Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

Trattamento di dati mancanti: un'applicazione all'analisi delle valutazioni degli studenti dell'Ateneo di Padova

Relatore Prof. Omar Paccagnella
Dipartimento di Scienze Statistiche

Laureando: Benedetta Ruffo
Matricola N 1130946

Anno Accademico 2017/2018

Indice

Elenco delle Tabelle

Elenco delle Figure

Introduzione

Nell'analisi di un dataset accade molto spesso che vi siano dati mancanti (NA - *not available*) per alcune variabili; le cause possono essere diverse e dipendono dalla natura stessa dei dati.

Da anni, grazie all'utilizzo di opportuni software per le analisi statistiche, è possibile studiare dataset di grandi dimensioni; l'alta dimensionalità dei dati però rende molto probabile la presenza di un numero elevato di dati mancanti.

Sono quindi state sviluppate numerose tecniche di analisi, 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), che prevede l'eliminazione di ogni unità statistica (i.e. riga del dataset) che contenga almeno un'informazione mancante. Il metodo però in molti casi non è adatto e comporta un'elevata perdita di informazioni, producendo spesso anche dei risultati distorti.

Negli ultimi anni sono stati introdotti metodi diversi al fine di ridurre la distorsione dei risultati e di trovare quello più adatto rispetto alla tipologia di NA presenti nel dataset.

In questa tesi verranno presentate varie tecniche di trattamento di dati mancanti, alcune delle quali verranno utilizzate e messe a confronto nell'analisi di un dataset contenente le risposte al questionario sulla valutazione della didattica dell'Ateneo di Padova, rilevate nell'anno accademico 2012/2013. In particolare, l'attenzione verrà posta sul metodo PMM (*Predictive Mean Matching*), molto utile data la scala di risposta delle domande.

Specificatamente l'interesse viene rivolto alla domanda riferita alla soddisfazione complessiva dell'attività didattica valutata, la quale contiene una percentuale cospicua di dati mancanti e questo permette il confronto tra i diversi metodi.

Il dataset ha un'evidente struttura gerarchica, in quanto le valutazioni sono annidate entro i corsi seguiti dagli studenti. Sfruttando questa struttura dei dati, in questo lavoro di tesi si vuole quindi introdurre un metodo PMM multilivello.

Attualmente la letteratura su questo tema è estremamente ridotta e tendenzialmente orientata a sfruttare la gerarchia dei dati con altri modi di imputazione; per tale motivo il metodo proposto verrà analizzato anche per mezzo di alcuni studi di simulazione.

Il lavoro è quindi strutturato in sei capitoli: nel Capitolo 1 viene presentato il problema dei *missing data* e le loro possibili strutture all'interno di un dataset; nel Capitolo 2 sono

invece descritti i metodi classici per trattare i dati mancanti, suddivisi in tre macro-classi: metodi di eliminazione, metodi di imputazione singola, imputazione multipla.
Nel Capitolo 3 viene presentato il dataset di interesse, tramite analisi descrittive, e viene successivamente realizzata un'analisi dei dati mancanti.
Il Capitolo 4 è dedicato alla presentazione dei modelli multilivello, data la struttura gerarchica dei dati, mentre nel Capitolo 5 si procede con lo studio della relazione tra la variabile di interesse e le variabili esplicative, tramite la stima di due modelli ad intercetta casuale, con la tecnica della *listwise deletion*; si verifica anche la presenza di una possibile distorsione da selezione (*selection bias*) che si potrebbe riscontrare se i rispondenti al questionario non venissero selezionati casualmente, tramite un modello di Heckman.
Il Capitolo 6 riguarda l'applicazione di alcuni dei metodi di imputazione proposti alla variabile di interesse, basandosi anche su modelli multilivello, nonché una proposta del metodo PMM multilivello per imputazione multipla.
Viene quindi realizzato uno studio di simulazione per analizzare le differenze tra alcuni dei metodi di imputazione applicati, da cui vengono tratte opportune considerazioni.
Infine un capitolo conclusivo riassume i principali risultati ottenuti, nonché evidenzia alcuni suggerimenti per sfruttare le soluzioni proposte in attività di ricerca future.

Capitolo 1

I dati mancanti

Il problema dei dati mancanti è rilevante nella ricerca empirica, in particolare nelle scienze economiche, in cui la somministrazione di questionari costituiti da più item è una delle tecniche più diffuse per la raccolta di dati e informazioni.
In letteratura non esiste un'unica tecnica o metodologia di approccio a questo problema, ma l'analisi di un dataset con osservazioni incomplete può essere affrontata secondo varie strategie, che possono condurre a risultati più o meno validi.
Prima di parlare dei metodi di trattamento dei dati mancanti è necessario soffermarsi su alcune questioni metodologiche di base come il concetto di non risposta, la struttura delle mancate risposte e il meccanismo generatore dei dati mancanti, che esprime il legame tra quest'ultimi ed i valori osservati delle variabili.

1.1 Il concetto di non risposta

Generalmente le metodologie statistiche standard sono sviluppate per analizzare dataset in forma rettangolare: le righe rappresentano le osservazioni, mentre le colonne sono le variabili misurate su ciascuna osservazione. Si può parlare di non risposta (Barcaroli et al., 1999) ogni qualvolta il valore di una variabile per una certa unità è:

- mancante, se non è stato possibile osservarlo;
- errato, se non è quello che nella realtà è posseduto dall'unità considerata. In particolare, il valore errato può essere ricondotto a tre tipologie:
  1) fuori dominio, se è esterno all'intervallo dei valori ammissibili;
  2) anomalo (*outlier*), se la risposta fornita dall'unità in esame si discosta molto da quella fornita da tutte le altre unità;
  3) incompatibile, se c'è contraddizione con i valori delle altre variabili rilevati sulla stessa unità.

La non risposta causa sia un incremento nella variabilità degli stimatori, dovuta ad una riduzione della base campionaria di analisi e/o all'applicazione di metodi per il

trattamento della stessa, sia stimatori distorti, se i rispondenti differiscono sistematicamente dai non rispondenti rispetto a certe caratteristiche di interesse.
Principalmente si distinguono due tipi di non risposta.

1) Non risposta totale (*unit non response*): si riferisce al tipo di non risposta in cui non si ha nessuna informazione disponibile (rilevata) per alcune unità campionarie. Le ragioni possono essere varie e dipendono dalle modalità di raccolta dei dati; alcune di esse possono essere: impossibilità di contatto, non reperibilità, inabilità a rispondere, rifiuto, questionario non restituito, etc.

2) Non risposta parziale (*item non response*): si intende la mancata risposta ad uno o più quesiti di un questionario e i motivi possono essere molteplici: l'intervistato non comprende il quesito, domanda troppo personale, rifiuto a rispondere, etc.
   Questo tipo di incompletezza risulta la più semplice da gestire, in quanto si dispone di una serie di informazioni sull'individuo in questione, il quale è presente nel dataset, ma con alcuni campi vuoti. Questo è il caso in cui si ha un dataset non rettangolare e dunque le analisi statistiche standard non sono più direttamente applicabili.

1.2 Meccanismi generatori di dati mancanti

Poiché la presenza di dati mancanti può influire in modo significativo sulle proprietà degli stimatori e pertanto può condurre a risultati inferenziali non corretti, è necessario stabilire quale sia il meccanismo generatore di tali dati, che definisce il legame tra essi ed i valori delle variabili osservate nel dataset.
Il concetto di generatore di dati mancanti è stato introdotto da Rubin nel 1976, il quale decise di trattare l'indicatore dei dati mancanti come una variabile casuale ed assegnare ad esso una distribuzione di probabilità.
Definendo con $Y = (y_{ij})$ il dataset completo (dove *i* si riferisce alle osservazioni e *j* alle variabili) 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 $f(M|Y, \phi)$, dove ϕ identifica un parametro (o un insieme di parametri) non noto.

Siano anche Y_{obs} la parte dei dati Y realmente osservati, e Y_{mis} la componente riferita a quelli mancanti (Little and Rubin, 2002).

Se la mancanza del dato non dipende dai valori di Y, né osservati né mancanti, cioè

$$f(M|Y,\phi) = f(M|\phi) \quad per\ ogni\ Y,\phi$$

i dati mancanti sono detti *Missing Completely At Random* (MCAR).
L'assunzione che i dati siano di questo tipo non presuppone che la loro struttura sia casuale, ma che la loro distribuzione non dipenda dal valore assunto dai dati stessi.
Questa prima tipologia è la più semplice da trattare in quanto se i dati sono mancanti in maniera completamente casuale possono essere ignorati e l'unica conseguenza sugli stimatori è la perdita di efficienza poiché si utilizzano $n_{obs} < n$ osservazioni, con il conseguente aumento di varianza degli stimatori.

Se la mancanza dei dati nel dataset Y dipende invece soltanto dalla componente Y_{obs} osservata e non da quella mancante, allora i dati si definiscono *Missing At Random* (MAR). Per questo tipo di dati la distribuzione condizionata di M dato Y è così definita:

$$f(M|Y,\phi) = f(M|Y_{obs},\phi) \quad per\ ogni\ Y_{mis},\phi.$$

In questo caso si parla di meccanismo di non risposta ignorabile, in quanto non serve specificare un modello di non risposta $f(M|Y,\phi)$ per ottenere valide inferenze sui parametri che si vogliono stimare.
Essendo comunque in presenza di dati mancanti, utilizzando $n_{obs} < n$, si ha una perdita di efficienza come nel caso MCAR, ma anche una distorsione degli stimatori, in quanto i dati mancanti non possono essere ignorati.

Il terzo tipo di meccanismo generatore è noto come *Missing Not At Random* (MNAR), che descrive un comportamento non casuale della presenza di dati mancanti e la distribuzione della matrice M dipende sia da Y_{mis} che da Y_{obs}:

$$f(M|Y,\phi) = f(M|Y_{obs},Y_{mis},\phi) \quad per\ ogni\ \phi.$$

Rispetto a quest'ultima tipologia si possono verificare due differenti situazioni: i dati mancanti dipendono da variabili completamente non osservate, oppure i dati mancanti di una certa variabile sono dovuti alla variabile stessa.

Si tratta di un meccanismo di non risposta non ignorabile e, data la differenza tra dati osservati e non, si ha una forte distorsione degli stimatori, il cui grado dipende dalla quantità di dati mancanti. Inoltre, analogamente ai casi precedenti, si ha una minor efficienza.

Si può quindi comprendere come la distinzione dei tipi di dati mancanti sia fondamentale: le tecniche per trattarli cambiano al variare della loro tipologia nel dataset in esame. Per esempio, la tecnica della *listwise deletion*, che consiste nell'eliminazione delle osservazioni del dataset che contengono almeno un valore mancante, restituisce buoni risultati nel caso in cui i dati sono di tipo MCAR e risultati distorti nel caso MAR e MNAR.

Identificare la natura del dato mancante risulta quindi essenziale, ma spesso difficile.

In letteratura infatti esistono dei test statistici per determinare se il meccanismo è di tipo MCAR, ma non vi sono test per verificare la struttura MAR e MNAR. L'unico modo per distinguere tra questi due tipi di meccanismi è sapere già la natura dei dati prima di analizzarli, grazie ad informazioni riguardanti le modalità, le tecniche e gli strumenti usati per la raccolta degli stessi dati.

Capitolo 2

Metodi per il trattamento di dati mancanti

La letteratura relativa all'analisi di dataset incompleti è relativamente recente ed offre numerose tecniche e metodologie per affrontare il problema dei dati mancanti e, in questo elaborato, si considerano tre macro-classi:

- Metodi di eliminazione
- Metodi di imputazione singola
- Imputazione multipla

2.1 Metodi di eliminazione

Un metodo semplice, indicato nei casi in cui l'ammontare dei dati mancanti è limitato ed il meccanismo che li ha generati è di tipo MCAR, è quello di cancellare le osservazioni (*case deletion*).
I metodi per far questo sono due: *listwise deletion* e *pairwise deletion.*

2.1.1 Listwise deletion

La *listwise deletion* (o analisi dei casi completi) è la tecnica più utilizzata e prevede, come già accennato in precedenza, l'eliminazione di tutte le righe, quindi tutte le osservazioni, che contengono tra le variabili almeno un valore mancante.
I vantaggi sono legati alla semplicità di esecuzione e alla comparabilità tra le statistiche univariate, in quanto realizzate tutte sullo stesso insieme di casi; gli svantaggi sono invece dati dalla potenziale perdita di informazione, dovuta ai casi scartati, che può essere tradotta in termini di perdita di precisione, ma anche di distorsione, qualora il meccanismo generatore dei dati mancanti non fosse MCAR ma soltanto MAR.

Questa tecnica quindi ha senso solo se l'imprecisione e la distorsione sono di entità ridotta, valori non sempre riconducibili soltanto alla proporzione di casi scartati sul totale del campione.

2.1.2 Pairwise deletion

Una valida alternativa al metodo della *listwise deletion*, che provoca la perdita di informazioni anche per le variabili in cui il dato non manca, è la *pairwise deletion* (o analisi dei casi disponibili).

Questa tecnica include tutte le unità statistiche per le quali la variabile di interesse è stata osservata, prevede la creazione di differenti dataset a seconda dei diversi studi che si vogliono realizzare e per ognuno si considerano solo le variabili di interesse per l'analisi, eliminando successivamente i valori mancanti.

Lo svantaggio principale 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 vantaggio è però quello di ridurre la distorsione delle stime rispetto alla *listwise deletion*, pur aumentando i costi computazionali.

2.2 Metodi di imputazione singola

Quando i dati mancanti non sono di tipo MCAR, bensì MAR, è opportuno sostituirli con appropriate funzioni dei dati effettivamente osservati (imputazione).

Si discutono quindi ora i metodi che imputano i valori mancanti; essi possono essere applicati per imputare un valore per ogni dato mancante (imputazione singola), o, in alcuni casi, per imputare più di un valore, per consentire un'appropriata valutazione dell'incertezza dell'imputazione (imputazione multipla).

L'imputazione singola è un metodo statistico usato con lo scopo di eliminare i valori mancanti all'interno di un dataset, sostituendoli con dei valori ammissibili per la variabile considerata, in modo da ripristinare la matrice dei dati completa.

Si tratta di una tecnica interessante poiché riduce la perdita di informazioni, ma nello stesso tempo è pericolosa (Dempster e Rubin, 1983). Il vantaggio deriva dal fatto che, una volta sostituiti i valori, consente di lavorare sul dataset come se fosse completo,

facilitando l'analisi dei dati e la presentazione dei risultati; il pericolo però è proprio quello di considerare i dati imputati come realmente osservati e trattarli come tali nelle analisi, le quali non tengono conto dell'incertezza dovuta all'ignoranza riguardo al vero valore assunto dalle variabili ove l'informazione è mancante (ciò comporta una riduzione della variabilità).

Per sostituire i valori mancanti è necessario definire dei criteri con cui imputare i dati ed esistono diverse tecniche per farlo; ognuna restituisce dei valori differenti.

Di seguito si riportano le principali:

- *Imputazione con la media*: il dato mancante viene sostituito con la media della variabile a cui appartiene, calcolata sul totale dei rispondenti. Questo metodo può essere utilizzato solo per le variabili quantitative ed utilizzato spesso per la sua semplicità, ma introduce una seria distorsione nella distribuzione della variabile creando un picco artificiale in corrispondenza del suo valor medio ed inoltre non dà buoni risultati nella stima della varianza (sottostima la variabilità).

- *Campionamento aleatorio:* il dato mancante è sostituito da un valore estratto in modo casuale da quelli disponibili per la variabile di riferimento.

- *Imputazione con regressione:* si tratta di un approccio basato sulle informazioni disponibili per le altre variabili che consiste nel regredire i valori osservati della variabile oggetto di imputazione sulle altre variabili (ausiliarie) e successivamente i valori mancanti vengono sostituiti con quelli predetti mediante l'equazione di regressione. Se la variabile di riferimento è quantitativa generalmente vengono utilizzati modelli di regressione lineare, mentre se è qualitativa si possono adottare modelli log-lineari o logistici. Le variabili ausiliarie possono essere sia di natura quantitativa che qualitativa.

  Il fatto di far uso di un numero elevato di variabili permette di ridurre, più che con altri metodi, le distorsioni generate dalle mancate risposte e le relazioni tra le variabili usate nel modello vengono ben preservate. Gli svantaggi invece sono l'introduzione di distorsione nella distribuzione della variabile (sebbene meno del metodo di imputazione con la media), distorsione nelle relazioni tra le variabili non utilizzate nel modello, il fatto di non preservare sufficientemente la variabilità

delle distribuzioni marginali, il rischio che possano essere imputati valori non reali (non ammissibili) e la forte influenza rispetto alla presenza di valori anomali.

- *Imputazione Hot-Deck*: i dati mancanti vengono forniti da un "donatore", ovvero un caso privo di dati mancanti, scelto, generalmente in modo casuale nella stessa base di dati, entro un insieme di casi simili a quello con dati mancanti.

- *Imputazione Cold deck:* sostituisce il valore mancante con un valore derivante da una sorgente esterna, come una precedente realizzazione della stessa variabile.

- *Imputazione con nearest neighbour*: il metodo definisce una nozione di distanza tra le osservazioni (basata sul tipo di covariate) e sceglie i valori imputati che provengono dall'unità più vicina all'osservazione con i valori mancanti, ossia il rispondente "più vicino". Quest'ultimo è determinato per mezzo di una funzione di distanza applicata alle variabili ausiliarie.

  La procedura è la seguente:

  1. calcolare la distanza (considerando i valori assunti sulle variabili ausiliarie) tra l'unità del campione con mancata risposta e tutte le altre unità senza dati mancanti usando un'appropriata *funzione di distanza*;
  2. determinare l'unità più vicina all'unità di interesse;
  3. utilizzare il valore dell'unità "più vicina" per effettuare l'imputazione.

  Le varianti di questo metodo possono essere ricondotte all'uso di differenti funzioni di distanza. Le funzioni generalmente usate sono:

  a) *distanza Euclidea*;

  b) *distanza ponderata*, in cui le variabili utilizzate nella funzione sono premoltiplicate per un peso rappresentativo della loro maggiore o minore importanza;

  c) *distanza di Mahalanobis*;

  d) *distanza Minmax*.

- *Predictive Mean Matching (PMM):* il valore mancante è imputato con il valore, previsto da un modello di regressione, di un'osservazione con covariate simili e complete.

  Si tratta di un metodo parzialmente parametrico che fa corrispondere il valore mancante al valore osservato con la media predittiva più vicina; in altri termini, non si cerca direttamente il valore da imputare, ma il donatore con le caratteristiche più simili a quelle dell'unità che presenta il valore mancante; combina quindi gli approcci della regressione lineare standard e dell'imputazione tramite *nearest neighbour*.

  Per descrivere il funzionamento generale del metodo si suppone che le variabili $\boldsymbol{Y}$ ($Y_1, \ldots, Y_p$) siano quelle che contengono i valori mancanti e che $\boldsymbol{X}$ sia un insieme di p variabili completamente osservate, che saranno utilizzate per imputare $\boldsymbol{Y}$.

  Il metodo è il seguente (van Buuren, 2012):

  1. si stima un modello di regressione di $\boldsymbol{Y}$ su $\boldsymbol{X}$;
  2. date le stime ottenute nel passo 1, si calcola la media predittiva, $\boldsymbol{Y^*} \equiv E(\boldsymbol{Y}|\boldsymbol{X})$, sia delle unità non complete (*riceventi*), sia delle unità con tutte le variabili osservate (*donatori*);
  3. per ogni unità ricevente (u_r) viene selezionata un'unità donatore (u_d) in modo da minimizzare la distanza di Mahalanobis definita come:

     $$D(u_d, u_r) \equiv (\boldsymbol{y_d^*} - \boldsymbol{y_r^*})^T \boldsymbol{S}^{-1} (\boldsymbol{y_d^*} - \boldsymbol{y_r^*}) ,$$

     dove y_d^*, y_r^* indicano le medie predittive stimate rispettivamente su donatore e ricevente e $\boldsymbol{S}$ è la matrice di varianza e covarianza residua del modello di regressione;
  4. ogni unità incompleta viene imputata trasferendo i valori delle variabili $\boldsymbol{Y}$ dal donatore selezionato al punto 3.

Estraendo proprio dai dati osservati, il metodo PMM preserva la distribuzione dei valori osservati nella parte mancante dei dati, il che rende questo metodo più robusto rispetto all'approccio completamente parametrico della regressione lineare.

I metodi di imputazione singola sono dunque interessanti poiché, sostituendo con un nuovo dato imputato il valore di ogni dato mancante, creano un dataset completo su cui è

possibile applicare le tecniche classiche di analisi, sfruttando anche le informazioni che i metodi di eliminazione avrebbero trascurato.
Nella maggior parte dei casi però le stime dei parametri risultano essere distorte, anche nel caso in cui i dati sono di tipo MCAR e questo problema, come accennato inizialmente, è da attribuire al fatto che le imputazioni generate con questi metodi non tengono conto della componente di incertezza nella fase di imputazione. La mancanza di variabilità dei valori imputati produce una sottostima delle deviazioni standard.

2.3 Imputazione multipla

L'imputazione multipla si distingue dall'imputazione singola perché tiene debitamente conto anche dell'incertezza sulle imputazioni, conducendo così a risultati inferenziali validi non solo in termini di stima puntuale dei parametri di interesse, ma anche in termini, per esempio, di intervalli di confidenza.
Essa sostituisce ciascun dato mancante con un certo numero di valori plausibili, rappresentando in questo modo l'incertezza sul vero valore da imputare.
Come nel caso dell'imputazione singola, l'assunzione su cui si basano le procedure di imputazione multipla è quella di un processo generatore dei dati mancanti casuale (MAR).
L'idea di questo metodo è stata proposta da Rubin (1978) ed è quella di generare più di un valore ($m>2$) da imputare per ogni dato mancante in modo che i dataset completi, su cui effettuare le analisi statistiche di interesse, siano m.
I risultati delle m analisi vengono poi combinati con regole tali che il risultato inferenziale finale tenga conto dell'incertezza causata dalla presenza di dati mancanti, stimata dalla variabilità anche tra le m uscite indipendenti.
Il processo di combinazione dei risultati è sempre lo stesso, a prescindere dall'analisi effettuata sugli m dataset completi.
Il principale svantaggio del metodo è pertanto il costo computazionale molto oneroso, poiché, dopo aver imputato i dati, è necessario ripetere le analisi m volte e combinare poi i risultati.
Il processo di imputazione multipla si può quindi riassumere in tre fasi:

1) la fase di imputazione, che crea m dataset completi sostituendo i valori mancanti;
2) la fase di analisi degli m dataset creati;

3) la fase di pooling, che unifica le diverse stime ottenute per ricavare un unico valore di stima per ogni parametro di interesse.

2.3.1 Fase di imputazione

In questa fase vengono imputati i valori mancanti e lo si può fare utilizzando le diverse tecniche adottate per l'imputazione singola, ripetendole per *m* volte in modo da creare gli *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 da usare dipendono dalla natura della variabile da imputare: ad esempio, per variabili binarie la tecnica più usata prevede l'uso della regressione logistica, mentre i metodi più utilizzati per imputare i valori di variabili continue sono quelli già riportati nel paragrafo relativo all'imputazione singola: imputazione con la media, campionamento aleatorio, imputazione con regressione lineare, metodo PMM.

2.3.2 Fase di analisi e fase di pooling

Una volta ottenuti gli *m* dataset è possibile applicarvi le opportune analisi statistiche.

Supponendo che θ sia un parametro incognito di interesse, al termine della procedura di inferenza sugli *m* dataset sono disponibili *m* coppie di valori composte dalla stima puntuale del parametro di interesse $\hat{\theta}_i$ e dalla stima della varianza dello stimatore, $\widehat{U}_i$ (*i*=1, …, *m*).

La stima puntuale di θ è data dalla media delle singole stime calcolate sulle *m* matrici complete:

$$\bar{\theta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i$$

La varianza della stima puntuale è invece data dalla somma di una componente di variabilità entro l'imputazione e da una componente di variabilità tra imputazioni.

La varianza entro imputazione, $\overline{U}$, può essere stimata come media delle varianze $\widehat{U}_i$:

$$\overline{U} = \frac{1}{m}\sum_{i=1}^{m}\widehat{U}_i$$

La varianza tra imputazioni, *B*, è invece calcolata in accordo alla seguente espressione:

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i - \bar{\theta})^2$$

La varianza totale associata a $\bar{\theta}$ è quindi ottenuta combinando le due componenti (Rubin, 1987):

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B = \bar{U} + B + B/m\,,$$

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

Capitolo 3

Il dataset: questionario per la valutazione della didattica

La misura più diffusa della qualità dell'insegnamento nell'ambito universitario è il questionario sulla valutazione della didattica, il quale è uno strumento utile alle università per far emergere il punto di vista degli studenti rispetto alle attività didattiche erogate, favorire la riflessione dei docenti e dei Consigli di Corso di Studio e innalzare la qualità degli insegnamenti per un generale miglioramento dell'offerta formativa e dei servizi per la didattica.

Ogni ateneo pubblico italiano, infatti, secondo le indicazioni dell'ANVUR (Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca), elabora e somministra al termine del ciclo di lezioni di ogni attività didattica, obbligatoria o facoltativa, un proprio questionario al fine di raccogliere le opinioni degli studenti riguardo la didattica universitaria.

L'Ateneo cui si fa riferimento in questa tesi è quello di Padova e precisamente il dataset in esame riguarda l'anno accademico 2012/13 e comprende non solo le risposte al questionario sulla valutazione della didattica da parte degli studenti, ma anche informazioni generali circa gli studenti iscritti, docenti, dati riassuntivi sulle carriere degli studenti, offerta didattica e caratteristiche degli insegnamenti, grazie all'unione di diversi database.

Tutte le analisi del presente capitolo e dei successivi vengono realizzate con il software statistico Stata.

3.1 Descrizione del dataset e selezione del campione

Il dataset è composto da 253318 record, uno per ogni questionario chiuso (ossia aperto dallo studente e completato interamente, sia con compilazione che con rifiuto alla

compilazione), somministrato nell'A.A. 2012/13. Sono incluse anche le valutazioni di corsi on-line.
Ogni questionario, studente, attività didattica e docente sono identificati da un codice, sfruttato per agganciare i vari database (studenti iscritti, offerta didattica, esami, carriera complessiva degli studenti) e creare quello completo usato per l'analisi.

Gli studenti sono suddivisi per tipo di corso di laurea di iscrizione (corsi singoli, laurea a ciclo unico, laurea magistrale, laurea triennale) nelle percentuali osservabili in Figura 3.1.

