misure di incertezza vengono sintetizzati in unica misura in maniera tale che il risultato inferenziale finale tenga conto della variabilità osservata nelle stime. Va precisato che, in questa procedura, un ruolo importante riveste la modalità con cui le m imputazioni vengono generate.
Per ulteriori approfondimenti sulle tecniche enunciate si può far riferimento a Rubin (1976) o Little e Rubin (2002).

Per quanto concerne i modelli *full probability* bayesiani, a differenza di quanto avviene per i metodi di imputazione multipla, in cui l'imputazione dei dati mancanti e l'analisi del modello sono effettuate in due momenti separati, per essi, invece, ciò avviene simultaneamente.
Si consideri un modello di regressione lineare, secondo l'approccio bayesiano, ossia:

$$y_i \sim N(\mu_i, \sigma^2) \tag{3.5}$$

$$\mu_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ki} \tag{3.6}$$

con *a priori* definite per i parametri $\beta_0, \beta_1, \ldots, \beta_p, \sigma^2$

Se si assume un meccanismo ignorabile, i valori di $\mathbf{y}_{miss}$ possono essere generati a partire dalla distribuzione predittiva a posteriori $f(\mathbf{y}_{miss}|\mathbf{y}_{obs}, \beta, \sigma^2)$ ottenendo inferenze valide per le stime dei parametri del modello, ossia $\boldsymbol{\beta}$ e σ^2. Se, invece, siamo in presenza di un meccanismo MNAR, si introduce, accanto al modello di regressione, il modello che tiene conto del meccanismo generatore del dato mancante, tipicamente:

$$d_i \sim Bernoulli(p_i) \tag{3.7}$$

$$link(p_i) = f(y_i, \xi) \tag{3.8}$$

con un'*a priori* definita sul parametro ξ.
d_i rappresenta, come sempre, l'indicatore della presenza \ assenza di missing e il link potrebbe essere, per esempio, una funzione logit o probit.
Nei casi MNAR, dunque, i due modelli vengono stimati congiuntamente.

Dopo questa breve disamina sui numerosi e diversi metodi di trattamento dei dati mancanti, nel prossimo paragrafo si focalizzerà l'attenzione su un particolare modello proposto, di recente, in letteratura da Holman e Glas (2005), interessante come richiamo teorico per le applicazioni effettuate nel presente lavoro di tesi.

3.3 Il modello di Holman & Glas

Il modello che verrà illustrato si inserisce nel solco dei lavori proposti da Moustaki (1996) (*vedi anche* Bartholomew e Knott (1999)), la quale ha sviluppato un modello generale *a classi latenti* e *a tratti latenti* per variabili miste; in tale contesto sono stati proposti metodi per il trattamento dei dati mancanti

con meccanismo non ignorabile. In Moustaki e Knott (2000), il dato mancante è trattato come una categoria di risposta aggiuntiva in un modello di risposta nominale; in O'Muircheartaigh e Moustaki (1999) si considera un modello a variabili latenti specificato da due dimensioni: la propensione alla risposta e l'abilità del soggetto.
E' proprio in riferimento a quest'ultimo approccio che si inserisce il modello di Holman e Glas (2005).
Nel contesto dell'*Educational Assessment* è assai frequente dover ipotizzare situazioni in cui il meccanismo generatore del dato mancante sia non ignorabile, ossia, è plausibile dover assumere che la mancata risposta dipenda, in molti casi, dall'abilità stessa del soggetto. Ha senso, quindi, ritenere che le due variabili latenti, una relativa all'abilità del soggetto e l'altra relativa alla sua propensione alla risposta, siano correlate.
Sulla base di quanto appena detto, il modello proposto da Holman e Glas (2005) definisce:

- un modello IRT a due parametri per la matrice indicatrice dei missing $\mathbf{D}$ la cui latente, che rappresenta la propensione alla risposta, è rappresentata da $\boldsymbol{\xi}$;
- un modello IRT a due parametri per la matrice dei dati $\mathbf{Y}$ la cui latente, che rappresenta l'abilità del soggetto, è rappresentata da $\boldsymbol{\theta}$.

Gli autori definiscono il modello MAR, in termini di verosimiglianza, come:

$$\prod_{ij} p(y_{ij}|d_{ij}, \theta_j, \alpha_i, \beta_i) p(d_{ij}|\xi_j, \gamma_i, \delta_i) g(\theta_j) g(\xi_j) \tag{3.9}$$

$p(y_{ij}|d_{ij}, \theta_j, \alpha_i, \beta_i)$ rappresenta il modello IRT a due parametri per i dati osservati (con α_i parametro di discriminazione e β_i parametro di difficoltà) mentre $p(d_{ij}|\xi_j, \gamma_i, \delta_i)$ rappresenta il modello IRT a due parametri per i dati mancanti (con γ_i parametro di discriminazione e δ_i parametro di difficoltà). $g(\theta_j)$ e $g(\xi_j)$ sono, come già specificato, le densità delle latenti.
Il modello, appena definito, consente di ignorare il meccanismo generatore del dato mancante ragion per cui, nel processo di stima MML (Marginal Maximum Likelihood), $p(d_{ij}|\xi_j, \gamma_i, \delta_i) g(\xi_j)$ può essere ignorato. Da notare che nel modello MAR le due latenti non sono affatto correlate.
Il secondo modello, definito MNAR, tiene conto, nella fase di stima, del meccanismo generatore del dato mancante e si caratterizza per il fatto che i parametri delle variabili latenti non sono indipendenti, ottenendo:

$$\prod_{ij} p(y_{ij}|d_{ij}, \theta_j, \alpha_i, \beta_i) p(d_{ij}|\xi_j, \gamma_i, \delta_i) g(\theta_j, \xi_j|\Sigma) \tag{3.10}$$

$g(.)$, in genere, è una densità normale multivariata con vettore delle medie nullo e matrice di varianze e covarianze Σ. In ottica bayesiana, $g(\theta_j, \xi_j|\Sigma)$ rappresenta la distribuzione *a priori* per le latenti.
In questo caso, l'inferenza sotto l'ipotesi di ignorabilità non è giustificata in quanto le distribuzioni *a priori* dei parametri del modello sono tutt'altro che indipendenti.
Risulta chiaro che, in tale ottica, ciò equivale a definire un modello IRT multidimensionale.
A riguardo, l'approccio proposto da Rose, von Davier e Xu (2010) risulta essere particolarmente interessante. Relativamente al trattamento dei dati mancanti attraverso i modelli IRT, con riferimento alla modellazione del caso MNAR con due dimensioni latenti, essi propongono di far riferimento alla classe di modelli presentati nell'articolo di Adams, Wilson e Wang (1997), punto di riferimento teorico per la maggior parte dei lavori nel contesto dei modelli IRT.

3.3.1 Modelli IRT multidimensionali

MRCMLM è l'acronimo di *Multidimensional random coefficient multinomial logit model*; proposto da Adams, Wilson e Wang (1997), esso può essere inteso anche come un'estensione al caso multidimensionale del modello di Rasch.
Si assuma che ci sia più di un tratto latente $\boldsymbol{\theta}$ e che l'item $i - esimo$ (con $i = 1, \ldots, I$) abbia $K_i + 1$ (con $k = 0, 1, \ldots, K_i$) categorie; sia Y_{ik} la variabile casuale tale che:

$$Y_{ik} = \begin{cases} 1 & \text{se la risposta all'item } i \text{ è uguale a } k \\ 0 & \text{altrimenti} \end{cases} \tag{3.11}$$

Il modello MRCMLM si scrive come:

$$P(Y_{ik} = 1; \mathbf{A}, \mathbf{B}, \boldsymbol{\delta}|\boldsymbol{\theta}) = \frac{exp(\mathbf{b}_{ik}\boldsymbol{\theta} + \mathbf{a}_{ik}^{'}\boldsymbol{\delta})}{\sum_{k=1}^{K_i}(\mathbf{b}_{ik}\boldsymbol{\theta} + \mathbf{a}_{ik}^{'}\boldsymbol{\delta})} \tag{3.12}$$

Le matrici $\mathbf{A}$ e $\mathbf{B}$ sono note, rispettivamente, come matrice di *scoring* e matrice *disegno* e vengono utilizzate con lo scopo di specificare la forma funzionale del modello; la matrice $\mathbf{A}$ descrive la relazione tra gli items e i parametri di difficoltà mentre la matrice $\mathbf{B}$ quella tra gli items e le dimensioni latenti. L'introduzione di queste due matrici rende il modello molto flessibile e in grado di rappresentare un'ampia classe di modelli appartenenti alla famiglia di Rasch.
Gli autori distinguono, nella trattazione, tra due diverse sottoclassi del modello, note rispettivamente come *Between-Item-multidimensional model* e

Within-Item-multidimensional model. Nel primo modello si hanno diversi tratti latenti unidimensionali e ogni item misura solo un tratto latente; nel secondo, invece, alcuni o tutti gli items sono legati a più di una dimensione latente.
Per gli scopi del presente lavoro, si farà riferimento al primo modello. La matrice di scoring e quella di disegno vengono adattate al tipo di modello in questione per cui, per items dicotomici, la probabilità di risposta corretta da parte del soggetto j all'item i semplifica in:

$$P(Y_{ij} = 1|\boldsymbol{\theta}_j) = \frac{exp(\mathbf{r_i'}\boldsymbol{\theta}_j + \beta_i)}{1 + exp(\mathbf{r_i'}\boldsymbol{\theta}_j + \beta_i)} \tag{3.13}$$

β_i rappresenta il parametro di difficoltà (o facilità) dell'item i e $\boldsymbol{\theta}_j$ il vettore dei parametri del soggetto j.
$r_i = (r_{i1}, r_{i2}, \ldots, r_{im})'$ dove

$$r_{im} = \begin{cases} 1 & \text{se l'item } i \text{ misura la dimensione } m \\ 0 & \text{altrimenti} \end{cases} \tag{3.14}$$

Se facciamo riferimento alle due variabili latenti introdotte in precedenza per il modello di Holman & Glas, ossia $\boldsymbol{\theta}$ (variabile latente dell'abilità) e $\boldsymbol{\xi}$ (variabile latente della propensione alla risposta), la differenza tra il modello *Between items* e quello *Within items* può essere facilmente intuita da quanto rappresentato in Figura 3.1 e 3.2.

Obiettivo del capitolo successivo sarà proprio la definizione di un modello multidimensionale *between items* per le latenti θ e ξ, al fine di modellizzare la situazione di *missing* non ignorabile, avvalendosi del linguaggio dei modelli grafici bayesiani ed estendendo l'applicazione ad un modello gerarchico a tre livelli.

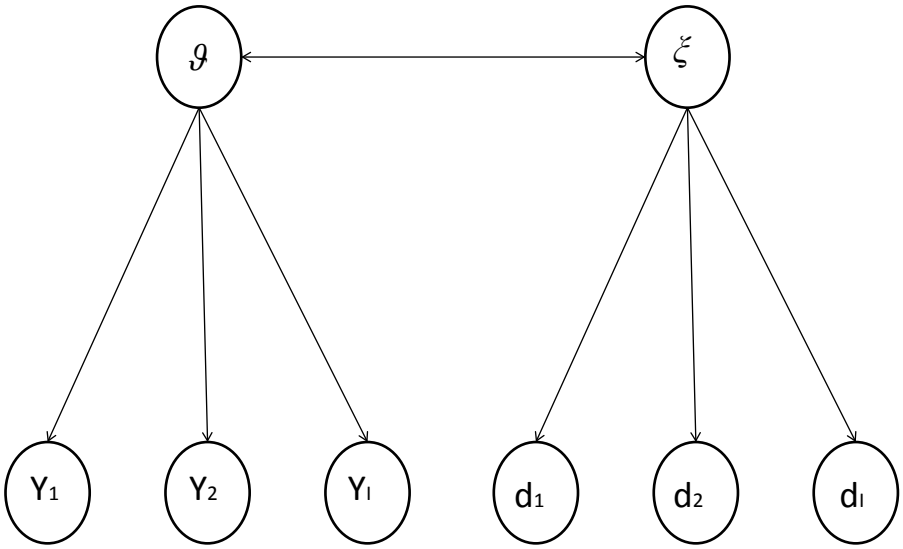


Figura 3.1: Esempio di Between-item-multidimensional model

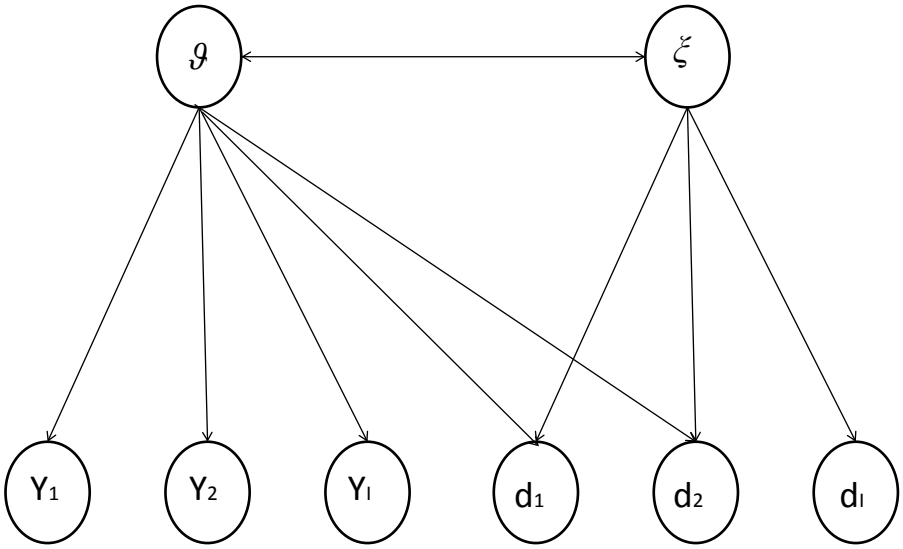


Figura 3.2: Esempio di Within-item-multidimensional model

Capitolo 4

Specificazione e caratteristiche del modello

4.1 Introduzione

Nel presente capitolo, il modello di Rasch gerarchico rappresenta l'elemento cardine della trattazione, con particolare riferimento alla sua definizione bidimensionale in ottica *between items*.
Come anticipato nel terzo capitolo, le due dimensioni latenti che caratterizzano il modello rappresentano, rispettivamente, l'abilità e la propensione alla risposta del soggetto. La prima è misurata dal gruppo di items somministrato tramite il questionario, la seconda dal gruppo di items, creato *ad hoc*, per la codifica della presenza\assenza del *missing* per ciascuno degli items del primo gruppo.
Obiettivo principale del lavoro è quello di descrivere i modelli IRT multilivello (bidimensionali) attraverso il linguaggio dei modelli grafici bayesiani, con l'indubbio vantaggio di comunicare la struttura "matematica" del modello nella modalità più intuitiva e versatile possibile, ossia attraverso la sua rappresentazione grafica.
Il secondo obiettivo concerne la possibilità di comparare tre diversi approcci al trattamento dei dati mancanti, mediante tre differenti specificazioni dei modelli appena descritti, al fine di individuare eventuali distorsioni nelle stime dei parametri, in caso di probabile violazione del princìpio di ignorabilità.
Il confronto fa riferimento ai seguenti tre modelli:

1. Modello **NIM** (*Non Ignorable Missing*).
   È il modello IRT bidimensionale *between items* che tiene conto del meccanismo generatore del dato mancante poiché considera due dimensio-

ni latenti: l'abilità e la propensione alla risposta. I parametri che le definiscono vengono, in questo caso, considerati non indipendenti.

2. Modello **IM** (*Ignorable Missing*).
   È il modello IRT unidimensionale che lascia la matrice dei dati incompleta; non modellizza il meccanismo generatore del dato mancante e di conseguenza considera, come unica variabile latente, l'abilità.

3. Modello **ZIM** (*Zero Imputation Missing*).
   È il modello IRT unidimensionale che considera la matrice dei dati completa, avendo precedentemente sostituito i dati mancanti con lo zero. Ovviamente, anche in questo caso, l'unica dimensione latente è l'abilità.

Il primo modello, come anticipato, ipotizza un meccanismo non ignorabile poiché considera le due dimensioni latenti non indipendenti; si avrà, quindi, un nodo bivariato latente sia al secondo che al terzo livello gerarchico.
Il secondo modello ipotizza un meccanismo di tipo ignorabile poiché sottintende l'indipendenza tra le latenti e ipotizza un meccanismo MAR, in quanto sfrutta la capacità del software di imputare i valori mancanti sulla base della distribuzione predittiva a posteriori (condizionata ai soli dati osservati e ai parametri non noti presenti nel modello).
Il terzo modello assume che tutte le risposte mancanti siano risposte sbagliate imputando lo "zero" al posto degli "NA" ed ottenendo, così, una matrice di dati completa. Quest'ultimo approccio trova riscontro in molte applicazioni pratiche.
La comparazione delle stime dei parametri dei modelli non sarà l'unico elemento su cui concentrarsi, ancora più informativa sarà la misura della correlazione tra le due dimensioni latenti nel modello bidimensionale. Quanto più forte sarà la correlazione tra di esse, tanto più consistente sarà la supposta deviazione dall'ipotesi di ignorabilità del meccanismo generatore del dato mancante.
Concludendo, va sottolineato che l'approccio gerarchico, o multilivello, rende più affascinante e complessa tale comparazione; la caratteristica dei dati "annidati" è molto diffusa in matrici di dati derivanti dal contesto educativo e non può essere ignorata.

Sembra opportuno far anticipare la descrizione e rappresentazione dei modelli grafici dalla trattazione delle loro caratteristiche analitiche e matematiche, per rendere più chiari ed evidenti i richiami teorici ai modelli visti nel secondo capitolo. In particolare, dalla lettura dei prossimi paragrafi, risulterà chiaro

che il contesto teorico di riferimento principale è quello dei modelli GLAMM, in una prospettiva bayesiana.
La trattazione parte con la definizione del modello unidimensionale per poi considerare la sua estensione al caso bidimensionale.

4.2 Il modello unidimensionale

Il paragrafo in questione descrive le caratteristiche del modello di Rasch unidimensionale a tre livelli e, quindi, definisce i modelli **IM** e **ZIM**.
Considerando tre livelli gerarchici e riprendendo la notazione solita, si indichi con $i = 1, \ldots, I$ le unità di primo livello (gli items), con $j = 1, \ldots, J$ le unità di secondo livello (gli studenti) e con $g = 1, \ldots, G$ le unità di terzo livello (le scuole o i gruppi territoriali, ad esempio).
Sia y_{ijg} la risposta dicotomica all'item i da parte dello studente j del gruppo g, che assume valore 1 in caso di risposta corretta, 0 altrimenti.
Assumendo che, per ogni terna (i, j, g), esse siano delle variabili causali bernoulliane indipendenti con probabilità di riposta corretta $p_{ijg} = P(y_{ijg} = 1)$, utilizzando la medesima notazione introdotta nel secondo capitolo, si ha:

$$y_{ijg} \sim Bernoulli(p_{ijg}) \tag{4.1}$$

$$logit(p_{ijg}) = \theta_g^{(3)} + \theta_{jg}^{(2)} - \beta_i \tag{4.2}$$

Poiché l'obiettivo è la costruzione di un modello generale, che si adatti a qualsiasi tipo di situazione reale, è risultato utile apportare piccole modifiche alla notazione usuale, in maniera tale che il modello possa trovare applicazione anche nei casi in cui gli items somministrati non siano uguali per tutti i rispondenti[1].
A tale proposito, come suggerito in Gelman e Hill (2006), è conveniente indicizzare le risposte individuali con $l = 1, \ldots, n$, in maniera tale che ad ogni risposta l sia associata una persona $j[l]$, un gruppo $g[l]$ ed un item $i[l]$.
In relazione al modello unidimensionale, il logit subisce la seguente modifica:[2]

$$y_l \sim Bernoulli(p_l) \tag{4.3}$$

$$logit(p_l) = \theta_{g[l]}^{(3)} + \theta_{j[l]}^{(2)} - \beta_{i[l]} \tag{4.4}$$

[1]Le modifiche notazionali si ispirano alla logica del linguaggio Bugs proprio per favorire la massima coincidenza tra enunciazione analitica del modello e rappresentazione grafica.

[2]In Openbugs non si può usare lo stesso indice per due livelli diversi per cui, nel logit del modello grafico, $g[l]$ sarà sostituito da $reg[l]$, $j[l]$ da $std[l]$ e $i[l]$ da $x[l]$, lasciando gli indici ijg per i cicli fuori dal logit.

dove:

- $\theta_j^{(2)}$ rappresenta l'abilità del soggetto j (annidato nel gruppo g); è un effetto random e, per esso, si assume che:

$$\theta_j^{(2)} \sim N(\gamma, \tau_{\theta^{(2)}})$$

- $\theta_g^{(3)}$ rappresenta l'effetto random per il gruppo g; per esso si assume che:

$$\theta_g^{(3)} \sim N(0, \tau_{\theta^{(3)}})$$

- β_i rappresenta l'effetto fisso, quindi il parametro di difficoltà per l'item i; per esso si assume che:

$$\beta_i \sim N(0, 0.0001)$$

.

Si assume, inoltre, che i due effetti random siano mutuamente indipendenti.

A questo punto sono doverose due precisazioni:

- È necessario definire le distribuzioni *a priori* per tutti i nodi stocastici del modello e, quindi, anche per γ, $\tau_{\theta^{(2)}}$ e $\tau_{\theta^{(3)}}$;
- Openbugs (così come Winbugs) definisce la distribuzione normale in termini di media μ e parametro di precisione τ, per cui la distribuzione *a priori* è attribuita all'inverso della varianza.

In questo studio, per l'effetto fisso γ si è deciso di adottare, come distribuzione *a priori*, la distribuzione normale con i seguenti parametri:

$$\gamma \sim N(0, 1)$$

In relazione all'altro effetto fisso β_i, va precisato che la scelta di considerare, per i parametri di difficoltà, una varianza elevata (10^4) è dettata dalla volontà di rendere le distribuzioni *a priori* non informative.
Ai parametri di precisione è stata attribuita una distribuzione *Gamma* per cui:

$$\tau_{\theta^{(2)}} \sim Gamma(0.001, 0.001)$$
$$\tau_{\theta^{(3)}} \sim Gamma(0.001, 0.001)$$

Anche la definizione dei parametri della distribuzione *Gamma* riflette la scelta di rendere le distribuzioni *a priori* non informative.
Essa, inoltre, gode della proprietà auspicabile di essere una distribuzione *coniugata* per il parametro di precisione di una distribuzione normale.[3]

4.3 Il modello bidimensionale

L'estensione del modello univariato al caso bivariato, in ottica *between items*, è di facile derivazione; oltre alla variabile risposta y_{ijg}, si considera la variabile dicotomica d_{ijg} che assume valore 1 se y_{ijg} è osservato, 0 altrimenti. Anche le d_{ijg} possono essere viste come variabili casuali bernoulliane indipendenti, con probabilità di riposta corretta $k_{ijg} = P(d_{ijg} = 1)$.
Facendo riferimento direttamente alla notazione più generale, il modello si caratterizza per la seguente formulazione:[4]:

$$\begin{array}{cc} \textit{Per la matrice } \mathbf{Y} & \textit{Per la matrice } \mathbf{D} \\ y_l \sim Bernoulli(p_l), & d_l \sim Bernoulli(k_l) \\ logit(p_l) = \theta^{(3)}_{1,g[l]} + \theta^{(2)}_{1,j[l]} - \beta_{i[l]}, & logit(k_l) = \theta^{(3)}_{2,g[l]} + \theta^{(2)}_{2,j[l]} - \delta_{i[l]}. \end{array} \tag{4.5}$$

dove:

- $\underline{\theta}^{(2)}_j \sim BVN(\underline{\gamma}, T_{\theta^{(2)}})$ è l'abilita dello studente j, annidato nel gruppo g, ed è un effetto random;
- $\underline{\theta}^{(3)}_g \sim BVN(\underline{0}, T_{\theta^{(3)}})$ è l'effetto random per il gruppo g;
- β_i e δ_i sono effetti fissi e rappresentano i parametri di difficoltà, i primi per il modello di misura su $\mathbf{Y}$, i secondi per il modello di misura su $\mathbf{D}$. Per entrambi si assume che:

$$\beta_i \sim N(0, 0.0001)$$
$$\delta_i \sim N(0, 0.0001)$$

[3]Si precisa che, nel caso dei modelli multilivello, molti seguono il suggerimento di A. Gelman di considerare distribuzioni *a priori* uniformi nell'intervallo (0,100) per la varianza.

[4]Si sottolinea un cambio di notazione rispetto ai simboli nelle Figure 3.2 e 3.3; la dimensione latente ξ viene qui sostituita dal simbolo θ_2, il cui pedice identifica la seconda dimensione latente. Tale cambiamento è dovuto alla necessità di rappresentare la relazione tra le due dimensioni mediante un unico nodo bivariato.

Anche per il vettore delle medie $\underline{\gamma}$ si è stabilito di considerare una distribuzione *a priori* normale di parametri:

$$\underline{\gamma} \sim BVN(\underline{0}, I_{2\times 2}) \tag{4.6}$$

Per l'inverso delle due matrici di varianza e covarianza è stata utilizzata la distribuzione di $Wishart$[5]:

$$T_{\theta^{(3)}} \sim Wishart(R, 2), con\ R = I_{2\times 2} \tag{4.7}$$
$$T_{\theta^{(2)}} \sim Wishart(S, 2), con\ S = I_{2\times 2} \tag{4.8}$$

Le matrici R ed S, così come indicato dagli autori del software Winbugs, vanno intese come una sorta di *prior guess* circa la magnitudo (l'ordine di grandezza) della matrice di varianza e covarianza.
Il secondo parametro rappresenta i gradi di libertà e definisce il livello di non informatività della distribuzione. Nel caso specifico, attribuire due gradi di libertà ad una matrice di ordine 2×2 equivale a definire una distribuzione *a priori* meno informativa possibile.

4.4 Identificabilità

I modelli, descritti nei precedenti paragrafi, presentano problemi di identificabilità. Per risolvere tale inconveniente, si ricorre all'approccio proposto da Chaimongkol, Huffer e Kamata (2007), i quali a loro volta fanno riferimento a Bafumi et al. (2005). Tale metodo stabilisce di lasciare liberi i parametri originari del modello e di introdurre nuove quantità (adjusted) che risultano essere ben identificate e che verranno, poi, utilizzate al posto dei parametri originari.
In relazione al modello unidimensionale, il logit viene ridefinito come segue:

$$logit(p_l) = \theta^{adj,(3)}_{g[l]} + \theta^{adj,(2)}_{j[l]} - \beta^{adj}_{i[l]} \tag{4.9}$$

dove:

$$\begin{aligned} \beta^{adj}_i &= \beta_i - \bar{\beta} \\ \theta^{adj,(3)}_g &= \theta^{(3)}_g - \bar{\theta}^{(3)} \\ \theta^{adj,(2)}_j &= \theta^{(2)}_j - \bar{\beta} + \bar{\theta}^{(3)} \end{aligned} \tag{4.10}$$

[5]Essa è considerata la versione multidimensionale della distribuzione *Gamma*.

In questo studio, si è reso necessario estendere tale parametrizzazione al caso bidimensionale, procedendo nel modo seguente:

$$logit(p_l) = \theta^{adj,(3)}_{1,g[l]} + \theta^{adj,(2)}_{1,j[l]} - \beta^{adj}_{i[l]} \tag{4.11}$$

$$logit(k_l) = \theta^{adj,(3)}_{2,g[l]} + \theta^{adj,(2)}_{2,j[l]} - \delta^{adj}_{i[l]}. \tag{4.12}$$

dove

$$\begin{aligned}
\beta^{adj}_i &= \beta_i - \bar{\beta} \\
\delta^{adj}_i &= \delta_i - \bar{\delta} \\
\theta^{adj,(3)}_{1,g} &= \theta^{(3)}_{1,g} - \bar{\theta}^{(3)}_1 \\
\theta^{adj,(3)}_{2,g} &= \theta^{(3)}_{2,g} - \bar{\theta}^{(3)}_2 \\
\theta^{adj,(2)}_{1,j} &= \theta^{(2)}_{1,j} - \bar{\beta} + \bar{\theta}^{(3)}_1 \\
\theta^{adj,(2)}_{2,j} &= \theta^{(2)}_{2,j} - \bar{\delta} + \bar{\theta}^{(3)}_2
\end{aligned} \tag{4.13}$$

4.5 I modelli grafici e la loro rappresentazione

Il software di riferimento per questo lavoro, Openbugs, consente di definire i modelli descritti sia sotto forma di codice Bugs sia sotto forma di modello grafico. Le due strategie sono equivalenti e ciò è reso ancora più evidente dal fatto che il software permette all'utente di generare, a partire dal modello grafico, il codice Bugs corrispondente, come se fosse stato programmato dall'utente.
Come consigliato dagli stessi autori del programma, è sempre bene, una volta definito il modello grafico, accertarsi che sia esatto verificando se il codice Bugs, ottenuto a partire da esso, sia corretto.
È possibile, ad esempio, costruire un modello grafico di riferimento e, successivamente, decidere di complicarlo inserendo ulteriori righe di comando nel codice da esso originato. Tutto ciò implica che alcune caratteristiche tipiche del linguaggio utilizzato si riflettano anche in fase di definizione del modello grafico.
Dopo la premessa analitica sui modelli in questione, è possibile, a questo punto, passare alla loro rappresentazione grafica.
Per i due modelli unidimensionali, il corrispondente modello grafico è rappresentato in Figura 4.1. Il codice ottenuto a partire da esso è riportato nell'appendice A.1.
Per il modello bidimensionale, il corrispondente modello grafico è riportato in Figura 4.2. Il codice ottenuto a partire da esso è riportato nell'appendice A.2.

