

Capitolo 7.

La qualità dei dati: l'imputazione dei dati mancanti

Outline

- il concetto di non risposta
- la notazione
- il meccanismo generatore dei dati mancanti
- il pattern dei dati mancanti
- i metodi di imputazione

7.1. Il concetto di non risposta

Con il termine non-risposta (*non response*) si intendono una moltitudine di situazioni in cui il dato non viene osservato.

Principalmente si distinguono due tipi di non risposta:

- *Non Risposta Totale (Unit Non-Response)*: non si ha nessuna informazione disponibile (rilevata) per unità campionarie “eligibili”. Le ragioni possono essere varie e dipendono ovviamente dalle modalità di raccolta dei dati.
- *Non Risposta Parziale (Item Non-Response)* le informazioni rilevate dal rispondente sono tali da essere ritenute accettabili per il data base, ma

alcune informazioni risultano mancanti. I motivi possono essere diversi.

Le metodologie che si adottano per trattare le due tipologie di mancata risposta sono sostanzialmente diverse. In questo capitolo ci occuperemo del **trattamento di non risposte parziali**.

Problemi creati dai MISSING DATA?

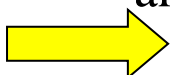
- un *incremento nella variabilità degli stimatori*, dovuta ad una riduzione della base campionaria di analisi e/o all'applicazioni di metodi per il trattamento della stessa,
- *stimatori distorti*, se i rispondenti differiscono sistematicamente dai non rispondenti rispetto a certe caratteristiche di interesse.

Example: on the following data if we estimate a regression model to predict **home ownership** based on **age** and **educational background**

Missing data imply loss of information

<i>Case</i>	<i>Age</i>	<i>Gender</i>	<i>Home</i>	<i>Educ</i>
<i>1</i>	.	<i>F</i>	<i>NO</i>	<i>16</i>
<i>2</i>	.	<i>M</i>	<i>NO</i>	.
<i>3</i>	<i>39</i>	<i>M</i>	.	<i>20</i>
<i>4</i>	.	<i>F</i>	<i>YES</i>	.
<i>5</i>	<i>20</i>	<i>M</i>	<i>YES</i>	<i>14</i>
<i>6</i>	<i>20</i>	<i>F</i>	<i>NO</i>	<i>10</i>
<i>7</i>	<i>37</i>	<i>M</i>	<i>YES</i>	<i>18</i>
<i>8</i>	<i>39</i>	<i>M</i>	<i>YES</i>	<i>20</i>

On this data (n=8) we would estimate a regression model to predict **Home ownership** based on **Age** and **Educational background**.

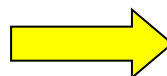


the total number of cases without missing, in the analysis, is reduced drastically (50%).

Missing data can lead to misleading results (bias)

<i>Case</i>	<i>Age</i>	<i>Gender</i>
<i>1</i>	<i>42</i>	<i>F</i>
<i>2</i>	<i>43</i>	<i>M</i>
<i>3</i>	<i>39</i>	<i>M</i>
<i>4</i>	<i>40</i>	<i>F</i>
<i>5</i>	<i>20</i>	<i>M</i>
<i>6</i>	<i>20</i>	<i>F</i>
<i>7</i>	<i>37</i>	<i>M</i>
<i>8</i>	<i>39</i>	<i>M</i>
<i>Average AGE=32,5</i>		

<i>Case</i>	<i>Age</i>	<i>Gender</i>
<i>1</i>	<i>.</i>	<i>F</i>
<i>2</i>	<i>.</i>	<i>M</i>
<i>3</i>	<i>39</i>	<i>M</i>
<i>4</i>	<i>.</i>	<i>F</i>
<i>5</i>	<i>20</i>	<i>M</i>
<i>6</i>	<i>20</i>	<i>F</i>
<i>7</i>	<i>37</i>	<i>M</i>
<i>8</i>	<i>39</i>	<i>M</i>
<i>Average AGE=27</i>		

 ***BIAS***

Besides theoretical problems, data analysts also meet a big practical problem when dealing with datasets affected by missing values: tools for effectively dealing with them are not readily available. (Paul D. Allison, 2001)

7.2. Questioni metodologiche preliminari

7.2.1 Notazione

- **Y** una matrice di dimensione $n \times p$ di dati non completamente osservata
- **Y_{obs}** la parte osservata di **Y**
- **Y_{mis}** la parte mancante di **Y**
- **R** la matrice di dimensione $n \times p$ degli indicatori di risposta, i cui elementi $R_{i,j}$ assumono valore **zero** se **Y** osservata, **uno** se **Y** mancante.

7.2.2 Meccanismo generatore dei dati mancanti

Il meccanismo generatore dei dati mancanti è MAR (*Missing At Random*) se la probabilità che una data osservazione sia mancante dipende da Y_{obs} ma non da Y_{mis} .

$$p(R | Y) = p(R | Y_{\text{obs}}) \text{ for all } Y$$

Ex.: Income is missing and marital status is fully observed. The probability of missing on income depends on marital status, but within each marital status the probability of missing on income does not depend on income. In such a case the nonresponse bias can be controlled by an analysis that stratifies on marital status.

Caso particolare di meccanismo MAR: è il meccanismo **MCAR** (*Missing Completely At Random*); in tal caso la probabilità che una data osservazione sia mancante non dipende né da Y_{obs} né da Y_{mis} , ovvero i dati mancanti sono semplicemente un campione casuale dei dati osservabili.

$$p(R | Y) = p(R) \text{ for all } Y \quad \text{“caso ideale!!”}$$

Se il processo generatore dei dati mancanti è MAR e il parametro del meccanismo generatore dei dati mancanti ed il parametro del modello sui dati completi sono distinti, allora il processo generatore dei dati mancanti è **ignorabile**.

