Università degli Studi di Padova
Corso di laurea in Statistica e Gestione delle Imprese

STIMATORI DELLA MEDIA DI UNA POPOLAZIONE PER DATI MANCANTI

Relatore: Prof. Giancarlo Diana
Dipartimento di Scienze Statistiche

Laureando: Nicola Olivieri

Anno Accademico 2011/2012

Indice

Introduzione

Il problema dei dati mancanti è abbastanza comune nella ricerca empirica, specialmente nelle scienze economico-sociali in cui la somministrazione di questionari è una delle tecniche più diffuse per la raccolta di dati e informazioni. Uno dei primi problemi che un ricercatore si trova ad affrontare, in fase di analisi dei risultati, è quello di un dataset incompleto e con errori. Questo accade generalmente perché chi compila il questionario non ne interpreta correttamente la struttura, commette accidentalmente qualche errore nel fornire le risposte, non vuole deliberatamente rispondere ad alcune domande, oppure a causa di un errore dello strumento di codifica, che dal supporto cartaceo deve trasferire i dati su supporto informatico, o di chi invece si occupa del data entry.

Non esiste in letteratura un'unica tecnica o una metodologia di approccio al problema di come tenere sotto controllo l'effetto dei dati mancanti. Tuttavia prima di parlare dei metodi di trattamento dei dati mancanti, ed in particolare dei metodi per la generazione delle imputazioni è opportuno soffermarsi sul concetto di non risposta.

Con il termine non-risposta (*non response*) si intendono una moltitudine di situazioni in cui il dato non viene osservato. In effetti si parla di non risposta ogni qualvolta non si riesce ad ottenere il dato su una o più variabili di interesse per una o più unità campionarie. La non risposta causa sia un incremento nella variabilità degli stimatori, dovuta ad una riduzione della base campionaria di analisi e/o all'applicazioni di metodi per il trattamento della stessa, sia stimatori distorti, se i rispondenti differiscono sistematicamente dai non rispondenti rispetto alle caratteristiche di interesse. Principalmente si distinguono due tipi di non risposta la non risposta totale (*Unit Non-Response*) e la non risposta parziale (*Item Non-Response*).

Non Risposta Totale: si riferisce al tipo di non risposta in cui non si ha nessuna informazione disponibile per unità campionarie eligibili. Le ragioni possono essere varie e dipendono ovviamente dalle modalità di raccolta dei dati, alcune possono

essere: impossibilità di contatto, non reperibilità, inabilità a rispondere, rifiuto, questionario non restituito.

Non Risposta Parziale: si riferisce al caso in cui le informazioni rilevate dal rispondente sono tali da essere ritenute accettabili per il data base, ma alcune informazioni risultano mancanti. I motivi possono essere diversi: l'intervistato considera il quesito non comprensibile, troppo personale oppure rifiuta categoricamente quesiti simili.

L'obiettivo di questo lavoro, è fare una rassegna delle tecniche utilizzate per gestire in modo efficace ed efficiente il problema delle non risposte. In particolare, si è deciso di esaminare la letteratura relativa ai metodi di stima della media di una popolazione del periodo 2005-2012, volendo dare un compendio dei più recenti sviluppi in materia.

Analizzando gli articoli pubblicati nel periodo scelto, si è notato che non esiste una simbologia univoca e universalmente utilizzata; si è ritenuto, quindi, opportuno uniformare il linguaggio per tutti i metodi di stima descritti, precisando il significato dei termini usati.

Il materiale raccolto è stato organizzato secondo due criteri fondamentali: il tipo di meccanismo generatore dei dati mancanti ('missing completely at random', o MCAR e 'missing at random', o MAR) e il metodo d'imputazione dei valori mancanti. Nel primo capitolo, diviso per meccanismo generatore dei dati mancanti, presentiamo cinque paragrafi riguardanti stimatori della media $\bar{Y}$: i primi due paragrafi ipotizzano l'utilizzo del meccanismo MCAR, mentre gli ultimi tre ipotizzano un meccanismo MAR. Nel paragrafo 1.2.1, le metodologie di stima vengono presentate in riferimento alla tipologia del dataset di dati mancanti. Nel paragrafo 1.2.2, presentiamo stimatori per rapporto, prodotto e regressione nel campionamento a due fasi; questi vengono analizzati su scenari diversi a seconda delle informazioni disponibili. Gli stimatori per rapporto e prodotto, sebbene siano meno efficienti dello stimatore per regressione, vengono comunque considerati per avere un quadro completo degli stimatori proposti in letteratura. Dal paragrafo 1.3.1, consideriamo i dati mancanti provenienti da un meccanismo di tipo MAR; il primo gruppo di stimatori presentato è quello degli stimatori doppiamente robusti. Questi implicano la modellazione sia della regressione della variabile di studio sulle variabili ausiliarie che della probabilità di risposta; sono stimatori consistenti di $\bar{Y}$ anche quando uno dei due modelli non è correttamente specificato. Nel paragrafo 1.3.2, consideriamo stimatori che si basano su un approccio semiparametrico; essi conciliano il metodo di

regressione non parametrica con gli approcci parametrici, condensando l'informazione contenuta nelle variabili ausiliarie attraverso una funzione parametrica, così da ridurre la dimensione per la regressione non parametrica successiva. Questo tipo di stima semiparametrica viene usata soprattutto negli studi con un elevato numero di variabili ausiliarie. Nell'ultimo paragrafo del capitolo uno, proponiamo stimatori che utilizzano metodi di aggiustamento per ponderazione. Essi consistono nell'aumentare i pesi delle unità rispondenti, in modo da compensare le mancate risposte delle altre unità. Nel paragrafo vengono considerati tre metodi: il metodo 'NWA diretto', la tecnica non parametrica di lisciamento di tipo Kernel e il metodo non parametrico di regressione polinomiale locale.

Nel secondo capitolo, diviso anch'esso per meccanismo generatore dei dati mancanti, presentiamo sei paragrafi riguardanti stimatori di $\overline{Y}$ che utilizzano metodi d'imputazione: i primi due paragrafi ipotizzano l'utilizzo del meccanismo MCAR, mentre gli ultimi quattro ipotizzano un meccanismo MAR. Nel primo paragrafo (2.2.1), presentiamo diversi stimatori che utilizzano metodi di imputazione per media, rapporto e regressione, confrontandone i relativi MSE. Tali stimatori vengono definiti ipotizzando diversi scenari, in relazione alla tipologia del dataset dei dati mancanti. Nel paragrafo 2.2.2, presentiamo alcune metodologie di imputazione nel campionamento stratificato e per cluster. Sotto campionamento stratificato, si è visto lo stimatore basato sull'imputazione multipla; quest'ultimo consiste nel creare un certo numero di valori imputati per ogni valore mancante e nel combinare i diversi dataset, completati separatamente, per ciascuno strato. Abbiamo, infine, descritto alcuni metodi per trattare le non risposte nel caso di campionamento per cluster. Si sono presentati modelli sia per il caso in cui si abbia a disposizione la variabile di studio e le informazioni sull'effetto dei cluster, sia se si dispone della variabile ausiliaria osservata per tutto il campione: in questo caso, si sfrutta l'informazione aggiuntiva per stimare $\bar{Y}$.

Se, invece, i dati mancanti seguono un meccanismo di tipo MAR, sono stati presentati quattro gruppi di stimatori che utilizzano le seguenti tecniche d'imputazione: l'imputazione 'nearest neighbor', l'imputazione tramite pseudo-verosimiglianza empirica, metodi di imputazione doppiamente robusti e l'imputazione ponderata. Il paragrafo 2.3.1 è dedicato al metodo NNI, il quale può essere più efficiente di altri metodi, come l'imputazione tramite media, quando la variabile ausiliaria fornisce informazione utile allo studio di y; inoltre, non assumendo un modello di regressione parametrica tra y e x, il metodo risulta essere

più robusto rispetto alle imputazioni per rapporto o regressione che sottintendono un modello di regressione lineare. Il paragrafo 2.3.2 è dedicato allo studio dei metodi d'imputazione basati sulla pseudo-verosimiglianza empirica; tra questi, abbiamo preso in considerazione la tecnica che utilizza la media di pseudo-verosimiglianza e la tecnica che utilizza l'imputazione casuale di pseudo verosimiglianza. Nel paragrafo 2.3.3, oggetti di studio sono i metodi d'imputazione doppiamente robusti, tra i quali ricordiamo: il metodo d'imputazione multipla, che utilizza per la selezione dei dataset imputati la tecnica NNI, e il metodo di regressione non parametrico. I metodi d'imputazione doppiamente robusti permettono agli stimatori di $\bar{Y}$ di essere asintoticamente non distorti ed efficienti quando almeno uno tra il modello di regressione e il modello per la probabilità di risposta è correttamente specificato. Infine, nell'ultimo paragrafo del secondo capitolo (2.3.4), si esamina l'approccio d'imputazione ponderata, basata sulle probabilità di risposta stimate; in questo caso, i valori imputati dipendono dai valori di una funzione Kernel.

Notazioni e assunzioni

Si definisca un campione *s*, di dimensione fissa *n* dalla popolazione finita *U* di *N* individui ($n < N$), come un sottoinsieme di *n* etichette che identificano gli elementi della popolazione. Una procedura di campionamento equivale a definire una regola per selezionare elementi di *U*; questa regola è descritta dal disegno campionario $d = (S_d, P_d)$ con probabilità di inclusione di primo e secondo ordine pari a π_i e π_{ij}, rispettivamente $(i = 1, \dots, N; j = 1, \dots, N; i \neq j)$. Quando le probabilità di inclusione di primo ordine sono tutte costanti, il campione *s* proviene da una campionamento casuale semplice senza reintroduzione (in inglese SRSWOR, da simple random sample without replacement). La procedura di selezione SRSWOR è di gran lunga la più importante nella pratica, in quanto viene utilizzata anche in processi di selezione intermedi di disegni più complessi (campionamento a due o più stadi).

Sia $y \in \mathbb{R}^+$, la variabile oggetto di studio, di media ignota: si dispone di un campione di *n* risposte, così indicato $y_1, \dots, y_n$. Per ciascuno degli *N* individui della popolazione, si può disporre, inoltre, di un'altra variabile ausiliaria correlata alla *y*. Siano, quindi, $x_1, \dots, x_n$ i valori della variabile ausiliaria.

All'interno del campione *s*, un sottogruppo di unità omette informazioni riguardo la caratteristica di interesse *y*, provocando una mancata risposta parziale. Per derivare tale fenomeno, si definisca una variabile indicatrice z_i per $i = 1, \dots, n$ che assume valori:

$$z_i = \begin{cases} 1 & \text{se } y_i \text{ è osservata} \\ 0 & \text{se } y_i \text{ non è osservata} \end{cases}$$

Il campione osservato può essere indicato con:

$$\left((y_1, x_1, z_1), (y_2, x_2, z_2), \dots, (y_n, x_n, z_n)\right)$$

Ciò non esclude il fatto che le metodologie di stima di $\bar{Y}$ presentate, utilizzino più variabili ausiliarie correlate alla variabile di studio.

Nel seguito, faremo uso di due meccanismi di generazione dei dati mancati: MCAR e MAR. Questa distinzione è dovuta al fatto che la probabilità di riposta può dipendere o meno dalle variabili ausiliarie correlate a y. Assumendo che gli indicatori di risposta siano variabili casuali indipendenti definiamo:

- 'missing completely at random' (MCAR), il meccanismo in cui la probabilità di risposta è indipendente sia dalla variabile di studio che dalla variabile ausiliaria associata; ossia, $Pr(z_i = 1|y_i, x_i) = p$ con $0 \leq p \leq 1$, per $i = 1, \dots, n$.
- 'missing at random' (MAR), il meccanismo in cui la probabilità di risposta dipende dalla variabile ausiliaria osservata per l'i-esima unità, ma non dalla variabile y; ossia, $Pr(z_i = 1|y_i, x_i) = Pr(z_i = 1|x_i) = p(x_i) = p_i$ con $0 \leq p_i \leq 1$, per $i = 1, \dots, n$.

Di seguito, presentiamo alcune quantità che caratterizzano la popolazione oggetto di studio, con lettere maiuscole, e le corrispondenti quantità campionarie, con lettere minuscole.

La media, varianza e varianza corretta di y sono indicate con:

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2 \qquad S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2$$

La covarianza, il coefficiente di correlazione, il coefficiente di regressione e il rapporto tra le medie delle variabili y e x della popolazione sono indicati rispettivamente con:

$$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)\left(x_i - \bar{X}\right) \qquad \rho = \frac{S_{xy}}{S_x S_y} \qquad \beta = \frac{S_{xy}}{S_x^2} \qquad \mathrm{R} = \frac{\bar{Y}}{\bar{X}}$$

Le rispettive quantità campionarie sono indicate con:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \hat{S}_y^2$$

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right) = \hat{S}_{xy} \qquad \hat{\rho} = \frac{s_{xy}}{s_x s_y} \qquad \hat{\beta} = \frac{s_{xy}}{s_x^2} \qquad \hat{\mathrm{R}} = \frac{\bar{y}}{\bar{x}}$$

Capitolo 1

Stimatori della media senza metodi di imputazione

1.1 Introduzione

L'obiettivo di questo capitolo, è presentare diverse metodologie di stima della media di una variabile di studio *y* in presenza di dati mancanti. Le metodologie di stima, qui esposte, non utilizzano metodi di imputazione e questo comporta l'eliminazione delle unità che non presentano risposta dai dati a disposizione. Gli stimatori proposti sono stati, inoltre, suddivisi a seconda del meccanismo di generazione dei dati mancanti: MCAR e MAR. Il paragrafo 1.2 è dedicato al caso in cui i dati mancanti siano del tipo MCAR: si contraddistinguono due filoni principali di stimatori, il primo riguardante gli stimatori che utilizzano tutta, o parte, dell'informazione disponibile; il secondo riguardante gli stimatori che usano un campione proveniente da un disegno in due fasi. Nel paragrafo 1.3 sono esposti tre gruppi di stimatori, ipotizzando un meccanismo di generazione di dati mancanti di tipo MAR. Il primo gruppo comprende stimatori doppiamente robusti; il secondo, stimatori che utilizzano un approccio semiparametrico e l'ultimo gruppo comprende stimatori basati su metodi di aggiustamento per ponderazione.

1.2 Dati mancanti MCAR

Se assumiamo che il meccanismo di generazione dei dati mancanti è di tipo MCAR, la probabilità che una data osservazione sia mancante, non dipende né dai valori della variabile di studio, né dai valori delle variabili ausiliarie. In questo caso, il campione incompleto, può essere considerato un sottocampione casuale del campione originario.

1.2.1 Stimatori che utilizzano tutta, o parte, dell'informazione disponibile

Consideriamo una classe generale g_y di stimatori della media $\bar{Y}$ della caratteristica di studio y sulla base di un campione s, di dimensione n, ottenuto a partire da un qualsiasi disegno di campionamento. Supponiamo che si possano presentare dati mancanti sia per la variabile di studio che per la variabile ausiliaria associata; sotto queste ipotesi, si può definire una classe di stimatori di Horvitz-Thompson che considera tutti i dati disponibili, e come tale preferibile agli stimatori usuali che rimuovono le osservazioni incomplete.

Si assuma di possedere un insieme $(n - p - q)$ di osservazioni complete all'interno del campione s e di avere a disposizione p osservazioni della caratteristica ausiliaria x ma non le corrispondenti osservazioni della caratteristica di studio y. Allo stesso modo, abbiamo una serie di q osservazioni della caratteristica y nel campione, ma i valori associati della caratteristica x sono mancanti. Inoltre p e q sono numeri (interi) che verificano le seguenti condizioni: $p > 0$ e $q < n/2$, .

Per comodità, separiamo le unità del campione s in tre insiemi disgiunti:

s_1			s_2			s_3		
y_1	...	y_{n-p-q}	Missing	...	Missing	y_{n-q+1}	...	y_n
x_1	...	x_{n-p-q}	$x_{n-p-q+1}$	...	x_{n-q}	Missing	...	Missing

Si consideri un disegno campionario generico $d = (S_d, P_d)$ con probabilità di inclusione del primo ordine π_k e del secondo ordine π_{kl} strettamente positive e definiamo i seguenti stimatori di Horvitz-Thompson, relativi ai tre insiemi s_1, s_2, s_3.

$$\bar{y}_{HT}^{(1)} = \sum_{i \in s_1} \frac{y_i}{N\pi_i} \qquad \bar{y}_{HT}^{(3)} = \sum_{i \in s_3} \frac{y_i}{N\pi_i} \quad \bar{x}_{HT}^{(1)} = \sum_{i \in s_1} \frac{x_i}{N\pi_i} \quad \bar{x}_{HT}^{(2)} = \sum_{i \in s_2} \frac{x_i}{N\pi_i} \tag{1.1}$$

A questo punto è possibile definire (nota $\bar{X}$) uno stimatore alle differenze, in funzione degli stimatori di H-T definiti precedentemente, appartenente ad una classe superiore g_y:

$$\bar{y}_{gd} = \bar{y}_{HT}^{(1,3)} + c(\overline{X} - \bar{x}_{HT}^{(1)}) + d(\overline{X} - \bar{x}_{HT}^{(2)}) \tag{1.2}$$

dove $\bar{y}_{HT}^{(1,3)}$ rappresenta lo stimatore di H-T calcolato sull'insieme $s_1 \cup s_3$. I valori dei parametri *c* e *d* che provvedono a minimizzare la varianza di stima si possono calcolare nel seguente modo: $(c_{opt}, d_{opt})' = \Sigma^{-1}\sigma$ e $Var_{min}(\bar{y}_{gd}) = Var\left(\bar{y}_{HT}^{(1,3)}\right) - \sigma'\Sigma^{-1}\sigma$, dove

$$\sum = \begin{pmatrix} V(\bar{x}_{HT}^{(1)}) & \mathrm{cov}(\bar{x}_{HT}^{(1)}, \bar{x}_{HT}^{(2)}) \\ \mathrm{cov}(\bar{x}_{HT}^{(1)}, \bar{x}_{HT}^{(2)}) & V(\bar{x}_{HT}^{(2)}) \end{pmatrix} \tag{1.3}$$

$$\sigma = \begin{pmatrix} \mathrm{cov}(\bar{y}_{HT}^{(1,3)}, \bar{x}_{HT}^{(1)}) & \mathrm{cov}(\bar{y}_{HT}^{(1,3)}, \bar{x}_{HT}^{(2)}) \end{pmatrix}' \tag{1.4}$$

I valori ottimi di *c* e *d* dipendono dalle caratteristiche della popolazione e per questo motivo lo stimatore differenza ottimale non può essere calcolato se non utilizzando i valori campionari. Sostituendoli in (1.2) si ha:

$$\bar{y}_{gd}^{*} = \bar{y}_{HT}^{(1,3)} + \hat{c}(\overline{X} - \bar{x}_{HT}^{(1)}) + \hat{d}(\overline{X} - \bar{x}_{HT}^{(2)}) \tag{1.5}$$

Possiamo calcolare, quindi, le stime delle varianze e covarianze degli stimatori di Horvitz-Thompson e successivamente $\hat{\Sigma}$ e $\hat{\sigma}$. Risulta quindi: $(\hat{c}, \hat{d}) = \hat{\Sigma}^{-1}\hat{\sigma}$.