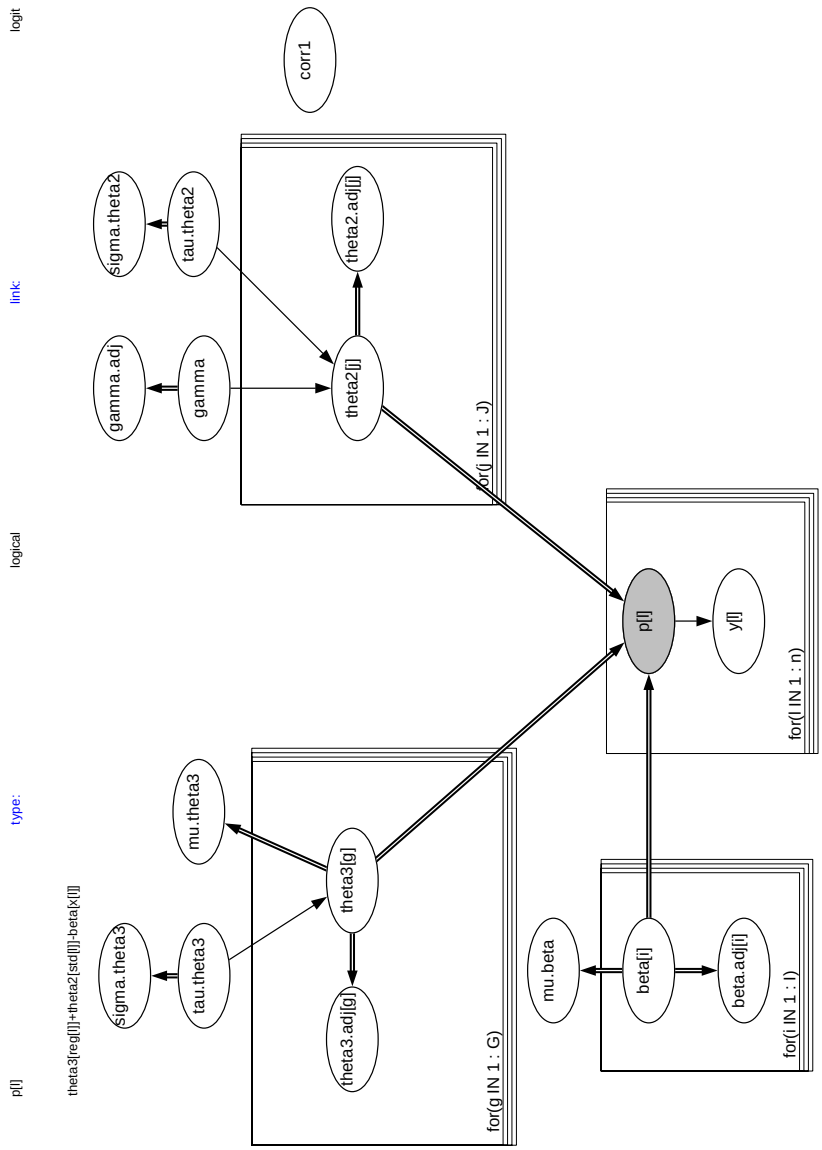


Figura 4.1: Modello grafico unidimensionale

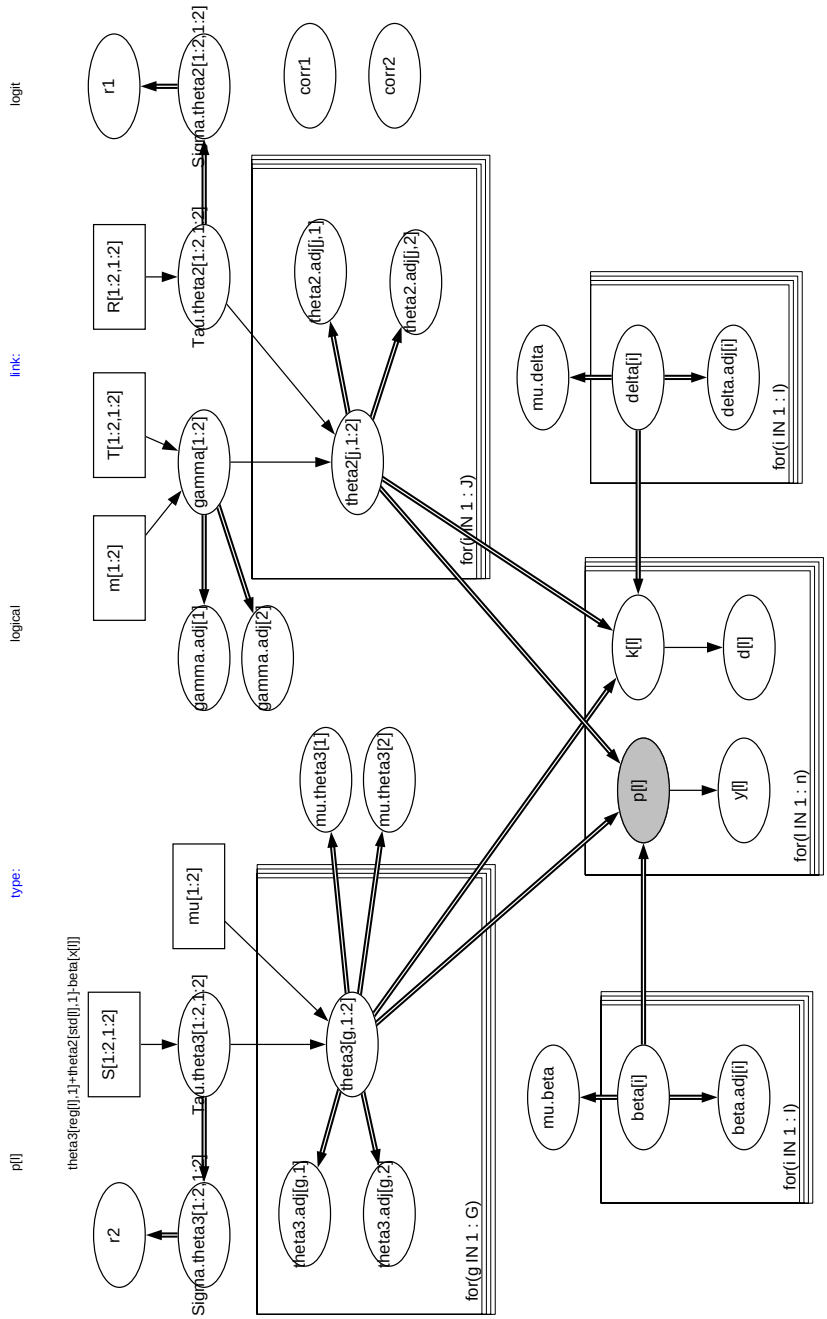


Figura 4.2: Modello grafico bidimensionale

Il nodo del logit, in grigio sia in Figura 4.1 che in Figura 4.2, è stato evidenziato di proposito in maniera tale da far comparire, nella parte alta del grafico, l'esplicitazione del nodo stesso.
Ad uno sguardo attento non sarà sfuggito l'introduzione nel logit degli indici "reg[l]", "std[l]" e "x[l]": essi vengono forniti in input, insieme alla matrice dei dati, al fine di consentire l'annidamento degli indici (i, j, g) così come anticipato nella trattazione analitica.

4.6 Decomposizione della varianza

Come più volte detto, individui che appartengono allo stesso gruppo sono più simili tra loro rispetto ad individui che appartengono a cluster differenti. Ciò implica che la variabilità nelle risposte sia dovuta e alle differenze "nei" gruppi e alle differenze "tra" i gruppi.
A tale proposito (un esempio è in Raudenbush, Johnson e Sampson (2003)) diventa interessante studiare non solo la correlazione tra le dimensioni (sia a livello studente che a livello di gruppo) ma, come da prassi nelle analisi multilivello, diventa utile calcolare un coefficiente noto con l'acronimo ICC[6], ossia *IntraClass Correlation Coefficient*, che calcola la quota di variabilità dovuta ai cluster.
In generale, per il modello bidimensionale, si hanno due ICC, una per ogni dimensione latente:

$$ICC_1 = \frac{\sigma^2_{1,\theta^{(3)}}}{(\sigma^2_{1,\theta^{(3)}} + \sigma^2_{1,\theta^{(2)}})} \tag{4.14}$$

$$ICC_2 = \frac{\sigma^2_{2,\theta^{(3)}}}{(\sigma^2_{2,\theta^{(3)}} + \sigma^2_{2,\theta^{(2)}})} \tag{4.15}$$

$\sigma^2_{1,\theta^{(3)}}$ indica la varianza relativa alla prima dimensione (l'abilità) a livello "gruppo"; $\sigma^2_{1,\theta^{(2)}}$ indica la varianza sempre relativa alla prima dimensione ma a livello "studente".
$\sigma^2_{2,\theta^{(3)}}$ indica la varianza relativa alla seconda dimensione (la propensione alla risposta) a livello "gruppo"; $\sigma^2_{2,\theta^{(2)}}$ indica la varianza sempre relativa alla seconda dimensione ma a livello "studente".
Come anticipato in questo paragrafo, importante ai fini dell'analisi è valutare il grado di correlazione tra le dimensioni latenti, per cui si introducono le

[6] L'acronimo è analogo a quello utilizzato per definire le Item Characteristic Curves ma non va confuso.

seguenti correlazioni:

$$\rho_{student} = \frac{\sigma^2_{12,\theta^{(2)}}}{\sqrt{\sigma^2_{1,\theta^{(2)}}\sigma^2_{2,\theta^{(2)}}}} \tag{4.16}$$

$$\rho_{group} = \frac{\sigma^2_{12,\theta^{(3)}}}{\sqrt{\sigma^2_{1,\theta^{(3)}}\sigma^2_{2,\theta^{(3)}}}} \tag{4.17}$$

$\sigma^2_{12,\theta^{(2)}}$ indica la covarianza tra le due dimensioni a livello "studente", mentre $\sigma^2_{12,\theta^{(3)}}$ indica la covarianza tra le due dimensioni a livello "gruppo".
Le correlazioni $\rho_{student}$ e ρ_{group} sono oggetto di particolare attenzione; qualora risultino statisticamente diverse da zero, si ha conferma del fatto che esiste una relazione tra abilità e propensione alla risposta.
Quanto più elevata è la loro magnitudo, tanto più consistente è la supposta deviazione dall'ipotesi di ignorabilità. La distorsione nelle stime dei parametri di interesse è direttamente proporzionale al numero di *missing* presenti nel dataset e al grado di correlazione tra le dimensioni latenti.
I quattro indici introdotti sono presenti, ovviamente, anche nel modello grafico: $\rho_{student}$ e ρ_{group} sono rappresentati dai nodi $r1$ ed $r2$ (presenti solo nel modello bidimensionale), mentre ICC_1 e ICC_2 sono rappresentati dai nodi $corr1$ e $corr2$.

4.7 Introduzione delle covariate nel modello bidimensionale

Come si avrà modo di approfondire nel capitolo successivo, è possibile inserire le covariate nei modelli descritti al fine di comprendere quali siano le caratteristiche che influenzano, in maniera significativa, il processo latente dell'abilità e quello della propensione alla risposta.
In relazione al solo modello bidimensionale, si è stabilito di introdurre le medesime covariate, a livello "studente", per entrambe le dimensioni. L'obiettivo è appunto verificare quale sia il ruolo di alcune caratteristiche individuali, in letteratura considerate rilevanti, per entrambi i processi.
A partire dal codice generato dal modello grafico, è possibile estendere lo stesso introducendo le covariate a livello studente: supponendo di considerare, ad esempio, due covariate $zeta1$ e $zeta2$, si ha la seguente modifica del

codice per il nodo bivariato $\underline{\gamma}$[7]:

$$gamma[j,1] \leftarrow b1[1] * zeta1[j] + b2[1] * zeta2[j] \tag{4.18}$$
$$gamma[j,2] \leftarrow b1[2] * zeta1[j] + b2[2] * zeta2[j] \tag{4.19}$$

Per ciascun regressore introdotto, la distribuzione *a priori* adottata è una $N(0, 0.0001)$.

4.8 DIfferential Item Functioning (DIF)

È ormai chiaro che, quando si utilizza un test per misurare l'abilità latente, condizione necessaria è che esso sia invariante rispetto alle caratteristiche della popolazione. La violazione di tale assunto è nota in letteratura come *Differential Item Functioning* (DIF) o come *item bias*.
Esso sarà oggetto di investigazione anche in questo studio, perciò diventa necessario approfondire la letteratura in merito.
Il DIF si manifesta quando, condizionatamente ad un certo livello di abilità, la probabilità di risposta corretta ad un item differisce in maniera sistematica tra gruppi di individui, omogenei secondo una certa caratteristica (di solito il sesso o l'appartenenza ad un gruppo etnico). Un item è considerato distorto, quindi, solo se il suo livello di difficoltà cambia tra individui che hanno lo stesso livello di abilità, ma che appartengono a gruppi differenti.
In riferimento a due gruppi, si è soliti indicare uno di essi con il termine di *reference group* e l'altro con il termine di *focal group*. È pratica diffusa designare, come *reference group*, il gruppo per il quale si immagina che il test sia vantaggioso.
Si distingue tra due diverse forme di DIF: *uniforme* e *non uniforme*.
Il DIF *uniforme* si realizza quando la probabilità di una risposta corretta da parte di un sottogruppo è sistematicamente maggiore o minore della medesima probabilità da parte dell'altro sottogruppo (di solito la sottopopolazione di riferimento), per tutti i livelli di abilità.
Il DIF *non uniforme* si realizza quando la probabilità di una risposta corretta per il *focal group* è maggiore di quella del *reference group* in corrispondenza dei livelli più bassi di abilità e minore per i livelli più alti, o viceversa. Due esempi a riguardo sono riportati in Figura 4.3 e 4.4.
Analizzando le ICC rappresentate è evidente che, in caso di DIF *uniforme*, le posizioni sono differenti ma la pendenza è la stessa; in caso di DIF *non*

[7] L'utilizzo del linguaggio Bugs invece del modello grafico è dettato dalla necessità di esplicitare il nodo bivariato $\underline{\gamma}$ nelle sue due componenti.

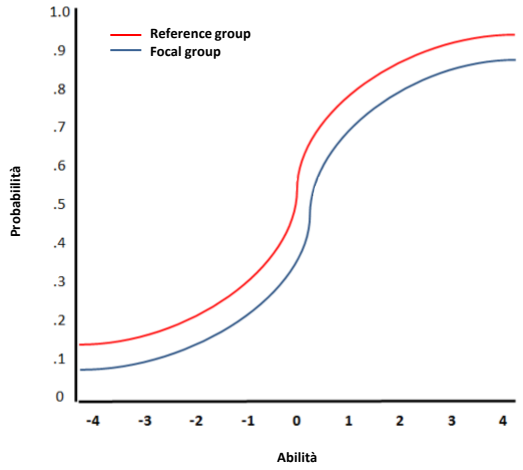


Figura 4.3: Esempio di DIF unifome

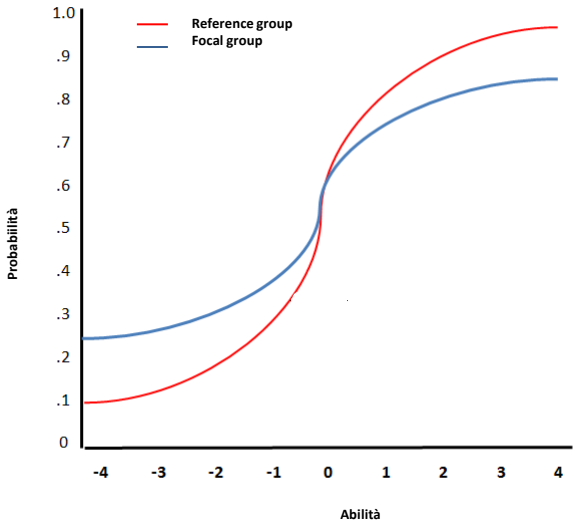


Figura 4.4: Esempio di DIF non unifome

uniforme, un item può simultaneamente variare sia in pendenza che in difficoltà, relativamente ai due gruppi.
Più in generale, quando le curve sono differenti, ma non si intersecano, siamo nel caso di DIF *uniforme*; in caso contrario, siamo in presenza di DIF *non uniforme*.
Le procedure più comuni per l'identificazione del DIF sono quelle di Mantel-Haenszel, dei modelli IRT e quelle basate sulla regressione logistica; i prossimi paragrafi ne descrivono gli aspetti essenziali per ognuno di essi.

4.8.1 La procedura di Mantel-Haenszel

La procedura di Mantel e Haenszel (1959), utilizzata per testare la dipendenza tra due variabili in una tabella di contingenza a tre vie, è stata modificata da Holland e Thayer (1988) per scopi di identificazione del DIF per items dicotomici.
L'abilità latente, rappresentata attraverso i punteggi totali, viene divisa in S intervalli al fine di definire un criterio di *matching* fra *reference* e *focal group*.
La procedura in questione valuta le differenze tra i due gruppi per tutti gli item, uno alla volta e per ogni intervallo $s = 1, \ldots, S$: ciò implica che ogni tabella di contingenza è del tipo 2 (gruppi) $\times$ 2 (risposta dicotomica) $\times$ s (punteggio).
Va sottolineato che la procedura prevede l'eliminazione degli items per i quali si hanno tutte risposte corrette o tutte sbagliate: essi sono troppo semplici o troppo difficili per tutti i rispondenti, indipendentemente dal gruppo di appartenenza, e non apportano informazioni utili ai fini dell'analisi.
Per un generico item, considerato l'$s - esimo$ intervallo di abilità, la tabella di contingenza si mostra così come definita nella Tabella 4.1.
Le righe corrispondono alla tipologia di gruppo di appartenenza (*reference* o *focal*) mentre le colonne corrispondono alle modalità di risposta (corretta o sbagliata).
È chiaro che le celle contengono le frequenze delle risposte in riferimento ai due gruppi, oltre alle marginali di riga, di colonna e alla frequenza totale.
In particolare A_s e B_s rappresentano, rispettivamente, le frequenze delle risposte corrette e non corrette per il gruppo di riferimento, mentre C_s e D_s rappresentano le medesime frequenze ma in riferimento al gruppo di studio.
n_{R_s} e n_{F_s} rappresentano, rispettivamente, il numero totale di persone nel gruppo di riferimento e in quello di studio, considerato il punteggio s.
n_{1_s} e n_{0_s} rappresentano il numero totale di risposte corrette e il numero totale di risposte sbagliate, sempre fissato il livello s.
N_s è, infine, il numero totale di persone che hanno ottenuto un punteggio totale uguale ad s.

Tabella 4.1: Tabella di contingenza M-H

Gruppo	corretta(1)	sbagliata(0)	Totale
Reference	A_s	B_s	n_{R_s}
focal	C_s	D_s	n_{F_s}
Totale	n_{1_s}	n_{0_s}	N_s

Il passo successivo consiste nel calcolare l'odd ratio stimato per l'$i-esimo$ item, ossia:

$$\widehat{\alpha}_{MH_i} = \frac{\sum_s^S A_s D_s / N_s}{\sum_s^S B_s C_s / N_s} \tag{4.20}$$

$\widehat{\alpha}_{MH_i}$ può assumere valori da 0 a $+\infty$ e quando si verifica $\widehat{\alpha}_{MH_i} = 1$ vi è assenza di DIF.
Valori dell'indice maggiori di 1 suggeriscono che il gruppo di riferimento performa meglio o trova l'item meno difficile rispetto all'altro gruppo. Al contrario, valori minori di 1 indicano che l'item è più facile per il gruppo di studio.
Tale indice viene proposto anche nella sua versione logaritmica, ossia:

$$\widehat{\beta}_{MH_i} = ln(\widehat{\alpha}_{MH_i}) \tag{4.21}$$

Questa seconda formulazione è particolarmente utile quando si vogliono effettuare delle comparazioni tra la procedura M-H e la regressione logistica, poiché l'indice così definito è equivalente a un generico coefficiente di regressione logistica.
È ovvio che $\widehat{\alpha}_{MH_i} = 1 \Rightarrow \widehat{\beta}_{MH_i} = 0$ e quindi indica assenza di DIF.
$\widehat{\alpha}_{MH_i} < 1 \Rightarrow \widehat{\beta}_{MH_i} < 0$ e quindi indica una distorsione a vantaggio del *focal group*.
$\widehat{\alpha}_{MH_i} > 1 \Rightarrow \widehat{\beta}_{MH_i} > 0$ e quindi indica una distorsione a vantaggio del *reference group*.
L'indice logaritmico può essere ulteriormente trasformato nella statistica MH-DIF$\widehat{D}_i$ al fine di valutare la magnitudo del DIF (Holland e Thayer 1985). Si ha quindi:

$$\widehat{D}_i = -2.35\widehat{\beta}_{MH_i} \tag{4.22}$$

Se $\widehat{D}_i$ assume valori positivi si ha indicazione di una possibile distorsione a vantaggio del gruppo di studio; in caso di valore negativo, la distorsione sarà a vantaggio del gruppo di riferimento.
La procedura di M-H è molto utilizzata nella pratica poiché ha il vantaggio di essere semplice da applicare; le critiche rivolte al metodo, invece, fanno

riferimento al fatto che non sia valido per investigare DIF *non uniformi* e al fatto che esso usi il punteggio totale come sostituto del tratto latente. In modelli IRT complessi, lo score osservato non è detto che sia il miglior surrogato della corrispondente abilità latente.

4.8.2 La regressione logistica

La procedura che fa riferimento alla regressione logistica è stata introdotta da Swaminathan e Rogers (1990) al fine di identificare la presenza di DIF in items dicotomici. Tale metodo utilizza, come variabile dipendente, la risposta all'item e come variabili indipendenti: il criterio di *matching* per le abilità (di solito il punteggio totale), il gruppo di appartenenza e l'interazione tra gruppo e abilità osservata.
La presenza del DIF è verificata testando la bontà di adattamento del modello che considera la variabile di appartenenza al gruppo e il termine di interazione.
Entrando nel dettaglio, per il soggetto $j - esimo$, si ha:

$$p_j \equiv P(Y_j = 1|\theta) = \frac{\exp(\eta)}{1 + exp(\eta)} \tag{4.23}$$

con $\eta = logit(p_j)$.
I tre modelli di interesse diventano:

$$1.\ \eta = \beta_0 + \beta_1\theta_j + \beta_2 G_j + \beta_3(\theta G)_j \tag{4.24}$$
$$2.\ \eta = \beta_0 + \beta_1\theta_j + \beta_2 G_j \tag{4.25}$$
$$3.\ \eta = \beta_0 + \beta_1\theta_j \tag{4.26}$$

θ_j rappresenta l'abilità osservata (anche detta *total score*), G_j la variabile indicatrice del gruppo di appartenenza e $(\theta G)_j$ l'effetto di interazione.
In caso di item non distorto, solo β_0 e β_1 devono essere diversi da zero; in caso di item con DIF *uniforme* si ha, invece, che $\beta_2 \neq 0$ ma $\beta_3 = 0$.
In caso di item con DIF *non uniforme* $\beta_3 \neq 0$, sia se β_2 è nullo o meno.
Il primo modello consente di testare sia la presenza di DIF *uniforme* che *non uniforme*, il secondo solo quello *uniforme* ed il terzo (il modello nullo) l'assenza di DIF.
Per ognuno dei tre modelli, la statistica Chi-quadrato viene calcolata a partire dal rapporto di log verosimiglianza. Per testare simultaneamente la presenza di DIF *uniforme* e\o *non uniforme*, basta confrontare il modello 3 con il modello 1 mediante la statistica G^2. Come è noto, essa è ancora una distribuzione Chi-quadrato con 2 gradi di libertà. Se il test non risulta statisticamente significativo, gli items non presentano DIF.

Se si confronta tale metodo con la procedura di M-H, si notano delle similarità. Entrambi ricorrono al punteggio totale come criterio di *matching*: va ricordato che non è sempre certo che i punteggi rappresentino bene la vera abilità latente se il numero di osservazioni per ogni soggetto è limitato. Swaminathan e Rogers (1990), a tale proposito, stabiliscono che entrambe le procedure restituiscono risultati migliori, oltre che simili, in caso di test con molti items e campioni elevati.
La procedura M-H è, inoltre, simile alla regressione logistica (per il DIF) quando, per quest'ultima, si assume la variabile dell'abilità discreta e non si inserisce il termine di interazione.
Va precisato che la procedura legata alla regressione logistica è, però, più flessibile poiché consente di inserire anche più di un criterio di *matching*, oltre che poter essere estesa al caso di items politomici.

4.8.3 Item response Theory

Quando i modelli dell'Item Response Theory vengono utilizzati per la valutazione del DIF, le differenze tra i parametri degli items (in relazione ai gruppi e ai parametri di abilità) diventano oggetto di interesse.
In generale, i metodi relativi a questa sezione si basano sul confronto delle ICC (*Item Characteristic Curves*) tra i gruppi testando, in un certo senso, la validità della proprietà di invarianza dei parametri, tipica dei modelli IRT.
Gli approcci implementati sono diversi: alcuni fanno riferimento all'area compresa tra le ICC (Raju 1988), altri testano l'uguaglianza tra i parametri delle ICC (Lord 1980), altri ancora effettuano test statistici sulla bontà di adattamento dei modelli (Thissen, Steinberg e Wainer 1988).
In riferimento ai primi due approcci, è necessario che gli item vengano calibrati, separatamente, per ciascuno dei due gruppi e, successivamente, trasformati mediante opportune procedure di "*linking*".
Proprio queste ultime assumono l'assenza di DIF per gli items utilizzati allo scopo: tale caratteristica, però, in genere, non è nota a priori.
Al fine di individuare il sottoinsieme di items non distorti con cui effettuare le operazioni di "*linking*", è necessaria una fase preliminare di selezione degli items con DIF.
Terminate le operazioni di "*linking*", è possibile procedere con il calcolo dell'area compresa tra le ICC, se si considera il primo approccio, o con il confronto dei parametri degli items di entrambi i gruppi, se si considera il secondo.
Il terzo approccio utilizza il test del rapporto di logverosimiglianza (G^2); prima di eseguire il test statistico, anche in questo caso, è necessaria una fase preliminare in cui viene individuato il sottoinsieme di items non distorti (detti *anchor items*). Ciò consente, nel momento successivo, di definire due

gruppi di items: il primo costituito dagli *anchor items* e il secondo dagli items di studio. Va sottolineato che, per il primo sottoinsieme, il vincolo imposto sui parametri è che siano uguali tra i gruppi.
A questo punto, il modello IRT di riferimento può essere applicato ad un particolare sottoinsieme, formato dagli items ancorati più un item di studio. Il modello viene applicato due volte, la prima non considerando il vincolo di uguaglianza (modello con DIF), la seconda considerando valido tale vincolo (modello per l'assenza di DIF).
Indicando con *model b* il modello senza vincolo di uguaglianza e con *model a* quello col vincolo, l'ipotesi di assenza di DIF (e quindi di invarianza dell'item) viene testata valutando:

$$G^2 = -2ln\frac{L(model\ a)}{L(model\ b)} = 2[\ell_{(b)} - \ell_{(a)}] \tag{4.27}$$

Il test è ripetuto per tutti gli items considerati sospetti.
Va specificato che i metodi sin qui introdotti vengono anche classificati come metodi IRT parametrici, per distinguerli da un altro approccio che non fa alcuna assunzione sulla forma funzionale delle ICC e che, per questo, è detto non parametrico (si veda SIBTEST).

4.8.4 I modelli multilivello per l'analisi del DIF

Prima di introdurre il modello gerarchico utilizzato nel presente lavoro, sembra interessante mostrare l'estensione del modello di Kamata (2001), introdotto nel secondo capitolo, per l'analisi del DIF.
Concentrando l'attenzione sul modello a due livelli gerarchici, per il primo livello non si hanno variazioni rispetto a quanto già visto, per il secondo livello invece si ha:

$$\begin{cases} \beta_{0j} = \gamma_{00} + u_{0j} \\ \beta_{1j} = \gamma_{10} + \gamma_{11}(Group)_j \\ \beta_{2j} = \gamma_{20} + \gamma_{21}(Group)_j \\ \beta_{3j} = \gamma_{30} \vdots \\ \beta_{(k-1)j} = \gamma_{(I-1)0} \end{cases}$$

In tale formulazione emerge che, mentre per gli item non distorti (quelli dal terzo in poi), vi è solo il termine di intercetta, per i due items distorti vi è anche la pendenza.
Qualora quest'ultima risulti statisticamente significativa (diversa da zero), si è in presenza di un item distorto. Va sottolineato che la variabile $(Group)_j$ può essere inserita anche per il termine d'intercetta al fine di valutare l'effetto

principale della variabile sulla probabilità di risposta.
L'analisi del DIF, per questi modelli, può applicarsi anche al caso di un modello lineare con tre livelli gerarchici, qualora si fosse interessati a vedere come il DIF vari tra i gruppi[8] (Kamata e Binici 2003).

Per quanto concerne i modelli sviluppati in questo capitolo, è possibile considerare una loro estensione all'analisi del DIF; il modello di riferimento rimane sempre:

$$\nu = \mathbf{x}'\boldsymbol{\beta} + \sum_{l=2}^{L} \mathbf{z}^{(l)'}\mathbf{u}^{(l)} \tag{4.28}$$

con $\mathbf{x}$ vettore che include anche l'intercetta; come mostrato in Chaimongkol (2005), esso non è altro che un modello di regressione logistica a più livelli.
In tale ottica, il modello di Rasch gerarchico può essere esteso inserendo ulteriori parametri fissi per l'analisi del DIF. Ciò implica che l'equazione(2.21) del secondo capitolo, ristretta a due soli livelli gerarchici, diventa:

$$log\left[\frac{p_{ij}}{(1-p_{ij})}\right] = \nu_{ij} = \beta_i + \alpha\, group_j + \lambda_i\, group_j + \theta_j^{(2)} \tag{4.29}$$

dove $group_j$ è una variabile dummy tale che:

$$group_j = \begin{cases} 1 & \text{se il soggetto } j \text{ appartiene al focal group} \\ 0 & \text{altrimenti} \end{cases} \tag{4.30}$$

Gli ulteriori due effetti fissi introdotti hanno il seguente significato:

- α rappresenta propriamente la differenza tra l'abilità del gruppo di studio e quella del gruppo di riferimento;
- λ_i è invece l'effetto di interazione tra item e caratteristiche del soggetto, ossia la magnitudo del DIF per l'$i-esimo$ item.

Dopo questa premessa, al fine di valutare l'impatto dei dati mancanti anche nell'analisi del DIF, non resta che passare alla definizione dei modelli grafici di riferimento.
Si procede mostrando il modello grafico per il DIF, sia nel caso unidimensionale che bidimensionale, questa volta considerando soltanto due livelli gerarchici, quindi solo l'effetto random a livello studente.

[8]Per *gruppi* qui si fa riferimento ai gruppi che compongono il terzo livello gerarchico. Non vanno confusi con *reference* e *focal group*.

4.9 Il modello DIF unidimensionale

In riferimento al modello unidimensionale a due livelli, la modifica è la seguente:

$$y_l \sim Bernoulli(p_l) \tag{4.31}$$

$$logit(p_l) = \theta^{(2)}_{j[l]} + \alpha\, group_{j[l]} - \beta_{i[l]} - \lambda_{i[l]}\, group_{j[l]} \tag{4.32}$$

dove:

- $\theta^{(2)}_j$ rappresenta l'abilità del soggetto j, è un effetto random e, per esso, si assume che:
$$\theta^{(2)}_j \sim N(\gamma, \tau_{\theta^{(2)}})$$
- β_i rappresenta l'effetto fisso, quindi il parametro di difficoltà per l'item i per il gruppo di riferimento e, per esso, si assume che:
$$\beta_i \sim N(0, 0.0001)$$
- α è un effetto fisso, è la differenza tra l'abilità media del *focal group* e quella del *reference group*;
- $group_j$ è la variabile dummy che assume valore uno per il *focal group*, zero altrimenti;
- λ_i è un effetto fisso e indica propriamente il DIF per l'item i. Il segno negativo che lo precede nel logit è motivato dal fatto che, in caso di valore positivo di λ_i, esso si interpreti come segnale di distorsione dell'item a svantaggio del *focal group*.

Per l'effetto fisso γ vale sempre che:

$$\gamma \sim N(0, 1)$$

Per il parametro di precisione si ha ancora:

$$\tau_{\theta^{(2)}} \sim Gamma(0.001, 0.001)$$

Per entrambi i nuovi parametri, α e λ_i, la distribuzione a priori scelta è una $N(0, 0.0001)$.
Anche per il modello in questione sussistono i medesimi problemi di identificabilità già visti nel paragrafo (4.2.2); in questo caso il logit viene ridefinito come segue:

$$logit(p_l) = \theta^{adj,(2)}_{j[l]} + \alpha^{adj}\, group_{j[l]} - \beta^{adj}_{i[l]} - \lambda^{adj}_{i[l]}\, group_{j[l]} \tag{4.33}$$

dove:

$$\begin{aligned}
\beta_i^{adj} &= \beta_i - \bar{\beta} \\
\theta_j^{adj,(2)} &= \theta_j^{(2)} - \bar{\beta} \\
\alpha^{adj} &= \alpha - \bar{\lambda} \\
\lambda_i^{adj} &= \lambda_i - \bar{\lambda}
\end{aligned} \tag{4.34}$$

Il modello grafico di riferimento è rappresentato in Figura 4.5[9]. Il codice ottenuto a partire da esso è riportato nell'appendice A.3.

4.10 Il modello DIF bidimensionale

L'estensione del modello bidimensionale all'analisi del DIF è di facile derivazione; semplificando ad un modello gerarchico a due livelli, si ottiene:

Per la matrice **Y**

$$\begin{aligned}
y_l &\sim Bernoulli(p_l) \\
logit(p_l) &= \theta_{1,j[l]}^{(2)} + \alpha_1\, group_{j[l]} - \beta_{i[l]} - \lambda_{i[l]}\, group_{j[l]}
\end{aligned} \tag{4.35}$$

Per la matrice **D**

$$\begin{aligned}
d_l &\sim Bernoulli(k_l) \\
logit(k_l) &= \theta_{2,j[l]}^{(2)} + \alpha_2\, group_{j[l]} - \delta_{i[l]} - \nu_{i[l]}\, group_{j[l]} \qquad (4.36) \\
& \qquad (4.37)
\end{aligned}$$

Il significato dei parametri è sempre il medesimo, anche in relazione ai nuovi parametri $\underline{\alpha}$, λ_i e ν_i.
Per i parametri di difficoltà vale sempre che:

$$\beta_i \sim N(0, 0.0001) \tag{4.38}$$
$$\delta_i \sim N(0, 0.0001) \tag{4.39}$$

Ad essi si aggiungono:

$$\underline{\alpha} \sim BVN(\underline{0}, I_{2\times 2}) \tag{4.40}$$
$$\lambda_i \sim N(0, 0.0001) \tag{4.41}$$
$$\nu_i \sim N(0, 0.0001) \tag{4.42}$$

[9] I nodi p_a e $pv[i]$ sono inseriti nel modello per il calcolo dei P-value.

