Università degli Studi di Firenze
Dipartimento di Statistica "G. Parenti"
Dottorato di Ricerca in Statistica Applicata
XX *ciclo* SECS-S/01

Imputazione multipla: metodologie e proposte per l'analisi di dati di reddito

Caterina Giusti

Tutor: **Prof. Bruno Chiandotto**

Co-tutor: **Dott. Orietta Luzi**

Coordinatore: **Prof. Guido Ferrari**

Desidero ringraziare il mio tutor, Prof. Bruno Chiandotto, per la sua preziosa guida durante i tre anni di dottorato. Un ringraziamento speciale va al Prof. Luigi Biggeri, grazie al quale ho potuto collaborare con ISTAT e conoscere il Prof. Rod Little, con cui ho avuto il piacere di lavorare. Presso l'ISTAT ringrazio, in particolare, la Dott.ssa Orietta Luzi, che ha saputo consigliarmi con grande disponibilità e puntualità, ed il Dott. Marco Di Marco, che ha messo a mia disposizione i dati dell'indagine Condizioni di Vita. Presso il Dipartimento di Statistica dell'Università di Firenze ringrazio tutti coloro che mi hanno aiutata e supportata; grazie, in particolare, al Prof. Guido Ferrari, coordinatore del dottorato, e al Prof. Andrea Giommi per le sue osservazioni e per aver riletto questo lavoro.

Durante questi tre anni ho avuto la fortuna di condividere molte esperienze di vita e di lavoro con i miei compagni di dottorato, Bruno, Eleonora, Federica, Giulia, Graziano, Roberta e Vincenzo. Un ringraziamento speciale va a Roberta, amica e collega davvero insostituibile.

Questo lavoro di tesi affronta il problema che nasce quando qualcosa di cui si ha bisogno è mancante. Il mio grazie più grande va a coloro che più di ogni altro non mi hanno fatto mancare affetto, rispetto, fiducia e comprensione: ai miei genitori, a Vittorio.

Caterina Giusti
31 Dicembre 2007

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

Le mancate risposte rappresentano una possibile fonte di errore in tutte le indagini campionarie. Se i dati rilevati attraverso un'indagine non sono completi, qualunque metodo si scelga per trattare i *missing values*, anche semplicemente ignorare la loro presenza, può influenzare il risultato delle analisi statistiche.

Negli ultimi due decenni al problema delle mancate risposte è stata dedicata un'attenzione crescente a ragione della diminuzione del livello di collaborazione degli intervistati in molti paesi sviluppati (de Leeuw and de Heer, 2002). Di conseguenza, intere monografie sono state dedicate al problema dei dati mancanti e varie, innovative metodologie per l'analisi di dati incompleti sono state proposte e sviluppate, anche grazie ai contemporanei progressi delle tecniche computazionali.

Da un punto di vista teorico, inoltre, il problema dei dati mancanti è più generale di quello delle mancate risposte nelle indagini campionarie: molti problemi statistici possono essere infatti formulati in termini di dati mancanti anche quando non vi è alcun insieme di dati incompleto (Little and Rubin, 2002). E' questo il caso, per esempio, dell'approccio controfattuale all'inferenza causale e dei modelli per variabili latenti (Gelman and Meng, 2004).

Oltre a sollevare importanti questioni teoriche, la presenza di dati mancanti comporta anche numerosi risvolti di carattere applicativo, e le due problematiche possono non coincidere: le soluzioni teoriche possono talvolta rivelarsi troppo complicate per essere applicate e, viceversa, è possibile che metodologie non ancora supportate da sufficienti giustificazioni teoriche si rivelino ottimali dal punto di vista pratico.

Il lavoro di tesi si colloca in questo interessante e dinamico quadro affrontando, sia da un punto di vista teorico che applicativo, il problema delle mancate risposte parziali nelle indagini campionarie, con particolare riferimento ai quesiti relativi a variabili di reddito. L'approccio scelto per trattare il problema è essenzialmente un approccio *da modello* (Little and Rubin, 1983). In particolare, i metodi implementati in questa tesi fanno riferimento

all'imputazione multipla, metodologia inizialmente proposta da Rubin (1978) nel contesto delle indagini campionarie, ma che sta attualmente riscuotendo un notevole successo anche per problemi statistici di tipo diverso.

Lo studio della distribuzione del reddito nella popolazione riveste un ruolo di importanza fondamentale per la comprensione di numerose dinamiche economiche e sociali; la diffusione del benessere e della povertà, per esempio, può influenzare la programmazione di interventi di politica sociale e consentire il confronto delle condizioni di vita in paesi diversi.

La necessità di rilevare il reddito attraverso indagini campionarie nasce per l'inadeguatezza dei dati provenienti da fonti fiscali o amministrative; le informazioni contenute nelle dichiarazioni dei redditi, per esempio, oltre a non essere rappresentative dell'intera popolazione, possono essere rese meno attendibili dal fenomeno dell'evasione. Inoltre, un'accurata conoscenza dei redditi e consumi delle famiglie piuttosto che dei singoli individui può difficilmente essere derivata da queste tipologie di dati.

Anche i redditi rilevati attraverso indagini campionarie non sono esenti da errori. La caratteristica più peculiare sono gli elevati tassi di mancata risposta: la percentuale di individui che si rifiuta di rispondere a quesiti relativi al reddito, per motivi di riservatezza, di ignoranza o altro, è spesso compresa tra il 20-40% (Heeringa et al., 2002). Si rende allora necessaria l'imputazione dei valori mancanti, ovvero la loro sostituzione con valori opportunamente scelti.

L'imputazione è il metodo solitamente utilizzato per trattare le mancate risposte a singoli quesiti; per le mancate risposte totali, invece, si fa usualmente ricorso a tecniche di ponderazione (Sarndal et al., 1992). L'analisi dei soli casi completi, senza l'imputazione dei valori mancanti, può rappresentare un buon metodo di analisi solo nelle situazioni in cui le mancate risposte sono poche e, inoltre, i dati mancanti possono essere verosimilmente ritenuti un sottoinsieme casuale di tutte le osservazioni. In questo caso, che riguarda pochissimi problemi reali, Rubin (1987) parla di dati *missing completely at random* (MCAR).

Nella maggior parte delle situazioni reali, invece, il motivo per cui i dati sono mancanti dipende da altre variabili osservate; per esempio, alcuni studi condotti utilizzando i dati dell'indagine Banca d'Italia sui Bilanci delle Famiglie hanno evidenziato come la probabilità di mancata risposta a quesiti di reddito possa essere più elevata per le famiglie residenti nelle aree urbane e nel nord Italia (D'Amuri and Fiorio, 2004). In questo caso i dati mancanti dipendono da altre variabili, che sono a disposizione del ricercatore: si dice allora che i dati sono *missing at random* (MAR).

La situazione più complicata, da cui secondo alcuni studiosi non si può prescindere quando ci si occupa di dati mancanti di reddito, è un'altra: il

motivo per cui un valore di reddito è mancante non è legato solo a caratteristiche note, ma anche a variabili non osservate, tra cui il reddito stesso. In queste situazioni si dice che i dati sono *missing not at random* (MNAR). Se per esempio ad un lavoratore autonomo viene chiesto il suo reddito medio mensile, è possibile che la probabilità che si rifiuti di rispondere sia in relazione positiva con il reddito stesso: più alto è il reddito e più la mancata risposta è probabile. Ma anche la situazione opposta è altrettanto verosimile: più basso è il reddito e più la mancata risposta è probabile; alcuni autori, infatti, parlano di una relazione a "forma di U", in cui le mancate risposte si collocano nelle code della distribuzione del reddito nella popolazione di riferimento (Lillard et al., 1986).

L'imputazione di valori MNAR presenta ancora notevoli problematiche. L'impiego di modelli MNAR, infatti, è sempre soggetto a forti elementi di soggettività e a numerosi problemi in fase di stima, che possono raggiungere un livello di complicazione ingiustificato (Little and Rubin, 2002). Chi può dirci se i redditi non osservati sono funzione dei loro stessi valori, dal momento che tali dati sono mancanti? E anche ammesso che sia noto che tale legame esiste, com'è possibile venirne a conoscenza con esattezza e riuscire a stimarne empiricamente la relazione? Una possibile risposta a queste domande può giungere da opportuni procedimenti "ad hoc". Per esempio, parte dei non rispondenti possono essere ricontattati e convinti a comunicare il loro reddito; oppure, si può effettuare il *matching* tra i dati provenienti dall'indagine e altri databases, per esempio quelli fiscali. Un'altra situazione di notevole interesse sono i campioni di tipo panel, in cui è possibile che un intervistato si rifiuti di comunicare il suo reddito in una data occasione ma non in una successiva; le mancate risposte sono legate d'altra parte anche a numerosi fattori di ordine psicologico e comportamentale, che possono non essere costanti nel tempo (Groves and Couper, 1998).

Le conclusioni cui si giunge anche in questo caso, tuttavia, possono non essere univoche. Ne è testimonianza, proprio in riferimento a dati mancanti di reddito, il dibattito che si è acceso tra gli studiosi statunitensi in merito alla Current Population Survey, indagine simile a quella sulle Forze Lavoro condotta in Italia. La relazione positiva tra reddito e mancate risposte, inizialmente individuata da alcuni autori, è stata successivamente smentita da altri, che hanno quindi suggerito l'imputazione del reddito secondo metodi MAR, facendo poi un'analisi della sensitività delle stime di interesse rispetto ad ipotesi MNAR.

I più moderni e sofisticati metodi di imputazione per dati multivariati si basano proprio sull'ipotesi di dati MAR: se tale ipotesi è corretta e se il procedimento di imputazione utilizza tutte le informazioni osservate, da cui i dati mancanti possono dipendere, allora non è necessario introdurre nel

procedimento una esplicita specificazione del meccanismo che ha generato le mancate risposte. Se i dati sono MNAR, invece, la differenza tra rispondenti e non rispondenti andrebbe esplicitamente modellata ed introdotta nel procedimento di imputazione.

In Italia non si è assistito, almeno per il momento, ad un dibattito altrettanto acceso come negli Stati Uniti sul meccanismo che genera le mancate risposte a quesiti di reddito. Questo può essere dovuto al fatto che in Italia l'imputazione delle mancate risposte parziali nelle indagini condotte su scala nazionale non ha una solida tradizione come negli Stati Uniti. Esistono però alcuni studi, condotti sui dati dell'indagine Bilanci delle Famiglie della Banca d'Italia, che sostengono l'esistenza di una relazione positiva tra le mancate risposte ed il reddito; tuttavia tali risultati sono stati derivati utilizzando un numero limitato di re-interviste, e c'è da chiedersi se non sarebbe possibile, disponendo degli stessi dati, giungere a conclusioni diverse, così come è avvenuto negli Stati Uniti. Sarebbe sicuramente interessante assistere in futuro a nuove ricerche su questo tema, per esempio utilizzando le numerose informazioni sui redditi individuali e familiari che provengono dall'indagine ISTAT sulle Condizioni di Vita, che appartiene al progetto europeo EU-SILC, la cui prima rilevazione è stata effettuata in Italia nel 2004. Utilizzando la componente panel del campione o effettuando il matching con i dati fiscali, procedura già utilizzata da ISTAT per i redditi dei lavoratori autonomi, potrebbe essere possibile verificare l'effettiva inadeguatezza dell'ipotesi MAR utilizzata nel procedimento di imputazione, correggendo a posteriori le stime di interesse.

Le difficoltà legate all'implementazione di una buona procedura di imputazione non sono legate tuttavia solamente alla comprensione del meccanismo che ha generato le mancate risposte. Anche la scelta della particolare tecnica di imputazione può essere basata su criteri diversi.

Qualsiasi tecnica di imputazione dipende da un modello, esplicitamente o meno. Per esempio, le tecniche di imputazione *da donatore* individuano delle celle di imputazione, definite in base ad informazioni osservate come genere, classe di età, ecc., e i valori mancanti per una unità vengono sostituiti con valori osservati per un'altra unità che appartiene alla stessa cella. In questo caso il modello implicito sottostante è che le mancate risposte siano casuali (MCAR) all'interno delle celle di imputazione. L'imputazione per regressione fa invece riferimento ad un modello esplicito: i dati mancanti vengono completati attraverso i valori previsti da un modello di regressione stimato utilizzando le osservazioni complete.

Qualunque siano il meccanismo che ha generato le mancate risposte ed il modello che caratterizza la tecnica di imputazione, se questa è stata opportunamente scelta è possibile ottenere un insieme di dati completo che, se ben

analizzato, porterà a risultati corretti. Analizzare bene un dataset imputato significa non dimenticare che alcuni dei valori erano inizialmente mancanti.

Per esempio, se le imputazioni vengono realizzate secondo l'ipotesi MAR e i dati così completati vengono analizzati attraverso i tradizionali stimatori per dati completi, non facendo alcuna distinzione tra dati osservati ed imputati, le componenti di varianza risulteranno in generale sottostimate; intuitivamente, ciò dipende dal fatto che l'imputazione di tipo MAR non genera informazioni che non siano già presenti nei dati osservati: di conseguenza la numerosità campionaria, anche se apparentemente maggiore, resta pari a quella che caratterizza i dati incompleti (Nielsen, 2003).

Essenzialmente, l'idea alla base dell'imputazione multipla è imputare non uno ma più valori per ciascun dato mancante; in questo modo si ottengono più datasets completati e lo stimatore di interesse viene applicato separatamente a ciascun datasets. Successivamente, utilizzando apposite regole di combinazione, è possibile ottenere un'unica inferenza che tiene in considerazione anche la variabilità *between* i datasets, che riflette l'incertezza legata al passo di imputazione dei dati.

In Italia l'imputazione multipla non è, almeno per il momento, una procedura standard. Negli Stati Uniti, invece, l'imputazione multipla secondo modelli di regressione multivariati basati sull'ipotesi MAR viene utilizzata per imputare i dati di reddito in alcune grandi indagini, la Consumer Expenditures Survey, condotta dall'U.S. Department of Labor, e la National Health Interview Survey, condotta dal National Center for Health Statistics. Probabilmente si dovrà attendere ancora un po' di tempo per capire se l'imputazione multipla potrà essere considerata un'utile soluzione anche per le indagini condotte nel nostro paese. Ultimamente nelle direttive EUROSTAT per l'imputazione dei dati mancanti nell'indagine EU-SILC qualcosa si sta muovendo in tale direzione: il metodo di imputazione attualmente in uso in ISTAT sono le regressioni sequenziali multivariate, lo stesso metodo impiegato negli Stati Uniti, anche se per realizzare imputazioni singole e non multiple.

Gli aspetti teorici e applicativi che si affrontano nella tesi sono molteplici. I dati utilizzati nelle applicazioni provengono da due indagini diverse, l'indagine ISTAT EU-SILC sulle Condizioni di Vita 2004 e l'indagine sulle Forze Lavoro 2002 del Comune di Firenze.

I dati dell'indagine EU-SILC si riferiscono a 24204 famiglie e 52509 individui; i questionari utilizzati sono due, uno familiare ed uno individuale, somministrato a tutti i componenti sopra i quindici anni delle famiglie entrate a far parte del campione. Tra le variabili rilevate numerose sono quelle riferite a componenti di reddito, sia a livello familiare che individuale. Queste componenti possono essere combinate in modo da ottenere interessanti stime,

come quella del reddito totale disponibile per le famiglie italiane durante il 2003.

I dati dell'indagine sulle Forze Lavoro del Comune di Firenze si riferiscono a quattro occasioni di intervista, Aprile, Luglio, Ottobre 2002 e Gennaio 2003, per un totale di 3209 intervistati. Il campione ha una struttura "panel-ruotato": gli individui estratti all'anagrafe comunale vengono intervistati per due occasioni consecutive, poi escono per due occasioni e successivamente vengono intervistati in altre due occasioni. Il questionario prevede un insieme di quesiti differenziati a seconda dello stato occupazionale, determinato in base ad una domanda-filtro iniziale. Per coloro che si dichiarano occupati, viene richiesto il reddito netto mensile medio da lavoro.

La tesi è organizzata in cinque capitoli, seguiti da alcune considerazioni conclusive.

Nel primo capitolo il problema delle mancate risposte nelle indagini campionarie viene collocato nel contesto dell'approccio all'inferenza da popolazioni finite basato su modello, e si introducono le definizioni formali che determinano l'ignorabilità del meccanismo di campionamento e di mancata risposta. Inoltre, si presentano i principali metodi per trattare le mancate risposte sotto l'ipotesi MAR, chiarendo in quali situazioni la struttura dei dati può consentire la semplificazione del procedimento di imputazione.

Nel secondo capitolo si introduce l'imputazione multipla, presentando le sue principali proprietà e giustificazioni teoriche, anche dal punto di vista dell'inferenza randomizzata. Vengono poi descritti i metodi di imputazione basati su modelli espliciti bayesiani attualmente più utilizzati; la presentazione di questi metodi tiene in considerazione innovazioni e problemi teorici che sono ancora esclusi dalle principali monografie che trattano il problema delle mancate risposte, ma che rappresentano lo *state of the art* dell'imputazione. Alla fine del capitolo si accenna alle possibili alternative all'imputazione multipla proposte in letteratura per calcolare varianze corrette partendo da datasets imputati singolarmente.

Nel terzo capitolo si presenta una rassegna dei principali risultati cui si è giunti, in Italia e negli Stati Uniti, in merito allo specifico problema delle mancate risposte parziali di reddito. Questa rassegna, sebbene non esaustiva, descrive le tecniche di imputazione attualmente utilizzate per le due principali indagini sul reddito condotte nel nostro paese, l'indagine sui Bilanci delle Famiglie della Banca d'Italia e l'indagine ISTAT sulle Condizioni di Vita. Vengono così messe in evidenza le principali lacune in merito al problema delle mancate risposte a quesiti di reddito, indicando gli aspetti cui sarebbe interessante dedicare una particolare attenzione per futuri contributi a questa problematica nel nostro paese.

Il quarto capitolo è dedicato ad alcune proposte ed analisi relative ai da-

ti di reddito provenienti dall'indagine ISTAT sulle Condizioni di Vita 2004. Dopo una descrizione del progetto EU-SILC e dei questionari dell'indagine, vengono presentati i tassi di mancata risposta delle variabili di reddito e le principali caratteristiche sia di queste che delle variabili osservate. Particolare attenzione, inoltre, è dedicata allo studio della struttura dei dati mancanti, che deve essere messa in relazione con la composizione delle famiglie intervistate. Viene poi descritta l'innovativa procedura di imputazione multipla proposta per i dati di reddito, che tiene in considerazione i numerosi fattori di complicazione presenti nei dati. L'analisi dei risultati ottenuti è accompagnata, inoltre, dalla presentazione e applicazione di una metodologia per la verifica informale dell'ipotesi MAR su cui si basa il procedimento di imputazione utilizzato.

Nel quinto capitolo l'attenzione si concentra sui dati provenienti dall'indagine sulle Forze Lavoro del Comune di Firenze e, in particolare, sull'implementazione di un'analisi di sensitività rispetto ad ipotesi MNAR per dati strutturati in modo complesso. Il "panel-ruotato" dell'indagine, infatti, presenta caratteristiche peculiari, in quanto il meccanismo che genera le mancate risposte può essere formalizzato come unione tra dati MCAR, MAR e MNAR. L'obiettivo è sfruttare la struttura dei dati mancanti per realizzare un'analisi di sensitività di tipo multivariato che non utilizzi ipotesi non verificabili, ma di immediata comprensione e valutazione. La proposta si inserisce in un contesto, quello delle analisi di sensitività per ipotesi MNAR, in cui le applicazioni vengono solitamente relegate a situazioni in cui la non risposta è univariata e non presenta fattori di complicazione.

Capitolo 1

Le mancate risposte nelle indagini campionarie

In questo capitolo si introduce la problematica delle mancate risposte nelle indagini campionarie, collocandosi nell'ambito dell'inferenza da popolazioni finite. Nei paragrafi 1.1.1 e 1.1.2 vengono definiti formalmente i concetti di dati mancanti *a caso*, *completamente a caso* e *non a caso*, assieme alla condizione di *ignorabilità* del meccanismo di mancata risposta. Nel paragrafo 1.1.3 si introducono i modelli non ignorabili e le relative problematiche di stima. I metodi solitamente utilizzati per trattare rispettivamente le mancate risposte totali, la ponderazione, e le mancate risposte parziali, l'imputazione, sono presentati nel paragrafo 1.2.1. Queste metodologie utilizzano quasi sempre l'ipotesi MAR. Nel paragrafo 1.2.2 si chiariscono dunque quali sono le situazioni in cui tale ipotesi può risultare più verosimile. Si introduce inoltre il concetto di *pattern* dei dati: è questa un'altra caratteristica che può rendere più semplice la derivazione di inferenze in presenza di dati mancanti.

1.1 L'inferenza da popolazioni finite in presenza di valori mancanti

La letteratura che si occupa dell'inferenza da indagini campionarie è molto complessa ed articolata, e continua ancora oggi a ricevere numerosi contributi. Alcuni autori hanno proposto dei possibili "schemi" per classificare i diversi approcci all'inferenza da popolazioni finite (Sarndal et al., 1992; Brewer and Sarndal, 1983; Cassel et al., 1977). In particolare, Little and Rubin (1983) e Rubin (1983) individuano una singola distinzione da considerarsi di fondamentale importanza, quella tra approccio *randomizzato* e approccio *basato su modello*.

Consideriamo un'indagine campionaria attraverso cui sono state selezionate n unità da una popolazione di N individui e indichiamo con $y_1, \ldots, y_J$ le variabili di interesse dell'indagine, corrispondenti per esempio a J quesiti di un questionario. Il valore della variabile j per l'individuo i, y_{ij}, sarà osservato solo se l'unità i appartiene al campione ovvero se $I_i = 1$, dove I è la variabile dicotomica di appartenenza al campione, definita per $i = 1, \ldots, N$ e che assume valore 1 quando l'unità viene campionata, 0 altrimenti. Nell'approccio randomizzato i valori y_{ij} appartenenti alla matrice $\boldsymbol{Y}_{(n \times J)}$ sono considerati fissi e l'inferenza si basa sulla distribuzione di probabilità introdotta dal *meccanismo di selezione* delle unità campionarie, ovvero da $P(\boldsymbol{IY})$. Nell'ambito dell'approccio randomizzato risulta possibile, sotto certe ipotesi, compiere inferenza per quantità di interesse della popolazione, come per esempio per la media campionaria di una variabile o in generale per una quantità $Q(\boldsymbol{Y})$.

Nell'approccio basato su modello, invece, oltre al meccanismo $P(\boldsymbol{I}|\boldsymbol{Y})$ viene introdotto anche un modello per $P(\boldsymbol{Y})$: i valori y_{ij} non sono più considerati quantità fisse ma realizzazioni di variabili casuali. In questo caso è il modello $P(\boldsymbol{Y})$ a costituire la base per il processo di inferenza, mentre il meccanismo di selezione $P(\boldsymbol{I}|\boldsymbol{Y})$ assume un ruolo più marginale (Little, 1982). Se il modello scelto per $\boldsymbol{Y}$ è di tipo bayesiano, ai parametri che caratterizzano $P(\boldsymbol{Y})$ viene assegnata una distribuzione a priori.

L'approccio basato su modello è legato al concetto di *superpopolazione*. E' proprio questo uno dei concetti più dibattuti nell'ambito della letteratura sull'inferenza da popolazioni finite; una volta decisa l'introduzione di una distribuzione $P(\boldsymbol{Y})$, tale modello può avere infatti diverse interpretazioni filosofiche (Cassel et al., 1977; Little, 1983; Thompson, 1997).

La principale attrattiva dell'approccio randomizzato è il fatto che le sue conclusioni non dipendono da un particolare modello scelto per i valori $\boldsymbol{Y}$; mentre la distribuzione di probabilità $P(\boldsymbol{I}|\boldsymbol{Y})$ è nota, la scelta di un modello $P(\boldsymbol{Y})$ introduce un elemento di soggettività. Tuttavia, l'"oggettività" dell'approccio randomizzato viene perduta quando si hanno delle deviazioni dal campionamento probabilistico, con l'introduzione di una qualche fonte di *errore* (Little, 1982).

In particolare, gli errori legati alle indagini campionarie vengono tradizionalmente suddivisi in due categorie (Sarndal et al., 1992): *errori campionari*, ovvero errori dovuti all'osservazione di un campione e non dell'intera popolazione, ed *errori non campionari*, che comprendono tutte le altre fonti di errore.

Gli errori appartenenti alla seconda categoria possono essere ulteriormente suddivisi in errori dovuti a *non osservazione* e errori *nelle osservazioni*. Il primo caso si riferisce all'impossibilità di ottenere alcuni dei dati di interesse, mentre il secondo caso comprende gli errori di misura, che possono

essere causati per esempio dal questionario utilizzato o dall'intervistatore, e gli errori di elaborazione, per esempio nelle trascrizioni e nell'*editing* dei dati rilevati.

Gli errori di non osservazione comprendono gli errori di mancata copertura (*undercoverage*) e le *mancate risposte* (*nonresponses*), che possono essere totali o parziali. Indicando nuovamente con y_{ij} il valore osservato per l'unità campionaria i relativamente alla variabile j, dove $i = 1, \ldots, n$ e $j = 1, \ldots, J$, si possono distinguere due casi:

- mancata risposta totale (*unit nonresponse*) per l'unità k: l'intero vettore $y_k = (y_{k1}, \ldots, y_{kJ})$ è mancante (*missing*);
- mancata risposta parziale (*item nonresponse*) per l'unità k: almeno uno ma non tutti i J elementi del vettore y_k sono mancanti.

Esempi di mancata risposta totale sono il rifiuto di partecipare all'intervista o la mancata riconsegna del questionario; le mancate risposte parziali si riferiscono tipicamente alla situazione in cui l'unità campionata partecipa all'indagine ma non fornisce risposta a qualcuno dei quesiti del questionario.

Come sottolineato da Rubin (1983), quando si hanno delle mancate risposte l'approccio randomizzato all'inferenza da popolazioni finite deve essere modificato, dal momento che il meccanismo di selezione noto, $P(\boldsymbol{I}|\boldsymbol{Y})$, non è più sufficiente per derivare inferenze per le quantità di interesse della popolazione $Q(\boldsymbol{Y})$. Si rende infatti necessaria l'introduzione di un modello per il *meccanismo di risposta*; utilizzando l'indicatore R_{ij} che assume valore 1 quando l'unità campionaria i risponde al quesito j, 0 altrimenti, tale modello prende la forma $P(\boldsymbol{R}|\boldsymbol{Y}, \boldsymbol{I})$. Solitamente, tuttavia, tale compito viene semplificato con l'introduzione dell'ipotesi di *ignorabilità* del meccanismo di mancata risposta; tale concetto, presentato in modo formale nei prossimi paragrafi, corrisponde ad assumere che il meccanismo $P(\boldsymbol{R}|\boldsymbol{Y}, \boldsymbol{I})$ possa non essere inserito nel procedimento di inferenza, sotto particolari ipotesi.

Alcuni autori (Oh and Scheuren, 1983) parlano per esempio di approccio *quasi randomizzato* quando il meccanismo di mancata risposta può essere considerato come un altro stadio di campionamento probabilistico. Solitamente questa ipotesi viene introdotta ipotizzando che il meccanismo di mancata risposta sia costante, e quindi ignorabile, all'interno di opportuni sottogruppi della popolazione. In questo caso il modello di mancata risposta, pur essendo presente, è di tipo *implicito*, motivo per cui l'approccio prende il nome di *quasi* randomizzato.

L'introduzione di un modello *esplicito* per il meccanismo di mancata risposta è concettualmente immediato, invece, nell'ambito dell'approccio basato su modello. In questo caso, i modelli per $\boldsymbol{Y}$ e per $\boldsymbol{I}$ devono semplicemente

essere estesi per includere anche il modello per $\boldsymbol{R}$. In particolare, autori come Rubin e Little (Rubin, 1983; Little, 1983) preferiscono, all'interno di questo approccio, la specificazione di un modello di tipo bayesiano. L'impostazione bayesiana comporta, nell'ambito delle mancate risposte, il calcolo della distribuzione *a posteriori* della quantità $Q(\boldsymbol{Y})$ condizionando per i valori osservati e per i modelli *a priori* ipotizzati per $P(\boldsymbol{Y})$, $P(\boldsymbol{I}|\boldsymbol{Y})$, $P(\boldsymbol{R}|\boldsymbol{Y},\boldsymbol{I})$ e per i relativi parametri. Proprio l'esplicito condizionamento rispetto ai valori osservati della popolazione è, secondo Rubin, la più diretta giustificazione dell'utilizzo di modelli bayesiani (Rubin, 1983).

L'imputazione multipla, che viene considerata in dettaglio nel prossimo capitolo, può essere considerata un possibile "compromesso" tra l'approccio randomizzato e l'approccio basato su modello bayesiano: il modello a posteriori bayesiano viene infatti utilizzato solo per generare le imputazioni, mentre per analizzare i datasets completati si ricorre solitamente alle tecniche proprie dell'inferenza randomizzata (Madow et al., 1983).

1.1.1 L'inferenza bayesiana per quantità della popolazione

Seguendo l'impostazione di Rubin (1987), consideriamo la notazione già precedentemente introdotta distinguendo però tra variabili completamente osservate e variabili con valori mancanti. In particolare, sia $\boldsymbol{X}$ la matrice delle variabili osservate, avente elementi x_{ik} dove $i = 1, \ldots, N$ e $k = 1, \ldots, K$, mentre $\boldsymbol{Y}$ rappresenta la matrice delle variabili con valori mancanti, avente elementi y_{ij} con $i = 1, \ldots, N$ e $j = 1, \ldots, J$. Per esempio, in un'indagine che ha come obiettivo la stima del reddito medio in una data popolazione le variabili x_{ik} includono solitamente informazioni come il genere e l'età degli intervistati, mentre tra le variabili y_{ij} vi è il reddito individuale, per il quale si hanno verosimilmente delle mancate risposte.

Come già specificato, la variabile dicotomica I_i rappresenta l'indicatore di inclusione nel campione: assume valore 1 per gli n individui della popolazione che entrano a far parte del campione, 0 altrimenti. Il valore assunto da $\boldsymbol{I}$ per una data unità campionaria è il risultato dello schema di campionamento scelto, ed è perciò solitamente osservato per tutti gli elementi della popolazione.

In presenza di non risposta occorre specificare anche la variabile dicotomica R_{ij} che assume valore 1 quando l'individuo i della popolazione risponde al quesito misurato dalla variabile Y_j, $j = 1, \ldots, J$, 0 altrimenti. L'ipotesi in questo caso è che R_{ij} sia nota quando $I_i = 1$, ovvero quando l'unità $i-esima$ della popolazione entra a far parte del campione. In particolare, $\boldsymbol{I}$ introdu-

ce una partizione dei valori $\boldsymbol{R}$ e $\boldsymbol{Y}$: $\boldsymbol{R} = (\boldsymbol{R}_{inc}, \boldsymbol{R}_{esc})$ e $\boldsymbol{Y} = (\boldsymbol{Y}_{inc}, \boldsymbol{Y}_{esc})$, dove il suffisso *inc* indica le unità incluse nel campione ($I_i = 1$) mentre *esc* indica quelle escluse dal campione ($I_i = 0$). Inoltre, i valori dell'indicatore $\boldsymbol{R}_{inc}$, sempre osservati, determinano un'ulteriore suddivisione dei valori $\boldsymbol{Y}$: $\boldsymbol{Y}_{inc} = (\boldsymbol{Y}_{oss}, \boldsymbol{Y}_{mis})$, dove il suffisso *oss* indica le unità campionarie per cui il valore della variabile Y_j è osservato ($I_i = 1$ e $R_{inc\,ij} = 1$), *mis* quelle per cui è missing ($I_i = 1$ e $R_{inc\,ij} = 0$). Allora, i valori non osservati di $\boldsymbol{Y}$ possono essere indicati con $\boldsymbol{Y}_{nos} = (\boldsymbol{Y}_{esc}, \boldsymbol{Y}_{mis})$.

Per semplificare la notazione, spesso i due indicatori $\boldsymbol{I}$ e $\boldsymbol{R}$ vengono "uniti" nell'unico indicatore $M_{ij} = I_i R_{ij}$, che prende il nome di indicatore dei valori mancanti, e che risulta definito per le variabili potenzialmente soggette a missing values (Little and Rubin, 2002). In particolare, M_{ij} assume valore 1 quando per l'individuo i il valore y_{ij} fa parte di $\boldsymbol{Y}_{oss}$, 0 altrimenti.

Per compiere inferenza per una qualche quantità di interesse $Q = Q(\boldsymbol{X}, \boldsymbol{Y})$ relativa alla popolazione è necessario introdurre una distribuzione di probabilità per gli indicatori $\boldsymbol{I}$ e $\boldsymbol{R}$. Per esempio la quantità di interesse potrebbe essere rappresentata dal reddito medio $\overline{\boldsymbol{Y}}_j$ per sottogruppi della popolazione definiti in base alle covariate osservate $\boldsymbol{X}$. In particolare, la distribuzione scelta per $P(I|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{R})$ rappresenta il meccanismo di selezione, mentre quella per $P(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{Y})$ il meccanismo di risposta. L'inferenza bayesiana per la quantità Q si baserà allora sulla sua distribuzione a posteriori, che condiziona per i valori osservati: $P(Q|\boldsymbol{X}, \boldsymbol{Y}_{oss}, \boldsymbol{R}_{inc}, \boldsymbol{I})$. Tale distribuzione può essere ricavata integrando le quantità non osservate, ovvero i valori $\boldsymbol{Y}_{nos}$ e $\boldsymbol{R}_{esc}$, dal rapporto tra le distribuzioni congiunte $P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{R}, \boldsymbol{I})$ e $P(\boldsymbol{X}, \boldsymbol{Y}_{oss}, \boldsymbol{R}_{inc}, \boldsymbol{I})$:

$$\begin{aligned} P(Q|\boldsymbol{X}, \boldsymbol{Y}_{oss}, \boldsymbol{R}_{inc}, \boldsymbol{I}) &= \int_{Y(Q)} P(\boldsymbol{Y}_{nos}|\boldsymbol{X}, \boldsymbol{Y}_{oss}, \boldsymbol{R}_{inc}, \boldsymbol{I})\, d\boldsymbol{Y}_{nos} = \\ &= \frac{\int P(\boldsymbol{X}, \boldsymbol{Y}) P(\boldsymbol{R}|X, \boldsymbol{Y}) P(\boldsymbol{I}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{R})\, d\boldsymbol{R}_{esc}}{\int\int P(\boldsymbol{X}, \boldsymbol{Y}) P(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{Y}) P(\boldsymbol{I}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{R})\, d\boldsymbol{R}_{esc} d\boldsymbol{Y}_{nos}} \end{aligned}$$

dove $Y(Q) = \boldsymbol{Y}_{nos}|Q(\boldsymbol{X}, \boldsymbol{Y}) = Q'$.

Il meccanismo di selezione del campione è detto *ignorabile* (Rubin, 1987) per i valori osservati $(\boldsymbol{X}, \boldsymbol{Y}_{oss}, \boldsymbol{R}_{inc}, \boldsymbol{I})$ se si ha:

$$\begin{aligned} P(Q|\boldsymbol{X}, \boldsymbol{Y}_{oss}, \boldsymbol{R}_{inc}, \boldsymbol{I}) &= \int P(\boldsymbol{Y}_{nos}|\boldsymbol{X}, \boldsymbol{Y}_{oss}, \boldsymbol{R}_{inc}, \boldsymbol{I})\, d\boldsymbol{Y}_{nos} \\ &= \frac{\int P(\boldsymbol{X}, \boldsymbol{Y}) P(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{Y})\, d\boldsymbol{R}_{esc}}{\int\int P(\boldsymbol{X}, \boldsymbol{Y}) P(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{Y})\, d\boldsymbol{R}_{esc}\, d\boldsymbol{Y}_{nos}} \end{aligned}$$

ovvero se:

$$P(\boldsymbol{Y}_{nos}|\boldsymbol{X}, \boldsymbol{Y}_{oss}, \boldsymbol{R}_{inc}, \boldsymbol{I}) = P(\boldsymbol{Y}_{nos}|\boldsymbol{X}, \boldsymbol{Y}_{oss}, \boldsymbol{R}_{inc}). \tag{1.1}$$

Tale definizione implica che la distribuzione a posteriori di $\boldsymbol{Y}_{nos}$, e quindi quella di Q, è la stessa per ogni meccanismo di selezione del campione ignorabile. In particolare, tutti i meccanismi probabilistici che non dipendono da valori non osservati, ovvero tali che $P(\boldsymbol{I}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{R}) = P(\boldsymbol{I}|\boldsymbol{X})$, soddisfano tale condizione. E' questo il caso degli schemi di campionamento più comunemente utilizzati per esempio nella statistica ufficiale, dove spesso si introduce una stratificazione della popolazione definita in base a caratteristiche completamente osservate.

La condizione (1.1) può essere specificata in alternativa come:

$$P(\boldsymbol{I}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{R}_{inc}) = P(\boldsymbol{I}|\boldsymbol{X},\boldsymbol{Y}_{oss},\boldsymbol{R}_{inc}). \tag{1.2}$$

Per dimostrare l'uguaglianza delle condizioni (1.1) e (1.2) basta osservare che:

$$\begin{aligned} & P(\boldsymbol{Y}_{nos}|\boldsymbol{X},\boldsymbol{Y}_{oss},\boldsymbol{R}_{inc},\boldsymbol{I}) = \frac{P(\boldsymbol{Y}_{nos},\boldsymbol{X},\boldsymbol{Y}_{oss},\boldsymbol{R}_{inc},\boldsymbol{I})}{P(\boldsymbol{X},\boldsymbol{Y}_{oss},\boldsymbol{R}_{inc},I)} = \\ = & \frac{P(\boldsymbol{I}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{R}_{inc})P(\boldsymbol{Y}_{nos}|\boldsymbol{Y}_{oss},\boldsymbol{X},\boldsymbol{R}_{inc})P(\boldsymbol{Y}_{oss},\boldsymbol{X},\boldsymbol{R}_{inc})}{P(\boldsymbol{I}|\boldsymbol{X},\boldsymbol{Y}_{oss},\boldsymbol{R}_{inc})P(\boldsymbol{Y}_{oss},\boldsymbol{X},\boldsymbol{R}_{inc})} = \\ = & \frac{P(\boldsymbol{I}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{R}_{inc})P(Y_{nos}|\boldsymbol{Y}_{oss},\boldsymbol{X},\boldsymbol{R}_{inc})}{P(\boldsymbol{I}|\boldsymbol{X},\boldsymbol{Y}_{oss},\boldsymbol{R}_{inc})} \end{aligned}$$

risulta uguale a $P(\boldsymbol{Y}_{nos}|\boldsymbol{X},\boldsymbol{Y}_{oss},\boldsymbol{R}_{inc})$, come richiesto dalla condizione (1.1), proprio se vale la condizione (1.2).

Una definizione analoga può essere introdotta relativamente all'ignorabilità del meccanismo di risposta.

Il meccanismo di risposta è ignorabile (Rubin, 1987) se:

$$\begin{aligned} P(\boldsymbol{Y}_{nos}|\boldsymbol{X},\boldsymbol{Y}_{oss},\boldsymbol{R}_{inc},\boldsymbol{I}) &= \frac{\int P(\boldsymbol{X},\boldsymbol{Y})P(\boldsymbol{I}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{R})\, d\boldsymbol{R}_{esc}}{\int\int P(\boldsymbol{X},\boldsymbol{Y})P(\boldsymbol{I}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{R})\, d\boldsymbol{R}_{esc}\, d\boldsymbol{Y}_{nos}} = \\ &= P(\boldsymbol{Y}_{nos}|\boldsymbol{X},\boldsymbol{Y}_{oss},\boldsymbol{I}) \end{aligned} \tag{1.3}$$

ovvero se la distribuzione a posteriori di $\boldsymbol{Y}_{nos}$, e quindi quella di Q, è la stessa per ogni meccanismo di risposta ignorabile.

Qualora anche il meccanismo di selezione del campione sia ignorabile, la condizione (1.3) diventa:

$$P(\boldsymbol{Y}_{nos}|\boldsymbol{X},\boldsymbol{Y}_{oss},\boldsymbol{R}_{inc},\boldsymbol{I}) = P(\boldsymbol{Y}_{nos}|\boldsymbol{X},\boldsymbol{Y}_{oss}). \tag{1.4}$$

che è equivalente alla condizione:

$$P(\boldsymbol{R}_{inc}|\boldsymbol{X},\boldsymbol{Y}) = P(\boldsymbol{R}_{inc}|\boldsymbol{X},\boldsymbol{Y}_{oss}) \tag{1.5}$$

Sotto la condizione (1.4), allora, la distribuzione a posteriori di $\boldsymbol{Y}_{nos}$, da cui è possibile ricavare quella di $\boldsymbol{Y}_{mis}$ e Q, è tale che:

$$P(\boldsymbol{Y}_{nos}|\boldsymbol{X}, \boldsymbol{Y}_{oss}, \boldsymbol{R}_{inc}, I) = P(\boldsymbol{Y}_{nos}|\boldsymbol{X}, \boldsymbol{Y}_{oss}) = \frac{P(\boldsymbol{X}, \boldsymbol{Y})}{\int P(\boldsymbol{X}, \boldsymbol{Y})\, d\boldsymbol{Y}_{nos}}. \quad (1.6)$$

Questo equivale a dire che si possono ignorare i meccanismi che generano i valori mancanti se valgono le condizioni (1.1) e (1.3). Come sottolineato da Rubin (1987), mentre la prima condizione è solitamente soddisfatta per i principali schemi di campionamento, la seconda condizione, relativa all'ignorabilità del meccanismo di risposta, non sarà in generale rispettata. Per esempio, nel caso in cui i valori mancanti riguardano variabili "delicate", come quelle che misurano quantità di reddito, è possibile che la probabilità di non fornire la risposta dipenda proprio dal valore del reddito.

1.1.2 L'inferenza parametrica con meccanismo di campionamento ignorabile

Le condizioni di ignorabilità del meccanismo di selezione del campione e del meccanismo di risposta possono essere espresse in termini di inferenza parametrica, bayesiana e non. In particolare, consideriamo il caso in cui si ipotizzi l'ignorabilità del meccanismo di selezione del campione, e concentriamo l'attenzione sulla condizione di ignorabilità del meccanismo di risposta. Consideriamo un modello parametrico per i dati $f(\boldsymbol{X}, \boldsymbol{Y}|\theta)$ e un modello per il meccanismo di risposta $g(\boldsymbol{M}|\boldsymbol{X}, \boldsymbol{Y}, \psi)$; nell'ottica bayesiana si specifica anche la distribuzione a priori per i parametri θ e ψ, che è in generale $p(\theta, \psi)$. Consideriamo inoltre il solo indicatore $\boldsymbol{M} = \boldsymbol{R}\,\boldsymbol{I}$ in quanto, avendo ipotizzato che il meccanismo di selezione del campione è ignorabile e che quindi l'indicatore $\boldsymbol{I}$ può essere escluso dall'analisi, i valori di $\boldsymbol{M}$ sono in relazione biunivoca con quelli di $\boldsymbol{R}_{inc}$. Inoltre, poichè non si considera il passo di campionamento, la partizione di interesse per i valori di $\boldsymbol{Y}$ è in questo caso $\boldsymbol{Y} = (\boldsymbol{Y}_{oss}, \boldsymbol{Y}_{mis})$.

In questa formulazione il meccanismo di mancata risposta è detto ignorabile per l'inferenza parametrica bayesiana se:

$$\begin{aligned} g(\boldsymbol{M}|\boldsymbol{X}, \boldsymbol{Y}, \psi) &= g(\boldsymbol{M}|\boldsymbol{X}, \boldsymbol{Y}_{mis}, \boldsymbol{Y}_{oss}, \psi) = g(\boldsymbol{M}|\boldsymbol{X}, \boldsymbol{Y}_{oss}, \psi) \quad (1.7) \\ p(\theta, \psi) &= p(\theta)p(\psi) \quad (1.8) \end{aligned}$$

La condizione (1.7) equivale a dire che i dati mancanti sono *mancanti a caso* (*missing at random*, MAR), mentre la condizione (1.8) corrisponde

all'indipendenza a priori dei parametri che caratterizzano rispettivamente il modello per i dati e quello per il meccanismo di mancata risposta[1].

Per vedere che le due condizioni (1.7) e (1.8) corrispondono all'ignorabilità del meccanismo di mancata risposta definito nell'ottica non parametrica di Rubin (1987) nel paragrafo 1.1.1 (condizione (1.4)), basta osservare che ponendo $\boldsymbol{Y}_{nos} = \boldsymbol{Y}_{mis}$ ed utilizzando il solo indicatore $\boldsymbol{M}$ tale condizione diventa:

$$P(\boldsymbol{Y}_{mis}|\boldsymbol{X}, \boldsymbol{Y}_{oss}, \boldsymbol{M}) = P(\boldsymbol{Y}_{mis}|\boldsymbol{X}, \boldsymbol{Y}_{oss}) \tag{1.9}$$

Allora, per dimostrare che tale condizione è sempre rispettata quando i dati sono MAR e i due parametri θ e ψ sono a priori indipendenti basta osservare che:

$$P(\boldsymbol{Y}_{mis}|\boldsymbol{X}, \boldsymbol{Y}_{oss}, \boldsymbol{M}) = \frac{\int\int f(\boldsymbol{X}, \boldsymbol{Y}|\theta) g(\boldsymbol{M}|\boldsymbol{X}, \boldsymbol{Y}, \psi) p(\theta, \psi)\, d\theta d\psi}{\int\int\int f(\boldsymbol{X}, \boldsymbol{Y}|\theta) g(\boldsymbol{M}|\boldsymbol{X}, \boldsymbol{Y}, \psi) p(\theta, \psi)\, d\theta\, d\psi\, d\boldsymbol{Y}_{mis}} =$$

$$\begin{aligned}
&\overset{MAR}{=} \quad \frac{\int\int f(\boldsymbol{X}, \boldsymbol{Y}|\theta) g(\boldsymbol{M}|\boldsymbol{X}, \boldsymbol{Y}_{oss}, \psi) p(\theta, \psi)\, d\theta d\psi}{\int\int\int f(\boldsymbol{X}, \boldsymbol{Y}|\theta) g(\boldsymbol{M}|\boldsymbol{X}, \boldsymbol{Y}_{oss}, \psi) p(\theta, \psi)\, d\theta\, d\psi\, d\boldsymbol{Y}_{mis}} = \\
&\overset{indip.}{=} \quad \frac{\int f(\boldsymbol{X}, \boldsymbol{Y}|\theta) p(\theta)\, d\theta}{\int\int f(\boldsymbol{X}, \boldsymbol{Y}|\theta) p(\theta)\, d\theta d\boldsymbol{Y}_{mis}} = \\
&= \quad P(\boldsymbol{Y}_{mis}|\boldsymbol{X}, \boldsymbol{Y}_{oss})
\end{aligned}$$

La definizione di ignorabilità normalmente utilizzata nella letteratura relativa all'inferenza parametrica in presenza di dati mancanti fa riferimento alle condizioni (1.7) e (1.8). Per esempio, quando tali condizioni sono soddisfatte Little and Rubin (2002) e Schafer (1997) parlano di *verosimiglianza ignorando il meccanismo di mancata risposta*:

$$L(\theta|\boldsymbol{Y}_{oss}, \boldsymbol{X}) \propto f(\boldsymbol{Y}_{oss}, \boldsymbol{X}|\theta) = \int f(\boldsymbol{Y}, \boldsymbol{X}|\theta)\, d\boldsymbol{Y}_{mis} \tag{1.10}$$

e, in ottica bayesiana, di *distribuzione a posteriori ignorando il meccanismo di mancata risposta*:

$$P(\theta|\boldsymbol{Y}_{oss}, \boldsymbol{X}) \propto L(\theta|\boldsymbol{Y}_{oss}, \boldsymbol{X}) p(\theta). \tag{1.11}$$

Questo significa che sotto le condizioni (1.7) e (1.8) si può compiere il processo di inferenza sul parametro della superpopolazione θ senza prendere in

[1]Nell'ottica non bayesiana la condizione (1.8) richiede che i parametri θ e ψ siano distinti, ovvero che lo spazio parametrico congiunto $\Omega_{\theta,\psi}$ sia uguale al prodotto dei due spazi parametrici, $\Omega_\theta \times \Omega_\psi$, (Little and Rubin, 2002).

considerazione il meccanismo di mancata risposta ed il suo parametro ψ, ma utilizzando la verosimiglianza (1.10) o, in ottica bayesiana, la distribuzione a posteriori (1.11). Quest'ultima operazione, in particolare, rappresenta un passo necessario anche quando si è interessati a compiere inferenza bayesiana per una quantità della popolazione $Q(\boldsymbol{Y})$. In particolare, Little (1982) sottolinea che i risultati derivati nell'ambito dell'inferenza parametrica bayesiana per i parametri che caratterizzano la superpopolazione, θ nelle espressioni precedenti, possono essere facilmente adattati per derivare inferenze bayesiane relative a quantità della popolazione, come per esempio $\overline{\boldsymbol{Y}}$. In termini "pratici" tuttavia, le espressioni (1.10) e (1.11) sono in generale funzioni molto complicate di θ e la loro trattazione richiede tecniche computazionali speciali (Schafer, 1997). Esistono però delle situazioni in cui tale compito risulta semplificato, per esempio quando il *pattern* dei dati mancanti assume forme particolari; tale concetto viene ripreso e specificato meglio nel proseguo di questo capitolo.

La condizione (1.7), che richiede che i dati siano mancanti a caso, può essere resa più restrittiva richiedendo che:

$$g(\boldsymbol{M}|\boldsymbol{X},\boldsymbol{Y},\psi) \;=\; g(\boldsymbol{M}|\boldsymbol{X},\boldsymbol{Y}_{mis},\boldsymbol{Y}_{oss},\psi) = g(\boldsymbol{M}|\psi) \qquad (1.12)$$

In questo caso si dice che i dati sono *mancanti completamente a caso* (*missing completely at random*, MCAR): il meccanismo che ha generato le mancate risposte non dipende da alcun valore $\boldsymbol{X}$ o $\boldsymbol{Y}$, ovvero i dati osservati sono un campione casuale dell'insieme delle osservazioni. L'ipotesi MCAR è solitamente ritenuta troppo restrittiva nelle maggiori situazioni reali (Little and Rubin, 2002); per tale motivo l'ipotesi normalmente utilizzata è quella di dati MAR, che come visto è sufficiente per la condizione di ignorabilità.