Formalmente l'assunzione di un meccanismo MAR implica che la distribuzione di \mathbf{R} può dipendere da Y_{obs} ma non da Y_{mis}

$$p(\mathbf{R}|Y_{obs}, Y_{mis}) = p(\mathbf{R}|Y_{obs})$$

Definizione formale di **ignorabilità** del meccanismo dei dati mancanti

Siano θ e ϕ rispettivamente i parametri del modello dei dati ed i parametri del meccanismo generatore dei dati mancanti,

- se tali parametri sono distinti, ovvero la conoscenza dell'uno non fornisce alcuna informazione sull'altro,
- il meccanismo è MAR



il meccanismo dei dati mancanti è ignorabile.

Tale assunzione permette di stimare il parametro incognito θ senza specificare la distribuzione del meccanismo generatore dei dati mancanti.

Meccanismo Missing Not At Random (MNAR):

se il processo generatore dei dati mancanti su Y dipende da Y (obs o mis).

Ex: Income is missing the probability of having missing data increases with the increasing of income. The probability of missing on income depends on income itself.

....more on mechanism

Exploring missing data mechanism

The missing data mechanism is important, because the **different types of missing data require different treatment**. For

- ▶ **MCAR**, analyzing only the complete cases will not result in biased parameter estimates (only too larger standard errors).
- ▶ **MAR** or **MNAR**, analyzing only complete cases can lead to biased parameter estimates.

7.2.3 Pattern dei dati mancanti

Il *pattern* dei dati mancanti (*missing data pattern*) è la combinazione degli stati di risposta (osservato o mancante) associato alla matrice dei dati \mathbf{Y} ; la matrice \mathbf{R} descrive il pattern dei dati mancanti.

Supponiamo di aver rilevato Y_1 , Y_2 , e Y_3 e di raggruppare le osservazioni in base ai comportamenti di risposta, formando otto gruppi distinti, omogenei al loro interno rispetto al comportamento di risposta.

Figura A1 : *Pattern dei dati mancanti ("X" indica valori osservati nel gruppo e "." indica dati mancanti)*

O bs	Y 1	Y 2	Y 3
1	X	X	X
2	X	X	.
3	X	.	X
4	X	.	.
5	.	X	X
6	.	X	.
7	.	.	X
8	.	.	.

Un caso particolare di pattern di dati mancanti è il **pattern monotono**. Siano Y_1, Y_2, \dots, Y_p le variabili rilevate (ordinate), si dice che il pattern dei dati mancanti è monotono quando il fatto che la

variabile Y_j è mancante per una certa unità implica che tutte le variabili che seguono Y_k , $k > j$, siano mancanti per tutte le unità. Alternativamente, quando una variabile Y_j è osservata per una particolare unità anche tutte le variabili antecedenti Y_k , $k < j$, risultano osservate per tutte le unità.

Figura A2: *Pattern dei dati mancanti monotono*

Gruppo	Y1	Y2	Y3
1	X	X	X
2	X	X	.
3	X	.	.

7.3. Imputazione dei dati mancanti

Per le non risposte parziali la procedura di compensazione comunemente usata è l'**imputazione**, che consiste nell'assegnazione di un valore sostitutivo del dato mancante, al fine di ripristinare la “completezza” della matrice dei dati.

Numerosi sono i metodi di imputazione proposti in letteratura, possiamo considerare tre classi di metodi:

- **metodi deduttivi**, nei quali il valore imputato è dedotto da informazioni o relazioni note;
- **metodi deterministici**, nei quali imputazioni ripetute per unità aventi “le stesse

caratteristiche” considerate producono sempre gli stessi valori imputati;

- **metodi stocastici**, nei quali imputazioni ripetute per unità aventi le stesse caratteristiche considerate possono produrre differenti valori imputati; per la presenza di una componente aleatoria, corrispondente ad uno schema probabilistico associato al particolare metodo d'imputazione prescelto.

Tutti i metodi di imputazione per le mancate risposte parziali (ad eccezione dei metodi deduttivi) si basano implicitamente o esplicitamente sull'assunzione che il meccanismo generatore dei dati mancanti sia MAR.

7.3.1 Metodi deduttivi

Metodo

Sfruttare le informazioni presenti nel data set in modo da poter dedurre il valore da sostituire al dato mancante

Campi di applicazione

Metodi che presuppongono la definizione di modelli di comportamento specifici del fenomeno in oggetto. L'applicazione del metodo è molto legata a valutazioni soggettive sul fenomeno oggetto di studio e spesso dipende dal grado di conoscenza del data set su cui si sta lavorando.

Esempi

- Il rispondente ha un'età inferiore ai 17 anni; se ne deduce che il valore da imputare alla variabile “patente di guida” è “NO”.
- Conoscendo il codice fiscale di un individuo, si possono ottenere anno di nascita, sesso, età.
- Conoscendo il reddito lordo ed il regime di tassazione/contribuzione, è possibile ottenere il reddito netto di un individuo.

Vantaggi

- Possibilità di imputare valori corrispondenti a quelli veri. Rappresenta la soluzione che più si

avvicina al ricontattare l'unità che presenta dati mancanti.

Svantaggi

- Possibilità di distorsioni nel caso in cui il metodo è applicato in maniera errata
- E' necessario avere un buon grado di conoscenza dei dati e delle relazioni esistenti al loro interno.