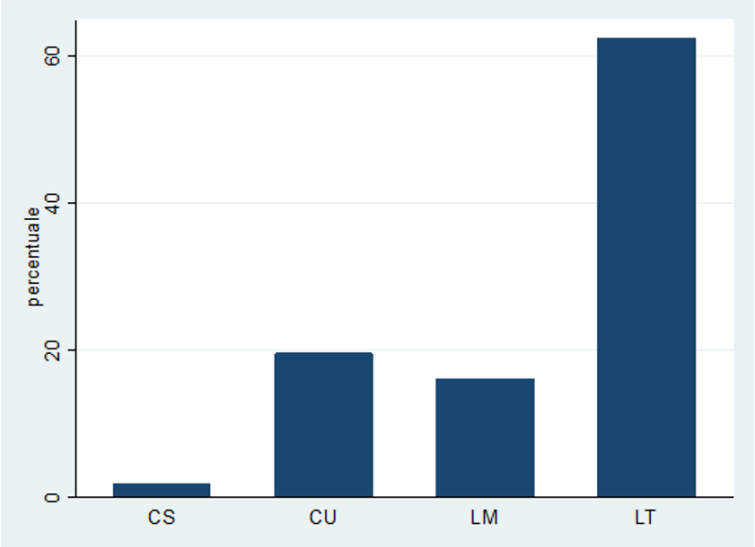

Figura 3.1: Percentuale studenti per tipo di corso di laurea di iscrizione

All'inizio del questionario vengono poste due domande filtro: la prima permette allo studente di decidere se procedere con la compilazione e la seconda chiede quale sia stata la frequenza alle lezioni dell'insegnamento a cui il questionario fa riferimento, per poter procedere con le successive domande.
Nel caso lo studente abbia frequentato meno del 30% delle lezioni, infatti, deve compilare solo una parte del questionario, composta da 7 domande, uguali per frequentanti e non frequentanti, più una domanda sul motivo della mancata frequenza; se lo studente dichiara di aver frequentato più del 30% delle lezioni, invece, può compilare l'intero questionario, composto da 18 domande (si veda Tabella 3.5).

Per quanto riguarda la prima domanda-filtro, i questionari compilati, da 42595 studenti diversi, sono 196103 (circa il 77% del totale), mentre 57215 sono stati chiusi con un rifiuto alla compilazione.

Per la seconda domanda-filtro, si riporta la Tabella 3.1 con le percentuali di frequenza alle lezioni per il diverso tipo di laurea a cui lo studente è iscritto.

Tabella 3.1: Distribuzione dei questionari compilati per percentuale di frequenza alle lezioni (%)

FREQUENZA ALLE LEZIONI	**TIPO LAUREA STUDENTE**				
	Corsi singoli	Laurea triennale	Laurea magistrale	Laurea ciclo unico	Totale
non frequentante	19.20	6.42	12.56	7.82	7.95
meno del 30%	6.33	3.01	2.77	2.33	2.91
tra il 30% e il 50%	9.44	4.81	4.24	3.39	4.54
tra il 50% e il 70%	18.92	11.31	11.38	10.00	11.23
più del 70%	46.11	74.44	69.06	76.46	73.37
Totale	100	100	100	100	100

Si osserva che, escludendo gli iscritti ai corsi singoli in quanto aventi caratteristiche a sé, la quota maggiore di non frequentanti si riscontra tra gli iscritti a lauree magistrali, mentre nelle lauree a ciclo unico vi è una quota maggiore di frequentanti (probabilmente per il fatto di avere l'obbligo di frequenza in certi corsi).

Nel dataset sono disponibili anche informazioni sul tipo di Corso di Laurea a cui appartiene l'insegnamento valutato: può accadere che uno studente inserisca nel proprio piano di studi un'attività didattica (AD) afferente ad un corso di laurea diverso da quello a cui è iscritto.

Tabella 3.2: Differenza tra Corso di Laurea dell'AD e dello studente

TIPO LAUREA AD	**TIPO LAUREA STUDENTE**					
	n.d.	Corsi singoli	Laurea triennale	Laurea magistrale	Laurea ciclo unico	Totale
n.d.	0	24	617	82	42	765
Laurea triennale	3	2527	**156659**	1555	437	161181
Laurea magistrale	7	1349	693	**39198**	98	41345
Laurea ciclo unico	1	689	185	23	**49129**	50027
Totale	11	4589	158154	40858	49706	253318

Dalla Tabella 3.2 risultano 244986 casi per cui il tipo di laurea dell'attività didattica corrisponde con quello dello studente, mentre per 7556 casi non c'è corrispondenza (circa il 3%). Per 776 casi invece non si può fare alcuna considerazione poiché 765 questionari sono relativi a corsi online e, nell'unione preliminare dei dataset, non hanno trovato corrispondenza con il database dell'offerta didattica; infine, ci sono 11 casi di errori di compilazione relativi a tre soggetti.

Andando ad osservare il numero di risposte date dagli studenti per ogni questionario si nota che questo va da un minimo di 0 ad un massimo di 18, ossia ci sono casi di questionari chiusi (da rispondenti) senza rispondere ad alcuna delle 18 domande.
In Tabella 3.3 si riporta la variabile che indica la risposta ad almeno una domanda del questionario in relazione a quella sul tipo di laurea di iscrizione (in cui sono presenti i già menzionati 11 valori mancanti riferiti ai questionari compilati erroneamente); si osserva una maggior frequenza di questionari senza alcuna risposta tra chi è iscritto a corsi singoli e a lauree a ciclo unico.

Tabella 3.3: Distribuzione di questionari con almeno una risposta per tipo di laurea

TIPO LAUREA	**Nessuna risposta**	**Almeno una risposta**	**Totale**
Corsi singoli	83 (2.38%)	3411 (97.62%)	3494
Laurea triennale	1962 (1.58%)	122475 (98.42%)	124437
Laurea magistrale	255 (0.76%)	33292 (99.24%)	33547
Laurea ciclo unico	819 (2.37%)	33795 (97.63%)	34614
Totale	**3119**	**192973**	**196092**

Nelle analisi successive si decide di ridurre il campione di questionari considerando soltanto quelli riferiti a Corsi di Studio afferenti a lauree triennali (per questioni di omogeneità del campione) ed escludendo quelli riconducibili a Corsi di Studio di tipo medico (aventi caratteristiche diverse e poco omogenee rispetto agli altri).
Vengono inoltre eliminati i 765 questionari relativi a corsi online, gli 11 compilati erroneamente, quelli chiusi da rispondenti senza aver risposto ad alcuna domanda e quelli aventi la stessa risposta a tutte domande, essendo probabilmente indice di poca sensibilità nel rispondere.
Inoltre, come verrà approfondito successivamente, la domanda del questionario di principale interesse per le analisi future è quella riferita alla soddisfazione complessiva

dello studente nei confronti dell'insegnamento frequentato e pertanto si è scelto di non tenere in considerazione i questionari in cui non viene fornita la risposta a questo item quando lo studente afferma di frequentare più del 30% delle lezioni; si tratta comunque di un numero di casi molto ridotto.
Dal dataset iniziale costituito da 253318 osservazioni si arriva dunque ad un campione di 112706 osservazioni.

3.1.1 Breve analisi del campione

Si presenta una breve analisi descrittiva del campione selezionato nella sua totalità, senza tener conto della suddivisione in rispondenti e non rispondenti, concentrandosi su caratteristiche di studenti e docenti.

Innanzitutto gli studenti che hanno aperto almeno un questionario (compilando ciascuno da 1 a 20 questionari diversi) sono 22002, di cui il 55% di genere femminile, con età media di circa 21 anni.
Il 30% risiede nella provincia di Padova, il 17.5% nella provincia di Vicenza, il 13% in quella di Treviso, l'11% in quella di Venezia e il 7% circa in quella di Verona; le restanti province di residenza hanno una percentuale di studenti sotto il 3%.
Per quanto riguarda i Corsi di Studio triennali, dalla Figura 3.2 si può osservare che quelli con il maggior numero di iscritti fanno riferimento ad Ingegneria, quindi a seguire a Scienze Matematiche/Fisiche/Naturali, Psicologia e Lettere e Filosofia.
Il 90% degli studenti inoltre risulta essere in corso.

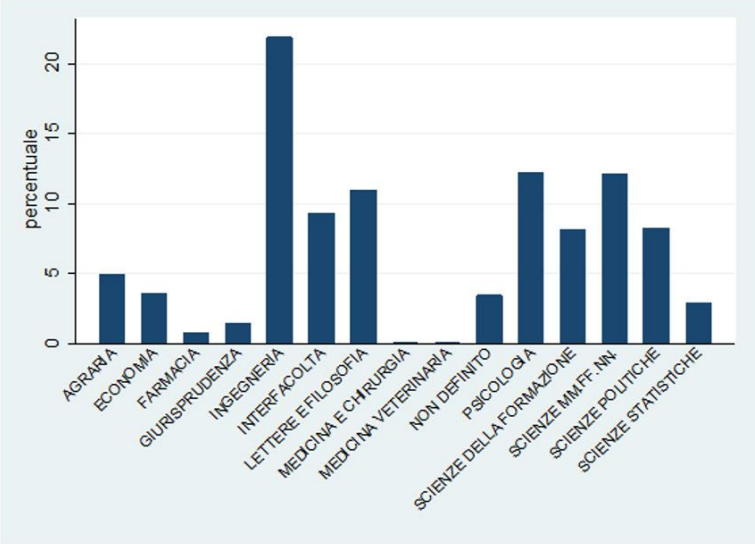

Figura 3.2: Percentuale studenti iscritti per facoltà

Per quanto riguarda i docenti, nel dataset vengono presi in considerazione non solo i titolari dei vari insegnamenti ma anche tutti quei docenti che erogano almeno 15 ore di lezione all'interno del corso.
In totale sono coinvolti 1562 docenti, di cui il 67% di genere maschile, con un'età media di 51 anni. Per quanto riguarda la loro posizione all'interno dell'università nell'anno accademico in esame, dalla Figura 3.3 si osserva che il 33% circa aveva il ruolo di ricercatore universitario, mentre il 29% di professore associato. Per coloro con ruolo non assegnato si presume si intendano collaboratori esterni, legati ad aziende o istituti di ricerca e che, per vari motivi, non rientrano nel ruolo di professore a contratto.

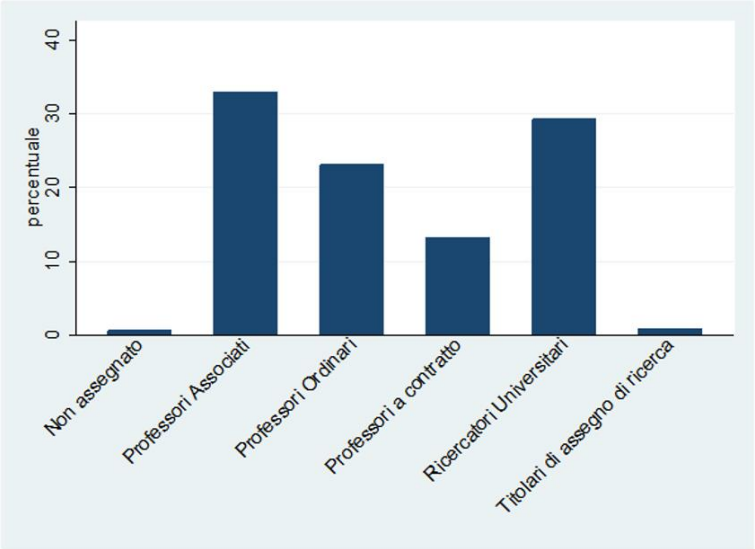

Figura 3.3: Percentuale docenti per ruolo accademico

Mediamente un docente svolge 54 ore per un insegnamento, ma molto dipende dal suo ruolo e dal tipo di corso in quanto, come già affermato, possono esserci corsi divisi in più parti, tenute da diversi docenti.
Dalla suddivisione in base al ruolo del docente risulta comunque che i professori associati e ordinari sono quelli che svolgono il maggior numero di ore in media; i titolari di assegni di ricerca invece ne svolgono meno (Tabella 3.4).

Tabella 3.4: Numero medio di ore insegnate per ruolo del docente

RUOLO DOCENTE	MEDIA ORE DOCENTE
Non assegnato	40
Professori associati	58
Professori ordinari	57
Professori a contratto	48
Ricercatori universitari	53
Titolari di assegno di ricerca	23

3.2 Analisi del questionario sulla valutazione della didattica

Si osserva innanzitutto che, nel campione selezionato, i questionari con risposta sono 89489, ossia circa l'80% del totale, mentre 23217 sono quelli consegnati dai non rispondenti (4263 studenti), a cui viene chiesto di fornire il motivo della non risposta tra le seguenti possibilità:

- perché in questo momento non ho tempo;
- perché ne ho già compilati troppi;
- perché non sono in grado di giudicare;
- perché ne ho già compilato uno per questo insegnamento;
- perché è inutile;
- perché non ci credo;
- perché il sistema web non funziona;
- perché non mi fido;
- perché i dati su insegnamento e docente non sono corretti;
- altro (risposta aperta).

Risulta che per l'8% circa dei casi non è stata data una motivazione per la non risposta, mentre tra i restanti la motivazione più usata è quella relativa al non funzionamento del sistema web.

Le 18 domande rivolte ai rispondenti sono riportate in Tabella 3.5 e riguardano aspetti organizzativi, soddisfazione complessiva, efficacia dell'attività didattica, contenuti e programma dei corsi, carico di lavoro percepito.

Tabella 3.5: Domande questionario valutazione della didattica

Domande questionario valutazione della didattica
1. All'inizio del corso gli obiettivi e i contenuti sono stati definiti in modo chiaro?
2. Le modalità d'esame sono state definite in modo chiaro?
3. Gli orari delle lezioni sono stati rispettati?
4. Le ore previste sono in numero adeguato per lo svolgimento del programma?
5. Le conoscenze preliminari possedute sono sufficienti per la comprensione degli argomenti?
6. Il docente stimola/motiva l'interesse verso la disciplina?
7. Il docente espone gli argomenti in modo chiaro?
8. Il materiale didattico consigliato è adeguato?
9. Il docente è disponibile nei confronti delle esigenze degli studenti?
10. Il docente è stato reperibile durante gli orari di ricevimento?
11. Esercitazioni/laboratori/seminari, se previsti, sono adeguati?
12. Le aule in cui si svolgono le lezioni sono adeguate?
13. I locali e le attrezzature per i laboratori sono adeguati?
14. Complessivamente quanto è soddisfatto del corso?
15. Il carico di studio richiesto è equilibrato rispetto al numero di CFU assegnati?
16. Indipendentemente da come si è svolto il corso, quanto era interessato ai suoi contenuti?
17. Quanto ritiene coerente l'insegnamento rispetto agli obiettivi del corso di studi?
18. Ritiene che l'insegnamento fornisca competenze adeguate in ambito lavorativo?

Si precisa che lo studente può rispondere ai quesiti utilizzando un punteggio su una scala da 1 a 10, dove 10 corrisponde alla situazione ottimale; in alcuni casi è presente l'opzione "non so/non pertinente".

Tra gli 89489 questionari di rispondenti, circa il 10% è riferito a studenti frequentanti meno del 30% delle lezioni e a costoro vengono rivolte soltanto le domande 8, 9, 10, 15, 16, 17, 18, in aggiunta a quella sul motivo della non frequenza.

Le possibili risposte a quest'ultima domanda sono:

- avevo già frequentato in precedenza;
- motivi di lavoro;
- l'orario delle lezioni era sovrapposto a quello di un altro insegnamento;
- ho perso l'interesse dopo le prime lezioni;
- ritengo non indispensabile la frequenza;
- altro.

Le principali risposte fornite sono motivi di lavoro (per il 33% dei casi), la sovrapposizione delle lezioni (16%) e la perdita di interesse, per il 14% circa; per il 21% dei casi la risposta risulta invece essere “altro”. Le restanti opzioni presentano percentuali molto basse.

3.2.1 Caratteristiche dei rispondenti e dei non rispondenti al questionario

Andando ad osservare le caratteristiche dei non rispondenti al questionario, si nota che il 51% è di genere femminile; essi sono principalmente iscritti ai Corsi di Studio di Lettere e Filosofia, Scienze Politiche e Scienze della Formazione.

Nella Tabella 3.6 si nota inoltre che circa il 44% dei questionari senza risposta sono legati a studenti iscritti al secondo anno e questo può essere dovuto al fatto che gli studenti del primo anno tendono ad essere più precisi, mentre al terzo anno arrivano quelli più motivati; il secondo anno presenta invece varie problematiche legate ad una sorta di selezione degli studenti, che porta poi ad una scrematura.

Tabella 3.6: Distribuzione studenti non rispondenti per anno di iscrizione

ANNO ISCRIZIONE	Freq.	%
1	7,828	33.72
2	10,202	43.94
3	5,187	22.34
Totale	23,217	100

Gli studenti rispondenti sono 17739, di cui il 55% di genere femminile, con età media di 20 anni.

Nella Figura 3.4 si riporta la suddivisione dei questionari compilati da questi studenti in base alla frequenza alle lezioni dell’insegnamento valutato.

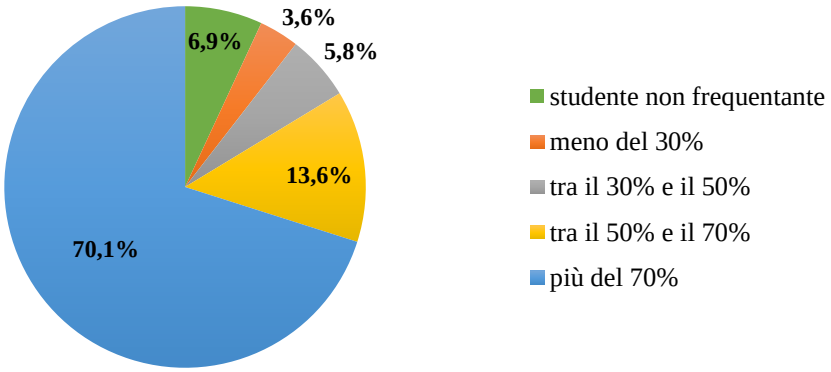

Figura 3.4: Distribuzione questionari compilati per frequenza delle lezioni

Si nota che circa il 70% dei questionari è stato compilato da studenti frequentanti più del 70% delle lezioni dell'attività didattica valutata, mentre l'11% circa è riferito a chi frequenta meno del 30% (compresi non frequentanti).

I Corsi di Studio che presentano il maggior numero di non frequentanti sono riconducibili a Giurisprudenza e a seguire Scienze Politiche e Psicologia, ed emerge anche come il numero di studenti non frequentanti si presenti in proporzione maggiore tra gli studenti fuori corso; questi infatti possono aver già seguito un insegnamento in precedenza senza averlo valutato ed essersi iscritti all'esame.

Maggior frequenza invece si riscontra per Scienze MM.FF.NN., Ingegneria e Scienze Statistiche.

Per quanto riguarda l'anno di corso in cui viene erogato l'insegnamento, la maggior parte di non frequentanti è presente tra i corsi del terzo anno, mentre i corsi del primo anno sono i più frequentati.

3.2.2 Analisi descrittive dei 18 item

Nella Tabella 3.7 vengono riportate alcune statistiche descrittive relative ai 18 item del questionario, con la percentuale di dati mancanti rispetto al campione totale; si osserva che gli item D10, D11 e D13 sono quelli che presentano la maggiore quantità di mancate risposte ed è comprensibile poiché sono domande a cui può rispondere solo chi ha direttamente sperimentato quanto richiesto, come il ricevimento del docente ed esercitazioni/laboratori/seminari.

Le domande con la percentuale più piccola di NA sono invece alcune di quelle rivolte anche ai non frequentanti: D08 (adeguatezza del materiale didattico), D09 (disponibilità del docente), D15 (carico di studio equilibrato rispetto ai CFU assegnati), D16 (interesse dello studente nei confronti dei contenuti del corso), D17 (coerenza dell'insegnamento rispetto agli obiettivi del Corso di Studi).

Gli item che hanno ricevuto mediamente i punteggi più bassi sono relativi alle conoscenze preliminari possedute per poter comprendere gli argomenti (D05), ritenute dall'11% di chi ha risposto non sufficienti (punteggio inferiore a 5), e al carico di studio richiesto rispetto al numero di CFU assegnati all'insegnamento (D15).

I punteggi più alti, superiori o uguali ad 8 in media, sono assegnati agli item D3, D10 e D17, inerenti rispettivamente al rispetto degli orari delle lezioni, alla reperibilità del docente durante gli orari di ricevimento e alla coerenza dell'insegnamento rispetto agli obiettivi del Corso di Studi.

Tabella 3.7: Statistiche descrittive dei 18 item

ITEM	NUMERO RISPOSTE	% NA	MEDIA	DEV. STD.
D01 OBIETTIVI	79030	29.88	7.78	1.815
D02 MODALITÀ	79042	29.87	7.90	1.885
D03 ORARI	79692	29.29	8.35	1.727
D04 NUMERO ORE	72723	35.48	7.58	1.938
D05 PRECONOSCENZE	79729	29.26	7.07	2.071
D06 STIMOLO	79752	29.24	7.29	2.191
D07 CHIAREZZA	79720	29.27	7.40	2.138
D08 MATERIALE	88903	21.12	7.28	2.034
D09 DISPONIBILE	88539	21.44	7.96	1.883
D10 RICEVIMENTO	38554	65.79	8.08	1.856
D11 LABORATORI	46992	58.31	7.54	2.000
D12 AULE	79770	29.22	7.35	2.127
D13 ATTREZZATURE	47207	58.11	7.40	2.044
D14 SODDISFAZIONE	80093	28.94	7.32	1.961
D15 CARICO	88998	21.04	7.19	2.032
D16 INTERESSE	89085	20.96	7.72	1.968
D17 COERENTE	86961	22.84	8.00	1.914
D18 UTILE	80854	28.26	7.38	2.056

Si riportano infine, in Tabella 3.8, le statistiche descrittive distinte per livello di frequenza alle lezioni e si nota che, all'aumentare della frequenza alle lezioni, aumentano i punteggi medi di tutti gli item.

Osservando i valori della deviazione standard si può dire che i giudizi dei non frequentanti, e più marcatamente di chi afferma di aver frequentato meno del 30% delle lezioni, sono i più eterogenei (valori più alti di deviazione) e questo può essere dovuto alla maggior eterogeneità dei componenti di questo gruppo rispetto agli altri (diverse motivazioni della non frequenza).

	non frequenta			frequenza<30%			30%<frequenza<50%			50%<frequenza<70%			frequenza>70%		
	% risp	media	dev. std.	% risp	media	dev. std.	% risp	media	dev. std.	% risp	media	dev. std.	% risp	media	dev. std.
D01 obiettivi							96.7%	6.93	2.022	98.1%	7.35	1.852	98.9%	7.93	1.757
D02 modalità							96.9%	7.18	2.065	98.4%	7.54	1.909	98.9%	8.03	1.842
D03 orari							99.3%	7.74	1.857	99.3%	7.99	1.773	99.6%	8.47	1.686
D04 numero ore							83.9%	7.09	1.957	89.2%	7.34	1.915	91.6%	7.66	1.932
D05 preconoscenze							99.5%	6.60	2.131	99.4%	6.80	2.086	99.6%	7.16	2.054
D06 stimolo							99.4%	6.25	2.342	99.5%	6.75	2.228	99.6%	7.48	2.128
D07 chiarezza							99.4%	6.52	2.253	99.5%	6.94	2.159	99.5%	7.57	2.093
D08 materiale	99.1%	7.01	1.891	99.4%	6.47	2.114	99.4%	6.71	2.068	99.3%	6.98	2.052	99.4%	7.46	2.008
D09 disponibile	95.6%	7.21	2.024	98.1%	6.84	2.173	99.2%	7.33	1.922	99.1%	7.66	1.871	99.2%	8.20	1.791
D10 ricevimento	41.9%	7.50	2.122	38.1%	7.07	2.193	45.4%	7.41	1.875	46.0%	7.72	1.807	42.8%	8.31	1.764
D11 laboratori							55.4%	6.94	2.041	58.6%	7.25	2.001	59.0%	7.64	1.983
D12 aule							99.6%	7.11	2.118	99.5%	7.19	2.134	99.6%	7.40	2.123
D13 attrezzature							58.7%	7.03	2.046	60.2%	7.20	2.001	58.7%	7.47	2.047
D14 soddisfazione							100.0%	6.27	2.116	100.0%	6.79	2.001	100.0%	7.51	1.892
D15 carico	98.6%	6.84	1.981	98.8%	6.61	2.054	99.4%	6.69	2.019	99.5%	6.95	1.982	99.6%	7.34	2.026
D16 interesse	98.5%	7.18	2.115	99.0%	6.46	2.371	99.6%	6.91	2.155	99.6%	7.31	2.024	99.7%	7.99	1.830
D17 coerente	91.6%	7.52	2.071	93.6%	7.15	2.271	97.2%	7.35	2.079	97.4%	7.63	1.981	97.9%	8.22	1.807
D18 utile	84.5%	6.92	2.148	87.3%	6.57	2.340	90.6%	6.87	2.142	91.2%	7.14	2.085	90.9%	7.56	1.991

Tabella 3.8: Punteggio medio, percentuale di rispondenti e deviazione standard di ciascun item, distinti per percentuale di frequenza alle lezioni dell'insegnamento valutato

3.2.3 Relazione tra item D14 e variabili esplicative

Esiste una grande letteratura che studia le determinanti osservabili delle valutazioni degli studenti, la quale ha trovato diversi fattori osservati correlati con i punteggi delle valutazioni (Marsh, 2007): il principale è il livello di rendimento degli studenti.
Gli studenti più bravi, come dimostrano i voti, tendono a valutare i corsi in modo più positivo; questo si può spiegare con il fatto che l'impressione che gli studenti hanno della qualità dell'insegnamento dovrebbe migliorare quando essi apprendono di più, e i voti sono un indicatore di apprendimento.
Il livello di rendimento inoltre può riflettere delle differenze preesistenti tra gli studenti.
Anche caratteristiche dei corsi, dei docenti e loro carriera ed altre caratteristiche osservabili degli studenti sembrano influenzare le valutazioni dei corsi, sebbene con effetti diversi (Marsh, 2007).
Questi fattori osservabili possono però spiegare solo una parte della variabilità dei punteggi delle valutazioni, in quanto esistono anche dei fattori non osservabili che li influenzano. Ad esempio, motivazioni intrinseche degli studenti possono avere un impatto, poiché è più probabile che gli studenti più motivati percepiscano la loro esperienza educativa come gratificante e quindi come più alta la qualità dell'insegnamento.

In base alle variabili disponibili nel dataset, l'interesse è proprio quello di capire quali tra esse, e con quale effetto, abbiano un legame con il punteggio assegnato dagli studenti alla domanda relativa alla soddisfazione complessiva dell'insegnamento seguito, ossia l'item D14.
Partendo con l'analizzare i punteggi assegnati a questa domanda, dalla Figura 3.5 si osserva che il punteggio 8 è quello che si presenta con maggior frequenza; la media è di 7.3 con una deviazione standard di 1.96; in generale si può quindi dire che il grado di soddisfazione degli studenti dell'Ateneo di Padova, senza distinzioni per Corso di Studio e frequenza, sia buono.

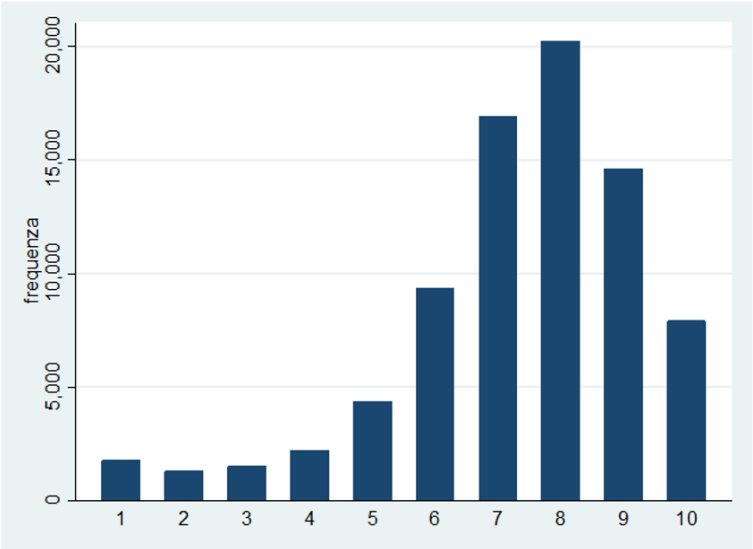

Figura 3.5: Istogramma con punteggi assegnati all'item D14

Le variabili che possono avere une relazione con la valutazione della soddisfazione vengono suddivise in:

- caratteristiche generali degli studenti: frequenza del corso, genere, età;
- carriera degli studenti: stato di iscrizione (in corso/fuori corso), media dei voti nell'A.A. 2012/13, numero medio di esami passati all'anno (nell'intera carriera), numero medio di CFU registrati all'anno (nell'intera carriera);
- caratteristiche del corso: corso obbligatorio, numero totale di ore, numero docenti coinvolti (uno o più di uno), locazione del corso (Padova o altra sede), insegnamento mutuato;
- caratteristiche generali dei docenti: genere, età;
- carriera dei docenti: posizione accademica (ruolo).

