UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

CORSO DI LAUREA TRIENNALE IN
STATISTICA E GESTIONE DELLE IMPRESE

TESI DI LAUREA

STIMATORI DELLA MEDIA DI UNA POPOLAZIONE PER DATI MANCANTI

Relatore: Ch.mo Prof. Giancarlo Diana

Laureando: Stefano Campigotto

ANNO ACCADEMICO 2010-2011

Alla mia famiglia,
in particolare a Federico

Indice

Introduzione

Le indagini spesso soffrono del problema della non risposta, ovvero della mancanza di un certo numero di dati rispetto alla numerosità campionaria programmata per la ricerca stessa. Davanti a questo problema, è naturale domandarsi se questo abbia effetto sull'attendibilità dei risultati finali dell'indagine, in particolare sulle statistiche campionarie usate per stimare i parametri d'interesse. Si possono presentare due casi: se le non risposte non intaccano i risultati, allora sono ignorate; se invece la mancanza di dati influenza i risultati finali, allora sorgerebbe il quesito su come limitare o eliminare gli effetti di tale problema.

Avere a che fare con i due casi, soprattutto il secondo, non è compito facile: si possono offrire una serie di soluzioni per aiutare a gestire in modo efficace ed efficiente il problema. L'obbiettivo di questo lavoro è fare una rassegna delle tecniche utilizzate per risolvere i casi appena detti. In particolare, tale lavoro raccoglie gli ultimi sviluppi relativamente ai metodi di stima della media della popolazione.

L'interesse verso l'organizzazione di un lavoro di questo tipo, trova giustificazione non solo nel fatto che il fenomeno delle non risposte è comune a tutti i tipi d'indagine, ma anche perchè nelle indagini odierne si ha un costante incremento del numero di non risposte. Di conseguenza, è diventato sempre più necessario avere a disposizione degli strumenti per risolvere i problemi che esse pongono.

Il problema delle non risposte, può essere analizzato da diversi punti di vista: il presente lavoro passa in rassegna tutti i possibili approcci alla stima della media; con questa finalità, si è deciso di esaminare la letteratura sull'argomento relativa al periodo 2005-2010, volendo dare in questo modo un compendio dei più recenti sviluppi in materia.

Analizzando gli articoli pubblicati nel periodo prescelto, ed inerenti al

tema trattato, si è notato che non esiste una simbologia univoca e universalmente utilizzata: si è ritenuto opportuno uniformare il linguaggio per tutti i metodi di stima descritti e precisare il significato di ogni termine utilizzato. Ciò dovrebbe favorire, anche, un impiego pratico del lavoro visto che le non risposte risultano un problema rilevante nella pratica statistica.

L'errore complessivo di una statistica campionaria, utilizzata come stimatore del corrispondente parametro della popolazione, si suddivide in campionario e non campionario: il primo è dovuto al fatto che non si osserva l'intera popolazione ma solo un sottoinsieme, cioè il campione; il secondo è dovuto a tutte quelle cause che non riguardano l'incompletezza della rilevazione. Ad esempio, tra gli errori non campionari vi sono quelli di copertura che dipendono dall'inadeguatezza della lista utilizzata per identificare gli elementi della popolazione. Tra gli errori non campionari, vi sono anche gli errori di non risposta: essi sono dovuti al fatto che, nella fase di rilevazione, alcune unità non vengono reperite o si rifiutano di rispondere per cui i dati raccolti sono incompleti. Vi sono due tipi di mancata risposta: totale o parziale. La mancata risposta totale è l'assenza completa di informazioni su alcune unità del campione; la mancata risposta parziale, è l'assenza di risposta ad alcuni quesiti.

L'interesse, più che attuale, nei confronti del fenomeno delle non risposte, è giustificato dalle conseguenze che esso ha sui risultati finali. Poichè tra gli obbiettivi di un'indagine vi è la stima di uno o più parametri, si valutano le conseguenze delle non risposte direttamente sugli stimatori e sulle loro proprietà. In tale situazione, le procedure statistiche standard per dati completi non possono essere immediatamente applicate per fare inferenza. E' problema rilevante allora, valutare le conseguenze delle non risposte in termini di effetti sullo stimatore: ad esempio, una conseguenza ovvia della non risposta è che la dimensione del campione effettivo è inferiore a quella programmata, il che può produrre un aumento della varianza dello stimatore. Tra l'altro, se il meccanismo che genera le mancate risposte è non casuale, le statistiche campionarie risultano distorte.

Vogliamo analizzare le tecniche di stima in presenza di dati mancanti e, per questo, il materiale recuperato in letteratura è stato organizzato secondo due criteri fondamentali: il tipo di meccanismo generatore delle non risposte ('missing completely at random', o MCAR; 'missing at random', o MAR) e

il metodo d'imputazione dei valori mancanti.

Partendo da questi concetti, abbiamo articolato la tesi nel seguente modo.

Il Capitolo 1 presenta una rassegna degli stimatori proposti in letteratura, a seconda del meccanismo generatore dei dati mancanti: nessun metodo d'imputazione viene considerato. All'interno della sezione 1.2, si mostrano gli stimatori nel caso MCAR: essi sono stati organizzati in due gruppi. Il primo gruppo mostra gli stimatori che si basano sul campionamento a due fasi; il secondo gruppo, mostra gli stimatori che utilizzano tutti i valori disponibili per ogni variabile osservata senza eliminare le osservazioni incomplete. All'interno della sezione 1.3, si presentano gli stimatori nel caso MAR: questi sono stati divisi in tre gruppi. Il primo gruppo comprende gli stimatori doppiamente robusti; il secondo gruppo, gli stimatori basati su un approccio semiparametrico alla stima; nel terzo gruppo sono presentati gli stimatori che fanno uso di metodi di aggiustamento per ponderazione.

Il Capitolo 2 presenta gli stimatori proposti in letteratura, a seconda del meccanismo generatore dei dati mancanti e del metodo d'imputazione che viene usato. Nella sezione 2.2, troviamo due gruppi di stimatori nel caso MCAR: il primo, riporta diversi stimatori a seconda che questi utilizzano l'imputazione per la media, per il rapporto, tramite differenze o per regressione; nel secondo gruppo, vengono presentati stimatori nel contesto del campionamento stratificato e per clusters, anche con l'utilizzo dell'imputazione multipla. Nella sezione 2.3, si mostrano gli stimatori nel caso MAR: in questo caso, si sono rilevati tre gruppi. Il primo gruppo contiene gli stimatori basati sull'imputazione 'nearest neighbor'; il secondo riporta gli stimatori che si basano sull'imputazione tramite pseudo-verosimiglianza; il terzo presenta una tecnica di stima basata sull'imputazione ponderata dei dati mancanti.

Notazioni e assunzioni

Nel presente lavoro si considera una popolazione finita U di N individui. Da questa popolazione si estrae un campione s di n ($n < N$) individui.

Tale campione è estratto secondo un disegno campionario $d = (S_d, P_d)$ con probabilità di inclusione di primo e secondo ordine pari a π_i e π_{ij}, rispettivamente ($i = 1, ..., N$, $j = 1, ..., N$, $i \neq j$). Quando le probabilità di inclusione nel campione s, sono tutte costanti allora saremo nel contesto del campione casuale semplice. Indicheremo con SRSWOR il campionamento casuale semplice senza reinserimento, che è la procedura di selezione più spesso usata nella pratica, anche in processi di selezione intermedi di disegni più complessi (tipo campionamento a due o più stadi). Qualora non si specificasse nulla, si sottointende che l'estrazione del campione s è effettuata con SRSWOR.

Sia $y \in \mathbb{R}^+$, la variabile quantitativa oggetto di studio di media ignota: si dispone di un campione di n risposte, così indicato $y_1, y_2, ..., y_n$. Si può disporre, per ciascuno degli N individui della popolazione, di un altro carattere quantitativo correlato con la variabile di studio y. Siano allora $x_1, ..., x_n$, i valori della variabile ausiliaria osservati contestualmente alla variabile di interesse y.

All'interno del campione, un sottogruppo di unità non risponde per la parte relativa alla caratteristica di interesse y. Si definisca allora a_i, per $i = 1, ..., n$, la variabile indicatrice che assume valori:

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

Il nostro scopo è quello di stimare la media di y, in questo contesto il campione è indicato con

$$((a_1, a_1 y_1, x_1), (a_2, a_2 y_2, x_2), ..., (a_n, a_n y_n, x_n,)).$$

In linea di massima considereremo una sola variabile ausiliaria correlata con la variabile di studio, senza per questo escludere la possibilità di considerare più variabili ausiliarie correlate con y.

Nel corso della tesi faremo uso, per la tipologia di mancate risposte, dei termini 'missing completely at random' e 'missing at random', indicati dalle sigle MCAR e MAR. Questa distinzione dipende dal fatto che la probabilità di risposta dell'i-esima unità del campione può dipendere o meno dalla variabile di studio e dalla variabile ausiliaria. Assumendo che, dato il campione s, gli indicatori di risposta siano variabili casuali indipendenti avremo che:

- 'missing completely at random' (MCAR), $P\left(\ a_i = 1|\ y_i,x_i\right) = p$ con $0 \leq p \leq 1$, per $i = 1, ..., n$; in questo caso, la probabilità di risposta è indipendente sia dalla variabile di studio che dalla variabile ausiliaria associata.
- 'missing at random' (MAR), $P\left(\ a_i = 1|\ y_i,x_i\right) = P\left(a_i = 1|x_i\right) = \phi(x_i) = \phi_i$ con $0 \leq \phi_i \leq 1$, per $i = 1, ..., n$; in questo caso, la probabilità di risposta è dipendente dalla variabile ausiliaria osservata per l'unità i-esima, ma indipendente dalla variabile di studio y.

Denoteremo nel seguito, con le lettere maiuscole, le caratteristiche della popolazione oggetto di studio mentre con le lettere minuscole le corrispondenti quantità nel campione. Le assunzioni che si descrivono, di seguito, per la variabile y valgono anche per qualsiasi altra variabile ausiliaria.

Si indicano rispettivamente media, varianza, e varianza corretta di y con

$$\overline{Y} = \tfrac{1}{N}\sum_{i=1}^{N} y_i \qquad \sigma_Y^2 = \tfrac{1}{N}\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2 \qquad S_y^2 = \tfrac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2.$$

La covarianza, il coefficiente di correlazione e il coefficiente di regressione tra le variabili y e x sono indicati rispettivamente con

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

Le quantità campionarie sono invece indicate con

$$\overline{y} = \tfrac{1}{n}\sum_{i=1}^{n} y_i \qquad s_y^2 = \tfrac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 = \widehat{S}_y$$

$$s_{xy} = \tfrac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)\left(x_i - \overline{x}\right) = \widehat{S}_{xy} \quad \widehat{\rho} = \tfrac{s_{xy}}{s_x s_y} \quad \widehat{\beta} = \tfrac{s_{xy}}{s_x^2}.$$

Un'altra quantità usata è il rapporto di popolazione,

$R = \frac{\overline{Y}}{\overline{X}}$.

Capitolo 1

Stimatori della media senza metodi d'imputazione

1.1 Introduzione

In questo capitolo, si presentano varie tecniche di stima della media di una variabile di studio y in presenza di dati mancanti, per una popolazione finita: i metodi di stima riguardano sia il contesto di dati mancanti del tipo MCAR sia quello di dati mancanti di tipo MAR. Queste tecniche sono accomunate dal fatto che non utilizzano metodi d'imputazione dei dati mancanti: in questo caso, il singolo valore o addirittura le unità che non presentano risposta sono eliminate dai dati a disposizione per derivare stime di parametri di popolazione. Di seguito, illustriamo i vari stimatori proposti in letteratura negli ultimi cinque anni, ponendo l'attenzione sul metodo di stima, sulla varianza dello stimatore e sulle proprietà rilevanti ai fini dell'utilizzo pratico.

1.2 Dati mancanti MCAR

Quando il meccanismo che genera i dati mancanti si assume essere MCAR, la probabilità di risposta è indipendente sia dal valore della variabile di studio che dal valore di tutte le altre variabili ausiliarie che entrano come informazione aggiuntiva sulle unità di interesse. In questo caso, i valori osservati della variabile y formano un sottocampione casuale dei valori già campionati.

1.2.1 Campionamento a due fasi

Assumiamo che dall'intera popolazione U, sia inizialmente estratto un campione s di dimensione $n < N$, secondo lo schema SRSWOR per stimare la media di y. Possiamo vedere il campione s che contiene non risposte come derivante da un capionamento a due fasi. Il campione di prima fase, contiene le n unità estratte inizialmente da U; il campione di seconda fase u, di dimensione $r < n$, può essere visto come un'ulteriore estrazione casuale dalle n unità che porta a selezionare gli r rispondenti.

Siano dunque le y_i, $i = 1, ..., n$ le risposte sulla variabile di studio y nel campione di dimensione n. La media campionaria dei rispondenti ($\overline{y}_r$) è data da

$$\overline{y}_r = \frac{\sum_{i=1}^{n} a_i y_i}{\sum_{i=1}^{n} a_i} \tag{1.1}$$

Si assume che le a_i, $i = 1, ..., n$ siano indipendenti tra loro e che il meccanismo di risposta sia uniforme, ovvero $P(a_i = 1) = p$ con $0 < p < 1$. Per quanto riguarda la varianza dello stimatore, si può dimostrare che la quantità $\overline{y}_r - \overline{Y}$ ha una distribuzione asintoticamente normale con media 0 e varianza $n^{-1}\left(p^{-1} - f\right)\sigma_Y^2$, dove $f = \frac{n}{N}$ rappresenta la frazione di campionamento. La quantità s_y^2 converge in probabilità a σ_Y^2, inoltre uno stimatore consistente della varianza asintotica è dato da $\left(r^{-1} - N^{-1}\right)s_{2y}^2$, dove r è il numero di rispondenti nel campione e s_{2y}^2 è la varianza campionaria dei rispondenti.

Nello stimare parametri di popolazione, come media o totale, è spesso utile ricorrere a informazioni ausiliarie per incrementare l'efficienza degli stimatori: anche nel contesto di uno schema di campionamento a due fasi propriamente utilizzato. Infatti, si sono studiati in letteratura, stimatori in situazioni in cui l'informazione sulla variabile ausiliaria è disponibile o meno per l'intero campione di fase e alcune osservazioni mancano sulla variabile di studio y.

Nel caso in cui $\overline{X}$ non sia nota, assumiamo che l'intera popolazione (indicata con U) sia divisa in due strati: il primo strato (indicato con U_1) di N_1 unità, che rispondono alla prima intervista nella seconda fase; l'altro strato (indicato con U_2) di N_2 unità, che non rispondono alla prima intervista nella

seconda fase ma rispondono alla seconda. Indichiamo il campione di prima e seconda fase con u' e u, rispettivamente: sia $u_1 = u \cap U_1$ e $u_2 = u \cap U_2$. Il sottocampione di u_2 lo indichiamo con u_{2r}.

Quindi, nella prima fase tutte le n' unità di u' forniscono informazioni sulla variabile x: in questa fase, la quantità ignota $\overline{X}$ è sostituita dalla sua stima non distorta basata su un campione numeroso ($n' >> n$). Nella seconda fase, nel campione u di dimensione n, ci sono n_1 unità (di u_1) che forniscono informazioni su y e n_2 (di u_2) non rispondono. Dagli n_2 non rispondenti, si estrae un sottocampione di r unità (di u_{2r}) tali che $r = n_2/k$, con $k > 1$.

Si definiscono in questo contesto le seguenti quantità campionarie che serviranno più tardi: $\overline{y}^* = (n_1/n)\,\overline{y}_1 + (n_2/n)\,\overline{y}_{2r}$, dove $\overline{y}_1 = \sum_{i=1}^{n_1} y_i/n_1$ e $\overline{y}_{2r} = \sum_{i=1}^{n_2} a_i y_i / {}^{n_2}_{i=1}a_i$ sono le medie di y per u_1 e per u_{2r}, rispettivamente. In modo analogo per x, si definiscono gli stimatori $\overline{x}^*$, $\overline{x}_1$, $\overline{x}_{2r}$, $\overline{x}$.

A questo punto, si possono considerare due differenti classi di stimatori sotto due diversi scenari.

Scenario 1. Informazione incompleta sia sulla variabile di studio y che sulla variabile ausiliaria x. Ovvero, considerato solo il campione di seconda fase u, si ha

$$\overbrace{\begin{array}{c} y_1 \cdots\ y_{n-n_2}, \text{missing} \cdots \text{missing} \\ x_1 \cdots\ x_{n-n_2}, \text{missing} \cdots \text{missing} \end{array}}^{u}$$

In questa situazione, uno stimatore per $\overline{Y}$ è

$$t^{(1)}_{(a)d} = \overline{y}^* \exp\left[\omega \left\{ \frac{(\overline{x}^* - \overline{x}')}{(\overline{x}^* + \overline{x}')} \right\}\right], \quad (1.2)$$

dove ω è uno scalare propriamente scelto. Per $\omega = -1, 1$,

$$t_{R1d} = \overline{y}^* \exp\left[\frac{(\overline{x}' - \overline{x}^*)}{(\overline{x}' + \overline{x}^*)}\right], \qquad (1.3)$$

$$t_{P1d} = \overline{y}^* \exp\left[\frac{(\overline{x}^* - \overline{x}')}{(\overline{x}^* + \overline{x}')}\right] \qquad (1.4)$$

la classe di stimatori (1.2) si riduce a (1.3) e (1.4), rispettivamente; la relativa varianza, al primo grado di approssimazione, risulta

$$Var\left(\ t_{R1d}\right) = \lambda^{*}\left[S_{y}^{2} + \left(R^{2}S_{x}^{2}/4\right)\left(1 - 4C\right)\right] + \lambda' S_{y}^{2} + \theta\left[S_{y_2}^{2} + \left(R^{2}S_{x_2}^{2}/4\right)\left(1 - 4C_{(2)}\right)\right] \qquad (1.5)$$

$$Var\left(\ t_{P1d}\right) = \lambda^{*}\left[S_{y}^{2} + \left(R^{2}S_{x}^{2}/4\right)\left(1 + 4C\right)\right] + \lambda' S_{y}^{2} + \theta\left[S_{y_2}^{2} + \left(R^{2}S_{x_2}^{2}/4\right)\left(1 + 4C_{(2)}\right)\right] \qquad (1.6)$$

dove $\lambda = \left(1 - f\right)/n$, $\theta = \left(N_2/N\right)\left(k - 1\right)/n$, $\lambda' = \left(1 - f'\right)/n'$, $f' = n'/N'$,

$\lambda^{*} = \left(\lambda - \lambda'\right)$, $R = \left(\overline{Y}/\overline{X}\right)$, $\beta_{(2)} = \left(S_{xy(2)}/S_{x_2}^{2}\right)$,

$$S_{xy(2)} = \sum_{i=1}^{N_2}\left(x_i - \overline{X}_2\right)\left(y_i - \overline{Y}_2\right)/\left(N_2 - 1\right),$$

$$S_{x_2}^{2} = \sum_{i=1}^{N_2}\left(x_i - \overline{X}_2\right)^{2}/\left(N_2 - 1\right),\ S_{y_2}^{2} = \sum_{i=1}^{N_2}\left(y_i - \overline{Y}_2\right)^{2}/\left(N_2 - 1\right),$$

$$\overline{X}_2 = \sum_{i=1}^{N_2}\left(x_i/N_2\right),\ \overline{Y}_2 = \sum_{i=1}^{N_2}\left(y_i/N_2\right),$$

$C = \left(\beta/R\right)$, $C_{(2)} = \left(\beta_{(2)}/R\right)$.