Quando invece il meccanismo di mancata risposta dipende anche da quantità non osservate, e quindi la formulazione $g(\boldsymbol{M}|\boldsymbol{X},\boldsymbol{Y}_{mis},\boldsymbol{Y}_{oss},\psi)$ non può essere semplificata, allora i dati mancanti sono detti *mancanti non a caso* (*missing not at random*, MNAR).

In questo caso non risulta possibile, volendo compiere inferenza sul parametro θ della popolazione, utilizzare la verosimiglianza semplificata (1.10) ma si deve considerare la verosimiglianza completa:

$$L(\theta,\psi|\boldsymbol{Y}_{oss},\boldsymbol{X},\boldsymbol{M}) \propto f(\boldsymbol{Y}_{oss},\boldsymbol{X},\boldsymbol{M}|\theta,\psi). \qquad (1.13)$$

1.1.3 L'inferenza parametrica con meccanismo di mancata risposta non ignorabile

La letteratura che si occupa dell'inferenza da modelli per i quali sia ipotizzabile l'ignorabilità del meccanismo di selezione del campione ma non quella del

meccanismo di mancata risposta fa riferimento a due possibili fattorizzazioni della distribuzione congiunta di $\boldsymbol{Y}$, $\boldsymbol{X}$ ed $\boldsymbol{M}$ (Little, 1993; Little and Rubin, 2002). La prima fattorizzazione corrisponde all'approccio dei *selection models* (Heckman, 1979):

$$f(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{M}|\theta, \psi) = f(\boldsymbol{Y}, \boldsymbol{X}|\theta) f(\boldsymbol{M}|\boldsymbol{Y}, \boldsymbol{X}, \psi). \tag{1.14}$$

Questa formulazione fa specifico riferimento al modello scelto per i dati, $f(\boldsymbol{Y}, \boldsymbol{X}|\theta)$, ed a quello per il meccanismo di mancata risposta, $f(\boldsymbol{M}|\boldsymbol{Y}, \boldsymbol{X}, \psi)$. Per esempio, il modello per i dati potrebbe fare riferimento ad una normale multivariata caratterizzata dal parametro $\theta = (\mu, \boldsymbol{\Sigma})$, mentre il meccanismo di mancata risposta potrebbe essere modellato con una distribuzione bernoulliana caratterizzata dal parametro ψ.

Al contrario, la fattorizzazione che fa riferimento ai *pattern mixture models* è del tipo (Little, 1993):

$$f(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{M}|\varphi, \upsilon) = f(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{M}, \varphi) f(\boldsymbol{M}|\upsilon). \tag{1.15}$$

In questo caso il modello di diretto interesse è quello per i dati, che condiziona per l'indicatore di mancata risposta $\boldsymbol{M}$; ciò significa, in pratica, che il modello $f(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{M}, \varphi)$ deriva dalle sue specificazioni all'interno degli strati definiti dal *pattern* dei dati mancanti, ovvero dall'indicatore $\boldsymbol{M}$, mentre la distribuzione di $\boldsymbol{M}$ modella l'incidenza dei *pattern*.

Tra le due parametrizzazioni non esiste una netta separazione. Come illustrato da Little (1993), infatti, le fattorizzazioni (1.14) e (1.15) possono essere viste come derivanti da un'unica tipologia di modelli, i *pattern-set mixture models*.

I modelli per meccanismi di risposta non ignorabile del tipo (1.14) e (1.15) presentano numerose problematiche che sono state affrontate, tra gli altri, da Little and Rubin (2002), Rubin (1987), Little (1993) e da Little (1995) nel contesto delle misure ripetute.

Tali problematiche fanno riferimento, innanzitutto, alla possibile non identificabilità del modello scelto. Per esempio, nel caso dei pattern mixture models (1.15) potrebbero non essere a disposizione osservazioni per stimare la distribuzione $f(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{M}, \varphi)$ nel *pattern* di dati completamente mancanti; una possibile soluzione, allora, è rappresentata dall'introduzione di una distribuzione a priori su tali parametri (Rubin, 1977). Altre soluzioni per i problemi di stima di questi modelli sono rappresentati dai *follow-ups*, ovvero dai dati ottenuti ricontattando alcuni dei non rispondenti alla prima intervista (Rubin, 1987; Little and Rubin, 2002), e dalla restrizione della stima a particolari strutture o meglio *pattern* di dati mancanti.

La specificazione di una distribuzione a priori per i parametri che caratterizzano le formulazioni (1.14) e (1.15) è uno degli elementi che rende necessaria un'analisi di *sensitività*. Quando si ipotizza una mancata risposta non ignorabile si dovrebbero infatti specificare diversi possibili modelli per tale mancata risposta, per poter valutare fino a che punto le inferenze risultano influenzate dalla scelta del modello. Il procedimento suggerito è quello di testare modelli diversi per il meccanismo di risposta, variando le variabili da cui questo dipende direttamente (Jansen et al., 2006; Little and Rubin, 2002).

L'imputazione multipla, oggetto del prossimo capitolo, può essere utilizzata per svolgere analisi di sensitività delle mancate risposte rispetto a meccanismi di tipo MNAR; tale approccio all'imputazione multipla viene preso in considerazione nell'applicazione relativa ai dati provenienti dall'indagine sulle Forze di Lavoro del Comune di Firenze nel capitolo 5. In questo senso l'imputazione multipla rappresenta un utile strumento per studiare la sensitività dei modelli di risposta non ignorabili (Madow et al., 1983).

1.2 Il trattamento delle mancate risposte

Un corretto trattamento delle mancate risposte nelle indagini campionarie dovrebbe coinvolgere più fasi della programmazione e implementazione dell'indagine stessa. In particolare, la non risposta dovrebbe essere considerata durante la pianificazione dell'indagine, la raccolta, l'analisi dei dati e la presentazione dei risultati (Madow et al., 1983). Sarndal et al. (1992) (capitolo 15) individuano la seguente classificazione delle strategie da utilizzare per trattare le mancate risposte nelle indagini campionarie:

- misure atte a ridurre il numero di mancate risposte prima e durante la raccolta dei dati;
- speciali tecniche di raccolta dei dati e di stima per ottenere stime non distorte in presenza di mancate risposte;
- ipotesi relative al meccanismo di risposta e alle relazioni tra le variabili, utilizzate per costruire stimatori che "correggano" per le mancate risposte.

Nel primo punto rientrano per esempio una corretta e pianificata realizzazione del questionario, la formazione degli intervistatori, durante la quale particolare attenzione può essere riservata ai quesiti particolarmente a rischio di non risposta, e i tentativi di ricontattare le unità inizialmente non intervistate.

Quando, nonostante gli accorgimenti, si ottengono comunque delle mancate risposte, una possibile soluzione è ricampionare un sottoinsieme dei non rispondenti cercando di ottenere da loro informazioni complete. Utilizzando tali dati aggiuntivi è possibile ottenere stime non distorte attraverso stimatori simili a quelli utilizzati per la stratificazione. Un'altra tecnica "speciale" è la somministrazione *casualizzata* dei quesiti particolarmente delicati; in questo caso l'obiettivo è proteggere l'identità dell'intervistato, sotto l'ipotesi che questo possa servire a ridurre fortemente o addirittura eliminare le mancate risposte.

Il principale problema di queste tecniche speciali è che possono essere molto costose e lente da implementare, specialmente nel caso di indagini campionarie con piano di campionamento complesso e che coinvolgono molte unità. Inoltre, senza introdurre particolari ipotesi sul meccanismo che ha generato le mancate risposte, possono presentarsi particolari problemi di stima. Ecco perchè il trattamento delle non risposte rende necessaria, in pratica, l'introduzione di un modello per le non risposte (paragrafo 1.1). In particolare, Sarndal et al. (1992) sottolineano che lo statistico dovrebbe cercare di formulare un buon modello per il meccanismo di risposta che venga utilizzato, assieme a variabili ausiliarie, per costruire stimatori che correggono per le mancate risposte.

1.2.1 Le tecniche di ponderazione e imputazione sotto l'ipotesi MAR

Supponiamo che i dati provenienti da un'indagine condotta secondo un piano di campionamento probabilistico vengano analizzati, e che l'interesse risieda nella stima della media nella popolazione di una data variabile $\boldsymbol{Y}_j = \boldsymbol{Y}$. Questo rientra nella problematica dell'inferenza per quantità della popolazione quando il meccanismo di campionamento è ignorabile (paragrafo 1.1.1). Lo stimatore Horvitz-Thompson utilizzato in assenza di mancata risposta pondera i valori y_i, $i = 1, \ldots, n$, rilevati per tale variabile con le probabilità individuali di inclusione nel campione π_i:

$$\widehat{\overline{\boldsymbol{Y}}} = \sum_{i=1}^{n} y_i \frac{1}{\pi_i} / \sum_{i=1}^{n} \frac{1}{\pi_i}. \tag{1.16}$$

Nel caso di disegno di campionamento complesso, per esempio a più stadi, i pesi π_i utilizzati nella (1.16) considerano tutte le fasi di campionamento.

In presenza di mancate risposte l'analizzatore potrebbe, in prima istanza, ipotizzare che coloro che hanno un valore mancante per la variabile $\boldsymbol{Y}$ sono uguali a coloro che hanno fornito una risposta, utilizzando lo stimatore (1.16)

per le sole unita con valori y_i osservati. Tale metodo di trattare le mancate risposte, che ha il pregio di essere molto semplice e di immediata applicazione, prende il nome di *analisi dei casi completi* (*complete case analysis*)[2]; l'ipotesi cui fa riferimento è che i dati siano mancanti completamente a caso (MCAR, paragrafo 1.12), ovvero che i rispondenti sono un sottoinsieme casuale degli intervistati. Questo metodo di trattare le mancate risposte, che può assumere la forma di *analisi dei casi disponibili* (*available case analysis*) nel caso di mancate risposte per più di una variabile[3], dovrebbe essere utilizzato solamente in situazioni in cui il numero di mancate risposte è limitato. Come sottolineato da Little and Rubin (2002), infatti, la perdita di informazione che si ha nell'utilizzare i soli dati disponibili comporta in generale una minor precisione quando i dati non sono MCAR.

Un'estensione più naturale dello stimatore (1.16) quando si hanno delle mancate risposte è l'inclusione delle probabilità di risposta ϕ_i, con $\phi_i = E(M_i|y_i, I_i)$, dove gli indicatori $\boldsymbol{M}$ e $\boldsymbol{I}$ sono gli stessi introdotti nella sezione 1.1.1. In questo caso:

$$\widehat{\overline{\boldsymbol{Y}}} = \sum_{i=1}^{n} y_i \frac{1}{\pi_i}\frac{1}{\phi_i} / \sum_{i=1}^{n} \frac{1}{\pi_i}\frac{1}{\phi_i}. \tag{1.17}$$

Per stimare i pesi ϕ_i, che saranno potenzialmente diversi per ogni unità i, si fa solitamente ricorso alle cosiddette *celle di ponderazione*: le unità campionarie vengono suddivise in K gruppi definiti in base alle covariate osservate, e la probabilità di risposta ϕ_k, ipotizzata costante per tutti gli individui in quel gruppo, viene calcolata come rapporto tra numero di unità rispondenti e numerosità campionaria nel gruppo $k - esimo$. L'utilizzo delle celle di ponderazione corrisponde ad un modello in cui si ipotizzino dati MAR (condizione (1.7)): il meccanismo che genera le mancate risposte può dipendere da variabili osservate, ma l'aver utilizzato tali variabili per formare le celle di ponderazione garantisce dati MCAR all'interno di ogni cella (Little and Rubin, 2002). Quando il numero delle covariate osservate è particolarmente elevato, Little (1986) suggerisce l'applicazione del concetto di *propensity score* (Rosenbaum and Rubin, 1983) alle mancate risposte; tale approccio viene

[2]La *complete case analysis* è il metodo di trattare i dati mancanti utilizzato dai più comuni software che svolgono analisi statistiche; ecco perchè spesso, nelle applicazioni in cui le mancate risposte vengono trattate con tecniche più sofisticate, i risultati della *complete case analysis* vengono comunque calcolati e presi come riferimento.

[3]Nella *available case analysis* le analisi univariate includono le osservazioni disponibili per ciascuna variabile, nonostante i possibili missing values per le altre variabili. Tale metodo, che ha il pregio di utilizzare tutte le informazioni disponibili, ha il difetto di introdurre una diversa numerosità tra le analisi univariate, oltre a quello di ipotizzare dati MCAR come l'analisi dei casi completi (Little and Rubin, 2002).

ripreso e considerato in dettaglio nel capitolo 4, relativamente ad un metodo di diagnostica per l'ipotesi MAR recentemente proposto nella letteratura relativa all'imputazione multipla.

L'utilizzo delle celle di ponderazione per trattare le mancate risposte è il metodo utilizzato da Oh and Scheuren (1983) e da loro definito *quasi randomizzato*; tale metodo è tra i più utilizzati per il trattamento delle mancate risposte totali, per esempio nelle indagini condotte in Italia dall'ISTAT. Nel caso di indagini con piano di campionamento complesso, il calcolo dei pesi può includere anche la *calibrazione* dei valori rispetto a totali noti, come nel caso dell'indagine ISTAT sulle Condizioni di Vita 2004, presentata nel capitolo 4.

D'altra parte, l'impiego di una grande quantità di variabili ausiliare per formare le celle di ponderazione può causare la formazione di celle di unità campionarie con numerosità troppo scarsa al loro interno; tale occorrenza andrebbe evitata in quanto può comportare valori esageratamente elevati per i pesi campionari (Oh and Scheuren, 1983). Inoltre, questo approccio presenta difficoltà aggiuntive quando le variabili da analizzare sono più di una e sono quindi possibili sia mancate risposte totali che parziali. In questo caso, se si è interessati a stimare il totale nella popolazione di più variabili $\boldsymbol{Y}_j$, si potrebbe pensare di costruire delle celle di ponderazione separatamente per ogni variabile, ma tale procedura può rivelarsi molto pesante. Ecco perchè, solitamente, le mancate risposte parziali vengono trattate attraverso le tecniche di imputazione: una volta imputati i valori mancanti, la stima delle quantità di interesse può essere ottenuta attraverso stimatori come il (1.16), che correggono per le mancate risposte totali.

Con l'imputazione ogni missing value per le variabili y_{ij}, $i = 1, \ldots, n$ e $j = 1, \ldots, J$, viene sostituito con un valore opportunamente scelto. Le tecniche di imputazione sono numerose e possiedono proprietà e caratteristiche molto diverse; mentre alcune si basano su ipotesi semplici ed intuitive, altre sono invece molto più raffinate. La principale caratteristica che accomuna tutte queste tecniche è la produzione di una matrice di dati completa. Questo permette l'utilizzo dei tradizionali metodi di analisi che non prevedono la presenza di dati mancanti, evitando l'eliminazione dall'analisi delle unità con informazioni incomplete, come avviene con la *complete case analysis*.

Anche alcuni metodi di imputazione si basano sulla costruzione di celle, che in questo caso prendono il nome di *celle di imputazione*; per esempio la tecnica dell'*hot-deck* prevede solitamente la sostituzione del valore mancante y_{ij} con il valore osservato $y_{i'j}$ dove i' indica un'unità appartenente alla stessa cella di i. In questo caso la modellazione *implicita* del meccanismo di mancata risposta viene utilizzata nell'ambito delle tecniche di imputazione piuttosto che di ponderazione. I metodi di imputazione *da donatore* come l'hot-deck

possono assumere strutture anche molto complesse; per esempio esistono una versione *sequenziale* e una *gerarchica* (Little and Rubin, 2002; Ford, 1983).

Altri metodi di imputazione utilizzano invece un modello *esplicito* per il meccanismo di risposta: per esempio, i valori imputati possono essere predetti da un modello di regressione in cui le variabili con valori mancanti vengono regredite sulle covariate osservate. Nel caso di una sola variabile continua $\boldsymbol{Y}_j = \boldsymbol{Y}$ soggetta a mancate risposte, il valore mancante per l'individuo $i - esimo$ può essere imputato attraverso la seguente regressione lineare:

$$\widehat{y}_i = \widehat{\beta}_0 + \sum_{t=1}^{K} \widehat{\beta}_k x_{ik}. \tag{1.18}$$

E' da sottolineare che se le covariate $\boldsymbol{X}$ sono tutte *dummy*, in pratica imputare tramite la (1.18) corrisponde ad imputare il valore medio di $\boldsymbol{Y}$ all'interno della cella di imputazione definita dalle variabili X; è chiaro dunque che anche ad una tecnica da donatore corrisponde un modello di imputazione implicito. L'ipotesi cui si fa riferimento con la regressione (1.18), dunque, è sempre quella di dati MAR. Se al valore previsto dalla regressione viene aggiunto un termine di errore stocastico, allora il metodo di imputazione corrisponde a compiere delle estrazioni dalla distribuzione predittiva dei valori mancanti. In questo caso si ha:

$$\widehat{y}_i = \widehat{\beta}_0 + \sum_{t=1}^{K} \widehat{\beta}_k x_{ik} + \epsilon_i. \tag{1.19}$$

Il valore ϵ_i può essere estratto da un variabile casuale normale con media 0 e varianza pari alla varianza residua della regressione oppure, nel caso in cui tale ipotesi distributiva non risulti verosimile, il residuo può essere estratto dai valori $\widehat{\epsilon_i}$ osservati per i rispondenti. Little (1988) suggerisce l'utilizzo del metodo (1.19) in quanto ha il vantaggio di preservare la distribuzione dei valori $\boldsymbol{Y}$, portando a stime non distorte di quantità diverse dalla media, come per esempio dei percentili. Inoltre, l'utilizzo della regressione piuttosto che di tecniche di ponderazione ha il pregio di poter utilizzare variabili continue come covariate, inserendo eventualmente solo gli effetti principali e non anche tutte le interazioni, come avviene invece, implicitamente, con l'utilizzo delle celle di imputazione. Per maggiori dettagli relativi alla descrizione ed al confronto tra le diverse tecniche di imputazione e ponderazione si rimanda a Little and Rubin (2002).

Uno dei principali vantaggi delle tecniche di imputazione, come già accennato, consiste nella produzione di una matrice dei dati completi, consentendo così l'utilizzo dei metodi di analisi classici, come per esempio dello stimatore

(1.16). Il principale svantaggio è che, una volta che i dati sono stati imputati, le normali tecniche di analisi applicate al dataset completato trattano i valori imputati come osservati, non tenendo in considerazione l'incertezza legata al procedimento di imputazione.

Questo comporta una sottostima della varianza per gli stimatori calcolati nel dataset completato; tra le possibili soluzioni per tale fenomeno vi è l'imputazione multipla. L'imputazione multipla sostituisce ad ogni dato mancante più di un valore imputato; se realizzata correttamente, ovvero secondo le procedure considerate in dettaglio nel prossimo capitolo, può portare a compiere inferenze corrette per quantità di interesse della popolazione.

1.2.2 L'ipotesi MAR e il *pattern* dei dati mancanti

Come più volte evidenziato nel corso di questo capitolo, due elementi che possono semplificare le inferenze in presenza di valori mancanti nei dati provenienti da indagini campionarie sono l'ipotesi MAR e il *pattern* dei dati.

L'ipotesi MAR (condizione (1.7)) riveste un ruolo particolarmente importante in quando è una delle condizioni necessarie per avere ignorabilità del meccanismo di mancata risposta. Schafer (1997) afferma che in teoria i dati dovrebbero essere ritenuti MAR solo quando sono *missing by design*, ovvero quando la loro presenza è già nota nella fase di implementazione dell'indagine. E' questo il caso, per esempio, degli esperimenti randomizzati sbilanciati e delle indagini in cui alcuni item vengono somministrati solo ad un sottocampione di unità. In quest'ultimo caso i dati mancati saranno MAR se il sottocampione è per esempio un campione casuale semplice, oppure se il disegno di campionamento impiegato per estrarre il sottocampione utilizza delle variabili completamente osservate. Il concetto di dati *missing by design* viene utilizzato nel capitolo 5 relativamente all'imputazione dei dati mancanti per l'indagine Forze Lavoro del Comune di Firenze.

I dati non saranno invece MAR, in generale, quando la loro presenza non era programmata; è questo il caso delle mancate risposte totali e parziali non preventivate nelle indagini campionarie. In queste situazioni l'ipotesi MAR può essere solamente ipotizzata, senza possibilità di testarla direttamente a meno che non si riesca successivamente ad ottenere il valore mancante, per esempio attraverso follow-ups o da fonti esterne. Schafer (1997) sottolinea che i concetti di MAR e di ignorabilità sono *relativi*, cioè definiti rispetto ad un insieme di covariate osservate: in molte situazioni lo *status* dei dati mancanti (MAR o altro) può cambiare se varia la definizione delle covariate osservate. Questo sottolinea l'importanza dell'individuazione e dell'inclusione delle variabili "predittive" della mancata risposta nei procedimenti di imputazione e, in particolare, nei modelli bayesiani di imputazione multipla,

che condizionano esplicitamente alle informazioni osservate. Ulteriori considerazioni sull'ipotesi MAR vengono presentate nel capitolo 3 relativamente a dati mancanti di reddito, per i quali questa ipotesi risulta particolarmente dibattuta.

Un altro aspetto importante che può talvolta semplificare le inferenze in presenza di non risposte è il *pattern* dei dati. Il *pattern* descrive la "struttura" dei dati mancanti e osservati in riferimento alla matrice dei dati completi. Specificatamente, il *pattern* dei dati mancati è definito dai valori M_{ij} (paragrafo 1.1.1) dell'indicatore di presenza/assenza del dato y_{ij} per l'individuo i relativamente alla variabile j. Se le variabili $\boldsymbol{Y}$ sono 5, tutte potenzialmente soggette a valori mancanti, due possibili *pattern* dei dati sono riportati nelle figure 1.1 e 1.2.

Y_1	Y_2	Y_3	Y_4	Y_5
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	1
0	0	0	0	1
0	0	0	1	1
0	0	0	1	1
0	0	1	1	1
0	0	1	1	1
0	1	1	1	1
0	1	1	1	1
0	1	1	1	1

Figura 1.1: Mancata risposta multivariata: *pattern monotono*.

Y_1	Y_2	Y_3	Y_4	Y_5
0	0	0	1	0
0	0	0	1	0
0	0	1	1	1
0	1	1	0	1
0	1	1	0	0
0	1	0	0	0
0	1	0	0	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	1
0	0	0	0	1
0	0	0	0	0

Figura 1.2: Mancata risposta multivariata: *pattern non monotono*.

Nel caso rappresentato in figura 1.1 si parla di *pattern* di dati mancanti *monotono*: le variabili possono essere disposte in un ordine tale che

$\boldsymbol{Y}_{j+1}, \ldots, \boldsymbol{Y}_k$ hanno valori mancanti per le osservazioni con $\boldsymbol{Y}_j$ mancante, $\forall\ j = 1, \ldots, k-1$. Alternativamente si può affermare che, in un *pattern* monotono, per una data osservazione i il valore $y_{i\,j+1}$ è osservato solo se y_{ij} è osservato, così che la variabile $\boldsymbol{Y}_1$ risulta più osservata della $\boldsymbol{Y}_2$, ecc.

La presenza di un *pattern* monotono in un problema multivariato con dati mancanti MAR può semplificare le analisi, in quanto consente la massimizzazione della verosimiglianza ignorabile (1.10) o l'estrazione dei parametri dalla distribuzione a posteriori (1.11) senza complicazioni computazionali eccessive.

Per esempio, assumendo che i dati seguano una distribuzione normale multivariata e che il *pattern* sia monotono, se esiste una parametrizzazione $\phi = \phi(\theta)$ tale che la verosimiglianza risulti fattorizzabile, allora si può scrivere:

$$\begin{aligned} \prod f(y_{i1}, \ldots, y_{iJ}|\phi) &= \prod f(y_{i1}|\phi_1) \prod f(y_{i2}|y_{i1}\phi_2) \ldots \\ &\ldots \prod f(y_{iJ}|y_{i1}, \ldots, y_{iJ-1}, \phi_J). \end{aligned} \tag{1.20}$$

Se $(y_{i1}, \ldots, y_{iJ})$ seguono una distribuzione normale multivariata allora $f(y_{ij}|y_{i1}, \ldots, y_{ij-1}, \phi_j)$ è normale con media lineare nelle variabili $y_{i1}, \ldots, y_{ij-1}$; in questo caso, Little and Rubin (2002) mostrano che le stime di massima verosimiglianza dei parametri ϕ_j possono essere calcolati facilmente applicando lo *sweep operator* alla verosimiglianza (1.10) così fattorizzata:

$$L(\phi|\boldsymbol{Y}_{oss}, \boldsymbol{X}) = \prod L(\phi_j|\boldsymbol{Y}_{oss}, \boldsymbol{X}) \tag{1.21}$$

con $j = 1, \ldots, J$.

In caso di analisi di tipo bayesiano, lo stesso metodo può essere utilizzato per estrarre i parametri ϕ_j dalla loro distribuzione a posteriori. In particolare, se la distribuzione a priori è fattorizzabile, ovvero se $p(\phi) = p_1(\phi_1), \ldots, p_J(\phi_J)$, allora la distribuzione a posteriori risulta anch'essa fattorizzabile in una struttura che Rubin (1987) definisce *monotone distinct*. Inferenze di tipo bayesiano su ϕ potranno allora essere realizzate utilizzando la seguente fattorizzazione della (1.11):

$$P(\phi_j|\boldsymbol{Y}_{oss}, \boldsymbol{X}) \propto \prod L(\phi_j|\boldsymbol{Y}_{oss}, \boldsymbol{X}) p_j(\phi_j) \tag{1.22}$$

con $j = 1, \ldots, J$.

Se il *pattern* dei dati assume invece una forma più generale, come nel caso della figura 1.2, allora le operazioni di massimizzazione della verosimiglianza e di estrazione dei parametri dalla loro distribuzione a posteriori richiedono solitamente tecniche computazionali più elaborate, come l'utilizzo di algoritmi di tipo EM o di tecniche MCMC. Tale tecniche vengono considerate in dettaglio nel capitolo 2 con riferimento alle tecniche di imputazione multipla.

Capitolo 2

L'imputazione multipla

In questo capitolo viene introdotta formalmente l'imputazione multipla. Nel paragrafo 2.1.1 si presentano le regole di combinazione originarie di Rubin, con le relative variazioni successivamente proposte anche da altri autori, mentre nel paragrafo 2.1.2 si descrivono le principali proprietà dell'imputazione multipla, dando anche una giustificazione di questo metodo dal punto di vista dell'inferenza randomizzata. Nel paragrafo 2.2 vengono presentate le metodologie attualmente più utilizzate per realizzare imputazioni multiple nel contesto dei modelli bayesiani. In particolare, vengono descritte la *data augmentation*, l'approccio *sequential regression multivariate imputation* e l'algoritmo *sampling importance resampling*. Nelle conclusioni alla fine del capitolo si accenna ai problemi di queste metodologie e ai metodi alternativi all'imputazione multipla per analizzare datasets imputati proposti in letteratura.

2.1 Perchè imputazione *multipla*?

L'imputazione dei valori mancanti è il metodo più diffuso per trattare le mancate risposte parziali nelle indagini campionarie. Come già discusso nel capitolo 1 (paragrafo 1.2.1), uno dei principali vantaggi dell'imputazione consiste nella produzione di una matrice di dati completa, che può essere analizzata utilizzando metodologie statistiche di tipo standard. Metodi di analisi che non prevedono l'imputazione dei valori mancanti, come per esempio l'utilizzo dell'algoritmo EM per massimizzare la verosimiglianza ignorabile (1.10), possono dare buoni risultati ma sono di difficile utilizzo in situazioni in cui le mancate risposte riguardano più variabili ed esistono altri fattori di complicazione. Inoltre, in molti casi le imputazioni vengono realizzate da coloro che si sono occupati dell'implementazione dell'indagine e della raccolta dei dati,

facendo sì che le imputazioni "racchiudano" informazioni aggiuntive non a disposizione degli utilizzatori finali dei dati. Questi due vantaggi delle tecniche di imputazione sono particolarmente evidenti nel contesto delle indagini di tipo complesso il cui obiettivo è la creazione di datasets di pubblico utilizzo, come per esempio nel caso delle indagini implementate dagli Istituti Nazionali di statistica.

E' proprio in questo contesto che nasce l'imputazione multipla, metodo inizialmente proposto da Rubin (1978) e ripreso in numerosi successivi lavori (Herzog and Rubin, 1983; Rubin, 1987, 1996). Lo scopo principale dell'imputazione multipla è correggere il principale svantaggio delle tecniche di imputazione singola.

Quando si analizza un dataset in cui i valori mancanti sono stati imputati singolarmente, l'analisi del dataset attraverso tecniche standard tratterà i valori imputati allo stesso modo dei valori osservati; questo significa che, anche in situazioni in cui il meccanismo che ha generato le mancate risposte è ignorabile, le inferenze basate sul dataset completato non tengono in considerazione la variabilità aggiuntiva dovuta alla presenza di valori originariamente mancanti.

Per esempio, riprendendo la simbologia utilizzata nel capitolo 1, sia Q la quantità di interesse della popolazione e sia $\widehat{Q}$ il suo stimatore campionario. Allora, solitamente le inferenze per Q vengono derivate dalla distribuzione $(Q-\widehat{Q}) \sim N(0,U)$ dove $U = \widehat{Var}(Q-\widehat{Q})$; per esempio, in presenza della sola variabile Y e nel caso di un campione casuale semplice di n osservazioni da una popolazione di N unità, le inferenze derivate dai dati completi si basano sulla distribuzione $(\overline{y}-\overline{Y}) \sim N(0, s^2(\frac{1}{n}-\frac{1}{N}))$, dove $\overline{y}$ e s^2 sono rispettivamente la media e la varianza campionaria della variabile Y, mentre $\overline{Y}$ rappresenta la media nella popolazione.

Supponiamo che solo n_1 valori campionari siano realmente osservati, a causa di un meccanismo di non risposta di tipo casuale, e che gli $n-n_1$ valori mancanti vengano imputati attraverso un qualsiasi modello di imputazione. Allora, la distribuzione di riferimento sarà sempre $(Q-\widehat{Q}) \sim N(0,U)$, con $\widehat{Q}$ e U calcolati nel dataset completato con le imputazioni. Per esempio, nel caso in cui ogni valore mancante venga imputato con un valore estratto a caso da quelli osservati in modo che $\overline{y}_1 = \overline{y}$ e $s_1^2 = s^2$ così da preservare le caratteristiche distributive dei dati, la varianza di $\overline{y} - \overline{Y} = \overline{y}_1 - \overline{Y}$ dovrebbe essere $s_1^2(\frac{1}{n_1}-\frac{1}{N})$ e non $s^2(\frac{1}{n}-\frac{1}{N}) = s_1^2(\frac{1}{n}-\frac{1}{N})$: quest'ultima quantità, essendo troppo piccola, determina per $Q = \overline{Y}$ la costruzione di intervalli di confidenza troppo stretti ed il calcolo di statistiche test troppo grandi (Rubin, 1987).

Una possibile soluzione consiste nell'imputare non uno ma m valori per ciascun missing, per esempio impiegando indipendentemente lo stesso mo-

dello già utilizzato per l'imputazione singola dei valori. In questo caso $\widehat{Q}_i$ e U_i, $i = 1, \ldots, m$, rappresentano i valori di $\widehat{Q}$ e U calcolati in ciascuno degli m dataset completati; la variabilità nei valori $\widehat{Q}_i$ e U_i riflette l'incertezza legata al procedimento di inferenza in presenza di dati mancanti.

Nei prossimi sottoparagrafi vengono presentate le principali proprietà e caratteristiche teoriche delle procedure di imputazione multipla, sottolineando gli aspetti essenziali.

2.1.1 Il procedimento di inferenza con imputazione multipla dei valori mancanti

La giustificazione teorica dell'imputazione multipla è stata data da Rubin (1987) nell'ambito dell'inferenza bayesiana per quantità di interesse della popolazione (paragrafo 1.1.1). La trattazione teorica di Rubin ipotizza l'ignorabilità del meccanismo di selezione del campione e del meccanismo di mancata risposta, concetti già introdotti nel primo capitolo, e assume inoltre che il modello bayesiano impiegato per realizzare le imputazioni coincida con il modello utilizzato per analizzare i dataset completati, cosa che può non essere vera soprattutto nelle situazioni in cui i due modelli vengono ipotizzati da due soggetti diversi. Tuttavia, Rubin (1987) fornisce indicazioni attraverso cui risulta possibile valutare la validità delle procedure di imputazione multipla anche nell'ottica dell'inferenza randomizzata, e accenna alle situazioni in cui i modelli di imputazione e di analisi non coincidono. Quest'ultima problematica, su cui ci si sofferma più in dettaglio nel prossimo paragrafo, ha rappresentato, almeno fino ad oggi, una delle principali critiche mosse contro l'imputazione multipla.

Nell'ambito dell'inferenza bayesiana per il parametro θ che caratterizza il modello per i dati, se si ipotizza che il meccanismo di campionamento ed il meccanismo di mancata risposta siano ignorabili (vedi paragrafo 1.1.2), e se si dispone di un insieme di m imputazioni multiple, derivate secondo uno specifico modello bayesiano, è possibile ottenere un'unica inferenza relativamente al parametro θ attraverso opportune regole di combinazione. Tale procedimento di inferenza richiede la stima della distribuzione a posteriori ignorando il meccanismo di campionamento e di mancata risposta, data dall'espressione (1.11) che risulta essere $P(\theta|\boldsymbol{Y}_{oss}) \propto L(\theta|\boldsymbol{Y}_{oss})p(\theta)$ senza esplicitare le variabili esplicative $\boldsymbol{X}$.

Allora, in presenza di valori mancanti si ha:

$$\begin{aligned} P(\theta|\boldsymbol{Y}_{oss}) &= \int P(\theta, \boldsymbol{Y}_{mis}|\boldsymbol{Y}_{oss})d\boldsymbol{Y}_{mis} = \\ &= \int P(\theta|\boldsymbol{Y}_{mis}, \boldsymbol{Y}_{oss})(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{oss})d\boldsymbol{Y}_{mis}. \end{aligned} \tag{2.1}$$

L'imputazione multipla approssima questo integrale nel modo seguente:

$$P(\theta|\boldsymbol{Y}_{oss}) = \frac{1}{m}\sum_{i=1}^{m} P(\theta|\boldsymbol{Y}_{mis}^{(i)}, \boldsymbol{Y}_{oss}) \tag{2.2}$$

dove $\boldsymbol{Y}_{mis}^{(i)}$, $i = 1, \ldots, m$ sono estrazioni di $\boldsymbol{Y}_{mis}$ dalla distribuzione predittiva a posteriori dei valori mancanti, $P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{oss})$. Utilizzando tale approssimazione la media e la varianza a posteriori di θ possono essere così ricavate:

$$\begin{aligned} E(\theta|\boldsymbol{Y}_{oss}) &= \int \theta P(\theta|\boldsymbol{Y}_{oss})d\theta = \\ &\approx \int \theta \frac{1}{m}\sum_{i=1}^{m} P(\theta|\boldsymbol{Y}_{mis}^{(i)}, \boldsymbol{Y}_{oss})d\theta = \\ &= \frac{1}{m}\sum_{i=1}^{m} \widehat{\theta}_i \end{aligned} \tag{2.3}$$

dove $\widehat{\theta}_i = E(\theta|\boldsymbol{Y}_{mis}^{(i)}, \boldsymbol{Y}_{oss})$ è la stima di θ nell'*i-esimo* dataset completato; inoltre:

$$\begin{aligned} E(\theta^2|\boldsymbol{Y}_{oss}) &= \int \theta^2 P(\theta|\boldsymbol{Y}_{oss})d\theta = \\ &\approx \int \theta^2 \frac{1}{m}\sum_{i=1}^{m} P(\theta|\boldsymbol{Y}_{mis}^{(i)}, \boldsymbol{Y}_{oss})d\theta = \\ &= \frac{1}{m}\sum_{i=1}^{m} (\widehat{\theta}_i + U_i) \end{aligned}$$

con $U_i = var(\theta|\boldsymbol{Y}_{mis}^{(i)}, \boldsymbol{Y}_{oss})$ è la stima della varianza nell'*i-esimo* dataset completato.

Combinando le due espressioni precedenti si ottiene:

$$\begin{aligned} var(\theta|\boldsymbol{Y}_{oss}) &\approx \frac{1}{m}\sum_{i=1}^{m}(\widehat{\theta}_i - \overline{\theta})^2 + \frac{1}{m}\sum_{i=1}^{m} U_i = \\ &\approx \left(\frac{m+1}{m}\right)\frac{1}{m}\sum_{i=1}^{m}(\widehat{\theta}_i - \overline{\theta})^2 + \frac{1}{m}\sum_{i=1}^{m} U_i = \\ &= \left(\frac{m+1}{m}\right) B + \overline{U} \end{aligned} \tag{2.4}$$

dove il fattore $\left(\frac{m+1}{m}\right)$ migliora l'approssimazione per m piccolo.

Quindi, si ha che la stima della media a posteriori del parametro di interesse che si ottiene con il procedimento di imputazione multipla è pari alla media delle m stime calcolate nei dataset imputati (espressione (2.3)), mentre la stima della varianza a posteriori è data dalla media delle m varianze a posteriori calcolate nei dataset completati più $\left(\frac{m+1}{m}\right)$ volte la varianza delle stime (espressione (2.4)).

Riassumendo ed utilizzando nuovamente la simbologia di Rubin (1987) in cui si fa riferimento ad una qualsiasi quantità di interesse della popolazione $Q = Q(\boldsymbol{Y})$, funzione di θ, la stima con imputazione multipla dei valori mancati è data da:

$$\widehat{Q}_{MI} = \sum_{i=1}^{m} \widehat{Q}_i/m \tag{2.5}$$

mentre la stima della sua varianza totale è data da:

$$\begin{aligned} T_{MI} &= \overline{U} + \left(\frac{m+1}{m}\right) B = \\ &= \sum_{i=1}^{m}\frac{U_i}{m} + (1+m^{-1})\sum_{k=1}^{m}\frac{(\widehat{Q}_k - \widehat{Q}_{MI})'(\widehat{Q}_k - \widehat{Q}_{MI})}{(m-1)}. \end{aligned} \tag{2.6}$$

Quando Q è scalare, la stima per intervallo ed i test di ipotesi basati sui risultati precedenti fanno riferimento alla distribuzione:

$$(Q - \widehat{Q}_{MI})T_{MI}^{-1} \sim t_\nu, \tag{2.7}$$

dove i gradi di libertà ν sono dati da:

$$\nu = (m-1)(1+r^{-1})^2, \tag{2.8}$$

con $r = \dfrac{(1+m^{-1})B}{\overline{U}}$ pari all'incremento relativo di varianza dovuto alla non risposta (Rubin, 1987). Infatti, $\overline{U}$ stima la varianza totale quando l'informazione mancante relativamente a Q è pari a 0, ovvero quando non ci sono dati mancanti e quindi $B = 0$, mentre $\left(\dfrac{m+1}{m}\right) B$ stima l'incremento della varianza dovuto ai dati mancanti (Schafer, 1997; Reiter and Raghunathan, 2007).

Attraverso il rapporto r ed i gradi di libertà ν è possibile calcolare la frazione d'informazione relativa a Q mancante a causa delle non risposte, ovvero la cosiddetta *frazione d'informazione mancante* (*fraction of missing information*):

$$\widehat{\lambda} = \frac{r + 2/(\nu+3)}{r+1}, \tag{2.9}$$

che può essere approssimata da:

$$\widehat{\lambda}_{approx} = \left(\frac{m+1}{m}\right) B/T_{MI}. \tag{2.10}$$

Questa quantità misura il contributo dei dati mancanti all'incertezza delle inferenze relative a Q; in caso di non risposta ignorabile senza variabili esplicative $\widehat{\lambda}$ è uguale al tasso di mancate risposte. Tuttavia, in situazioni più generali solitamente $\widehat{\lambda}$ è minore di questa quantità a causa delle correlazioni esistenti tra le variabili con valori mancanti e le variabili osservate; infatti, l'informazione mancante non dipende solamente dal tasso di mancata risposta ma anche dall'informazione "incorporata" nel modello di imputazione (Schenker et al., 2006). La quantità $\widehat{\lambda}$, che può essere calcolata relativamente a qualunque parametro di interesse, viene utilizzata e commentata nelle analisi svolte nei capitoli 4 e 5.

E' stato dimostrato che la stima T_{MI} (2.6) può risultare distorta (Reiter and Raghunathan, 2007). In ogni caso, secondo Rubin nei dataset più comuni le proprietà di T_{MI} per la costruzione di intervalli di confidenza sono più importanti delle proprietà asintotiche: varie applicazioni hanno infatti mostrato una sostanziale correttezza degli intervalli di confidenza costruiti utilizzando l'imputazione multipla e la quantità T_{MI} per una pluralità di parametri e quantità di interesse.

Quando $\boldsymbol{Q}$ è invece un vettore di k componenti, si può testare l'ipotesi $\boldsymbol{Q} = \boldsymbol{Q}_0$ utilizzando le versioni multivariate delle espressioni precedenti. In particolare, la naturale trasposizione della statistica test precedente al caso multivariato sarebbe la seguente:

$$D = (\hat{\boldsymbol{Q}} - \boldsymbol{Q}_0)^t T_{MI}^{-1} (\hat{\boldsymbol{Q}} - \boldsymbol{Q}_0)/k.$$

Tuttavia, questo test non è affidabile quando $k > m$ e m è piccolo, motivo per cui in questi casi Rubin (1987) propone l'utilizzo della statistica test:

$$\widetilde{D} = \frac{(\boldsymbol{Q}_0 - \hat{\boldsymbol{Q}})^t \overline{U}^{-1} (\boldsymbol{Q}_0 - \hat{\boldsymbol{Q}})}{[(1+r)k]},$$

dove r dipende dalla media degli elementi della diagonale di $B\overline{U}^{-1}$ dato che risulta pari a:

$$r = \frac{(1+m^{-1})\,\mathrm{tr}(B\overline{U}^{-1})}{k}$$

dove tr(.) è la traccia della matrice. Il $p-value$ è allora:

$$Prob = \{F_{k,\omega} > \widetilde{D}\},$$

dove $F_{k,\omega}$ è la distribuzione F con k e ω gradi di libertà con:

$$\begin{aligned} \omega &= 4 + [k(m-1) - 4]\left(1 + \frac{a}{r}\right)^2, \\ a &= 1 - \frac{2}{k[m-1]}; \end{aligned}$$

quando $k(m-1) > 4$, mentre se $k(m-1) \leq 4$, $\omega = \frac{(k+1)\nu}{2}$. Rubin (1987) e Li, Raghunathan and Rubin (1991) forniscono motivazioni per il test statistico e per la sua distribuzione di riferimento, mostrando come questo test ha proprietà ottimali quando m è piccolo.

In alternativa, sono state proposte altre due procedure per compiere inferenza quando Q è multivariato. Li, Meng, Raghunathan and Rubin (1991) hanno proposto l'utilizzo del test di Wald multivariato per dati completi in ciascun dataset, ovvero del test che non considera la varianza *between* le imputazioni, calcolando poi la statistica test finale combinando i p-values degli m test; Meng and Rubin (1992) hanno proposto la medesima combinazione per i likelihood-ratio test per dati completi condotti negli m dataset imputati. Per una rassegna di tali proposte si rimanda anche a Schafer (1997) e Reiter and Raghunathan (2007).

Nei casi in cui la numerosità campionaria risulti piuttosto ridotta, Reiter and Raghunathan (2007) riportano varie proposte, alcune delle quali recentissime, attraverso cui è possibile correggere i gradi di libertà delle statistiche test sia nel caso in cui Q è univariato che multivariato. In tale lavoro, inoltre, sono riportate le regole di combinazione per compiere inferenza quando l'imputazione multipla viene utilizzata in contesti statistici diversi da quello

in cui è nata. Come già precedentemente accennato, infatti, l'imputazione multipla nasce per il trattamento delle mancate risposte in indagini con numerosità campionaria elevata i cui risultati sono destinati ad un vasto pubblico; negli ultimi anni, tuttavia, sta riscuotendo un notevole successo anche per scopi diversi, come per esempio per il trattamento delle mancate risposte in piccoli campioni, per la protezione di dati sensibili e per la correzione del *measurement error* (Reiter and Raghunathan, 2007; Gelman and Meng, 2004). In questi contesti è necessario modificare le regole di combinazione originariamente proposte da Rubin (1987) in quanto le quantità note cui ci si condiziona possono essere diverse.

2.1.2 Le proprietà dell'imputazione multipla

La giustificazione delle regole di combinazione presentate nel paragrafo precedente viene svolta da Rubin (1987) sotto l'ipotesi che le m imputazioni multiple siano estrazioni dalla distribuzione predittiva a posteriori dei valori mancanti, considerando prima m infinito e poi finito. Tali giustificazioni utilizzano l'ipotesi di ignorabilità del meccanismo di selezione del campione e del meccanismo di mancata risposta, condizioni che permettono di lavorare con la distribuzione a posteriori della statistica di interesse relativamente ai dati completi, senza tenere in considerazione la distinzione tra valori imputati ed osservati[1].

Nella sua complessa ed esaustiva trattazione sull'imputazione multipla, Rubin (1987) va oltre la giustificazione teorica delle regole di combinazione (2.5) e (2.6), implementando anche la valutazione delle inferenze, precedentemente introdotte in ottica bayesiana, dal punto di vista dell'inferenza randomizzata. Come specifica meglio anche in successivi lavori, infatti, Rubin è fermamente convinto che il concetto di "validità statistica" sia, nell'ambito delle databases condivisi ed analizzati da molti utilizzatori, un concetto di tipo frequentista, in cui la distribuzione randomizzata è introdotta dal noto meccanismo di selezione del campione, dato un certo meccanismo di non risposta (Rubin, 1996). Tali valutazioni servono anche per definire alcune caratteristiche desiderabili che le procedure di imputazione multipla dovrebbero possedere.

[1] Più precisamente, Rubin (1987) distingue tra la distribuzione a posteriori per la quantità di interesse Q basata sui dati *completi* e quella basata sui dati *completati*: le condizioni che consentono l'utilizzo della prima, che non richiede il condizionamento all'indicatore di mancata risposta R (paragrafo 1.1.2), sono l'ignorabilità dei due meccanismi che generano le mancate risposte. Alternativamente, quando il solo meccanismo di selezione del campione è ignorabile, è comunque possibile che la *completed-data* e la *complete-data posterior distributions* coincidano, per esempio quando si ha una numerosità campionaria elevata.

Il principale risultato è il seguente (Rubin, 1987):

1. se l'inferenza condotta sui dataset completi è un'inferenza valida in ottica randomizzata in assenza di mancate risposte,

2. se il metodo di imputazione è proprio,

allora in campioni numerosi le inferenze cui si giunge tramite le procedure di imputazione multipla sono valide nell'ottica dell'inferenza randomizzata, almeno quando il numero di imputazioni è elevato. Quando invece m è piccolo, le procedure sono "quasi" valide nell'ottica dell'inferenza randomizzata, e la loro performance può essere valutata attraverso opportuni calcoli.

I punti sopra esposti sono definiti in modo esaustivo in Rubin (1987); una "traduzione" di tali definizioni in termini meno complicati, riportata di seguito, è in Rubin (1996) e Schafer (1997).

Innanzitutto, Rubin definisce le condizioni in cui si ha validità delle inferenze nell'ambito della randomizzazione, in assenza di mancate risposte (punto 1). In questo caso, considerando ancora Q la quantità di interesse della popolazione, $\widehat{Q}$ il suo stimatore campionario e $U = \widehat{Var}(Q - \widehat{Q})$, siano $\boldsymbol{X}$ e $\boldsymbol{Y}$ le informazioni provenienti dalla popolazione, che in questo caso devono essere considerate quantità fisse, mentre è l'indicatore di inclusione nel campione I ad introdurre la distribuzione di probabilità di riferimento. Le due condizioni per avere validità delle stime in ottica randomizzata sono allora:

$$E(\widehat{Q}|\boldsymbol{X}, \boldsymbol{Y}) = Q \tag{2.11}$$

$$E(U|\boldsymbol{X}, \boldsymbol{Y}) = var(\widehat{Q}|\boldsymbol{X}, \boldsymbol{Y}). \tag{2.12}$$

Per avere invece imputazioni multiple proprie (punto 2) le statistiche calcolate negli m datasets, $\widehat{Q}_i$ e U_i, devono essere approssimativamente non distorte per le loro analoghe nel caso di dati completi; considerando le medie di tali statistiche per m grande e considerando che in questo caso la distribuzione di probabilità dipende dall'indicatore di mancata risposta R, mentre X, Y e I sono quantità fisse, questo significa:

$$E(\widehat{Q}_\infty|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{I}) = \widehat{Q} \tag{2.13}$$

$$E(\overline{U}_\infty|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{I}) = U. \tag{2.14}$$

Inoltre, per avere una procedura di imputazione multipla propria le varianze e covarianze delle stime $\widehat{Q}_i$ devono essere approssimativamente non distorte per la varianza di $\widehat{Q}_\infty$ nell'ottica dell'inferenza randomizzata rispetto alle mancate risposte:

$$E(B_\infty|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{I}) = var(\widehat{Q}_\infty|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{I}). \tag{2.15}$$

E' interessante notare come la condizione (2.13) rappresenti l'analogo della (2.11) per la validità randomizzata: entrambe le condizioni richiedono infatti la non distorsione dello stimatore (rispettivamente $\widehat{Q}_\infty$ e $\widehat{Q}$) per la loro quantità di interesse ($\widehat{Q}$ e Q), rispetto alla distribuzione indotta rispettivamente dal meccanismo di non risposta e dal meccanismo di selezione del campione. Allo stesso modo, l'espressione (2.15) rappresenta l'analogo della (2.14).

