Politecnico di Milano

Scuola di Ingegneria Industriale e dell'Informazione

Corso di Laurea Magistrale in Ingegneria Matematica

Metodi di imputazioni per dati mancanti: applicazione al dataset INVALSI

Relatore:
Prof. Anna Maria Paganoni
Co-Relatore:
Chiara Masci

Tesi di Laurea Magistrale di:
Matteo Rivolta
Matr. 800989

Anno Accademico 2014/2015

Indice

Elenco delle figure

Elenco delle tabelle

Sommario

L'obiettivo di questo lavoro è studiare i metodi di imputazione per trattare i dati mancanti. Il dataset analizzato contiene informazioni su più di 500,000 bambini al primo anno di scuola media, nell'anno scolastico 2012/2013, fornite dall'Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e di Formazione (INVALSI). L'interesse è studiare i numerosi dati mancanti presenti che riducono notevolmente le informazioni a disposizione. Utilizzando le diverse tecniche di imputazione si sostituiscono i valori mancati con dei valori verosimili in modo da costruire dataset completi. Sfruttando i modelli lineari a effetti misti è possibile studiare le relazioni tra il voto del test e le caratteristiche dello studente e della scuola. Dopo aver adattato i modelli ai dataset imputati con i diversi metodi, si confrontano le stime dei parametri per verificare se esse risultano statisticamente identiche o diverse al variare del metodo di imputazione. I risultati di tutti i metodi mostrano che ci sono grandi differenze tra le tre aree geografiche Nord, Centro e Sud Italia caratterizzate da effetti scuola e caratteristiche rilevanti degli studenti molto differenti. Tramite cross-validazione si sono confrontati gli errori di previsione dei modelli realizzati sui dataset imputati per valutare il miglior metodo di imputazione per questo dataset.

Introduzione

Lo scopo di questa tesi è studiare le diverse tecniche con le quali è possibile trattare i *missing data* (dati mancanti).
Nell'analisi di un dataset accade molto spesso che vi siano dati mancanti (NA) per alcune variabili; le cause possono essere diverse e dipendono dalla natura stessa dei dati.
Da pochi anni grazie all'utilizzo di computer e software per le analisi statistiche, è possibile studiare dataset di grandi dimensioni. L'alta dimensionalità dei dati però rende pressoché certa la presenza di un numero elevato di dati mancanti.
Sono quindi state sviluppate numerose tecniche di analisi dati che hanno un'efficacia differente a seconda della natura degli NA contenuti nel dataset oggetto di studio.
Il classico e più diffuso approccio ai dati mancanti è la *listwise deletion*(LD). Questo metodo prevede l'eliminazione di ogni unità statistica (i.e riga del dataset) che contenga almeno un'informazione mancante. Il metodo in molti casi non è adatto e comporta un'elevata perdita di informazioni, producendo dei risultati molto distorti.
Negli ultimi anni sono stati introdotti metodi diversi per trattare i *missing data* al fine di ridurre la distorsione dei risultati e di trovare il metodo più adatto rispetto alla tipologia di dati mancanti presenti nel dataset.
In questa tesi verranno presentate le differenti tecniche di imputazione di dati mancanti e verranno utilizzate e messe a confronto nell'analisi di un dataset contenente i risultati della prova di matematica degli studenti delle classi prima media di tutta Italia dell'anno scolastico 2012/2013 creata dall' Istituto Nazionale di Valutazione del Sistema Educativo di Istruzione e Formazione (INVALSI); il dataset contiene molte osservazioni e molti NA, e ciò permette il confronto tra le diverse metologie di studio sui *missing data*.
La tesi è così strutturata: nel Capitolo 1 è presentato il problema dei *mis-*

sing data e le possibili strutture dei dati mancanti all'interno di un dataset; nel Capitolo 2 sono descritte le tecniche classiche con cui trattare i valori NA, i metodi di eliminazione e i metodi di imputazione singola; il Capitolo 3 introduce l'algoritmo Expectation Maximization (EM) che stima i parametri della distribuzione dei dati; nel Capitolo 4 sono descritte le fasi che caratterizzano le tecniche di imputazione multipla dei dati mancanti; nel Capitolo 5 viene descritto il dataset INVALSI; nel Capitolo 6 viene introdotto un test per la scelta del metodo con cui trattare i valori mancanti, test che permette di definire la natura del dato mancante; nel Capitolo 7 si adottano i modelli lineari a effetti misti per spiegare la valutazione della prova INVALSI a livello studente con effetto casuale a livello scuola, confrontando i risultati ottenuti a valle delle diverse tecniche di imputazione. Nel Capitolo 8 sono definiti i metodi di imputazione per dati raggruppati. Tutte le analisi e i modelli sono stati realizzati con il software statistico R [25].

Capitolo 1

Il problema dei *missing data*

La presenza dei dati mancanti all'interno di un dataset è da sempre stato un grande problema per gli statistici. Oggi lo sviluppo delle tecnologie e dei software statistici ha reso possibile lo studio di dataset di grandi dimensioni. La grandezza dei dataset tuttavia rende pressoché certa la presenza dei valori mancanti (NA) tra le osservazioni. Per questa ragione è divenuto necessario studiare nuovi metodi per poter risolvere questa problematica.

In passato la tecnica che veniva utilizzata era quella di tralasciare le osservazioni che contenevano NA e realizzare le analisi sui dati completi rimanenti: questo semplice approccio causa la perdita di molte informazioni e porta molto spesso a risultati sbagliati e quindi a conclusioni errate, poiché non tiene in considerazione la natura stessa del dato mancante.
Durante tutta la trattazione è utilizzata la seguente notazione: con $\mathbf{x}$ è indicata la variabile aleatoria e con x le realizzazioni per le unità osservate, anche se a volte i due concetti verranno erroneamente scambiati.

1.1 Diverse tipologie di *missing data*

La presenza di dati mancanti può influire significativamente sulle proprietà degli stimatori (media, varianza, quantili, parametri dei coefficienti di regressione) e pertanto può condurre a risultati inferenziali non corretti.
In questo contesto è determinante stabilire se il meccanismo che ha generato i valori mancanti è di tipo casuale, nonché analizzare le possibili relazioni tra i valori mancanti e i valori assunti dalle variabili per cui si dispone l'os-

servazione.

Definiti con Y=(y_{ij}) i dati e con M=(M_{ij}) la matrice costituita da 1 se il dato y_{ij} è mancante e 0 altrimenti, la natura della mancanza è caratterizzata dalla distribuzione condizionata di M dato Y ,cioè f(M|Y,ϕ), dove ϕ identifica un parametro (o un insieme di parametri) che descrive la relazione tra la matrice M e i dati (vedi [8, 22]).
Se la mancanza del dato non dipende dai valori dei dati Y, né osservati né mancanti, cioè

$$f(M|Y,\phi) = f(M|\phi) \qquad \textit{per ogni } Y, \phi \tag{1.1}$$

i dati mancanti sono detti *Missing Completely At Random* (MCAR). L'assunzione che i dati siano di questo tipo non presuppone che il pattern sia casuale, ma che la distribuzione dei dati mancanti non dipenda dai valori assunti dai dati stessi.
Sia ora Y_{obs} la parte di dati Y che sono realmente osservati, e invece Y_{mis} la componente che descrive quelli mancanti. Se la mancanza dei dati nel dataset Y dipende soltanto dalla componente Y_{obs} osservata, e non da quella mancante, allora i dati si definiscono *Missing At Random* (MAR). Per dati di questo tipo la distribuzione condizionata di M dato Y è così definita

$$f(M|Y,\phi) = f(M|Y_{obs,}\phi) \qquad \textit{per ogni } Y_{mis}, \phi. \tag{1.2}$$

L'ultima tipologia di missing è quella dei *Missing Not At Random*(MNAR) che descrive un comportamento non casuale della presenza degli NA: la distribuzione della matrice M dei missing dipende dai dati mancanti Y_{mis}.

$$f(M|Y,\phi) = f(M|Y_{obs}, Y_{mis}, \phi) \qquad \textit{per ogni } \phi. \tag{1.3}$$

Questa distinzione dei tipi di *missing data* è fondamentale: le tecniche per trattare i dati mancanti cambiano al variare della tipologia di dato mancante nel dataset in esame. Per esempio la tecnica della *listwise deletion*, che consiste nella eliminazione delle osservazioni del dataset che contengono almeno un NA, restituisce buoni risultati nel caso in cui i dati sono di tipo MCAR e risultati distorti nel caso MAR e MNAR.
L'importanza quindi di identificare la natura del dato mancante risulta essere fondamentale per la determinazione della tecnica migliore da utilizzare.

Questa problematica però è difficile da risolvere: in letteratura esistono dei test statistici per determinare se la tipologia di *missing* è di tipo MCAR, ma non vi sono test per verificare la struttura di MAR e MNAR. L'unico modo per distinguere tra questi due tipi di dati è sapere già la natura dei dati prima di analizzarli grazie a informazioni fornite dall'esperto che li ha raccolti.

Capitolo 2

Tecniche classiche

2.1 Metodi di eliminazione

I metodi di eliminazione sono i metodi più frequentemente utilizzati quando si ha a che fare con dataset che presentano dati mancanti. Questi approcci sono quelli che sono stati implementati nella maggior parte delle funzioni dei software statistici. In questa tipologia di metodo figurano due diversi approcci: la *listwise deletion* e la *pairwise deletion*.

2.1.1 Listwise deletion

La *listwise deletion* (o *Complete Case Analysis*) è la tecnica più naturale e più spesso utilizzata: essa prevede, come accennato in precedenza, l'eliminazione di tutte le righe, quindi di tutte le osservazioni, che contengono tra le variabili almeno un valore NA.
lI vantaggio di questo metodo è la semplicità poichè le analisi statistiche possono essere applicate senza modifiche ad un dataset di dimensione ridotta ma completo.
Lo svantaggio principale è invece la perdita di informazioni: se il numero di osservazioni complete è assai piccolo il dataset si riduce notevolmente in dimensione e ciò potrebbe portare a dei risultati di stima distorti e dunque a conclusioni errate. A causa della facilità d'uso e del ridotto costo computazionale questa tecnica è stata utilizzata per diversi decenni e ancora oggi è una delle più diffuse nelle comunità scientifiche.
Questa tecnica può essere considerata valida nel solo caso in cui i *missing*

data sono di tipo MCAR, condizione molto rara nelle situazioni reali.

2.1.2 Pairwise deletion

Il metodo appena descritto della *listwise deletion* è molto dispendioso poiché fa perdere informazioni anche per quelle variabili in cui il dato non manca. Una valida alternativa a questa perdita eccessiva di informazioni è costituita dalla tecnica della *pairwise deletion* (o *Available Case Analysis*), la quale include tutte le unità statistiche per le quali la variabile di interesse è stata osservata.
Il metodo prevede la creazione di differenti dataset a seconda dei diversi studi che si vogliono realizzare e per ognuno di essi si considerano solo le variabili di interesse per l'analisi, eliminando successivamente i valori NA.
Il principale svantaggio di questo metodo è che il campione varia al variare dei dataset creati e a seconda delle variabili considerate. La variabilità nella base del campione crea notevoli problemi in quanto non rende possibile l'utilizzo di semplici strumenti per la verifica della corretta costruzione dei dataset.
Il metodo ha però il vantaggio di ridurre la distorsione sulle stime rispetto alla *listwise deletion*, pur aumentando i costi computazionali.

2.2 Metodi di imputazione singola

L'imputazione singola è un metodo statistico che cerca di eliminare i valori mancanti all'interno di un dataset, sostituendo gli NA con dei valori ammissibili per la variabile considerata.
Il metodo di imputazione singola è una tecnica interessante per trattare i *missing data* poiché riduce la perdita di informazioni ma, come riportano Dempster and Rubin [7]:

"The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases."

Il vantaggio principale del metodo è che, una volta sostituiti i valori e creato il dataset completo, l'analisi statistica può proseguire utilizzando tecniche e modelli per i dataset completi. Lo svantaggio invece consiste nel fatto che i dati imputati vengono poi considerati come realmente osservati e trattati come tali nello studio, ciò comporta una riduzione della variabilità.
Per sostituire alle variabili i valori mancanti è necessario definire dei criteri con cui imputare i dati [22]. Le tecniche sviluppate sono diverse e ognuna restituisce un suo valore di stima differente:

- *Media*: la media dei valori osservati per ogni variabile va a sostituire i valori mancanti della stessa variabile;
- *Campionamento aleatorio*: il valore NA è sostituito da un valore estratto in modo casuale da quelli disponibili per la variabile;
- *Regressione*: il valore mancante di una variabile di una particolare unità statistica è sostituita dal risultato di una regressione tra le variabili osservate della stessa unità;
- *Regressione Stocastica*: il valore è imputato come nel caso della regressione con l'aggiunta di un residuo che va a descrivere l'incertezza nella previsione;
- *Hot Deck*: sostituisce i valori NA con quelli di un'istanza simile che è stata realmente osservata (per esempio il valore imputato viene campionato da un sottocampione del dataset con caratteristiche simili all'osservazione in cui manca il dato);
- *Cold Deck*: sostituisce il valore mancante con un valore derivante da una sorgente esterna, come una precedente realizzazione della stessa variabile;
- *Nearest Neighbor*: questo metodo definisce una nozione di distanza tra le osservazioni (basata sulla tipologia delle covariate) e sceglie i valori imputati che provengono dalla unità più vicina all'osservazione con i valori mancanti;
- *Predictive Mean Matching*: il valore mancante è imputato con il valore previsto da un modello di regressione di un osservazione con covariate simili e complete.

I metodi di imputazioni singola sono dunque dei metodi interessanti poiché, sostituendo con un nuovo dato imputato il valore di ogni *missing data*, creano un dataset completo su cui è possibile applicare le tecniche classiche di analisi, e inoltre permettono l'uso di quelle informazioni che i metodi di eliminazione avrebbero trascurato.
Nella maggior parte dei casi però le stime dei parametri prodotte risultano essere distorte, anche nel caso di dati di tipo MCAR. Il motivo di questa distorsione nelle stime dei parametri è da attribuire al fatto che le imputazioni generate con questi metodi non tengono in considerazione una componente di incertezza nella fase di imputazione. La mancanza di variabilità dei valori imputati produce una sottostima delle deviazioni standard.

Capitolo 3

Algoritmo Expectation Maximization (EM)

Un modo alternativo per trattare i dataset incompleti, a causa della presenza di NA, è l'algoritmo di Expectation Maximization (EM) [3, 8]. Il metodo si basa sulla nozione di stima di massima verosimiglianza (MLE) di θ, parametro incognito della distribuzione dei dati.
In generale dato uno spazio campionario X, sia $x \in X$ una osservazione estratta dalla densità $f(\mathrm{x}|\theta)$ che dipende dal parametro θ, che in generale identifica i parametri di media e varianza della distribuzione; si definisce "funzione verosimiglianza" di θ data la singola osservazione x la funzione

$$L(\theta|x) \propto f(x|\theta) \tag{3.1}$$

Quando il campione è costituito da n osservazioni indipendenti allora la funzione di verosimiglianza si fattorizza

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} L(\theta|x_i). \tag{3.2}$$

Poiché i valori della verosimiglianza sono molto prossimi allo 0 e per semplificare il calcolo della derivata per le stime di massima verosimiglianza dei parametri è opportuno trasformare la funzione con una trasformazione logaritmica e studiare dunque quella che è chiamata *log-verosimiglianza*.

$$\ell(\theta, \mathbf{x}) = \sum_{i=1}^{n} logL(\theta, x_i) \tag{3.3}$$

3.1 L'algoritmo

L'algoritmo EM è un algoritmo iterativo per la stima di massima verosimiglianza nel caso di problemi di *missing data*. La semplicità del metodo è la ragione per cui è divenuto molto popolare nello studio di questi tipologie di problemi.
L'idea dell'algoritmo alterna una fase di previsione (Expectation) che calcola il valore atteso della log-verosimiglianza del parametro condizionata ai dati completi e alle stime precedenti del parametro, e una fase di massimizzazione (Maximization) che sfrutta i dati appena aggiornati dal passo E per trovare il nuovo valore stimato di massima verosimiglianza del parametro.
Al generico passo q l'algoritmo può essere così descritto:

1. se q=0, la stima $\theta^{(q)}$ del parametro θ è realizzata tramite le informazioni disponibili dal dataset;
2. $Q(\theta|\theta^{(q)})$=E$[l(\theta|Y_{obs}, Y_{mis})|Y_{obs},\theta^{(q)}]$;
3. $\theta^{(q+1)}$ = arg max_θ Q$(\theta|\theta^{(q)})$.

Il procedimento viene ripetuto finché la differenza tra le ultime iterazioni non raggiunge la soglia di tolleranza prefissata, a dimostrazione che l'algoritmo è arrivato a convergenza.
Anche questo metodo ha dei punti di forza e di debolezza: il vantaggio principale dell'algoritmo sta nel fatto che ad ogni iterazione viene aumentata la log-verosimiglianza $\ell(\theta|Y_{obs})$, e se la log-verosimiglianza è limitata, la successione $\ell(\theta^{(t)}|Y_{obs})$ converge ad un valore stazionario; lo svantaggio, invece, è che il tasso di convergenza può essere molto lento quando si hanno molti *missing data*.

3.2 Il passo E e il passo M dell'algoritmo EM

Il passo E utilizza gli elementi del vettore delle medie e la matrice di covarianza per costruire un insieme di equazioni di regressione che prevedono i valori delle variabili incompleti dalle variabili osservate. Lo scopo di questo passo è prevedere i valori mancanti in un modo che ricorda l'imputazione di regressione stocastica, ma senza imputare realmente gli NA.
Il passaggio M successivo applica le formule standard per dati completi ai

dati appena "creati" per aggiornare le stime del vettore delle medie e della matrice di varianza e covarianza. Le nuove stime dei parametri sono passate al successivo passo E, dove viene costruita una nuova serie di equazioni di regressione per prevedere di nuovo i *missing data*.
L'algoritmo EM ripete questi due passi fino a che la media e la matrice di covarianza non variano per più passi consecutivi, e a quel punto l'algoritmo converge alle stime di massima verosimiglianza. É molto importante ribadire che l'algoritmo EM non imputa o sostituisce i dati mancanti: piuttosto esso utilizza tutti i dati a disposizione per fornire le stime del vettore delle medie e della matrice di varianza e covarianza delle variabili del dataset.
L'obiettivo di ogni iterazione dell'algoritmo è quello di regolare i valori dei parametri nella direzione che aumenta il valore della log-verosimiglianza, cioè portare la successione di parametri stimati a convergere alla stima MLE del parametro. La procedura di regressione ad ogni passo E fa proprio questo, e le stime dei parametri aggiornate ad ogni passo M producono un valore di log-verosimiglianza superiore alle stime del passo M precedente. Quando la differenza tra le stime consecutive scende sotto ad una soglia di tolleranza, il processo iterativo si arresta e l'algoritmo è arrivato a convergenza.
Sia $\theta^{(t)}$ il valore di stima corrente del parametro θ. Il passo E di EM trova la log-verosimiglianza prevista con dati completi Y se θ assume valore $\theta^{(t)}$:

$$Q(\theta|\theta^{(t)}) = \int l(\theta|Y_{mis}, Y_{obs}) f(Y_{mis}|Y_{obs}, \theta = \theta^{(t)}) dY_{mis} \tag{3.4}$$

Il passo M invece determina il $\theta^{(t+1)}$ massimizzando la log-verosimiglianza prevista coi dati completi:

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)}), \tag{3.5}$$

3.3 Applicazione di EM al caso a due variabili

Nella precedente descrizione dell'algoritmo è stato presentato il modo in cui i passi agiscono. Per illustrare in modo più dettagliato il funzionamento e i passi dell'algoritmo viene presentato un caso bivariato, con una delle variabili che risulta essere incompleta [8]. In questa sezione X indicherà la variabile completa e Y la variabile incompleta. É un semplice caso di stima della media, ma l'idea alla base può essere estesa senza particolari problemi

all'analisi multivariata (oggetto della prossima sessione).
Con i dataset completi, le seguenti formule generano le stime di massima verosimiglianza della media, varianza e covarianza

$$\hat{\mu}_Y = \frac{1}{N}\sum Y \tag{3.6}$$

$$\hat{\sigma}_Y^2 = \frac{1}{N}(\sum Y^2 - \frac{(\sum Y)^2}{N}) \tag{3.7}$$

$$\hat{\sigma}_{X,Y} = \frac{1}{N}(\sum XY - \frac{\sum X \sum Y}{N}) \tag{3.8}$$

Lo scopo del passo E è di prevedere i valori mancanti in modo che il passagio M dell'algoritmo possa usarli per generare le stime dei parametri. Ad ogni iterazione il passo E crea un'equazione di regressione che predice i valori mancanti dalle variabili di cui si hanno a disposizione le osservazioni, sfruttando le nuove stime dei parametri prodotte dal passo M dell'iterazione precedente. Nel caso di dataset incompleto bivariato, le equazioni necessarie sono

$$\hat{\beta}_1 = \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_X^2} \tag{3.9}$$

$$\hat{\beta}_0 = \hat{\mu}_Y - \hat{\beta}_1\hat{\mu}_X \tag{3.10}$$

$$\hat{\sigma}_{Y|X}^2 = \hat{\sigma}_Y^2 - \hat{\beta}_1{}^2\hat{\sigma_X}{}^2 \tag{3.11}$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{3.12}$$

dove $\hat{\beta}_0$ e $\hat{\beta}_1$ sono rispettivamente i coefficienti dell'intercetta e della pendenza, $\hat{\sigma}_{X,Y}^2$ è la varianza dei residui della regressione di Y su X, e $\hat{Y}_i$ è il valore predetto di Y per un dato valore di X_i.
Il passo E dell'algoritmo sostituisce le componenti mancanti delle statistiche sufficienti $\sum Y^2$, $\sum Y$ e $\sum XY$ con i loro valori attesi. EM utilizza le informazioni delle altre variabili e sostituisce i valori mancanti condizionandoli alle variabili osservate.
É opportuno ricordare nuovamente che il passo E dell'algoritmo non imputa i dati, ma sostituisce i valori per il calcolo della media, varianza e covarianza.