Per il vettore delle medie $\underline{\gamma}$ vale ancora:

$$\underline{\gamma} \sim BVN(\underline{0}, I_{2\times 2}) \tag{4.43}$$

Per l'inverso della matrice di varianza e covarianza vale sempre:

$$T_{\theta^{(2)}} \sim Wishart(S, 2), con\ S = I_{2\times 2} \tag{4.44}$$

Visti i problemi di indentificabilità, complicati dall'introduzione di nuovi parametri, si ha che:

$$logit(p_l) = \theta^{adj,(2)}_{1,j[l]} + \alpha^{adj}_1\, group_{j[l]} - \beta^{adj}_{i[l]} - \lambda^{adj}_i\, group_{j[l]} \tag{4.45}$$

$$logit(k_l) = \theta^{adj,(2)}_{2,j[l]} + \alpha^{adj}_2\, group_{j[l]} - \delta^{adj}_{i[l]} - \nu^{adj}_i\, group_{j[l]} \tag{4.46}$$

dove

$$\begin{aligned}
\beta^{adj}_i &= \beta_i - \bar{\beta} \\
\delta^{adj}_i &= \delta_i - \bar{\delta} \\
\alpha^{adj}_1 &= \alpha - \bar{\lambda} \\
\alpha^{adj}_2 &= \alpha - \bar{\nu} \\
\lambda^{adj}_i &= \lambda_i - \bar{\lambda} \\
\nu^{adj}_i &= \nu_i - \bar{\nu} \\
\theta^{adj,(2)}_{1,j} &= \theta^{(2)}_{1,j} - \bar{\beta} \\
\theta^{adj,(2)}_{2,j} &= \theta^{(2)}_{2,j} - \bar{\delta}
\end{aligned}$$

Il modello grafico di riferimento è rappresentato in Figura 4.6[10] . Il codice ottenuto a partire da esso è riportato nell'appendice A.4.

Dopo aver introdotto i modelli di riferimento, a questo punto, non resta che procedere con l'applicazione di questi ultimi ai dati reali, in particolare ad un sottoinsieme dei dati dell'indagine PISA 2006.
L'analisi dei risultati ottenuti sarà oggetto del prossimo capitolo.

[10] I nodi p_a e $pv[i]$ sono inseriti nel modello per il calcolo del P-value.

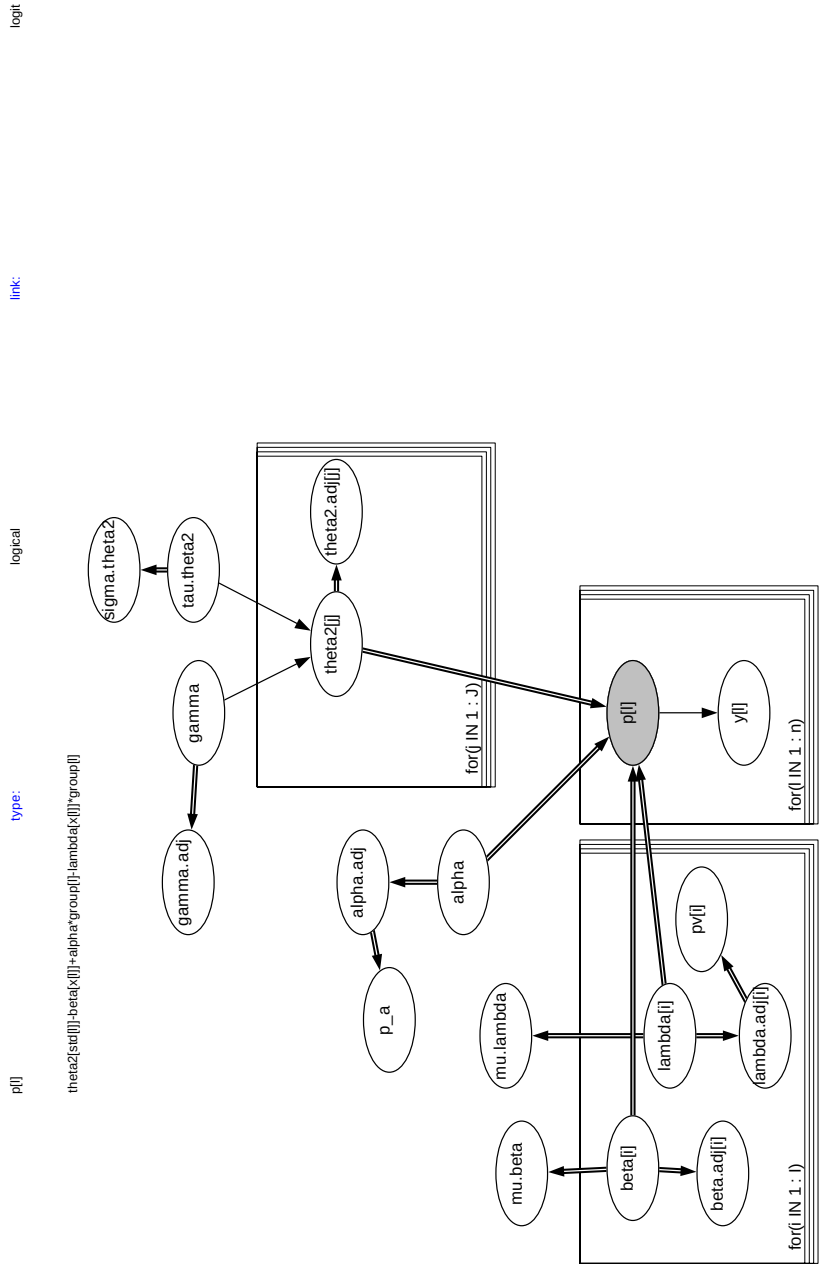


Figura 4.5: Modello grafico unidimensionale per il DIF

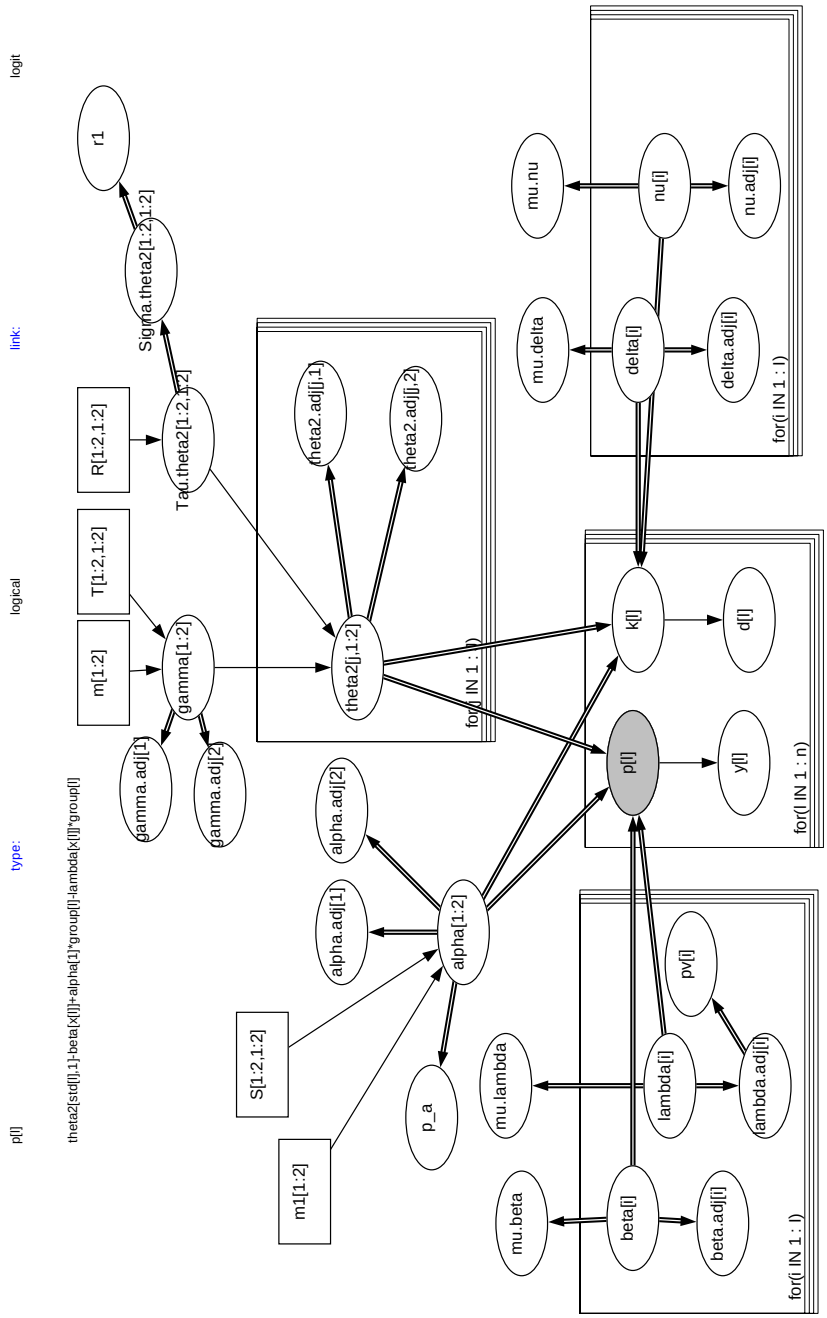


Figura 4.6: Modello grafico bidimensionale per il DIF

Capitolo 5

Applicazione del modello ai dati reali

5.1 Introduzione

Scopo del presente capitolo è l'applicazione dei modelli gerarchici ai dati reali, al fine di comparare le stime derivanti dai tre diversi approcci al trattamento dei dati mancanti.
L'applicazione riguarda un campione dei dati dell'indagine PISA[1] 2006. Tale scelta è motivata dalla considerazione secondo la quale, nel contesto educativo, ha senso supporre un pattern di dati mancanti tutt'altro che casuale; è lecito, insomma, ritenere che a maggiore abilità corrisponda una maggiore propensione alla risposta e che la distribuzione dei *missing data* nel campione dipenda dall'abilità stessa che si vuole stimare.
Prima di passare alla descrizione dettagliata del campione di riferimento per l'analisi, è assolutamente necessario dare un quadro generale dei risultati dell'indagine PISA 2006, con un focus sui dati italiani relativamente alla *literacy* matematica.
La comparazione dei sistemi scolastici di ben 57 Paesi del mondo mette in luce le peculiarità e le caratteristiche di ciascuno di essi: differenze demografiche e territoriali, contraddizioni interne a ciascun sistema scolastico, così come punti di forza e di debolezza, emergono con chiarezza ed evidenza dall'analisi di questa mole ingente di informazioni.

[1] *Programme for International Student Assessment.*

5.1.1 L'indagine PISA 2006: uno sguardo ai risultati italiani

PISA è un indagine internazionale promossa dall'OCSE[2] e realizzata in Italia dall'INVALSI[3], presso il quale è costituito il Centro nazionale PISA.
L'obiettivo è accertare le competenze dei quindicenni[4] scolarizzati nelle seguenti tre aree di apprendimento:

- Lettura
- Scienze
- Matematica

Il primo ciclo risale all'anno 2000 e la periodicità dell'indagine è triennale.
I ricercatori PISA hanno stabilito che, per ogni ciclo, ci sia un focus particolare su una delle tre aree: in PISA 2000 l'area principale d'indagine è stata la lettura, in PISA 2003 la matematica, in PISA 2006 le scienze, in PISA 2009 ancora la lettura mentre, in PISA 2012, la matematica e il problem solving.
I Paesi coinvolti sono cinquantasette, in particolare trenta di essi sono Paesi membri dell'OCSE, i restanti ventisette sono Paesi partner.[5]
Hanno partecipato all'indagine 400.000 studenti, campione rappresentativo di quasi 20 milioni di quindicenni scolarizzati. Per ciascun Paese, il numero di studenti sottoposti a test è stato variabile, fra le 4.500 e le 10.000 unità, appartenenti ad almeno 150 scuole.

I questionari Le prove sono state organizzate in 13 cluster (sette di scienze, due di lettura e quattro di matematica); questi, a loro volta, secondo uno schema di rotazione, sono stati distribuiti in 13 fascicoli o *booklet*, contenenti, ognuno, solo quattro cluster (col vincolo che ce ne fosse sempre almeno uno di scienze).
Ogni singolo test, su supporto cartaceo, ha richiesto un impegno di due ore

[2] *Organizzazione per la Cooperazione e lo Sviluppo Economico.*

[3] *Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e di Formazione.*

[4] Età che comprende, per la maggioranza dei Paesi, la fascia dell'istruzione obbligatoria.

[5] **Paesi OCSE**

Australia, Austria, Belgio, Canada, Corea, Danimarca, Finlandia, Francia, Germania, Giappone, Grecia, Irlanda, Islanda,Italia, Lussemburgo, Messico, Norvegia, Nuova Zelanda, Paesi Bassi, Polonia, Portogallo, Regno Unito, Rep. Ceca, Rep. Slovacca, Spagna, Stati Uniti, Svezia, Svizzera, Turchia, Ungheria.

Paesi partner

Argentina, Azerbaijan, Brasile, Bulgaria, Cile, Colombia, Croazia, Estonia, Giordania, Hong Kong-Cina, Indonesia, Israele, Kyrghizistan, Lettonia, Liechtenstein, Lituania, Macao-Cina, Montenegro, Qatar, Romania, Russia, Serbia, Slovenia, Taiwan-Cina, Thailandia, Tunisia, Uruguay.

per la sua compilazione. Il sistema descritto ha generato *missing da disegno* poiché, appunto, non tutti gli studenti hanno risposto a tutte le domande della rilevazione.
La prova cognitiva è stata strutturata in quesiti, sia in forma aperta che chiusa, organizzati per "unità"; queste ultime sono state strutturate in maniera tale da introdurre uno stimolo iniziale, composto in genere da un testo verbale (corredato da figure, grafici o tabelle), seguito da una serie di domande riferite ad esso.
Va ricordato che, non solo i fascicoli sono stati assegnati in maniera casuale a ciascuno studente, ma sono stati somministrati a un numero sufficiente a garantire stime adeguate, sia in riferimento al livello di preparazione complessiva degli studenti di ciascun Paese, sia in riferimento a singole subpopolazioni.
Al fine di ottenere informazioni ulteriori, sia sulle variabili di contesto che di processo, sono stati somministrati anche i seguenti questionari:

- **Questionario Studente**:
  gli studenti hanno fornito risposte utili alla raccolta di informazioni relative al background familiare e sociale; è stato possibile, inoltre, ottenere informazioni aggiuntive sulle caratteristiche del corso di studi seguito, sui loro atteggiamenti nei confronti della scuola e dell'apprendimento delle scienze, sulla motivazione nei confronti dello studio di tale disciplina.
  Una parte dei quesiti, poi, ha riguardato l'uso delle tecnologie dell'informazione e della comunicazione (TIC).
  Il test ha richiesto 30 minuti per la compilazione.

- **Questionario Scuola**:
  i dirigenti scolastici hanno risposto a domande sulle caratteristiche strutturali ed organizzative della scuola di appartenenza;

- **Questionario Genitori:**
  i genitori hanno risposto a domande utili per ottenere informazioni specifiche sull'educazione scientifica degli studenti all'interno dei contesti familiari[6].

[6] Il questionario genitori è stato somministrato in 16 Paesi, tra cui l'Italia.

Obiettivi dell'indagine In un quadro articolato come questo, gli obiettivi principali riguardano[7]:

- la produzione di indicatori del rendimento scolastico degli studenti quindicenni, in funzione della comparazione dei sistemi scolastici dei Paesi membri dell'organizzazione;
- l'individuazione delle caratteristiche dei sistemi scolastici virtuosi[8] al fine di ottenere indicazioni relative all'efficacia delle politiche scolastiche nazionali;
- la produzione dei dati sui risultati dei sistemi di istruzione in maniera regolare, favorendone in tal modo il monitoraggio; le serie storiche dei dati a disposizione possono rappresentare un utile strumento per l'orientamento delle politiche educative e scolastiche.

La valutazione delle conoscenze e delle abilità assume, in PISA, una connotazione ampia e pregnante. L'obiettivo è soprattutto quello di capire come, e se, gli studenti siano in grado di servirsi delle conoscenze acquisite per affrontare le normali sfide quotidiane e concrete.
La rilevazione si basa su una concezione dinamica dell'*apprendimento per tutta la vita* (**lifelong learning**), secondo la quale le abilità e le conoscenze necessarie per valutare scelte e prendere decisioni, in un mondo in continuo mutamento, si acquisiscono lungo l'intero arco della vita.
Un'attenzione particolare viene rivolta, in un certo senso, al *futuro*, a ciò di cui avranno bisogno gli studenti e a cosa saranno in grado di fare con quanto assimilato fino a quel momento. L'indagine ha un duplice fine quindi: valutare le competenze degli studenti ma, allo stesso tempo, valutare anche la loro capacità di contestualizzare ed applicare, negli ambiti della vita reale, le proprie conoscenze ed esperienze.

5.1.2 Il campione italiano

Il campione di studenti quindicenni scolarizzati italiani è composto da 21.773 unità, in 806 scuole[9], stratificato per macroaree geografiche (Nord Ovest, Nord Est, Centro, Sud, Sud Isole) e per indirizzi di studio (Licei, Istituti

[7] INVALSI (2008)

[8] Si fa riferimento ai Paesi che hanno conseguito i risultati migliori, in termini di livello medio delle prestazioni e di dispersione dei punteggi.

[9] L'analisi dei dati è stata poi condotta su 799 scuole poiché, per le restanti, non sono stati raggiunti i livelli minimi di partecipazione previsti.

tecnici, Istituti professionali, Scuole medie, Formazione professionale)[10].
Va precisato che il campione è rappresentativo della popolazione degli studenti quindicenni che frequentano ciascuno degli indirizzi di studio, non degli indirizzi di studio nel loro complesso.
Il campione italiano è rappresentativo di 11 regioni (Basilicata, Campania, Emilia Romagna, Friuli Venezia Giulia, Liguria, Lombardia, Piemonte, Puglia, Sardegna, Sicilia, Veneto) e delle due province autonome di Bolzano e di Trento.

5.1.3 Literacy matematica

Come anticipato, è bene dare uno sguardo ai risultati ottenuti dal campione italiano, relativamente alla competenza matematica. Innanzitutto, è necessario capire cosa si intenda con essa, prendendo spunto dalla definizione che ne dà l'INVALSI:

> la capacità di un individuo di individuare e comprendere il ruolo che la matematica gioca nel mondo reale, di operare valutazioni fondate e di utilizzare la matematica e confrontarsi con essa in modi che rispondono alle esigenze della vita di quell'individuo in quanto cittadino impegnato, che riflette e che esercita un ruolo costruttivo.

Dalla definizione appena citata, emerge chiaramente l'importanza dell'uso funzionale della matematica; diventa essenziale che lo studente sia in grado di utilizzare gli strumenti propri della disciplina per trovare le giuste soluzioni ai problemi della vita reale. La matematica non deve, quindi, essere intesa come una mera capacità di calcolo (sebbene non si possa prescindere dal possedere una solida base di conoscenze tecniche).

I quesiti utilizzati sono stati, in totale, 48. Ogni prova matematica ha fatto riferimento a tre dimensioni:

- i concetti matematici[11] che devono essere richiamati per la risoluzione del quesito;

[10] «Nel campione sono state incluse le scuole professionali delle Province autonome di Bolzano e di Trento, coerentemente con la definizione della popolazione oggetto di indagine in PISA. In alcune regioni sono stati inclusi nel campione anche gli studenti quindicenni che nel 2006 frequentavano i corsi di formazione professionale attivati in anticipazione della Legge 53/2003. L'indirizzo di studio "Formazione professionale" comprende sia le scuole professionali delle province di Bolzano e Trento, sia questi corsi professionali attivati a livello regionale». INVALSI (2008). *Le competenze in scienze lettura e matematica degli studenti quindicenni.* URL: `http://www.invalsi.it/invalsi/ri/pisa2006.php?page=pisa2006_it_05`

[11] Quantità, spazio e forma, cambiamento e relazioni, incertezza.

- le competenze, e quindi le abilità matematiche[12], necessarie per la risoluzione del quesito;
- il contesto[13] a cui il quesito si riferisce.

L'indagine prevede una scala delle competenze composta da 6 livelli, partendo dal presupposto che esse possano essere possedute con un diverso grado di padronanza; gli stessi quesiti sono costruiti con differenti livelli di difficoltà, dai più semplici, dove si utilizzano domande con applicazione al contesto familiare, ai più complessi, in cui lo studente deve essere in grado di gestire problemi molto più articolati.

Punteggi ottenuti Il punteggio medio OCSE per la matematica è pari a 498 (DS[14] 92); come è noto l'Italia ha una performance inferiore a tale valore e, precisamente, pari a 462 (DS 96).
I Paesi più virtuosi, che presentano punteggi significativamente più alti della media OCSE, sono: l'Australia (520; DS 88), il Belgio (520; DS 106), il Canada (527; DS 86), la Corea (547; DS 93), la Finlandia (548; DS 81), il Giappone (523; DS 91), la Nuova Zelanda (522; DS 93), i Paesi Bassi (531; DS 89), la Svizzera (530; DS 97).
Tra i paesi partner spiccano i risultati di Hong Kong (547; DS 93), Liechtenstein (525; DS 93), Macao (525; DS 84), Taiwan (549; DS 103).
Per quanto concerne la differenza di genere, fatta eccezione per Belgio, Corea e Liechtenstein, l'Italia e gli altri Paesi presentano una differenza di punteggio statisticamente significativa, a favore dei maschi. Per l'Italia si ha: maschi (470; ES[15] 2,9), femmine (453; ES 2,7).
Per quanto concerne la distribuzione delle frequenze percentuali in riferimento ai 6 livelli della scala di competenza, i risultati italiani sono i seguenti:

- **Sotto il Livello 1**: 13,5% degli studenti (media OCSE 7,7);
- **Livello 1**: 19,3% degli studenti (media OCSE 13,6);
- **Livello 2**: 25,5% degli studenti (media OCSE 21,9);
- **Livello 3**: 22,1% degli studenti (media OCSE 24,3);

[12] I raggruppamenti per competenze si distinguono in: raggruppamento della Riproduzione (semplici operazioni matematiche); raggruppamento delle Connessioni (integrare e collegare elementi di diverse aree di contenuto); raggruppamento della Riflessione (sviluppare ed applicare strategie in problemi più complessi).

[13] Vengono prese in considerazione quattro diverse situazioni: Personale; Scolastica/Occupazionale; Pubblica; Scientifica.

[14] Deviazione Standard

[15] Errore Standard

- **Livello 4**: 13,3% degli studenti (media OCSE 19,1);
- **Livello 5**: 5,0% degli studenti (media OCSE 10,0);
- **Livello 6**: 1,3% degli studenti (media OCSE 3,3).

I ricercatori PISA hanno individuato il livello 2 come livello base di competenza matematica; per il caso italiano ben il 32,8% non raggiunge tale livello di sufficienza, a fronte di una media OCSE corrispondente pari a 21,3%.
I più virtuosi sono solo il 6,3% del campione, a fronte di una corrispondente media OCSE del 13,3%.
In generale, va sottolineato che il punteggio italiano, non solo è ben al di sotto della media OCSE, è anche inferiore di 4 punti rispetto al punteggio medio ottenuto nel 2003 (sebbene tale differenza non sia statisticamente significativa).
L'Italia, inoltre, ha ottenuto, insieme alla Grecia, (459 punti, ma con trend positivo) il punteggio più basso tra i Paesi europei dell'area OCSE.
Per contro, in riferimento alla medesima area, la Finlandia si conferma la migliore, così come nel 2003, seguita dai Paesi Bassi e dalla Svizzera.
Ai fini dell'analisi, è importante tener presente anche un altro elemento: la varianza tra le scuole.
In particolare, per l'Italia risulta che ben il 52,1% della varianza totale è spiegata dalla variabilità tra le scuole, a fronte di una media OCSE pari al 33,1%.
Il fatto che il nostro Paese presenti punteggi medi inferiori alla media OCSE, e variabilità tra le scuole elevata, mette in rilievo come il nostro sistema scolastico non sia pienamente in grado di fornire un livello di qualità diffusa tra le scuole. Le differenze territoriali, più in generale, si ripercuotono anche in relazione al sistema educativo della nazione.

Le differenze tra macroaree

Le contraddizioni interne al nostro Paese si rendono evidenti quando si analizzano i risultati dei punteggi ottenuti per macroarea geografica e tipologia di scuola.
Il Nord presenta rendimenti nettamente superiori a quelli del Sud, così come i Licei rispetto agli Istituti professionali.
Dalla Figura 5.1 emerge un quadro chiaro e sintetico della situazione per macroarea geografica: il Nord Est ha gli studenti più virtuosi, mostrando una percentuale pari al 13,4% per gli ultimi due livelli della scala, percentuale analoga a quella della media OCSE.
Il Nord Ovest si classifica alla seconda posizione, con una percentuale pari

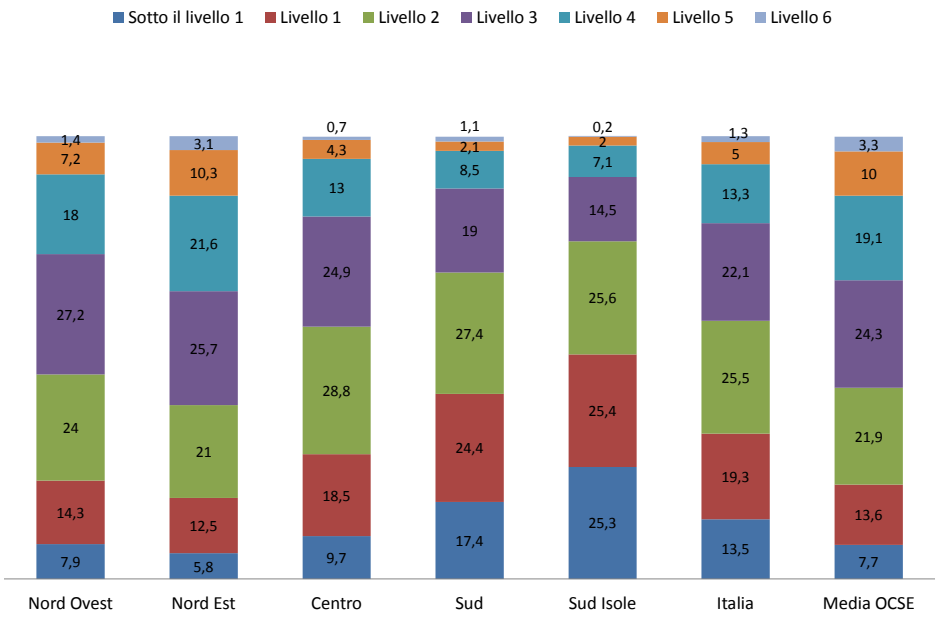


Figura 5.1: Percentuale di studenti a ciascun livello della scala complessiva di literacy matematica, per area geografica.

all'8,6% negli ultimi due livelli della scala; questo dato è positivo in relazione alla media nazionale (6,3%), negativo rispetto alla maggior parte dei Paesi dell'Unione Europea.
Risulta abbastanza evidente che il Centro, il Sud e il Sud Isole hanno, in relazione ai due livelli più alti, percentuali inferiori alla media nazionale, pari al 6,3%.
Dal momento che il livello 2 è considerato la soglia di sufficienza, i dati relativi ai primi due livelli della scala sono particolarmente interessanti.
La classifica è la seguente: Nord Ovest 22,2%; Nord Est 18,3%, Centro 28,2%, Sud 41,8%, Sud Isole 50,7%. I dati sono davvero poco confortanti per le regioni del Sud, in particolare per il Sud Isole.
Scendendo ad un livello di dettaglio maggiore di quello rappresentato in Figura 5.1, le uniche due regioni ad ottenere punteggi superiori alla media OCSE (498) sono: il Veneto (518) e il Friuli-Venezia Giulia (513). Seguono Lombardia (497), Emilia-Romagna (494), Piemonte (492). Presentano valori al di sotto della media la Liguria e tutte le regioni del Sud.
Per quanto riguarda il medesimo confronto con riferimento alla tipologia di scuola, dalla Figura 5.2 si nota che, per gli ultimi due livelli 5 e 6, i Licei

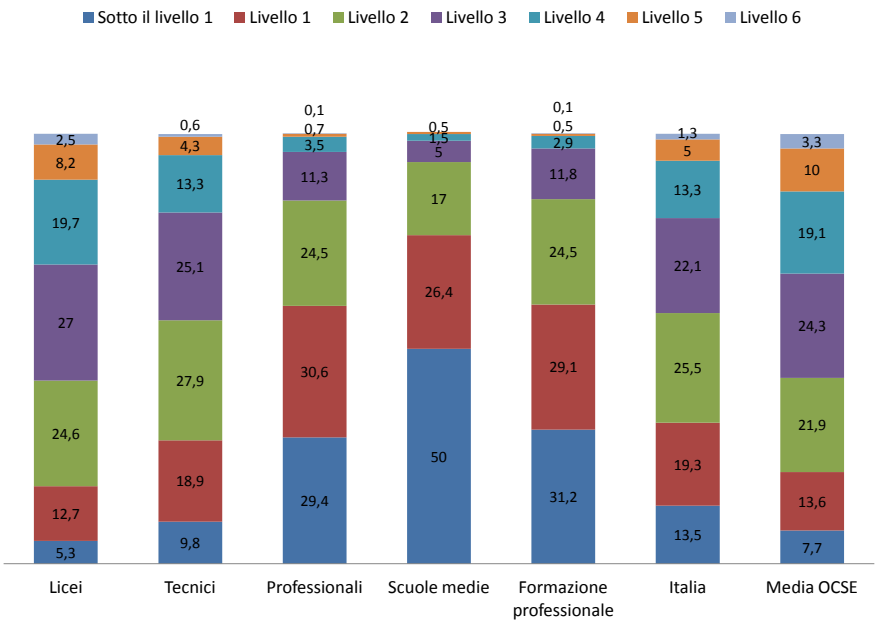


Figura 5.2: Percentuale di studenti a ciascun livello della scala complessiva di literacy matematica, per tipo di scuola.

hanno la percentuale più consistente, circa il 10%; pessima è la corrispondente performance degli Istituti professionali. A conferma di ciò, ai primi due livelli della scala, le percentuali più consistenti si registrano proprio per questi ultimi (circa il 60%).
I ricercatori precisano che tali basse performance non sono la conseguenza del cattivo funzionamento degli istituti professionali quanto il frutto di una migrazione degli studenti meno capaci, alla fine delle scuole medie, verso questa tipologia di scuola.
È evidente che i risultati ottenuti comportino una riflessione sui futuri percorsi del sistema scolastico italiano, suggerendo linee guida per il miglioramento di un sistema educativo poco omogeneo al suo interno e con standard di qualità inferiori alla media degli altri Paesi europei e non.

A questo punto non resta che dedicarsi alla descrizione delle caratteristiche del campione dei dati PISA 2006 utilizzato in questa applicazione.

5.2 Descrizione del campione utilizzato nell'applicazione

Per l'applicazione dei modelli descritti nel precedente capitolo, si è deciso di considerare i dati italiani dell'indagine PISA 2006, relativi al secondo *booklet* e con riferimento ai soli items di matematica. Esso consta di 1666[16] studenti per un totale di 12 items dicotomici.
I gruppi, che compongono il terzo livello della gerarchia, fanno riferimento alla stratificazione territoriale in Regioni\Province autonome stabilite in PISA 2006, riproposta con le relative frequenze assolute nella Tabella 5.1:

Tabella 5.1: Ripartizione dei gruppi nel dataset di studio.

