Università degli Studi di Padova

Facoltà di Scienze Statistiche

Corso di Laurea
in Statistica Popolazione e Società

Applicazione di un algoritmo EM: trattamento dei dati mancanti in una regressione logistica

Relatore:
Dott.
Stefano Mazzuco

Laureanda:
Daniela Marchetti

Anno Accademico 2009/2010

Si ritiene la cosa non spiegata e oscura
più importante di quella spiegata e chiara
(Friedrich Nietzsche)

Indice

Introduzione

'Bere per ubriacarsi: è la moda shock dei giovanissimi': questo è il titolo del comunicato stampa dell'Istituo Superiore di Sanità del 12/04/07.
Un fenomeno di tale portata e importanza sociale merita di essere indagato con attenzione, utilizzando tutti i metodi statistici necessari, prendendo in considerazione anche il fatto che *'i dati statistici sono esposti al rischio di* errori di rilevazione*, nel senso che i dati rilevati possono differire dalla realtà rappresentata a causa di imperfezioni nel processo di rilevazione'* (Fabbris (1998) [12]).
Nello specifico sono due le tipologie principali di *errori di rilevazione*: la mancata risposta e l'errore di rilevazione. In questa tesi si pone l'attenzione sul fenomeno delle non risposte, che sono effetto di distorsione in quanto è possibile che la tendenza a non fornire la risposta sia maggiore in certe unità, tendenzialmente diverse da quelle che collaborano all'indagine.

L'obiettivo della tesi è quello di studiare il fenomeno del *binge drinking*, esplorando la possibilità di correggere l'eventuale distorsione causata dalla presenza di dati mancanti, introducendo modelli ner il caso in cui la non risposta non sia ignorabile.

La tesi viene così articolata: nel primo capitolo abbiamo introdotto il problema dei dati mancanti, elencando una classificazione formale dei principali metodi proposti in letteratura per superare l'incompletezza delle rilevazioni. La presentazione dei vari casi e metodi non vuole essere esaustiva, ma vuole sottolineare l'importanza di tenere presente il problema dei dati mancanti, riflettendo prima di decidere come trattarli, sia che la nostra scelta sia di ignorarli, sia che venga presa la decisione di trattarli con qualche metodo proposto dalla letteratura.

Nel secondo capitolo abbiamo descritto l'algortimo EM, uno strumento per il calcolo

di stime di massima verosimiglianza nel caso di dati incompleti. Dapprima si cerca di dare un'idea intuitiva di questo algoritmo, poi si cerca di dare una formalizzazione adeguata. Saranno proposte varie tipologie dell'algoritmo e delle sue applicazioni; questo perchè si vuole mostrare la flessibilità di questo strumento e la possibilità di adattarlo a diverse situazioni.

Introduciamo nel terzo capitolo il fenomeno del *binge drinking* (l'abitudine di consumare quantità eccessive, convenzionalmente 6 o più bicchieri, di bevande alcoliche), presentando inizialmente parte della letteraruta esistente e in seguito riportando uno studio sulle possibili cause di tale fenomeno, utilizzando i dati dell'Indagine Multiscopo sulle Famiglie -Aspetti della vita Quotidiana- dell'anno 2005 (ISTAT), ignorando però la presenza dei dati mancanti.

Nel quarto capitolo ritorniamo allo studio delle cause del *binge drinking*, prendendo in considerazione anche la presenza dei dati mancanti. Presentiamo un'applicazione dell'algoritmo EM sullo studio in questione. Verranno confrontati i risultati avuti prima e dopo aver preso in considerazione i dati mancanti, mostrando così il miglioramento ottenuto avendo inserito l'applicazione dell'algoritmo.