La scelta dello valore ottimo di ω che garantisce minima varianza dello stimatore (1.2), risulta essere:

$$\omega = -\left(2D^{**}/D^{*}\right) = \omega_0,$$

dove

$$D^{*} = \left[\lambda^{*}S_{x}^{2} + \theta S_{x_2}^{2}\right],$$

$$D^{**} = \left[\lambda^{*}CS_{x}^{2} + \theta C_{(2)}S_{x_2}^{2}\right].$$

Sostituendo ω_0 al posto di ω, nello stimatore (1.2), ottengo lo stimatore ottimo di $\overline{Y}$

$$t_{(\omega_0)d}^{(1)} = \overline{y}^{*}\exp\left[\omega_0\left\{\frac{\left(\overline{x}^{*} - \overline{x}'\right)}{\left(\overline{x}^{*} + \overline{x}'\right)}\right\}\right], \qquad (1.7)$$

con relativa varianza minima

$$\min Var\left(t^{(1)}_{(\omega)d}\right) = Var\left(t^{(1)}_{(\omega_0)d}\right) = \left[\lambda S_y^2 + \theta S_{y_2}^2 - R^2\left(D^{**^2}/D^*\right)\right]. \quad (1.8)$$

In pratica l'ottimo valore ω_0 non è noto, ma si può sostituire con la sua stima consistente $\widehat{\omega}_0 = -\left(2\widehat{D}^{**}/\widehat{D}^*\right)$, dove

$$\widehat{D}^* = \left[\lambda^* \widehat{S}_x^2 + \theta \widehat{S}_{x_2}^2\right],$$

$$\widehat{D}^{**} = \left(\lambda^* \widehat{S}_{xy} + \theta \widehat{S}_{xy(2)}\right)/\widehat{R},$$

$\widehat{R} = (\overline{y}^*/\overline{x})$ e le stime $\widehat{S}_x^2$, $\widehat{S}_{x_2}^2$, $\widehat{S}_{xy}$, $\widehat{S}_{xy(2)}$, sono basate sui dati campionari disponibili. Ora, lo stimatore ottimo $t^{(1)}_{(\widehat{\omega}_0)d}$, con al posto di ω_0 la quantità $\widehat{\omega}_0$, avrà varianza al primo ordine di approssimazione ancora pari alla (1.8).

Confrontando gli stimatori, in termini di varianze asintotiche, si può vedere che $t^{(1)}_{(\omega)d}$ è più efficiente di t_{R1d} se

(i) $-2\left(1+2C\right) < \omega < 2$ e $-2\left(1+2C_{(2)}\right) < \omega < 2$,
oppure $2 < \omega < -2\left(1+2C\right)$ e $2 < \omega < -2\left(1+2C_{(2)}\right)$.

Analogamente, $t^{(1)}_{(\omega)d}$ è più efficiente di t_{P1d} se

(ii) $-\left(1+4C\right) < \omega < 1$ e $-\left(1+4C_{(2)}\right) < \omega < 1$,
oppure $1 < \omega < -\left(1+4C\right)$ e $1 < \omega < -\left(1+4C_{(2)}\right)$.

Scenario 2. Informazione incompleta sulla variabile di studio y e completa informazione sulla variabile ausiliaria x. Ovvero, considerato solo il campione di seconda fase u, si ha

$$\overbrace{\begin{array}{llll} y_1 & \cdots & y_{n-n_2}, \text{missing} & \cdots \text{missing} \\ x_1 & & \cdots & x_n \end{array}}^{u}$$

In questa situazione, la $\overline{Y}$ può essere stimata usando l'informazione sulle $(n_1 + r)$ unità rispondenti per la variabile y e completa informazione per la variabile ausiliaria x dalle n unità campionate .

Stimatore della media di y è

$$t^{(2)}_{(b)d} = \overline{y}^* \exp\left[b\left\{\frac{(\overline{x}-\overline{x}')}{(\overline{x}+\overline{x}')}\right\}\right], \qquad (1.10)$$

dove b è un opportuno scalare scelto. Per $b = -1, 1$

$$t_{R2d} = \overline{y}^* \exp\left[\frac{(\overline{x}'-\overline{x})}{(\overline{x}'+\overline{x})}\right], \qquad (1.11)$$

$$t_{P2d} = \overline{y}^* \exp\left[\frac{(\overline{x}-\overline{x}')}{(\overline{x}+\overline{x}')}\right] \qquad (1.12)$$

lo stimatore (1.10) si riduce agli stimatori (1.11) e (1.12), rispettivamente. In questo caso gli stimatori rapporto e prodotto avranno varianza:

$$Var\,(\ t_{R2d}) = \lambda S_y^2 + \theta S_{y_2}^2 + \lambda^* \left(R^2 S_x^2/4\right)(1 - 4C) \qquad (1.13)$$

$$Var\,(\ t_{P2d}) = \lambda S_y^2 + \theta S_{y_2}^2 + \lambda^* \left(R^2 S_x^2/4\right)(1 + 4C). \qquad (1.14)$$

Il valore ottimo di b, che rende minima la varianza dello stimatore (1.10), è $b = -2C = b_0$: sostituendo al posto di b, la quantità b_0 otteniamo la minima varianza dello stimatore

$$Var\left(\ t^{(2)}_{(b_0)d}\right) = \left[\lambda^* S_y^2 (1 - \rho^2) + \theta S_{y_2}^2 + \lambda' S_y^2\right].$$

Poichè il valore ottimo b_0 in pratica non è noto, possiamo sostituirlo con una sua stima consistente $\widehat{b}_0$ basata sui dati campionari e definire lo stimatore $t^{(2)}_{(\widehat{b}_0)d}$ in questo modo:

$$t^{(2)}_{(\widehat{b}_0)d} = \overline{y}^* \exp\left[\widehat{b}_0 \left\{(\overline{x} - \overline{x}') \,/\, (\overline{x} + \overline{x}')\right\}\right],$$

dove $\widehat{b}_0 = -2\widehat{C}$, $\widehat{C} = \left(\widehat{\beta}/\widehat{R}\right)$. Si può dimostrare che, al primo grado di approssimazione, la varianza risulta ancora essere

$$Var\left(\ t^{(2)}_{(\widehat{b}_0)d}\right) = \left[\lambda^* S_y^2 (1 - \rho^2) + \theta S_{y_2}^2 + \lambda' S_y^2\right]. \qquad (1.15)$$

Un confronto tra gli stimatori porta a concludere che $t^{(2)}_{(b)d}$ è più efficiente di t_{R2d} se

(i) $-1 < b < (1-4C)$ oppure $(1-4C) < b < -1$.

Analogamente, $t^{(2)}_{(b)d}$ è più efficiente di t_{P2d} se

(ii) $-(1+4C) < b < 1$ oppure $1 < b < -(1+4C)$.

Sotto lo stesso disegno di campionamento a due fasi sono stati proposti, in letteratura, altri tipi di stimatori di tipo rapporto, prodotto e regressione: si considera la usuale situazione in cui l'informazione sulla variabile ausiliaria x è completamente disponibile per tutto il campione di n unità di seconda fase; mentre tra le n unità mancano informazioni sulla variabile di studio y.

Definiamo, allora, i seguenti stimatori per la media della variabile di studio y

$$t^{(\alpha_2)}_{(\alpha_1)} = \overline{y}^* \left(\frac{\overline{x}}{\overline{x}^*}\right)^{\alpha_1} \left(\frac{\overline{x}'}{\overline{x}}\right)^{\alpha_2} \tag{1.16}$$

$$t_d = \overline{y}^* + d_1 (\overline{x} - \overline{x}^*) + d_2 (\overline{x}' - \overline{x}), \tag{1.17}$$

dove gli α_i e i d_i $(i = 1, 2)$ sono costanti scelte in modo opportuno. Per $\alpha_1 = 1$, $\alpha_2 = 2$ lo stimatore $t^{(\alpha_2)}_{(\alpha_1)} \to t^{(r)}_d$ mentre per $\alpha_1 = -1$, $\alpha_2 = -2$ lo stimatore $t^{(\alpha_2)}_{(\alpha_1)} \to t^{(p)}_d$

$$t^{(r)}_d = \overline{y}^* \left(\frac{\overline{x}'}{\overline{x}^*}\right) \left(\frac{\overline{x}'}{\overline{x}}\right), \tag{1.18}$$

$$t^{(p)}_d = \overline{y}^* \left(\frac{\overline{x}^*}{\overline{x}'}\right) \left(\frac{\overline{x}}{\overline{x}'}\right), \tag{1.19}$$

La varianza degli stimatori indicati, risulta essere esatta per lo stimatore (1.17) e approssimata (al primo grado di approssimazione) per gli altri stimatori (1.16), (1.18) e (1.19). Rispettivamente, avremo

$$\begin{aligned} Var(t_d) = &\left(\tfrac{1}{n} - \tfrac{1}{n\prime}\right) \left\{S^2_y + d_2 S^2_x (d_2 - 2\beta)\right\} + \\ &\tfrac{(N_2/N)(k-1)}{n} \left\{S^2_{y_2} + d_1 S^2_{x_2} \left(d_1 - 2\beta_{(2)}\right)\right\} + \left(\tfrac{1}{n\prime} - \tfrac{1}{N}\right) S^2_y, \end{aligned} \tag{1.20}$$

$$Var\left(t_d^{(r)}\right) = \left(\tfrac{1}{n} - \tfrac{1}{n\prime}\right)\left\{S_y^2 + 4RS_x^2\left(R - \beta\right)\right\} + \tfrac{(N_2/N)(k-1)}{n}\left\{S_{y_2}^2 + RS_{x_2}^2\left(R - 2\beta_{(2)}\right)\right\} + \left(\tfrac{1}{n\prime} - \tfrac{1}{N}\right)S_y^2, \tag{1.21}$$

$$Var\left(t_d^{(p)}\right) = \left(\tfrac{1}{n} - \tfrac{1}{n\prime}\right)\left\{S_y^2 + 4RS_x^2\left(R + \beta\right)\right\} + \tfrac{(N_2/N)(k-1)}{n}\left\{S_{y_2}^2 + RS_{x_2}^2\left(R + 2\beta_{(2)}\right)\right\} + \left(\tfrac{1}{n\prime} - \tfrac{1}{N}\right)S_y^2, \tag{1.22}$$

$$Var\left(t_{(\alpha_1)}^{(\alpha_2)}\right) = \left(\tfrac{1}{n} - \tfrac{1}{n\prime}\right)\left\{S_y^2 + R\alpha_2 S_x^2\left(R\alpha_2 - 2\beta\right)\right\} + \tfrac{(N_2/N)(k-1)}{n}\left\{S_{y_2}^2 + RS_{x_2}^2\alpha_1\left(R\alpha_1 - 2\beta_{(2)}\right)\right\} + \left(\tfrac{1}{n\prime} - \tfrac{1}{N}\right)S_y^2. \tag{1.23}$$

Con riferimento agli stimatori (1.16) e (1.17), la varianza minima viene raggiunta con specifici valori dei parametri (α_1, α_2) e (d_1, d_2), rispettivamente. Infatti, si può dimostrare che i valori ottimi sono $\left(\beta_{(2)}, \beta\right)$ per t_d e $\left(\beta_{(2)}/R, \beta/R\right)$ per $t_{(\alpha_1)}^{(\alpha_2)}$. Così, sostituendo questi valori all'interno delle formule degli stimatori si ottengono gli stimatori ottimi nelle classi $t_{(\alpha_1)}^{(\alpha_2)}$ e t_d. In formule

$$t_{(d_{10})}^{(d_{20})} = \overline{y}^* + \beta_{(2)}\left(\overline{x} - \overline{x}^*\right) + \beta\left(\overline{x}\prime - \overline{x}\right), \tag{1.24}$$

$$t_{(\alpha_{10})}^{(\alpha_{20})} = \overline{y}^*\left(\tfrac{\overline{x}}{\overline{x}^*}\right)^{\beta_{(2)}/R}\left(\tfrac{\overline{x}\prime}{\overline{x}}\right)^{\beta/R}. \tag{1.25}$$

Si può dimostrare che lo stimatore ottimo (1.24) è non distorto e invece lo stimatore ottimo (1.25) è distorto. L'esatta varianza dello stimatore $t_{(d_{10})}^{(d_{20})}$ è data da

$$Var\left(t_{(d_{10})}^{(d_{20})}\right) = \left(\tfrac{1}{n\prime} - \tfrac{1}{N}\right)S_y^2 + \left(\tfrac{1}{n} - \tfrac{1}{n\prime}\right)S_y^2\left(1 - \rho^2\right) + \tfrac{(N_2/N)(k-1)}{n}S_{y_2}^2\left(1 - \rho_2^2\right), \tag{1.26}$$

che è la varianza minima dello stimatore (1.17), dove $\rho_2 = \left(S_{xy(2)}/S_{y_2}S_{x_2}\right)$ è il coefficiente di correlazione tra y e x nel gruppo di non risposta della popolazione. D'altra parte, usando $(\alpha_{10}, \alpha_{20})$ nello stimatore (1.16), otteniamo ancora la (1.26) come la minima varianza approssimata.

Si può osservare, dalle espressioni degli stimatori ottimi $t_{(d_{10})}^{(d_{20})}$ e $t_{(\alpha_{10})}^{(\alpha_{20})}$, che si possono usare solo se i valori di β, $\beta_{(2)}$ e R sono noti. Queste quantità, possono essere già note con una buona precisione da indagini precedenti: se non sono note in nessun modo, allora è consigliabile usare i loro stimatori consistenti basati sui dati campionari a disposizione. Siano perciò,

$\widehat{\beta} = \frac{s^*_{xy}}{s^{*2}_x}$,

$\widehat{\beta}_{(2)} = \frac{s_{xy(2)}}{s^2_{x_2}}$

$\widehat{R} = \frac{\overline{y}^*}{\overline{x}^*}$,

gli stimatori consistenti di β, $\beta_{(2)}$ e R rispettivamente, dove

$s_{xy(2)} = \frac{1}{(r-1)} \sum_{i=1}^{r} (x_i - \overline{x}'_2)(y_i - \overline{y}'_2)$,

$s^2_{x_2} = \frac{1}{(r-1)} \sum_{i=1}^{r} (x_i - \overline{x}'_2)^2$,

$s^*_{xy} = \frac{1}{(n-1)} (\sum_{i=1}^{n_1} x_i y_i + r \sum_{i=1}^{r} x_i y_i - n\overline{x}\overline{y}^*)$,

$s^{*^2}_x = \frac{1}{(n-1)} (\sum_{i=1}^{n_1} x_i^2 + r \sum_{i=1}^{r} x_i^2 - n\overline{x}\overline{x}^*)$.

In questo modo, usando $\widehat{\beta}$, $\widehat{\beta}_{(2)}$ e $\widehat{R}$ negli stimatori $t^{(d_{20})}_{(d_{10})}$ e $t^{(\alpha_{20})}_{(\alpha_{10})}$ otteniamo gli stimatori consistenti di $\overline{Y}$ di forma

$$t_{lrd} = \overline{y}^* + \widehat{\beta}_{(2)} (\overline{x} - \overline{x}^*) + \widehat{\beta} (\overline{x}' - \overline{x}) \tag{1.28}$$

$$\widehat{t_e} = \overline{y}^* \left(\frac{\overline{x}}{\overline{x}^*}\right)^{\widehat{\beta}_{(2)}/\widehat{R}} \left(\frac{\overline{x}'}{\overline{x}}\right)^{\widehat{\beta}/\widehat{R}}; \tag{1.29}$$

si può dimostrare che, al primo grado di approssimazione,

$Var(t_{lrd}) = Var\left(\widehat{t_e}\right) = Var\left(t^{(d_{20})}_{(d_{10})}\right) = Var\left(t^{(\alpha_{20})}_{(\alpha_{10})}\right)$.

Confrontiamo ora gli stimatori tra di loro, in termini di varianze: si può vedere che

$Var\left(t^{(r)}_d\right) - Var(t) > 0$ a meno che $R = \beta/2$ e $R = \beta_{(2)}$,

$Var\left(t^{(p)}_d\right) - Var(t) > 0$ a meno che $R = -\beta/2$ e $R = -\beta_{(2)}$,

con $t = \left(t_{lrd}\ ,\ \widehat{t}_e\right)$. Sotto queste condizioni, gli stimatori (1.28) e (1.29) sono più efficienti degli stimatori di tipo rapporto e prodotto (1.18) e (1.19), mostrati per primi.

Consideriamo ulteriori stimatori nella situazione in cui, sotto le medesime ipotesi illustrate sopra, l'informazione sulla variabile x sia disponibile per tutte le unità del campione e $\overline{X}$ sia nota, ma vi siano informazioni mancanti per la variabile di studio y.

Gli stimatori hanno forma

$$t_{R4} = \overline{y}^* \left(\frac{\overline{X}}{\overline{x}^*}\right) \left(\frac{\overline{X}}{\overline{x}}\right), \tag{1.30}$$

$$t_{P4} = \overline{y}^* \left(\frac{\overline{x}^*}{\overline{X}}\right) \left(\frac{\overline{x}}{\overline{X}}\right), \tag{1.31}$$

$$t_g = \overline{y}^* \left(\frac{\overline{X}}{\overline{x}^*}\right)^{\alpha_1} \left(\frac{\overline{X}}{\overline{x}}\right)^{\alpha_2} \tag{1.32}$$

$$t_d = \overline{y}^* + d_1 \left(\overline{x} - \overline{x}^*\right) + d_2 \left(\overline{X} - \overline{x}\right), \tag{1.33}$$

dove le α_i $(i = 1, 2)$ e le d_i $(i = 1, 2)$ sono costanti propriamente scelte. La varianza risulta essere esatta per lo stimatore (1.33) e per gli altri stimatori (1.30), (1.31) e (1.32), l'approssimazione di primo ordine risulta

$$\begin{aligned} Var\left(t_d\right) = {} & \left(\tfrac{1-f}{n}\right) \left\{S_y^2 + d_2 S_x^2 \left(d_2 - 2\beta\right)\right\} + \\ & \tfrac{(N_2/N)(k-1)}{n} \left\{S_{y_2}^2 + d_1 S_{x_2}^2 \left(d_1 - 2\beta_{(2)}\right)\right\}, \end{aligned} \tag{1.34}$$

$$\begin{aligned} Var\left(t_{R4}\right) = {} & \left(\tfrac{1-f}{n}\right) \left\{S_y^2 + 4RS_x^2 \left(R - \beta\right)\right\} + \\ & \tfrac{(N_2/N)(k-1)}{n} \left\{S_{y_2}^2 + RS_{x_2}^2 \left(R - 2\beta_{(2)}\right)\right\}, \end{aligned} \tag{1.35}$$

$$\begin{aligned} Var\left(t_{P4}\right) = {} & \left(\tfrac{1-f}{n}\right) \left\{S_y^2 + 4RS_x^2 \left(R + \beta\right)\right\} + \\ & \tfrac{(N_2/N)(k-1)}{n} \left\{S_{y_2}^2 + RS_{x_2}^2 \left(R + 2\beta_{(2)}\right)\right\}, \end{aligned} \tag{1.36}$$

$$\begin{aligned} Var\left(t_g\right) = {} & \left(\tfrac{1-f}{n}\right) \left\{S_y^2 + R\theta S_x^2 \left(\theta - 2\beta\right)\right\} + \\ & \tfrac{(N_2/N)(k-1)}{n} \left\{S_{y_2}^2 + RS_{x_2}^2 \alpha_1 \left(R\alpha_1 - 2\beta_{(2)}\right)\right\}, \end{aligned} \tag{1.37}$$

dove $\theta = \alpha_1 + \alpha_2$.