3.4 EM per dati multivariati

La precedente descrizione dell'algoritmo nel caso bivariato è relativamente semplice perchè i *missing data* sono tutti concentrati su di un'unica variabile. Nel caso di dati multivariati applicare EM risulta più complesso, poiché il passo E richiede un'equazione di regressione o un insieme di equazioni per ogni *pattern* di dati mancanti.
Nonostante l'aumento della complessità, la logica di base di EM rimane la stessa descritta nel paragrafo precedente e richiede soltanto dei dettagli aggiuntivi.
Per semplicità, l'algoritmo viene applicato ad un dataset con tre variabili, ma il ragionamento può essere esteso ad un numero arbitrario di variabili.
É importante sottolineare come il passo M dell'algoritmo non cambia. Infatti il passo adotta la tecnica classica per dati completi di massimizzazione della log-verosimiglianza per poi aggiornare i parametri, mentre il passo E varia la sua struttura.

Il dataset considerato contiene tre variabili X, Y e Z, e l'applicazione del passo E dell'algoritmo richiede le seguenti statistiche sufficienti: $\sum X$, $\sum X^2$, $\sum Y$, $\sum Y^2$, $\sum Z$, $\sum Z^2$, $\sum XY$, $\sum XZ$ e $\sum YZ$ (le stesse che sono state calcolate nel caso bivariato). Come in precedenza il passo propone di sostituire le componenti mancanti con i valori attesi delle statistiche sufficienti, ma ciò richiede un insieme di equazioni di regressione per ogni *pattern* di *missing data*. In questo caso i *pattern* possibili sono sette: solo la variabile X disponibile e le altre due mancanti, X e Y osservate e Z mancante, ecc.
Si prenda in considerazione il sottocaso in cui la variabile Y contiene gli NA. Il *pattern* di dati mancanti ora ha due variabili complete, dunque l'equazione di regressione multipla che genera le previsioni è

$$\hat{Y}_{i|X,Z} = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i \tag{3.13}$$

dove $\hat{Y}_{i|X,Z}$ è la previsione di Y per il caso i condizionato a X e Z. Analogamente al caso bivariato, la previsione per $\sum Y^2$ implica un valore previsto al quadrato e una stima della varianza dei residui della regressione di Y sui valori di X e Z, cioè $\hat{Y}^2_{i|X,Z} + \hat{\sigma}^2_{Y|X,Z}$.
Analogamente considerando la variabile Z si ottiene la seguente regressione

multipla

$$\hat{Z}_{i|X,Y} = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Y_i \tag{3.14}$$

e il valore atteso sostituisce le componenti mancanti di $\sum Z$, $\sum XZ$ e $\sum YZ$.

Ogni *pattern* di dati mancanti richiede quindi il suo insieme di regressioni e valori attesi, ma la logica segue quella dell'analisi bivariata. L'unica differenza si ha quando il *pattern* ha due o più dati mancanti. Per esempio se X è l'unica variabile osservata

$$\hat{Y}_{i|X} = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{3.15}$$

$$\hat{Z}_{i|X} = \hat{\beta}_2 + \hat{\beta}_3 X_i \tag{3.16}$$

la statistica prodotto $\sum YZ$ richiede una ulteriore previsione $(\hat{Y}_{i|X})(\hat{Z}_{i|X})+ \hat{\sigma}^2_{Y,Z|X}$ dove i due termini tra parentesi tonde sono calcolate grazie alle regressioni appena descritte, e $\hat{\sigma}_{Y,Z|X}$ è la covarianza dei residui tra le variabili, cioè $\hat{\sigma}_{Y,Z|X} = \hat{\sigma}_{Y,Z}$ - $\hat{\beta}_2\hat{\beta}_3\hat{\sigma}_X$.

L'estensione ai dati multivariati del passo E richiede diverse regressioni per ogni *pattern* di *missing data*: il numero elevato di variabili incomplete nel dataset contribuisce ad aumentare il numero di *pattern* e quindi il numero di regressioni necessarie al calcolo dei valori attesi.

Capitolo 4

Imputazione Multipla

L'algoritmo dell'Expectation Maximization descritto nel capitolo precedente permette di stimare i parametri della distribuzione dei dati sfruttando le informazioni dalle osservazioni disponibili nel dataset. L'algoritmo stima i valori mancanti presenti nel dataset in modo da poter poi calcolare le stime dei parametri della distribuzione dei dati.
L'imputazione multipla, invece, è un metodo che basa il suo procedimento sulla imputazione singola descritta nel Capitolo 2.2. Analogamente ad essa utilizza le informazioni disponibili nel dataset per imputare i valori mancanti del dataset, in modo così da poter studiare dataset completi con le tecniche classiche. Il metodo dell'imputazione multipla eredita i vantaggi che caratterizzano l'imputazione singola, riducendone notevolmente gli svantaggi [8],[20],[22],[27],[31],[33].
L'idea principale alla base dell'imputazione multipla è che ogni dato mancante è imputato m volte in modo da creare m dataset completi. A differenza dell'imputazione singola, imputare diverse volte lo stesso valore serve ad aggiungere variabilità alle stime poiché permette di creare m dataset completi con valori imputati diversi. Il principale svantaggio del metodo è il costo computazionale molto oneroso, poiché dopo aver imputato i dataset è necessario ripetere le analisi m volte e combinare successivamente i risultati.
Il processo di imputazione multipla è costituito da tre differenti fasi: (1) la fase di imputazione che crea i dataset completi sostituendo i valori NA presenti, (2) la fase di analisi degli m dataset creati e infine (3) la fase di *pooling* che unifica le diverse stime ottenute dalle analisi per ottenere un unico valore di stima per ogni parametro d'interesse.
É importante ricordare che lo scopo principale dell'analisi statistica è deter-

minare i parametri che descrivono la distribuzione dei dati. L'imputazione multipla non è altro che uno strumento matematico che rende possibile il raggiungimento di questo scopo. Un altro aspetto da sottolineare è che l'imputazione multipla e l'algoritmo EM producono asintoticamente gli stessi risultati. Il fatto che le due procedure, di cui una sola sostituisce i valori nel dataset, siano intercambiabili dimostra che l'imputazione non è un metodo problematico nonostante l'elevato costo computazionale.
Il metodo dell'imputazione multipla risulta essere molto interessante poichè, a differenza di altre tecniche di imputazione, tiene conto anche dell'incertezza associata al valore mancante.

4.1 Fase di imputazione

Questa è la fase del metodo che imputa i *missing data* e che crea le differenze tra le diverse tecniche. Un primo modo per imputare i dati mancanti consiste nell'adottare le migliori tecniche descritte per le imputazioni singole e ripeterle per ognuno degli m dataset. In questo modo si ha la possibilità di realizzare la stessa analisi su dataset leggermente diversi e ottenere diverse stime dei parametri di interesse. La media di queste stime poi fornisce la stima finale del parametro.
I metodi usati si differenziano a seconda della natura della variabile da imputare: per variabili binarie, ad esempio, la tecnica di imputazione più usata prevede l'utilizzo della regressione logistica. I metodi più utilizzati per le imputazioni di variabili di tipo numerico sono quelli già descritti nell'imputazione singola: il metodo *mean* che sostituisce il valore NA con la media della variabile; il metodo *sample* che estrae in modo casuale un valore osservato della variabile e lo sostituisce al dato mancante; il metodo di regressione della variabile con il valore NA sulle altre variabili osservate.
Un metodo molto interessante è il *predictive mean matching*(PMM) che è molto usato per le variabili numeriche che non sono normalmente distribuite. Per descrivere il funzionamento del metodo si suppone che la variabile $\mathbf{Y}$ sia quella che contiene gli NA, e che $\mathbf{X}$ sia un insieme di p variabili completamente osservate che saranno utilizzate per imputare $\mathbf{Y}$. Il metodo agisce in questo modo [34]:

1. Per i casi $\mathbf{Y}_{obs}$, si realizza una regressione lineare di $\mathbf{Y}_{obs}$ su $\mathbf{X}$ producendo i coefficienti $\mathbf{b}$, σ e $\hat{\epsilon}$ attraverso il metodo dei minimi quadrati;

2. Si calcola $\sigma^{*2} = \hat{\epsilon}^T \hat{\epsilon}/A$, con A variabile χ^2 con $n_{obs} - p$ gradi di libertà;

3. Si campiona $\mathbf{b}^*$ in modo casuale dalla distribuzione normale multivariata centrata in $\mathbf{b}$ e con matrice di varianza e covarianza $\sigma^{*2}(\mathbf{X}^T\mathbf{X})^{-1}$;

4. Utilizzando $\mathbf{b}^*$, si generano le previsioni dei valori $\mathbf{Y}$ per tutti i casi, sia dove manca sia dove è osservata;

5. Per i casi $\mathbf{Y}_{mis}$, si identifica un insieme di casi possibili tra quelli osservati $\mathbf{Y}_{obs}$ i cui valori previsti con il modello di regressione $\hat{\mathbf{Y}}_{obs}$ sono simili al valore previsto nei casi $\hat{\mathbf{Y}}_{mis}$;

6. Tra i valori trovati al passo precedente sceglierne casualmente uno e sostituirlo al valore NA;

7. Ripetere i passi da 2 a 5 per ognuno degli m dataset.

Le tecniche più recenti sono state sviluppate in ambito bayesiano. In particolare queste tecniche si focalizzano sulla stima della distribuzione a posteriori dei dati per poi imputare i dati mancanti da essa. Per la stima dei parametri della distribuzione a posteriori viene in aiuto l'algoritmo EM descritto in precedenza, mentre per estrarre i nuovi valori dalla distribuzione a posteriori l'algoritmo più comunemente utilizzato è il Monte Carlo Markov Chain (MCMC).

Un'ulteriore tecnica per imputare i dati mancanti sfrutta l'algoritmo **EM con *bootstrap***. Sebbene richieda delle forti assunzioni teoriche per l'applicazione (quali la normalità dei dati e l'ipotesi che i dati siano MAR), esso restituisce dei buoni risultati anche nel caso in cui esse non siano strettamente soddisfatte.
Si consideri un dataset incompleto costituito da n osservazioni da un vettore aleatorio $\mathbf{Y} = (y_1, ..., y_p)$ p-dimensionale e sia

$$\mathbf{Y} \sim N(\mu, \Sigma) \tag{4.1}$$

dove $\mu \in \mathbb{R}^n$ e $\Sigma \in \mathbb{R}^{n \times n}$. Il dataset contiene dei valori NA, quindi sarà usata la notazione $Y = (Y_{obs}, Y_{mis})$ per evidenziare i due diversi tipi di dato. L'algoritmo EM con *bootstrap* può essere diviso in tre fasi: una prima fase di *bootstrapping*, una fase in cui è applicato l'algoritmo EM e una fase di imputazione dei dati mancanti.

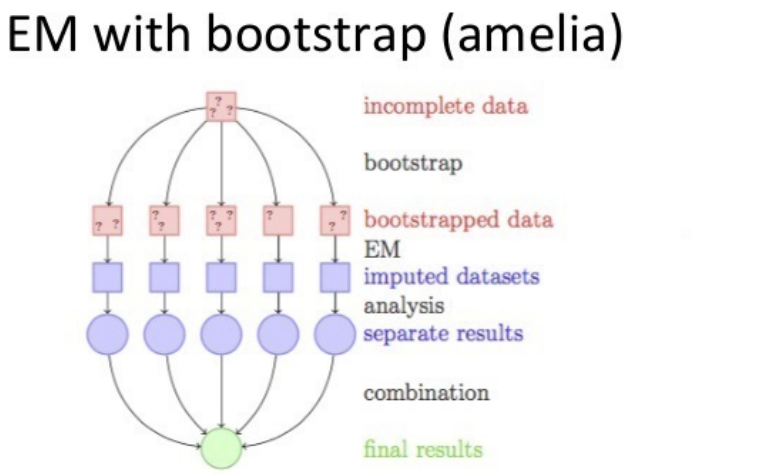

Figura 4.1: Schema del funzionamento del metodo di imputazione EM con bootstrap.

Ricordando che l'algoritmo ha l'obiettivo di creare m dataset a partire da un dataset incompleto Y, è necessario descrivere i dettagli delle sue fasi per comprenderne meglio il funzionamento:

Fase *bootstrap*. Questa prima fase dell'algoritmo consiste nella creazione degli m dataset. Ognuno di essi è generato usando la tecnica bootstrap che consiste nell'estrarre un campione in modo casuale da Y con reimmissione. É importante sottolineare come questa tecnica non crea dataset completi e ciò rende necessario il secondo passo.

Fase EM. In questo passo è applicato l'algoritmo EM descritto nei capitoli precedenti: per ogni dataset creato col metodo *bootstrap* esso stima la media $\tilde{\mu}_i$ e la matrice di varianza e covarianza $\tilde{\Sigma}_i$.

Fase di imputazione. L'ultima fase prevede la creazione dei dataset imputati che poi saranno analizzati. I valori sono generati in modo casuale dalla distribuzione a posteriori $P(Y_{mis}|Y_{obs}, \tilde{\mu}_i, \tilde{\Sigma}_i)$ per ogni i= 1,2,..,m.

4.2 Fase di analisi e fase di *pooling*

Ottenuti gli m dataset dalla fase precedente è ora possibile analizzarli e estrarre le informazioni e le stime dei parametri di interesse da ognuno di essi. In questo modo si ottengono m stime di uno stesso parametro, uno per ogni dataset imputato, che vengono mediate in un'unica stima:

$$\tilde{\theta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}_i. \tag{4.2}$$

La varianza della stima puntuale invece è la media delle varianze stimate all'interno di ogni set di dati imputato più la varianza campionaria nelle stime puntuali tra i set di dati (moltiplicata per un fattore che corregge il bias perché $m < \infty$). Quindi per calcolare la deviazione standard è necessario calcolare la varianza *within* nei dataset e quella *between* tra i dataset:

$$V_w = \frac{1}{m} \sum_{i=1}^{m} diag(\hat{U}_i) \tag{4.3}$$

dove $\hat{U}_i$ è una stima della matrice di varianza e covarianza del parametro $\hat{\theta}_i$, mentre la varianza tra i dataset imputati è così definita

$$V_b = \frac{1}{m-1} \sum_{i=1}^{m} (\tilde{\theta} - \hat{\theta}_i)(\tilde{\theta} - \hat{\theta}_i)^T \tag{4.4}$$

La varianza totale quindi risulta essere

$$V_t = V_w + V_b + \frac{V_b}{m}, \tag{4.5}$$

dove il terzo termine tiene conto del fatto che il numero delle imputazioni utilizzate è in numero finito: in pratica può essere considerato come un errore di simulazione e dunque non è presente se il numero di simulazioni è molto grande.

Capitolo 5

Il Dataset INVALSI

La prova INVALSI è un test standardizzato creato dall'Istituto Nazionale di Valutazione del Sistema Educativo di Istruzione e di Formazione (INVALSI), somministrato agli studenti di vari ordini e gradi di scuola, che rappresenta uno strumento utile per una valutazione più accurata e complessiva della preparazione degli studenti di tutta Italia. É uno strumento utile per i responsabili del sistema educativo italiano (come Ministero dell'istruzione, istituti di ricerca, uffici scolastici) per studiare le politiche di intervento individuando le potenzialità e i limiti del sistema didattico ed educativo grazie alle analisi dei risultati ottenuti nel test. Il test prevede una prova scritta di matematica e una di italiano; lo studio e le analisi dei capitoli successivi sono state realizzate sulle valutazioni degli studenti di prima media nel test di matematica dell'anno scolastico 2012/2013.

5.1 Descrizione del dataset

Le tecniche e i metodi descritti nei capitoli precedenti sono stati applicati ad un dataset che contiene informazioni sulle prove INVALSI degli studenti della prima media di tutta Italia.
Il dataset contiene 509371 sudenti e ognuno di essi è descritto dai valori di 41 variabili aleatorie. Ogni studente è identificato da un codice ID che è costituito da tre parti: una prima parte identifica la scuola, una seconda parte la classe d'appartenenza e una terza parte identifica lo studente. Si vengono quindi a definire tre possibili livelli di analisi dei dati che permettono di confrontare i risultati delle prove tra gli studenti di una stessa classe,

tra le classi di una stessa scuola, o tra le scuole di Italia.

Le covariate presenti nel dataset sono anch'esse caratterizzate da questa suddivisione in livelli (si veda Tabella 5.1).
Considerando le **caratteristiche del singolo studente**, le informazioni disponibili riguardano il sesso (*FEMMINA* assume valore 1 se lo studente è femmina e 0 se maschio), lo stato di immigrazione (Italiano, immigrato di prima generazione (*IMM1*) o immigrato di seconda generazione (*IMM2*)), se lo studente ha anticipato l'iscrizione alla scuola (*ANTICIPATARIO*, cioè se ha iniziato la prima elementare all'età di cinque anni, invece che, come avviene per la maggioranza, all'età di sei anni) o se ha posticipato l'inizio della scuola (*POSTICIPATARIO*, questo è il caso di studenti ripetenti o di studenti immigrati che iniziano la scuola un anno dopo). Il dataset contiene informazioni anche sullo stato familiare: una variabile binaria assume valore 1 se lo studente non vive con entrambi i genitori (*No_genitori*) e un'altra indica se ha fratelli (*Si_fratelli*). Inoltre una variabile importante fornisce informazioni sulla condizione socio-economica e culturale della famiglia, l'indicatore chiamato Economic and Social Cultural Status (*ESCS*), creato in analogia a quello proposto dall'Organisation for Economic Co-operation and Development (OECD) in base al titolo di studio, la professione dei genitori e il possesso di certi beni in casa (per esempio il numero di libri). Una volta definito questo indicatore per tutti gli studenti, esso viene standardizzato in modo che abbia media zero e varianza uno. Il valore minimo e massimo osservati nel dataset sono -3.11 e 2.67. In generale gli alunni con un indicatore *ESCS* superiore a 2 sono in un'ottima condizione socio-economica. Infine ci sono le variabili che registrano le valutazioni degli studenti nella prova INVALSI: $MATH\_5$ contiene la valutazione della prova di matematica dell'anno precedente; $MATH$ la valutazione della prova dell'anno corrente; e $MATH\_corr$ che contiene le valutazioni degli alunni corrette per il *cheating*. Quest'ultima variabile corregge le valutazioni considerando eventuali suggerimenti da parte degli insegnanti o possibili copiature.
Come anticipato in precedenza, il dataset permette di esplorare diverse **caratteristiche a livello classe**, tra le quali i valori medi delle variabili degli studenti appartenenti a quella determinata classe (per esempio: la media di classe dell'indicatore *ESCS*, la proporzione di studenti immigrati nella classe, ecc.).

Per quanto riguarda le variabili che descrivono le **caratteristiche della scuola**, è presente nel dataset una variabile binaria *Tempo Pieno* che definisce lo svolgimento temporale delle lezioni (le scuole a tempo pieno svolgono attività educative nel pomeriggio e non prevedono lezioni il Sabato, mentre quelle tradizionali terminano le lezioni all'orario di pranzo tutti i giorni dal lunedì a sabato compreso). Come succede a livello classe altre variabili a livello scuola sono date dalla media dei valori delle variabili a livello studente che frequentano quella determinata scuola: per esempio si avranno la percentuale media di immigrati nella scuola, l'$ESCS$ medio della scuola, ecc.
É opportuno precisare che le variabili a livello classe e a livello scuola non sono le semplici medie dei valori assunti dalle variabili corrispondenti a livello studente considerando le sole unità statistiche presenti nel dataset. Ciò accade poiché non tutti gli studenti della classe hanno svolto la prova: per esempio il numero di studenti della classe (n_stud_cla) non coincide con il numero di studenti della stessa classe presenti nel dataset (analogamente per la scuola).
Sono presenti inoltre due importanti variabili binarie che permettono di distinguere (i) le scuole private da quelle pubbliche ($PARITARIA$), e gli (ii) Istituti Comprensivi (IC), che includono sia le scuole elementari che le scuole medie all'interno dellla stessa struttura. Quest'ultima variabile è importante perchè permette di capire se la continuità dello stesso ambiente educativo influisce in modo positivo o negativo sulla valutazione degli studenti nel test INVALSI. Alcune variabili di tipo numerico forniscono informazioni sulle dimensioni della scuola (numero di studenti, numero di classi, numero medio di studenti per classe). Infine, due variabili binarie permettono di localizzare la scuola nel Centro o nel Sud Italia: questa divisione è stata introdotta poichè precedenti studi in letteratura hanno dimostrato che gli studenti del Nord Italia tendono ad avere punteggi più elevati rispetto ai loro coetanei nelle altre regioni [2]. Le scuole e le classi sono identificate da un codice che, analogamente al codice studente, permette di distinguere le diverse classi e scuole in modo anonimizzato.