Capitolo 1

I dati mancanti

Il problema dei dati mancanti è abbastanza comune nella ricerca empirica, specialmente nelle scienze sociali, nell'ambito delle quali il tentativo di misurazione di quantità non direttamente osservabili avviene attraverso la somministrazione di test o questionari costituiti da più item. Infatti, se parliamo di dati di tipo socio-demografico, l'assenza di dati mancanti diventa un evento quasi impossibile. Quando in un insieme di dati vi sono dei valori mancanti possono sorgere numerosi problemi, ciò è dovuto al fatto che i metodi classici di analisi statistica sono stati sviluppati per analizzare matrici rettangolari (con tutti i valori presenti).
Ulteriori problemi sono legati alla perdita di correttezza ed efficienza delle stime. Infatti, diminuendo la numerosità campionaria si ha un aumento della varianza e, dunque, una perdita di efficienza; mentre la perdita di correttezza (quindi la distorsione della stima) può verificarsi se i dati mancanti sono tendenzialmente diversi da quelli osservati.
Diventa quindi necessario prendere in considerazione la distribuzione di tali dati mancanti e il possibile legame con i valori osservati: in poche parole bisogna capire se i dati sono mancanti casualmente oppure no. Si inizia così a parlare di *meccanismo generatore dei dati mancanti* per indicare la modalità con cui i dati non sono stati rilevati. Come vedremo nel seguito, prima di attuare procedure inferenziali classiche, sarà necessario tener conto del meccanismo generatore di dati mancanti per poter fare un'inferenza che dia stime corrette ed efficienti.

1.1 Origini dei dati mancanti

Le cause che conducono all'incompletezza dell'informazione sono numerose e diverse, ma si possono riassumere in tre grandi categorie:

1. Mancata copertura
2. Mancate risposte totali (unit nonresponse)
3. Mancate risposte parziali (item nonresponse)

La prima causa, la mancata copertura, è la tipologia più difficile da individuare. Infatti, questo tipo di dato 'mancante' non compare nel dataset e, quindi, per scoprirlo è necessario uno studio approfondito delle modalità di rilevazione dei dati. Siamo nel caso in cui alcuni individui, appartenenti alla popolazione obiettivo, vengono completamente o parzialmente esclusi dalla lista di campionamento. Queste unità hanno quindi probabilità nulla o inferiore al dovuto di essere selezionate. Le cause possono riguardare omissioni nel preparare le liste della popolazione, come nel caso di campionamento per area in cui vengono escluse alcune zone, oppure, la cattiva qualità delle liste di campionamento, per esempio l'elenco telefonico nelle indagini CATI.
Oltre ad essere un tipo di incompletezza abbastanza difficile da individuare, è altrettanto complicato trattare la mancanta copertura: l'unico modo è cercare di compensare le informazioni non raccolte con alcune simili provenienti da fonti esterne.

La seconda causa, la mancata risposta totale (o meglio, unit nonresponse), è data dalla situazione in cui l'unità campionata non fornisce alcuna risposta. Questo può essere determinato da varie situazioni, come l'impossibilità dell'intervistato a cooperare, il contatto non riuscito oppure dallo smarrimento del questionario.
La possibilità di individuare la presenza di questo tipo di missing, è legata alla modalità di costruzione del dataset: può essere che l'individuo venga eliminato dal dataset (e ci riportiamo al caso precedente) oppure che sia riportato con tutti i campi vuoti.
L'effetto della mancata risposta totale da parte di singole unità può essere grave, soprattutto quando le persone non intervistate sono in qualche modo diverse da quelle intervistate: questo può causare forti distorsioni nelle stime dei parametri delle quantità di interesse.

La terza, ed ultima, causa è la mancata risposta parziale. Con questa espressione si intende la mancata risposta ad uno o più quesiti di un questionario. L'unità campionata collabora con l'intervista ma non fornisce risposta ad alcune delle domande che gli vengono proposte. Questo può essere determinato, oltre che dal rifiuto e dall'incapacità dell'intervistato al rispondere, anche dalla mancata registrazione della risposta o dall'eliminazione del dato perché incoerente rispetto ad altre risposte.
Tale tipo di incompletezza risulta la più semplice da gestire, in quanto si dispone di una serie di informazioni sull'individuo in questione. Questo individuo è sicuramente presente nel dataset ma alcuni campi saranno vuoti: è il caso in cui si dispone di un dataset non rettangolare (in quanto contiene delle celle vuote) e, dunque, le analisi statistiche tradizionali non sono più direttamente applicabili.