Minimizzando la varianza (1.34) rispetto a d_1 e d_2, otteniamo i valori ottimi di d_1 e d_2, che sono $d_1 = \beta_{(2)}$ e $d_2 = \beta$. Sostituendo questi valori all'interno dello stimatore (1.33) si ottiene lo stimatore

$$t_d^{(0)} = \overline{y}^* + \beta_{(2)} \left(\overline{x} - \overline{x}^*\right) + \beta \left(\overline{X} - \overline{x}\right), \tag{1.38}$$

che è lo stimatore con minima varianza nella classe degli stimatori t_d, la quale è pari a

$$Var\left(t_d^{(0)}\right) = \left[\left(\tfrac{1-f}{n}\right) S_y^2 \left(1-\rho^2\right) + \tfrac{(N_2/N)(k-1)}{n} S_{y_2}^2 \left(1-\rho_2^2\right)\right] =$$

$$= \; Var_{\min}\left(t_d\right). \tag{1.39}$$

Analogamente, i valori di α_1 e α_2 che portano allo stimatore ottimo $t_g^{(0)}$ con minima varianza nella classe degli stimatori t_g sono pari a

$$\alpha_1 = \beta_{(2)}/R$$

$$\alpha_2 = \left(\beta - \beta_{(2)}\right) R^{-1}.$$

La forma dello stimatore sarà dunque

$$t_g^{(0)} = \overline{y}^* \left(\frac{\overline{X}}{\overline{x}^*}\right)^{\left(\beta_{(2)}/R\right)} \left(\frac{\overline{X}}{\overline{x}}\right)^{\left(\beta-\beta_{(2)}\right)R^{-1}} \tag{1.40}$$

e si può dimostrare che

$$Var\left(t_g^{(0)}\right) = Var_{\min}\left(t_g\right) = Var_{\min}\left(t_d\right) = Var\left(t_d^{(0)}\right).$$

Si noti che, gli stimatori (1.38) e (1.40), si usano nella pratica solo se $\beta, \beta_{(2)}$ e R sono quantità note a priori. Se sono note, lo sono per ricerche passate o attraverso i dati. Se non sono note, una possibilità è stimarle con i dati campionari in possesso del ricercatore. Siano allora, $\widehat{\beta} = s_{xy}^*/s_x^2$, $\widehat{\beta}_{(2)} = s_{xy(2)}/s_{x_2}^2$ e $\widehat{R} = \overline{y}^*/\overline{X}$ gli stimatori consistenti di $\beta, \beta_{(2)}$ e R, rispettivamente. Sostituendo $\widehat{\beta}, \widehat{\beta}_{(2)}$ e $\widehat{R}$ al posto dei rispettivi $\beta, \beta_{(2)}$ e R, otteniamo questi stimatori

$$t_{lr} = \overline{y}^* + \widehat{\beta}_{(2)} \left(\overline{x} - \overline{x}^*\right) + \widehat{\beta} \left(\overline{X} - \overline{x}\right), \tag{1.41}$$

$$\widehat{t}_g^{(0)} = \overline{y}^* \left(\frac{\overline{X}}{\overline{x}^*}\right)^{\left(\widehat{\beta}_{(2)}/\widehat{R}\right)} \left(\frac{\overline{X}}{\overline{x}}\right)^{\left(\widehat{\beta}-\widehat{\beta}_{(2)}\right)\widehat{R}^{-1}}. \qquad (1.42)$$

Al primo grado di approssimazione, si dimostra che

$$var(t_{lr}) = var\left(\widehat{t}_g^{(0)}\right) = V_{\min}(t_d) = var(t_d^{(0)}).$$

Si può raccomandare, in questo caso, l'uso degli stimatori $t_d^{(0)}$,$t_g^{(0)}$,$\widehat{t}_g^{(0)}$ e t_{lr} rispetto a t_{R4} e t_{P4}.

In definitiva, con questi stimatori, si è cercato di affrontare il problema della stima di $\overline{Y}$ attraverso il campionamento a due fasi. Oltre a questo strumento, si è utilizzata anche l'informazione ausiliaria, la quale può contribuire a migliorare l'efficienza degli stimatori della media.

1.2.2 Stimatori che utilizzano tutta l'informazione disponibile

Consideriamo, ora, due tipi di stimatori di $\overline{Y}$ sulla base di un campione casuale formato secondo un qualsiasi disegno di campionamento. Supponiamo che i dati siano mancanti sia per la variabile di studio che per le variabili ausiliarie: definiremo allora stimatori che considerano tutti i dati disponibili per ogni variabile, in contrasto con quelli usuali che 'amputano' le osservazioni incomplete.

Assumiamo che siano disponibili un insieme di $(n-p-q)$ osservazioni complete tra le n unità selezionate. Oltre a queste, sono disponibili le osservazioni sulla variabile ausiliaria x di p unità nel campione, ma le corrispondenti osservazioni sulla variabile y sono mancanti. Analogamente, abbiamo un insieme di q osservazioni sulla caratteristica y nel campione ma i valori associati sulla caratteristica x sono mancanti. Inoltre, p e q sono numeri interi che soddisfano la condizione $p > 0$ e $q > 0$.

Per semplicità, 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, ma y_i mancanti$\}$,
$s_3 = \{i \in s / y_i$ sono disponibili, ma x_i mancanti$\}$.

La struttura dei dati campionari è la quindi la seguente:

$$\overbrace{\begin{matrix} y_i \cdots y_{n-q-p} \\ x_1 \cdots x_{n-q-p} \end{matrix}}^{s_1} \quad \overbrace{\begin{matrix} \text{missing} \cdots \text{missing} \\ x_{n-p-q+1} \cdots x_{n-q} \end{matrix}}^{s_2} \quad \overbrace{\begin{matrix} y_{n-q+1} \cdots y_n \\ \text{missing} \cdots \text{missing} \end{matrix}}^{s_3}$$

Si considera, un disegno campionario generale $d = (S_d, P_d)$, le cui probabilità di inclusione π_i e π_{ij} sono assunte essere strettamente positive. Definiamo i seguenti stimatori di Horvitz-Thompson, relativi ai tre insiemi s_1, s_2, s_3 per le due variabili y e x (1.43)

$$\overline{y}_{HT}^{(1)} = \sum_{i \in s_1} \frac{y_i}{N\pi_i}, \qquad \overline{y}_{HT}^{(3)} = \sum_{i \in s_3} \frac{y_i}{N\pi_i}, \ \overline{x}_{HT}^{(1)} = \sum_{i \in s_1} \frac{x_i}{N\pi_i} \text{ e } \overline{x}_{HT}^{(2)} = \sum_{i \in s_2} \frac{x_i}{N\pi_i}.$$

A questo punto, supposto che sia $\overline{X}$ nota, consideriamo un primo tipo di stimatore alle differenze di $\overline{Y}$

$$\overline{y}_{gd} = \overline{y}_{HT}^{(1,3)} + c\left(\overline{X} - \overline{x}_{HT}^{(1)}\right) + d\left(\overline{X} - \overline{x}_{HT}^{(2)}\right), \tag{1.44}$$

dove $\overline{y}_{HT}^{(1,3)}$ è lo stimatore di Horvitz-Thompson basato sul campione $s_1 \cup s_3$.

I valori di c e d che forniscono la minima varianza dello stimatore (1.44), sono pari a $(c_{opt}, d_{opt})' = \Sigma^{-1}\sigma$, dove

$$\Sigma = \begin{pmatrix} V\left(\overline{x}_{HT}^{(1)}\right) & cov\left(\overline{x}_{HT}^{(1)}, \overline{x}_{HT}^{(2)}\right) \\ cov\left(\overline{x}_{HT}^{(1)}, \overline{x}_{HT}^{(2)}\right) & V\left(\overline{x}_{HT}^{(2)}\right) \end{pmatrix} \text{ e} \tag{1.45}$$

$$\sigma = \left(cov\left(\overline{y}_{HT}^{(1,3)}, \overline{x}_{HT}^{(1)}\right), cov\left(\overline{y}_{HT}^{(1,3)}, \overline{x}_{HT}^{(2)}\right)\right)'. \tag{1.46}$$

La minima varianza avrà espressione:

$$V_{\min}\left(\overline{y}_{gd}\right) = V\left(\overline{y}_{HT}^{(1,3)}\right) - \sigma'\Sigma^{-1}\sigma.$$

Poichè c_{opt} e d_{opt} dipendono dalle caratteristiche della popolazione non nota, lo stimatore ottimo alle differenze non può essere direttamente usato. Usando i valori campionari, possiamo calcolare $\widehat{\Sigma}$ e $\widehat{\sigma}$ dagli stimatori delle

varianze e covarianze degli stimatori di Horvitz-Thompson (i quali esistono perchè abbiamo assunto le π_{ij} strettamente positive $\forall i,j \in U$) e quindi lo stimatore alle differenze assume forma (1.47)

$$\overline{y}^*_{gd} = \overline{y}^{(1,3)}_{HT} + \widehat{c}\left(\overline{X} - \overline{x}^{(1)}_{HT}\right) + \widehat{d}\left(\overline{X} - \overline{x}^{(2)}_{HT}\right), \text{ essendo } \left(\widehat{c}, \widehat{d}\right)' = \widehat{\Sigma}^{-1}\widehat{\sigma}.$$

Nel caso in cui il disegno di campionamento è un SRSWOR, allora gli stimatori di Horvitz-Thompson sono le medie campionarie, e si ha

$$\overline{y}^*_g = \overline{y}^{(1,3)} + \widehat{c}\left(\overline{X} - \overline{x}^{(1)}\right) + \widehat{d}\left(\overline{X} - \overline{x}^{(2)}\right), \tag{1.48}$$

dove

$$\overline{y}^{(1,3)} = \frac{1}{n-p}\sum_{i\in s_1\cup s_3} y_i,\ \overline{x}^{(1)} = \frac{1}{n-p-q}\sum_{i\in s_1} x_i,\ \overline{x}^{(2)} = \frac{1}{p}\sum_{i\in s_2} x_i.$$

In questo caso, possiamo dedurre le espressioni delle varianze e covarianze dello stimatore (1.48). Se consideriamo p e q costanti, le espressioni risultano

$$Var\left(\overline{y}^{(1)}\right) = S^2_y\left(\frac{1}{n-p-q} - \frac{1}{N}\right), \tag{1.49}$$

$$Var\left(\overline{y}^{(1,3)}\right) = S^2_y\left(\frac{1}{n-p} - \frac{1}{N}\right), \tag{1.50}$$

$$Var\left(\overline{x}^{(1)}\right) = S^2_x\left(\frac{1}{n-p-q} - \frac{1}{N}\right), \tag{1.51}$$

$$Var\left(\overline{x}^{(2)}\right) = S^2_x\left(\frac{1}{p} - \frac{1}{N}\right), \tag{1.52}$$

$$Cov\left(\overline{x}^{(1)}, \overline{x}^{(2)}\right) = \begin{cases} \left(\frac{1}{n-p-q} - \frac{1}{N}\right) S^2_x & \text{se } n-p-q \geqslant p, \\ \left(\frac{1}{p} - \frac{1}{N}\right) S^2_x & \text{se } n-p-q < p, \end{cases} \tag{1.53}$$

$$Cov\left(\overline{y}^{(1,3)}, \overline{x}^{(1)}\right) = \left[\frac{1}{n-q} - \frac{1}{N}\right] S_{xy}, \tag{1.54}$$

$$Cov\left(\overline{y}^{(1,3)}, \overline{x}^{(2)}\right) = \left[\frac{1}{n-q} - \frac{1}{N}\right] S_{xy}. \tag{1.55}$$

Le varianze e covarianze possono essere stimate facilmente dal campione, e così otteniamo semplici stimatori per la matrice Σ e il vettore σ. Quindi,

nella stima di $\overline{Y}$ entra anche l'informazione sulla varianza della variabile ausiliaria. Se consideriamo p e q variabili casuali, le corrispondenti espressioni si ottengono con un procedimento simile, sostituendo $\frac{1}{p}, \frac{1}{q}, \frac{1}{n-p}$,... con i loro valori attesi facendo riferimento alle variabili casuali p e q a valori positivi interi.

Si può dimostrare che, sotto campionamento casuale semplice, gli stimatori (1.44) e (1.47) sono asintoticamente non distorti e normali.

Un altro stimatore di $\overline{Y}$,della stessa classe di (1.44), ma di tipo rapporto è

$$\overline{y}_r = \overline{y}_{HT}^{(1,3)} \left(\frac{\overline{X}}{\overline{x}_{HT}^{(1)}} \right)^{\alpha_1} \left(\frac{\overline{X}}{\overline{x}_{HT}^{(2)}} \right)^{\alpha_2}. \tag{1.56}$$

I valori di α_1 e α_2 che minimizzano la varianza dello stimatore sono $(\alpha_1, \alpha_2)'_{opt} = C^{-1}C_0$, dove

$$C = \begin{pmatrix} V\left(\overline{x}_{HT}^{(1)}\right) R^2 & Cov\left(\overline{x}_{HT}^{(1)}, \overline{x}_{HT}^{(2)}\right) R^2 \\ Cov\left(\overline{x}_{HT}^{(1)}, \overline{x}_{HT}^{(2)}\right) R^2 & V\left(\overline{x}_{HT}^{(2)}\right) R^2 \end{pmatrix}, \tag{1.57}$$

$$C_0 = \left(Cov\left(\overline{y}_{HT}^{(1,3)}, \overline{x}_{HT}^{(1)}\right) R, Cov\left(\overline{y}_{HT}^{(1,3)}, \overline{x}_{HT}^{(2)}\right) R \right)', \tag{1.58}$$

$$R = \frac{\overline{y}_{HT}^{(1,3)}}{\overline{X}}.$$

Si può vedere che, le espressioni di questi valori ottimi di solito includono quantità ignote: possiamo allora usare la stessa procedura dello stimatore alle differenze e ottenere il seguente stimatore

$$\overline{y}_r^* = \overline{y}_{HT}^{(1,3)} \left(\frac{\overline{X}}{\overline{x}_{HT}^{(1)}} \right)^{\widehat{\alpha}_1} \left(\frac{\overline{X}}{\overline{x}_{HT}^{(2)}} \right)^{\widehat{\alpha}_2}, \tag{1.59}$$

dove i valori $\widehat{\alpha}_1$ e $\widehat{\alpha}_2$ si ottengono usando i valori campionari. Nel caso di SRSWOR, possiamo usare le espressioni delle varianze e covarianze discusse sopra. In questo caso, lo stimatore alle differenze in (1.44) è più efficiente di quello di tipo rapporto in (1.56), poichè è lo stimatore con varianza minima della classe.

Sotto le medesime ipotesi e struttura del campione, si può considerare uno stimatore basato sulla pseudo-verosimiglianza empirica che può essere

usato con qualsiasi disegno di campionamento a probabilità variabili. Tale stimatore usa tutte le informazioni campionarie su y e x, e ha forma

$$\overline{y}_{PE\alpha} = \alpha \overline{y}_{PE}^{(1)} + (1-\alpha)\, \overline{y}_{w}^{(3)}, \tag{1.60}$$

dove α è una costante opportuna con $0 < \alpha < 1$. Questo stimatore è composto da due stimatori, che sono $\overline{y}_{PE}^{(1)}$ e $\overline{y}_{w}^{(3)}$. Il primo, $\overline{y}_{PE}^{(1)}$, è lo stimatore di massima verosimiglianza di $\overline{Y}$ basato sulla pseudo-verosimglianza empirica, dato dall'espressione

$$\overline{y}_{PE}^{(1)} = \sum_{i \in s_1} \widehat{p}_i^{(1)} y_i, \tag{1.61}$$

dove $\widehat{p}_i^{(1)}$ massimizza $l\left(p^{(1)}\right) = \sum_{i \in s_1} d_i^{(1)} \log p_i^{(1)}$ sotto i vincoli

$$\sum_{i \in s_1} p_i^{(1)} = 1 \quad \left(0 \leqslant \ p_i^{(1)} \leqslant 1\right)$$

$$\sum_{i \in s_1} p_i^{(1)} \left(x_i - \overline{X}\right) = 0.$$

Le quantità $d_i^{(1)}$ valgono $d_i^{(1)} = 1/\pi_i^{(1)}$ (dove $\pi_i^{(j)}$ sono le probabilità di inclusione di primo ordine di s_j, per $j = 1,2,3$) . Considerando il metodo dei moltiplicatori di Lagrange, si può dimostrare che $\widehat{p}_i^{(1)}$ si trovano risolvendo

$\widehat{p}_i^{(1)} = \frac{d_i^{(1)^*}}{1+\lambda\left(x_i - \overline{X}\right)}$ per $i \in s_1$, dove $d_i^{(1)^*} = d_i^{(1)} / \sum_{i \in s_1} d_i^{(1)}$ e

dove il moltiplicatore di Lagrange λ è la soluzione di

$$\sum_{i \in s_1} \frac{d_i^{(1)^*}\left(x_i - \overline{X}\right)}{1+\lambda\left(x_i - \overline{X}\right)} = 0.$$

Si noti che tutte queste espressioni sono definite per $i \in s_1$.
Il secondo stimatore $\overline{y}_{w}^{(3)}$, invece, è definito come

$$\overline{y}_{w}^{(3)} = \sum_{i \in s_3} d_i^{(3)^*} y_i, \tag{1.62}$$

dove

$d_i^{(3)^*} = d_i^{(3)} / \sum_{i \in s_3} d_i^{(3)}$

Uno stimatore della classe $\overline{y}_{PE\alpha}$ usa allora tutte le informazioni disponibili per y da s_1 e s_3 . I valori delle x da s_2 non sono usati nella stima, poichè y è mancante per $i \in s_2$. Questo potrebbe anche peggiorare la stima ma, si può dimostrare che, questo stimatore è tanto efficiente quanto gli stimatori che usano informazioni da ciascun insieme s_1, s_2 e s_3 .

Lo stimatore ottimo della classe degli stimatori del tipo (1.60), si ottiene con il valore ottimale di α che minimizza la varianza asintotica. In questo caso, il valore si può esprimere

$$\alpha_{opt}=\frac{N^*-L^*}{M^*+N^*-2L^*}, \tag{1.63}$$

dove

$$M^* = Var\left(\overline{y}_w^{(1)}\right) + \beta^2 Var\left(\overline{x}_w^{(1)}\right) - 2\beta Cov\left(\overline{x}_w^{(1)}, \overline{y}_w^{(1)}\right),$$

$$N^* = Var\left(\overline{y}_w^{(3)}\right),$$

$$L^* = Cov\left(\overline{y}_w^{(3)}, \overline{y}_w^{(1)}\right) - \beta Cov\left(\overline{x}_w^{(1)}, \overline{y}_w^{(3)}\right).$$

Si ricorda inoltre, che le quantità $\overline{y}_w^{(1)}$ e $\overline{x}_w^{(1)}$ valgono rispettivamente

$\overline{y}_w^{(1)} = \sum_{i\in s_1} d_i^{(1)^*} y_i,$

$\overline{x}_w^{(1)} = \sum_{i\in s_1} d_i^{(1)^*} x_i.$

Sostituendo (1.63) in (1.60), si ottiene lo stimatore $\overline{y}_{PE\alpha_{opt}}$ con varianza asintotica minima, pari a

$$Var(\overline{y}_{PE\alpha_{opt}}) = \alpha_{opt}^2 M^* + (1-\alpha_{opt})^2 N^* + 2\alpha_{opt}(1-\alpha_{opt}) L^*. \tag{1.64}$$

Il valore ottimo di α dipende da parametri di popolazione non noti che possono essere stimati dai dati campionari. Con queste stime, otteniamo una stima $\widetilde{\alpha}_{opt}$ di α_{opt}, che porta al seguente stimatore

$$\widetilde{\overline{y}}_{PE\alpha_{opt}} = \widetilde{\alpha}_{opt}\overline{y}_{PE}^{(1)} + (1-\widetilde{\alpha}_{opt})\,\overline{y}_w^{(3)}. \tag{1.65}$$

Si può dimostrare che $\widetilde{\overline{y}}_{PE\alpha_{opt}}$ è asintoticamente non distorto.

Abbiamo usato il metodo della pseudo-verosimiglianza empirica per stimare la $\overline{Y}$ quando le osservazioni sono mancanti o per la variabile di studio o per la variabile ausiliaria. Il vantaggio pratico del metodo esposto è che ha vaste applicazioni, infatti può essere esteso anche a parametri di popolazione diversi dalla media tipo rapporti, varianze e quantili.

Indichiamo un altro metodo di stima con il relativo stimatore alle differenze, che usa tutte le informazioni disponibili, in cui si assume ancora che le informazioni siano mancanti in modo non simulataneo per la y e x. Supponiamo però, di avere a disposizione due variabili quantitative ausiliarie x e z. Si assumono note le relative medie di popolazione $\overline{X}$ e $\overline{Z}$. Lo scopo è sempre quello di stimare la media di popolazione $\overline{Y}$. Con il campione s, si osservano i valori di tre variabili, (x_i, y_i, z_i), $i = 1, ..., n$.