Il dataset contiene inoltre alcune variabili che possono essere considerate le variabili "output": tra esse vi sono la valutazione del test di matematica della prova INVALSI ($MATH$) e il voto ufficiale della prova ($MATH\_corr$). Quest'ultima è la vera variabile risposta del dataset: essa

contiene la valutazione finale della prova in seguito alla correzione del voto $MATH$ a causa del *cheating*, basandosi sulla variabilità della percentuale di risposte corrette nella classe e i tipi di risposte sbagliate. In aggiunta a queste due variabili, che descrivono la valutazione dei test per i singoli alunni nell'anno corrente, è presente la valutazione della prova della quinta elementare. I punteggi delle prove INVALSI della quinta elementare sono utilizzati come controllo nel modello multilivello per stimare l'effetto casuale della scuola. É infatti noto che l'educazione è un processo cumulativo: ciò che è stato appreso in un anno ha effetti sui risultati dell'anno successivo.
Sfortunatamente ci sono moltissimi dati mancanti nelle valutazioni della prova della quinta elementare. La perdita di tali dati è dovuta probabilmente ad una mancanza nel passaggio di informazione dalle scuole elementari alle scuole medie. Per trattare questi dati mancanti sono state utilizzate le diverse tecniche descritte nei capitoli precedenti, a partire dalla *listwise deletion* fino ai diversi metodi di imputazione.
Nella Tabella 5.1 sono descritte le variabili presenti nel dataset e le percentuali di dati mancanti.

Tabella 5.1: Variabili del dataset e percentuali di NA

Tipo	*Variabile*	*Media*	*% di NA*
	Livello Studente		
-	CODICE_STUDENTE	-	-
(S/N)	FEMMINA	48.93%	0.15%
(S/N)	IMM1	6%	8.11%
(S/N)	IMM2	5%	8.11%
numerica	ESCS	0.14	8.29%
(S/N)	ANTICIPATARIO	2.2%	0.15%
(S/N)	POSTICIPATARIO	7.3%	0.15%
(S/N)	No_genitori	13.7%	3.63%
(S/N)	Si_fratelli	84.2%	3.54%
numerica	MATH	46.16	-
numerica	MATH_corr	44.93	0.02%
numerica	MATH_5	70.29	46.48%
	Livello Classe		
-	CODICE_CLASSE	-	-
%	ESCS_medio		8.05%
%	Perc_femmine	43.56%	0.02%
%	Perc_IMM1	4.85%	7.95%
%	Perc_IMM2	4.35%	7.95%
%	Perc_Anticipatari	1.95%	0.02%
%	Perc_Posticipatari	6.38%	0.02%
%	Perc_DISABILI	5.42%	-
numerica	n_stud_cla	23.14	-
(S/N)	tempo_pieno	2.0%	7.95%
	Livello Scuola		
-	CODICE_SCUOLA	-	-
numerica	N_CLASSI	6.48	-
numerica	ESCS_scu_medio	0.13	5.83%
%	Perc_femmine_scu	43.23	-
%	Perc_IMM1_scu	4.73	5.87%
%	Perc_IMM2_scu	4.23	5.87%
%	Perc_Anticipatari_scu	1.93	-
%	Perc_Posticipatari_scu	6.36	-
numerica	N_stud_scu	149.4	-
numerica	N_Stud_medio	22.67	-
(S/N)	NORD	46.3%	-
(S/N)	CENTRO	17.7%	-
(S/N)	SUD	39.0%	-
(S/N)	PARITARIA	2.92%	-
(S/N)	IC	60.64%	-

Capitolo 6

Test MCAR

Una prima analisi descrittiva del dataset, si veda Tabella 5.1, evidenzia come la percentuale di valori NA si concentri soprattutto a livello studente e in particolare sulla variabile $MATH\_5$, che definisce la valutazione della prova INVALSI di matematica della quinta elementare. La presenza numerosa di questi valori mancanti necessita di uno studio per definirne la natura stessa.
La determinazione della natura dei dati mancanti è un passaggio necessario poiché a seconda della tipologia di *missing data* varia la scelta della tecnica più adatta con cui trattarli. L'obiettivo di questo capitolo dunque è capire in quale delle tre tipologie di *missing data* descritte nel Capitolo 1 (MAR, MCAR, MNAR) si inserisce il dataset INVALSI.
La tipologia di *missing data* è il fattore che determina il metodo con cui trattare i dati: se i dati sono di tipo MCAR, cioè se la distribuzione degli NA all'interno del dataset non dipende né dai valori osservati né dai valori mancanti, allora non è necessario che i dati vengano imputati ed è sufficiente utilizzare la tecnica della *listwise deletion*; se invece la distribuzione dei *missing data* dipende dai valori osservati dei dati e dunque i dati sono MAR, la tecnica della *listwise deletion* non è la più adatta poiché restituisce risultati distorti. Anche nel caso di dati MNAR il metodo di eliminazione delle osservazioni incomplete si rivela essere poco efficace.

6.1 Descrizione del test

L'identificazione della natura del dato mancante è dunque un problema a cui prestare molta attenzione poiché un errore nell'assegnazione della na-

tura del *missing* può portare a risultati molto distorti. Prima di occuparsi dunque dei metodi per imputare gli NA è quindi necessario determinare se il dataset in esame è di tipo MCAR, MAR, o MNAR.
L'ipotesi che i dati siano di tipo MNAR è chiaramente da escludere poichè, per come è costruito il dataset, la mancanza del valore di una o più variabili non ha una motivazione particolare: la maggior parte delle variabili registrate sono state ottenute, presumibilmente, da questionari somministrati alle famiglie degli studenti e l'assenza di un valore di esse è da attribuire probabilmente alla volontà dei genitori di non dichiarare quella particolare informazione. Il significato delle variabili presenti nel dataset dunque non permette di trovare una ragione valida alla mancanza dei valori e ciò fa cadere l'ipotesi che i dati siano di tipo MNAR.
Esclusa la possibilità che i dati siano MNAR, resta da verificare se essi siano MAR o MCAR. In letteratura esistono alcuni test che valutano l'ipotesi che i dati siano di tipo MCAR [21], purtroppo però non sono disponibili test per la verifica di ipotesi MAR.

6.1.1 Test per la verifica della natura del dato mancante

Il test, che mette a confronto l'ipotesi nulla H_0:"i dati sono MCAR" e l'ipotesi alternativa H_1:"i dati non sono MCAR", è stato realizzato tramite la funzione *TestMCARNormality* del pacchetto *MissMech* di R [17]. É molto importante sottolineare che, grazie alla precedente esclusione dell'ipotesi MNAR, se il test rifiuta H_0 allora è possibile affermare che i dati sono di tipo MAR.
La funzione *TestMCARNormality* realizza il test valutando se le matrici di varianza e covarianza nei differenti *missing data pattern* sono uguali o diverse [19]: se le matrici di varianza e covarianza nei diversi *pattern* di dati mancanti risultano essere uguali, allora i dati sono assunti MCAR, se invece le matrici di varianza e covarianza sono differenti al variare dei *missing pattern*, allora è rifiutata l'ipotesi di dati MCAR. La funzione realizza due test: il primo è il test di *Hawkins* [12] modificato per dati non completi, mentre il secondo è un test non parametrico [16].

I Test di omoschedasticità e normalità

Il test di Hawkins [12] è un test di omogeneità delle covarianze e di normalità multivariata. Utilizzando il test in combinazione con il test non parametrico si può fare inferenza riguardo l'omoschedasticità, la distribuzione MCAR dei dati mancanti e la normalità multivariata di un insieme di dati. Se i dati assumono distribuzione normale il rifiuto dell'ipotesi nulla del test di Hawkins implica la eteroschedasticità delle varianze. Se però la distribuzione del campione non è nota, il rifiuto dell'ipotesi nulla del test può essere dovuta alla non-gaussianità o alla non-omogeneità delle matrici di varianza e covarianza o ad entrambi le condizioni. In generale se i dati non sono gaussiani è opportuno agire in questo modo: per prima cosa si applica il test di Hawkins e se il test non rifiuta l'ipotesi nulla allora sono accettate sia l'ipotesi di omogeneità delle covarianze che di gaussianità dei dati; se invece il test di Hawkins rifiuta l'ipotesi nulla H_0 è necessario sfruttare il test non parametrico per la verifica dell'omoschedasticità. Se quest'ultimo test non rifiuta l'ipotesi nulla, allora si può concludere che i dati non sono gaussiani; se invece l'ipotesi nulla è rifiutata si può affermare che le covarianze non sono omogenee.

Test di omoschedasticità sotto l'ipotesi di normalità

Si assume che i dati Y siano non completi e che vi siano g diversi *pattern* di dati mancanti. Si assume inoltre che Y_{ij} sia indipendente da Y_{ik} per ogni i e $j \neq k$ e si denota con p il numero delle variabili del dataset

$$Y_{ij} \sim N_p(\mu_i, \Sigma_i), \quad i = 1, .., g, \;\; j = 1, .., n_i. \tag{6.1}$$

Il test però funziona se applicato a dataset completi e dunque è necessario imputare i dati mancanti per proseguire.
Siano μ_i e Σ_i definiti in questo modo in accordo con le parti osservate e mancanti dei dati e p_i il numero delle variabili osservate nel *pattern* i

$$\mu_i = \begin{bmatrix} \mu_{o,i} \\ \mu_{m,i} \end{bmatrix}, \Sigma_i = \begin{bmatrix} \Sigma_{oo,i} & \Sigma_{om,i} \\ \Sigma_{mo,i} & \Sigma_{mm,i} \end{bmatrix} \tag{6.2}$$

la distribuzione condizionata di $Y_{mis,ij}$ dati $Y_{obs,ij}$, μ_i e Σ_i è

$$Y_{mis,ij}|(Y_{obs,ij},\mu_i,\Sigma_i) \sim N_{p-p_i}(\mu_{m,i}+\Sigma_{mo,i}\Sigma_{oo,i}^{-1}(Y_{obs,ij}-\mu_{o,i}),\Sigma_{mm,i}-\Sigma_{mo,i}\Sigma_{oo,i}^{-1}\Sigma_{om,i}) \tag{6.3}$$

dove la media in (6.3) è il valore atteso condizionato $\mathrm{E[Y_{mis,ij}|Y_{obs,ij}]}$ e la varianza è la varianza condizionata $\mathrm{Var}[Y_{mis,ij}|Y_{obs,ij}]$. In accordo con quanto proposto da Jamshidian e Jalal [16], la distribuzione condizionata appena calcolata è utilizzata per imputare in modo casuale i valori mancanti del dataset. Per l'imputazione però è necessario che siano noti i valori di μ_i e Σ_i. Nella maggior parte delle applicazioni μ_i e Σ_i non sono noti. Poiché l'obiettivo principale è quello di verificare l'ipotesi nulla di uguaglianza delle matrici di covarianza nei gruppi, si assume la media uguale a μ per tutti i g gruppi e la matrice di varianza e covarianza Σ uguali per tutti i gruppi e si stimano μ e Σ con il metodo della massima verosimiglianza [15]. Se le medie non sono uguali, allora possono essere utilizzate le stime MLE di μ_i per ogni gruppo. Siano $\hat{\mu}$ e $\hat{\Sigma}$ le stime di massima verosimiglianza di μ e Σ, rispettivamente. Nella fase di imputazione sono usate queste quantità al posto di μ_i e Σ_i nella distribuzione condizionata per imputare le $Y_{mis,i}$ per ogni i generando valori casuali da una normale multivariata con media e covarianza in (6.3).

Una volta imputati i dati mancanti si può utilizzare il test di Hawkins: il test considera il nuovo dataset completo $\mathbf{X}$ (n x p) costituito dai g gruppi, dove $\mathbf{X}_{ij}$ identifica il j-esimo caso dell' i-esimo gruppo, con j =1,..,n_i e i=1,..,g. Inoltre il test assume che $\mathbf{X}_{ij}$ abbia una distribuzione normale p-variata con media μ_i e covarianza Σ_i. Per testare l'ipotesi di uguaglianza delle matrici di varianza e covarianza nei g gruppi viene calcolata la statistica F_{ij} corrispondente al caso j nel gruppo i definita in questo modo

$$F_{ij} = \frac{(n-g-p)n_iV_{ij}}{p\{(n_i-1)(n-g)-n_iV_{ij}\}}, \quad dove \quad V_{ij} = (\mathbf{X}_{ij}-\bar{\mathbf{X}}_i)^T S^{-1}(\mathbf{X}_{ij}-\bar{\mathbf{X}}_i) \tag{6.4}$$

con $\bar{\mathbf{X}}_i$ e S che rispettivamente corrispondono alla media campionaria del gruppo i e la matrice di varianza e covarianza generale "pooled". Hawkins mostrò che, sotto l'ipotesi di normalità e di uguaglianza delle covarianze, i valori F_{ij} seguono una distribuzione F di Snedecor con gradi di libertà p e $n-g-p$. Propose inoltre di calcolare la statistica $A_{ij} = \mathrm{P}[\mathcal{F} > F_{ij}]$, come la probabilità che una variabile aleatoria $\mathcal{F}$ con distribuzione di Fisher-Snedecor con gradi di libertà p e $n-g-p$ sia maggiore del valore

della statistica F_{ij}. Sotto l'ipotesi di distribuzione normale omoschedastica dei dati la statistica A_{ij} è distribuita come una variabile uniforme in (0,1). In seguito a ciò Hawkins propose di verificare dunque la omoschedasticità delle matrici di varianza e covarianza dei g gruppi tramite la verifica di uniformità della statistica A_{ij}. In pratica se A_{ij} non risulta essere distribuita uniformemente allora è rifiutata l'ipotesi di omoschedasticità o quella di normalità dei dati. Calcolati i valori A_{ij} per ogni gruppo i, i p-value ottenuti dai test di uniformità di A_{ij} sono combinati in un unico valore P_T (vedi [9]), p-value del test che verifica l'uniformità di tutti gli A_{ij}, dove P_T assume valore

$$P_T = \sum_{i=1}^{g}(-2logP_i) \sim \chi^2_{2g}. \tag{6.5}$$

Per calcolare i *p-value* dei test per ogni gruppo Jamshidian e Jalal proposero di utilizzare il test di Neyman [24] per la verifica dell' ipotesi di uniformità del 1937. La statistica del test di Neyman è

$$N_{ik} = \sum_{l=1}^{k}\{n_i^{-1/2}\sum_{j=1}^{n_i}\pi_l(A_{ij})\}^2 \qquad i = 1, ..., g, \tag{6.6}$$

dove π_1,π_2,..,π_k sono polimomi ortonormali su (0,1) scelti come funzioni di base: π_l è un polinomio di grado l ortogonale a tutti i polinomi di grado precedente

$$\int_0^1 \pi_l(y)\pi_k(y)\,dy = 0, \quad \forall k < l \tag{6.7}$$

incluso il polinomio costante $\pi_0(y)$=1, e normalizzati a uno

$$\int_0^1 \pi_l^2(y)\,dy = 1. \tag{6.8}$$

Un polinomio di grado l contiene l+1 coefficienti, genera l+1 equazioni (6.7) e quindi è univocamente determinato. La funzione *TestMCARNormality* implementata nel pacchetto **MissMech** utilizza k pari a 4 come suggerito da Jamshidian e Jalal. Per valori grandi della statistica N_{ik} l'ipotesi nulla di uniformità viene rifiutata. Per ottenere i *p-value* P_i sono simulati un gran numero di valori di N_{ik} simulando un gran numero di A_{ij} da una distribuzione uniforme in (0,1) e calcolando la percentuale di N_{ik} simulati che sono più grandi dei valori N_{ik} ottenuti dai dati. Nei casi dove il numero n_i è grande la statistica N_{ik} ha approssimativamente una distribuzione χ^2

con k gradi di libertà.

Test non parametrico di Jamshidian e Jalal

Il test non parametrico di Jamshidian e Jalal per la verifica dell'ipotesi di omoschedasticità delle matrici di covarianza dei gruppi non assume alcuna forma della distribuzione dei dati (a differenza del test di Hawkins). Per realizzare il test però sono necessari i valori delle statistiche F_{ij} che sono calcolate in precedenza nel test di Hawkins. Naturalmente a differenza del test di Hawkins, poiché la distribuzione dei dati non è nota a priori, anche la distribuzione delle statistiche F_{ij} non sarà nota. Jamshidian e Jalal però hanno dimostrato che se i dati hanno una densità della forma $f(\mathbf{Y}_{ij};\Sigma_i,\, \theta)$ e gli n_i sono pressoché uguali o molto grandi allora sotto l'assunzione di omoschedasticità delle matrici di varianza e covarianza la distribuzione di F_{ij} per tutti i g gruppi deve essere identica. Sfruttando questa informazione la realizzazione del test per l'omoschedasticità richiede due importanti attenzioni: come nel test di Hawkins il calcolo delle statistiche F_{ij} richiede che i dati mancanti vengano prima imputati, questa volta però senza che sia fatta alcuna assunzione sulla distribuzione dei dati; in secondo luogo deve essere impiegato il test $k - sample$ per testare l'uguaglianza delle distribuzioni delle statistiche F_{ij} nei diversi gruppi.

Il metodo di imputazione proposto da Jamshidian e Jalal [16] segue la linea introdotta da Srivastava e Dolatabadi [29, 30] che assume solo l'indipendenza delle osservazioni da caso a caso e la continuità della loro distribuzione cumulativa. In pratica i valori imputati sono ottenuti aggiungendo un appropriato errore aleatorio ai migliori predittori lineari delle osservazioni mancanti. La realizzazione del metodo di imputazione richiede però le stime delle medie e delle covarianze delle variabili; in questo caso le stime sono ottenute dai casi completamente osservati e i valori ottenuti saranno ragionevoli se la quantità di casi completi è in numero sufficiente. La funzione *TestMCARNormality* utilizza questo metodo di imputazione e assume implicitamente che le variabili siano linearmente correlate.

Senza perdite di generalità, si può assumere che il primo gruppo ha n_1 osservazioni tutte completamente osservate (con $n_1 > p$). Siano ora $\bar{\mathbf{Y}}_1$ e $\mathbf{S}_1$ rispettivamente la media e la covarianza ottenute dagli n_1 casi completi.

Se n_1 ha valore piccolo si può sostituire $\bar{\mathbf{Y}}_1$ e $\mathbf{S}_1$ con le stime di massima verosimiglianza $\hat{\mu}$ e $\hat{\Sigma}$. Come nel paragrafo precedente si identificano con $Y_{obs,ij}$ e $Y_{mis,ij}$ i dati realmente osservati e quelli mancanti per il j-esimo caso nel i-esimo gruppo, e quindi si divide anche $\bar{\mathbf{Y}}_1$ e $\mathbf{S}_1$ in questo modo

$$\bar{Y}_1 = \begin{bmatrix} \bar{Y}_{o,1} \\ \bar{Y}_{m,1} \end{bmatrix}, S_1 = \begin{bmatrix} S_{oo,1} & S_{om,1} \\ S_{mo,1} & S_{mm,1} \end{bmatrix}. \tag{6.9}$$

Il miglior predittore lineare per il valore mancante $Y_{mis,ij}$ è dato da

$$\hat{Z}_{mis,ij} = \bar{Y}_{m,1} + S_{mo,1} S_{oo,1}^{-1} (Y_{obs,ij} - \bar{Y}_{o,1}). \tag{6.10}$$

La covarianza condizionata di $\hat{Z}_{mis,ij}$ dato Σ è approssimativamente

$$\frac{1}{n_1}(\Sigma_{mm,i} - \Sigma_{mo,i}\Sigma_{oo,i}^{-1}\Sigma_{om,i}), \tag{6.11}$$

che è più piccola di un fattore $1/n_1$ della varianza condizionata di $Y_{mis,ij}$. Dunque accade che $\hat{Z}_{mis,ij}$ ha meno variabilità di $Y_{mis,ij}$ e quindi non risulta essere appropriato per imputare i valori mancanti di Y_{ij}. Per rimediare a questo inconveniente Srivastava [30] propose di calcolare i seguenti residui per i casi completi:

$$e_j = \left(\frac{n_1}{n_1 - 1}\right)^{1/2} (Y_{1j} - \bar{Y}_1), j = 1, .., n_1. \tag{6.12}$$

Un campione di dimensione $n - n_1$ è estratto con reimmissione da questi residui. Indicando gli elementi del campione con e^*_{ij} , con i=1,..,g, e j=1,..,n_i; la media condizionata e la covarianza condizionata di e^*_{ij} dati i casi completi di Y_1 sono rispettivamente 0 e S_1. Sfruttando i residui appena calcolati si ottengono le seguenti quantità

$$\eta^*_{ij} = e^*_{m,ij} - S_{mo,1} S_{oo,1}^{-1} e^*_{o,ij}, \tag{6.13}$$

dove e^*_{ij}=($e^*_{o,ij}$,$e^*_{m,ij}$) è diviso nella parte osservata e in quella non osservata. Dunque ora un appropriato valore imputato al dato $Y_{mis,ij}$ è ottenuto da

$$\hat{Y}_{mis,ij} = \hat{Z}_{mis,ij} + \eta^*_{ij}. \tag{6.14}$$

La covarianza dei valori imputati $\hat{Y}_{mis,ij}$ sarà più simile alla covarianza di

$Y_{mis,ij}$, se fossero osservati, per valori grandi di n_i.
Ottenuti i valori imputati si può utilizzare il nuovo dataset imputato per testare l'uguaglianza delle distribuzioni delle statistiche F_{ij} tra i g gruppi. Per verificare l'uguaglianza delle F_{ij} si utilizza il test non parametrico $k-$ *sample* di Scholz e Stephens [28](basato sul test di Anderson-Darling [4]). Il test utilizza una statistica della forma $T = \frac{1}{N}\sum_{i=1}^{g} T_i$ con

$$T_i = \frac{1}{n_i}\sum_{j=1}^{N-1}\frac{(NM_{ij} - jn_i)^2}{j(N-j)}, \tag{6.15}$$

dove $N = \sum_{i=1}^{g} n_i$ è la dimensione del campione "pooled" di F_{ij} e M_{ij} è il numero di osservazioni nell'i-esimo campione che non superano il j-esimo ordine statistico nel campione pooled di F_{ij}.
Se i due test appena descritti rifiutano in successione l'ipotesi nulla, allora le covarianze non sono uguali nei diversi *pattern* di dati mancanti e dunque, nel contesto del problema dei *missing data*, significa che i dati non sono MCAR.