Regioni\Province aut.	Frequenza ass.
Basilicata	117
Campania	110
Emilia Romagna	115
Friuli Venezia Giulia	120
Liguria	133
Lombardia	119
Piemonte	110
Prov. aut. di Bolzano	162
Prov. di Trento	129
Puglia	113
Sardegna	104
Sicilia	104
Veneto	126
Resto d'Italia	104
Totale	1666

In riferimento ai 12 item studiati, sulla base del grafico a torta in Figura 5.3, risulta una percentuale di dati mancanti pari al 17% del totale delle risposte. Il dato è, però, troppo generale; è più interessante considerare le medesime frequenze in relazione a ciascun item, come mostra la Figura 5.4 La percentuale più elevata di dati mancanti si registra per l'item 5, con il 41%; se, ad essa, si somma la percentuale delle risposte sbagliate pari al 55%, risulta evidente che l'item in questione è da osservare con un certo interesse essendo risultato, ai più, molto complesso.

[16]Nel modello con le covariate, il dataset considera 1629 studenti poiché sono stati eliminati i records per i quali vi erano dati mancanti in relazione alle covariate d'interesse.

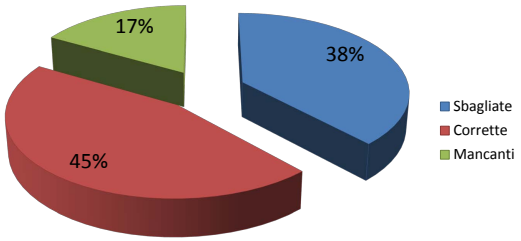


Figura 5.3: Ripartizione delle frequenze percentuali delle risposte.

Il secondo item, con una percentuale consistente pari al 39%, è l'item 12; la somma della percentuale di *missing* e di risposte sbagliate rappresenta ben il 74% del totale delle risposte.
Segue l'item 10 con il 37% e, anche in questo caso, con un massiccio numero di risposte sbagliate (45%).
L'item 11 ha una percentuale pari al 30% di dati mancanti, sebbene sia da notare che, in questo caso, la ripartizione delle frequenze è a favore delle risposte corrette.
Un altro item, con una percentuale consistente di *missing value*, è l'item 7 con il 24%; la somma dei dati mancanti e delle risposte sbagliate rappresenta il 73% del totale delle risposte.
In ultimo si segnala l'item 4, con una percentuale di poco superiore al dieci per cento e con un elevato numero di risposte corrette.
Da questo primo sguardo descrittivo dei dati, sembra che la presenza di *missing* si accompagni ad un'altrettanta elevata percentuale di risposte sbagliate, quindi ad item percepiti dallo studente come difficili, oltre che ad items cosiddetti "*not reached*", non completati per mancanza di tempo (non a caso ultimi in ordine di lettura).
Sembra plausibile, sulla base di questa prima disamina, ipotizzare un meccanismo del dato mancante MNAR dal momento che i dati suggeriscono come la mancata risposta si associ, frequentemente, ad item con livelli di difficoltà più elevati e, quindi, in un certo senso, faccia riferimento a livelli di abilità più bassi.
Particolarmente interessante è sembrato valutare la ripartizione delle percentuali di risposta in relazione al gruppo di appartenenza. I grafici in Figura 5.5, 5.6 e 5.7 confermano, con assoluta coerenza rispetto ai risultati generali

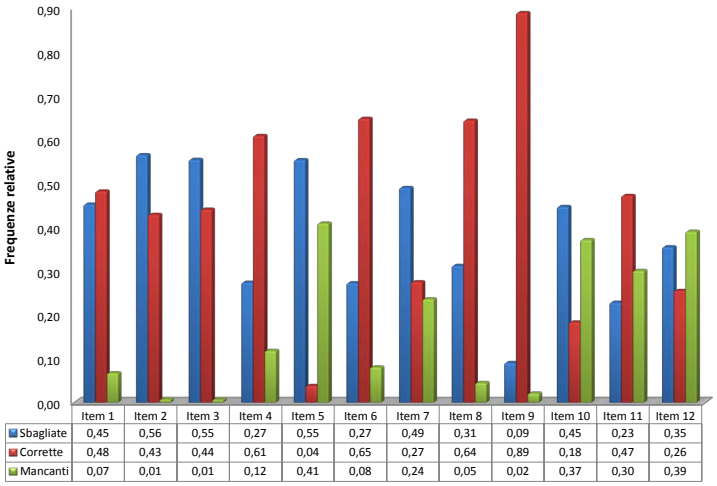


Figura 5.4: Ripartizione delle frequenze relative delle risposte fra i 12 item.

dell'indagine PISA, le performance negative delle regioni del Sud Italia; cosa ancora più interessante è che lo stesso atteggiamento è riscontrabile per la distribuzione di frequenza dei dati mancanti.
Quest'ultima considerazione fa riflettere sul fatto che, da questa nuova prospettiva, i dati possano suggerire una situazione di tipo MAR, in quanto la distribuzione dei dati mancanti sembra dipendere dalla struttura "geografica", dalla stratificazione territoriale dei gruppi. Ma ciò sarà oggetto di analisi dei prossimi paragrafi.
Per concludere la descrizione del dataset di riferimento, si riporta l'elenco delle covariate utilizzate in questo studio:

- Genere (1 se Maschio);
- Immigrato (1 se nato all'estero);
- Tipologia di scuola (1 se privata);
- Indice di status socio-economico-culturale (ESCS)[17]
- Area Geografica di appartenenza[18] (1 se Sud).

[17]È un indicatore PISA composito che integra tre tipi di variabili: livello di istruzione dei genitori; la loro professione; la presenza in casa di beni materiali e culturali e di risorse di tipo educativo. L'indice è standardizzato sui parametri dei Paesi OCSE: ha quindi media 0 e varianza unitaria.
[18]Essa è stata creata dicotomizzando la variabile relativa alla macroarea geografica.

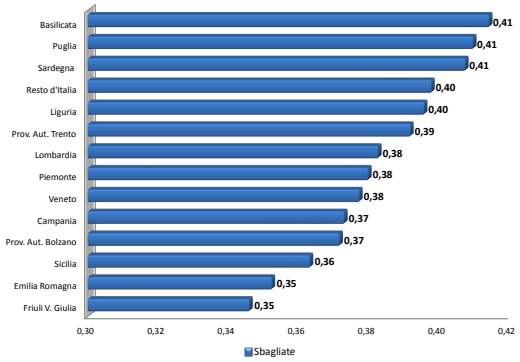


Figura 5.5: Frequenze relative delle risposte sbagliate per gruppo di appartenenza (in ordine decrescente).

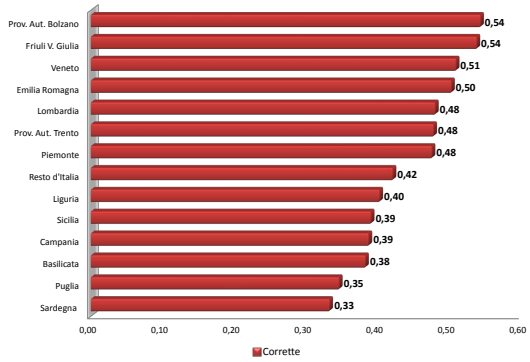


Figura 5.6: Frequenze relative delle risposte corrette per gruppo di appartenenza (in ordine decrescente).

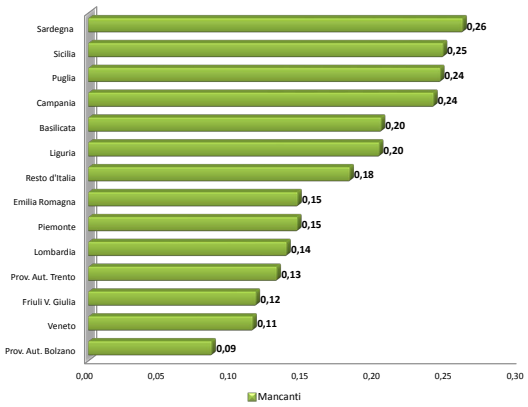


Figura 5.7: Frequenze relative delle risposte mancanti per gruppo di appartenenza (in ordine decrescente).

5.3 Risultati: i modelli a confronto

Nei paragrafi che seguono, la logica della trattazione impone di considerare, in primis, il grado di correlazione tra abilità e propensione alla risposta (nel modello NIM) al fine di avere una misura del grado di deviazione dall'ipotesi di ignorabilità del meccanismo generatore del dato mancante.
Ne segue che, qualora le due correlazioni, a livello studente e a livello gruppo, risultino statisticamente significative, il modello NIM diventa il modello di paragone rispetto al quale confrontare le stime ottenute per gli altri due modelli che, per costruzione, assumono l'ignorabilità del meccanismo generatore.
In sintesi, l'obiettivo è investigare quali conseguenze e distorsioni nelle stime possano verificarsi quando, evidenziata una correlazione statisticamente significativa tra le latenti, si decida di considerare modelli che suppongono, al contrario, un'ipotesi di ignorabilità.
Prima di passare all'analisi vera e propria dei risultati, al fine di comprendere come questi siano stati ottenuti, è necessaria una piccola digressione sui metodi di analisi utilizzati per verificare la convergenza degli algoritmi MCMC alla distribuzione target.
La possibilità di utilizzare tali algoritmi risiede proprio nel fatto che, dopo una fase transiente, la catena campioni effettivamente dalla distribuzione target, l'*a posteriori* appunto.
In generale, non si è mai certi della convergenza dell'algoritmo utilizzato, piuttosto si dice che ha raggiunto la convergenza all'iterazione T se NON c'è evidenza che questo non sia avvenuto e, in tal caso, è possibile utilizzare i risultati della catena per t>T.
La diagnostica di convergenza viene utilizzata proprio a tale scopo e si compone di ausili grafici e di opportuni test. In questo studio ci si è avvalsi dei seguenti strumenti implementati in Openbugs:

- **La statistica BGR**.
  Calcola la statistica di Gelman e Rubin, così come modificata da Brooks e Gelman (1998). L'idea alla base è quella di considerare un minimo di 2 catene, con valori iniziali molto diversi, e valutare la convergenza confrontando la variabilità *within chain* e *between chain*, relativamente alla seconda metà delle catene.
  La procedura è la seguente: sia M il numero delle catene e 2T la lunghezza di ciascuna di esse; si considera come misura della variabilità *a posteriori* l'ampiezza dell'intervallo di credibilità al $100(1-\alpha)\%$ (in Openbugs $\alpha = 0.2$).
  Per le ultime T iterazioni e per ciascuna catena, viene calcolato l'inter-

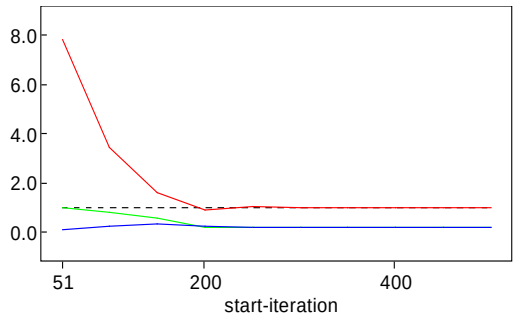


Figura 5.8: Output relativo alla statistica di Gelman e Rubin.

vallo di credibilità empirico; successivamente viene calcolata l'ampiezza media degli intervalli tra le M catene, indicata con **W**. Infine, viene calcolata l'ampiezza degli intervalli basata su tutti gli MT campioni "pooled", indicata con **B**.
Il rapporto $\mathbf{R} = \mathbf{B}/\mathbf{W}$ tende a 1 se la convergenza è stata raggiunta.
Openbugs produce un output come quello in Figura 5.8 che permette di esaminare il comportamento dell'indice **R** rispetto alle iterazioni effettuate. **R** è rappresentato dalla linea rossa, **B** dalla linea verde e **W** dalla linea di colore blu[19]. Brooks e Gelman (1998) sottolineano che è importante, non solo che il rapporto **R** converga approssimativamente ad 1, ma che anche **B** e **W** siano stabili.
Nel grafico mostrato, ad esempio, la convergenza viene raggiunta dopo 250 iterazioni. Le precedenti possono essere scartate poiché rappresentano la cosiddetta fase di *burn in*.

- **History plot, autocorrelation function e stime kernel di densità**.
  Nella Figura 5.9 sono riportati esempi di output generati da Openbugs.
  Il primo dei grafici in alto, l'history plot, riporta la serie storica dei valori campionati per due catene. In caso di convergenza, come nell'esempio, esse mostrano un buon grado di "mixing" e una certa stabilità intorno alla moda della distribuzione.
  Il secondo grafico, a sinistra, riporta i valori dell'autocorrelazione fino

[19] **B** e **W** sono normalizzati così da poterli rappresentare nella stessa scala di **R**.

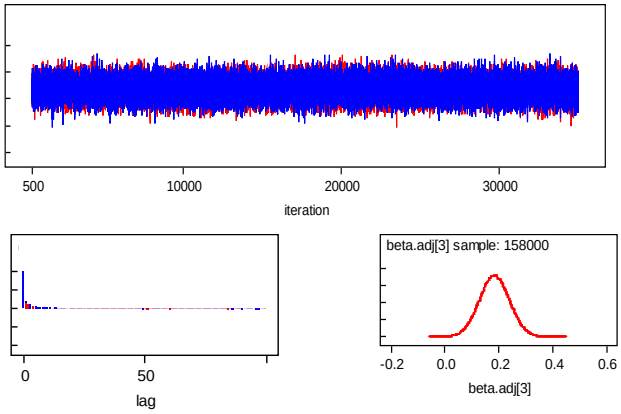


Figura 5.9: History plot, Autocorrelation function e stime kernel di densità.

al *lag* 50: una forte autocorrelazione del primo ordine appare scontata ma, se a lag crescenti questa non diminuisce rapidamente, ci si muove troppo lentamente negli spazi parametrici e si supera lentamente la fase transiente; la situazione ideale è una autocorrelazione che decresca rapidamente dopo il primo lag. In caso contrario, è possibile campionare con un "thinning interval", ovvero si accetta un valore campionato scartandone un certo numero pari all'intervallo stabilito (2,3,....).
Il terzo grafico riporta semplicemente la stima kernel della densità a posteriori del parametro.

Il package CODA, implementato in R, è stato particolarmente utile per effettuare ulteriori verifiche circa la stazionarietà della catena e l'accuratezza delle stime dei quantili, in particolar modo per i modelli riferiti al DIF, per i quali è stata considerata una singola catena (e quindi il test BGR non è applicabile). In appendice C sono riportate la stima di densità a posteriori per i principali parametri di interesse.
Il dettaglio, per i modelli utilizzati, è il seguente:

- modello NIM:
  due catene di 180.000 iterazioni con burn in di 22.000 e thin di 2;

- modello IM:
  due catene di 120.000 iterazioni e burn in di 10.000;

- modello ZIM:
  due catene di 120.000 con burn in di 6.000;
- modello NIM con covariate:
  due catene di 120.000 iterazioni e burn in di 5.000;
- modello DIF NIM:
  una catena di 40.000 iterazioni con burn in di 5.500;
- modello DIF IM:
  una catena di 30.000 iterazioni e burn in di 1.000;
- modello DIF ZIM:
  una catena di 30.000 iterazioni e burn in di 1.000.

.

5.3.1 Modello NIM: analisi della correlazione.

Oggetto privilegiato d'indagine, come anticipato, è la correlazione tra le dimensioni, sia al secondo che al terzo livello gerarchico. Nelle Tabelle 5.2 e 5.3, oltre ai valori della matrice di varianza e covarianza, sono riportate le correlazioni tra le latenti.
In particolare, la media a posteriori del coefficiente di correlazione, a livello studente, è pari a 0.684 mentre l'altra, a livello di gruppo, è pari a 0.673.
La cosa più importante da valutare è, però, se questi parametri siano statisticamente diversi da zero; gli intervalli di credibilità al 95%, non includendo lo zero, sostengono l'ipotesi alternativa: è, quindi, possibile affermare che esiste una relazione statistica tra abilità e propensione alla risposta.
Va sottolineato che la significatività statistica della correlazione deve evidenziarsi per entrambi i livelli gerarchici considerati.
A questo punto, verificata la presenza del supposto legame tra le due dimensioni latenti, è opportuno ed interessante analizzare quanto la forza di questa relazione sia tale da determinare distorsioni evidenti nelle stime dei parametri di interesse.
A tal proposito va ricordato che due sono i fattori determinanti in questo contesto: la forza della correlazione e il numero di dati mancanti nel campione. È la combinazione di questi fattori a determinare il livello di distorsione.
Prima di procedere con il confronto tra il modello NIM, che assurge a modello di riferimento nello studio, e gli altri due modelli, è bene completare il quadro relativo alla variabilità commentando i risultati della Tabella 5.4: l'ICC, per entrambe le dimensioni, ha valori della media a posteriori simili, consentendo

di affermare che la quota di variabilità dovuta ai gruppi (in questo caso alle regioni) è circa pari al 15%.

Tabella 5.2: Modello NIM: Media a posteriori, deviazione standard e intervalli di credibilità per la matrice di varianza e covarianza (e per il coefficiente di correlazione) dell'effetto random di secondo livello.

	Media	S. D.	$2.5pc$	Mediana	$97.5pc$
$\sigma^2_{1,\theta^{(2)}}$	1.249	0.078	1.103	1.247	1.407
$\sigma_{12,\theta^{(2)}}$	1.492	0.097	1.309	1.489	1.688
$\sigma^2_{2,\theta^{(2)}}$	3.814	0.238	3.369	3.807	4.302
$\rho_{student}$	0.684	0.027	0.630	0.685	0.735

Tabella 5.3: Modello NIM: Media a posteriori, deviazione standard e intervalli di credibilità per la matrice di varianza e covarianza (e per il coefficiente di correlazione) dell'effetto random di terzo livello.

	Media	S. D.	$2.5pc$	Mediana	$97.5pc$
$\sigma^2_{1,\theta^{(3)}}$	0.235	0.108	0.103	0.210	0.508
$\sigma_{12,\theta^{(3)}}$	0.275	0.157	0.076	0.242	0.666
$\sigma^2_{2,\theta^{(3)}}$	0.683	0.318	0.294	0.611	1.493
ρ_{group}	0.673	0.154	0.297	0.700	0.891

Tabella 5.4: Modello NIM: Media a posteriori, deviazione standard e intervalli di credibilità per l'Intraclass Correlation Coefficient (sia per la prima che per la seconda dimensione)

	Media	S. D.	$2.5pc$	Mediana	$97.5pc$
ICC_1	0.155	0.056	0.076	0.144	0.291
ICC_2	0.148	0.055	0.072	0.138	0.281

A corredo di quanto appena detto, si riportano, nella tabella riassuntiva 5.5, anche le varianze e l'ICC dei due modelli univariati. Si noti che la quota di variabilità dovuta ai cluster è leggermente più bassa, per entrambi i modelli, rispetto a quella del modello bivariato.

Tabella 5.5: Modelli IM e ZIM: Tabella riassuntiva per le varianze e l'Intraclass Correlation Coefficient.

		Media	S. D.	$2.5pc$	Mediana	$97.5pc$
Modello IM	$\sigma^2_{\theta^{(2)}}$	1.150	0.074	1.011	1.148	1.299
	$\sigma^2_{\theta^{(3)}}$	0.137	0.072	0.052	0.120	0.321
	ICC_1	0.104	0.046	0.043	0.095	0.219
Modello ZIM	$\sigma^2_{\theta^{(2)}}$	1.577	0.088	1.411	1.575	1.757
	$\sigma^2_{\theta^{(3)}}$	0.230	0.117	0.092	0.203	0.526
	ICC_1	0.124	0.051	0.055	0.114	0.251

Per effettuare i confronti tra le stime dei parametri dei modelli, considerando come termine di riferimento il modello bivariato NIM, è utile avvalersi, oltre che dei risultati numerici, anche dell'ausilio grafico.
Ai valori delle medie a posteriori, contenuti nelle Tabelle 5.6 e 5.7, si affiancano i grafici, detti "caterpillar", delle Figure 5.10, 5.11, 5.13 e 5.14, che rendono più agevoli i confronti e consentono una visione chiara delle posizioni relative dei parametri, essendo questi ultimi ordinati in relazione al proprio rango.
È, cioè, possibile valutare graficamente le graduatorie di difficoltà degli items e quelle di abilità dei gruppi.

5.3.2 Confronto tra i parametri di difficoltà

Partendo dall'analisi dei parametri di difficoltà, si considerino le Figure 5.10 e 5.11: i caterpillar definiscono la propria graduatoria dei parametri di difficoltà (in deviazione dalla media), ordinati dal più facile al più difficile.
Considerando il caterpillar del modello NIM, è possibile individuare quali siano gli items "più estremi": l'item 9, ad esempio, risulta particolarmente facile, infatti è abbastanza distaccato dal gruppo costituito dagli items 6, 8, 4, 11 che presentano, comunque, segno negativo.
In maniera speculare, spicca, come item particolarmente complesso, il quesi-

to numero 5 seguito, con un certo distacco, dal quesito 10.
Continuano poi la graduatoria gli item 12 e 7 ed, infine, gli items 2 e 3, questi con valori positivi molto più vicini allo zero. Proprio a quest'ultimo gruppo è da associare l'item 1, il quale pur avendo segno negativo, è comunque prossimo a valori intorno allo zero.
Passando al confronto con la graduatoria prodotta dal modello IM, utilizzando il solo approccio grafico della Figura 5.10, sembra che le posizioni relative degli items siano pressoché uguali, fatta eccezione per un piccolo "*switch*" di posizione per due items contigui, ossia il 4 e l' 8.
Diventa, a questo punto, necessario, per analizzare eventuali differenze e distorsioni, ricorrere al confronto delle medie a posteriori contenute nella Tabella 5.6[20] e considerare il grafico in Figura 5.12; è sufficiente scorrere la prima e la terza colonna della Tabella 5.6 per rendersi conto che i due modelli producono stime più o meno differenti tra loro; al fine di valutare la variazione, in logit, per ciascun parametro, si è deciso di standardizzare tali valori[21] e di calcolarne poi la differenza in valore assoluto. Il grafico (linea rossa), in Figura 5.12, riporta i valori di tali differenze in ordine crescente mentre, sull'asse delle ascisse, è possibile individuare l'item di riferimento.
Risulta evidente come la variazione delle stime sia trasversale e non interessi esclusivamente gli items con più risposte mancanti: considerare un modello con ipotesi di ignorabilità comporta, quindi, variazioni nelle stime della maggior parte dei parametri coinvolti e può determinare cambiamenti di graduatoria tra gli items.
Volendo cogliere una tendenza generale, fatta eccezione per l'item 11, per tutti gli altri vale che il modello IM tende a sovrastimare la difficoltà per gli items con segno negativo e a sottostimare la difficoltà di quelli con segno positivo.
Operando il medesimo confronto tra modello NIM e modello ZIM, dalla valutazione del caterpillar emerge che ci sono due "*switch*" di posizione che interessano, di nuovo, item contigui tra di loro: in particolare le coppie di items (1,11) e (7,12). Non meraviglia ciò dal momento che gli items 11 e 12 sono tra gli items con percentuale di dati mancanti consistente; l'inserimento degli zeri ha determinato un aumento del valore del logit (li ha resi più difficili).
Un altro elemento interessante è notare come il secondo ed il terzo items si trovino ora, a sinistra, rispetto alla linea della media, sono cioè considerati più semplici rispetto al modello NIM.

[20] In Appendice B, per ciascun modello, sono riportate le tabelle complete dei parametri di difficoltà, che includono gli intervalli di credibilità.

[21] Essendo già in deviazione dalla media, si è semplicemente diviso tali valori per la deviazione standard.

Il grafico (linea blu) in Figura 5.12 riporta il trend delle differenze (in valore assoluto) tra le medie a posteriori standardizzate riferite al modello NIM e ZIM.
Va assolutamente sottolineato come il modello ZIM sia quello che produce distorsioni più consistenti delle stime dei parametri; la deviazione dal valore presente per il modello bivariato è, di gran lunga, più evidente rispetto al modello IM.
Questo implica che la considerazione di definire la non risposta come una risposta certamente errata produce le distorsioni maggiori nelle stime dei parametri.
È ovvio che la tendenza generale di questo terzo modello è quella di determinare valori del logit molto più alti per i parametri con un numero consistente di dati mancanti, effettuando di fatto una sovrastima degli stessi; per contro, tende a sottostimare gli items di bassa e media difficoltà (fatta eccezione per l'item 11).
Se si dà un ultimo sguardo ai dati, confrontando tutte e tre le stime dei parametri, in maniera del tutto generale, sembrerebbe che il modello NIM giochi una posizione mediana, ossia i valori delle sue stime sono sempre comprese nell'intervallo di valori con estremi le stime degli altri due modelli.

Tabella 5.6: Confronto tra le Medie a posteriori dei parametri di difficoltà dei tre modelli

	Modello NIM		Modello IM		Modello ZIM	
	Media	S. D.	Media	S. D.	Media	S. D.
β_1^{adj}	−0.121	0.057	−0.086	0.0564	−0.293	0.056
β_2^{adj}	0.241	0.056	0.294	0.055	−0.008	0.056
β_3^{adj}	0.182	0.055	0.236	0.055	−0.070	0.056
β_4^{adj}	−0.978	0.062	−0.971	0.062	−0.980	0.057
β_5^{adj}	3.397	0.129	3.311	0.129	3.571	0.127
β_6^{adj}	−1.117	0.062	−1.072	0.062	−1.202	0.059
β_7^{adj}	0.806	0.064	0.786	0.064	0.889	0.061
β_8^{adj}	−0.991	0.060	−0.941	0.059	−1.178	0.058
β_9^{adj}	−2.871	0.089	−2.799	0.089	−3.065	0.084
β_{10}^{adj}	1.397	0.073	1.337	0.073	1.556	0.069
β_{11}^{adj}	−0.662	0.067	−0.727	0.068	−0.239	0.056
β_{12}^{adj}	0.717	0.069	0.633	0.069	1.020	0.062

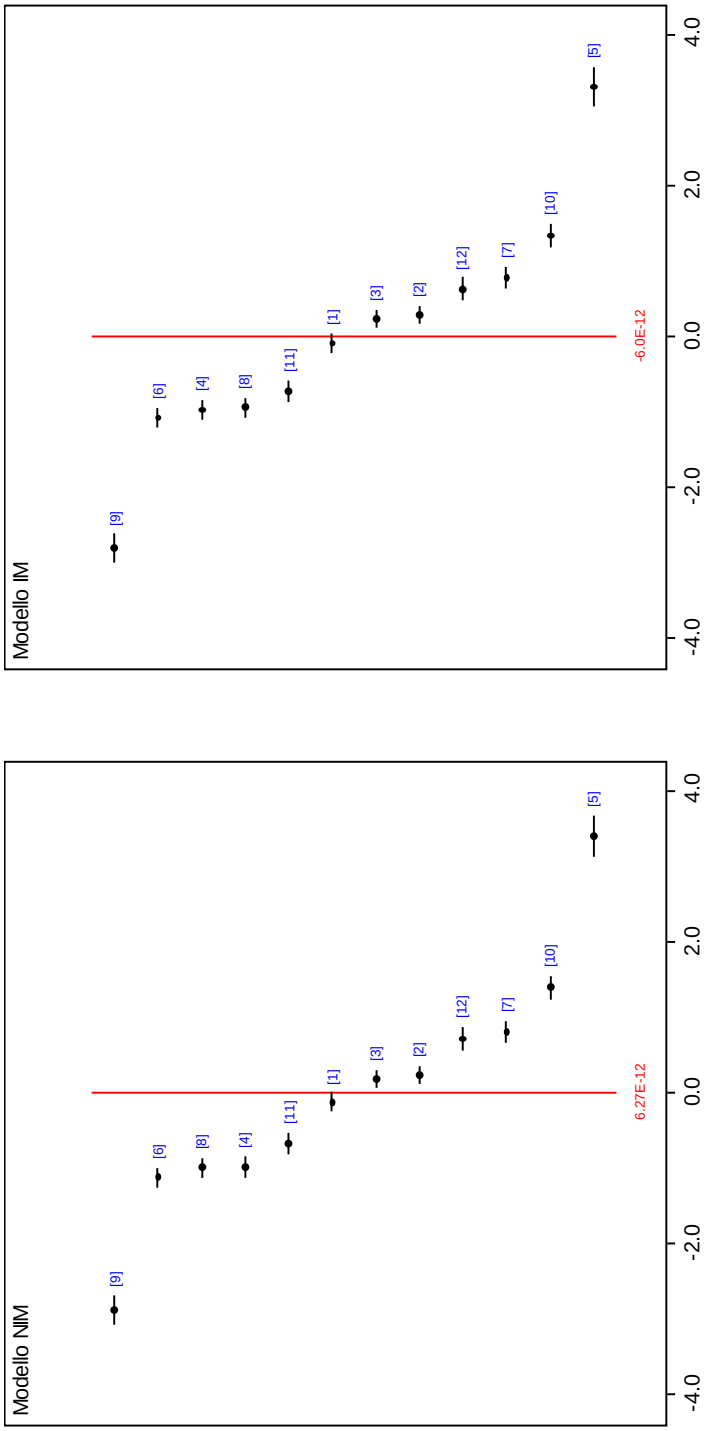


Figura 5.10: Caterpillar dei parametri di difficoltà, ordinati per rango, per i modelli NIM e IM

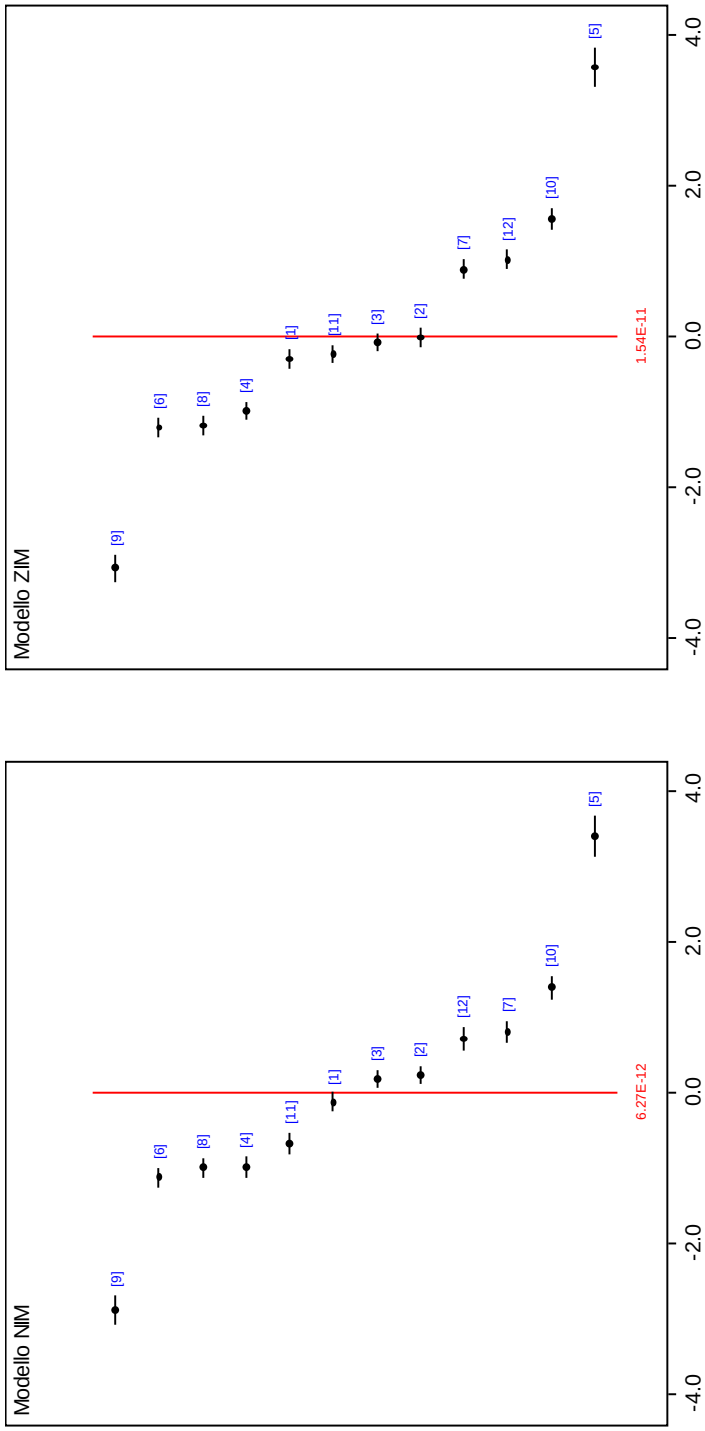


Figura 5.11: Caterpillar dei parametri di difficoltà, ordinati per rango, per i modelli NIM e ZIM