Se il disegno di campionamento considerato è di tipo SRSWOR, gli stimatori di H-T sono rappresentati da semplici medie e lo stimatore alle differenze diventa:

$$\bar{y}_{gd}^{*} = \bar{y}^{(1,3)} + \hat{c}(\overline{X} - \bar{x}^{(1)}) + \hat{d}(\overline{X} - \bar{x}^{(2)}) \tag{1.6}$$

dove $\bar{y}^{(1,3)} = \frac{1}{n-p}\sum_{i \in s1 \cup s2} y_i$, $\bar{x}^{(1)} = \frac{1}{n-p-q}\sum_{i \in s1} x_i$, $\bar{x}^{(2)} = \frac{1}{p}\sum_{i \in s2} x_i$

In questo caso, possiamo esplicitare l'espressione per la varianze e covarianze degli stimatori esaminati. Se consideriamo p e q costanti le espressioni sono:

$$Var(\bar{y}^{(1)}) = S_y^2\left(\frac{1}{n-p-q}-\frac{1}{N}\right), \tag{1.7}$$

$$Var(\bar{y}^{(1,3)}) = S_y^2\left(\frac{1}{n-p}-\frac{1}{N}\right), \tag{1.8}$$

$$Var(\bar{x}^{(1)}) = S_x^2\left(\frac{1}{n-p-q}-\frac{1}{N}\right), \tag{1.9}$$

$$Var(\bar{x}^{(2)}) = S_x^2\left(\frac{1}{p}-\frac{1}{N}\right), \tag{1.10}$$

$$Cov(\bar{x}^{(1)},\bar{x}^{(2)}) = \begin{cases} S_x^2\left(\frac{1}{n-p-q}-\frac{1}{N}\right) & \text{se } n-p-q \geq p \\ S_x^2\left(\frac{1}{p}-\frac{1}{N}\right) & \text{se } n-p-q \leq p \end{cases} \tag{1.11}$$

$$Cov(\bar{y}^{(1,3)},\bar{x}^{(1)}) = \left[\frac{1}{n-q}-\frac{1}{N}\right]S_{xy} \tag{1.12}$$

$$Cov(\bar{y}^{(1,3)},\bar{x}^{(2)}) = \left[\frac{1}{n-q}-\frac{1}{N}\right]S_{xy} \tag{1.13}$$

Le varianze e covarianze possono essere facilmente stimate dal campione, così da ottenere le stime della matrice $\hat{\Sigma}$ e del vettore $\hat{\sigma}$. Se consideriamo p e q variabili casuali, le corrispondenti espressioni sono ottenute sostituendo a $\frac{1}{p}, \frac{1}{q}\frac{1}{n-p}, \ldots$ i loro valori attesi.

Analizzando le proprietà degli stimatori proposti è possibile verificare che, quando $N \to \infty$ e il campionamento è di tipo SRSWOR, gli stimatori sono, sotto condizioni di regolarità, asintoticamente normali e non distorti.

Un particolare stimatore all'interno della classe g_y, equivalente al precedente in quanto ad efficienza, è quello di tipo rapporto:

$$\bar{y}_r = \bar{y}^{(1,3)}\left(\frac{\bar{X}}{\bar{x}^{(1)}}\right)^{\alpha_1}\left(\frac{\bar{X}}{\bar{x}^{(2)}}\right)^{\alpha_2} \tag{1.14}$$

Seguendo una procedura equivalente a quella usata per lo stimatore alle differenze, otteniamo i valori di α_1 e α_2 che minimizzano la varianza, $(\alpha_1, \alpha_2)_{opt} = C^{-1}C_0$, dove

$$C = \begin{pmatrix} V(\bar{x}^{(1)})R^2 & Cov(\bar{x}^{(1)}, \bar{x}^{(2)})R^2 \\ Cov(\bar{x}^{(1)}, \bar{x}^{(2)})R^2 & V(\bar{x}^{(2)})R^2 \end{pmatrix}, \tag{1.15}$$

$$C_0 = \begin{pmatrix} Cov(\bar{y}^{(1,3)}, \bar{x}^{(1)})R & Cov(\bar{y}^{(1,3)}, \bar{x}^{(2)})R \end{pmatrix}' \tag{1.16}$$

dove, in questo caso, R vale $\frac{\bar{y}^{(1,3)}}{\bar{X}}$.

Generalmente le espressioni di questi valori ottimali, dipendono da quantità di popolazione, ma, applicando la stessa procedura usata per lo stimatore alle differenze si ottiene il seguente stimatore ottimale:

$$\bar{y}_r^* = \bar{y}^{(1,3)} \left(\frac{\bar{X}}{\bar{x}^{(1)}} \right)^{\hat{\alpha}_1} \left(\frac{\bar{X}}{\bar{x}^{(2)}} \right)^{\hat{\alpha}_2} \tag{1.17}$$

dove i valori di $\hat{\alpha}_1$ e $\hat{\alpha}_2$ possono essere ottenuti su base campionaria.

Si propone, ora, uno stimatore basato sulla pseudo verosimiglianza empirica che lavora sotto le medesime ipotesi e con la stessa struttura del campione precedente:

s_1			s_2			s_3		
y_1	...	y_{n-p-q}	Missing	...	Missing	y_{n-q+1}	...	y_n
x_1	...	x_{n-p-q}	$x_{n-p-q+1}$	...	x_{n-q}	Missing	...	Missing

Iniziamo col definire lo stimatore basato sull'intero campione *s* per poi ricondurci al nostro caso di studio. Lo stimatore è strutturato nel seguente modo:

$$\bar{y}_{PE} = \sum_{i \in s} \hat{p}_i y_i \tag{1.18}$$

Per trovare $\hat{p}_i$ si massimizza la funzione di pseudo log-verosimiglianza empirica $l(p) = \sum_{i \epsilon s} d_i log p_i$, vincolata dalle seguenti condizioni:

$$\sum p_i = 1 \qquad (0 \le p_i \le 1), \tag{1.19}$$

$$\sum p_i u_i = 0, \tag{1.20}$$

dove $d_i = 1/\pi_i$ e u_i sono quantità note, con $N^{-1} \sum_{i=1}^{N} u_i = 0$.

Si possono proporre diverse espressioni per u_i, ma qui si considera la più comune: $u_i = x_i - \bar{X}$, che può essere giustificata considerando una relazione lineare tra y e x. Usando i moltiplicatori di Lagrange si può vedere che

$$\hat{p}_i = \frac{d_i^*}{1+\lambda u_i}, \quad \text{per i} \in \text{s}, \tag{1.21}$$

dove $d_i^* = d_i / \sum_{j\epsilon s} d_j$, e il moltiplicatore di Lagrange, λ, è soluzione di

$$\sum_{i\epsilon s} \frac{d_i^* u_i}{1+\lambda u_i} = 0 \quad \forall \text{u}_i \tag{1.22}$$

Torniamo ora a considerare i tre insiemi s_1, s_2, s_3 da cui eravamo partiti. Consideriamo, quindi, i seguenti stimatori per rapporto.

$$\bar{y}_w^1 = \sum_{i\in s_1} d_i^{*1} y_i; \quad \bar{y}_w^3 = \sum_{i\in s_3} d_i^{*3} y_i; \quad \bar{y}_w^{13} = \sum_{i\in s_1 \cup s_3} d_i^{*13} y_i; \tag{1.23}$$

$$\bar{x}_w^1 = \sum_{i\in s_1} d_i^{*1} x_i; \quad \bar{x}_w^2 = \sum_{i\in s_2} d_i^{*2} x_i; \quad \bar{x}_w^{12} = \sum_{i\in s_1 \cup s2} d_i^{*12} x_i; \tag{1.24}$$

dove

$$d_i^{*1} = \frac{d_i^1}{\sum_{j\in s1} d_j^1}, \quad d_i^{*2} = \frac{d_i^2}{\sum_{j\in s2} d_j^2}, \quad d_i^{*3} = \frac{d_i^3}{\sum_{j\in s3} d_j^3}, \tag{1.25}$$

$$d_i^{*12} = \frac{d_i^{12}}{\sum_{j\in s1\cup s2} d_j^{12}}, \quad d_i^{*13} = \frac{d_i^{13}}{\sum_{j\in s1\cup s3} d_j^{13}}, \tag{1.26}$$

$$d_i^1 = 1/\pi_i^1, \quad d_i^2 = 1/\pi_i^2, \quad d_i^3 = 1/\pi_i^3, \quad d_i^{12} = 1/\pi_i^{12}, \quad d_i^{13} = 1/\pi_i^{13}. \tag{1.27}$$

Le quantità $\pi_i^1, \pi_i^2, \pi_i^3, \pi_i^{12}, \pi_i^{13}$, sono rispettivamente le probabilità di inclusione di primo ordine dei campioni $s_1,\ s_2,\ s_3,\ s_1 \cup s_2,\ s_1 \cup s_3$. Notiamo che quando $u_i = 0$, si ha $\hat{p}_i = d_i^*$ e lo stimatore di massima pseudo verosimiglianza empirica è uguale allo stimatore per rapporto definito precedentemente, il quale non utilizza l'informazione della variabile ausiliaria x.

Consideriamo lo stimatore di massima pseudo verosimiglianza empirica di $\bar{Y}$:

$$\bar{y}_{PE}^1 = \sum_{i\in s1} \hat{p}_i^1 y_i \tag{1.28}$$

dove $\hat{p}_i^1$ massimizza $l(p^1) = \sum_{i\epsilon s_1} d_i^1 log\ p_i^1$. Considerando il metodo dei moltiplicatori di Lagrange, $\hat{p}_i^1$ è dato da (1.21) e (1.22) dopo aver sostituito d_i con d_i^1 in (1.21) e (1.22). Si evidenzia che tutte queste nuove espressioni sono valide per $i\epsilon s_1$.

Lo stimatore $\bar{y}_{PE}^{1}$ non utilizza l'informazione fornita dai campioni s_2 e s_3 ed è per questo motivo che è utile definire un nuovo stimatore che consideri anche queste informazioni. Dal fatto che, per la variabile di interesse, si abbiano n – p – q valori, il nuovo vettore dei pesi $\hat{p}_i^{12}$ deve essere definito con dimensione n – p – q. Quindi, il nuovo stimatore è dato da

$$\bar{y}_{PE}^{12} = \sum_{i \in s1} \hat{p}_i^{12} y_i \tag{1.29}$$

dove i $\hat{p}_i^{12}$ sono ottenuti come i $\hat{p}_i^{1}$ (che ha dimensione n – p – q); usiamo inoltre il moltiplicatore di Lagrange λ^{12} basato sul campione s_1 e s_2 in (1.21). La quantità λ^{12} è ottenuta da (1.22) dopo aver sostituito d_i con d_i^{12} in (1.22).

Sfortunatamente, lo stimatore proposto $\bar{y}_{PE}^{12}$ non usa l'informazione per la variabile di studio y fornita dal campione s_3. Per risolvere questo problema consideriamo una classe di stimatori, che usa tutta l'informazione della variabile y inclusa in s_1 e s_3.

$$\bar{y}_{PE\alpha} = \alpha \bar{y}_{PE}^{1} + (1-\alpha)\bar{y}_{w}^{3} \tag{1.30}$$

dove α è una costante che assume valori nell'intervallo (0,1). Lo stimatore $\bar{y}_{w}^{3}$ è definito in (1.23).

Lo stimatore ottimo nella classe proposta è lo stimatore definito da (1.30) con un valore α_{opt} che minimizza la varianza asintotica, la quale può essere scritta come

$$V(\bar{y}_{PE\alpha}) = \alpha_{opt}{}^{2} M^{*} + (1-\alpha_{opt})^{2} N^{*} + 2\alpha_{opt}(1-\alpha_{opt}) L^{*} \tag{1.31}$$

dove

$$M^{*} = V(y_w^1) + B^2 V(\bar{x}_w^1) - 2BCov(y_w^1, \bar{x}_w^1), \tag{1.32}$$

$$N^{*} = V(\bar{y}_w^3), \tag{1.33}$$

$$L^{*} = Cov(y_w^1, \bar{y}_w^3) - BCov(\bar{x}_w^1, \bar{y}_w^3). \tag{1.34}$$

e

$$\alpha_{opt} = \frac{N^{*} - L^{*}}{M^{*} + N^{*} - 2L^{*}} \tag{1.35}$$

Una caratteristica interessante dello stimatore basato sulla pseudo verosimiglianza empirica è che i pesi risultanti sono sempre positivi e per questo la tecnica è generalmente applicabile anche alla stima di parametri diversi dalla media di popolazione come per esempio i quantili.

Un'altra situazione considerata in letteratura è relativa al caso in cui i dati mancano separatamente e non simultaneamente per tutte le caratteristiche. Assumiamo la presenza di due variabili quantitative ausiliarie x e z di cui conosciamo le medie $\bar{X}, \bar{Z}$. In pratica, i dati mancanti possono essere presenti in entrambe le variabili sopracitate e per questo motivo si utilizza una struttura per i dati non disponibili come la seguente:

s_1	s_2	s_3	s_4
$y_1 \cdots y_{n-p-q-k}$ $x_1 \ldots x_{n-p-q-k}$ $z_1 \cdots z_{n-p-q-k}$	$y_{n-p-q-k+1} \cdots y_{n-q-k}$ $x_{n-p-q-k+1} \ldots x_{n-q-k}$ Missing .... Missing	$y_{n-q-k+1} \cdots y_{n-k}$ Missing .... Missing $z_{n-q-k+1} \cdots z_{n-k}$	Missing...Missing $x_{n-k+1} \ldots x_n$ $z_{n-k+1} \cdots z_n$

Le informazioni ausiliarie disponibili possono essere utilizzate la costruzione dello stimatore:

$$\hat{\bar{Y}}_{DAI} = \bar{y}_{AI} + B' \begin{pmatrix} \bar{X} - \bar{x}_{AI} \\ \bar{Z} - \bar{z}_{AI} \end{pmatrix} \tag{1.36}$$

dove $\bar{y}_{AI}, \bar{x}_{AI}, \bar{z}_{AI}$, sono stimatori di Horvitz-Thompson calcolati per le rispettive variabili e rappresentati come:

$$\bar{y}_{AI} = \sum_{i \in s1 \cup s2 \cup s3} \frac{y_i}{\pi_i}, \qquad \bar{x}_{AI} = \sum_{i \in s1 \cup s2 \cup s4} \frac{x_i}{\pi_i}, \qquad \bar{z}_{AI} = \sum_{i \in s1 \cup s3 \cup s4} \frac{z_i}{\pi_i}. \tag{1.37}$$

Minimizzando la varianza dello stimatore (1.36) rispetto a B si consegue che il valore ottimo $B_{opt} = \Sigma^{-1}\sigma$ con

$$\sum = \begin{pmatrix} V(\bar{x}_{AI}) & Cov(\bar{x}_{AI}, \bar{z}_{AI}) \\ Cov(\bar{x}_{AI}, \bar{z}_{AI}) & V(\bar{z}_{AI}) \end{pmatrix} \quad \text{e} \quad \sigma = \begin{pmatrix} Cov(\bar{y}_{AI}, \bar{x}_{AI}) \\ Cov(\bar{y}_{AI}, \bar{z}_{AI}) \end{pmatrix}. \tag{1.38}$$

Si noti che l'espressione approssimata dello stimatore dipende dai valori della varianza e covarianza tra le medie degli stimatori, calcolate in s_1, s_2, s_3 e s_4. Usando, quindi, i valori campionari di Σ e σ e sostituendoli in (1.36) si ottiene lo stimatore ottimo di $\bar{Y}$:

$$\hat{\bar{Y}}_{DAIopt} = \bar{y}_{AI} + \hat{\Sigma}^{-1}\hat{\sigma}' \begin{pmatrix} \bar{X} - \bar{x}_{AI} \\ \bar{Z} - \bar{z}_{AI} \end{pmatrix} \tag{1.39}$$

Nel caso in cui il disegno di campionamento sia un SRSWOR, le espressioni di $\hat{\Sigma}$ *e* $\hat{\sigma}$ sono le stime delle seguenti varianze e covarianze: (1.40), (1.41), (1.42), (1.43).

$$V(\bar{x}_{AI}) = \frac{S_x^2}{n-q}\left(1-\frac{n-q}{N}\right), \quad V(\bar{z}_{AI}) = \frac{S_z^2}{n-p}\left(1-\frac{n-p}{N}\right),$$

$$Cov(\bar{x}_{AI}, \bar{z}_{AI}) = \frac{(n-p-q)^2}{(n-q)(n-p)} Cov(\bar{x}_{s_1 \cup s_4}, \bar{z}_{s_1 \cup s_4}) + \frac{p(n-p-q)}{(n-q)(n-p)} Cov(\bar{x}_{s_2}, \bar{z}_{s_1 \cup s_4})$$
$$+ \frac{q(n-p-q)}{(n-q)(n-p)} Cov(\bar{x}_{s_1 \cup s_4}, \bar{z}_{s_3}) + \frac{pq}{(n-q)(n-p)} Cov(\bar{x}_{s_2}, \bar{z}_{s_3})$$

$$Cov(\bar{y}_{AI}, \bar{x}_{AI}) = \frac{(n-q-k)^2}{(n-k)(n-q)} Cov(\bar{y}_{s_1 \cup s_2}, \bar{x}_{s_1 \cup s_2}) + \frac{k(n-q-k)}{(n-k)(n-q)} Cov(\bar{y}_{s_1 \cup s_2}, \bar{x}_{s_4})$$
$$+ \frac{q(n-q-k)}{(n-k)(n-q)} Cov(\bar{y}_{s_3}, \bar{x}_{s_1 \cup s_2}) + \frac{qk}{(n-k)(n-q)} Cov(\bar{y}_{s_3}, \bar{x}_{s_4})$$

$$Cov(\bar{y}_{AI}, \bar{z}_{AI}) = \frac{(n-p-k)^2}{(n-k)(n-p)} Cov(\bar{y}_{s_1 \cup s_3}, \bar{z}_{s_1 \cup s_3}) + \frac{k(n-p-k)}{(n-k)(n-p)} Cov(\bar{y}_{s_1 \cup s_3}, \bar{z}_{s_4})$$
$$+ \frac{p(n-p-k)}{(n-k)(n-p)} Cov(\bar{y}_{s_2}, \bar{z}_{s_1 \cup s_3}) + \frac{pk}{(n-k)(n-p)} Cov(\bar{y}_{s_2}, \bar{z}_{s_4})$$

Il metodo proposto può essere esteso facilmente al caso di più di due variabili ausiliarie; inoltre, si dimostra che $\hat{\bar{Y}}_{DAIopt}$ è più efficiente rispetto al tradizionale stimatore della media.

1.2.2 Campionamento a due fasi

Supponiamo di estrarre un campione *s* contenente dati mancanti. In situazioni semplici, possiamo vedere il campione appena estratto come un campione proveniente da un disegno a due fasi. Nella prima fase, il campione contiene tutte le unità estratte col meccanismo casuale semplice. Verrà, quindi, eseguito in seconda fase un esperimento Bernoulliano per selezionare un sottocampione che rappresenterà i soggetti rispondenti.

Siano y_i, $i \epsilon s$ i valori della variabile di interesse delle unità rispondenti nel campione; sia z_i la variabile indicatrice che prende valori 1 o 0 a seconda che l'unità abbia risposto o meno. Allora, la media campionaria dei rispondenti è data:

$$\bar{y}_r = \frac{\sum_{i \in s} z_i y_i}{\sum_{i \in s} z_i} \tag{1.44}$$

Assumiamo che gli z_i siano indipendenti tra loro e che il meccanismo di risposta sia bernoulliano ($P(z_i = 1) = p$ con $0 < p < 1$). Effettuando analisi asintotiche è possibile concludere che la quantità $\bar{y}_r - \bar{Y}$ è asintoticamente normale con media 0 e varianza $n^{-1}(p^{-1} - f)\sigma_y^2$, dove $f = \frac{n}{N}$ rappresenta la frazione di campionamento. Notiamo inoltre che s_y^2 converge in probabilità a σ_y^2 e che uno stimatore consistente della varianza asintotica è dato da $n^{-1}(r^{-1} - N^{-1})s_{2y}^2$, dove r è il numero di rispondenti all'interno del campione e s_{2y}^2 è la varianza campionaria dei rispondenti.

In molte situazioni di importanza pratica il problema di stimare la media di una popolazione assume maggiore rilevanza quando la media della popolazione della variabile ausiliaria x è sconosciuta e si è in presenza di dati mancanti. In una situazione di questo tipo, riprendendo il campione precedente si definisce il campionamento a due fasi come segue:

(i) Si seleziona un campione di grandi dimensioni di dimensione n' usando il meccanismo SRSWOR e si osserva la variabile x per queste unità;

(ii) Dalle unità selezionate in prima fase (n'), si considera un campione di seconda fase di dimensione n usando sempre lo stesso disegno campionario. Le unità rispondenti sono n_1 e le non rispondenti le rimanenti n_2 ($n_1 + n_2 = n$).