Software

Nessun software generalizzato per database permette di effettuare automaticamente imputazioni deduttive. Servono algoritmi “ad hoc”

7.3.2 Metodi deterministici

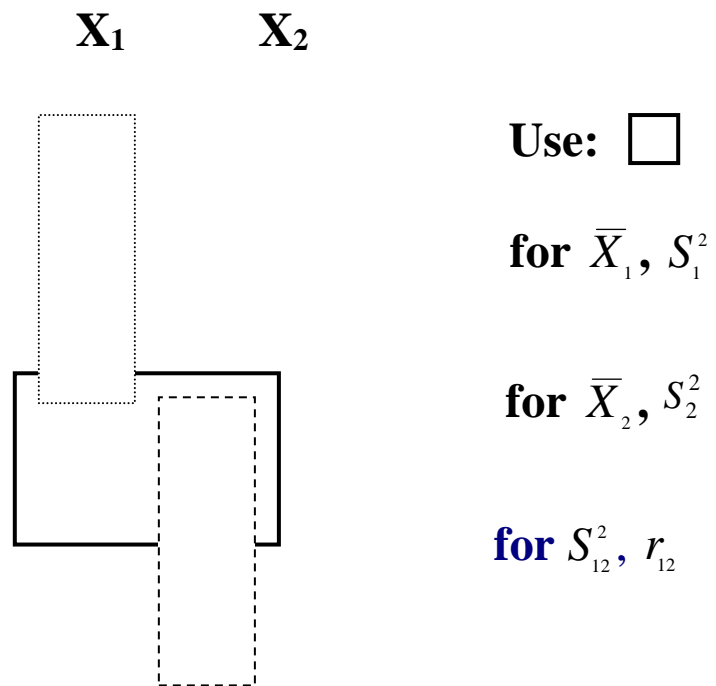
- Casi Completi (Complete Cases, CC)
- Casi Disponibili (Available Cases, AC)
- Imputazione deterministica di valori

medi (medie non condizionate e condizionate)

Il metodo CC

Metodo

Il metodo CC elimina tutte le osservazioni in cui è presente almeno un dato mancante, quindi riconduce la matrice dei dati ad una matrice completa riducendo il numero delle osservazioni



Campi di applicazione Tutti, ma attenzione.

Vantaggi

- Semplicità e possibilità di ottenere una matrice di dati rettangolare sulla quale applicare metodi standard.

Svantaggi

- riduce la base campionaria e quindi rende inefficienti le stime;
- se il meccanismo generatore dei dati mancanti non è MCAR, ma soltanto MAR, con questo metodo si introduce distorsione nelle stime.

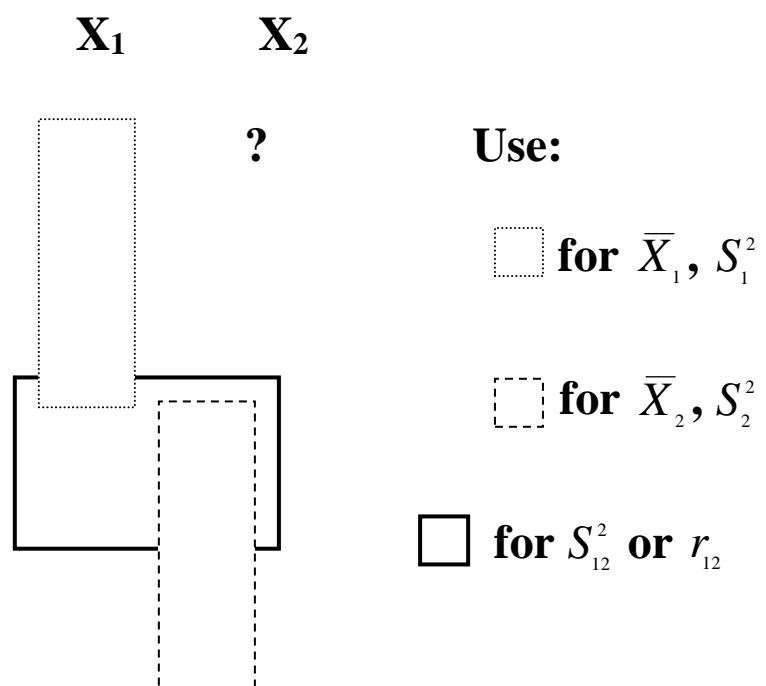
Software

Nessun software specifico, questo metodo è quello utilizzato da tutti i *software standard* nell'applicazione di metodi statistici.

Il metodo AC

Metodo

Il metodo non prevede nessuna operazione da svolgere a priori, dipende dal tipo di analisi. Ad esempio si usano tutte le osservazioni disponibili su ogni variabile per la stima di media e varianza, mentre per la covarianza si utilizzano le coppie di osservazioni complete.



Campi di applicazione Tutti, ma attenzione.

Vantaggi

- Semplicità per ogni tipo di analisi si utilizzano tutti i dati completi a disposizione

Svantaggi

- Per ogni tipo di analisi si possono avere basi campionarie diverse (può portare a correlazioni fuori range o a matrici di correlazioni non positive definite);
 - riduce la base campionaria e quindi rende inefficienti le stime;
 - se il meccanismo generatore dei dati mancanti non è MCAR, ma soltanto MAR, con questo metodo si introduce distorsione nelle stime.