Quello che si vuole fare preliminarmente in questa sezione è vedere se sono presenti evidenti differenze tra le valutazioni al variare del valore assunto da alcune delle variabili sopracitate; questo viene ottenuto principalmente per mezzo di box plot.
Si precisa che, solo nell'ambito di queste analisi descrittive, le variabili continue sono state trasformate in categoriali tramite una suddivisione dei loro valori in classi, per facilitarne l'interpretazione.
A supporto dell'analisi grafica, si riportano i valori delle medie del punteggio dell'item D14 per i gruppi definiti dalle categorie delle variabili.

Infine, per le variabili dicotomiche viene riportato il risultato del 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. Per le variabili categoriali si applica invece il test di Kruscal-Wallis. Questi test, basati sui ranghi, sono un'alternativa non parametrica dei test t e ANOVA (analisi della varianza ad una via), in quanto i dati non risultano avere distribuzione Normale; quest'ultima ipotesi è stata verificata tramite il test di Shapiro-Wilk, ma la non normalità era già evidente osservando l'asimmetria nel box plot ed i valori di media e mediana.
Data l'elevata dimensione del dataset comunque è emerso come i risultati ottenuti con i test di Wilcoxon e Kruscal-Wallis portino alle stesse conclusioni dei test t e ANOVA.

Caratteristiche generali degli studenti

Nella Figura 3.6 si può notare come i giudizi di studenti di genere maschile e femminile non sembrano mostrare evidenti differenze, mentre la variabile sull'età, come preannunciato, è stata resa categoriale tramite una suddivisione in cinque classi: 17-19, 20, 21, 22, 23-78 anni. Questa divisione dovrebbe rispecchiare l'anno di iscrizione dello studente; le prime tre risultano quindi essere più numerose.
Sia dalla Tabella 3.9 che dalla Figura 3.7 emerge che le classi più giovani, ossia fino ai 22 anni, sono quelle che assegnano punteggi più bassi alla soddisfazione, senza palesi differenze tra loro; con la quinta classe invece emerge una differenza più evidente, soprattutto per quanto riguarda il valor medio.
Per quanto riguarda la frequenza, come già visto precedentemente, si nota un aumento di soddisfazione tra i frequentanti più del 70% delle lezioni (media e mediana sono superiori, come si osserva in Tabella 3.9 e Figura 3.8).

Tabella 3.9: Medie item D14 per caratteristiche generali degli studenti

STUDENTE - GENERALE	Variabile	Media
genere	maschio	7.28
	femmina	7.35
età	17-19	7.25
	20	7.28
	21	7.33
	22	7.38
	23+	7.66
frequenza	più del 70%	7.51
	tra il 30 e il 50%	6.26
	tra il 50 e il 70%	6.79

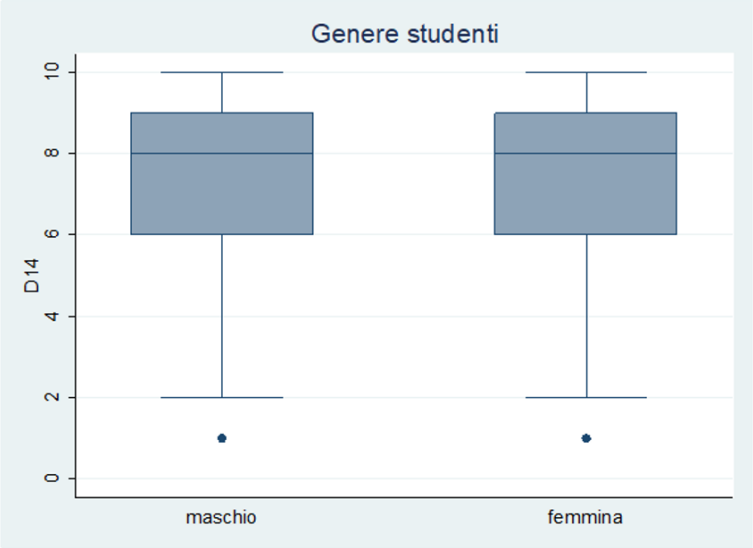

Figura 3.6: Box plot della variabile D14 al variare del genere dello studente

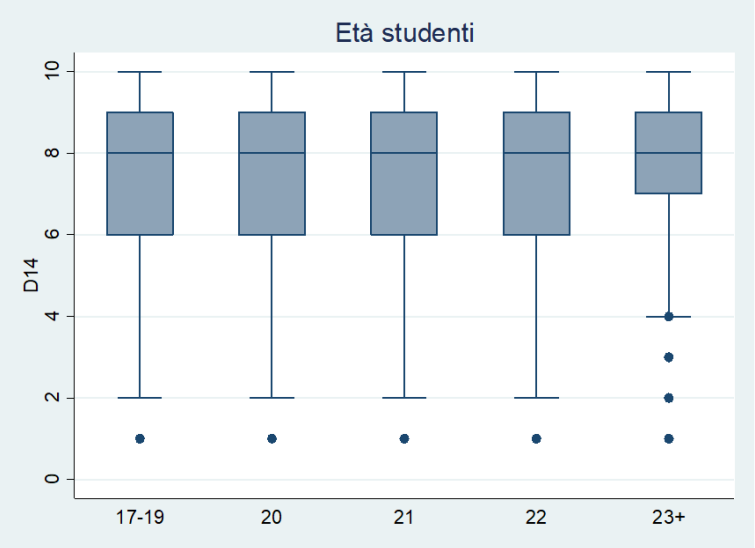

Figura 3.7: Box plot della variabile D14 al variare dell'età dello studente

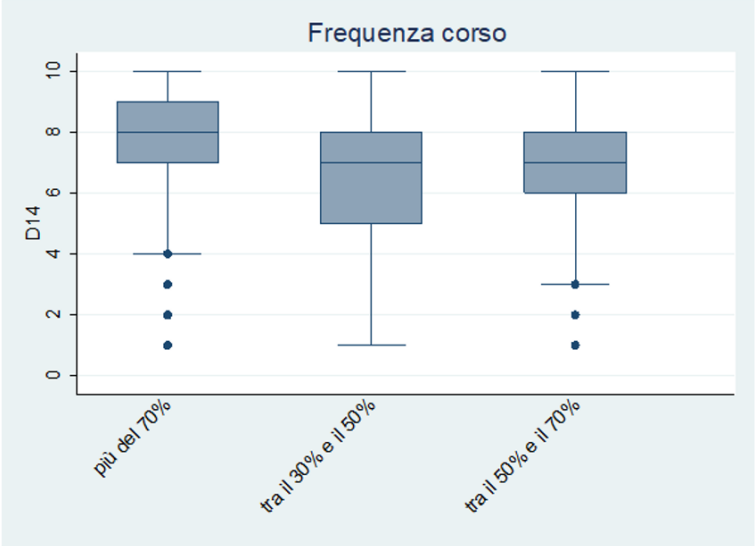

Figura 3.8: Box plot della variabile D14 al variare della frequenza alle lezioni

Gli esiti del test di Wilcoxon e Kruscal-Wallis, riportati in Tabella 3.10, confermano la presenza di una differenza statisticamente significativa tra le distribuzioni delle valutazioni nei gruppi definiti dalle variabili.

Tabella 3.10: Test di Wilcoxon e Kruscal-Wallis per caratteristiche generali degli studenti

		Wilcoxon		**Kruscal-Wallis**	
		z	p-value	χ^2	p-value
STUDENTE - GENERALE	genere	-5.495	0.000	-	-
	età	-	-	262.756	0.000
	frequenza	-	-	3276.472	0.000

Carriera degli studenti

Sulla carriera degli studenti si tengono in considerazione le variabili relative allo stato di iscrizione (in corso o fuori corso), alla media dei voti nell'A.A. 2012/13 e ai CFU registrati mediamente in un anno (nell'intera carriera), quindi il loro rendimento.

Per quanto riguarda la prima variabile, si consideri che per studenti in corso si intendono tutti quelli iscritti regolarmente ai tre anni, mentre tra gli studenti fuori corso sono compresi sia i "fuori corso" effettivi, ossia chi non si è laureato entro il terzo anno di iscrizione e quindi risulta ancora iscritto al terzo anno, ma anche chi è ripetente dei primi due anni.

Dalla Tabella 3.11 e dalla Figura 3.9 si osserva che i punteggi degli studenti "in corso" sembrano inferiori rispetto a quelli assegnati da studenti "fuori corso", probabilmente per il fatto che i secondi potrebbero aver compilato il questionario dopo aver riseguito qualche insegnamento e quindi potrebbero ritenersi più soddisfatti rispetto alla volta precedente, avendo anche maggiori conoscenze.

La variabile relativa alla media dei voti registrati è stata resa categoriale con la costruzione di tre classi: [18, 23), [23, 27), [27, 30], che si possono intendere come media "bassa", "media", "alta".

Per questa variabile vi sono dei valori mancanti che si riferiscono a coloro che non hanno registrato alcun esame nell'A.A. 2012/13 o a chi ha registrato esami che hanno come valutazione finale una prova senza voto.

In Figura 3.10 si nota come all'aumentare della media dei voti aumenti la soddisfazione degli studenti; questo a conferma di quanto già affermato, ossia che i voti sono indicatore di apprendimento e che quanto più uno studente riesce ad apprendere tanto più avrà una buona impressione della qualità dell'insegnamento. Anche i valori delle medie dei punteggi nelle tre classi, in Tabella 3.11, lo ribadiscono.

Infine, una laurea triennale normalmente prevede che lo studente raggiunga la quota di 180 CFU per conseguire il diploma e quindi mediamente uno studente dovrebbe acquisire 60 CFU all'anno con un numero variabile di esami (ci sono esami da pochi crediti e altri da molti crediti).

È stata pertanto creata una variabile che indica il numero di CFU acquisiti mediamente in un anno dallo studente e per semplicità quest'ultima è stata divisa in classi: 0-30, 30-60, più di 60. Nella Figura 3.11 si osserva che chi registra meno di 30 crediti in un anno risulta

meno soddisfatto rispetto a chi ne registra di più. Tra la seconda e terza classe non emergono differenze: nella Tabella 3.11 infatti le medie di quest'ultime sono uguali.

Tabella 3.11: Medie item D14 per caratteristiche della carriera dello studente

CARRIERA STUDENTE	Variabile	Media
stato iscrizione	in corso	7.31
	fuori corso	7.64
media voti A.A. 2012/13	bassa	7.15
	media	7.29
	alta	7.48
crediti annui	0-30	7.23
	31-60	7.33
	60+	7.33

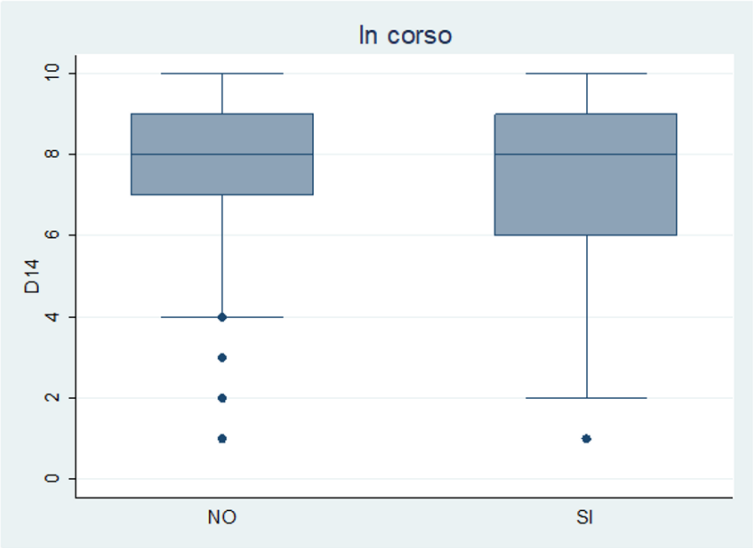

Figura 3.9: Box plot della variabile D14 al variare dello stato di iscrizione dello studente

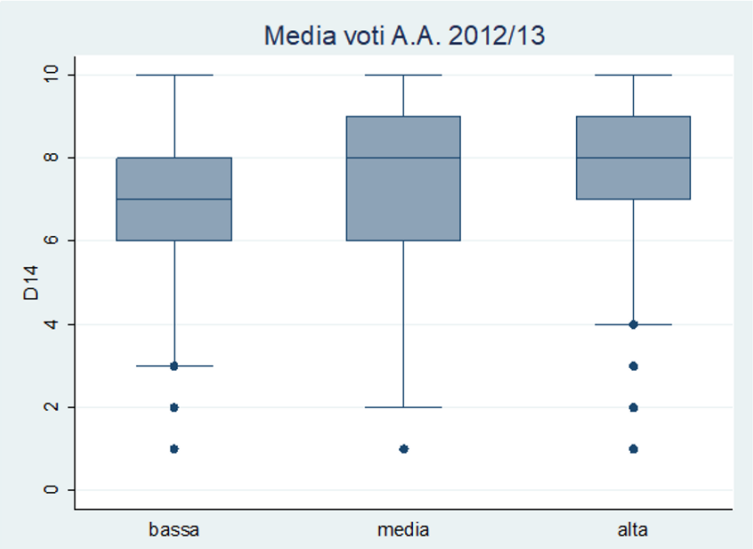

Figura 3.10: Box plot della variabile D14 al variare della media dei voti dello studente

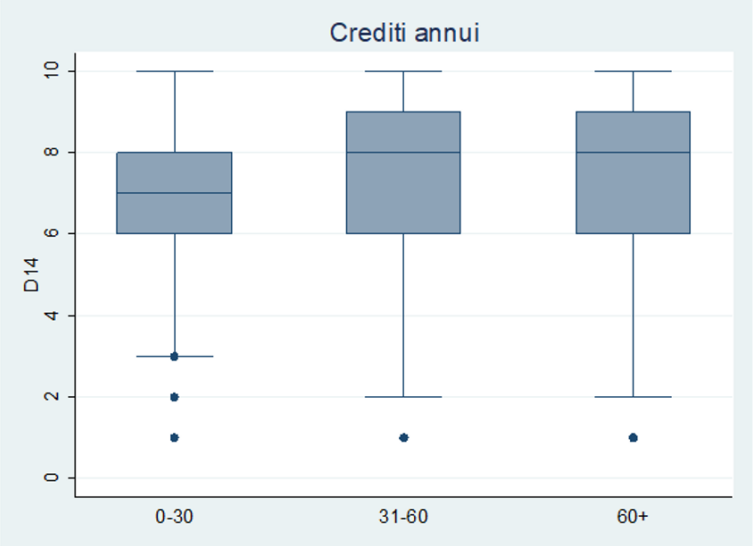

Figura 3.11: Box plot della variabile D14 al variare dei CFU registrati

I test di Wilcoxon e Kruscal-Wallis in Tabella 3.12 rifiutano l'ipotesi nulla di uguaglianza delle distribuzioni delle valutazioni nei gruppi definiti dalle variabili in esame.

Tabella 3.12: Test di Wilcoxon e Kruscal-Wallis per caratteristiche della carriera dello studente

		Wilcoxon		**Kruscal-Wallis**	
		z	p-value	χ^2	p-value
CARRIERA STUDENTE	stato iscrizione	8.233	0.000	-	-
	media voti A.A. 2012/13	-	-	464.77	0.000
	crediti annui	-	-	39.726	0.000

Caratteristiche del corso

Per quanto riguarda le caratteristiche dei corsi, dalla Figura 3.12 sembra che ai corsi obbligatori vengano assegnati punteggi leggermente più bassi, in quanto si suppone che un corso opzionale, scelto dallo studente, possa soddisfare di più rispetto ad uno "imposto". La media in Tabella 3.13 riporta una differenza minima.

La sede dove si svolge il corso invece non sembra avere effetti sui punteggi dati all'item.

Altra caratteristica dei corsi di studio è la possibilità di compresenza di classi di studenti diverse, dovuta a corsi mutuati, e, osservando la Figura 3.14 e la media in Tabella 3.13, sembra emergere una maggior soddisfazione degli studenti nei confronti dei corsi non mutuati.

In particolare, andando a guardare i punteggi medi (non riportati) dati anche agli altri item, maggior apprezzamento si riscontra nelle domande riferite all'adeguatezza di esercitazioni/laboratori/seminari e delle aule, poiché si presume che i corsi mutuati coinvolgano molti più studenti e quindi possa risultare più difficile seguire in maniera adeguata esercitazioni e altre attività e le aule possano essere non idonee.

Un corso, infine, può anche essere tenuto da più di un docente ed è stata creata una variabile che indica che il 28% circa dei corsi (su un totale di 2920) è tenuto da più docenti: emerge una netta maggior soddisfazione nei confronti di corsi legati ad un unico docente (Figura 3.15).

Tabella 3.13: Medie item D14 per caratteristiche del corso

CORSO	Variabile	Media
corso obbligatorio	SI	7.25
	NO	7.36
locazione	Padova	7.32
	altra sede	7.30
insegnamento mutuato	SI	7.24
	NO	7.37
numero docenti coinvolti	uno	7.39
	più di uno	7.10

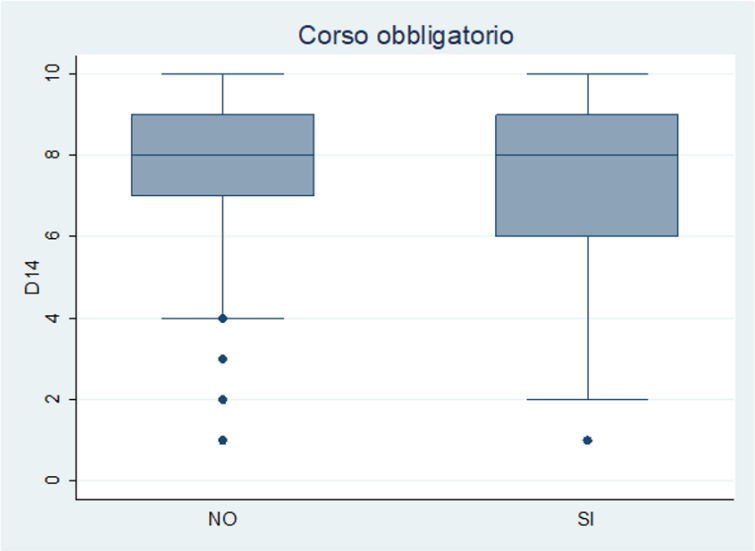

Figura 3.12: Box plot della variabile D14 al variare del tipo di corso: obbligatorio o meno

Figura 3.13: Box plot della variabile D14 al variare della sede del corso

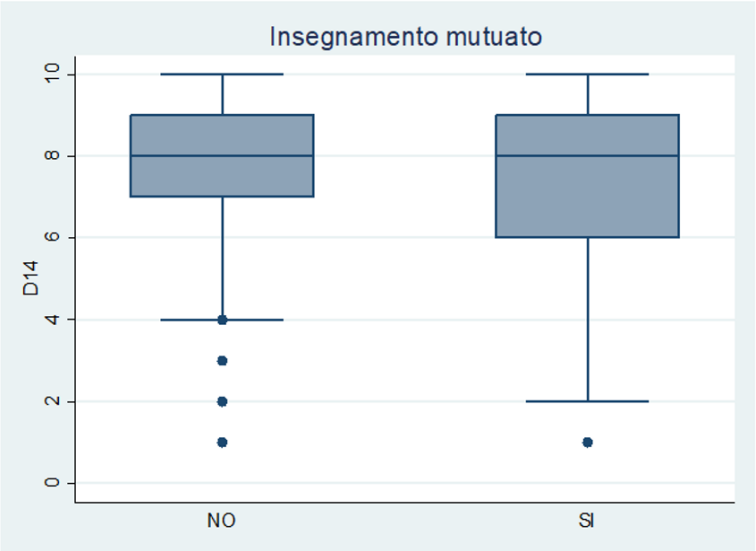

Figura 3.14: Box plot della variabile D14 al variare del tipo di corso: mutuato o meno

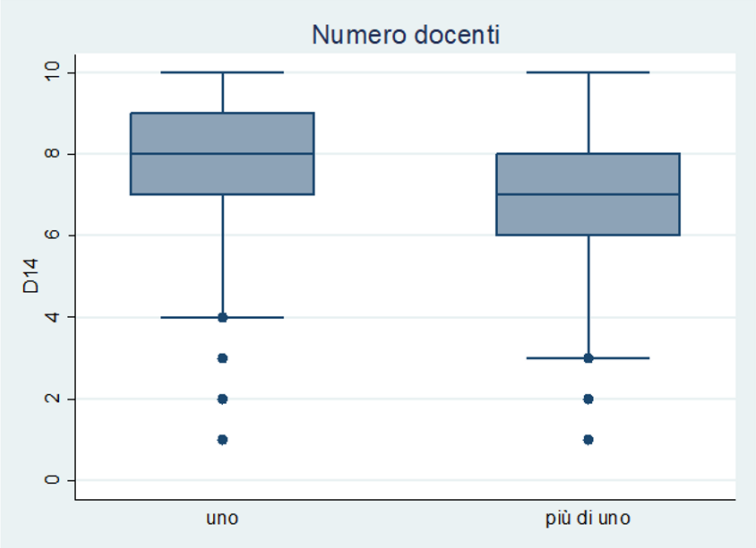

Figura 3.15: Box plot della variabile D14 in base al numero di docenti coinvolti (uno o più di uno)

Gli esiti del test di Wilcoxon, riportati in Tabella 3.14, permettono di affermare che vi è una differenza statisticamente significativa tra le distribuzioni delle valutazioni nei gruppi definiti dalle varie variabili, avendo ottenuto dei p-value pari a zero.

Tabella 3.14: Test di Wilcoxon per caratteristiche del corso

		Test Wilcoxon	
		z	p-value
CORSO	corso obbligatorio	8.515	0.000
	sede Padova	-4.397	0.000
	corso condiviso	11.069	0.000
	più docenti	18.971	0.000

Caratteristiche dei docenti

Infine si osserva la relazione della variabile di interesse con le caratteristiche dei docenti, ossia genere, età e ruolo accademico, per capire se la tipologia di docente ha effetto sul punteggio dato dagli studenti alla soddisfazione del rispettivo insegnamento.

Innanzitutto dalla Figura 3.16 emerge come le valutazioni dei corsi tenuti da docenti di genere femminile sembrino essere leggermente più alte e meno variabili rispetto a quelle di corsi tenuti da docenti di genere maschile. Le medie in Tabella 3.15 non presentano però una marcata differenza.

Anche l'età dei docenti è stata suddivisa in tre classi: 30-45, 46-60, 61+ anni.

Dal rispettivo box plot in Figura 3.17 si nota che i docenti con età più giovane (prime due classi) sono legati a corsi che hanno ottenuto punteggi più alti rispetto ai corsi tenuti da docenti più anziani.

Andando ad osservare anche il valore medio delle valutazioni, risulta più alto quello della prima classe di età (Tabella 3.15).

Tutto ciò si rispecchia nell'analisi della relazione con il ruolo del docente, da cui, dalla Figura 3.18 e dalla Tabella 3.15, emerge che i ricercatori universitari sono i docenti che sembrano legati a corsi con punteggi più alti, a seguire i professori associati e ordinari e ciò può essere collegato al fatto che, come emerso da ulteriori analisi, il 68% dei ricercatori risultano appartenere alla classe di età 30-45 e il 30% alla classe 46-60, cioè quelle che hanno riscontrato punteggi più alti.

La relazione tra il ruolo dei docenti e la soddisfazione quindi potrebbe dipendere dall'età. Altra motivazione connessa all'età però potrebbe essere relativa al fatto che i docenti più giovani sono soprattutto coinvolti in corsi opzionali.

Infine, i titolari di assegni di ricerca hanno i punteggi più bassi, probabilmente per la poca esperienza.

Tabella 3.15: Medie item D14 per caratteristiche dei docenti

DOCENTE	Variabile	Media
genere	maschio	7.30
	femmina	7.35
età	30-45	7.44
	46-60	7.35
	61+	7.03
ruolo	Professori associati	7.31
	Professori ordinari	7.27
	Professori a contratto	7.19
	Ricercatori universitari	7.45
	Titolari di assegno di ricerca	6.64

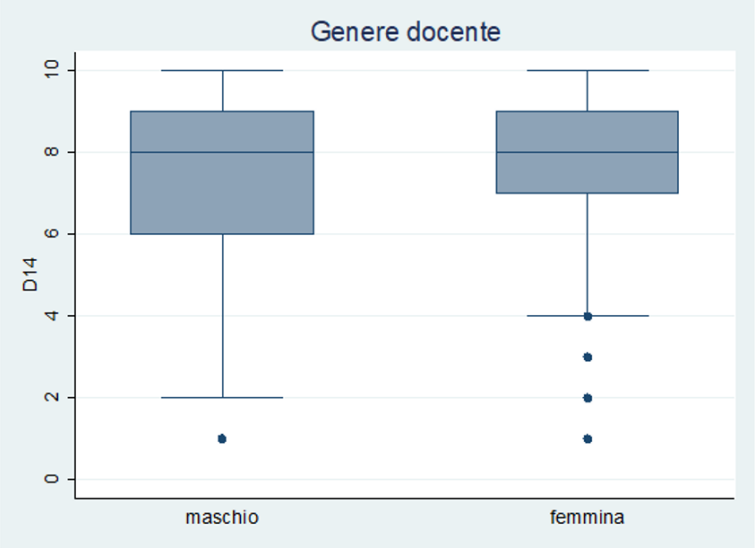

Figura 3.16: Box plot della variabile D14 al variare del genere del docente

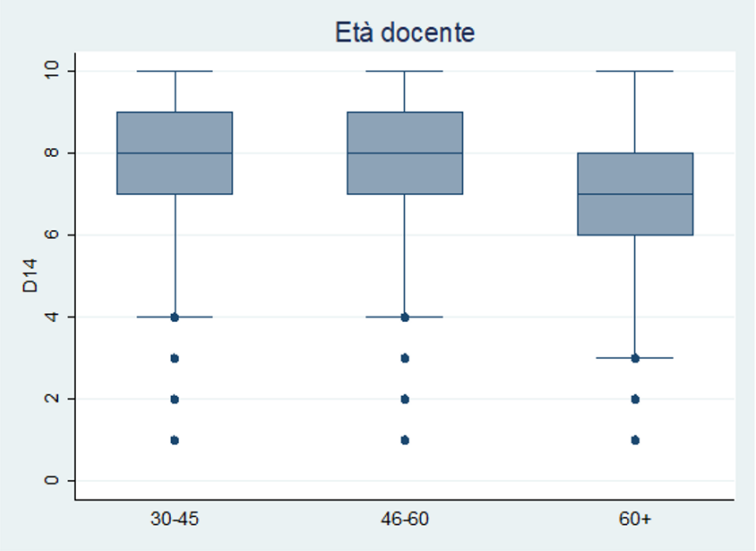

Figura 3.17: Box plot della variabile D14 al variare dell'età del docente

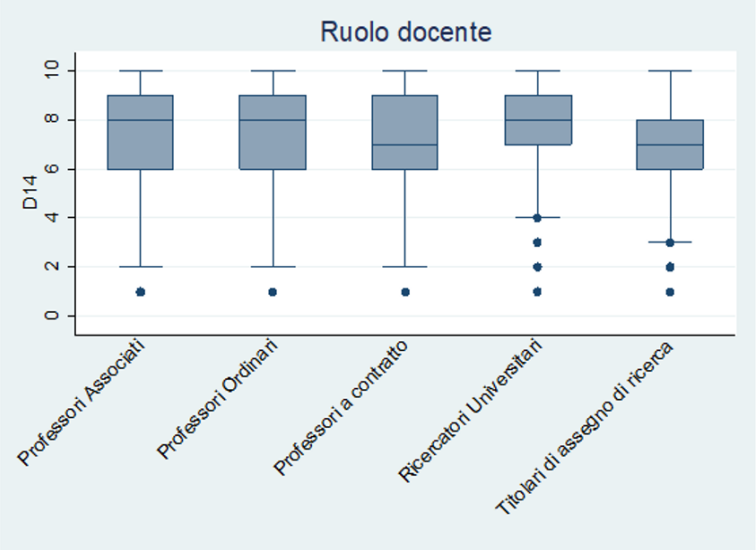

Figura 3.18: Box plot della variabile D14 al variare del ruolo del docente

Il test di Wilcoxon applicato alla variabile sul genere del docente porta ad un rifiuto dell'ipotesi nulla al 5%; i test di Kruscal-Wallis riferiti ad età e ruolo sono invece significativi all'1% (Tabella 3.16).