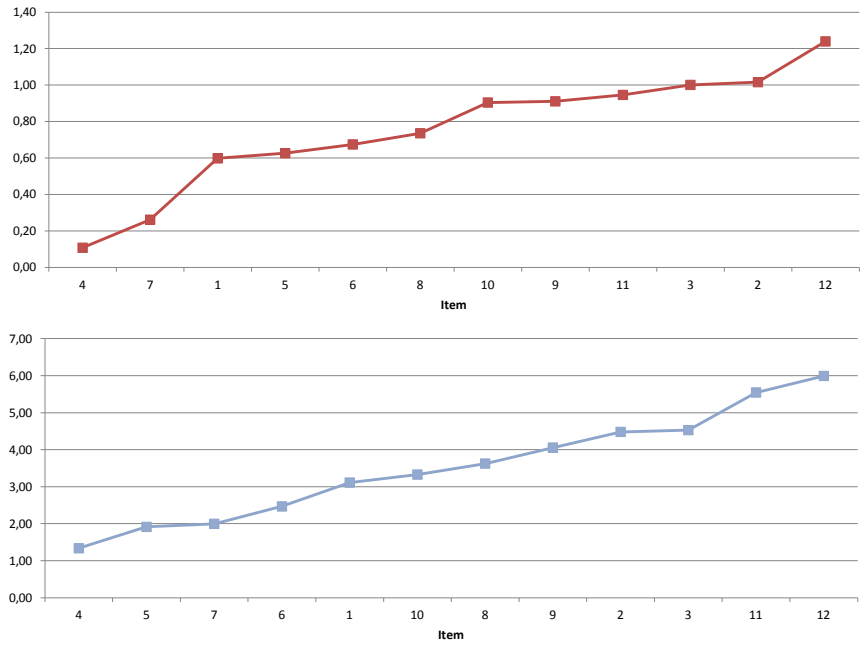


Figura 5.12: Trend delle differenze in valore assoluto tra le medie a posteriori (standardizzate) dei parametri di difficoltà del modello NIM e del modello IM (linea rossa), del modello NIM e ZIM (linea blu)

5.3.3 Parametri di abilità di gruppo

L'osservazione dei ranghi, per tutti e tre i caterpillar (*vedi* Figure 5.13 e 5.14), è in quasi perfetta consonanza coni risultati ottenuti per l'intero campione italiano[22].
I dati confermano che le regioni del Sud e la Liguria hanno performance negative, inferiori alla media dei gruppi, per la literacy matematica, mentre performance di segno positivo si riscontrano per le regioni del Nord Ovest ed, in modo particolare, per le regioni del Nord Est. Ancora una volta la distribuzione delle competenze si mostra strettamente connessa alla stratificazione territoriale, a conferma di un sistema scolastico fortemente caratterizzato da criticità e sperequazioni territoriali anche molto accentuate.
Focalizzando l'attenzione sul confronto tra i modelli, prima tra NIM e IM, emerge che non ci sono stravolgimenti nella graduatoria delle abilità, a meno di uno *switch* tra abilità contigue (gruppo 11 e gruppo 6); in realtà vi è un

[22]In questo studio non sono stati usati i pesi campionari.

cambiamento di posizione anche per i gruppi 7 e 8, ma di entità trascurabile (per i tre gruppi contigui (8, 7, 12), in realtà, in entrambi i modelli, le loro posizioni sono pressoché allineate).
Per quanto riguarda, invece, il cambio di posizione per il gruppo 6 (Liguria) e 11 (Sicilia), esso è motivato dal fatto che, nel modello bivariato, il gruppo 11 subisce una variazione del logit verso valori ancora più negativi (si ricorda che la Sicilia è la seconda regione per percentuale di risposte mancanti).
Si noti, poi, che il modello NIM pone su un livello paritario il punteggio del Veneto e dell'Emilia Romagna, al contrario del modello IM, che pone il Veneto in posizione più bassa rispetto all'Emilia Romagna (si ricorda che il Veneto è, tra le regioni virtuose, quella con la percentuale più bassa di *missing*). Il modello NIM, inoltre, riduce anche la distanza tra la prima e la seconda classificata.
Quanto illustrato è assolutamente deducibile dal confronto delle medie a posteriori della Tabella 5.7[23] e dall'andamento del trend delle differenze (in valore assoluto) standardizzate, rappresentato in Figura 5.15 (linea rossa).
Il confronto delle stime mette in evidenza come il modello NIM operi una sorta di ponderazione delle stesse: in particolare, le variazioni più consistenti si verificano per le regioni con comportamenti estremi, ed opposti, in relazione alla mancata risposta. Le regioni più virtuose, con percentuali basse di dati mancanti e abilità elevate, sono quelle che ottengono i benefici più netti, vedendosi spostare le stime verso valori più alti; al contrario, le regioni con percentuali elevate di *missing* e basse performance, vedono le loro stime ridursi ancora di più.
Aumenta, in un certo senso, la distanza tra il gruppo dei virtuosi e il gruppo dei mediocri; il modello IM pare non cogliere tali sfumature.
In relazione al confronto con il modello ZIM, per quanto attiene alla graduatoria delle abilità in Figura 5.14, si nota una piccola variazione nei ranghi delle regioni contigue (dal punto di vista dei punteggi) 11, 6, 3: mentre nel modello NIM i gruppi 6 e 3 sono sostanzialmente allineati, nel modello ZIM il gruppo 3 arretra di posizione.
Il modello ZIM, inoltre, pone la regione 13 in una posizione di relativo vantaggio rispetto alla regione 3, mentre nel modello NIM esse hanno rango uguale; situazione analoga si verifica per le due regioni più virtuose, con vantaggio della regione 1 nel modello ZIM.
Il confronto delle stime, in Tabella 5.7, mette in evidenza come il modello ZIM renda ancora più acute le differenze; esso tende ad attribuire, in maniera più consistente del modello NIM, punteggi più bassi per le regioni meno

[23] In Appendice B, per ciascun modello, sono riportate le tabelle complete dei parametri di abilità di gruppo, che includono gli intervalli di credibilità.

virtuose e punteggi più alti per le regioni con competenze elevate e pochi valori mancanti.
Anche per questo confronto è possibile valutare le differenze, in valore assoluto, delle medie a posteriori standardizzate sulla base del grafico prodotto in Figura 5.15 (linea blu).
In generale, anche per i parametri di abilità, si verifica che le stime, ottenute con il modello che tiene conto del meccanismo generatore del dato mancante, siano sempre comprese tra i valori prodotti dagli altri due modelli.
Le stime del modello NIM sembrano essere una mediazione tra due diverse ed opposte situazioni: una che tende a sovrastimare i valori negativi dell'abilità e a sottostimare quelli positivi, l'altra che tende a fare il contrario, a rendere cioè ancora più acute le differenze.

Tabella 5.7: Confronto tra le Medie a posteriori,degli effetti random di terzo livello dei tre modelli

	Modello NIM		Modello IM		Modello ZIM	
	Media	S. D.	Media	S. D.	Media	S. D.
$\theta_1^{adj,(3)}$	0.495	0.102	0.431	0.098	0.619	0.110
$\theta_2^{adj,(3)}$	−0.302	0.118	−0.295	0.114	−0.360	0.125
$\theta_3^{adj,(3)}$	−0.213	0.122	−0.171	0.116	−0.310	0.129
$\theta_4^{adj,(3)}$	0.327	0.118	0.324	0.114	0.361	0.127
$\theta_5^{adj,(3)}$	0.507	0.116	0.464	0.112	0.568	0.125
$\theta_6^{adj,(3)}$	−0.218	0.111	−0.213	0.107	−0.253	0.118
$\theta_7^{adj,(3)}$	0.171	0.115	0.138	0.110	0.225	0.124
$\theta_8^{adj,(3)}$	0.163	0.119	0.144	0.114	0.199	0.128
$\theta_9^{adj,(3)}$	−0.499	0.120	−0.457	0.117	−0.573	0.129
$\theta_{10}^{adj,(3)}$	−0.588	0.127	−0.510	0.124	−0.667	0.136
$\theta_{11}^{adj,(3)}$	−0.232	0.125	−0.183	0.120	−0.308	0.132
$\theta_{12}^{adj,(3)}$	0.164	0.111	0.141	0.107	0.216	0.119
$\theta_{13}^{adj,(3)}$	0.326	0.113	0.284	0.108	0.403	0.122
$\theta_{14}^{adj,(3)}$	−0.101	0.123	−0.100	0.118	−0.119	0.131

* 1. Prov. aut. Bolzano; 2. Basilicata; 3. Campania; 4. Emilia Romagna; 5. Friuli V. Giulia; 6. Liguria; 7. Lombardia; 8. Piemonte; 9. Puglia; 10. Sardegna; 11. Sicilia; 12. Prov. aut. Trento; 13. Veneto; 14. Resto d'Italia.

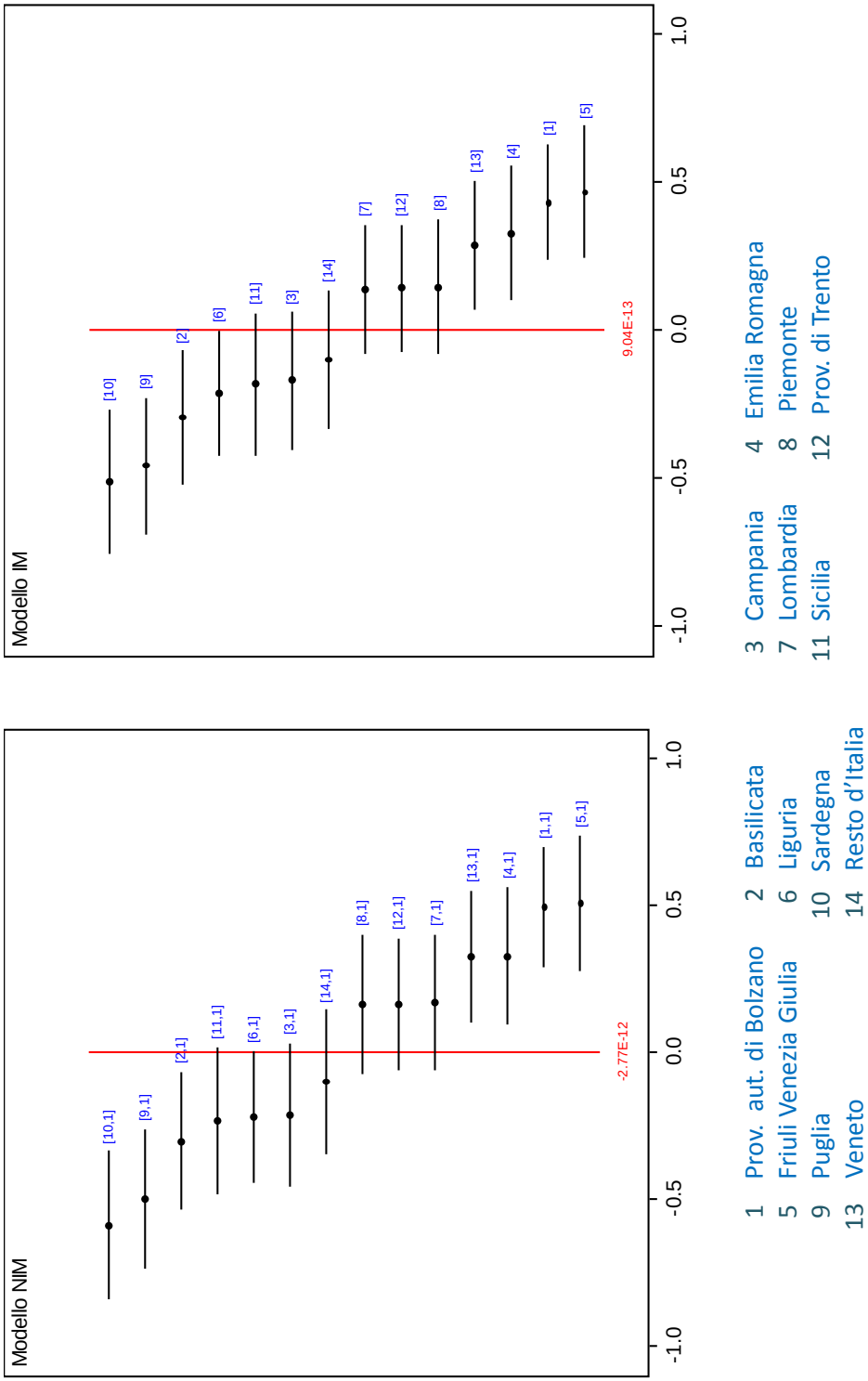


Figura 5.13: Caterpillar dei parametri di abilità di, ordinati per rango, per i modelli NIM e IM

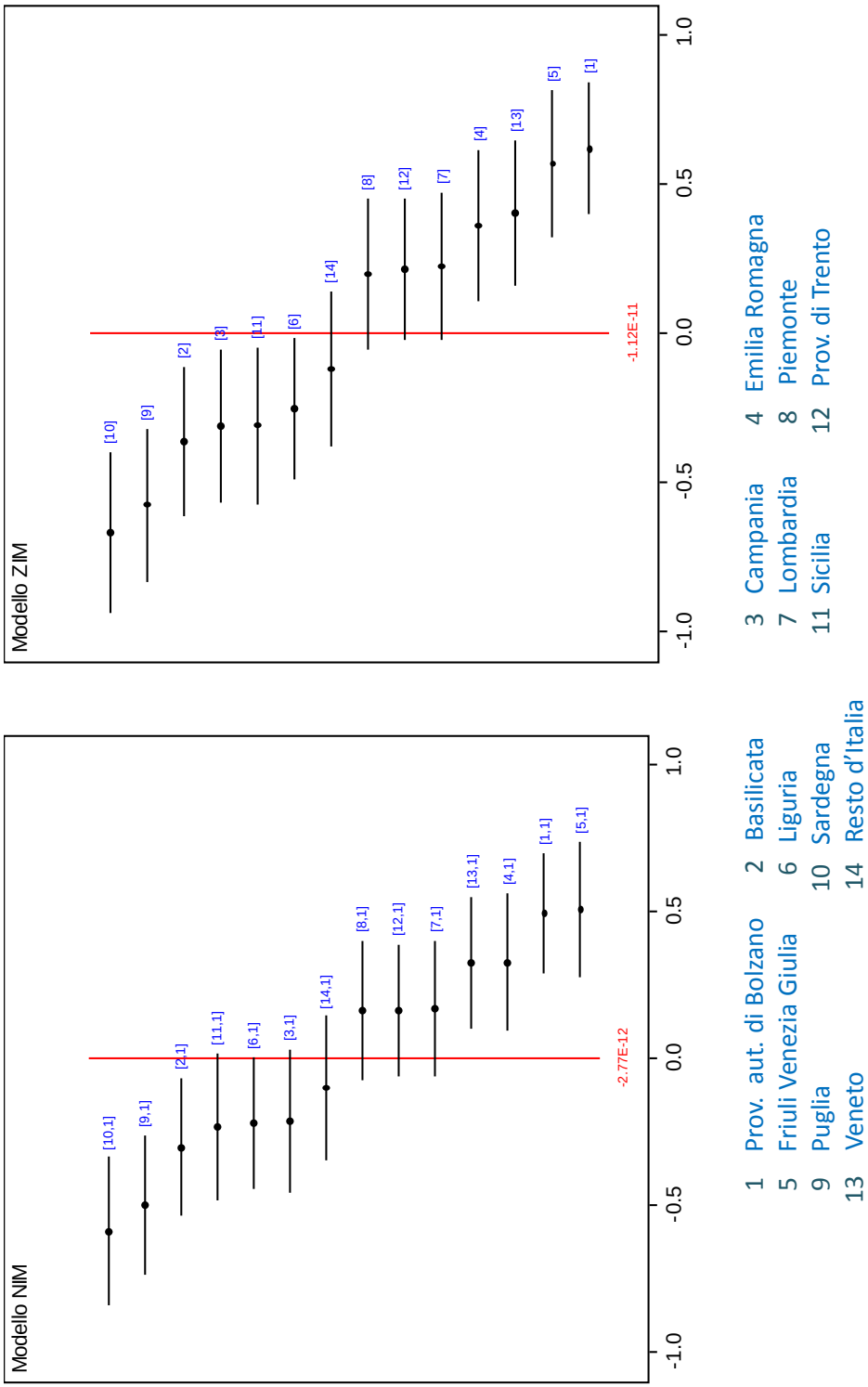


Figura 5.14: Caterpillar dei parametri di abilità di gruppo, ordinati per rango, per i modelli NIM e ZIM

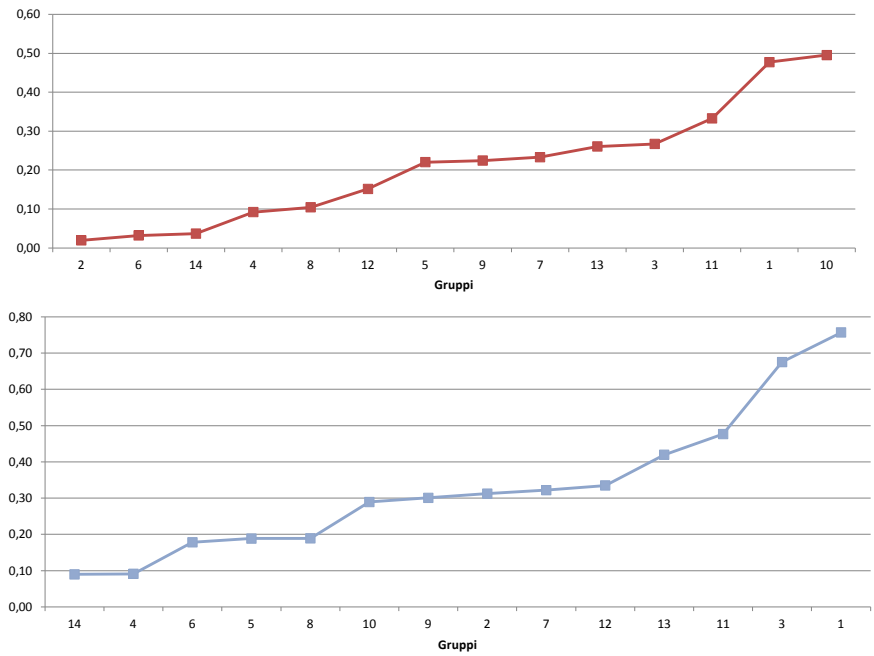


Figura 5.15: Trend delle differenze in valore assoluto tra le medie a posteriori (standardizzate) dei parametri di abilità di gruppo del modello NIM e del modello IM (linea rossa), del modello NIM e ZIM(linea blu)

5.4 Modello NIM con covariate

Al fine di valutare l'impatto di alcune caratteristiche dello studente sui processi di apprendimento (e, in questo caso, anche sul processo della mancata risposta), è interessante considerare il ruolo delle covariate introdotte nel modello. La scelta di considerare due soli livelli gerarchici è motivata dal fatto che si è voluto inserire, come regressore, la variabile "Area Geografica di appartenenza", per cui la presenza del terzo livello gerarchico, rappresentato dal raggruppamento territoriale, sarebbe ridondante.
Le covariate di riferimento sono:

- Genere (1 se Maschio);
- Immigrato (1 se nato all'estero);
- Indice di status socio-economico-culturale (ESCS);

- Tipologia di scuola (1 se privata);
- Area Geografica di appartenenza (1 se Sud).

Nella Tabella 5.8 vengono riportate le medie a posteriori ottenute per ciascun regressore e i relativi intervalli di credibilità al 95%. Nessuno di questi ultimi contiene lo zero per cui i coefficienti sono, tutti, statisticamente significativi.
I risultati delle medie a posteriori (espresse in logit) per ciascun regressore, relativamente a ciascuna dimensione, non destano sorprese in quanto concordanti con i risultati ottenuti in letteratura.
Il quadro che ne deriva è abbastanza chiaro: i valori negativi del logit, per tre dei cinque regressori, evidenziano come le performance migliori, sia in termini di abilità che di propensione alla risposta, siano associate a studenti non stranieri, a coloro che frequentano le scuole pubbliche e a gli studenti del Nord.
I due logit positivi sottolineano che gli studenti maschi hanno performance superiori rispetto alle proprie coetanee e che il contesto sociale e familiare ha un peso non trascurabile sulla formazione dello studente. Più è alto tale indicatore (ESCS), maggiore è l'abilità e la propensione alla risposta dello studente.
Questo studio mette in evidenza la stretta relazione tra abilità e propensione alla risposta attraverso l'utilizzo delle covariate; ciò che influenza l'abilità influenza anche la propensione alla risposta: i segni, infatti, sono sempre concordanti per le due dimensioni.
L'ultimo regressore, riferito all'area geografica di appartenenza, statisticamente significativo e negativo, è un'ulteriore evidenza e conferma del fatto che la stratificazione territoriale non è meramente geografica, anzi rispecchia declinazioni culturali e contrapposizioni sociali ben più profonde e radicate.
In appendice B, le Tabelle B.7 e B.8, riferite alla matrice di varianza e covarianza e ai parametri di difficoltà, sono riportate a completamento dell'analisi e non evidenziano differenze sostanziali rispetto al modello NIM *unconditional*. In appendice A.5, invece, è riportato il codice Bugs utilizzato.

5.5 Analisi del DIF: confronto tra modelli

Come già ampiamente trattato nel quarto capitolo, tra gli obiettivi di questo studio, vi è anche quello di confrontare i tre modelli, oggetto di interesse, in relazione al *Differential Item Functioning*.
Tale analisi vuole valutare se i tre diversi approcci al trattamento dei dati mancanti abbiano delle ripercussioni in relazione all'individuazione di eventuali items distorti.

Tabella 5.8: Modello NIM con covariate: media a posteriori, deviazione standard e intervalli di credibilità per i regressori.

	Media	S. D.	2.5pc	Mediana	97.5pc
$b_{gen,1}$	0.226	0.065	0.098	0.227	0.353
$b_{gen,2}$	0.318	0.112	0.098	0.318	0.536
$b_{imm,1}$	−0.672	0.153	−0.972	−0.672	−0.373
$b_{imm,2}$	−0.692	0.254	−1.193	−0.691	−0.194
$b_{escs,1}$	0.281	0.035	0.211	0.281	0.350
$b_{escs,2}$	0.467	0.061	0.350	0.467	0.586
$b_{sc,1}$	−0.659	0.143	−0.939	−0.658	−0.380
$b_{sc,2}$	−0.776	0.239	−1.245	−0.776	−0.307
$b_{ag,1}$	−0.680	0.069	−0.815	−0.680	−0.545
$b_{ag,2}$	−1.302	0.118	−1.535	−1.301	−1.073

Anche per l'analisi del DIF si considera un modello a due livelli, dal momento che si è deciso di utilizzare come *reference* e *focal group*, rispettivamente, il Nord e il Sud. La scelta è tutt'altro che casuale, è anzi motivata e suffragata dai precedenti risultati, i quali mettono in evidenza contrapposizioni e differenze tra i gruppi in questione.
Il primo dato analizzato fa riferimento al parametro α^{adj}; dalla Tabella 5.9 è possibile notare che, per tutti i modelli considerati, esso assume un valore (in logit) negativo e statisticamente significativo: la lettura di questo dato è immediata in quanto testimonia, ancora una volta, il gap tra le abilità dei due gruppi, a sfavore ovviamente delle regioni del Sud. Da notare che, anche in questo caso, il valore del parametro del modello bivariato è compreso tra i valori dei parametri degli altri due modelli.

Tabella 5.9: Media a posteriori, deviazione standard e intervalli di credibilità per il parametro α^{adj} nei tre modelli

	Media	S. D.	2.5pc	Mediana	97.5pc	P-Value
Modello NIM						
α^{adj}	−0.618	0.075	−0.767	−0.618	−0.471	0.000
Modello IM						
α^{adj}	−0.571	0.073	−0.713	−0.571	−0.429	0.000
Modello ZIM						
α^{adj}	−0.786	0.078	−0.939	−0.786	−0.633	0.000

Al fine di valutare se l'assunzione di ignorabilità del meccanismo generatore del dato mancante abbia conseguenze, relativamente all'individuazione degli items distorti, è bene confrontare i valori dell'effetto fisso λ^{adj} per tutti e tre i modelli. Più che la loro magnitudo, è interessante vedere quali di essi risultino statisticamente significativi, ossia quali intervalli di credibilità non contengano lo zero.
Dall'analisi delle Tabelle 5.10, 5.11 e 5.12 emerge chiaramente come i tre modelli percepiscano la presenza di items distorti in maniera differente. Gli asterischi, in corrispondenza del P-value inferiore a 0.05, evidenziano gli items con DIF.
In particolare, il modello NIM individua due items distorti, il 6 e il 7; essendo il valore del parametro positivo, la distorsione è a sfavore del focal group, nel senso che gli studenti del Sud percepiscono i due quesiti come più difficili, a parità di livello di abilità.
Per il modello IM, gli items distorti, oltre a quelli individuati dal modello bivariato, sono anche il 3 e l'11. Per il primo la distorsione sfavorisce il gruppo di studio, per il secondo sfavorisce il gruppo di riferimento.
Il modello ZIM è quello che più si allontana dalle due situazioni precedenti. Per esso, gli item con DIF sono l'1, il 2 e il 10; i primi due con distorsione a sfavore del gruppo del Nord, il terzo a sfavore del gruppo del Sud.
In generale, appare evidente che i tre modelli definiscano profili diversi per questo genere di analisi e che l'individuazione del DIF sia sensibile alle diverse ipotesi formulate per il meccanismo generatore del dato mancante.
A corredo dell'analisi, in appendice B, sono riportate le tabelle relative ai parametri di difficoltà β^{adj}[24] e alle varianze e covarianze dei tre modelli.

5.6 Conclusioni

Al termine di questo studio, è opportuno verificare se, e in che misura, gli obiettivi di partenza siano stati raggiunti.
Si è dimostrato, in primis, che è possibile utilizzare il linguaggio dei modelli grafici bayesiani per rappresentare e, quindi, definire un modello IRT bivariato (oltre che univariato). La flessibilità e la potenza del linguaggio grafico hanno permesso di rendere più chiari ed evidenti i rapporti di indipendenza (condizionata) propri della classe dei modelli di Rasch.
L'estensione degli stessi, al caso multilivello, ha aggiunto un grado di com-

[24] Si ricorda che essi sono gli effetti fissi, i parametri di difficoltà per il gruppo di riferimento, ossia il Nord in questo caso.

plessità all'analisi, necessario affinché i modelli in questione possano trovare la più larga applicazione e diffusione.
Un secondo obiettivo, connesso al precedente, è stato quello di utilizzare tali modelli per definire tre diversi approcci al trattamento dei dati mancanti.
Il modello di Rasch, esteso al caso bidimensionale, è stato definito con lo scopo di considerare un modello che tenesse conto del meccanismo generatore del dato mancante, che supponesse, quindi, una violazione del principio di ignorabilità, così come introdotto da Rubin.
In contrapposizione ad esso, sono stati definiti altri due modelli (con acronimi IM e ZIM), l'uno che ha ipotizzato un meccanismo di tipo MAR poiché ha sfruttato la capacità del software di imputare i valori mancanti sulla base della distribuzione predittiva a posteriori, condizionata ai soli dati osservati e ai parametri non noti del modello, l'altro che ha proceduto con l'imputazione dello zero al posto del dato mancante.
Nel modello bivariato NIM, la correlazione tra la prima dimensione, l'abilità, e la seconda dimensione, la propensione alla risposta, è risultata statisticamente significativa, sia per il secondo che per il terzo livello gerarchico.
Ciò ha confortato l'ipotesi iniziale secondo la quale, in matrici di dati come quelle derivanti dal contesto educativo, ha senso supporre un meccanismo generatore del dato mancante tutt'altro che casuale; ha acquistato vigore la tesi secondo la quale a maggiore abilità corrisponde maggiore propensione alla risposta.
Dal momento che l'obiettivo specifico dell'analisi è il confronto delle stime derivanti dai tre modelli, e in particolare l'individuazione di distorsioni delle stesse in caso di dati MNAR, il modello NIM è stato considerato come modello di paragone.
Come argomentato nel presente capitolo, i risultati hanno evidenziato che le variazioni nelle stime sia dei parametri di difficoltà che dei parametri di abilità dei due modelli univariati rispetto al modello bivariato, sono trasversali nel senso che interessano tutti gli items e i gruppi coinvolti.
Va sottolineato che il grado distorsione è, però, strettamente connesso a due fattori: la forza del legame tra le dimensioni latenti e il numero dei dati mancanti nel campione. Nel caso specifico, la media a posteriori della correlazione tra le dimensioni, per entrambi i livelli, è risultata circa pari al 68%, quindi di medio-alta intensità; la percentuale di missing totali è risultata, invece, non molto consistente, poiché pari al 17%.
In relazione alle stime ottenute sui dati analizzati, è emerso che il modello NIM, che tiene conto del meccanismo generatore del dato mancante, occupa una posizione mediana: restituisce cioè stime dei parametri che sono sempre comprese tra i valori prodotti dagli altri due modelli.
In relazione ai parametri di difficoltà, considerando sempre come riferimento

le stime del modello bivariato, si nota, infatti, che il modello IM tende a sovrastimare i valori degli items con logit negativo e a sottostimare quelli con logit positivo, a differenza del modello ZIM, che assume il comportamento opposto.
In relazione alle stime dei parametri di abilità di gruppo, il modello NIM effettua una sorta di ponderazione delle stesse, migliorando le stime dei gruppi più virtuosi (con punteggi elevati e pochi dati mancanti) e riducendo ulteriormente quelle dei gruppi mediocri (con punteggi bassi e molti dati mancanti), rispetto al modello IM che considera un meccanismo ignorabile.
È, però, certo che, dei due modelli univariati, quello che produce distorsioni più evidenti è il modello ZIM, che considera come errate le mancate risposte: esso acuisce le distanze tra le posizioni estreme, attribuendo punteggi ancor più elevati del modello NIM alle regioni virtuose, declassando ulteriormente le regioni con basse performance e numero di dati mancanti elevato.
In ultimo, è bene porre l'accento sul fatto che, pur non avendo una percentuale totale di dati mancanti consistente, ciò che è emerso dallo studio è che le distorsioni delle stime, in alcuni casi, hanno generato dei cambi di posizione nella graduatoria dei ranghi, sia di difficoltà che di abilità, seppure tra posizione contigue.
Come obiettivo futuro di ricerca vi è proprio l'applicazione di tali modelli a casi in cui la percentuale di dati mancanti sia molto più consistente, al fine di investigare quali ulteriori ripercussioni si possano registrare in relazione alle stime di interesse; se, per esempio, si possano verificare variazioni nei ranghi molto più consistenti.
A corredo dell'analisi, il modello bivariato a due livelli ha permesso di considerare quali caratteristiche dello studente avessero un impatto significativo sia sul processo dell'abilità che su quello della mancata risposta.
I risultati hanno confermato che vi è una differenza statisticamente significativa di genere, a favore dei maschi, e che gli studenti delle scuole pubbliche, delle regioni del Nord e non immigrati, sono quelli che ottengono i punteggi più elevati per la matematica. È ancora emerso che il contesto socio - familiare ha le sue ripercussioni sullo sviluppo delle competenze matematiche dello studente.
L'ultima parte del lavoro ha considerato l'analisi del DIF relativamente al gruppo degli studenti del Nord e del Sud.
Anche in relazione a tale studio, è emerso che, a seconda dell'approccio considerato per il trattamento dei dati mancanti, l'individuazione di eventuali items distorti, per i due gruppi, è differente.
Le ripercussioni, circa il non considerare il meccanismo generatore del dato mancante, sono rese evidenti, quindi, anche per l'analisi del *Differential Item Functioning.*

In estrema sintesi, la ricerca ha voluto sottolineare due elementi fondamentali: l'utilizzo del linguaggio grafico come alternativa equivalente per la modellizzazione statistica e l'importanza dell'analisi e trattamento della mancata risposta in matrici di dati per le quali è logico ritenere che il pattern dei *missing* non sia casuale.