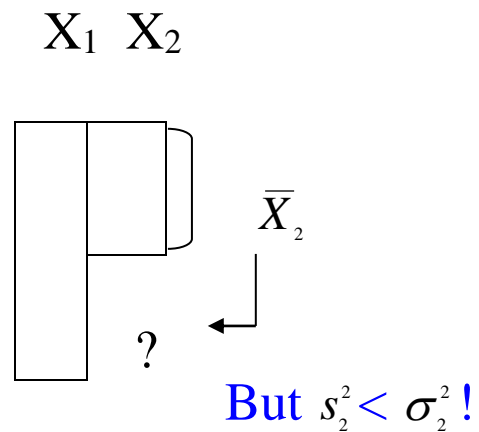
Software: Nessun software specifico.

Imputazione deterministica con media

(mean imputation overall)

Metodo

Con questo metodo si sostituiscono tutte le mancate risposte nella variabile y con un unico valore, la media calcolata sul totale dei rispondenti. E' un metodo che può essere utilizzato solo per le variabili quantitative (per le variabili qualitative al posto del valor medio si può imputare la moda).



Campi di applicazione

E' consigliabile utilizzare questo metodo solo nei casi in cui: il numero dei dati mancanti per ciascuna variabile è esiguo; lo scopo dell'analisi è limitato alla stima di medie e totali; sembrano esistere poche relazioni tra le variabili.

Vantaggi

- Preserva la media dei rispondenti.
- Facile da applicare e da spiegare.

Svantaggi

- Introduce una seria distorsione nella distribuzione della variabile, creando un picco artificiale in corrispondenza del suo valor medio.
- Non dà buoni risultati nella stima della varianza.
- Provoca distorsioni nelle relazioni tra le variabili.

Imputazione deterministica con medie condizionate

Metodo

Si divide il campione totale in classi di imputazione in base ai valori assunti da prefissate variabili ausiliarie considerate esplicative di y e si calcola la media dei rispondenti (r) della variabile y all'interno di ogni classe. Ciascuna media viene poi assegnata ai valori mancanti in unità appartenenti alla stessa classe: $y_{mhi} = \bar{y}_{rh}$, per l' i -esimo non rispondente della classe h ($h = 1, 2 \dots H$).

Per individuare la migliore classificazione possibile possono essere utilizzati diversi metodi di analisi multivariata.

Campi di applicazione

L'applicazione di questo metodo può essere utile nei casi in cui: l'obiettivo dell'analisi è rappresentato dalla stima di medie e aggregati; sembrano esistere poche relazioni tra le variabili

Vantaggi

- Può ridurre le distorsioni generate dalle mancate risposte (se la scelta delle classi di imputazione è stata effettuata in modo appropriato).
- Rapido, semplice da applicare e da spiegare, una volta definite le classi di imputazione.

Svantaggi

- Introduce distorsioni (sebbene in maniera meno evidente del metodo overall) nella distribuzione della variabile, creando una serie di picchi artificiali in corrispondenza della media di ciascuna classe.
- Provoca un'attenuazione della varianza della distribuzione dovuta al fatto che i valori imputati riflettono solo la parte di variabilità tra le classi (*between*) ma non quella all'interno delle classi (*within*).
- Provoca distorsioni nelle relazioni tra le variabili non considerate per la definizione delle classi di imputazione.

Software

Una volta determinate le classi, il metodo può essere facilmente applicato attraverso un qualsiasi programma di analisi o software per database.

Imputazione con regressione

(Predictive regression imputation)

Metodo

Si utilizzano i valori dei rispondenti per stimare i parametri della regressione per la variabile di studio y su prefissate variabili ausiliarie considerate esplicative di y . Le determinazioni della y sono, poi, imputate come valori stimati dell'equazione di regressione:
$$y_{mi} = \beta_{r0} + \sum_j \beta_{rj} z_{mij} .$$

Anche questo metodo può richiedere la *suddivisione in classi delle unità*. Infatti diversi modelli possono essere necessari in ogni classe, in quanto (soprattutto per variabili di tipo economico)

le relazioni tra y e le covariate possono cambiare molto da strato a strato.

Campi di applicazione

Si adatta a situazioni in cui la variabile sulla quale effettuare l'imputazione è quantitativa oppure binaria oltre che naturalmente essere fortemente correlata con altre variabili. E' meno adatto a situazioni in cui le variabili qualitative presentano numerose modalità.

Vantaggi

- Si può fare uso di un numero elevato di variabili, sia quantitative che qualitative, in modo da ridurre, più che con altri metodi, le distorsioni generate dalle mancate risposte.

- Preserva bene le relazioni delle variabili usate nel modello.

Svantaggi

- Introduce distorsioni nella distribuzione della variabile (sebbene meno del metodo di imputazione per medie condizionate).
- Metodo deterministico, non preserva sufficientemente la variabilità delle distribuzioni marginali.
- Provoca distorsioni nelle relazioni tra le variabili non utilizzate nel modello.
- E' necessario mettere a punto un modello diverso per ogni variabile sulla quale si intende effettuare imputazioni.

- Nel caso in cui si applica il metodo suddividendo in classi le unità, è necessario stimare molti modelli diversi tra loro, tanti quante sono le celle di imputazione.

- Può richiedere il possesso di conoscenze tecniche molto specifiche per la messa a punto di modelli appropriati.

- Metodo parametrico, richiede assunzioni sulle distribuzioni delle variabili.

- C'è il rischio che possano essere imputati valori non reali.

- È fortemente influenzato dalla presenza di dati anomali.

Software

La maggior parte dei pacchetti statistici per l'analisi dei dati fornisce routine generalizzate per la costruzione di modelli di regressione più o meno complessi, il che facilita lo sviluppo di programmi che implementano tale metodo.

Imputazione dal più vicino donatore (Nearest-Neighbour imputation)

Metodo

Si sostituisce ogni dato mancante con il valore del rispondente “più vicino”. Quest’ultimo è determinato per mezzo di una funzione di distanza applicata alle variabili ausiliarie.