Tabella 3.16: Test di Wilcoxon e Kruscal-Wallis per caratteristiche dei docenti

		Wilcoxon		**Kruscal-Wallis**	
		z	p-value	χ^2	p-value
DOCENTE	genere	-2.270	0.023	-	-
	età	-	-	367.516	0.000
	ruolo	-	-	230.148	0.000

Dai risultati ottenuti fino a questo momento sembra dunque che tutte le variabili coinvolte influenzino la soddisfazione degli studenti; si andrà poi nei prossimi capitoli ad approfondire il tutto tramite la stima di opportuni modelli.

3.3 Analisi dei dati mancanti

Come si è potuto osservare in Tabella 3.7, nel dataset è presente una considerevole quantità di dati mancanti, dovuta alla non risposta da parte degli studenti all'intero questionario sulla valutazione della didattica, per gli svariati motivi già elencati, oppure soltanto ad alcune delle domande, per la mancata frequenza.

Concentrandosi sulla domanda di interesse, relativa a quanto lo studente si ritenesse complessivamente soddisfatto di un insegnamento, il tasso di non risposta è del 29% circa e la scelta di non rispondere o non frequentare e quindi non rispondere, può essere dovuta a diversi fattori.

Per concludere la fase di analisi esplorativa del dataset si vuole dunque cercare di capire quale meccanismo possa aver generato i dati mancanti, per poi applicare gli opportuni metodi di trattamento in una fase successiva.

Il meccanismo *Missing Completely At Random* (MCAR) è l'unica tipologia che permette di essere verificata tramite dei test (ad esempio, test chi quadrato di Little, 1988)[1], ma nel caso in esame non si ritiene che i dati mancanti abbiano una natura completamente casuale, data la tipologia del questionario e, quindi, la struttura del dataset.

Resterebbe dunque da capire se il meccanismo generatore dei dati sia MAR o MNAR.

Si ricorda che i dati si definiscono *Missing Not At Random* quando la probabilità di avere un valore mancante per una variabile è legata ai valori della variabile stessa, oppure dipende da variabili completamente non osservate.

Nel caso in esame questo "problema" sorge quando si osservano le valutazioni solo per un sottoinsieme di studenti che non sono selezionati a caso in base a variabili osservabili e/o sono condizionati a quest'ultime, ma la selezione è legata a fattori non osservabili come motivazioni personali o opinioni sull'anonimato del questionario, ad esempio.

[1] Il test di Little è nato per testare l'ipotesi MCAR per dati quantitativi multivariati. Nel caso in esame si lavora nell'ambito univariato (con una sola variabile avente NA), quindi il test non è adatto.

Il meccanismo *Missing At Random* invece si verifica quando la probabilità di avere dati mancanti per una variabile è legata ad alcune variabili osservate nell'analisi, ma non ai valori della variabile stessa.
Il problema pratico del meccanismo MAR è però proprio la mancanza di un modo per confermare che la probabilità che vi siano dati mancanti per una variabile è esclusivamente funzione di altre variabili osservate (Enders, 2010).
Come per il MAR, non esiste un metodo esatto per verificare che i dati mancanti siano generati da un meccanismo MNAR; questo poiché i dati necessari per eventuali test sono, per definizione, mancanti (van Buuren, 2012).

Per il momento si considerano separatamente i valori completi e mancanti della variabile di interesse e si applica un test t, o test F, per esaminare le differenze nelle medie di gruppo sulle altre variabili osservate del dataset.
Se i casi con valori mancanti sono sostanzialmente diversi in media da quelli con valori osservati ci si trova nella situazione di dati mancanti non ignorabili e quindi l'ipotesi MAR, che supporta i vari metodi di imputazione singola e multipla, può essere valida.
Il meccanismo MNAR verrà verificato in un secondo momento, in maniera alternativa.
In Tabella 3.17 si riportano dunque i valori delle percentuali di rispondenti solo per le variabili in cui, tramite il test t o test F, sono emerse differenze significative tra i campioni statistici individuati dalle loro categorie rispetto a quella di interesse, ossia una variabile indicatrice della risposta all'item D14 che vale 1 se lo studente ha risposto e 0 altrimenti.
Per le restanti variabili non sono state osservate evidenti discrepanze.

Tabella 3.17: Valori delle percentuali di rispondenti per le variabili selezionate con test t/F

Variabili	**Modalità**	**% rispondenti item D14**
stato iscrizione	fuori corso	55.87
	in corso	71.85
corso obbligatorio	NO	68.1
	SI	77.29
età studenti	17-19	81.45
	20	68.43
	21	70.04
	22	61.14
	23+	57.18
media voti 2012/13	bassa	69.1
	media	70.38
	alta	73.97
CFU	0-30	68.29
	31-60	68.41
	60+	72.94
periodo erogazione corso	primo	73.28
	secondo	68.37

Si osserva che le caratteristiche che discriminano maggiormente tra rispondenti e non rispondenti sono il fatto di essere in corso, di valutare un corso obbligatorio, ma anche l'età dello studente: più avanza più la percentuale di rispondenti cala, come ci si poteva aspettare.

Per quanto riguarda la carriera dello studente sono importanti la media dei voti e i crediti formativi acquisiti: il loro aumento comporta una maggior percentuale di rispondenti, indicando la maggior puntualità degli studenti più meritevoli.

Da notare che, in aggiunta alle usuali caratteristiche di corsi, studenti e docenti, è stata considerata anche la variabile riferita al periodo di erogazione dell'insegnamento, la cui utilità sarà spiegata nei capitoli successivi.

Per il momento è sufficiente notare che il tasso di risposta per i corsi del primo periodo (primo semestre/trimestre) è significativamente più alto che per i corsi tenuti nel secondo periodo dell'anno accademico; questo è probabilmente il risultato del fatto che durante questo periodo gli studenti possono avere iniziato la pausa estiva o, nel caso di studenti dell'ultimo anno, possono essersi laureati e aver lasciato l'università.

Capitolo 4

I dati gerarchici

Dopo aver osservato, tramite le analisi descrittive, l'esistenza di una relazione tra il punteggio assegnato alla soddisfazione complessiva di un corso e le caratteristiche di corsi, studenti e docenti, si passa all'implementazione di opportuni modelli per studiare l'entità di questo legame.

Per far questo è necessario introdurre i modelli multilivello, per tenere conto della struttura gerarchica dei dati, relativa all'appartenenza delle valutazioni dell'item D14 ad ogni singola attività didattica tenuta da un docente.

Nelle strutture multilivello, l'"incompletezza" dovuta ai valori mancanti può facilmente diventare un problema complesso e avere un effetto potenzialmente negativo sulla validità delle inferenze, a meno che non si intervenga in modo opportuno.

Ignorare i valori mancanti infatti può condurre ad ottenere stime imprecise delle relazioni tra le variabili (Yucel, 2008).

I *missing data* possono essere presenti in qualsiasi livello di una struttura gerarchica e, nel caso in esame, si considerano in quelle che verranno definite unità di primo livello, come verrà approfondito nel corso del capitolo.

4.1 Struttura di tipo gerarchico

Un modello statistico nasce con lo scopo di rappresentare la realtà in maniera semplice e parsimoniosa, per interpretare, prevedere e simulare dei fenomeni reali. Per poter far questo è necessario individuare e studiare gli aspetti presenti nei dati, in modo da poter usare i metodi statistici più opportuni per raggiungere l'obiettivo preposto.

La struttura dei dati può essere semplice o complessa: la prima non implica particolari tipi di dipendenze o raggruppamenti delle osservazioni, mentre la seconda indica una suddivisione delle unità statistiche in sottoinsiemi.

In particolare, sempre più spesso e in vari ambiti disciplinari, si analizzano fenomeni con una struttura gerarchica, in cui i dati si riferiscono ad uno o più livelli di osservazione/appartenenza: individuale, familiare, territoriale, sociale e così via. Lo

studio delle relazioni tra l'individuo e il contesto che lo circonda può essere pertanto ricondotto all'analisi di fenomeni a struttura gerarchica (Ruscone, 2011).

Le strutture complesse possono essere divise in *nested* e *non-nested*.

Una struttura *nested* è quella in cui la gerarchia comporta l'esistenza di sottoinsiemi nidificati che contengono sottogruppi definiti a livelli inferiori. Una caratteristica dei dati con struttura *nested* è che gli individui che fanno parte del medesimo gruppo sono più somiglianti fra loro rispetto a quelli appartenenti a gruppi diversi. Ad esempio, gli studenti con attitudini e motivazioni affini si trovano ad essere riuniti nelle stesse scuole a seguito di processi di selezione oppure, anche nel caso in cui il raggruppamento venga fatto senza tenere in considerazione le caratteristiche degli individui, gli alunni della stessa scuola condividono la stessa realtà e subiscono le medesime influenze; le persone che vivono nella stessa area geografica o amministrativa sono soggette alle stesse politiche locali e manifestano uno stile di vita e un comportamento più simile rispetto a persone residenti in contesti differenti (Ruscone, 2011).

Si osservi che le strutture *nested* sono in genere indicate anche con la denominazione alternativa di "gerarchiche". I dati hanno struttura di tipo gerarchico se le entità appartengono a gruppi che a loro volta possono essere contenuti in altri gruppi di ampiezza/livello superiore.

I dati hanno invece struttura *non-nested* quando la condizione di contenimento dei livelli più bassi non è soddisfatta. Nella struttura *non-nested* non è cioè definibile una partizione.

Nel caso in esame la struttura gerarchica dei dati si può delineare tramite lo schema di rappresentazione grafica a due livelli in Figura 4.1, proposta da Brown et al. (2001): i box rappresentano i livelli ai quali le unità vengono classificate, mentre la relazione gerarchica esistente viene evidenziata da una freccia.

Si comprende dunque che le valutazioni degli studenti dell'Ateneo di Padova (livello-1) sono *nested* nelle attività didattiche legate ad ogni singolo docente (livello-2); ciò significa per esempio che, se un corso è legato a due docenti, agli studenti sono sottoposte a valutazione entrambe le parti.

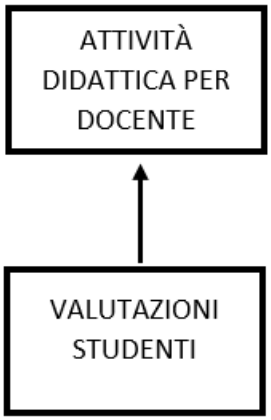

Figura 4.1: Schema di rappresentazione grafica a due livelli della struttura gerarchica dei dati in esame

Ogni gruppo tende quindi a differenziarsi dagli altri in quanto la valutazione del singolo studente viene influenzata dal gruppo di appartenenza e le caratteristiche del gruppo vengono influenzate dagli studenti che lo compongono (Paccagnella, 2006).

4.2 Modelli multilivello

I modelli multilivello costituiscono lo strumento più adatto per trattare le informazioni presenti all'interno di strutture gerarchiche: tengono conto, in maniera esplicita, sia della presenza di relazioni tra le variabili appartenenti ad uno specifico livello, sia delle relazioni tra i differenti livelli, considerando in tal modo l'effetto netto sulle unità e le interazioni presenti.

Una tipologia specifica delle tecniche multilivello riguarda i modelli ad intercetta casuale (*random intercept model*), oppure i modelli a pendenza casuale (*random slope model*), come si osserverà successivamente.

In seguito verrà considerata la presenza di solamente due livelli, coerentemente con il caso in esame, ma il tutto può essere generalizzato anche a più livelli.

Considerando un campione di numerosità N e J gruppi di numerosità n_j, tale che $N = \sum_{j=1}^{J} n_j$, si definisce:

- j = indice delle unità a livello-2, $j=1, 2, \ldots, J$
- i = indice delle unità a livello-1 all'interno del gruppo j, $i=1, 2, \ldots, n_j$
- Y_{ij} = variabile dipendente a livello-1 (si riferisce all'unità *i-esima* del gruppo *j-esimo*)

- X_{ij} = variabile esplicativa a livello-1 che esprime una caratteristica dell'unità *i-esima* del gruppo *j-esimo*
- W_j = variabile esplicativa a livello-2 (riferita al gruppo *j-esimo*). Può essere di due tipi:
  - Z_j = variabile che esprime una caratteristica propria del gruppo *j-esimo*
  - $\bar{X}_{.j}$ = media di gruppo, cioè il valore medio all'interno del gruppo *j-esimo* di tutte le caratteristiche X_{ij}.
- ε_{ij} = effetto non osservato specifico a livello-1
- U_j = effetto non osservato specifico a livello-2

Il modello multilivello può essere rappresentato mediante un'equazione di primo livello,

$$Y_{ij} = \alpha_j + \beta_j X_{ij} + \varepsilon_{ij} \tag{4.1}$$

che rappresenta, in base ai coefficienti α_j e β_j, la relazione esistente tra le variabili riferite all'unità statistica di primo livello, nella quale il termine di errore ε_{ij} si assume con distribuzione Normale di media pari a zero e varianza costante pari a σ_ε^2, e mediante le equazioni di secondo livello

$$\alpha_j = \gamma_{00} + \gamma_{01} W_j + u_{oj} \tag{4.2}$$

$$\beta_j = \gamma_{10} + \gamma_{11} W_j + u_{1j} \tag{4.3}$$

che rappresentano, in base ai coefficienti γ_{00}, γ_{01}, γ_{10}, γ_{11}, le relazioni esistenti tra la variabile osservata (W) su ciascun gruppo (unità di secondo livello) e i coefficienti α_j, β_j inseriti nell'equazione di primo livello.

In particolare "$\gamma_{00} + \gamma_{01} W_j$" indica il valore atteso dell'intercetta per gruppi con caratteristiche W_j, mentre u_{oj} rappresenta la deviazione del gruppo j da questo valore atteso. Analogo per l'espressione (4.3).
I termini di errore u_{oj} e u_{1j} si assumono Normali con media zero e varianze costanti, σ_{u0}^2 e σ_{u1}^2, in generale correlati tra loro e indipendenti dal termine di errore ε_{ij}:

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

$$Cov(\varepsilon_{ij}, U_j) = 0$$

Combinando il tutto in un'unica equazione si ottiene:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}W_j + \gamma_{11}W_jX_{ij} + u_{oj} + u_{1j}X_{ij} + \varepsilon_{ij} \qquad \text{(4.1 bis)}$$

dove:

- $\gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}W_j + \gamma_{11}W_jX_{ij}$ costituisce la parte deterministica del modello, poiché non contiene alcuna variabile casuale;
- $u_{oj} + u_{1j}X_{ij} + \varepsilon_{ij}$ costituisce la parte casuale del modello, poiché contiene gli effetti casuali a tutti i livelli e le interazioni con le variabili osservate.

Il modello proposto nella sua formulazione più generale prende il nome di *modello a pendenza casuale*, ma può essere semplificato assumendo che i gruppi si differenzino tra loro soltanto rispetto al valore atteso della variabile dipendente e non anche rispetto alla pendenza di ogni singola retta.

Questo caso particolare del modello a due livelli in cui solo l'intercetta α_j può assumere dei valori diversi in funzione del gruppo di appartenenza è noto come *modello ad intercetta casuale.*

Esso può essere utilizzato per valutare le differenze esistenti tra i gruppi di unità statistiche, considerando le equazioni (4.1) e (4.2), e modificando l'equazione (4.3):

$$Y_{ij} = \alpha_j + \beta_j X_{ij} + \varepsilon_{ij} \qquad (4.4)$$

$$\alpha_j = \gamma_{00} + \gamma_{01}W_j + u_{oj} \qquad (4.5)$$

$$\beta_j = \gamma_{10} + \gamma_{11}W_j \qquad (4.6)$$

Sulla base di quest'ultima formulazione si ricava che la varianza può essere scomposta come la somma delle varianze a livello 1, ossia tra le unità all'interno dei gruppi (σ_ε^2 -varianza *within*-), e a livello 2, ossia tra i gruppi (σ_{u0}^2 -varianza *between*-), nel seguente modo:

$$Var(Y_{ij}|X_{ij}, W_j) = Var(u_{oj}) + Var(\varepsilon_{ij}) = \sigma_\varepsilon^2 + \sigma_{u0}^2 \qquad (4.7)$$

La covarianza tra due individui *i* e *i'* appartenenti allo stesso gruppo *j* è uguale alla varianza di u_{oj}:

$$Cov\left(Y_{ij}, Y_{i'j} \middle| X_{ij}, X_{i'j}, W_j\right) = Var\left(u_{oj}\right) = \sigma_{u0}^2 \quad i \neq i' \tag{4.8}$$

Il rapporto

$$\rho\left(Y_{ij}, Y_{i'j} | X_{ij}, X_{i'j}, W_j\right) = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_{\varepsilon}^2} \tag{4.9}$$

è il coefficiente di correlazione intraclasse (ICC), il quale varia tra 0 e 1, e rappresenta una misura che giustifica il ricorso al modello gerarchico. Un valore molto basso infatti, non segnalando la presenza di correlazione all'interno dei gruppi, suggerisce di evitare la modellizzazione a più livelli e di ricorrere a tradizionali modelli di regressione ad un solo livello. All'aumentare dell'ICC aumenta il contributo esplicativo dovuto alla struttura gerarchica.

Questo coefficiente fornisce una misura di omogeneità all'interno di uno stesso gruppo e rappresenta la proporzione di varianza residua spiegata dal raggruppamento; misura quindi la parte di variabilità dovuta all'effetto del raggruppamento e quella derivante dalla dipendenza tra osservazioni raggruppate in unità dello stesso livello (Kreft e De Leeuw, 1998).

Approfondendo la relazione tra le componenti di varianza e le variabili esplicative si ricorda che in un modello di regressione standard la varianza del termine di errore ha il significato di varianza residua, cioè varianza non spiegata dai regressori e generalmente l'inserimento di una nuova variabile comporta una riduzione del suo valore, la cui entità dipende dal potere esplicativo della variabile inserita.

In un modello multilivello la situazione è più complessa per il fatto che, come evidenziato, la varianza viene scomposta nella componente *between*, cioè la varianza non spiegata dai regressori e che è attribuibile agli effetti casuali, ovvero alla struttura gerarchica, e la componente *within*, ovvero la varianza residua in senso stretto, che non è spiegata né dai regressori né dall'appartenenza ai gruppi, ma che è legata alla variabilità individuale. L'effetto dell'inserimento di nuove variabili sulle componenti di varianza dipende dal tipo di variabile (Longford, 1993):

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

- variabile individuale (livello 1): come è naturale attendersi, l'inserimento di una variabile individuale riduce la varianza *within*, ma la direzione del suo effetto sulla componente *between* non è determinabile a priori. Infatti, bisogna pensare che la componente *between* è una misura del grado di eterogeneità dei gruppi non spiegata dai regressori e che l'inserimento di una nuova variabile individuale può sia aumentare che diminuire la misura di tale eterogeneità non spiegata.

4.2.1 Centratura delle variabili

La centratura di una variabile in un modello multilivello ha effetti molto diversi rispetto al modello di regressione lineare, in base a come la variabile viene centrata e al tipo di modello gerarchico (Paccagnella, 2006).
Un modello multilivello senza alcuna variabile centrata è detto *raw score model*.

Le variabili possono essere centrate in diversi modi:

- centratura rispetto alla media complessiva: $\ddot{X}_{ij} = X_{ij} - \bar{X}_{..}$
- centratura rispetto alla media di gruppo: $\tilde{X}_{ij} = X_{ij} - \bar{X}_{.j}$

Centrare rispetto alla media complessiva porta solo ad una riparametrizzazione del modello, il quale risulta essere equivalente ad un *raw score model*; lo stesso non vale per un modello contenente una variabile centrata rispetto alla media di gruppo, la quale cambia il modello stimato e quindi anche l'interpretazione delle stime.
Si è detto che le osservazioni di uno stesso gruppo sono generalmente più simili rispetto ad osservazioni di altri gruppi e la centratura delle variabili può aiutare a studiare come l'appartenenza ad un gruppo influenzi il comportamento individuale.
A questo proposito si introducono gli effetti contestuali, che misurano proprio quanto il comportamento individuale tenda a variare con il background (caratteristiche osservabili del gruppo). La propensione di un individuo a comportarsi in un certo modo varia con la media delle caratteristiche del gruppo.

Facendo l'esempio di un *random intercept model* con variabili di gruppo, si ha:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \delta\, \bar{X}_{.j} + u_{oj} + \varepsilon_{ij} \qquad (4.10)$$

dove δ indica l'effetto contestuale.

Le stime di tale modello però possono non essere accurate, in quanto esso potrebbe soffrire di alta collinearità, dovuta alla relazione tra X_{ij} e $\bar{X}_{.j}$ (Aitkin e Longford, 1986).

Considerando le deviazioni dalla media di gruppo:

$$(X_{ij} - \bar{X}_{.j})$$

Il modello

$$Y_{ij} = \gamma_{00} + \gamma_{10w}(X_{ij} - \bar{X}_{.j}) + \delta_B \bar{X}_{.j} + u_{oj} + \varepsilon_{ij} \qquad (4.11)$$

è una semplice riparametrizzazione del (4.10) e non soffre più di collinearità in quanto le quantità sono per costruzione incorrelate.

Per testare la presenza di effetti contestuali dunque:

- in (4.10) si testa la significatività di δ e la presenza di collinearità può influenzare i risultati;
- in (4.11) si testa la significatività di $(\delta_B - \gamma_{10w})$, cioè che i due parametri siano statisticamente uguali.

4.2.2 Metodo di stima

Il principale metodo di stima usato nelle analisi multilivello è la massima verosimiglianza (*Maximum Likelihood* - ML) e ne esistono due varianti (Bryk e Raudenbush, 1992): *Full Maximum Likelihood* (FML) e *Restricted Maximum Likelihood* (REML).

Il primo tipo è caratterizzato dal fatto che la stima dei coefficienti di regressione e delle componenti di varianza avviene in due fasi distinte, mentre il secondo tipo consente di stimare simultaneamente tutti i parametri.

Dalla letteratura sull'argomento risulta che questi due tipi di stimatori non differiscono tanto per i valori delle stime dei parametri relativi ai coefficienti (γ) di regressione, quanto per quelli della stima delle componenti di varianza ($\sigma_\varepsilon^2, \sigma_{u0}^2$); in particolare, con lo stimatore REML tali componenti vengono stimate tenendo conto di un numero inferiore di gradi di libertà rispetto a quello che caratterizza lo stimatore FML, che tende a sottostimare le componenti di varianza di una certa quantità, la quale diminuisce con l'aumento del numero delle unità di secondo livello (Snijders e Bosker, 1999) e che è ritenuta trascurabile oltre un certo valore del numero *J* dei gruppi (*J* > 30).

Inoltre, dai risultati ottenuti mediante diversi studi di simulazione (Kreft e De Leeuw, 1998), non è stato possibile individuare il metodo migliore di stima delle componenti di varianza poiché è stato osservato che lo stimatore REML, pur essendo corretto, è meno efficiente dello stimatore FML.

Capitolo 5

Applicazioni empiriche

5.1 Stima di modelli multilivello

In questo Capitolo si vogliono proporre delle analisi multivariate sulla soddisfazione degli studenti, sfruttando la struttura gerarchica che si è visto caratterizzare il dataset in esame. Le unità di primo livello saranno quindi le valutazioni degli studenti, mentre le unità di secondo livello saranno le attività didattiche legate ad ogni docente.

Sebbene sia stato suggerito come presumibilmente il meccanismo generatore dei dati mancanti non sia di tipo MCAR, come punto di partenza, e di confronto, nella fase di stima, si procede con la stima di due modelli ad intercetta casuale ottenuti con la tecnica utilizzata più comunemente, ossia la *listwise deletion*; vengono quindi eliminate tutte le righe contenenti un valore mancante nella variabile di interesse.

I modelli sono realizzati con il metodo FML, dato che, a differenza del metodo REML, produce degli indicatori di bontà del modello, come la devianza.

Per ottenere le stime il software sfrutta l'algoritmo *Expectation Maximization.*

5.1.1 Modello ad intercetta casuale con listwise deletion – I^{o} livello

Il primo modello proposto studia la relazione tra la soddisfazione degli studenti nei confronti dei corsi frequentati e le variabili a livello 1, ossia relative proprio agli studenti, con l'aggiunta di un effetto casuale per gruppo, ossia le attività didattiche per docente. In questo modo si può rilevare se e come il punteggio assegnato dagli studenti all'item 14 sia legato alle caratteristiche degli studenti e quanta variabilità vi sia tra i gruppi.

I principali risultati sono riportati in Tabella 5.1.

*Tabella 5.1: Stime del modello ad intercetta casuale con variabili di primo livello dopo LD. Livello di significatività: *** p-value<0.01, ** p-value<0.05, * p-value<0.1*

STIME EFFETTI FISSI				
Caratteristiche	**Variabili**	**Stima**	**Std. Error**	
STUDENTE-GENERALE	frequenza 30-50%	-1.072	0.026	***
	frequenza 50-70%	-0.622	0.018	***
	genere: femmina	-0.037	0.014	***
	età	0.034	0.002	***
STUDENTE-CARRIERA	in corso	-0.243	0.036	***
	media dei voti	0.022	0.003	***
	numero medio esami	-0.060	0.012	***
	CFU	0.004	0.001	***
	numero esami * numero crediti	0.0004	0.000	***
	costante	6.622	0.093	***
EFFETTI CASUALI				
	Varianza between	0.992		
	Varianza within	2.721		
	ICC	0.267		
DIMENSIONI				
	Numero osservazioni	77177		
	Numero gruppi	1925		
	Media oss. per gruppo	40		
TEST				
	Wald χ^2_9	2917.45		
	Prob > χ^2	0.000		
	LR χ^2_{bar} (01)	18464.73		
	Prob ≥ χ^2_{bar}	0.000		

In totale l'analisi comprende 1925 unità di secondo livello, con una media di 40 osservazioni per gruppo; in fase di preparazione del dataset si è deciso di eliminare i gruppi con meno di cinque osservazioni.

