Imputazione multipla di dati mancanti: breve introduzione

Multiple imputation for missing data: a brief introduction

Michela Baccini

Dipartimento di statistica «G. Parenti», Università di Firenze, Unità di biostatistica, CSPO, Firenze; e-mail: mbaccini@inwind.it

Introduzione

Il problema dei dati mancanti è rilevante nella ricerca empirica e l'analisi di un data set con osservazioni incomplete può essere affrontata secondo varie strategie che possono condurre a risultati più o meno validi.

Una pratica piuttosto diffusa nel caso di variabili esplicative con dati mancanti consiste nel fare un'analisi stratificata rispetto alla variabile esplicativa in questione, aggiungendo uno strato per i valori mancanti. Questo metodo è assolutamente sconsigliabile perché conduce a risultati distorti.

La soluzione standard usata dai software statistici in presenza di dati mancanti consiste nell'esclusione dall'analisi di tutte le osservazioni incomplete. Questa procedura ha due inconvenienti: a meno che il meccanismo che ha generato i dati mancanti non sia completamente casuale, i risultati dell'analisi sui soli dati completi sono distorti; inoltre, le stime sono inefficienti a causa della riduzione della base campionaria.

Un approccio alternativo e molto diffuso consiste nell'imputare al posto dei valori mancanti in una variabile la media della variabile stessa (imputazione non condizionata) o una sua media condizionata (imputazione condizionata), ottenuta, per esempio, come valore predetto da un'analisi di regressione. Mentre l'imputazione non condizionata, parimenti all'analisi sulle sole osservazioni complete, può condurre a risultati distorti, l'imputazione condizionata riduce questo problema. Tuttavia la sostituzione del valore medio (sia esso marginale o condizionato) alle osservazioni mancanti causa una sottostima della variabilità dei risultati delle analisi condotte sui dati imputati. La sottostima della variabilità ha una duplice origine:

- i dati mancanti di una variabile sono sostituiti da un valore medio, con la conseguenza che la variabilità marginale delle variabili completate e dei risultati di eventuali analisi multivariate risulta inferiore a quella effettiva;
- i dati imputati vengono trattati come veri e pertanto le analisi statistiche non tengono conto dell'incertezza dovuta all'ignoranza riguardo al vero valore assunto dalle variabili ove l'informazione è mancante.

Il problema relativo al primo punto è in parte colmato se i valori mancanti di ciascuna variabile vengono sostituiti da valori estratti casualmente da una distribuzione (imputazione stocastica singola). In questo modo si ottengono risultati più soddisfacenti in termini di distribuzione dei dati completati, ma gli errori standard delle stime sono ancora troppo piccoli, in quanto anche l'imputazione stocastica singola tratta i dati imputati come veri.

L'imputazione multipla si distingue dall'imputazione singola perché tiene debitamente conto anche dell'incertezza sulle imputazioni (secondo punto), conducendo così a risultati inferenziali validi non solo in termini di stima puntuale dei parametri d'interesse, ma anche in termini, per esempio, di intervalli di confidenza. Essa sostituisce ciascun dato mancante con un certo numero di valori plausibili, rappresentando in questo modo l'incertezza sul vero valore da imputare.

Esistono diverse funzioni per l'imputazione multipla di dati mancanti implementate nei principali software per l'analisi statistica (per esempio la *proc mi* di SAS o la funzione *ice* di STATA). Tuttavia, la ricerca è ancora attiva riguardo alle procedure di imputazione multipla più appropriate da utilizzare in situazioni complesse.

L'assunzione su cui di solito si basano le procedure d'imputazione multipla è quella di processo generatore dei dati mancanti casuale (Missing at Random, MAR). Il meccanismo generatore dei dati mancanti è MAR se, condizionatamente ai dati osservati, la probabilità che un'osservazione sia mancante non dipende dal valore (non osservato) che essa assume. Un caso particolare di meccanismo MAR è il meccanismo di censura completamente casuale in cui i dati mancanti sono un campione casuale dei dati osservabili (MCAR). L'assunzione MAR è particolarmente conveniente perché implica che l'imputazione possa essere effettuata senza specificare un modello per il meccanismo generatore del dato mancante.

Qui descriveremo il razionale del metodo d'imputazione multipla e le sue principali caratteristiche, con riferimento a situazioni in cui l'ipotesi MAR è soddisfatta. Se l'assunzione MAR non è sostenibile, l'imputazione multipla è ancora possibile, ma più complicata poiché è necessario formulare ipotesi sul meccanismo di generazione dei dati mancanti e tenerne conto in fase di imputazione.

Il metodo di imputazione multipla

L'idea di base del metodo di imputazione multipla è quella di generare più di un valore ($m > 2$) da imputare per ogni dato mancante campionando da un'opportuna distribuzione, in modo che i data set completati siano m. Su ciascuno di essi sono quindi effettuate le analisi statistiche pianificate utilizzando software standard. I risultati delle m analisi vengono poi combinati con regole tali che il risultato inferenziale finale tenga conto dell'incertezza causata dalla presenza di dati mancanti, stimata dalla variabilità tra le m uscite indipendenti.