La procedura è la seguente:

1. Calcolare la distanza (considerando i valori assunti sulle variabili ausiliarie, poiché in genere i dati vengono stratificati) tra l’unità del campione con mancata risposta e tutte le altre unità rispondenti usando un’appropriata *funzione di distanza*.

2. Determinare l'unità più vicina all'unità di interesse.

3. Utilizzare il valore dell'unità "più vicina" per effettuare l'imputazione.

Quando si usa una sola variabile ausiliaria si può ordinare il campione in base ai valori da essa assunti; in questo caso ogni donatore è selezionato calcolando la più piccola differenza assoluta tra non rispondente e rispondenti. Quando, invece, sono disponibili molte variabili ausiliarie possono essere trasformate tutte nei loro ranghi.

a) Le varianti di questo metodo possono essere ricondotte all'uso di differenti funzioni di distanza.

A seconda dell'utilizzo che viene fatto dei donatori selezionati, si possono distinguere *due versioni* del metodo:

- Ogni donatore viene usato per ogni valore mancante nel recipiente;
- Uno stesso donatore viene usato per tutti i valori mancanti nel recipiente.

Campi di applicazione

Il metodo è particolarmente adatto nel caso di :
indagini dove la percentuale delle mancate risposte è esigua; indagini su larga scala in cui trovare un donatore per molte variabili simultaneamente sia più agevole, con notevoli vantaggi in termini di qualità dei risultati; indagini con informazioni di carattere

quantitativo utilizzabili nelle funzioni di distanza; indagini in cui esistano relazioni fra variabili difficilmente esplicabili mediante “modelli” (statistici, economici etc.) e sia al contempo necessario preservare la variabilità delle distribuzioni marginali e congiunte.

Si sconsiglia, invece, l’utilizzo del metodo nel caso di: indagini con un numero elevato di mancate risposte (specialmente se una stessa risposta risulta mancante per una grossa percentuale di casi); indagini nelle quali si hanno solo informazioni di carattere quantitativo; indagini di piccole dimensioni.

Vantaggi

- Garantisce, in buona misura, il mantenimento delle relazioni tra variabili anche all'interno di data sets complessi, specialmente nei casi in cui uno stesso donatore è utilizzato per predire simultaneamente molte mancate risposte.
- Potenzialmente è in grado di gestire simultaneamente le informazioni relative ad un numero elevato di variabili.

Svantaggi

- Può provocare distorsioni di varia entità nella distribuzione delle variabili, sebbene i valori imputati includano una parte “residuale” implicitamente osservata nei donatori. In tal senso,

la qualità delle imputazioni dipende dalla “ricchezza” del serbatoio dei donatori.

- Richiede una preparazione dei dati tale da assicurare che le variabili non abbiano effetti diseguali sulle misure di distanza.

Software

Per i casi più semplici (ad esempio una sola variabile ausiliaria) si può utilizzare un qualsiasi package statistico dotato di una funzione di ordinamento. Per i casi più complessi è necessario avere a disposizione delle macro per il calcolo della funzione di distanza, anche se in alcuni package sono già predisposte le più comuni funzioni di distanza.

7.3.3 Metodi stocastici

- Imputazione stocastica singola
- Imputazione multipla

Imputazione stocastica singola

Si ottengono risultati più soddisfacenti in termini di distribuzioni marginali dei dati completati e si possono ridurre le distorsioni sulle associazioni tra caratteri. Gli errori standard calcolati sono più veritieri di quelli calcolati con metodi deterministici, ma sono ancora troppo piccoli, infatti non si tiene assolutamente conto dell'incertezza associata al dato imputato anzi, si

tratta esattamente come se fosse un dato vero (rilevato). Vediamo alcuni metodi.

Example: Regress Y2 on Y1 from complete cases

$$\hat{y}_{i,2} = \hat{E}(X_{i,2} | X_{i,1}) + r_i$$

where

$$r_i \approx N(0, s_{22|1}^2)$$

Imputazione con donatore casuale all'interno delle classi (Random donor imputation within classes)

Metodo

Creazione di classi di imputazione all'interno delle quali poi si sostituiscono i dati mancanti con quelli disponibili selezionati casualmente all'interno della medesima classe.

I migliori risultati si ottengono selezionando i donatori mediante un campionamento senza ripetizione all'interno delle classi.

Esistono diverse versioni del metodo a seconda che:

- le imputazioni tengono conto o meno dei vincoli;
- le imputazioni siano di tipo sequenziale (dato un record con più valori mancanti, viene utilizzato un donatore diverso per ogni mancata risposta) o congiunto (dato un record con più valori mancanti, viene utilizzato un solo donatore per integrarne simultaneamente le mancate risposte).

Campi di applicazione

Questo metodo va usato possibilmente nei casi in cui si lavora con data set di grosse dimensioni (in modo di avere molti donatori), ma con relativamente poche variabili (per ridurre l'entità delle distorsioni delle relazioni).

Vantaggi

- Il valore sostituito al posto del dato mancante è un valore “reale”.
- In genere il donatore proviene da un’unità “simile”, a differenza di quanto accade imputando senza classi di imputazione.
- Uno stesso donatore viene utilizzato una sola volta.
- Se uno stesso donatore è usato per imputare tutte le mancate risposte parziali di un record, vengono preservate le relazioni fra le variabili.
- Maggiore è il numero di classi, maggiori sono le possibilità di imputare un valore da un’unità vicina.