Innanzitutto dall'output del modello emerge come il test di Wald rifiuti l'ipotesi nulla di non significatività congiunta di tutte le variabili eccetto la costante; anche il test log-rapporto di verosimiglianza, che verifica l'uguaglianza tra il modello multilivello ed il rispettivo modello di regressione lineare, ossia l'assenza di eterogeneità non osservata, ha un p-value pari a zero, confermando la validità del modello in esame.

Nella Tabella 5.1 si nota come tutte le variabili risultino essere significative ad un livello dell'1%, quindi, a conferma di quanto emerso nelle analisi descrittive, le caratteristiche individuali degli studenti influenzano in qualche modo la soddisfazione degli stessi verso una determinata attività didattica.

Si noti che è stato inserito un effetto di interazione tra il numero di esami ed il numero crediti annui, a causa del segno opposto delle stime dei coefficienti delle due singole variabili.

Osservando i valori delle stime, a parità di altre condizioni, emerge:

- una minor soddisfazione per coloro che frequentano meno del 70% delle lezioni: chi frequenta più del 70% (categoria di riferimento) si presume segua meglio il corso e quindi possa giudicarlo in modo più positivo;
- una minor soddisfazione degli studenti di genere femminile;
- all'aumentare di un anno di età aumenta la soddisfazione di 0.034 punti e ciò, non considerando i casi estremi, potrebbe essere legato all'anno di corso di iscrizione: si presume che l'avanzare dell'età corrisponda all'avanzare dell'anno di corso e quindi che uno studente possa essere più preparato, abbia maggior consapevolezza e conoscenza dei meccanismi di valutazione e inoltre ci possa essere stata una sorta di selezione che faccia proseguire gli studenti più capaci o motivati;
- l'essere in corso è legato ad un calo del punteggio della soddisfazione di 0.24 punti e può essere dovuto al fatto che gli studenti fuori corso potrebbero riseguire alcune attività, con un bagaglio di conoscenze maggiore, arrivando a dare valutazioni più alte;
- il punteggio dato alla soddisfazione è associato positivamente, seppur in leggera quantità, alla media dei voti degli studenti registrati nell'A.A. 2012/2013: più alta è la media dei voti, più alti sono i punteggi assegnati;
- le stime dei coefficienti legati al numero medio di esami registrati all'anno e ai crediti acquisiti mediamente in un anno hanno segni opposti: l'aumento dei crediti annui è associato ad un aumento della soddisfazione, mentre l'aumento degli esami è associato ad una diminuzione della stessa; il coefficiente legato all'interazione risulta significativo e positivo e, calcolando gli effetti marginali per diversi valori delle due variabili, è emerso come la soddisfazione degli studenti sia influenzata positivamente se vengono registrati esami più consistenti dal punto di vista dei crediti. Per uno studente, quindi, ciò che è veramente

importante è il numero di CFU da acquisire più che il numero di esami, per cui è preferibile fare pochi esami con tanti CFU, piuttosto che tanti esami con pochi CFU.

Per quanto riguarda le componenti di varianza si osserva che, con le sole variabili riferite agli studenti, quindi di primo livello, il valore del coefficiente di correlazione intraclasse è pari a 0.267, ad indicare cioè che il 27% circa della variabilità totale è spiegata dalla variabilità tra i gruppi; l'elevato valore denota la presenza di molta variabilità tra le attività didattiche.
Per questo motivo si ritiene necessario inserire delle variabili di secondo livello, che contribuiscano a spiegare le differenze tra i gruppi e quindi a ridurre la componente *between* (e di conseguenza l'ICC).

5.1.2 Modello ad intercetta casuale con listwise deletion – IIº livello

Il secondo modello proposto considera dunque, oltre alle variabili individuali riferite agli studenti, una serie di variabili che esprimono caratteristiche proprie dei gruppi, ossia riferite ai docenti che conducono i corsi (attività didattiche) ed ai corsi stessi.
Vengono inserite inoltre le medie di gruppo, cioè il valore medio all'interno di ogni gruppo di tutte le caratteristiche degli studenti (di primo livello), in modo da valutare gli effetti contestuali, ossia se il comportamento individuale degli studenti tende a variare con la media delle caratteristiche del gruppo.
Infine, i Corsi di Studio non sono intesi come un ulteriore livello ma vengono inseriti come variabili di secondo livello di controllo, di cui non si riportano le stime dato il loro elevato numero.
Si riportano i risultati in Tabella 5.2.

*Tabella 5.2: Stime del modello ad intercetta casuale con variabili di secondo livello dopo LD. Livello di significatività: *** p-value<0.01, ** p-value<0.05, * p-value<0.1*

STIME EFFETTI FISSI				
Caratteristiche	**Variabili**	**Stima**	**Std. Error**	
STUDENTE-GENERALE	frequenza 30-50%	-1.062	0.026	***
	frequenza 50-70%	-0.617	0.018	***
	genere: femmina	-0.045	0.015	***
	età	0.033	0.002	***
STUDENTE-CARRIERA	in corso	-0.214	0.036	***
	media dei voti	0.018	0.003	***
	numero medio esami	-0.061	0.012	***
	CFU	0.004	0.001	***
	numero esami * numero crediti	0.0004	0.000	***
STUDENTE-GENERALE (media)	frequenza 30-50%	-3.329	0.537	***
	frequenza 50-70%	-1.832	0.368	***
	genere: femmina	-0.414	0.209	**
	età	0.022	0.258	
STUDENTE-CARRIERA (media)	in corso	-0.248	0.27	
	media dei voti	0.025	0.031	
	numero medio esami	-0.114	0.062	
	CFU	0.011	0.008	
CORSO	corso obbligatorio	-0.028	0.04	
	numero di ore	0.058	0.14	
	più di un docente	-0.077	0.05	
	sede Padova	2.213	1.896	
	insegnamento mutuato	-0.138	0.054	**
DOCENTE-GENERALE	genere: femmina	-0.055	0.053	
	età	-0.016	0.029	***
DOCENTE-CARRIERA	professore associato	-0.004	0.065	
	professore ordinario	0.061	0.08	
	professore a contratto	0.027	0.079	
	titolare assegno di ricerca	-0.409	0.236	*
	costante	7.691	0.945	***
EFFETTI CASUALI				
	Varianza between	0.832		
	Varianza within	2.716		
	ICC	0.235		
DIMENSIONI				
	Numero osservazioni	77177		
	Numero gruppi	1925		

Media oss. per gruppo	40
TEST	
Wald χ^2_{109}	3397.3
Prob > χ^2	0.000
LR χ^2_{bar} (01)	14719
Prob ≥ χ^2_{bar}	0.000

Dalla Tabella 5.2 si osserva innanzitutto come le stime dei parametri associati alle variabili di primo livello siano rimaste pressoché invariate rispetto al precedente modello, anche in termini di significatività.

Rispetto a quanto emerso nelle analisi descrittive le caratteristiche legate al corso non risultano essere significative, ad eccezione della variabile relativa all'insegnamento mutuato, che risulta avere un effetto negativo: il fatto che l'attività didattica sia mutuata porta ad una diminuzione di 0.14 del punteggio dato alla soddisfazione complessiva, a parità di altre condizioni.

Anche le caratteristiche dei docenti hanno un effetto debole, soltanto l'età è fortemente significativa con segno negativo. Il ruolo di titolare di assegno di ricerca è invece significativo al 10%, con effetto negativo: i docenti con questa posizione sono legati ad attività didattiche aventi punteggio di 0.41 punti inferiore rispetto a quello delle attività legate ai ricercatori (modalità di riferimento), ceteris paribus.

Per quanto riguarda le altre variabili di secondo livello, cioè le medie di gruppo, sono riportati direttamente i valori degli effetti contestuali (nel modello sono state inserite le variabili centrate) e si nota che risultano significativi soltanto quelli relativi alla frequenza e al genere; si può quindi affermare che, a parità di altre condizioni:

- appartenere a classi composte in prevalenza da studenti frequentanti tra il 30% e il 50% delle lezioni o tra il 50 e il 70% comporta l'assegnazione di un minor punteggio alla soddisfazione complessiva, rispetto all'appartenere a classi con prevalenza di frequentanti più del 70% delle lezioni;
- l'appartenere ad una classe composta in prevalenza da studentesse, rispetto a classi con maggioranza di individui di genere maschile, implica una diminuzione della soddisfazione dei rispondenti; questo può essere dovuto ad un diverso modo di lavorare delle studentesse, le quali tendono ad interagire di più, facendo emergere maggiormente delle criticità nella classe, con una conseguente valutazione inferiore del corso.

Le restanti medie di gruppo, riferite sostanzialmente alla carriera dello studente, e quindi indici della qualità della classe relativa ad ogni insegnamento, non sono statisticamente significative ed è stato osservato che, provando a togliere le variabili indicatrici dei Corsi di Studio, la non significatività è proprio legata alla presenza di quest'ultime; senza esse le medie di gruppo accorperebbero delle caratteristiche dei corsi, per cui inserendole perdono di significatività.
In sostanza, la struttura di ogni Corso di Studio si riflette sulla composizione delle varie classi, per cui gli effetti contestuali perdono importanza.

Osservando le componenti della varianza si nota che il valore della varianza *between* è calato, con la conseguente diminuzione dell'ICC, che rimane comunque elevato, pari a 0.23; ciò significa che le variabili di secondo livello inserite non sono state in grado di spiegare sufficientemente l'eterogeneità tra i gruppi.
Sarebbe quindi necessario l'inserimento di ulteriori variabili di secondo livello, magari riferite ai docenti, poiché, per quanto visto fino a questo punto, le differenze tra le attività didattiche sembrano leggermente più legate a quest'ultimi che ai corsi stessi. Un'idea potrebbe essere quella di inserire alcune caratteristiche soggettive dei docenti, ricavate dal questionario PRODID (Preparazione alla Professionalità Docente e Innovazione Didattica), ossia uno strumento nato per analizzare le pratiche didattiche svolte dai docenti, le loro credenze ed i loro bisogni, nonché la loro attitudine nei confronti dell'insegnamento e della ricerca.
Tuttavia in questa sede non si procede con tale possibilità, volendo concentrarsi nel trattamento dei dati mancanti a livello di studenti.
Le stime dei primi due modelli, sfruttando la tecnica della *listwise deletion*, infatti potrebbero essere poco precise e distorte, data la perdita di informazione dovuta ai casi scartati e la non ragionevolezza di un'ipotesi MCAR.

Prima di passare alla ragionevole ipotesi di un meccanismo MAR, con l'applicazione delle tecniche di imputazione, si vuole verificare la possibile presenza di distorsione dovuta alla selezione non casuale dei *missing data*, sottostante al meccanismo MNAR; nel caso ci si trovasse in questa situazione infatti non avrebbe senso proseguire con l'imputazione, in quanto si otterrebbero risultati statisticamente non affidabili.

5.2 Approccio per verificare la presenza di un meccanismo MNAR

In questa sezione si pone dunque l'attenzione sulla possibile distorsione da selezione (*selection bias*) che si verifica quando gli studenti che scelgono di partecipare al questionario non sono selezionati casualmente; in questo caso si tratta di dati mancanti generati da un meccanismo *Missing Not At Random*.

Come già accennato, nel caso dei questionari per la valutazione della didattica questo fatto si verifica quando si osservano le valutazioni solo per un sottoinsieme di studenti che non sono selezionati casualmente in base a caratteristiche osservabili, come genere, rendimento, percorso e/o sono condizionati ad esse, ma dipendono da fattori non osservabili, come motivazioni personali o preoccupazione sull'anonimato del questionario (Goos and Salomons, 2016).

Dalle analisi descrittive, infatti, è emersa una differenza tra rispondenti e non rispondenti basata su alcune caratteristiche osservate, ma, data la tipologia di questionario, non si può escludere che la distinzione sia dovuta anche a caratteristiche non osservabili.

J. Heckman si è occupato di questo problema, sviluppando un modello di selezione che include sia la dimensione osservabile sia quella non osservabile.

5.2.1 Modello di Heckman

Heckman (1979) propone un modello diviso in due parti, poiché combina un modello di regressione lineare (sulla variabile di interesse) con un'equazione di "selezione" aggiuntiva che prevede la probabilità che si verifichi un certo evento, ossia la compilazione del questionario, nel caso in esame; si vuole tenere conto dell'autoselezione, ottenendo un risultato non inficiato da *selection-bias*.

Equazione di selezione

Uno studente decide se rispondere al questionario, e in particolare all'item D14 di interesse, in base alla sua utilità netta, $Y_1^* \in (-\infty, +\infty)$, derivata dalla risposta. Questa utilità è determinata da un vettore di covariate, X_1, e dai loro coefficienti, β_1, nonché da un termine aggiuntivo di errore ε_1, per mezzo di una relazione lineare:

$$Y_1^* = X_1\beta_1 + \varepsilon_1 \quad (5.1)$$

Se $Y_1^* \geq 0$ lo studente risponde alla domanda di interesse, mentre se $Y_1^* < 0$ lo studente non risponde.

Tuttavia Y_1^* non è direttamente osservata, ma si ha solo una variabile che indica la risposta o meno dello studente; in pratica si osserva una variabile Y_1 che ha valore 1 se lo studente risponde all'item D14 e 0 altrimenti:

$$Y_1 = I(Y_1^* \geq 0) = I(X_1\beta_1 + \varepsilon_1 \geq 0) \quad (5.2)$$

dove $I(\cdot)$ denota la funzione indicatrice e ε_1 si assume distribuito come una Normale di media zero.
I coefficienti β_1 possono essere stimati dal campione completo di tutti gli studenti, rispondenti e non, ossia $Y_1 = 1$ e $Y_1 = 0$, con le loro caratteristiche osservate, X_1 (Goos e Salomons, 2016).

Equazione di regressione

La valutazione della soddisfazione di uno studente è una variabile continua Y_2^*, che dipende da un set di covariate, X_2, e dai loro coefficienti, β_2, nonché da un termine aggiuntivo ε_2:

$$Y_2^* = X_2\beta_2 + \varepsilon_2 \quad (5.3)$$

Tuttavia, si osserva Y_2^* solo per l'insieme di studenti che decidono di rispondere alla domanda, cioè $Y_2^* = Y_2$ se $Y_1 = 1$.
L'equazione usata per stimare β_2 è dunque:

$$(Y_2 \mid Y_1 = 1) = (X_2\beta_2 + \varepsilon_2 | \varepsilon_1 \geq -X_1\beta_1) \quad (5.4)$$

Nella matrice di variabili X_2 sono compresi tutti quei fattori osservati che possono influenzare la valutazione di uno studente ma, ripetendo quanto già detto, dato che le variabili osservate possono spiegare solo una parte della variabilità delle valutazioni, ci possono essere dei fattori non osservabili che le influenzano, contenuti in ε_2.

Un meccanismo MNAR implica che i dati e la probabilità di avere valori mancanti abbiano una distribuzione congiunta, ossia i dati portano informazione sulla probabilità di avere valori mancanti, e viceversa (Enders, 2010). Il modello di selezione incorpora questa dipendenza tramite una distribuzione normale bivariata per i termini di errore, nel modo seguente:

$$\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\end{bmatrix} \sim N\left(\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix}\right)$$

dove σ_{12} è la covarianza tra i termini di errore ed è particolarmente importante poiché rappresenta il meccanismo con cui il modello per dati mancanti regola la distorsione nel modello principale (sui punteggi delle valutazioni).

Per approfondire, l'equazione (5.1) è in realtà un modello probit standard, che descrive la scelta di rispondere o meno alla domanda di interesse, e per questo motivo è necessario imporre una condizione di normalizzazione, che solitamente corrisponde a $\sigma_1^2 = 1$.

Il punteggio atteso, condizionale al fatto che uno studente abbia risposto, è dato da

$$\begin{aligned} E[Y_2 \,|Y_1 = 1] &= X_2\beta_2 + E[\varepsilon_2|Y_1 = 1] \\ &= X_2\beta_2 + E[\varepsilon_2|\varepsilon_1 > -X_1\beta_1] \\ &= X_2\beta_2 + \frac{\sigma_{12}}{\sigma_1^2}\, E[\varepsilon_1|\varepsilon_1 > -X_1\beta_1] \\ &= X_2\beta_2 + \sigma_{12}\frac{\phi(X_1\beta_1)}{\Phi(X_1\beta_1)} \end{aligned} \qquad (5.5)$$

dove l'ultima uguaglianza usa la proprietà $\sigma_1^2 = 1$ e l'espressione del valore atteso di una distribuzione normale standard troncata (dove ϕ indica la funzione di densità e Φ la funzione di ripartizione).

Si noti che $\sigma_{12} = \rho_{12}\sigma_2$, dove ρ_{12} è il coefficiente di correlazione tra i due errori.

Dall'espressione (5.5) segue direttamente che il punteggio atteso condizionale è pari a $X_2\beta_2$ solo se $\sigma_{12} = \rho_{12} = 0$. Dunque, se i termini di errore delle due equazioni sono incorrelati l'equazione del punteggio può essere stimata consistentemente usando i minimi quadrati ordinari, cioè senza ricorrere al modello di selezione. Se $\sigma_{12} \neq 0$ lo stimatore OLS è affetto da una distorsione dovuta alla selezione campionaria.

Il termine $\lambda = \frac{\phi(X_1\beta_1)}{\Phi(X_1\beta_1)}$ è noto come inverso del rapporto di Mill, anche chiamato lambda di Heckman (Verbeek, 2010).

Si arriva quindi alla conclusione che la correlazione tra ε_1 e ε_2 quantifica la grandezza della violazione del meccanismo MAR. In particolare, una correlazione pari a zero indica che i dati sono di tipo MAR, poiché non esiste alcuna relazione tra la variabile dipendente e la probabilità di avere dati mancanti, dopo aver controllato per un insieme di caratteristiche osservabili.

Al contrario, una correlazione significativamente diversa da zero suggerisce che i dati siano di tipo MNAR (Enders, 2010).

Osservare il coefficiente di correlazione può quindi essere un modo per testare i meccanismi MAR/MNAR.

Stima del modello

Esistono due approcci principali per stimare il modello di Heckman: la tecnica della massima verosimiglianza e un metodo a due passi.

In questo caso si decide di stimare il modello con il metodo di massima verosimiglianza, che produce stimatori consistenti e asintoticamente efficienti, con distribuzione asintotica normale.

Il modello può essere interpretato come composto da due parti: la prima descrive il problema della scelta binaria. Il contributo dell'osservazione *i-esima* alla funzione di verosimiglianza è la probabilità di osservare $y_{1i} = 1$ o $y_{1i} = 0$.

La seconda parte descrive la distribuzione dei punteggi per gli studenti che hanno risposto, e per questo il contributo alla verosimiglianza è dato da $f(y_{2i}|y_{1i} = 1)$.

La funzione di log-verosimiglianza è dunque:

$$\log L(\beta, \sigma_2^2, \sigma_{12}) = \sum_{i \in I_0} \log P(y_{1i} = 0) + \sum_{i \in I_1} [\log f(y_{2i}|y_{1i} = 1) + \log P(y_{1i} = 1)] \tag{5.6}$$

La componente di scelta binaria è standard; più complicata è invece la distribuzione condizionale di y_{2i} dato $y_{1i} = 1$. Per questo motivo è più frequente scomporre in maniera diversa la distribuzione congiunta di y_{1i} e y_{2i}, sfruttando la seguente proprietà:

$$f(y_{2i}|y_{1i} = 1)P(y_{1i} = 1) = P(y_{1i} = 1|y_{2i})f(y_{2i})$$

La funzione di log-verosimiglianza può quindi essere riscritta come:

$$\log L(\beta, \sigma_2^2, \sigma_{12}) = \sum_{i \in I_0} \log P(y_{1i} = 0) + \sum_{i \in I_1} [\log f(y_{2i}) + \log P(y_{1i} = 1|y_{2i})] \qquad (5.7)$$

Massimizzando tale funzione rispetto ai parametri ignoti si ottengono stimatori consistenti e asintoticamente efficienti, con distribuzione asintotica normale (Verbeek, 2010).

Passando alla pratica, è necessario specificare che, affinché il modello di selezione non sia debolmente identificato, è preferibile introdurre l'uso di uno strumento, ossia una variabile che influenzi la probabilità di rispondere all'item D14, ma non il punteggio della valutazione. In altre parole, lo strumento deve essere contenuto in X_1 ma non in X_2. Tutte le altre variabili osservate sono solitamente incluse in entrambe le equazioni.
Si stima dunque un modello di Heckman per la valutazione della soddisfazione, con l'aggiunta dell'equazione relativa alla probabilità di risposta. Come strumento si è scelto di inserire il periodo di erogazione del corso: il primo periodo si riferisce al primo semestre o primo/secondo trimestre, ed assume valore pari a 1, mentre il secondo periodo (che assume valore 0) riguarda un'attività erogata nel secondo semestre o terzo trimestre. Questa scelta è conseguenza dell'osservazione, in Tabella 3.18, di una differenza significativa tra la percentuale di rispondenti nei due periodi, presumibilmente dovuta al fatto che al termine del secondo periodo dell'anno accademico gli studenti abbiano iniziato la pausa estiva, con conseguente minor disponibilità di rispondere, o, nel caso di studenti all'ultimo anno, alcuni possano essersi laureati e aver lasciato l'università. Non c'è invece motivo di pensare che la valutazione degli studenti possa dipendere dal periodo in cui un insegnamento viene erogato, rendendo la variabile potenzialmente un valido strumento utilizzabile.
In Tabella 5.3 viene presentato l'output del modello.

Tabella 5.3: Stima del modello di Heckman

Caratteristiche	Variabili	Eq. di selezione		Eq. di regressione	
		Stima	Std. Error	Stima	Std. Error
STUDENTE-GENERALE	frequenza 30-50%	-	-	-1.300***	0.029
	frequenza 50-70%	-	-	-0.748***	0.020
	genere: femmina	0.024***	0.009	0.071***	0.015
	età	-0.213***	0.012	0.407***	0.024
STUDENTE-CARRIERA	in corso	0.281***	0.019	-0.302***	0.039
	media dei voti	0.025***	0.002	0.033***	0.003
	numero medio esami	-0.055***	0.006	-0.078***	0.010
	CFU	0.006***	0.001	0.002*	0.001
	numero esami * numero crediti	0.0002***	0.000	0.0004***	0.000
CORSO	corso obbligatorio	0.193***	0.010	-0.032*	0.017
	numero di ore	0.387***	0.026	0.249***	0.044
	più di un docente	-0.009	0.011	-0.355***	0.018
	sede Padova	-0.090***	0.011	0.041**	0.018
	insegnamento mutuato	-0.008	0.009	-0.112***	0.014
DOCENTE-GENERALE	genere: femmina	-0.004	0.009	0.013	0.015
	età	-0.004***	0.001	-0.017***	0.001
DOCENTE-CARRIERA	professore associato	0.037***	0.011	0.021	0.019
	professore ordinario	0.028*	0.014	0.081***	0.024
	professore a contratto	-0.072***	0.014	-0.082***	0.024
	titolare assegno di ricerca	0.098**	0.046	-0.783***	0.074
STRUMENTO	corso nel primo periodo	0.151***	0.008	-	-
	costante	-0.213***	0.060	7.126***	0.111
	Osservazioni	108908			
	χ^2	5034.69			
	ρ	0.026			
	Log-likelihood Ratio test	0.51			

Innanzitutto si noti che le variabili sulla frequenza delle lezioni sono le uniche presenti solo nell'equazione di regressione e non su quella di selezione, in quanto tutti i rispondenti hanno frequentato più del 30% delle lezioni.
Le stime dei parametri legati alle covariate risultano essere quasi totalmente significative in entrambe le equazioni, compreso lo strumento inserito solo nell'equazione di selezione, ad indicare l'influenza dei fattori osservati nella scelta di non rispondere.
Concentrandosi nell'ultima parte dell'output, sono riportate due statistiche che forniscono informazioni circa l'entità della distorsione causata dalla selezione.
Si osservi il valore del test log-rapporto di verosimiglianza (*log-likelihood ratio test*), che confronta la log-verosimiglianza del modello di selezione completo con la somma delle log-verosimiglianze delle equazioni di selezione e di regressione stimate separatamente: il valore ottenuto è pari a 0.51 ed è non significativo, ad indicare che l'ipotesi nulla di non distorsione da selezione può essere accettata.
La seconda statistica è già stata citata e si riferisce al coefficiente di correlazione tra i termini di errore delle due equazioni, chiamato ρ; esso è molto basso e non significativamente diverso da zero, per cui si arriva alla conclusione che i punteggi delle valutazioni all'item relativo alla soddisfazione degli studenti verso un determinato insegnamento non siano distorti a causa della selezione dei rispondenti ed i dati mancanti non siano presumibilmente generati da un meccanismo MNAR.

Capitolo 6

Imputazione dei dati mancanti

In questo capitolo si considera l'ipotesi che i dati mancanti siano di tipo MAR e si andranno a trattare con alcune delle tecniche di imputazione esposte nel Capitolo 2.
Oltre che nella fase di stima, anche in quella di imputazione si decide di tenere conto della struttura gerarchica dei dati. Si realizza dunque un confronto tra i risultati ottenuti sotto questa assunzione e quelli ricavati senza tenerne conto: nel caso gerarchico i metodi diventano computazionalmente più complicati da applicare, pertanto si vuole vedere se, basandosi su una semplice regressione lineare, si possono ottenere risultati soddisfacenti o se invece la quantità di informazione aggiuntiva considerata dal modello multilivello conduce a performance migliori.
In primo luogo si decide di applicare tre tecniche di imputazione all'item relativo alla soddisfazione complessiva considerando la struttura uni-livello del dataset (non gerarchica), e di osservare i possibili cambiamenti nella sua distribuzione, rispetto a quella originale riferita a 80093 osservazioni: si tratta quindi di imputare 32613 valori, per arrivare alla totalità del dataset.
Ogni tecnica di imputazione, basandosi su approcci diversi, produce diversi valori imputati e tanto meno l'aggiunta dei nuovi valori distorce la distribuzione osservata dei dati, tanto più la tecnica si può considerare valida.

Le tre tecniche utilizzate sono:

- imputazione con la media;
- imputazione con regressione;
- Predictive Mean Matching (PMM).

Tranne la prima, applicata semplicemente sostituendo i valori mancanti con la media calcolata sul totale delle risposte date, le altre due si basano su una regressione lineare della variabile di interesse (item D14) con le variabili esplicative menzionate in precedenza, ossia quelle relative a caratteristiche del corso (corso obbligatorio, numero totale di ore, numero docenti coinvolti, locazione del corso, insegnamento mutuato), dello studente (frequenza, genere, età), della carriera (CFU acquisiti all'anno, media dei voti

nell'A.A. 2012/13, numero medio di esami passati all'anno, stato di iscrizione), del docente (genere, età), della carriera dei docenti (ruolo) , nonché i Corsi di Studio.
In particolare, come descritto nel Capitolo 2, il PMM associa l'utilizzo di questa regressione con la tecnica del *nearest neighbour*, in modo da imputare il valore mancante con il valore osservato avente media predittiva, estratta dal modello di regressione, "più vicina".
Per far questo è stato implementato un programma che riuscisse a trovare, per ogni unità ricevente (con valore mancante), un donatore tra le 80093 possibili unità osservate, in modo da minimizzare la distanza di Mahalanobis tra le rispettive medie predette. Ogni *missing data* è stato poi sostituito con il valore osservato nel donatore scelto (si veda il procedimento descritto nel Capitolo 2).