Ma per quali motivi un individuo dovrebbe rifiutarsi di collaborare con un indagine? Sono diversi gli aspetti che influenzano questo evento, riportando Bosio [2] alcuni sono: il contesto sociale (livello di urbanizzazione, l'adesione al valore della privacy), l'oggetto e gli scopi della ricerca, il proponente (un committente più autorevole comporta maggior collaborazione), caratteristiche del disegno di ricerca (CATI, CAPI, lunghezza del questionario). Ma il fenomeno dell'item nonresponse può essere strettamente legato anche alla tipologia di domanda in questione: quesiti 'sensibili' sono maggiormente legati alla presenza di dati mancanti: un individuo può sentirsi violato nella sua privacy, oppure, più semplicemente, non vuole rischiare di dare una cattiva immagine di sé. Bosco [1] riporta: *'gli argomenti più sensibili sono quelli riguardanti i comportamenti sessuali, le preferenze politiche, il reddito, l'uso di alcolici o altre sostanze psicoattive.'* Continua affermando che, ogni volta che percepiamo la domanda come intrusiva, imbarazzante o addirittura 'minacciosa', ci candidiamo implicitamente a camuffare (oppure omettere) la nostra risposta in funzione della desiderabilità sociale.

1.2 Meccanismi generatori di dati mancanti

Precedentemente è stata usata l'espressione *meccanismo generatore di dati mancanti* in riferimento alla modalità con cui i dati sono mancanti.
I meccanismi possibili sono diversi, ognuno è dotato di caratteristiche proprie che deter-

minano il tipo di distorsione ottenibile sulle stime e, quindi, sull'inferenza.
Proprio per questo, i metodi di correzione delle stime sono strettamente legati alla tipologia di distribuzione dei dati mancati e, di conseguenza, diventa essenziale capire qual è il meccanismo più plausibile.
È importante sottolineare che è molto improbabile poter dire con certezza quale sia l'esatto meccanismo che genera i missing: si è obbligati a fare delle assunzioni a riguardo, che dovranno risultare plausibili ma che difficilmente saranno verificabili.

Usando la stessa notazione riportata da Little e Rubin [13] definiamo con

- Y=$[y_{ij}]$ il dataset completo (i si riferisce all'osservazione e j alla variabile)
- Y=(Y_{obs},Y_{mis}) dove Y_{obs} sono le osservazioni di Y osservate e Y_{mis} quelle mancanti
- M=$[m_{ij}]$ la matrice che indica se y_{ij} è mancate (m_{ij}=1) o è osservato (m_{ij}=0)
- f ($M|Y$,ϕ) la funzione di densità che caratterizza la distribuzione di dati mancanti, dove ϕ è un insieme di parametri ignoti
- f ($Y|\theta$) la funzione di densità di Y, dove θ è un insieme dei parametri ignoti (su cui si vuole fare inferenza)

Usando queste notazioni si ha che

$$f(Y, M|\theta, \phi) = f(Y|\theta)f(M|Y, \phi),$$

quindi, marginalizzando rispetto a Y_{mis}:

$$f(Y_{obs}, M|\theta, \phi) = \int f(Y|\theta)f(M|Y, \phi)dY_{mis}, \tag{1.1}$$

Se i dati mancanti non dipendono dai valori del dataset Y, comprensivo di valori osservati e valori mancanti, e quindi, se

$$f(M|Y, \phi) = f(M|\phi) \text{ per qualsiasi } y_{ij} \in Y, \phi \tag{1.2}$$

i dati si definiscono mancanti in modo del tutto casuale (*missing completely at random* MCAR).

Applicando una condizione più restrittiva rispetto al caso precedente: se i valori mancanti dipendono dai soli valori osservati e, quindi

$$f(M|Y,\phi) = f(M|Y_{obs},\phi) \text{ per qualsiasi } y_{ij} \in Y_{mis},\phi \tag{1.3}$$

siamo nella situazione in cui i dati sono *missing at random* (MAR).