Assumiamo la seguente struttura di dati mancanti (DM) sulle tre variabili,

$$\overbrace{\begin{matrix} y_1 & \cdots & y_{n-p-q-k} \\ x_1 & \cdots & x_{n-p-q-k} \\ z_1 & \cdots & z_{n-p-q-k} \end{matrix}}^{s_1} \quad \overbrace{\begin{matrix} y_{n-p-q-k+1} & \cdots & y_{n-q-k} \\ x_{n-p-q-k+1} & \cdots & x_{n-q-k} \\ DM & \cdots & DM \end{matrix}}^{s_2}$$

$$\overbrace{\begin{matrix} y_{n-q-k+1} & \cdots & y_{n-k} \\ DM & \cdots & DM \\ z_{n-q-k+1} & \cdots & z_{n-k} \end{matrix}}^{s_3} \quad \overbrace{\begin{matrix} DM & \cdots & DM \\ x_{n-k+1} & \cdots & x_n \\ z_{n-k+1} & \cdots & z_n \end{matrix}}^{s_4}$$

I tradizionali metodi di stima, richiedono le osservazioni complete di (y, x, z). Questo riduce la numerosità del campione da n a $n-p-q-k$ e così aumenta l'errore campionario. Tuttavia, l'informazione ausiliaria disponibile può essere usata per costruire il seguente stimatore alle differenze

$$\overline{y}_{DAI} = \overline{y}_{AI} + B' \begin{pmatrix} \overline{X} - \overline{x}_{AI} \\ \overline{Z} - \overline{z}_{AI} \end{pmatrix}, \tag{1.66}$$

dove $\overline{y}_{AI}$, $\overline{x}_{AI}$, $\overline{z}_{AI}$ sono gli stimatori di Horvitz-Thompson basati sulle relative variabili, cioè

$$\overline{y}_{AI} = \sum_{i \in s_1 \cup s_2 \cup s_3} \frac{y_i}{\pi_i}, \quad \overline{x}_{AI} = \sum_{i \in s_1 \cup s_2 \cup s_4} \frac{x_i}{\pi_i}, \quad \overline{z}_{AI} = \sum_{i \in s_1 \cup s_3 \cup s_4} \frac{z_i}{\pi_i}. \tag{1.67}$$

La minimizzazione della varianza dello stimatore (1.66) rispetto a B porta al valore ottimo $B_{opt} = \Sigma^{-1}\sigma$ con

$$\Sigma = \begin{pmatrix} Var\left(\overline{x}_{AI}\right) & Cov\left(\overline{x}_{AI}, \overline{z}_{AI}\right) \\ Cov\left(\overline{x}_{AI}, \overline{z}_{AI}\right) & Var\left(\overline{z}_{AI}\right) \end{pmatrix} \text{ e } \quad \sigma = \begin{pmatrix} Cov\left(\overline{y}_{AI}, \overline{x}_{AI}\right) \\ Cov\left(\overline{y}_{AI}, \overline{z}_{AI}\right) \end{pmatrix}. \tag{1.68}$$

Si noti che l'espressione dello stimatore $\overline{y}_{DAI}$ dipende dai valori, generalmente ignoti, delle varianze e covarianze degli gli stimatori basati su s_1,s_2,s_3 e s_4. Usando i valori campionari, possiamo calcolare $\widehat{\Sigma}$ e $\widehat{\sigma}$ dagli stimatori delle varianze e covarianze degli stimatori di Horvitz-Thompson: otteniamo lo stimatore così migliorato

$$\overline{y}_{DAI\mathbf{opt}} = \overline{y}_{AI} + \widehat{\Sigma}^{-1}\widehat{\sigma}' \begin{pmatrix} \overline{X} - \overline{x}_{AI} \\ \overline{Z} - \overline{z}_{AI} \end{pmatrix}. \tag{1.69}$$

Nel caso in cui il campione sia stato formato con un disegno del tipo SRSWOR, possiamo indicare questi stimatori (1.70; 1.71; 1.72; 1.73)

$$Var\left(\overline{x}_{AI}\right) = \tfrac{S_x^2}{n-q}\left(1 - \tfrac{n-q}{N}\right),\ Var\left(\overline{z}_{AI}\right) = \tfrac{S_z^2}{n-p}\left(1 - \tfrac{n-p}{N}\right),$$

$$\begin{aligned} Cov\left(\overline{x}_{AI}, \overline{z}_{AI}\right) = {} & \tfrac{(n-p-q)^2}{(n-q)(n-p)} Cov\left(\overline{x}^{(1,4)}, \overline{z}^{(1,4)}\right) + \tfrac{p(n-p-q)}{(n-q)(n-p)} Cov\left(\overline{x}^{(2)}, \overline{z}^{(1,4)}\right) + \\ & \tfrac{q(n-p-q)}{(n-q)(n-p)} Cov\left(\overline{x}^{(1,4)}, \overline{z}^{(3)}\right) + \tfrac{pq}{(n-q)(n-p)} Cov\left(\overline{x}^{(2)}, \overline{z}^{(3)}\right) \end{aligned}$$

$$\begin{aligned} Cov\left(\overline{y}_{AI}, \overline{x}_{AI}\right) = {} & \tfrac{(n-q-k)^2}{(n-k)(n-q)} Cov\left(\overline{y}^{(1,2)}, \overline{x}^{(1,2)}\right) + \tfrac{k(n-q-k)}{(n-k)(n-q)} Cov\left(\overline{y}^{(1,2)}, \overline{x}^{(4)}\right) + \\ & \tfrac{q(n-q-k)}{(n-k)(n-q)} Cov\left(\overline{x}^{(1,2)}, \overline{y}^{(3)}\right) + \tfrac{kq}{(n-q)(n-k)} Cov\left(\overline{x}^{(4)}, \overline{y}^{(3)}\right), \end{aligned}$$

$$\begin{aligned} Cov\left(\overline{y}_{AI}, \overline{z}_{AI}\right) = {} & \tfrac{(n-p-k)^2}{(n-k)(n-p)} Cov\left(\overline{y}^{(1,3)}, \overline{z}^{(1,3)}\right) + \tfrac{k(n-p-k)}{(n-k)(n-p)} Cov\left(\overline{y}^{(1,3)}, \overline{z}^{(4)}\right) + \\ & \tfrac{p(n-p-k)}{(n-k)(n-p)} Cov\left(\overline{z}^{(1,3)}, \overline{y}^{(2)}\right) + \tfrac{kp}{(n-p)(n-k)} Cov\left(\overline{y}^{(2)}, \overline{z}^{(4)}\right). \end{aligned}$$

Anche in questo caso, nella stima di $\overline{Y}$, entrano le varianze delle due variabili ausiliarie. Ricordiamo che le varianze e le covarianze degli stimatori di Horvitz-Thompson possono essere stimate dal campione. Inoltre, il metodo esposto, può essere esteso facilmente al caso di più di due variabili ausiliarie.

Si dimostra che l'uso delle osservazioni incomplete, porta a un miglioramento in efficienza rispetto a quei stimatori di $\overline{Y}$ che usano solo dati completi.

1.3 Dati mancanti MAR

Quando si assume che il meccanismo che genera i dati mancanti sia del tipo MAR, la probabilità di risposta delle unità è indipendente dalla variabile di studio ma dipendente dalla variabile ausiliaria che si osserva sulle unità di interesse. Il pattern che definisce il meccanismo di non risposta, è ricostruibile o prevedibile sulla base delle variabili ausiliarie coinvolte nell'indagine (piuttosto che dalla specifica variabile di studio per la quale mancano alcune osservazioni). Quando i dati mancanti si possono ragionevolmente assumere del tipo MAR, i popolari approcci alla stima della media includono stimatori basati su modelli di regressione per la relazione tra la variabile di studio e le ausiliarie e metodi che utilizzano modelli per la probabilità di risposta $p(x_i)$ che si è osservata, come ad esempio i metodi di ponderazione delle risposte con l'inverso delle probabilità. Questi metodi richiedono, rispettivamente, la corretta specificazione del modello di regressione e/o di quello per la probabilità.

1.3.1 Stimatori doppiamente robusti

In letteratura è stata proposta una classe di stimatori di $\overline{Y}$, basata sulla ponderazione per l'inverso delle probabilità, che implica la modellazione sia della regressione della variabile di studio sulle ausiliarie che della probabilità di risposta. Gli stimatori in questa classe sono definiti doppiamente robusti, nel senso che sono consistenti per $\overline{Y}$ perfino se uno dei due modelli (ma non entrambi insieme) non è correttamente specificato. Se invece, entrambi i modelli non sono correttamente specificati, si può dimostrare che lo stimatore doppiamente robusto risulta decisamente distorto.

Consideriamo n unità estratte a caso da una popolazione: i dati effettivamente osservati sono indipendenti, identicamente distribuiti e consistono in $(a_i, a_i y_i, x_i)$ $(i = 1, ..., n)$, dove $\mathbf{x}$ è un vettore di ausiliarie che covariano con la variabile di studio y. La probabilità di risposta, dipendente dalle ausiliarie, la indicheremo nel seguito con $\phi(\mathbf{x})$. Di solito $\phi(\mathbf{x})$ è ignota e si deve assumere per questa un modello parametrico.

Un primo tipo di stimatore doppiamente robusto per $\overline{Y}$, ha forma

$$\overline{y} = n^{-1} \sum_{i=1}^{n} \left\{ \frac{a_i y_i}{\phi(x_i)} - \frac{a_i - \phi(x_i)}{\phi(x_i)} m\left(x_i, \widehat{\beta}\right) \right\}, \tag{1.74}$$

dove $m\left(\mathbf{x}, \widehat{\beta}\right)$ stima il modello di regressione della variabile risposta sulle ausiliarie, $m\left(\mathbf{x}, \beta\right)$. La stima di β, viene fatta usando le osservazioni dalle unità che rispondono $\{i : a_i = 1\}$; inoltre, il tipo di stima di β può determinare (i) la proprietà di robustezza doppia per lo stimatore $\overline{y}$ e, (ii) se la probabilità di risposta $\phi(\mathbf{x})$ è correttamente specificata, la più piccola varianza asintotica tra gli stimatori della forma (1.74) anche se $m\left(\mathbf{x}, \beta\right)$ non è corretto.

L'uso di tecniche di stima di β familiari, tipo stime ai minimi quadrati ordinari o ai minimi quadrati iterati o stime derivanti dalla minimizzazione della varianza empirica di (1.74), porta a stimatori di $\overline{Y}$ che raggiungono una delle due condizioni (i) o (ii), ma non entrambe. Per soddisfarle simultaneamente, consideriamo $\widehat{\beta}_{opt}$ come soluzione di

$$n^{-1} \sum_{i=1}^{n} \frac{a_i}{\phi(x_i)} \frac{1-\phi(x_i)}{\phi(x_i)} \left\{y_i - m\left(x_i, \beta\right)\right\} m_\beta\left(x_i, \beta\right) \; = \; 0, \tag{1.75}$$

dove $m_\beta\left(\mathbf{x}, \beta\right) = \vartheta / \vartheta\beta \left\{m\left(\mathbf{x}, \beta\right)\right\}$. Questa stima, può essere vista come stima ai minimi quadrati ponderati con pesi $\left\{1 - \phi(x_i)\right\} / \phi^2(x_i)$ e lo stimatore (1.74) con $\widehat{\beta} = \widehat{\beta}_{opt}$ è doppiamente robusto e ha minima varianza asintotica perfino se $m\left(\mathbf{x}, \beta\right)$ non è correttamente specificato.

Nella pratica però, potrebbe essere supposto un modello per la probabilità di risposta parametrico del tipo $\phi(\mathbf{x}, \gamma)$ invece di $\phi(\mathbf{x})$, che non implicava parametri ignoti. In questo caso, non possiamo usare i risultati sopra riportati perchè c'è un effetto di stima di γ che deve essere considerato: infatti, sotto il modello $\phi(\mathbf{x}, \gamma)$ con γ stimato, non è garantita la minima varianza asintotica.

In analogia con (1.75), si può stimare β in questo contesto, risolvendo in (β,c) (1.76)

$$\sum_{i=1}^{n} \frac{a_i}{\phi(x_i,\widehat{\gamma})} \frac{1-\phi(x_i,\widehat{\gamma})}{\phi(x_i,\widehat{\gamma})} \left\{ \frac{\phi_\gamma(x_i,\widehat{\gamma}) m_\beta(x_i,\beta^*)}{1-\phi(x_i,\widehat{\gamma})} \right\} \left\{ y_i - m\left(x_i, \beta\right) - c' \frac{\phi_\gamma(x_i,\widehat{\gamma})}{1-\phi(x_i,\widehat{\gamma})} \right\} = 0$$

dove $\phi_\gamma(\mathbf{x}, \gamma) = \vartheta / \vartheta\gamma \left\{\phi(\mathbf{x}, \gamma)\right\}$, $\widehat{\gamma}$ converge in probabilità a qualche γ^* e β^* è il limite in probabilità di un qualunque stimatore $\widehat{\beta}$. Indicando con $\widehat{\beta}_{2opt}$ la soluzione della (1.76) e, sostituendo la $\widehat{\beta}_{2opt}$ in (1.74) al posto di $\widehat{\beta}$ ottengo uno stimatore per la media di y che è soddisfa entrambe le condizioni (i) e (ii).

Si può dimostrare che, per lo stimatore (1.74) usando al posto di $\widehat{\beta}$ sia $\widehat{\beta}_{opt}$ che $\widehat{\beta}_{2opt}$, la varianza asintotica si approssima con la tecnica di stima di tipo sandwich; il relativo stimatore della varianza sarà consistente, per la vera varianza, anche se uno o entrambi tra i modelli $m(\mathbf{x}, \beta)$ e $\phi(\mathbf{x}, \gamma)$ sono non correttamente specificati.

Vale la pena notare che, se il modello di regressione della risposta è corretto ma il modello per la probabilità di risposta non lo è, il tentativo di migliorare l'efficienza degli stimatori è vano.

Indichiamo altri due stimatori doppiamente robusti, uno ottenuto calibrando i coefficienti in un modello lineare per la probabilità di risposta l'altro considerando un termine aggiuntivo nello stimatore basato sulla verosimiglianza.

Abbiamo, nel primo caso

$$\overline{y}_{LIK2} = n^{-1} \sum_{i=1}^{n} \left\{ \frac{a_i y_i}{\psi\left(x_i, \widetilde{\lambda}_{step2}\right)} \right\}, \qquad (1.77)$$

dove $\psi\left(x_i, \widetilde{\lambda}_{step2}\right)$ è una funzione lineare della probabilità di risposta con coefficienti $\widetilde{\lambda}_{step2} = (\lambda_1\ ,\ \lambda_2)$, stimati con la calibrazione.

Lo stimatore (1.77) è doppiamante robusto, localmente e intrinsecamente efficiente. Diverse scelte di stima di $m(\mathbf{x}, \beta)$, portano a specifiche versioni di $\overline{y}_{LIK2}$: infatti si indicano con $\overline{y}_{LIK2,OLS}$,$\overline{y}_{LIK2,WLS}$ e $\overline{y}_{LIK2,RV}$ le versioni di $\overline{y}_{LIK2}$ corrispondenti a $m\left(\mathbf{x}, \widehat{\beta}_{OLS}\right)$, $m\left(\mathbf{x}, \widehat{\beta}_{WLS}\right)$ e $m\left(\mathbf{x}, \widehat{\beta}_{RV}\right)$.
Le stime per β, sono rispettivamente ai minimi quadrati ordinari, minimi quadrati ponderati con pesi $\phi^{-1}(\mathbf{x}, \widehat{\gamma})$ e minimi quadrati pesati con diversi pesi $\phi^{-1}(\mathbf{x}, \widehat{\gamma})(\phi^{-1}(\mathbf{x}, \widehat{\gamma}) - 1))$. Lo stimatore $\overline{y}_{LIK2,RV}$, diversamente dai due precedenti, è ancora più efficiente localmente.

Il secondo stimatore, basato su un metodo alternativo per rendere doppiamente robusto lo stimatore di Tan (2006), prevede un termine aggiuntivo

$$n^{-1} \sum_{i=1}^{n} \left\{ \frac{a_i}{\psi\left(x_i, \widehat{\lambda}\right)} - 1 \right\} m\left(x_i, \widehat{\beta}\right).$$

Lo stimatore risultante è, appunto, doppiamente robusto oltre che localmente e intrinsecamente efficiente.

Questi stimatori, oltre a essere doppiamente robusti, sono localmente e intrisecamente efficienti: intrinsecamente, perchè se la probabilità di risposta è correttamente specificata, ciascun stimatore è asintoticamente efficiente nella classe degli stimatori che usano la stessa stima del modello $m(\mathbf{x}, \beta)$; localmente, perchè sono almeno asintoticamente efficienti quanto lo stimatore che usa la vera probabilità di risposta e stima in modo ottimale $m(\mathbf{x}, \beta)$. Inoltre lo stimatore $\overline{y}_{LIK2}$ è anche legato al campione, nel senso che vengono escluse le stime che stanno fuori dal range del campione. Riguardo a questo si può dire che, per quanto riguarda lo stimatore (1.77), nessun altro stimatore doppiamente robusto gode di queste quattro proprietà tutte insieme.

1.3.2 L'approccio semiparametrico

L'approccio semiparametrico alla stima di $\overline{Y}$, consiste nel conciliare il metodo di regressione non parametrica con gli approcci model-based. Adottiamo la regressione non parametrica per ridurre la dipendenza dal modello, e usiamo l'informazione a priori su $E(Y|X)$ per migliorare l'efficienza. Poichè le variabili ausiliarie (in questo contesto assumiamo che il vettore delle variabili ausiliarie sia $\mathbf{x} = (x_1, ..., x_d)'$) contengono informazione sulla variabile di studio y, usiamo una funzione parametrica $S = S(\mathbf{x})$ per riassumere questa informazione. La media condizionata di Y dato S è indicata da $m(S) = E(Y|S)$: come vedremo, $m(S)$ sarà stimata con una regressione non parametrica.

Allora, supponiamo che S, sia una funzione continua da $\mathbb{R}^d$ in $\mathbb{R}$ tale che $S = S(\mathbf{x})$ è univariata e $S_i = S(x_i)$. Poichè $E\{m(S)\} = E(Y)$, lo stimatore di $\overline{Y}$ avrà forma

$$\overline{y} = n^{-1} \sum_{i=1}^{n} \widehat{m}(S_i), \qquad (1.78)$$

dove $\widehat{m}(s)$ è una stima di $m(s)$ ottenuta con la regressione non parametrica ($s = S(\mathbf{x})$). Questo metodo di stima (1.78), si dice stima semiparametrica con riduzione della dimensione, ed è consistente per qualsiasi S: inoltre, è consistente quando o la probabilità di risposta è correttamente specificata o vale la relazione $E(Y|X) = l(S)$, per qualche funzione l; l'efficienza ottimale è raggiunta quando entrambe le condizioni sono soddisfatte.

Una stima di $m(s)$, $\widehat{m}(s)$, deriva da una regressione non parametrica pesata con l'inverso delle probabilità di risposta

$$\sum_{j=1}^{n} \frac{a_j}{\phi(x_j)} K_h(S_j - s) \left\{ y_j - \alpha_0 - \alpha_1 (S_j - s) \right\}^2, \qquad (1.79)$$

dove $\alpha = \alpha(s)$ varia con s, $K_h(\cdot) = h^{-1} K(\cdot/h)$ con K funzione Kernel (funzione positiva e continua) e h il parametro di lisciamento; la stima di $m(s)$ ottenuta minimizzando (1.79), è $\widehat{m}(s) = \widehat{\alpha}_0$.

Se nella (1.79) si pone $\alpha_1 = 0$, si ottiene come stima

$$\widehat{m}(s) = \frac{\sum_{j=1}^{n} \frac{a_j}{\phi(x_j)} K_h(S_j - s) y_j}{\sum_{j=1}^{n} \frac{a_j}{\phi(x_j)} K_h(S_j - s)}. \qquad (1.80)$$

Si può dimostrare che, lo stimatore (1.78), ha distribuzione asintotica normale con varianza asintotica pari a

$$Var(\overline{y}) = n^{-1} \left(Var(y) + E\left[\{\phi(\mathbf{x})^{-1} - 1\} Var(Y|S) \right] \right);$$

inoltre, lo stimatore in questione non è molto sensibile alla scelta del parametro di lisciamento h, il che rende la scelta di h non critica.

Quando ϕ è nota, o può essere specificata correttamente, c'è poco interesse riguardo la scelta di S, poichè la (1.78) è consistente per qualunque S. Comunque sia tale stimatore, è il più efficiente se vale la relazione $E(Y|X) = l(S)$. Nel caso in cui, ci sia informazione insufficiente per garantire la corretta specificazione di ϕ, la specificazione di S è cruciale ma non unica: $\overline{y}$ è consistente a patto che $E(Y|X) = l(S)$. Perciò, la miglior scelta di S è quella che soddisfa tale condizione. Inoltre S, non ha bisogno di essere completamente specificata poichè la regressione non parametrica di $m(S)$ non ha bisogno di l.

Per questo metodo di stima, l'uso della regressione non parametrica porta alla robustezza dello stimatore anche in assenza di una corretta specificazione del modello ed elimina il problema della dimensionalità del vettore delle ausiliarie.