In secondo luogo si tiene conto della struttura gerarchica delle osservazioni per imputare i valori mancanti: come già sottolineato, si vuole valutare se i metodi di imputazione offrono risultati migliori, o comunque diversi, rispetto a quelli conseguiti senza tenere conto della gerarchia stessa.
Le tre tecniche menzionate vengono quindi estese in questa direzione.
La prima si basa sul calcolo della media dei punteggi assegnati all'item D14 all'interno di ogni gruppo, quindi i valori mancanti delle unità vengono sostituiti con la media relativa al gruppo a cui esse appartengono. La seconda e terza tecnica funzionano come in precedenza, con la differenza di stimare un modello gerarchico ad intercetta casuale tra la variabile di interesse e le stesse variabili esplicative utilizzate per la regressione, con l'aggiunta delle medie di gruppo delle variabili di primo livello, cioè relative agli studenti.
Per essere più chiari, il metodo PMM multilivello può essere riassunto nei seguenti passi:

1. si stima un modello ad intercetta casuale tra l'item D14 e le variabili esplicative (con aggiunta degli effetti contestuali);
2. date le stime ottenute nel passo 1 si calcola la media predittiva, sia delle unità riceventi sia di quelle donatrici, composta da una parte fissa (per ogni osservazione) e da una casuale riferita ai residui di ogni gruppo (u_{oj}).
3. per ogni unità ricevente viene selezionata un'unità donatrice in modo da minimizzare la distanza di Mahalanobis tra le rispettive medie predette;
4. ogni unità incompleta viene imputata trasferendole il punteggio appartenente al donatore selezionato al punto 3.

Si specifica che in letteratura non è presente questa applicazione del PMM tramite modelli multilivello, la quale quindi viene introdotta, come preludio per ulteriori sviluppi, in questa sede[2].

Nelle Figure 6.1 e 6.2 si riportano i grafici relativi alla distribuzione dei valori delle medie predette dai modelli (gerarchico e non), su cui si basano le imputazioni; si osserva una maggior variabilità in quelle prodotte dal modello multilivello, in quanto si tiene conto della variabilità tra i gruppi.

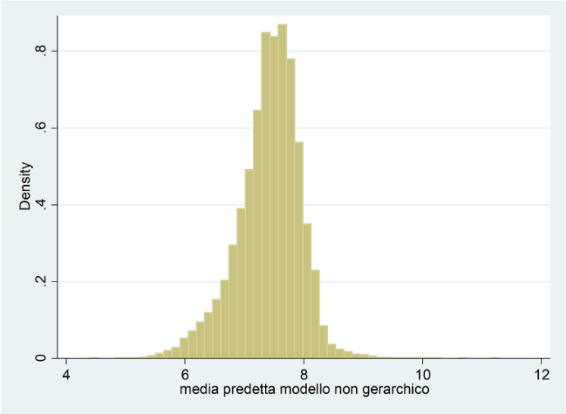

Figura 6.1: Distribuzione media predetta modello non gerarchico

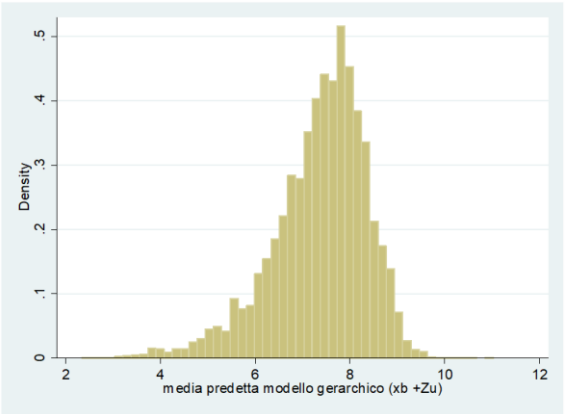

Figura 6.2: Distribuzione media predetta modello gerarchico

In Tabella 6.1 si presentano le statistiche descrittive riferite all'item D14, per i diversi metodi; nella prima colonna vengono riportate quelle riferite alla variabile originale, prima dell'imputazione.

Successivamente sono riportati anche gli istogrammi (Figure 6.4 - 6.9) dove si osserva la distribuzione dei punteggi assegnati all'item D14 ottenuta a seguito delle diverse imputazioni. In Figura 6.3 invece vi è la distribuzione della variabile originale.

[2] Vink et al. (2015) hanno introdotto il Partitioned Predictive Mean Matching, in cui la struttura gerarchica viene sfruttata per migliorare il metodo PMM, ma nessun modello multilivello viene direttamente stimato.

Tabella 6.1: Statistiche descrittive valori completi item D14, per diversi metodi di imputazione

		STATISTICHE DESCRITTIVE D14 IMPUTATA					
	ORIGINALE	**MEAN**	**REG**	**PMM**	**MEAN GER.**	**REG GER.**	**PMM GER.**
Oss.	80093	112706	112706	112706	112503	112503	112503
Mediana	8	7.32	7.72	8	7.56	7.76	8
Media	7.32	7.32	7.41	7.36	7.29	7.36	7.35
Dev. Std.	1.96	1.65	1.67	1.95	1.75	1.73	1.94
Min	1	1	1	1	1	1	1
Max	10	10	11.29	10	10	11.03	10

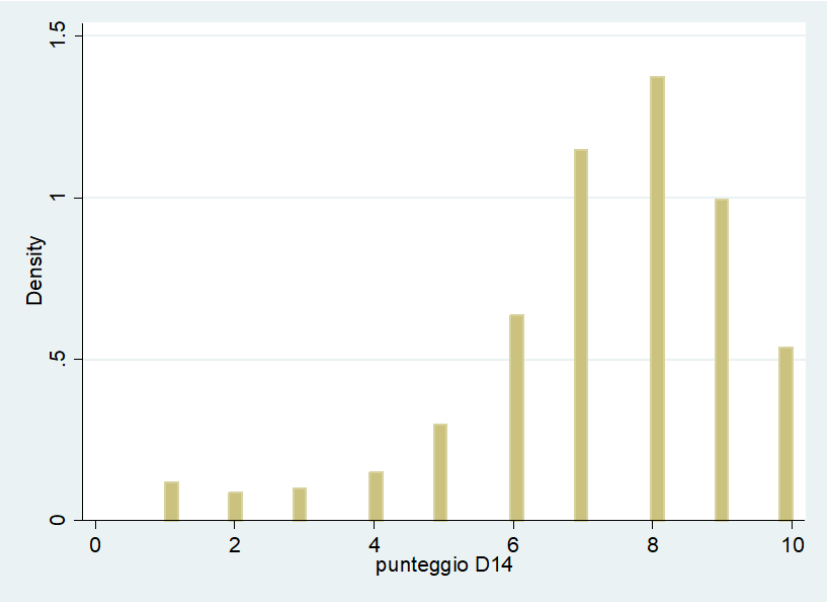

Figura 6.3: Distribuzione punteggi assegnati alla variabile D14 originale

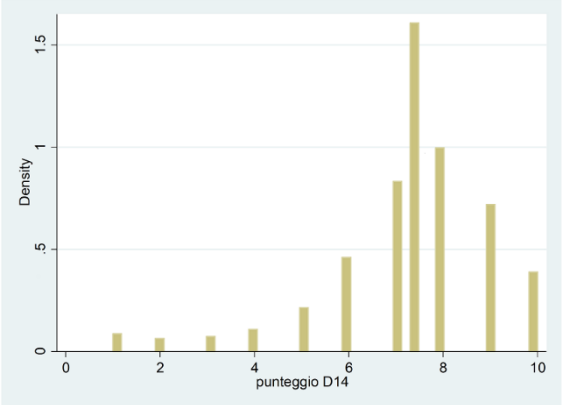

Figura 6.4: Distribuzione punteggi assegnati alla variabile D14 imputata con metodo della media - caso non gerarchico

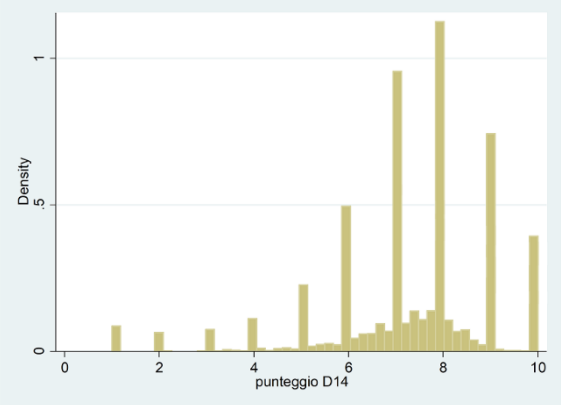

Figura 6.5: Distribuzione punteggi assegnati alla variabile D14 imputata con metodo della media - caso gerarchico

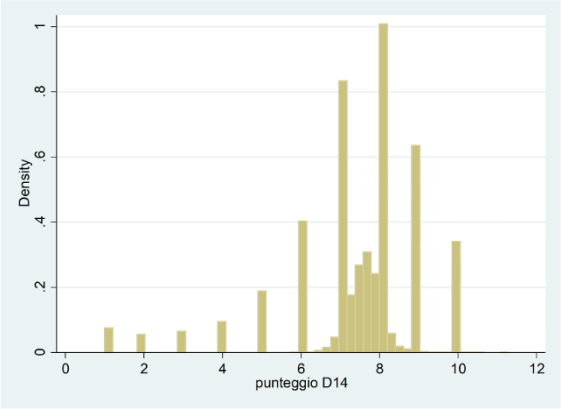

Figura 6.6: Distribuzione punteggi assegnati alla variabile D14 imputata con metodo della regressione - caso non gerarchico

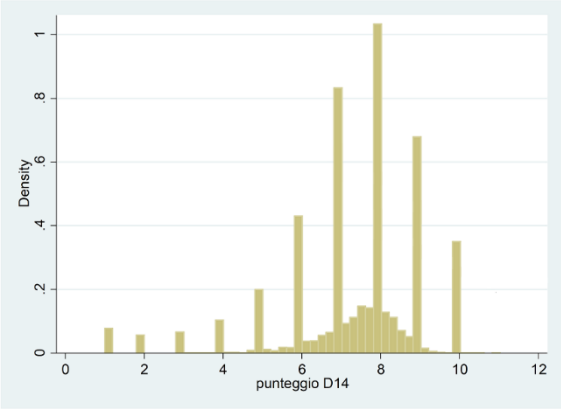

Figura 6.7: Distribuzione punteggi assegnati alla variabile D14 imputata con metodo della regressione - caso gerarchico

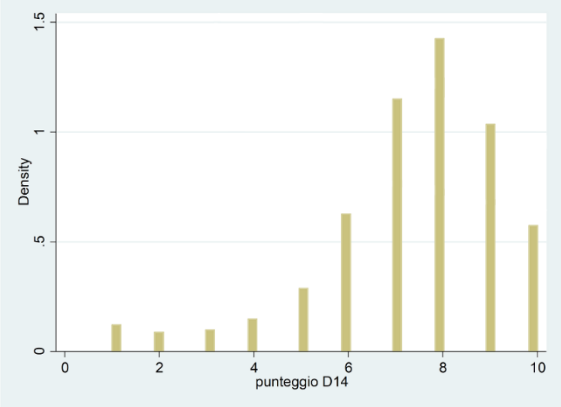

Figura 6.8: Distribuzione punteggi assegnati alla variabile D14 imputata con metodo PMM - caso non gerarchico

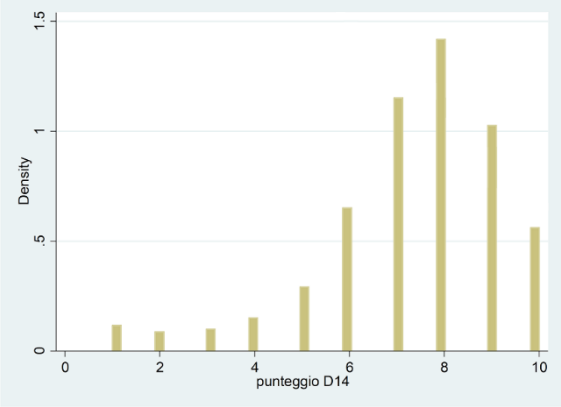

Figura 6.9: Distribuzione punteggi assegnati alla variabile D14 imputata con metodo PMM - caso gerarchico

Per quanto riguarda le tecniche applicate senza tener conto della gerarchia dei dati, l'imputazione tramite la media calcolata sul totale dei rispondenti produce una distorsione nella distribuzione della variabile, creando un picco in corrispondenza del suo valor medio (Figura 6.4) e inoltre si verifica una sottostima della variabilità con una deviazione standard più bassa dell'originale (Tabella 6.1).

Anche l'imputazione tramite regressione produce distorsione nella distribuzione della D14 (Figura 6.6), dato che i valori mancanti sono sostituiti con quelli predetti dalla regressione; si verifica inoltre l'imputazione di valori non ammissibili, fuori dal range [1,10], come si può vedere osservando il valore massimo nella Tabella 6.1. Anche in questo caso la variabilità è sottostimata.

La tecnica del PMM è quella che produce risultati più vicini all'originale: mediana e deviazione standard sono uguali, così come il range di valori; solo la media è leggermente

più alta. La distribuzione dei valori osservati viene quindi preservata nella parte mancante dei dati, osservando la Figura 6.8, e questo pertanto sembra essere, tra i tre, il metodo più efficace.

Sull'imputazione conseguita tenendo conto della struttura gerarchica dei dati, nella Tabella 6.1 si nota che in totale si ottengono 112503 osservazioni, cioè 203 in meno rispetto ai casi precedenti, e questo avviene poiché queste osservazioni appartengono a gruppi in cui tutte le unità hanno un valore mancante nell'item D14 e pertanto non si è potuta calcolare la media all'interno del gruppo (per l'imputazione tramite media) e la media predetta dal modello gerarchico sottostante l'imputazione (per le restanti tecniche).
Il metodo della media e della regressione anche in questo caso producono distorsione nella distribuzione della variabile (Figura 6.5 e Figura 6.7) e la seconda anche un valore di massimo non ammissibile; in entrambe la variabilità viene sottostimata (Tabella 6.1).
Per essere precisi, in entrambi i casi (gerarchico e non), nemmeno entro il range di valori ammissibili le tecniche della media e della regressione producono valori imputati accettabili, ossia numeri interi da 1 a 10, ma punteggi intermedi, non effettivamente utilizzabili dagli studenti in sede di valutazione.
Per questo motivo i punteggi imputati vengono arrotondati all'intero più vicino e si riportano nuovamente gli istogrammi della distribuzione della variabile con i nuovi valori (Figure 6.10-6.13).
Si osserva come la distorsione nelle distribuzioni sia mantenuta, con una forma più appuntita (sono evidenti dei picchi più marcati).

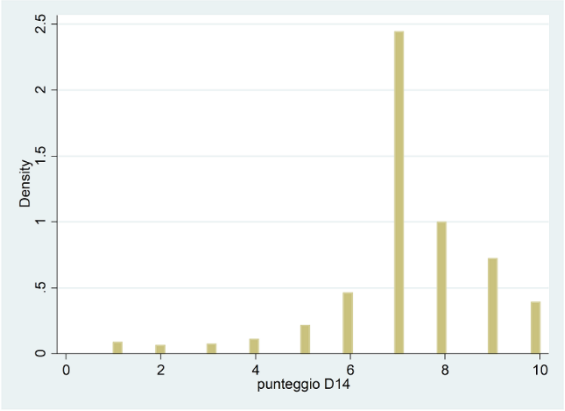

Figura 6.10: Distribuzione punteggi assegnati alla variabile D14 imputata con metodo della media (arrotondati) - caso non gerarchico

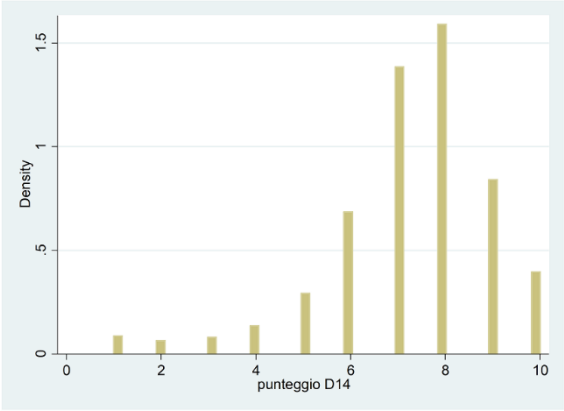

Figura 6.11: Distribuzione punteggi assegnati alla variabile D14 imputata con metodo della media (arrotondati) - caso gerarchico

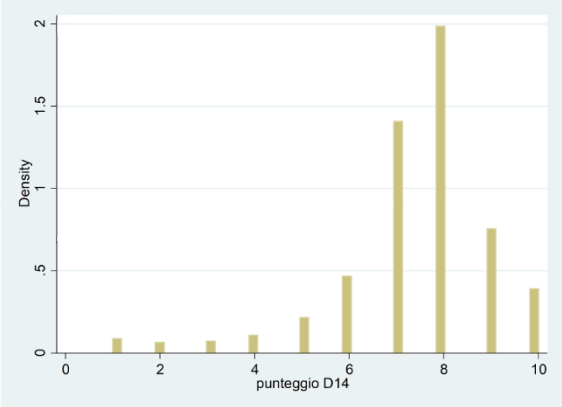

Figura 6.12: Distribuzione punteggi assegnati alla variabile D14 imputata con metodo della regressione (arrotondati)- caso non gerarchico

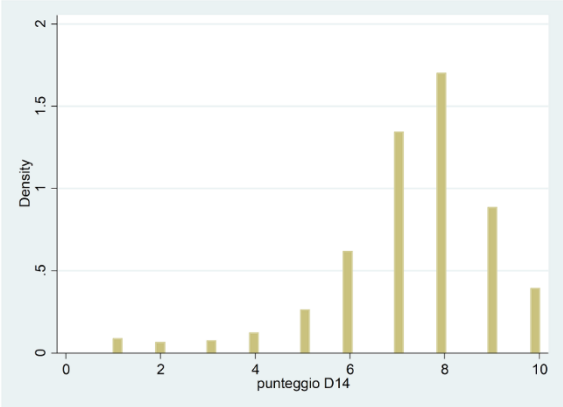

Figura 6.13: Distribuzione punteggi assegnati alla variabile D14 imputata con metodo della regressione (arrotondati)- caso gerarchico

Dalla Figura 6.9 invece emerge come il metodo PMM produca risultati praticamente uguali alla variante non gerarchica. Si tratta di un risultato un po' inatteso poiché, dato che la relativa media predetta tiene conto della variabilità tra i gruppi, ci si aspettava un'imputazione "migliore" o comunque leggermente diversa dal caso non gerarchico: con il calcolo del *nearest neighbour* quindi le imputazioni nei due casi portano circa allo stesso esito.

Fino a questo punto si può dunque affermare che, nonostante il metodo PMM applicato a dati gerarchici funzioni bene, tutto sommato, sfruttando una semplice regressione lineare, si potrebbero ottenere gli stessi soddisfacenti risultati a livello marginale.

A tal proposito, in Tabella 6.2 si riporta un confronto tra i valori imputati dal PMM "gerarchico" e "non gerarchico", in cui si può chiaramente vedere come le quantità marginali siano all'incirca le stesse.

Per analizzare il livello di concordanza tra i singoli punteggi, in Tabella 6.3 si riportano le percentuali per riga della Tabella 6.2, per vedere in che modo ogni singolo punteggio imputato dal PMM non gerarchico sia stato classificato nel caso gerarchico; si osserva, per esempio, in quale percentuale tutti gli NA imputati con valore 1 dal metodo non gerarchico vengano distribuiti con l'altro approccio, e così via per ogni punteggio.

Tabella 6.2: Confronto punteggi imputati tra metodo PMM non gerarchico e gerarchico

		D14 PMM gerarchico											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Tot**	**%**
D14 PMM non gerarchico	**1**	320	3	3	10	22	51	84	106	53	44	696	**2.13**
	2	2	232	3	7	19	35	42	68	45	23	476	**1.46**
	3	3	2	245	7	14	31	61	62	58	27	510	**1.56**
	4	9	8	5	389	26	45	87	88	100	39	796	**2.44**
	5	8	9	15	26	656	119	181	205	165	81	1465	**4.49**
	6	42	25	26	42	95	1669	381	498	390	207	3375	**10.35**
	7	63	70	57	84	155	457	3336	1015	764	457	6458	**19.80**
	8	84	45	93	122	255	658	1032	4746	1079	580	8694	**26.66**
	9	53	54	49	101	195	526	788	1098	3055	484	6403	**19.63**
	10	28	33	31	56	106	289	486	660	512	1539	3740	**11.47**
	Tot	612	481	527	844	1543	3880	6478	8546	6221	3481	32613	**100**
	%	**1.88**	**1.47**	**1.62**	**2.59**	**4.73**	**11.90**	**19.86**	**26.20**	**19.08**	**10.67**	**100**	

Tabella 6.3: Percentuali per riga della Tabella 6.2

		D14 PMM gerarchico										
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Tot**
D14 PMM non gerarchico	**1**	**45.98**	0.43	0.43	1.44	3.16	7.33	12.07	15.23	7.61	6.32	100
	2	0.42	**48.74**	0.63	1.47	3.99	7.35	8.82	14.29	9.45	4.83	100
	3	0.59	0.39	**48.04**	1.37	2.75	6.08	11.96	12.16	11.37	5.29	100
	4	1.13	1.01	0.63	**48.87**	3.27	5.65	10.93	11.06	12.56	4.90	100
	5	0.55	0.61	1.02	1.77	**44.78**	8.12	12.35	13.99	11.26	5.53	100
	6	1.24	0.74	0.77	1.24	2.81	**49.45**	11.29	14.76	11.56	6.13	100
	7	0.98	1.08	0.88	1.30	2.40	7.08	**51.66**	15.72	11.83	7.08	100
	8	0.97	0.52	1.07	1.40	2.93	7.57	11.87	**54.59**	12.41	6.67	100
	9	0.83	0.84	0.77	1.58	3.05	8.21	12.31	17.15	**47.71**	7.56	100
	10	0.75	0.88	0.83	1.50	2.83	7.73	12.99	17.65	13.69	**41.15**	100

Se a livello marginale i due metodi sono quasi del tutto equivalenti, la concordanza tra i singoli punteggi risulta essere del 50% circa e in certi casi addirittura a punteggi bassi di un metodo corrispondono punteggi alti dell'altro, e viceversa.

Con una concordanza di questo tipo e dato che in entrambi i casi le strutture di correlazione vengono rispettate (data la forma delle distribuzioni), si può preliminarmente preferire il metodo PMM gerarchico, in quanto tiene conto della forte variabilità tra gruppi, quindi di maggiore informazione.

Per approfondire si osservano anche le distanze di Mahalanobis tra i valori delle medie predette di unità riceventi e donatrici ed emerge come i casi di concordanza dei punteggi

presentino per la maggior parte delle volte (il 78%) una distanza pari a zero, sia con il metodo gerarchico che con il non gerarchico.
Nei casi in cui la distanza è zero significa che viene scelta come unità donatrice la prima unità avente distanza nulla; tuttavia, non è da escludere che nel corso del ciclo implementato ci possano essere altre unità donatrici aventi distanza pari a zero, che però non vengono scelte solo a causa dell'ordine con cui vengono calcolate le distanze. In questo caso potrebbe essere utile utilizzare più *nearest neighbours*, ossia anziché soltanto uno, considerare per ogni unità ricevente, ad esempio, cinque donatori ed estrarne poi uno casualmente.

6.1 Imputazione con metodo PMM - 5 nearest neighbours

La procedura per l'implementazione del metodo del *Predictive Mean Matching* con cinque *nearest neighbours* (NN) è la medesima fatta nel caso di un NN, con una modifica nella selezione del donatore finale; essa risulta essere computazionalmente più complicata da effettuare. Per riepilogare, consiste nel:

1. stimare un modello di regressione lineare/multilivello di Y (punteggio item D14) su X (insieme delle usuali variabili esplicative) e, solo nel caso gerarchico, si aggiungono le medie di gruppo;
2. date le stime ottenute nel punto 1, si calcola la media predetta sia per le unità complete (donatori) che per quelle con valore mancante (riceventi);
3. per ogni unità ricevente vengono selezionati cinque donatori, in modo da minimizzare la distanza di Mahalanobis tra le medie predette; vengono quindi scelti i cinque donatori "più vicini";
4. in maniera casuale viene selezionato un donatore tra i cinque;
5. ogni unità incompleta viene imputata trasferendo il valore della variabile Y dal donatore selezionato al punto 4.

Questa procedura permette di gestire il problema legato alle distanze nulle tra le medie predette dell'unità ricevente e di alcuni donatori: nel caso in cui ci fosse più di un donatore avente media predetta con distanza pari a zero rispetto alla media dell'unità ricevente, non si seleziona necessariamente il primo nell'ordine di analisi, escludendo

automaticamente tutti gli altri, ma si considerano i primi cinque, selezionandone poi uno casualmente.
Si vuole vedere, in questo modo, se e come cambia la concordanza tra i punteggi dei metodi PMM gerarchico e non gerarchico.

In Tabella 6.4 si riportano le statistiche descrittive delle nuove imputazioni, in aggiunta a quelle riferite al caso di un NN, in cui si osserva un po' più di differenza, in termini di media e deviazione standard, tra i risultati del PMM gerarchico e non gerarchico con cinque NN; la procedura quindi ha rispettato le aspettative, confermando la "distorsione" delle imputazioni dovuta alla selezione di un solo *nearest neighbour*, causata dalle distanze di Mahalanobis nulle.

Tabella 6.4: Statistiche descrittive valori completi dell'item D14 con metodo PMM (1 e 5 nearest neighbours) nel caso gerarchico e non gerarchico

	Originale	**PMM non gerarchico 1 NN**	**PMM gerarchico 1 NN**	**PMM non gerarchico 5 NN**	**PMM gerarchico 5NN**
Oss.	80093	112706	112503	112706	112503
Mediana	8	8	8	8	8
Media	7.32	7.36	7.35	7.31	7.36
Dev. Std.	1.96	1.95	1.95	1.97	1.94
Min	1	1	1	1	1
Max	10	10	10	10	10

Nelle Figure 6.14 e 6.15 si osservano gli istogrammi dei punteggi della variabile imputata, in cui continua ad essere preservata la forma della distribuzione sia tra i due metodi, che con la variabile originale (Figura 6.3).