Tabella 5.10: Modello DIF NIM: media a posteriori, deviazione standard e intervalli di credibilità per il parametro λ^{adj}

	Media	S. D.	$2.5pc$	Mediana	$97.5pc$	P-value
λ_1^{adj}	-0.194	0.117	-0.422	-0.193	0.036	0.098
λ_2^{adj}	-0.112	0.115	-0.336	-0.113	0.114	0.331
λ_3^{adj}	0.203	0.116	-0.022	0.203	0.431	0.077
λ_4^{adj}	0.095	0.127	-0.154	0.095	0.339	0.454
λ_5^{adj}	-0.448	0.276	-0.987	-0.450	0.103	0.106
λ_6^{adj}	0.324	0.125	0.078	0.323	0.570	0.010*
λ_7^{adj}	0.294	0.138	0.023	0.293	0.564	0.035*
λ_8^{adj}	-0.004	0.121	-0.242	-0.004	0.231	0.971
λ_9^{adj}	-0.087	0.176	-0.433	-0.087	0.257	0.619
λ_{10}^{adj}	0.315	0.170	-0.014	0.315	0.648	0.061
λ_{11}^{adj}	-0.257	0.143	-0.538	-0.258	0.0224	0.071
λ_{12}^{adj}	-0.129	0.151	-0.424	-0.128	0.166	0.391

Tabella 5.11: Modello DIF IM: media a posteriori, deviazione standard e intervalli di credibilità per il parametro λ^{adj}

	Media	S. D.	$2.5pc$	Mediana	$97.5pc$	P-value
λ_1^{adj}	−0.180	0.117	−0.408	−0.179	0.049	0.124
λ_2^{adj}	−0.078	0.114	−0.301	−0.077	0.145	0.492
λ_3^{adj}	0.232	0.116	0.006	0.231	0.461	0.044*
λ_4^{adj}	0.092	0.128	−0.156	0.093	0.344	0.470
λ_5^{adj}	−0.462	0.276	−0.992	−0.466	0.087	0.097
λ_6^{adj}	0.345	0.126	0.098	0.346	0.590	0.006*
λ_7^{adj}	0.313	0.137	0.044	0.314	0.583	0.023*
λ_8^{adj}	0.026	0.121	−0.208	0.026	0.262	0.834
λ_9^{adj}	−0.079	0.176	−0.425	−0.078	0.271	0.654
λ_{10}^{adj}	0.270	0.170	−0.062	0.269	0.605	0.109
λ_{11}^{adj}	−0.306	0.143	−0.587	−0.305	−0.027	0.031*
λ_{12}^{adj}	−0.175	0.151	−0.472	−0.175	0.118	0.245

Tabella 5.12: Modello DIF ZIM: media a posteriori, deviazione standard e intervalli di credibilità per il parametro λ^{adj}

	Media	S. D.	$2.5pc$	Mediana	$97.5pc$	P-value
λ_1^{adj}	−0.236	0.116	−0.465	−0.236	−0.010	0.041*
λ_2^{adj}	−0.250	0.116	−0.478	−0.249	−0.022	0.033*
λ_3^{adj}	0.073	0.117	−0.155	0.074	0.304	0.534
λ_4^{adj}	0.076	0.117	−0.156	0.0762	0.304	0.512
λ_5^{adj}	−0.396	0.270	−0.914	−0.403	0.145	0.151
λ_6^{adj}	0.221	0.119	−0.013	0.220	0.455	0.064
λ_7^{adj}	0.191	0.132	−0.066	0.191	0.455	0.146
λ_8^{adj}	−0.150	0.118	−0.384	−0.150	0.081	0.200
λ_9^{adj}	−0.205	0.165	−0.522	−0.207	0.121	0.216
λ_{10}^{adj}	0.484	0.159	0.176	0.484	0.799	0.002*
λ_{11}^{adj}	0.116	0.116	−0.112	0.115	0.344	0.319
λ_{12}^{adj}	0.075	0.133	−0.186	0.074	0.336	0.578

Appendice A

Codice Bugs per i modelli grafici

A.1 Modello unidimensionale

```
model {

    for( l in 1 : n ) {
            y[l] ~ dbern(p[l])
            logit(p[l]) <- theta3[reg[l]] + theta2[std[l]] - beta[x[l]]
                }

    for( i in 1 : I ) {
        beta[i] ~ dnorm(0.0, 1.0E-4)
        beta.adj[i] <- beta[i] - mean(beta[ ])
                }

    for( j in 1 : J ) {
        theta2[j] ~ dnorm(gamma, tau.theta2)
        theta2.adj[j] <- theta2[j] - mu.beta + mu.theta3
                }

    for( g in 1 : G ) {
        theta3[g] ~ dnorm(0.0, tau.theta3)
        theta3.adj[g] <- theta3[g] - mean(theta3[ ])
                }

    gamma ~ dnorm(0.0, 1)
    gamma.adj <- gamma - mu.beta + mu.theta3
    mu.beta <- mean(beta[ ])
    mu.theta3 <- mean(theta3[ ])
    tau.theta2 ~ dgamma(0.001, 0.001)
    tau.theta3 ~ dgamma(0.001, 0.001)
    sigma.theta2 <- 1 / tau.theta2
    sigma.theta3 <- 1 / tau.theta3
    corr1 <- sigma.theta3 / (sigma.theta3 + sigma.theta2)

        }
```

A.2 Modello bidimensionale

```
model{

    for( l in 1 : n ) {
        y[l] ~ dbern(p[l])
        d[l] ~ dbern(k[l])
        logit(p[l]) <- theta3[reg[l] , 1] + theta2[std[l] , 1] - beta[x[l]]
        logit(k[l]) <- theta3[reg[l] , 2] + theta2[std[l] , 2] - delta[x[l]]
                }

    for( i in 1 : I ) {
        beta[i] ~ dnorm(0.0, 1.0E-4)
        delta[i] ~ dnorm(0.0, 1.0E-4)
        beta.adj[i] <- beta[i] - mean(beta[ ])
        delta.adj[i] <- delta[i] - mean(delta[ ])
                }

    for( j in 1 : J ) {
        theta2[j , 1:2] ~ dmnorm(gamma[1:2], Tau.theta2[1:2 , 1:2])
        theta2.adj[j,1] <- theta2[j,1] - mu.beta + mu.theta3[1]
        theta2.adj[j,2] <- theta2[j,2] - mu.delta + mu.theta3[2]
                }

    for( g in 1 : G ) {
        theta3[g , 1:2] ~ dmnorm(mu[1:2], Tau.theta3[1:2 , 1:2])
        theta3.adj[g,1] <- theta3[g,1] - mean(theta3[ ,1])
        theta3.adj[g,2] <- theta3[g,2] - mean(theta3[ ,2])
                }


    gamma[1:2] ~ dmnorm(m[1:2], T[1:2 , 1:2])
    gamma.adj[1] <- gamma[1] - mu.beta + mu.theta3[1]
    gamma.adj[2] <- gamma[2] - mu.delta + mu.theta3[2]
    mu.beta <- mean(beta[ ])
    mu.delta <- mean(delta[ ])
    mu.theta3[1] <- mean(theta3[ , 1])
    mu.theta3[2] <- mean(theta3[ , 2])
    Tau.theta2[1:2 , 1:2] ~ dwish(R[1:2 , 1:2], 2)
    Tau.theta3[1:2 , 1:2] ~ dwish(S[1:2 , 1:2], 2)
    Sigma.theta2[1:2 , 1:2] <- inverse(Tau.theta2[1:2 , 1:2])
    Sigma.theta3[1:2 , 1:2] <- inverse(Tau.theta3[1:2 , 1:2])
    r1 <- Sigma.theta2[1 , 2] / sqrt(Sigma.theta2[1 , 1]) * sqrt(Sigma.theta2[2 , 2])
    r2 <- Sigma.theta3[1 , 2] / sqrt(Sigma.theta3[1 , 1]) * sqrt(Sigma.theta3[2 , 2])
    corr1 <- Sigma.theta3[1 , 1] / (Sigma.theta2[1 , 1] + Sigma.theta3[1 , 1])
    corr2 <- Sigma.theta3[2 , 2] / (Sigma.theta2[2 , 2] + Sigma.theta3[2 , 2])
}
```

A.3 Modello unidimensionale per il DIF

```
model {

   for( l in 1 : n ) {
         y[l] ~ dbern(p[l])
         logit(p[l]) <- theta2[std[l]] - beta[x[l]] + alpha * group[l] - lambda[x[l]] * group[l]
            }

   for( i in 1 : I ) {
      beta[i] ~ dnorm(0.0, 1.0E-4)
      lambda[i] ~ dnorm(0.0, 1.0E-4)
      beta.adj[i] <- beta[i] - mean(beta[ ])
      lambda.adj[i] <- lambda[i] - mean(lambda[])
      pv[i] <- step(lambda.adj[i])
            }

   for( j in 1 : J ) {
      theta2[j] ~ dnorm(gamma, tau.theta2)
      theta2.adj[j] <- theta2[j] - mu.beta
            }

   alpha ~ dnorm(0.0, 1.0E-4)
   alpha.adj <- alpha - mu.lambda
   gamma ~ dnorm(0.0, 1)
   gamma.adj <- gamma - mu.beta
   mu.beta <- mean(beta[ ])
   mu.lambda <- mean(lambda[])
   tau.theta2 ~ dgamma(0.001, 0.001)
   sigma.theta2 <- 1 / tau.theta2
   p_a <- step(alpha.adj)
      }
```

A.4 Modello bidimensionale per il DIF

```
model{

   for( l in 1 : n ) {
      y[l] ~ dbern(p[l])
      d[l] ~ dbern(k[l])
      logit(p[l]) <-  theta2[std[l] , 1] - beta[x[l]] + alpha[1] * group[l] - lambda[x[l]] * group[l]
      logit(k[l]) <- theta2[std[l] , 2] - delta[x[l]]+ alpha[2] * group[l] - lambda1[x[l]] * group[l]
            }

   for( i in 1 : I ) {
      beta[i] ~ dnorm(0.0, 1.0E-4)
      delta[i] ~ dnorm(0.0, 1.0E-4)
      lambda[i] ~ dnorm(0.0, 1.0E-4)
      lambda1[i] ~ dnorm(0.0, 1.0E-4)
      beta.adj[i] <- beta[i] - mean(beta[ ])
      delta.adj[i] <- delta[i] - mean(delta[ ])
      lambda.adj[i] <- lambda[i] - mean(lambda[])
      lambda1.adj[i] <- lambda1[i] - mean(lambda1[])
      pv[i] <- step(lambda.adj[i])
                  }

   for( j in 1 : J ) {
      theta2[j , 1:2] ~ dmnorm(gamma[1:2], Tau.theta2[1:2 , 1:2])
      theta2.adj[j,1] <- theta2[j,1] - mu.beta
      theta2.adj[j,2] <- theta2[j,2] - mu.delta
            }

   alpha[1:2] ~ dmnorm(m1[1:2], S[1:2,1:2])
   alpha.adj [1]<- alpha[1] - mu.lambda
   alpha.adj[2] <- alpha[2] - mu.lambda1
   gamma[1:2] ~ dmnorm(m[1:2], T[1:2 , 1:2])
   gamma.adj[1] <- gamma[1] - mu.beta
   gamma.adj[2] <- gamma[2] - mu.delta
   mu.beta <- mean(beta[ ])
   mu.delta <- mean(delta[ ])
   mu.lambda <- mean(lambda[])
   mu.lambda1 <- mean(lambda1[])
   Tau.theta2[1:2 , 1:2] ~ dwish(R[1:2 , 1:2], 2)
   Sigma.theta2[1:2 , 1:2] <- inverse(Tau.theta2[1:2 , 1:2])
   r1 <- Sigma.theta2[1 , 2] /( sqrt(Sigma.theta2[1 , 1]) * sqrt(Sigma.theta2[2 , 2]))
   p_a <- step(alpha.adj[1])
}
```

A.5 Modello bidimensionale con le covariate

```
model{

    for( l in 1 : n ) {
        y[l] ~ dbern(p[l])
        d[l] ~ dbern(k[l])
        logit(p[l]) <- theta2[std[l] , 1] - beta[x[l]]
        logit(k[l]) <- theta2[std[l] , 2] - delta[x[l]]
                }

    for( i in 1 : I ) {
        beta[i] ~ dnorm(0.0, 1.0E-4)
        delta[i] ~ dnorm(0.0, 1.0E-4)
        beta.adj[i] <- beta[i] - mean(beta[ ])
        delta.adj[i] <- delta[i] - mean(delta[ ])
                }

    for( j in 1 : J ) {
        theta2[j , 1:2] ~ dmnorm(gamma[j,1:2], Tau.theta2[1:2 , 1:2])
        theta2.adj[j,1] <- theta2[j,1] - mu.beta
        theta2.adj[j,2] <- theta2[j,2] - mu.delta
        gamma[j,1] <-  b1[1]*sex[j]+b2[1]*imm[j]+b3[1]*(escs[j]-mean(escs[]))+b4[1]*sc[j]+b5[1]*ma[j]
        gamma[j,2]<-   b1[2]*sex[j]+b2[2]*imm[j]+b3[2]*(escs[j]-mean(escs[]))+b4[2]*sc[j]+b5[2]*ma[j]
        gamma.adj[j,1] <- gamma[j,1] - mu.beta
        gamma.adj[j,2] <- gamma[j,2] - mu.delta
                }

    for( m in 1 : 2) {
    b1[m] ~ dnorm(0,0.0001)
    b2[m] ~ dnorm(0,0.0001)
    b3[m] ~ dnorm(0,0.0001)
    b4[m] ~ dnorm(0,0.0001)
    b5[m] ~ dnorm(0,0.0001)
    }

    mu.beta <- mean(beta[ ])
    mu.delta <- mean(delta[ ])
    Tau.theta2[1:2 , 1:2] ~ dwish(R[1:2 , 1:2], 2)
    Sigma.theta2[1:2 , 1:2] <- inverse(Tau.theta2[1:2 , 1:2])
    r1 <- Sigma.theta2[1 , 2] / (sqrt(Sigma.theta2[1 , 1]) * sqrt(Sigma.theta2[2 , 2]))
}
```

Appendice B

Tabelle aggiuntive

Tabella B.1: Modello NIM: media a posteriori, deviazione standard e intervalli di credibilità per i parametri di difficoltà

	Media	S. D.	$2.5pc$	Mediana	$97.5pc$
β_1^{adj}	-0.121	0.057	-0.232	-0.121	-0.009
β_2^{adj}	0.241	0.056	0.132	0.241	0.351
β_3^{adj}	0.182	0.056	0.073	0.182	0.291
β_4^{adj}	-0.978	0.062	-1.101	-0.978	-0.857
β_5^{adj}	3.397	0.129	3.149	3.394	3.655
β_6^{adj}	-1.117	0.0618	-1.238	-1.117	-0.996
β_7^{adj}	0.806	0.064	0.681	0.806	0.932
β_8^{adj}	-0.991	0.060	-1.108	-0.991	-0.875
β_9^{adj}	-2.871	0.089	-3.048	-2.871	-2.699
β_{10}^{adj}	1.397	0.073	1.255	1.397	1.541
β_{11}^{adj}	-0.662	0.067	-0.794	-0.661	-0.531
β_{12}^{adj}	0.717	0.069	0.582	0.717	0.853

Tabella B.2: Modello NIM: media a posteriori, deviazione standard e intervalli di credibilità per l'effetto random di terzo livello

	Media	S. D.	$2.5pc$	Mediana	$97.5pc$
$\theta_{1,1}^{adj,(3)}$	0.495	0.102	0.294	0.494	0.694
$\theta_{2,1}^{adj,(3)}$	-0.302	0.118	-0.531	-0.302	-0.071
$\theta_{3,1}^{adj,(3)}$	-0.213	0.122	-0.452	-0.213	0.027
$\theta_{4,1}^{adj,(3)}$	0.327	0.118	0.096	0.326	0.560
$\theta_{5,1}^{adj,(3)}$	0.507	0.116	0.279	0.507	0.735
$\theta_{6,1}^{adj,(3)}$	-0.218	0.111	-0.437	-0.218	-0.001
$\theta_{7,1}^{adj,(3)}$	0.171	0.115	-0.054	0.171	0.397
$\theta_{8,1}^{adj,(3)}$	0.163	0.119	-0.069	0.162	0.396
$\theta_{9,1}^{adj,(3)}$	-0.499	0.120	-0.735	-0.498	-0.263
$\theta_{10,1}^{adj,(3)}$	-0.588	0.127	-0.838	-0.588	-0.338
$\theta_{11,1}^{adj,(3)}$	-0.232	0.125	-0.478	-0.232	0.012
$\theta_{12,1}^{adj,(3)}$	0.164	0.111	-0.054	0.164	0.382
$\theta_{13,1}^{adj,(3)}$	0.326	0.113	0.105	0.326	0.548
$\theta_{14,1}^{adj,(3)}$	-0.101	0.123	-0.342	-0.101	0.140

* 1. Prov. aut. Bolzano; 2. Basilicata; 3. Campania; 4. Emilia Romagna; 5. Friuli V. Giulia; 6. Liguria; 7. Lombardia; 8. Piemonte; 9. Puglia; 10. Sardegna; 11. Sicilia; 12. Prov. aut. Trento; 13. Veneto; 14. Resto d'Italia.

Tabella B.3: Modello IM: media a posteriori, deviazione standard e intervalli di credibilità per i parametri di difficoltà

	Media	S. D.	$2.5pc$	Mediana	$97.5pc$
β_1^{adj}	-0.086	0.056	-0.197	-0.086	0.024
β_2^{adj}	0.294	0.055	0.187	0.294	0.403
β_3^{adj}	0.236	0.055	0.128	0.236	0.344
β_4^{adj}	-0.971	0.0621	-1.093	-0.971	-0.850
β_5^{adj}	3.311	0.129	3.063	3.308	3.569
β_6^{adj}	-1.072	0.062	-1.193	-1.072	-0.951
β_7^{adj}	0.786	0.064	0.662	0.786	0.912
β_8^{adj}	-0.941	0.059	-1.058	-0.941	-0.826
β_9^{adj}	-2.799	0.089	-2.977	-2.799	-2.626
β_{10}^{adj}	1.337	0.073	1.194	1.337	1.481
β_{11}^{adj}	-0.727	0.068	-0.860	-0.727	-0.595
β_{12}^{adj}	0.633	0.069	0.498	0.633	0.769

Tabella B.4: Modello IM: media a posteriori, deviazione standard e intervalli di credibilità per l'effetto random di terzo livello

	Media	S. D.	$2.5pc$	Mediana	$97.5pc$
$\theta_1^{adj,(3)}$	0.431	0.099	0.239	0.430	0.624
$\theta_2^{adj,(3)}$	-0.295	0.114	-0.519	-0.294	-0.073
$\theta_3^{adj,(3)}$	-0.171	0.116	-0.399	-0.170	0.056
$\theta_4^{adj,(3)}$	0.324	0.114	0.103	0.323	0.549
$\theta_5^{adj,(3)}$	0.464	0.112	0.246	0.464	0.686
$\theta_6^{adj,(3)}$	-0.213	0.107	-0.424	-0.213	-0.005
$\theta_7^{adj,(3)}$	0.138	0.110	-0.077	0.138	0.354
$\theta_8^{adj,(3)}$	0.144	0.114	-0.079	0.144	0.369
$\theta_9^{adj,(3)}$	-0.457	0.117	-0.687	-0.456	-0.230
$\theta_{10}^{adj,(3)}$	-0.510	0.124	-0.753	-0.509	-0.270
$\theta_{11}^{adj,(3)}$	-0.183	0.120	-0.418	-0.183	0.052
$\theta_{12}^{adj,(3)}$	0.141	0.107	-0.068	0.141	0.351
$\theta_{13}^{adj,(3)}$	0.284	0.108	0.074	0.284	0.498
$\theta_{14}^{adj,(3)}$	-0.100	0.118	-0.331	-0.100	0.130

* 1. Prov. aut. Bolzano; 2. Basilicata; 3. Campania; 4. Emilia Romagna; 5. Friuli V. Giulia; 6. Liguria; 7. Lombardia; 8. Piemonte; 9. Puglia; 10. Sardegna; 11. Sicilia; 12. Prov. aut. Trento; 13. Veneto; 14. Resto d'Italia.

Tabella B.5: Modello ZIM: media a posteriori, deviazione standard e intervalli di credibilità per i parametri di difficoltà

	Media	S. D.	$2.5pc$	Mediana	$97.5pc$
β_1^{adj}	-0.293	0.056	-0.403	-0.293	-0.183
β_2^{adj}	-0.008	0.056	-0.118	-0.008	0.102
β_3^{adj}	-0.070	0.056	-0.180	-0.070	0.041
β_4^{adj}	-0.980	0.057	-1.092	-0.980	-0.868
β_5^{adj}	3.571	0.127	3.330	3.569	3.824
β_6^{adj}	-1.202	0.059	-1.317	-1.202	-1.088
β_7^{adj}	0.889	0.061	0.769	0.888	1.009
β_8^{adj}	-1.178	0.058	-1.293	-1.178	-1.064
β_9^{adj}	-3.065	0.084	-3.233	-3.064	-2.902
β_{10}^{adj}	1.556	0.069	1.421	1.556	1.693
β_{12}^{adj}	-0.239	0.056	-0.347	-0.239	-0.129
β_{11}^{adj}	1.020	0.062	0.898	1.020	1.142

Tabella B.6: Modello ZIM: media a posteriori, deviazione standard e intervalli di credibilità per l'effetto random di terzo livello

	Media	S. D.	$2.5pc$	Mediana	$97.5pc$
$\theta_1^{adj,(3)}$	0.618	0.110	0.403	0.619	0.836
$\theta_2^{adj,(3)}$	-0.360	0.125	-0.606	-0.360	-0.115
$\theta_3^{adj,(3)}$	-0.310	0.129	-0.561	-0.310	-0.059
$\theta_4^{adj,(3)}$	0.361	0.127	0.112	0.361	0.609
$\theta_5^{adj,(3)}$	0.568	0.125	0.324	0.568	0.814
$\theta_6^{adj,(3)}$	-0.253	0.118	-0.486	-0.253	-0.021
$\theta_7^{adj,(3)}$	0.225	0.124	-0.017	0.225	0.469
$\theta_8^{adj,(3)}$	0.199	0.128	-0.051	0.199	0.449
$\theta_9^{adj,(3)}$	-0.573	0.129	-0.827	-0.573	-0.321
$\theta_{10}^{adj,(3)}$	-0.667	0.136	-0.934	-0.666	-0.402
$\theta_{11}^{adj,(3)}$	-0.308	0.132	-0.567	-0.307	-0.050
$\theta_{12}^{adj,(3)}$	0.216	0.119	-0.017	0.216	0.451
$\theta_{13}^{adj,(3)}$	0.403	0.122	0.165	0.403	0.642
$\theta_{14}^{adj,(3)}$	-0.119	0.131	-0.375	-0.119	0.139

* 1. Prov. aut. Bolzano; 2. Basilicata; 3. Campania; 4. Emilia Romagna; 5. Friuli V. Giulia; 6. Liguria; 7. Lombardia; 8. Piemonte; 9. Puglia; 10. Sardegna; 11. Sicilia; 12. Prov. aut. Trento; 13. Veneto; 14. Resto d'Italia.

Tabella B.7: Modello NIM con covariate: media a posteriori, deviazione standard e intervalli di credibilità per i parametri di difficoltà.

	Media	S. D.	$2.5pc$	Mediana	$97.5pc$
β_1^{adj}	-0.120	0.057	-0.232	-0.120	-0.009
β_2^{adj}	0.245	0.056	0.135	0.244	0.355
β_3^{adj}	0.191	0.0558	0.082	0.191	0.300
β_4^{adj}	-0.964	0.062	-1.087	-0.964	-0.843
β_5^{adj}	3.385	0.129	3.138	3.383	3.643
β_6^{adj}	-1.111	0.062	-1.234	-1.111	-0.989
β_7^{adj}	0.818	0.064	0.692	0.818	0.944
β_8^{adj}	-1.003	0.060	-1.122	-1.003	-0.886
β_9^{adj}	-2.904	0.092	-3.088	-2.904	-2.726
β_{10}^{adj}	1.399	0.073	1.256	1.398	1.542
β_{11}^{adj}	-0.651	0.068	-0.784	-0.650	-0.518
β_{12}^{adj}	0.717	0.069	0.581	0.717	0.852

Tabella B.8: Modello NIM con covariate: media a posteriori, deviazione standard e intervalli di credibilità per la matrice di varianza e covarianza (e per il coefficiente di correlazione) dell'effetto random

	Media	S. D.	$2.5pc$	Mediana	$97.5pc$
$\sigma^2_{1,\theta^{(2)}}$	1.081	0.071	0.948	1.079	1.225
$\sigma_{12,\theta^{(2)}}$	1.275	0.089	1.106	1.273	1.453
$\sigma^2_{2,\theta^{(2)}}$	3.520	0.226	3.097	3.513	3.984
$\rho_{student}$	0.654	0.029	0.595	0.655	0.709

Tabella B.9: Modello DIF NIM: media a posteriori, deviazione standard e intervalli di credibilità per i parametri di difficoltà

	Media	S. D.	$2.5pc$	Mediana	$97.5pc$
β_1^{adj}	-0.047	0.071	-0.185	-0.047	0.091
β_2^{adj}	0.284	0.070	0.148	0.284	0.421
β_3^{adj}	0.109	0.069	-0.026	0.109	0.245
β_4^{adj}	-1.013	0.080	-1.171	-1.012	-0.857
β_5^{adj}	3.525	0.157	3.229	3.522	3.842
β_6^{adj}	-1.251	0.082	-1.412	-1.251	-1.094
β_7^{adj}	0.713	0.078	0.561	0.713	0.869
β_8^{adj}	-0.986	0.077	-1.140	-0.986	-0.835
β_9^{adj}	-2.827	0.120	-3.065	-2.826	-2.599
β_{10}^{adj}	1.319	0.085	1.154	1.319	1.485
β_{11}^{adj}	-0.579	0.081	-0.741	-0.578	-0.422
β_{12}^{adj}	0.753	0.082	0.594	0.753	0.914

Tabella B.10: Modello DIF IM: media a posteriori, deviazione standard e intervalli di credibilità per i parametri di difficoltà

	Media	S. D.	$2.5pc$	Mediana	$97.5pc$
β_1^{adj}	-0.016	0.070	-0.153	-0.016	0.121
β_2^{adj}	0.328	0.069	0.192	0.328	0.462
β_3^{adj}	0.156	0.070	0.020	0.156	0.292
β_4^{adj}	-1.003	0.080	-1.161	-1.003	-0.847
β_5^{adj}	3.443	0.154	3.147	3.441	3.753
β_6^{adj}	-1.213	0.082	-1.374	-1.213	-1.055
β_7^{adj}	0.689	0.078	0.538	0.689	0.842
β_8^{adj}	-0.949	0.077	-1.101	-0.948	-0.800
β_9^{adj}	-2.759	0.119	-2.997	-2.757	-2.528
β_{10}^{adj}	1.271	0.084	1.108	1.271	1.435
β_{11}^{adj}	-0.628	0.081	-0.790	-0.628	-0.469
β_{12}^{adj}	0.682	0.083	0.522	0.682	0.844

Tabella B.11: Modello DIF ZIM: media a posteriori, deviazione standard e intervalli di credibilità per i parametri di difficoltà

	Media	S. D.	$2.5pc$	Mediana	$97.5pc$
β_1^{adj}	-0.208	0.071	-0.346	-0.207	-0.069
β_2^{adj}	0.080	0.071	-0.058	0.080	0.217
β_3^{adj}	-0.100	0.070	-0.237	-0.100	0.036
β_4^{adj}	-1.015	0.074	-1.162	-1.015	-0.870
β_5^{adj}	3.684	0.152	3.396	3.681	3.992
β_6^{adj}	-1.300	0.077	-1.454	-1.300	-1.149
β_7^{adj}	0.825	0.074	0.683	0.825	0.970
β_8^{adj}	-1.120	0.075	-1.268	-1.121	-0.972
β_9^{adj}	-2.975	0.113	-3.203	-2.974	-2.757
β_{10}^{adj}	1.421	0.081	1.262	1.421	1.579
β_{11}^{adj}	-0.285	0.070	-0.423	-0.285	-0.150
β_{12}^{adj}	0.994	0.075	0.847	0.994	1.140

Tabella B.12: Modelli DIF: tabella riassuntiva della Media a posteriori, deviazione standard e intervalli di credibilità per le varianze (nel modello DIF NIM è riportata la matrice di varianza e covarianza nonché il coefficiente di correlazione) dell'effetto random.