1.3.3 Metodi di aggiustamento per ponderazione

Numerose altre procedure sono disponibili per compensare la mancanza di dati: tra queste, gli aggiustamenti per ponderazione. Essi consistono nell'aumentare i pesi delle unità che rispondono, in modo da compensare le mancate risposte delle altre unità. In particolare, moltiplicando l'inverso della probabilità di risposta per il peso di campionamento del rispondente, si riesce a ridurre la distorsione dovuta alla non risposta. Poichè la vera probabilità di risposta è di solito non nota, si stima: in questo caso, il metodo viene chiamato 'NWA diretto' e ha la forma (con N nota)

$$\overline{y}_e = \frac{1}{N} \sum_{i \in s} \frac{1}{\pi_i} \frac{a_i}{\widehat{\phi}_{i|s}} y_i, \qquad (1.81)$$

dove $\widehat{\phi}_{i|s} = \widehat{P}(a_i = 1 | i \in s)$ è la stima della probabilità di risposta dell'unità i-esima, che entra nel campione s.

Lo stimatore (1.81) è chiamato stimatore NWA diretto, perchè entra nello stimatore una probabilità di risposta stimata e non viene usata alcuna variabile ausiliaria.

La stima della varianza dello stimatore $\overline{y}_e$, è $\widehat{Var}(\overline{y}_e) = \widehat{Var}_{e1} + \widehat{Var}_{e2}$, dove

$$\widehat{Var}_{e1} = \sum_{i \in s} \sum_{j \in s} \Omega_{ij} \widehat{\eta}_i \widehat{\eta}_j, \qquad (1.82)$$

con Ω_{ij} coefficienti e

$$\widehat{\eta}_i = k_i \pi_i \widehat{\phi}_i \widehat{h}_i' \widehat{\gamma}_n + \frac{a_i}{\widehat{\phi}_i} (y_i - k_i \pi_i \widehat{\phi}_i \widehat{h}_i' \widehat{\gamma}_n),$$

$$\widehat{\gamma}_n = \left\{ \sum_{i \in s_r} k_i (1 - \widehat{\phi}_i) \widehat{h}_i \widehat{h}_i' \right\}^{-1} \sum_{i \in s_r} \pi_i^{-1} (\widehat{\phi}_i^{-1} - 1) \widehat{h}_i y_i,$$

$$s_r = \{ i \in s : a_i = 1 \},$$

$$\widehat{h}_i = \vartheta / \vartheta \gamma \left\{ \log it \left(\phi(z_i, \widehat{\gamma}) \right) \right\},$$

supponendo che la probabilità di risposta sia una funzione che dipenda dalla variabile ausiliaria z_i con parametro γ, che di solito viene stimato. Il

valore k_i, rappresenta il peso dell'unità i-esima nell'equazione di stima per γ: infatti, per $k_i = 1$ la stima di γ, è l'usuale stima di massima verosimiglianza. La seconda componente è

$$\widehat{Var}_{e2} = \tfrac{1}{N^2} \sum_{i \in s_r} \pi_i^{-1} \widehat{\phi}_i^{-2} (1 - \widehat{\phi}_i)(y_i - k_i \pi_i \widehat{\phi}_i \widehat{h}_i' \widehat{\gamma}_n)^2. \qquad (1.83)$$

Quando la frazione di campionamento (n/N) è molto piccola, $\widehat{Var}_{e2}$ è di un ordine più piccolo di $\widehat{Var}_{e1}$, e così può essere trascurata. Questo stimatore della varianza di (1.81), è non distorto.

Quando sono disponibili delle variabili ausiliarie per tutto il campione, correlate con la variabile di studio, si può migliorare il metodo di NWA diretto e in questa situazione si parla di ponderazione per regressione. A tal fine, supponiamo che ci sia una variabile ausiliaria x_i, correlata alla variabile di studio y_i. Lo stimatore per regressione ponderato per dati mancanti, avrà forma

$$\overline{y}_{re} = \overline{y}_e + (\overline{x}_n - \overline{x}_e)' \, \widehat{\beta}_e, \qquad (1.84)$$

dove

$$\overline{x}_n = \tfrac{1}{N} \sum_{i \in s} \pi_i^{-1} x_i, \; \overline{x}_e = \tfrac{1}{N} \sum_{i \in s} \pi_i^{-1} \widehat{\phi}_i^{-1} a_i x_i,$$

$$\widehat{\beta}_e = (\sum_{i \in s} \pi_i^{-1} \widehat{\phi}_i^{-1} a_i x_i x_i')^{-1} \sum_{i \in s} \pi_i^{-1} \widehat{\phi}_i^{-1} a_i x_i y_i.$$

Lo stimatore per regressione NWA riesce a migliorare l'efficienza riferita allo stimatore diretto NWA in modo significativo, se la variabile di studio y_i è bene approssimata da una combinazione lineare di x_i. Lo stimatore (1.84) è approssimativamente non distorto e, ha varianza

$$Var(\overline{y}_{re}) = E \left\{ \sum_{i \in s} \Omega_{ii} y_i^2 + \sum_{i \neq j} \sum_{i,j \in s} \Omega_{ij} y_i y_j \right\} + \qquad (1.85)$$
$$N^{-2} E \left\{ \sum_{i \in s} \tfrac{1}{\pi_i^2} \tfrac{(1-\phi_i)}{\phi_i} (y_i - x_i' \beta_N - k_i \pi_i \phi_i h_{i0}' \alpha_N)^2 \right\},$$

dove

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

h_{i0} è il valore di $h_i(\gamma)$ calcolato in $\gamma = \gamma_s^0$ (vero valore del parametro). Inoltre,

$$\alpha_N = \left\{\sum_{i=1}^{N} k_i \pi_i \phi_i (1-\phi_i) h_{i0} h_{i0}'\right\}^{-1} \sum_{i=1}^{N} (1-\phi_i) h_{i0} (y_i - x_i' \beta_N);$$

quindi, dato β_N, la scelta di α_N minimizza la varianza in (1.85) dello stimatore per regressione NWA.

Lo stimatore NWA diretto (1.81), è meno interessante per l'uso pratico, perchè i pesi non sommano a 1 e può essere molto instabile quando i $\widehat{\phi}_i$ sono vicini a zero. Uno stimatore più utile è

$$\overline{y}_{e2} = \frac{\sum_{i\in s} \pi_i^{-1} \widehat{\phi}_i^{-1} a_i y_i}{\sum_{i\in s} \pi_i^{-1} \widehat{\phi}_i^{-1} a_i}, \qquad (1.86)$$

ossia lo stimatore di Hàyek applicato alla stima NWA. La stima della varianza dello stimatore (1.86), si può derivare usando i risultati relativi allo stimatore per regressione NWA.

Negli stimatori del tipo NWA fin qui considerati, la probabilità di risposta è stimata usando la massima verosimiglianza. Questi stimatori sono efficienti e hanno una ridotta distorsione anche grazie l'uso di informazione ausiliaria che entra nel modello della probabilità della risposta.

Ci sono altri metodi per stimare le probabilità di risposta, in particolare si può stimarle con metodi non parametrici: questi richiedono solo che, le probabilità in questione, siano collegate alle variabili ausiliarie da una funzione lisciata ma non specificata. Per esempio, le probabilità di risposta possono essere stimate con tecniche di lisciamento di tipo Kernel.

Riconsiderando gli stimatori del tipo

$$\overline{y}_{\pi\widehat{\phi}} = \frac{1}{N} \sum_{i\in s_r} \frac{\widehat{\phi}_i^{-1} y_i}{\pi_i} \qquad (1.87)$$

$$\overline{y}_{rat,\widehat{\phi}} = \frac{\sum_{i\in s_r} \pi_i^{-1} \widehat{\phi}_i^{-1} y_i}{\sum_{i\in s_r} \pi_i^{-1} \widehat{\phi}_i^{-1}}; \qquad (1.88)$$

in questo caso, si può pensare di stimare $\phi(x_i)$ usando lo stimatore per regressione di tipo Kernel: questo metodo ha il pregio che le probabilità di risposta non devono essere necessariamente uguali per tutte le osservazioni e che la forma della funzione $\phi(\cdot)$ non deve essere specificata. Lo stimatore che consideriamo è definito nel modo seguente

$$\widehat{\phi}_i = \frac{\sum_{i,j\in s} a_j(x_j - x_i)K_h\pi_j^{-1}}{\sum_{i,j\in s}(x_j - x_i)K_h\pi_j^{-1}}, \qquad (1.89)$$

dove $K_h(\cdot) = h^{-1}K(\cdot/h)$, con $K(\cdot)$ funzione Kernel e h un parametro di lisciamento. Questo stimatore può essere calcolato purchè le x_i siano osservate per tutto il campione. Evidentemente, questa procedura permette di stimare ϕ_i, $i = 1, ..., N$, senza dover specificare una forma parametrica per la probabilità di risposta; inoltre, purchè $\phi(\cdot)$ sia una funzione continua e lisciata, lo stimatore di tipo kernel (1.89) può essere usato per correggere lo stimatore della media in caso di dati mancanti.

Questo stimatore è consistente per $\overline{Y}$: si può dimostrare che la tecnica di aggiustamento per regressione di Kernel, avente pesi finali che sommano a uno, è il modo più sicuro per ridurre la distorsione e la varianza rispetto a uno stimatore aggiustato che usa semplicemente probabilità di risposta stimate.

Poichè le varianze asintotiche degli stimatori (1.87) e (1.88), usando (1.89), sono troppo scomode da stimare si può usare per esempio, il metodo jack-knife, si ottiene così una buona stima della varianza dello stimatore corretto di tipo Kernel.

In generale, per questo tipo di stimatori basati su aggiustamenti non parametrici, un punto fondamentale è la scelta del grado del polinomio e del parametro di lisciamento. Dalla letteratura è noto che, polinomi di grado elevato, riducono la distorsione ma aumentano la varianza dello stimatore così sono consigliati polinomi di grado 1 o 2, come buon compromesso. Per quanto riguarda la scelta del parametro di lisciamento h, una regola è quella di considerare valori per h che sono tra il 20% e il 50% del range della variabile x.

Un'altra strada per stimare in modo non parametrico la probabilità di risposta $\phi(x_i)$, è attraverso una regressione polinomiale locale: questo metodo, comparato al lisciamento di Kernel, migliora l'approssimazione locale.

Tra l'altro, è molto più diffuso come metodo di lisciamento nella pratica, con applicazioni disponibili nella maggior parte dei programmi di statistica.

La procedura che mostriamo di seguito, detta regressione polinomiale locale, può essere descritta come segue: sia $K(\cdot)$ funzione Kernel e h il suo parametro. Definiamo la matrice di dimensione $n \times (g+1)$ (si suppone che la variabile ausiliaria x sia disponibile per tutta il campione)

$$X_s = \begin{bmatrix} 1 & (x_1 - x_i) & \cdots & (x_1 - x_i)^g \\ \vdots & \vdots & & \vdots \\ 1 & (x_n - x_i) & \cdots & (x_n - x_i)^g \end{bmatrix},$$

e la matrice $n \times n$

$W_s = diag\left\{\frac{1}{\pi_j h} K\left(\frac{x_j - x_i}{h}\right) : j \in s\right\}$

$A_s = (a_j : j \in s)'$,

allora uno stimatore di regressione polinomiale locale di grado g, della probabilità $\phi(x_i)$ è dato da

$$\widehat{\phi}_i^0 = e_1' \widehat{T}_s^{-1} \widehat{t}_s. \qquad (1.90)$$

dove e_j rappresenta la j-esima colonna della matrice identità di ordine $g+1$, e

$$\left(\widehat{T}_s, \widehat{t}_s\right) = (X_s' W_s X_s, X_s' W_s A_s),$$

(con $\widehat{T}_s$ invertibile). Un caso speciale della (1.90), si ottiene considerando $g = 0$, che corrisponde allo stimatore (1.89) visto in precedenza.

Il caso in cui la matrice $\widehat{T}_s$ non sia invertibile costituisce un problema, che è superato dallo stimatore modificato

$$\widehat{\phi}(x_i, g, h) = e_1' \left(\widehat{T}_s + diag\left\{\frac{\delta_1}{N}\right\}\right)^{-1} \widehat{t}_s,\ i \in s; \qquad (1.91)$$

Con δ_1, una opportuna costante positiva piccola. I termini di ordine δ_1/N aggiunti alla diagonale principale di $\widehat{T}_s$ sono sufficienti per garantire l'invertibilità della risultante matrice per ogni h. Di conseguenza, $\widehat{\phi}(x_i, g, h)$ sarà ben definita per ogni $i \in s$.

Comunque, un'altra difficoltà che si può incontrare è la possibilità che (1.91) vada ad assumere valori vicini a zero: per ovviare a questo problema, limitiamo $\widehat{\phi}(x_i, g, h)$ lontano da zero considerando lo stimatore

$$\widehat{\phi}_i = \max\left\{\widehat{\phi}(x_i, g, h), \delta_2(Nh)^{-1}\right\}, \qquad (1.92)$$

per qualche costante $\delta_2 > 0$.

Anche per il metodo della regressione polinomiale locale, la stima diretta della varianza degli stimatori (1.87) e (1.88) risulta impraticabile quindi la soluzione adottata è la stessa del metodo di Kernel: si fa ricorso a metodi tipo 'jackknife'.

Consideriamo un altro tipo di metodo di aggiustamento per ponderazione: esso consiste nel classificare rispondenti e non rispondenti in celle di aggiustamento secondo l'informazione ausiliaria disponibile. Il meccanismo è il seguente: dato un campione di n unità estratte attraverso un campione casuale semplice, si suddividono rispondenti e non rispondenti in C celle di aggiustamento basate sulla variabile ausiliaria x. Sia n_{ac} il numero degli individui campionati con $a = a_i = 0, 1$; $x = c = 1, ..., C$; $n_{+c} = n_{1c} + n_{0c}$ il numero delle unità campionate presenti nella cella c; $n_0 = \sum_{c=1}^{C} n_{0c}$ e $n_1 = \sum_{c=1}^{C} n_{1c}$ il numero totale di non rispondenti e rispondenti; infine $p_c = n_{+c}/n$ e $p_{1c} = n_{1c}/n_1$ le proporzioni degli elementi campionati e dei rispondenti nella cella c. Consideriamo come stimatore di $\overline{Y}$, la media pesata

$$\overline{y}_w = \sum_{c=1}^{C} p_c \overline{y}_{1c} = \sum_{c=1}^{C} w_c p_{1c} \overline{y}_{1c}, \qquad (1.93)$$

che pesa i rispondenti nella cella c con l'inverso del tasso di risposta $w_c = p_c/p_{1c}$. A questo punto, consideriamo la varianza dello stimatore (1.93), ipotizzando il seguente modello: supponiamo che, condizionatamente al campione di dimensione n, le unità campionate abbiano distribuzione multinomiale sulla tabella di contingenza $(C \times 2)$ basata sulla classificazione di a_i e x. Le probabilità di cella sono $P(a = 1, x = c) = \phi\pi_{1c}$; $P(a = 0, x = c) = (1 - \phi)\pi_{0c}$, dove $\phi = P(a = 1)$ è la probabilità marginale di risposta. La distribuzione condizionata di x, dato $a = 1$ e n_1 è multinomiale con probabilità di cella pari a $P(x = c|a = 1) = \pi_{1c}$; e la distribuzione marginale di x, dato n, è multinomiale con probabilità di cella

pari a $P(x = c) = \phi\pi_{1c} + (1 - \phi)\pi_{0c} = \pi_c$. Assumiamo anche che, la distribuzione condizionata di y dato $a = a_i$, $x = c$ abbia media μ_{ac} e varianza costante pari a σ^2. La media di y per rispondenti e non rispondenti sarà

$$\overline{y}_1 = \sum_{c=1}^{C} \pi_{1c}\mu_{1c} \qquad \overline{y}_0 = \sum_{c=1}^{C} \pi_{0c}\mu_{0c},$$

rispettivamente, e la media totale di y è $\overline{y} = \phi\mu_1 + (1 - \phi)\mu_0$. Indicando con $\widetilde{\mu}_1 = \sum_{c=1}^{C} \pi_c\mu_{1c}$ la media dei rispondenti corretta, la varianza di (1.93) è

$$Var(\overline{y}_w) = (1 + \lambda)\,\sigma^2/n_1 + \sum_{c=1}^{C} \pi_c\,(\mu_{1c} - \widetilde{\mu}_1)^2\,/n, \tag{1.94}$$

dove

$\lambda = \sum_{c=1}^{C} \pi_{1c}\left((\pi_c/\pi_{1c} - 1)^2\right)$.

Nel nostro contesto di dati mancanti MAR, si può dimostrare che $\widetilde{\mu}_1 = \overline{y}$.

Si deve notare che, la ponderazione, è efficiente solo per la variabile di studio y che è associata alla variabile di aggiustamento di cella x, altrimenti incrementa solo la varianza senza una riduzione compensativa della distorsione. Inoltre, per una variabile y che è associata alla variabile x, la ponderazione aumenta la precisione e diminuisce la anche distorsione se la variabile di aggiustamento di cella è legata alla mancata risposta.

1.4 Conclusioni

Le questioni salienti presentate in questo capitolo riguardano la stima di $\overline{Y}$, nel caso in cui il meccanismo che genera i dati mancanti sia del tipo MAR e MCAR. Nel caso di dati mancanti del tipo MCAR, due filoni principali si contraddistinguono tra le tecniche di stima: il primo basato sull'utilizzo del campionamento a due stadi, permette di stimare la media in situazioni in cui l'informazione sulla variabile ausiliaria è completamente disponibile per l'intero campione di seconda fase e nella stessa fase alcune unità non

rispondono per la y. Nel campione di prima fase, si recupera una stima consistente della media della variabile ausiliaria (se non è gia nota) che entrerà nello stimatore finale.

La seconda tecnica di stima, si basa sul fatto che l'uso anche delle osservazioni incomplete porta a un incremento di efficienza rispetto a quegli stimatori che usano solo osservazioni complete. Sono stati così definiti stimatori che considerano tutti i valori disponibili per ogni variabile: la prima classe di stimatori risulta funzione di stimatori tipo Horvitz-Thompson; la seconda classe deriva da un metodo di stima che usa la pseudo-verosimiglianza empirica; la terza è ancora funzione di altri stimatori (ancora Horvitz-Thompson), però usa due variabili ausiliarie invece di una.

Se i dati mancanti seguono un meccanismo di tipo MAR, sono stati presentati tre gruppi di stimatori accomunati da caratteristiche rilevanti ai fini del metodo di stima: stimatori doppiamente robusti, stimatori basati su un approccio semiparametrico e stimatori che si basano su aggiustamenti per ponderazione.

Gli stimatori doppiamente robusti, pur basandosi sulla ponderazione per l'inverso delle probabilità, implicano la modellazione sia della regressione della variabile di studio sulle ausiliarie che della probabilità di risposta. Essi hanno la caratteristica di essere doppiamente robusti, nel senso che sono stimatori consistenti di $\overline{Y}$ anche quando uno dei due modelli non è correttamente specificato. Invece, nei metodi più tradizionali di aggiustamento della stima della media per ponderazione, si modella solo la probabilità di risposta in modo parametrico o non parametrico; in questo modo le unità sono ponderate con l'inverso della loro probabilità di inclusione, stimata come il prodotto della probabilità di selezione per la probabilità di risposta. In questo ambito, vi è anche la possibilità di utilizzare l'informazione della variabile ausiliaria per creare delle celle di aggiustamento; i pesi poi vengono calcolati in modo proporzionale all'inverso del tasso di risposta nella cella.

Gli stimatori che si basano su un approccio semiparametrico, conciliano il metodo di regressione non parametrica con gli approcci parametrici. Vi è la condensazione dell'informazione contenuta nelle variabili ausiliarie attraverso una funzione parametrica, così da ridurre la dimensione per la regressione non parametrica successiva. In questo modo, la robustezza dello stimatore è garantita e si elimina il problema della dimensionalità del vettore delle

ausiliarie. Questo tipo di stima semiparametrica viene usata soprattutto negli studi con un elevato numero di variabili ausiliarie.

1.5 Nota bibliografica

In questo capitolo, si sono visti vari metodi per stimare la media di una variabile di studio y nel caso di dati mancanti: nessun metodo di imputazione, per dati mancanti, è stato preso in considerazione nella stima. Abbiamo suddiviso gli stimatori tra quelli che assumono un meccanismo di dati mancanti di tipo MCAR e quelli che assumono un meccanismo MAR.