(iii) Dalle n_2 unità non rispondenti si seleziona un sottocampione di dimensione r ($r = n_2 / k$, $k > 1$) usando il meccanismo SRSWOR e si osserva la variabile y per queste unità.

Hansen e Hurwitz definirono lo stimatore di $\bar{Y}$ come:

$$\bar{y}^* = (n_1 / n)\bar{y}_1 + (n_2 / n)\bar{y}_2' \tag{1.45}$$

dove $\bar{y}_i = \sum_{i=1}^{n_1} \frac{y_i}{n_1}$ *e* $\bar{y}_2' = \sum_{i=1}^{r} \frac{y_i}{r}$ sono le medie di y della porzione dei rispondenti del campione e del sottocampione. Lo stimatore è non distorto con varianza data da

$$V(\bar{y}^*) = \lambda S_y^2 + \theta S_{y_2}^2 \tag{1.46}$$

dove (1.47)

$$W_2 = N_2 / N \ , \ \lambda = (1-f)/n \ , \ \theta = W_2(k-1)/n \ , \ \bar{Y}_2 = \sum_{i=1}^{N} \frac{y_i}{N_2} \ , \ S_{y_2}^2 = \sum_{i=1}^{N_2} \frac{(y_i - \bar{Y})^2}{(N_2 - 1)}$$

Questo stimatore può essere migliorato con l'utilizzo di informazioni ausiliarie sulla popolazione. Quando si conosce $\bar{X}$ e le informazioni sulle variabili y e x per le *n* unità selezionate sono incomplete, è possibile definire uno stimatore rapporto come il seguente:

$$T_{R1} = \bar{y}^* \left(\frac{\bar{X}}{\bar{x}^*} \right) \tag{1.48}$$

dove $\bar{x}^* = \left(\frac{n_1}{n}\right)\bar{x}_1 + \left(\frac{n_2}{n}\right)\bar{x}_2'$, $\bar{x}_1 = \sum_{i=1}^{n_1} \frac{x_i}{n_1}$, $\bar{x}_2' = \sum_{i=1}^{r} \frac{x_i}{r}$, e $\bar{Y}^*$ è definito come $\bar{X}^*$.

Uno stimatore per rapporto alternativo, usando $\bar{X}$, avendo informazioni complete per la variabile ausiliaria x ed incomplete sulla variabile di studio y, è dato da:

$$T_{R2} = \bar{y}^* \left(\frac{\bar{X}}{\bar{x}} \right) \tag{1.49}$$

Si definiscono, ora, due stimatori prodotto della media della popolazione in presenza di non risposte.

$$T_{P1} = \bar{y}^* \left(\frac{\bar{x}^*}{\bar{X}} \right) \quad \text{e} \quad T_{P2} = \bar{y}^* \left(\frac{\bar{x}}{\bar{X}} \right) \tag{1.50}$$

Tuttavia, non è rara una situazione in cui non siano note informazioni sulla media dalla variabile ausiliaria x; in questo caso gli stimatori presentati non possono essere utilizzati e il parametro ignoto $\bar{X}$ è sostituito da un suo stimatore non distorto costruito su un campione preliminare di grande dimensione n'.

Considerando i campioni definiti precedentemente per il disegno a due fasi si esaminano le seguenti quantità: n' osservazioni di x per il campione di prima fase, n_1 osservazioni di y per le unità rispondenti prese dal campione di seconda fase e r osservazioni di y selezionate dalle n_2 unità non rispondenti. Sia $\bar{x}'$ la media campionaria della variabile ausiliaria x basata sul campione di prima fase. Usando queste informazioni, nel seguito si propongono due classi di stimatori della media di popolazione considerando due differenti scenari.

Scenario 1. La media di popolazione $\bar{X}$ è ignota e le informazioni della variabile di studio y e della variabile ausiliaria x sono incomplete.

In questa situazione, per stimare $\bar{X}$ usiamo (n_1 + r) unità rispondenti per y, x e $\bar{x}'$. Con queste premesse, si può definire una versione generalizzata dello stimatore per rapporto di $\bar{Y}$:

$$t_{(a)d}^{(1)} = \bar{y}^* \exp\left[a \left\{ \frac{(\bar{x}^* - \bar{x}')}{(\bar{x}^* + \bar{x}')} \right\} \right] \tag{1.51}$$

dove a è uno scalare opportunamente scelto. La varianza dello stimatore generalizzato approssimata al primo ordine è data da: (1.52)

$$Var(t_{(a)d}^{(1)}) = \lambda^* \left[S_y^2 + (aR^2 S_x^2 / 4)(a + 4C) \right] + \lambda' S_y^2 + \theta \left[S_{y_2}^2 + (aR^2 S_{x_2}^2 / 4)(a + 4C_{(2)}) \right]$$

dove

$$\begin{gathered} \lambda' = (1-f')/n', \;\; f' = n'/N', \;\; \lambda^* = \left[(1/n) - (1/n') \right], \;\; \beta_{(2)} = (S_{xy(2)} / S_{x_2}^2) \\ S_{xy(2)} = \sum_{i=1}^{N_2} (x_i - \bar{X}_2)(y_i - \bar{Y}_2)/(N_2 - 1), \;\; S_{x_2}^2 = \sum_{i=1}^{N_2} (x_i - \bar{X}_2)/(N_2 - 1) \\ C = (\beta / R), \;\; C_{(2)} = (\beta_{(2)} / R) \end{gathered} \tag{1.53}$$

Lo stimatore rapporto proposto è più efficiente dello stimatore $\bar{y}^*$. Da (1.46), è possibile verificare che la differenza $\left[Var(\bar{y}^*) - Var\left(t_{(a)d}^{(1)} \right) \right]$, è positiva se:

$$\begin{aligned} &\text{o} \quad -4C < a < 0 \quad \text{e} \quad -4C_{(2)} < a < 0 \\ &\text{o} \quad 0 < a < -4C \quad \text{e} \quad 0 < a < -4C_{(2)} \end{aligned} \tag{1.54}$$

Scenario 2. La media di popolazione $\bar{X}$ è ignota, la variabile di studio y presenta informazioni incomplete a differenza della variabile ausiliaria x. In questo caso, la versione generalizzata dello stimatore per rapporto è così definita:

$$t_{(b)d}^{(2)} = \bar{y}^* \exp\left[b \left\{ \frac{(\bar{x} - \bar{x}')}{(\bar{x} + \bar{x}')} \right\} \right] \tag{1.55}$$

dove b è uno scalare opportunamente scelto. La varianza dello stimatore approssimato al primo ordine è dato da:

$$Var(t_{(b)d}^{(2)}) = \lambda S_y^2 + \theta S_{y_2}^2 + \lambda^* (bR^2 S_x^2 / 4)(b + 4C) \tag{1.56}$$

Anche in questo caso, lo stimatore proposto è più efficiente dello stimatore usuale $\bar{y}^*$. Da (1.46) è possibile verificare che la differenza $\left[Var(\bar{y}^*) - Var\left(t_{(b)d}^{(2)} \right) \right]$, è positiva se:

$$\text{o} \quad -4C < b < 0 \quad \text{o} \quad 0 < b < -4C \tag{1.57}$$

Minimizzando l'espressione della varianza (1.52) rispetto al parametro a troviamo che il valore ottimo del parametro è dato da

$$a = -(2D^{**} / D^*) = a_0 \tag{1.58}$$

dove $D^* = \left[\lambda^* S_x^2 + \theta S_{x_2}^2 \right]$ e $D^{**} = \left[\lambda^* C S_x^2 + \theta C_{(2)} S_{x_2}^2 \right]$.

Sostituendo a con a_0 in $t_{(a)d}^{(1)}$, lo stimatore ottimo della media di popolazione diventa

$$t_{(a_0)d}^{1} = \bar{y}^{*} \exp\left[a_0 \left\{ \frac{(\bar{x}^{*} - \bar{x}')}{(\bar{x}^{*} + \bar{x}')} \right\} \right]. \tag{1.59}$$

Inserendo, quindi, (1.58) in (1.52), si ottiene che la minima varianza di $t_{(a)d}^{(1)}$ è

$$\min \mathrm{Var}(t_{(a)d}^{1}) = Var(t_{(a_0)d}^{1}) = \left[\lambda S_y^2 + \theta S_{y_2}^2 - R^2 (D^{**2} / D^{*}) \right]. \tag{1.60}$$

Il valore ottimo di a_0 è però ignoto. Tuttavia, per ottenere lo stimatore sostituiamo a_0 con una sua stima consistente $\hat{a}_0 = -\left(2\widehat{D}^{**}/\widehat{D}^{*}\right)$ dove $\widehat{D}^{*} = \left[\lambda^{*}\hat{S}_x^2 + \theta\hat{S}_{x_2}^2\right]$ e $\widehat{D}^{**} = \left[\lambda^{*}\hat{S}_{yx}^2 + \theta\hat{S}_{yx(2)}^2\right]/\hat{R}$.

La varianza di $t_{(b)d}^{(2)}$ viene minimizzata se b = - 2C. Sostituendo questo risultato a (1.56), si ottiene che la minima varianza dello stimatore risulta essere

$$\min \mathrm{Var}(t_{(b)d}^{2}) = \left[\lambda^{*} S_y^2 (1 - \rho^2) + \theta S_{y_2}^2 + \lambda' S_y^2 \right]. \tag{1.61}$$

Come nel caso precedente, il valore di b_0 è ignoto ed è possibile sostituirlo con una sua stima consistente costruita sui dati campionari, definendo lo stimatore

$$t_{(\hat{b}_0)d}^{2} = \bar{y}^{*} \exp\left[\hat{b}_0 \left\{ \frac{(\bar{x} - \bar{x}')}{(\bar{x} + \bar{x}')} \right\} \right] \tag{1.62}$$

dove $\hat{b}_0 = -2\hat{C},\ \hat{C} = (\hat{\beta} / \hat{R})$

Sotto lo stesso disegno di campionamento a due fasi è stato proposto, in letteratura, un altro tipo di stimatore di tipo rapporto e uno stimatore per regressione: si considera la situazione in cui l'informazione sulla variabile ausiliaria x è completamente disponibile per tutto il campione di n unità di seconda fase; mentre mancano informazioni sulla variabile di studio y. Definiamo, quindi, i seguenti stimatori di $\bar{Y}$:

$$t_{(\alpha 1)}^{(\alpha 2)} = \bar{y}^{*} \left(\frac{\bar{x}}{\bar{x}^{*}} \right)^{\alpha 1} \left(\frac{\bar{x}'}{\bar{x}} \right)^{\alpha 2} \tag{1.63}$$

e

$$t_d = \bar{y}^{*} + d_1(\bar{x} - \bar{x}^{*}) + d_2(\bar{x}' - \bar{x}), \tag{1.64}$$

dove α_i e d_i (i = 1, 2) sono costanti opportunamente scelte. La varianza esatta dello stimatore (1.64) e la varianza approssimata al primo ordine per lo stimatore (1.63), sono rispettivamente date da:

$$Var(t_d) = \left[\left(\frac{1}{n} - \frac{1}{n'}\right)\left\{S_y^2 + d_2 S_x^2 (d_2 - 2\beta)\right\} + \frac{(N_2 / N)(k-1)}{n}\left\{S_{y_2}^2 + d_1 S_{x_2}^2 (d_1 - 2\beta_{(2)})\right\} + \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2\right], \tag{1.65}$$

$$Var(t_{(\alpha 1)}^{(\alpha 2)}) = \left[\left(\frac{1}{n} - \frac{1}{n'}\right)\left\{S_y^2 + R\alpha_2 S_x^2 (R\alpha_2 - 2\beta)\right\} + \frac{(N_2 / N)(k-1)}{n}\left\{S_{y_2}^2 + R S_{x_2}^2 \alpha_1 (R\alpha_1 - 2\beta_{(2)})\right\} + \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2\right]. \tag{1.66}$$

Siamo in grado di minimizzare la (1.65) e (1.66) rispetto a (d_1, d_2) e (α_1, α_2) con

$$\left.\begin{aligned} d_1 &= \beta_{(2)} = d_{10} \\ d_2 &= \beta = d_{20} \end{aligned}\right\} \tag{1.67}$$

e

$$\left.\begin{aligned} \alpha_1 &= \beta_{(2)} / R = \alpha_{10} \\ \alpha_2 &= \beta / R = \alpha_{20} \end{aligned}\right\} \tag{1.68}$$

Quindi, sostituendo (1.67) in (1.64) e (1.68) in (1.63), si possono ottenere gli stimatori ottimali delle classi t_d *e* $t_{(\alpha_1)}^{(\alpha_2)}$:

$$t_{(d10)}^{(d20)} = \bar{y}^* + \beta_{(2)}(\bar{x} - \bar{x}^*) + \beta(\bar{x}' - \bar{x}) \tag{1.69}$$

$$t_{(\alpha 10)}^{(\alpha 20)} = \bar{y}^* \left(\frac{\bar{x}}{\bar{x}^*}\right)^{\beta_{(2)}/R} \left(\frac{\bar{x}'}{\bar{x}}\right)^{\beta/R} \tag{1.70}$$

È facile verificare che lo stimatore ottimo (1.69) è non distorto mentre lo stimatore (1.70) presenta distorsione; proprio per questo motivo lo stimatore per regressione è preferibile agli stimatori per rapporto fin qui esaminati. La varianza esatta di (1.69) è data da:

$$Var\left(t_{(d10)}^{(d20)}\right) = \left[\left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right) S_y^2 (1 - \rho^2) + \frac{(N_2 / N)(k-1)}{n} S_{y_2}^2 (1 - \rho_2^2)\right] \tag{1.71}$$

dove $\rho_2 = (S_{xy(2)} / S_{x_2} S_{y_2})$ è il coefficiente di correlazione tra la variabile y e x nel gruppo delle non risposte.

Usando (1.68) in (1.66) si ha che la varianza di (1.70) approssimata al primo ordine è la stessa dello stimatore (1.69).

Si osserva da (1.69) e (1.70) che gli stimatori ottimali possono essere utilizzati solo quando i valori esatti di β, β_2 e R sono noti. Tuttavia, se ciò non accade, è consigliabile sostituire tali parametri con delle stime consistenti basate sui dati campionari.

Siano

$$\hat{\beta} = \frac{s^*_{xy}}{s^{*2}_x} \qquad \hat{\beta}_{(2)} = \frac{s_{xy(2)}}{s^2_{x_2}} \tag{1.72}$$

$$\hat{R} = \frac{\bar{y}^*}{\bar{x}^*} \tag{1.73}$$

stimatori consistenti di β, β_2 e R; sostituendoli in (1.69) e (1.70) si ottengono due stimatori consistenti e ottimi di $\bar{Y}$.

$$t_{lrd} = \bar{y}^* + \hat{\beta}_{(2)}(\bar{x} - \bar{x}^*) + \hat{\beta}(\bar{x}' - \bar{x}) \tag{1.74}$$

$$\hat{t}_e = \bar{y}^* \left(\frac{\bar{x}}{\bar{x}^*}\right)^{\hat{\beta}_{(2)}/\hat{R}} \left(\frac{\bar{x}'}{\bar{x}}\right)^{\hat{\beta}/\hat{R}} \tag{1.75}$$

1.3 Dati mancanti MAR

Se assumiamo che il meccanismo di generazione dei dati mancanti è di tipo MAR, la probabilità che una data osservazione sia mancante, non dipende dai valori della variabile di studio, ma è dipendente dai valori delle variabili ausiliarie. Questo tipo di meccanismo coinvolge sia la modellazione della regressione di y sulle variabili ausiliarie associate, che la modellazione della probabilità di risposta. Per questo motivo, possiamo dire che i dati mancanti provenienti da un meccanismo MAR rispecchiano molto più la realtà rispetto ai dati mancanti di tipo MCAR.

1.3.1 Stimatori doppiamente robusti

Allo scopo di stimare $\bar{Y}$, è stata introdotta in letteratura una classe di stimatori basata sull'inverso della probabilità che coinvolge sia la modellazione della regressione di y sulle variabili ausiliarie associate, che la modellazione della probabilità di risposta.

Gli stimatori di questa classe sono doppiamente robusti in quanto sono consistenti per la vera media di popolazione persino se uno dei due modelli illustrati non è specificato (ma non entrambi).

Supponiamo, ora, che Y_i non sia disponibile per tutti i soggetti; i dati attuali osservati sono strutturati nel seguente modo: (z_iY_i , z_i , X_i) (i = 1,…, *n*), dove $z_i = 1$ o 0 a seconda se Y_i viene osservata o meno. Infine, si assume che i dati mancanti siano di tipo MAR in quanto Y_i e z_i sono indipendenti tra loro.

La probabilità di risposta data da $Pr(z_i = 1 \mid X)$ viene indicata con $p_0(x_i)$; solitamente risulta sconosciuta ed è per questo motivo che si ricorre ad un modello parametrico come rappresentato nella seguente scrittura: p(X, γ). A questo punto possiamo considerare una classe di stimatori doppiamente robusti di $\bar{Y}$:

$$\hat{\mu}_{DR} = n^{-1}\sum_{i=1}^{n}\left\{\frac{z_i Y_i}{p(X_i\ ,\ \hat{\gamma})} - \frac{z_i - p(X_i\ ,\ \hat{\gamma})}{p(X_i\ ,\ \hat{\gamma})} m(X_i\ ,\ \hat{\beta})\right\} \tag{1.76}$$

dove $m(X_i\ ,\hat{\beta})$ stima il modello di regressione della variabile risposta sulle variabili ausiliarie $m(X_i\ ,\beta)$ e $\hat{\gamma}$ è lo stimatore di massima verosimiglianza di γ.

Supponiamo che la probabilità di risposta sia correttamente specificata dal modello $p(X) = p_0(X)$, a differenza di m(X , β) che può non esserlo. Esaminiamo in seguito le metodologie per stimare il parametro β per ottenere uno stimatore di $\bar{Y}$ che sia (i) doppiamente robusto e (ii), se la probabilità di risposta è specificata, abbia la più piccola varianza asintotica tra tutti gli stimatori nella forma (1.76).

Consideriamo lo stimatore

$$n^{-1}\sum_{i=1}^{n}\left\{\frac{z_i Y_i}{p_0(X_i)} - \frac{z_i - p_0(X_i)}{p_0(X_i)} m(X_i\ ,\ \beta^*)\right\} \tag{1.77}$$

dove β^* è il limite in probabilità di $\hat{\beta}$. L'obiettivo identificare il valore β^*, e quello della stima corrispondente $\hat{\beta}$, tale per cui la varianza di (1.77) sia minimizzata. Sia $m_\beta(X\ ,\beta) = \partial/\partial\beta\{m(X\ ,\beta)\}$, notiamo che la varianza minima si ottiene scegliendo un valore di β^* che sia soluzione di

$$E\left[\frac{1-p_0(X)}{p_0(X)}\left\{m_0(X) - m(X,\beta^*)\right\} m_\beta(X,\beta)\right] = 0 \tag{1.78}$$

indicandolo con β^*_{opt}.

Consideriamo l'ordinario stimatore dei minimi quadrati per β, $\hat{\beta}_1$, ad esempio, risolvendo

$$n^{-1}\sum_{i=1}^{n} z_i \left\{Y - m\left(X_i\ ,\ \beta\right)\right\}\ m_\beta(X_i\ ,\ \beta) = 0 \tag{1.79}$$

Se la probabilità di risposta è specificata, mentre il modello $m(X\ ,\ \beta) \neq m_0(X)$ per ogni β, allora la parte sinistra dell'equazione (1.78) converge in probabilità a

$$E\Big[\, p_0(X)\{m_0(X) - m(X,\beta)\}\, m_\beta(X,\beta)\Big]. \tag{1.80}$$

Quindi $\hat{\beta}_1$ converge in probabilità a β_1 in modo tale che (1.80) risulti pari a zero; ciò non esclude il fatto che confrontando (1.80) con (1.78) risulti $\beta_1 \neq \beta^*_{opt}$. Se, invece, la probabilità di risposta non è specificata, mentre il modello di regressione è corretto, la parte sinistra dell'equazione (1.79) converge ancora in probabilità a (1.80); $\beta_1 = \beta_0$, in modo tale che $\hat{\beta}_1$ converga in probabilità a β_0. Gli stimatori (1.77), usando $\hat{\beta}_1$, sono doppiamente robusti ma non raggiungono la minima varianza quando il modello di regressione non è specificato completamente. Nel caso in cui volessimo stimare β minimizzando la varianza empirica di (1.77), ci accorgeremo che lo stimatore non sarà più doppiamente robusto.