6.2 Risultato del Test MCAR sul dataset INVALSI

Prima di proseguire con l'uso delle tecniche di imputazione dei dati, è indispensabile dunque applicare il test per la verifica che i dati siano di tipo MCAR al dataset INVALSI per poter giustificare l'uso dei metodi di imputazione. Se il test dovesse restituire che i dati sono MCAR allora basterebbe trattare i valori NA presenti tramite la classica tecnica della *listwise deletion*.
L'applicazione della funzione *TestMCARNormality* al dataset INVALSI ha causato problemi di allocazione di memoria. Il numero elevato di osservazioni del campione (509371) e il numero elevato di variabili per ogni studente (41) fanno sì che il numero di *pattern* possibili nel dataset sia molto alto, e dunque il numero di gruppi g di cui il test deve confrontare la matrice di covarianza sono molti. Il test di Hawkins valuta anche la normalità dei dati e quindi sono state escluse dal test le molte variabili binarie presenti nel dataset. Ulteriori problemi di memoria non hanno reso possibile la

realizzazione del test su un campione più numeroso di 10000 osservazioni scelte in modo casuale.
Il test di omogeneità delle matrici di varianza e covarianza realizzato sulle sole variabili numeriche riconosce 8 *missing data pattern* sulle 10000 osservazioni del campione.
Il primo test, quello di Hawkins, restituisce un *p-value* uguale a 0. L'interpretazione di questo valore è che o l'ipotesi di normalità dei dati o l'ipotesi di omoschedasticità delle matrici di covarianza o entrambi le ipotesi sono rifiutate con un livello di significatività al 5%. Assumendo che i dati siano distribuiti in modo gaussiano il test rifiuta comunque l'ipotesi che i dati siano MCAR ad un livello di significatività del 5%.
Il test non parametrico di omoschedasticità, che come precedentemente detto non assume nessuna distribuzione dei dati, restituisce un *p-value* pari a 5.29×10^{-14}. Questo valore molto piccolo fa rifiutare nuovamente l'ipotesi nulla che i dati siano di tipo MCAR.
I test realizzati dunque restituiscono dei risultati che sono quelli attesi. Il test è realizzato su un sottocampione del dataset completo e risulta che esso non è di tipo MCAR. É possibile quindi estendere i risultati a tutto il campione e concludere che i dati non sono MCAR. Il rifiuto dell'ipotesi di MCAR permette di non considerare la tecnica *listwise deletion* come la più adatta allo studio del dataset e di sfruttare le tecniche di imputazione descritte nel Capitolo 4. Questo conclusione porta a considerare i dati MAR o MNAR: la costruzione del dataset e la natura delle covariate porta a escludere la possibilità che i dati siano MNAR poiché non vi sono ragioni specifiche per giustificare la mancanza di alcuni dati.

Capitolo 7

Modello

L'interesse nello studio di questo dataset sta nell'analisi delle relazioni tra il voto della prova INVALSI nella classe prima media e le altre variabili. In particolare si vuole analizzare se la valutazione della prova è influenzata in qualche modo dalle informazioni che descrivono gli studenti, le classi e le scuole. La variabile di interesse è $MATH\_corr$, che corrisponde al voto della prova INVALSI di matematica degli studenti della prima media delle scuole italiane normalizzata e corretta per il *cheating*. I valori di questa variabile sono stati corretti a seguito di possibili imbrogli da parte degli studenti o da parte degli insegnanti che hanno suggerito qualche risposta.
Lo scopo di questa analisi è individuare quali caratteristiche dello studente hanno un effetto positivo o negativo sulle realizzazioni della variabile $MATH\_corr$ e stimare gli impatti della scuola sugli studenti, in modo da valutare se una scuola ha un particolare effetto positivo o negativo sul rendimento dello studente. Per questa ragione da qui in avanti l'analisi si è focalizzata sulle variabili che definiscono le caratteristiche dello studente e le variabili *NORD*, *CENTRO* e *SUD*, che definiscono la locazione della scuola frequentata, il *CODICE_SCUOLA* e la variabile *IC* che distingue gli istituti comprensivi da quelle scuole che non lo sono.
Il modo per capire queste dipendenze è dato dai modelli multilivello lineari a effetti misti (vedi [10]) che consentono di decomporre la variabilità totale in parti che variano nei diversi livelli di studio. Prima di descrivere il modello e analizzarne i risultati è interessante vedere se sono presenti evidenti differenze tra le valutazioni di matematica al variare del valore assunto dalle variabili binarie. Questa differenza può essere visualizzata realizzando dei boxplot.

Come si può notare dal boxplot in Figura 7.1, i voti delle prove svolte

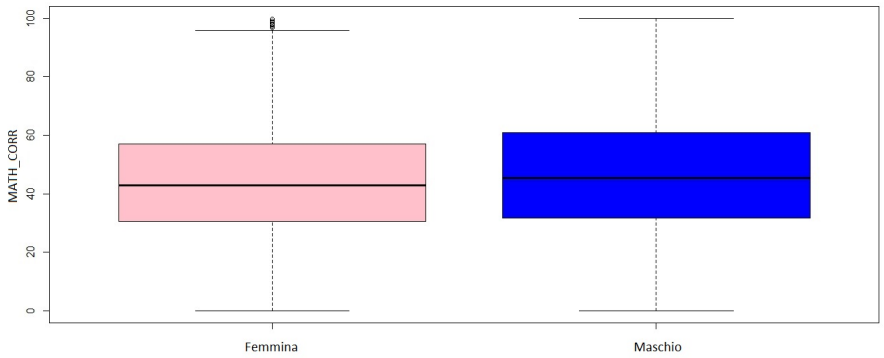

Figura 7.1: Boxplot della variabile $MATH\_corr$ al variare del sesso.

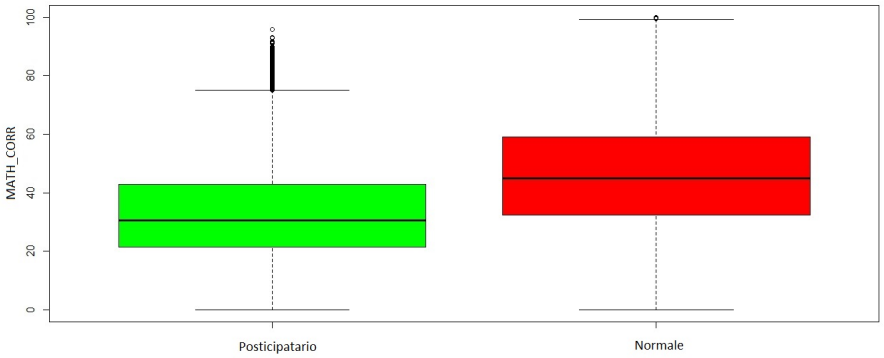

Figura 7.2: Boxplot della variabile $MATH\_corr$ al variare della variabile $POSTICIPATARIO$.

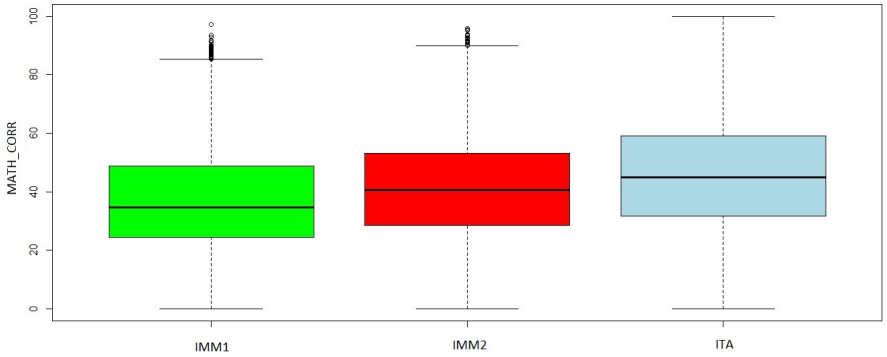

Figura 7.3: Boxplot della variabile $MATH\_corr$ al variare della nazionalità.

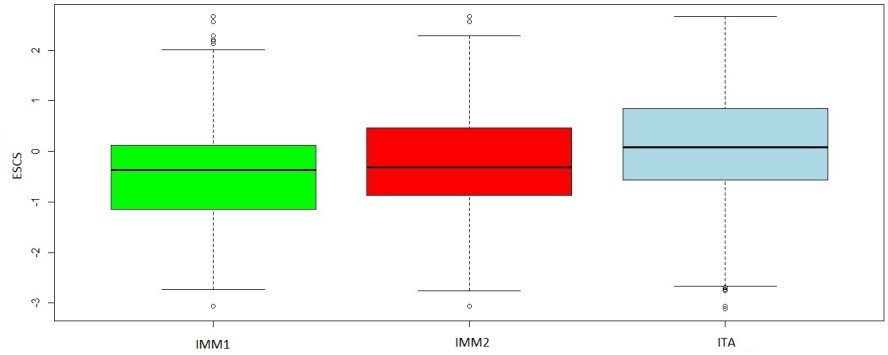

Figura 7.4: Boxplot della variabile $ESCS$ al variare della nazionalità.

dagli studenti maschi sono leggermente più alti di quelli delle loro coetanee. A causa dell'elevata dimensione del dataset non è stato possibile verificare la normalità dei dati e dunque è stato applicato il test di Wilcoxon, ossia un test non parametrico per verificare, in presenza di valori ordinali provenienti da una distribuzione continua, se due campioni statistici provengono dalla stessa popolazione. L'esito del test ha permesso di affermare che vi è differenza statisticamente significativa tra le distribuzioni dei voti nei due sessi, avendo ottenuto un *p-value* di 2.2×10^{-16}. Nel grafico 7.2 si può notare come gli alunni che hanno iniziato la scuola un anno dopo, poiché immigrati, o gli alunni che hanno dovuto ripetere l'anno, abbiano un voto molto più basso dei loro compagni di classe (anche in questo caso il *p-value* del test di Wilcoxon è 2.2×10^{-16}). Per quanto riguarda gli studenti stranieri, negli ultimi due boxplot è ben evidente che la prima e la seconda generazioni di immigrati hanno dei voti molto più bassi rispetto agli studenti italiani; il test di Wilcoxon restituisce un risultato concorde all'analisi puramente grafica. L'ultimo grafico in Figura 7.4 descrive la condizione socio-economica delle famiglie degli studenti analizzando il valore della variabile ESCS per gli studenti italiani e stranieri. Si può ben notare come le distribuzioni della variabile siano molto diverse nei tre casi: le famiglie degli studenti immigrati di prima generazione hanno una condizione socio-economica ben più bassa di quella delle famiglie degli studenti immigrati di seconda generazione; questi ultimi a loro volta si trovano in condizioni inferiori a quelli delle famiglie degli studenti italiani. Il test di Kruskal-Wallis, realizzato a supporto dell'analisi puramente grafica, ha evidenziato con un *p-value* pari a 2.2×10^{-16} che la media della variabile $ESCS$ è differente se lo studente

è italiano o immigrato di prima e seconda generazione.

7.1 Modello a livello studente

Il primo modello proposto studia l'efficacia della scuola considerando le variabili a livello studente con l'aggiunta di un effetto casuale scuola. In questo modo si può rilevare come la variabile risposta *MATH_corr*, che definisce le abilità matematiche acquisite, sia legata alle caratteristiche degli studenti, e quale sia il valore aggiunto che la scuola dà a questo punteggio. Si adotta perciò un modello lineare a effetti misti a due livelli, in cui lo studente i (primo livello) è annidato nella scuola j (secondo livello) [1, 23]. Il modello usato è il seguente

$$y_{ij} = \beta_0 + \sum_{k=1}^{K} \beta_k x_{kij} + b_j + \epsilon_{ij} \tag{7.1}$$

$$b_j \sim N(0, \sigma_b^2), \quad \epsilon_{ij} \sim N(0, \sigma_\epsilon^2) \tag{7.2}$$

dove

- y_{ij} è il voto del test di matematica dello studente i nella scuola j;
- x_{kij} è il corrispondente valore del k-esimo predittore a livello studente;
- β=(β_0,..,β_k) è il vettore (K+1) dimensionale dei parametri da stimare;
- b_j è l'effetto casuale della j-esima scuola (si assume distribuita come una gaussiana ed indipendente dagli altri predittori che sono inclusi nel modello);
- ϵ_{ij} è l'errore gaussiano con media nulla;

Questo modello, che coinvolge il livello studente ed il livello scuola, è stato applicato ai dati originali.

7.1.1 Modello a livello studente con effetto casuale scuola con *listwise deletion*

Il modello appena descritto è stato applicato al dataset realizzato con la tecnica più comunemente utilizzata della *listwise deletion*. I risultati del

modello riportati in Tabella 7.1 si riferiscono alle stime dei coefficienti dopo l'esclusione dei valori mancanti.

Quasi tutte le variabili risultano significative, ad eccezione di *POSTICIPATARIO* e *Si_fratelli* che hanno un *p-value* alto, e inoltre una correlazione pressoché nulla con la risposta. In accordo con quanto visto graficamente nel boxplot di Figura 7.1, il voto della prova di matematica è più alto per gli studenti maschi rispetto alle femmine. Le variabili che definiscono se l'alunno è immigrato o figlio di genitori immigrati contribuiscono negativamente al voto della prova. Il coefficiente positivo della variabile $MATH\_5$ suggerisce che lo studente con un alto risultato della prova INVALSI nella quinta elementare continuerà ad avere buoni risultati anche nella classe prima media. Un importante risultato riguarda le aree geografiche: rispetto al Nord, frequentare una scuola nel Sud Italia comporta una riduzione del

Effetti Fissi		
Covariate	*Stima*	*Deviazione Standard*
Intercetta	10.558 ***	0.198
FEMMINA	-2.106 ***	0.053
IMM1	-1.097 ***	0.151
IMM2	-1.970 ***	0.129
CENTRO	-2.606 ***	0.262
SUD	-6.384 ***	0.207
ANTICIPATARIO	-0.486 ***	0.201
POSTICIPATARIO	-2.674 *	0.178
ESCS	2.460 ***	0.029
No_genitori	-1.334 ***	0.081
Si_fratelli	0.089	0.073
MATH_5	0.568 ***	0.001
Effetto Casuale		
σ_b	5.145	
σ_ϵ	13.521	
Dimensioni		
Numero di osservazioni	259757	
Numero di gruppi	4119	
VPC	14.43%	

Tabella 7.1: Stime del modello (7.1) livello studente con effetto aleatorio scuola dopo LD. Gli asterischi individuano i diversi livelli di significatività: . $0.01 <$ *p-value* < 0.1; * $0.001 <$ *p-value* $<$ 0.01; ** $0.0001 <$ *p-value* < 0.001; *** *p-value* < 0.0001.

voto medio di matematica di più di 6 punti, mentre nel Centro la riduzione è di 2 punti. Questo importante risultato ha portato a considerare le tre arre geografiche separatamente.
Un ulteriore indice di fondamentale importanza è il VPC (*Variance Partition Coefficient*) che indica la percentuale di variabilità totale spiegata dagli effetti aleatori del modello

$$VPC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\epsilon^2}. \tag{7.3}$$

Il 14.44% della varianza totale del modello è spiegata dalla varianza dell'effetto aleatorio scuola. Questo suggerisce che il livello di preparazione non è omogeneo in tutte le scuole d'Italia.

7.1.2 Modello a livello studente con effetto casuale scuola per dati imputati

Il modello (7.1) è stato applicato al dataset creato dopo l'utilizzo della *listwise deletion*. Avendo però dimostrato, tramite il test di omogeneità delle covarianze, che i dati mancanti sono di tipo MAR, si adottano i metodi di imputazione descritti nel Capitolo 4 per imputare i *missing data* e ottenere le nuove stime dei coefficienti del modello.
Si può ben notare come i risultati in Tabella 7.2 siano diversi a seconda del metodo di imputazione utilizzato. Le stime sono state calcolate tramite tre diversi pacchetti di R: le stime della *listwise deletion* sono state realizzate con il pacchetto **ForImp** [5], la tecnica che usa l'algoritmo *Expectation-Maximization con bootstrap* (EMB) (vedi Capitolo 4) tramite il pacchetto **amelia** [13] e le altre tecniche con il pacchetto **mice** [32]. Il metodo *sample*, che imputa il dato mancante con un campionamento casuale dai dati osservati, e il metodo *mean*, che invece sostituisce la media della variabile al posto del valore NA, restituiscono dei valori di stima diversi da quelli degli altri metodi. Questa situazione è dovuta alla semplicità dei metodi applicati, i quali imputano i valori mancanti di una variabile senza tenere in considerazione le relazioni che la variabile stessa ha con le altre covariate del dataset.
Gli altri tre metodi invece sfruttano le interazioni tra le variabili per imputare i dati, realizzando delle regressioni lineari della variabile di cui manca il valore sulle altre covariate osservate nel caso *reg*, realizzando la stessa re-

Tabella 7.2: Stime del modello (7.1) livello studente con effetto aleatorio scuola per dati dopo LD e imputati con i metodi *EMB*, *Mean*, *Sample*, *Reg*, *Reg_Bay* e *PMM*. Gli asterischi individuano i diversi livelli di significatività: . 0.01 < *p-value* < 0.1; * 0.001 < *p-value* < 0.01; ** 0.0001 < *p-value* < 0.001; *** *p-value* < 0.0001.

Stime Effetti Fissi							
	LD	*EMB*	*Sample*	*Mean*	*Reg*	*Reg Bay*	*PMM*
Intercetta	10.558 ***	11.836 ***	30.830 ***	12.448 ***	11.845 ***	11.883 ***	13.186 ***
FEMMINA	-2.106 ***	-1.879 ***	-2.480 ***	-2.236 ***	-1.887 ***	-1.883 ***	-1.915 ***
IMM1	-1.097 ***	-0.975 ***	-2.694 ***	-2.313 ***	-0.948 ***	-0.941 ***	-0.943 ***
IMM2	-1.970 ***	-1.418 ***	-2.634 ***	-2.091 ***	-1.408 ***	-1.389 ***	-1.462 ***
CENTRO	-2.606 ***	-2.523 ***	2.480 ***	-2.534 ***	-2.527 ***	-2.513 ***	-2.531 ***
SUD	-6.384 ***	-7.347 ***	-7.853 ***	-7.857 ***	-7.363 ***	-7.376 ***	-7.297 ***
ANTICIPATARIO	-0.486 ***	-0.290	-0.407	-0.347	-0.272	-0.203	-0.294 *
POSTICIPATARIO	-2.674 *	-5.549 ***	-8.695 ***	-8.431 ***	-5.523 ***	-5.522 ***	-5.503 ***
ESCS	2.460 ***	2.640 ***	3.426 ***	3.304 ***	2.643 ***	2.644 ***	2.691 ***
No_genitori	-1.334 ***	-1.257 ***	-1.973 ***	-1.812 ***	-1.270 ***	-1.276 ***	-1.313 ***
Si_fratelli	0.089	-0.020	-0.145	-0.123	-0.03	-0.032	-0.019
MATH_5	0.568 ***	0.539 ***	0.272 ***	0.532 ***	0.540 ***	0.539 ***	0.521 ***
Effetti Casuali							
σ_b	5.145	5.491	6.435	6.494	5.272	5.243	5.226
σ_ϵ	13.521	13.951	15.789	15.161	13.940	13.948	14.031
VPC	12.64%	13.41%	14.24%	15.50%	12.51%	12.38%	12.18%

gressione ma in ambito bayesiano nel caso *reg_ bay*, e campionando in modo casuale dalla distribuzione a posteriori nel caso *PMM* (vedi Capitolo 4.1 per i dettagli). Le stime prodotte da questi ultimi tre metodi sono coerenti tra loro e ben diverse da quelle ottenute con la semplice *listwise deletion*.
Considerando numericamente i valori ottenuti dalle stime dei predittori, si può notare come le variabili che identificano la zona d'Italia in cui la scuola è ubicata abbiano dei valori importanti e negativi: il fatto di frequentare una scuola nel meridione abbassa il voto medio della prova di più di 7 punti e di circa 2.5 punti per chi vive nel centro Italia. Le analisi dei risultati delle variabili *FEMMINA* e *ESCS* restano identiche a quelle esposte nei paragrafi precedenti. La variabile *POSTICIPATARIO* contribuisce negativamente alla valutazione: lo studente che deve ripetere la classe o che inizia l'anno dopo, come per gli studenti immigrati, ha un punteggio medio ridotto di 5.5 punti rispetto ai suoi compagni. Come visto anche in precedenza la variabile *MATH_ 5* contribuisce con un coefficiente pari a 0.5 circa alla valutazione finale della prova.
La varianza degli effetti casuali dovuti alla scuola non assume valori molto diversi a seconda del metodo utilizzato per imputare i dati: ad esclusione dei due metodi banali *mean* e *sample* in cui la deviazione standard dell'effetto aleatorio scuola è maggiore di 6, per gli altri metodi il valore è pressochè identico e pari a 5.2. In modo analogo si comporta l'indice VPC che evidenzia come il 12% circa della varianza totale è spiegata dall'effetto aleatorio scuola.

7.2 Test sui coefficienti del modello tra le aree

Le valutazioni fatte nel paragrafo precedente riguardo alla significatività dell'area geografica d'Italia in cui si frequenta la scuola hanno portato a pensare di poter dividere il dataset totale in tre sottocampioni corrispondenti alle aree Nord, Centro e Sud Italia. Applicando successivamente lo stesso modello (7.1) ai tre dataset è possibile confrontare le stime dei coefficienti di regressione e verificare l'effettiva differenza dei risultati. Il solo confronto dei valori dei coefficienti sul singolo valore numerico però non tiene in considerazione la distribuzione del parametro. Sfuttando una tecnica di ricampionamento si possono generare diverse stime dei parametri in modo da avere a disposizione un campione numeroso con cui realizzare

dei test per confrontare i coefficienti delle diverse aree.