Allora, la condizione di validità dell'imputazione multipla dal punto di vista della randomizzazione dell'inferenza quando m è grande (ovvero la conclusione che deriva dai punti 1 e 2), può essere derivata combinando le condizioni precedenti:

$$E(\widehat{Q}_\infty|\boldsymbol{X},\boldsymbol{Y}) = E[(E(\widehat{Q}_\infty|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{I}))|X,Y] = E(\widehat{Q}|\boldsymbol{X},\boldsymbol{Y}) = Q$$

e

$$\begin{aligned}
E(T_\infty|\boldsymbol{X},\boldsymbol{Y}) &= E(\overline{U}_\infty|\boldsymbol{X},\boldsymbol{Y}) + E(B_\infty|\boldsymbol{X},\boldsymbol{Y}) = \\
&= E[E(\overline{U}_\infty|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{I})|X,Y] + E[E(B_\infty|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{I})|\boldsymbol{X},\boldsymbol{Y}] = \\
&= E(U|\boldsymbol{X},\boldsymbol{Y}) + E[var(\widehat{Q}_\infty|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{I})|\boldsymbol{X},\boldsymbol{Y}] = \\
&= var(\widehat{Q}|\boldsymbol{X},\boldsymbol{Y}) + E[var(\widehat{Q}_\infty|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{I})|\boldsymbol{X},\boldsymbol{Y}] = \\
&= var[E(\widehat{Q}_\infty|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{I})|\boldsymbol{X},\boldsymbol{Y}] + E[var(\widehat{Q}_\infty|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{I})|\boldsymbol{X},\boldsymbol{Y}] = \\
&= var(\widehat{Q}_\infty|\boldsymbol{X},\boldsymbol{Y}).
\end{aligned}$$

Allora, questo implica che vale approssimativamente la condizione:

$$(Q - \widehat{Q}_\infty) \sim N(0, T_\infty) \tag{2.16}$$

dove $T_\infty = \overline{U}_\infty + B_\infty$, condizione che corrisponde ad affermare che le inferenze che si ottengono dalla procedura di imputazione multipla risultano, per m grande, approssimativamente valide per l'inferenza randomizzata.

Anche realizzando un numero finito di imputazioni, i risultati cui si giunge possono essere molto soddisfacenti, soprattutto quando la *fraction of missing information* non è elevata. Infatti, l'efficienza relativa di un stima puntuale basata su m imputazioni rispetto ad una basata su un numero infinito di imputazioni può essere approssimativamente misurata attraverso la quantità $(1 + \lambda/m)^{-1}$, con λ pari alla *fraction of missing information*. Per esempio, dalla tabella seguente, ripresa da Schafer and Olsen (1998), si evince che per $\lambda = 0.3$ (30% di informazione mancante), con $m = 5$ imputazioni multiple proprie si raggiunge già il 94% di efficienza.

Dal punto di vista pratico richiedere un'imputazione multipla propria significa richiedere che essa incorpori un'appropriata variabilità tra le ripetizioni all'interno di un modello; come sottolineato in Rubin (1996), infatti, la

Tabella 2.1: Efficienza delle stime ottenute con l'imputazione multipla, per numero di imputazioni m e *fraction of missing information* λ (valori percentuali).

	λ				
m	0.1	0.3	0.5	0.7	0.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96
∞	100	100	100	100	100

condizione più importante per avere imputazione multipla propria è la (2.15). Questo è solitamente soddisfatto quando le imputazioni derivano da un modello bayesiano esplicito, ovvero sono estrazioni dalla distribuzione predittiva a posteriori dei valori mancanti, in quanto per ogni imputazione viene anche estratto, dalla sua distribuzione a posteriori, un valore del parametro che caratterizza il modello.

Un esempio di modello di imputazione esplicito è il seguente. Data una variabile univariata $\boldsymbol{Y}$, soggetta a non risposta ignorabile, e un insieme di variabili esplicative $\boldsymbol{X}$ completamente osservate, un semplice modello esplicito bayesiano per $\boldsymbol{X}$ e $\boldsymbol{Y}$ è il modello di regressione lineare normale. In particolare, si ha:

$$Y_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2),$$

con $\theta = (\boldsymbol{\beta}, log(\sigma))$, $\boldsymbol{\beta}$ vettore $(q\times 1)$ e σ scalare. Per il parametro θ si ipotizza una distribuzione a priori impropria, $p(\theta) \propto$ *costante*. Queste specificazioni fanno parte del cosiddetto "passo di modellazione", in cui si sceglie uno specifico modello per i dati, che rappresenta il primo passo da eseguire quando si imputa utilizzando un modello esplicito. Indicando il numero di rispondenti con n_{oss}, $n_{oss} > q$, il numero di non rispondenti risulta pari a $n_{mis} = n - n_{oss}$; $\boldsymbol{X} = (\boldsymbol{X}_{oss}, \boldsymbol{X}_{mis})$ e $\boldsymbol{X} = (\boldsymbol{Y}_{oss}, \boldsymbol{Y}_{mis})$ rappresentano le corrispondenti partizioni per le variabili $\boldsymbol{X}$ e $\boldsymbol{Y}$.

I successivi passi di stima e di imputazione, che derivano da calcoli bayesiani standard per il modello lineare normale (Gelman et al., 2004), sono allora i seguenti:

1. calcolo di $V = (\boldsymbol{X}'_{oss}\boldsymbol{X}_{oss})^{-1}$, $\widehat{\boldsymbol{\beta}} = \boldsymbol{V}\boldsymbol{X}'_{oss}\boldsymbol{Y}_{oss}$ e di $\widehat{\boldsymbol{Y}}_{oss} = \boldsymbol{X}_{oss}\widehat{\beta}$;

2. estrazione di una variabile casuale g da una distribuzione χ^2 con $n_{oss}-q$ gradi di libertà;

3. calcolo di $\sigma_*^2 = (\boldsymbol{Y}_{oss} - \widehat{\boldsymbol{Y}}_{oss})'(\boldsymbol{Y}_{oss} - \widehat{\boldsymbol{Y}}_{oss})/g$;

4. estrazione di una vettore $\boldsymbol{Z}$ di q elementi dalla distribuzione Normale $N(0, I_q)$;

5. calcolo di $\widehat{\boldsymbol{\beta}}_* = \widehat{\boldsymbol{\beta}} + \sigma_* \boldsymbol{V}^{1/2} \boldsymbol{Z}$, con $\boldsymbol{V}^{1/2}$ matrice triangolare[2], radice quadrata della $\boldsymbol{V}$;

6. calcolo dei valori previsti $\boldsymbol{Y}_{mis} = \boldsymbol{X}_{mis}\widehat{\boldsymbol{\beta}}_* + \upsilon_i \sigma_*$ per $i = \ldots, n_{mis}$, con i valori υ_i estratti indipendentemente da una variabile casuale normale $N(0, I_q)$.

Per ottenere m imputazioni multiple i passi 2-6 devono essere ripetuti m volte. In particolare, i passi 2-5 consentono di estrarre valori dalla distribuzione a posteriori di $\boldsymbol{\beta}$, mentre al passo 6 vengono realizzate le imputazioni.

Imputazioni multiple che fanno riferimento ad un metodo proprio posso essere realizzate, in generale, anche utilizzando un modello bayesiano approssimato che incorpori in modo appropriato la variabilità between. Inoltre, è talvolta possibile rendere proprio o approssimativamente tale anche un metodo di imputazione multipla che consista in ripetizioni di un modello di imputazione implicito, come per esempio l'hot-deck. In questo caso Rubin (1987) suggerisce l'utilizzo dell'*approximate Bayesian Bootstrap*; tale metodo fa precedere l'estrazione dei donatori all'interno delle celle di imputazione da un'estrazione con reimmissione dei valori osservati, incorporando in questo modo la variabilità mancante.

Un'altra utile indicazione per ottenere imputazioni proprie in senso generale riguarda la raccomandazione di introdurre come variabili esplicative nel modello di imputazione tutte le caratteristiche relative al disegno di campionamento, come per esempio gli indicatori di stratificazione, clusterizzazione e i pesi campionari (Rubin, 1996). Infatti, il pericolo per chi compie le imputazioni è piuttosto la non inclusione di predittori importanti che l'inclusione di troppe variabili esplicative: la possibile perdita di precisione che si può avere includendo predittori non importanti è un prezzo solitamente molto piccolo da pagare per avere una generale validità delle analisi svolte sui datasets completati (Rubin, 1996; Collins et al., 2001).

In ogni caso, sapere se un metodo di imputazione sia tecnicamente proprio o meno assume spesso una rilevanza relativa rispetto a sapere che esso ha un comportamento ottimale (Schafer, 1997). Inoltre, Rubin è fermamente

[2]La matrice $\boldsymbol{V}^{1/2}$ può essere ricavata, per esempio, attraverso la decomposizione di Cholesky.

convinto che anche in situazioni in cui è noto che l'imputazione multipla è meno efficiente rispetto ad altre procedure "ad hoc" per il trattamento delle mancate risposte, spesso gli sforzi aggiuntivi necessari per realizzare tali procedure possono non essere giustificati: l'imputazione multipla può non essere la soluzione ideale per tutti i problemi di dati mancanti, ma è la soluzione più generale e "user friendly" che sia stata proposta per risolverli (Rubin, 2003). Ciò è quasi universalmente accettato nel contesto dei dati provenienti da grandi indagini (Nielsen, 2003), che come sottolineato più volte è il contesto in cui nasce l'imputazione multipla e nel quale vengono realizzate le due applicazioni presentate in questa tesi (capitoli 4 e 5).

Esistono poi altre considerazioni che possono servire da linee guida per determinare se un dato procedimento di imputazione multipla possa essere considerato ottimale o meno. Per esempio, tutte le considerazioni precedenti relative alla validità delle procedure di imputazione multipla erano basate sull'assunzione di una sostanziale uguaglianza tra il modello ipotizzato dal soggetto che realizza le imputazioni e quello ipotizzato dall'utilizzatore dei dati completati con le imputazioni. Tuttavia, specialmente in situazioni in cui questi due soggetti sono distinti e non sono in contatto l'uno con l'altro, i due modelli ipotizzati possono in realtà essere diversi. La compatibilità tra il modello di imputazione e il modello di analisi, definita *congeniality* da Meng (1994), ha rappresentato uno degli aspetti più discussi relativamente all'imputazione multipla (Fay, 1992; Rubin, 1996; Meng, 2002). Anche in questo caso, così come per la definizione di metodo di imputazione proprio, l'analisi formale del problema dell'*uncongeniality* può risultare molto complicata: per esempio, quando le imputazioni vengono realizzate utilizzando un modello bayesiano mentre le analisi condotte sui datasets completati sono *design-based*, è necessario formulare quest'ultima analisi in termini di modello bayesiano, compito che può rilevarsi impossibile per alcune tipologie di analisi (Meng, 2002). Da un punto di vista pratico, la raccomandazione principale consiste nell'utilizzare un metodo di imputazione più generale possibile, che possa risultare compatibile con una pluralità di analisi che gli utilizzatori finali saranno presumibilmente interessati a svolgere (Rubin, 1996; Schafer, 2003). In questo modo colui che realizza le imputazioni non corre il rischio di imporre delle restrizioni su parametri che saranno poi oggetto dell'interesse dell'analizzatore, fatto questo che potrebbe portare a compiere inferenze erronee. Quando invece il modello di imputazione utilizza delle informazioni aggiuntive e tali informazioni non sono a disposizione dell'analizzatore dei dati, che utilizza un modello più generale, le inferenze che si ottengono sono più efficienti di quanto atteso; in questo caso Rubin (1996) parla di *superefficency*.

Si riassumono adesso le principali linee guida, già incontrate in questo e

nel precedente capitolo, attraverso cui è possibile implementare una buona procedura di imputazione multipla utilizzando un modello bayesiano, anche approssimato, sotto l'ipotesi che i dati siano mancanti a caso (MAR):

- individuare le variabili che possono spiegare la presenza delle mancate risposte, così da poterle introdurre come variabili esplicative nel modello di imputazione: come già sottolineato (paragrafo 1.2.2), questo può aiutare a rendere più plausibile l'ipotesi MAR anche in contesti in cui potrebbe sembrare particolarmente critica;

- introdurre nel modello di imputazione anche le variabili relative al disegno di campionamento utilizzato per raccogliere i dati: tale passo può aiutare a rendere propria la procedura di imputazione multipla in situazioni in cui la verifica diretta di tale proprietà non risulta possibile;

- considerare, sempre in termini di variabili esplicative da introdurre nel modello di imputazione, quali potranno essere le stime ed analisi di interesse degli analizzatori dei datasets imputati; il rilascio dei datasets imputati dovrebbe inoltre essere sempre accompagnato da una descrizione della procedura di imputazione, così che l'utilizzatore finale possa sapere se il suo modello di analisi presenta differenze importanti rispetto a quello di imputazione;

- generando le imputazioni multiple secondo più modelli per la mancata risposta risulta possibile studiare la sensitività delle inferenze rispetto al modello scelto: per far questo basta ripetere il procedimento di imputazione più volte, andando a verificare poi i cambiamenti in termini delle inferenze finali ottenute.

Prima di passare all'illustrazione delle metodologie attraverso cui è possibile realizzare imputazioni multiple secondo modelli bayesiani è importante sottolineare un'ultima "proprietà" dell'imputazione multipla. Non bisogna dimenticare, infatti, che tutto quanto è stato finora considerato, ipotesi, proprietà e criticità dell'imputazione multipla, riguarda solo la parte mancante dei dati, e non quella osservata: per questo motivo le stime derivate attraverso l'imputazione multipla possiedono una sostanziale robustezza rispetto ad eventuali mis-specificazioni del modello di imputazione, specialmente quando l'informazione mancante è limitata (Madow et al., 1983; Rubin, 1996; Schafer, 1997).

2.2 Metodi bayesiani per realizzare imputazioni multiple

In questa sezione si considerano alcuni metodi per produrre imputazioni che fanno riferimento ai modelli bayesiani. Si è scelto di concentrarci su questi metodi in quanto sono quelli attualmente più utilizzati sia per la loro vasta applicabilità (dati mancanti multivariati), sia perchè vengono implementati da appositi software che stanno conoscendo una vasta diffusione.

Tutti i metodi presentati si basano sull'ipotesi di dati mancanti a caso (MAR); le espressioni di riferimento[3] sono dunque quelle del paragrafo 1.1.2. Inoltre, se non diversamente specificato, il *pattern* dei dati mancanti è ipotizzato di tipo generale.

Anche se i metodi che si presentano nascono nello specifico contesto dell'imputazione multipla, il loro utilizzo può essere finalizzato alla realizzazione di imputazioni singole; questo è quello che avviene, per esempio, per l'imputazione dei dati mancanti di reddito dell'indagine EU-SILC da parte di ISTAT, come viene specificato meglio nei capitoli 3 e 4.

A conferma che i metodi qui illustrati rappresentano lo *state of the art* in termini di imputazione, nell'ultimo paragrafo del capitolo vengono illustrati i più recenti studi, alcuni dei quali ancora in fase di implementazione, che stanno estendendone l'applicabilità e le giustificazioni teoriche.

2.2.1 La *data augmentation*

La *data augmentation* (Tanner and Wong, 1987) è un metodo iterativo per la simulazione di distribuzioni a posteriori. L'idea base della data augmentation è quella di risolvere un problema complesso in presenza di dati mancanti risolvendo iterativamente problemi trattabili con dati completi (Schafer, 1997). Questo significa che i dati osservati vengono *aumentati*, ovvero i dati mancanti vengono sostituiti con appropriati valori, in modo da rendere più facile, per esempio, il calcolo della distribuzione a posteriori di un parametro di interesse. Ecco perchè la data augmentation nasce naturalmente nel contesto dei dati mancanti e, in particolare, in quello delle mancate risposte.

In pratica il punto di partenza della data augmentation nel contesto delle mancate risposte è l'espressione (2.1): la quantità di interesse è la distribuzione a posteriori del parametro che caratterizza il modello per i dati, $p(\theta|\boldsymbol{Y}_{oss})$, che può risultare difficile da derivare o simulare; tuttavia, disponendo dei valori $\boldsymbol{Y}_{mis}$, la distribuzione $p(\theta|\boldsymbol{Y}_{oss}, \boldsymbol{Y}_{mis})$ risulterà in genere molto più facile

[3] Rispetto alle formulazioni del paragrafo 1.1.2 in questo capitolo le covariate osservate $\boldsymbol{X}$ non vengono esplicitate, ovvero sono da considerarsi incluse in $\boldsymbol{Y}_{oss}$.

da trattare. Se per esempio si dispone di m valori (imputazioni multiple) $\boldsymbol{Y}_{mis}$, allora la distribuzione $p(\theta|\boldsymbol{Y}_{oss})$ può essere approssimata attraverso l'espressione (2.2).

Estrarre i valori $\boldsymbol{Y}_{mis}$ dalla distribuzione condizionata $P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{oss})$ può non essere facile (Kong et al., 1994). Allora, la data augmentation prevede il seguente procedimento iterativo: dato un valore provvisorio per il parametro θ, $\theta^{(t)}$, si estre un valore $\boldsymbol{Y}_{mis}$ dalla distribuzione predittiva a posteriori:

$$\boldsymbol{Y}_{mis}^{(t+1)} \sim P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{oss}, \theta^{(t)}) \tag{2.17}$$

e poi, condizionando rispetto a $\boldsymbol{Y}_{mis}^{(t+1)}$, si estrae un nuovo valore per θ dalla sua distribuzione a posteriori con dati completi:

$$\theta^{(t+1)} \sim P(\theta|\boldsymbol{Y}_{oss}, \boldsymbol{Y}_{mis}^{(t+1)}). \tag{2.18}$$

La ripetizione dei passi (2.17) e (2.18) partendo da un valore iniziale $\theta^{(0)}$ genera una sequenza stocastica $\left\{(\theta^{(t)}, \boldsymbol{Y}_{mis}^{(t)}) : t = 1, 2, \ldots\right\}$ la cui distribuzione stazionaria è $P(\theta, \boldsymbol{Y}_{mis}|\boldsymbol{Y}_{oss})$, mentre le sottosequenze $\left\{\theta^{(t)} : t = 1, 2, \ldots\right\}$ e $\left\{\boldsymbol{Y}_{mis}^{(t)} : t = 1, 2, \ldots\right\}$ hanno rispettivamente come distribuzione stazionaria $P(\theta|\boldsymbol{Y}_{oss})$ e $P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{oss})$. Quindi, per valori elevati di t, $\theta^{(t)}$ e $\boldsymbol{Y}_{mis}^{(t)}$ possono essere considerate come estrazioni approssimate rispettivamente da $P(\theta|\boldsymbol{Y}_{oss})$ e $P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{oss})$ (Schafer, 1997). La data augmentation può essere ripetuta indipendentemente m volte per ottenere m estrazioni da queste due distribuzioni: i valori $\boldsymbol{Y}_{mis}$ che si ottengono sono imputazioni multiple dei valori mancanti estratte dalla loro distribuzione predittiva a posteriori (Little and Rubin, 2002).

Un esempio di utilizzo della data augmentation è in Schafer (1997) (capitolo 3). Data una variabile casuale normale $Y \sim N(\mu, \sigma)$ di cui si osservano $n_{oss} = n - n_{mis}$ valori, ipotizzando la distribuzione a priori $p(\mu, \sigma) \propto \sigma^{-1}$, le distribuzioni a posteriori sono le seguenti:

$$\begin{aligned} \mu|\sigma, \boldsymbol{Y}_{oss} &\sim N(\overline{y}_{oss}, \sigma/n_{oss}) \\ \sigma|\boldsymbol{Y}_{oss} &\sim (n_{oss} - 1)S_{oss}^2\chi_{n_{oss}-1}^{-2} \end{aligned}$$

con S_{oss}^2 varianza campionaria dei valori osservati $y_1, \ldots, y_{n_{oss}}$. La data augmentation prevede allora i seguenti passi:

$$y_i^{(t+1)}|\mu^{(t)}, \sigma^{(t)}, \boldsymbol{Y}_{oss} \sim N(\mu^{(t)}, \sigma^{(t)}),$$

indipendentemente per $i = 1, \ldots, n_{mis}$, e:

$$\begin{aligned} \mu^{(t+1)}|\sigma^{(t)}, \boldsymbol{Y} = (\boldsymbol{Y}_{oss}, \boldsymbol{Y}_{mis}^{(t+1)}) &\sim N(\overline{y}, \sigma^{(t)}/n) \\ \sigma^{(t+1)}|\boldsymbol{Y} = (\boldsymbol{Y}_{oss}, \boldsymbol{Y}_{mis}^{(t+1)}) &\sim (n-1)S^2\chi_{n-1}^{-2}. \end{aligned}$$

Tanner and Wong (1987) definiscono *I-step* (Imputation step) il passo (2.17), *P-step* (Posterior step) il passo (2.18). Questo mette in evidenza lo stretto legame esistente tra la data augmentation e l'algoritmo EM (Expectation-Maximization). Questo algoritmo può essere utilizzato per massimizzare la verosimiglianza dei dati osservati $L(\theta|\boldsymbol{Y}_{oss})$ in presenza di valori mancanti. In particolare, poichè la verosimiglianza dei dati completi $L(\theta|\boldsymbol{Y})$ è più facile da massimizzare (tale massimizzazione corrisponde a calcolare la stima ML in assenza di dati mancanti), nell'*E-step* l'algoritmo EM "riempie" i dati mancanti $\boldsymbol{Y}_{mis}$, così come avviene in modo stocastico nell'*I-step* della data augmentation. Questo corrisponde a calcolare la log-verosimiglianza attesa dei dati completi, $l(\theta|\boldsymbol{Y})$, mediando rispetto alla distribuzione predittiva dei valori mancanti $P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{oss}, \theta)$ per $\theta = \theta^{(t)}$:

$$Q(\theta|\theta^{(t)}) = \int l(\theta|\boldsymbol{Y}) P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{oss}, \theta^{(t)}) d\boldsymbol{Y}_{mis}. \tag{2.19}$$

Nell'*M-step*, poi, si determina il valore $\theta^{(t+1)}$ che massimizza tale log-verosimiglianza:

$$Q(\theta^{(t+1)}|\theta^{(t)}) = Q(\theta|\theta^{(t)}),\ \forall\theta. \tag{2.20}$$

Ripetendo iterativamente i passi (2.19) e (2.20) partendo da un valore mancante $\theta^{(0)}$ si ottiene la sequenza $\left\{\theta^{(t)} : t = 1, 2, \ldots\right\}$, che converge ad un punto stazionario della log-verosimiglianza dei dati osservati: in situazioni ottimali tale punto è un massimo globale e l'EM fornisce la stima di massima verosimiglianza (MLE) di θ, ovvero il massimo della $l(\theta|\boldsymbol{Y}_{oss})$ (Schafer, 1997). A differenza di quanto avviene con la data augmentation, il calcolo dello standard error della stima ML ottenuta attraverso l'algoritmo EM richiede ulteriori calcoli, implementabili attraverso vari metodi (Tanner, 1996); questo deriva dal fatto che nell'*E-step* vengono calcolati dei valori attesi condizionati, a differenza di quanto avviene nell'*I-step* della data augmentation, dove vengono realizzate delle estrazioni dalla distribuzione predittiva a posteriori dei valori mancanti (Little and Rubin, 2002).

Inoltre, è interessante notare che la data augmentation è, in pratica, un caso particolare di Gibbs sampler, metodo iterativo di simulazione che appartiene alle tecniche Markov Chain Monte Carlo (Casella and George, 1992). Nel caso di due variabili causali (X, Y), attraverso il Gibbs sampler risulta possibile estrarre un campione dalla distribuzione $f(x)$ utilizzando in realtà le distribuzioni condizionate $f(x|y)$ e $f(y|x)$, che in molte situazioni possono rilevarsi trattabili quando la $f(x)$ non lo è. Per far questo il Gibbs sampler genera una sequenza di variabili casuali $\{Y'_0, X'_0, Y'_1, X'_1, \ldots, Y'_k, X'_k\}$. Partendo da un valore iniziale $Y'_0 = y'_0$, tale sequenza viene ottenuta iterativamente

generando valori dalle distribuzioni:

$$\begin{aligned} X'_j &\sim f(x|Y'_j = y'_j) \\ Y'_{j+1} &\sim f(Y|X'_j = x'_j). \end{aligned}$$

Sotto condizioni generali di regolarità, la distribuzione delle variabili X'_j converge a $f(x)$ per $k \to \infty$: ovvero, per k abbastanza grande il valore finale della sequenza di variabili casuali, $X'_j = x'_j$, è un'estrazione dalla distribuzione marginale di X, $f(x)$ (Casella and George, 1992).

In generale è possibile implementare versioni multivariate del Gibbs sampler, per estrarre valori da una distribuzione $(X_1, \ldots, X_p)$ difficile da trattare; per $p = 2$, ponendo $X_1 = \boldsymbol{Y}_{mis}$ e $X_2 = \theta$ e condizionando per $\boldsymbol{Y}_{oss}$, il Gibbs sampler corrisponde essenzialmente alla data augmentation. Avendo maggiore flessibilità, il Gibbs sampler è utilizzabile in problemi di dati mancanti più generali della data augmentation (Little and Rubin, 2002).

Quando si utilizzano metodi iterativi come la data augmentation, particolare attenzione andrebbe dedicata allo studio dell'effettiva convergenza dell'algoritmo. In particolare, in presenza di dati mancanti la velocità di convergenza dell'algoritmo è legata in modo inverso alla *fraction of missing information* (Schafer, 1997).

Ci sono poi altre situazioni in cui la procedura di imputazione può risultare facilitata. Come già illustrato nel capitolo 1 (paragrafo 1.2.2), quando i dati mancanti seguono un pattern monotono e la distribuzione a priori del parametro che caratterizza il modello è fattorizzabile, le inferenze di tipo bayesiano possono basarsi sulla fattorizzazione (1.22), rendendo non necessario l'utilizzo di procedure di tipo iterativo come la data augmentation. Infatti, in questo caso il procedimento di imputazione risulta "scomponibile" in più procedimenti di imputazione univariati, ciascuno dei quali condiziona solo rispetto ad informazioni osservate (Rubin, 1987). Se il pattern dei dati non è esattamente monotono ma può diventarlo "riempiendo" una piccola porzione dei dati mancanti, allora è possibile ottenere le imputazioni multiple implementando la *monotone data augmentation* (Schafer, 1997): in ciascuno degli I-step vengono imputati solamente i valori necessari a rendere il pattern monotono, facendo sì che la convergenza venga raggiunta molto più velocemente rispetto alla data augmentation classica.

Per realizzare l'imputazioni multipla di dati mancanti, la data augmentation si basa solitamente sull'ipotesi che i dati seguano una distribuzione normale multivariata: ovviamente nella pratica esistono numerose situazioni in cui le variabili di interesse sono categoriche. Anche se è stato dimostrato attraverso varie applicazioni e simulazioni (Schafer and Olsen, 1998; Bernaards et al., 2007) che l'imputazione di variabili categoriche attraverso modelli continui, seguita dall'arrotondamento dei valori imputati alla categoria

più vicina, può dare buoni risultati, in generale tale procedura non risulta ottimale. Ecco perchè, per evitare le complicazioni legate all'implementazione della data augmentation per variabili categoriche e di tipo misto (Schafer, 1997), negli ultimi anni si è assistito allo sviluppo ed alcla rapida diffusione di algoritmi in grado di realizzare l'imputazione multipla multivariata per insiemi di variabili con caratteristiche distributive anche molto diverse tra loro.

2.2.2 L'approccio *Sequential Regression Multivariate Imputation*

Diversi autori sono concordi nell'attribuire la prima implementazione di un metodo di imputazione *variable by variable* a Kennickell (1991). A partire da quel lavoro sono state proposte varie implementazioni di questo approccio all'imputazione dei dati mancati che, pur differenziandosi per alcuni dettagli, ne condividono gli aspetti essenziali[4].

In particolare, il contesto in cui nascono gli approcci *variable by variable* è quello dei datasets provenienti da indagini di tipo complesso, che raccolgono numerose informazioni modellabili attraverso variabili con caratteristiche distributive anche molto diverse tra loro. Per questo tipo di dati ipotizzare un modello multivariato per tutte le variabili, come succede con la data augmentation, può risultare un compito molto complicato. Da questo nasce l'idea di risolvere il problema di imputazione multivariato dividendolo in tanti problemi univariati (Van Buuren et al., 2006).

Nell'approccio di Raghunathan et al. (2001), il metodo di imputazione univariato per ciascuna delle variabili da imputare è un modello di regressione multiplo che comprende come variabili esplicative le informazioni osservate e le altre variabili imputate. A seconda delle caratteristiche della variabile da imputare, il modello di regressione può essere lineare, logistico (a due livelli o generalizzato), ecc. I valori imputati vengono quindi estratti, per ogni variabile, dalla distribuzione predittiva a posteriori specificata dal particolare modello di regressione scelto.

Più in dettaglio, dato un insieme di variabili osservate $\boldsymbol{X}$ e di variabili con valori mancanti $(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_k)$ con *pattern* monotono (paragrafo 1.2.2), la loro distribuzione congiunta viene fattorizzata nel modo seguente:

$$\begin{aligned} f(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_k | \boldsymbol{X}, \theta_1, \theta_2, \ldots, \theta_k) &= f_1(\boldsymbol{Y}_1 | \boldsymbol{X}, \theta_1) f_2(Y_2 | \boldsymbol{X}, \boldsymbol{Y}_1, \theta_2) \ldots \\ &\ldots \quad f_k(\boldsymbol{Y}_k | \boldsymbol{X}, \boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_{k-1}, \theta_k) \end{aligned}$$

[4]Alcuni dei software packages che utilizzano questo approccio sono presentati in Raghunathan et al. (1998); Van Buuren and Oudshoorn (1999); Royston (2005).

dove le f_j, $j = 1, \ldots, k$, sono le distribuzioni condizionate e θ_j i parametri che le caratterizzano. Ciascuna delle distribuzioni condizionate viene modellata attraverso il modello di regressione più appropriato, e le imputazioni vengono estratte dalla distribuzione predittiva corrispondente, che deriva anche da una distribuzione a priori solitamente non informativa per il parametro θ_j. Per esempio, il primo passo consiste nel regredire la variabile con il minor numero di valori mancanti, $\boldsymbol{Y}_1$, rispetto alle covariate osservate $\boldsymbol{X}$; se $\boldsymbol{Y}_1$ è continua e viene modellata attraverso una regressione lineare, la distribuzione da cui vengono estratte le imputazioni è la stessa già presentata nel paragrafo 2.1.2. Se invece la variabile Y_1 è binaria 0/1, la regressione scelta è quella logistica:

$$logit[P(\boldsymbol{Y}_1 = 1|\boldsymbol{X})] = \boldsymbol{X}\boldsymbol{\beta}.$$

In questo caso i passi da compiere per estrarre le imputazioni sono i seguenti:

1. calcolo di $\widehat{\boldsymbol{\beta}}$, stima di massima verosimiglianza di $\boldsymbol{\beta}$, e della sua matrice di varianza e covarianza asintotica $\boldsymbol{V}$;

2. generazione di un vettore di valori $\boldsymbol{Z}$ da una $N(0, I_q)$, dove q è il numero di predittori $\boldsymbol{X}$;

3. calcolo dei valori $\widehat{\boldsymbol{\beta}}_* = \widehat{\boldsymbol{\beta}} + \boldsymbol{T}\boldsymbol{Z}$, dove $\boldsymbol{T}$ deriva dalla decomposizione di Cholesky di $\boldsymbol{V}$;

4. data $\boldsymbol{X}_{mis}$, parte di $\boldsymbol{X}$ che corrisponde alle osservazioni con $\boldsymbol{Y}$ mancante, si calcolano i valori $p_* = [1 + exp(-\boldsymbol{X}_{mis}\widehat{\boldsymbol{\beta}}_*)]$;

5. generazione di n_{mis} valori $\boldsymbol{U}$ da una distribuzione uniforme sull'intervallo $(0, 1)$ e imputazione dei valori per cui $\boldsymbol{Y}_1$ è mancante secondo lo schema seguente: valore 1 se $\boldsymbol{U} \leq p_*$, 0 altrimenti.

Con questo approccio le imputazioni corrispondono ad estrazioni approssimate dalla distribuzione predittiva a posteriori dei valori mancanti, in quanto le estrazioni di $\boldsymbol{\beta}$ avvengono da un'approssimazione asintotica della vera distribuzione a posteriori. Nel prossimo paragrafo viene illustrato un algoritmo che può essere utilizzato per realizzare estrazioni dalla distribuzione a posteriori effettiva. Una volta realizzate le imputazioni per la variabile $\boldsymbol{Y}_1$, la seconda variabile con il minor numero di valori mancanti, $\boldsymbol{Y}_2$, viene regredita sui valori $\boldsymbol{X}$ e $\boldsymbol{Y}_1$ imputati, secondo il più appropriato modello di regressione, e così via. Altre specificazioni per i modelli di regressione sono illustrate in Raghunathan et al. (2001); alcune di queste vengono riprese e presentate nel capitolo 4.

Se il *pattern* dei dati è monotono, una volta che il metodo ha completato il primo ciclo delle variabili, le imputazioni realizzate sono estrazioni approssimate dalla distribuzione congiunta fattorizzata. Quando il *pattern* non è monotono, invece, per realizzare le imputazioni è necessario un algoritmo iterativo come il Gibbs sampler, come già visto nel paragrafo precedente. Le estrazioni dei valori mancanti della variabile Y_j al round $(t+1)$ dovrebbero avvenire dalla distribuzione:

$$f_j(Y_j|\theta_1^{(t+1)}, Y_1^{(t+1)}, \ldots, \theta_j^{(t+1)}, \theta_{j+1}^{(t)}, Y_{j+i}^{(t)}, \ldots, \theta_k^{(t)} Y_k^{(t)}, X) \tag{2.21}$$

Nell'algoritmo dell'approccio *SRMI* di Raghunathan et al. (2001), invece, i valori mancanti della variabile Y_j vengono estratti, al round $(t+1)$, dalla distribuzione:

$$g_j(Y_j|Y_1^{(t+1)}, Y_2^{(t+1)}, \ldots, Y_{j-i}^{(t+1)}, Y_{j+i}^{(t)}, \ldots, Y_k^{(t)}, X, \varphi_j) \tag{2.22}$$

dove g_j è determinata dal modello di regressione scelto per Y_j e φ_j rappresenta i parametri di tale regressione.

Questa distribuzione condizionata rappresenta un'approssimazione del vero e proprio Gibbs sampler che dovrebbe essere implementato; in pratica, se tutte le variabili da imputare sono continue e ciascun modello di regressione condizionato è un modello lineare normale con varianza costante, l'algoritmo converge ad una distribuzione predittiva congiunta normale multivariata, con distribuzioni a priori improprie per media e varianza (Raghunathan et al., 2001).

Per esempio, nel caso di variabili continue Y_j si specifica per ciascuna delle distribuzioni univariate un modello di regressione lineare avente come covariate le variabili completamente osservate, $\boldsymbol{X} = (X_1, \ldots, X_p)$, e i valori imputati per le altre variabili, $\boldsymbol{Y}_{mis(-j)} = \big(y_{mis(1)}, \ldots, y_{mis(j-1)}, y_{mis(j+1)}, \ldots, y_{mis(k)}\big)$. Allora si ha:

$$\begin{aligned} y_{mis(j)}|\boldsymbol{X}, \boldsymbol{Y}_{mis(-j)} &\sim N(\boldsymbol{X}_{mis(-j)}\beta_j, \sigma_j^2 \boldsymbol{I}_n) \; j = 1, \ldots, k \\ \pi\left(\beta_j, log(\sigma_j^2)\right) &\propto 1 \end{aligned}$$

dove $\boldsymbol{X}_{mis(-j)} = (1, \boldsymbol{X}, \boldsymbol{Y}_{mis(-j)})$ e $\beta_j = (\beta_{0(j)}, \ldots, \beta_{p+k-1(j)})$.

L'algoritmo SRMI prevede allora i seguenti passi:

1. per $j = 1, \ldots, k$, scegliere dei valori iniziali per $\beta_j^{(0)}, \sigma_j^{2(0)}$ e per $y_{mis(j)}^{(0)}$;
2. per $j = 1, \ldots, k$, date le estrazioni $\beta_j^{(t)}, \sigma_j^{2(t)}$ e $y_1^{(t+1)}, \ldots, y_{j-1}^{(t+1)}, y_{j+1}^{(t)}, \ldots, y_k^{(t)}$ all'iterazione t, le nuove estrazioni per i valori mancanti y_j e per i parametri della regressione all'iter $t+1$ si ottengono dalle distribuzioni seguenti:

$$\left(\sigma_j^{2(t+1)}|y_j^{(t)}, \boldsymbol{X}, \boldsymbol{Y}_{(-j)}^{t+j}, \beta_{(j)}^{(t)}\right) \sim$$

$$\sim Inv-\chi^2\left(n-p-k, \frac{1}{n-p-k}\left\|y_j^{(t)}-\boldsymbol{X}_{(-j)}^{(t+j)}\widehat{\beta}^{(t+1)_{(j)}}\right\|^2\right),$$

$$\left(\beta_{(j)}^{(t+1)}|y_j^{(t)}, x_1, \ldots, x_p, \boldsymbol{Y}_{(-j)}^{t+j}, \sigma_j^{2(t+1)}\right) \sim N\left(\widehat{\beta}_{(j)}^{(t+1)}, \sigma_j^{2(t+1)}\left(\boldsymbol{X}_{(-j)}^{(t+j)'}\boldsymbol{X}_{(-j)}^{(t+j)}\right)^{-1}\right),$$

$$\left(y_{ij}^{(t+1)}|x_1, \ldots, x_k, \boldsymbol{Y}_{(-j)}^{t+j}, \beta_{(j)}^{(t+1)}, \sigma_j^{2(t+1)}\right) \sim N\left(\boldsymbol{X}_{i(-j)}^{(t+j)}\widehat{\beta}_{(j)}^{(t+1)}, \sigma_j^{2(t+1)}\right),$$

dove

$$\begin{aligned}
\boldsymbol{Y}_{(-j)}^{t+j} &= \left(y_1^{(t+1)}, \ldots, y_{j-1}^{(t+1)}, y_{j+1}^{(t)}, \ldots, y_k^{(t)}\right), \\
\beta_{(j)}^{(t)} &= \left(\boldsymbol{X}_{(-j)}^{(t+j)'}\boldsymbol{X}_{(-j)}^{(t+j)}\right)^{-1}\boldsymbol{X}_{(-j)}^{(t+j)'}y_j^{(t)} \\
\boldsymbol{X}_{(-j)}^{(t+j)} &= \left(1, y_1, \ldots, y_k, \boldsymbol{Y}_{(-1)}^{t+j}\right)
\end{aligned}$$

e $\boldsymbol{X}_{i(-j)}^{(t+j)}$ è l'*i-esima* riga della matrice $\boldsymbol{X}_{(-j)}^{(t+j)}$.

In alcune situazioni è possibile che le estrazioni dalle distribuzioni condizionate di tipo (2.22) non convergano ad una distribuzione congiunta stazionaria, come in una normale sequenza Gibbs sampler (Raghunathan et al., 2001; Van Buuren and Oudshoorn, 1999), dando vita ad un Gibbs sampler "inconsistente". Ciò è vero, in particolare, quando le distribuzioni univariate (2.22) comprendono fattori di complicazione come limiti superiori ed inferiori. Tuttavia, numerosi studi hanno dimostrato che l'utilizzo di distribuzione condizionate incompatibili, seppure preoccupante dal punto di vista teorico, ai fini dell'imputazione di dati mancanti consente di ottenere risultati molto buoni (Van Buuren et al., 2006, 1999; Heeringa et al., 2002). Come sottolineato da Gelman and Raghunathan (2001), lo studio delle distribuzioni condizionate è uno di quelle aree della statistica in cui la teoria non è ancora al passo con la pratica.

2.2.3 Un algoritmo non iterativo: il *Sampling Importance Resampling*

Un metodo non iterativo che può essere utilizzato per realizzare imputazioni multiple è il *Sampling Importance Resampling* (SIR). Tale metodo, inizialmente proposto da Rubin (1987), continua ad essere studiato quale metodo alternativo alle tecniche iterative MCMC per l'estrazione dei valori mancanti dalla loro distribuzione a posteriori (Kong et al., 1994; Li, 2004; Tian et al.,

2007). Inoltre, il metodo è stato anche utilizzato in combinazione con le tecniche considerate nei paragrafi precedenti, come per esempio per effettuare estrazioni esatte dalle distribuzioni a posteriori derivanti dal modello di regressione logistico nell'approccio SRMI di Raghunathan et al. (2001).

Per implementare l'algoritmo SIR è necessario disporre di un'approssimazione della distribuzione dei valori di interesse $P(\boldsymbol{Y}_{mis}, \theta|\boldsymbol{X}, \boldsymbol{Y}_{oss})$:

$$\widetilde{P}(\boldsymbol{Y}_{mis}, \theta|\boldsymbol{X}, \boldsymbol{Y}_{oss}) = \widetilde{P}(\theta|\boldsymbol{X}, \boldsymbol{Y}_{oss})\widetilde{P}(\boldsymbol{Y}_{mis}|\boldsymbol{X}, \boldsymbol{Y}_{oss}, \theta)$$

per poter calcolare gli *importance ratios*:

$$r(\boldsymbol{Y}_{mis}, \theta) \propto P(\boldsymbol{Y}|\boldsymbol{X}, \theta)/\widetilde{P}(\boldsymbol{Y}_{mis}, \theta|X, \boldsymbol{Y}_{oss})$$

per tutti i $(\boldsymbol{Y}_{mis}, \theta)$ nei punti osservati $(\boldsymbol{X}, \boldsymbol{Y}_{oss})$.

L'algoritmo prevede infatti i seguenti passi (Rubin, 1987):

1. estrarre dalla distribuzione approssimata $\widetilde{P}(\boldsymbol{Y}_{mis}, \theta|\boldsymbol{X}, \boldsymbol{Y}_{oss})$ M valori di $(\boldsymbol{Y}_{mis}, \theta)$ con $M > m$;

2. calcolare gli importance ratios $r(\boldsymbol{Y}_{mis}, \theta) = (r_1, \ldots, r_M)$ per ciascuno degli M valori estratti;

3. estrarre m valori dagli M estratti al passo 1 con probabilità proporzionale a $r_1, \ldots, r_M$.

Per $M/m \to \infty$, gli m valori di $(\boldsymbol{Y}_{mis}, \theta)$ estratti al passo 3 hanno probabilità $P(\boldsymbol{Y}_{mis}, \theta|X, \boldsymbol{Y}_{oss})$; si ha infatti:

$$\frac{\widetilde{P}(\boldsymbol{Y}_{mis}, \theta|X, \boldsymbol{Y}_{oss})r(\boldsymbol{Y}_{mis}, \theta)}{\int\int \widetilde{P}(\boldsymbol{Y}_{mis}, \theta|X, \boldsymbol{Y}_{oss})r(\boldsymbol{Y}_{mis}, \theta)d\boldsymbol{Y}_{mis}d\theta} =$$
$$= \frac{P(\boldsymbol{Y}|\boldsymbol{X}, \theta)P(\theta)}{\int\int P(\boldsymbol{Y}|\boldsymbol{X}, \theta)P(\theta)d\boldsymbol{Y}_{mis}d\theta} = P(\boldsymbol{Y}_{mis}, \theta|X, \boldsymbol{Y}_{oss})$$

Quindi, l'algoritmo SIR è utile, in generale, in quelle situazioni in cui la distribuzione da cui si vogliono campionare dei valori è difficile da trattare, ma si dispone di una buona approssimazione della distribuzione stessa. Per l'applicazione a problemi di dati mancanti un fattore che può influire in positivo su questa approssimazione, giustificando l'utilizzo di un rapporto M/m piuttosto basso, è un basso valore della *fraction of missing information* (Rubin, 1987; Little and Rubin, 2002).

2.3 Conclusioni

In questo capitolo si sono presentate la principali caratteristiche e proprietà dell'imputazione multipla, seguendo l'impostazione di Rubin (1987). Nella trattazione sono stati sottolineati, oltre ad approfondimenti di tipo teorico, anche gli aspetti di tipo applicativo, che vengono ripresi nei capitoli seguenti della tesi.

Per quanto riguarda i metodi per realizzare imputazioni multiple utilizzando modelli bayesiani che si sono presentati, come già detto sebbene tali metodi siano molto utilizzati nella pratica, anche nell'ambito di indagini di pubbliche (Schenker et al., 2006), presentano ancora delle problematiche di tipo teorico.

In particolare, nel caso degli approcci che utilizzando la specificazione delle distribuzioni condizionate univariate attraverso opportuni modelli di regressione, sono attualmente in fase di elaborazione metodologie che garantiscano la convergenza degli algoritmi di tipo Gibbs sampling anche in caso di pattern non monotono. In particolare, ciò può risultare possibile se il dataset viene scomposto in sottoinsiemi di *blocchi monotoni* di dati, e le imputazioni vengono estratte, iterativamente, all'interno di tali blocchi (Rubin et al., 2004).

Un'altra problematica riguarda il fatto che i metodi precedentemente presentati sono stati concepiti per datasets le cui osservazioni possono essere considerate indipendenti tra loro. Tuttavia, in molte situazioni pratiche possono essere presenti nei dati delle strutture di correlazione tali da giustificare, per esempio, l'analisi dei dati completati attraverso modelli di regressione multilivello. In questo caso sarebbe opportuno che anche il modello utilizzato per realizzare le imputazioni presentasse la medesima struttura di correlazione; come sottolineato nel corso di questo capitolo, infatti, l'uguaglianza del modello di imputazione con quello di analisi può garantire la validità delle inferenze ottenute attraverso l'imputazione multipla. Per questo motivo sono attualmente in fase di implementazione versioni degli algoritmi sequential regression che contengano opportuni modelli di regressione multilivello (Yucel et al., 2006).

Infine, occorre puntualizzare che nella trattazione di questo capitolo si è volutamente scelto di presentare soltanto l'imputazione multipla come metodo per ottenere inferenze corrette partendo da datasets imputati. Altri metodi sono possibili. In particolare, sono stati proposti vari metodi per calcolare correttamente gli errori standard delle stime di interesse utilizzando metodi di imputazione singola, come per esempio l'hot-deck. Tali soluzioni, che comprendono per esempio l'utilizzo di metodi di ricampionamento come il *bootstrap* e il *jackknife* e di stimatori a due stadi (Rao and Shao, 1992; Fay,

1996; Lee et al., 2002; Little and Rubin, 2002), possono essere particolarmente utili a quelle istituzioni, per le quali, come avviene in Italia per ISTAT, la realizzazione di imputazioni multiple non rappresenta per il momento una metodologia di tipo standard.

Capitolo 3

L'imputazione di dati di reddito

Obiettivo principale di questo capitolo è presentare un panoramica su come è stato affrontato, fino a questo momento, il problema delle mancate risposte di reddito in alcune indagini campionarie realizzate nel nostro Paese. Nel primo paragrafo (3.1) viene introdotta e giustificata l'esigenza di rilevare le informazioni relative a quantità di reddito da indagine, mentre nel paragrafo 3.2 si presenta un'interessante dibattito nato negli Stati Uniti relativamente all'appropriatezza dell'utilizzo dell'ipotesi MAR per i dati di reddito della Current Population Survey. Infine, nel paragrafo 3.3 si presenta lo "stato dell'arte" per l'imputazione dei dati di reddito in Italia, con particolare riferimento all'indagine ISTAT sulle Condizioni di Vita e all'indagine Banca d'Italia sui Bilanci delle Famiglie.

3.1 La rilevazione del reddito attraverso indagini campionarie

Lo studio della distribuzione del reddito nella popolazione riveste un ruolo di importanza fondamentale per la comprensione di numerose dinamiche economiche e sociali. Solitamente le grandezze di reddito cui si fa riferimento sono due, il reddito individuale e il reddito familiare, quest'ultimo costituito dalla somma dei redditi individuali dei componenti di una stessa famiglia. Sulla base del reddito individuale risulta possibile, per esempio, classificare il livello di reddito secondo il tipo di attività lavorativa, l'età, il genere ed altre caratteristiche individuali. Il livello di reddito delle famiglie, invece, consente lo studio delle condizioni economico-sociali della popolazione, come la diffusione del benessere e della povertà, fattori che influenzano la programmazione degli interventi di politica sociale. Il reddito è infatti uno degli elementi utilizzati per costruire gli indicatori di povertà e disuguaglianza, che possono

servire anche per compiere confronti tra il livello di benessere economico in Paesi diversi.

In Italia le fonti in grado di fornire informazioni sul reddito individuale dei cittadini sono, innanzitutto, quelle di natura fiscale ed amministrativa. Tuttavia, le dichiarazioni fiscali non coprono tutta la popolazione, per esempio non riguardano i cittadini il cui reddito è sotto la soglia minima del reddito imponibile, e sono soggette al fenomeno dell'evasione fiscale, che in Italia viene solitamente indicato come rilevante soprattutto per i lavoratori autonomi. Anche i redditi rilevati da fonte amministrativa, per esempio dagli istituti previdenziali, non sono rappresentativi di tutta la popolazione; inoltre, così come avviene per i redditi da fonte fiscale, questi redditi sono solitamente messi a disposizione degli utilizzatori finali solo a certi livelli di aggregazione, e corredati da un insieme di informazioni che possono non essere esaustive ai fini di studi volti a determinare le cause e le caratteristiche peculiari dei livelli di reddito nella popolazione o in suoi sottoinsiemi. Ciò è vero, soprattutto, quando l'unità di interesse è la famiglia piuttosto che il singolo individuo, e sono necessarie informazioni di vario genere sia sui singoli componenti che sulla famiglia nel suo complesso (caratteristiche dell'abitazione, del principale percettore di reddito, dei consumi e dei beni posseduti, ecc.).

Da queste e da altre motivazioni nasce l'esigenza di rilevare il reddito familiare ed individuale attraverso indagine campionaria. La rilevazione del reddito da indagine presenta numerose problematiche: tra queste vi sono la definizione della popolazione di riferimento, la selezione del campione, le definizioni utilizzate per le grandezze di reddito, l'*under-reporting* e, ovviamente, le mancate risposte.

Le mancate risposte totali e parziali in indagini che hanno come principale obiettivo la rilevazione di informazioni sui redditi presentano problematiche peculiari che comprendono e in un certo senso superano tutte quelle presentate nei capitoli precedenti. In particolare, le mancate risposte a singoli quesiti di reddito, essendo *item nonresponses*, vengono solitamente trattate attraverso l'imputazione dei valori (paragrafo 1.2.1).