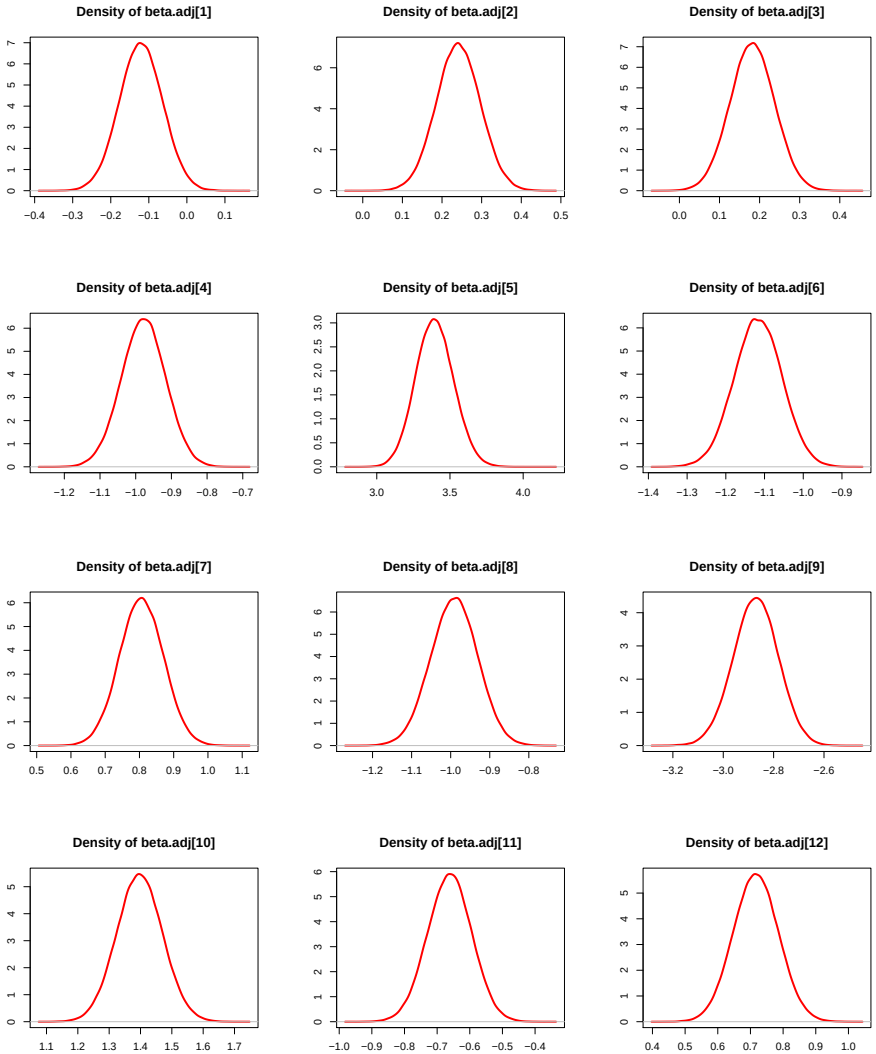
		Media	S. D.	$2.5pc$	Mediana	$97.5pc$
Modello NIM	$\sigma^2_{1,\theta^{(2)}}$	1.282	0.081	1.131	1.279	1.446
	$\sigma_{12,\theta^{(2)}}$	1.560	0.100	1.371	1.557	1.762
	$\sigma^2_{2,\theta^{(2)}}$	3.996	0.252	3.527	3.987	4.514
	$\rho_{student}$	0.690	0.026	0.638	0.690	0.738
Modello IM	$\sigma^2_{\theta^{(2)}}$	1.181	0.075	1.040	1.179	1.332
Modello ZIM	$\sigma^2_{\theta^{(2)}}$	1.620	0.089	1.452	1.617	1.802

Appendice C

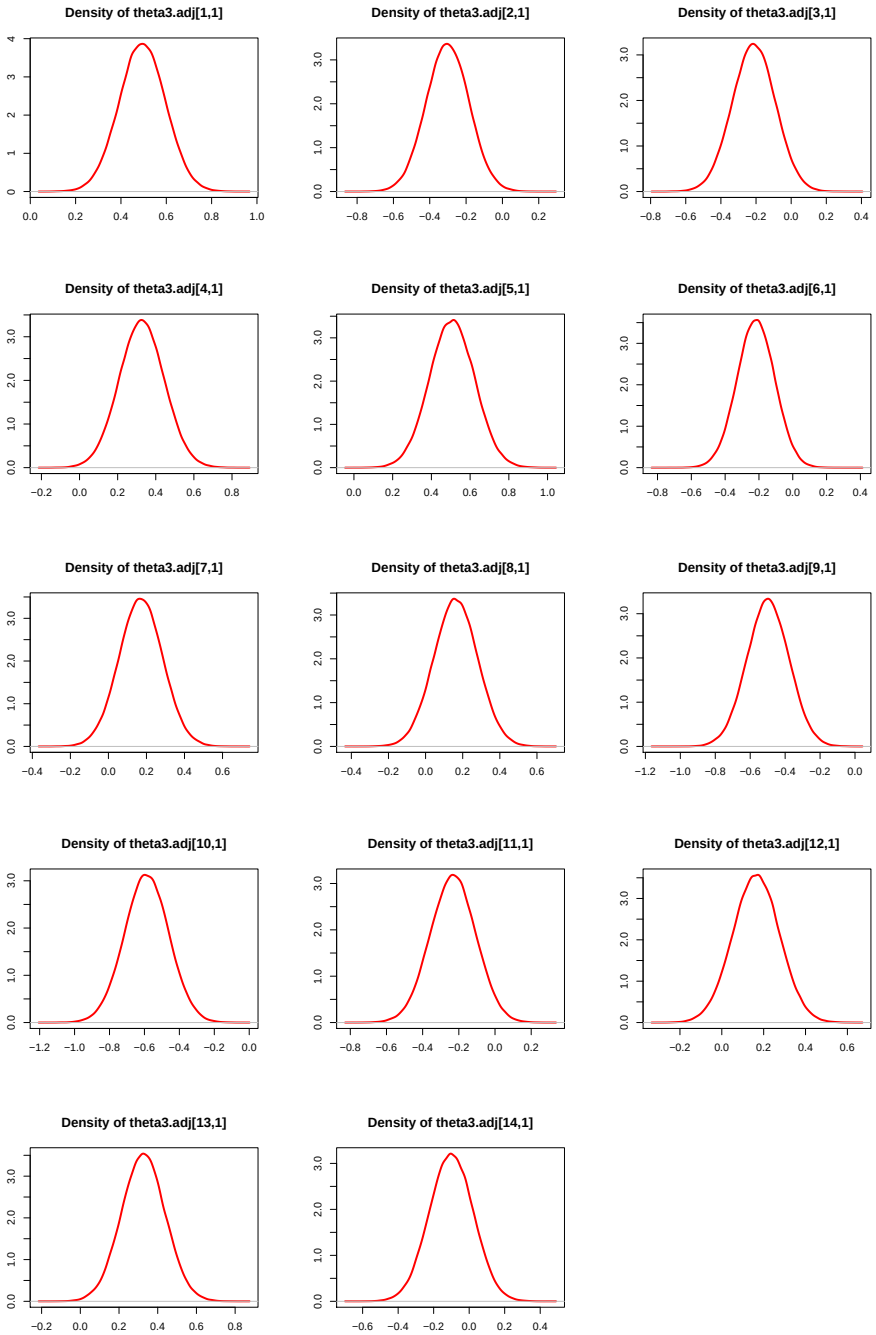
Stime kernel di Densità a posteriori

C.1 Modello NIM

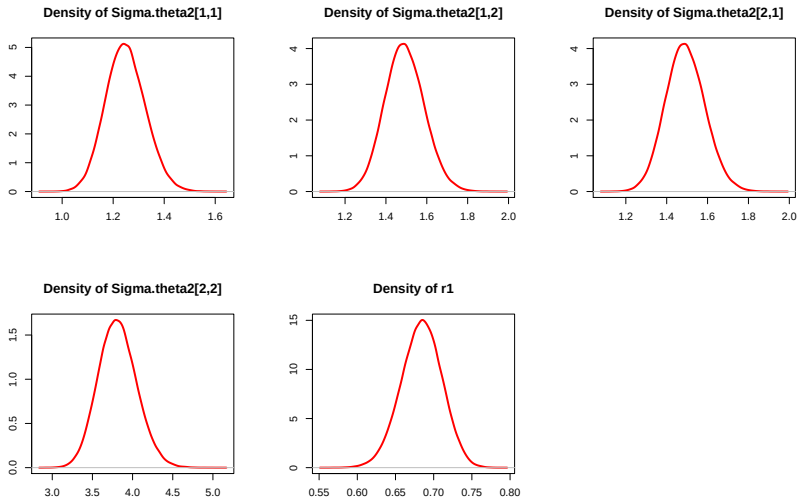
Parametri di difficoltà β^{adj}



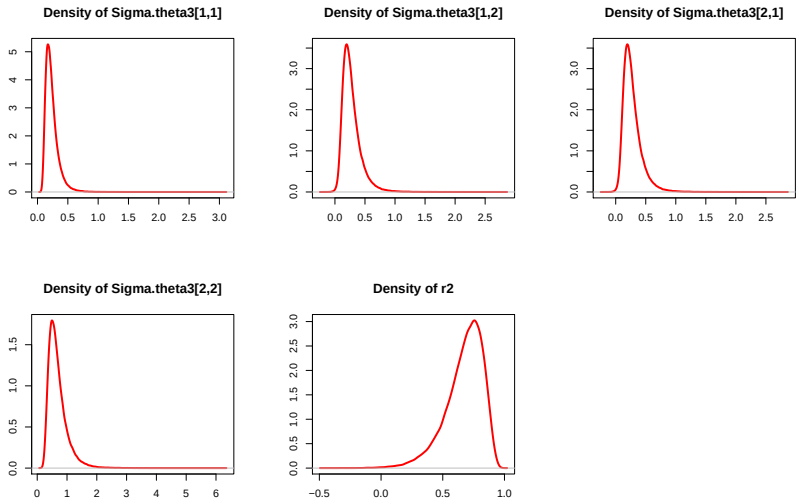
parameri di abilità $\theta^{(3),adj}$



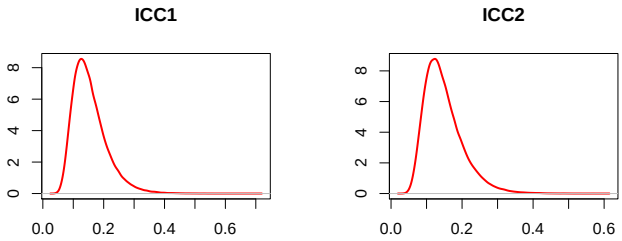
Matrice di varianza e covarianza di $\theta^{(2),adj}$ e coefficiente di correlazione $\rho_{student}$



Matrice di varianza e covarianza di $\theta^{(3),adj}$ e coefficiente di correlazione ρ_{group}

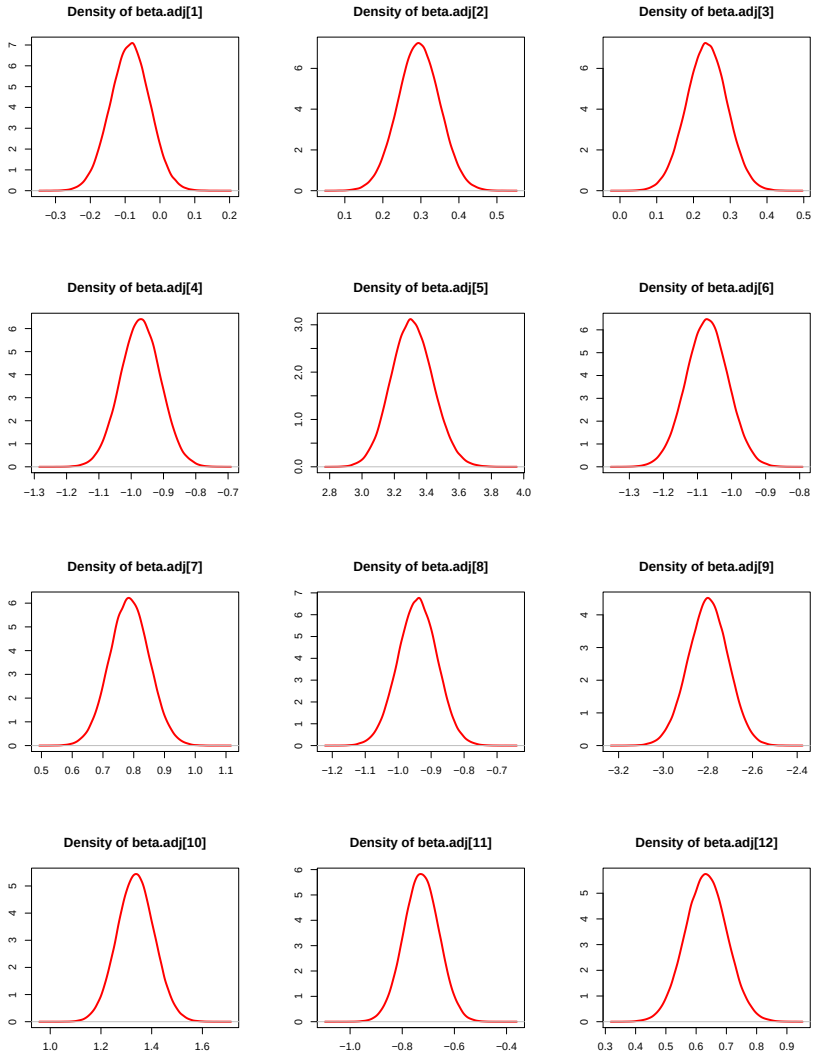


ICC_1 e ICC_2

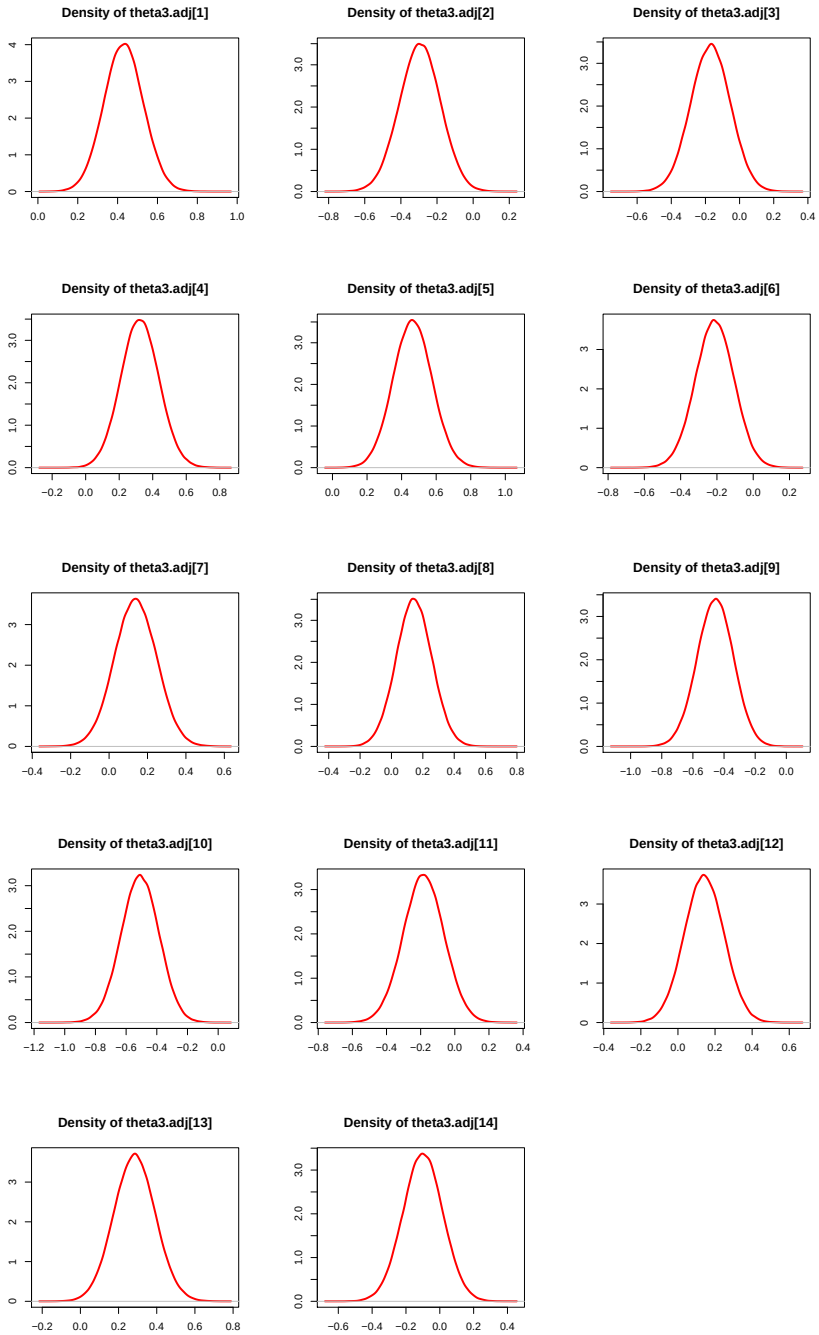


C.2 Modello IM

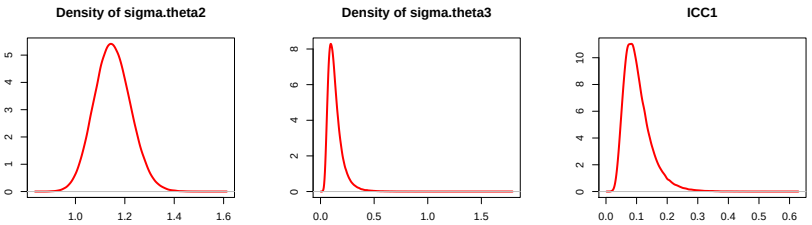
Parametri di difficoltà β^{adj}



Parametri di abilità $\theta^{(3),adj}$

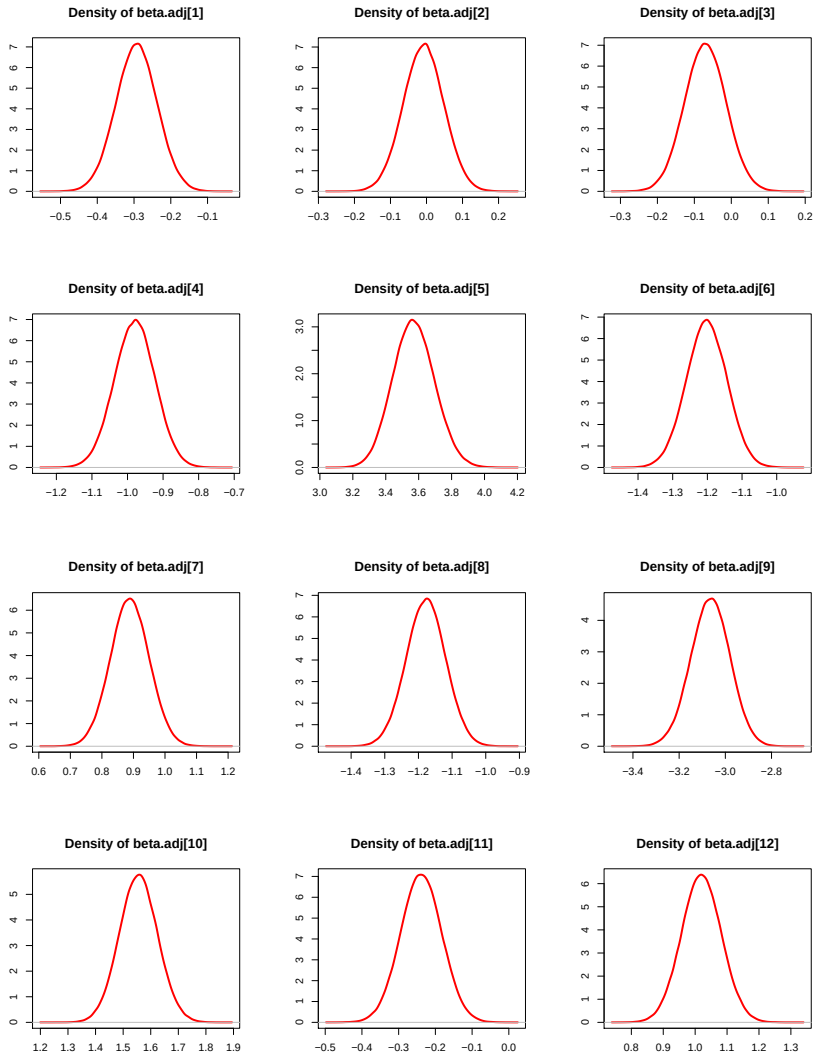


Varianza di $\theta^{(2),adj}$, varianza di $\theta^{(3),adj}$ e ICC

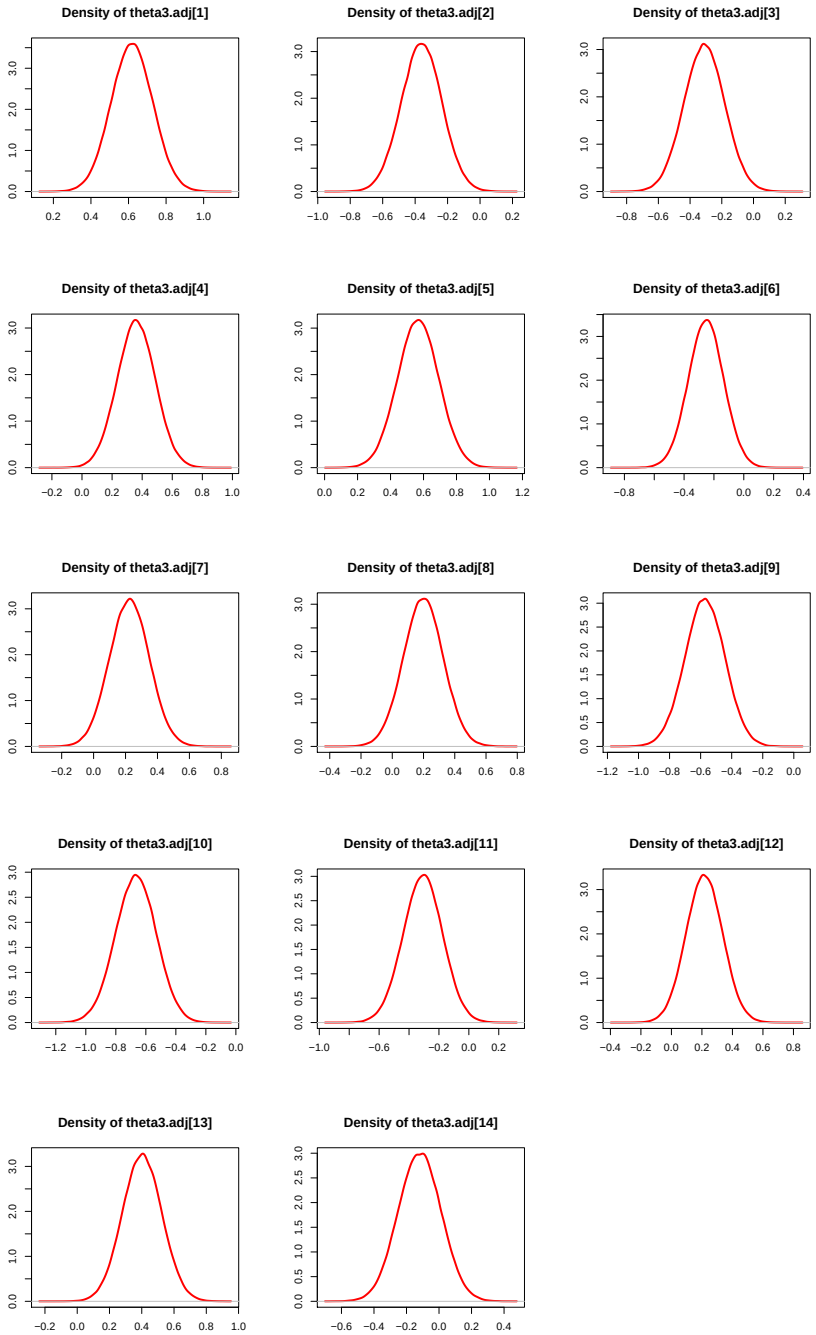


C.3 Modello ZIM

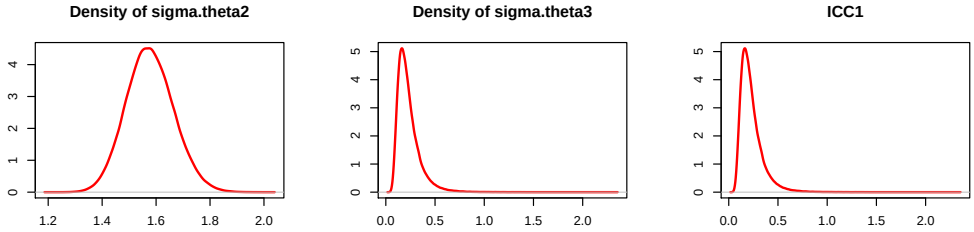
Parametri di difficoltà β^{adj}



Parametri di abilità $\theta^{(3),adj}$

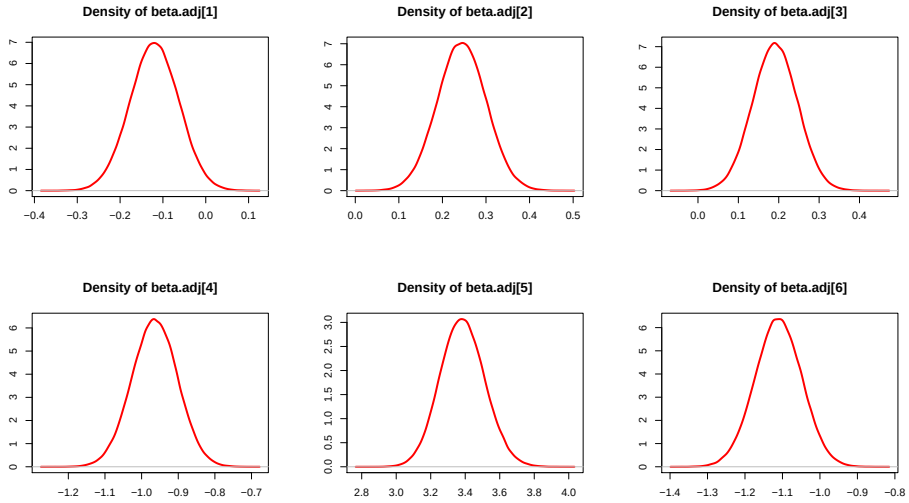


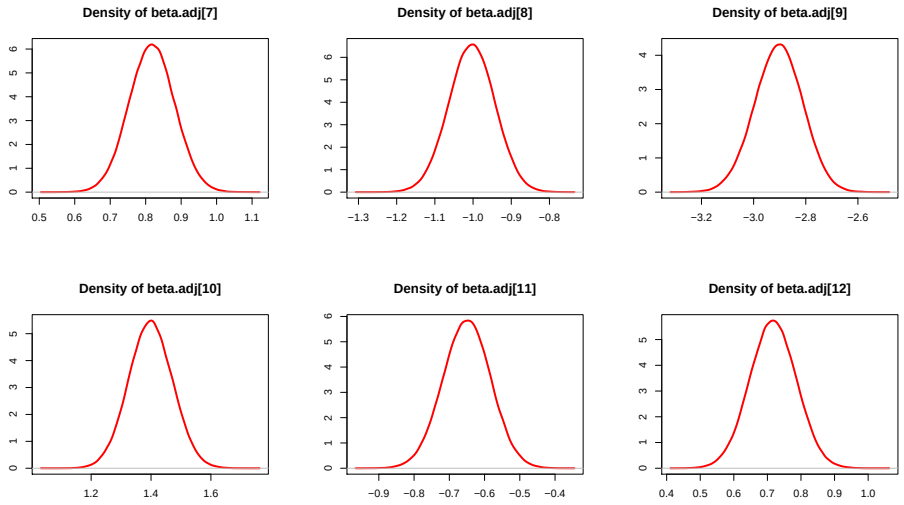
varianza di $\theta^{(2),adj}$, varianza di $\theta^{(3),adj}$ e ICC



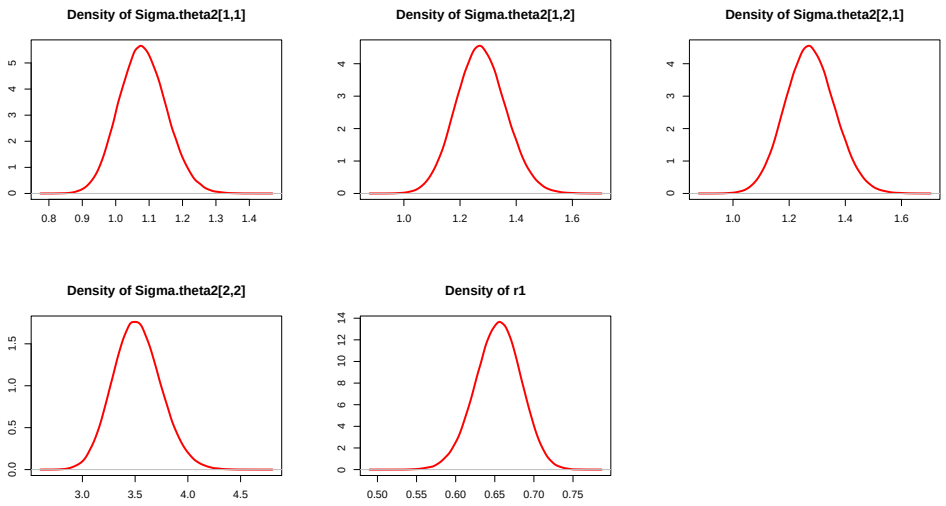
C.4 Modello NIM con covariate

Parametri di difficoltà β^{adj}

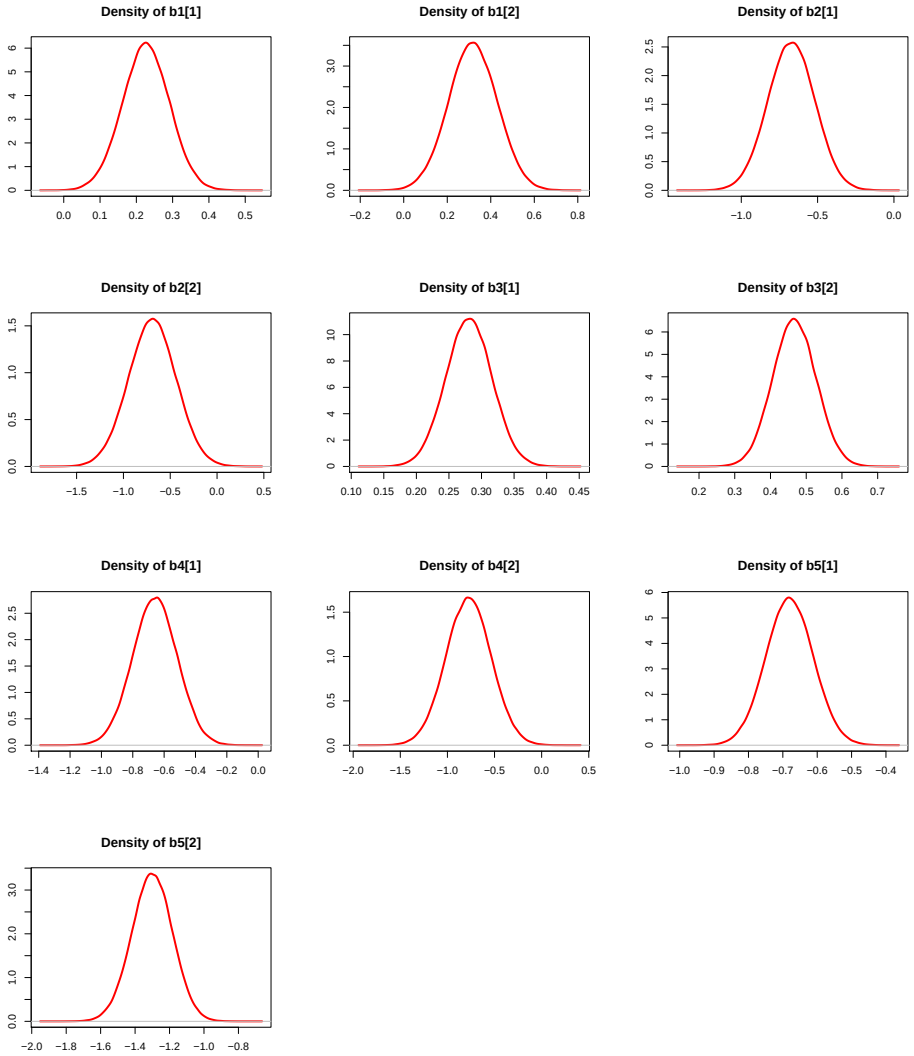




Matrice di varianza e covarianza di $\theta^{(2),adj}$ e coefficiente di correlazione $\rho_{student}$