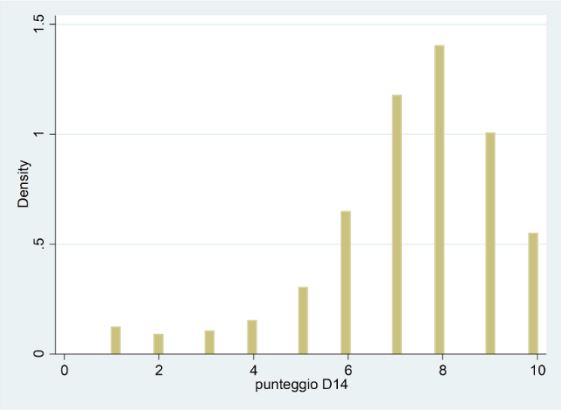

Figura 6.14: Distribuzione punteggi assegnati alla variabile D14 imputata con metodo PMM - caso non gerarchico 5NN

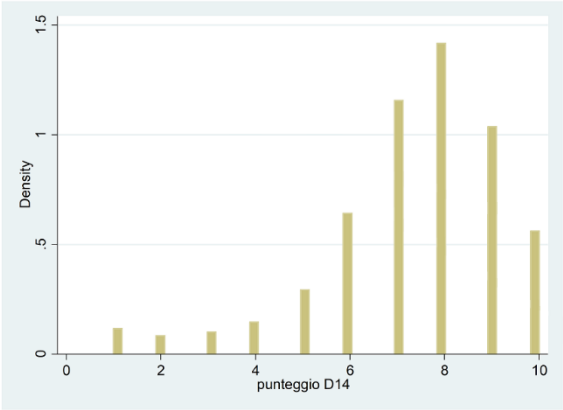

Figura 6.15: Distribuzione punteggi assegnati alla variabile D14 imputata con metodo PMM - caso gerarchico 5NN

La concordanza tra i punteggi, a conferma della maggior differenza già emersa con le statistiche descrittive, risulta essere ora pari al 17% circa, quindi molto più bassa del 50% ottenuto in precedenza; sono leggermente diverse anche le quantità marginali, con punteggi imputati più alti per il PMM gerarchico, il quale infatti ha una media leggermente più alta (Tabella 6.5).

Tabella 6.5: Confronto punteggi imputati tra metodo PMM nel caso non gerarchico e caso gerarchico con 5NN

		D14 PMM gerarchico 5 NN											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Tot**	**%**
D14 PMM non gerarchico 5 NN	**1**	15	9	18	12	43	88	148	230	145	98	806	**2.47**
	2	8	11	8	14	28	72	108	164	111	46	570	**1.75**
	3	12	11	7	16	35	74	131	186	107	68	647	**1.98**
	4	11	9	12	27	48	107	167	260	184	87	912	**2.80**
	5	33	30	28	46	69	203	361	480	352	191	1793	**5.50**
	6	75	48	74	87	171	419	777	953	754	411	3769	**11.56**
	7	153	84	101	178	358	772	1368	1813	1358	718	6903	**21.17**
	8	154	122	126	181	400	912	1587	2095	1634	922	8133	**24.94**
	9	107	89	98	151	289	643	1226	1545	1124	629	5901	**18.09**
	10	51	40	67	77	136	374	658	805	639	332	3179	**9.75**
	Tot	619	453	539	789	1577	3664	6531	8531	6408	3502	32613	**100**
	%	**1.90**	**1.39**	**1.65**	**2.42**	**4.84**	**11.23**	**20.03**	**26.16**	**19.65**	**10.74**	**100**	

6.2 Imputazione multipla

Un'altra possibile soluzione è l'implementazione di un'imputazione multipla nel caso gerarchico, sfruttando la stima della distribuzione dei residui di gruppo, u_{oj}, in modo da ottenere *m* possibili variabili imputate con il metodo PMM. Per far questo è stato scritto appositamente un programma, computazionalmente impegnativo, che ripetesse per *m* volte quanto fatto singolarmente, con l'eccezione di sfruttare di volta in volta un diverso valore di media predetta.

Procedendo con questa opzione si è deciso di porre *m* = 5 e, una volta stimato l'usuale modello gerarchico ad intercetta casuale, si estrae la previsione lineare degli effetti fissi, per ogni osservazione.

Sapendo che i residui di secondo livello, u_{oj}, hanno distribuzione Normale, di media zero e varianza σ^2_{u0}, si estraggono casualmente cinque valori da questa distribuzione: osservazioni appartenenti allo stesso gruppo hanno quindi lo stesso valore predetto dell'effetto casuale.

Sommando poi ogni effetto fisso a quello casuale si ottengono cinque diverse medie predette per ogni osservazione ed ognuna viene coinvolta nell'applicazione del metodo PMM per imputare i *missing data*.

Si riportano in la Tabella 6.6 le statistiche descrittive relative alle cinque imputazioni.

Tabella 6.6: Statistiche descrittive D14 – Imputazione multipla con m=5 (i=1, …, 5)

	STATISTICHE DESCRITTIVE IMPUTAZIONE MULTIPLA				
	i=1	**i=2**	**i=3**	**i=4**	**i=5**
Oss.	112503	112503	112503	112503	112503
Mediana	8	8	8	8	8
Media	7.330	7.329	7.322	7.329	7.335
Dev. Std.	1.964	1.964	1.961	1.962	1.959
Min	1	1	1	1	1
Max	10	10	10	10	10

Dai valori in Tabella 6.6 emerge subito la forte somiglianza, o meglio uguaglianza, tra i cinque casi, nonostante i valori delle medie predette con cui sono stati ricavate le imputazioni fossero abbastanza differenti.

Tuttavia, osservando i donatori, nei cinque casi è presente una concordanza del 38% (ossia l'unità ricevente ha lo stesso donatore nelle cinque imputazioni), con il 40% di

concordanza tra i punteggi. Marginalmente quindi le cinque applicazioni si equivalgono, mentre si presentano differenze nei singoli punteggi imputati.

Con riferimento alla sola prima imputazione, dalla Figura 6.16 si può vedere come la distribuzione della variabile venga preservata; lo stesso risultato è emerso per le altre imputazioni (i = 2, 3, 4, 5), di cui quindi non si riportano i grafici.

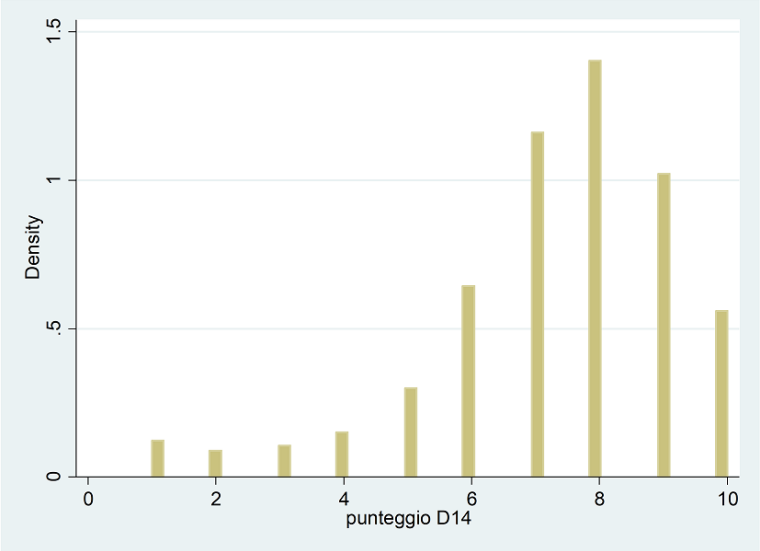

Figura 6.16: Distribuzione punteggi assegnati alla variabile D14 con imputazione multipla per i=1 (metodo PMM gerarchico)

Confrontando questi risultati con quelli relativi all'imputazione "non gerarchica" applicata in precedenza (con un *nearest neighbour*), dalla Tabella 6.7 emerge una concordanza tra i punteggi imputati del 35%, più bassa rispetto al 50% precedente (osservato nella Tabella 6.3). Un calo della concordanza tuttavia era da attendersi, data la variazione del tipo di soluzione proposta.

Si noti che la Tabella 6.7 è riferita soltanto alla prima imputazione e, dato che è stata verificata all'incirca l'uguaglianza con le restanti quattro, non si riportano anche le relative tabelle.

Tabella 6.7: Confronto punteggi imputati tra metodo PMM nel caso non gerarchico e caso gerarchico con imputazione multipla (i=1)

		D14 PMM gerarchico i=1											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Tot**	**%**
D14 PMM non gerarchico	**1**	194	20	13	13	30	68	134	106	75	43	696	**2.13**
	2	11	125	6	18	19	53	63	75	62	44	476	**1.46**
	3	10	11	136	13	22	46	76	106	57	33	510	**1.56**
	4	10	8	8	194	34	70	175	161	92	44	796	**2.44**
	5	24	24	23	34	417	136	233	313	173	88	1465	**4.49**
	6	59	45	70	75	144	1086	587	603	472	234	3375	**10.35**
	7	100	74	112	128	241	586	2405	1399	927	486	6458	**19.80**
	8	164	98	120	161	387	751	1431	3509	1276	797	8694	**26.66**
	9	94	73	98	151	263	570	972	1237	2324	621	6403	**19.63**
	10	66	28	58	65	143	339	580	743	655	1063	3740	**11.47**
	Tot	732	506	644	852	1700	3705	6656	8252	6113	3453	32613	**100**
	%	**2.24**	**1.55**	**1.97**	**2.61**	**5.21**	**11.36**	**20.41**	**25.30**	**18.74**	**10.59**	**100**	

Nonostante l'introduzione di questa sorta di imputazione multipla abbia dei vantaggi e abbia portato a risultati attesi e validi, non si ritiene che essi apportino un miglioramento tale da poter essere preferita ai metodi precedenti, anche per il fatto di essere computazionalmente più complicata da realizzare; a parità di risultati si decide quindi di privilegiare le soluzioni meno onerose.

6.3 Studio di simulazione

Con le analisi finora realizzate non è possibile definire il metodo di imputazione migliore, ma soltanto notare come il metodo del *Predictive Mean Matching* sia il più efficace, in quanto i risultati prodotti non hanno provocato una distorsione nella distribuzione della variabile originale. In particolare, la sua applicazione con un modello multilivello è parsa la più completa, poiché ha tenuto conto dell'informazione relativa alla forte eterogeneità tra i gruppi.

Per poter dire che un metodo di imputazione è preferibile rispetto ad un altro però sarebbe necessario osservare quanto i valori imputati si avvicinino alla realtà e per far questo si è pensato di introdurre uno studio di simulazione, per poter imputare un certo numero di valori noti della variabile di interesse e quindi attuare poi un confronto tra valori reali e valori imputati.

Si è deciso di focalizzare tale studio sul metodo PMM ad imputazione singola, tralasciando l'imputazione multipla per i motivi espressi al termine del precedente paragrafo; il lavoro però può, ad esempio, fornire le basi per approfondimenti anche in quell'ambito.

Il procedimento è il seguente, implementato tramite un programma scritto appositamente a partire dalla sintassi del software Stata:

1. dal dataset sulle valutazioni degli studenti a disposizione si estrae casualmente un numero arbitrario di osservazioni (50000) dall'insieme di rispondenti all'item D14;
2. dalle 50000 osservazioni si eliminano casualmente 15000 valori della risposta all'item D14; questa quantità corrisponde al 30% del totale, ossia all'incirca la stessa percentuale di dati mancanti osservata nel dataset originale;
3. si imputano i 15000 valori con il metodo PMM, gerarchico e non gerarchico, con uno e cinque *nearest neighbours*;
4. per ognuno si calcola la percentuale di concordanza dei valori imputati rispetto ai valori esatti e noti dell'item D14 e su quelli esatti con una discordanza di un punto, in negativo e in positivo (punteggio esatto +1 o -1); si realizzano poi dei box plot per riassumere graficamente i risultati.

Di norma sarebbe opportuno fare 1000 simulazioni, tuttavia, essendo il programma computazionalmente complesso, soprattutto per quanto riguarda l'applicazione dei metodi con 5 *nearest neighbours*, in cui ogni ciclo ha dei tempi abbastanza lunghi, si è

deciso di realizzarne 250; i risultati ottenuti, come si vedrà, saranno comunque abbastanza chiari, quindi si presume che l'aumento del numero di simulazioni non avrebbe portato particolari miglioramenti.

Inoltre, è necessario specificare che la procedura implementata simula un meccanismo MCAR come generatore dei dati mancanti, in quanto essi vengono "creati" in maniera completamente casuale; ciò può sembrare in contraddizione con quanto affermato finora, ma, come già ricordato, questo studio di simulazione vuole essere inteso come punto di partenza, da cui poter prendere spunto per ulteriori lavori futuri, in cui potrà essere migliorato ed arricchito.

In questa sede si vuole capire se, partendo con una simulazione di questo tipo, si riesce già ad individuare in maniera abbastanza netta il metodo di imputazione più efficace e capire quale effetto abbia nelle stime dei modelli.

6.3.1 Risultati

In Figura 6.17 si riportano i box plot relativi al confronto esatto tra valori imputati e valori reali del punteggio dell'item D14, per i metodi PMM gerarchico e PMM non gerarchico, con un *nearest neighbour*. Emerge quindi la percentuale di concordanza delle due tecniche, con una netta differenza tra esse: il metodo applicato considerando la gerarchia dei dati, come ci si aspettava, sembra funzionare meglio, confermando che l'informazione aggiuntiva di cui esso tiene conto permette un maggior avvicinamento dei valori imputati a quelli reali.

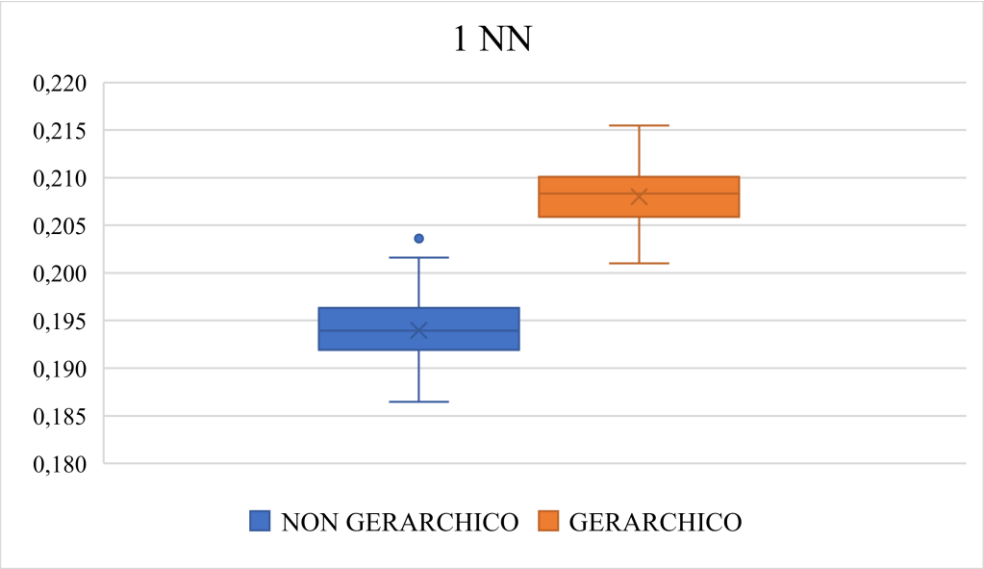

Figura 6.17: Confronto esatto tra valori imputati e reali dell'item D14 nei due metodi (1 NN)

Si è deciso di andare a verificare anche la concordanza con un margine di errore di un punto, in positivo o in negativo.

Partendo con un confronto interno nei due metodi, nelle Figure 6.18 e 6.19 non emerge una differenza evidente tra l'errore unitario in negativo e in positivo, in entrambi i casi; ciò significa che i due metodi non tendono a sovrastimare o sottostimare in maniera particolare, ma "sbagliano" di un punto, in positivo e in negativo, all'incirca nella stessa misura.

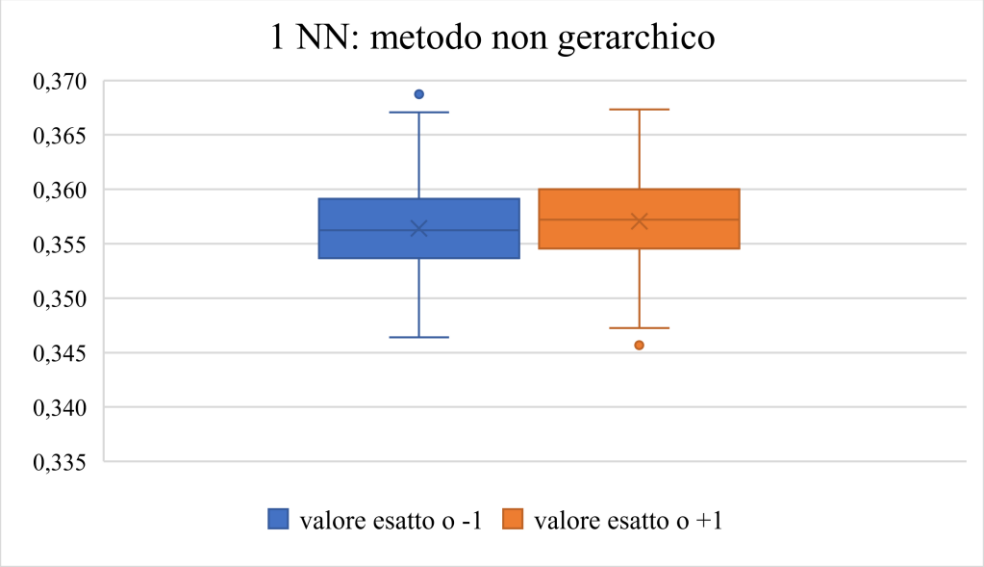

Figura 6.18: Concordanza con errore di un punto in positivo o in negativo – metodo non gerarchico

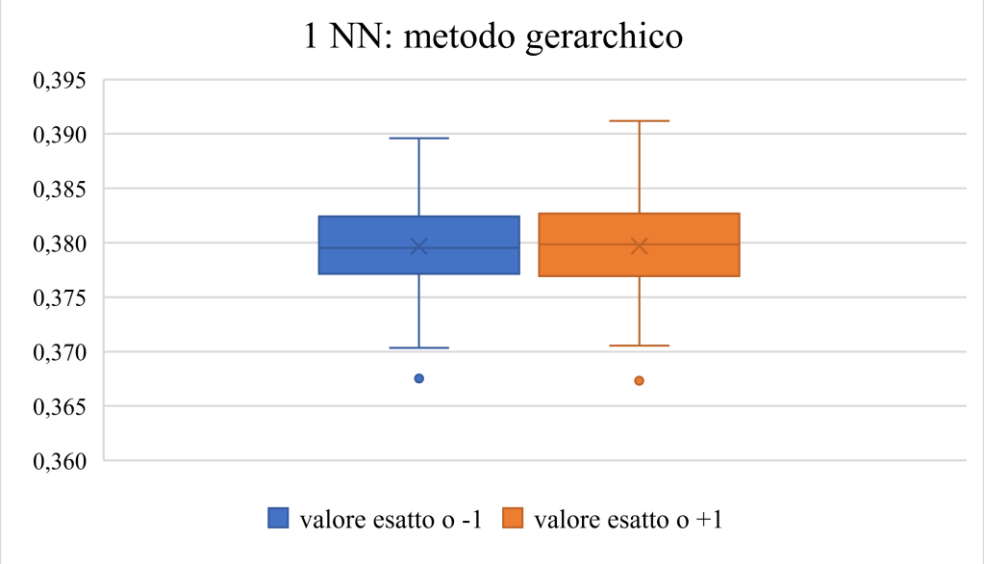

Figura 6.19: Concordanza con errore di un punto in positivo o in negativo – metodo gerarchico

Dal confronto tra i valori imputati dai due metodi, con la possibilità di errore di un punto in negativo (Figura 6.20) o in positivo (Figura 6.21), si osserva invece una differenza marcata, ancor più di quella presente in Figura 6.17.

Si evince infatti che il metodo PMM gerarchico tende a sottostimare o sovrastimare di un punto il punteggio esatto in percentuale maggiore rispetto al PMM non gerarchico; quest'ultimo dunque imputa i punteggi commettendo errori quantitativamente maggiori di un punto.

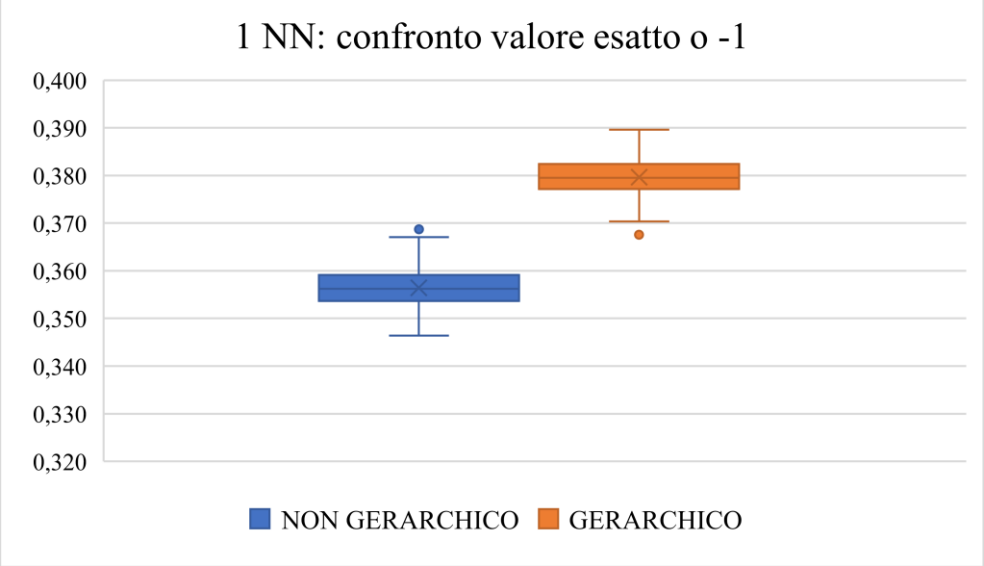

Figura 6.20: Confronto esatto o con errore di un punto in negativo tra i due metodi

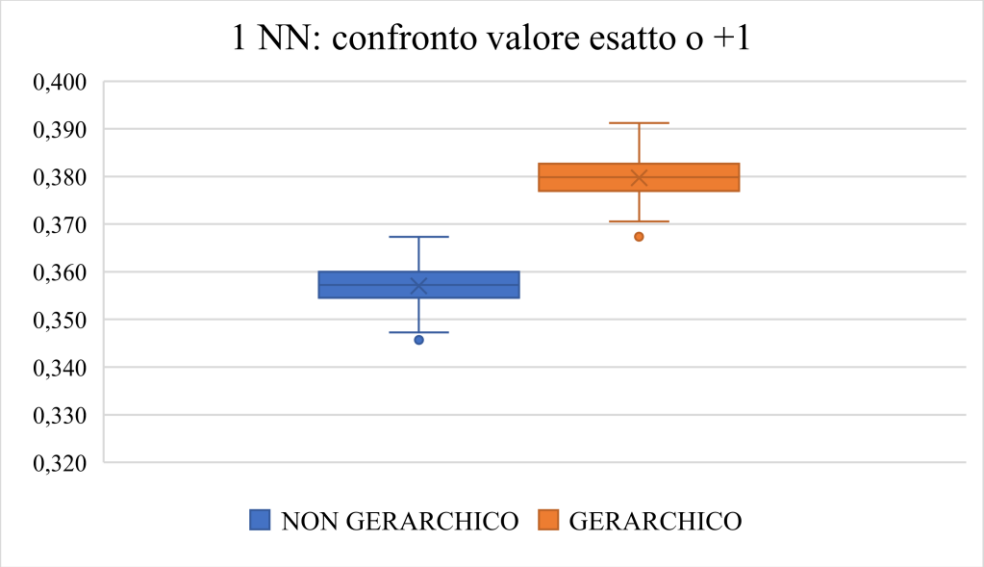

Figura 6.21: Confronto esatto o con errore di un punto in positivo tra i due metodi

Infine, si riportano in Figura 6.22 i box plot per il confronto tra i due metodi tenendo conto della concordanza esatta, con un margine di errore unitario sia in positivo che in negativo (valore reale del punteggio item D14 ± 1): si osserva come la differenza tra PMM gerarchico e non gerarchico in questo caso sia ancora più marcata.

Numericamente si parla di un distacco del 3%, che è un valore relativamente basso ma comunque atteso; l'insieme di variabili inserite nel modello sottostante l'imputazione gerarchica infatti potrebbe essere ampliato, inserendo per esempio una serie di caratteristiche relative ai docenti, ricavabili, come già detto, dal questionario PRODID, o semplicemente alcune variabili relative agli studenti, che in sede di interpretazione non avrebbero un significato ragionevole, ma porterebbero informazione aggiuntiva nel modello gerarchico (si pensi ad esempio alla provincia di residenza). In questo modo la differenza tra i due metodi potrebbe diventare quantitativamente più rilevante, con un aumento della concordanza tra i valori imputati dal PMM gerarchico ed i valori reali.

La variabilità in entrambi i casi è comunque molto bassa, a conferma della validità dei metodi.

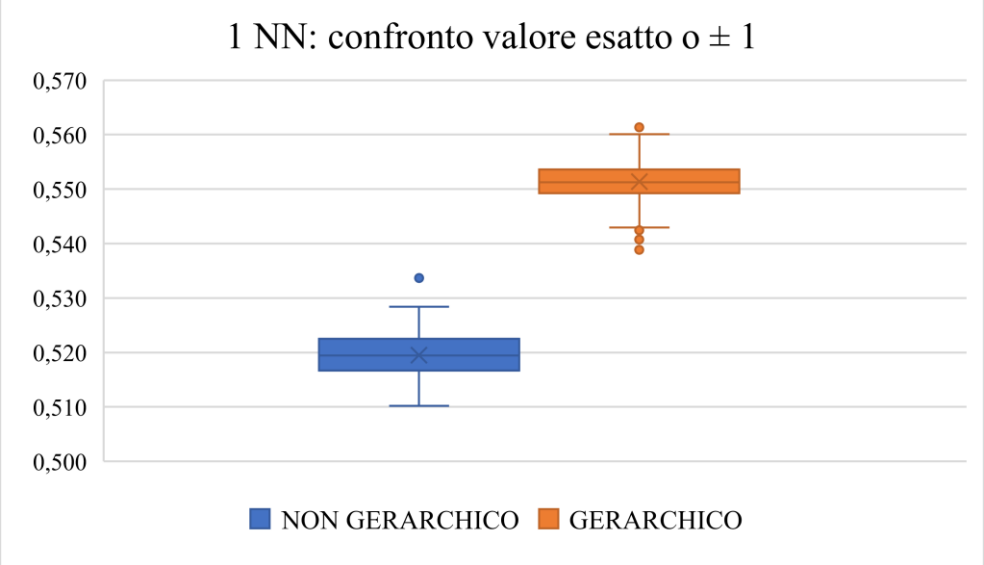

Figura 6.22: Confronto esatto o con errore di un punto in positivo e in negativo tra i due metodi

Per quanto riguarda le simulazioni relative alle imputazioni ottenute con il metodo PMM con cinque *nearest neighbours*, dalle Figure 6.23 e 6.24 si può osservare come venga preservata la miglior qualità di imputazione del metodo basato su un modello gerarchico e come, anche in questo caso, non vi sia un metodo che sovrastima o sottostima di un punto in modo evidente.