Si prendano in considerazione due modelli lineari indipendenti applicati a due dataset diversi ma con le stesse covariate

$$\textbf{Modello 1} : y_{ij} = \beta_0 + \sum_{k=1}^{K} \beta_k x_{kij} + b_j + \epsilon_{ij} \tag{7.4}$$

$$\textbf{Modello 2} : \tilde{y}_{ij} = \alpha_0 + \sum_{k=1}^{K} \alpha_k x_{kij} + \tilde{b}_j + \tilde{\epsilon}_{ij} \tag{7.5}$$

Il test ha lo scopo di confrontare i coefficienti dei due modelli e per far ciò le loro stime vengono assunte per ipotesi distribuite secondo una gaussiana.

$$\hat{\beta}_k \sim N(\beta_k, \sigma^2_{\beta_k}) \qquad \hat{\alpha_k} \sim N(\alpha_k, \sigma^2_{\alpha_k}) \qquad \hat{\alpha_k} \perp \hat{\beta}_k \;\; \forall k \tag{7.6}$$

dove α_k e β_k sono i veri valori dei coefficienti e $\hat{\alpha_k}$ e $\hat{\beta}_k$ le loro stime. Successivamente si estraggono M stime dai dati usati per realizzare il modello 1 e M dai dati usati per il modello 2 utilizzando queste ultime per realizzare le M stime dei coefficienti α_k e β_k. Dunque ora si avrà

$$\hat{\vec{\beta}}^{(1)}, \hat{\vec{\beta}}^{(2)}, ..., \hat{\vec{\beta}}^{(M)} \quad da \;\; N(\vec{\beta}, \Sigma^2_{\vec{\beta}}) \tag{7.7}$$

$$\hat{\vec{\alpha}}^{(1)}, \hat{\vec{\alpha}}^{(2)}, ..., \hat{\vec{\alpha}}^{(M)} \quad da \;\; N(\vec{\alpha}, \Sigma^2_{\vec{\alpha}}). \tag{7.8}$$

Il test che si vuole realizzare è dunque un test di Hotelling sul confronto tra le medie di due popolazioni gaussiane con varianza incognita e diversa, cioè con ipotesi nulla H_0: $\vec{\alpha}$=$\vec{\beta}$ e ipotesi alternativa $H_1 : \vec{\alpha} \neq \vec{\beta}$.

La realizzazione di questo test ha lo scopo di confrontare le medie delle stime dei coefficienti di uno stesso modello applicato a due diversi dataset. L'interesse è quello di valutare se vi è evidenza statistica che le stime siano diverse nelle tre aree d'Italia. Il test è stato realizzato ponendo M uguale a 1000: in questo modo si campionano in modo casuale 1000 osservazioni tra quelle a disposizione e si ottengono 1000 stime dei coefficienti del modello. Per il confronto sono stati utilizzati i dataset completi ottenuti tramite le tecniche di imputazione: per ogni area geografica è stato utilizzato il dataset generato con la tecnica d'imputazione della regressione lineare semplice. Per prima cosa è necessario verificare l'ipotesi di normalità dei coefficien-

ti stimati tramite il metodo *bootstrap*. Per far ciò è stato utilizzato uno Shapiro test multivariato che ha restituito esito positivo per tutte le stime realizzate. I dataset completi utilizzati per il confronto sono stati imputati tramite la tecnica di regressione lineare. I risultati dei confronti a due a due delle medie delle 1000 stime dei coefficienti del modello restituscono dei *p-value* pari a 2.2×10^{-16}, che consentono di rifiutare l'ipotesi nulla di uguaglianza del vettore delle medie nelle diverse aree geografiche. Il test appena realizzato è supportato dal test Manova, il quale confronta i vettori delle medie nei diversi dataset. L'esito del test conferma quanto emerso dal precedente test di Hotelling. Il test Manova inoltre permette di confrontare ogni componente del vettore delle medie, in modo da evidenziare le stime delle variabili che hanno un comportamento diverso nelle tre diverse aree geografiche. Dal test si evince che tutte le stime dei coefficienti sono diverse nel Nord, Centro e Sud e ciò giustifica nuovamente la scelta di divisione del dataset nelle tre aree.

7.3 Modello a livello studente con effetto casuale scuola per aree geografiche

Dal test del capitolo precedente emerge come la differenza tra le diverse aree geografiche d'Italia sia molto marcata. Considerando separatamente le aree si può analizzare quali aspetti influenzino in positivo o in negativo le valutazioni degli studenti e vedere se l'effetto scuola sia veramente diverso a seconda della zona, dimostrando che vi sono differenze nel sistema educativo Italiano.
Una prima idea della distribuzione del voto della prova di matematica nelle diverse aree geografiche è visibile nei boxplot in Figura 7.5.
Analizzando i boxplot emerge che non vi è una significativa differenza tra il voto della prova nel Nord e nel Centro Italia, mentre nel Sud la distribuzione è visibilmente inferiore con una varianza molto più alta. Effettuando un test anova sulla variabile *MATH_corr* si ottiene un *p-value* inferiore a 2×10^{-16}, che indica una netta diversità del valor medio del voto nelle tre aree d'Italia.

Il modello realizzato per ogni area geografica è il medesimo presentato

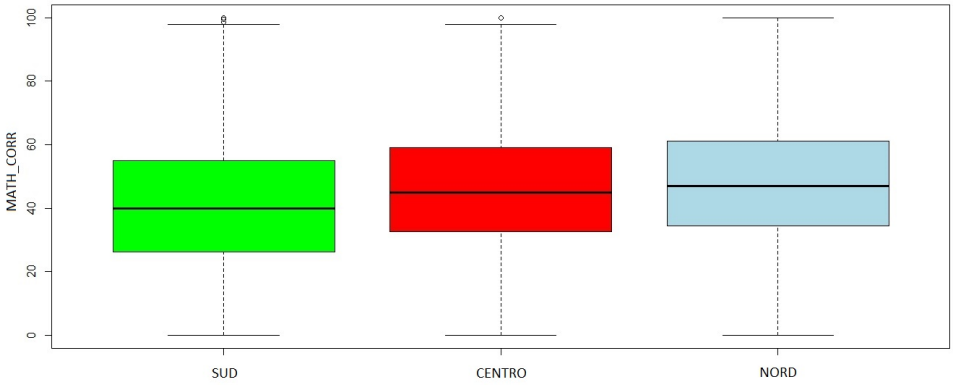

Figura 7.5: Boxplot della variabile $MATH\_corr$ al variare dell'area geografica

in precedenza

$$y_{ij}^{(R)} = \beta_0^{(R)} + \sum_{k=1}^{K} \beta_k^{(R)} x_{kij}^{(R)} + b_j^{(R)} + \epsilon_{ij}^{(R)} \tag{7.9}$$

dove l'apice R={Nord, Centro, Sud} serve per differenziare le aree geografiche.

La numerosità nei tre dataset è differente: per quanto riguarda il Nord e il Sud si hanno a disposizione rispettivamente 220299 e 198854 oseervazioni, per il Centro invece il numero di dati disponibili è molto inferiore e pari a 90218. Anche le percentuali di dati mancanti nei tre dataset sono ben diverse: nel Nord Italia le variabili *ESCS*, *IMM1* e *IMM2* contengono circa l'11% di valori NA, nel Centro circa lo 0.35%, mentre nel Sud lo 0.65%; la covariata che contiene più valori mancanti è *MATH_ 5* con una percentuale di 38% nel Nord, 46% nel Centro e 56% nel Sud. Queste percentuali evidenziano come la mancanza delle informazioni è in qualche modo connessa con l'area geografica d'Italia.

I dati mancanti nei tre dataset Nord, Centro e Sud sono stati trattati nello stesso modo che nel dataset completo: inizialmente sono stati realizzati i modelli dopo l'applicazione della *listwise deletion* e successivamente sono stati imputati i dati nei tre dataset con le diverse tecniche di imputazione. Le stime dei parametri del modello 7.1 ottenute dopo l'uso della *listwise deletion* sono presenti nella prima colonna delle Tabelle 7.3 - 7.5. Le altre colonne delle tabelle contengono le stime degli stessi parametri del mo-

dello prodotte dopo aver imputato i dati mancanti con le diverse tecniche descritte nel Capitolo 4.

La prima osservazione che è opportuno fare riguarda il coefficiente *VPC*: nelle tre aree esso assume valori molto diversi tra loro, evidenziando che il contributo della scuola varia a seconda della zona geografica. La variabilità tra le scuole è molto più alta nel Sud che al Nord, dove il *VPC* assume valori rispettivamente di circa il 20% e circa il 6%. Nel Nord Italia frequentare una scuola rispetto ad un'altra sembra avere poca influenza sul voto della prova. La situazione è pressochè analoga nel caso del Centro, dove però l'indice *VPC* assume valori intorno al 10%. Per quanto riguarda il Sud Italia la situazione è differente: con un indice di *VPC* pari al 20% l'effetto scuola è molto significativo. Ciò dimostra che l'insegnamento nel meridione non è omogeneo e dunque si possono riscontrare grandi differenze nel rendimento degli studenti di scuole diverse.
Analizzando invece le stime dei coefficienti del modello è importante notare come *ESCS* influenzi positivamente l'esito della prova in tutte e tre le aree ma con maggior peso nel Sud Italia, suggerendo che la situazione socio-economica e culturale della famiglia sia molto rilevante soprattutto nel mezzogiorno.
I coefficienti associati alle variabili *IMM1* e *IMM2*, che identificano rispettivamente lo studente immigrato o figlio di immigrati, hanno peso diverso nei tre modelli: le variabili pesano di più in modo negativo per le scuole del Nord rispetto al resto d'Italia e ciò può essere dovuto al fatto che nel Centro e nel Sud vi sono meno immigrati.
Un'ulteriore considerazione da fare riguarda la stima del coefficiente di *MATH_5*, che indica il voto della prova INVALSI di quinta elementare, il cui valore decresce dal Nord al Sud. Questo risultato evidenzia una maggiore continuità del rendimento scolastico nel Nord Italia.
I restanti coefficienti degli effetti fissi hanno un valore simile nelle tre aree.
Nei grafici in Figura 7.6 e 7.7 sono rappresentate le distribuzioni dei valori osservati, in blu, e le distribuzioni dei valori imputati, in rosso. Si può notare molto chiaramente come il metodo *mean*, che sostituisce il valore mancante con la media della variabile, modifica enormemente la distribuzione dei valori imputati del voto della prova di quinta spostandola verso il valor medio. La distribuzione dei valori imputati con il metodo *PMM* invece mantiene pressochè identitica la distribuzione della variabile; questa è una caratteristica del metodo poiché esso imputa i valori molto realisticamente

Tabella 7.3: Stime del modello (7.4) livello studente con effetto aleatorio scuola per il Nord dopo LD e imputati con i metodi *EMB*, *Mean*, *Sample*, *Reg*, *Reg_Bay* e *PMM*. Gli asterischi individuano i diversi livelli di significatività: . o.01 < *p-value* < 0.1; * 0.001 < *p-value* < 0.01; ** 0.0001 < *p-value* < 0.001; *** *p-value* < 0.0001.

Stime Effetti Fissi (NORD)							
	LD	*EMB*	*Sample*	*Mean*	*Reg*	*Reg Bay*	*PMM*
Intercetta	1.389 ***	2.129 ***	22.452 ***	3.721 ***	2.252 ***	2.309 ***	3.329 ***
FEMMINA	-1.6996 ***	-1.580 ***	-2.475 ***	-2.103 ***	-1.568 ***	-1.610 ***	-1.646 ***
IMM1	-0.911 ***	-1.086 ***	-3.281 ***	-3.134 ***	-1.113 ***	-1.137 ***	-1.135 ***
IMM2	-1.978 ***	-1.802 ***	-3.375 ***	-2.987 ***	-1.752 ***	-1.845 ***	-1.809 ***
ANTICIPATARIO	-1.913 ***	-1.588 ***	-2.983 ***	-2.358 ***	-1.599 ***	-1.765 ***	-1.844 ***
POSTICIPATARIO	-2.401 ***	-4.591 ***	-9.493 ***	-8.889 ***	-4.623 ***	-4.608 ***	-4.462 ***
ESCS	2.460 ***	2.002 ***	2.907 ***	2.705 ***	2.015 ***	2.016 ***	2.054 ***
No_genitori	-1.229 ***	-1.320 ***	-2.105 ***	-1.950 ***	-1.347 ***	-1.326 ***	-1.319 ***
Si_fratelli	0.166	0.085	-0.132	-0.088	-0.062	0.056	0.101
MATH_5	0.697 ***	0.680 ***	0.397 ***	0.661 ***	0.679 ***	0.679 ***	0.664 ***
Effetti Casuali							
σ_b	3.631	3.202	3.958	3.961	3.228	3.222	3.213
σ_ϵ	12.425	12.627	15.118	14.214	12.618	12.615	12.697
VPC	7.87%	6.04%	6.41%	7.08%	6.14%	6.12%	6.01%

Tabella 7.4: Stime del modello (7.4) livello studente con effetto aleatorio scuola per il Centro dopo LD e imputati con i metodi *EMB*, *Mean*, *Sample*, *Reg*, *Reg_Bay* e *PMM*. Gli asterischi individuano i diversi livelli di significatività: . $0.01 <$ *p-value* < 0.1; * $0.001 <$ *p-value* < 0.01; ** $0.0001 <$ *p-value* < 0.001; *** *p-value* < 0.0001.

Stime Effetti Fissi (CENTRO)							
	LD	*EMB*	*Sample*	*Mean*	*Reg*	*Reg Bay*	*PMM*
Intercetta	8.021 ***	9.112 ***	28.071 ***	9.606 ***	9.078 ***	9.248 ***	10.399 ***
FEMMINA	-2.657 ***	-2.605 ***	-3.299 ***	-3.026 ***	-2.563 ***	-2.602 ***	-2.644 ***
IMM1	-0.738 ***	-0.729 ***	-2.293 ***	-1.940 ***	-0.769 ***	-0.671 ***	-0.760 **
IMM2	-1.127 ***	-0.434 ***	-1.321 ***	-0.887 ***	-0.353	-0.373	-0.481
ANTICIPATARIO	-0.284	-0.602	-1.224 ***	-1.060 *	-0.642	-0.652	-0.588
POSTICIPATARIO	-1.692 ***	-5.092 ***	-7.898 ***	-7.827 ***	-5.022 ***	-5.180 ***	-5.260 ***
ESCS	2.460 ***	2.522 ***	3.498 ***	3.187 ***	2.015 ***	2.554 ***	2.573 ***
No_genitori	-1.338 ***	-1.072 ***	-1.731 ***	-1.565 ***	-1.083 ***	-1.040 ***	-1.109 ***
Si_fratelli	0.031	-0.095	-0.144	-0.146	-0.109	-0.091	-0.092
MATH_5	0.571 ***	0.548 ***	0.279 ***	0.539 ***	0.548 ***	0.546 ***	0.529 ***
Effetti Casuali							
σ_b	4.521	4.079	4.875	4.960	4.161	4.135	4.215
σ_ϵ	13.526	13.769	15.663	13.814	13.813	13.828	13.906
VPC	10.05%	8.06%	8.83%	9.83%	8.32%	8.20%	8.41%

Tabella 7.5: Stime del modello (7.4) livello studente con effetto aleatorio scuola per il Sud dopo LD e imputati con i metodi *EMB*, *Mean*, *Sample*, *Reg*, *Reg_Bay* e *PMM*. Gli asterischi individuano i diversi livelli di significatività: . 0.01 < *p-value* < 0.1; * 0.001 < *p-value* < 0.01; ** 0.0001 < *p-value* < 0.001; *** *p-value* < 0.0001.

Stime Effetti Fissi (SUD)							
	LD	*EMB*	*Sample*	*Mean*	*Reg*	*Reg Bay*	*PMM*
Intercetta	16.835 ***	17.029 ***	31.501 ***	16.686 ***	16.904 ***	16.900 ***	18.325 ***
FEMMINA	-2.140 ***	-1.766 ***	-1.978 ***	-1.904 ***	-1.777 ***	-1.762 ***	-1.802 ***
IMM1	0.392	0.858	-2.293	0.513	0.880 **	1.095 *	0.647
IMM2	-0.658	-0.197	-1.321	-0.348	-0.263	-0.250	-0.362
ANTICIPATARIO	0.066	0.079	0.334	0.232	0.056	0.058	0.016
POSTICIPATARIO	-1.692 ***	-6.797 ***	-8.157 ***	-8.117 ***	-6.722 ***	-6.811 ***	-6.807 ***
ESCS	-3.885 ***	3.336 ***	3.773 ***	3.822 ***	3.339 ***	3.338 ***	3.414 ***
No_genitori	-3.179 ***	-1.408 ***	-1.927 ***	-1.829 ***	-1.340 ***	-1.359 ***	-1.440 ***
Si_fratelli	0.015	-0.031	-0.072	-0.036	-0.049	0.005	-0.033
MATH_5	0.387 ***	0.354 ***	0.144 ***	0.354 ***	0.356 ***	0.355 ***	0.335 ***
Effetti Casuali							
σ_b	7.357	8.273	9.204	9.153	8.247	8.257	8.276
σ_ϵ	16.301	15.332	15.663	15.993	15.319	15.311	15.371
VPC	20.20%	21.48%	24.17%	24.67%	22.47%	22.53%	22.47%

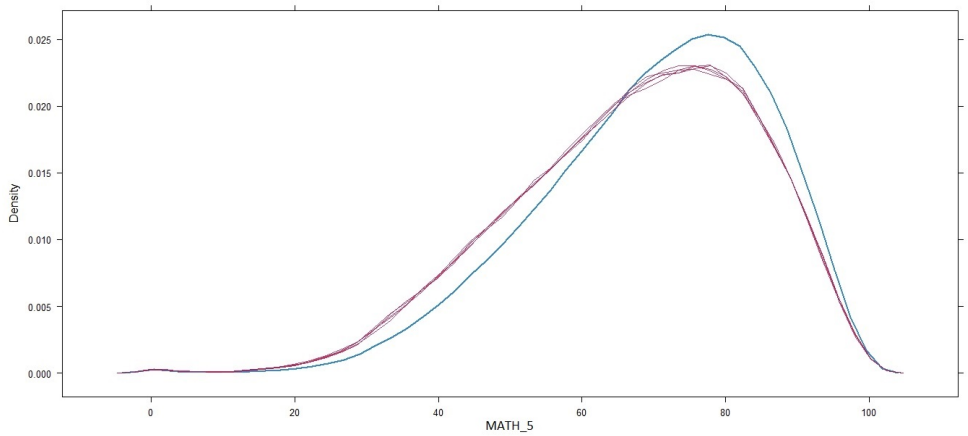

Figura 7.6: Distribuzione della variabile *MATH_ 5* per il dataset Nord: in blu la distribuzione dei valori osservati, in rosso le distribuzioni dei valori imputati con il metodo *PMM*.

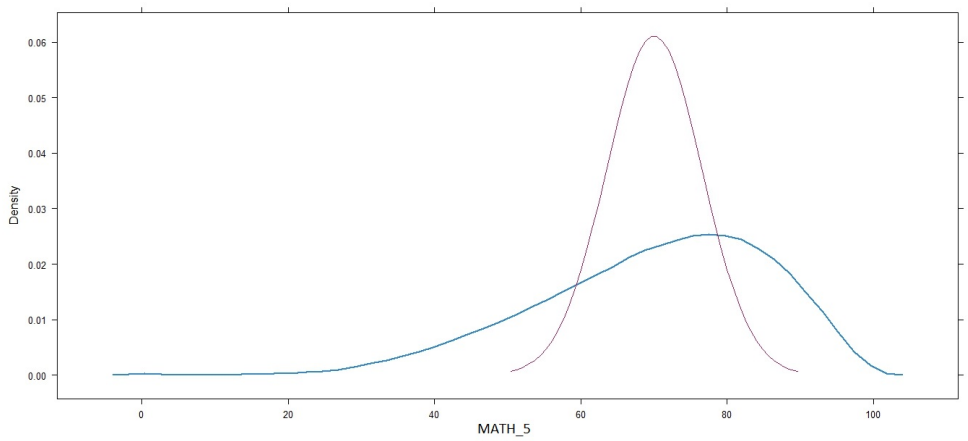

Figura 7.7: Distribuzione della variabile *MATH_ 5* per il dataset Nord: in blu la distribuzione dei valori osservati, in rosso le distribuzioni dei valori imputati con il metodo *mean*.

mantenendo invariata la distribuzione della variabile.

7.4 Test sui coefficienti del modello nelle aree

I risultati descritti nel capitolo precedente permettono di confrontare le stime dei parametri di uno stesso modello applicato a dataset diversi. In questo modo è possibile evidenziare dipendenze diverse delle stesse covariate sulla variabile risposta al variare del dataset. Attraverso il test descritto nel Capitolo 7.2 risulta che le stime dei parametri del modello di regressione

(7.1) sono diverse nei tre dataset. Il confronto tra i vettori dei parametri di stima è stato realizzato sui dataset completi i cui valori mancanti sono stati imputati tramite il metodo di regressione lineare, cioé ogni valore mancante è generato tramite una regressione lineare sulle altre covariate del dataset.