Svantaggi

- Per ottenere un'imputazione da casi vicini è necessario un numero molto elevato di classi di imputazione, che comporta la messa a punto di complicate strategie di stratificazione.
- Possibile perdita di dettaglio nella formazione delle classi di imputazione dovuta alla eventuale conversione di dati continui in gruppi discreti di dati.

Software

Una volta determinate le classi, è possibile creare programmi in grado di effettuare imputazioni casuali sfruttando sistemi per la gestione di database oppure

utilizzare i moduli presenti all'interno dei più diffusi software generalizzati.

Imputazione con regressione casuale (Random regression imputation)

Metodo

Questa tecnica costituisce la versione stocastica dell'imputazione con regressione esposta in precedenza, in cui i valori imputati sono sempre stimati con l'equazione di regressione nella quale si aggiunge, però, la componente residuale e_{mi} . In questo tipo di modello sono cruciali le assunzioni per la determinazione dei termini residui e_{mi} .

A tale proposito sono state proposte le seguenti soluzioni:

- 1) ipotizzare che i residui abbiano una distribuzione normale e rispettino il requisito di

omoschedasticità e sceglierli, a caso, dalla distribuzione con media zero e varianza uguale a quella residua della regressione;

2) ipotizzare che i residui provengano dalla stessa distribuzione non specificata dei rispondenti e selezionarli casualmente dai residui di questi ultimi;

3) infine, se si hanno dubbi sulla linearità e sull'additività delle componenti del modello di regressione, si scelgono da quei rispondenti con valori simili nelle variabili ausiliarie.

Campi di applicazione

Si adatta a situazioni in cui la variabile sulla quale effettuare l'imputazione è quantitativa oppure binaria oltre che naturalmente essere fortemente

correlata con altre variabili. E' meno adatto, invece, a situazioni in cui le variabili qualitative presentano numerose modalità.

Vantaggi

- Si può fare uso di un numero elevato di variabili, sia quantitative che qualitative, in modo da ridurre, più che con altri metodi, la distorsioni generate dalle mancate risposte.
- I valori imputati non generano distorsioni nella distribuzione della variabile.
- Rispetto alla versione deterministica del metodo preserva meglio la variabilità della distribuzione.

Svantaggi

- Provoca distorsioni nelle relazioni tra le variabili non utilizzate nel modello.
- E' necessario mettere a punto un modello diverso per ogni variabile sulla quale si intende effettuare imputazioni.
- Nel caso si utilizzino classi di imputazione, deve essere stimato un modello di verso per ogni classe.
- Può richiedere il possesso di conoscenze tecniche molto specifiche per la messa a punto di modelli appropriati.
- C'è il rischio che possano essere imputati valori non reali.

- È fortemente influenzato dalla presenza di dati anomali.

Software

Per effettuare imputazioni con questo metodo è necessario avere a disposizione un generatore di numeri casuali in aggiunta al software richiesto per l'imputazione con regressione.

Summarising on (single) imputation

Imputations should:

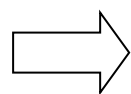
- condition on observed variables
- be multivariate to preserve associations between missing variables

- generally be draws rather than means

- **Key problem:**

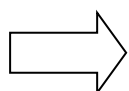
single imputations do not account for imputation uncertainty in se's

- Imputation “makes up” the missing data



treats imputed values as the truth

- For statistical inference (standard errors, P-Values, confidence intervals) need methods that account for imputation error.



Multiple Imputation (Rubin, 1987)

Imputazione multipla

Il metodo più utilizzato per ottenere risultati inferenziali validi nelle ricerche applicate in presenza di dati mancanti.

L'idea di base: generare più di un valore ($m > 2$) da imputare per ogni dato mancante, in modo che le matrici dei dati completi da analizzare con metodi e software standard siano m . I risultati delle m analisi distinte vengono poi combinati con opportune regole in modo da produrre risultati inferenziali che tengano conto dell'incertezza causata dai dati mancanti.

Vediamo a questo punto quali sono i metodi per generare le imputazioni, a questo proposito si deve sottolineare che i metodi per la generazione delle imputazioni dipendono dal tipo di pattern dei dati mancanti.

- Nel caso di pattern monotoni, si possono applicare sia metodi parametrici come modelli di regressione (Rubin 1987) e stime di massima verosimiglianza, in particolare Anderson (1957) propose di fattorizzare la funzione di verosimiglianza per formulare in forma esplicitamente stimatori di massima verosimiglianza; sia metodi non parametrici come il propensity scores (Rubin 1987).

- Per data set con pattern di dati mancanti qualsiasi si ricorre a metodi iterativi. La tecnica generale per determinare stime di massima verosimiglianza per modelli parametrici in caso di dati incompleti è l'algoritmo EM (Dempster, Laird, Rubin, 1977), in alternativa si ricorre a metodi MCMC (Schafer 1997), in questo caso le imputazioni multiple vengono generate come valori simulati da una distribuzione predittiva a posteriori per i dati mancanti.

- Una ulteriore alternativa è quella di ricondurre il data set con dati mancanti secondo un pattern qualsiasi ad un pattern monotono con metodi MCMC e poi applicare alla matrice parzialmente completata metodi propri dei pattern monotoni.

7.4 EVENTUALI APPROFONDIMENTI¹

Sequential Regression Imputation Method

(*Survey Methodology*, June 2001).

Y1	Y2	Yk	X
?		?		
	?			
		?	?	
?				
	?			
			?	
?		?		

X is the set of completely observed variables

Round 0: variables are ordered according to the number of missing, the variable with the fewest

¹ Questa sezione costituisce un **approfondimento**.

number of missing values, say Y_1 , the following Y_2 and so on.