Tuttavia, come evidenziato nei capitoli precedenti, i metodi di imputazione attualmente più utilizzati, sia in ottica di imputazione singola che di imputazione multipla, ipotizzano che i dati siano mancanti a caso, assunzione che può risultare particolarmente forte proprio nel caso di dati mancanti di reddito. Inoltre, se il modello di imputazione condiziona rispetto a tutte le informazioni che si assume abbiano potere esplicativo sulla mancata risposta, talvolta l'ipotesi MAR può essere resa più plausibile (paragrafo 1.2.2). D'altra parte un test formale per tale ipotesi può essere realizzato solo se è possibile disporre, in un secondo momento o da fonte alternativa di dati, dei valori di reddito inizialmente mancanti, il che risulta spesso difficile da

ottenere. Questo è risultato possibile, negli Stati Uniti, per i dati di reddito provenienti dalla Current Population Survey, in merito ai quali è nato un acceso dibattito tra gli studiosi americani, i cui principali risultati vengono presentati nel prossimo paragrafo e da cui possono essere tratte utili indicazioni di interesse generale.

Un'altra caratteristica peculiare che rende problematica l'imputazione delle mancate risposte di reddito è il fatto che il tasso di mancate risposte risulta quasi sempre particolarmente elevato, spesso compreso tra il 20-40% (Heeringa et al., 2002), e solitamente molto maggiore rispetto a quello degli altri quesiti presenti nella stessa indagine.

Secondo Juster and Smith (1997) ci sono tre tipologie di problematiche cognitive che possono influire sugli alti tassi di mancate risposte per le variabili di reddito. Innanzitutto, è possibile che il rispondente non conosca la risposta, specialmente quando questa consiste nella somma di diverse voci di reddito; poi, se anche l'intervistato ha un'idea dell'ammontare richiesto, è possibile che si rifiuti di comunicarlo per non riportare un'inesattezza, ritenendo che l'intervistatore voglia sapere l'ammontare esatto. Infine, il rifiuto di rispondere può essere legato alla volontà di non far conoscere il reddito in questione, ritenuto una variabile delicata, troppo personale da poter essere rilevata anche per il timore che possano esserci ripercussioni di tipo fiscale.

Per cercare di ridurre questo tipo di mancate risposte, grande attenzione è stata dedicata negli ultimi anni alla pianificazione delle indagini (per esempio facendo precedere l'intervista da una lettera in cui si spiegano i fini puramente statistici dell'indagine, e si ricordano le normative in materia di privacy) e all'implementazione del questionario, utilizzando quindi le tecniche "preventive" delle mancate risposte già viste nel paragrafo 1.2.

Inoltre, in molti questionari è stata introdotta, nel caso dei quesiti di reddito, la possibilità di collocare la propria risposta in una classe di valori (*bracketed response*), a fronte del rifiuto di fornire un preciso ammontare; questo può aiutare l'intervistato se non conosce l'ammontare esatto e può garantirgli una maggiore riservatezza (se per esempio per i redditi più bassi e più elevati si predispongono delle classi aperte di valori) (Juster and Smith, 1997). I dati che sono una mistura di valori puntuali, classi di valori e valori mancanti sono stati definiti *coarsened data* (Heeringa et al., 2002).

Infine, nel caso di indagini di interesse nazionale, spesso viene prevista la possibilità di integrare i dati di reddito provenienti dall'indagine con le informazioni di origine fiscale ed amministrativa; questo tipo di procedura viene attualmente utilizzata, per esempio, per il reddito da lavoro autonomo rilevato dall'indagine ISTAT sulle Condizioni di Vita, presentata nel proseguo del capitolo e oggetto delle elaborazioni del capitolo 4.

3.2 Le mancate risposte di reddito: MAR o MNAR? Il caso della Current Population Survey negli Stati Uniti

Negli Stati Uniti la discussione sulla natura del meccanismo che genera le mancate risposte a quesiti di reddito è nata negli anni ottanta in merito alla Current Population Survey (CPS), indagine campionaria condotta a cadenza mensile dall'U.S. Census Bureau, il cui scopo principale è raccogliere informazioni sul reddito e sull'occupazione. In particolare, la rilevazione realizzata nel mese di marzo prevede la somministrazione di quesiti aggiuntivi relativamente al reddito degli intervistati in tutto il precedente anno di calendario[1] (David et al., 1986).

Il Census Bureau ha iniziato ad imputare i dati mancanti relativi al reddito attraverso un metodo da donatore, l'hot-deck, a partire dal 1962 (Lillard et al., 1986). Da allora numerosi studiosi hanno analizzato questi dati sia per valutare la bontà del metodo di imputazione utilizzato, sia per verificare l'ipotesi, sottostante alla tecnica hot-deck, che i dati mancanti siano MAR. Per svolgere quest'ultimo tipo di analisi è stato sfruttato il *matching* dei dati provenienti dall'indagine CPS con quelli di natura fiscale forniti dall'Internal Revenue Service (IRS). I risultati cui sono giunti gli studiosi americani in merito all'ipotesi MAR per i dati di reddito dell'indagine CPS sono in parte contrastanti.

Nel loro studio Greenlees et al. (1982) hanno testato l'ipotesi che la probabilità della mancata risposta a quesiti di reddito dipendesse dal reddito stesso, utilizzando il link tra i dati CPS e IRS. La loro conclusione è quella di una forte relazione negativa tra le due grandezze, ovvero di una probabilità di risposta più bassa all'aumentare del reddito. Tali autori hanno quindi suggerito l'utilizzo di un metodo di imputazione non ignorabile, appartenente alla classe dei *selection models*: il modello proposto per la modellazione della probabilità di risposta è una funzione logistica del reddito e di altre variabili. Per stimare tale modello gli autori propongono un metodo di massima verosimiglianza, che si basa sull'ipotesi che i residui siano normalmente distribuiti; tale modello, quindi, può risultare soggetto alle critiche e alle problematiche

[1] Al contrario del patrimonio, che è uno stock, il reddito è una grandezza di flusso. Questo significa che il reddito può essere definito come incremento o decremento, espresso in termini monetari, della ricchezza di un soggetto in un determinato periodo di tempo; senza il riferimento ad un preciso orizzonte temporale, quindi, il reddito non avrebbe senso. Ecco perchè nei questionari delle indagini che rilevano le componenti del reddito individuali e familiari deve essere sempre specificato il preciso periodo a cui l'intervistato deve riferirsi.

di stima già presentate in merito ai modelli non ignorabili nel paragrafo 1.1.3, e sottolineate da Little (1988).

Lillard et al. (1986) hanno evidenziato, attraverso un'analisi descrittiva dei dati CPS relativi ad un sottogruppo della popolazione, l'esistenza di un legame non monotono tra la probabilità di risposta a quesiti di reddito ed il reddito stesso. In particolare, nel loro lavoro viene messa in evidenza una caratteristica "forma ad U" per tale relazione, suggerendo una più alta probabilità di mancata risposta sia tra i percettori dei redditi più elevati che di quelli più bassi. In tale lavoro le probabilità più alte di mancate risposte ai quesiti di reddito vengono classificate anche in base all'attività lavorativa del percettore e ad altre caratteristiche individuali. In particolare, gli autori individuano nei percettori dei redditi più bassi e con contatti sporadici o irregolari con il mondo del lavoro un maggior tasso di mancata risposta non solo ai quesiti di reddito ma anche a tutti gli altri (*general nonreporters*), mentre per i percettori dei redditi più elevati le mancate risposte sono peculiari proprio dei quesiti di reddito (*specific nonreporters*). Gli autori hanno quindi suggerito che il metodo di imputazione utilizzato per il reddito, per esempio per la costruzione delle celle di imputazione hot-deck, dovrebbe tenere in considerazione anche le caratteristiche che possono determinare l'appartenenza ad uno dei suddetti gruppi, come per esempio il livello di istruzione, ipotizzato più basso per i general nonreporters.

Entrambi i lavori precedenti sono stati ripresi e in un certo senso confutati nell'articolo di David et al. (1986). In questo caso, sempre sfruttando il *matching* tra i dati CPS e quelli IRS, ma lavorando con tutti i dati e non con un loro sottoinsieme, è stata evidenziata la sostanziale non distorsione della distribuzione dei valori di reddito ottenuti attraverso diverse tecniche di imputazione di tipo MAR, tra cui modelli di regressione con residui casuali e lo stesso hot-deck utilizzato dal Census Bureau. Questo lavoro ha rappresentato dunque un "rilancio" per le tecniche di imputazione che utilizzano l'ipotesi MAR anche per dati mancanti di reddito, sebbene ciò richieda un'accurata implementazione delle tecniche stesse, come per esempio l'utilizzo di modelli multivariati quando le grandezze di reddito rilevate sono più di una.

Little (1988), riprendendo ancora i risultati dei precedenti studi, ha sottolineato l'importanza di utilizzare come predittori nei modelli di regressione per l'imputazione del reddito anche variabili che potrebbero essere considerate non *esogene* secondo studi comportamentali di tipo econometrico. Per esempio, anche una variabile relativa alla grandezza dell'abitazione, se disponibile, dovrebbe essere utilizzata come esplicativa: l'esogeneità è irrilevante per l'imputazione, dal momento che il fine è prevedere, non compiere inferenza causale. Secondo Little è dunque particolarmente importante costruire buoni modelli di imputazione che utilizzino l'ipotesi di dati di reddito MAR:

il punto principale non è se la non risposta al reddito è non casuale, ma se la non risposta al reddito è non casuale dopo aver condizionato per l'informazione di tutte le covariate; David et al. (1986), sottolinea Little (1988), non hanno trovato alcuna evidenza a supporto di quest'ultimo punto.

A sostegno di queste ultime conclusioni in merito all'imputazione di dati di reddito vi è il fatto che l'imputazione multipla di questi dati secondo modelli di regressione basati sull'ipotesi MAR sta conoscendo negli Stati Uniti una rapida diffusione. Ne sono un esempio la Consumer Expenditures Survey, condotta dall'U.S. Department of Labor, e la National Health Interview Survey, condotta dal National Center for Health Statistics.

3.3 Il trattamento delle mancate risposte di reddito in Italia

In Italia non si è assistito, almeno per il momento, ad un dibattito acceso sul meccanismo che genera le mancate risposte a quesiti di reddito come quello presentato nel paragrafo precedente. Questo può essere presumibilmente dovuto al fatto che in Italia l'imputazione delle mancate risposte parziali nelle indagini condotte su scala nazionale non ha una solida tradizione come negli Stati Uniti. Questa problematica ha tuttavia raccolto un interesse crescente negli ultimi anni, comportando un progressivo "aggiornamento" delle procedure di imputazione alla luce delle nuove proposte apparse nella letteratura; tuttavia, almeno per il momento, in Italia per nessuna indagine condotta su larga scala e i cui dati siano di pubblico utilizzo i dati mancanti vengono trattati attraverso l'imputazione multipla dei valori.

Questo è vero, in particolare, per le due principali indagini italiane che hanno come principale obiettivo la misura del reddito delle famiglie italiane: l'indagine ISTAT sulle Condizioni di Vita, che fa riferimento al progetto europeo SILC (Statistics on Income and Living Conditions), e l'indagine sui Bilanci delle Famiglie della Banca d'Italia.

L'indagine ISTAT sulle Condizioni di Vita ed il relativo progetto EU-SILC vengono descritte nel dettaglio nel capitolo 4, dove viene implementata una procedura di imputazione multipla per i dati di reddito provenienti dalla rilevazione 2004, la prima effettuata in Italia.

Per il momento si specifica soltanto che questa indagine, che nel 2004 ha raggiunto 24204 famiglie e 52509 individui, ha come principale obiettivo la produzione di dati comparabili tra i Paesi europei relativamente alla distribuzione del reddito ed al livello e composizione della povertà e dell'esclusione sociale (Parlamento Europeo, 2003*a*). Le variabili di reddito rilevate sono

numerose, e consentono di costruire le variabili e gli indicatori di diretto interesse di EUROSTAT.

Le mancate risposte parziali per le variabili di reddito[2] raggiungono in certi casi valori superiori al 20%: per esempio, la percentuale di mancate risposte è pari al 21% circa per il reddito individuale da lavoro autonomo[3], mentre scende all'8% circa per quello relativo ai lavoratori dipendenti.

L'imputazione di questi valori mancanti da parte di ISTAT viene realizzata, secondo quanto già suggerito da EUROSTAT per l'European Community Household Panel, indagine sostituita dalla EU-SILC, attraverso il metodo delle regressioni sequenziali multivariate (paragrafo 2.2.2) (EUROSTAT, 2001). La descrizione delle variabili utilizzate come esplicative nelle regressioni è a disposizione degli utilizzatori dei dati (Vitaletti, 2005). Tale metodo viene utilizzato per l'imputazione singola dei valori, e l'analisi dei datasets completati non pone alcuna enfasi sulle correzioni necessarie per ottenere standard error non sottostimati. L'imputazione multipla, come già sottolineato, ad oggi non è considerata una possibile soluzione per il trattamento delle mancate risposte nell'ambito della statistica ufficiale in Italia. Il rilascio di più di un dataset finale comporterebbe infatti una corrispondente diffusione sia delle giustificazioni teoriche dell'imputazione multipla, sia della spiegazione delle regole di combinazione attraverso cui gli utilizzatori finali possono ottenere le statistiche di interesse. I risultati presentati nel capitolo 4 sono in grado di fornire una prima risposta sull'effettiva rilevanza dell'introduzione di un metodo di imputazione multipla per i valori di reddito dell'indagine ISTAT EU-SILC.

L'altra indagine di carattere nazionale in grado di fornire informazioni sui redditi delle famiglie italiane è l'indagine della Banca d'Italia sui Bilanci delle Famiglie. Questa indagine nasce negli anni '60 con lo specifico obiettivo di raccogliere informazioni sui redditi e risparmi delle famiglie italiane; il campione è composto da circa 8000 famiglie, e l'indagine viene realizzata ogni due anni.

Il tasso di mancate risposte totali che caratterizza questa indagine è piuttosto elevato; in particolare, il tasso di risposta è compreso tra il 58% ed il 36% per le indagini condotte tra il 1995 ed il 2004 (Banca d'Italia, 2006);

[2]Le mancate risposte totali, come viene spiegato più in dettaglio nel capitolo 4, vengono trattate attraverso tecniche di ponderazione; nel caso della prima rilevazione EU-SILC, inoltre, è stata effettuata la calibrazione rispetto a totali noti derivati da fonte alternativa (ISTAT, 2006).

[3]Come già accennato, i dati relativi al reddito da lavoro autonomo provenienti dall'indagine sono stati completati, in caso di *match* positivo, con dati di origine fiscale; nonostante ciò la percentuale di mancate risposte è risultata comunque piuttosto elevata, pari al 21.21%.

questo trend decrescente è in linea con quelli riscontrati negli altri Paesi sviluppati nello stesso periodo di tempo (de Leeuw and de Heer, 2002). Le mancate risposte totali vengono trattate attraverso tecniche di ponderazione; tali tecniche tengono in considerazione che la probabilità di mancata partecipazione della famiglia può dipendere da alcune covariate osservate, come per esempio il titolo di studio del capofamiglia.

Le mancate risposte parziali, che riguardano variabili come le integrazioni non monetarie dei lavoratori dipendenti, i proventi dei lavoratori autonomi ed il valore delle aziende riguardano, in media, meno del 7 per cento dei casi rilevati. Per il trattamento delle mancate risposte parziali la Banca d'Italia utilizza modelli di regressione con residuo casuale, al fine di evitare un'eccessiva concentrazione delle imputazioni attorno al valore medio; le componenti casuali vengono estratte da una variabile casuale normale con media zero e varianza pari a quella stimata per il modello (Banca d'Italia, 2006). Non è chiaramente indicato, tuttavia, quali siano le variabili utilizzate come predittori nei modelli di regressione, e non viene proposto un particolare metodo per calcolare gli standard errors delle stime.

I dati dell'indagine sui Bilanci delle Famiglie sono stati utilizzati in vari studi, nel corso degli ultimi anni, che hanno messo in evidenza alcune caratteristiche di particolare interesse relativamente alle mancate risposte di reddito. Per l'immediato futuro c'è da attendersi che anche i dati provenienti dall'indagine ISTAT sulle Condizioni di Vita motiveranno altrettanti studi ed analisi.

In particolare, D'Alessio and Faiella (2002) hanno condotto uno studio sulle mancate risposte totali, evidenziando una relazione positiva tra il livello di reddito e la non risposta per famiglie inizialmente non rispondenti ma successivamente ricontattate: le famiglie con redditi più elevati hanno una maggiore propensione a non partecipare all'indagine. Come riportato da Brandolini (1999), la correlazione negativa tra tasso di risposta e livello di reddito ha comportato, in studio condotto sui dati del 1987, una correzione verso l'alto della stima del reddito familiare pari al 5%. Inoltre, relativamente a dati del 1989, l'*under-reporting* dei lavoratori autonomi è stato stimato attorno al 20%. Tali risultati derivano tuttavia da un numero limitato di re-interviste e non è chiaro quale fosse il modello utilizzato per realizzare le imputazioni; c'è da chiedersi quindi se non sarebbe possibile, disponendo degli stessi dati, giungere a conclusioni diverse, così come è avvenuto negli Stati Uniti per i dati della CPS.

Nello studio di Quintano et al. (2001) è stata invece simulata, nei dati già completati provenienti dall'indagine sui Bilanci delle Famiglie 1998, una percentuale pari al 20% di mancate risposte per la variabile relativa al reddito familiare annuale. La simulazione, basata su un modello MAR in cui il

reddito dipende dai consumi della famiglia (ad un maggior consumo corrisponde una maggiore probabilità di mancata risposta) ed un modello in cui il reddito dipende, sempre positivamente, dalla ricchezza della famiglia, ha evidenziato una sostanziale robustezza[4] dei due modelli testati nel ricostruire i missing values simulati. Le differenze principali sono state riscontrate solo in merito all'ultimo decile della distribuzione del reddito familiare annuale, e per la correlazione di tale variabile con altre variabili economiche. I risultati che è possibile dedurre da studi di simulazione come questo sono sicuramente molto interessanti, ed in linea con quanto suggerito anche per testare la robustezza rispetto ad ipotesi MNAR.

Indicazioni interessanti relativamente al meccanismo di mancata risposta che caratterizza l'indagine sui Bilanci delle Famiglie sono giunte, inoltre, anche da vari studi che hanno sfruttando la componente longitudinale del campione. A partire dal 1989 nel campione dell'indagine della Banca d'Italia è stata infatti introdotta una parte panel, inizialmente volontaria e successivamente casualizzata a partire dall'indagine del 1995. Sfruttando questa componente panel è possibile valutare l'impatto delle mancate risposte nel campione complessivo; D'Amuri and Fiorio (2004) riportano uno studio della Banca d'Italia del 1992 secondo cui le mancate risposte, in base all'analisi dell'*attrition*, risultavano più elevate tra le famiglie residenti nelle aree urbane e nel nord Italia.

In merito a questo è interessante sottolineare che anche il campione dell'indagine ISTAT sulle Condizioni di Vita comprenderà, a partire dalla seconda rilevazione, una componente longitudinale (Parlamento Europeo, 2003*a*). Utilizzando tale componente sarà possibile, in futuro, ottenere informazioni in merito alle caratteristiche che contribuiscono a determinare la mancata partecipazione delle famiglie. Inoltre, le informazioni relative al reddito ottenute in un'occasione precedente o successiva a quella caratterizzata da mancata risposta potrebbero essere utilizzate per l'imputazione dei valori mancanti; per esempio l'imputazione realizzata attraverso un modello di regressione che condiziona rispetto al reddito dichiarato in un'altra occasione è in grado di sfruttare questa importante informazione.

3.4 Conclusioni

L'imputazione dei dati mancanti di reddito presenta problematiche notevoli; una delle difficoltà principali è la possibile non ignorabilità del meccanismo che causa i valori mancanti. D'altra parte anche tale affermazione risulta

[4]Il metodo di imputazioni utilizzato per ri-completare il dataset è l'imputazione multipla basata su modelli di regressione.

difficilmente verificabile: gli stessi dati fiscali, che potrebbero essere utilizzati per testare l'ipotesi MAR per i redditi provenienti da indagini campionarie, sono soggetti al fenomeno dell'evasione e a difficoltà di *matching*.

Anche negli Stati Uniti, dove il dibattito è iniziato molto prima che in Italia, gli studiosi non sono giunti ad una conclusione definitiva; inoltre, è possibile che ciò che risulta vero per una particolare indagine non lo sia invece per altre, dal momento che il campione, il questionario, il metodo di somministrazione, ecc. sono tutti elementi che possono influenzare il meccanismo di risposta.

Il punto importante, come sottolineato da vari autori, è l'utilizzo di un buon modello di imputazione: se tale modello condiziona rispetto a tutte le informazioni osservate, e tali informazioni fanno parte o sono correlate con le variabili del sottostante e sconosciuto meccanismo di mancata risposta, allora è possibile che l'ipotesi MAR sia approssimativamente valida anche per i dati di reddito. Ecco perchè nel capitolo 4 viene implementata una procedura di imputazione multipla per i dati mancanti di reddito dell'indagine ISTAT EU-SILC che pone grande attenzione alla selezione delle variabili esplicative. Inoltre, viene presentata ed applicata una metodologia che, confrontando la distribuzione dei valori imputati e osservati per una variabile di reddito, a parità di tutte le altre variabili osservate o imputate, può fornire utili indicazioni in merito all'ipotesi MAR. Un altro obiettivo particolarmente rilevante del lavoro presentato nel capitolo 4, inoltre, è la valutazione di quanto l'imputazione multipla dei valori mancanti di reddito possa contribuire a misurare l'incertezza delle stime di interesse, rispetto all'imputazione singola attualmente utilizzata dall'ISTAT. In questo senso, quindi, il lavoro vuole essere una prima risposta al possibile dubbio in merito all'utilizzo dell'imputazione multipla in indagini di tipo "ufficiale", procedura come più volte sottolineato già in uso negli Stati Uniti, ma ancora lontana dall'effettiva applicabilità nel nostro Paese.

Oltre a costruire un buon modello di imputazione, che possa rendere più plausibile l'ipotesi di dati di reddito MAR, sarebbe opportuno andare a verificare l'effetto sulle stime di interesse di possibili spostamenti rispetto a tale ipotesi. Anche in questo caso l'imputazione multipla può rivelarsi un utile strumento: l'analisi di sensitività condotta nel capitolo 5 viene realizzata proprio attraverso una procedura di imputazione multipla. In particolare, piuttosto che utilizzare modelli non ignorabili, che possono essere opinabili e difficili da stimare, la metodologia proposta ipotizza deviazioni dall'ipotesi MAR di facile comprensione ed in grado di fornire utili ed immediate conclusioni anche in situazioni complesse come nel caso del "panel-ruotato" dell'indagine Forze Lavoro del Comune di Firenze.

Capitolo 4

L'imputazione dei dati di reddito dell'indagine ISTAT sulle Condizioni di Vita EU-SILC 2004

In questo capitolo viene proposta una procedura per l'imputazione multipla dei dati mancanti di reddito per l'indagine ISTAT sulle Condizioni di Vita 2004, indagine che rientra nel progetto europeo EU-SILC (paragrafo 4.1). Dopo un'ampia descrizione degli obiettivi e dei questionari dell'indagine (paragrafi 4.2 e 4.3), l'attenzione si concentra sui dati, con la descrizione delle percentuali di mancate risposte, delle caratteristiche distributive delle variabili, del pattern dei dati, ecc. (paragrafi 4.4 e 4.5). La procedura di imputazione multipla proposta per i dati di reddito è presentata nel paragrafo 4.6, assieme ai risultati delle analisi condotte sui datasets imputati, che considerano anche la *available case analysis*. Infine, nel paragrafo 4.7 si presenta e si applica una diagnostica per testare non formalmente l'ipotesi MAR. Alcune considerazioni conclusive chiudono il capitolo.

4.1 Il progetto EU-SILC

Il progetto EU-SILC (European Union - Statistics on Income and Living Conditions) è definito dal regolamento N. 1177/2003 del Parlamento Europeo e del Consiglio dell'Unione Europea del giugno 2003. Scopo del progetto è la necessità, già riscontrata nei Consigli Europei degli anni 2000 e 2001, di poter "disporre di dati comparabili e tempestivi sia trasversali che longitudinali sulla distribuzione del reddito, nonchè sul livello e sulla composizione della

povertà e dell'esclusione sociale, per poter effettuare comparazioni attendibili e pertinenti tra gli Stati membri" (Parlamento Europeo, 2003*a*). Il progetto EU-SILC sostituisce, riprendendone e arricchendone parte delle finalità, l'European Community Household Panel, indagine longitudinale svolta tra il 1994 e il 2001, e progettata per raccogliere informazioni comparabili a livello europeo sulle condizioni di vita delle famiglie e sulle misure di politica economica e sociale a livello comunitario (EUROSTAT, 2003).

Il regolamento europeo N. 1177/2003 individua quale popolazione di riferimento per le statistiche del progetto EU-SILC tutte le famiglie ed i loro componenti di sedici anni e più residenti nel territorio degli Stati membri al momento della rilevazione dei dati. Agli Stati viene chiesto, a meno di particolari deroghe, di iniziare la prima rilevazione trasversale e longitudinale a partire dal 2004. La modalità di selezione del campione probabilistico, rappresentativo a livello nazionale, è a discrezione dei vari paesi, con il vincolo che la componente longitudinale comprenda un minimo di quattro anni e che il periodo di riferimento del reddito sia di dodici mesi.

Relativamente alle "variabili target primarie", ovvero le variabili che ogni paese deve necessariamente trasmettere ad EUROSTAT, queste possono essere suddivise in quattro tipologie (EUROSTAT, 2006):

- variabili misurate a livello familiare (tra cui alcune componenti di reddito);
- informazioni sulla numerosità e sulla composizione della famiglia;
- reddito e altre variabili "base" (istruzione, stato occupazionale, ecc.) misurate a livello individuale e solitamente aggregate a livello familiare;
- variabili rilevate e analizzate a livello individuale (condizioni di salute, informazioni dettagliate sul lavoro svolto, ecc.).

Tali aspetti sono stati rilevati attraverso l'utilizzo di un questionario familiare, somministrato ad un solo membro di ogni famiglia entrata a far parte del campione, possibilmente il membro responsabile per l'abitazione, e di un questionario individuale, somministrato a tutti i componenti sopra i 16 anni. Oltre all'indagine diretta, ai singoli Stati è stata concessa l'autonomia di attingere alcune informazioni da fonti esterne, come altre indagini campionarie o archivi amministrativi (Parlamento Europeo, 2003*a*).

Altri aspetti più tecnici relativi all'implementazione pratica del progetto EU-SILC sono contenuti in cinque regolamenti specifici prodotti dalla Commissione Europea. In particolare, il regolamento N. 1981/2003 dell'ottobre 2003 attua il regolamento 1177/2003 relativamente agli aspetti della rilevazione sul campo e delle procedure di imputazione. In tale regolamento i dati

mancanti vengono classificati in quattro diverse tipologie, definite anche in relazione alla struttura delle variabili di interesse:

- errori di copertura e selezione del campione;
- non risposta totale da parte di una unità (famiglie e/o individui);
- non risposta parziale di una unità (famiglie e/o individui).

Per trattare questi dati mancati il regolamento stabilisce la possibilità di affrontare due diversi approcci: l'imputazione, attraverso cui le informazioni mancanti vengono generate in base a relazioni statistiche interne all'insieme dei dati, e la modellizzazione, che utilizza invece informazioni esterne all'insieme dei dati (Parlamento Europeo, 2003*b*).

Relativamente alle procedure adottate, l'indicazione è quella di utilizzare metodi che conservino la variazione e la correlazione tra le variabili, attraverso la presenza di una componente erratica, e che tengano in considerazione la struttura di correlazione delle variabili stesse, preferendo quindi gli approcci multivariati.

Più in particolare, in una successiva documentazione (EUROSTAT, 2006) i metodi di correzione suggeriti per le prime due tipologie di mancata risposta sono l'uso di pesi e di tecniche di calibrazione, mentre per compensare le mancate risposte parziali vengono suggerite le tecniche di imputazione. La scelta delle particolari tecniche è lasciata a discrezione nazionale. L'ultima direttiva di carattere generale relativa ai dati mancanti è che "le variabili di reddito devono avere un *imputation factor* (...) questo è un numero positivo, risultato della divisione tra il valore rilevato all'intervista e il valore registrato nel database; variabili totalmente imputate hanno *imputation factor* pari a zero (il valore rilevato è nullo), variabili non imputate hanno *imputation factor* pari ad uno (valore rilevato e registrato sono uguali)" (EUROSTAT, 2006).

Se anche le precedenti documentazioni della Commissione Europea non indicano un preciso metodo di imputazione per le variabili di reddito dei questionari EU-SILC, va detto che nel caso dell'European Community Household Panel, indagine sostituita dal progetto EU-SILC, EUROSTAT suggeriva l'uso delle regressioni sequenziali multivariate (paragrafo 2.2.2) implementate dal software IVEware (EUROSTAT, 2001).

4.2 L'indagine ISTAT sulle Condizioni di Vita

Già dal 2004, primo anno di vita di EU-SILC, l'Italia ha partecipato al progetto con l'idagine ISTAT sulle Condizioni di Vita. Il disegno di campionamento e la strategia per la costruzione delle stime trasversali utilizzate nel nostro paese vengono descritti in dettaglio nella pubblicazione ISTAT (2006), in cui vengono anche presentati i risultati di questa prima indagine EU-SILC .

In particolare, la popolazione oggetto di indagine è l'insieme delle famiglie residenti in Italia e degli individui maggiori di quindici anni che le compongono[1]. Le famiglie intervistate nel 2004 sono state 24204, composte da un totale di 61429 individui, di cui 52509 sopra i quindici anni.

La rilevazione del 2004 non costituisce solo la prima componente trasversale EU-SILC ma anche la prima componente longitudinale, dal momento che nel nostro Paese si è deciso di adottare un'unica rilevazione integrata per le due componenti. L'indagine 2004 verrà dunque integrata con le indagini successive secondo un campione di tipo panel ruotato, rappresentato in figura 4.1. Il campione relativo ad ogni occasione di indagine (T, t+1, t+2, t+3 e t+4, ecc.) sarà costituito da quattro gruppi di rotazione (A, B, C, ecc.), ciascuno dei quali rimarrà nel campione per quattro anni successivi. I dati che si analizzano nei prossimi paragrafi, provenendo dalla prima rilevazione, hanno solo carattere trasversale: nella figura 4.1 possono essere pensati come provenienti dai gruppi "A4", "B3", "C2" e "D1" ponendo il tempo T=2004.

	A	B	C	D	E	F	G	H	I
T	A4	B3	C2	D1					
t+1		B4	C3	D2	E1				
t+2			C4	D3	E2	F1			
t+3				D4	E3	F2	G1		
t+4					E4	F3	G2	H1	
t+5						F4	G3	H2	I1

Figura 4.1: Schema di rotazione indagine EU-SILC.

Per quanto riguarda lo schema di selezione delle unità campionarie, per la prima rilevazione EU-SILC l'ISTAT ha utilizzato uno schema di tipo com-

[1]La popolazione di indagine definita dall'ISTAT è più ampia di quella richiesta dalla Commissione Europea, comprendendo anche gli individui tra quindici e sedici anni. Tale scelta è dovuta a motivi di comparabilità con altre indagini campionarie svolte in Italia, e comporta una differenziazione in sede di produzione delle stime a seconda dell'utilizzatore finale.

plesso basato sulle regioni geografiche italiane quali domini territoriali di studio. In particolare, i comuni sono stati suddivisi in due gruppi:

- comuni Auto-Rappresentativi (AR), ovvero comuni con maggiore dimensione demografica;

- comuni Non Auto-Rappresentativi (NAR).

Successivamente i comuni sono stati stratificati in modo differenziato a seconda del gruppo di appartenenza:

- campionamento ad uno stadio, considerando ogni comune come strato a sè, per i comuni di tipo AR: le unità primarie di campionamento sono costituite dalle famiglie anagrafiche, estratte dall'anagrafe del comune stesso, e tutti i componenti le famiglie sono oggetto di indagine;

- campionamento a due stadi per i comuni di tipo NAR: le unità primarie di campionamento sono costituite dai comuni, stratificati a seconda della dimensione demografica, mentre le unità secondarie sono le famiglie, e tutti i componenti delle famiglie sono oggetto di indagine.

Il procedimento di stratificazione dei comuni tiene in considerazione, oltre alla dimensione demografica, anche altre condizioni, come per esempio la selezione di un numero minimo di comuni in ciascuno strato di comuni NAR; per maggiori dettagli in merito al procedimento di stratificazione si rimanda a ISTAT (2006).

Il disegno di campionamento di tipo complesso appena descritto ha comportato, in fase di costruzione delle stime trasversali relative all'anno 2004, l'utilizzo della procedura di stima generalizzata utilizzata dall'ISTAT per tutte le indagini campionarie. Tale procedura utilizza tecniche di calibrazione (Deville and Sarndal, 1992) per produrre un unico coefficiente di riporto all'universo, sia a livello individuale che familiare, che produce stime coerenti con totali noti (ISTAT, 2006). Per l'indagine 2004, in particolare, poichè tutti i membri sopra i 15 anni della stessa famiglia entrano a far parte del campione, i pesi trasversali individuali sono uguali al corrispondente peso familiare (EUROSTAT, 2006).

Tra gli elementi considerati nella complessa procedura di calcolo dei coefficienti di riporto all'universo vi è anche la mancata risposta totale. Relativamente a tale aspetto, ISTAT utilizza solitamente l'ipotesi che il meccanismo che genera le mancate risposte totali sia ignorabile all'interno di opportune celle di ponderazione, utilizzando così un modello *implicito* per il meccanismo di mancata risposta totale (paragrafo 1.2.1). Nel caso di EU-SILC le

caratteristiche considerate sono la dimensione demografica del comune, la cittadinanza dell'individuo di riferimento, la regione di residenza ed il numero di componenti la famiglia. Se la numerosità in una cella è scarsa, la procedura prevede il collassamento con le celle attigue, seguito da un successivo passo di riproporzionamento (ISTAT, 2006).

Per la prima indagine EU-SILC tuttavia, la sola procedura di suddivisione in celle di ponderazione non sembrava garantire l'ipotesi di ignorabilità del meccanismo di mancata risposta totale, date alcune differenze riscontrate tra le stime totali provvisorie e quelle ricavate da fonti alternative utilizzate in fase di validazione dei dati. Per evitare possibili distorsioni è stato dunque deciso di introdurre un primo passo di calibrazione utilizzando come fonte esterna i totali derivati dalla Rilevazione Continua sulle Forze Lavoro relativa allo stesso periodo di indagine. I totali considerati sono la distribuzione delle famiglie per numero di componenti, la distribuzione per sesso, età e posizione nella professione (ISTAT, 2006).

L'applicazione ai dati EU-SILC di questo capitolo riguarda le mancate risposte a singoli item, che vengono imputate attraverso metodi basati su modello *esplicito*; tale procedura viene realizzata senza tenere in considerazione le mancate risposte totali, che vengono compensate nella fase di analisi finale dei dati utilizzando i pesi ISTAT.

4.3 Il questionario dell'indagine sulle Condizioni di Vita 2004

L'indagine ISTAT sulle Condizioni di Vita 2004 è caratterizzata da un questionario molto articolato, somministrato durante i primi mesi del 2004 con modalità "faccia a faccia". In particolare, dato l'interesse a rilevare sia caratteristiche delle famiglie che dei singoli componenti, il questionario è stato suddiviso in due distinti moduli[2]: il questionario familiare, sottoposto ad un solo membro di ogni famiglia, ed il questionario individuale, somministrato a tutti i maggiori di quindici anni appartenenti alle famiglie intervistate. Inoltre, ogni rilevatore disponeva anche di un modulo attraverso cui ricostruire la composizione familiare e registrare le caratteristiche base (sesso, data di nascita, ecc.) di ogni componente.

Il periodo di riferimento per i quesiti relativi al reddito, sia nel questionario familiare che in quello individuale, sono gli ultimi 12 mesi precedenti il momento dell'intervista o l'anno solare 2003.

[2]I questionari dell'indagine ISTAT Condizioni di Vita sono consultabili sul sito internet dell'Istituto, www.istat.it.

Il questionario familiare è suddiviso nelle seguenti quattro sezioni:

- La casa e la zona di abitazione;
- Affitto e subaffitto;
- Case di proprietà;
- La situazione economica.

La prima sezione ha come oggetto le condizioni di vita della famiglia relativamente all'abitazione; i quesiti riguardano infatti la tipologia di casa, le sue caratteristiche ed eventuali problematiche, per esempio umidità, danneggiamento al tetto, ecc., unitamente alle eventuali problematiche della zona di residenza (inquinamento, criminalità, ecc.). Inoltre, vengono richieste le spese sostenute negli ultimi 12 mesi per il pagamento di tutte le utenze, ovvero acqua, gas, legna, elettricità, nettezza urbana, telefono fisso, e per eventuali lavori di riparazione. Relativamente a tali spese, alla famiglia viene chiesto se ha potuto disporre di contributi pubblici per sostenerle. Inoltre, per meglio comprendere il livello di benessere, sempre in questa sezione viene richiesto il possesso o meno di una lista di beni, tra cui lavatrice, televisione a colori, personal computer, ecc.

Nella sezione dedicata alle famiglie che vivono in una casa in affitto, i quesiti di maggior interesse riguardano l'ammontare dell'affitto e l'eventuale difficoltà da parte della famiglia nel far fronte a tale spesa. Per le famiglie che vivono in una casa di proprietà, invece, l'interesse principale riguarda la contrazione di un mutuo, gli interessi pagati per il mutuo e l'eventuale difficoltà a sostenerne il pagamento. In entrambe queste due sezioni, inoltre, è prevista una domanda finale in cui si richiede se la famiglia ha potuto disporre, durante il 2003, di contributi pubblici per sostenere le spese relative all'affitto o al mutuo.

Infine, nell'ultima sezione del questionario familiare, per capire l'eventuale disagio economico, viene chiesto se la famiglia può permettersi alcuni beni non necessari, come per esempio una settimana di ferie all'anno, e se si è trovata in difficoltà negli ultimi dodici mesi per sostenere le spese per vestiti, cibo, scuola, ecc.

Tra le informazioni rilevate attraverso questo questionario vi sono anche le variabili target EUROSTAT riferite al reddito. Sono queste le variabili che l'ISTAT deve necessariamente trasmettere ad EUROSTAT, in quanto andranno a costituire uno degli elementi di confronto con gli altri paesi europei.

Alcune delle variabili target EUROSTAT misurate a livello familiare sono costituite da singoli quesiti del questionario familiare o da loro combinazioni:

- Benefici per la casa = Contributi pubblici per spese per la casa (escluso affitto) + Contributi pubblici per affitto + Contributi pubblici per interessi sul mutuo;
- Esclusione sociale = Reddito minimo vitale * numero mesi di percezione.

Altre variabili target riferite alle famiglie sono invece derivabili come somma a livello familiare di grandezze rilevate a livello dei singoli componenti attraverso il questionario individuale.

Il questionario individuale dell'indagine 2004, molto articolato, è suddiviso nelle seguenti sezioni:

- Dati anagrafici;
- Istruzione;
- Condizioni di salute;
- Lavoro e non lavoro;
- Attività lavorativa principale;
- Redditi correnti da lavoro dipendente;
- Attività lavorativa svolta in passato;
- Informazioni sul lavoro svolto;
- Condizione nella professione;
- Redditi da lavoro dipendente percepiti nel 2003;
- Redditi da lavoro autonomo percepiti nel 2003;
- Pensioni e indennità, assegni o pensioni di invalidità, inabilità o per infortuni sul lavoro percepite nel 2003;
- Altre informazioni relative al 2003.

Nelle prime tre sezioni del questionario vengono raccolte informazioni relative alla cittadinanza dell'intervistato, al livello di istruzione e alle eventuali esperienze di formazione ancora in corso, alle condizioni di salute e all'eventuale difficoltà a sostenere spese relative a visite mediche. Con le sezioni successive, invece, l'attenzione si sposta sulla condizione lavorativa e, quindi, sulle caratteristiche della maggiore fonte di sostentamento dell'intervistato.

Dopo aver definito la condizione lavorativa dell'individuo al momento dell'intervista, il questionario procede con quesiti relativi al lavoro svolto, per coloro che risultano occupati, o sull'ultimo lavoro svolto, per coloro che risultano disoccupati ma che sono stati occupati in passato. Inoltre, a coloro che si dichiarano impiegati con un lavoro alle dipendenze viene chiesto il reddito mensile lordo assieme ad alcune caratteristiche del contratto, mentre un'altra sezione specifica registra la condizione lavorativa dell'intervistato in tutti i dodici mesi del 2003 fino a giungere al momento dell'intervista, per poter verificare anche eventuali cambiamenti della condizione occupazionale.

Le sezioni del questionario individuale attraverso cui vengono rilevate le più importanti variabili di reddito sono le sezioni "Redditi da lavoro dipendente percepiti nel 2003", "Redditi da lavoro autonomo percepiti nel 2003" e "Pensioni e indennità, assegni o pensioni di invalidità, inabilità o per infortuni sul lavoro percepite nel 2003". Ciascuna di queste tre sezioni è preceduta da una domanda filtro che richiede all'intervistato se durante l'anno solare 2003 abbia percepito o meno il particolare tipo di reddito in questione; le domande di ogni sezione sono quindi sottoposte solo a coloro che dichiarano di aver percepito quel particolare tipo di reddito.

Nella sezione relativa ai lavoratori dipendenti vengono rilevate tutte le componenti di reddito da lavoro dipendente percepite durante il 2003, compresi i compensi aggiuntivi, le liquidazioni e gli assegni dello Stato. Per il lavoratori autonomi la sezione prevede quesiti relativi alle perdite e ai guadagni nel 2003, e agli eventuali assegni statali. Infine, la sezione dedicata a coloro che hanno percepito un reddito di tipo pensionistico è più articolata, in quanto tale tipologia di reddito comprende le pensioni sociali, le pensioni di vecchiaia, di reversibilità, di invalidità e altri eventuali assegni familiari e sussidi.

Infine, anche l'ultima sezione del questionario individuale prevede alcuni quesiti relativi a componenti di reddito individuali, quali per esempio le eventuali borse di studio o di lavoro, i versamenti di denaro ricevuti o effettuati al di fuori della propria famiglia, i guadagni da risparmi investiti in titoli, azioni, ecc. e dall'affitto di terreni o fabbricati.

Le numerose variabili di reddito rilevate a livello individuale possono formare variabili target EUROSTAT sia a livello individuale che familiare, come somma di una stessa variabile per tutti gli individui componenti la famiglia; per queste ultime si hanno le relazioni seguenti:

- Reddito da proprietà = Guadagni da affitti o subaffitti (var. familiare) + somma a livello familiare dei guadagni da terreni o fabbricati ;
- Profitti da capitale = somma a livello familiare dei guadagni da risparmi;

- Indennità per famiglia/figli = Assegno di sostegno per almeno 3 figli minori (var. familiare) * numero mesi di percezione + Assegno di maternità (var. familiare) + somma a livello familiare degli assegni familiari per disoccupati * numero mesi di percezione + somma a livello familiare degli assegni familiari per cassaintegrati * numero mesi di percezione + somma a livello familiare degli assegni familiari per lavoratori dipendenti * numero mesi di percezione + somma a livello familiare degli assegni familiari per lavoratori autonomi * numero mesi di percezione + somma a livello familiare degli assegni familiari per pensionati * numero mesi di percezione;

- Trasferimenti di denaro ricevuti = somma a livello familiare dei versamenti da persone fuori della famiglia;

- Trasferimenti di denaro versati = somma a livello familiare dei versamenti a persone fuori della famiglia;

- Tasse sul patrimonio = somma a livello familiare dell'Imposta Comunale sugli Immobili;

- Reddito dei minori di 15 anni = somma a livello familiare del reddito dei minori di 15 anni in famiglia;

- Pagamenti per tasse = somma a livello familiare dei pagamento da dichiarazione dei redditi - somma a livello familiare dei rimborsi da dichiarazione dei redditi;

La relazioni tra le variabili target individuali EUROSTAT e le singole variabili direttamente presenti nel questionario individuale EU-SILC sono invece le seguenti:

- Reddito da lavoro dipendente = retribuzione mensile netta * numero mesi di percezione + compensi aggiuntivi * numero mesi di percezione + altri compensi aggiuntivi + arretrati da lavoro;

- Contributi per pensioni private = Contributi versati per pensione integrativa * mesi di versamento;

- Guadagni o perdite da lavoro autonomo = reddito totale da lavoro autonomo;

- Pensione privata = Pensione integrativa * numero mesi di percezione;

- Benefici per i sopravvissuti = Pensione di reversibilità * numero mesi di percezione;

- Indennità per motivi di studio = Borsa di studio * numero mesi di percezione;
- Benefici per invalidità = Pensione di invalidità * numero mesi di percezione + Assegni di accompagnamento * numero mesi di percezione;
- Benefici di disoccupazione = Indennità disoccupazione * numero mesi di percezione + Cassa integrazione * numero mesi di percezione + Borsa lavoro * numero mesi di percezione + Liquidazioni da lavoro (per disoccupazione);
- Benefici per anzianità = Pensione di anzianità * numero mesi di percezione + Pensione sociale * numero mesi di percezione + Liquidazioni da lavoro (per pensione);

Anche da questa breve presentazione dei questionari dell'indagine sulle Condizioni di Vita si comprende come il contenuto informativo di questa indagine sia sicuramente molto ampio, soprattutto relativamente alle componenti di reddito sia familiari che individuali. E' chiaro dunque che qualsiasi problematica relativa ai dati provenienti dall'indagine, come le mancate risposte, non può non tenere in considerazione la complessa ed articolata struttura dei dati stessi.

4.4 I dati mancanti dell'indagine sulle Condizioni di Vita 2004

I risultati che vengono presentati nei prossimi paragrafi si riferiscono ai dati dell'indagine ISTAT sulle Condizioni di Vita 2004. Occorre precisare che tali dati sono quelli presenti nel file "standard" prodotto da ISTAT, e sono microdati anonimi a livello individuale e familiare, non grezzi ma già elaborati.

I dati risultavano infatti già imputati relativamente a tutte le mancate risposte a singoli item, sia livello familiare che individuale. Tuttavia, utilizzando gli *imputation factors* (paragrafo 4.1) è stato possibile, attraverso programmazioni "ad hoc", ricostruire parte dei dati originariamente mancanti.

In particolare, l'imputation factor era presente nel file di dati per le variabili target EUROSTAT presentate nel paragrafo precedente, tutte relative a dati di reddito. Proprio attraverso tali factors, e conoscendo la relazione tra ciascuna variabile target ed i singoli quesiti del questionario familiare e individuale è stato possibile ricostruire i dati mancanti per questi quesiti.

Come illustrato nel paragrafo precedente, il valore registrato in una generica variabile target EUROSTAT $y_{target\ i}$ riferita al reddito è calcolata come somma di j componenti di reddito rilevate con appositi quesiti del questionario; non tenendo in considerazione i fattori di complicazione come il livello di somministrazione dei quesiti del questionario (familiare o individuale), una possibile rappresentazione è la seguente:

$$(y_{target\ i\ registrata}) = (y_{item\ 1\ registrata}) + \ldots + (y_{item\ j\ registrata}).$$

Invece, per i valori originariamente rilevati al momento dell'intervista si ha:

$$(y_{target\ i\ rilevata}) = (y_{item\ 1\ rilevata}) + \ldots + (y_{item\ j\ rilevata}).$$

L'imputation factor è invece rappresentabile come:

$$IF(y_{target\ i}) = (y_{target\ i\ rilevata})/(y_{target\ i\ registrata}).$$

Quindi, se per esempio il valore registrato in una variabile target è 10000 euro ed il corrispondete valore dell'imputation factor è 0.8, dalle relazioni precedenti si ricava:

$$(y_{target\ i\ rilevata}) = 0.8 * 10000 = 8000 = (y_{item\ 1\ rilevata}) + \ldots + (y_{item\ j\ rilevata}).$$

In questo caso, allora, la procedura di ricostruzione dei valori mancanti prevede la "conservazione" dei valori $(y_{item\ t\ registrata})$, $t = 1, \ldots, j$, la cui somma risulta pari a 8000 euro, e la cancellazione dei restanti. Nel caso invece di imputation factor pari a 0 oppure ad 1, la procedura prevede la cancellazione rispettivamente di tutti oppure di nessun valore $y_{item\ k\ registrata}$, con $k = 1, \ldots, j$. Ovviamente tale procedura è risultata più complessa, ma ugualmente efficace, nel caso in cui la variabile target risultasse composta dalla somma a livello familiare di variabili rilevate a livello individuale.

Con tale procedimento è risultato possibile ricostruire i valori mancanti per i singoli quesiti del questionario relativi a valori di reddito; nel presente lavoro viene presa in considerazione soltanto questa tipologia di mancata risposta parziale; gli altri quesiti dei questionari vengono invece considerati come completamente osservati.

Le percentuali di mancate risposte per le variabili target e per i singoli quesiti dei questionari, ricostruite attraverso la procedura appena descritta, sono riportate nelle tabelle 4.1, 4.2 e 4.3[3].

[3]Bisogna specificare che esistono delle differenze tra le variabili delle tabelle 4.2 e 4.3 ed i quesiti presenti nei questionari individuali e familiari. In particolare, per la variabile

Come si vede, le percentuali di mancata risposta risultano molto variabili. In particolare, a livello individuale la percentuale più elevata di missing values si ha per la variabile "Borsa lavoro" (48.67%) mentre a livello familiare per "Contributi pubblici per affitto" (81.53%); tali variabili non riguardano tuttavia un numero molto elevato di osservazioni (rispettivamente pari a 113 e 157). E' da sottolineare che le percentuali di valori mancanti risultano particolarmente elevate, a livello individuale, per le variabili relative alla sezione "Altre informazioni relative al 2003" e "Redditi da lavoro autonomo".

Una particolare precisazione riguarda la variabile "Reddito complessivo" della sezione "Redditi da lavoro autonomo" della tabella 4.2. Tale variabile risultava già composta da tre quesiti del questionario e dai dati fiscali provenienti da fonte esterna, senza possibilità di risalire alla sua struttura originaria. Come già accennato nel capitolo 3, infatti, nella procedura di imputazione del reddito da lavoro autonomo l'ISTAT ha effettuato un'operazione di recupero delle informazioni da una fonte di dati alternativa (ISTAT, 2006), procedura resa possibile dai regolamenti EU-SILC, per poi ricorrere alla normale procedura di imputazione utilizzata per tutte le altre variabili di reddito solo nei casi di mancato *matching* con i dati fiscali (che risulta comunque piuttosto alto, pari al 21.21%, come si vede nella tabella 4.2).