Infine, si definiscono *not missing at random* (NMAR) se la distribuzione di M dipende sia da Y_{mis} che da Y_{obs}:

$$f(M|Y,\phi) = f(M|Y_{obs},Y_{mis},\phi) \text{ per qualsiasi } \phi \tag{1.4}$$

1.2.1 Missing completely at random

È l'ipotesi più facile da trattare, ma è anche quella più difficilmente riscontrabile in situazioni concrete: la probabilità di osservare una risposta mancante è indipendente sia dalla parte osservata che da quella non osservata dell'insieme di dati completo.
Il termine ha un preciso significato: pensando al dataset come ad un'ampia matrice di dati, i valori mancanti sono casualmente distribuiti attraverso la matrice.
Negli studi sulle famiglie questo accade raramente dato che, per esempio, individui appartenenti a minoranze sociali, persone con alti redditi, soggetti con un basso livello d'istruzione e persone soggette a sindromi di depressione o di ansia, sono meno propense a rispondere a tutti gli item di un questionario.
Esemplificando nel caso di due variabili causali: si consideri Y_1 completamente osservata e Y_2 con alcuni valori mancanti. Se la probabilità che un dato sia mancante non dipende né da Y_1 né da Y_2 allora i dati sono mancanti in modo completamente casuale (MCAR). Rubin dimostra che, se si effettuano inferenze basate sulla distribuzione campionaria di Y, si può ignorare il meccanismo che causa i dati mancanti solo assumendo che i dati siano MCAR. Questo si ha perchè, utilizzando la (1.1) e la (1.2) si ottiene:

$$f(Y_{obs},M|\theta,\phi) = f(M|\phi)\int f(Y|\theta)dY_{mis},$$

con il risultato che

$$f(Y_{obs}|\theta) = \int f(Y|\theta)dY_{mis}.$$

In conlcusione, sotto tale assunzione, l'unica conseguenza sugli stimatori è la minor efficienza in quanto si è costretti ad utilizzare solo n_{obs} con $n_{obs} < n$, e di conseguenza aumenta la varianza dello stimatore.

1.2.2 Missing at random

Si parla di dati MAR, o non risposta ignorabile, quando la probabilità di osservare una risposta mancante dipende soltanto dalla parte osservata dell'insieme di dati.
Utilizzando le due variabili Y_1, completamente osservata, e Y_2, con alcuni valori mancanti, siamo nella situazione *missing at random* se la variabile Y_1 rappresenta un 'elemento esplicativo' della presenza o meno dell'informazione Y_2. Una variabile è considerata 'elemento esplicativo' quando aiuta a spiegare se un soggetto risponderà o meno ad un quesito.
Sottolineiamo che la probabilità di non risposta è legata solamente alla variabile che fa da elemento esplicativo, e non dal valore stesso del dato mancante. Quindi, la probabilità che Y_2 sia mancante dipende solamente da Y_1, e non dal valore stesso di Y_2. Molti 'elementi esplicativi' vengono inclusi negli studi sulle famiglie di grande scala, tra i quali i più comuni sono: il livello d'istruzione, la razza, l'età, il sesso ed indicatori di benessere psico-sociale. L'assunto per i valori MAR è valido solo se il modello dei dati mancanti è condizionatamente casuale, dati i valori osservati nelle variabili considerate 'elementi esplicativi'.
Per definire questa situazione si può parlare di meccanismo di non risposta *ignorabile* in quanto non serve specificare un modello di non risposta $f\ (M|Y,\phi)$ per ottenere valide inferenze (basate sulla verosimiglianza) riguardo θ[1].
Per definizione

$$f(Y|\theta) = f(Y_{mis}|Y_{obs},\theta)f(Y_{obs}|\theta) \tag{1.5}$$

Utilizzando la (1.3) e la (1.5) si ottiene che

$$f(Y_{obs},M|\theta,\phi) = f(Y_{obs}|\theta)f(M|Y_{obs}\phi) \tag{1.6}$$

Se si è interessati a fare inferenza su θ, si considera solo il termine $f(Y_{obs}|\theta)$ e si può tralasciare $f(M|Y_{obs}\phi)$, poichè non porta informazioni su θ. Questo vale se θ e ϕ sono

[1]Da notare che è il meccanismo di non risposta, non il dato mancante, che può essere ignorato.

distinti. Dalla (1.6) si vede anche che Y_{obs} e M non sono indipendenti e quindi che la densità di Y_{obs} dato M dipende dal modello di non risposta.
Oltre a tutto questo non bisogna dimenticare che, essendo presenti dei dati mancanti, siamo costretti ad utilizzare n_{obs} con $n_{obs} < n$, e di conseguenza aumenta la varianza dello stimatore.
In conclusione, una situazione di tipo MAR porta a due problemi: una minor efficienza dello stimatore (come nel meccanismo MCAR), ma, problema ben più grave, la distorsione dello stimatore (in quanto i dati mancanti non possono essere ignorati).