Per soddisfare (i) e (ii) simultaneamente, prendiamo in considerazione la soluzione di

$$n^{-1}\sum_{i=1}^{n} \frac{z_i}{p_0(X_i)} \frac{1 - p(X_i)}{p(X_i)} \left\{Y_i - m(X_i\ ,\ \beta)\right\}\ m_\beta(X_i\ ,\ \beta) = 0 \tag{1.81}$$

chiamata $\hat{\beta}_3$, che può essere vista come stima ai minimi quadrati ponderati con pesi pari a $\{1 - p(X_i)\}/p^2(X_i)$. Adottando lo stesso procedimento visto in precedenza, si può verificare che lo stimatore (1.77) per $\hat{\beta} = \hat{\beta}_3$ risulta doppiamente robusto e rispetta la proprietà di minima varianza asintotica persino se il modello di regressione non è specificato.

Nella pratica, si potrebbe supporre parametrico il modello della probabilità di risposta, $p(X\ ,\ \gamma)$. In questo caso possiamo utilizzare i risultati precedenti direttamente per trovare uno stimatore della media di y nella forma (1.76), dove $\hat{\gamma}$ è la stima di massima verosimiglianza per la regressione binaria, che soddisfa le condizioni (i) e (ii). In analogia a (1.81), si propone una stima di β risolvendo congiuntamente l'equazione seguente in (β, c): (1.82)

$$\sum_{i=1}^{n}\left[\frac{z_i}{p(X_i\ ,\ \hat{\gamma})} \frac{1 - p(X_i\ ,\ \hat{\gamma})}{p(X_i\ ,\ \hat{\gamma})} \left\{\frac{m_\beta(X_i\ ,\ \beta^*)p_\gamma(X_i\ ,\ \hat{\gamma})}{1 - p(X_i\ ,\ \hat{\gamma})}\right\} \left\{Y_i - m(X_i\ ,\ \beta) - c' \frac{p_\gamma(X_i\ ,\ \hat{\gamma})}{1 - p(X_i\ ,\ \hat{\gamma})}\right\}\right] = 0$$

Con un argomento del tutto analogo a quello riportato seguendo (1.81), quando la probabilità di risposta è corretta, ma $m(X, \beta)$ non lo è, si nota che la soluzione di

(1.82) $\hat{\beta}_4$, converge in probabilità a β_{opt}^{**}. Viceversa, la quantità alla sinistra di (1.82) converge in probabilità a zero quando (β, c) = (β_0, 0). Prendendo $\hat{\beta} = \hat{\beta}_4$ in (1.76), lo stimatore rispetta entrambe le proprietà (i) e (ii).

Indichiamo altri tre stimatori doppiamente robusti con la caratteristica di essere legati al campione; questo significa che vengono escluse le stime che stanno fuori dal range del campione.

Il primo stimatore proposto è

$$\tilde{\mu}_{LIK} = n^{-1}\sum_{i=1}^{n}\frac{z_i Y_i}{p(X_i\ ,\ \tilde{\gamma})} \tag{1.83}$$

il quale, oltre a rispettare le due caratteristiche sopracitate, è localmente ed intrinsecamente efficiente. Lo stimatore proposto è doppiamente robusto perché

$$\tilde{E}\left\{\frac{z}{p(X,\tilde{\gamma})}\hat{m}(X)\right\} = \tilde{E}\left\{\hat{m}(X)\right\} \tag{1.84}$$

e quindi $\tilde{\mu}_{LIK}$ è del tipo $\hat{\mu}\{p(\cdot\,,\tilde{\gamma}),\hat{m}\}$, ossia nella tipica forma degli stimatori doppiamente robusti. Consideriamo, infine, $\tilde{\gamma}$ come una trasformazione della stima di massima verosimiglianza fatta sul modello lineare della probabilità di risposta. Ciò non implica il fatto che non permangano problemi circa l'esistenza e il calcolo di $\tilde{\gamma}$. In primo luogo, è difficile caratterizzare condizioni in cui esiste una soluzione sottoposta al vincolo $p(X_i\,,\gamma) > 0$. In secondo luogo, si può presentare il problema che non esistano soluzioni o ci siano soluzioni multiple; la difficoltà sta nel fatto di selezionare $\tilde{\gamma}$ tra tutte le possibili soluzioni.

Il secondo stimatore proposto tenta di risolvere i problemi appena elencati utilizzando una robustificazione della stima basata sulla verosimiglianza, che consiste nel calibrare i coefficienti del modello lineare esteso alla probabilità di risposta. Lo stimatore risulta conveniente nel calcolo ed è basato sulla massimizzazione a due step di funzioni concave. Inoltre, lo stimatore è localmente ed intrinsecamente efficiente, è legato al campione, e migliora ulteriormente in termini di efficienza nel caso in cui la funzione di regressione è opportunamente stimata. Non esistono altri stimatori doppiamente robusti che rispettano queste quattro proprietà simultaneamente. Lo stimatore descritto è nella forma:

$$\tilde{\mu}_{LIK2} = n^{-1}\sum_{i=1}^{n}\frac{z_i Y_i}{p(X_i\ ,\ \tilde{\gamma}_{step2})} \tag{1.85}$$

Diverse scelte del modello di regressione $\hat{m} = (x_i, \beta)$, portano a specifiche versioni di $\tilde{\mu}_{LIK2}$: si indicano con $\tilde{\mu}_{LIK2,OLS}, \tilde{\mu}_{LIK2,WLS}$ e $\tilde{\mu}_{LIK2,RV}$ le versioni corrispondenti a $\hat{m} = (x_i, \hat{\beta}_{OLS})$, $\hat{m} = (x_i, \hat{\beta}_{WLS})$, $\hat{m} = (x_i, \hat{\beta}_{RV})$. Le stime per β, sono rispettivamente ai minimi quadrati, ai minimi quadrati ponderati con pesi $p^{-1}(x_i, \hat{\gamma})$ e ai minimi quadrati ponderati con pesi diversi $p^{-1}(x_i, \hat{\gamma})(p^{-1}(x_i, \hat{\gamma}) - 1)$.

Il terzo ed ultimo stimatore si ottiene robustificando lo stimatore proposto da Tan (2006), aggiungendo il seguente termine:

$$n^{-1}\sum_{i=1}^{n}\left[\left\{z_i / p(x_i, \hat{\gamma}) - 1\right\}\hat{m}(x_i)\right] \tag{1.86}$$

Gli stimatori proposti sono intrinsecamente efficienti perché se la probabilità di risposta è specificata, ogni stimatore è asintoticamente efficiente nella classe degli stimatori che utilizzano lo stesso modello di regressione m(X , β). Sono, inoltre, localmente efficienti perché l'efficienza, almeno asintoticamente, è uguale a quella degli stimatori che usano la vera probabilità di risposta.

1.3.2 L'approccio semiparametrico

In letteratura, si utilizza un approccio semiparametrico per stimare la media della popolazione $\bar{Y}$ quando si vuole riconciliare l'approccio regressivo non parametrico con gli approcci model-based. Si adotta una regressione non parametrica per diminuire la dipendenza dal modello specificato mentre si utilizza l'informazione data a priori dal modello su E(Y | X) per migliorare l'efficienza. Siccome la covariata contiene informazioni riguardanti la variabile di studio y, si utilizza una funzione parametrica S = S(X) per riassumere tale informazione.

Indichiamo con S una funzione continua da $\mathbb{R}^d$ *a* $\mathbb{R}$, tale che S = S(X) sia univariata; la media condizionata di Y dato S viene indicata con m(S) = E(Y | S). Poiché E(Y) = E{E(Y | S)}, per una arbitraria S, $n^{-1}\sum_{i=1}^{n} m(S_i)$ risulta non distorto per il parametro θ = E(Y), dove m(S) è stimata con una regressione non parametrica univariata di Y verso S. Assunto ciò, possiamo definire lo stimatore

$$\hat{\theta} = n^{-1}\sum_{i=1}^{n}\hat{m}(S_i) \tag{1.87}$$

Rispetto all'approccio basato sulla regressione non parametrica, l'utilizzo di una funzione indicizzata riduce la dimensione nella regressione da *d* a 1. Rispetto, invece,

agli approcci model-based, lo stimatore semiparametrico di dimensione ridotta risulta essere più robusto nel caso in cui il modello non sia completamente specificato.

La stima non parametrica di m(S) non può essere semplicemente eseguita sui dati completi, ma deve tenere conto anche dei dati mancanti generati con un meccanismo MAR. Un modo per affrontare quest'ultimo problema è utilizzare una regressione non parametrica pesata con l'inverso delle probabilità di risposta. Per ogni *s*, si inserisce la media ponderata di regressione lineare centrata in s con $\alpha = (\alpha_0, \alpha_1)'$ minimizzando

$$\sum_{j=1}^{n} \frac{z_i}{p(X_j)} K_h(S_j - s)\left\{Y_j - \alpha_0 - \alpha_1(S_j - s)\right\}^2 \qquad (1.88)$$

dove $\alpha = \alpha(s)$ varia con *s*, $K_h(u) = h^{-1}K(u/h)$, K è la funzione Kernel ed è generalmente una funzione di densità simmetrica e $h = h_n$ è la larghezza di banda lisciata. La stima di m(s) è $\hat{m}(s) = \hat{\alpha}_0$ e per ogni *s*, α può essere stimato da (1.88) attraverso la regressione ai minimi quadrati pesati. Dimostrazione del fatto è data da: $\hat{m}(s) = \sum_{j=1}^{n} w_j y_j / \sum_{j=1}^{n} w_j$ con $w_j = \{z_j / p(X_j)\} K_h(S_j - s)\{A_{n,2} - A_{n,1}(S_j - s)\}$ e $A_{n,l} = n^{-1}\sum_{j=1}^{n}\{z_j / p(X_j)\} K_h(S_j - s)(S_j - s)^l$ per $l = 1, 2$

Se ci restringiamo alla condizione $\alpha_1 = 0$ in (1.88), si ottiene la stima Nadaraya-Watson basata sull'inverso della probabilità di m(s):

$$\hat{m}(s) = \sum_{j=1}^{n} \frac{z_j}{p(X_j)} K_h(S_j - s) Y_j \Big/ \sum_{j=1}^{n} \frac{z_j}{p(X_j)} K_h(S_j - s). \qquad (1.89)$$

Si può dimostrare che conoscendo la probabilità di risposta p(X), la varianza asintotica dello stimatore (1.87) è data da:

$$n^{-1}\left\{Var(Y) + E\left[\left(p(X)^{-1} - 1\right) Var(Y \mid S)\right]\right\} \qquad (1.90)$$

In conclusione possiamo dire che quando si conosce o è possibile specificare correttamente la probabilità di risposta p(X), c'è poco interesse per quello che concerne la scelta di S, poiché (1.88) è consistente per qualunque S. Quando, invece, l'informazione non permette di definire correttamente p(X), la formulazione di S non è unica: $\hat{\theta}$ è consistente a patto che E(Y | X) = m(s).

1.3.3 Metodi di aggiustamento per ponderazione

Il problema dei dati mancanti, può essere risolto utilizzando una metodologia di aggiustamento basata sulla ponderazione delle risposte. Questi adeguamenti hanno lo scopo di aumentare i pesi delle unità che rispondono al sondaggio in modo da compensare le unità che causano una non risposta. Nel seguito analizzeremo quattro tecniche di aggiustamento.

La prima tecnica consiste nel classificare le unità rispondenti e non in celle di aggiustamento secondo le informazioni della variabile ausiliaria note per tutte le unità del campione. Il peso dei dati mancanti viene calcolato proporzionalmente all'inverso del tasso di risposta presente all'interno della cella.

Assumiamo che le unità rispondenti e le unità non rispondenti possano essere classificate in C celle di aggiustamento basate sulla covariata X. Sia n_{mc} il numero di individui campionati dove, $z_i = 0,1;\ X = c = 1,\ldots,C;\ n_{+c} = n_{0c} + n_{1c}$ identifica il numero di individui campionati presenti all'interno delle celle c; $n_0 = \sum_{c=1}^{C} n_{0c}$ e $n_1 = \sum_{c=1}^{C} n_{1c}$ individuano il numero totale dei non rispondenti e dei rispondenti; infine, $\phi_c = n_{+c} / n$ e $\phi_{1c} = n_{1c} / n_0$ rappresentano le proporzioni degli elementi campionati e dei rispondenti nella cella *c*. Consideriamo, quindi, come stimatore di $\overline{Y}$, la media pesata:

$$\bar{y}_w = \sum_{c=1}^{C} \phi_c \bar{y}_{1c} = \sum_{c=1}^{C} w_c \phi_{1c} \bar{y}_{1c} \qquad (1.91)$$

che pesa i rispondenti nella cella c con l'inverso del tasso di risposta $w_c = \phi_c / \phi_{1c}$. La varianza dello stimatore (1.91) verrà calcolata prendendo in considerazione il seguente modello. Si supponga, condizionatamente alla numerosità campionaria n, che le unità campionate abbiano distribuzione multinomiale sulla tabella di contingenza di grandezza (C x 2) basata sulla classificazione di z_i e X, con le probabilità di cella pari a $\Pr(z_i = 1, X = c) = p\pi_{1c}$; $\Pr(z_i = 0, X = c) = (1-p)\pi_{0c}$ dove $p = \Pr(z_i = 1)$ è la probabilità marginale di risposta. La distribuzione condizionata di X, dato $z_i = 1$ e n_0, è multinomiale con probabilità di cella pari a $\Pr(X = c \mid z_i = 1) = \pi_{1c}$; mentre la distribuzione marginale di X dato n è multinomiale con indice *n* e probabilità di cella pari a $\Pr(X = c) = p\pi_{1c} + (1-p)\pi_{0c} = \pi_c$. Assumiamo che la distribuzione condizionata di Y dato $z_i = z, X = c$ abbia media μ_{zc} e varianza costante σ^2. La media di Y per i rispondenti e dei non rispondenti è rispettivamente:

$$\mu_1 = \sum_{c=1}^{C} \pi_{1c}\mu_{1c} \ , \ \mu_0 = \sum_{c=1}^{C} \pi_{0c}\mu_{0c} \tag{1.92}$$

e la media totale di Y è $\mu = p\mu_1 + (1-p)\mu_0$.

Con questo modello, la media condizionata di Y e la varianza di (1.91) dato ϕ_c sono rispettivamente $\sum_{c=1}^{C} \phi_c\mu_{1c}$ e $\sigma^2 \sum_{c=1}^{C} \phi_c^2/n_{1c}$. Quindi la distorsione di (1.91) è $b(\bar{y}_w) = \sum_{c=1}^{C} \pi_c(\mu_{1c} - \mu_c)$ dove π_c e μ_c sono la proporzione di popolazione e la media di Y nella cella c.

La varianza di (1.91), indicando con $\tilde{\mu}_1 = \sum_{c=1}^{C} \pi_c\mu_{1c}$ la media corretta dei rispondenti, è approssimativamente data da:

$$Var\left(\bar{y}_w\right) = \frac{\left(1+\lambda\right)\sigma^2}{n_1} + \frac{\sum_{c=1}^{C} \pi_c \left(\mu_{1c} - \tilde{\mu}_1\right)^2}{n} \tag{1.93}$$

dove $\lambda = \sum_{c=1}^{C} \pi_{1c}[(\pi_c/\pi_{1c} - 1)^2]$.

Nel caso in cui i risultati della variabile di studio y non sono correlati alle celle di aggiustamento, si ha che le variabili ausiliarie riferite alle celle peggiorano la predizione di y, in quanto si verifica un aumento della varianza dei pesi senza esserci alcuna riduzione della distorsione. Se, invece, la variabile di aggiustamento X non è correlata ai dati mancanti, i pesi tendono a non avere alcun impatto sulla distorsione, ma riducono la varianza nella misura in cui X è un buon predittore dei risultati.

Una seconda tecnica riguardante l'aggiustamento per ponderazione consiste nel moltiplicare per l'inverso delle probabilità di risposta i pesi dei rispondenti campionati. Siccome la vera probabilità di risposta è solitamente sconosciuta, può essere usata una sua stima per correggere la distorsione dei dati mancanti. Quando siamo in questa situazione, il metodo viene chiamato "NWA diretto". Si dimostra che utilizzare una stima della probabilità di risposta risulta essere più efficiente che impiegare la vera probabilità di risposta quando i parametri vengono stimati con il metodo della massima verosimiglianza.

Sia $p_{i\,|\,s} = Pr(z_i = 1 \mid i\epsilon s)$ la probabilità di risposta per le unità campionate. Se si conosce tale probabilità, la media della popolazione può essere stimata senza distorsione nel seguente modo:

$$\bar{y}_d = \frac{1}{N}\sum_{i\in s} \frac{1}{\pi_i}\frac{z_i}{p_{i\,|\,s}} y_i \tag{1.94}$$

Quando $p_{i\,|\,s}$ non è disponibile, si utilizza una stima $\hat{p}_{i\,|\,s}$ ottenuta dal modello specificato per la probabilità di risposta. Sia

$$\bar{y}_e = \frac{1}{N}\sum_{i\in s}\frac{1}{\pi_i}\frac{z_i}{\hat{p}_{i\,|\,s}}y_i \tag{1.95}$$

lo stimatore NWA diretto che utilizza una stima della probabilità di risposta e non usufruisce di variabili ausiliarie. La probabilità di risposta è modellata parametricamente come $p_{i\,|\,s} = p(v_i;\alpha_s^0)$, dove $p(v_i;\cdot)$ è una funzione continua con parametro $\alpha = \alpha_s^0$ e v_i è un vettore di variabili ausiliarie osservate sia per i rispondenti che per i non rispondenti . Il valore α_s^0 viene stimato da $\hat{\alpha}$, la quale è l'unica soluzione dell'equazione $\frac{\partial}{\partial\alpha}\sum_{i\in s}k_i\{z_i ln(p_{i\,|\,s}) + (1-z_i)ln(1-p_{i\,|\,s})\} = 0$ dove k_i è il peso delle unità campionate. Quando $k_i = 1$, la soluzione dell'equazione è la stima di massima verosimiglianza per α_s^0.