4.5 Le caratteristiche dei dati

L'imputazione dei dati mancanti per le variabili di reddito presenta molte interessanti problematiche sia a livello teorico che applicativo. Nel presente lavoro vengono affrontati soltanto alcuni di questi aspetti, lasciando i restanti a futuri studi ed approfondimenti; alcune considerazioni in merito a tali sviluppi futuri sono presentate nell'ultimo paragrafo di questo capitolo.

Innanzitutto, l'approccio scelto per l'imputazione dei dati mancanti è il metodo delle regressioni sequenziali multivariate (paragrafo 2.2.2), implementato attraverso il modulo "IMPUTE" del pacchetto IVEware (Raghunathan et al., 1998). In particolare, l'interesse si è concentrato sull'implementazione di imputazioni multiple per gli item di reddito utilizzando la stessa

relativa al quesito sul reddito corrente da lavoro dipendente, rilevata per gli occupati al momento dell'intervista, non è risultato possibile risalire ai valori mancanti originari, mentre per la variabile relativa agli interessi sul mutuo si è scelto di non ricostruire i valori mancanti in quanto l'imputazione di tale variabile é quasi sempre basata su calcoli esatti di matematica finanziaria che utilizzano le altre informazioni disponibili (capitale prestato, tasso di interesse, durata, ecc.). Infine, non risultavano disponibili le risposte ai quesiti richiedenti la collocazione del reddito in fasce di valori, in caso di risposta "non so" alla richiesta di un valore preciso per alcune componenti di reddito, e al quesito sul reddito familiare totale nel questionario familiare.

Tabella 4.1: Variabili target EUROSTAT relative al reddito: percentuale di valori mancanti.

Variabile	% valori mancanti
A livello individuale	
Reddito da lavoro dipendente	8.49
Contributi per pensioni	9.59
Guadagni o perdite da lavoro autonomo	21.21
Pensione privata	22.47
Benefici per i sopravvissuti	6.26
Indennità per motivi di studio	21.73
Benefici per invalidità	6.65
Benefici per vecchiaia	3.51
Benefici di disoccupazione	17.13
A livello familiare	
Reddito da proprietà	9.56
Profitti da capitale	25.07
Indennità per famiglia/figli	11.28
Benefici per la casa	60.63
Esclusione sociale	12.69
Trasferimenti di denaro ricevuti	21.31
Trasferimenti di denaro versati	10.83
Reddito dei minori di 16 anni	25.68
Tasse sul patrimonio	17.95
Pagamenti per tasse	12.72

Tabella 4.2: Variabili di reddito del questionario individuale: risposte dovute, valori mancanti e percentuale di valori mancanti.

Variabile	Risposte dovute	Valori mancanti	% valori mancanti
Sezione redditi da lavoro dipendente			
Retribuzione mensile netta	18730	1684	8.99
Compensi aggiuntivi (mensili)	2009	140	6.97
Altri compensi aggiuntivi	8794	398	4.53
Arretrati da lavoro	617	93	15.07
Liquidazioni da lavoro	1277	178	13.94
Assegni familiari (mensili)	3785	511	13.50
Sezione redditi da lavoro autonomo			
Reddito complessivo	8907	1889	21.21
Assegni familiari (mensili)	189	57	30.16
Indennità per maternità	65	28	43.08
Sezione redditi da pensioni			
Pensione sociale (mensile)	750	69	9.20
Pensione di anzianità (mensile)	11812	476	4.03
Pensione di reversibilità (mensile)	4304	121	2.81
Pensione di invalidità (mensile)	4119	129	3.13
Assegni di accompagnamento (mensili)	1119	21	1.88
Assegni familiari (mensili)	1037	11	1.06
Pensione integrativa (mensile)	178	40	22.47
Sezione altre informazioni relative al 2003			
Indennità disoccupazione (mensili)	769	144	18.73
Assegni familiari per disoccupati (mensili)	116	13	11.21
Cassa integrazione (mensile)	232	62	26.72
Assegni familiari per cassaintegrati (mensili)	38	2	5.26
Borsa lavoro (mensile)	113	55	48.67
Borsa di studio (mensile)	359	78	21.73
Versamenti a persone fuori dalla famiglia	1160	122	10.52
Versamenti da persone fuori della famiglia	1298	277	21.34
Contributi versati per pensione integrativa (mensile)	4128	396	9.59
Guadagni da risparmi	19583	5093	26.01
Guadagni da terreni o fabbricati	1970	195	9.90
Imposta Comunale sugli Immobili	26677	4814	18.05
Rimborso da dichiarazione dei redditi	8410	620	7.37
Pagamento da dichiarazione dei redditi	4774	808	16.93

Tabella 4.3: Variabili di reddito del questionario familiare: risposte dovute, valori mancanti e percentuale di valori mancanti.

Variabile	Risposte dovute	Valori mancanti	% valori mancanti
Sezione casa e zona di abitazione			
Contributi pubblici per spese per la casa	150	68	45.33
Sezione famiglie in affitto			
Contributi pubblici per affitto	157	128	81.53
Sezione famiglie con casa di proprietà			
Contributi pubblici per interessi sul mutuo	101	57	56.44
Sezione situazione economica			
Reddito minimo vitale (mensile)	268	34	12.69
Assegno di sostegno per almeno 3 figli minori (mens.)	159	18	11.32
Assegno di maternità	174	10	5.75
Guadagni da affitti o subaffitti	222	10	4.50
Reddito dei minori di 15 anni in famiglia	39	10	25.64

metodologia e software attualmente impiegati da ISTAT per procedure di imputazione singola. L'interesse principale, in questa ottica, è stato verificare se ed in quale misura per l'indagine EU-SILC l'imputazione multipla, approccio non ancora utilizzato in Italia nell'ambito delle statistiche ufficiali (capitolo 3), potesse comportare una differenza notevole o meno rispetto all'utilizzo dell'imputazione singola per il calcolo dei valori medi dei redditi nella popolazione e del loro errore standard.

Il metodo di imputazione scelto si basa sull'ipotesi che i dati siano mancanti a caso (MAR, paragrafi 1.1.2 e 1.2.2). Come già ampiamente discusso (capitolo 3), questa ipotesi può risultare particolarmente restrittiva proprio nel caso di variabili di reddito. Tuttavia, dato il gran numero di variabili esplicative presenti nell'indagine sulle Condizioni di Vita, in questo capitolo viene presa in considerazione solo l'ipotesi MAR e viene presentata, relativamente ad un sottoinsieme di dati, una metodologia recentemente proposta che può servire per testare tale ipotesi.

4.5.1 Il *pattern* dei dati

Un primo elemento da tenere in considerazione nella procedura di imputazione è la struttura dei dati. Nel caso specifico dell'indagine sulle condizioni di vita, le variabili di reddito che presentano dati mancanti sono in totale 38, di cui 30 rilevate a livello individuale, 8 a livello familiare. Per analizzare il *pattern* delle mancate risposte consideriamo la matrice degli indicatori di mancata risposta M_{pij}, di dimensioni pari alla matrice dei dati, e i cui elementi assumono valore pari ad 1 se la variabile p relativa all'individuo i della fami-

glia j è mancante, 0 altrimenti (paragrafo 1.2.2). Ignorando per il momento la presenza di filtri per i quesiti relativi al reddito, la matrice M_{pij} può essere schematicamente rappresentata come in figura 4.2. Tale rappresentazione è basata sugli individui indipendentemente dalla famiglia di appartenenza: ciò significa che individui della stessa famiglia possono appartenere a *pattern* diversi.

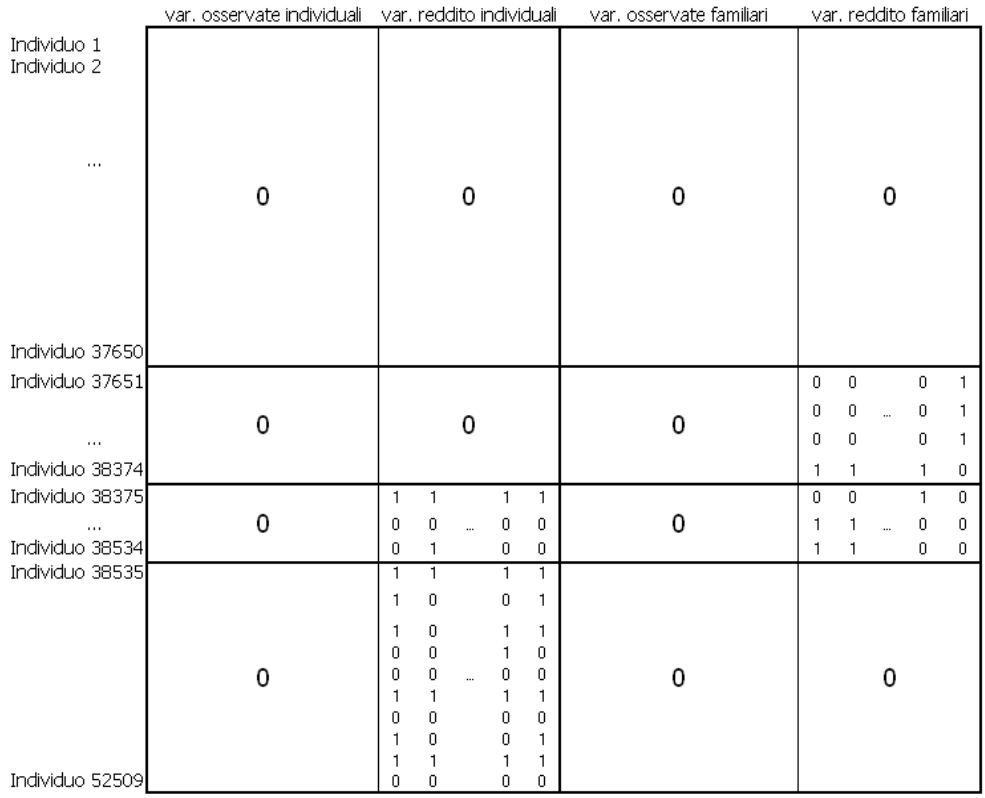

Figura 4.2: Rappresentazione schematica del *pattern* dei dati mancanti.

Come si vede dalla figura 4.2, sono possibili quattro tipologie di *pattern*. Il primo *pattern* riguarda 37650 individui per i quali non si ha nessun valore mancante; nel secondo *pattern* invece, per ciascuno dei 724 individui si ha almeno un valore mancante per una delle variabili di reddito familiari: ovviamente, individui appartenenti alla stessa famiglia avranno uguali valori M_{pij} per le variabili familiari. Nel terzo *pattern*, cui appartengono 160 individui, si osserva almeno un missing value per una delle variabili di reddito individuali e per una delle variabili di reddito familiari, mentre per i restanti individui (13975) si ha almeno un missing per le variabili di reddito individuali, mentre le altre sono tutte completamente osservate.

Il *pattern* dei dati mancanti risulta quindi piuttosto complesso, anche se sembra possibile concludere che i valori mancanti riguardano solo un percentuale limitata delle osservazioni. In ogni caso, la struttura dei dati evidenziata nella figura 4.2 dovrebbe servire da guida per l'implementazione di una procedura di imputazione che possieda "caratteristiche desiderabili".

Innanzitutto il metodo di imputazione deve essere implementato in modo che per individui appartenenti alla stessa famiglia vengano imputati valori uguali per le variabili di reddito familiari. Inoltre, relativamente alla scelta delle variabili esplicative, sarebbe auspicabile che il livello di reddito individuale, e quindi l'imputazione delle sue singoli componenti, tenesse in considerazione il livello di reddito familiare e viceversa, dal momento che, per questa indagine, esiste una diretta relazione tra le componenti di reddito familiari e individuali (paragrafo 4.3).

Il metodo di imputazione multipla che viene proposto in questo capitolo cerca di affrontare queste problematiche, tenendo in considerazione anche altri fattori di complicazione che caratterizzano le variabili di reddito.

4.5.2 Le variabili di reddito

Le variabili di reddito che presentano valori mancanti sono tutte misurate su scala continua. E' chiaro dunque che il vantaggio di utilizzare il metodo delle regressioni sequenziali multivariate risulta in questo caso attenuato, dal momento che si rende necessario l'impiego del solo modello di regressione lineare[4]. In pratica, tuttavia, l'utilizzo di tale metodo di imputazione attraverso il software IVEware è comunque particolarmente indicato in quanto consente di trattare due fattori di complicazione presenti nei dati.

In particolare, IVEware prevede la possibilità di adattare ciascun modello di regressione al corretto sottoinsieme di osservazioni quando nei dati sono presenti dei filtri, e di estrarre i valori mancanti da distribuzioni predittive troncate quando per i valori stessi esistano dei limiti inferiori e superiori da rispettare[5].

Per quanto riguarda il primo dei due fattori di complicazione, ciascuna delle variabili di reddito è preceduta da un filtro, ovvero da un quesito relativo all'acquisizione o meno del reddito stesso: il modello di regressione per imputare i valori mancanti andrà dunque adattato solamente alle osserva-

[4]Come già sottolineato nel paragrafo 2.2.2, senza fattori di complicazione particolari, se tutte le variabili da imputare sono continue e ciascun modello di regressione condizionato è un modello lineare normale con varianza costante, l'algoritmo SRMI converge ad una distribuzione predittiva congiunta normale multivariata, con distribuzioni a priori improprie per media e varianza (Raghunathan et al., 2001).

[5]I comandi cui si fa riferimento sono *restrict* e *bounds* (Raghunathan et al., 1998).

zioni filtrate. Una possibile rappresentazione schematica dei filtri presenti nei dati è nella figura 4.3. Tale figura fa riferimento al *pattern* dei dati, già rappresentato nella figura 4.2; in questo caso però non si fa distinzione tra variabili rilevate a livello individuale o familiare. Si vede allora che una data variabile p potrà essere osservata o mancante per l'individuo i appartenente alla famiglia j ($M_{pij} = 0$ oppure $M_{pij} = 1$) solo se il relativo quesito è stato effettivamente somministrato; in caso contrario si hanno degli *skip patterns*, ovvero dei "salti" nel *pattern* dei dati, rappresentati in figura dalle aree ombreggiate.

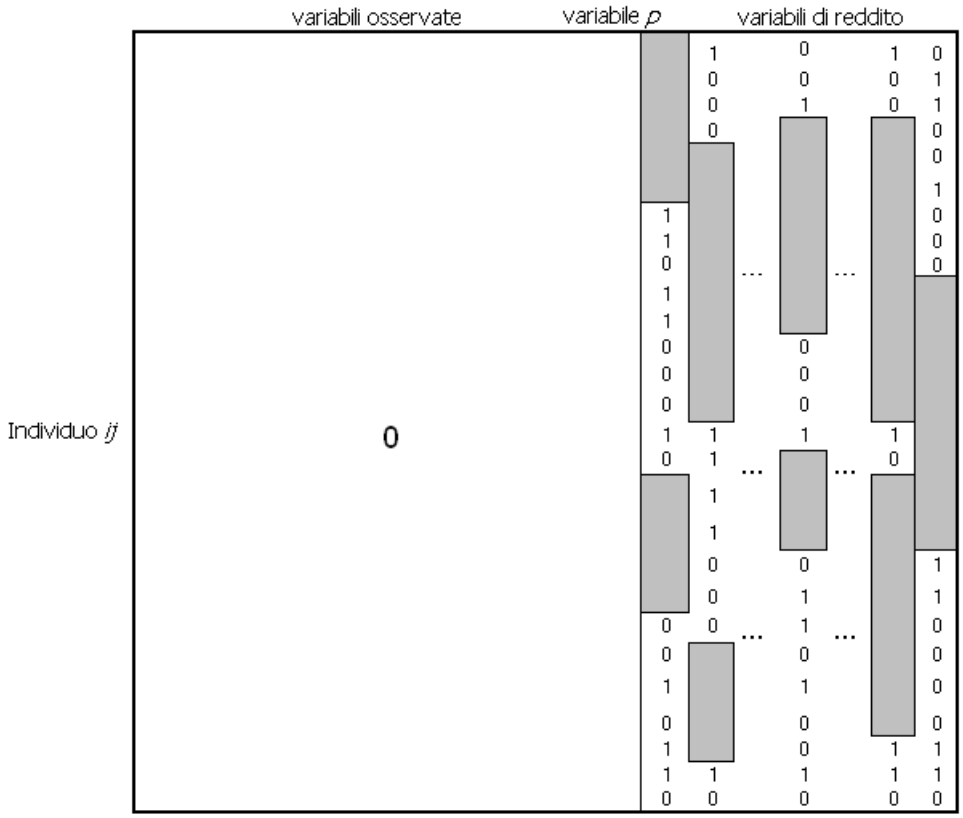

Figura 4.3: Rappresentazione schematica dei filtri presenti nel *pattern* dei dati.

La presenza del filtro comporta anche il fatto che le variabili di reddito assumano valori strettamente superiori a zero. Le uniche eccezioni sono costituite dalla variabile "Guadagni da risparmi", per la quale non esiste una vera e propria domanda filtro e che quindi può assumere anche il valore zero, e dalla variabile "Reddito da lavoro autonomo" che, per definizione, essendo

calcolata come differenza tra i guadagni e le perdite da lavoro autonomo può assumere non solo valore zero ma anche valori negativi.

Per quanto riguarda invece il valore massimo delle variabili, il metodo delle regressioni sequenziali può portare ad imputare valori talvolta molto elevati rispetto al valore massimo osservato[6]. Invece, specie per alcune variabili di reddito, la logica suggerisce che esiste un limite massimo oltre il quale il valore non può essere considerato plausibile. Poichè non è risultato possibile individuare per ciascuna variabile tale limite superiore da fonte esterna, sono stati calcolati dei limiti basandosi sui valori osservati.

In particolare, per ciascuna delle variabili di reddito è stato calcolato il limite massimo oltre il quale le osservazioni sono da considerarsi anomale secondo la metodologia Hidiroglou-Berthelot per variabili positive (Hunt et al., 2003) utilizzata da ISTAT nella fase di *editing* dei dati. Tale procedura prevede i seguenti calcoli:

$$Y_{HB} = \begin{cases} \frac{Y - Me_y}{Y} & 0 < Y < Me_y \\ \frac{Me_y - Y}{Me_y} & Y \geq Me_y \end{cases}$$

$$\begin{aligned} Y_{min\,HB}(k) &= \frac{Q1_y Me_y}{Q1_y + k(Me_y - Q1_y)} \qquad 0 < Y < Me_y \\ Y_{max\,HB}(k) &= Me_y + k(Q3_y - Me_y) \qquad Y \geq Me_y \end{aligned}$$

dove $Q1_y$, Me_y e $Q3_y$ sono rispettivamente il primo quartile, la mediana ed il terzo quartile osservati per la variabile Y, mentre k è una costante che determina l'ampiezza dell'intervallo di accettazione dei valori. I valori soglia massimi, calcolati secondo tale procedura per $k = 20$[7], sono riportati per tutte le variabili assieme al primo quartile, alla mediana e al terzo quartile nelle tabelle 4.4 e 4.5; tali valori si sono tradotti in vincoli per il massimo delle variabili nella fase di imputazione dei dati[8]. Questa procedura permette di evitare l'imputazione di valori anomali, ed evita inoltre che gli *outliers* di una variabile vengano utilizzati per l'imputazione di un'altra dal metodo delle regressioni sequenziali.

[6] Autori come Abayomi et al. (2007) affermano che i valori anomali imputati per alcune variabili potrebbero non essere tali ma dipendere da un meccanismo di risposta non completamente a caso; dato però che il meccanismo non è solitamente noto, è consuetudine introdurre dei limiti superiori, posti solitamente pari al massimo valore osservato.

[7] Tale valore viene suggerito da Hunt et al. (2003) per individuare tutte quelle osservazioni che richiedono una particolare attenzione nella fase di editing dei dati.

[8] Nel caso della variabile "Reddito da lavoro autonomo" prima della procedura è stata effettuata una traslazione dei valori, mentre per la variabile "Guadagni da risparmi" non è stato considerato il valore 0. Inoltre, per alcune variabili il limite superiore utilizzato è stato aumentato, in quanto nel dataset erano presenti dei valori osservati, non eliminati da ISTAT nella fase di *editing* dei dati, superiori alla soglia HB.

Tabella 4.4: Variabili di reddito del questionario individuale: primo quartile, mediana e terzo quartile dei valori osservati, e soglia massima secondo la procedura Hidiroglou-Berthelot.

Variabile	Primo quartile	Mediana	Terzo quartile	Soglia massima
Sezione redditi da lavoro dipendente				
Retribuzione mensile netta	850	1080	1300	5480
Compensi aggiuntivi	70	120	300	3720
Altri compensi aggiuntivi	800	1060	1800	15860
Arretrati da lavoro	300	590	1000	8790
Liquidazioni da lavoro	500	1400	3950	52400
Assegni familiari (mensili)	25	60	120	1260
Reddito complessivo	5954	12042	20437	179952
Assegni familiari (mensili)	28	68	140	1518
Indennità per maternità	1600	2000	2820	18400
Sezione redditi da pensioni				
Pensione sociale (mensile)	229	292	378	2005
Pensione di anzianità (mensile)	455	764	1108	7633
Pensione di reversibilità (mensile)	281	449	647	4396
Pensione di invalidità (mensile)	224	402	526	2878
Assegni di accompagnamento (mensili)	431.19	431.19	431.19	431.19
Assegni familiari (mensili)	20	30	50	430
Pensione integrativa (mensile)	300	429	600	3849
Sezione altre informazioni relative al 2003				
Indennità disoccupazione (mensili)	290	470	700	5070
Assegni familiari per disoccupati (mensili)	50	114	200	1834
Cassa integrazione (mensile)	550	700	830	3300
Assegni familiari per cassaintegrati (mensili)	24.5	90	135	990
Borsa lavoro (mensile)	250	465	620	3565
Borsa di studio (mensile)	110	200	650	9200
Versamenti a presone fuori dalla famiglia	1000	2400	4500	44400
Versamenti da persone fuori della famiglia	1000	3000	5000	43000
Contributi per pensione integrativa (mensile)	83	103	150	1043
Guadagni da risparmi	78	162	512	7162
Guadagni da terreni o fabbricati	915	2400	4800	50400
Imposta Comunale sugli Immobili	86	159	300	2979
Rimborso da dichiarazione dei redditi	140	250	500	10250
Pagamento da dichiarazione dei redditi	150	400	1249	17380

Tabella 4.5: Variabili di reddito del questionario familiare: primo quartile, mediana e terzo quartile dei valori osservati, e soglia massima secondo la procedura Hidiroglou-Berthelot.

Variabile	Primo quartile	Mediana	Terzo quartile	Soglia massima
Sezione casa e zona di abitazione				
Contributi pubblici per spese per la casa	190	300	525	4800
Sezione famiglie in affitto				
Contributi pubblici per affitto	150	200	250	1200
Sezione famiglie con casa di proprietà				
Contributi pubblici per interessi sul mutuo	410	688	1195	10828
Sezione situazione economica				
Reddito minimo vitale	280	600	1100	10600
Assegno di sostegno per almeno 3 figli minori	110	120	248	2680
Assegno di maternità	1000	1200	1340	4000
Guadagni da affitti o subaffitti	1000	1800	3438	34560
Reddito dei minori di 15 anni in famiglia	1000	2000	3600	34000

In pratica l'introduzione del limite massimo e minimo per le variabili nella procedura di imputazione corrisponde ad utilizzare dei modelli di regressione lineari *troncati*.Estrarre i valori dei parametri direttamente dalla loro distribuzione a posteriori con verosimiglianza normale troncata può essere complicato, ma risulta in generale più semplice per un dato valore dei parametri. IVEware utilizza l'algoritmo SIR (Sampling Importance Resampling, paragrafo 2.2.3) per estrarre i valori dei parametri dalla loro distribuzione a posteriori effettiva (Raghunathan et al., 2001).

In particolare, alcuni valori di prova dei parametri vengono estratti dalla distribuzione a posteriori senza i limiti; poi, ad ogni valore di prova viene associato un *importance ratio*, dato dal rapporto della densità a posteriori troncata rispetto a quella non troncata, valutate entrambe nel valore estratto. Alla fine si estraggono i valori del parametro con probabilità proporzionali agli *importance ratios*, secondo la procedura già descritta nel paragrafo 2.2.3.

Infine, sempre per quanto riguarda le variabili di reddito, ci si è chiesti quale trasformazione dei loro valori potesse rendere più plausibile l'ipotesi di normalità dei modelli di regressione utilizzati. L'implementazione di procedure di imputazione multipla per variabili continue non normalmente distribuite è un argomento che sta recentemente ricevendo sempre più attenzione in letteratura; le distribuzioni considerate sono, per esempio, la Weibull, la Beta e la famiglia di trasformazioni *gh* di Tukey (Demirtas and Hedeker, 2007; He and Raghunathan, 2006).

Per semplicità, nella presente applicazione sono state considerate solamente le trasformazioni appartenenti alla famiglia delle *power trasformations*

(Hoaglin et al., 1983) attraverso il calcolo, per ciascuna variabile, del valore λ che rende la distribuzione della variabile più vicina a quella normale. Poichè il valore λ risultava prossimo a 0 per la grande maggioranza delle variabili, in pratica la trasformazione scelta è stata, per tutte le variabili, quella logaritmica. E' da sottolineare che, nonostante tale scelta rappresenti un'approssimazione specialmente per alcune variabili, l'utilizzo dei limiti massimi (tabelle 4.4 e 4.5) ha evitato l'imputazione di valori eccessivamente elevati; come indicato da He and Raghunathan (2006), infatti, l'utilizzo della trasformazione logaritmica può portare talvolta ad imputare valori troppo grandi.

Nelle figure 4.4-4.11 sono rappresentati gli istogrammi dei valori prima e dopo la trasformazione logaritmica per alcune delle variabili di reddito, ed in particolare per le variabili rispettivamente con il maggior ed il minor numero di osservazioni nelle quattro sezioni del questionario individuale contenenti questo tipo di variabili (Redditi da lavoro dipendente, Redditi da lavoro autonomo, Redditi da pensioni e Altre informazioni relative al 2003). Come si vede, l'approssimazione alla distribuzione risulta in generale buona, specialmente per le variabili con un maggior numero di osservazioni[9].

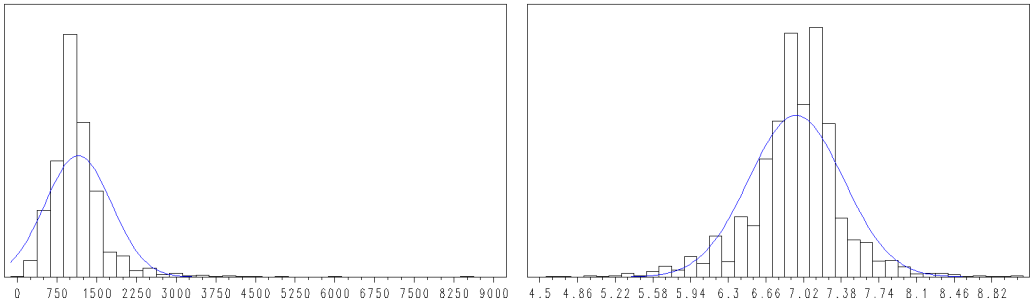

Figura 4.4: Istogramma dei valori osservati e dei valori in scala logaritmica per la variabile "Retribuzione mensile netta" (N=17046).

Le variabili "Assegni di accompagnamento per pensionati" e "Contributi versati per pensione integrativa" presentavano invece una distribuzione particolare. In questo caso infatti nell'istogramma dei valori osservati (figure 4.12 e 4.13) si evidenziava la concentrazione di più del 20% delle osservazioni su due valori singoli, pari rispettivamente a 431.19 euro per gli assegni di accompagnamento e 100 euro per i contributi per pensione integrativa. Per

[9]Come già detto, per trasformare in scala logaritmica la variabile "Reddito complessivo da lavoro autonomo", che assumeva anche valori negativi, si è effettuata una traslazione dei valori stessi.

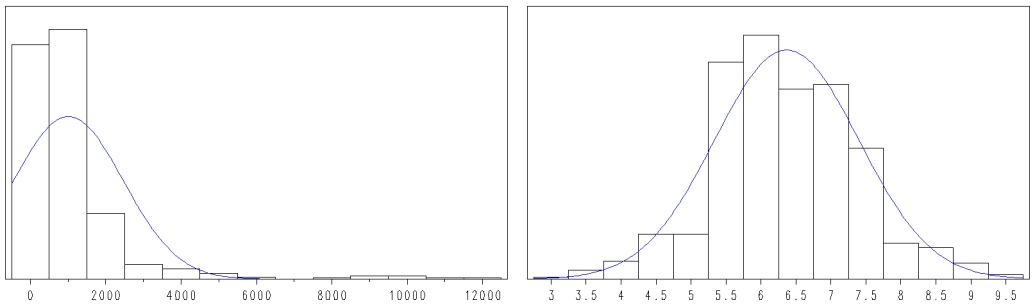

Figura 4.5: Istogramma dei valori osservati e dei valori in scala logaritmica per la variabile "Arretrati da lavoro" (N=524).

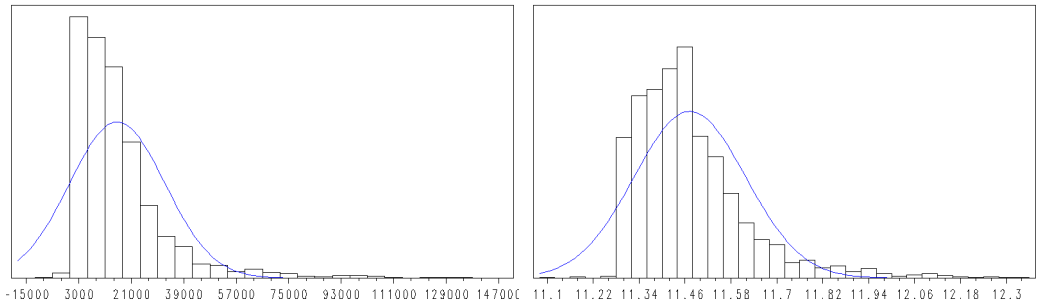

Figura 4.6: Istogramma dei valori osservati e dei valori in scala logaritmica per la variabile "Reddito complessivo da lavoro autonomo" (N=7018).

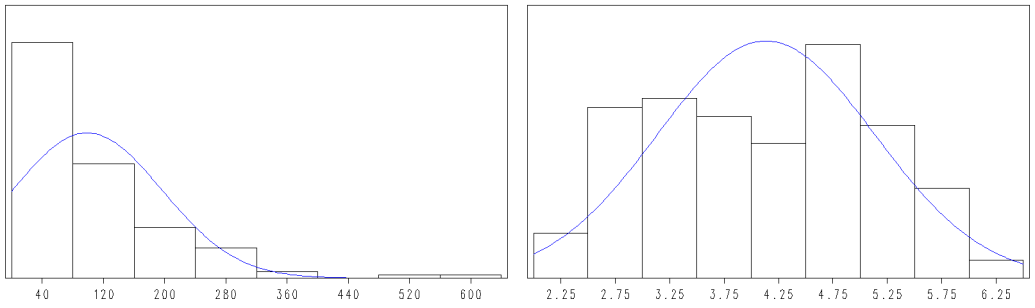

Figura 4.7: Istogramma dei valori osservati e dei valori in scala logaritmica per la variabile "Assegni familiari per lavoratori autonomi" (N=132).

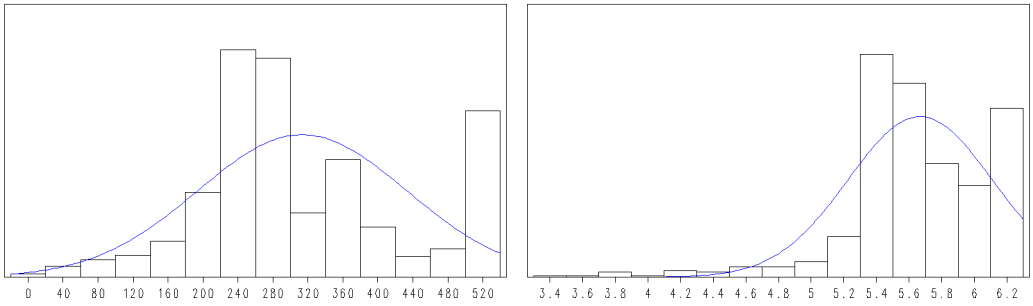

Figura 4.8: Istogramma dei valori osservati e dei valori in scala logaritmica per la variabile "Pensione sociale" (N=681).

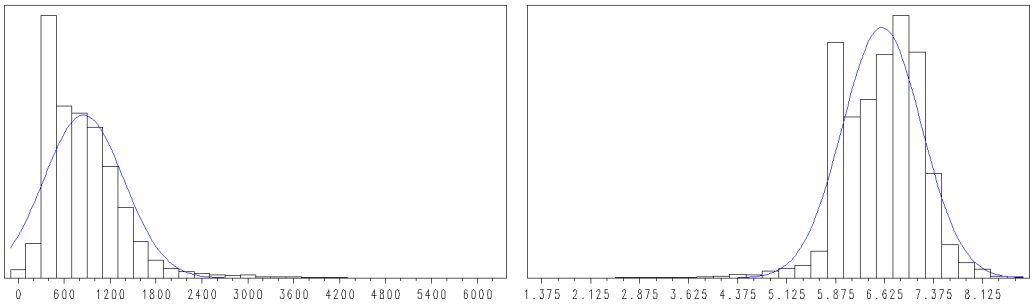

Figura 4.9: Istogramma dei valori osservati e dei valori in scala logaritmica per la variabile "Pensione di anzianità" (N=11336).

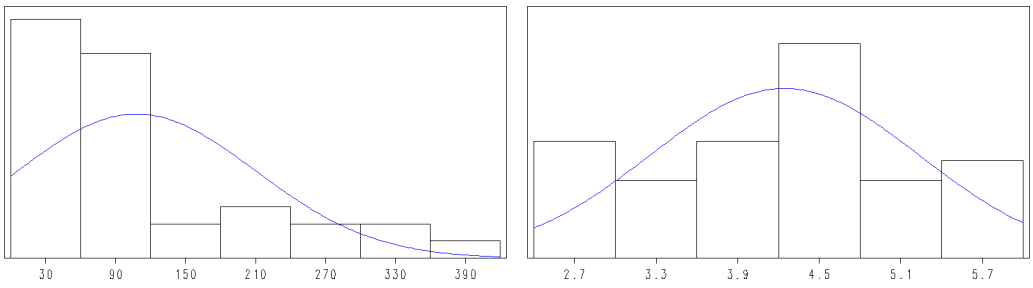

Figura 4.10: Istogramma dei valori osservati e dei valori in scala logaritmica per la variabile "Assegni familiari per cassaintegrati" (N=36).

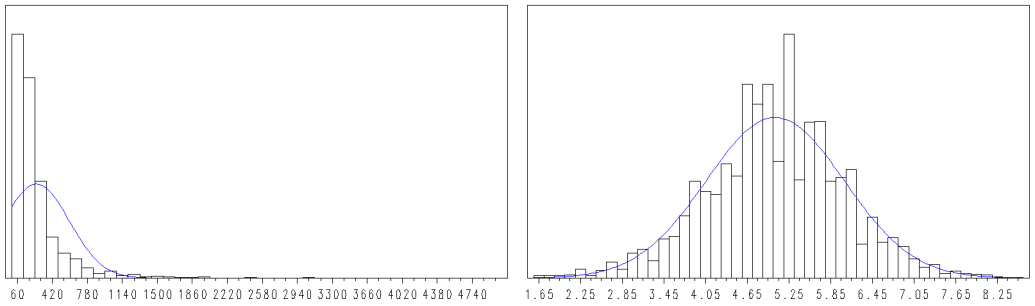

Figura 4.11: Istogramma dei valori osservati e dei valori in scala logaritmica per la variabile "Imposta Comunale sugli Immobili" (N=21863).

queste variabili la trasformazione dei valori ha previsto l'assegnazione del valore 0 alle osservazioni concentrate nel valore "speciale", e la trasformazione in scala logaritmica di tutti gli altri valori. Questo perchè l'imputazione con le regressioni sequenziali di queste variabili non utilizza una semplice regressione lineare, ma una procedura in due passi: prima si imputa l'appartenenza o meno alla categoria "speciale", posta pari al valore 0, attraverso un modello di regressione logistica e poi, in caso di non appartenenza a tale categoria, il valore mancante viene imputato attraverso un modello di regressione lineare per il logaritmo come avviene per tutte le altre variabili. Indicando con Y_i le osservazioni per la variabile con valori mancanti "concentrati" nel valore 0 e con X_i i valori dello covariate, i passi dell'algoritmo sono in questo caso i seguenti:

$$\begin{aligned} logit(P[Y_i > 0|\boldsymbol{X}_i, \boldsymbol{\beta}^{(1)}]) &= \boldsymbol{X}_i^{(1)'}\boldsymbol{\beta}^{(1)} \ [passo1] \\ P[ln(Y_i)|Y_i > 0, \boldsymbol{X}_i, \boldsymbol{\beta}^{(2)}, \sigma^2] &= N[(\boldsymbol{X}_i^{(2)'})\boldsymbol{\beta}^{(2)}, \sigma^2] \ [passo2] \end{aligned}$$

con distribuzione a priori $P(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \sigma^2) \propto 1/\sigma^2$. Per l'estrazione del parametro $\boldsymbol{\beta}^{(1)}$ il software IVEware utilizza l'approssimazione alla distribuzione normale per grandi campioni; tale approssimazione può essere evitata ricorrendo, anche in questo caso, all'algoritmo SIR (Raghunathan et al., 2001).

Questa procedura è stata utilizzata anche per la variabile "Guadagni da risparmi", trattando come valore "speciale" il valore zero.

Infine, gli istogrammi dei valori osservati e trasformati attraverso la funzione logaritmo per le variabili di reddito rilevate a livello familiare e rispettivamente con il numero maggiore e minore di osservazioni sono nelle figure 4.14 e 4.15.

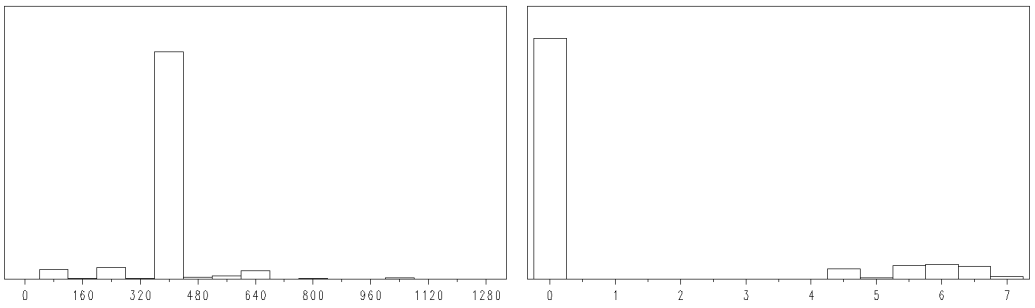

Figura 4.12: Istogramma dei valori osservati e dei valori in scala logaritmica per la variabile “Assegni di accompagnamento per pensionati” (N=1098).

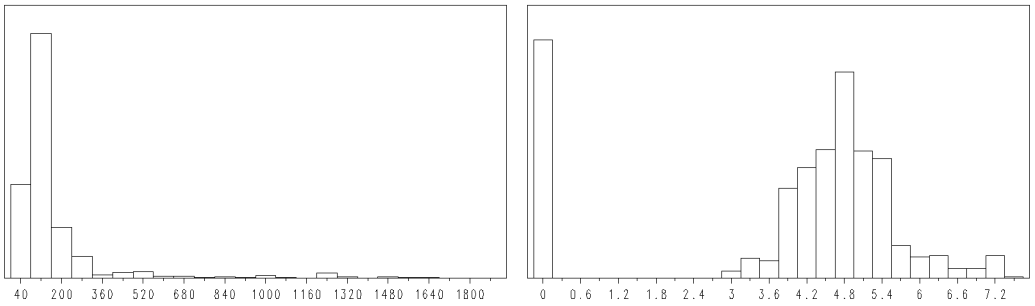

Figura 4.13: Istogramma dei valori osservati e dei valori in scala logaritmica per la variabile “Contributi versati per pensione integrativa” (N=3732).

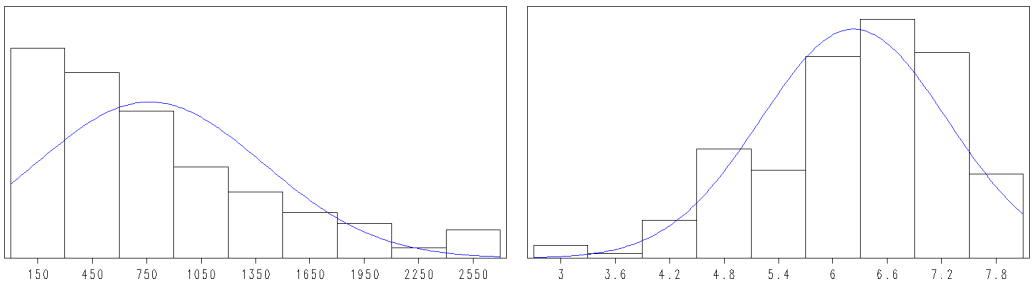

Figura 4.14: Istogramma dei valori osservati e dei valori in scala logaritmica per la variabile “Reddito minimo vitale” (N=234).

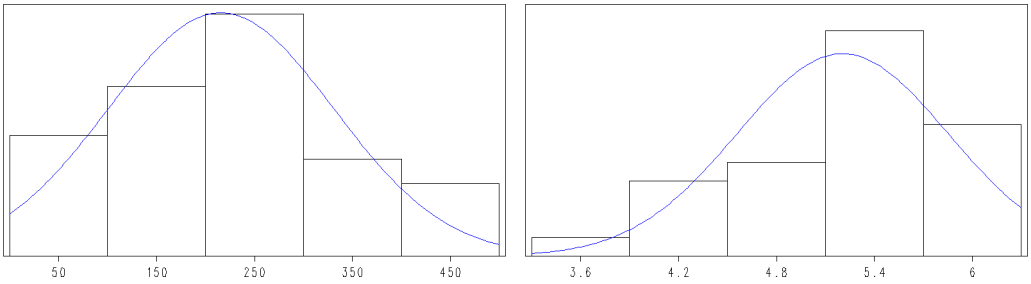

Figura 4.15: Istogramma dei valori osservati e dei valori in scala logaritmica per la variabile "Contributi pubblici per l'affitto" (N=29).

4.5.3 Le variabili osservate

Come già sottolineato, i questionari dell'indagine sulle Condizioni di Vita consentono la raccolta di molte informazioni, a livello individuale e familiare, che possono essere utilizzate come variabili esplicative nei modelli di regressione del processo di imputazione. Dato che il metodo di imputazione scelto ipotizza dati MAR, la scelta più ovvia è stata quella di utilizzare nelle regressioni un gran numero di predittori, proprio per rendere questa ipotesi più verosimile (paragrafo 1.2.2). Quando il numero delle variabili utilizzabili è molto alto, tuttavia, è consigliabile effettuare una selezione delle variabili stesse. L'introduzione di un numero di variabili esplicative troppo elevato nei modelli di regressione può infatti creare problemi di collinearità ed è, in generale, non necessario: l'incremento della varianza spiegata nella regressione lineare è solitamente trascurabile, per esempio, quando i migliori 15-20 predittori sono già stati inclusi nel modello (Van Buuren et al., 1999).

In particolare, la scelta delle variabili esplicative da utilizzare nei modelli di regressione utilizzati per imputare dovrebbe avvenire sulla base di una pluralità di elementi (paragrafo 2.1.2). Con specifico riferimento alle variabili di reddito del questionario, le variabili esplicative considerate appartengono ai seguenti gruppi:

- variabili del modello di analisi finale: a livello individuale sono state considerate le caratteristiche secondo le quali viene solitamente classificato il reddito, ovvero sesso, età, livello di istruzione, condizione e posizione lavorativa, ripartizione geografica; a livello familiare, invece, numero di componenti, tipo di abitazione, titolo di godimento dell'abitazione, presenza di minori di 15 anni, ripartizione geografica;

- variabili del modello di mancata risposta: poichè nella regressione per imputare una singola componente di reddito gli altri redditi sono au-

tomaticamente inclusi tra i predittori per la natura multivariata del metodo di imputazione, sono stati considerati, in particolare, gli indicatori relativi al disegno di campionamento, ovvero a strati, clusters e pesi campionari come suggerito in Rubin (1996);

- variabili che spiegano in modo considerevole la varianza delle variabili da imputare: per individuare questo tipo di variabili sono state calcolate le correlazioni, separatamente per le sezione dei questionari, tra le variabili di reddito e le altre informazioni osservate.

Alla fine, per l'imputazione delle variabili di reddito è stato selezionato un insieme di 14 variabili esplicative a livello familiare e di 22 a livello individuale. Le variabili utilizzate a livello familiare sono le seguenti:

- Caratteristiche territoriali: macro-ripartizione geografica (nord-est, nord-ovest, centro, sud, isole), almeno un problema nella zona di residenza (tra inquinamento, rumori e criminalità);

- Caratteristiche della famiglia e dell'abitazione: numero di componenti la famiglia, tipologia dell'abitazione, titolo di godimento dell'abitazione, superficie dell'abitazione, assenza di almeno un elemento tra gabinetto interno all'abitazione, vasca da bagno o doccia e acqua calda, presenza di almeno uno tra i problemi al tetto, umidità nei muri, scarsa luminosità, possesso di almeno un bene tra lavatrice, televisore a colori, frigorifero;

- Difficoltà economiche dichiarate dalla famiglia: ritardo nei pagamenti nel pagare l'affitto, il mutuo, le bollette o altri debiti, mancanza di soldi per comprare il cibo, i vestiti, pagare le spese per la scuola, pagare le spese per i trasporti o pagare le tasse, difficoltà nell'arrivare a fine mese con i redditi disponibili, gravosità delle spese per l'affitto o per il mutuo;

- Spese annuali per l'abitazione sostenute dalla famiglia: spese sostenute per tutte le utenze (gas, acqua, elettricità, ecc.).

A livello individuale invece:

- Caratteristiche territoriali: macro-ripartizione geografica (nord-est, nord-ovest, centro, sud, isole), problemi nella zona di residenza della famiglia (inquinamento, rumori, criminalità, ecc.);

- Caratteristiche della famiglia e dell'abitazione: numero di componenti la famiglia, tipologia dell'abitazione, titolo di godimento dell'abitazione;

- Difficoltà economiche dichiarate dalla famiglia: mancanza di soldi per comprare il cibo, i vestiti, pagare le spese per la scuola, pagare le spese per i trasporti o pagare le tasse, difficoltà nell'arrivare a fine mese con i redditi disponibili, gravosità delle spese per l'affitto o per il mutuo;

- Spese annuali per l'abitazione sostenute dalla famiglia: spese sostenute per tutte le utenze (gas, acqua, elettricità, ecc.).

- Caratteristiche individuali: sesso, età, stato civile, livello di istruzione, condizioni generali di salute, necessità di ricorrere al dentista o a visite specialistiche ed eventuale motivazione finanziaria per il mancato ricorso;

- Caratteristiche della professione: settore di attività, ore di lavoro settimanali, cambiamento di lavoro negli ultimi 12 mesi, posizione nella professione, età in cui si è iniziato a lavorare, numero di anni di lavoro, possesso di risparmi;

Tutte le variabili, con la sola eccezione delle spese sostenute dalla famiglia per il pagamento delle utenze, hanno carattere categorico. Per verificare la bontà delle variabili esplicative scelte e l'eventuale omissione di predittori importanti, sono state adattate delle imputazioni di prova lasciando selezionare le variabili esplicative da IVEware. Con IVEware esiste infatti la possibilità di realizzare la selezione *stepwise*, specificando il numero massimo di predittori che deve essere utilizzato per ogni variabile oppure indicando l'incremento marginale minimo per l'R^2 che ogni variabile deve superare per entrare nel modello[10]. L'utilizzo di queste opzioni ha consentito la verifica del maggior potere predittivo delle variabili già selezionate rispetto a quelle non incluse nell'analisi.

4.6 Imputazione multipla dei dati di reddito: un approccio iterativo

La scelta del particolare metodo di implementazione della procedura di imputazione multipla si è basata sullo studio del *pattern* dei dati e anche sulle caratteristiche delle variabili dell'indagine. Come detto, le variabili di reddito rilevate a livello individuale vengono aggregate per costruire variabili target a livello familiare; questo suggerisce la possibilità di derivare una misura del

[10]Le opzioni di IVEware per implementare i due metodi di selezione *stepwise* delle variabili sono *maxpred* e *minrsqd* (Raghunathan et al., 1998).

reddito familiare totale come somma dei redditi individuali di tutti i componenti della famiglia, dove ciascun reddito individuale può essere costruito come somma[11] delle singole voci di reddito del questionario individuale.

In particolare, indicando con $Y_{pij}, p = 1, ..., 30$ le variabili di reddito individuali e con $Y_{t.j}, t = 1, ..., 8$ le variabili di reddito familiari, si sono definiti i seguenti redditi:

- Reddito individuale totale= $\sum_{p=1}^{30} Y_{pij} = Y_{.ij}$;
- Reddito familiare totale = $\sum_{t=1}^{8} Y_{t.j} + \sum_{i=1}^{ncomp_j} Y_{.ij} = Y_{..j}^{fam} + Y_{..j}^{ind} = Y_{..j}^{tot}$.

Come già accennato, una considerazione di particolare interesse per l'imputazione delle singole variabili di reddito è il fatto che, per esempio, i redditi percepiti a livello familiare, come i contributi e gli assegni dello Stato, possono *dipendere* dall'ammontare dei redditi individuali dei componenti della famiglia e, viceversa, i redditi dei singoli individui possono *dipendere* dal percepimento di alcuni redditi familiari.

Queste riflessioni hanno suggerito l'implementazione di una procedura di imputazione multipla di tipo *iterativo*, in grado di sfruttare le dipendenze appena evidenziate e di gestire la struttura in due livelli del dataset. Il procedimento di imputazione lavora in due diversi step, uno familiare ed uno individuale, ed i risultati vengono poi ricomposti in modo da ricostruire la struttura dei dati.

Per imputare i dati mancanti a livello individuale erano possibili due tipi di approcci: un approccio gerarchico *esplicito*, attraverso regressioni multilivello, e un approccio gerarchico *implicito*, in cui le regressioni individuali non introducono un termine casuale di livello familiare.