Il processo di combinazione dei risultati è sempre lo stesso,

a prescindere dall'analisi effettuata sugli m data set completati. Supponiamo che Q sia il parametro incognito di interesse; al termine della procedura di inferenza sugli m data set, sono disponibili m coppie di valori composte dalla stima puntuale del parametro di interesse $\hat{Q}_i$ e dalla stima della varianza dello stimatore, $\hat{U}_i$ (i=1, 2, ..., m). In accordo alla procedura d'imputazione multipla, la stima puntuale di Q è data dalla media delle singole stime calcolate sulle m matrici completate:

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

La varianza dello stimatore $\bar{Q}$ sarà la somma di una componente di variabilità entro imputazione (l'unica di cui terremmo conto in una procedura d'imputazione singola) e da una componente di variabilità tra imputazioni. La varianza entro imputazione, $\bar{U}$, può essere stimata come media delle varianze $\hat{U}_i$:

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

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

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

La varianza totale associata a $\bar{Q}$ è quindi ottenuta combinando le componenti e $\bar{U}$ e B (Rubin 1987):

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

Sulla base della stima puntuale $\bar{Q}$ e della varianza stimata T, è possibile ottenere stime intervallari e fare test statistici sul parametro d'interesse. Per sottoporre a verifica l'ipotesi nulla H_0: $Q = Q_0$, si utilizza la statistica test $(\bar{Q} - Q_0) / \sqrt{T}$, approssimativamente distribuita come una t di Student con v gradi di libertà, dove $v = (m\text{-}1)(1 + \bar{U} / [(1 + m^{-1}) B])^2$.
L'intervallo di confidenza, per esempio al 90%, per Q ha la seguente espressione: $\bar{Q} \pm t_{v,\,0,95} \sqrt{T}$.
L'efficienza della procedura d'imputazione cresce al crescere di m e dipende ovviamente dalla frazione di informazione mancante. Tuttavia la teoria dell'imputazione multipla suggerisce che usualmente 3-5 imputazioni garantiscono ottimi risultati in termini di efficienza.

Generazione delle *m* imputazione

L'aspetto più complicato della procedura di imputazione multipla riguarda la generazione degli m valori con cui imputare ciascun dato mancante.[2]
L'imputazione multipla ha una naturale interpretazione secondo il paradigma bayesiano, in base al quale i valori da imputare sono estratti dalla distribuzione predittiva a posteriori dei dati mancanti, dati i dati osservati, $P(y_{mis} \mid y_{obs})$.
In generale, un campione da questa distribuzione può essere ottenuto per via numerica utilizzando algoritmi MCMC, una volta specificato un modello parametrico per i dati completi (osservati e mancanti) e una distribuzione a priori sui parametri (θ) di questo modello. E' possibile definire un modello parametrico multivariato per i dati completi, $P(Y_1, Y_2, ..., Y_K \mid \theta)$, oppure specificare separatamente le distribuzioni condizionate di ciascuna variabile date tutte le altre, $P(Y_K \mid Y_1, ... Y_{k-1}, Y_{k+1}, ... Y_K, \theta_K)$, per esempio attraverso modelli di regressione. Questo secondo approccio ha il grosso vantaggio di scomporre il problema della specificazione di un'unica distribuzione multivariata in K problemi univariati più flessibili e facili da trattare. Tuttavia è affetto dal problema noto come incompatibilità: potrebbe non esistere una distribuzione congiunta cui corrispondano le distribuzioni condizionate specificate.
Nella situazione speciale in cui il pattern di dati mancanti è monotono, il problema dell'incompatibilità non sussiste e l'algoritmo di imputazione è semplificato. Il pattern di dati mancanti si dice monotono se le variabili possono essere ordinate in modo tale che, quando la variabile Y_j è mancante per una certa unità, tutte le variabili che seguono (Y_k, $k > j$), sono mancanti per tutte le unità. In questo caso, per ciascuna variabile, seguendo l'ordine del pattern, si possono imputare i dati mancanti campionando dalla distribuzione predittiva condizionata alle sole variabili precedenti (quelle con un numero minore di dati mancanti). Un solo ciclo su tutte le variabili è sufficiente per ottenere un data set completato.
Nell'ambito di applicazioni complesse, sono oggetto di studio metodi che traggono vantaggio dalla scomposizione di un pattern di dati mancanti qualsiasi in una parte monotona e una parte che «rompe» la monotonicità, per ridurre il problema dell'incompatibilità.
Un ultimo aspetto rilevante relativo alla specificazione del modello sui dati completi, riguarda il numero di variabili da coinvolgere nella procedura di imputazione. Per esempio, quando si specificano le distribuzioni condizionate attraverso modelli di regressione, l'insieme dei predittori dovrebbe essere il più ricco possibile. Infatti, l'utilizzo di molti predittori rende l'assunzione MAR più plausibile. Inoltre, un modello di imputazione ricco produce data set completati nei quali è preservato un maggior numero di associazioni; questo fa sì che il loro utilizzo sia appropriato per analisi statistiche aventi obiettivi inferenziali differenti (in questo senso, il metodo di imputazione non condiziona l'analisi successiva e viceversa).

Conflitti di interesse: nessuno

Bibliografia

1. Rubin, DB, *Multiple Imputation for Nonresponse in Surveys*, 1987, New York, Wiley.
2. Schafer, JL, *Analysis of Incomplete Multivariate Data*, 1997, New York, Chapman and Hall.