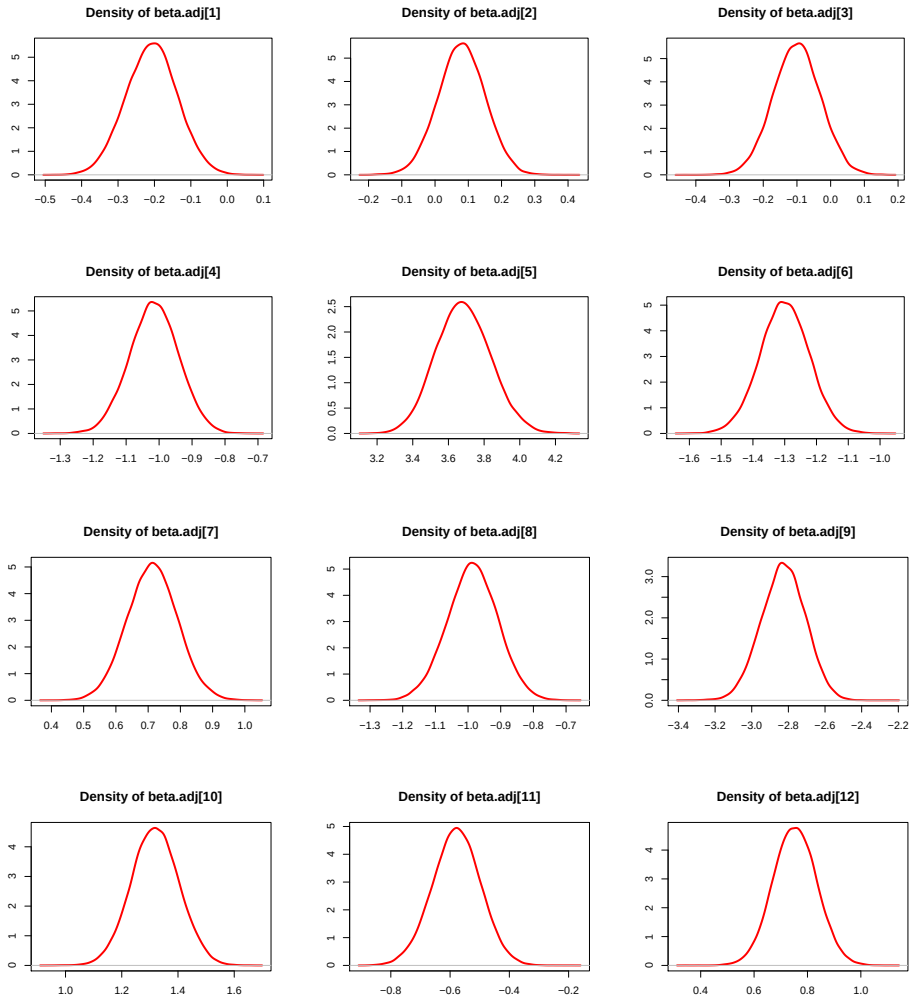


Regressori b_1, b_2, b_3, b_4, b_5

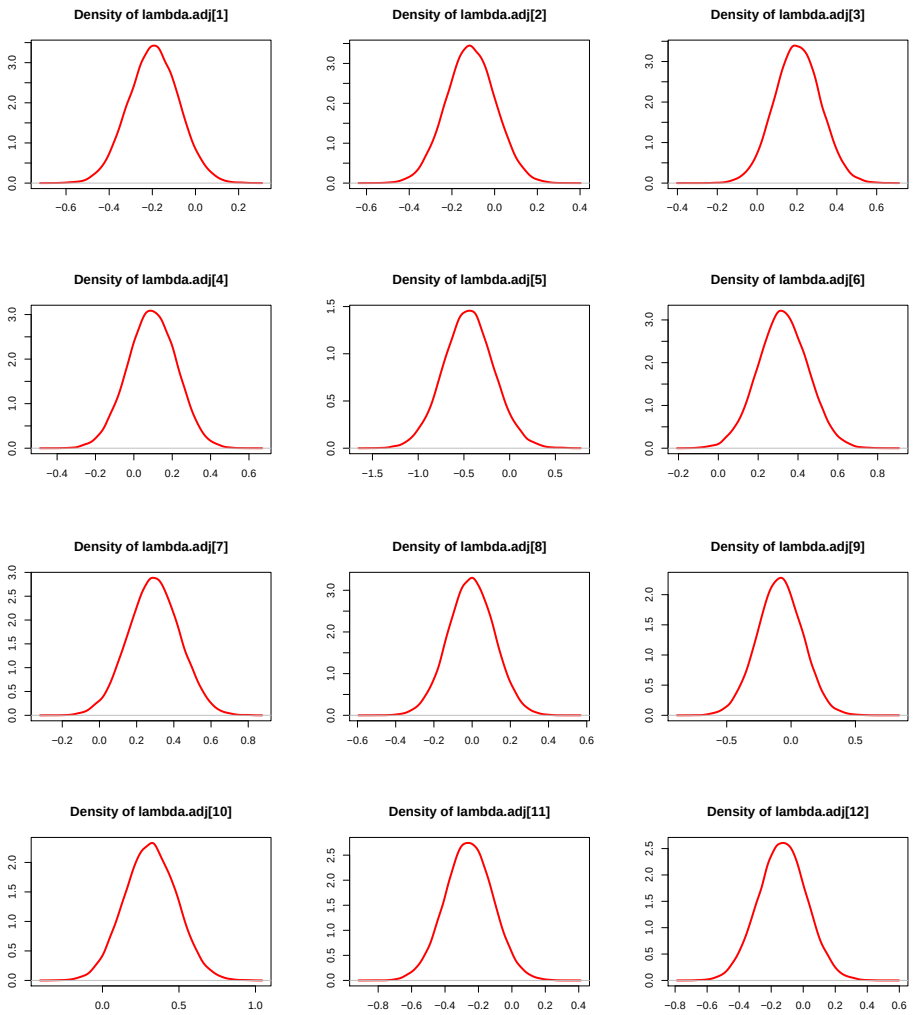


C.5 Modello DIF - NIM

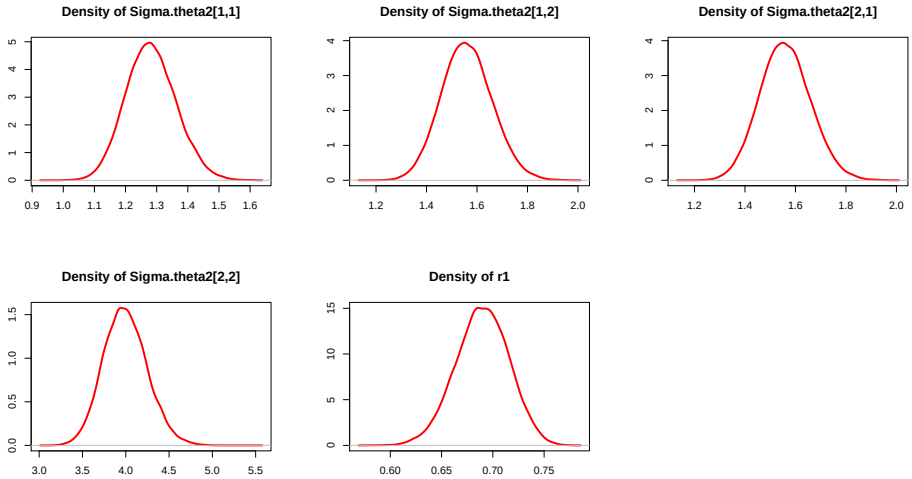
Parametri di difficoltà β^{adj}



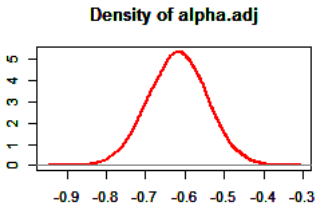
Parametri λ^{adj}



Matrice di varianza e covarianza di $\theta^{(2),adj}$ e coefficiente di correlazione $\rho_{student}$

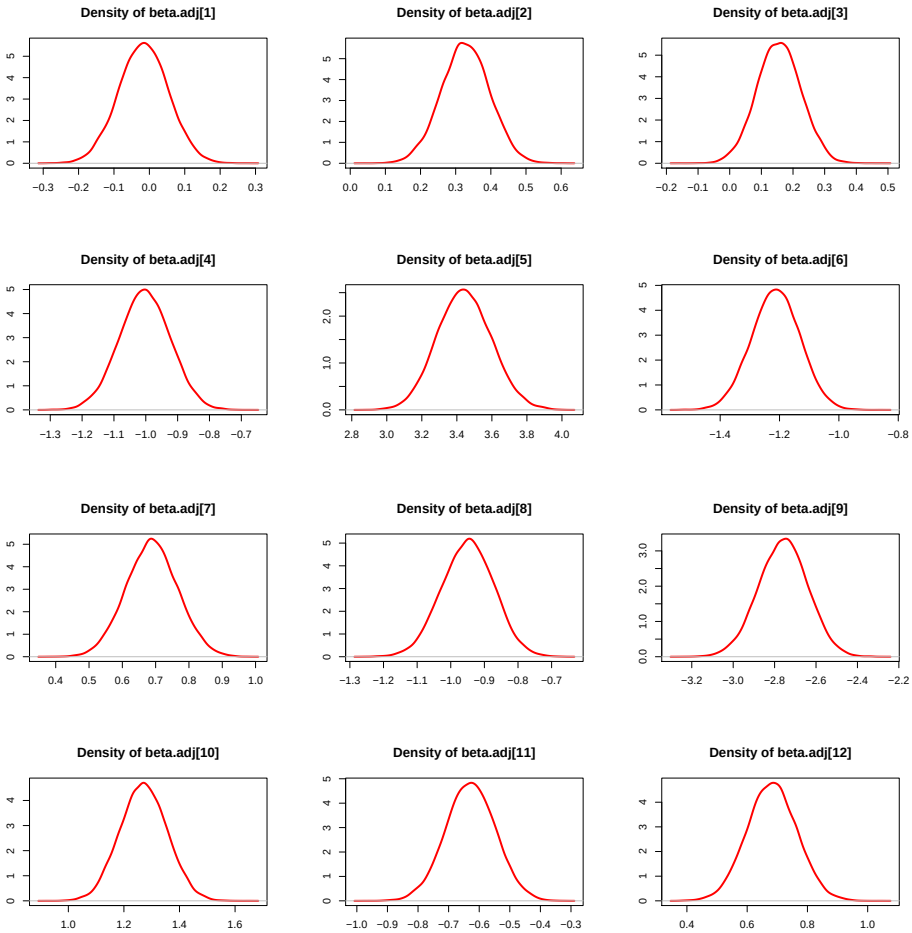


Parametro α_{adj}

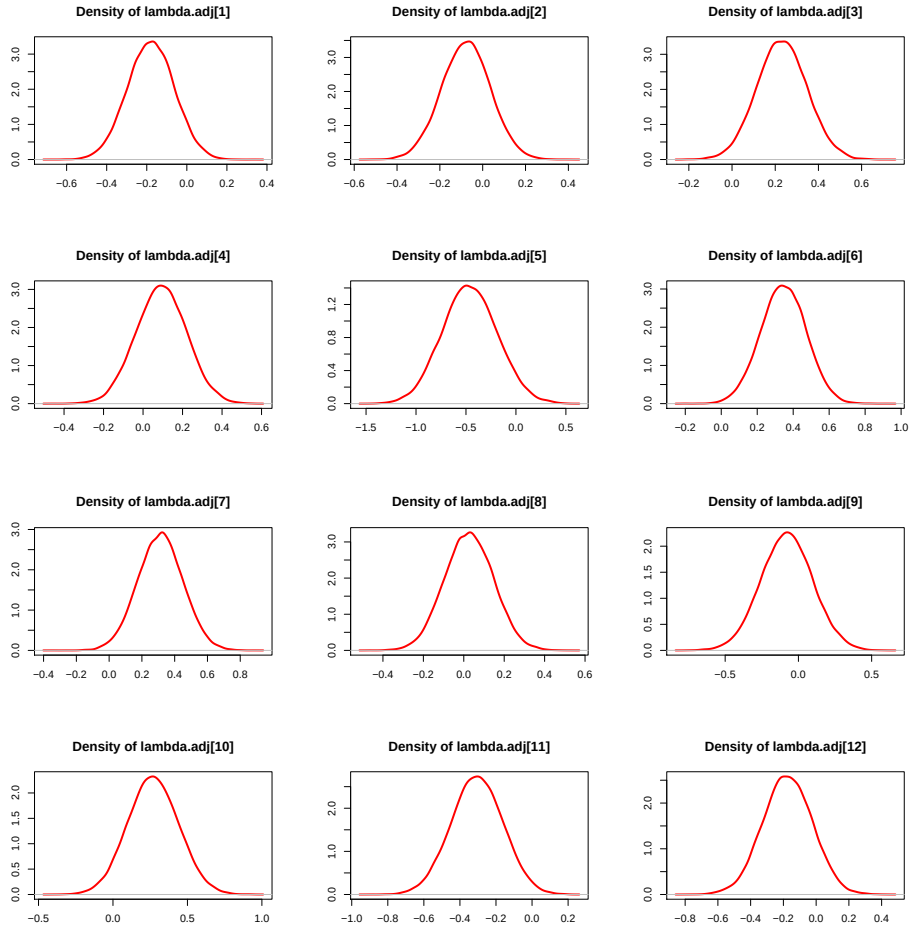


C.6 Modello DIF - IM

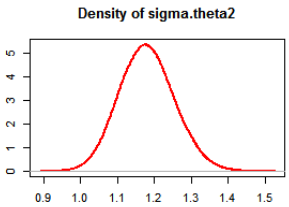
Parametri di difficoltà β^{adj}



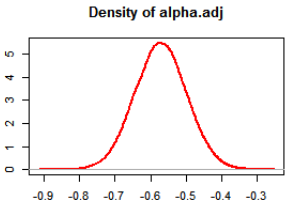
Parametri λ^{adj}



Varianza di $\theta^{(2),adj}$

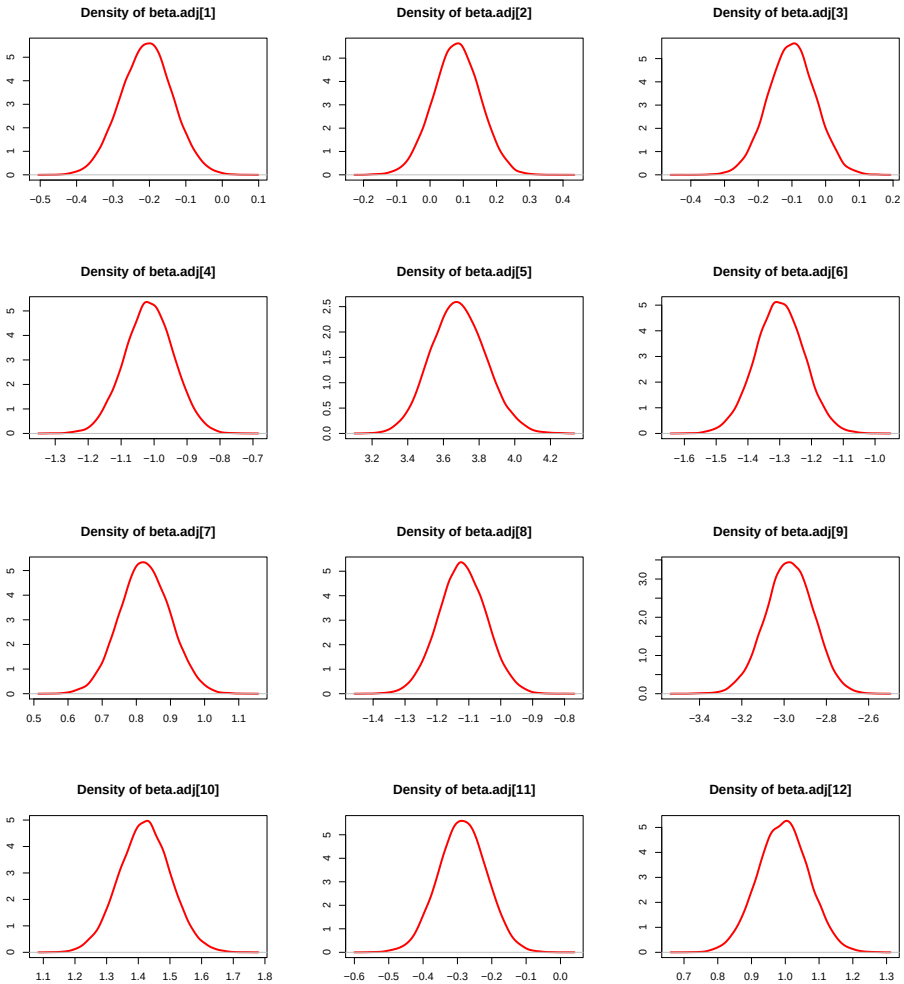


Parametro α_{adj}

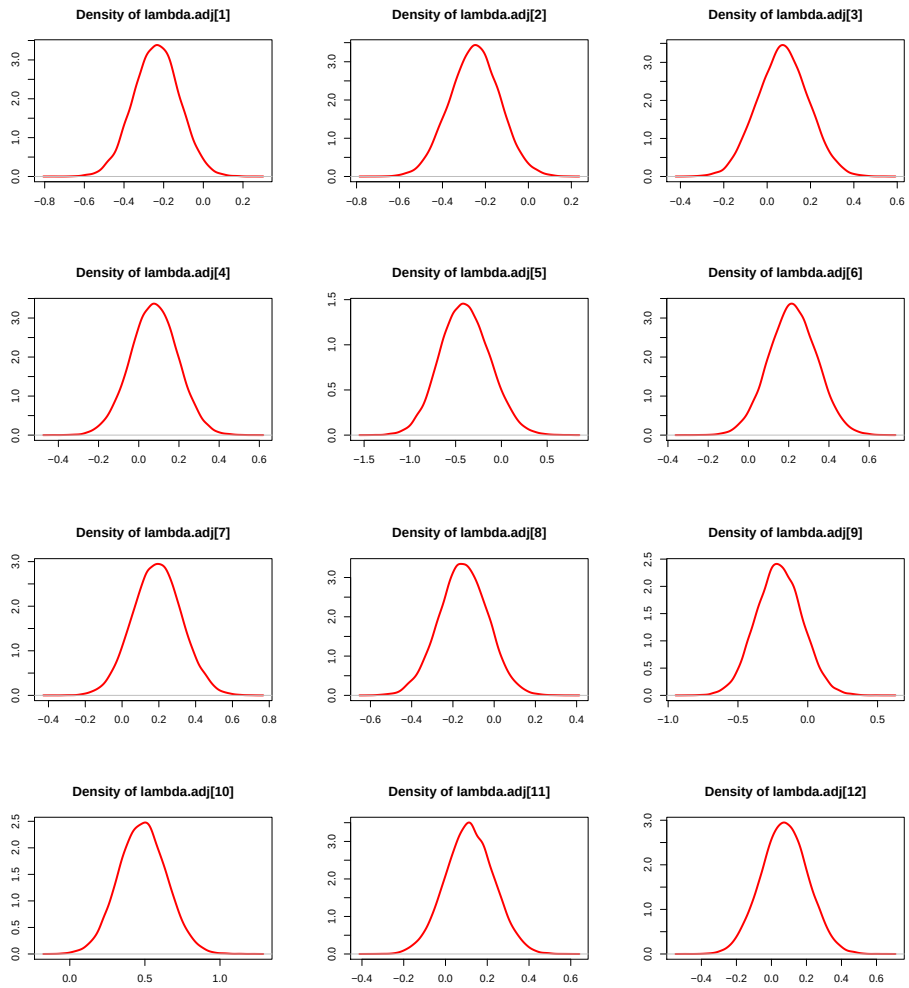


C.7 Modello DIF - ZIM

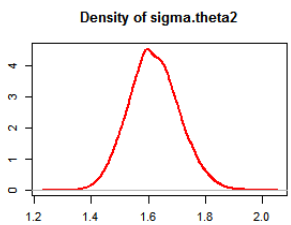
Parametri di difficoltà β^{adj}



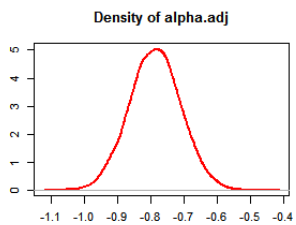
Parametri λ^{adj}



Varianza di $\theta^{(2),adj}$



Parametro α_{adj}



Bibliografia

Adams, R. J. e M. R. Wilson (1996). "Formulating the Rasch model as a mixed coefficients multinomial model." In: *Objective measurement: Theory into practice,* a cura di Jr. Engelhard e M. R. Wilson. Vol. 3, pp 143 -166. Norwood, New Jersey: Ablex.

Adams, R. J., M. R. Wilson e W. Wang (1997). "The multidimensional random coefficients multinomial logit model." In: *Applied Psychological Measurement,* 21, pp. 1–23.

Adams, R. J., M. R. Wilson e M.L. Wu (1997). "Multilevel item response modelling: An approach to errors in variables regression." In: *Journal of Educational and Behavioral Statistics,* 22, pp. 47–76.

Albert, J. H. (1992). "Bayesian estimation of normal ogive item response curves using Gibss Sampling". In: *Journal of Educational Statistics,* 17, pp 251–269.

Andrich, D. (1978). "A rating formulation for ordered response categories." In: *Psychometrika,* 43, pp. 561–73.

Bafumi, J. et al. (2005). "Pratical issues in implementing and understanding bayesian ideal point estimation." In: *Political analysis* 13), pp 171–187.

Baker, F. e S. Kim (2004). *Item Response Theory: Parameter Estimation Techniques.* New York, NY: Marcel Dekker.

Bartholomew, D.J. e M. Knott (1999). *Latent variable models and factor analysis,* 2nd edition. London: Oxford University Press.

Beguin, A.A. e C.A.W. Glas (2001). "MCMC estimation and some modelfit analysis of multidimensional IRT models." In: *Psychometrika,* 66, 541–562.

Bock, R. D. e M. F. Zimowski (1997.). "Multiple group IRT." In: *Handbook of modern item response theory.* A cura di W. van der Linden e R. Hambleton (Eds.) New York: Springer Verlag., pp 433–448.

Bolt, D.M. e V.F. Lall (2003). "Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo." In: *Applied Psychological Measurement,* 29, 395–414.

Bradlow, E.T., H. Wainer e X. Wang (1999). "A Bayesian random effects model for testlets." In: *Psychometrika,* 64, 153–168.

Brooks, S. e A. Gelman (1998). "Methods for monitoring convergence of iterative simulations." In: *Journal of Computational and Graphical Statistics,* 7, pp. 434–455.

Chaimongkol, S. (2005). "Modeling Differential Item Functioning (DIF) using Multilevel Logistic Regression Models: A Bayesian Perspective," tesi di dott. THE FLORIDA STATE UNIVERSITY.

Chaimongkol, S., F. Huffer e A. Kamata (2007). "An explanatory differential item functioning (DIF) model by the WinBUG 1.4". In: *Songklanakarin J. Sci. Technol.,* 29(2), pp. 449–458.

Cheong, Y. F. e S. W. Raudenbush (2000). "Measurement and structural models for children's problem behaviors." In: *Psychological Methods,* 5, pp. 477–495.

Cho, S.J. (2007). "A Multilevel Mixture IRT Model for DIF Analysis". Tesi di dott. University of Georgia.

Cox, D. D. R. e N. Wermuth (1996). *Multivariate Dependencies: Models, Analysis and Interpretation.* Chapman & Hall.

Darroch, J. N., S. L. Lauritzen e T. P. Speed (1980). "Markov fields and log-linear interaction models for contingency tables." In: *the Annals of Statistics,* 8, pp. 522–539.

Dempster, A. P. (1972). "Covariance Selection." In: *biometrics,* 28, pp. 157–175.

Dempster, A. P., N. M. Laird e D. B. Rubin (1977). "Maximum likelihood from incomplete data via EM algorithm." In: *Journal of Royal Statistical Society. Serie B (Methodological),* 39(1), pp 1–38.

Fischer, G. H. (1995). "The linear logistic test model." In: *Rasch Models: Foundations, Recent Developments, and Applications.* A cura di G. H. Fischer e I. W. Molenaar (Eds.) New York: Springer Verlag., pp. 131–155.

Fox, J.-P. (2004). "Applications of multilevel IRT modeling." In: *School Effectiveness and School Improvement,* 15, pp. 261–280.

— (2010). *Bayesian item response modeling: Theory and applications.* New York: Springer.

Fox, J.-P. e C.A.W. Glas (2001). "Bayesian estimation of multilevel IRT models using Gibbs sampling." In: *Psychometrika,* 66, 271–288.

— (2003). "Bayesian modeling of measurement error in predictor variables using item response theory." In: *Psychometrika,* 68 (2), pp. 169–191.

Gelfand, A. E. e A. F. M. Smith (1990). "Sampling-based approaches to calculating marginal densities." In: *Journal of the American Statistical Association* 85, pp. 398–409.

Gelman, A e J. Hill (2006). *Data Analysis Using Regression And Multilevel/Hierarchical Models,* Cambridge Univ Pr.

Geman, S. e D Geman (1984). "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images." In: *IEEE Transactions on Pattern Analysis and Machine intelligence* 6, pp. 721–741.

Gibbs, W. (1902). *Elementary principles of statistical mechanics.* New Haven, Connecticut: Yale University Press.

Gilks, W. R., S. Richardson e D. J. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice.* London: Chapman e Hall.

Gori, E., M. Sanarico e G Plazzi (2005). "La valutazione e la misurazione nelle scienze sociali: oggettività specifica, statistiche sufficienti e modello di Rasch." In: *Non Profit,* 3, pp 605–644.

Haberman, S. J. (1902). *The Analysis of Frequency Data.* Chicago, Illinois: University of Chicago Press.

Hastings, W. K. (1970). "Monte Carlo sampling-based methods using Markov chains and their applications." In: *Biometrika* 57, pp. 97–109.

Hedeker, D. e R. D. Gibbons (1993). "MIXOR: A computer program for mixed-effects ordinal probit and logistic regression analysis." Unpublished manuscript, University of Illinois, Chicago.

Holland, P. W. e S. T. Thayer (1985). *An alternative definition of the ETS delta scale of item difficulty.* Rapp. tecn. (ETS-RR-94-13). Princeton, NJ; Educational Testing Service.

— (1988). "Differential item performance and the Mantel-Haenszel procedure." In: *Test Validity.* A cura di H. Wainer e H. Braun (Eds.). Hillsdale, NJ: Lawrence Erlbaum., pp. 129–145.

Holman, R. e C. A. W. Glas (2005). "Modelling nonignorable missing data mechanisms with item response theory models." In: *British Journal of Mathematical and Statistical Psychology.* 58, pp 1–17.

INVALSI (2008). *Le competenze in scienze lettura e matematica degli studenti quindicenni.* URL: `http://www.invalsi.it/invalsi/ri/pisa2006.php?page=pisa2006_it_05`.

Jiao, H. et al. (2012). "A multilevel testlet Model for Dual Local Dependence". In: *Journal of Educational Measurement* 49(1), pp 82–100.

Johnson, M.S. e S. Sinharay (2005). "Calibration of polytomous item families using Bayesian hierarchical modeling." In: *Applied Psychological Measurement,* 29, 369–400.

Kamata, A. (2001). "Item analysis by the hierarchical generalized linear model." In: *Journal of Educational Measurement,* 38, 79–93.

Kamata, A. (2002). "Procedure to perform item response analysis by hierarchical generalized linear model." In: Paper presented at the 2002 annual meeting of the American Educational Research Association, New Orleans, LA.

Kamata, A. e S. Binici (2003). "Random Effect DIF Analysis via Hierarchical Generalized Linear Modeling." In: Paper presented at the biannual meeting of Psychometric Society, Sardinia, Italy.

Kamata, A. e Y.F. Cheong (2007). "Multilevel Rasch Models." In: *Multivariate and Mixture Distribution Rasch Models,* a cura di Claus H. Carstensen. Statistics for Social and Behavioral Sciences. Springer. Cap. 14, pp. 217–232.

Kim, I. (2007). "A Comparison of a Bayesian and Maximum Likelihood Algorithms for Estimation of a Multilevel IRT Model". Tesi di dott. University of Georgia.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford, United Kingdom: Clarendon Press.

Lauritzen, S. L. e D. J. Spiegelhalter (1988). "Local computations with probabilities on graphical structures and their application to expert systems (with discussion)." In: *Journal of the Royal Statistical Society,* B 50, pp. 157–224.

Little, R. J. A. e D. B. Rubin (2002). *Statistical analysis with missing data.* 2nd ed. Hoboken, NJ: Wiley-Interscience.

Lord, F. M. e M. R. Novick (1968). *Statistical Theories of Mental Test Scores.* Reading, Massachusetts: Addison-Wesley Pub. Co.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Maier, K. (2001). "A Rasch Hierarchical Measurement Model." In: *Journal of Educational and Behavioral statistics* 26(3), pp 307–330.

Mantel, N. e W. Haenszel (1959). "Statistical aspects of the analysis of data from retrospective studies of disease." In: *Journal of the National Cancer Institute,* 22, pp 719–748.

Masters, G. M. (1982). "A Rasch model for partial credit scoring." In: *Psychometrika,* 47, 149–174.

Masters, G. N. e B. D. Wright (1997). "The partial credit model." In: *Handbook of Modern Item Response Theory.* A cura di W. J. van der Linden e R. K. Hambleton (Eds.). New York: Springer, pp. 101–121.

Metropolis, N. et al. (1953). "Equations of state calculations by fast computing machines." In: *Journal of Chemical Physics* 21, pp. 1087–1091.

Mislevy, R. J. (1986). "Bayes modal estimation in item response models." In: *Psychometrika,* 51, pp. 177–195.

Mislevy, R. J. (1987). "Exploiting auxiliary information about examinees in the estimation of item parameters." In: *Applied Psychological Measurement.* 11, pp 81–91.

Moustaki, I. (1996). "A latent Trait and a latent class model for mixed observed variables." In: *British Journal of Mathematical and Statistical Psychology,* 49, pp 313–334.

Moustaki, I. e M. Knott (2000). "Weighting for item non-response in attitude scales using latent variable models with covariates." In: *Journal of the Royal Statistical Society. Serie A,* 163(3), pp 445–459.

Neal, R. M. (1997). *Markov chain Monte Carlo methods based on slicing the density function.* Rapp. tecn. Dept. of Statistics.

O'Muircheartaigh, C. e I. Moustaki (1999). "Symmetric pattern models: a latent variable approach to item nonresponse in attitude scales." In: *Journal of Royal Statistical Society. Serie A,* 162 (2), pp 177–194.

Park, C. e D. M. Bolt (2008). "Application of multilevel IRT to investigate cross-national skill profiles on TIMSS 2003". In: *Issues and methodologies in large-scale assessment.* IERI monograph series. Cap. 4, pp 71–96.

Patz, R. e B. Junker (1999a). "A straightforward approach to Markov chain Monte Carlo methods for item response models." In: *Journal of Educational and Behavioral Statistics,* 24, 146–178.

— (1999b). "Applications and extensions ofMCMC in IRT: Multiple item types, missing data, and rated responses." In: *Journal of Educational and Behavioral Statistics,* 24, 342–366.

Pearl, J. (1986a). "A constraint propagation approach to probabilistic reasoning." In: *Uncertainty in Artifcial Intelligence,* a cura di L. N. Kanal e J. F. Lemmer (Eds.). Amesterdam, The Netherlands. North-Holland, pp. 357–70.

— (1986b). "Fusion, propagation and structuring in belief networks." In: *Artificial Intelligence,* 29, pp. 241–88.

— (1988). *Probabilistic Inference in Intelligent Systems.* San Mateo, CA: Morgan Kaufmann Publishers.

Pearl, J. e A. Paz (1987). "Graphoids: a graph based logic for reasoning about relevancy relations." In: *Advances in Artificial Intelligence-II,* a cura di B. D. Boulay, D. Hogg e L. Steel. North-Holland, Amesterdam, pp. 357–63.

Raju, N.S. (1988). "The area between two item characteristic curves." In: *Psychometrika,* 53, pp. 495–502.

Rasch, G. (1960). *Probabilistic Models for some Intelligence Tests and Attainment Tests.* Copenhagen: Danish Institute for Educational Research.

Raudenbush, S. W. e A. S. Bryk (2002). *Hierarchical linear models (2nd ed.).* Thousand Oaks, CA: Sage.

Raudenbush, S. W., C. Johnson e R. J. Sampson (2003). "A multivariate, multilevel Rasch model for self-reported criminal behavior." In: *Sociological Methodology,* 33(1), pp 169–211.

Rigdon, S. E. e R. K. Tsutakawa (1983). "Parameter estimation in latent trait models." In: *Psychometrika,* 48, pp. 567–574.

Rijmen, F. et al. (2003). "A nonlinear mixed model framework for item response theory." In: *Psychological Methods,* 8(2), 185–205.

Rose, N., M. von Davier e X. Xu (2010). *Modeling Nonignorable Missing Data with Item Response theory(IRT),* rapp. tecn. RR-10-11. ETS.

Rubin, D. B. (1976). "Inference and missing data." In: *Biometrika,* 63(3), pp 581–592.

Spiegelhalter, D. J. et al. (1996.). *BUGS 0.5 examples.* Vol. 1. Cambridge, England: University of Cambridge, Institute of Public Health, Medical Research Council Biostatistics Unit.

Swaminathan, H. e J. A. Gifford (1982). "Bayesian estimation in the Rasch model." In: *Journal of Educational Statistics,* 7, pp. 175–192.

— (1985). "Bayesian estimation in the two parameter logistic model." In: *Psychometrika,* 50, pp. 349–364.

Swaminathan, H. e H.J. Rogers (1990). "Detecting differential item functioning using logistic regression procedures." In: *Journal of Educational Measurement,* 27, pp. 361–370.

Thissen, D., L. Steinberg e H. Wainer (1988). "Use of item response theory in the study of group differences in trace lines." In: *Test Validity.* A cura di H. Wainer e H. Braun (Eds.). Hillsdale, NJ: Lawrence Erlbaum, pp. 147–169.

Wermuth, N. e S. L. Lauritzen (1983). "Graphical and recursive models for contingency tables." In: *Biometrika,* 70, pp. 537–552.

— (1989). "Graphical models for associations between variables, some of which are qualitative and some quantitative." In: *the Annals of Statistics,* 17, pp. 31–57.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Chichester, United Kingdom: John Wiley e Sons.

Wold, H. D. A. (1954). "Causality and Econometrics." In: *Econometrica,* 28, pp. 443–63.

Wright, B. e G. Masters (1982). *Rating scale analysis.* Mesa Press.

Wright, S. (1923). "The Theory of path coefficient: a reply to Nile's criticism." In: *Genetics.* 8, pp. 239–55.

— (1934). "The method of path coeffcients." In: *Annals of Mathematical Statistics,* 5, pp. 161–215.