Assumendo che la probabilità di risposta non dipende dalle osservazioni campionate, $p_{i\,|\,s} = p_i$. La varianza stimata di (1.95) può essere divisa in due componenti: $Var(\bar{y}_e) = \widehat{Var}_{e1} + \widehat{Var}_{e2}$. La prima componente, supposto $s_z = \{i \in s: z_i = 1\}$, viene rappresentata dall'espressione

$$\widehat{Var}_{e1} = \sum_{i\epsilon s}\sum_{j\epsilon s}\Omega_{ij}\hat{\eta}_i\hat{\eta}_j \tag{1.96}$$

dove

$$\hat{\eta}_i = k_i\pi_i\hat{p}_i\hat{h}_i'\hat{\alpha}_n + \frac{z_i}{\hat{p}_i}\left(y_i - k_i\pi_i\hat{p}_i\hat{h}_i'\hat{\alpha}_n\right)$$

$$\hat{\alpha}_n = \left\{\sum_{i\in s_z}k_i(1-\hat{p}_i)\hat{h}_i\hat{h}_i'\right\}^{-1}\sum_{i\in s_z}\pi^{-1}(\hat{p}_i^{-1}-1)\hat{h}_i y_i$$

e $\hat{h}_i = \partial/\partial\hat{\alpha}\{logit(p(v_i;\hat{\alpha}))\}$. La seconda componente è data da:

$$\widehat{Var}_{e2} = \frac{1}{N^2}\sum_{i\epsilon s_z}\pi_i^{-1}\hat{p}_i^{-2}(1-\hat{p}_i)\left(y_i - k_i\pi_i\hat{p}_i\hat{h}_i'\hat{\alpha}_N\right)^2 \tag{1.97}$$

Finora, abbiamo ipotizzato la probabilità di riposta essere in funzione alla variabile ausiliaria v_i. Supponiamo ora che, oltre a v_i, vi è un'altra variabile x_i correlata a y. Se la variabile x viene osservata in tutto il campione, possiamo definire uno stimatore per regressione NWA, costruito come:

$$\bar{y}_{re} = \bar{y}_e + \left(\bar{x}_n - \bar{x}_e\right)' \hat{\beta}_e \tag{1.98}$$

dove

$$\bar{x}_n = N^{-1} \sum_{i \in s} \pi_i^{-1} x_i$$

$$\hat{\beta}_e = \left(\sum_{i \in s} \pi_i^{-1} \hat{p}_i^{-1} z_i x_i x_i' \right)^{-1} \sum_{i \in s} \pi_i^{-1} \hat{p}_i^{-1} z_i x_i y_i$$

$$\bar{x}_e = N^{-1} \sum_{i \in s} \pi_i^{-1} \hat{p}_i^{-1} z_i x_i$$

Lo stimatore proposto risulta più efficiente dello stimatore diretto NWA se la variabile di studio y è approssimata da una combinazione lineare di x_i. La varianza stimata di (1.98) viene rappresentata come segue:

$$E\left(\sum_{i \in s} \Omega_{ii} y_i^2 + \sum_{i \neq j} \sum_{i,j \in s} \Omega_{ij} y_i y_j \right) + N^{-2} E\left(\sum_{i \in s} \frac{1}{\pi_i^2} \frac{1-p_i}{p_i} \left(y_i - x_i' \beta_N - k_i \pi_i p_i h_{i0}' \alpha_N \right)^2 \right) \tag{1.99}$$

dove

$$\alpha_N = \left\{ \sum_i^N k_i \pi_i p_i (1-p_i) h_{i0} h_{i0}' \right\}^{-1} \sum_i^N (1-p_i) h_{i0} (y_i - x_i' \beta_N)$$

$$\beta_N = \left(\sum_{i=1}^N x_i x_i' \right)^{-1} \sum_{i=1}^N x_i y_i$$

e h_{i0} è il valore assunto da h_i calcolato in $\alpha = \alpha_s^0$.

Lo stimatore NWA diretto è meno interessante ai fini pratici in quanto la somma dei pesi può essere diversa da 1. Per risolvere tale problema, Hájek propose uno stimatore alternativo a (1.95) definito come:

$$\bar{y}_{e2} = \frac{\sum_{i \in s} \pi_i^{-1} \hat{p}_i^{-1} z_i y_i}{\sum_{i \in s} \pi_i^{-1} \hat{p}_i^{-1} z_i} \tag{1.100}$$

La varianza asintotica di (1.100) può essere derivata con la metodologia usata per lo stimatore per regressione NWA.

La terza tecnica da noi analizzata consiste nello stimare la probabilità di risposta con metodi non parametrici. Questi metodi richiedono che le probabilità di risposta siano collegate alle variabili ausiliarie da una funzione regolare, ma non specificata. Nel seguito prenderemo in considerazione la tecnica di lisciamento basata su funzioni kernel.

Consideriamo, quindi, due tipi di stimatori della media $\bar{Y}$ nel caso in cui non si conoscano le vere probabilità di risposta:

$$\bar{y}_{\pi\hat{p}} = \frac{1}{N}\sum_{i\in s_r} w_i \hat{p}_i^{-1} y_i \tag{1.101}$$

e una versione aggiustata per rapporto

$$\bar{y}_{rat,\hat{p}} = \frac{\sum_{i\in s_r} w_i \hat{p}_i^{-1} y_i}{\sum_{i\in s_r} w_i \hat{p}_i^{-1}} \tag{1.102}$$

dove $s_r = \{i \epsilon s : z_i = 1\}$ e $w_i = \pi_i^{-1}$.

Prenderemo, ora, in considerazione la stima di p_i utilizzando la regressione kernel. Questo metodo presenta il vantaggio di differenziare le probabilità di risposta per tutte le osservazioni, e non richiede la specificazione della funzione p(·). Per il processo di stima, ci sono diverse scelte della funzione kernel; in particolare studiamo la funzione definita come segue:

$$\hat{p}_i = \frac{\sum_{j\in s} w_j K_h\left(x_j - x_i\right) z_j}{\sum_{j\in s} w_j K_h\left(x_j - x_i\right)} \tag{1.103}$$

dove $K_h(\cdot) = h^{-1}K(\cdot/h)$, con K(·) funzione Kernel continua e positiva e h parametro di lisciamento.

È possibile ottenere una stima della varianza dello stimatore (1.101) solamente se si utilizzano le seguenti ipotesi. Partiamo dal presupposto che la popolazione U possa essere incorporata in una successione crescente di popolazioni finite $\{U_v\}_{v=1}^{\infty}$ dove la v-esima popolazione ha dimensione N_v. Definiamo $y_v = \left(y_1, \ldots, y_{N_v}\right)'$, vettore di osservazioni della caratteristica di interesse e $x_v = \left(x_1, \ldots, x_{N_v}\right)'$, il corrispondente vettore per la variabile ausiliaria. Per ogni v, viene selezionato un campione s_v di numerosità n_v. Si consideri, quindi, una procedura di replicazione che produca L_v repliche, per ogni elemento della successione di popolazioni, delle seguente quantità:

$$\hat{m}_{iv}^{(l)} \equiv \left(\hat{m}_{1iv}^{(l)}, \hat{m}_{2iv}^{(l)}\right)' = \frac{1}{N_v h_v} \sum_{j \in s_v} w_j^{(l)} K\left(\frac{x_j - x_i}{h_v}\right)(z_j, 1)' \qquad (1.104)$$

$$\hat{m}_{iv}^{*(l)} = \left(\hat{m}_{1iv}^{*(l)}, \hat{m}_{2iv}^{(l)}\right)'$$

dove $\hat{m}_{1iv}^{*(l)} = max\left\{\hat{m}_{1iv}^{(l)}, (N_v h_v)^{-1}\delta\right\}$ e δ è una costante fissata. La varianza stimata di replicazione è definita come:

$$\widehat{Var}\left(\bar{y}_{\pi\hat{p}v}\right) = \sum_{l=1}^{L} c_{lv}\left(\bar{y}_{\pi\hat{p}}^{(l)} - \bar{y}_{\pi\hat{p}}\right)^2 \qquad (1.105)$$

dove

$$\bar{y}_{\pi\hat{p}}^{(l)} = \frac{1}{N} = \sum_{i \in s_v} w_i^{(l)} y_i z_i \frac{\hat{m}_{2iv}^{(l)}}{\hat{m}_{1iv}^{*(l)}}$$

e $\hat{m}_{1iv}^{*(l)}$ *e* $\hat{m}_{2iv}^{(l)}$ sono definiti in (1.104).

Utilizzando le ipotesi precedentemente fatte sulla popolazione U, possiamo definire un ultimo metodo di aggiustamento per stimare la probabilità di risposta $p(x_i)$ basato sulla regressione polinomiale locale; questo metodo, comparato al lisciamento della funzione kernel, migliora l'approssimazione locale.

Sia $K(\cdot)$ una funzione kernel continua e positiva e h_v il suo parametro di lisciamento. Definiamo la matrice basata sul campione di dimensioni $n_v \times (k+1)$

$$\boldsymbol{X}_{s_i} = \begin{pmatrix} 1 & (x_1 - x_i) & \dots & (x_1 - x_i)^k \\ \vdots & \vdots & & \vdots \\ 1 & (x_{n_v} - x_i) & \dots & (x_n - x_i)^k \end{pmatrix},$$

la matrice di dimensioni $n_v \times n_v$

$$\boldsymbol{W}_{s_i} = diag\left\{\frac{1}{\pi_j h_v} K\left(\frac{x_j - x_i}{h_v}\right) : j \in s_v\right\},$$

e il vettore di variabili indicatrici di risposta $\boldsymbol{Z}_s = \left(Z_j : j \in s_v\right)'$. A questo punto, un possibile stimatore basato sulla regressione polinomiale locale di grado k di $p_i = p(x_i)$ è dato da

$$\hat{p}_i^0 = e_1' \hat{\boldsymbol{T}}_{si}^{-1} \hat{\boldsymbol{t}}_{si} \qquad (1.106)$$

dove e_j denota la j-esima colonna della matrice identità di ordine k+1 e

$$\left(\hat{\boldsymbol{T}}_{si}, \hat{\boldsymbol{t}}_{si}\right) = \left(\boldsymbol{X}'_{s_i} \boldsymbol{W}_{s_i} \boldsymbol{X}_{s_i}, \boldsymbol{X}'_{s_i} \boldsymbol{W}_{s_i} \boldsymbol{R}_s\right)$$

assumendo che $\hat{\boldsymbol{T}}_{si}$ sia invertibile. In particolare, quando $\hat{\boldsymbol{T}}_{si}$ è singolare, si può definire una procedura per assicurare che $\hat{p}_i^0$ sia ben definito, garantendo che la scelta della larghezza di banda sia sufficiente a contenere almeno *k+1* valori di z_j nell'intervallo $[x_i - h_v,\ x_i + h_v]$, per ogni $i \in s_v$. Se questa finestra non contiene

abbastanza indicatori risposta, si adotta un approccio che definisce lo stimatore per regressione polinomiale locale, basato sul campione, di grado k di $p_i = p(x_i)$

$$\hat{p}\left(x_i\ ,k\ ,\mathrm{h}_v\right)=\boldsymbol{e}_1'\left(\hat{\boldsymbol{T}}_{si}+diag\left\{\frac{\delta_1}{N_v}\right\}\right)^{-1}\hat{\boldsymbol{t}}_{si}\ ,\ i\in s_v \tag{1.107}$$

dove δ_1 è una piccola costante positiva. I termini di ordine δ_1/N_v aggiunti alla diagonale principale di $\widehat{\boldsymbol{T}}_{si}$ sono sufficienti a rendere la matrice risultante invertibile per ogni h_v. Tuttavia, anche utilizzando (1.107), non si esclude la possibilità che lo stimatore assuma valori vicini a zero; per ovviare a questo problema consideriamo lo stimatore

$$\hat{p}_i=\max\left\{\hat{p}\left(x_i\ ,k\ ,\mathrm{h}_v\right)\ ,\ \delta_2\left(N_v h_v\right)^{-1}\right\} \tag{1.108}$$

con δ_2 costante maggiore di zero.

Il calcolo della stima della varianza per gli stimatori (1.101) e (1.102) viene eseguito con metodi analoghi a quelli utilizzati nella tecnica di lisciamento basata su funzione kernel.

1.4 Nota bibliografica

In questo capitolo, si sono analizzati vari metodi per stimare la media di una variabile in presenza di dati mancanti non utilizzando metodi di imputazione. Per quanto riguarda i dati mancanti di tipo MCAR, in letteratura si può fare riferimento ad un gruppo di articoli che trattano stimatori che utilizzano tutta, o parte dell'informazione disponibile: l'articolo di Rueda, Gonzàlez e Arcos (2006); l'articolo di Rueda, Muñoz, Berger, Arcos e Martínez (2007) se lo stimatore è basato sulla pseudo-verosimiglianza empirica; la pubblicazione di Gonzàlez, Rueda e Arcos (2008) quando si hanno a disposizione due variabili quantitative ausiliarie. Sempre nel contesto di dati mancanti MCAR, si sono presentati stimatori costruiti a partire da un campionamento a due fasi, trattati nelle pubblicazioni: Chen e Rao (2007); Singh e Kumar (2008); Singh, Kumar e Kozak (2010).

Per quanto riguarda i dati di tipo MAR, in letteratura si possono consultare due articoli che fanno riferimento agli stimatori doppiamente robusti: Cao, Tsiatis e Davidian (2009) e Tan (2010). Se si segue, invece, un approccio semiparametrico alla stima, si può considerare l'articolo di Hu, Follmann e Qin (2010). Sempre nel contesto di dati mancanti di tipo MAR, si sono presentati stimatori che utilizzano metodi di aggiustamento per ponderazione: se la probabilità di risposta viene stimata in modo parametrico si consideri l'articolo di Kim e Kim (2007), se si utilizza un metodo di stima della probabilità di risposta non parametrico, un utile approfondimento è reperibile nelle pubblicazioni di Silva e Opsomer (2006) e (2009). Infine, il metodo di aggiustamento per ponderazione che utilizza 'celle di aggiustamento' è reperibile nell'articolo di Little e Vartivarian (2006).

Capitolo 2

Stimatori della media utilizzando metodi di imputazione

2.1 Introduzione

Nel seguente capitolo, si presentano alcune metodologie di stima della media di una variabile di studio y utilizzando tecniche di imputazione. Per quanto riguarda le mancate risposte parziali la procedura di compensazione comunemente usata è l'imputazione, che consiste nell'assegnazione di un valore sostitutivo del dato mancante, al fine di ripristinare la "completezza" dei dati. Alcuni metodi per il trattamento dei dati mancanti sono molto semplici e, possono essere utilizzati se la proporzione dei dati mancanti è molto ridotta, altri metodi sono piuttosto complessi e richiedono competenze specifiche sul problema. Numerosi sono i metodi di imputazione proposti in letteratura per predire valori sostitutivi per le mancate risposte parziali. In questo capitolo i metodi sono stati suddivisi a seconda che il meccanismo di generazione dei dati mancanti sia di tipo MCAR o MAR. Il paragrafo 2.2 è dedicato al caso in cui i dati mancanti siano del tipo MCAR: si contraddistinguono due filoni principali di procedure di imputazione, il primo riguardante l'imputazione per media, rapporto, differenza e regressione; il secondo riguardante l'imputazione nel campionamento stratificato e per cluster. Nel paragrafo 2.3 sono esposte, invece, quattro classi di metodi di imputazione, ipotizzando un meccanismo MAR. La prima classe relativa all'imputazione 'nearest neighbor'; la seconda, relativa all'imputazione tramite pseudo-verosimiglianza empirica; la terza,

riguardante metodi di imputazione doppiamente robusti e l'ultima classe relativa all'imputazione ponderata.

2.2 Dati mancanti MCAR

Ricordiamo che, in questo caso, la probabilità di risposta per ogni unità è indipendente oltre che dal valore della variabile di studio anche dal valore delle variabili ausiliarie.

2.2.1 Imputazione per media, rapporto, differenza e regressione

Da alcuni recenti studi in letteratura sul problema riguardante la presenza di non risposte in indagini campionarie, emergono stimatori efficienti per la media della popolazione che utilizzano tecniche di imputazione dei dati mancanti per media e rapporto.

È noto che adoperando l'imputazione per media, la media campionaria e la sua varianza possono essere definite rispettivamente come:

$$\bar{y}_s = \frac{1}{r}\sum_{i=1}^{r} y_i \tag{2.1}$$

$$Var\left(\bar{y}_s\right) = \left(\frac{1}{r} - \frac{1}{N}\right) S_y^2 \tag{2.2}$$

Dove, r è il numero di unità rispondenti tra le n unità campionate con meccanismo SRSWOR. Si osservi che la media campionaria è uno stimatore non distorto per $\bar{Y}$ in modo che $Var(\bar{y}_s) = MSE(\bar{y}_s)$.

Nel caso in cui utilizzassimo l'imputazione per rapporto, potremmo definire lo stimatore rapporto di $\bar{Y}$ e l'errore quadratico medio (MSE) rispettivamente come:

$$\bar{y}_r = \frac{\bar{y}_s}{\bar{x}_s}\bar{x} \tag{2.3}$$

$$MSE\left(\bar{y}_r\right) \cong \left(\frac{1}{r} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{r} - \frac{1}{N}\right)\left(R^2 S_x^2 - 2RS_{xy}\right) \tag{2.4}$$

dove $\bar{x} = \sum_{i=1}^{n} x_i / n$ è la media campionaria della variabile ausiliaria (di cui conosciamo tutte le osservazioni) e $\bar{x}_s = \sum_{i=1}^{r} x_i / r$ è la media campionaria delle r unità della variabile ausiliaria.

Confrontando (2.2) con (2.4), si può facilmente vedere che l'imputazione per rapporto è più efficiente dell'imputazione per media quando:

$$R < 2\beta, \quad \text{per R} > 0 \qquad (2.5)$$
$$R > 2\beta, \quad \text{per R} < 0$$

Enunceremo ora una serie di stimatori di $\bar{Y}$, con il relativo errore quadratico medio e confrontandoli tra loro. Si parta col definire uno stimatore per un campione completo:

$$\bar{y}_{KC} = \frac{\bar{y} + \hat{\beta}\left(\bar{X} - \bar{x}\right)}{\bar{x}} \bar{X} \qquad (2.6)$$

A questo punto, se si vuole tenere conto dei metodi di imputazione precedentemente osservati, è possibile modificare lo stimatore (2.6) come segue:

$$\bar{y}_{pr1} = \frac{\bar{y}_s + \hat{\beta}\left(\bar{X} - \bar{x}\right)}{\bar{x}} \bar{X} \qquad (2.7)$$

con MSE pari a

$$MSE\left(\bar{y}_{pr1}\right) \cong \left(\frac{1}{r} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{N}\right) S_x^2 (R^2 - \beta^2)$$

Un secondo stimatore della media è il seguente:

$$\bar{y}_{pr2} = \frac{\bar{y}_s + \hat{\beta}\left(\bar{X} - \bar{x}_s\right)}{\bar{x}_s} \bar{X} \qquad (2.8)$$

con MSE pari a

$$MSE\left(\bar{y}_{pr2}\right) \cong \left(\frac{1}{r} - \frac{1}{N}\right) \left(S_y^2 - \beta S_{xy} + R^2 S_x^2\right)$$

Se la media di popolazione della variabile ausiliaria, $\bar{X}$, è sconosciuta, si può proporre un terzo stimatore definito come:

$$\bar{y}_{pr3} = \frac{\bar{y}_s + \hat{\beta}\left(\bar{X} - \bar{x}_s\right)}{\bar{x}_s} \bar{x} \qquad (2.9)$$

con MSE pari a

$$MSE\left(\bar{y}_{pr3}\right) \cong \left(\frac{1}{r} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{r} - \frac{1}{n}\right) \left[(R + \beta)^2 S_x^2 - 2(R + \beta) S_{xy}\right]$$

Confrontando l'errore quadratico medio degli stimatori fin qui proposti possiamo affermare che quelli definiti in (2.7), (2.8) e (2.9) sono più efficienti dello stimatore che utilizza l'imputazione per media se vale la relazione

$$R^2 < \beta^2$$

Siamo in grado di definire ulteriori condizioni di efficienza comparando i vari stimatori; per esempio, (2.7) è più efficiente di (2.3) quando

$$\left(\frac{1}{n}-\frac{1}{N}\right)S_x^2(R^2-\beta^2)-\left(\frac{1}{r}-\frac{1}{n}\right)\left(R^2S_x^2-2RS_{xy}\right)<0,$$

mentre lo stimatore (2.8) è più efficiente dello stimatore (2.3) se

$$\left(\frac{1}{r}-\frac{1}{N}\right)S_x^2(R^2-\beta^2)-\left(\frac{1}{r}-\frac{1}{n}\right)\left(R^2S_x^2-2RS_{xy}\right)<0$$

Infine, il terzo stimatore (2.9) è più efficiente dello stimatore (2.3) se

$$S_{xy}(2R-\beta)<0$$

In questa sezione, verrà esaminato più nello specifico il metodo di imputazione per rapporto.

Supponiamo esistano r osservazioni complete $(y_1,x_1),(y_2,x_2),\dots,(y_r,x_r)$ e $(n-r)$ osservazioni incomplete $x_1^*,x_2^*,\dots,x_{n-r}^*$. Il campione verrà quindi diviso in due gruppi, uno avente dimensione r chiamato s_1 e l'altro di dimensione $(n-r)$ chiamato s_2.