The sequence of imputations is determined by the following factorisation:

$$[Y_1|X] [Y_2|X, Y_1] \dots [Y_k|X, Y_1, \dots, Y_{k-1}]$$

Round 1: starts regressing the variable with the fewest number of missing values, say Y_1 , on X , and imputing the missing values with the appropriate regression model.

Round 2: After Y_1 has been completed, the variable with the fewest number of missing values is considered, say Y_2 ; observed Y_2 values are regressed on (X, Y_1) and the missing values are imputed, and so on.

.....

Round c: The imputation process is then repeated, modifying the predictor set to include X and all the Y variables already imputed.



The data set is completed

The imputation process is then repeated, modifying the predictor set to include all the Y variables except the one used as the dependent variable

Repeated cycles continue for a pre-specified number of rounds, or until stable imputed values occur (convergence in distribution).

The form of model regression depends on the nature of Y: a general linear regression for continuous variables, a logistic regression for binary variables, etc.

Inferenza combinata da archivi imputati

Se m è il numero di imputazioni effettuate e Q il parametro incognito della distribuzione al termine della procedura di imputazione si hanno m coppie di valori composte dalla stima puntuale del parametro di interesse e la stima della varianza del parametro stesso. Siano \hat{Q}_i e \hat{U}_i la stima puntuale del parametro e la varianza stimata relativamente alla matrice di dati della i -esima imputazione $i=1, 2, \dots, m$, la stima puntuale di Q relativa alle imputazioni effettuate è data dalla media delle singole stime calcolate sulle m matrici completate.

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

Denominata \bar{U} la *within-imputation variance*

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

e B la *between-imputation variance*

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

si determina la varianza totale associata a \bar{Q} (Rubin 1987)

Le procedure inferenziali si basano sulla statistica $(Q - \bar{Q})T^{-(1/2)}$ che approssimativamente si distribuisce come una *t-Student* con v_m gradi di libertà (Rubin 1987), dove

$$v_m = (m-1) \left[1 + \frac{\bar{U}}{(1+m^{-1})B} \right]^2$$

Nel caso in cui i gradi di libertà dei dati completi v_0 sia piccolo e che la proporzione dei dati mancanti sia modesta può accadere che i gradi di libertà calcolati v_m risultino maggiori di v_0 , in questo caso la letteratura (Barnard e Rubin, 1999) raccomanda di utilizzare i gradi di libertà aggiustati:

$$v_m^* = \left[\frac{1}{v_m} + \frac{1}{\hat{v}_{obs}} \right]^{-1}$$

Dove

$$\hat{v}_{obs} = (1 - \gamma)v_0(v_0 + 1)/(v_0 + 3)$$

$$\gamma = (1 + m^{-1})B/T .$$

Note that the MI procedure uses the adjusted degrees of freedom, v_m^* , for inference.

I gradi di libertà v_m dipendono dal numero di imputazioni m e dal rapporto:

$$r = \frac{(1 + m^{-1})B}{\bar{U}}$$

Il rapporto r è definito incremento relativo della varianza dovuto alla non risposta (Rubin 1987); quando non ci sono dati mancanti B è zero e quindi anche r è zero. Per un numero di imputazioni elevato e/o r piccolo il numero di gradi di libertà

diviene elevato e la distribuzione di $(Q - \bar{Q})r^{-1/2}$ diviene approssimativamente normale.

Un'altra statistica di cui tenere conto è la frazione di informazione mancante relativamente a Q :

$$\lambda = \frac{r + 2/(v + 3)}{r + 1}$$

Le statistiche r and λ sono utili nelle procedure di diagnostica per stabilire in quale misura i dati mancanti contribuiscono all'incertezza sul parametro Q .

Efficienza dell'imputazione multipla

L'efficienza della procedura di imputazione cresce al crescere di m , ovviamente la procedura di imputazione sarebbe pienamente efficiente se il numero di imputazioni fosse infinito. Il fatto che m sia nella realtà un valore finito rende inferiore l'efficienza, l'efficienza relativa (*relative*

efficiency, RE) è approssimativamente funzione di m e λ (Rubin 1987, p. 114).

$$RE = \left(1 + \frac{\lambda}{m}\right)^{-1}$$

Nella tabella seguente sono riportati i valori di efficienza relativa per combinazioni significative di m e λ . E' evidente che quando la frazione di informazione mancante è piccola è necessario un numero esiguo di imputazioni.

Tab. A3 : *Efficienza relativa*

		λ				
m		10%	20%	30%	50%	70%
3	0.96 77	0.93 75	0.90 91	0.85 71	0.81 08	
5	0.98 04	0.96 15	0.94 34	0.90 91	0.87 72	
10	0.99 01	0.98 04	0.97 09	0.95 24	0.93 46	
20	0.99 50	0.99 01	0.98 52	0.97 56	0.96 62	

Summarising

Multiple Imputation (MI): three steps

1. **Imputation or Fill-in Phase:** Missing values are imputed, forming a complete data set. This process is repeated m times.
2. **Analysis Phase:** Each of the m complete data sets is then analyzed using a statistical model
3. **Pooling Phase:** The parameter estimates (e.g. coefficients and standard errors) obtained from each analyzed data set are then combined for inference.

Per i più interessati

Da

<http://www.restore.ac.uk/PEAS/imputation.php>

Table 6.1 summarises some of the procedures available for handling missing data in the packages featured on this site.