Un altro modo per poter sfruttare questo test è confrontare le stime dei parametri del modello ottenute al variare del metodo di imputazione utilizzato sullo stesso dataset. L'esito di questi confronti permette di verificare se le stime cambiano notevolmente al variare del metodo di imputazione scelto.
Per ogni area geografica è stato realizzato un test Manova per confrontare insieme i vettori delle stime ottenuti con i diversi metodi: i *p-value* pari a 2.2×10^{-16} in tutte e tre le aree permettono di affermare che le stime non sono identiche.
In seguito a questo primo test sono stati realizzati numerosi test di Hotelling per confrontare a due a due i vettori delle stime dei parametri del modello. I risultati dei test sono analoghi per le tre aree: i vettori delle stime dei parametri ottenuti sui dataset imputati con il metodo EM con *bootstrap*, con la regressione lineare e con la regressione lineare Bayesiana sono risultati essere identici. Questo risultato garantisce che i tre metodi di imputazione in considerazione generano valori imputati molto simili. Applicando il test ai metodi che imputano i dati tramite un campionamento casuale, *sample*, e tramite il valore medio, *mean*, risulta che le stime ottenute non sono statisticamente uguali: questo risultato è dovuto alla semplicità dei due metodi che non permette di generare valori imputati realistici.

Capitolo 8

Metodi di imputazione per dati raggruppati

Nei capitoli precedenti sono stati analizzati i risultati del modello (7.1) in seguito all'utilizzo di tecniche diverse di imputazione per i dati mancanti. La generazione dei valori che sostituiscono gli NA del dataset non ha tenuto in considerazione l'appartenenza degli studenti alle diverse scuole. La suddivisione in scuole differenti è stata infatti introdotta nel modello solo successivamente tramite l'utilizzo dell'effetto casuale b_j.
L'interesse ora è quindi valutare come le stime dei parametri del modello di regressione a livello studente cambiano se i numerosi valori imputati del dataset sono generati considerando la divisione degli alunni nelle scuole.
Il modo per imputare i dati mancanti con l'informazione della scuola di appartenenza dello studente è semplicemente aggiungere la variabile *CODICE_SCUOLA* nel dataset. In questo modo si sfrutta il metodo Gibbs sampler della funzione **mice** per imputare i dati condizionati alla scuola e il modello multilivello lineare successivamente applicato cambia la varianza dell'errore a seconda della scuola frequentata [18]. Il Gibbs sampler è un algoritmo MCMC (Markov Chain Monte Carlo) che permette di ottenere una sequenza di campioni casuali da una distribuzione di probabilità multivariata. Ad ogni iterazione l'algoritmo genera per ogni variabile un campione dalla distribuzione condizionata sui valori delle restanti variabili. La sequenza di campioni costituisce una catena di Markov e la distribuzione stazionaria di tale catena è la distribuzione congiunta cercata.
Il metodo si applica in modo semplice alle tecniche di imputazione *sample* e *mean* e le stime del modello di regressione non si discostano molto

dalle stime ottenute senza considerare il raggruppamento per scuola. Per quanto riguarda i metodi di regressione lineare classica e bayesiana e l'algoritmo *predictive mean matching* subentrano molte difficoltà nella fase di imputazione. Il maggior costo computazionale di questi metodi, l'elevato numero di scuole (5311) e di variabili da imputare non rendono possibile la realizzazione delle imputazioni condizionate alla variabile che identifica le scuole.
Per imputare i valori NA condizionatamente al codice scuola sono stati quindi utilizzati altri software adatti all'imputazione a più livelli: MLwiN [26], REALCOM Impute [6, 11] e WinMICE [14]. Analogamente a R anche questi programmi restituiscono una serie di errori di allocazione memoria poiché i costi computazionali dei metodi di imputazione utilizzati sono molto elevati.
Analizzando successivamente in dettaglio i dati per giustificare perché i metodi non riescono a realizzare le imputazioni condizionate alla scuola frequentata si sono riscontrate altre problematiche: per molte scuole infatti non è presente nessuna osservazione per la variabile *MATH_5*; per molte altre invece vi sono solo poche osservazioni. Queste condizioni non rendono possibile agli algoritmi *Reg*, *Reg_Bay* e *PMM* di imputare i dati mancanti. Una possibile soluzione per evitare questi problemi è eliminare le scuole che rientrano in queste tipologie e poi eseguire gli algoritmi sulle scuole rimanenti.
Ricordando il significato della variabile *MATH_5*, che definisce il punteggio della prova nella classe quinta elementare, si può chiaramente intuire che non occorre imputare la variabile in funzione delle scuole in tutti i casi: infatti il voto della prova è associato alla scuola elementare e non alla scuola media di cui si dispone il codice scuola. L'imputazione di *MATH_5* condizionata alla scuola è ragionevole solo nel caso in cui le scuole sono degli istituti comprensivi (*IC*), cioé le scuole elementari e medie sono ubicate nella stessa struttura e dunque il codice scuola è il medesimo. Poiché la variabile che definisce il punteggio della prova INVALSI della quinta elementare contiene il 46.48% di NA e poiché l'interesse è studiare la relazione tra questa variabile e il voto della prova nelle medie *MATH_corr* è opportuno focalizzare l'attenzione sull'imputazione di questa variabile.
Per questi diversi motivi sono state tralasciate le scuole non IC per le quali il voto della prova *MATH_5* (da imputare) non è condizionato al codice scuola e da qui in avanti le analisi proseguono su un dataset ridotto di

2883 istituiti comprensivi con 230312 studenti. L'interesse è studiare se le stime del modello (7.1) variano condizionando le imputazioni alla divisione in gruppi del dataset.

8.1 Partioned Predictive Mean Matching

Nelle analisi per dati raggruppati l'effetto dovuto al raggruppamento è assunto come effetto casuale del modello. La scelta di ignorare però gli effetti dei gruppi durante la fase di imputazione potrebbe portare a sottostimare la correlazione tra i gruppi nel dataset completo. Idealmente i parametri del modello all'interno dei gruppi possono variare in modo differente durante l'imputazione. In alcuni casi la variabile che identifica i gruppi è inserita come effetto fisso: questa scelta è ottimale soltanto quando il numero di gruppi è piccolo e quindi è possibile utilizzare alcune variabili binarie per definire l'appartenenza ai gruppi. Nel caso di un elevato numero di gruppi, come nel dataset INVALSI, questa soluzione non è applicabile.
Una soluzione a questa problematica è sfruttare un'estensione dell'algoritmo *PMM* descritto nel Capitolo 4.1 chiamato *Partitioned Predictive Mean Matching* [35]. L'algoritmo *PMM* applicato al dataset che contiene il codice scuola può essere utilizzato per generare le imputazioni tenendo conto della struttura multilivello: infatti il pacchetto **mice** di R sfrutta il metodo Fully Conditionally Specification in cui la distribuzione di ogni variabile condizionata alle altre è usata come punto iniziale del metodo Gibbs Sampler.
L'idea alla base dell'algoritmo è suddividere le scuole in gruppi in modo da applicare l'algoritmo *PMM* a dataset più piccoli. L'algoritmo consiste in:

- Partizionare i dati in P parti più piccole di dimensione approssimativamente uguale.
- Eseguire l'algoritmo *PMM* per ogni parte di dataset.
- Combinare le P parti di dataset imputati.

Per la fase di stima è necessario che le scuole siano interamente contenute in una sola delle P parti e non divise tra le diverse parti.
Prima di applicare il metodo è stato necessario dividere il nuovo dataset con solo gli istituti comprensivi in gruppi di scuole. La divisione in gruppi

di 5 scuole però non rende nuovamente possibile l'utilizzo dell'algoritmo: durante la fase di imputazione si riscontrano problemi sul condizionamento delle matrici necessarie all'imputazione. Anche riducendo ulteriormente il numero di scuole per gruppo i problemi non si risolvono. L'unica possibile soluzione per imputare i *missing data* degli studenti condizionatamente alle scuole frequentate è imputare i valori di una scuola alla volta: in pratica si utilizza l'algoritmo *Partitioned Predictive Mean Matching* per gruppi formati da una scuola soltanto.
L'idea di questo metodo è stata utilizzata anche per le altre tecniche di imputazione. Le scuole sono state partizionate in gruppi e ogni gruppo è stato imputato separatamente: anche in questi casi però si presentano problemi nella fase di imputazione. Per questo motivo, in modo analogo al metodo *PPMM*, l' imputazione è stata effettuata una scuola alla volta anche per i metodi *Mean*, *Sample*, *Reg*, *Reg_Bay* ed *EMB*. Nonostante questa semplificazione nella fase di imputazione, la tecnica che utilizza la regressione lineare per imputare i dati mancanti non converge e dunque non sono presenti i risultati delle stime prodotte da questo metodo per l'imputazione condizionata al codice scuola.

8.2 Il dataset IC: modello e imputazioni

Il dataset considerato per questa fase di analisi è quello costituito dalle scuole che sono Istituti Comprensivi (IC). Come descritto nella sezione precedente, l'algoritmo *PPMM* non è in grado di imputare per più gruppi di scuole e dunque è stato applicato ad una scuola alla volta. Questo utilizzo però non permette l'imputazione dei valori mancanti per quelle scuole che hanno poche (o nessuna) osservazioni. Per questo motivo è necessario escludere queste ultime dal dataset per ottenere il dataset completo su cui applicare il modello lineare a effetti misti (7.1). Il dataset di cui sono stati imputati i dati quindi è definito da 222638 osservazioni raggruppate in 2767 scuole.

L'interesse di questa parte di analisi, come già anticipato, è valutare le stime dei coefficienti del modello (7.1) in seguito all'applicazione delle tecniche di imputazione al dataset contenente il codice identificativo della scuola e al dataset senza di esso: se le stime dei parametri del modello sono differenti allora le imputazioni condizionate al codice scuola e quelle otte-

nute senza il condizionamento al raggruppamento in scuole restituiscono dei valori imputati diversi.

Come si può notare dai risultati in Tabella 8.1 e 8.2 l'interpretazione delle stime ottenute dal modello è analoga a quella presentata nel Capitolo 7.1, dove le variabili che caratterizzano la zona geografica della scuola risultano essere significative, così come la variabile che definisce il sesso dello studente, il grado di immigrazione e quelle che definiscono se lo studente ha anticipato o posticipato l'inizio della scuola.
Le stime dei parametri del modello ottenute applicando lo stesso modello lineare a effetti misti ai due dataset con solo gli istituti comprensivi, imputati condizionando i valori al codice scuola e imputati senza di esso, sembrano significativamente diverse (vedi Tabella 8.1 e 8.2). Per verificare se le stime ottenute sono statisticamente differenti si utilizza il test descritto nel Capitolo 7.2. Utilizzando una tecnica di ricampionamento si generano numerose stime dei parametri in modo da poter realizzare il test per il confronto dei vettori delle medie.
Il test di Hotelling dunque è stato utilizzato per confrontare se lo stesso metodo di imputazione applicato ai due dataset restituisce diverse stime dei parametri. Il test restituisce *p-value* molto inferiori al 5% per tutti i metodi utilizzati ad eccezione del metodo che imputa i dati mancanti con la media: questi risultati permettono di affermare che i vettori delle stime dei parametri del modello (7.1) ottenuti dopo l'imputazione dei dati mancanti del dataset condizionato alla scuola e quello non condizionato sono differenti tra loro e dunque anche i valori imputati differiscono a parità di metodo di imputazione. Queste considerazioni supportano l'argomentazione che l'imputazione dei valori mancanti condizionatamente al codice scuola sia una scelta più adeguata: il motivo risiede nel fatto che, come è risultato dalle analisi dei capitoli precedenti, frequentare una scuola piuttosto che un'altra può comportare un notevole cambiamento nella valutazione dello studente.
Lo stesso test è stato utilizzato in modo analogo al Capitolo 7.4 per confrontare i vettori delle stime dei parametri del modello ottenute al variare del metodo di imputazione utilizzato sullo stesso dataset.
I confronti a due a due delle stime dei parametri restituiscono tutti il medesimo esito: i vettori delle stime dei parametri ottenuti al variare del metodo di imputazione sono tutti diversi tra loro sia per i dati condizionati al codi-

Tabella 8.1: Stime del modello (7.1) livello studente con effetto aleatorio scuola per il dataset IC imputati con i metodi *EMB*, *Mean*, *Sample*, *Reg*, *Reg_Bay* e *PMM* condizionatamente al *CODICE_SCUOLA*. Gli asterischi individuano i diversi livelli di significatività: . 0.01 $<$ *p-value* $<$ 0.1; * 0.001 $<$ *p-value* $<$ 0.01; ** 0.0001 $<$ *p-value* $<$ 0.001; *** *p-value* $<$ 0.0001.

Stime Effetti Fissi						
	EMB	*Sample*	*Mean*	*Reg*	*Reg Bay*	*PMM*
Intercetta	8.757 ***	21.285 ***	9.590 ***	-	47.522 ***	10.957 ***
FEMMINA	-1.838 ***	-2.256 ***	-2.030 ***	-	-2.772 ***	-1.933 ***
IMM1	-1.122 ***	-2.758 ***	-2.230 ***	-	-3.718 ***	-1.426 ***
IMM2	-1.699 ***	-2.749 ***	-2.222 ***	-	-3.796 ***	-1.963 ***
CENTRO	-2.615 ***	-2.652 ***	-2.658 ***	-	-2.612 ***	-2.753 ***
SUD	-7.268 ***	-7.865 ***	-7.833 ***	-	-7.857 ***	-7.675 ***
ANTICIPATARIO	-1.108 ***	-1.497 ***	-1.336 ***	-	-1.962 ***	-1.282 ***
POSTICIPATARIO	-4.999 ***	-8.045 ***	-8.138 ***	-	-7.692 ***	-5.806 ***
ESCS	2.383 ***	3.255 ***	2.877 ***	-	4.149 ***	2.564 ***
No_genitori	-1.279 ***	-1.740 ***	-1.597 ***	-	-2.252 ***	-1.437 ***
Si_fratelli	0.136	0.025	-0.004	-	0.100	0.108
MATH_5	0.587 ***	0.411 ***	0.576***	-	0.039 ***	0.557 ***
Effetti Casuali						
σ_b	5.185	5.807	5.929	-	6.052	6.126
σ_ϵ	13.410	14.816	14.239	-	15.940	13.777
VPC	13.00%	13.31%	14.77%	-	12.59%	16.51%

Tabella 8.2: Stime del modello (7.1) livello studente con effetto aleatorio scuola per il dataset IC imputati con i metodi *EMB*, *Mean*, *Sample*, *Reg*, *Reg_Bay* e *PMM* senza il condizionamento al *CODICE_SCUOLA*. Gli asterischi individuano i diversi livelli di significatività: . $0.01 < p\text{-}value < 0.1$; * $0.001 < p\text{-}value < 0.01$; ** $0.0001 < p\text{-}value < 0.001$; *** $p\text{-}value < 0.0001$.

Stime Effetti Fissi						
	EMB	*Sample*	*Mean*	*Reg*	*Reg Bay*	*PMM*
Intercetta	8.803 ***	21.237 ***	9.590 ***	8.861 ***	8.839 ***	10.015 ***
FEMMINA	-1.796 ***	-2.257 ***	-2.030 ***	-1.844 ***	-1.834 ***	-1.877 ***
IMM1	-1.215 ***	-2.781 ***	-2.230 ***	-1.182 ***	-1.125 ***	-1.214 ***
IMM2	-1.695 ***	-2.742 ***	-2.222 ***	-1.719 ***	-1.686 ***	-1.706 ***
CENTRO	-2.607 ***	-2.642 ***	-2.658 ***	-2.610 ***	-2.606 ***	-2.611 ***
SUD	-7.291 ***	-7.872 ***	-7.833 ***	-7.308 ***	-7.289 ***	-7.269 ***
ANTICIPATARIO	-1.084 ***	-1.527 ***	-1.336 ***	-1.148 ***	-1.127 ***	-1.177 ***
POSTICIPATARIO	-4.942 ***	-8.040 ***	-8.138 ***	-4.982 ***	-5.111 ***	-4.995 ***
ESCS	2.398 ***	3.251 ***	2.877 ***	2.375 ***	2.383 ***	2.428 ***
No_genitori	-1.302 ***	-1.783 ***	-1.597 ***	-1.289 ***	-1.269 ***	-1.277 ***
Si_fratelli	0.046	0.019	-0.004	0.042	0.034	0.048
MATH_5	0.587 ***	0.411 ***	0.576***	0.587 ***	0.587 ***	0.570 ***
Effetti Casuali						
σ_b		5.809	5.929	5.177	5.184	5.129
σ_ϵ		14.809	14.239	13.423	13.420	13.491
VPC		13.33%	14.77%	12.95%	12.98%	12.63%

ce scuola sia per quelli non condizionati. Questi risultati indicano che per questi due dataset metodi diversi di imputazione generano stime diverse dei coefficienti del modello. Per quanto riguarda il confronto tra le stime prodotte dai metodi *mean* e *sample* e quelle ottenute con gli altri metodi il risultato che i vettori delle stime dei parametri non siano uguali non è sorprendente: a causa della semplicità di questi due metodi non è garantito infatti che i valori generati siano realistici. Dunque poiché i risultati sono diversi a seconda del metodo di imputazione scelto saranno considerate più appropriate le stime associate al metodo *PPMM* poiché esso genera valori imputati più realistici degli altri metodi ed inoltre è il metodo più adatto all'imputazione di dati multilivello [35].

8.2.1 Test coefficienti del modello tra le aree

Nella sezione precedente è emerso che ancora una volta le valutazioni della prova INVALSI degli studenti è molto influenzata dalla zona geografica dove lo studente frequenta l'Istituto Comprensivo. Per questa ragione è stato diviso nuovamente il dataset nelle tre aree geografiche d'Italia in modo da poter verificare che le stime dei parametri del modello siano differenti per le tre aree dopo aver imputato i dati condizionando le variabili al codice scuola. Di seguito sono riportati i valori delle stime dei coefficienti del modello (7.4) realizzato per le tre aree.

I risultati sono molto differenti soprattuto per quanto riguarda le variabili che definiscono il grado di immigrazione dello studente e quella associata al voto della prova INVALSI della classe quinta elementare. Le stime dei coefficienti delle variabili *IMM1* e *IMM2* decrescono muovendosi dal Nord al Sud d'Italia: come già evidenziato questo comportamento è dovuto al fatto che la percentuale di studenti immigrati è più alta nel settentrione rispetto che nel meridione e dunque l'influenza è maggiore nel Nord. Un andamento crescente è invece riscontrato nel coefficiente VPC che è enormemente più alto nel Sud. Questo valore molto alto nel Sud evidenzia come la varianza associata all'effetto aleatorio scuola contribuisce in percentuale maggiore alla definizione della varianza totale, evidenziando differenti valutazioni al variare della scuola in quella zona geografica.

Sfruttando il test descritto nel Capitolo 7.2 è possibile confrontare le stime del modello al variare dell'area geografica. Confrontando a parità di

Tabella 8.3: Stime del modello (7.4) livello studente per le scuole nel Nord imputate condizionatamente al codice scuola con i metodi *EMB*, *Mean*, *Sample*, *Reg*, *Reg_Bay* e *PMM*. Gli asterischi individuano i diversi livelli di significatività: . $0.01 < p\text{-}value < 0.1$; * $0.001 < p\text{-}value < 0.01$; ** $0.0001 < p\text{-}value < 0.001$; *** $p\text{-}value < 0.0001$.

Stime Effetti Fissi (NORD)						
	EMB	*Sample*	*Mean*	*Reg*	*Reg Bay*	*PMM*
Intercetta	1.371 ***	14.007 ***	2.712 ***	-	22.290 ***	2.421 ***
FEMMINA	-1.545 ***	-2.072 ***	-1.792 ***	-	-2.246 ***	-1.657 ***
IMM1	-1.270 ***	-3.227 ***	-2.585 ***	-	-2.720 ***	-1.390 ***
IMM2	-1.813***	-3.205 ***	-2.566 ***	-	-2.911***	-2.097 ***
ANTICIPATARIO	-1.368 ***	-2.648***	-2.254 ***	-	-2.944 ***	-1.840 ***
POSTICIPATARIO	-4.164 ***	-8.577 ***	-8.575 ***	-	-6.142 ***	-5.399 ***
ESCS	1.911 ***	2.842 ***	2.453 ***	-	2.959***	2.078 ***
No_genitori	-1.229 ***	-1.755 ***	-1.583 ***	-	-1.692 ***	-1.395 ***
Si_fratelli	0.177	-0.022	-0.001	-	0.199	0.192
MATH_5	0.693 ***	0.516 ***	0.676 ***	-	0.398***	0.664 ***
Effetti Casuali						
σ_b	3.359	3.614	3.691	-	3.592	3.819
σ_ϵ	12.418	14.100	13.387	-	14.093	12.759
VPC	6.81%	6.16%	7.06%	-	6.10%	8.22%

Tabella 8.4: Stime del modello (7.4) livello studente per le scuole del Centro imputate condizionatamente al codice scuola con i metodi *EMB*, *Mean*, *Sample*, *Reg*, *Reg_Bay* e *PMM*. Gli asterischi individuano i diversi livelli di significatività: . $0.01 < p\text{-}value < 0.1$; * $0.001 < p\text{-}value < 0.01$; ** $0.0001 < p\text{-}value < 0.001$; *** $p\text{-}value < 0.0001$.