L'approccio gerarchico esplicito suggerirebbe, per questa particolare applicazione, la modifica del metodo delle regressioni sequenziali implementato da IVEware per le variabili di reddito misurate a livello individuale con l'introduzione di corrispondenti regressioni gerarchiche sequenziali. Tale approccio, come già accennato nel paragrafo 2.3, è attualmente in corso di implementazione attraverso la creazione di un nuovo software (Yucel et al., 2006), che prevede l'introduzione, in ciascuna delle regressioni utilizzate per generare le imputazioni delle variabili individuali, di un residuo di livello familiare. Per esempio, indicizzando con i il livello individuale e con j quello

[11] Le somme tengono in considerazione il segno delle componenti di reddito; in particolare, nel computo del reddito individuale totale le quantità relative ai versamenti *a* persone fuori dalla famiglia, all'Imposta Comunale sugli Immobili, ai contributi versati per pensioni integrative e ai pagamenti supplementari per dichiarazione dei redditi vengono sommate alle altre componenti con segno negativo.

familiare, il modello di regressione per le variabili continue $\boldsymbol{Y}$ che condiziona rispetto a quelle osservate $\boldsymbol{X}$ sarebbe il seguente:

$$\begin{aligned} Y_{ij} &= X'_{ij}\beta + \gamma_i b_i + \epsilon_{ij} \\ b_i &\sim N(0, \Sigma_b) \\ \epsilon_{ij} &\sim N(0, \sigma^2). \end{aligned}$$

L'utilizzo di questo tipo di modelli consentirebbe di tenere in considerazione la correlazione intrafamiliare esistente tra le variabili durante il processo di imputazione dei dati.

Per quanto riguarda la presente applicazione è da sottolineare, tuttavia, che il numero medio di componenti per famiglia è piuttosto ridotto; infatti, i 52509 individui cui è stato somministrato il questionario individuale appartengono a 24204 famiglie, con una media di 2.15 individui per famiglia. Questo potrebbe implicare una sostanziale "irrilevanza" dell'esplicita introduzione della struttura familiare nel procedimento di imputazione, anche se tale aspetto andrà verificato attraverso futuri approfondimenti.

Tali riflessioni hanno suggerito per il momento l'utilizzo di un metodo di imputazione di tipo gerarchico *implicito*, che potesse tra l'altro essere realizzato attraverso il software IVEware. In particolare, è stato scelto un approccio *iterativo* in grado di imputare le variabili di reddito di livello individuale condizionando anche per il corrispondente reddito familiare e, viceversa, di imputare le variabili di reddito di livello familiare condizionando anche per la somma dei redditi degli individui appartenenti alla famiglia. In pratica questo approccio, suggerito in un recente articolo da Schenker et al. (2006) per l'imputazione di dati mancanti di reddito provenienti da un indagine di tipo complesso come l'EU-SILC, frammenta il processo di imputazione multipla per gestire la struttura gerarchica dei redditi. In particolare, la procedura di imputazione per ottenere un primo dataset completo è la seguente:

1. imputazione delle variabili di reddito Y_{pij} misurate a livello individuale utilizzando come esplicative le variabili individuali e familiari osservate;

2. calcolo della somma dei redditi individuali $Y^{ind}_{.j} = \sum_{i=1}^{ncomp_j} \sum_{p=1}^{30} Y_{pij}$ per ogni famiglia;

3. imputazione delle variabili di reddito $Y_{t.j}$ misurate a livello familiare utilizzando come esplicative le variabili familiari osservate ed il reddito $Y^{ind}_{.j}$;

4. calcolo del reddito familiare $Y^{fam}_{.j} = \sum_{t=1}^{8} Y_{t.j}$;

5. nuova imputazione delle variabili di reddito Y_{pij} misurate a livello individuale utilizzando come esplicative non solo le variabili individuali e familiari osservate ma anche il reddito familiare imputato al passo precedente;

6. ripetizione dei passi 2-5 per 5 cicli.

Per realizzare m imputazioni multiple dei dati mancanti, questa procedura ciclica è stata ripetuta in modo indipendente per m volte.

In pratica, ogni imputazione singola viene realizzata durante i passi 3 e 5 dell'ultimo ciclo della procedura, mentre i cicli precedenti servono per realizzare delle imputazioni provvisorie per calcolare le variabili di reddito $Y_{\cdot j}^{ind}$ e $Y_{\cdot j}^{fam}$. Dal punto di vista computazionale, quindi, le imputazioni a livello familiare e individuale avvengono in due *step* diversi: le imputazioni individuali vengono realizzate adattando le regressioni su un dataset in cui si ha una riga per ogni individuo, mentre per le imputazioni familiari le regressioni vengono implementate in un dataset con una sola osservazione per ogni famiglia.

Le imputazioni multiple delle variabili di reddito dell'indagine sono state realizzate applicando la procedura iterativa appena presentata per 10 volte, ottenendo 10 dataset completi. Le trasformazioni, i limiti e gli eventuali valori "speciali" considerati per le variabili di reddito oggetto di imputazione sono quelli già presentati nel paragrafo 4.5.2.

Per quanto riguarda le variabili esplicative considerate nelle regressioni, oltre a quelle già precedentemente selezionate e verificate attraverso le specifiche opzioni di IVEware (paragrafo 4.5.3), sono state inserite le variabili di reddito costruite nei vari passi della procedura in più cicli (paragrafo 4.6). Inoltre, nel caso delle imputazioni delle variabili di reddito misurate a livello familiare, è stata inserita come variabile esplicativa non solo la somma dei redditi individuali, $Y_{\cdot j}^{ind}$, ma anche il sesso e il tipo di lavoro del maggior percettore in famiglia (eseguendo l'ordinamento dei dati all'interno di ogni famiglia secondo tale criterio ad ogni passo della procedura).

Infine, prima di accettare le imputazioni ottenute attraverso il metodo e le scelte appena descritte, si è ritenuto opportuno eseguire alcune verifiche sull'effettiva convergenza dell'algoritmo di imputazione. In particolare, le singole procedure di imputazione a livello familiare ed individuale sono state ripetute con molte iterazioni per 20 volte, utilizzando ogni volta un diverso *random seed*. In questo modo si è verificato empiricamente che nelle 20 imputazioni l'algoritmo non causasse l'estrazione di valori imputati estranei ai valori osservati (Raghunathan et al., 2001). Quest'ultimo di tipo di errore è stato verificato attraverso il calcolo di alcune statistiche di interesse per le variabili imputate, ma anche attraverso una comparazione grafica

della distribuzione dei valori imputati e di quelli osservati per ciascuna delle variabili. Tali confronti grafici, sebbene possano fornire utili indicazioni, andrebbero tuttavia condotti condizionando per le variabili osservate, come viene illustrato nel paragrafo 4.7 di questo capitolo.

Le 10 imputazioni multiple finali sono state realizzate con numero di iterazioni pari a 10. Il numero di imputazioni multiple scelte, come verrà meglio evidenziato nel prossimo paragrafo, ha consentito la verifica della bassa varianza *between* delle stime per quasi tutte le variabili imputate.

4.6.1 L'analisi dei dataset imputati

Dopo aver realizzato l'imputazione dei valori mancanti, in ciascuno dei 10 dataset completati sono state calcolate le stime pesate ed i relativi errori standard (*standard errors*, s.e.) per tutte le varabili di reddito, sia individuali che familiari. I pesi utilizzati per calcolare tali stime tengono conto della stratificazione e clusterizzazione delle osservazioni, e correggono per le mancate risposte totali (paragrafo 4.2); gli standard errors sono stati calcolati attraverso le usuali procedure di linearizzazione utilizzate dai più comuni software nel caso di disegni campionari complessi (Woodruff, 1971; SAS, 1999).

Le $m = 10$ stime delle medie e degli standard errors sono poi state combinate attraverso le regole di Rubin (paragrafo 2.1.1), ottenendo i risultati nelle tabelle 4.6 e 4.8. Per poter confrontare le stime con quelle derivanti dalla *available case analysis*, gli stessi risultati sono stati calcolati ignorando, per ciascuna variabile, le osservazioni con valori missing (tabelle 4.7 e 4.9).

Confrontando le medie pesate calcolate prima e dopo l'imputazione multipla dei valori mancanti si nota come soltanto alcuni valori risultano modificati in modo rilevante. Ciò è in linea, innanzitutto, con i diversi tassi di mancata risposta delle variabili di reddito, già precedentemente riportati nelle tabelle 4.2 e 4.3, e anche con il corrispondente numero di osservazioni da cui è stata ricavata ciascuna delle stime pesate, indicato nelle tabelle 4.6-4.9.

Si nota infatti che le differenze maggiori per le medie pesate si osservano proprio per alcune delle variabili con tasso di mancata risposta o numero di osservazioni mancanti maggiore; è il caso delle variabili individuali "Borsa lavoro", "Borsa di studio", "Versamenti a/da persone fuori dalla famiglia" e "Guadagno da risparmi", e delle variabili familiari relative ai contributi pubblici per spese, affitto e mutuo, e "Redditi dei minori di 15 anni in famiglia". Per quasi tutte le altre variabili, invece, la stima puntuale della media prima e dopo il procedimento di imputazione risulta molto simile; per queste variabili potrebbe essere particolarmente interessante andare a svolgere analisi di

Tabella 4.6: Variabili di reddito individuali: numero di osservazioni, medie pesate e relativi standard errors con imputazione multipla dei valori mancanti.

Variabile	Osserv. (N)	Media	Standard Error
Sezione redditi da lavoro dipendente			
Retribuzione mensile netta	18730	1147	5.28
Compensi aggiuntivi	2009	351	13.31
Altri compensi aggiuntivi	8794	1917	65.93
Arretrati da lavoro	617	998	42.08
Liquidazioni da lavoro	1277	4232	271.40
Assegni familiari (mensili)	3785	93	1.92
Sezione redditi da lavoro autonomo			
Reddito complessivo	8907	16776	228.78
Assegni familiari (mensili)	189	156	11.46
Indennità per maternità	65	2645	209.15
Sezione redditi da pensioni			
Pensione sociale (mensile)	750	321	4.30
Pensione di anzianità (mensile)	11812	867	4.83
Pensione di reversibilità (mensile)	4304	507	4.47
Pensione di invalidità (mensile)	4119	436	5.11
Assegni di accompagnamento (mensili)	1119	418	2.94
Assegni familiari (mensili)	1037	55	2.73
Pensione integrativa (mensile)	178	513	13.47
Sezione altre informazioni relative al 2003			
Indennità disoccupazione (mensili)	769	517	21.28
Assegni familiari per disoccupati (mensili)	116	152	9.19
Cassa integrazione (mensile)	232	659	19.20
Assegni familiari per cassaintegrati (mensili)	38	97	2.90
Borsa lavoro (mensile)	113	599	40.35
Borsa di studio (mensile)	359	548	37.93
Versamenti a persone fuori dalla famiglia	1160	3211	96.76
Versamenti da persone fuori della famiglia	1298	3957	110.57
Contributi per pensione integrativa (mensile)	4128	164	3.78
Guadagni da risparmi	19583	583	13.44
Guadagni da terreni o fabbricati	1970	4257	129.76
Imposta Comunale sugli Immobili	26677	251	2.83
Rimborso da dichiarazione dei redditi	8410	447	7.13
Pagamento da dichiarazione dei redditi	4774	1282	38.17
Reddito individuale totale	52509	11948	67.25
Reddito totale dipendenti	18730	14494	86.34
Reddito totale pensionati	17334	11117	53.12

Tabella 4.7: Variabili di reddito individuali: numero di osservazioni, medie pesate e relativi standard errors senza imputazione dei valori mancanti.

Variabile	Osserv. (N)	Media	Standard Error
Sezione redditi da lavoro dipendente			
Retribuzione mensile netta	17046	1149	5.61
Compensi aggiuntivi	1869	347	13.74
Altri compensi aggiuntivi	8396	1934	68.93
Arretrati da lavoro	524	945	41.48
Liquidazioni da lavoro	1099	4099	279.39
Assegni familiari (mensili)	3274	94	2.05
Sezione redditi da lavoro autonomo			
Reddito complessivo	7018	16676	263.85
Assegni familiari (mensili)	132	118	10.53
Indennità per maternità	37	2551	111.72
Sezione redditi da pensioni			
Pensione sociale (mensile)	681	319	4.39
Pensione di anzianità (mensile)	11336	865	4.83
Pensione di reversibilità (mensile)	4183	508	4.45
Pensione di invalidità (mensile)	3990	436	5.15
Assegni di accompagnamento (mensili)	1098	418	2.93
Assegni familiari (mensili)	1026	55	2.75
Pensione integrativa (mensile)	138	479	9.55
Sezione altre informazioni relative al 2003			
Indennità disoccupazione (mensili)	625	507	22.28
Assegni familiari per disoccupati (mensili)	103	149	9.36
Cassa integrazione (mensile)	170	659	13.61
Assegni familiari per cassaintegrati (mensili)	36	100	3.06
Borsa lavoro (mensile)	58	516	24.85
Borsa di studio (mensile)	281	441	9.63
Versamenti a persone fuori dalla famiglia	1038	3094	76.67
Versamenti da persone fuori della famiglia	1021	3844	107.60
Contributi per pensione integrativa (mensile)	3732	164	4.15
Guadagni da risparmi	14490	623	17.60
Guadagni da terreni o fabbricati	1775	4213	129.96
Imposta Comunale sugli Immobili	21863	257	3.27
Rimborso da dichiarazione dei redditi	7790	447	7.57
Pagamento da dichiarazione dei redditi	3966	1242	42.72
Reddito individuale totale	39130	10797	74.02
Reddito totale dipendenti	16791	14409	92.18
Reddito totale pensionati	16552	11106	53.06

Tabella 4.8: Variabili di reddito familiari: numero di osservazioni, medie pesate e relativi standard errors con imputazione multipla dei valori mancanti.

Variabile	Osserv. (N)	Media	Standard Error
Sezione casa e zona di abitazione			
Contributi pubblici per spese per la casa	150	871	75.16
Sezione famiglie in affitto			
Contributi pubblici per affitto	157	316	10.21
Sezione famiglie con casa di proprietà			
Contributi pubblici per interessi sul mutuo	101	1115	80.01
Sezione situazione economica			
Reddito minimo vitale (mensile)	268	875	33.44
Assegno di sostegno per almeno 3 figli minori	159	297	28.80
Assegno di maternità	174	1268	36.68
Guadagni da affitti o subaffitti	222	2661	136.11
Reddito dei minori di 15 anni in famiglia	39	3249	759.69
Reddito familiare totale	24204	25400	156.90

Tabella 4.9: Variabili di reddito familiari: numero di osservazioni, medie pesate e relativi standard errors senza imputazione dei valori mancanti.

Variabile	Osserv. (N)	Media	Standard Error
Sezione casa e zona di abitazione			
Contributi pubblici per spese per la casa	82	476	21.72
Sezione famiglie in affitto			
Contributi pubblici per affitto	29	205	11.82
Sezione famiglie con casa di proprietà			
Contributi pubblici per interessi sul mutuo	44	782	19.69
Sezione situazione economica			
Reddito minimo vitale (mensile)	234	802	27.14
Assegno di sostegno per almeno 3 figli minori	141	288	23.57
Assegno di maternità	164	1267	37.99
Guadagni da affitti o subaffitti	212	2625	144.59
Reddito dei minori di 15 anni in famiglia	29	2508	378.79
Reddito familiare totale	20042	20567	149.61

sensitività rispetto ad ipotesi non a caso (MNAR), oppure cercare di testare l'ipotesi MAR secondo la procedura presentata in seguito.

Le singole variabili di reddito misurate a livello individuale possono essere sommate tra loro per ottenere il reddito complessivo da lavoro dipendente, da pensione ed il reddito individuale totale; la differenza maggiore per la media pesata prima e dopo l'iputazione dei valori si osserva proprio per quest'ultima grandezza.

Per quanto riguarda invece il confronto tra gli standard errors delle stime, per facilitare tale comparazione è stato calcolato il rapporto tra gli standard errors delle medie pesate senza imputazioni e i corrispondenti standard errors ottenuti dopo l'applicazione della procedura di imputazione multipla. I risultati sono nella prima colonna delle tabelle 4.10 e 4.12.

In teoria gli standard errors delle stime senza imputazione dei valori mancanti dovrebbero essere maggiori di quelli ottenuti con le imputazioni; l'informazione aggiuntiva apportata delle imputazioni dovrebbe infatti tradursi in una maggiore precisione delle stime. Tuttavia, tale confronto va effettuato con cautela dal momento che le stime degli standard errors che si ottengono con la *available case analysis* possono essere distorte quando i dati non sono MCAR (Schenker et al., 2006). In questa applicazione il rapporto tra gli standard errors risulta quasi sempre molto vicino ad 1, tranne i casi in cui le stime puntali indicavano già una sostanziale differenza tra i due metodi di trattamento delle mancate risposte. E' questo il caso delle variabili "Indennità per maternità", "Borsa da lavoro", "Borsa di studio" e "Versamenti a/da persone fuori dalla famiglia". A livello familiare, in modo simile, il rapporto tra lo standard error senza imputazione dei valori mancanti e con imputazione multipla risulta lontano da 1 per quelle variabili i cui risultati medi erano piuttosto diversi, ovvero le variabili relative ai contributi pubblici per spese, affitto e mutuo, e "redditi dei minori di 15 anni in famiglia",.

Il confronto degli standard errors è stato esteso anche al caso in cui i missing values vengano imputati una sola volta attraverso l'implementazione di un solo iter completo della procedura iterativa di imputazione multipla. Per completezza è stato riportato il rapporto tra lo standard error minore e quello maggiore ottenuti considerando ciascuna delle 10 imputazioni multiple come imputazioni singole, rispetto agli standard errors ottenuti con imputazione multipla (seconda e terza colonna delle tabelle 4.10 e 4.12). In questo caso la teoria suggerisce che gli standard errors ottenuti con imputazione singola dovrebbero essere inferiori a quelli ottenuti con imputazione multipla, dal momento che i primi non tengono in considerazione l'incertezza dovuta all'imputazione dei valori, trattandoli come osservati. In pratica, questa sottostima dipende dalla varianza delle stime ottenute con l'imputazione multipla: se la varianza *between* è bassa, la varianza con imputazione multi-

pla corrisponde alla media delle varianze con imputazione singola (paragrafo 2.1.1).

Imputando una sola volta i valori con la procedura iterativa di imputazione si vede che, in 10 distinte applicazioni, per tutte le variabili si è ottenuto almeno uno standard error maggiore di quello con imputazione multipla ed almeno uno standard error inferiore; in generale la variabilità del rapporto tra le due quantità (s.e. imputazione singola / s.e. imputazione multipla) tende a essere lontano da 1 (maggiore o minore) per le variabili già messe in evidenza dall'analisi dei valori medi.

Sembra dunque confermato che la differenza tra compiere una *available case analysis* o imputare i valori mancanti, con imputazione multipla o singola, risulta particolarmente differenziato solo per alcune variabili. Per altre, invece, lavorare con i soli dati disponibili non sembra comportare una particolare differenza in termini di media pesata e relativo standard error.

Per quanto riguarda poi le variabili di reddito di livello individuale costruite, gli effetti risultano piuttosto mitigati (tabella 4.11); in questo caso è stata calcolata anche la *fraction of missing information* (paragrafo 2.1.1).

Quest'ultima quantità risulta inferiore rispetto al tasso di mancata risposta per le prime due variabili (9% e 1% circa contro percentuali di mancata risposta pari rispettivamente al 34% e 11%), mentre per la variabile relativa al reddito totale per pensionati la percentuale di mancate risposte risulta maggiore (5% contro una *fraction of missing information* pari al 7% circa). Questo potrebbe essere legato all'alta variabilità, sotto il modello di imputazione proposto, dello standard error della variabile "Pensione integrativa" (tabella 4.10). Infine, svolgendo lo stesso tipo di analisi per la variabile di reddito costruita a livello familiare, si osserva che il rapporto tra lo standard error ottenuto senza imputazione dei valori mancanti e con l'imputazione multipla è pari a 0.95, mentre lo stesso rapporto utilizzando l'imputazione singola varia tra 0.99 e 1.02. La scarsa rilevanza della varianza *between* è quindi confermata dalla *fraction of missing information*, pari al 6% contro un tasso di mancata risposta osservato del 17.2%. La scarsa variabilità tra le imputazioni per la media pesata di molte delle variabili considerate è quindi dovuta alla poca informazione mancate, che deriva verosimilmente sia dai ridotti tassi di mancata risposta che dall'informazione apportata dal modello di imputazione.

Per approfondire l'analisi delle variabili dopo il procedimento di imputazione, oltre alle media sono state calcolare anche le mediane delle singole variabili[12]. I quantili della distribuzione delle variabili imputate sono solita-

[12]Per ottenere la stima dello standard error della mediana in ogni singolo dataset completato è stata utilizzata la formula approssimata suggerita in Schafer (1997) (capitolo 4),

Tabella 4.10: Variabili di reddito individuali. Rapporto tra standard errors delle stime: s.e. senza imputare/s.e. con imputazione multipla, s.e. minimo con una imputazione/s.e. con imputazione multipla, s.e. massimo con una imputazione/s.e. con imputazione multipla.

Variabile	No imp. /MI	Min 1imp. /MI	Max 1imp. /MI
Sezione redditi da lavoro dipendente			
Retribuzione mensile netta	1.06	0.97	1.02
Compensi aggiuntivi	1.03	0.97	1.06
Altri compensi aggiuntivi	1.05	1.00	1.00
Arretrati da lavoro	0.99	0.89	1.17
Liquidazioni da lavoro	1.03	0.94	1.20
Assegni familiari (mensili)	1.07	0.95	1.24
Sezione redditi da lavoro autonomo			
Reddito complessivo	1.15	0.96	1.03
Assegni familiari (mensili)	0.92	0.78	1.22
Indennità per maternità	0.53	0.44	1.21
Sezione redditi da pensioni			
Pensione sociale (mensile)	1.02	0.92	1.08
Pensione di anzianità (mensile)	1.00	0.98	1.02
Pensione di reversibilità (mensile)	0.99	0.97	1.06
Pensione di invalidità (mensile)	1.01	0.98	1.04
Assegni di accompagnamento (mensili)	0.99	0.98	1.04
Assegni familiari (mensili)	1.01	1.00	1.01
Pensione integrativa (mensile)	0.71	0.46	1.68
Sezione altre informazioni relative al 2003			
Indennità disoccupazione (mensili)	1.05	0.91	1.09
Assegni familiari per disoccupati (mensili)	1.02	0.92	1.07
Cassa integrazione (mensile)	0.71	0.71	1.42
Assegni familiari per cassaintegrati (mensili)	1.05	1.00	1.00
Borsa lavoro (mensile)	0.62	0.68	1.44
Borsa di studio (mensile)	0.25	0.32	1.60
Versamenti a persone fuori dalla famiglia	0.79	0.77	1.49
Versamenti da persone fuori della famiglia	0.97	0.79	1.21
Contributi per pensione integrativa (mensile)	1.10	0.99	1.02
Guadagni da risparmi	1.31	0.97	1.03
Guadagni da terreni o fabbricati	1.00	0.94	1.07
Imposta Comunale sugli Immobili	1.16	0.98	1.03
Rimborso da dichiarazione dei redditi	1.06	0.98	1.07
Pagamento da dichiarazione dei redditi	1.12	0.96	1.06

Tabella 4.11: Variabili di reddito individuali composte. Rapporto tra standard errors delle stime: s.e. senza imputare/s.e. con imputazione multipla, s.e. minimo con una imputazione/s.e. con imputazione multipla, s.e. massimo con una imputazione/s.e. con imputazione multipla, *fraction of missing information*.

Variabile	No imp. /MI	Min 1imp. /MI	Max 1imp. /MI	Fraction mis.info
Reddito individuale totale	1.10	0.97	1.02	0.087
Reddito totale dipendenti	1.07	1.01	1.02	0.011
Reddito totale pensionati	1.00	0.99	1.02	0.074
Reddito familiare totale	0.95	0.99	1.02	0.060

Tabella 4.12: Variabili di reddito familiari. Rapporto tra standard errors delle stime: s.e. senza imputare/s.e. con imputazione multipla, s.e. minimo con una imputazioni/s.e. con imputazione multipla, s.e. massimo con una imputazioni/s.e. con imputazione multipla.

Variabile	No imp. /MI	Min 1imp. /MI	Max 1imp. /MI
Sezione casa e zona di abitazione			
Contributi pubblici per spese per la casa	0.29	0.73	1.16
Sezione famiglie in affitto			
Contributi pubblici per affitto	1.16	0.91	1.07
Sezione famiglie con casa di proprietà			
Contributi pubblici per interessi sul mutuo	0.25	0.71	1.20
Sezione situazione economica			
Reddito minimo vitale (mensile)	0.81	0.66	1.64
Assegno di sostegno per almeno 3 figli minori	0.82	0.80	1.08
Assegno di maternità	1.04	0.98	1.04
Guadagni da affitti o subaffitti	1.06	1.00	1.00
Reddito dei minori di 15 anni in famiglia	0.50	0.76	1.09

mente più sensibili al particolare modello di imputazione scelto, in quanto il loro valore è determinato da tutta la distribuzione dei valori.

Tabella 4.13: Mediane per alcune variabili di reddito individuali e familiari, senza e con imputazione multipla dei valori mancanti.

Variabile	Osserv. (N)	Mediana	Standard Error
Senza imputazione			
Reddito totale dipendenti	16791	13380	84.00
Reddito totale pensionati	16552	9985	73.68
Reddito individuale totale	39130	9792	70.24
Reddito familiare totale	20042	17204	104.25
Imputazione multipla			
Reddito totale dipendenti	18730	13200	36.27
Reddito totale pensionati	17334	9903	72.92
Reddito individuale totale	52509	10726	52.37
Reddito familiare totale	24204	21294	127.10

Tabella 4.14: Confronto tra gli s.e. della mediana per alcune variabili di reddito individuali e familiari

Variabile	No imp. /MI	Min 1imp. /MI	Max 1imp. /MI
Reddito totale dipendenti	2.32	1.34	0.76
Reddito totale pensionati	1.01	0.98	0.94
Reddito individuale totale	1.34	1.05	0.83
Reddito familiare totale	0.82	1.01	0.92

Per i valori delle mediane si osservano infatti differenze più elevate rispetto a quanto avveniva per le medie; nelle tabelle 4.13 e 4.14 sono riportati i valori ed i confronti degli standard errors relativamente alle variabili costruite a livello individuale e familiare. In particolare, si noti che la differenza maggiore si osserva per la mediana del reddito individuale totale (tabella 4.13): ciò dipende soprattutto dalla elevata variabilità della stima della mediana per le variabili di reddito della sezione "Altre informazioni relative al 2003", fenomeno già in parte presente nella stima delle medie. Anche il valore degli standard errors delle mediane risulta particolarmente variabile a seconda del metodo di trattamento delle mancate risposte; questo vuol dire, quindi, che

che utilizza la distribuzione empirica della variabile di interesse $\widehat{F}$ e le relative statistiche d'ordine $\xi^{(i)}$ con $i = 1, \ldots, m$. La stima del quantile p_i è la $i - esima$ statistica di ordine con $i = p(m+1)$ se questo è un intero, oppure $\widehat{F}^{-1}(p) = (1-c)\xi^{(i_1)} + c\xi^{(i_2)}$, con $c = p(m+1) - i_1$.

si possono avere differenze piuttosto elevate nella precisione della stima della mediana a seconda che si utilizzi una *available case analysis* o l'imputazione (singola o multipla) dei valori. E' da sottolineare, tuttavia, che i valori degli standard errors nella tabella 4.14 potrebbero risentire della formula approssimata utilizzata per calcolarli; sarebbe sicuramente interessante ricalcolare tali standard errors utilizzando in ogni dataset tecniche di ricampionamento come il *bootstrap*.

Per vedere se le distribuzioni dei valori delle singole componenti di reddito risultano modificate in seguito all'imputazione dei valori mancanti, nelle figure 4.16-4.27 sono riportati gli istogrammi e le stime kernel delle densità dei valori osservati e dei valori completati con le imputazioni[13] per le stesse variabili già considerate nel paragrafo 4.5.2. Si vede allora che l'introduzione dei valori mancanti modifica la distribuzione dei valori soprattutto per le variabili con maggior numero di missing values, come nel caso delle variabile familiare "Contributi pubblici per l'affitto" (figura 4.27) per la quale la percentuale di mancante risposte è superiore all'80%. In tutti gli altri casi, invece, le distribuzioni appaiono sostanzialmente immutate anche dopo l'introduzione dei valori imputati; è chiaro tuttavia che il confronto, se di interesse sostanziale, dovrebbe basarsi su opportuni test statistici. Dalle figure 4.22 e 4.23 si vede inoltre che la presenza di valori "speciali" nelle variabili è stata correttamente riprodotta dal modello di imputazione attraverso la procedura in due step descritta nel paragrafo 4.5.2.

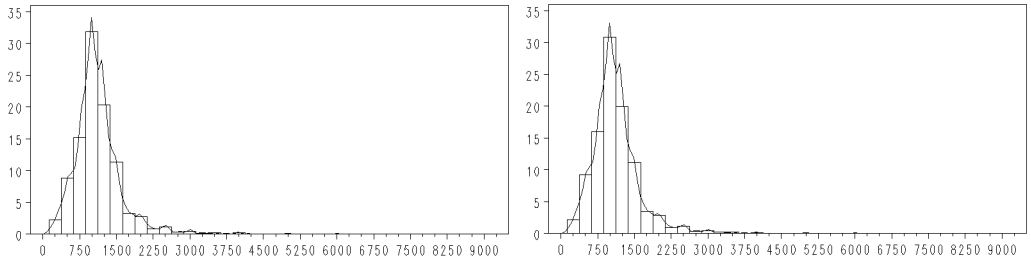

Figura 4.16: Istogramma dei valori prima (N=17046) e dopo l'imputazione (N=18730) per la variabile "Retribuzione mensile netta".

Se si vanno a considerare le stime puntuali riportate da ISTAT nella pubblicazione ufficiale ISTAT (2006) relativamente alle grandezze di reddito costruite a livello individuale e familiare, si può avere un'indicazione delle differenze apportate dalla procedura di imputazione multipla proposta in questa tesi. Come si vede dalla tabella 4.15, le stime ottenute attraverso

[13]Gli istogrammi si riferiscono ad uno dei 10 datasets completati, selezionato a caso.

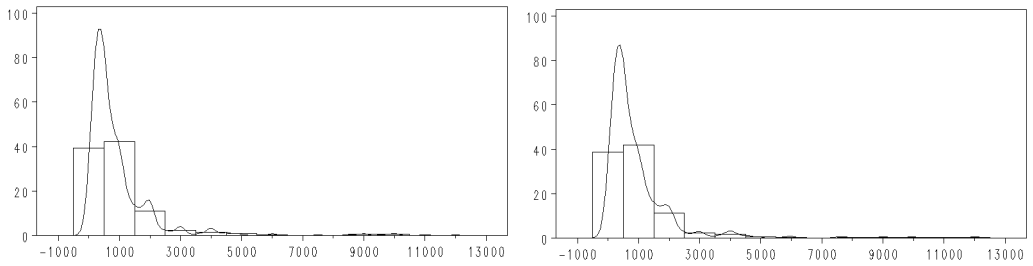

Figura 4.17: Istogramma dei valori prima (N=524) e dopo l'imputazione (N=617) per la variabile "Arretrati da lavoro").

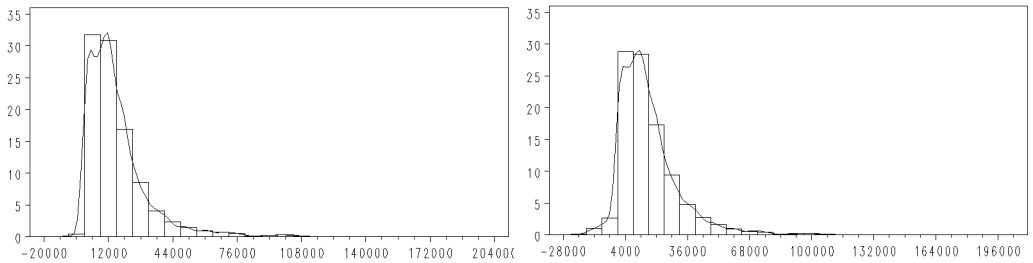

Figura 4.18: Istogramma dei valori prima (N=7018) e dopo l'imputazione (N=8907) per la variabile "Reddito totale da lavoro autonomo".

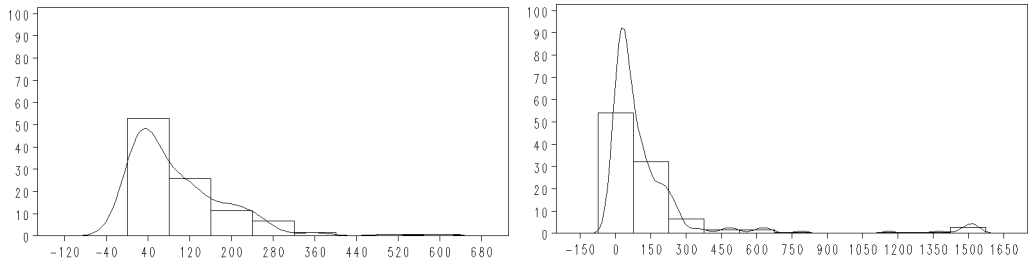

Figura 4.19: Istogramma dei valori prima (N=132) e dopo l'imputazione (N=189) per la variabile "Assegni familiari per lavoratori autonomi".

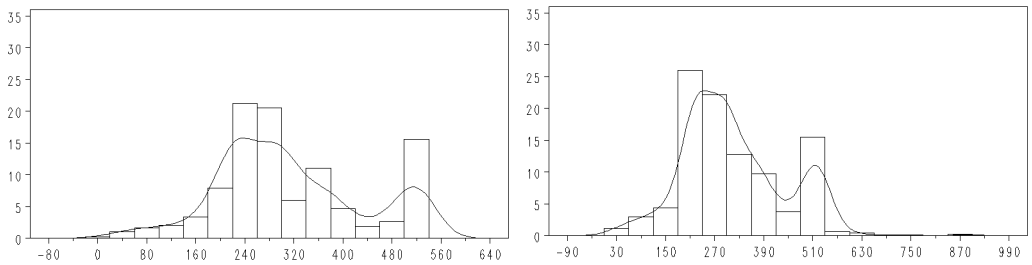

Figura 4.20: Istogramma dei valori prima (N=681) e dopo l'imputazione (N=750) per la variabile "Pensione sociale".

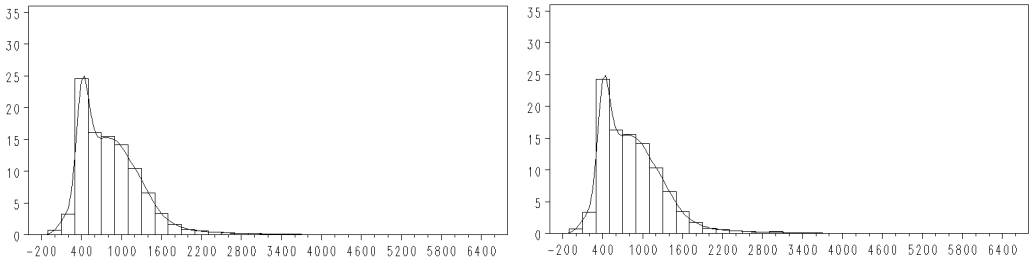

Figura 4.21: Istogramma dei valori prima (N=11336) e dopo l'imputazione (N=11812) per la variabile "Pensione di anzianità".

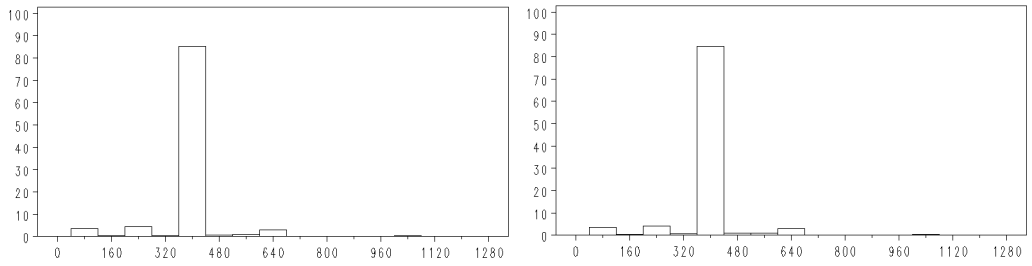

Figura 4.22: Istogramma dei valori prima (N=1098) e dopo l'imputazione (N=1119) per la variabile "Assegni di accompagnamento".

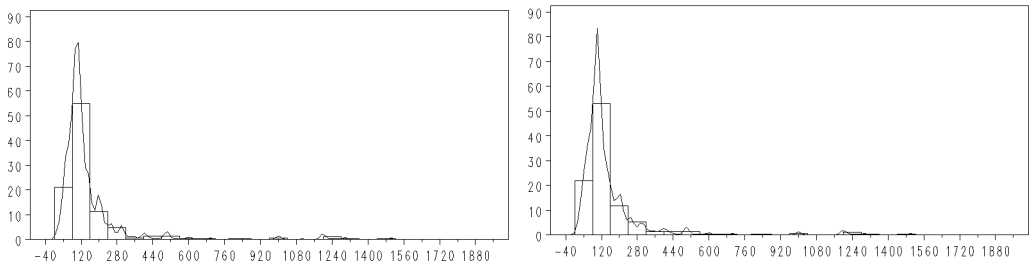

Figura 4.23: Istogramma dei valori prima (N=3732) e dopo l'imputazione (N=4128) per la variabile "Contributi per pensione privata".

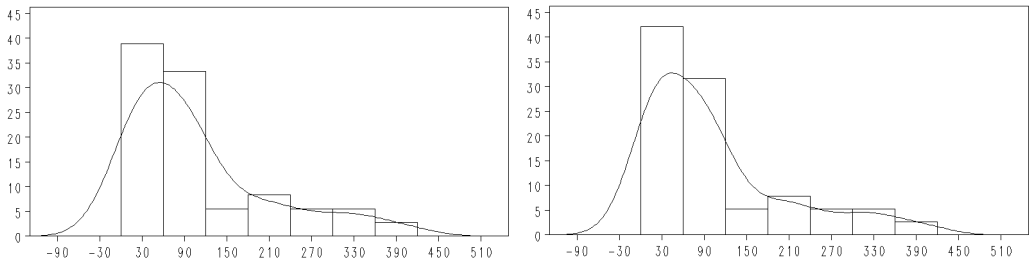

Figura 4.24: Istogramma dei valori prima (N=36) e dopo l'imputazione (N=38) per la variabile "Assegni familiari per cassaintegrati".

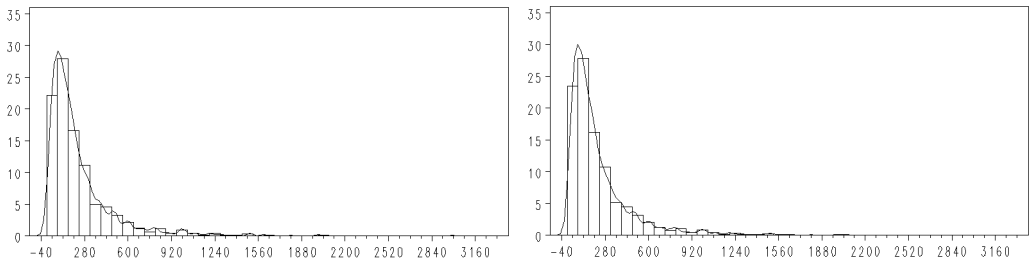

Figura 4.25: Istogramma dei valori prima (N=21863) e dopo (N=26677) l'imputazione per la variabile "Imposta Comunale sugli Immobili".

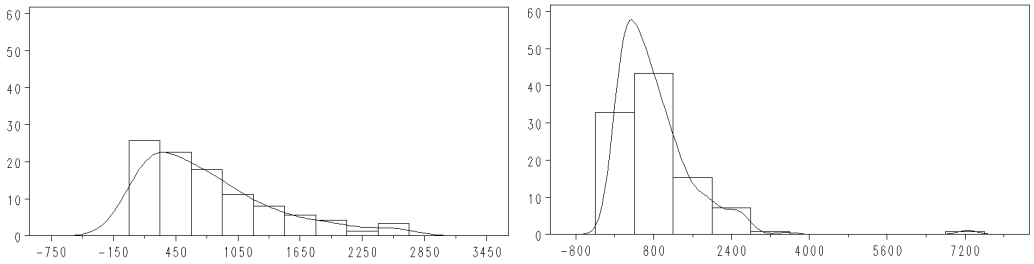

Figura 4.26: Istogramma dei valori prima (N=234) e dopo l'imputazione (N=268) per la variabile "Reddito minimo vitale".

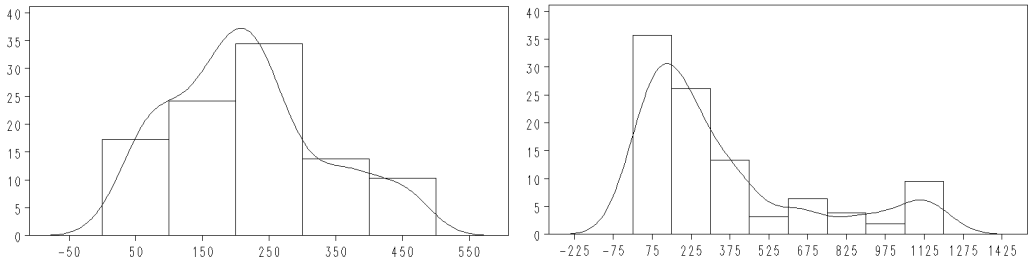

Figura 4.27: Istogramma dei valori prima (N=29) e dopo l'imputazione (N=157) per la variabile "Contributi pubblici per l'affitto".

la procedura di imputazione multipla iterativa risultano maggiori rispetto a quelle ISTAT, con la sola eccezione del reddito individuale da pensione. Come detto, l'imputazione singola delle variabili di reddito è stata effettuata da ISTAT attraverso lo stesso algoritmo utilizzato nella tesi, l'algoritmo Sequential Regression Multivariate Imputation, ma senza la procedura iterativa tra i due livelli di analisi e con un diverso insieme di predittori. In particolare, ISTAT ha specificato modelli separati per ciascuna sezione dei questionari, utilizzando comunque una grande quantità di predittori nelle regressioni (Vitaletti, 2005); le differenze sembrano quindi dovute principalmente alla procedura iterativa proposta nella tesi, che utilizza le componenti di reddito familiari per imputare quelle di livello individuale e vicerversa.

Tabella 4.15: Confronto tra alcune stime puntuali ottenute attraverso le imputazioni multiple ISTAT e le imputazioni multiple della tesi.

Variabile	Imputazioni ISTAT	Imputazioni tesi
Reddito familiare totale (media)	24951	25400
Reddito familiare totale (mediana)	20034	21294
Reddito individuale da lavoro dipendente	14289	14494
Reddito individuale da lavoro autonomo	15787	16776
Reddito individuale da pensione	11131	11117

Oltre al calcolo delle medie e delle mediane delle singole componenti di reddito individuali e familiari, si è ritenuto opportuno svolgere un'analisi di regressione utilizzando i dati a disposizione, per vedere in quale modo la procedura di imputazione proposta potesse avere effetto anche su tale tipo di analisi. In particolare, si è scelto di adattare una regressione logistica al livello familiare in cui la variabile risposta dicotomica è la difficoltà, dichiarata dalla famiglia, di arrivare a fine mese con i redditi a disposizione[14].

Come variabili esplicative sono state considerate le caratteristiche della famiglia e anche le caratteristiche individuali del maggior percettore in famiglia; in particolare, è stata introdotta come esplicativa anche la variabile "reddito familiare totale", ovvero $Y_{..j}^{fam} = \sum_{t=1}^{8} Y_{t.j} + \sum_{i=1}^{ncomp_j} \sum_{p=1}^{30} Y_{pij}$, suddivisa in quattro fasce definite in base ai quantili della distribuzione osservata. Si noti che l'analisi di regressione, più che ad uno specifico scopo interpretativo, vuole servire come strumento per leggere gli effetti della procedura di imputazione sulle stime dei coefficienti e degli errori standard.

[14]Tale variabile risposta è stata definita raggruppando le modalità di risposta "con grande difficoltà", "con difficoltà", "con qualche difficoltà" e, all'opposto, le modalità "con una certa facilità", "con facilità", "con molta facilità".

I risultati dell'applicazione del modello di regressione logistica[15] utilizzando la *complete case analysis* e l'imputazione multipla iterata dei dati mancanti sono riportati, rispettivamente, nelle tabelle 4.16 e 4.17.

Tabella 4.16: Regressione logistica per la difficoltà dichiarata dalla famiglia di arrivare a fine mese: risultati senza imputazione dei valori mancanti.

Effetto	Stima	Standard Error
Intercetta	0.78	0.0591
Appartamento (Villa o villetta rif.)	0.10	0.0319
2 componenti in famiglia (1 comp. rif.)	-0.08	0.0425
3 componenti in famiglia	0.18	0.0484
4 componenti in famiglia	0.37	0.0529
5 o più componenti in famiglia	0.63	0.0796
Affitto (proprietà rif.)	0.79	0.0487
Residenza nel nord-est (nord-ovest rif.)	-0.17	0.0405
Residenza nel centro	0.47	0.0428
Residenza nel sud	0.83	0.0515
Residenza sulle isole	0.74	0.0701
Reddito familiare totale 9500-15000 euro ($\leq$ 9500 rif.)	0.44	0.0456
Reddito familiare totale 15000-25000 euro	0.16	0.0452
Reddito familiare totale $\geq$ 25000 euro	-0.54	0.0455
Principale percettore femmina (maschio rif.)	0.07	0.0337
Principale percettore autonomo (dipendente rif.)	-0.30	0.0484
Principale percettore disoccupato	0.54	0.1176
Principale percettore in altra condiz. lavorativa	-0.09	0.0379
Principale percettore diplomato (istruzione inferiore rif.)	-0.52	0.0356
Principale percettore laureato	-1.30	0.0535

Come si vede dal confronto dei risultati, alcune stime degli effetti risultano modificate; questo si verifica in particolare per le categorie del reddito familiare totale, variabile che in effetti racchiude tutti i valori mancanti. L'imputazione di questi dati fa sì che per le categorie del reddito familiare totale si ottengano stime coerenti con quanto ci si attenderebbe: passando da una fascia di reddito a quella maggiore l'effetto sulla risposta aumenta in segno negativo, ovvero la probabilità di trovarsi in difficoltà economiche a fine mese diminuisce all'aumentare del reddito. Questo non avviene, invece, per le stime della regressione che utilizza i soli dati completi. Anche per le variabili relative al numero dei componenti appartenenti alla famiglia si osservano delle differenze, e questo è attribuibile al fatto che la *complete case*

[15]Il modello di regressione è stato stimato attraverso la PROC LOGISTIC del software SAS. In caso di mancate risposte alle singole variabili utilizzate nel modello di regressione le corrispondenti osservazioni vengono eliminate dall'analisi.

Tabella 4.17: Regressione logistica per la difficoltà dichiarata dalla famiglia di arrivare a fine mese: risultati con imputazione multipla dei valori mancanti.

Effetto	Stima	Standard Error
Intercetta	1.51	0.0646
Appartamento (Villa o villetta rif.)	0.13	0.0222
2 componenti in famiglia (1 comp. rif.)	0.33	0.0412
3 componenti in famiglia	0.82	0.0442
4 componenti in famiglia	1.04	0.0455
5 o più componenti in famiglia	1.23	0.0542
Affitto (proprietà rif.)	0.71	0.0353
Residenza nel nord-est (nord-ovest rif.)	-0.17	0.0285
Residenza nel centro	0.44	0.0299
Residenza nel sud	0.69	0.0356
Residenza sulle isole	0.60	0.0487
Reddito familiare totale 9500-15000 euro ($\leq$ 9500 rif.)	-0.34	0.0647
Reddito familiare totale 15000-25000 euro	-0.82	0.0639
Reddito familiare totale $\geq$ 25000 euro	-1.73	0.0656
Principale percettore femmina (maschio rif.)	0.03	0.0217
Principale percettore autonomo (dipendente rif.)	-0.44	0.0352
Principale percettore disoccupato	0.47	0.0645
Principale percettore in altra condiz. lavorativa	-0.23	0.0261
Principale percettore diplomato (istruzione inferiore rif.)	-0.49	0.0242
Principale percettore laureato	-1.16	0.0379

analysis, eliminando le osservazioni con valori mancanti del reddito, modifica anche la composizione delle famiglie.

Per quanto riguarda invece gli standard errors, come già fatto relativamente alla stima delle medie e delle mediane, sono stati calcolati i rapporti nella tabella 4.18, in cui compare anche la *fraction of missing information*.

La *fraction of missing information* è vicina a zero per gli effetti diversi dal reddito familiare totale che non risultano direttamente influenzati dalla composizione della famiglia, ovvero gli effetti relativi alla residenza e al titolo di godimento dell'abitazione: ciò significa che le non risposte per le variabili di reddito non comportano una perdita di informazione per queste covariate. Negli altri casi, invece, la *fraction of missing information* assume valori maggiori.

Coerentemente con questo e con i risultati precedenti, per tutti gli effetti non relativi al reddito i rapporti degli standard errors maggiore e minore ottenuti con imputazione singola rispetto a quello con imputazione multipla sono tutti vicini ad 1; questo significa, come già precedentemente sottolineato, che la variabilità *between* è praticamente nulla. Per gli effetti relativi al reddito, invece, i due rapporti sono sempre minori di 1, anche quando si considera lo standard error maggiore ottenuto con imputazione singola.

Infine, i rapporti tra gli standard errors che si ottengono senza imputazione e con imputazione multipla sono maggiori di 1 per tutte le covariate diverse dal reddito familiare: ciò deriva dal fatto che la *complete case analysis* utilizza una minor numero di osservazioni; anche in questo caso bisogna ricordare che gli standard errors ottenuti con i soli dati completi possono essere distorti all'allontanarsi dall'ipotesi MCAR. Ciò non avviene invece per gli effetti relativi al reddito, per i quali risulta determinante la variabilità *between* le imputazioni multiple, che incrementa la stima del relativo standard error.