Per quanto riguarda i dati mancanti di tipo MCAR, in letteratura si può fare riferimento anzitutto a un primo gruppo di articoli che trattano l'uso del campionamento a due fasi per il metodo di stima: si può approfondire nelle pubblicazioni di Chen e Rao (2007); Singh e Kumar (2008); Singh, Kumar e Kozak (2010). Sempre nel contesto di dati mancanti del tipo MCAR, abbiamo presentato stimatori che utilizzano tutti i valori disponibili per ciascuna variabile: in tal caso si può fare riferimento agli articoli di Rueda, Gonzàlez e Arcos (2006); all'articolo di Rueda, Muñoz, Berger, Arcos e Martínez (2007) se lo stimatore è basato sulla pseudo-verosimiglianza empirica; alla pubblicazione di González, Rueda, Arcos (2008) quando si hanno a disposizione due variabili quantitative ausiliarie.

Per i dati mancanti di tipo MAR, si può fare riferimento agli stimatori doppiamente robusti: per questo tipo di stimatori, in letteratura si possono consultare gli articoli di Cao, Tsiatis e Davidian (2009) oppure di Tan (2010). Se si segue, invece, un approccio semiparametrico alla stima, si può fare riferimento all'articolo di Hu, Follmann e Qin (2010). Inoltre, se si considerano i metodi di aggiustamento per ponderazione: con stima parametrica della probabilità di risposta si può fare riferimento all'articolo di Kim e Kim (2007); se si utilizza un metodo di stima della probabilità di tipo non parametrico, utile approfondimento si può avere nell'articolo di Silva e Opsomer (2006) e (2009). Infine, il metodo di aggiustamento per ponderazione che usa celle di aggiustamento è reperibile nell'articolo di Little e Vartivarian (2006).

Capitolo 2

Stimatori della media con metodi d'imputazione

2.1 Introduzione

Nel seguente capitolo, si presentano tecniche di stima di $\overline{Y}$ in presenza di dati mancanti: sia nel contesto di dati mancanti MCAR sia MAR. Queste tecniche sono accomunate dal fatto che utilizzano metodi d'imputazione per trovare 'sostituti' delle osservazioni mancanti. Trattando questi valori imputati come vere osservazioni, si possono condurre le analisi statistiche usando anche le procedure standard sviluppate per dati senza osservazioni mancanti. Di seguito, illustriamo i vari stimatori proposti in letteratura negli ultimi cinque anni, ponendo l'attenzione sul metodo di stima, sulla varianza dello stimatore e sulle proprietà rilevanti ai fini dell'utilizzo pratico.

2.2 Dati mancanti MCAR

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

2.2.1 Imputazione con media, rapporto, differenze e tramite regressione

Consideriamo, inizialmente, stimatori per $\overline{Y}$ che utilizzano tecniche di imputazione dei dati mancanti per media o rapporto. Assumiamo che le osservazioni siano mancanti sulla variabile di studio y, mentre tutte le osservazioni sulla variabile ausiliaria x sono disponibili. E' noto dalla letteratura che, con il metodo di imputazione basato sulla media, la stima di $\overline{Y}$ è

$$\overline{y}_s = \frac{1}{r}\sum_{i=1}^{r} y_i, \qquad (2.1)$$

con varianza pari a

$$Var(\overline{y}_s) = (\tfrac{1}{r} - \tfrac{1}{N})S_y^2, \qquad (2.2)$$

dove r è il numero delle unità rispondenti fra le n unità campionate. Sappiamo che (2.1) è uno stimatore non distorto di $\overline{Y}$. Analogamente, con il metodo di imputazione tramite rapporto, è noto lo stimatore di tipo rapporto di $\overline{Y}$

$$\overline{y}_r = \frac{\overline{y}_s}{\overline{x}_s}\overline{x}, \qquad (2.3)$$

con MSE (mean squared error) uguale a

$$MSE(\overline{y}_r) \cong \left(\tfrac{1}{r} - \tfrac{1}{N}\right) S_y^2 + \left(\tfrac{1}{r} - \tfrac{1}{n}\right)(S_x^2 R^2 - 2RS_{xy}). \qquad (2.4)$$

La $\overline{x}_s$,

$$\overline{x}_s = \frac{1}{r}\sum_{i=1}^{r} x_i, \qquad (2.5)$$

è la media campionaria delle r osservazioni di x. Confrontando (2.2) e (2.4), possiamo facilmente vedere che il metodo di imputazione tramite rapporto è più efficiente del metodo di imputazione per media se vale

$R < 2\beta$, per $R > 0$

$R > 2\beta$, per $R < 0$.

Kadilar and Cingi (2004) proposero il seguente stimatore di $\overline{Y}$, sotto campionamento casuale semplice

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

dove $\widehat{\beta}$ è il coefficiente di regressione tra y e x stimato sul campione completo, calcolato col metodo dei minimi quadrati; si noti che $\overline{X}$ è assunta nota. Lo stimatore (2.6), può essere modificato per tener conto del metodo di imputazione tramite rapporto e del metodo di imputazione per la media. Si ottiene così

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

che ha MSE pari a

$$MSE(\overline{y}_{pr1}) \cong \left(\tfrac{1}{r} - \tfrac{1}{N}\right) S_y^2 + \left(\tfrac{1}{n} - \tfrac{1}{N}\right) S_x^2(R^2 - \beta^2). \qquad (2.8)$$

Citiamo un secondo stimatore, di forma

$$\overline{y}_{pr2} = \frac{\overline{y}_s+\widehat{\beta}\left(\overline{X}-\overline{x}_s\right)}{\overline{x}_s}\overline{X}, \qquad (2.9)$$

con relativo MSE pari a

$$MSE(\overline{y}_{pr2}) \cong \left(\tfrac{1}{r} - \tfrac{1}{N}\right) \left(S_y^2 - \beta S_{xy} + R^2 S_x^2\right). \qquad (2.10)$$

Se $\overline{X}$ è ignota, indichiamo un terzo stimatore

$$\overline{y}_{pr3} = \frac{\overline{y}_s+\widehat{\beta}(\overline{x}-\overline{x}_s)}{\overline{x}_s}\overline{x}, \qquad (2.11)$$

con MSE pari a

$$MSE(\overline{y}_{pr3}) \cong \left(\tfrac{1}{r} - \tfrac{1}{N}\right) S_y^2 + \left(\tfrac{1}{r} - \tfrac{1}{n}\right) [S_x^2(R+\beta)^2 - 2(R+\beta)S_{xy}]. \qquad (2.12)$$

Facendo un confronto tra gli MSE degli stimatori fin qui enunciati, si può dimostrare che tutti e tre gli stimatori (2.7), (2.9) e (2.11) sono più efficienti di (2.1) se vale la condizione

$$R^2 < \beta^2; \qquad (2.13)$$

si dimostra anche che lo stimatore in (2.7), è più efficiente dello stimatore di tipo rapporto in (2.3), quando vale

$$\left(\tfrac{1}{n} - \tfrac{1}{N}\right) S_x^2(R^2 - \beta^2) - \left(\tfrac{1}{r} - \tfrac{1}{n}\right)(R^2 S_x^2 - 2RS_{xy}) < 0. \qquad (2.14)$$

Il secondo stimatore in (2.9), è più efficiente dello stimatore di tipo rapporto in (2.3), quando vale la seguente relazione

$$\left(\tfrac{1}{r} - \tfrac{1}{N}\right) S_x^2(R^2 - \beta^2) - \left(\tfrac{1}{r} - \tfrac{1}{n}\right)(R^2 S_x^2 - 2RS_{xy}) < 0. \qquad (2.15)$$

Il terzo stimatore in (2.11), è più efficiente dello stimatore di tipo rapporto in (2.3), quando vale la seguente relazione

$$S_{xy}(2R - \beta) < 0. \quad (2.16)$$

Consideriamo ora l'uso del metodo di imputazione tramite rapporto.

Si consideri un campione casuale s di n unità: supponiamo che ci siano r osservazioni complete $(y_1, x_1), (y_2, x_2), ..., (y_r, x_r)$ e $(n - r)$ osservazioni incomplete $x_1^*, ..., x_{n-r}^*$. Così il campione comprende due gruppi: uno di dimensione r indicato da s_1 e l'altro di dimensione $(n - r)$ indicato da s_2 $(s = s_1 \cup s_2)$.

Prima di tutto, ci serve il seguente stimatore che scarta tutte le osservazioni incomplete nel campione:

$$\overline{y}_r = \tfrac{1}{r} \sum_{i=1}^{r} y_i. \qquad (2.17)$$

In questo caso, il data set completo è indicato da

$$z_i = \begin{cases} y_i & \text{se} \quad i \in s_1 \\ \widetilde{y}_i & \text{se} \quad i \in s_2 \end{cases}. \qquad (2.18)$$

La quantità $\overline{Y}$, è stimata da

$$t = \tfrac{1}{n} \sum_{i=1}^{n} z_i = \tfrac{1}{n} \left(\sum_{i \in s_1} y_i + \sum_{i \in s_2} \widetilde{y}_i \right), \qquad (2.19)$$

dove $\widetilde{y}_i$ indica il valore imputato di y_i, corrispondente all'osservazione x_i^*.

Se si usa il metodo di imputazione per rapporto, vi sono due semplici scelte di $\widetilde{y}_i$,

$$\widetilde{y}_i = \overline{y}\left(\frac{\overline{X}}{\overline{x}}\right) \qquad (2.20)$$

$$\widetilde{y}_i = \overline{y}\left(\frac{n\overline{X}}{r\overline{x}+(n-r)\overline{x}^*}\right), \qquad (2.21)$$

dove $\overline{x}_1 = \frac{1}{r}\sum_{i=1}^{r} x_i$, $\overline{x}_2 = \frac{1}{n-r}\sum_{i=1}^{n-r} x_i^*$ e $\overline{X}$ è ancora supposta nota. Se non fosse nota, possiamo definire i valori imputati come

$$\widetilde{y}_i = \overline{y}\left(\frac{x_i^*}{\overline{x}}\right). \qquad (2.22)$$

Sulla stessa linea, possiamo indicare un altro insieme di valori imputati

$$\widetilde{y}_i = \overline{y}\left(\frac{nx_i^*}{r\overline{x}+(n-r)\overline{x}^*}\right). \qquad (2.23)$$

Utilizzando (2.20)-(2.23) in (2.19), otteniamo i seguenti quattro stimatori di $\overline{Y}$

$$t_1 = \overline{y}\left[\frac{r\overline{x}+(n-r)\overline{X}}{n\overline{x}}\right], \qquad (2.24)$$

$$t_2 = \overline{y}\left[\frac{r^2\overline{x}+n(n-r)\overline{X}+r(n-r)\overline{x}^*}{rn\overline{x}+n(n-r)\overline{x}^*}\right], \qquad (2.25)$$

$$t_3 = \overline{y}\left[\frac{r\overline{x}+(n-r)\overline{x}^*}{n\overline{x}}\right], \qquad (2.26)$$

$$t_4 = \overline{y}\left[\frac{r^2\overline{x}+(n+r)(n-r)\overline{x}^*}{rn\overline{x}+n(n-r)\overline{x}^*}\right], \qquad (2.27)$$

Gli stimatori t_1, t_2, t_3 e t_4 sono basati sull'imputazione per rapporto delle osservazioni mancanti: i primi due richiedono la conoscenza di $\overline{X}$, diversamente dagli altri due.

Poichè questi stimatori sono generalmente distorti, consideriamo i loro MSE per il confronto. Introduciamo anche alcune quantità utili

$$\theta = \frac{\overline{Y}S_x}{\overline{X}S_y}, \;\; f_k = E_p\left(\frac{1}{n-p}\right)\left(\frac{p}{n}\right)^k, \;\; l_k = \frac{1}{n}E_p\left(\frac{p}{n}\right)^k,$$

dove E_p indica l'aspettativa rispetto alla variabile casuale p, a valori interi positivi, e k è un numero intero positivo fissato. Si assume che il coefficiente

di correlazione ρ sia positivo, in quanto questo è un requisito basilare per l'applicazione del metodo del rapporto.

A questo punto, confrontiamo gli MSE dei quattro stimatori distorti. Quando $\overline{X}$ è noto, poichè t_1 ignora le coppie (y_i, x_i) incomplete di dati mentre t_2 le include negli stimatori, si può dimostrare che t_2 ha minor distorsione e anche minor MSE rispetto a t_1, sotto condizione che

$$2\rho < \left(\frac{f_2 - l_2}{f_1 - l_1}\right)\theta. \qquad (2.28)$$

Se nella (2.28) vale il segno $>$, lo stimatore t_1 ha minor MSE, sebbene maggior distorsione dello stimatore t_2.

Quando $\overline{X}$ è ignoto, consideriamo gli altri due stimatori t_3 e t_4 che sono basati su tutte le osservazioni disponibili nel campione. Entrambi sono distorti, ma t_3 possiede minor MSE di t_4 quando vale

$$2\rho > \left(\frac{f_1 - l_1 + l_2}{f_1 - l_1}\right)\theta; \qquad (2.29)$$

se vale (2.29) con verso opposto della disuguaglianza, lo stimatore t_4 è più efficiente di t_3.

Infine, esaminiamo il ruolo dell'informazione su $\overline{X}$ nella formulazione degli stimatori. Si può notare che gli stimatori t_3 e t_4 possono essere considerati come derivanti dalla sostituzione di $\overline{x}^*$ al posto di $\overline{X}$ in t_1 e t_2, rispettivamente. Così, confrontando t_1 con t_3, osserviamo che entrambi gli stimatori hanno la stessa distorsione ma t_1 ha sempre minor MSE di t_3. Allo stesso modo, se confrontiamo t_2 con t_4, entrambi gli stimatori sono distorti ma t_2 ha minor MSE di t_4 a condizione che $l_1 > 2l_2$. Quando $l_1 \leqslant 2l_2$, lo stimatore t_2 basato su $\overline{X}$ non è preferibile.

Consideriamo, ora, un metodo di imputazione basato su stimatori indiretti dei valori disponibili: la procedura consiste nel far uso dell'informazione disponibile dalle osservazioni incomplete per migliorare la precisione di questi stimatori.

Assumiamo che un insieme di $(n - p - q)$ osservazioni complete siano disponibili tra le n unità. Oltre a queste, sono disponibili le osservazioni sulla variabile ausiliaria x di p unità, ma le corrispondenti osservazioni sulla variabile y sono mancanti. Analogamente, abbiamo un insieme di q osservazioni sulla caratteristica y nel campione senza che siano disponibili i corrispondenti

valori sulla caratteristica x. Inoltre, p e q sono assunti come valori interi che soddisfano la condizione $p, q > 0$. Per semplicità, separiamo le unità del campione s in tre insiemi disgiunti,

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

La struttura dei dati campionari è la quindi la seguente:

$$\begin{array}{ccc} \overbrace{\begin{array}{c} y_i \cdots y_{n-q-p} \\ x_1 \cdots x_{n-q-p} \end{array}}^{s_1} & \overbrace{\begin{array}{c} \text{missing} \cdots \text{missing} \\ x_{n-p-q+1} \quad \cdots \quad x_{n-q} \end{array}}^{s_2} & \overbrace{\begin{array}{c} y_{n-q+1} \quad \cdots \quad y_n \\ \text{missing} \cdots \text{missing} \end{array}}^{s_3} \end{array}$$

Quando viene applicato un metodo di imputazione, l'insieme dei dati completi è specificato da

$$z_i = \begin{cases} y_i & \text{se } i \in s_1 \cup s_3 \\ y_i & \text{se } i \in s_2 \end{cases} \qquad (2.31)$$

dove $\widetilde{y_i}$ è il valore imputato, e le stime necessarie possono essere calcolate da questi dati. Così, secondo il nostro obbiettivo, si può indicare la stima di $\overline{Y}$ con

$$\overline{y}_{imp} = \frac{1}{N} \sum_{i \in s} \frac{z_i}{\pi_i}. \qquad (2.32)$$

I metodi di imputazione comunemente usati comprendono l'imputazione per la media, come abbiamo visto precedentemente. Utilizzando metodi di stima indiretta, si possono utilizzare i tradizionali stimatori 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 sarà ridotta di conseguenza.

Gli stimatori di Horvitz-Thompson, relativi ai tre campioni s_1, s_2 e s_3, sono (2.33)

$$\overline{y}_{HT}^{(1)} = \sum_{i \in s_1} \frac{y_i}{N\pi_i}, \qquad \overline{y}_{HT}^{(3)} = \sum_{i \in s_3} \frac{y_i}{N\pi_i}, \ \overline{x}_{HT}^{(1)} = \sum_{i \in s_1} \frac{x_i}{N\pi_i}, \ \overline{x}_{HT}^{(2)} = \sum_{i \in s_2} \frac{x_i}{N\pi_i}.$$

Consideriamo le seguenti classi di stimatori, che includono tutte le osservazioni disponibili

$$\widehat{y}_{r2} = \frac{\alpha_r \ \overline{y}_{HT}^{(3)} + (1-\alpha_r)\overline{y}_{HT}^{(1)}}{b_r \overline{x}_{HT}^{(2)} + b_r \overline{x}_{HT}^{(1)}} \overline{X}, \qquad (2.34)$$

$$\widehat{y}_{d2} = \alpha_d \ \overline{y}_{HT}^{(1)} + (1-\alpha_d)\overline{y}_{HT}^{(3)} + \left(\overline{X} - \left(b_d \overline{x}_{HT}^{(1)} + (1-b_d)\overline{x}_{HT}^{(2)}\right)\right), \quad (2.35)$$

$$\widehat{y}_{reg2} = \alpha_{reg} \ \overline{y}_{HT}^{(1)} + (1-\alpha_{reg})\overline{y}_{HT}^{(3)} + \beta \left[\overline{X} - \left(b_{reg} \overline{x}_{HT}^{(1)} + (1-b_{reg})\overline{x}_{HT}^{(2)}\right)\right] \qquad (2.36)$$

.

Gli stimatori (2.34), (2.35) e (2.36), sono di tipo rapporto, alle differenze e per regressione, rispettivamente.

Nel caso del coefficiente di regressione, se β non è noto, due possibili stimatori di β possono essere

$$\widehat{\beta}_1 = \frac{\widehat{Cov}_{i \in s_1}(x,y)}{\widehat{Var}_{i \in s_1}(x)} \quad \text{e} \quad \widehat{\beta}_2 = \frac{\widehat{Cov}_{i \in s_1}(x,y)}{\widehat{Var}_{i \in s_1 \cup s_2}(x)},$$

che generano i due stimatori per regressione $\widehat{y}_{reg21}$ e $\widehat{y}_{reg22}$. I valori ottimi di α e b che minimizzano l'errore di stima, sono funzione delle varianze e covarianze tra gli stimatori di Horvitz-Thompson. Così si ottengono approssimazioni dei valori ottimi ($\widetilde{\alpha}_r$, $\widetilde{b}_r$, $\widetilde{\alpha}_d$, $\widetilde{b}_d$,$\widetilde{\alpha}_{reg}$ e $\widetilde{b}_{reg}$) che ci permettono di enunciare gli stimatori corrispondenti $\widetilde{y}_{r2}$, $\widetilde{y}_{d2}$ e $\widetilde{y}_{reg2}$, i quali hanno la medesima distribuzione asintotica di (2.34), (2.35) e (2.36).

A questo punto, si analizzano le seguenti procedure d'imputazione:

- procedura basata su uno stimatore per rapporto: in questa situazione, si specifica l'intero data set usando lo stimatore $\widetilde{y}_{r2}$, per il valore imputato.
- procedura basata su uno stimatore alle differenze: in questa situazione, si specifica l'intero data set usando lo stimatore $\widetilde{y}_{d2}$, per il valore imputato.
- procedura basata su uno stimatore per regressione (β ignoto): in questa situazione, si hanno due stimatori per regressione, usando $\widehat{\beta}_1$ e $\widehat{\beta}_2$. Così abbiamo due procedure d'imputazione, con $\widetilde{y}_{reg21}$ e $\widetilde{y}_{reg22}$.

Dopo aver usato uno di questi metodi d'imputazione e specificato il data set completo, si può usare lo stimatore in (2.32) per stimare $\overline{Y}$.

I metodi d'imputazione mostrati, sono un pò più complessi da applicare, però producono un incremento di efficienza rispetto alle tecniche tradizionali (ad esempio, imputazione per la media).