Stime Effetti Fissi (CENTRO)						
	EMB	*Sample*	*Mean*	*Reg*	*Reg Bay*	*PMM*
Intercetta	7.666 ***	20.167 ***	8.331 ***	-	39.173 ***	10.086 ***
FEMMINA	-2.439 ***	-3.032 ***	-2.753 ***	-	-3.327 ***	-2.490 ***
IMM1	-0.824 ***	-1.906 ***	-1.559 ***	-	-2.334 ***	-1.047 **
IMM2	-1.090 ***	-1.727 ***	-1.365 ***	-	-2.082***	-1.348
ANTICIPATARIO	-1.420	-1.564 ***	-1.586 *	-	-1.896 ***	-1.640
POSTICIPATARIO	-4.910 ***	-7.565 ***	-7.548 ***	-	-6.757 ***	-5.590 ***
ESCS	2.416 ***	3.269***	2.906 ***	-	3.964 ***	2.595 ***
No_genitori	-1.175 ***	-1.674***	-1.467 ***	-	-1.967 ***	-1.321 ***
Si_fratelli	0.218	-0.027	-0.146	-	-0.113	-0.020
MATH_5	0.566 ***	0.391 ***	0.557 ***	-	0.121 ***	0.534 ***
Effetti Casuali						
σ_b	4.187	4.437	4.554	-	4.540	4.869
σ_ϵ	13.653	14.938	14.435	-	15.673	13.997
VPC	8.59%	8.10%	9.05%	-	7.74%	10.79%

Tabella 8.5: Stime del modello (7.4) livello studente per le scuole del Sud imputate condizionatamente al codice scuola con i metodi *EMB*, *Mean*, *Sample*, *Reg*, *Reg_Bay* e *PMM*. Gli asterischi individuano i diversi livelli di significatività: . $0.01 < p\text{-}value < 0.1$; * $0.001 < p\text{-}value < 0.01$; ** $0.0001 < p\text{-}value < 0.001$; *** $p\text{-}value < 0.0001$.

Stime Effetti Fissi (SUD)						
	EMB	*Sample*	*Mean*	*Reg*	*Reg Bay*	*PMM*
Intercetta	14.510 ***	24.962 ***	14.873 ***	-	41.071 ***	15.501 ***
FEMMINA	-1.678 ***	-1.798 ***	-1.748 ***	-	-1.901 ***	-1.768 ***
IMM1	0.223	-0.179	-0.036	-	-0.199 *	-0.050
IMM2	-0.548	-0.416	-0.318	-	-0.649	-0.362
ANTICIPATARIO	-0.935	-0.905	-0.834	-	-0.991	-0.866
POSTICIPATARIO	-6.325 ***	-7.558 ***	-7.666 ***	-	-7.302 ***	-6.375 ***
ESCS	3.145 ***	3.743 ***	3.487 ***	-	4.208 ***	3.255***
No_genitori	-1.456 ***	-1.971 ***	-1.846 ***	-	-2.477 ***	-1.714 ***
Si_fratelli	0.054	0.112	0.033	-	0.149	0.178
MATH_5	0.393 ***	0.238 ***	0.384 ***	-	0.008 ***	0.377 ***
Effetti Casuali						
σ_b	7.937	8.632	8.630	-	8.779	8.637
σ_ϵ	14.793	15.599	15.312	-	16.049	15.018
VPC	22.35%	23.44%	24.10%	-	23.03%	24.85%

metodo di imputazione le stime ottenute nel Nord, Centro e Sud tramite un test Manova si evince che i tre vettori di stima sono tutti differenti. Il risultato dei test di confronto a coppie delle stime ottenute, sempre a parità di metodo di imputazione, è che le stime sono differenti nelle tre zone per tutti i metodi.
La realizzazione dei test di confronto delle stime ottenute al variare del metodo di imputazione all'interno della stessa area geografica invece restitutiscono dei risultati ben diversi da quelli emersi nel Capitolo 7.4. In questo caso, in cui i valori mancanti sono stati imputati per le sole scuole IC e condizionatamente al codice scuola, non vi sono metodi per i quali le stime sono statisticamente uguali. Questo risultato garantisce che i metodi diversi imputano diversamente i valori mancanti e di conseguenza restituiscono stime differenti dei parametri del modello. In questa situazione il solo modo di scegliere il metodo di imputazione più adeguato è quello di escludere quelli più semplici (*Mean* e *Sample*), che restituiscono stime più distorte, e preferire quelli più complessi (*EMB*, *Reg_Bay* e *PMM*) ottenuti partizionando le scuole e realizzare cross-validazione per calcolare gli errori di previsione del modello (si veda Capitolo 9).
Analogamente a quanto svolto nell'intero dataset contenente i soli Istituti Comprensivi è opportuno realizzare un confronto tra le stime del modello ottenute dopo l'imputazione dei dati condizionatamente alle scuole e quelle ottenute senza condizionare ad esse per le tre aree geografiche Nord, Centro e Sud. In questo modo è possibile verificare se imputare condizionando al codice scuola restituisce risultati ben differenti rispetto a non considerare la scuola frequentata dall'alunno. Se le stime ottenute condizionando alla scuola son diverse allora il condizionamento può essere considerato più adatto e le stime risultanti meno distorte. Analogamente a quanto svolto nell'intero dataset contenente i soli istituti comprensivi è opportuno realizzare un confronto tra le stime del modello ottenute dopo l'imputazione dei dati condizionatamente alle scuole e quelle ottenute senza condizionare ad esse per le tre aree geografiche Nord, Centro e Sud. In questo modo è possibile verificare se imputare condizionando al codice scuola restituisce risultati ben differenti rispetto a non condisiderare la scuolafrequentata dall'alunno. Se le stime ottenute condizionando alla scuola son diverse allora il condizionamento può essere considerato più adatto e le stime risultanti meno distorte.

I risultati emersi nel Capitolo 7.3 si ripropongono anche in questo da-

Tabella 8.6: Stime del modello (7.4) livello studente per le scuole nel Nord imputate senza condizionamento al codice scuola con i metodi *EMB*, *Mean*, *Sample*, *Reg*, *Reg_Bay* e *PMM*. Gli asterischi individuano i diversi livelli di significatività: . $0.01 < p\text{-}value < 0.1$; * $0.001 < p\text{-}value < 0.01$; ** $0.0001 < p\text{-}value < 0.001$; *** $p\text{-}value < 0.0001$.

Stime Effetti Fissi (NORD)						
	EMB	*Sample*	*Mean*	*Reg*	*Reg Bay*	*PMM*
Intercetta	1.417 ***	14.154 ***	2.712 ***	1.485 ***	1.488 ***	2.504 ***
FEMMINA	-1.538 ***	-2.094 ***	-1.792 ***	-1.547 ***	-1.532 ***	-1.583 ***
IMM1	-1.147 ***	-3.253 ***	-2.585 ***	-1.158 ***	-1.130 ***	-1.175 ***
IMM2	-1.721***	-3.188 ***	-2.566 ***	-1.776 ***	-1.755***	-1.824 ***
ANTICIPATARIO	-1.792 ***	-2.734***	-2.254 ***	-1.758 ***	-1.786 ***	-1.837 ***
POSTICIPATARIO	-4.262 ***	-8.550 ***	-8.575 ***	-4.413 ***	-4.417 ***	-4.257 ***
ESCS	1.907 ***	2.857 ***	2.453***	1.915 ***	1.923***	1.959 ***
No_genitori	-1.221 ***	-1.774 ***	-1.583 ***	-1.248***	-1.229 ***	-1.293 ***
Si_fratelli	0.124	-0.038	-0.001	0.112	0.078	0.099
MATH_5	0.693 ***	0.514 ***	0.676***	0.692 ***	0.692***	0.678***
Effetti Casuali						
σ_b	3.337	3.607	3.691	3.338	3.353	3.359
σ_ϵ	12.432	14.096	13.387	12.427	12.426	12.505
VPC	6.72%	6.14%	7.06%	6.73%	6.78%	6.73%

Tabella 8.7: Stime del modello (7.4) livello studente per le scuole del Centro imputate senza condizionamento al codice scuola con i metodi *EMB*, *Mean*, *Sample*, *Reg*, *Reg_ Bay* e *PMM*. Gli asterischi individuano i diversi livelli di significatività: . $0.01 < p\text{-}value < 0.1$; * $0.001 < p\text{-}value < 0.01$; ** $0.0001 < p\text{-}value < 0.001$; *** $p\text{-}value < 0.0001$.

Stime Effetti Fissi (CENTRO)						
	EMB	*Sample*	*Mean*	*Reg*	*Reg Bay*	*PMM*
Intercetta	7.983 ***	19.962 ***	8.331 ***	8.124***	8.046 ***	9.286 ***
FEMMINA	-2.465 ***	-3.026 ***	-2.753 ***	-2.496 ***	-2.494 ***	-2.574 ***
IMM1	-0.586 ***	-2.002 ***	-1.559 ***	-0.745 ***	-0.786 ***	-0.879 **
IMM2	-0.904 ***	-1.687 ***	-1.365 ***	-1.022	-0.991***	-1.049
ANTICIPATARIO	-0.834	-1.783 ***	-1.586 *	-1.481	-1.475 ***	1.477
POSTICIPATARIO	-4.869 ***	-7.468***	-7.548 ***	-4.807 ***	-4.851 ***	-4.849***
ESCS	2.424 ***	3.256***	2.906 ***	2.438 ***	2.435 ***	2.497 ***
No_genitori	-0.930 ***	-1.591***	-1.467 ***	-1.144 ***	-1.0788 ***	-1.185 ***
Si_fratelli	0.109	0.047	0.078		0.069	0.120
MATH_5	0.563 ***	0.393 ***	0.557 ***	0.562***	0.563 ***	0.546 ***
Effetti Casuali						
σ_b	4.168	4.382	4.554	4.131	4.147	4.136
σ_ϵ	13.656	14.952	14.435	13.704	13.643	13.760
VPC	8.52%	7.90%	9.05%	8.33%	8.46%	8.28%

Tabella 8.8: Stime del modello (7.4) livello studente per le scuole del Sud imputate senza condizionamento al codice scuola con i metodi *EMB*, *Mean*, *Sample*, *Reg*, *Reg_Bay* e *PMM*. Gli asterischi individuano i diversi livelli di significatività: . $0.01 < p\text{-}value < 0.1$; * $0.001 < p\text{-}value < 0.01$; ** $0.0001 < p\text{-}value < 0.001$; *** $p\text{-}value < 0.0001$.

Stime Effetti Fissi (SUD)						
	EMB	*Sample*	*Mean*	*Reg*	*Reg Bay*	*PMM*
Intercetta	14.724 ***	24.142 ***	14.873 ***	14.610 ***	14.740 ***	16.070 ***
FEMMINA	-1.693 ***	-1.808 ***	-1.748 ***	-1.714 ***	-1.719 ***	-1.743 ***
IMM1	0.405	-0.214	-0.036	0.114 **	0.309 *	0.021
IMM2	-0.395	-0.426	-0.318	-0.250	-0.324	-0.295
ANTICIPATARIO	-0.796	-0.860	-0.834	-0.790	-0.754	-0.803
POSTICIPATARIO	-6.335 ***	-7.545 ***	-7.666 ***	-6.147***	-6.232***	-6.249***
ESCS	3.164 ***	3.751 ***	3.487 ***	3.117***	3.132 ***	3.224***
No_genitori	-1.567 ***	-1.960 ***	-1.846 ***	-1.504***	-1.488 ***	-1.649 ***
Si_fratelli	0.053	0.046	0.033	0.063	0.001	-0.001
MATH_5	0.390 ***	0.237 ***	0.384 ***	0.392 ***	0.391 ***	0.372 ***
Effetti Casuali						
σ_b	7.955	8.622	8.630	7.923	7.988	8.041
σ_ϵ	14.822	15.604	15.312	14.784	14.804	14.896
VPC	22.36%	23.38%	24.10%	22.31%	22.55%	22.56%

taset ridotto: le stime sono diverse al variare dell'area. Sfruttando il test descritto nel Capitolo 7.2, che restituisce *p-value* dell'ordine di 10^{-16}, a parità di metodo di imputazione utilizzato le stime dei parametri del modello (7.4) sono diverse nelle tre aree geografiche.
La variabile *Si_fratelli* risulta non essere significativa in nessuna delle tre aree mentre *ANTICIPATARIO* è significativa solo nel Nord. Nel Sud le variabili che definiscono il grado di immigrazione non sono significative per il modello, come osservato già nel Capitolo 7.3. É importante invece commentare i valori dei parametri della variabile *MATH_5* la cui stima nel Nord assume valore 0.69 circa, decresce a 0.56 nel Centro, fino ad arrivare a 0.39 del Sud. Questo andamento decrescente del parametro definisce un differente peso della variabile associata alla valutazione della prova nell'anno precedente. Nel Sud Italia il voto della quinta elementare contribuisce molto meno, quasi la metà rispetto al Nord, alla valutazione della prova della prima media. Questo valore è quello che maggiormente giustifica lo studio separato delle aree geografiche, poiché conferma come il percorso formativo degli studenti nel Sud è più discontinuo rispetto al resto d'Italia.
Anche dal confronto delle stime prodotte da metodi diversi applicati ai dati della stessa area geografica emerge che non vi sono metodi che generano stime statisticamente uguali. Questo risultato è spiegato dal fatto che i metodi diversi di imputazione imputano in modo molto diverso i valori mancanti e dunque le successive stime del modello lineare a effetti misti applicato sono diverse.
A questo punto è interessante valutare le stime prodotte dal modello per i tre dataset al variare del condizionamento alla scuola nella fase di imputazione. Per questa ragione è necessario confrontare queste stime a parità di metodo di imputazione utilizzato e a parità di dataset. Per quanto riguarda il dataset che descrive i dati degli studenti del Nord Italia risulta che dopo aver imputato i dati tramite l'algoritmo *EMB*, *Mean* e *Sample* condizionando i dati alla scuola e non condizionandoli le stime ottenute sono significativamente uguali; per i metodi di *regressione bayesiana* e *PMM* invece risultano essere diverse. Gli stessi risultati si possono riscontrare nel Sud Italia. Nel Centro invece anche il metodo *EMB* produce risultati diversi a seconda che si condizioni o meno al codice scuola.
La spiegazione per cui i metodi *Sample*, che campiona in modo casuale da quelli osservati, e *Mean*, che sostituisce il valore mancante con il valore medio della variabile, hanno stime statisticamente uguali condizionando al

codice scuola o non condizionando è che i valori imputati non sono molto differenti. Gli altri metodi invece producono stime diversi nei due casi, a dimostrazione che i valori imputati sono diversi.
Il risultato che le stime dei parametri del modello sono differenti per tutte e tre le aree dopo aver imputato i dati mancanti condizionatamente alla scuola rispetto a non considerarla è un indicazione che considerare il codice identificativo della scuola nella fase di imputazione è la scelta più adatta: infatti in questo maniera i valori mancanti degli studenti di una scuola sono imputati tenendo in considerazione soltanto quelli osservati degli studenti della stessa scuola, in modo da considerare separatamente le diverse scuole.

Capitolo 9

Confronto validità dei metodi

I risultati di stima dei parametri del modello (7.1) ottenuti per il dataset completo, Capitolo 7, e per il dataset degli Istituti Comprensivi , Capitolo 8, sono stati confrontati tramite il test di Hotelling, che permette di verificare tramite una tecnica di ricampionamento se i vettori delle medie delle stime ottenuti con due metodi diversi sono statisticamente uguali. Questo test permette soltanto di confrontare se i metodi diversi di imputazione utilizzati producono gli stessi valori di stime dei parametri, ma non consente di valutare la bontà dei differenti metodi. Per questa ragione è necessario introdurre un nuovo strumento per valutare la bontà dei metodi di imputazione in modo da poter stabilire quale metodo sia più adatto. Lo strumento che è stato utilizzato per confrontare la validità dei diversi metodi di imputazione è l'errore di previsione associato al modello lineare a effetti misti.
Questo confronto è stato realizzato sul dataset contente solo gli Istituti Comprensivi, poiché l'interesse dell'analisi è studiare il percorso formativo degli alunni valutando se imputare principalmente i numerosi valori mancanti della variabile *MATH_5* condizionandoli alla scuola o non condizionandoli cambia notevolmente le stime dei parametri del modello.
Per poter confrontare dunque le stime dei parametri ottenute al variare del metodo di imputazione tramite l'errore di previsione del modello è stato necessario selezionare un campione del dataset IC su cui valutare l'errore. La percentuale di dati scelti per essere utilizzati come *test set* è il 25%, mentre il restante 75% è stato utilizzato come *training set* per realizzare il modello lineare.

La procedura di calcolo degli errori di previsione è la seguente:

- Selezionare in modo casuale una percentuale p di righe del dataset che saranno usate come *test set*;
- Adottare il modello (7.1) sui dati rimanenti (*training set*);
- Tramite le stime dei parametri calcolate al passo precedente realizzare le previsioni sul *test set*;
- Calcolare l'errore di previsione medio.

L'errore di previsione è stato calcolato come *Mean Square Error* (MSE)

$$MSE = E[(\hat{Y}_i - Y_i)^2] \tag{9.1}$$

dove $\hat{Y}_i$ è il valore di previsione di *MATH_corr* calcolato dal modello per lo studente i del *test set* e Y_i è invece il valore di *MATH_corr* disponibile dal dataset completo per lo stesso studente.
La procedura di calcolo appena descritta è stata iterata 100 volte per poi calcolare una media degli MSE. Gli errori di previsione sono stati calcolati per ogni dataset imputato con metodi diversi sugli stessi studenti: ad ogni iterazione vengono campionate le righe da estrarre da ogni dataset imputato con metodi diversi, vengono calcolati i modelli e infine gli errori di previsione. In questo modo è possibile confrontare gli errori di previsione prodotti a parità di studenti del *test set*.
Lo strumento appena descritto è stato utilizzato per valutare i diversi metodi di imputazione applicati allo stesso dataset. Inizialmente è stato sfruttato per valutare le imputazioni del dataset IC generato condizionatamente alla scuola frequentata e senza condizionare.

Come si può notare in Tabella 9.1 per il dataset costituito dagli Istituti Comprensivi, imputato senza condizionare i dati mancanti al codice che identifica la scuola, l'algoritmo *PMM* restituisce un errore di previsione minore rispetto agli altri metodi. Questo risultato garantisce che le stime dei parametri del modello (7.1) calcolate dopo aver imputato i dati con il metodo *Predictive Mean Matching* sono meno distorte.
La situazione invece è ben differente nel caso in cui i valori mancanti siano imputati condizionandoli all'istituto comprensivo. L'algoritmo *PMM* risulta essere comunque uno degli algoritmi più validi con un errore di previsione relativamente basso, ma il metodo *Expectation Maximization* con *Bootstrap*

Tabella 9.1: Errori di previsione calcolati per il dataset IC imputato con metodi diversi condizionando alla scuola e non condizionando

Errore di Previsione Dataset IC					
Condizionato alla scuola					
EMB	*Sample*	*Mean*	*Reg*	*Reg Bay*	*PMM*
262.94	222.70	205.83	-	258.61	192.72
Non condizionato alla scuola					
EMB	*Sample*	*Mean*	*Reg*	*Reg Bay*	*PMM*
182.35	221.97	205.83	182.83	182.63	184.51

si rivela essere il più adatto nell'imputazione condizionata insieme ai metodi di regressione lineare e di regressione lineare bayesiana, contrariamente a quanto emerso nel caso condizionato in cui risultavano essere i metodi peggiori. Per quanto riguarda i metodi *Mean* e *Sample* gli errori di previsione calcolati sono molto alti, a dimostrazione che i metodi sono banali e le stime distorte.

In modo del tutto analogo a quanto descritto, gli errori di previsione sono stati calcolati sui dataset divisi per area geografica. I risultati in Tabella 9.2 mostrano gli errori di previsione al variare dell'area geografica e al variare del condizionamento alla scuola e al metodo di imputazione utilizzato. Per tutte e tre le aree geografiche il metodo di imputazione condizionato al codice scuola che restituisce l'errore di previsione più basso è l'*EMB*, mentre per i dataset generati senza condizionamento i metodi *EMB*, *Reg*, *Reg_bay* e *PMM* restituiscono errori pressoché uguali evidenziando la validità di tutti e quattro i metodi.

É importante evidenziare come l'errore di previsione sia molto differente nei tre dataset, aumentando notevolmente da Nord a Sud. La spiegazione di questa crescita risiede nel fatto che la componente dovuta all'effetto aleatorio delle scuole al Sud contribuisce maggiormente alla varianza totale del modello e dunque i valori previsti dal modello si discostano di più dal valore realmente osservato o imputato della variabile *MATH_corr* nel dataset. Inoltre anche la percentuale di valori NA cresce da Nord a Sud e contribuisce all'aumento dell'errore di previsione (si veda Sezione 7.3 per le percentuali di valori NA stratificate per le aree).

Un risultato sorprendente è quello che evidenzia come gli errori di previsione

sui dataset non condizionati al codice scuola siano inferiori ai corrispettivi errori calcolati sui dataset condizionati. Questo risultato non è in accordo con l'utilizzo dell'imputazione condizionata al codice scuola: l'imputazione condizionata è stata introdotta per far sì che i valori mancanti di uno studente di una determinata scuola venissero imputati tenendo in considerazione i soli studenti della medesima scuola, avendo verificato che vi è un contributo importante della scuola frequentata alla valutazione della prova (soprattutto nel Sud Italia). Analizzando in dettaglio l'errore di previsione associato al metodo *EMB* si può notare come in tutte e tre le aree esso assuma all'incirca lo stesso valore sia per il dataset condizionato che per quello non condizionato alla scuola. Questo risultato suggerisce che il metodo *Expectation Maximization* con *Bootstrap* è quello più adatto per l'imputazione di questi dataset divisi per area geografica poiché è l'unico metodo "robusto" rispetto al condizionamento al codice scuola.
É opportuno sottolineare che il metodo *Expectation Maximization* con *Bootstrap* non è il metodo più corretto in assoluto. Il metodo si è rivelato essere il migliore nell'analisi di questo particolare dataset, ma ciò non significa che restituisca errori di previsione sempre più bassi degli altri metodi se applicato a dati differenti.