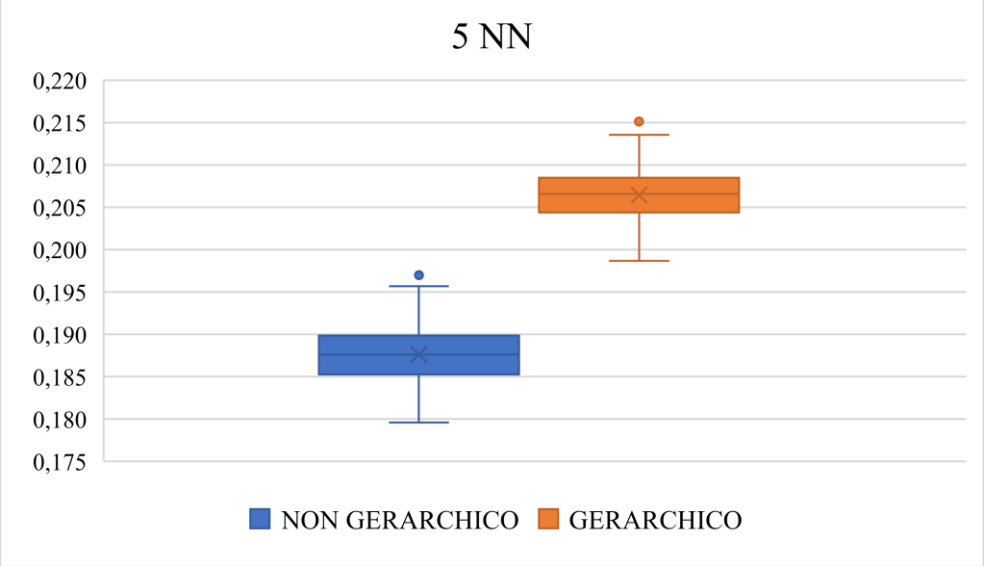

Figura 6.23: Confronto esatto tra valori imputati e reali dell'item D14 nei due metodi (5 NN)

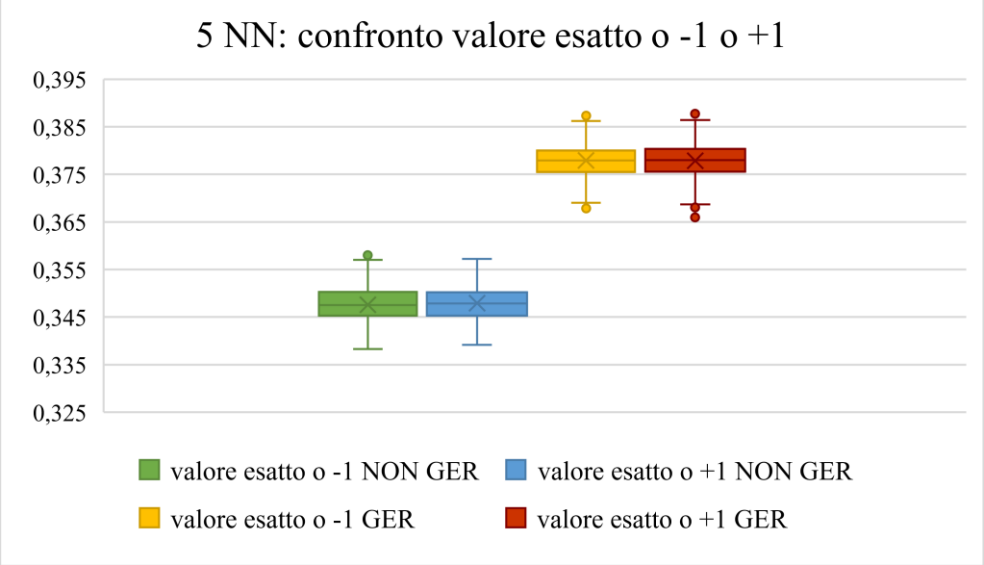

Figura 6.24: Confronto esatto o con errore di un punto in positivo o negativo tra i due metodi

Una volta appurato il fatto che le imputazioni migliorano, ossia i valori imputati si avvicinano di più a quelli reali, tenendo conto della struttura gerarchica dei dati (sfruttando quindi un modello multilivello nell'applicazione del metodo *Predictive Mean Matching)*, è di interesse fare un confronto tra l'uso di uno e cinque *nearest neighbours*, per capire se le ipotesi fatte al di fuori dello studio di simulazione trovino riscontro.
Nelle Figure 6.25 e 6.26 si riportano i box plot per le concordanze esatte o con un margine di errore di un punto, da cui emerge una differenza leggermente più marcata tra metodo gerarchico e non gerarchico con cinque *nearest neighbours* rispetto a quella osservata nel caso di un *nearest neighbour* (Figura 6.22), ossia del 4% circa, confermando quanto evidenziato in fase di imputazione: il fatto di considerare cinque donatori e sceglierne poi uno casualmente, anziché direttamente uno, con i conseguenti problemi legati alle distanze di Mahalanobis nulle, consente di marcare la differenza tra i due metodi, favorendo maggiormente quello gerarchico.
Tuttavia, l'utilizzo di un solo *nearest neighbour* permette di avere una percentuale di concordanza tra valori imputati e valori reali leggermente più alta rispetto a quella ottenuta con cinque NN, sia nel caso gerarchico che non.
I cinque NN quindi non apportano un miglioramento dal punto di vista della qualità delle imputazioni, ma sono utili nell'evidenziare la discrepanza tra metodo gerarchico e non gerarchico.

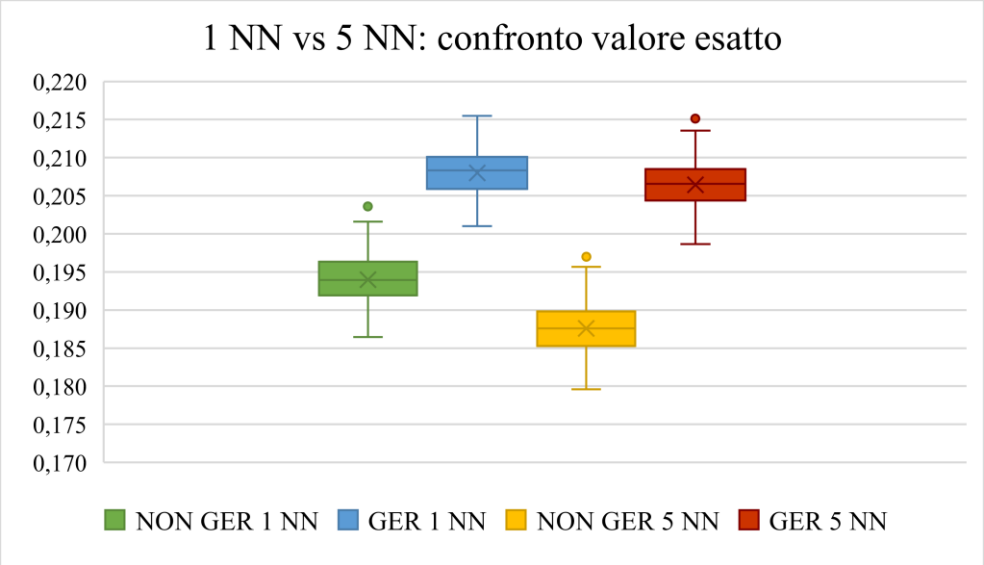

Figura 6.25: Confronto con valore esatto tra metodi con 1 NN e 5 NN

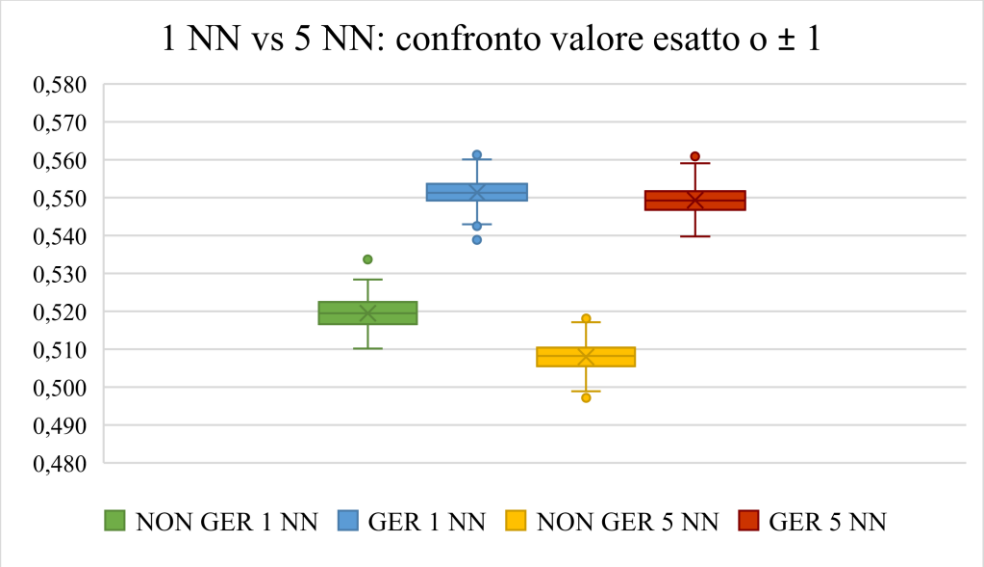

Figura 6.26: Confronto con valore esatto o con errore di un punto tra metodi con 1 NN e 5 NN

6.4 Modello ad intercetta casuale per dati imputati

Lo studio di simulazione ha permesso di capire come il metodo *Predictive Mean Matching*, applicato tramite un modello gerarchico e con l'utilizzo di un *nearest neighbour*, sia il migliore, ossia quello che produce valori imputati più vicini a quelli reali dell'item D14. L'utilizzo di cinque NN, come sottolineato in precedenza, oltre ad essere computazionalmente più oneroso da attuare, non comporta miglioramenti sulla qualità delle imputazioni.

Per concludere la fase di analisi si decide dunque di stimare un modello multilivello ad intercetta casuale sulla variabile imputata con il metodo scelto; il nuovo modello si baserà quindi su un numero più elevato di osservazioni e si vuole vedere quanto le stime si discostano da quelle del modello basato sulla tecnica della *listwise deletion* (LD) (Tabella 5.2).

Le variabili esplicative sono le medesime e in Tabella 6.8 si riporta l'output; per facilitare il confronto, nell'ultima colonna sono riportati i risultati principali ottenuti con LD.

*Tabella 6.8: Stime del modello ad intercetta casuale su item D14 imputato con PMM gerarchico e 1 NN. Livello di significatività: *** p-value<0.01, ** p-value<0.05, * p-value<0.1*

STIME EFFETTI FISSI					
Caratteristiche	**Variabili**	**Stima**	**Std. Error**		**Stima LD**
STUDENTE-GENERALE	frequenza 30-50%	-1.063	0.025	***	-1.062
	frequenza 50-70%	-0.614	0.017	***	-0.617
	genere: femmina	-0.040	0.012	***	-0.045
	età	0.034	0.002	***	0.033
STUDENTE-CARRIERA	in corso	-0.254	0.027	***	-0.214
	media dei voti	0.014	0.002	***	0.018
	numero medio esami	-0.029	0.009	***	-0.061
	CFU	0.003	0.001	***	0.004
	numero esami * numero crediti	0.0002	0.000	***	0.0004
STUDENTE-GENERALE (media)	frequenza 30-50%	-2.923	0.526	***	-3.329
	frequenza 50-70%	-1.731	0.359	***	-1.832
	genere: femmina	-0.386	0.198	**	-0.414
	età	0.004	0.243		0.022
STUDENTE-CARRIERA (media)	in corso	-0.272	0.258		-0.248
	media dei voti	0.032	0.029		0.025
	numero medio esami	-0.156	0.060	***	-0.114
	CFU	0.014	0.008	*	0.011
CORSO	corso obbligatorio	-0.029	0.034		-0.028
	numero di ore	-0.013	0.138		0.058
	più di un docente	-0.057	0.047		-0.077
	sede Padova	1.719	1.101		2.213
	insegnamento mutuato	-0.149	0.052	***	-0.138
DOCENTE-GENERALE	genere: femmina	-0.055	0.053		-0.055
	età	-0.016	0.003	***	-0.016
DOCENTE-CARRIERA	professore associato	0.000	0.064		-0.004
	professore ordinario	0.076	0.080		0.061
	professore a contratto	-0.001	0.078		0.027
	titolare assegno di ricerca	-0.418	0.236	*	-0.409
	costante	8.083	0.899	***	7.691
EFFETTI CASUALI					
	Varianza between	0.874			0.832
	Varianza within	2.671			2.716
	ICC	0.247			0.235
DIMENSIONI					
	Numero osservazioni	108280			77177
	Numero gruppi	1931			1925

Media oss. per gruppo	56	40
TEST		
Wald χ^2_{110}	4096.46	3397.27
Prob > χ^2	0.000	0.000
LR χ^2_{bar} (01)	23448.3	14719
Prob ≥ χ^2_{bar}	0.000	0.000

Innanzitutto si specifica che non sono stati riportati nuovamente i livelli di significatività per il modello con LD in quanto essi coincidono con quelli del nuovo modello, ad eccezione di quelli riferiti alla media per gruppo del numero medio di esami e dei CFU acquisiti mediamente in un anno; quest'ultimi nel modello con LD non risultavano significativi.

Dalla Tabella 6.8 emerge come le stime delle variabili statisticamente significative del nuovo modello si avvicinino molto a quelle ottenute con la tecnica della *listwise deletion* e in alcuni casi coincidano.

Senza entrare nel merito dell'interpretazione delle stime, ciò che deve essere evidenziato è proprio la differenza tra la significatività delle medie di gruppo: con LD queste vengono calcolate su un numero di osservazioni inferiore rispetto a quello su cui si basano nel modello sottostante l'imputazione; è quindi ragionevole pensare che il metodo PMM gerarchico (1 NN) sia utile in quanto, coinvolgendo più osservazioni per ogni gruppo, fa emergere maggiormente gli effetti contestuali (più accurati) e quindi l'eterogeneità tra i gruppi.

Infatti, nonostante la differenza tra il numero di gruppi sia minima e vi siano due variabili significative in più rispetto a LD (quindi la variabilità tra i gruppi dovrebbe essere maggiormente spiegata), l'ICC risulta più alto: l'imputazione quindi mette maggiormente in evidenza la variabilità tra i gruppi.

Invece, la quasi assenza di differenza tra le stime delle variabili individuali porta a pensare che la distanza tra l'ipotesi MAR e l'ipotesi MCAR non sia molto marcata.

Conclusioni

Questo elaborato si è focalizzato nell'analisi dei dati mancanti presenti in una delle domande contenute nel questionario per la valutazione della didattica, compilato nell'anno accademico 2012/2013 dagli studenti dell'Ateneo di Padova, estendendo in ambito multilivello un metodo di imputazione molto utile per questo tipo di variabili; la domanda di interesse è relativa a quanto uno studente si ritenesse soddisfatto del corso frequentato (item D14), avente il 29% di risposte mancanti.

Inizialmente ci si è concentrati sulla definizione di dato mancante e sulla sua natura; in particolare, a seconda della tipologia di *missing data*, possono essere sfruttati diversi metodi per il loro trattamento.

Se essi vengono generati da un meccanismo MCAR, ossia non dipendono né dai valori osservati né da quelli mancanti, l'utilizzo della tecnica *listwise deletion* è giustificato.

Nel caso in cui la distribuzione dei dati mancanti dipenda dai valori osservati invece i dati si definiscono generati da un meccanismo MAR; in questo caso la *listwise deletion* potrebbe produrre risultati distorti, rendendo necessario l'utilizzo di tecniche di imputazione.

Nel caso in esame è stata ritenuta poco plausibile l'ipotesi MCAR, data la struttura del questionario, e quindi del dataset, e sono stati tenuti in considerazione i meccanismi MAR e MNAR; secondo quest'ultimo i *missing data* dipendono sia dai fattori osservati che da fattori non osservati e vi sarebbe una sorta di distorsione dovuta alla selezione non casuale dei valori mancanti, che porterebbe le stime dei parametri basate sulle tecniche di imputazione a non essere affidabili.

Come punto di partenza, la tecnica della *listwise deletion* è stata comunque sfruttata per stimare due modelli multilivello ad intercetta casuale, per descrivere il punteggio assegnato all'item D14 in funzione di una serie di variabili esplicative (con e senza variabili riferite ai corsi e ai docenti).

Data la possibile distorsione delle stime prodotte, prima di procedere con l'imputazione, è stata verificata la possibile distorsione da selezione dovuta ad un meccanismo MNAR, tramite un modello di Heckman; dalle analisi descrittive, infatti, era emersa una differenza tra rispondenti e non rispondenti basata su alcune caratteristiche osservate. Tuttavia, tale

differenza potrebbe essere dovuta anche a caratteristiche non osservabili, come motivazioni personali degli studenti.

Dal modello è emerso un valore molto basso del coefficiente di correlazione tra i termini di errore delle due equazioni utilizzate, portando all'esclusione dell'ipotesi MNAR e procedendo dunque con l'imputazione sotto ipotesi MAR.

Sono state applicate tre tecniche di imputazione sfruttando, oltre che un semplice modello di regressione lineare, anche il modello multilivello; le tecniche utilizzate sono l'imputazione con la media, con la regressione e con il metodo *Predictive Mean Matching* (PMM), noto in letteratura per associare l'utilizzo della regressione lineare con la tecnica del *nearest neighbour* (nel caso in esame estratto tramite il calcolo della distanza di Mahalanobis). Ciò che è stato introdotto in questo lavoro è proprio una sua applicazione tramite un modello multilivello.

Dai risultati è emerso come l'imputazione con media e con regressione, sia nel caso gerarchico sia in quello non gerarchico, abbiano condotto ad una distorsione nella distribuzione della variabile, rispetto all'originale, ad una sottostima della variabilità e, nel caso della regressione, anche all'imputazione di valori non ammissibili.

Il metodo computazionalmente più complesso, ossia il PMM, è invece apparso più efficace, in quanto la distribuzione dei valori osservati è stata preservata nella parte mancante dei dati. Inoltre, è emersa un'equivalenza, dal punto di vista marginale, tra i punteggi imputati dal metodo basato sulla regressione e quello basato sul modello gerarchico, mentre la concordanza tra i singoli punteggi nei due casi è risultata del 50%.

L'ampliamento del metodo tramite il calcolo di cinque *nearest neighbours*, per ovviare al problema legato alle distanze di Mahalanobis nulle nella ricerca dell'unità donatrice, ha permesso di incrementare il distacco tra metodo gerarchico e non gerarchico, con un calo della concordanza tra i punteggi (17%), pur mantenendo inalterata la distribuzione della variabile.

In alternativa è stata anche introdotta una particolare versione di imputazione multipla "gerarchica", sfruttando la stima della distribuzione dei residui di gruppo, con cui sono state ottenute cinque diverse variabili imputate con il metodo PMM. Quest'ultime hanno riportato una concordanza nei donatori del 38%, con un 40% di concordanza tra i punteggi; marginalmente invece è emersa una concordanza quasi totale. Inoltre, la distribuzione della variabile è stata preservata e, nel confronto con i risultati del metodo

PMM non gerarchico con un *nearest neighbour*, è risultata una concordanza tra i punteggi imputati del 35%, quindi inferiore rispetto al 50% ottenuto con l'imputazione singola.
La procedura è risultata quindi interessante e valida dal punto di vista delle imputazioni prodotte, ma computazionalmente molto impegnativa, con tempi di calcolo troppo lunghi per essere preferita a metodi meno onerosi e comunque validi.
La fase di imputazione si è conclusa con uno studio di simulazione, concentrato dunque sul metodo PMM ad imputazione singola, gerarchico e non gerarchico, con uno e cinque *nearest neighbours*, e si è voluto analizzare quanto i valori imputati dai vari metodi si avvicinassero ai valori reali della variabile in oggetto. È stata presa in considerazione anche la possibilità di un errore unitario, in positivo e/o in negativo.
Nel confronto tra i valori imputati ed il punteggio esatto dell'item D14, sia con uno che con cinque *nearest neighbours*, è emerso un netto distacco tra il metodo gerarchico e non gerarchico; il primo è quello che ha condotto ad una performance migliore, confermando che l'informazione aggiuntiva di cui tiene conto porta ad un'imputazione più efficace.
Tenendo conto anche del possibile errore unitario il distacco si fa ancora più marcato.
Ciò che si è notato è il fatto che utilizzare cinque *nearest neighbours* non comporti un miglioramento nella qualità delle imputazioni, ma, come già sottolineato al di fuori dello studio di simulazione, contribuisca soltanto ad incrementare la differenza tra metodo gerarchico e non gerarchico, favorendo il primo.
Emerge quindi come il metodo *Predictive Mean Matching*, nella versione con modello gerarchico e 1 NN, sia quello che porta a risultati più soddisfacenti.
Lo studio di simulazione introdotto ha comunque dei limiti.
Il primo tra questi è legato al fatto che la procedura implementata ha simulato un meccanismo MCAR come generatore dei dati mancanti, in quanto essi sono stati creati in maniera completamente casuale, e questo sembrerebbe essere in contraddizione con l'ipotesi ritenuta più ragionevole, ossia MAR.
Tale studio però deve essere inteso come punto di partenza per ulteriori sviluppi e miglioramenti futuri. Uno di questi, ad esempio, può essere quello di simulare un meccanismo MAR, vincolando la "creazione" dei valori mancanti a fattori osservabili presenti nel dataset: anziché utilizzare una probabilità di estrazione degli NA uguale per tutti gli individui, si potrebbe sfruttare una probabilità proporzionale ad una loro caratteristica osservata, magari una di quelle risultate discriminanti tra rispondenti e non

all'item D14. In questo modo si creerebbe una situazione di maggior similitudine con il dataset originale, con risultati più coerenti.

Un altro limite può essere legato al numero di variabili esplicative inserite nel modello sottostante l'imputazione gerarchica e un logico miglioramento è quello di aggiungere ulteriori variabili, per esempio relative a caratteristiche soggettive dei docenti (dal questionario PRODID - Preparazione alla Professionalità Docente e Innovazione Didattica) oppure ad altre caratteristiche degli studenti, le quali non avrebbero un significato ragionevole dal punto di vista interpretativo, ma porterebbero informazione aggiuntiva contribuendo ad incrementare il distacco con il modello di regressione.

Infine, per quanto riguarda il numero di simulazioni, sarebbe stato ottimale farne almeno 1000 ma, dato che il programma realizzato è risultato essere computazionalmente complesso, con tempi di esecuzione abbastanza lunghi, si è deciso di farne 250; tuttavia, a posteriori, vista la bontà dei risultati, si può dire che l'aumento del numero di simulazioni presumibilmente non avrebbe comportato notevoli miglioramenti.

Tutte queste possibilità vengono lasciate come spunti per lavori futuri.

Per concludere, la variabile imputata con il metodo prescelto, nel dataset originale, è stata sfruttata per la stima di un modello ad intercetta casuale e le stime ottenute sono state confrontate con quelle del modello realizzato inizialmente tramite *listwise deletion*.

È emerso come l'imputazione sia utile nel far emergere maggiormente la variabilità tra i gruppi e gli effetti contestuali: l'unica differenza nella significatività delle stime infatti è stata riscontrata nelle variabili relative a medie di gruppo.

L'assenza di differenza tra le stime delle variabili individuali potrebbe indicare che il meccanismo MAR non è poi così distante da quello MCAR.

Bibliografia

[1] Aitkin M. e Longford N.T. (1986) *Statistical modelling issues in school effectiveness studies (with discussion)*. Journal of the Royal Statistical Society A 149, 1-43.

[2] Baccini M. (2008) *Imputazione multipla di dati mancanti*. Strumenti e Metodi 162-163.

[3] Barcaroli G., D'Aurizio L., Luzi O., Manzari A., Pallara A. (1999) *Metodi e software per il controllo e la correzione dei dati.* ISTAT, Roma.

[4] Bassi F., Grilli L., Paccagnella O, Rampichini C e Varriale R. (2017) *New Insights on Students Evaluation of Teaching in Italy*.

[5] Berg N. (2005) *Non-Response Bias*. Encyclopedia of Social Measurement, Volume 2, Elsevier Inc.

[6] Boscaino G. e Sulis I. *L'imputazione dei dati mancanti: un confronto tra diversi approcci*. Articolo scientifico Dipartimento di Metodi Quantitativi per le Scienze Umane, Università degli Studi di Palermo.

[7] Bryk A.S. e Raudenbush S.W. (1992) *Hierarchical Linear Models: Applications and data analysis methods*. CA: Sage Publications, Inc.

[8] Cappuccio N. e Orsi R. (2005) *Econometria.* Il Mulino.

[9] Chemolli E. e Pasini M. (2007) *I dati mancanti*. DiPAV, 20, 51-56.

[10] Craig K. Enders (2010) *Applied Missing Data Analysis*. Guilford Publications.

[11] Dalla Zuanna G., Clerici R., Paccagnella O., Paggiaro A., Martinoia S., Pierobon S. (2016) *Evaluation research in education: a survey among professors of University of Padua*. Excellence and Innovation in Learning and Teaching 1, 17-34.

[12] Dempster A. e Rubin D. B. (1983) *Incomplete data in sample surveys.* Sample surveys, 2:3–10.

[13] Di Manno R. *Metodi per l'imputazione di dati mancanti in dataset di variabili nominali*. Tesi di dottorato Università di Roma tre.

[14] Felisatti E. e Serbari A. (2014) *The professional development of teachers: from teachers practices and beliefs to new strategies at the university of Padua*. Procedings of the ICED conference Educational development in a changing world, Stockholm.

[15] Ghilardi G. e Orsini N. (2002) *Modelli lineari ad intercetta casuale, stimatori e valutazione di sistemi formativi.* STATISTICA, anno LXII, n. 4.

[16] Goldstein H. (2011) *Multilevel Statistical Models*. 4th Edition, Wiley Publication.

[17] Goos M. e Salomons A. (2016) *Measuring teaching quality in higher education: assessing selection bias in course evaluations*. Springer.

[18] Heckman J. J. (1979) *Sample Selection Bias as a Specification Error.* Econometrica 47(1): 153- 161.

[19] Hox J. J. (2010) *Multilevel Analysis: Techniques and Applications*. Quantitative methodology series, Second Edition.

[20] Kreft I. e De Leeuw J. (1998) *Introducing Multilvel Modeling.* Sage Publication.

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

[22] Little R. J. A. e Rubin D. B. (2002) *Statistical Analysis with Missing Data*. Wiley Interscience.

[23] Longford N.T. (1993) *Random Coefficient Models.* Oxford, Clarendon Press.

[24] Marsh H. W. (2007) *Students' evaluations of university teaching: A multidimensional perspective, Chapter 9*. The scholarship of teaching and learning in higher education: an evidence-based perspective (pp.319–384). Springer.

[25] Morris T. P., I. R. White e P. Royston (2014) *Tuning multiple imputation by predictive mean matching and local residual draws*. BMC Medical Research Methodology 14: 75.

[26] O'Connell A. e McCoach D. (2008) *Multilevel Modeling of Educational Data*. Information Age Publishing.

[27] Paccagnella O. (2006) *Centering or not Centering in Multilevel Models? The Role of the Group Mean and the Assessment of Group Effects*. Evaluation Review 30, 66-85.

[28] Rampichini, C., Grilli, L. e Petrucci, A. (2004) *Analysis of university course evaluations: from descriptive measures to multilevel models*. Statistical Methods & Applications, **13**, pp 357–373.

[29] Recai M. Yucel (2008) *Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response*. Phil. Trans. R. Soc. 366, 2389–2403.

[30] Ruscone N. M. (2011) *Modelli gerarchici: aspetti metodologici e ambiti di applicazione.* Tesi di dottorato.

[31] Ruscone N. M. (2012) *Utilizzo dell'ICC come indicatore dell'esistenza di una struttura gerarchica*.

[32] Snijders T.A.B. e Bosker R.J. (2012) *Multilevel analysis: an introduction to basic and advanced multilevel modelling.* Sage, Londra.

[33] Stephen W. e Anthony S. (2002) *Hierarchical Linear Models, Applications and Data Analysis Methods,* Second Edition.

[34] Stef van Buuren (2011) *Multiple Imputation of Multilevel Data*. TNO Quality of Life, Department of Statistics, Leiden and University of Utrecht, The Netherlands.

[35] Stef van Buuren (2012) *Flexible imputation of missing data*. New York, NY: Chapman & Hall.

[36] Verbeek M. (2006) *Econometria*. Zanichelli.

[37] Wild C. *The Wilcoxon Rank-Sum Test*. University of Auckland.