2.2.2 Imputazione nel campionamento stratificato e per clusters

Consideriamo ora l'imputazione multipla, nel contesto del campionamento stratificato. Essa si basa sull'idea di creare γ valori imputati per ogni valore mancante e combinare i γ data set completati, secondo la formula di combinazione di Rubin (1996).

Nel campionamento stratificato, consideriamo H campioni casuali semplici dagli H strati di dimensioni N_h, $h = 1, ..., H$, da una popolazione finita di dimensione N. Assumiamo, in ogni strato, un meccanismo di non risposta del tipo MCAR: questo significa che gli indicatori di risposta nello strato h, $a_{h,1}, ..., a_{h,N_h}$, sono indipendenti con $P_{ha} = P(a_{h,i} = 1)$. Il metodo di imputazione, che può essere uno qualsiasi, è basato sui dati osservati all'interno dello strato. Siano $y_{h,obs}$ i valori osservati dal campione rispondente s_{ha} di dimensione n_{ha} dallo strato h, $y_{h,obs} = (y_i : i \in s_{ha})$. Allora, un valore imputato $\widetilde{y}_i$ nello strato h, è estratto a caso tra i valori di $y_{h,obs}$. Lo stimatore basato sul campione completo è l'usuale media pesata delle medie di strato $\overline{y}_{strat} = \sum_{h=1}^{H} N_h \overline{y}_h / N$, dove $\overline{y}_h = \sum_{i \in s_h} y_i / n_h$ e s_h è il campione dallo strato h di dimensione n_h. La varianza dello stimatore è

$$Var(\overline{y}_{strat}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 S_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right),$$

con

$$S_h^2 = \frac{1}{(N_h - 1)} \sum_{i=1}^{N_h} (y_i - \overline{Y}_h)^2$$

varianza dello strato h. Qui, $\overline{Y}_h$ è la media di popolazione dello strato h. Sia $\overline{y}_{ha}$ la media campionaria dello strato h, e

$$\widehat{S}_{ha}^2 = \frac{1}{n_{ha} - 1} \sum_{i \in s_{ha}} (y_i - \overline{y}_{ha})^2$$

la varianza campionaria. Lo stimatore, basato sull'imputazione, è dato da

$$\overline{y}_{strat,IMP} = \frac{1}{N} \sum_{h=1}^{H} N_h \overline{y}_h^*, \qquad (2.37)$$

dove

$\overline{y}_h^* = (\sum_{i \in s_{ha}} y_i + \sum_{i \in (s_h - s_{ha})} \widetilde{y}_i)/n_h.$

Poichè siamo nella situazione di imputazione multipla, indichiamo con $\overline{y}_{strat,IMP,i}$ le γ repliche d'imputazione di (2.37), per $i = 1, ..., \gamma$. Lo stimatore finale, combinazione dei γ stimatori, è dato da

$$\overline{\overline{y}}_{strat,IMP} = \sum_{i=1}^{\gamma} \frac{\overline{y}_{strat,IMP,i}}{\gamma}. \qquad (2.38)$$

Nell'imputazione multipla, la formula di combinazione di Rubin, permette di determinare la stima della varianza totale di (2.38)

$$\widehat{Var}(\overline{\overline{y}}_{strat,IMP}) = \sum_{h=1}^{H} \overline{V}_h^* + \sum_{h=1}^{H} \left(\frac{1}{1-f_h} + \frac{1}{\gamma}\right) B_h^*, \qquad (2.39)$$

dove

$$f_h = (n_h - n_{ha})/n_h \qquad (2.40)$$

è il tasso di non risposta nello strato h; $\overline{V}_h^*$, detta varianza entro-imputazioni dello strato h, è la media dei γ valori

$\widehat{V}_h^* = \left(\frac{N_h}{N}\right)^2 \widehat{S}_{h*}^2 (\frac{1}{n_h} - \frac{1}{N_h})$,

dove

$\widehat{S}_{h*}^2 = \frac{1}{n_h - 1} \left(\sum_{i \in s_{ha}} (y_i - \overline{y}_h^*)^2 + \sum_{i \in (s_h - s_{ha})} (\widetilde{y}_i - \overline{y}_h^*)^2\right)$.

La quantità B_h^*, detta componente tra-imputazioni, è data da

$B_h^* = \frac{1}{\gamma - 1} \sum_{i=1}^{\gamma} \left(\overline{y}_{h,i}^* - \overline{\overline{y}}_h^*\right)^2$,

con $\overline{\overline{y}}_h^* = \sum_{i=1}^{\gamma} (\overline{y}_{h,i}^*)/\gamma$. Si può notare che la varianza in (2.39), dello stimatore (2.38), dipende dalla quantità (2.40): maggiore è il tasso di non risposta negli strati, più peso c'è sulla componente B_h^* e maggiore è la varianza dello stimatore.

Presentiamo ora, un metodo di stima di $\overline{Y}$, nel caso di campionamento per clusters assumendo ancora un meccanismo dei dati mancanti di tipo MCAR. Nel campionamento per clusters, il campione si forma in due stadi: nel primo, le unità campionate sono i clusters che contengono le unità che saranno estratte nel secondo stadio. Questo tipo di disegno campionario, viene usato per motivi economici: è necessario quando non è disponibile una lista attendibile delle unità di secondo stadio della popolazione.

Dentro l'i-esimo cluster campionato, sia s_i il campione di secondo stadio di dimensione $m_i \geqslant 2$. Per la j-esima unità campionata ($j \in s_i$), il peso di campionamento w_{ij} è costruito a partire dalla specificazione del disegno campionario.

Sia y_{ij} il valore di y dell'unità j-esima nel cluster i, adottando un approccio di imputazione basato su un modello si assume che ciascuna y_{ij} sia una variabile casuale con

$$y_{ij} = \mu_i + b_i + e_{ij}. \qquad (2.41)$$

La quantità μ_i in (2.41), è un parametro ignoto; b_i è un effetto casuale a livello di cluster che non è osservato, con media 0 e varianza finita; e_{ij} è un effetto casuale all'interno del cluster che non è osservato, con media 0 e varianza finita; si assume inoltre che, b_i e e_{ij} siano indipendenti. La distribuzione di y_{ij} può variare con i, j.

Assumiamo che a_{ij}, sia la variabile indicatrice di risposta per y_{ij} ($a_{ij} = 1$ se y_{ij} è un rispondente, $a_{ij} = 0$ altrimenti) e che ogni cluster ha almeno un rispondente, cioè per un qualsiasi $i \in s$, almeno un a_{ij} è pari a 1.

Si ipotizza che, la non risposta dipenda da un effetto casuale a livello di cluster, cioè

$$P_m(\mathbf{a}_i|\mathbf{y}_i,b_i) = P_m(\mathbf{a}_i|b_i)\ i \in s, \qquad (2.42)$$

dove P_m è la probabilità rispetto al modello (2.41); $\mathbf{a}_i$ è il vettore contenente a_{ij} e $\mathbf{y}_i$ è il vettore contenente y_{ij}. Poichè b_i è un effetto non osservato, il meccanismo di non risposta è del tipo MCAR: cioè, condizionatamente a b_i, $\mathbf{y}_i$ e $\mathbf{a}_i$ sono indipendenti.

In queste condizioni, si esegue un'imputazione all'interno di ogni cluster per ottenere uno stimatore non distorto di $\overline{Y}$. Infatti, se imputiamo un non rispondente y_{ij} nel cluster i, con la media di cluster

$$\widetilde{y}_i = \frac{\sum\limits_{j\in s_i} a_{ij} w_{ij} y_{ij}}{\sum\limits_{j\in s_i} a_{ij} w_{ij}}\,; \tag{2.43}$$

allora, lo stimatore di $\overline{Y}$ è

$$\overline{y}_c = \sum_{i\in s}\sum_{j\in s_i} a_{ij}\overline{w}_{ij} y_{ij}, \tag{2.44}$$

con

$$\overline{w}_{ij} = w_{ij}\left(\frac{\sum\limits_{j\in s_i} w_{ij}}{\sum\limits_{j\in s_i} a_{ij} w_{ij}}\right).$$

Si può dimostrare che $E_m(\overline{y}_c) = E_m(y) = \overline{Y}$, dove E_m è il valore atteso rispetto a P_m. Poichè l'imputazione è fatta all'interno di ogni cluster, lo stimatore in (2.44) sembra inefficiente quando alcuni m_i sono molto piccoli. Questa preoccupazione, non sussiste nel caso in cui $w_{ij} = w_i$ per tutti i j (e.g. la probabilità di inclusione è la stessa per tutte le unità di uno stesso campione di secondo stadio). Quando $w_{ij} = w_i$ per tutti i j, l'imputazione che porta allo stimatore (2.44) è fatta su un più ampio gruppo G_l, che comprende i clusters che hanno la stessa dimensione m_i e lo stesso tasso di risposta intra-cluster $\overline{a}_i$. Rappresentiamo il gruppo di clusters G_l come

$$G_l = \left\{ i \in s : m_i = m, \overline{a}_i = m_i^{-1} \textstyle\sum_{j\in s_i} a_{ij} \right\},$$

e, quindi, l'imputazione è fatta all'interno di ogni gruppo G_l quando vale $w_{ij} = w_i$ per tutti i j. Ad esempio, un non rispondente in s_i, è imputato con la media campionaria dei rispondenti in G_l pari a

$$\overline{y}_{G_l} = \frac{\sum\limits_{i\in G_l}\sum\limits_{j\in s_i} a_{ij} w_{ij} y_{ij}}{\sum\limits_{i\in G_l}\sum\limits_{j\in s_i} a_{ij} w_{ij}}. \tag{2.45}$$

Nel caso in cui $w_{ij} = w_j$ per tutti i, l'imputazione viene fatta sulla base di un gruppo di clusters che condividono la stessa quantità $E_m(\mathbf{y}_i|\mathbf{a}_i)$.

La varianza dello stimatore (2.44), oppure una sua approssimazione (per $n \to \infty$), si calcola tramite il metodo jackknife corretto, come descritto in Rao e Shao (1992).

Senza l'assunzione che ciascun cluster campionato abbia almeno un rispondente, lo stimatore (2.44) può non essere calcolato a meno che non siano fatte

altre assunzioni. Con il meccanismo di risposta (2.42), quando tutte le osservazioni in un cluster sono mancanti, non si può recuperare informazione dagli altri clusters a meno che non si assuma ad esempio, un modello per imputare i valori per i clusters senza rispondenti.

Il modello d'imputazione può essere il modello a effetto casuale (2.41), con il meccanismo di risposta basato su un effetto casuale (2.42). Naturalmente se (2.41) e (2.42) non valgono, lo stimatore (2.44) è distorto.

In questo caso si è usato il metodo di imputazione tramite media, ma estensioni ad altri metodi d'imputazione sono fattibili. Se c'è una variabile ausiliaria x, i cui valori sono tutti osservati, il risultato mostrato può essere esteso all'imputazione tramite regressione modificando così il modello

$$y_{ij} = \alpha + \beta x_{ij} + b_i + e_{ij}. \qquad (2.46)$$

Quando la probabilità di risposta di y_{ij}, dipende dall'effetto casuale specifico di cluster b_i, il meccanismo di non risposta non è MAR poichè gli effetti b_i non sono osservati. Di conseguenza, il modello (2.46) porta a stimatori distorti. Per correggere la distorsione, modifichiamo il modello (2.46) in modo da introdurre il tasso stimato di risposta nei clusters.

Il modello sarà

$$[y_{ij}|x_{ij}, \alpha, b_i, \delta, \beta, e_{ij}, \sigma^2] = N(b_i + \delta\overline{a}_i + \alpha + \beta x_{ij} + e_{ij}, \sigma^2) \qquad (2.47)$$

dove $N(\cdot)$ indica una distribuzione normale. Tale modello include il tasso di risposta $\overline{a}_i$ di ogni cluster come una variabile ausiliaria aggiuntiva rispetto a x.

Un approccio più rigoroso per correggere la distorsione, porta al seguente modello parametrico alternativo:

$$[y_{ij}|b_i, \chi_i, \delta, \alpha, \beta, e_{ij}, \sigma^2] = N\left(b_i + \delta\chi_i + \alpha + \beta x_{ij} + e_{ij}, \sigma^2\right)$$

$$[z_{ij}|\chi_i, \alpha_1, \beta_1] = N\left(\chi_i + \alpha_1 + \beta_1 x_{ij}, 1\right)$$

$$a_{ij} = \begin{cases} 1 & \text{se } z_{ij} > 0 \\ 0 & \text{se } z_{ij} < 0 \end{cases}. \qquad (2.48)$$

La z_{ij} è una variabile latente che determina lo stato della risposta del soggetto j nel cluster i. Se $z_{ij} > 0$, il soggetto risponde; altrimenti, il soggetto

non risponde. Gli effetti casuali b_i e χ_i modellano le correlazioni intra-cluster. Si possono comuque estendere i modelli appena descritti anche al caso di un vettore $\mathbf{x}$ di variabili ausiliarie ($\mathbf{x} = (x_1, ..., x_d)'$). Secondo i modelli (2.47) e (2.48), gli a_{ij} e gli y_{ij} sono indipendenti all'interno dei clusters, dopo aver condizionato rispetto alle ausiliarie e ai χ_i: questo porta a concludere che da un meccanismo non MAR ci siamo ricondotti a un modello di mancata risposta di tipo MAR.

Sotto questi modelli, lo stimatore di $\overline{Y}$, può essere ottenuto tramite imputazione multipla; anzitutto, formiamo Γ data set imputati, completando i valori mancanti di y tramite Γ estrazioni indipendenti dalla distribuzione di (2.47) o (2.48). Per il γ-esimo data set imputato, la stima di $\overline{Y}$ è

$$\overline{y}_\gamma = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{y_{ij}^{(\gamma)}}{\pi_{ij}}}{\sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{1}{\pi_{ij}}} \qquad (2.49)$$

dove $y_{ij}^{(\gamma)}$ è il valore di y_{ij} osservato o imputato. Così, uno stimatore consistente di $\overline{Y}$ è

$$\overline{y} = \frac{1}{\Gamma} \sum_{\gamma=1}^{\Gamma} \overline{y}_\gamma \qquad (2.50)$$

e la sua varianza è

$$Var(\overline{y}) = \frac{1}{\Gamma} \sum_{\gamma=1}^{\Gamma} V_\gamma + \frac{1}{\Gamma - 1} \sum_{\gamma=1}^{\Gamma} \left(\overline{y}_\gamma - \overline{y}\right)^2, \qquad (2.51)$$

dove V_γ è la varianza dello stimatore in (2.49) calcolata per il γ-esimo data set.

Abbiamo descritto alcuni metodi per trattare i dati mancanti nel caso di campionamento per clusters in cui si tiene conto sia delle ausiliarie che dell'informazione dei clusters stessi. Se si ha a disposizione solamente la variabile di studio e informazioni sull'effetto dei clusters, il modello (2.41) è il più appropriato nel caso di un campionamento per clusters a due stadi. Invece, se si dispone anche di variabili ausiliarie completamente osservate, i modelli (2.47) e (2.48) permettono di sfruttare l'informazione aggiuntiva per stimare $\overline{Y}$. In tutti e tre i modelli (2.41), (2.47) e (2.48), si assume che la probabilità di risposta dipenda dall'effetto del cluster. Se le variabili

ausiliarie sono rilevabili per tutti, rispondenti e non (per esempio dai dati del censimento), allora possiamo confrontare la distribuzione delle variabili ausiliarie tra non rispondenti e rispondenti all'interno del cluster. Se non vi è alcuna differenza sistematica, è probabile che la non risposta dipenda dalle specifiche caratteristiche di cluster.

2.3 Dati mancanti MAR

Il meccanismo che genera i dati mancanti del tipo MAR, prevede che la probabilità di risposta delle unità sia indipendente dalla variabile di studio ma dipendente dalla variabile ausiliaria osservata. Il meccanismo di non risposta, è ricostruibile o prevedibile dalle variabili ausiliarie coinvolte nell'indagine (piuttosto che dalla variabile di studio y).

2.3.1 Imputazione 'nearest neighbor'

Consideriamo un campione del tipo $(y_1, x_1), ..., (y_n, x_n)$, con i valori osservati $y_1, ..., y_r$ (rispondenti), mancanti $y_{r+1}, ..., y_n$ (non rispondenti) e osservati $x_1, ..., x_n$. Il metodo di imputazione 'nearest neighbor' (NNI), imputa un valore mancante y_j con y_i $(1 \leq i \leq r)$ ed i è l'unità più vicina a j, secondo la variabile x. Quindi, i soddisfa la condizione $|x_i - x_j| = \min_{1 \leq l \leq r} |x_l - x_j|$. Inoltre, NNI è spesso messa in atto dividendo, inizialmente, il campione in diverse 'classi d'imputazione' e poi trovando i valori sostitutivi all'interno di queste classi.

L'imputazione nearest neighbor, permette anzitutto di imputare un non rispondente con un rispondente dalla stessa variabile; i valori imputati non sono valori del tutto costruiti e, anche se possono non essere perfetti sostituti, sono valori sensati. Il metodo NNI può essere più efficiente di altri, tipo l'imputazione tramite media, quando la variabile x fornisce informazione ausiliaria utile. Inoltre, questo metodo non assume un modello di regressione parametrica tra y e x, e quindi è più robusto rispetto a metodi tipo l'imputazione tramite rapporto e regressione che sono basati su un modello di regressione lineare.

Sia s un campione di dimensione n e sia w_i, il peso per l'unità i-esima, ovvero l'inverso della probabilità che l'unità i sia campionata $(w_i = \pi_i^{-1})$.

La popolazione U, è divisa in κ (intero fissato) classi d'imputazione: all'interno di ogni classe le osservazioni $(a_i, a_i y_i, x_i)$, $i = 1, ..., n$, sono indipendenti e identicamente distribuite (i.i.d.) con $P(a_i = 1|y_i, x_i, \kappa) = P(a_i = 1|x_i, \kappa)$. Le osservazioni di classi d'imputazione diverse, sono ancora indipendenti e la NNI viene effettuata all'interno di ciascuna classe κ. Sebbene le osservazioni siano assunte i.i.d. all'interno di una classe, il meccanismo di risposta è MAR poichè $P(a_i = 1|x_i)$ dipende dalla x.

Le classi di imputazione sono costruite usando una variabile categoriale, i cui valori sono osservati per tutte le unità del campione; per esempio, nel campionamento stratificato, gli strati o unioni di strati sono usate spesso come classi d'imputazione.

Sebbene il metodo NNI si basi su un approccio da modello, l'assunzione del modello è nonparametrica: infatti assumiamo solo che le osservazioni siano i.i.d. all'interno di ciascuna classe d'imputazione. Questa condizione è molto più debole dell'assunzione di un modello parametrico lineare per $E(y|x)$, la quale è tipicamente usata nell'imputazione per regressione.

Nella classe d'imputazione κ, sia a_κ l'insieme di indici per i rispondenti a y e $\overline{a}_\kappa$ l'insieme degli indici per i non rispondenti (con $s_\kappa = a_\kappa \cup \overline{a}_\kappa$). La dimensione del campione s_κ nella classe κ-esima è n_κ, mentre la dimensione di popolazione di U_κ è N_κ ($N = \sum_\kappa N_\kappa$).