La differenza significativa degli errori di previsione nei due casi di imputazione condizionata e non condizionata può essere casuata dall'aver selezionato solo gli Istituti Comprensivi che possono manifestare dei comportamenti problematici non noti. Per questo motivo è stato calcolato l'errore di previsione su un campione di 2000 scuole selezionate in modo casuale dal dataset iniziale (eliminando quelle che non si possono imputare per l'elevato numero di NA presenti) e imputate con il metodo *EMB* per verificare se vi è una differenza notevole di errore tra il dataset imputato con la scuola e quello imputato senza di essa. Il *Mean Square Error* calcolato tramite cross-validazione sul nuovo campione imputato condizionatamente alla scuola risulta essere 191.82, mentre per quello ottenuto senza scuola è 192.73. I due valori sono molto simili a differenza di quelli presenti in Tabella 9.1: questo risultato evidenzia come i dati riferiti agli Istituti Comprensivi possano avere qualche problematica nella fase di imputazione condizionata al codice che identifica la scuola. Ciò è dovuto anche al fatto che l'imputazione condizionata ai gruppi non è una tecnica ancora completamente sviluppata; infatti molte scuole devono essere eliminate se

contengono troppi valori mancanti poichè i metodi non possono ottenere informazioni da altri gruppi per imputare i valori NA.

Tabella 9.2: Errori di previsione calcolati per i dataset IC divisi nelle aree geografiche imputati con metodi diversi condizionando alla scuola e non condizionando

Errore di Previsione Nord IC					
Condizionato alla scuola					
EMB	*Sample*	*Mean*	*Reg*	*Reg Bay*	*PMM*
155.96	200.30	181.47	-	202.71	165.06
Non condizionato alla scuola					
EMB	*Sample*	*Mean*	*Reg*	*Reg Bay*	*PMM*
156.55	200.81	181.47	156.37	156.51	158.70
Errore di Previsione Centro IC					
Condizionato alla scuola					
EMB	*Sample*	*Mean*	*Reg*	*Reg Bay*	*PMM*
189.52	228.45	211.32	-	250.26	199.07
Non condizionato alla scuola					
EMB	*Sample*	*Mean*	*Reg*	*Reg Bay*	*PMM*
188.74	227.20	211.32	188.26	190.28	191.01
Errore di Previsione Sud IC					
Condizionato alla scuola					
EMB	*Sample*	*Mean*	*Reg*	*Reg Bay*	*PMM*
223.36	248.36	238.88	-	262.96	229.91
Non condizionato alla scuola					
EMB	*Sample*	*Mean*	*Reg*	*Reg Bay*	*PMM*
222.60	248.23	238.88	223.21	223.65	225.57

Conclusioni

Questo lavoro eslpora il campo poco conosciuto dell'analisi dei dati mancanti di un dataset. Il dataset preso in considerazione contiene informazioni sulla prova INVALSI di matematica di 5311 classi prime medie d'Italia nell'anno scolastico 2012/2013.
L'analisi iniziale ha interessato la definizione della natura dei dati mancanti presenti nel dataset. A seconda della tipologia di *missing data* le scelte delle tecniche con cui trattare i dati variano notevolmente. Se i valori NA sono distribuiti in modo tale che la legge dei dati mancanti non dipenda né dai valori osservati né da quelli mancanti i dati sono definiti MCAR e l'utilizzo della tecnica *listwise deletion*, che prevede l'eliminazione di tutte le unità statistiche con almeno un NA, è giustificato. Nel caso in cui la distribuzione dei dati mancanti dipenda dai valori osservati invece i dati si dicono MAR. Per questo tipo di dati la *listwise deletion* restituisce risultati molto distorti e dunque è necessario utilizzare i metodi di imputazione.
Dopo aver constatato che i dati mancanti nel dataset INVALSI sono di tipo MAR, sono stati utilizzati diversi metodi di imputazione multipla per sostituire i valori NA a livello studente.
In seguito è stato applicato un modello lineare ad effetti misti con effetto aleatorio scuola per descrivere il voto della prova del test in funzione delle caratteristiche dello studente. Le stime del modello cambiano al variare del metodo di imputazione utilizzato, ma la significatività delle variabili resta invece invariata: l'indice socio-economico *ESCS*, il voto della prova dell'anno precedente e soprattutto l'area geografica d'Italia dove la scuola è ubicata sono caratteristiche molto significative. Per queste ragioni i dati sono stati divisi in base alla locazione della scuola nelle tre macro-aree geografiche Nord, Centro e Sud ed è stato realizzato un modello per ogni area.
Dopo aver imputato i dati mancanti con i diversi metodi è risultato che le stime dei parametri del modello sono molto differenti nelle tre aree. L'effetto più rilevante è quello dovuto all'effetto aleatorio scuola: frequentare

una scuola piuttosto che un'altra non contribuisce molto alla valutazione della prova per gli studenti del Nord e del Centro Italia,mentre nel Sud più del 20% della varianza del voto è dovuta alla scuola frequentata.
In seguito a questa analisi è stato selezionato un dataset ridotto contenente solo gli istituti comprensivi, con l'obiettivo di imputare i dati mancanti condizionatamente alla scuola. La scelta di selezionare solo gli IC è stata fatta poiché essi comprendono nella stessa struttura sia le scuole elementari che le scuole medie e dunque l'imputazione della variabile *MATH_5* condizionatamente al codice scuola è corretta soltanto per questi tipi di scuole. Confrontando le stime prodotte dal modello applicato al dataset ottenuto imputando i dati mancanti condizionandoli alla scuola e a quello senza condizionarli risulta che esse sono statisticamente diverse a parità di metodo utilizzato. Questi risultati validano l'utilizzo dell'imputazione condizionata ai gruppi rimarcando quanto emerso dal modello che l'effetto scuola contribuisce in modo significativo alla valutazione della prova. Inoltre anche in questo caso risultano essere significative le variabili che definiscono l'area geografica della scuola e dunque il dataset è stato suddiviso nelle tre aree e analizzato separatamente. Le stime prodotte dai modelli nelle tre aree sono diverse tra loro sia per i dataset imputati condizionando alla scuola sia per quelli imputati senza scuola, sottolineando nuovamente la necessità di considerare separatamente le aree. A parità di metodo di imputazione le stime dei parametri prodotte condizionando e non condizionando alla scuola risultano essere diverse solo per i metodi più complessi e onerosi *Reg_Bay* e *PMM* in tutte e tre le aree, mentre quelle ottenute con i metodi più banali *Mean* e *Sample* risultano essere statisticamente uguali. Queste conclusioni portano a considerare validi i metodi più complessi computazionalmente tra cui principalmente il metodo *PMM*.
Successivamente, è stato introdotto uno strumento per la valutazione della bontà dei metodi di imputazione che si basa sul confronto degli errori di previsione del modello lineare. Gli errori di previsione sono stati calcolati per i dataset divisi per area geografica, sia per i dati imputati condizionatamente alla scuola sia per quelli non condizionati. Sorprendentemente gli errori di previsione associati ai dataset imputati senza condizionamento alla scuola risultano essere più piccoli di quelli calcolati imputando con il codice scuola. L'unico metodo che lascia pressoché invariato l'errore di previsione nei due dataset in tutte e tre le aree geografiche è l'*EMB* che risulta essere anche quello a cui è associato l'errore di previsione più basso

e dunque può essere considerato il metodo migliore per l'imputazione di questo particolare dataset.

Nelle analisi realizzate in questo lavoro sono state considerate solo le variabili che descrivono le caratteristiche dello studente più alcune variabili associate alla scuola. Questa scelta è stata fatta principalmente perché la variabile con più valori NA è a livello studente ($MATH\_5$) ed è una variabile molto importante per studiare l'andamento scolastico degli alunni. Le informazioni riguardanti le classi possono essere oggetto di studi futuri sulle influenze dei diversi insegnanti sugli alunni valutando in questo modo se vi è una disparità di voti tra classi della stessa scuola. Un'altra interessante analisi si potrebbe realizzare per confrontare le regioni della stessa area geografica per evidenziare se vi è omogeneità nelle valutazioni tra esse. Ciò permetterebbe di identificare quali contribuiscono in modo positivo o negativo alle stime dei parametri del modello e prendere gli opportuni provvedimenti per migliorare il rendimento delle scuole della regione.

Codici R

In questo capitolo sono riportati i codici R utilizzati per realizzare i risultati descritti.

Test per la determinazione della natura dei dati mancanti

```
library(MissMech)
#seleziono un sotttocampione su cui realizzare il test
outMCAR<-TestMCARNormality(data=mate[sample(nrow(mate),
           10000,rep=FALSE),])
outMCAR
```

Imputazioni e modello

```
library(mice)
library(nlme)
library(Amelia)

lme_completo = lme(MATH_corr~FEMMINA+S1+S2+CENTRO
                +SUD+ANTICIPATARIO+POSTICIPATARIO
                +ESCS+No_genitori+Si_fratelli+ MATH_5,data = data,
                random=~1|CODICE_SCUOLA,na.action = na.exclude)
summary(lme_completo)

#####MULTIPLE IMPUTATION####
data_imp<-data[ , -c(1,2,3,13:24,27:39)]
attach(data_imp)

#SAMPLE
imp_sample<-mice(data_imp,m=5,method="sample")
fit_sample<- with(imp_sample,lme(MATH_corr~FEMMINA+S1+S2
             +CENTRO+SUD+ANTICIPATARIO
             +POSTICIPATARIO+ESCS+No_genitori+Si_fratelli+MATH_5,
             random = ~ 1 | CODICE_SCUOLA)
```

```
summary(pool(fit_sample))

#MEAN
imp_mean<-mice(data_imp,m=5,method=c("logreg","logreg","logreg",
               "logreg","logreg","mean","logreg","logreg","mean",
               "␣","␣","mean","␣"))
fit_mean<- with(imp_mean,lme(MATH_corr~FEMMINA+S1+S2
                 +CENTRO+MEZZOGIORNO+ ANTICIPATARIO
                 +POSTICIPATARIO+ESCS+ No_genitori+Si_fratelli+ MATH_5,
                  random = ~ 1 | CODICE_SCUOLA)
summary(pool(fit_mean))

#REGRESSIONE LINEARE NON BAYESIANA
data_imp1<-data[ , -c(1,2,3,13:24,27:39,41)]
attach(data_imp1)
imp_reg<-mice(data_imp1,m=5,method=c("logreg","logreg","logreg",
              "logreg","logreg","norm.nob","logreg","logreg",
              "norm.nob","␣","␣","norm.nob"))
fit_reg<-with(imp_reg, lme(MATH_corr~FEMMINA+S1+S2+
            CENTRO+SUD+ANTICIPATARIO+POSTICIPATARIO+ESCS+No_genitori
           +Si_fratelli+MATH_5,random = ~ 1 | CODICE_SCUOLA)
summary(pool(fit_reg))

#REGRESSIONE LINEARE BAYESIANA
data_imp1<-data[ , -c(1,2,3,13:24,27:39,41)]
imp_reg_bay<-mice(data_imp1,m=5,method=c("logreg","logreg","logreg",
                  "logreg","logreg","norm","logreg","logreg","norm",
                  "␣","␣","norm"))
fit_reg_bay<-with(imp_reg_bay, lme(MATH_corr~FEMMINA+S1+S2
                +CENTRO+SUD+ANTICIPATARIO+POSTICIPATARIO+ESCS
                +No_genitori+Si_fratelli+MATH_5,
                 random = ~ 1 | CODICE_SCUOLA)
summary(pool(fit_reg_bay))

#PMM
imp_pmm<-mice(data_imp1,m=5,method=c("logreg","logreg","logreg",
            "logreg","logreg","pmm","logreg","logreg","pmm",
            "␣","␣","pmm"))
fit_pmm<-with(imp_pmm, lme(MATH_corr~FEMMINA+S1+S2
```

```
            +CENTRO+SUD+ANTICIPATARIO+POSTICIPATARIO+ESCS
            +No_genitori+Si_fratelli+MATH_5,
            random = ~ 1 | CODICE_SCUOLA)
summary(pool(fit_pmm))

#EMB
imp.amelia <- amelia(data.imp)

lme.amelia<-lapply(imp.amelia$imputations,
        function(i) lme(MATH_corr~FEMMINA+S1+S2+CENTRO+SUD
        +ANTICIPATARIO+POSTICIPATARIO+ESCS+No_genitori+Si_fratelli
        +MATH_5,random = ~ 1 | CODICE_SCUOLA,data = i))

summary(lme.amelia)
```

Errori di previsione per il dataset contenente gli IC del Nord

```
library(mice)
library(nlme)
library(Amelia)

#seleziono 25% di dati
perc=0.25
for(i in 1:100){
  indici <- sample(1:nrow(dati_completi_nord_mean),
  round(nrow(dati_completi_nord_mean)*perc))
  #mean scuola
  test_set_nord_mean<-dati_completi_nord_mean[indici, ]
  training_set_nord_mean<-dati_completi_nord_mean[-indici,]
  fit_training_nord_mean=lme(MATH_corr~FEMMINA+S1+S2+ANTICIPATARIO
                        +POSTICIPATARIO+ESCS+No_genitori+Si_fratelli
                        +MATH_5,data = training_set_nord_mean,
                        random=~ 1|CODICE_SCUOLA)
  summary(fit_training_nord_mean)
  pred_nord_mean<- predict(fit_training_nord_mean,
                        newdata = test_set_nord_mean)
  MSE_nord_mean<-mean((test_set_nord_mean[9]-pred_nord_mean)^2)
  errore_nord_MSE_scuola[i,1]<-MSE_nord_mean
  #sample scuola
  test_set_nord_sample<-dati_completi_nord_sample[indici, ]
  training_set_nord_sample<-dati_completi_nord_sample[-indici,]
```

```
fit_training_nord_sample=lme(MATH_corr~FEMMINA+S1+S2+ANTICIPATARIO
                    +POSTICIPATARIO+ESCS+No_genitori+Si_fratelli
                    +MATH_5,data = training_set_nord_sample,
                    random=~ 1|CODICE_SCUOLA)
summary(fit_training_nord_sample)
pred_nord_sample<- predict(fit_training_nord_sample,
                   newdata = test_set_nord_sample)
MSE_nord_sample<-mean((test_set_nord_sample[9]-pred_nord_sample)^2)
errore_nord_MSE_scuola[i,2]<-MSE_nord_sample
#reg_bay scuola
test_set_nord_reg_bay<-dati_completi_nord_reg_bay[indici, ]
training_set_nord_reg_bay<-dati_completi_nord_reg_bay[-indici,]
fit_training_nord_reg_bay=lme(MATH_corr~FEMMINA+S1+S2+ANTICIPATARIO
                 +POSTICIPATARIO+ESCS+No_genitori+Si_fratelli
                 +MATH_5,data = training_set_nord_reg_bay,
                 random=~ 1|CODICE_SCUOLA)
summary(fit_training_nord_reg_bay)
pred_nord_reg_bay<- predict(fit_training_nord_reg_bay,
                newdata = test_set_nord_reg_bay)
MSE_nord_reg_bay<-mean((test_set_nord_reg_bay[9]-pred_nord_reg_bay)^2)
errore_nord_MSE_scuola[i,3]<-MSE_nord_reg_bay
#pmm scuola
test_set_nord_ppmm<-dati_completi_nord_ppmm[indici, ]
training_set_nord_ppmm<-dati_completi_nord_ppmm[-indici,]
fit_training_nord_ppmm= lme(MATH_corr~FEMMINA+S1+S2+ANTICIPATARIO
              +POSTICIPATARIO+ESCS+No_genitori+Si_fratelli
              +MATH_5, data = training_set_nord_ppmm,
              random=~ 1|CODICE_SCUOLA)
summary(fit_training_nord_ppmm)
pred_nord_ppmm<- predict(fit_training_nord_ppmm,
            newdata = test_set_nord_ppmm)
MSE_nord_ppmm<-mean((test_set_nord_ppmm[9]-pred_nord_ppmm)^2)
errore_nord_MSE_scuola[i,4]<-MSE_nord_ppmm
#emb scuola
test_set_nord_emb<-dati_completi_nord_emb[indici, ]
training_set_nord_emb<-dati_completi_nord_emb[-indici,]
fit_training_nord_emb= lme(MATH_corr~FEMMINA+S1+S2+ANTICIPATARIO
            +POSTICIPATARIO+ESCS+No_genitori+Si_fratelli
            +MATH_5, data = training_set_nord_emb,
```

```
                random=~ 1|CODICE_SCUOLA)
  summary(fit_training_nord_emb)
  pred_nord_emb<- predict(fit_training_nord_emb,
              newdata = test_set_nord_emb)
  MSE_nord_emb<-mean((test_set_nord_emb[9]-pred_nord_emb)^2)
  errore_nord_MSE_scuola[i,5]<-MSE_nord_emb
  ######################################
  #analogamente per il dataset imputato senza scuola
  ...

 }
colMeans(errore_nord_MSE_scuola)
colMeans(errore_nord_MSE_NO_scuola)
```

Bibliografia

[1] Tommaso Agasisti, Francesca Ieva, and Anna Maria Paganoni. Heterogeneity, school-effects and achievement gaps across italian regions: further evidence from statistical modeling. *Submitted [online] http://mox. polimi. it/it/progetti/pubblicazioni/quade rni/07-2014. pdf*, 6:121–158, 2014.

[2] Tommaso Agasisti and Giorgio Vittadini. Regional economic disparities as determinants of student's achievement in italy. *Research in Applied Economics*, 4(2):33, 2012.

[3] Paul D. Allison. Handling missing data by maximum likelihood. In *SAS global forum*, volume 312, 2012.

[4] Theodore W. Anderson and Donald A. Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954.

[5] Alessandro Barbiero, Pier Alda Ferrari, and Giancarlo Manzi. *ForImp: Imputation of Missing Values Through a Forward Imputation Algorithm*, 2015. R package version 1.0.3.

[6] James R. Carpenter, Harvey Goldstein, Michael G. Kenward, et al. Realcom-impute software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software*, 45(5):1–14, 2011.

[7] A. Dempster and Donald Rubin. Incomplete data in sample surveys. *Sample surveys*, 2:3–10, 1983.

[8] Craig K. Enders. *Applied missing data analysis*. Guilford Publications, 2010.

[9] Ronald Aylmer Fisher. Statistical methods for research workers. 1934.

[10] Harvey Goldstein. *Multilevel models in education and social research.* Oxford University Press, 1987.

[11] Harvey Goldstein, Fiona Steele, Jon Rasbash, and Christopher Charlton. Realcom: methodology for realistically complex multilevel modelling. *Bristol: Centre for Multilevel Modelling, Graduate School of Education, University of Bristol*, 2008.

[12] Douglas M. Hawkins. A new test for multivariate normality and homoscedasticity. *Technometrics*, 23(1):105–110, 1981.

[13] James Honaker, Gary King, Matthew Blackwell, et al. Amelia ii: A program for missing data. *Journal of statistical software*, 45(7):1–47, 2011.

[14] G.W. Jacobusse. Winmice user's manual. *TNO Quality of Life, Leiden. URL http://www. multiple-imputation. com*, 2005.

[15] Mortaza Jamshidian and Peter M. Bentler. Ml estimation of mean and covariance structures with missing data using complete data routines. *Journal of Educational and Behavioral Statistics*, 24(1):21–24, 1999.

[16] Mortaza Jamshidian and Siavash Jalal. Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. *Psychometrika*, 75(4):649–674, 2010.

[17] Mortaza Jamshidian, Siavash Jala Jalal, and Camden Jansen. Missmech: an r package for testing homoscedasticity, multivariate normality, and missing completely at random (mcar). *Journal of Statistical Software*, 56(6), 2014.

[18] Rafa M. Kasim and Stephen W. Raudenbush. Application of gibbs sampling to nested variance components models with heterogeneous within-group variance. *Journal of Educational and Behavioral Statistics*, 23(2):93–116, 1998.

[19] Kevin H. Kim and Peter M. Bentler. Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika*, 67(4):609–623, 2002.

[20] Brendan Klick. Missing data in health research: The good, the bad and the ugly. 2007.

[21] Roderick J.A. Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202, 1988.

[22] Roderick J.A. Little and Donald B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[23] Chiara Masci, Francesca Ieva, Tommaso Agasisti, and Anna Maria Paganoni. Bivariate multilevel models for the analysis of mathematics and reading pupils' achievements. 2015.

[24] Jerzy Neyman. "smooth" test for goodness of fit. *Scandinavian Actuarial Journal*, 1937(3-4):149–199, 1937.

[25] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[26] Jon Rasbash, William Browne, Harvey Goldstein, Min Yang, Ian Plewis, Michael Healy, Geoff Woodhouse, David Draper, Ian Langford, and Toby Lewis. A user's guide to mlwin. 2000.

[27] Donald Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

[28] Fritz W. Scholz and Michael A. Stephens. K-sample anderson–darling tests. *Journal of the American Statistical Association*, 82(399):918–924, 1987.

[29] Muni S. Srivastava and Mohammad Dolatabadi. Multiple imputation and other resampling schemes for imputing missing observations. *Journal of Multivariate Analysis*, 100(9):1919–1937, 2009.

[30] Muni Shanker Srivastava and M.S. Srivastava. *Methods of multivariate statistics*, volume 1. Wiley-Interscience New York, 2002.

[31] Stef Van Buuren et al. Multiple imputation of multilevel data. *Handbook of advanced multilevel analysis*, pages 173–196, 2011.

[32] Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45(3), 2011.

[33] Stef Van Buuren and Karin Oudshoorn. Flexible multivariate imputation by mice. *Leiden, The Netherlands: TNO Prevention Center*, 1999.

[34] Gerko Vink, Laurence E. Frank, Jeroen Pannekoek, and Stef van Buuren. Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, 68(1):61–90, 2014.

[35] Gerko Vink, Goran Lazendic, and Stef van Buuren. Partioned predictive mean matching as a large data multilevel imputation technique. *Psychological Test and Assessment Modeling*, 57(4):577–594, 2015.