Nel momento in cui le osservazioni incomplete non vengono scartate, il data set completo viene definito da

$$h_i=\begin{cases}y_i & \text{se i}\in s_1\\ \tilde{y}_i & \text{se i}\in s_2\end{cases} \tag{2.10}$$

e la media della popolazione viene stimata da

$$t=\frac{1}{n}\sum_{i=1}^{n}h_i=\frac{1}{n}\left(\sum_{i\in s_1}y_i+\sum_{i\in s_2}\tilde{y}_i\right) \tag{2.11}$$

dove $\tilde{y}_i$ indica il valore imputato della variabile di studio corrispondente all'osservazione x_i^*. Se utilizziamo l'imputazione per rapporto, esistono due scelte di $\tilde{y}_i$:

$$\tilde{y}_i=\bar{y}\left(\frac{\bar{X}}{\bar{x}}\right) \tag{2.12}$$

$$\tilde{y}_i=\bar{y}\left(\frac{n\bar{X}}{r\bar{x}+(n-r)\bar{x}^*}\right) \tag{2.13}$$

dove $\bar{x}=\sum_{i=1}^{r}x_i/r$ e $\bar{x}^*=\sum_{i=1}^{n-r}x_i^*/(n-r)$. Se non si conosce il vero valore di $\bar{X}$, i valori imputato possono essere definiti come:

$$\tilde{y}_i=\bar{y}\left(\frac{x_i^*}{\bar{x}}\right) \tag{2.14}$$

Sulla stessa linea, si propone un altro insieme di valori imputati

$$\tilde{y}_i = \bar{y}\left(\frac{nx_i^*}{(n-r)\bar{x} + r\bar{x}^*} \right) \tag{2.15}$$

Utilizzando le espressioni (2.12) – (2.15) in (2.11), si derivano i seguenti quattro stimatori di $\bar{Y}$:

$$t_1 = \bar{y}\left[\frac{r\bar{x} + (n-r)\bar{X}}{n\bar{x}} \right], \tag{2.16}$$

$$t_2 = \bar{y}\left[\frac{r^2\bar{x} + n(n-r)\bar{X} + (n-r)r\bar{x}^*}{rn\bar{x} + n(n-r)\bar{x}^*} \right], \tag{2.17}$$

$$t_3 = \bar{y}\left[\frac{r\bar{x} + (n-r)\bar{x}^*}{n\bar{x}} \right], \tag{2.18}$$

$$t_4 = \bar{y}\left[\frac{r^2\bar{x} + (n+r)(n-r)\bar{x}^*}{rn\bar{x} + n(n-r)\bar{x}^*} \right]. \tag{2.19}$$

Per analizzare la distorsione dei quattro stimatori proposti e per confrontare la loro efficienza dobbiamo innanzitutto definire tre quantità:

$$\varphi = \frac{\bar{Y}S_x}{\bar{X}S_y}, \qquad f_k = E_r\left(\frac{1}{r}\right)\left(\frac{n-r}{n}\right)^k, \qquad g_k = \frac{1}{n}E_r\left(\frac{n-r}{n}\right)^k,$$

dove E_r indica il valore atteso rispetto alla variabile casuale r (considerando valori positivi) e k è un intero fissato maggiore di zero. Si ipotizza, inoltre, che il coefficiente di correlazione ρ sia positivo, come si richiede per l'applicazione del metodo per rapporto. È semplice ora dimostrare che gli stimatori (2.16) – (2.19) sono generalmente distorti tranne nel caso in cui $\rho = \varphi$.

Nel seguito, confronteremo l'MSE dei quattro stimatori. Quando $\bar{X}$ è noto, abbiamo a disposizione i due stimatori t_1 e t_2, di cui il primo ignora le coppie incomplete di osservazioni, mentre il secondo le incorpora nella formulazione dello stimatore. Ora è facile vedere che (2.17) non ha solo minor distorsione, ma anche un più piccolo errore quadratico medio rispetto a (2.16) purché

$$2\rho < \left(\frac{f_2 - g_2}{f_1 - g_1}\right)\varphi$$

Quando, invece, $\bar{X}$ non è noto, abbiamo a disposizione i due stimatori t_3 e t_4. Entrambi si basano sulle intere osservazioni disponibili ma t_3 possiede un minor errore quadratico medio se

$$2\rho > \left(\frac{f_1 - g_1 + g_2}{f_1 - g_1}\right)\varphi$$

Ricordando che il requisito per l'utilizzo della tecnica d'imputazione per rapporto è $2\rho > \varphi$ e osservando che la quantità $(f_1 - g_1 + g_2)/(f_1 - g_1)$ non può essere inferiore a uno, si verifica che (2.19) ha prestazioni migliori di (2.18) quando

$$\varphi < 2\rho < \left(\frac{f_1 - g_1 + g_2}{f_1 - g_1}\right)\varphi$$

Infine, si può constatare che (2.18) e (2.19) si derivano facilmente da (2.16) e (2.17) sostituendo $\bar{x}^*$a $\bar{X}$, rispettivamente. Quindi, confrontando t_1 con t_3, si osserva che entrambi gli stimatori hanno stessa distorsione ma il primo presenta un minore MSE. Allo stesso modo, se si confronta t_2 con t_3, entrambi gli stimatori presentano medesima distorsione, ma t_2 presenta un minore MSE, a condizione che $g_1 > 2g_2$.

Un ulteriore metodo proposto in letteratura per l'imputazione dei dati mancanti si basa sull'utilizzo di stimatori indiretti che fanno uso delle informazioni disponibili di osservazioni incomplete. Questo metodo è in grado di migliorare la precisione degli stimatori.

Partendo da una situazione analoga presentata nella parte iniziale del primo capitolo, si assuma di possedere un insieme *(n – p – q)* completo di osservazioni selezionate all'interno del campione e di avere a disposizione *p* osservazioni della caratteristica ausiliaria *x* ma non le corrispondenti della caratteristica di studio *y*. Allo stesso modo, abbiamo una serie di *q* osservazioni della caratteristica *y* nel campione, ma i valori associati della caratteristica *x* sono mancanti. Inoltre *p* e *q* sono numeri interi che verificano le seguenti condizioni: $p > 0$ e $q < n/2$.

Per comodità, separiamo le unità del campione *s* in tre insiemi disgiunti:

$s_1 = \{i \in s \,/\, x_i, y_i$ sono disponibili$\}$,
$s_2 = \{i \in s \,/\, x_i$ sono disponibili, y_i non sono disponibili$\}$,
$s_3 = \{i \in s \,/\, y_i$ sono disponibili, x_i non sono disponibili$\}$.

s_1			s_2			s_3		
y_1	...	y_{n-p-q}	Missing	...	Missing	y_{n-q+1}	...	y_n
x_1	...	x_{n-p-q}	$x_{n-p-q+1}$	...	x_{n-q}	Missing	...	Missing

Quando viene applicato un metodo di imputazione, l'insieme completo dei dati viene definito come:

$$h_i = \begin{cases} y_i & \text{se i} \in \text{s}_1 \cup s_3 \\ \tilde{y}_i & \text{se i} \in \text{s}_2 \end{cases}$$

dove $\tilde{y}_i$ sono i valori imputati. Lo stimatore di $\bar{Y}$ che utilizza questo data-set è:

$$\bar{y}_{imp} = \frac{1}{N}\sum_{i\in s}\frac{h_i}{\pi_i} \tag{2.20}$$

I metodi di imputazione comunemente usati includono l'imputazione per media, come precedentemente visionato. Impiegando metodi di stima indiretti, è possibile utilizzare gli stimatori tradizionali per rapporto, alle differenze e per regressione della media. Tuttavia, se una grande proporzione di dati è mancante, gli stimatori usuali saranno basati su un campione relativamente piccolo e la loro precisione verrà ridotta di conseguenza.

Gli stimatori di Horvitz-Thompson basati sugli insiemi s_1, s_2 e s_3 sono: (2.21)

$$\bar{y}^1_{HT} = \frac{1}{N}\sum_{i\in s_1}\frac{y_i}{\pi_i},\ \bar{y}^3_{HT} = \frac{1}{N}\sum_{i\in s_3}\frac{y_i}{\pi_i},\ \bar{\text{x}}^1_{HT} = \frac{1}{N}\sum_{i\in s_1}\frac{x_i}{\pi_i},\ \bar{\text{x}}^2_{HT} = \frac{1}{N}\sum_{i\in s_2}\frac{x_i}{\pi_i}$$

Di seguito, si propongono tre classi di stimatori, rispettivamente, per rapporto, alle differenze e per regressione, che considerano le informazioni disponibili provenienti da osservazioni incomplete.

$$y_{r2} = \frac{\alpha_r \bar{y}^3_{HT} + (1-\alpha_r)\bar{y}^1_{HT}}{\psi_r \bar{x}^2_{HT} + \psi_r \bar{x}^1_{HT}}\bar{X}\ , \tag{2.22}$$

$$\bar{\text{y}}_{d2} = \alpha_d \bar{y}^1_{HT} + (1-\alpha_d)\bar{y}^3_{HT} + \left[\bar{X} - \left(\psi_d \bar{x}^1_{HT} + (1-\psi_d)\bar{x}^2_{HT}\right)\right]\ , \tag{2.23}$$

$$\bar{\text{y}}_{\text{Re}g2} = \alpha_{reg}\bar{y}^1_{HT} + (1-\alpha_{reg})\bar{y}^3_{HT} + \beta\left[\bar{X} - \left(\psi_{reg}\bar{x}^1_{HT} + (1-\psi_{reg})\bar{x}^2_{HT}\right)\right]\ . \tag{2.24}$$

Nel caso in cui nello stimatore per regressione non si conosce β, è possibile stimarlo in due modi:

$$\hat{\beta}_1 = \frac{Cov_{i\in s_1}(x,y)}{Var_{i\in s_1}} \quad \text{e} \quad \hat{\beta}_2 = \frac{Cov_{i\in s_1}(x,y)}{Var_{i\in s_1 \cup s_2}(x)}$$

generando due diversi stimatori a seconda della stima di β utilizzata chiamati $\bar{y}_{Reg21}$ e $\bar{y}_{Reg22}$.

Per ottenere gli stimatori con minimo MSE, si devono conoscere i coefficienti ottimali $\alpha_{r_{opt}}$, $\psi_{r_{opt}}$, $\alpha_{d_{opt}}$, $\psi_{d_{opt}}$, $\alpha_{reg_{opt}}$, $\psi_{reg_{opt}}$. Sfortunatamente, i valori ottimali dipendono dalle varianze e covarianze teoriche tra gli stimatori di Horvitz-

Thompson, che risultano generalmente sconosciuti. Tuttavia, i valori possono essere stimati mediante metodi replicativi ottenendo valori approssimati, $\tilde{\alpha}_r$, $\tilde{\psi}_r$, $\tilde{\alpha}_d$, $\tilde{\psi}_d$, $\tilde{\alpha}_{reg}$ e $\tilde{\psi}_{reg}$, i quali ci permettono di enunciare gli stimatori corrispondenti $\tilde{y}_{r2}$, $\tilde{y}_{d2}$ e $\tilde{y}_{Reg2}$.

A questo punto, analizziamo i seguenti metodi di imputazione:

- procedura basata su uno stimatore per rapporto: in questa situazione, si specifica il data set completo usando lo stimatore (2.22) per il valore imputato;
- procedura basata su uno stimatore alle differenze: in questa circostanza, si specifica il data set completo facendo uso di (2.23) per il valore imputato;
- procedura basata su uno stimatore per regressione con β ignoto: in questo caso, si propongono due stimatori per regressione usando due possibili stime del coefficiente di regressione. Quindi deriviamo due procedure di imputazione, facendo uso di $\bar{y}_{Reg21}$ e $\bar{y}_{Reg22}$ per i valori imputati.

2.2.2 Utilizzo di imputazione multipla nel campionamento stratificato e metodi di imputazione nel campionamento per cluster.

Il più popolare approccio in letteratura per affrontare il problema dell'impatto della mancata risposta sulla precisione delle stime nelle indagini statistiche è probabilmente quello dell'imputazione multipla, introdotta da Rubin (1978, 1987). Esso consiste essenzialmente nella ripetizione del processo di imputazione η volte e, conseguentemente, nella generazione di un insieme di η data-set completi.

Nel contesto del campionamento stratificato, consideriamo una popolazione finita di N unità, suddivisa in H strati di numerosità N_h, $h = 1, \dots, H$. Assumiamo, all'interno di ogni strato un meccanismo di non risposta di tipo MCAR, ossia, gli indicatori di riposta nello strato h, $z_{h,1}, \dots, z_{h,N_h}$, sono indipendenti con $p_{hz} = P(z_{h,i} = 1)$.

Sia $y_{h,obs} = (y_i : i \in s_{hz})$ la parte osservata dal campione di rispondenti s_{hz} di numerosità n_{hz} dallo strato h. Proprio da $y_{h,obs}$, viene prelevato casualmente un valore imputato y_i^*. Lo stimatore che fa uso dei dati campionati completi è l'usuale media pesata delle medie di strato:

$$\bar{Y}_{strat} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h \tag{2.25}$$

dove $\bar{y}_h = \sum_{s_h} y_i / n_h$, s_h è il campione di strato h di numerosità n_h. La varianza dello stimatore appena esposto è definita come

$$Var\left(\bar{Y}_{strat}\right) = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2 S_h^2\left(\frac{1}{n_h} - \frac{1}{N_h}\right) \tag{2.26}$$

con $S_h^2 = \sum_{i=1}^{N_h}(y_i - \bar{Y}_h)^2 /(N_h - 1)$ la varianza della popolazione nello strato h.

Sia $\bar{y}_{hz}$ la media campionaria dello strato h e $\hat{S}_{hz}^2 = \frac{1}{n_{hz}-1}\sum_{i\in s_{hz}}(y_i - \bar{y}_{hz})^2$ la varianza campionaria; allora, lo stimatore basato sull'imputazione è definito da:

$$\bar{Y}_{strat}^{*} = \sum_{h=1}^{H}\frac{N_h}{N}\bar{y}_h^{*} \tag{2.27}$$

dove

$$\bar{y}_h^{*} = \frac{\left(n_{hz}\bar{y}_{hz} + \sum_{i\in(s_h - s_{hz})} y_i^{*}\right)}{n_h} \tag{2.28}$$

Infine, considerando le ψ repliche dello stimatore (2.27) e indicandole con $\bar{Y}_{strat,i}^{*}$, per $i =, \ldots, \psi$, possiamo definire lo stimatore combinato

$$\bar{\bar{Y}}_{strat}^{*} = \sum_{i=1}^{\psi}\frac{\bar{Y}_{strat,i}^{*}}{\psi} \tag{2.29}$$

La stima della varianza di (2.29) definita dalla formula di combinazione è data da

$$\widehat{Var}\left(\bar{\bar{Y}}_{strat}^{*}\right) = \sum_{h=1}^{H}\bar{V}_h + \sum_{h=1}^{H}\left(\frac{1}{1-f_h} + \frac{1}{\psi}\right)B_h^{*} \tag{2.30}$$

dove $f_h = (n_h - n_{hz})/n_h$ e $\bar{V}_h^{*}$ è la media di ψ valori della varianza stimata basata sull'imputazione $\hat{V}_h^{*} = \left(\frac{N_h}{N}\right)^2 \hat{S}_{h*}^2\left(\frac{1}{n_h} - \frac{1}{N_h}\right)$ dove $\hat{S}_{h*}^2 = \frac{1}{n_h - 1}\left(\sum_{i\in s_{hz}}(y_i - \bar{y}_h^{*})^2 + \sum_{i\in(s_h - s_{hz})}(y_i^{*} - \bar{y}_h^{*})^2\right)$. La componente intra-imputazione è definita come $B_h^{*} = \sum_{i=}^{\psi}\left(\bar{y}_{h,i}^{*} - \bar{\bar{y}}_h^{*}\right)^2/(\psi - 1)$ dove $\bar{\bar{y}}_h^{*} = \sum_{i=1}^{\psi}\bar{y}_{h,i}^{*}/\psi$.

Il tasso di risposta $1 - f_h$ presenta una peso elevato se il tasso di non risposta è grande e/o la stima della varianza di (2.27) è elevata.

Si propongono nel seguito metodologie e tecniche per stimare $\bar{Y}$ nel caso in cui ci troviamo di fronte ad un campionamento per cluster assumendo che il meccanismo di non risposta sia del tipo MCAR. Nel campionamento per cluster, il campione si forma in due stadi: le unità campionate del primo stadio sono cluster contenenti unità campionate in seconda fase.

Sia s un campione di cluster di dimensione n. All'interno dell'i-esimo gruppo campionato, sia s_i il campione di seconda fase di dimensione $m_i > 2$. Per unità campionata $j \in s_i$, il peso di campionamento w_{ij} è costruito in modo tale che quando si presentano dati mancanti, $\hat{Y} = \sum_{i\in s}\sum_{j\in s_i} w_{ij}y_{ij}$ sia uno stimatore non distorto di $\bar{Y}$.

Assumiamo che ogni y_{ij}, a livello di popolazione, sia una varabile casuale con

$$y_{ij} = \mu_i + b_i + e_{ij} \tag{2.31}$$

dove μ_i è la media ignota di y, b_i è un effetto casuale non osservato a livello di clusters con media 0 e varianza finita (tutte le unità nello stesso cluster presentano stesso b_i) ed e_{ij} è l'effetto casuale non osservato all'interno dei cluster con media 0 e varianza finita. Si assuma inoltre che i due effetti siano indipendenti tra loro.

Infine, indichiamo con z_{ij} le variabili indicatrici di risposta per y_{ij} ($z_{ij} = 1$ quando y_{ij} è un rispondente e $z_{ij} = 0$ se y_{ij} non lo è) supponendole definite per ogni unità della popolazione. Ipotizziamo, a questo punto, il seguente meccanismo di risposta basato su effetti casuali non ignorabili:

$$P_m(z_i|b_i, y_i) = P_m(z_i|b_i), \quad i\epsilon s \tag{2.32}$$

dove P_m è la probabilità di risposta rispetto al modello (2.31), z_i è il vettore contenente gli z_{ij} e y_i il vettore contenente gli y_{ij}. Eseguendo, per ogni cluster, un'imputazione degli y_{ij} non rispondenti con la media di cluster $\sum_{j\epsilon s_i} z_{ij}w_{ij}y_{ij} / \sum_{j\epsilon s_i} z_{ij}w_{ij}$, è possibile specificare uno stimatore non distorto di $\bar{Y}$:

$$\hat{Y}_c = \sum_{i\in s}\sum_{j\in s_1} z_{ij}\bar{w}_{ij}y_{ij} \tag{2.33}$$

con

$$\bar{w}_{ij} = w_{ij}\left(\sum_{j\in s_1} w_{ij} \Big/ \sum_{j\in s_1} z_{ij}w_{ij}\right) \tag{2.34}$$

Dal momento che l'imputazione avviene all'interno di ogni gruppo, lo stimatore definito da (2.33) sembra inefficiente quando la dimensione di alcuni cluster è piccola. Questa preoccupazione, tuttavia, non sussiste nel caso in cui $w_{ij} = w_i$ per ogni *j* (ad esempio, quando il campione di secondo stadio presenta uguale probabilità di inclusione). In questo caso, l'imputazione che porta all'utilizzo di (2.33) viene calcolata su un insieme molto più grande G_l definito come:

$$G_l = \{i \in s : m_i = m\ ,\ \bar{z}_i = k/m\}, \quad l = (k\ ,\ m),\ k \le m \tag{2.35}$$

dove $\bar{z}_i = m_i^{-1}\sum_{j\in s_1} z_{ij}$ rappresenta il tasso di risposta intra-cluster. Da notare che l'insieme G_l è formato dal gruppo di cluster campionati aventi stessa dimensione e $\bar{z}_i = k$. Sotto queste ultime ipotesi, possiamo definire il valore imputato con la media campionaria dei rispondenti in G_l, come:

$$\tilde{y}_{G_l} = \frac{\sum_{i\in G_l}\sum_{j\in s_1} z_{ij}w_{ij}y_{ij}}{\sum_{i\in G_l}\sum_{j\in s_1} z_{ij}w_{ij}} \tag{2.36}$$

Per il calcolo della varianza stimata di (2.33) possiamo applicare il metodo jacknife corretto solamente quando: la dimensione del campione di prima fase è grande, $m_i \leq m$ per ogni i con m fissato e il rapporto n/N è piccolo.