	SAS	SPSS	Stata	R
missing value patterns	MI	MVA	nmissing (dm67) mvmissing(dm91)	md.pattern(mice) prelim.norm(norm)
repeated measures analysis	MIXED not GLM (*)		?	pan
single imputation		MVA	impute uvis(ice)	em.norm da.norm
multiple imputation	MI IMPUTE (IVEWARE)	MVA with EM algorithm	ice (ice)	norm (norm) mice (mice)
post-imputation	MIANALYZE		micombine (r-buddy.gif " alt='warn') mifit and others (st0042)	glm.mids.pool(mice) mi.inference(norm)

(* PROC GLM in SAS does listwise deletion and so does not allow for missing values. This is also true of SPSS repeated measures analyses)

Items in (**brackets**) indicate that the item is a set of contributed procedures. In particular the following research groups have provided routines and their web sites are helpful.

- IVEWARE software for SAS developed by a group at the University of Michigan. This is a set of SAS macros runs a chained equation analysis in SAS. It can also be run as a stand-alone package.
- The MICE library of functions for Splus/R has been written by a group at the University of Leiden to implement chained methods.
- Chained equations have been implemented in Stata by Patrick r-buddy.gif" alt='warn' of the MRC Clinical Trials Unit in London. The original procedure was called mvis and is described in the Stata journal (Royston, P. 2004. *Multiple imputation of missing values. Stata Journal 4: 227-241.*). A more recent version called ice is now available (Royston, P. (2005), *Multiple imputation of missing values: update, Stata Journal 5, 188-201*). Both can be downloaded from the Stata journal by searching net resources for **mvis** and for **ice** respectively.
- Methods based on the multivariate normal distribution have been developed by Jo Schafer of Penn State University using his program NORM can be run as a stand alone resource and is implemented in SAS and in R/Splus. There appear to

be problems with the current implementation in R that are being taken up with the authors.









The SPSS Missing Value Analysis (MVA) software has been criticised in an article in the *American Statistician* [von Hippel P, Volume 58\(2\),160-164](#). The MVA procedure provides two options. The first is a regression method that uses only the observed data in the imputations and the second is based on the normal distribution and resembles the first step of the NORM package. Neither are proper imputations.

Other specialised software for imputation, such as SOLAS, has to be purchased separately and are not featured on the PEAS site. SOLAS links to SPSS and implements various methods, including imputation using a nested procedure. The [SOLAS web site](#) has useful advice on imputation practicalities, and it has now been extended to cover multiple imputation procedures.

The programs [MLWin](#) and [BUGS](#) can be used for imputation. Carpenter and Kenward's [missing data](#) site has details.

Despite having been written a few year's ago, an article by Horton and Lipsitz (*Multiple imputation in practice: comparison of software packages for regression models with missing variables. The American Statistician 2001;55(3):244-254.*) that can be accessed on the [web](#), has lots of useful practical advice on imputation software.

Weblinks

#	Web Link	Hits
	 IVEWARE homepage	
1	Home page of SAS software for multiple imputation using a sequence of regression models	243
	 ICE - imputation by chained equations in Stata	
2	The latest version of Patrick Royston's implementation in Stata of imputation by chained equations (ICE) can be downloaded from his website.	234
	 FAQs for ICE in Stata	
3	A collection of frequently asked questions regarding ICE in Stata maintained at the MRC's Biostatistics Unit website.	233
	 www.multiple-imputation.com	
4	Contains information on multiple imputation software from a variety of authors. Software for Multiple Imputation Using Chained Equations (MICE) can be downloaded from here.	261
	 AMELIA software	
5	Uses EM algorithm with importance sampling. They are similar to the usual posterior-imputation, but computationally quicker.	247
	 WinBUGS	
6	WinBUGS - freely available software for fitting models within the Bayesian paradigm, using MCMC.	451
	 MLwiN	
7	Website for multi-level modelling software MLwiN, hosted at the Centre for Multi-level Modelling in Bristol, UK	244
	 REALCOM	
8		269

REALCOM - freely available Windows software for 'Developing multilevel models for REAListically COMplex social science data', from the Centre for Multi-level Modelling in Bristol, UK

 [REALCOM Impute](#)

9 The free REALCOM Impute package, from researchers at the Centre for Multilevel Modelling (University of Bristol), performs multiple-imputation using multi-level (random-effects) models. 99

Riferimenti Bibliografici classici

- Rubin, D.B. (1976), "Inference and Missing Data." *Biometrika* 63: 581-592.
- Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York : Wiley.
- Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.
- Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, New York : Chapman and Hall.
- Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J., Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27, 85-95.
- Rubin, Donald B.; Little, Roderick J. A. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley. [ISBN 0-471-18386-5](#)
- Enders, Craig K. (2010). *Applied Missing Data Analysis* (1st ed.). New York: Guildford Press. [ISBN 978-1-60623-639-0](#)
- Allison, Paul D. (2001). *Missing Data* (1st ed.). Thousand Oaks: Sage Publications, Inc. [ISBN 978-0761916727](#).
- Schafer, J. L.; Graham, J. W. (2002). "Missing data: Our view of the state of the art". *Psychological Methods* 7 (2): 147–177. [doi:10.1037/1082-989X.7.2.147](#). [PMID 12090408](#).– [edit](#)
- Graham, John W. (2009). "Missing Data Analysis: Making It Work in the Real World". *Annual review of psychology* 60: 549–576.

Per bibliografia recentissima e spunti interessanti vedere

<http://www.missingdata.org.uk/>

<http://www.stat.psu.edu/~jls/index.html>

<http://sitemaker.umich.edu/rlittle/home>

http://www.src.isr.umich.edu/content.aspx?id=research_themes_methodology_108

<http://www.multiple-imputation.com/>