Tabella 4.18: Regressione logistica: confronto degli standard error e *fraction of missing information.*

Variabile	No imp. /MI	Min 1imp. /MI	Max 1imp. /MI	Fraction mis.info
Intercetta	0.92	0.93	0.94	0.127
Appartamento (Villa o villetta rif.)	1.44	1.00	1.00	0.002
2 componenti in famiglia (1 comp. rif.)	1.03	0.98	0.98	0.041
3 componenti in famiglia	1.09	0.97	0.97	0.054
4 componenti in famiglia	1.16	0.97	0.97	0.056
5 o più componenti in famiglia	1.47	0.97	0.97	0.057
Affitto (proprietà rif.)	1.38	1.00	1.00	0.007
Residenza nel nord-est (nord-ovest rif.)	1.42	0.99	1.00	0.008
Residenza nel centro	1.43	0.99	1.00	0.008
Residenza nel sud	1.44	0.99	1.00	0.010
Residenza sulle isole	1.44	0.99	0.99	0.014
Reddito familiare totale 9500-15000 euro ($\leq$ 9500 rif.)	0.70	0.88	0.90	0.218
Reddito familiare totale 15000-25000 euro	0.71	0.87	0.90	0.215
Reddito familiare totale $\geq$ 25000 euro	0.69	0.86	0.88	0.244
Principale percettore femmina (maschio rif.)	1.55	1.00	1.00	0.005
Principale percettore autonomo (dipendente rif.)	1.37	0.99	0.99	0.017
Principale percettore disoccupato	1.82	0.99	1.00	0.011
Principale percettore in altra condiz. lavorativa	1.45	0.99	0.99	0.015
Principale percettore diplomato (istruzione inferiore rif.)	1.47	1.00	1.00	0.004
Principale percettore laureato	1.41	1.00	1.00	0.003

4.7 Alcune diagnostiche per la verifica delle imputazioni

La pratica diffusasi negli ultimi anni per il trattamento delle mancate risposte è l'utilizzo di modelli che ipotizzino dati MAR, come il metodo delle regressioni sequenziali multivariate. Le imputazioni ottenute attraverso tali modelli, sebbene imperfette, possono essere una buona approssimazione specialmente quando la *fraction of missing information* è bassa (Abayomi et al., 2007).

A sostegno dell'impiego di modelli di imputazione MAR vi è anche il fatto che risulta spesso impossibile testare tale ipotesi dal momento che, per testarla formalmente, occorrerebbe disporre non solo dei dati osservati ma anche di quelli mancanti che però, per definizione, non sono disponibili (capitolo 3). Recentemente tuttavia, alcuni autori hanno iniziato ad affrontare questa interessante problematica in relazione alle procedure di imputazione multipla (Abayomi et al., 2007; Raghunathan and Bondarenko, 2007). L'idea comune a questi autori è di sostituire ad un formale test dell'ipotesi MAR, quando questo è impossibile da realizzare, altri tipi di verifiche. In particolare, sono stati proposti test di tipo *esterno*, che confrontano la distribuzione dei valori osservati con quella dei valori imputati, e test di tipo *interno* relativamente al modello di imputazione utilizzato. Le verifiche proposte hanno il pregio di poter essere applicate con software e procedure *standard*, e anche senza l'esatta conoscenza del modello di imputazione utilizzato. Ciò è particolarmente utile nelle situazioni in cui l'utilizzatore dei dati completati è un soggetto diverso da colui che ha realizzato le imputazioni.

Per quanto riguarda i test di tipo *esterno*, questi si basano sostanzialmente sul confronto della distribuzione dei valori imputati con la distribuzione di quelli osservati, al fine di individuare potenziali problemi o suggerire modifiche per il modello di imputazione. Per esempio, data una variabile continua, l'uguaglianza della distribuzione dei valori osservati con quelli imputati può essere testata attraverso il test non parametrico di Kolmogorov-Smirnov, e visualmente, disegnando le densità empiriche. Bisogna specificare, però, che l'eventuale disuguaglianza delle densità marginali dei valori osservati ed imputati non comporta automaticamente la presenza di errori nel modello di imputazione con ipotesi MAR. Infatti, l'uguaglianza delle densità marginali deve essere rispettata sotto l'ipotesi MCAR, più restrittiva della MAR; è chiaro, tuttavia, che evidenti disuguaglianze delle distribuzioni marginali possono comunque servire per evidenziare situazioni da approfondire (Abayomi et al., 2007).

Per testare l'ipotesi MAR, invece, il confronto tra le distribuzioni dei va-

lori osservati ed imputati per una variabile dovrebbe essere svolto *ceteris paribus*, ovvero condizionando per uguali valori delle altre variabili (Raghunathan and Bondarenko, 2007). In particolare, le variabili cui condizionare il confronto possono essere le stesse utilizzate nel modello di imputazione, se il modello di analisi coincide con quello di imputazione, oppure un insieme di variabili alternativo, nel caso in cui l'interesse sia testare l'ipotesi MAR per un modello diverso. Questa situazione rientra nel più vasto concetto di *uncongeniality*, ovvero di non affinità tra modello di imputazione e successiva procedura di analisi (paragrafo 2.1.2).

Per quanto riguarda invece i test di tipo *interno*, nel caso in cui la procedura di imputazione utilizzi delle regressioni è possibile ricorrere ad alcune delle diagnostiche classiche normalmente utilizzate per questo tipo di modelli. Per esempio, un'analisi dei residui potrebbe suggerire una procedura di rifinimento in cui i valori imputati vengano corretti o calibrati (Abayomi et al., 2007).

4.7.1 Un'applicazione ai dati EU-SILC

Per capire la potenzialità delle diagnostiche per l'ipotesi MAR di tipo *interno* è stata svolta un'applicazione ai dati di reddito EU-SILC imputati attraverso il metodo delle regressioni sequenziali presentato nei paragrafi precedenti.

Per questa prima analisi sono stati confrontate le distribuzioni dei valori imputati e dei valori osservati per la variabile "Assegni familiari ricevuti dai lavoratori dipendenti". Nella figura 4.28 sono riportati gli istogrammi, contenenti anche alcune statistiche base, relativi ai valori osservati e a quelli imputati provenienti da uno dei 10 dataset completati.

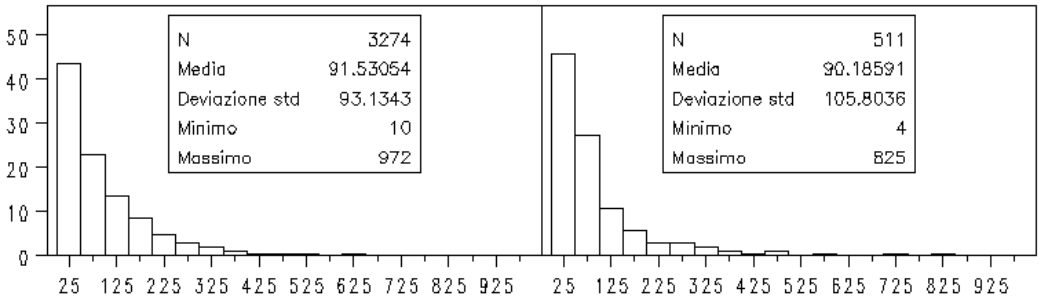

Figura 4.28: Assegni familiari per lavoratori dipendenti: distribuzione dei valori osservati (a sinistra) e dei valori imputati (a destra).

Per svolgere un test più formale sull'uguaglianza delle distribuzioni si è scelto di utilizzare il test non parametrico di Kolmogorov-Smirnov per due

campioni (Gibbons and S., 1992), che confronta le distribuzioni empiriche dei valori osservati ed imputati in ciascun dataset. Il valore asintotico della statistica di Kolmogorov-Smirnov ed il p-value per il confronto a coppie sono riportati nella tabella 4.19 separatamente per ciascuno dei 10 dataset completati.

Tabella 4.19: Valore asintotico della statistica di Kolmogorov-Smirnov e p-value per il confronto a coppie delle distribuzioni marginali dei valori osservati ed imputati per la variabile "Assegni familiari ricevuti dai lavoratori dipendenti" (per dataset).

Dataset Completato	Statistica K-S	p-value K-S
Dataset 1	1.8993	0.0015
Dataset 2	2.1885	0.0001
Dataset 3	1.7989	0.0031
Dataset 4	2.2951	0.0001
Dataset 5	1.6368	0.0094
Dataset 6	1.9518	0.0010
Dataset 7	2.0482	0.0005
Dataset 8	2.0753	0.0004
Dataset 9	1.6302	0.0098
Dataset 10	2.1873	0.0001

Come si vede, il test rifiuta l'ipotesi nulla di uguaglianza tra le distribuzioni empiriche dei valori osservati ed imputati in tutti i dataset completati. Tuttavia, come già accennato, per testare l'ipotesi MAR il confronto tra le distribuzioni dei valori imputati e osservati dovrebbe essere eseguita condizionando per le covariate utilizzate nel modello di imputazione. Per far questo, si è scelto di utilizzare l'approccio del *propensity score* proposto da Raghunathan and Bondarenko (2007).

Il *propensity score* (Rosenbaum and Rubin, 1983) è stato introdotto e discusso nella letteratura relativa alle mancate risposte da Little (1986). In tale articolo, Little parla di *nonresponse propensity* riferendosi alla probabilità di appartenere al gruppo dei non rispondenti date le covariate osservate; tale metodo di classificazione può essere utilizzato per costruire *celle di imputazione* (paragrafo 1.2.1).

Per formalizzare il problema, utilizziamo la variabile indicatrice delle mancate risposte M_{pi}, che assume valore 1 quando la variabile $p = 1, ..., P$ relativa all'individuo $i = 1, ..., n$ è missing, 0 altrimenti; inoltre, sia $\boldsymbol{Y}_{obs,p}$ il vettore dei valori osservati per la variabile p, $\boldsymbol{Y}_{mis,p}$ il vettore dei valori man-

canti. Se per esempio $\boldsymbol{Y}_{obs,p} = y_{pi}$ per $i = 1, ..., n_i$ mentre $\boldsymbol{Y}_{mis,p} = y_{pi}$ per $i = n_i + 1, ..., n$, allora $M_{pi} = 1$ per $i = 1, ..., n_i$, $M_{pi} = 0$ per $i = n_i + 1, ..., n$.

Indichiamo poi con $\boldsymbol{Y}_{obs,-p}$ l'insieme dei valori osservati per tutte le variabili esclusa la *p-esima*, ovvero l'insieme dei vettori $\boldsymbol{Y}_{obs,p}$ con $p = 1, ...p-1, p+1, ..., P$. La probabilità di osservare un valore mancante per la variabile p può allora essere espressa, per ogni i, in funzione dei dati osservati attraverso la *nonresponse propensity* $e_{obs,-p} = P(\boldsymbol{M}_p = 1|\boldsymbol{Y}_{obs,-p})$. $e_{obs,-p}$ rappresenta una compattazione efficiente delle covariate che può essere utilizzata per confrontare gli esiti tra i due "trattamenti" $M_p = 1$ e $M_p = 0$ aggiustando per $\boldsymbol{Y}_{obs,-p}$. Sotto l'ipotesi MAR, allora, le distribuzioni dei valori osservati e mancanti per la variabile p, $\boldsymbol{Y}_{obs,p}$ e $\boldsymbol{Y}_{mis,p}$, dovrebbero essere simili condizionando per la *nonresponse propensity* $e_{obs,-p}$ (Raghunathan and Bondarenko, 2007). Infatti, l'ipotesi di ignorabilità del meccanismo di mancata risposta, assumendo $e_{obs,-p} > 0$ per ogni p, corrisponde alla condizione $\boldsymbol{Y}_p \perp \boldsymbol{M}_p | e_{obs,-p}$, ovvero all'indipendenza dei valori di Y da M (si veda la condizione (1.9)), condizionando per la *nonresponse propensity* (Little, 1986)
.

Da un punto di vista applicativo, la *nonresponse propensity* può essere stimata per ogni individuo adattando una regressione logistica per la variabile dicotomica $\boldsymbol{M}_p$ separatamente in ciascuno dei dataset completati, includendo tra i predittori i valori osservati $\boldsymbol{Y}_{obs,-p}$ e quelli imputati $\boldsymbol{Y}_{mis,-p}$. Poichè l'imputazione multipla produce m valori imputati per ciascun dato mancante, un unico valore $e_{obs,-p}$ si può ottenere, per m abbastanza grande, attraverso la seguente approssimazione (Raghunathan and Bondarenko, 2007):

$$\widehat{e}_{obs,-p} = \widehat{P}(\boldsymbol{M}_p = 1|\boldsymbol{Y}_{obs,-p}) = \sum_{m=1}^{M} \widehat{P}(\boldsymbol{M}_p = 1|\boldsymbol{Y}_{obs,-p}, \boldsymbol{Y}^m_{mis,-p})$$

Calcolata tale quantità per ciascuno degli n individui, un possibile metodo per testare l'ipotesi MAR consiste nel confrontare le distribuzioni e le principali statistiche descrittive dei valori osservati $\boldsymbol{Y}_{obs,p}$ e di quelli imputati $\boldsymbol{Y}_{mis,p}$ all'interno dei quantili della *nonresponse propensity* $\widehat{e}_{obs,-p}$, per ciascuno dei dataset imputati. In alternativa, per le variabili continue un utile metodo consiste nel regredire i dati completati $\boldsymbol{Y}_p$ su $\widehat{e}_{obs,-p}$ in ogni dataset, andando poi a confrontare la distribuzione dei residui tra i rispondenti e i non rispondenti. In questo modo risulta possibile svolgere un unico test d'ipotesi, invece che più test all'interno dei vari quantili.

Per applicare la procedura di diagnostica alla variabile "Assegni familiari ricevuti dai lavoratori dipendenti", per ciascuno dei 10 dataset è stata stimata, come primo passo, una regressione lineare dei valori $\boldsymbol{Y}_{ass}$ sulla *nonresponse propensity* $\widehat{e}_{obs,-ass}$, calcolata utilizzando come variabili esplicative

della regressione logistica per $\boldsymbol{M}_{ass}$ le stesse variabili impiegate nel modello di imputazione. La distribuzione dei residui di questa regressione nel gruppo dei rispondenti e dei non rispondenti sono state poi confrontate tra loro; in questo caso, ripetendo il test di Kolmogorov-Smirnov, si sono ottenuti i risultati riportati nella tabella 4.20.

Tabella 4.20: Valore asintotico della statistica di Kolmogorov-Smirnov e p-value per il confronto a coppie tra i residui dei rispondenti e non rispondenti alla variabile "Assegni familiari ricevuti dai lavoratori dipendenti", condizionando per la *nonresponse propensity* (per dataset).

Dataset Completato	Statistica K-S	p-value K-S
Dataset 1	1.3369	0.0561
Dataset 2	1.4383	0.0319
Dataset 3	1.2715	0.0788
Dataset 4	1.5798	0.0136
Dataset 5	1.3372	0.0559
Dataset 6	1.3111	0.0642
Dataset 7	1.5113	0.0208
Dataset 8	1.5684	0.0146
Dataset 9	1.3871	0.0426
Dataset 10	1.4901	0.0236

Condizionando per la *nonresponse propensity*, dunque, non si rifiuta l'ipotesi di uguaglianza della distribuzione dei valori tra il gruppo dei rispondenti e dei non rispondenti. Quindi, la conclusione cui si giunge è che l'ipotesi MAR per la variabile "Assegni familiari ricevuti dai lavoratori dipendenti" può essere accettata relativamente alle variabili utilizzate nel modello di imputazione.

Questo metodo di diagnostica , sebbene non sia un formale test d'ipotesi, ha il pregio di essere di facile applicazione ed utilizzo, permettendo la verifica della ragionevolezza dei valori imputati rispetto all'ipotesi MAR. Inoltre, la procedura può essere applicata anche quando non si è a conoscenza dell'esatto modello di imputazione, per esempio per verificare un'ipotesi MAR relativa al modello ipotizzato dall'analizzatore dei dati. Tuttavia, il metodo proposto può risultare sensibile rispetto alla scelta di come confrontare i valori dei rispondenti con quelli dei non rispondenti condizionando per la *nonresponse propensity*; in particolare, potrebbe essere erroneo supporre l'esistenza di una relazione lineare tra i valori della variabile considerata e la *nonresponse propensity*. Nel futuro sarebbe dunque interessante andare a ripetere questo

tipo di diagnostica utilizzando, per esempio, una regressione non parametrica basata su una *penalized spline* (Zheng and Little, 2003; Jo et al., 2007).

4.8 Conclusioni

In questo capitolo è stata proposta una procedura di imputazione multipla per i valori mancanti di reddito dell'indagine ISTAT Condizioni di Vita 2004. La procedura utilizza il metodo delle regressioni sequenziali multivariate in modo iterativo, così da sfruttare le componenti di reddito osservate a livello familiare per imputare le componenti di reddito a livello individuale e viceversa.

La procedura di imputazione proposta, oltre a gestire la struttura in due livelli dei dati, ha il pregio di dedicare grande attenzione alla scelta delle covariate, aspetto molto importante per le procedure di imputazione multipla; una buona scelta delle informazioni cui condizionarsi può rendere più plausibile, tra l'altro, l'ipotesi che i dati siano mancanti "a caso" (MAR). Per esempio, sono state considerate quali variabili esplicative per le regressioni del modello di imputazione anche le caratteristiche legate al disegno di campionamento dell'indagine e, come detto, le informazioni relative al livello familiare per l'imputazione delle variabili individuali e viceversa.

Attraverso la procedura di imputazione proposta sono stati imputati 10 valori per ciascuno dei missing values; le analisi condotte sui 10 dataset completati hanno riguardato le medie pesate ed i relativi standard errors per ciascuna delle 38 variabili di reddito, così come il calcolo delle mediane e la visualizzazione grafica dell'effetto delle imputazioni per alcune delle variabili. Inoltre, si è svolta un'analisi di regressione confrontando anche in questo caso le stime ottenute utilizzando le sole osservazioni complete ed i dati completati con le imputazioni.

In generale, tutte queste analisi hanno mostrato che sia la qualità che l'effetto delle imputazioni risultano variabili a seconda della particolare componente di reddito considerata, in quanto le percentuali di mancate risposte ed il numero di osservazioni risultano molto diversificate.

In ogni caso, andando a considerare non solo le singole componenti di reddito ma anche le loro aggregazioni nelle variabili "reddito individuale totale" e "reddito familiare totale", relativamente alle medie si è evidenziata una variabilità *between* le stime ottenute con l'imputazione multipla molto bassa. Questo significa che la correzione apportata dall'imputazione multipla al calcolo degli standard errors delle medie risulta piuttosto limitata, indicando una sostanziale indifferenza rispetto all'impiego di procedure di imputazione singola. Per quanto riguarda invece il valore della stima pun-

tuale, la procedura di imputazione comporta una correzione verso l'alto della media pesata sia rispetto alle stime ottenute da ISTAT con la propria procedura di imputazione singola, sia rispetto a quanto si ottiene con i soli casi completi.

Variazioni maggiori sia per le stime puntuali che per gli standard errors si osservano invece relativamente al calcolo delle mediane delle variabili di interesse; ciò può essere dovuto al fatto che le mediane dipendono da tutta la distribuzione dei valori osservati o imputati. Nel caso poi dell'analisi di regressione considerata, i coefficienti risultano coerenti con le ipotesi a priori solo effettuando l'imputazione dei valori; le differenze che si osservano tra imputazione multipla ed imputazione singola risultano invece più limitate. In ogni caso, su questi risultati influisce molto il fatto che i valori mancanti riguardano un sottoinsieme ristretto dei dati.

Infine, è stato proposto un metodo di diagnostica attraverso cui risulta possibile "testare" l'ipotesi MAR, anche se non formalmente, separatamente per tutte le variabili imputate, condizionando per la *nonresponse propensity*.

Capitolo 5

Un'analisi di sensitività per i dati di reddito dell'indagine Forze Lavoro del Comune di Firenze

In questo capitolo si analizzano i dati dell'indagine Forze Lavoro del Comune di Firenze, caratterizzata da una struttura "panel-ruotato". Dopo aver presentato le percentuali di mancata risposta e le stime del reddito medio che si ottengono analizzando le sole osservazioni disponibili (paragrafo 5.2), viene presentata la formalizzazione del meccanismo di risposta MAR e la relativa procedura di imputazione (paragrafo 5.3.1). Nel paragrafo 5.3.2 viene realizzata un'analisi di sensitività rispetto a due ipotesi MNAR, analizzando le modifiche apportate alle distribuzioni di reddito di interesse.

5.1 I dati mancanti di reddito nelle indagini sulle Forze di Lavoro

I dati mancanti per quesiti relativi al reddito rappresentano una problematica molto importante per le indagini sulle Forze di Lavoro. Tale problema è stato affrontato per la prima volta da un punto di vista metodologico negli Stati Uniti, relativamente all'*Income Supplement* della Current Population Survey, dove fu proposto l'utilizzo di tecniche di imputazione di tipo *hot deck*. In particolare, l'hot-deck della Current Population Survey utilizza delle celle di imputazione definite in base alle informazioni disponibili per rispondenti e non rispondenti, imputando ogni valore mancante di reddito utilizzando quello di un rispondente estratto a caso nella stessa cella del non rispondente.

Questo metodo di imputazione ipotizza che i dati siano mancanti a caso (MAR, paragrafi 1.1.2 e 1.2.2), ovvero che il loro essere mancanti dipenda solo da caratteristiche osservate e non dai valori mancanti delle variabili di reddito stesse.

Come già discusso nel paragrafo 3.2, quest'ipotesi è stata criticata da molti studiosi, secondo cui la probabilità di osservare una non risposta è maggiore per gli individui con reddito basso o elevato rispetto a quanto avviene per gli individui il cui reddito si colloca nella parte centrale della distribuzione del reddito. In particolare, Lillard et al. (1986) suggerirono un modello per il reddito di tipo *non a caso* (MNAR) basato sui modelli inizialmente proposti da autori come Heckman (Heckman, 1979). La loro conclusione fu che i valori di reddito imputati dalla tecnica hot-deck della Current Population Survey erano gravemente sottostimati. Tuttavia, i metodi di imputazione di tipo MNAR sono stati a loro volta fortemente criticati per la sensitività rispetto alle ipotesi strutturali, e un'applicazione basata sul *matching* dei dati della Current Population Survey con quelli dell'Internal Revenue Service ha evidenziato come non esista una prova certa dell'infondatezza dell'ipotesi MAR (David et al., 1986).

Nonostante questo, è possibile che ipotizzando un modello di tipo MAR per i dati di reddito si ottengano delle distorsioni, e ciò è particolarmente vero nelle situazioni in cui le covariate osservate in grado di caratterizzare le differenze tra rispondenti e non rispondenti sono poche.

Il trattamento delle mancate risposte di tipo MNAR presenta problemi metodologici notevoli (paragrafo 1.1.3), in quanto spesso non si dispone dei dati empirici in grado di cogliere le differenze tra rispondenti e non rispondenti. Molti autori hanno perciò suggerito che l'approccio scientifico più adeguato è rappresento dallo studio della sensitività rispetto a dati MNAR, andando a considerare l'effetto di una gamma di differenze plausibili tra rispondenti e non rispondenti dopo aver aggiustato per le covariate disponibili. Applicazioni di questo tipo tendono però a considerare solo le situazioni in cui la non risposta è univariata. In questo capitolo viene invece svolta un'analisi di sensitività per non risposta a dati di reddito di tipo MNAR multivariati, nell'ambito di un'indagine con uno schema di tipo "panel-ruotato". Questa applicazione presenta dunque varie interessanti fattori di complicazione.

In particolare, il pattern dei dati mancanti è di tipo multivariato, dal momento che il reddito viene rilevato ad ogni occasione d'indagine, ovvero ogni tre mesi; oltre all'ammontare del reddito anche il percepimento dello stesso deve essere considerato in ogni occasione, dal momento che il reddito deve essere posto pari a zero per i disoccupati; lo schema dell'indagine è di tipo panel ruotato, ovvero gli individui vengono intervistati in alcune occasioni ma non in altre; la rotazione del panel provoca dei missing "strutturali" di tipo

MAR; oltre a questi missing strutturali ci sono dei dati mancanti dovuti alle non risposte per valori di reddito, potenzialmente MNAR. In entrambi i casi, inoltre, per alcuni individui con valori di reddito mancanti l'informazione sul reddito è disponibile relativamente ad altre occasioni d'indagine, mentre per altri non si ha alcuna informazione sul reddito. Quindi, la ricchezza delle informazioni osservate risulta differenziata tra gli individui, e tale aspetto deve essere considerato nello svolgere l'analisi di sensitività.

In questo capitolo viene descritta un'analisi che si pone l'obiettivo di affrontare queste problematiche. Come primo passo i dati mancanti di reddito vengono imputati in tutte le occasioni di indagine con imputazione multipla attraverso il metodo delle regressioni sequenziali multivariate (paragrafo 2.2.2), metodo che consente il condizionamento rispetto alle covariate osservate; tra le informazioni osservate vi è anche, per alcuni individui, il reddito rilevato in altre occasioni. Successivamente vengono considerate due diverse forme di analisi di sensitività relativamente ai missing di reddito potenzialmente MNAR.

Contrariamente ad altri approcci, i metodi proposti possono essere implementati in modo relativamente semplice e sono in grado di fornire utili informazioni relativamente al potenziale impatto di deviazioni dall'ipotesi MAR per i dati di reddito mancanti.

5.2 L'indagine Forze Lavoro del Comune di Firenze

L'indagine sulle Forze Lavoro del Comune di Firenze costituisce una base di dati molto importante relativamente la tasso di occupazione, alla proporzione di persone in cerca di lavoro e al reddito dei lavoratori.

L'indagine prevede ogni anno quattro occasioni di indagine (in Aprile, Luglio, Ottobre e Gennaio) al fine di produrre stime trimestrali. Il campione è di tipo stratificato: gli individui vengono estratti all'anagrafe comunale di Firenze all'interno degli strati definiti in base al genere, la classe di età e la zona di residenza in Firenze.

Inoltre, l'indagine ha un disegno di tipo "panel-ruotato" in cui ciascun soggetto entra nel campione per due occasioni consecutive, esce per le successive due e entra di nuovo per due occasioni. Tale rotazione viene effettuata assegnando casualmente i soggetti a un determinato "gruppo panel" nel momento dell'estrazione del campione; in particolare in ogni occasione d'indagine un quarto del campione è alla prima intervista, un quarto alla seconda, un quarto alla terza e un quarto alla quarta intervista. Questo fa sì che vi sia

una sovrapposizione pari al 50% del campione a distanza di 3 e 12 mesi, e una sovrapposizione del 25% dopo 9 e 15 mesi. Tuttavia, se si considerano per esempio quattro occasioni di indagine il cui periodo di riferimento del reddito è lo stesso anno solare, solitamente il numero di individui intervistati in una sola occasione è superiore a quello determinato dallo schema di rotazione, a causa dell'impossibilità di ricontattare alcuni soggetti. In queste situazioni vengono intervistati dei sostituiti che appartengono allo stesso strato della popolazione.

In ciascuna delle quattro occasioni di indagine considerate in questa applicazione (Aprile, Luglio, Ottobre 2002 e Gennaio 2003) sono state intervistate circa 1200 persone. A seconda del "gruppo panel" di appartenenza, ciascun soggetto è stato intervistato una o due volte (tabella 5.1). Il numero totale dei rispondenti nelle quattro occasioni è pari a 3209.

Tabella 5.1: Numero di rispondenti, per gruppo panel.

Gruppo panel	Aprile 2002	Luglio 2002	Ottobre 2002	Gennaio 2003
Gruppo 1	529	0	0	0
Gruppo 2	0	330	0	0
Gruppo 3	0	0	285	0
Gruppo 4	0	0	0	482
Gruppo 5	234	0	0	234
Gruppo 6	437	437	0	0
Gruppo 7	0	433	433	0
Gruppo 8	0	0	479	479
Totale	1200	1200	1197	1195

Il questionario dell'indagine Forze Lavoro del Comune di Firenze inizia con un quesito sulla condizione occupazionale. Un individuo viene considerato occupato in una data occasione di indagine se si dichiara tale oppure se ha svolto ore di lavoro durante la settimana precedente all'intervista; lo stato occupazionale così definito costituisce un filtro per i quesiti successivi. Il questionario procede infatti con quesiti relativi al lavoro svolto e al reddito per le persone occupate, mentre a coloro che non risultano occupati vengono somministrate domande relative alla ricerca del lavoro.

In questo capitolo si affronta il problema dei dati mancanti per i quesiti relativi allo stato occupazionale e al reddito da lavoro per le persone occupate. In particolare, quando una persona viene intervistata i quesiti relativi allo stato occupazionale sono sempre osservati. A coloro che risultano occupati viene successivamente somministrato il quesito: "Mi può dire qual è il suo reddito netto medio mensile?", e per tale domanda si hanno dei missing values. I redditi considerati nell'indagine sono dunque solo quelli derivanti dal

lavoro corrente. Poichè gli altri quesiti non vengono presi in considerazione nella procedura di imputazione, come spiegato meglio nei prossimi paragrafi, in questo caso non risulta interessante procedere ad una completa descrizione del questionario dell'indagine[1].

La tabella 5.2 riporta, separatamente per ciascun gruppo panel, il numero di persone occupate (N) e la corrispondente percentuale di valori mancanti per il quesito relativo al reddito.

Tabella 5.2: Numero di persone occupate (N) e percentuale di valori mancanti per il reddito medio mensile, per gruppo panel.

Gruppo	Aprile 2002		Luglio 2002		Ottobre 2002		Gennaio 2003	
Panel	N	% missing	N	% missing	N	% missing	N	% missing
Gruppo 1	286	31.47	0	0	0	0	0	0
Gruppo 2	0	0	195	37.95	0	0	0	0
Gruppo 3	0	0	0	0	174	36.21	0	0
Gruppo 4	0	0	0	0	0	0	272	39.34
Gruppo 5	118	31.36	0	0	0	0	119	26.05
Gruppo 6	244	24.59	245	31.43	0	0	0	0
Gruppo 7	0	0	239	38.49	239	36.82	0	0
Gruppo 8	0	0	0	0	263	36.50	264	31.44
Totale oss.	648	28.86	679	35.79	676	36.54	655	33.74

Le percentuali di mancate risposte per il reddito medio mensile sono molto alte se confrontate con quelle di tutti gli altri quesiti dell'indagine, per i quali si hanno percentuali sempre inferiori al 3%. Questo è in linea con quanto avviene per molte altre indagini in cui vengono rilevate informazioni relative a redditi, patrimoni e variabili di tipo finanziario (Heeringa et al., 2002). E' da sottolineare che i valori zero in tabella 5.2 derivano dalla rotazione del panel; se gli individui entrati nel campione fossero stati tutti intervistati nelle quattro occasioni di indagine, al posto di tali zero avremmo potuto osservare i corrispondenti valori N e le relative percentuali di mancate risposte di reddito. Quindi, lo schema di rotazione rappresenta una fonte di dati mancati sia per lo stato occupazionale che per il reddito medio mensile, fonte che va ad aggiungersi al meccanismo che genera le mancate risposte degli individui intervistati.

Nel caso in cui, invece, il quesito relativo al reddito non venga somministrato ad un intervistato perchè questo non risulta occupato, allora il corrispondente valore del reddito non deve essere considerato mancante ma deve essere posto pari a zero.

[1]Il questionario è consultabile nella pubblicazione del Comune di Firenze a cura di Giommi et al. (2003), disponibile sul sito internet dell'Ufficio di Statistica http://statistica.comune.fi.it/.

Sia $Z_{ihj} = 0, 1$ $(i = 1, ..., n_h,\ h = 1, ..., H,\ j = 1, ..., J)$ l'indicatore dello stato occupazionale per il soggetto i appartenente allo strato h nell'occasione di indagine j. Y_{ihj} è il corrispondente reddito medio mensile in euro; se un individuo non è occupato ($Z_{ihj} = 0$), allora il reddito è pari a zero ($Y_{ihj} = 0$).

X_{ihp} $(p = 1, ..., P)$ è il vettore contenente le informazioni personali, costanti tra le occasioni di indagine, e precisamente il genere, la classe di età, il numero di componenti della famiglia, la zona di residenza nel Comune di Firenze, il livello di istruzione e lo stato civile. Tutte queste variabili sono categoriche[2].

Infine, sia w_h il peso campionario per tutti gli individui appartenenti allo strato h che, come già accennato, risulta definito da tre delle covariate osservate X_{ihp}: genere, classe di età e zona di residenza.

Per quanto riguarda invece le mancate risposte, sia T_{ihj} l'indicatore di mancata risposta per lo stato occupazionale Z_{ihj}: $T_{ihj} = 1$ se lo stato occupazionale è mancante, ovvero se l'individuo i dello strato h non viene intervistato nell'occasione j a causa della rotazione del panel. Sia invece M_{ihj} l'indicatore di mancata risposta per il reddito, con $M_{ihj} = 1$ se Y_{ihj} è missing, o altrimenti. Allora, M_{ihj} può assumere valore 1 a causa della rotazione del panel (l'individuo ih non viene intervistato nell'occasione j), oppure se la persona viene intervistata ma si rifiuta di rispondere. Le informazioni X_{ihp} sono invece considerate pienamente osservate, sebbene qualche valore sia mancante anche per queste variabili, con percentuali sempre inferiori al 2%. Tali valori mancanti vengono imputati attraverso lo stesso modello utilizzato per il reddito. Infine, i pesi w_h sono completamente osservati.

Stime trimestrali del reddito medio mensile possono essere ottenute utilizzando i soli casi disponibili per ciascuna delle occasioni j:

$$\widehat{\overline{Y}}_{..j} = \frac{\sum_{h=1}^{H} \sum_{i=1}^{n_h} Y_{ihj} Z_{ihj} w_h}{\sum_{h=1}^{H} \sum_{i=1}^{n_h} Z_{ihj} w_h} \tag{5.1}$$

La stima del relativo standard error può essere calcolata utilizzando le usuali procedure di linearizzazione utilizzate nel caso di disegni campionari complessi.

Oltre alle stime trimestrali del reddito medio mensile potrebbe essere interessante calcolarne il valore con un periodo di riferimento pari all'intero anno

[2]Per quanto riguarda le modalità delle variabili X_{ihp}, si ha: genere=(0=femmina, 1=maschio), classe di età=(1=0-14, 2=15-24, 3=25-34, 4=35-44, 5=45-54, 6=55-64, 7=65-74, 8=più di 75), numero di componenti della famiglia=(1,2,3,4,5,6=6 o più), zona di residenza in Firenze=(1,...,20), livello di istruzione=(0=nessuno, 1=scuola elementare, 2=scuola media inferiore, 3=diploma professionale, 4=scuola superiore, 5=laurea di primo livello, 6=laurea di secondo livello, 7=dottorato), stato civile=(1=mai sposato, 2=sposato, 3=divorziato, 4=vedovo).

2002. Tale stima può essere calcolata come media delle $\widehat{\overline{Y}}_{..j}$ tra le occasioni; tale metodo tuttavia non tiene in considerazione il diverso numero di occupati e di mancate risposte in ciascuna delle occasioni. In alternativa è possibile calcolare per ciascun individuo la media dei valori di reddito osservati, considerando solo coloro risultati occupati in almeno una delle occasioni:

$$\widehat{\overline{Y}}_{ih.} = \frac{\sum_{j=1}^{4} Y_{ihj} Z_{ihj}}{\sum_{j=1}^{4} Z_{ihj}}, \qquad \sum_{j=1}^{4} Z_{ihj} > 0 \tag{5.2}$$

e ottenere la stima totale come:

$$\widehat{\overline{Y}} = \frac{\sum_{h=1}^{H} \sum_{i=1}^{n_h} \widehat{\overline{Y}}_{ih.} w_h}{\sum_{h=1}^{H} \sum_{i=1}^{n_h} w_h}. \tag{5.3}$$

I risultati di queste *available case analyses* sono riportati nella tabella 5.3.

Tabella 5.3: Stima del reddito mensile medio (in euro) per gli occupati durante il 2002, per periodo di riferimento del reddito (ipotesi MCAR).

Periodo di riferimento	N	Reddito mensile medio	Standard Error
$\widehat{\overline{Y}}_{..1}$=Primo trimestre	461	1195.2	31.3
$\widehat{\overline{Y}}_{..2}$=Secondo trimestre	436	1186.8	26.6
$\widehat{\overline{Y}}_{..3}$=Terzo trimestre	429	1309.0	33.3
$\widehat{\overline{Y}}_{..4}$=Quarto trimestre	434	1234.3	26.8
$\widehat{\overline{Y}}$=Anno 2002	1327	1221.2	22.7

Dai risultati in tabella 5.3 si vede che la stima del reddito mensile cresce negli ultimi due quarti dell'anno, specialmente nel terzo. Il valore più basso è quello del secondo quarto, per il quale vengono utilizzati i dati rilevati a Luglio.

Questi risultati sono stati calcolati facendo l'assunzione che i dati siano mancanti completamente a caso (MCAR), ovvero che la loro mancanza non sia legata ai valori di reddito non osservati e nemmeno alle covariate osservate. Questa ipotesi sarebbe ammissibile per i valori mancanti dovuti alla rotazione del panel, ma risulta particolarmente forte nel caso in cui le mancate risposte siano dovute al rifiuto di rispondere. Ecco perchè risulta più appropriato, generalmente, calcolare stime consistenti utilizzando l'ipotesi MAR, più debole della MCAR, secondo la quale la distribuzione condizionata degli indicatori delle mancate risposte può dipendere dai dati osservati.

In questo caso l'ipotesi MAR può essere così formalizzata:

$$Pr(T_{ihj} = 1|Y_{hij}, Z_{ihj}, X_{hip}, \phi) = Pr(T_{ihj} = 1|Y_{\text{oss},hij}, Z_{\text{oss},ihj}, X_{ihp}, \phi) \quad (5.4)$$

$$Pr(M_{ihj} = 1|Y_{hij}, Z_{ihj}, X_{hip}, \psi) = Pr(M_{ihj} = 1|Y_{\text{oss},hij}, Z_{\text{oss},ihj}, X_{ihp}, \psi) \quad (5.5)$$

per tutti gli $Y_{\text{oss},ihj}, Z_{\text{oss},ihj}, X_{ihp}, \phi$ e ψ.

In queste formule $Y_{\text{oss},ihj}, Z_{\text{oss},ihj}$ rappresentano le componenti osservate delle variabili Y_{ihj}, Z_{ihj}, mentre $Y_{\text{mis},ihj}, Z_{\text{mis},ihj}$ quelle mancanti, secondo la simbologia già utilizzata nei capitoli precedenti.

5.3 La procedura di imputazione multipla

5.3.1 Imputazione multipla con ipotesi MAR

In questa sezione viene presentata una procedura di imputazione multipla per i dati mancanti relativi allo stato occupazionale e al reddito mensile, $Y_{\text{mis},ihj}$ e $Z_{\text{mis},ihj}$, sotto l'ipotesi che tutti i dati siano MAR.

Attraverso l'imputazione multipla vengono prodotti m dataset completi, con i valori mancanti sostituiti da estrazioni dalla loro distribuzione predittiva a posteriori sotto un particolare modello di imputazione (capitolo 2). Per trattare la natura multivariata dei dati mancanti e osservati, e per condizionare rispetto a tutte le informazioni, anche per questa applicazione si utilizza l'approccio delle regressioni sequenziali multivariate (Raghunathan et al., 2001; Van Buuren and Oudshoorn, 1999). Tale approccio evita la specificazione di un modello congiunto multivariato per le variabili, compito che può essere particolarmente arduo quando le covariate sono molte e hanno forme distributive diverse (paragrafo 2.2.2).

Nel metodo di imputazione utilizzato non viene fatta distinzione, per il momento, tra un valore di reddito Y_{ihj} mancante perchè l'individuo i dello strato h non è stato intervistato nell'occasione j oppure perchè l'individuo è stato intervistato ma si è rifiutato di rispondere. Lo schema del modello di imputazione è rappresentato nella tabella 5.4.

Come si vede, l'ipotesi MAR relativa alle variabili Y_{ihj} condiziona rispetto a Z_{ihj} e, quindi, al suo indicatore T_{ihj}: questo significa che il modello per Y_{ihj} è MAR sia quando lo stato occupazionale è osservato ($T_{ihj} = 0$) sia quando è missing ed è stato imputato il valore $Z_{ihj} = 1$.

Sotto l'ipotesi MAR l'imputazione multipla delle variabili Y e Z attraverso il metodo delle regressioni sequenziali procede così. Viene scelto un modello di regressione per ciascuna delle variabili con valori mancanti: in

Tabella 5.4: Schema del modello di imputazione per il reddito sotto l'ipotesi MAR.

Gruppo	Aprile 2002		Luglio 2002		Ottobre 2002		Gennaio 2003	
Panel	Z_{hi1}	Y_{hi1}	Z_{hi2}	Y_{hi2}	Z_{hi3}	Y_{hi3}	Z_{hi4}	Y_{hi4}
Gruppo 1	Oss	Oss/MAR	MAR	MAR	MAR	MAR	MAR	MAR
Gruppo 2	MAR	MAR	Oss	Oss/MAR	MAR	MAR	MAR	MAR
Gruppo 3	MAR	MAR	MAR	MAR	Oss	Oss/MAR	MAR	MAR
Gruppo 4	MAR	MAR	MAR	MAR	MAR	MAR	Oss	Oss/MAR
Gruppo 5	Oss	Oss/MAR	MAR	MAR	MAR	MAR	Oss	Oss/MAR
Gruppo 6	Oss	Oss/MAR	Oss	Oss/MAR	MAR	MAR	MAR	MAR
Gruppo 7	MAR	MAR	Oss	Oss/MAR	Oss	Oss/MAR	MAR	MAR
Gruppo 8	MAR	MAR	MAR	MAR	Oss	Oss/MAR	Oss	Oss/MAR

questo caso una regressione logistica per la variabile dummy relativa allo stato occupazionale e una regressione lineare per il logaritmo del reddito. Per i parametri di queste regressioni vengono scelte delle distribuzioni a priori non informative. Durante il primo passo della procedura viene stimata una regressione di $Z_{\text{oss},ihj}$ sulle covariate X_{ihp}, e i valori mancanti vengono imputati dalla corrispondente distribuzione predittiva a posteriori; poi, viene stimata la regressione di $\ln(Y_{\text{oss},ihj})$ su X_{ihp} e sui dati completati nel passo precedente, Z_{ihj}, e anche i valori $Y_{\text{mis},ihj}$ vengono imputati. In questa applicazione, se $Z_{ihj} = 0$, ovvero se l'individuo non è occupato, il corrispondente valore imputato per Y_{ihj} assume valore 0.

Questa procedura viene poi ripetuta durante più cicli, utilizzando ogni volta come predittori le covariate osservate e i valori imputati nel ciclo precedente; il procedimento si ferma non appena si ottengono imputazioni stabili per tutte le variabili. Poichè il missing dei dati non è monotono tale procedura di imputazione utilizza un algoritmo di tipo Gibbs sampling (paragrafo 2.2.2).

Se questo intero procedimento viene ripetuto m volte si ottengono m dataset completi. Poi, su ogni dataset vengono condotte le analisi di interesse ed i risultati vengono combinati con le regole di Rubin (paragrafo 2.1.1).

Relativamente all'utilizzo come variabili esplicative delle due grandezze interessate dai missing values, è stato ipotizzato che la distribuzione condizionata di ciascun ammontare di reddito in una data occasione d'indagine dipende solo dal reddito e dallo stato occupazionale dell'occasione o delle due occasioni più vicine. Questo vuol dire che per un dato individuo il reddito relativo al mese di Aprile dipende solo dalle informazioni sul reddito e sullo stato occupazionale in Luglio, quello di Luglio solo dalle informazioni in Aprile e Ottobre, quello in Ottobre dalle informazioni in Luglio e Gennaio e, infine, il reddito in Gennaio dipende solo dal reddito e dallo stato occu-

pazionale in Ottobre. E' comunque chiaro che, nonostante tali restrizioni, la procedura iterata del metodo di imputazione fa sì che alla fine tutte le imputazioni si influenzino a vicenda.

In particolare, all'iterazione t dell'algoritmo le imputazioni per il logaritmo del reddito $\ln(Y_{ihj})$ relativo all'individuo i dello strato h nell'occasione j vengono estratte dalle distribuzioni:

$$
\begin{aligned}
& f\left[\ln(Y_{ih1})|Z_{ih1}^{t}, \ln(Y_{ih2})^{t-1}, Z_{ih2}^{t-1}, X_{ihp}, \sigma_{11}^{t}, \beta_{1}^{t}, T_{ih1}, M_{ih1}, T_{ih2}, M_{ih2}\right] \\
& f\left[\ln(Y_{ih2})|\ln(Y_{ih1})^{t}, Z_{ih1}^{t}, Z_{ih2}^{t}, \ln(Y_{ih3})^{t-1}, Z_{ih3}^{t-1}, X_{ihp}, \sigma_{22}^{t}, \beta_{2}^{t}, \ldots\right. \\
& \qquad \left.\ldots T_{ih1}, M_{ih1}, T_{ih2}, M_{ih2}, T_{ih3}, M_{ih3}\right] \\
& f\left[\ln(Y_{ih3})|\ln(Y_{ih2})^{t}, Z_{ih2}^{t}, Z_{ih3}^{t}, \ln(Y_{ih4})^{t-1}, Z_{ih4}^{t-1}, X_{ihp}, \sigma_{33}^{t}, \beta_{3}^{t}, \ldots\right. \\
& \qquad \left.\ldots T_{ih2}, M_{ih2}, T_{ih3}, M_{ih3}, T_{ih4}, M_{ih4}\right] \\
& f\left[\ln(Y_{ih4})|Z_{ih4}^{t}, \ln(Y_{ih3})^{t-1}, Z_{ih3}^{t-1}, X_{ihp}, \sigma_{44}^{t}, \beta_{4}^{t}, T_{ih3}, M_{ih3}, T_{ih4}, M_{ih4}\right]
\end{aligned}
$$

In queste espressioni $Z_{ihj}^{t} = Z_{ihj}$ e $Y_{ihj}^{t} = Y_{ihj}$ se i valori sono osservati ($T_{ihj} = 1$, $M_{ihj} = 1$), e il condizionamento rispetto a Z_{ihj} determina il valore $Y_{ihj} = 0$ se $Z_{ihj} = 0$. Le distribuzioni sono ipotizzate normali, e le corrispondenti distribuzioni a priori dei parametri sono non informative, ovvero del tipo $g(\beta_j, \sigma_{jj}) = \sigma_{jj}^{-1/2}$.

L'approccio all'imputazione multipla attraverso il metodo delle regressioni sequenziali non è esente da problemi (paragrafo 2.2.2). Come già evidenziato, vari autori hanno tuttavia mostrato che il metodo funziona bene nella pratica (Van Buuren et al., 2006; Heeringa et al., 2002); per tale motivo le imputazioni ottenute sono state verificate prima di procedere con le analisi.

In particolare, sono stati imputati $m = 25$ datasets attraverso il pacchetto "Ice" del software Stata (Royston, 2005). Valori inferiori di m sono solitamente sufficienti quanto il tasso di mancata risposta è molto basso (Rubin, 1987): in questo caso è stato necessario utilizzare un m più elevato in quanto la rotazione del panel induce un alto tasso di mancata risposta (tabella 5.5). Il numero $m = 25$ ha garantito stime stabili per la componente between della varianza derivante dal procedimento di imputazione multipla.

Lo schema di imputazione MAR (equazioni (5.4) e (5.5)) ha richiesto la scelta di un insieme di covariate X a cui condizionarsi nel modello di imputazione. Per non complicare troppo il modello si è scelto di utilizzare solo le covariate costanti durante tutte le occasioni di indagine, ovvero quelle già presentate nel paragrafo 5.2, non utilizzando invece informazioni come il tipo di lavoro e il tipo di contratto, che essendo rilevate attraverso il questionario possono variare tra le occasioni e presentare a loro volta dei valori mancanti. Anche queste covariate potrebbero comunque essere inserite nel modello senza variazioni sostanziali.

Come detto, particolare attenzione è stata dedicata alla verifica della convergenza dell'algoritmo di imputazione. Per quanto riguarda lo stato occupazionale, i valori imputati risultano molto influenzati da quelli osservati. Per esempio, se un soggetto è stato intervistato in due occasioni ed è risultato (non) occupato in entrambe, allora il suo stato occupazionale è stato imputato come (non) occupato anche nelle altre due occasioni nel 95% dei casi (valore medio tra le 25 imputazioni multiple). Quanto invece lo stato occupazionale variava tra le due occasioni, le imputazioni risultano anch'esse più variabili. Infine, quando un solo stato occupazionale risultava osservato, quello stesso stato occupazionale è stato imputato per le restanti 3 occasioni di indagine nell'85% dei casi. Il numero medio di persone occupate tra le 25 imputazioni e le corrispondenti percentuali di valori di reddito missing sono riportate nella tabella 5.5. Ovviamente, quando lo stato occupazionale è missing per la rotazione del panel, il reddito corrispondente è sempre mancate (percentuali pari al 100%). Si vede allora che le percentuali totali di valori mancanti per il reddito sono molto elevate, sempre intorno al 75%.

Tabella 5.5: Numero di persone occupate (N) e percentuale di valori mancanti per il reddito medio mensile, valori medie tra i 25 datasets.

Gruppo	Aprile 2002		Luglio 2002		Ottobre 2002		Gennaio 2003	
Panel	N	% missing	N	% missing	N	% missing	N	% missing
Gruppo 1	286	31.47	291	100	290	100	288	100
Gruppo 2	192	100	195	37.95	192	100	191	100
Gruppo 3	168	100	172	100	174	36.21	166	100
Gruppo 4	267	100	272	100	271	100	272	39.34
Gruppo 5	118	31.36	120	100	120	100	119	26.05
Gruppo 6	244	24.59	245	31.43	246	100	244	100
Gruppo 7	235	100	239	38.49	239	36.82	239	100
Gruppo 8	260	100	264	100	263	36.50	264	31.44
Totale	1771	73.91	1797	75.79	1795	76.10	1783	75.66

Per quanto riguarda l'imputazione del reddito, per effettuare un controllo dei valori imputati si sono considerati gli individui intervistati in due occasioni (gruppi panel 5, 6, 7 e 8); per questi si è andati a confrontare la relazione esistente tra le coppie di valori di reddito osservati, disponibili per alcuni soggetti, con la relazione tra il valore di reddito osservato e quello imputato per gli individui che nelle due occasioni avevano dichiarato un solo valore di reddito.

Alcuni di questi scatterplots[3] sono riportati nella figura 5.1; come si vede,

[3]Per avere una migliore comparazione, negli scatterplots sono stati esclusi i valori di reddito osservati che superavano i 5000 euro.

la correlazione positiva tra i valori osservati sembra essere preservata anche dalle imputazioni, e questo risultato è piuttosto stabile tra le imputazioni.