Il metodo appena proposto risulta applicabile se esiste almeno un rispondente in ogni cluster; in caso contrario lo stimatore (2.33) non può essere calcolato.

Un'estensione del metodo di imputazione appena presentato è possibile nel caso in cui si abbia a disposizione una variabile ausiliaria x con tutti i relativi valori osservati. I risultati precedenti possono essere estesi al metodo di imputazione per regressione modificando il modello (2.31) in

$$y_{ij} = \alpha + \beta x_{ij} + b_i + e_{ij} \tag{2.37}$$

Quando la probabilità di risposta dipende dall'effetto casuale specifico di cluster b_i, il meccanismo generatore di dati mancanti è di tipo MCAR, poiché gli effetti casuali che caratterizzano i cluster non sono osservati. Di conseguenza, la definizione del modello (2.37) porta alla formulazione di stimatori distorti. Per correggere la distorsione, si propone un approccio informale che consiste nel modificare il modello (2.37) permettendo alla media di cluster di dipendere da una stima del tasso di risposta di cluster $\bar{z}_i$. Assumendo, per semplicità, una relazione lineare e definendo la funzione $g_1(x_{ij}) = \alpha + \beta x_{ij}$, si ottiene il modello approssimato:

$$\left[y_{ij} \mid x_{ij},\ b_i,\ \delta,\ \alpha, \beta, e_{ij},\ \sigma^2 \right] = N\left(b_i + \delta \bar{z}_i + g_1\left(x_{ij} \right) + e_{ij}\ ,\ \sigma^2 \right) \tag{2.38}$$

che considera il tasso di risposta $\bar{z}_i$, una variabile ausiliaria aggiuntiva e dove $N(\cdot)$ rappresenta una distribuzione normale. La funzione $g_1(x_{ij})$ determina il grado di dipendenza della media di y dalla covariata x.

Un approccio più rigoroso per correggere la distorsione consiste nel considerare il seguente modello parametrico:

$$\begin{aligned}
&\left[y_{ij} \mid x_{ij},\ b_i,\ \delta,\ \chi_i,\ \alpha, \beta, e_{ij},\ \sigma^2 \right] = N\left(b_i + \delta \chi_i + g_1\left(x_{ij} \right) + e_{ij}\ ,\ \sigma^2 \right) \\
&\left[a_{ij} \mid x_{ij},\ \chi_i,\ \gamma_0,\ \gamma_1, e_{ij} \right] = N\left(\chi_i + g_2\left(x_{ij} \right) + e_{ij}\ ,\ 1 \right) \\
&z_{ij} = \begin{cases} 1 & \text{se } a_{ij} > 0 \\ 0 & \text{se } a_{ij} < 0 \end{cases}
\end{aligned} \tag{2.39}$$

dove z_{ij} è l'indicatore di risposta e a_{ij} è una variabile latente che determina lo stato della risposta del soggetto j nel cluster i (se $a_{ij} > 0$ il soggetto risponde, altrimenti il

soggetto non risponde); gli effetti casuali $b_i \, e \, \chi_i$ modellano la correlazione intra-cluster. La funzione $g_2(x_{ij}) = \gamma_0 + \gamma_1 x_{ij}$ determina in che modo la media della variabile latente a_{ij} e, quindi, la probabilità di risposta, dipende dalla covariata x. Per questo motivo, se vogliamo mantenerci in un contesto di dati mancanti MCAR, la funzione $g_2(x_{ij})$ deve essere costante (e.g., $\gamma_1 = 0$).

Con i modelli definiti in (2.38) e (2.39), possiamo ottenere uno stimatore di $\bar{Y}$ utilizzando il metodo di imputazione multipla. Innanzitutto, formiamo η datasets imputati, completando i valori mancanti di y tramite η estrazioni indipendenti dalle distribuzioni dei modelli (2.38) o (2.39). Per il k-esimo dataset imputato, lo stimatore di $\bar{Y}$ è dato da:

$$\hat{\bar{Y}}_k = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i} \frac{y_{ij}^{(k)}}{\pi_{ij}}}{\sum_{i=1}^{n}\sum_{j=1}^{m_i} \frac{1}{\pi_{ij}}} \tag{2.40}$$

dove $y_{ij}^{(k)}$ è il valore osservato o imputato di y_{ij}. Quindi, uno stimatore consistente di $\bar{Y}$ è definito da:

$$\hat{\bar{Y}} = \frac{1}{k}\sum_{k=1}^{\eta} \hat{\bar{Y}}_k \tag{2.41}$$

avente varianza

$$Var\left(\hat{\bar{Y}}\right) = \frac{1}{k}\sum_{k=1}^{\eta} V_k + \frac{1}{k-1}\sum_{k=1}^{\eta}\left(\hat{\bar{Y}}_k - \hat{\bar{Y}}\right)^2 \tag{2.42}$$

dove V_k è la varianza di $\hat{\bar{Y}}_k$ calcolata per il k-esimo dataset imputato.

I modelli (2.37), (2.38) e (2.39) possono essere applicati anche se la variabile di studio y dipende da un vettore di variabili ausiliarie $\boldsymbol{x} = (x_1, \ldots, x_d)'$.

2.3 Dati mancanti MAR

Ricordiamo che il meccanismo generante dati mancanti del tipo MAR, prevede che la probabilità di risposta delle unità sia indipendente dalla variabile di studio *y*, ma dipendente dalle variabili ausiliarie associate.

2.3.1 Imputazione 'nearest neighbor'

Consideriamo un campione bivariato $(x_1, y_1), \dots, (x_n, y_n)$ strutturato nel seguente modo:

$y_1, y_2, \dots\dots\dots\dots, y_r$	$Missing, \dots\dots\dots, Missing$
$x_1, x_2, \dots\dots\dots\dots, x_r$	$x_{r+1}, x_{r+2}, \dots\dots\dots\dots, x_n$

Il metodo di imputazione nearest neighbor (NNI), attribuisce ad un valore mancante y_j, un valore osservato y_i $(1 \leq i \leq r)$ tale che la distanza della variabile x tra i due indici sia minima; ossia *i* soddisfa la condizione $|x_i - x_j| = \min_{1 \leq l \leq r} |x_l - x_j|$. La tecnica NNI viene spesso utilizzata dividendo, in primo luogo, il campione in diverse 'classi di imputazione' e, infine, trovando i valori da imputare all'interno di queste classi.

Il metodo NNI presenta due caratteristiche interessanti: innanzitutto NNI risulta più efficiente dei comuni metodi di imputazione, come per esempio l'imputazione per media e hot-deck, quando la variabile x fornisce informazioni ausiliarie utili per lo studio di y. Infine, NNI non assumendo un modello di regressione parametrico tra y ed x, è più robusto dei metodi di imputazione per rapporto e per regressione basati su un modello di regressione lineare.

Prima di definire gli stimatori di $\bar{Y}$, è utile fare alcune assunzioni di base. Sia w_i il peso dell'unità *i* appartenente al campione. La popolazione U viene divisa in K 'classi di imputazione': all'interno di una classe k, le osservazioni (x_i, y_i, z_i) sono indipendenti ed identicamente distribuite con $P(z_i = 1|x_i, y_i, k) = P(z_i = 1|x_i, k)$. Sebbene le osservazioni all'interno di una classe siano assunte i.i.d., il meccanismo di generazione dei dati mancanti è di tipo MAR in quanto $P(z_i = 1|x)$ dipende dalla covariata x.

Nella classe di imputazione k, sia z_k l'insieme degli indici delle unità rispondenti a y e $\bar{z}_k$ l'insieme degli indici delle unità non rispondenti, tale che $s_k = z_k \cup \bar{z}_k$. La dimensione del campione s_k nella classe k-esima è n_k, mentre la dimensione di U_k è N_k.

Con queste assunzioni, possiamo definire uno stimatore di $\bar{Y}$ come: (2.43)

$$\bar{y}_{NNI(1)} = \frac{1}{N}\sum_{k=1}^{K}\left(\sum_{i\in z_k} w_i y_i + \sum_{i\in z_k} w_i \tilde{y}_i\right) = \sum_{k=1}^{K}\frac{N_k}{N}\left(\sum_{i\in z_k} \bar{w}_{k,i} y_i + \sum_{i\in z_k} \bar{w}_{k,i} \tilde{y}_i\right)$$

dove $\tilde{y}_i$ è il valore imputato per il dato mancante y_i, $i \in \bar{z}_k$, e $\bar{w}_{k,i} = w_i/N_k$ quando $i \in s_k$. Se N è ignota, una sua stima consistente è data da $\hat{N} = \sum_{i\in s} w_i$; in questo caso, possiamo definire il seguente stimatore per rapporto di $\bar{Y}$:

$$\bar{y}_{NNI(2)} = \frac{1}{\hat{N}}\sum_{k=1}^{K}\left(\sum_{i\in z_k} w_i y_i + \sum_{i\in z_k} w_i \tilde{y}_i\right) = \frac{\bar{y}_{NNI(1)}}{\frac{\hat{N}}{N}} \qquad (2.44)$$

Le proprietà asintotiche dello stimatore (2.44) possono essere derivate dallo stimatore (2.43); per questo motivo, nel seguito, ci limiteremo a considerare lo stimatore $\bar{y}_{NNI(1)}$.

Assumendo valide le condizioni precedenti, si può dimostrare che

$$\frac{\sqrt{n}\left(\bar{y}_{NNI(1)} - \bar{Y}\right)}{\sigma} \xrightarrow{d} N(0,1) \qquad (2.45)$$

per ogni $\sigma > 0$, dove $\xrightarrow{d}$ rappresenta la convergenza in distribuzione.

Una versione semplificata della stima della varianza di (2.43) ottenuta col metodo jackknife corretto, è:

$$\widehat{Var}\left(\bar{y}_{NNI(1)}\right) = \sum_{k=1}^{K}\frac{1}{n_k(n_k-1)N^2}\sum_{j\in s_k}\left(n_k w_j \tilde{y}_j - \bar{y}_k\right)^2 \qquad (2.46)$$

dove

$\bar{y}_k = \sum_{i\in z_k}\left(1 + d_i^{(k)}\right) w_i y_i$,

$d_i^{(k)} = \sum_{j\in z_k}\left(w_j/w_i\right) d_{ij}$, con $d_{ij} = 1$ se i è l'unità più vicina a j, $d_{ij} = 0$ altrimenti;

$\tilde{y}_j = y_j + d_j^{(k)} g_j^{(k)}\left(y_j - \left(y_{jk1} + y_{jk2}\right)/2\right)$ se $j \in z_k$ e $\tilde{y}_j =$ valore imputato di y_j se $j \in \bar{z}_k$.

Infine, $g_j^{(k)} = \left[\sqrt{6\left(d_j^{(k)}\right)^2 + 6d_j^{(k)} + 4} - 2\right] \Big/ 3d_j^{(k)}$ $\left(g_j^{(k)} = 0 \; se \; d_j^{(k)} = 0\right)$ e

$jk1$ e $jk1$ sono le due unità più vicine a j in z_k.

2.3.2 Imputazione tramite pseudo-verosimiglianza empirica

Lo scopo di questo paragrafo è presentare metodi di stima e di imputazione basati sulla pseudo-verosimiglianza empirica, ipotizzando un meccanismo di generazione dei dati mancanti MAR.

La popolazione U viene suddivisa in H strati con N_h unità presenti nell'h-esimo strato. Si supponga di selezionare $n_h \geq 2$ unità dallo strato h attraverso un qualunque disegno di campionamento. A seconda del piano di campionamento scelto, si identificano i pesi $w_{hi} = (N\pi_{hi})^{-1}$, dove π_{hi} rappresenta la probabilità di inclusione nel campione dell'unità i dello strato h. Consideriamo, inoltre, le seguenti ipotesi asintotiche: $n_h \to \infty$ e $n_h/N_h \to 0$ per ogni h.

Sia Y la variabile di studio e A una variabile categoriale ausiliaria che assume valori in $\{a_1, \dots, a_\tau\}$ (τ costante fissata). Assumendo che Y abbia una distribuzione marginale non parametrica ignota F_h, possiamo scrivere la funzione di probabilità parametrica come:

$$P_h\left(A = a \mid Y = y\right) = f_h\left(y, z, \beta\right) \tag{2.47}$$

dove β è un vettore di parametri ignoti e f_h è una funzione nota. Senza perdita di generalità, ipotizziamo che in uno strato h, le prime r_h unità campionate siano rispondenti e il resto delle $n_h - r_h$ unità siano non rispondenti. Quindi, il dataset osservato è:

$$\left\{\left(Y_{hi}, A_{hi}\right),\ i = 1, \dots, r_h\right\} \cup \left\{A_{hi},\ i = r_h + 1, \dots, \mathrm{n}_h\right\},\quad h = 1, \dots, H \tag{2.48}$$

Utilizzando gli stimatori di massima verosimiglianza, consideriamo le seguenti procedure di imputazione:

1. Imputazione tramite media basata sulla pseudo-verosimiglianza. Per ogni unità non rispondente nello strato h, con $A = a_j$, il valore imputato di y è lo stimatore della media
$$\tilde{\tilde{y}}_{hj} = \sum_{i=1}^{r_h} \hat{p}_{hi} f_h\left(y_{hi}, a_j, \hat{\beta}\right) Y_{hi} \Big/ \sum_{i=1}^{r_h} \hat{p}_{hi} f_h\left(y_{hi}, a_j, \hat{\beta}\right) \tag{2.49}$$
2. Imputazione casuale basata sulla pseudo-verosimiglianza. Ogni unità non rispondente nello strato h, con $A = a_j$, viene imputata utilizzando un campione casuale con reinserimento estratto dall'insieme dei rispondenti dello strato h; la probabilità che un y_{hi} venga selezionato è pari a $f_h(y_{hi}, a_j, \hat{\beta})\hat{p}_{hi} / \sum_{i=1}^{r_h} f_h(y_{hi}, a_j, \hat{\beta})\hat{p}_{hi}$, $i = 1, \dots, r_h$.

Sia $p_{hi} = d\ F_h(y_{hi})$ tale che $p_{hi} \geq 0,\ \sum_{i=1}^{r_h} p_{hi} = 1,\ \sum_{i=1}^{r_h} p_{hi} f_h(y_{hi}, a_j, \beta) = \pi_{hj}$; utilizzando la tecnica dei moltiplicatori di Lagrange, è possibile derivare che

$$p_{hi} = \frac{w_{hi}}{\sum_{i=1}^{n_h} w_{hi} - \sum_{j=1}^{\tau} \frac{c_{hj}}{\pi_{hj}} f_f\left(y_{hi}, a_j, \beta\right)} \tag{2.50}$$

dove $c_{hj} = \sum_{i=r_h+1}^{n_h} w_{hi} I_{\{A_{hi}=a_j\}}$ e I_A è la funzione indicatrice dell'evento A.

La stima di p_{hi} presente in (2.49) viene calcolata utilizzando una stima consistente di π_{hj} $\left(\pi_{hj} = P_h(Z = z_j)\right)$, definita come:

$$\hat{\pi}_{hj} = \sum_{i=1}^{n_h} w_{hi} I_{\{Z_{hi}=z_j\}} \Big/ \sum_{i=1}^{n_h} w_{hi} \tag{2.51}$$

e usando una stima di β, ottenuta massimizzando la seguente funzione di pseudo verosimiglianza empirica (2.52)

$$l\left(\beta, \hat{\pi}\right) = \sum_{h=1}^{H} \left[\sum_{i=1}^{r_h} w_{hi} \log \left(\frac{w_{hi} f_h\left(y_{hi}, a_{hi}, \beta\right)}{\sum_{i=1}^{n_h} w_{hi} - \sum_{j=1}^{\tau} \frac{c_{hj}}{\hat{\pi}_{hj}} f_h\left(y_{hi}, a_j, \beta\right)} \right) + \sum_{j=1}^{\tau} c_{hj} \log\left(\hat{\pi}_{hj}\right) \right]$$

rispetto a β.

Dopo aver effettuato le eventuali imputazioni dei dati mancanti, la $\bar{Y}$ può essere stimata come:

$$\bar{y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} \tilde{y}_{hi} \tag{2.53}$$

dove $\tilde{y}_{hi} = y_{hi}$ se y_{hi} è un'unità rispondente, altrimenti $\tilde{y}_{hi}$ rappresenta il valore imputato dell'unità non rispondente y_{hi}.

Si può dimostrare che lo stimatore (2.53), basato sulle due procedure di imputazione descritte precedentemente, è consistente ed asintoticamente normale. Non viene presentata la varianza asintotica dello stimatore, in quanto, essendo di difficile derivazione, può essere calcolata solamente applicando il metodo bootstrap.

2.3.3 Metodi di imputazione doppiamente robusti.

Un metodo in grado di ridurre il problema della dimensionalità, che limita l'utilizzo delle attuali tecniche di imputazione, è il metodo non parametrico di imputazione multipla. Questa procedura è doppiamente robusta e differisce dagli stimatori per calibrazione, in quanto evita la ponderazione per probabilità inversa, basandosi unicamente sull'imputazione utilizzando due modelli di lavoro.

Nella prima parte del paragrafo, si introducono i modelli di lavoro; quindi, vengono descritte in dettaglio le procedure doppiamente robuste di imputazione.

Per poter utilizzare pienamente l'informazione apportata dalla variabile ausiliaria x allo scopo di definire un set imputato per ogni osservazione di y mancante, consideriamo due modelli. Il primo riguarda la regressione sulla variabile risposta y,

$$E(Y \mid X_0) = m(X_0\ ,\ \beta) \tag{2.54}$$

dove $m(\cdot,\cdot)$ è una specifica funzione a valori reali lisciata; X_0 è un set di p_1 osservazioni delle variabili ausiliarie e $\beta = (\beta_1, \dots, \beta_{p_1})'$ è il vettore dei coefficienti di regressione. Se il modello (2.54) è correttamente specificato e il meccanismo di generazione dei dati mancanti è di tipo MAR, si può ottenere una riduzione della distorsione. Il secondo modello è costruito per predire i valori della variabile indicatrice di risposta z_i, ed è rappresentato come:

$$E(z_i \mid X_m) = p(X_m\ ,\ \alpha) \tag{2.55}$$

dove p(·,·) è una specifica funzione a valori reali lisciata; X_m è un set di p_2 osservazioni delle variabili ausiliarie e $\alpha = (\alpha_1, \dots, \alpha_{p_2})'$ è il vettore dei coefficienti di regressione.

Siano $G_1 = m(X_0, \beta)$ e $G_2 = p(X_m, \alpha)$. Dopo aver stimato i parametri dei modelli (2.54) e (2.55) con il metodo della massima verosimiglianza, le stime dei predittori possono essere indicate come: $(\hat{G}_1, \hat{G}_2) = \{m(X_0, \hat{\beta}), p(X_m, \hat{\alpha})\}$. Le stime $\hat{\alpha}\ e\ \hat{\beta}$ convergono in probabilità rispettivamente ai valori dei veri parametri $\alpha^0\ e\ \beta^0$. La strategia proposta è in grado di ridurre la multi-dimensionalità di X, rendendo bi-dimensionale il predittore $G \equiv (G_1, G_2)$.

Per stabilizzare l'imputazione, il predittore G viene standardizzato con la sua media e la sua deviazione standard; il risultato viene indicato con $S \equiv (S_1, S_2)$. Dato S, per ogni osservazione con y mancante, si crea un set imputato di risposte osservate simili tra loro. In particolare, S viene utilizzato per selezionare il set imputato calcolando le distanze tra le osservazioni come:

$$d(i,j)=\left\{\omega_0\left[S_1(i)-S_1(j)\right]^2+\omega_m\left[S_2(i)-S_2(j)\right]^2\right\}^{1/2} \tag{2.56}$$

dove ω_0 *e* ω_m rappresentano i pesi non negativi per i predittori (2.54) e (2.55), rispettivamente, con $\omega_0+\omega_m=1$. Per ogni osservazione i con y mancante, la distanza (2.56) viene usata per definire un set di K soggetti $\left(R_K(i)\right)$ che presentano minor distanza (d) dall'osservazione *i*.