Dopo l'uso di NNI, un possibile stimatore per $\overline{Y}$, è

$$\overline{y}_{(1)} = \frac{1}{N} \sum_\kappa \left(\sum_{i \in a_\kappa} w_i y_i + \sum_{i \in \overline{a}_\kappa} w_i \widetilde{y}_i \right) = \sum_\kappa \frac{N_\kappa}{N} \left(\sum_{i \in a_\kappa} \overline{w}_{\kappa,i} y_i + \sum_{i \in \overline{a}_\kappa} \overline{w}_{\kappa,i} \widetilde{y}_i \right) \quad (2.52)$$

,

dove $\widetilde{y}_i$ è il valore imputato per il non rispondente ($i \in \overline{a}_\kappa$) e $\overline{w}_{\kappa,i} = w_i / N_\kappa$ (quando $i \in s_\kappa$). Se N è ignota, una sua stima consistente è $\widehat{N} = \sum_{i \in s} w_i$. In questo caso, possiamo stimare $\overline{Y}$ con lo stimatore di tipo rapporto

$$\overline{y}_{(2)} = \frac{1}{\widehat{N}} \sum_\kappa \left(\sum_{i \in a_\kappa} w_i y_i + \sum_{i \in \overline{a}_\kappa} w_i \widetilde{y}_i \right) = \frac{\overline{y}_{(1)}}{\frac{\widehat{N}}{N}}. \quad (2.53)$$

Le proprietà asintotiche dello stimatore (2.53), derivano da quelle dello stimatore in (2.52), perciò ci concentreremo su $\overline{y}_{(1)}$. Si può dimostrare che,

$$\frac{\sqrt{n}(\overline{y}_{(1)} - \overline{Y})}{\sigma} \longrightarrow_d N(0, 1), \quad (2.54)$$

per qualche $\sigma > 0$, dove $\rightarrow_d$ è la convergenza in distribuzione. Uno stimatore per la varianza di (2.52), è

$$\widehat{V}(\overline{y}_{(1)}) = \sum_{\kappa} \frac{1}{n_\kappa(n_\kappa - 1)N^2} \sum_{j \in s_\kappa} (n_\kappa w_j \widetilde{y}_j - \overline{y}_\kappa)^2, \qquad (2.55)$$

dove

$$\overline{y}_\kappa = \sum_{i \in a_\kappa} (1 + d_i^{(\kappa)}) w_i y_i,$$

$$d_i^{(\kappa)} = \sum_{j \in \overline{a}_\kappa} \left(\frac{w_j}{w_i}\right) d_{ij};$$

$d_{ij} = 1$ se i è l'unità più vicina a j, e $d_{ij} = 0$ altrimenti. Inoltre,

$\widetilde{y}_j = y_j + d_j^{(\kappa)} g_j^{(\kappa)} (y_j - (y_{j\kappa 1} + y_{j\kappa 2})/2)$ se $j \in a_\kappa$

$\widetilde{y}_j =$ il valore imputato di y_j se $j \in \overline{a}_\kappa$; con

$$g_j^{(\kappa)} = \frac{\left[\sqrt{6(d_j^{(\kappa)})^2 + 6d_j^{(\kappa)} + 4} - 2\right]}{3d_j^{(\kappa)}} \; (g_j^{(\kappa)} = 0 \text{ se } d_j^{(\kappa)} = 0)$$

e $jk1$, $jk2$ le due unità più vicine di j in a_κ.

A questo punto, sfruttando i risultati in (2.54) e (2.55), possiamo indicare un intervallo di confidenza di livello $1 - \alpha$ per $\overline{Y}$:

$$\left[\overline{y}_{(1)} - z_{1-\alpha/2}\sqrt{\widehat{V}(\overline{y}_{(1)})}, \overline{y}_{(1)} + z_{1-\alpha/2}\sqrt{\widehat{V}(\overline{y}_{(1)})}\right], \qquad (2.56)$$

dove α è il livello di confidenza prefissato e $z_{1-\alpha/2}$ è il quantile $1 - \alpha/2$ della distribuzione normale standard.

Per l'intervallo di confidenza (2.56), vale la condizione

$$P\left(\overline{y}_{(1)} - z_{1-\alpha/2}\sqrt{\widehat{V}(\overline{y}_{(1)})} \leq \overline{Y} \leq \overline{y}_{(1)} + z_{1-\alpha/2}\sqrt{\widehat{V}(\overline{y}_{(1)})}\right) \longrightarrow 1 - \alpha,$$

per $n/N \rightarrow 0$.

2.3.2 Imputazione tramite pseudo-verosimiglianza

Presentiamo metodi d'imputazione basati sulla pseudo-verosimiglianza empirica sotto assunzione di distribuzione marginale non parametrica per y.

Dopo l'imputazione, la stima del parametro di interesse avviene trattando i valori imputati come valori osservati e usando la formule usuali nel caso di assenza di dati mancanti. Sia U la popolazione suddivisa in H strati con N_h unità nell'h-esimo strato. Assumiamo che $n_h \geqslant 2$ unità sono campionate dallo strato h, secondo un certo disegno campionario, in modo indipendente tra gli strati. Secondo il disegno campionario, i pesi sono $w_{hi} = (N\pi_{hi})^{-1}$, $i = 1, ..., n_h$, $h = 1, ..., H$, dove π_{hi} è la probabilità che l'i-esima unità nello strato h entri nel campione. Sia inoltre z una variabile ausiliaria categoriale, che assume valori in $\{z_1, ..., z_\tau\}$ (τ intero fissato). All'interno dello strato h, assumiamo che i valori (y, z) siano casuali e y abbia una distribuzione marginale non parametrica ignota F_h; inoltre si suppone la seguente funzione di probabilità parametrica

$$P_h = (Z = z|Y = y) = f_h(y, z, \beta), \qquad (2.57)$$

dove β è un vettore di parametri ignoti e f_h è una funzione nota. Per ogni unità, il valore di z è sempre osservato mentre ci possono essere osservazioni mancanti tra i valori di y. Infatti si assume che nello strato h, le prime r_h unità rispondono e le restanti $n_h - r_h$ non rispondono. I dati si presentanto in tal modo

$$\{(y_{hi}, z_{hi}), i = 1, ..., r_h\} \cup \{z_{hi}, i = r_h + 1, ..., n_h\},\ h = 1, ..., H.$$

Sia $\widetilde{y}_{hi} = y_{hi}$ se y_{hi} è una risposta e sia $\widetilde{y}_{hi}$ il valore imputato se y_{hi} non è osservato. Dopo l'imputazione, la $\overline{Y}$ viene stimata da

$$\overline{y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi}\widetilde{y}_{hi}. \qquad (2.58)$$

Usando gli stimatori di massima verosimiglianza empirica, consideriamo le seguenti due procedure d'imputazione:

1. Imputazione tramite media di pseudo-verosimiglianza. Per ogni non rispondente nello strato h con $Z = z_j$, il valore imputato per y è lo stimatore della media

$$\overline{y}_{hj} = \frac{\sum_{i=1}^{r_h} \widehat{p}_{hi} f_h(y_{hi}, z_j, \widehat{\beta}) y_{hi}}{\sum_{i=1}^{r_h} \widehat{p}_{hi} f_h(y_{hi}, z_j, \widehat{\beta})}. \qquad (2.59)$$

2. Imputazione casuale di pseudo-verosimiglianza. Ciascun non rispondente nello strato h con $Z = z_j$, è imputato utilizzando un campione casuale con reinserimento estratto da tutti i rispondenti dello stesso strato; la probabilità di ciascun y_{hi} di essere selezionato è

$$\frac{\widehat{p}_{hi} f_h(y_{hi}, z_j, \widehat{\beta}) y_{hi}}{\sum_{i=1}^{r_h} \widehat{p}_{hi} f_h(y_{hi}, z_j, \widehat{\beta})}, \qquad (2.60)$$

per $i = 1, ..., r_h$. La quantità $\widehat{p}_{hi}$ in (2.59) e (2.60) è pari a

$$\widehat{p}_{hi} = \frac{w_{hi}}{\sum_{i=1}^{n_h} w_{hi} - \sum_{j=1}^{\tau} \frac{\alpha_{hj}}{\widehat{\pi}_{hj}} f_h(y_{hi}, z_j, \widehat{\beta})}, \qquad (2.61)$$

per $i = 1, ..., r_h$, $h = 1, ..., H$, dove $\alpha_{hj} = \sum_{i=r_h+1}^{n_h} w_{hi} I_{\{Z_{hi}=z_j\}}$ con $I_{\{Z\}}$ la funzione indicatrice dell'evento Z. Nella (2.61), una stima consistente di π_{hj} ($\pi_{hj} = P_h(Z = z_j)$, all'interno dello strato h) è

$$\widehat{\pi}_{hj} = \frac{\sum_{i=1}^{n_h} w_{hi} I_{\left\{Z_{hi}=z_j\right\}}}{\sum_{i=1}^{n_h} w_{hi}}. \qquad (2.62)$$

A questo punto, la stima $\widehat{\beta}$ di β, si ottiene massimizzando la seguente pseudo-verosimiglianza empirica (2.63)

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

rispetto a β.

Si può dimostrare che, lo stimatore in (2.58) basato sulle due procedure d'imputazione appena descritte, è consistente e asintoticamente normale, infatti si ha

$$\sqrt{n}(\overline{y} - \overline{Y}) \longrightarrow_d N(0, \sigma^2),$$

dove σ^2 è la varianza asintotica, che si stima usando la procedura bootstrap.

Questo tipo di imputazione, attraverso la pseudo-verosimiglianza, utilizza tutti i rispondenti con una ponderazione appropriata.

Mostriamo un altro metodo di imputazione dei dati mancanti basato sulla pseudo-verosimiglianza empirica, e lo stimatore di $\overline{Y}$ che ne risulta.

Supponiamo di avere un campione $(a_{i,}a_i y_i, x_i)$, $i = 1, ..., n$: i dati mancanti sono relativi alla variabile y, mentre osserviamo tutti i valori per la x. In questo caso, indichiamo con $\mathbf{x}$ il vettore delle variabili ausiliarie che influenzano la variabile di studio y. Uno stimatore di $\overline{Y}$ è

$$\overline{y} = \tfrac{1}{n} \sum_{i=1}^{n} [a_i y_i + (1 - a_i)\widetilde{y}_i] \qquad (2.64)$$

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

Consideriamo ora un metodo d'imputazione in cui il valore di y_i mancante viene sostituito con

$$\widetilde{y}_i = m(x_i,\widehat{\beta}) + \sum_{i=1}^{n} a_i \widehat{q}_i \left[y_i - m(x_i,\widehat{\beta}) \right]. \qquad (2.65)$$

La quantità $m(x_i,\widehat{\beta})$, rappresenta una stima del modello di $E(y_i|x_i)$: si ha che $\widehat{\beta}$ stima il vettore di parametri ignoti β usando i dati $(y_i, x_i, a_i = 1)$, $i = 1, ..., n$. La stima di β, viene fatta sotto il modello di regressione imposto $m(x_i,\beta)$ per $E(y_i|x_i)$. La quantità q_i è la probabilità condizionata di (y_i, x_i), dato $a_i = 0$ e viene stimata attraverso un approccio di pseudo-verosimiglianza empirica.

Sostituendo (2.65) in (2.64), si ottiene il seguente stimatore di $\overline{Y}$

$$\overline{y}_{ER} = \tfrac{1}{n} \sum_{i=1}^{n} \left[a_i y_i + (1 - a_i) m(x_i,\widehat{\beta}) \right] + \left(1 - \tfrac{n_r}{n}\right) \sum_{i=1}^{n} a_i \widehat{q}_i \left[y_i - m(x_i,\widehat{\beta}) \right], \qquad (2.66)$$

dove $n_r = \sum_{i=1}^{n} a_i$ è il numero dei rispondenti.

Si può aggiungere che lo stimatore (2.66) è uno stimatore doppiamente robusto, poichè si può dimostrare che è consistente per $\overline{Y}$, se almeno uno tra $m(x_i,\beta)$ e $\phi(x_i)$ è correttamente specificato. A prescindere dal fatto che il modello $m(x_i,\beta)$ sia correttamente specificato, se il modello $\phi(x_i)$ è corretto allora

$$\sqrt{n}(\overline{y}_{ER} - \overline{Y}) \longrightarrow_d N(0, \sigma^2), \qquad (2.67)$$

per qualche $\sigma > 0$. La stima della varianza asintotica di (2.66) viene calcolata tramite il metodo bootstrap. Con questa stima della varianza si può costruire un intervallo di confidenza per $\overline{Y}$, grazie al risultato asintotico in (2.67).

La tecnica d'imputazione presentata in (2.65) viene detta imputazione doppiamente robusta al contrario dell'imputazione per regressione, che è una tecnica non robusta nel caso di non corretta specificazione del modello di regressione.

2.3.3 Imputazione ponderata

Sia $\mathbf{x} = (x_1, ..., x_d)'$, un vettore di variabili ausiliarie osservate per tutto il campione che influenzano la variabile di studio y. Il campione osservato è $(a_{i,}a_i y_i, x_i)$, $i = 1, ..., n$: i dati incompleti sono presenti solo per la variabile y. Uno stimatore per $\overline{Y}$ è

$$\overline{y}_{WI} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{y}_i, \qquad (2.68)$$

e, usa l'imputazione ponderata con $\widetilde{y}_i$ che vale

$$\widetilde{y}_i = \frac{a_i y_i}{\widehat{\phi}(x_i)} + \left(1 - \frac{a_i}{\widehat{\phi}(x_i)}\right) \widehat{m}_b(x_i),\ i = 1, ..., n. \qquad (2.69)$$

La quantità $\widehat{m}_b(\mathbf{x})$ è una versione troncata dello stimatore di $m(\mathbf{x}) = E(Y|X)$, cioè

$$\widehat{m}_b(\mathbf{x}) = \frac{\left(nh^d\right)^{-1} \sum_{i=1}^{n} a_i y_i K_h(x_i - \mathbf{x})}{\max\left\{b, \left(nh^d\right)^{-1} \sum_{i=1}^{n} a_i K_h(x_i - \mathbf{x})\right\}}, \qquad (2.70)$$

dove ciascuna $h = h_n$ e $b = b_n$, è una successione di costanti positive tendente a zero; invece $K_h(\cdot) = K(\cdot/h)$, con $K(\cdot)$ una funzione Kernel. La quantità $\widehat{\phi}(\mathbf{x})$, è uno stimatore di $\phi(\mathbf{x})$ dato da

$$\widehat{\phi}(\mathbf{x}) = \frac{\sum_{i=1}^{n} a_i L_\alpha(x_i - \mathbf{x})}{\max\left\{1, \sum_{i=1}^{n} L_\alpha(x_i - \mathbf{x})\right\}}, \qquad (2.71)$$

dove $\alpha = \alpha_n$ è una successione di costanti positive tendente a zero, mentre $L_\alpha(\cdot) = L(\cdot/\alpha)$ con $L(\cdot)$ funzione Kernel.

In questo caso, l'imputazione permette di lavorare con un data set completo sostituendo i valori mancanti di y_i con $\widehat{m}_b(x_i)$. La stima di $\overline{Y}$, viene calcolata a partire dai dati imputati $\{(\widetilde{y}_i, a_i); 1 \leq i \leq n\}$ dove $\widetilde{y}_i$ deriva dalla (2.69).

Si può provare che, lo stimatore in (2.68) ha distribuzione limite

$$\frac{\sqrt{n}(\overline{y}_{WI}-\overline{Y})}{\widehat{Var}^{1/2}} \longrightarrow_d N(0,1),$$

dove $\widehat{Var}$ è lo stimatore consistente della varianza asintotica di $\overline{y}_{WI}$ che vale

$$\widehat{Var}(\overline{y}_{WI}) = \frac{1}{n}\sum_{i=1}^{n}(\widetilde{y}_i - \overline{y}_{WI})^2. \tag{2.72}$$

Usando questo risultato, un intervallo di confidenza per $\overline{Y}$ basato sull'approssimazione normale è

$$\left[\overline{y}_{WI} - z_{1-\alpha/2}\sqrt{\widehat{Var}/n}, \overline{y}_{WI} + z_{1-\alpha/2}\sqrt{\widehat{Var}/n}\right], \tag{2.73}$$

dove $z_{1-\alpha/2}$ è il quantile $1-\alpha/2$ della distribuzione normale standard e $1-\alpha$ è il livello di confidenza dell'intervallo in (2.73).

2.4 Conclusioni

Le questioni principali presentate in questo capitolo riguardano la stima di $\overline{Y}$, a seconda del meccanismo che genera i dati mancanti, del tipo d'imputazione usata e del tipo di campionamento.

Nel caso di un meccanismo MCAR, si sono presentati tre stimatori adattando lo stimatore considerato in Kadilar and Cingi (2004) al metodo di imputazione tramite rapporto e per la media; sotto opportune condizioni, si è mostrato a livello teorico che questi stimatori sono più efficienti dei tradizionali stimatori basati sull'imputazione tramite media o rapporto. In aggiunta a partire da quattro diversi metodi d'imputazione per rapporto, abbiamo enunciato quattro stimatori di $\overline{Y}$ e mostrato i relativi MSE.

Sempre sotto assunzione MCAR, viene presentato un metodo di imputazione basato sugli stimatori indiretti delle osservazioni disponibili: in

questo caso, i tradizionali stimatori per rapporto, alle differenze e per regressione della media sono utilizzati come valori imputati. Questi metodi sono più complessi da applicare, però producono un incremento di efficienza rispetto alle tecniche d'imputazione tradizionali. Sotto campionamento stratificato, si è visto lo stimatore basato sull'imputazione multipla, la quale si basa sull'idea di creare un certo numero di valori imputati per ogni valore mancante e combinare i diversi data set completati separatamente per ciascun strato. Abbiamo descritto alcuni metodi per trattare le non risposte nel caso di campionamento per clusters, in cui si tiene conto sia delle variabili ausiliarie che dell'informazione dei clusters stessi. Si sono presentati modelli sia per il caso in cui si ha a disposizione solamente la variabile di studio e le informazioni sull'effetto dei clusters, sia se si dispone anche di variabili ausiliarie osservate per tutto il campione: in questo caso, si sfrutta l'informazione aggiuntiva per stimare $\overline{Y}$.
Se invece i dati mancanti seguono un meccanismo di tipo MAR, sono stati presentati tre gruppi di stimatori accomunati dallo stesso tipo d'imputazione usata: l'imputazione nearest neighbor, l'imputazione basata sulla pseudo-verosimiglianza e l'imputazione ponderata. Il metodo NNI può essere più efficiente di altri, come l'imputazione tramite media, quando la variabile ausiliaria fornisce informazione utile; inoltre non si assume un modello di regressione parametrica tra y e x, quindi è più robusto rispetto a metodi basati sull'imputazione per rapporto e regressione che sottointendono un modello di regressione lineare. Tra i metodi d'imputazione basati sulla pseudo-verosimiglianza empirica, si hanno quelli che usano la media di pseudo-verosimiglianza e quelli che usano l'imputazione casuale di pseudo-verosimiglianza; in aggiunta, si è mostrata l'imputazione doppiamente robusta, la quale permette allo stimatore di $\overline{Y}$ di essere asintoticamente non distorto ed efficiente quando almeno uno tra il modello di regressione e il modello per la probabilità di risposta è corretto. Infine si è visto l'approccio dell'imputazione ponderata, basata sulle probabilità di risposta stimate: in questo caso, i valori imputati dipendono dai valori della funzione Kernel.

2.5 Nota bibliografica

In questo capitolo, si sono indicati diversi metodi per stimare la media di una variabile di interesse y nel caso di dati mancanti: gli stimatori sono stati suddivisi per tecniche d'imputazione e per meccanismo di non risposta assunto a priori.

Per un meccanismo di non risposta di tipo MCAR: se l'imputazione è tramite media e rapporto, in letteratura si può fare riferimento all'articolo di Kadilar e Cingi (2008) oppure si può approfondire la sola imputazione per rapporto nella pubblicazione di Toutenburg, Srivastava e Shalabh (2008); un articolo a cui fare riferimento, nel caso si utilizzano le tecniche d'imputazione basate sugli stimatori indiretti, è quello di Rueda, Gonzàlez e Arcos (2005). Sempre sotto assunzione MCAR, abbiamo mostrato l'imputazione multipla nel contesto del campionamento stratificato: per approfondimenti si può far riferimento all'articolo di Bjørnstad (2007); nel contesto del campionamento per clusters, 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 usa l'imputazione 'nearest neighbor', ulteriori approfondimenti si possono trovare nell'articolo di Shao e Wang (2008); per quanto riguarda le tecniche d'imputazione che usano la pseudo-verosimiglianza, è possibile riferirsi alle pubblicazioni di Fang, Hong e Shao (2009); Qin, Shao e Zhang (2008). In merito all'imputazione ponderata che usa la 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] Diana, G. e Perri, F. (2010) Improved estimators of the population mean for missing data, *Communications in Statistics-Theory and Methods,* 39, 18, pp. 3245-3251.

[7] 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.

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

[9] 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.

[10] Kadilar, 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.

[11] 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.

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

[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) "A regression approach to the estimation of the finite population mean in the presence of non-response", *Australian & New Zealand Journal of Statistics*, 50, 4, pp. 395-408.

[20] Singh, H. P. e Kumar, S. (2010a) "Estimation of mean in presence of non-response using two phase sampling scheme", *Statistical Papers,* 51, 3, pp. 559-582.

[21] Singh, H. P., Kumar, S. e Kozak, 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.

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

[23] 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.

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

[25] 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.