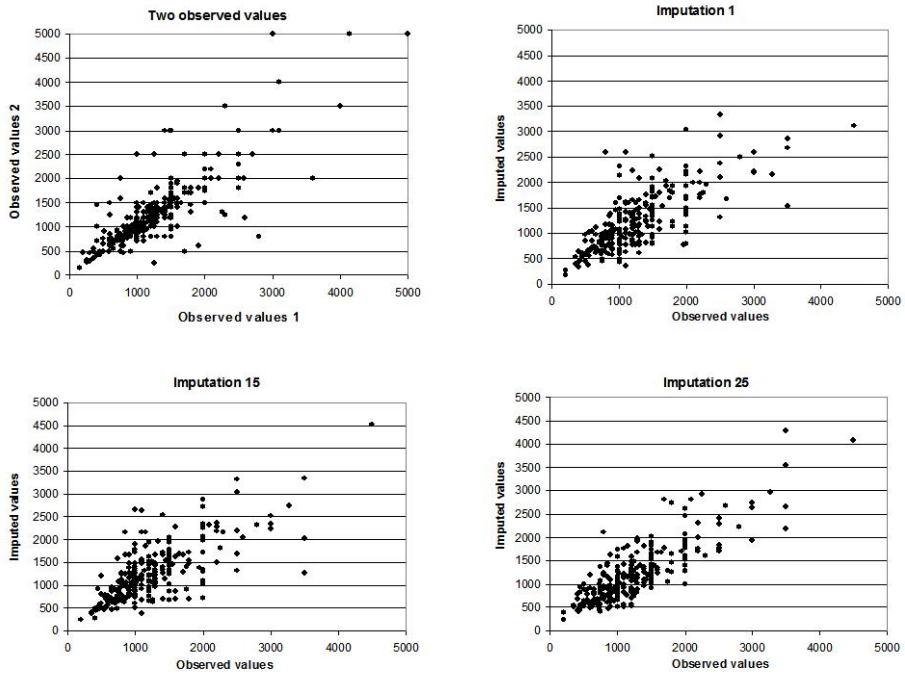

Figura 5.1: Scatterplots dei valori di reddito osservati ed imputati.

Dopo queste verifiche sono state ricalcolate le stime di interesse (5.1) e (5.3), già considerate sotto l'ipotesi MCAR, per valutare le eventuali differenze rispetto all'ipotesi MAR. Inoltre, come risultato del metodo di imputazione impiegato, è stata definita una stima del reddito totale annuale. Più in particolare, prendendo in considerazione gli individui occupati in tutte le quattro occasioni di indagine ($Z_{hij} = 1 \ \forall j = 1, ..., 4$), le stime individuali del reddito annuale nel 2002 sono date da:

$$\widehat{Y}_{hi\ 2002} = \sum_{j=1}^{4} Y_{hij} * 3. \tag{5.6}$$

Allora, la stima totale del reddito annuale in tutta la popolazione si può ricavare come:

$$\widehat{Y}_{2002} = \frac{\sum_{h=1}^{H} \sum_{i=1}^{n_h} \widehat{Y}_{hi\ 2002} w_h}{\sum_{h=1}^{H} \sum_{i=1}^{n_h} w_h} \tag{5.7}$$

Utilizzando le regole di Rubin, le stime dei valori medi di reddito, trimestrali ed annuali, calcolati in ciascuno dei 25 datasets sono state combinate

tra loro; poichè si sono presi in considerazione anche alcuni percentili delle distribuzioni, in questo caso le varianze sono state calcolate utilizzando la tecnica del *bootstrap*, estraendo 200 campioni da ciascuno dei datasets con un campionamento casuale semplice negli strati.

Inoltre, per ciascuna delle stime è stata calcolata la *fraction of missing information* (paragrafo 2.1.1), una misura del grado in cui i valori mancanti contribuiscono all'incertezza delle inferenze relative alla quantità di interesse.

Tabella 5.6: Stime del reddito medio mensile in euro durante il 2002 (ipotesi MAR).

Stima	$\widehat{\overline{Y}}_{..1}$	$\widehat{\overline{Y}}_{..2}$	$\widehat{\overline{Y}}_{..3}$	$\widehat{\overline{Y}}_{..4}$
N	1771	1797	1795	1783
Reddito medio mensile	1198.5	1201.2	1244.6	1236.9
Standard Error	28.6	22.2	25.0	29.5
% missing info.	69.01	54.95	49.18	75.89

I risultati delle stime trimestrali sotto l'ipotesi MAR sono riportati nella tabella 5.6. Rispetto ai risultati ottenuti con l'ipotesi MCAR (tabella 5.3) le differenze tra i redditi mensili medi riferiti ai diversi trimestri si riducono leggermente. In ogni caso le stime puntuali più elevate continuano ad essere quelle relative agli ultimi due trimestri, e questo nonostante il corrispondente numero di persone occupate non sia superiore (tabella 5.5). La *fraction of missing information* risulta molto differenziata tra le stime; attraverso un'analisi più approfondita si è verificato che i valori elevati relativi al primo e quarto trimestre dipendono da alcuni valori di reddito particolarmente elevati riportati dagli intervistati in Aprile e Gennaio. Nelle altre due occasioni la *fraction of missing information* risulta invece molto inferiore rispetto al tasso di mancate risposte (tabella 5.5); ciò dipende dall'informazione contenuta nelle regressioni del modello di imputazione.

I risultati delle due stime riferite all'intero anno 2002, $\widehat{\overline{Y}}$ e $\widehat{Y}_{2002}$, sono riportati nella tabella 5.7. Relativamente a queste grandezze si sono calcolate anche la mediane ed il ventesimo ed ottantesimo percentile.

La stima del reddito medio mensile relativa a tutto il 2002 risulta sotto l'ipotesi MAR inferiore a quella sotto l'ipotesi MCAR (tabella 5.3), anche se la differenza non è significativa andando a considerare il relativo standard error.

Tabella 5.7: Reddito mensile riferito a tutto il 2002 e reddito annuale in euro (ipotesi MAR).

Stima	$\widehat{\overline{Y}}$ (reddito mensile)	$\widehat{Y}_{2002}$ (reddito annuale)
N	1968	1585
Media	1199.5	14874.3
S.E. media	17.2	235.0
% missing info.	45.62	48.30
Mediana	1047.3	13331.0
S.E. mediana	14.6	208.7
20esimo percentile	764.4	9704.3
S.E. 20esimo percentile	16.3	188.0
80esimo percentile	1530.0	18773.8
S.E. 80esimo percentile	26.6	321.1

5.3.2 Analisi di sensitività per deviazioni dall'ipotesi MAR

In questo paragrafo viene descritto il procedimento attraverso cui i risultati ottenuti con l'analisi MAR vengono modificati per studiarne la sensitività rispetto a meccanismi di mancata risposta non MAR. Per questa indagine i valori di reddito che risultano mancanti a causa della rotazione del panel non possono essere considerati MNAR, mentre questo è plausibile per i missing values dovuti a rifiuto di rispondere. In particolare, l'analisi di sensitività che viene presentata introduce degli *offsets* ai valori di reddito imputati sotto ipotesi MAR che erano originariamente mancati per rifiuto di rispondere; la grandezza di questi offsets è una frazione predeterminata della deviazione standard residua del valore mancante, specificata meglio più avanti, e in quanto tale dipende dalle informazioni osservate, che possono essere anche molto differenziate tra gli individui. Nel caso in cui un individuo sia stato intervistato in due occasioni (gruppi panel 5, 6, 7 e 8) e sia risultato impiegato in entrambe le occasioni, è infatti possibile che si sia rifiutato di dichiarare il suo reddito due volte, una sola volta o nessuna.

Per esempio, nel gruppo panel numero 5 il reddito è sempre mancante nella seconda e terza occasione a causa della rotazione del panel. Questo determina quattro possibili pattern per M_{ihj}, l'indicatore delle mancate risposte del reddito: 0110, 1110, 0111, 1111. I soggetti appartenenti al pattern 0110 hanno dichiarato il loro reddito nelle due occasioni in cui sono stati intervistati (prima e quarta), mentre quelli nel pattern 1110 si sono rifiutati di dichiarare il loro reddito alla prima ma non alla quarta occasione, e così

via.

Nell'analisi di sensitività, come detto, si è scelto di applicare l'offset solo nel caso delle "vere" non risposte, ovvero quando il reddito è mancante per rifiuto a rispondere. Quando invece il reddito risulta mancante per la rotazione del panel, l'ipotesi resta quella di dati MAR, come indicato nella tabella 5.8. Questa analisi di sensitività segue dunque le indicazioni di Little (2005) di specificare l'ipotesi MNAR in base al motivo per cui il valore risulta mancante.

Tabella 5.8: Schema di imputazione del reddito sotto le ipotesi MAR e MNAR.

Gruppo	Aprile 2002		Luglio 2002		Ottobre 2002		Gennaio 2003	
Panel	Z_{hi1}	Y_{hi1}	Z_{hi2}	Y_{hi2}	Z_{hi3}	Y_{hi3}	Z_{hi4}	Y_{hi4}
Gruppo 1	Oss	Oss/MNAR	MAR	MAR	MAR	MAR	MAR	MAR
Gruppo 2	MAR	MAR	Oss	Oss/MNAR	MAR	MAR	MAR	MAR
Gruppo 3	MAR	MAR	MAR	MAR	Oss	Oss/MNAR	MAR	MAR
Gruppo 4	MAR	MAR	MAR	MAR	MAR	MAR	Oss	Oss/MNAR
Gruppo 5	Oss	Oss/MNAR	MAR	MAR	MAR	MAR	Oss	Oss/MNAR
Gruppo 6	Oss	Oss/MNAR	Oss	Oss/MNAR	MAR	MAR	MAR	MAR
Gruppo 7	MAR	MAR	Oss	Oss/MNAR	Oss	Oss/MNAR	MAR	MAR
Gruppo 8	MAR	MAR	MAR	MAR	Oss	Oss/MNAR	Oss	Oss/MNAR

Nel modello MNAR le equazioni 5.4 e 5.5 diventano:

$$Pr(T_{ihj} = 1|Y_{hij}, Z_{ihj}, X_{hip}, \phi) = Pr(T_{ihj} = 1|Y_{\text{oss},hij}, Z_{\text{oss},ihj}, X_{ihp}, \phi)$$

$$Pr(M_{ihj} = 1|Y_{ihj}, Z_{ihj}, X_{ihp}, \psi) = \begin{cases} Pr(M_{ihj} = 1|Y_{\text{oss},hij}, Z_{\text{oss},ihj}, X_{ihp}, \psi) \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{se } T_{ihj} = 1 \\ Pr(M_{ihj} = 1|Y_{hij}, Z_{\text{oss},ihj}, X_{ihp}, \psi) \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{se } T_{ihj} = 0 \end{cases}$$

per tutti gli $Y_{ihj}, Z_{\text{oss},ihj}, X_{ihp}, \phi$ e ψ.

Relativamente alla specificazione dell'ipotesi MNAR sono stati considerati due diversi modelli, indicati con MNAR_1 e MNAR_2. In entrambi i modelli i valori di reddito ottenuti sotto l'ipotesi MAR vengono modificati con l'aggiunta di una quantità al loro valore in scala logaritmica: introdurre questo tipo di offset corrisponde ad ipotizzare che il reddito sia sottostimato sotto l'ipotesi MAR. Per quanto riguarda la quantità aggiunta, come detto per il logaritmo di ogni valore imputato con ipotesi MAR si è considerata una frazione della deviazione standard residua della regressione sugli altri valori imputati e sui valori osservati:

$$\ln(y_{ihj,\text{MNAR}_1}) = \ln(y_{ihj,\text{MAR}}) + k * \widehat{\sigma}_{jj}. \tag{5.8}$$

Più specificamente, $\widehat{\sigma}_{11}$ proviene dalla regressione di $\ln(y_{ih1,\text{MAR}})$ su X_{ihp} e Z_{ihj}, $\widehat{\sigma}_{22}$ dalla regressione di $\ln(y_{ih2,\text{MAR}})$ su $\ln(y_{ih1,\text{MNAR}_1}), X_{ihp}$ e Z_{ihj}, $\widehat{\sigma}_{33}$ dalla regressione di $\ln(y_{ih3,\text{MAR}})$ su $\ln(y_{ih1,\text{MNAR}_1})$,$\ln(y_{ih2,\text{MNAR}_1})$,$X_{ihp}$ e Z_{ihj} ed infine $\widehat{\sigma}_{44}$ proviene dalla regressione di $\ln(y_{ih4,\text{MAR}})$ su $\ln(y_{ih1,\text{MNAR}_1})$, $\ln(y_{ih2,\text{MNAR}_1})$, $\ln(y_{ih3,\text{MNAR}_1})$ e su X_{ihp} e Z_{ihj}.

Così facendo la distribuzione del reddito sotto l'ipotesi MAR risulta traslata, senza che questa modifica faccia parte dell'algoritmo di imputazione; questo fa sì che l'incremento introdotto non risulti amplificato dalle iterazioni dello schema di imputazione (Van Buuren et al., 1999). Per quanto riguarda k, i valori scelti per studiare la sensitività delle stime rispetto al modello MNAR_1 sono $k = 0.2, 0.5, 0.7$.

Il modello MNAR alternativo che è stato considerato, MNAR_2, è uguale al MNAR_1, a parte il fatto che in questo caso l'offset viene aggiunto solo quando non sono disponibili valori di reddito osservati. Per esempio, per un individuo che è stato intervistato alla prima e seconda occasione (gruppo panel numero 6 nella tabella 5.8) ed è risultato impiegato in entrambe sono possibili i seguenti quattro pattern per M_{ihj}: 0011, 1011, 0111, 1111. Il modello MNAR_2 modifica i valori ottenuti sotto l'ipotesi MAR solo nel pattern 1111, ovvero quando il reddito è mancate nelle due occasioni di intervista. Il modello MNAR_1, invece, modifica i valori con l'offset anche nei pattern 1011 e 0111, quando è presente un valore osservato di reddito.

Quindi, il meccanismo MNAR_2 considera i valori di reddito che sono missing per rifiuto a rispondere come mancanti a caso (MAR) per i soggetti che hanno dichiarato almeno una volta il loro reddito; il meccanismo MNAR_2 è dunque più simile al MAR rispetto al meccanismo MNAR_1. I risultati dei modelli MNAR_1 e MNAR_2 rappresentano, per ogni valore di k, una possibile combinazione di un meccanismo di tipo MNAR con uno MAR.

Le stime del reddito medio mensile nei quattro trimestri del 2002 sotto le ipotesi MNAR_1 e MNAR_2 e per i valori $k = 0.2, 0.5, 0.7$ sono riportate nella tabella 5.9.

Anche in questo caso, come già per i risultati sotto l'ipotesi MAR (tabella 5.6), le differenze del reddito medio non sono molto accentuate rispetto ai valori ottenuti con ipotesi MCAR (tabella 5.3). Tuttavia, con entrambi i modelli MNAR sembra evidenziarsi ulteriormente il distacco tra le stime relative ai primi due trimestri del 2002 e quelle relative agli ultimi due; in particolare, le stime $\widehat{\overline{Y}}_{..3}$ e $\widehat{\overline{Y}}_{..4}$ risultano molto simili tra loro per ogni valore di k.

Inoltre, come era logico attendersi, il meccanismo MNAR_1 ha un impatto più forte nel modificare le stime rispetto a quello MNAR_2; anche in questo caso, tuttavia, i risultati sono comunque molto simili se si tengono in

Tabella 5.9: Stime del reddito medio mensile in euro (ipotesi MNAR_1 e MNAR_2).

	Stima	$\widehat{\overline{Y}}_{..1}$ (First quarter)	$\widehat{\overline{Y}}_{..2}$ (Second quarter)	$\widehat{\overline{Y}}_{..3}$ (Third quarter)	$\widehat{\overline{Y}}_{..4}$ (Fourth quarter)
MNAR_1, $k = 0.2$	N	1771	1797	1795	1783
	Reddito medio mensile	1208.2	1208.3	1251.1	1247.1
	Standard Error	29.0	22.5	25.0	29.8
	% missing info.	69.39	55.29	48.78	76.01
MNAR_1, $k = 0.5$	N	1771	1797	1795	1783
	Reddito medio mensile	1224.1	1219.7	1261.4	1263.6
	Standard Error	29.7	23.1	25.0	30.5
	% missing info.	69.83	55.72	48.07	76.07
MNAR_1, $k = 0.7$	N	1771	1797	1795	1783
	Reddito medio mensile	1235.6	1227.9	1268.7	1275.5
	Standard Error	30.3	23.5	25.0	31.0
	% missing info.	70.03	55.98	47.70	76.01
MNAR_2, $k = 0.2$	N	1771	1797	1795	1783
	Reddito medio mensile	1205.5	1205.6	1249.0	1244.5
	Standard Error	28.9	22.3	24.0	29.7
	% missing info.	69.37	55.35	48.84	75.99
MNAR_2, $k = 0.5$	N	1771	1797	1795	1783
	Reddito medio mensile	1217.1	1212.6	1256.1	1256.7
	Standard Error	29.4	22.6	25.0	30.2
	% missing info.	69.83	55.96	48.24	76.05
MNAR_2, $k = 0.7$	N	1771	1797	1795	1783
	Reddito medio mensile	1225.4	1217.5	1261.0	1265.6
	Standard Error	29.8	22.8	25.0	30.6
	% missing info.	70.06	56.36	47.82	76.01

considerazione i relativi standard errors.

Le stime annuali $\widehat{\overline{Y}}$ e $\widehat{Y}_{2002}$ sotto le due ipotesi MNAR sono riportate nelle tabelle 5.10 e 5.11. Il valore del reddito mensile riferito all'intero anno 2002 è uguale a quello ottenuto sotto l'ipotesi MCAR, 1221 euro, se si ipotizza il modello MNAR_1 con $k = 0.5$, mentre è superiore per $k = 0.7$ sotto entrambe le ipotesi MNAR.

Tabella 5.10: Stime del reddito mensile ed annuale riferite all'intero 2002 in euro (ipotesi MNAR_1).

	MNAR_1					
	$k = 0.2$		$k = 0.5$		$k = 0.7$	
Stima	$\widehat{\overline{Y}}$	$\widehat{Y}_{2002}$	$\widehat{\overline{Y}}$	$\widehat{Y}_{2002}$	$\widehat{\overline{Y}}$	$\widehat{Y}_{2002}$
N	1968	1585	1968	1585	1968	1585
Media	1207.8	14975.2	1221.2	15138.1	1230.9	15255.4
S.E. media	17.4	237.0	17.7	240.4	18.9	243.0
% missing info.	45.92	48.43	46.38	48.57	48.67	46.72
Mediana	1081.2	13404.9	1091.8	13535.9	1099.0	13637.5
S.E. mediana	14.6	220.8	15.0	224.1	15.6	218.2
20esimo percentile	769.9	9771.4	777.7	9855.3	783.4	9914.0
S.E. 20esimo percentile	16.5	183.9	15.6	192.8	16.1	197.8
80esimo percentile	1542.2	18924.4	1560.2	19148.0	1572.3	19315.5
S.E. 80esimo percentile	27.5	324.0	27.8	337.0	27.3	349.5

Tabella 5.11: Stime del reddito mensile ed annuale riferite all'intero 2002 in euro (ipotesi MNAR_2).

	MNAR_2					
	$k = 0.2$		$k = 0.5$		$k = 0.7$	
Stima	$\widehat{\overline{Y}}$	$\widehat{Y}_{2002}$	$\widehat{\overline{Y}}$	$\widehat{Y}_{2002}$	$\widehat{\overline{Y}}$	$\widehat{Y}_{2002}$
N	1968	1585	1968	1585	1968	1585
Media	1205.4	14943.5	1215.0	15055.1	1221.9	15135.3
S.E. Media	17.4	236.2	17.6	238.3	17.7	240.0
% missing info.	46.01	48.52	46.63	48.84	47.05	49.04
Mediana	1080.0	13382.2	1087.0	13477.9	1091.7	13546.7
S.E. Mediana	14.9	217.3	15.2	219.0	15.5	222.0
20esimo percentile	768.8	9752.5	774.2	9814.6	778.1	9854.3
S.E. 20esimo percentile	16.3	185.6	15.5	188.1	15.6	189.6
80esimo percentile	1539.2	18870.6	1550.6	19025.8	1559.6	19135.6
S.E. 80esimo percentile	27.7	320.8	28.0	340.1	27.7	346.8

Per quanto riguarda invece la stima del reddito annuale, questa supera i 15000 euro sotto entrambi i modelli MNAR per $k = 0.5$ e 0.7. In ogni caso,

specie considerando i valori degli standard errors, si può concludere che in questa applicazione le differenze tra i valori di reddito considerando mancate risposte di tipo MNAR piuttosto che MAR risultano piuttosto ridotte; ovvero, le stime del reddito sono piuttosto insensibili alle deviazioni MNAR considerate.

5.4 Conclusioni

In questo capitolo è stato descritto l'utilizzo dell'imputazione multipla per imputare i dati mancanti di reddito in un'indagine con un disegno di tipo "panel-ruotato" in cui il percepimento del reddito ed il valore del reddito stesso sono mancanti sia nelle occasioni in cui l'individuo non viene intervistato, che per rifiuto o inabilità nel rispondere.

Inoltre è stata descritta un'analisi di sensitività per deviazioni dall'ipotesi MAR basata su *offsets* applicati alle imputazioni del modello MAR, nella forma di frazioni k della deviazione standard residua del modello log-normale.

L'analisi di sensitività ha indicato che i valori di reddito in questa applicazione sono piuttosto robusti per una gamma di valori k. Questo approccio ha il vantaggio di essere semplice e trasparente, dal momento che le deviazioni dall'ipotesi MAR sono facilmente comprensibili e non dipendono da ipotesi strutturali complesse, come avviene solitamente nei *selection models*.

In particolare, in questo caso specificare un offset è più realistico che cercare di stimare ipotesi strutturali dal momento che l'evidenza proveniente dai dati relativamente a deviazioni dall'ipotesi MAR è molto limitata. La proposta di utilizzare quale offset una frazione della deviazione standard residua possiede l'utile proprietà che gli offset tengono in considerazione la relazione con le covariate osservate ed imputate nelle altre occasioni di indagine, aspetto questo particolarmente importante nella presente applicazione. Modellando il reddito sulla scala logaritmica, inoltre, l'offset può essere interpretato approssimativamente come una variazione percentuale sulla scala di origine, ed è quindi facile da interpretare.

Conclusioni

Le problematiche metodologiche ed applicative tuttora irrisolte relativamente all'imputazione di dati mancanti di reddito sono molte.

Una delle ipotesi più criticate è quella relativa all'ignorabilità del meccanismo che genera le mancate risposte. Nel capitolo 3 si sono presentate le conclusioni, tra loro contrastati, cui sono giunti gli studiosi americani in merito ai dati provenienti dalla Current Population Survey. Anche se ipotizzare che i dati di reddito siano mancanti "a caso", ovvero che non esista nessun legame tra la probabilità che il dato sia mancante ed il valore del reddito stesso, può sembrare molto restrittivo, in alcuni casi non è stata individuata alcuna evidenza empirica contro tale ipotesi. Probabilmente è anche per questo motivo che l'imputazione multipla per dati mancanti di reddito secondo modelli di regressione basati sull'ipotesi MAR sta conoscendo negli Stati Uniti una rapida diffusione. In Italia invece, almeno per il momento, l'imputazione multipla non è mai stata utilizzata per trattare le mancate risposte in indagini di tipo "ufficiale"; inoltre, in Italia non si è assistito ad un dibattito altrettanto acceso relativamente al meccanismo che genera le mancate risposte a quesiti di reddito.

Come evidenziato nel corso dei capitoli 1 e 2, l'ipotesi di dati MAR può essere resa più plausibile attraverso una scelta accurata delle variabili osservate da inserire nel procedimento di imputazione. Inoltre, come suggerito da vari autori, l'utilizzo di ipotesi di tipo MAR andrebbe sempre accompagnato, se possibile, da una sua verifica; questo può risultare fattibile, tuttavia, solo se è possibile venire a conoscenza, in un secondo momento, dei dati inizialmente mancanti, per esempio attraverso re-interviste o da fonti alternative di dati. Negli altri casi, poichè la stima di modelli di tipo MNAR può risultare spesso complicata, una procedura opportuna riguarda l'implementazione di studi della sensitività rispetto a deviazioni dall'ipotesi MAR.

Nelle due applicazioni di questa tesi le variabili utilizzate come esplicative nel procedimento di imputazione sono state accuratamente selezionate, tenendo in considerazione anche le caratteristiche del *pattern* dei dati mancanti. In entrambe le applicazioni della tesi, inoltre, si è scelto di realizzare

l'imputazione multipla dei valori mancanti. L'imputazione multipla è stata inizialmente proposta da Rubin (1978) come metodo per ottenere inferenze corrette per datasets provenienti da grandi indagini campionarie. Nei capitoli 1 e 2 della tesi si sono presentate le principali caratteristiche e proprietà dell'imputazione multipla, collocandola nell'ottica originaria di Rubin, ovvero quella dell'inferenza per popolazioni finite basata su modello. In particolare, nel capitolo 2 sono stati presentati alcuni metodi che consentono di realizzare imputazioni multiple secondo modelli bayesiani; tra questi vi è anche l'approccio delle regressioni sequenziali multivariate. Tale metodo, utilizzato nelle due applicazioni della tesi, non specifica un unico modello per tutti i dati, che sarebbe spesso molto difficile da ipotizzare, ma lavora *variable by variable*. Nonostante il suo ampio utilizzo, per il momento questo approccio viene presentato solo parzialmente nelle monografie dedicate all'imputazione dei dati mancanti; ecco perchè nella tesi si è scelto di presentarlo dettagliatamente, assieme alle sue principali problematiche che sono attualmente oggetto della ricerca statistica.

Nel capitolo 4, relativamente all'indagine Condizioni di Vita 2004 è stata realizzata, per la prima volta, l'imputazione multipla dei valori mancanti di reddito, che vengono attualmente imputati da ISTAT attraverso un procedura di imputazione singola. In questo senso uno dei principali obiettivi è stato il fornire una prima risposta sull'effettiva rilevanza dell'introduzione di una procedura di imputazione multipla per i valori di reddito di questa indagine.

In particolare, le componenti di reddito vengono rilevate a livello individuale e familiare, e contengono numerosi fattori di complicazione; la percentuale di mancate risposte, piuttosto variabile, è spesso superiore al 20%. Per gestire la struttura in due livelli dei dati è stata proposta un'innovativa procedura di imputazione di tipo *iterativo*, che ha il pregio di utilizzare le componenti di reddito di livello familiare per imputare quelle di livello individuale e viceversa. Questo risulta particolarmente importante nel caso dell'indagine Condizioni di Vita in quanto le componenti di reddito rilevate ai due diversi livelli dipendono le une dalle altre.

I risultati presentati nel capitolo 4 mostrano come la realizzazione delle imputazioni multiple comporti variazioni di lieve entità per la media pesata e per il relativo standard error delle variabili "Reddito individuale totale" e "Reddito familiare totale" nel 2003. Rispetto all'analisi dei soli casi completi e alle stime calcolate da ISTAT le due stime puntuali risultano corrette verso l'alto, mentre la variabilità *between* i dieci datasets imputati è sostanzialmente nulla. Per quanto riguarda la stima delle componenti di varianza non vi è quindi una differenza sostanziale tra imputare più volte oppure una sola. Questo è probabilmente dovuto ai bassi tassi di mancata risposta per alcune componenti di reddito, al fatto che i tassi più alti riguardano spes-

so un numero totale non elevato di osservazioni, ma anche alle informazioni apportate dal modello iterativo di imputazione: tutto questo rende molto bassa, nel complesso, la *fraction of missing information*, ovvero la frazione d'informazione relativa alla stima di interesse che risulta mancante a causa delle non risposte.

Le conclusioni cui si giunge sono in parte diverse, tuttavia, se si considerano le singole componenti delle due suddette variabili, e se si è interessati alle mediane piuttosto che alle medie. In questo caso infatti i risultati che si ottengono analizzando i soli dati completi, oppure realizzando l'imputazione singola o multipla dei valori mancanti risultano maggiormente differenziati. Anche eventuali analisi di regressione condotte sui dati provenienti dall'indagine, per esempio per comprendere le determinanti delle difficoltà economiche delle famiglie italiane, si sono mostrate sensibili alla scelta del trattamento delle mancate risposte. Queste indicazioni sono sicuramente interessanti nell'ottica di future eventuali modifiche della procedura di imputazione singola attualmente utilizzata da ISTAT.

Il metodo utilizzato per realizzare le imputazioni, l'approccio *sequential regression multivariate imputations* di Raghunathan et al. (2001), si basa sull'ipotesi che i dati mancanti siano MAR. Per cercare di testare tale ipotesi relativamente alle componenti di reddito imputate, alla fine del capitolo 4 è stata presentata una procedura per confrontare, separatamente per ogni variabile, la distribuzione dei valori imputati e di quelli osservati condizionando per la *nonresponse propensity*. Ciò corrisponde a testare, anche se non formalmente, l'ipotesi MAR; in particolare, le variabili cui si sceglie di condizionarsi possono coincidere con quelle utilizzate nel modello di imputazione ma possono anche essere diverse, situazione che può verificarsi quando i soggetti che realizzano le imputazioni e analizzano i dati completati sono distinti. Nel capitolo 4 la procedura di diagnostica è stata applicata ad una delle componenti di reddito, "Assegni familiari per lavoratori dipendenti", per la quale si accetta l'ipotesi MAR condizionando rispetto alle stesse variabili utilizzate nel modello di imputazione. Questa procedura, sebbene necessiti di ulteriori approfondimenti, risulta particolarmente indicata proprio per indagini condotte su larga scala, in quanto utilizza metodi grafici e test di immediata comprensione. L'impiego di tale metodologia per tutte le componenti di reddito rilevate dall'indagine potrebbe fornire utili indicazioni su quali sono le variabili per le quali l'ipotesi MAR risulta più critica.

Nel capitolo 5, invece, l'utilizzo dell'imputazione multipla è stato finalizzato all'implementazione di un'analisi di sensitività di tipo multivariato. Come sottolineato da alcuni autori, l'imputazione multipla rappresenta un metodo piuttosto efficace e relativamente semplice per valutare la sensitività rispetto a modelli non ignorabili.

In particolare, il pattern dei dati osservati e mancanti dell'indagine Forze Lavoro del Comune di Firenze, caratterizzato da un "panel-ruotato", ha determinato anche per la seconda applicazione la scelta del metodo di imputazione delle regressioni sequenziali multivariate. Nonostante vi sia un solo quesito relativo al reddito, la struttura ed il questionario dell'indagine rendono multivariato il problema di imputazione, con alte percentuali di valori mancanti. Oltre alle effettive non risposte, infatti, la rotazione del panel induce dei valori mancanti di tipo "strutturale" sia per il quesito sul reddito che per la domanda-filtro relativa allo stato occupazionale; in pratica quindi le mancate risposte possono essere pensate come derivanti dall'unione di meccanismi MCAR, MAR e MNAR. Inoltre, la componente longitudinale del campione fa sì che le informazioni disponibili per i singoli individui risultino piuttosto differenziate. L'utilizzo di un metodo di imputazione basato su modelli di regressione ha consentito il condizionamento al diverso insieme di covariate disponibili, che per alcuni soggetti comprendevano anche il reddito in una occasione di indagine precedente o successiva. I dati mancanti sono stati inizialmente formalizzati come MAR, sia che derivassero dalla rotazione del panel che da vere e proprie non risposte; successivamente è stata implementata un'analisi di sensitività di tipo multivariato relativamente ai veri *missing values*. Questo tipo di analisi, che vengono suggerite da numerosi autori, vengono solitamente realizzate in contesti piuttosto semplici e per dati mancanti univariati; in questo senso l'analisi di sensitività implementata nella tesi risulta particolarmente interessante ed innovativa nel contesto dei datasets con dati mancanti multivariati e che presentano numerosi fattori di complicazione.

Le ipotesi di tipo MNAR sono state introdotte come variazioni delle imputazioni realizzate con ipotesi MAR; queste variazioni, che prendono la forma di *offsets* aggiunti ai dati, rappresentati da frazioni della deviazione residua del modello per il logaritmo del reddito, sono di facile comprensione e realizzazione ed evitano le problematiche dei *selection models*. Per le stime del reddito medio nel Comune di Firenze nei quattro trimestri e nell'intero anno 2002 si è evidenziata una sostanziale robustezza rispetto alle variazioni MNAR considerate. Le variazioni introdotte, infatti, modificano solo marginalmente le stime, che oltre ai valori medi comprendono le mediane e alcuni percentili delle distribuzioni di reddito.

I risultati cui si è giunti in questa tesi possono essere visti come punto di partenza di molteplici nuovi sviluppi, sia relativamente allo studio del meccanismo che genera le mancante risposte di reddito che alla scelta del particolare modello di imputazione.

Relativamente ai dati provenienti dall'indagine Condizioni di Vita, per esempio, sarebbe interessante valutare la procedura di imputazione multipla

proposta attraverso studi di simulazione che introducano dei missing values in sottoinsiemi originariamente completi di dati. Le mancate risposte potrebbero essere simulate secondo meccanismi MAR e MNAR; in questo modo, applicando il modello di imputazione multipla ai dati cancellati, sarebbe possibile valutare se tale modello è in grado di preservare le principali caratteristiche distributive delle variabili di interesse. Risulterebbe possibile, inoltre, verificare l'effettiva dipendenza dei valori mancanti da tutte le variabili esplicative utilizzate, ottenendo alcune indicazioni relativamente all'ipotesi MAR. Per il momento è stato proposto un metodo di diagnostica attraverso cui risulta possibile testare non formalmente l'ipotesi MAR, separatamente per ciascuna variabile imputata, condizionando per la *nonresponse propensity*.

Sarebbe poi interessante confrontare le imputazioni ottenute con il metodo delle regressioni sequenziali con quelle di un metodo *da donatore*; il calcolo degli standard errors attraverso metodi di ricampionamento potrebbe essere utilizzato per verificare l'effettiva irrilevanza, per alcune delle stime di interesse, della realizzazione di imputazioni multiple.

Poichè nei casi in cui si disponeva di un limitato numero di osservazioni l'ipotesi di normalità della trasformata logaritmica dei valori osservati si è rivelata un po' più critica rispetto alle altre variabili, nel futuro sarebbe interessante utilizzare trasformazioni alternative e valutarne l'effetto sulle imputazioni.

Infine, l'analisi di sensitività per ipotesi MNAR realizzata nel capitolo 5 potrebbe essere applicata ai dati dell'indagine ISTAT, considerando anche eventuali modifiche. In particolare, nell'applicazione ai dati provenienti dall'indagine Forze Lavoro i valori sono stati perturbati attraverso un incremento, ipotizzando che la presenza dei missing values sia in relazione positiva con il vero valore del reddito. Questo tipo di analisi di sensitività potrebbe essere modificato utilizzando deviazioni di tipo diverso; per esempio si potrebbe aumentare la deviazione standard della distribuzione predittiva del logaritmo del reddito sotto l'ipotesi che i valori di reddito dei non rispondenti siano più dispersi di quelli predetti dal modello MAR.

Bibliografia

Abayomi, K., Gelman, A. and Levy, M. (2007), 'Diagnostics for Multivariate Imputations', *SSRN eLibrary* .

Banca d'Italia (2006), I bilanci delle famiglie italiane nell'anno 2004, Supplementi al bollettino statistico, Anno XVI - 17 Gennaio 2006.

Bernaards, C. A., Belin, T. R. and Schafer, J. L. (2007), 'Robustness of a Multivariate Normal Approximation for Imputation of Incomplete Binary Data', *Statistics in Medicine* **26**, 1368–1382.

Brandolini, A. (1999), The Distribution of Personal Income in Post-War Italy: Source Description, Data Quality, and the Time Pattern of Income Inequality, Temi di discussione, number 350, Banca d'Italia.

Brewer, K. R. W. and Sarndal, C. E. (1983), Six Approaches to Enumerative Survey Sampling, *in* W. G. Madow and I. Olkin, eds, 'Incomplete Data in Sample Surveys - Volume 3', Academic Press, New York, pp. 363–368.

Casella, G. and George, E. I. (1992), 'Explaining the Gibbs Sampler', *The American Statistician* **46**, 167–174.

Cassel, C. M., Sarndal, C. E. and Wretman, J. (1977), *Foundations of Inference in Survey Sampling*, Wiley, New York.

Collins, L. M., Schafer, J. L. and Kam, C. M. (2001), 'A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures', *Psychological Methods* **6**, 330–351.

D'Alessio, G. and Faiella, I. (2002), Non-response Behaviour in the Bank of Italy's Survey on Household Income and Wealth, Temi di discussione, number 462, Banca d'Italia.

D'Amuri, F. and Fiorio, C. (2004), Work Income Tax Evasion in Italy: Analysis of Redistributive Effects, XVI Conferenza, Società Italiana di Economia Pubblica.

David, M., Little, R. J. A., Samuhel, M. E. and Triest, R. K. (1986), 'Alternative Methods for CPS Income Imputation', *Journal of the American Statistical Association* **81**, 29–41.

de Leeuw, E. and de Heer, W. (2002), Trends in Household Survey Nonresponse: A Longitudinal and International Comparison, *in* R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little, eds, 'Survey Nonresponse', Wiley, New York, pp. 357–371.

Demirtas, H. and Hedeker, D. (2007), 'Imputing Continuous Data under some Non-Gaussian Distributions', *Statistica Neerlandica* .

Deville, J. C. and Sarndal, C. E. (1992), 'Calibration Estimators in Survey Sampling', *Journal of the American Statistical Association* **87**, 376–382.

EUROSTAT (2001), Imputation of income in the ECHP, DOC.PAN 164/2001-12, European Commission - EUROSTAT.

EUROSTAT (2003), European Community Household Panel - User manual, DOC.PAN 168/2003-12, European Commission - EUROSTAT.

EUROSTAT (2006), EU-SILC user database description, Version 2004-1 from 24-05-06, EU-SILC/BB D(2005), European Commission - EUROSTAT.

Fay, R. E. (1992), 'When are inferences from multiple imputation valid?', *Proceedings of the Survey Research Methods Section, American Statistical Association* pp. 227–232.

Fay, R. E. (1996), 'Alternative Paradigms for the Analysis of Imputed Survey Data', *Journal of the American Statistical Association* **91**, 490–498.

Ford, B. L. (1983), An overview of Hot-Deck Procedures, *in* W. G. Madow and I. Olkin, eds, 'Incomplete Data in Sample Surveys - Volume 2', Academic Press, New York, pp. 363–368.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004), *Bayesian Data Analysis*, Chapman & Hall, New York.

Gelman, A. and Meng, X. L. (2004), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, Wiley, Chichester.

Gelman, A. and Raghunathan, T. E. (2001), 'Discussion of Arnold et al. Conditional specified distributions', *Statistical Science* **16**, 268–269.

Gibbons, J. D. and S., C. (1992), *Non Parametric Statistical Inference*, Marcel Dekker Inc., New York.

Giommi, A., Innocenti, R., Rocco, E. and Sifone, M. (2003), Indagine sperimentale sulle forze di lavoro. Rapporto Aprile 2002-Gennaio 2003, La statistica per la città, Comune di Firenze.

Greenlees, J. S., Reece, W. S. and Zieschang, K. D. (1982), 'Imputation of Missing Values When the Probability of Response Depends On the Variable Being Imputed', *Journal of the American Statistical Association* **77**, 251–261.

Groves, R. M. and Couper, M. P. (1998), *Nonresponse in Household Interview Surveys*, Wiley, New York.

He, Y. and Raghunathan, T. E. (2006), 'Tukey's gh Distribution for Multiple Imputation', *The American Statistician* **60**, 251–256.

Heckman, J. J. (1979), 'Sample Selection Bias as a Specification Error', *Econometrica* **47**, 153–161.

Heeringa, S. G., Little, R. J. A. and Raghunathan, T. E. (2002), Multivariate Imputation of Coarsened Survey Data on Household Wealth, *in* R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little, eds, 'Survey Nonresponse', Wiley, New York, pp. 357–371.

Herzog, T. N. and Rubin, D. B. (1983), Using Multiple Imputations to Handle Nonresponse in Sample Surveys, *in* W. G. Madow and I. Olkin, eds, 'Incomplete Data in Sample Surveys - Volume 2', Academic Press, New York, pp. 376–379.

Hoaglin, D. C., Monsteller, F. and Tukey, J. W. (1983), *Understanding Robust and Exploratory Data Analysis*, Wiley.

Hunt, J. W., S., J. J. and King, C. S. (2003), Detecting outliers in the monthly retail trade survey using the Hidiroglou-Berthelot method, Proceedings of the survey research methods section, American Statistical Association.

ISTAT (2006), Reddito e condizioni di vita, Collana informazioni, n.31, Istituto Nazionale di Statistica.

Jansen, I., Hens, N., Molenberghs, G., Aerts, M., Verbeke, G. and Kenward, M. G. (2006), 'The nature of sensitivity in monotone missing not at random models', *Computational Statistics and Data Analysis* **50**, 830–858.

Jo, C., Simpson, P. M. and Gossett, J. M. (2007), 'Regression Splines with Longitudinal Data', *SAS Global Forum 2001. Paper 143-2007* .

Juster, F. T. and Smith, J. P. (1997), 'Improving the Quality of Economic Data: Lessons from the HRS and AHEAD', *Journal of the American Statistical Association* **92**, 1268–1278.

Kennickell, A. B. (1991), 'Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation', *Proceedings of the American Statistical Association, Survey Research Methods Section* .

Kong, A., Liu, J. S. and Wong, W. H. (1994), 'Sequential Imputations and Bayesian Missing Data Problems', *Journal of the American Statistical Association* **89**, 278–288.

Lee, H., Rancourt, E. and Sarndal, C. E. (2002), Variance Estimation from Survey Data under Single Imputation, *in* R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little, eds, 'Survey Nonresponse', Wiley, New York, pp. 357–371.

Li, K. H. (2004), The Sampling/Importance Resampling Algorithm, *in* A. Gelman and X. L. Meng, eds, 'Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspective', Wiley, New York, pp. 265–276.

Li, K. H., Meng, X. L., Raghunathan, T. E. and Rubin, D. B. (1991), 'Significance Levels from Repeated p-values with Multiply Imputed Data', *Statistica Sinica* **1**, 65–92.

Li, K. H., Raghunathan, T. E. and Rubin, D. B. (1991), 'Large-Sample Significance Levels from Multiply Imputed Data using Moment-based Statistics and a F Reference Distribution', *Journal of the American Statistical Association* **86**, 1065–1073.

Lillard, L., Smith, J. P. and Welch, F. (1986), 'What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation', *Journal of Political Economy* **94**, 489–506.

Little, R. J. A. (1982), 'Models for Nonresponse in Sample Surveys', *Journal of the American Statistical Association* **77**, 237–250.

Little, R. J. A. (1983), The Ignorable Case, *in* W. G. Madow and I. Olkin, eds, 'Incomplete Data in Sample Surveys - Volume 2', Academic Press, New York, pp. 376–379.

Little, R. J. A. (1986), 'Survey Nonresponse Adjustments for Estimates of Means', *International Statistical Review* **54**, 139–157.

Little, R. J. A. (1988), 'Missing-Data Adjustments in Large Surveys', *Journal of Business and Economic Statistics* **6**, 287–296.

Little, R. J. A. (1993), 'Pattern-Mixture Models for Multivariate Incomplete Data', *Journal of the American Statistical Association* **88**, 125–134.

Little, R. J. A. (1995), 'Modeling the Drop-Out Mechanism in Repeated-Measures Studies', *Journal of the American Statistical Association* **90**, 1112–1121.

Little, R. J. A. and Rubin, D. B. (1983), Discussion of Six Approaches to Sample Surveys, by Brewer and Sarndal, *in* W. G. Madow and I. Olkin, eds, 'Incomplete Data in Sample Surveys - Volume 3', Academic Press, New York, pp. 376–379.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, Wiley, New York.

Madow, W. G., Nisselson, H. and Olkin, I. (1983), *Incomplete Data in Sample Surveys - Volume 1*, Academic Press, New York.

Meng, X. L. (1994), 'Multiple-Imputation Inferences with Uncongenial Sources of Input', *Statistical Science* **9**, 538–573.

Meng, X. L. (2002), A Congenial Overview and Investigation of Multiple Imputation Inferences under Uncongeniality, *in* R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little, eds, 'Survey Nonresponse', Wiley, New York, pp. 357–371.

Meng, X. L. and Rubin, D. B. (1992), 'Performing Likelihood Ratio Tests with Multiply-Imputed Data sets', *Biometrika* **79**, 103–111.

Nielsen, S. F. (2003), 'Proper and Improper Multiple Imputation', *International Statistical Review* **71**, 593–607.

Oh, H. L. and Scheuren, F. J. (1983), Weighting Adjustment for Unit Nonresponse, *in* W. G. Madow and I. Olkin, eds, 'Incomplete Data in Sample Surveys - Volume 2', Academic Press, New York, pp. 363–368.

Parlamento Europeo (2003*a*), Regolamento (CE) N. 1177/2003 del Parlamento Europeo e del Consiglio del 16 giugno 2003, Gazzetta Ufficiale dell'Unione Europea, Parlamento Europeo.

Parlamento Europeo (2003*b*), Regolamento (CE) N. 1181/2003 della Commissione del 21 ottobre 2003, Gazzetta Ufficiale dell'Unione Europea, Parlamento Europeo.

Quintano, C., Castellano, R. and Regoli, A. (2001), 'How to improve the quality of the income variable in a household survey. A simulation study through multiple imputation', *International Conference on Quality in Official Statistics, Stockholm, Sweden* .

Raghunathan, T. and Bondarenko, I. (2007), 'Diagnostics for Multiple Imputations', *SSRN eLibrary* .

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J. and Solenberger, P. (2001), 'A Multivariate Technique for Multiply Imputing Missing Values using a Sequence of Regression Models', *Survey Methodology* **27**, 85–95.

Raghunathan, T. E., Solenberger, P. W. and Van Hoewyk, J. (1998), IVEware: Imputation and Variance Estimation Software: Installation Instructions and User Guide, Technical report, Survey Research Center - University of Michigan.

Rao, J. N. K. and Shao, J. (1992), 'Jackknife Variance Estimation with Survey Data under Hot-deck Imputation', *Biometrika* **79**, 811–822.

Reiter, J. P. and Raghunathan, T. E. (2007), 'The Multiple Adaptations of Multiple Imputation', *Journal of the American Statistical Association* **102**, 1462–1471.

Rosenbaum, P. R. and Rubin, D. B. (1983), 'The Central Role of the Propensity Score in Observational Studies for Causal Effects', *Biometrika* **70**, 41–55.

Royston, P. (2005), 'Multiple imputation of missing values: update', *The Stata Journal* **2**, 188–201.

Rubin, D. B. (1977), 'Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys', *Journal of the American Statistical Association* **72**, 538–543.

Rubin, D. B. (1978), 'Multiple Imputations in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse', *Proceedings of the Survey Research Methods Section of the American Statistical Association* pp. 20–34.

Rubin, D. B. (1983), Conceptual Issues in the Presence of Nonresponse, *in* W. G. Madow and I. Olkin, eds, 'Incomplete Data in Sample Surveys - Volume 2', Academic Press, New York, pp. 376–379.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Sample Surveys*, Wiley, New York.

Rubin, D. B. (1996), 'Multiple Imputation After 18+ Year', *Journal of the American Statistical Association* **91**, 473–489.

Rubin, D. B. (2003), 'Discussion on Multiple Imputation', *International Statistical Review* **71**, 619–625.

Rubin, D. B., Cook, S., Yu, Y., Frangakis, C., Li, F., Mealli, F. and Baccini, M. (2004), Multiple Imputation for AVA Clinical Trials, Progress report, Harvard University.

Sarndal, C. E., Swensson, B. and Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.

SAS (1999), SAS user's guide, SAS Institute Inc.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, Chapman & Hall, New York.

Schafer, J. L. (2003), 'Multiple Imputation in Multivariate Problems when the Imputation and Analysis Models Differ', *Statistica Neerlandica* **57**, 19–35.

Schafer, J. L. and Olsen, M. K. (1998), 'Multiple Imputation for Multivariate Missing-Data Problems: a Data Analyst's Perspective', *Multivariate Behavioral Research* **33**, 545–571.

Schenker, N., Raghunathan, T. E., Chiu, P., Makuc, D. M., Zhang, G. and Cohen, A. J. (2006), 'Multiple Imputation of Missing Income Data in the National Health Interview Survey', *Journal of the American Statistical Association* **101**, 924–933.

Tanner, M. A. (1996), *Tools for Statistical Inference*, Springer, New York.

Tanner, M. A. and Wong, W. H. (1987), 'The Calculation of Posterior Distributions by Data Augmentation', *Journal of the American Statistical Association* **82**, 528–540.

Thompson, M. E. (1997), *Theory of Sample Surveys*, Chapman & Hall, New York.

Tian, G., Tan, M. and Ng, K. W. (2007), 'An Exact Non-Iterative Sampling Procedure for Discrete Missing Data Problems', *Statistica Neerlandica* **61**, 232–242.

Van Buuren, S., Boshuizen, H. C. and Knook, D. L. (1999), 'Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis', *Statistics in Medicine* **18**, 681–694.

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M. and Rubin, D. (2006), 'Fully Conditional Specification in Multivariate Imputation', *Journal of Statistical Computation and Simulation* **76**, 1049–1064.

Van Buuren, S. and Oudshoorn, K. (1999), Flexible multivariate imputation by MICE, Netherlands Organization for Applied Scientific Reserch (TNO) 54, Princeton University.

Vitaletti, S. (2005), Correzioni e imputazioni delle informazioni sui redditi, Seminario, ISTAT.

Woodruff, R. S. (1971), 'A Simple Method for Approximating the Variance of a Complicated Estimate', *Journal of the American Statistical Association* **66**, 411–414.

Yucel, R. M., Raghunathan, T. E. and Schenker, N. (2006), 'SHRIMP: sequential hierarchical regression imputations', *International Conference on Health Policy Research, Boston, USA* .

Zheng, H. and Little, R. J. A. (2003), 'Penalized Spline Model-Based Estimation of the Finite Population Total from Probability-Proportional-To-Size Samples', *Journal of Official Statistics* **19**, 99–117.