Definiti, quindi, i datasets imputati, proponiamo uno stimatore di imputazione multipla per il parametro di interesse $\bar{Y}$. Nell'l-esimo dataset, definito $R_K(i)$ per ogni osservazione i con dato mancante, si estrae casualmente un y_i^* da $R_K(i)$ per sostituirlo all'i-esimo dato mancante. Ripetiamo questa procedura per tutte le osservazioni che presentano mancata risposta e definiamo $\{\tilde{y}_i(l)=z_iy_i+(1-z_i)y_i^*(l)\ \ (i=1,\dots,n)\}$ e $\hat{y}(l)=\sum_{i=1}^n\tilde{y}_i(l)$ rispettivamente, l'l-esimo dataset imputato e l'associato stimatore per media. Lo stimatore finale di imputazione multipla per $\bar{Y}$ è

$$\hat{y}_{MI}=\frac{1}{L}\sum_{l=1}^{L}\hat{y}(l) \tag{2.57}$$

Consideriamo, ora, lo stimatore (2.57) calcolato utilizzando $G^0=\{m(X_0,\beta^0),p(X_m,\alpha^0)\}$, chiamato $\hat{y}_{MI}^0$. Presi $\mu(G^0)=E(Y|G^0)$, $p_g(G^0)=Pr(z_i=1|G^0)$ e $\sigma^2(G^0)=Var(Y|G^0)$, se (2.54) e (2.55) sono correttamente specificati, allora $\sqrt{n}(\hat{y}_{MI}^0-\bar{Y})$ ha distribuzione asintotica normale con media 0 e varianza

$$\sigma_{MI}^2=\sigma_1^2+\sigma_2^2+\sigma_3^2+2\sigma_{23}^2 \tag{2.58}$$

dove

$$\sigma_1^2=Var\left[\mu\left(G^0\right)\right],\ \ \sigma_2^2=E\left[Var\left(z_i\left\{Y-\mu\left(G^0\right)\right\}|G^0\right)\right],$$

$$\sigma_3^2=E\left[\frac{\left[1-p_g\left(G^0\right)\right]^2}{p_g\left(G^0\right)^2}Var\left[z_i\left\{Y-\mu\left(G^0\right)\right\}|G^0\right]\right],$$

$$\sigma_{23}=E\left[\frac{1-p_g\left(G^0\right)}{p_g\left(G^0\right)}Var\left[z_i\left\{Y-\mu\left(G^0\right)\right\}|G^0\right]\right].$$

Utilizzando $\hat{\alpha}$ *e* $\hat{\beta}$, e argomenti simili a quelli del teorema appena enunciato, è semplice dimostrare che $\hat{y}_{MI}$ ha la stessa distribuzione asintotica di $\hat{y}_{MI}^0$. Si può scrivere, quindi, la varianza asintotica di $\hat{y}_{MI}$ come

$n^{-1}\big(Var(Y) + E\big[\{p_g(G^0)^{-1} - 1\}\sigma^2(G^0)\big]\big)$. Quando entrambi i modelli sono correttamente specificati, la varianza asintotica di $\hat{y}_{MI}$ si riduce a $n^{-1}\big(Var(Y) + E\big[\{p_g(X)^{-1} - 1\}\sigma^2(X)\big]\big)$.

Nell'ultima parte del paragrafo, analizziamo una procedura di imputazione doppiamente robusta che fa uso del metodo di regressione non parametrico.

Consideriamo un campione casuale $(y_i, x_i, z_i), i = 1, \dots, n$ con dati mancanti relativi solamente alla variabile di studio y. Un classico stimatore di $\bar{Y}$ è dato da:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}\left[z_i y_i + (1 - z_i)\tilde{y}_i\right] \tag{2.59}$$

dove $\tilde{y}_i$ è il valore imputato per y_i.

Quando X_i è univariata, $E(y_i|\, x_i)$ può essere effettivamente stimata usando un qualsiasi metodo di regressione non parametrico. Quando, invece, X_i è multidimensionale, gli stimatori non parametrici non sono efficienti a causa della dimensionalità. Se imponiamo un modello parametrico $E(y_i|\, x_i) = \mathrm{m}(\mathrm{x_i}, \beta)$, dove $\mathrm{m}(\cdot,\cdot)$ è una funzione nota e β è un vettore di parametri ignoti, è possibile imputare un valore mancante di y con $\mathrm{m}(\mathrm{x_i}, \hat{\beta})$, dove $\hat{\beta}$ è una stima di β utilizzando $(y_i, x_i, z_i = 1)$, $i = 1, \dots, n$. Questo approccio non è robusto se siamo in presenza di errata specificazione del modello di regressione.

Un altro approccio per imputare i valori mancanti di y, consiste nell'utilizzare una stima di $E(y|z_i = 0)$, indicata da $\hat{y}_{z_i=0}$. Notiamo che:

$$E(y|z_i = 0) = \frac{\iint y Pr(z_i = 0|x) dF(y,x)}{1 - p} \tag{2.60}$$

dove $F(y,x)$ è la distribuzione congiunta di (y,x) e $p = Pr(z_i = 1)$. Quindi $E(y|z_i = 0)$ può essere stimata da:

$$\hat{y}_{z_i=0} = \frac{\iint y\widehat{Pr}(z_i = 0|x) d\hat{F}(y,x)}{1 - \hat{p}} = \sum_{i=1}^{n} z_i \hat{q}_i y_i \tag{2.61}$$

dove $\hat{q}_i$ è la probabilità condizionata di (y_i, x_i) dato $z_i = 0$ e viene stimata, come pure $\widehat{Pr}(z_i = 0|x), \hat{F}, \hat{p}$, attraverso un approccio di pseudo-verosimiglianza empirica. A differenza della precedente metodologia di imputazione, non occorre specificare un modello per $E(y|x)$; questo rende l'approccio robusto se si è in presenza di errata specificazione del modello di regressione. Un aspetto negativo nell'utilizzare (2.61) è

che i valori mancanti di y con differenti valori della variabile ausiliaria sono tutti imputati con lo stesso valore, rendendo la tecnica non efficiente.

Per derivare una procedura di imputazione che sia doppiamente robusta, si possono unire i due metodi descritti precedentemente, imputando gli y mancanti con:

$$\tilde{y}_i = m\left(x_i, \hat{\beta}\right) + \sum_{i=1}^{n} z_i \hat{q}_i \left[y_i - m\left(x_i, \hat{\beta}\right) \right] \tag{2.62}$$

Lo stimatore risultante di $\bar{Y}$ è:

$$\begin{aligned} \hat{y}_{ER} &= \frac{1}{n} \sum_{i=1}^{n} \left[z_i y_i + (1 - z_i) m\left(x_i, \hat{\beta}\right) \right] \\ &+ \left(1 - \frac{n_r}{n}\right) \sum_{i=1}^{n} z_i \hat{q}_i \left[y_i - m\left(x_i, \hat{\beta}\right) \right] \end{aligned} \tag{2.63}$$

$$\begin{aligned} &= \frac{1}{n} \sum_{i=1}^{n} \left[z_i y_i + (1 - z_i) \hat{y}_{z_i=0} \right] + \sum_{i=1}^{n} \left[(1 - z_i) m\left(x_i, \hat{\beta}\right) \right] \\ &- \left(1 - \frac{n_r}{n}\right) \sum_{i=1}^{n} z_i \hat{q}_i m\left(x_i, \hat{\beta}\right) \end{aligned} \tag{2.64}$$

dove $n_r = \sum_{i=1}^{n} z_i$ è il numero delle y_i osservate. Per quel che riguarda la doppia robustezza di (2.63), notiamo che se il modello di regressione $\mathrm{m}(\mathrm{x_i}, \beta)$ è corretto e il secondo termine dello stimatore (2.63) converge a 0, allora $\hat{y}_{ER}$ converge a $\bar{Y}$, indipendentemente dal fatto che $\widehat{Pr}(z_i = 0|x)$ sia corretta. Se il modello associato a $Pr(z_i = 0|x)$ è specificato correttamente e la somma degli ultimi due termini di (2.64) convergono a 0, l'espressione (2.64) converge a $\bar{Y}$ indipendentemente dal fatto che $\mathrm{m}(\mathrm{x_i}, \beta)$ sia corretto.

Si può dimostrare, inoltre, che se il modello associato a $Pr(z_i = 0|x)$ è correttamente specificato, allora

$$\sqrt{n}\left(\hat{y}_{ER} - \bar{Y}\right) \xrightarrow{d} N\left(0, \sigma^2\right) \tag{2.65}$$

per $\sigma > 0$. La stima della varianza asintotica di (2.63) viene calcolata tramite il metodo bootstrap.

2.3.4 Imputazione ponderata

In questa ultima sezione, viene presentato un metodo di imputazione ponderata basato sul rapporto di log-verosimiglianza empirica aggiustato con fattori correttivi.

Sia $x = (x_1, \dots, x_d)'$ un vettore di variabili ausiliarie osservate per tutto il campione. Il campione osservato è $(y_i, x_i, z_i),\ i = 1, \dots, n$ ed è caratterizzato da dati mancanti solo per la variabile di studio y.

Per costruire la funzione rapporto di verosimiglianza empirica per $\bar{Y}$, si può applicare un'imputazione per regressione basata su funzioni kernel per introdurre le variabili ausiliarie:

$$\tilde{y}_i = z_i y_i + (1 - z_i)\hat{m}_b(x_i), \quad i = 1, \dots, n \tag{2.66}$$

dove $\hat{m}_b(x)$ è una versione troncata dello stimatore di $m(x) = E(Y|X)$ ed è definito come:

$$\hat{m}_b(x) = \frac{(nh^d)^{-1} \sum_{i=1}^{n} z_i y_i K_h(X_i - x)}{\max\left\{b,\ (nh^d)^{-1} \sum_{i=1}^{n} z_i K_h(X_i - x)\right\}} \tag{2.67}$$

Nell'espressione (2.67), $h = h_n$ e $b = b_n$ sono sequenze di costanti positive tendenti a zero, mentre $K_h(\cdot) = K(\cdot/h)$ e $K(\cdot)$ è una funzione kernel. Usando (2.66) è possibile costruire uno stimatore del rapporto di log-verosimiglianza empirica, chiamato $\tilde{l}(\bar{Y})$. Si può dimostrare che la distribuzione asintotica di $\tilde{l}(\bar{Y})$ non è un chi-quadrato standard ma un chi-quadrato non centrale con un grado di libertà. Per questo motivo, dobbiamo aggiustare $\tilde{l}(\bar{Y})$ con fattori correttivi oppure calcolando direttamente il rapporto di log-verosimiglianza empirica ponderato, $\hat{l}(\bar{Y})$.

Se scegliamo di adottare il secondo metodo di correzione, per ridurre la distorsione $\hat{m}_b(X_i) - m(X_i)$ apportata da $\tilde{y}_i$, utilizziamo un approccio di imputazione ponderata che introduce una nuova variabile ausiliaria y_i^* dipendente dalla stima della probabilità di risposta $\hat{p}(X_i)$ e definita come:

$$y_i^* = \frac{z_i y_i}{\hat{p}(X_i)} + \left(1 - \frac{z_i}{\hat{p}(X_i)}\right)\hat{m}_b(X_i) \tag{2.68}$$

dove

$$\hat{p}(X_i) = \frac{\sum_{i=1}^{n} z_i L_a(X_i - x)}{\max\left\{1,\ \sum_{i=1}^{n} L_a(X_i - x)\right\}} \tag{2.69}$$

Nell'espressione (2.69), $a = a_n$ è una sequenza positiva di costanti tendente a zero, mentre $L_a(\cdot) = L(\cdot/a)$ e $L(\cdot)$ è una funzione kernel. Possiamo, quindi, definire $\hat{l}(\bar{Y})$ come:

$$\hat{l}\left(\bar{Y}\right) = -2\max\sum_{i=1}^{n}\log\left(np_i\right) \tag{2.70}$$

dove il massimo è preso da tutte le serie non negative di numeri $p_1, \ldots, p_n$ che sommano a uno e tali che $\sum_{i=1}^{n} p_i y_i^* = \bar{Y}$. Applicando il metodo dei moltiplicatori di Lagrange, quando $min_{1\le i\le n}(y_i^*) < \bar{Y} < max_{1\le i\le n}(y_i^*)$, $\hat{l}(\bar{Y})$ può essere rappresentata come:

$$\hat{l}\left(\bar{Y}\right) = 2\sum_{i=1}^{n}\log\left(1+\lambda\left(y_i^* - \bar{Y}\right)\right) \tag{2.71}$$

dove $\lambda = \lambda(\bar{Y})$ è la soluzione dell'equazione

$$\sum_{i=1}^{n}\frac{y_i^* - \bar{Y}}{1+\lambda\left(y_i^* - \bar{Y}\right)} = 0 \tag{2.72}$$

Dopo aver imputato i valori dei dati mancanti, proponiamo due stimatori di $\bar{Y}$: il primo è definito come

$$\hat{y}_{WR} = \frac{1}{n}\sum_{i=1}^{n}\tilde{y}_i \tag{2.73}$$

mentre il secondo, utilizzante l'imputazione ponderata, è definito come

$$\hat{y}_{WI} = \frac{1}{n}\sum_{i=1}^{n}y_i^* \tag{2.74}$$

Si può provare che l'espressione seguente, posto $\hat{\theta} = \hat{y}_{WR}\ o\ \hat{y}_{WI}$, ha distribuzione asintotica

$$\frac{\sqrt{n}\left(\hat{\theta} - \bar{Y}\right)}{\widehat{Var}^{1/2}\left(\hat{\theta}\right)} \xrightarrow{d} N(0,1)$$

dove $\widehat{Var}^{1/2}\left(\hat{\theta}\right)$ è lo stimatore consistente della varianza asintotica di $\hat{\theta}$ che vale:

$$\widehat{Var}^{1/2}\left(\hat{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - \hat{\theta}\right)^2 \tag{2.75}$$

con $\hat{y}_i = \tilde{y}_i\ o\ y_i^*$.

2.4 Nota bibliografica

Nel secondo capitolo, si sono analizzati alcuni metodi per stimare la media di una variabile in presenza di dati mancanti utilizzando metodi d'imputazione.

Per quanto riguarda i dati mancanti di tipo MCAR, informazioni utili riguardanti l'imputazione tramite media e rapporto, vengono fornite nell'articolo di Kadilar e Cingi (2008), oppure è possibile approfondire la sola imputazione per rapporto nella pubblicazione di Toutenburg, Srivastava e Shalabh (2008). Un articolo da prendere in considerazione se si utilizzano le tecniche d'imputazione basate sugli stimatori indiretti, è quello di Rueda, Gonzàlez e Arcos (2005). Sempre sotto assunzione MCAR, si è presentata l'imputazione multipla nel contesto del campionamento stratificato esaminando l'articolo di Bjørnstad (2007); se il campionamento è per cluster, si rimanda agli articoli di Shao (2007) e Yuan e Little (2008).

Per il meccanismo di non risposta MAR, è stato presentato uno stimatore che utilizza l'imputazione 'nearest neighbor' esaminando l'articolo di Shao e Wang (2008); per quanto riguarda le tecniche d'imputazione che usano la pseudo-verosimiglianza empirica, è possibile riferirsi alla pubblicazione di Fang, Hong e Shao (2009). Sotto assunzione MAR, si sono presentati, inoltre, metodi d'imputazione doppiamente robusti analizzando: l'articolo Long, Hsu e Yisheng (2012) e, se si utilizza il metodo di regressione non parametrico, l'articolo Qin, Shao e Zhang (2008). In merito all'imputazione ponderata basata su una funzione Kernel, un utile approfondimento si può trovare nell'articolo di Xue (2009).

Bibliografia

[1] Bjørnstad, J. F. (2007) “Non-Bayesian multiple imputation”, *Journal of Official Statistics*, 23, 4, pp. 433-452.

[2] Cao, W., Tsiatis, A. A. e Davidian, M. (2009) “Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data”, *Biometrika*, 96, 3, pp. 723-734.

[3] Chen, J. e Rao, J. N. K. (2007) “Asymptotic normality under two-phase sampling designs”, *Statistica Sinica*, 17, 3, pp. 1047-1064.

[4] Da Silva, D. N. e Opsomer, J. D. (2006) “A kernel smoothing method of adjusting for unit non-response in sample surveys”, *The Canadian Journal of Statistics*, 34, 4, pp. 563-579.

[5] Da Silva, D. N. e Opsomer, J. D. (2009)”Nonparametric propensity weighting for survey nonresponse through local polynomial regression”, *Survey Methodology*, 35, 2, pp. 165-176.

[6] Fang, F. Hong, Q. e Shao, J. (2009) “A pseudo empirical likelihood approach for stratified samples with nonresponse”, *The Annals of Statistics*, 37, 1, pp. 371-393.

[7] Gonzàlez, S., Rueda, M. M. e Arcos, A. (2008) “An improved estimator to analyse missing data”, *Statistical Papers*, 49, 4, pp. 791-796.

[8] Hu, Z., Follmann, D. A. e Qin, J. (2010) "Semiparametric dimension reduction estimation for mean response with missing data", *Biometrika*, 97, 2, pp. 305-319.

[9] Kadilr, C. e Cingi, H. (2008) "Estimators for the population mean in the case of missing data", *Communications in Statistics-Theory and Methods*, 37, 14, pp. 2226-2236.

[10] Kim, J. K. e Kim, J. J. (2007) "Nonresponse weighting adjustment using estimated response probability", *The Canadian Journal of Statistics*, 35, 4, pp. 501-514.

[11] Little, R. J. e Vartivarian, S. (2005) "Does weighting for nonresponse increase the variance of survey means?", *Survey Methodology*, 31, 2, pp. 161-168.

[12] Long, Q., Hsu, C. H. e Yisheng, L. (2012) "Doubly robust nonparametric multiple imputation for ignorable missing data", *Statistica Sinica*, 22, 1, pp. 149-172.

[13] Qin, J., Shao, J. e Zhang, B. (2008) "Efficient and doubly robust imputation for covariate-dependent missing responses", *Journal of the American Statistical Association*, 103, 482, pp. 797-810.

[14] Rueda, M. M., Gonzàlez, S. e Arcos, A. (2005) "Indirect methods of imputation of missing data based on available units", *Applied Mathematics and Computation*, 164, 1, pp. 249-261.

[15] Rueda, M. M., Gonzàlez, S. e Arcos, A. (2006) "A general class of estimators with auxiliary information based on available units", *Applied Mathematics and Computation*, 175, 1, pp. 131-148.

[16] Rueda, M. M., Muñoz, J. F., Berger, Y. G., Arcos, A. e Martínez, S. (2007) "Pseudo empirical likelihood method in the presence of missing data", *Metrika*, 65, 3, pp. 349-367.

[17] Shao, J. (2007) "Handling survey nonresponse in cluster sampling", *Survey Methodology*, 33, 1, pp. 81-85.

[18] Shao, J. e Wang, H. (2008) "Confidence intervals based on survey data with nearest neighbor imputation", *Statistica Sinica*, 18, 1, pp. 281-297.

[19] Singh, H. P. e Kumar, S. (2008) "Estimation of mean in presence of non-response using two phase sampling scheme", *Statistical Papers*, 51, 3, pp. 559-582.

[20] Singh, H. P., Kumar, S. e Kozac, M. (2010b) "Improved estimation of finite-population mean using sub-sampling to deal with non response in two-phase sampling scheme", *Communications in Statistics-Theory and Methods*, 39, 5, pp.791-802.

[21] Tan, Z. (2010) "Bounded, efficient and doubly robust estimation with inverse weighting", *Biometrika*, 97, 3, pp. 661-682.

[22] Toutenburg, H., Srivastava, V. K. e Shalabh (2008) "Amputation versus imputation of missing values through ratio method in sample surveys", *Statistical Papers*, 49, 2,pp. 237-247.

[23] Xue, L. (2009) "Empirical likelihood confidence intervals for response mean with data missing at random", *Scandinavian Journal of Statistics*, 36, 4, pp. 671-685.

[24] Yuan, Y e Little, R. J. A. (2008) "Model-based inference for two-stage cluster samples subject to nonignorable item nonresponse", *Journal of Official Statistics*, 24, 2, pp.193-211.