1.2.3 Not missing at random

Si parla di dati NMAR, o non risposta non ignorabile, quando la probabilità di risposta dipende sia dai dati osservati che da quelli non osservati. In questo caso i dati mancanti non sono più *at random*: certi valori di y hanno più probabilità di altri di essere osservati.
Usando la variabile Y_1, per esempio, potrebbero essere osservati solo i valori positivi di Y_1: $f(M|Y_1, \phi) = I(y_{i1} > 0)$. Avremo quindi una distorsione dello stimatore in quanto i dati non osservati sono fortemente diversi da quelli osservati (nell'esempio riportato abbiamo solo valori positivi e non quelli negativi). Il grado di distorsione dipenderà dalla quantità di dati mancanti.
Inoltre, analogamente ai casi precedenti, avremo una minor efficienza dello stimatore.
In questo caso il meccanismo di non risposta dovrà essere tenuto esplicitamente in considerazione se si vogliono fare valide inferenze su θ.

1.3 Metodi per trattare i dati mancanti

A questo punto sorge spontanea una domanda: è possibile ottenere stime valide in presenza di dati mancanti? Negli ultimi anni sono stati sviluppati metodi statistici per il trattamento di questa problematica e, in generale, i metodi proposti in letteratura per l'analisi dei dati in presenza di osservazioni mancanti possono essere classificati in quattro gruppi:

- Metodi basati sulle sole unità osservate

- Metodi di imputazione
- Procedure di ponderazione
- Metodi basati sui modelli

Verrà di seguito riportata una breve descrizione di ogni gruppo, per un'esauriente trattazione si veda Little e Rubin [13].

1.3.1 Metodi basati sulle sole unità osservate

Le procedure basate sull'analisi delle sole unità, senza dati mancanti, sono quelle, ovviamente, più semplici. In queste procedure vengono ignorate tutte le osservazioni parziali, e vengono calcolate le stime di interesse sul dataset completo. Conseguenza naturale di questo metodo è la maggior facilità di trattazione, seguita però da una perdita potenziale di informazioni.
Questa probabile perdita ha due dimensioni: la diminuzione di precisione, dovuta alla maggiore varianza derivante da una più bassa numerosità campionaria, e la distorsione, dovuta al fatto che i dati mancanti potrebbero essere sostanzialmente diversi da quelli osservati. Per questo motivo questa procedura si utilizza solo sotto l'assunzione di dati di tipo MCAR, in questo modo gli stimatori si possono considerare non distorti.
Riguardo la perdita di efficienza, è ragionevole pensare che, in presenza di campioni sufficientemente grandi e ridotta presenza di dati mancanti, questa non sia eccessiva. In ogni caso, a volte può essere più conveniente 'accontentarsi' di questo tipo di procedura (in quanto semplice) e delle possibili distorsioni che ne derivano, piuttosto di utilizzare procedure più complesse che comunque non garantiscono un minor grado di distorsione.

1.3.2 Metodi di imputazione

I metodi di imputazione vengono usati prevalentemente nel caso di mancante risposte parziali, e consistono nel sostituire i valori mancanti con valori opportunamente calcolati. Bisogna sottolineare che imputare i dati può essere una tecnica vantaggiosa ma, al tempo stesso, pericolosa. Riportando Dempster & Rubin (1983):

> 'The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases.'

Quindi, il vantaggio deriva dal fatto che si può analizzare il dataset interamente, senza scartare nessuna unità. Proprio per questo, sono metodi frequentemente utilizzati e risultano particolarmente attraenti perché abbastanza semplici e intuitivi. Il pericolo è quello di dimenticarsi che il dataset è stato 'ricostruito' e, quindi, si possono considerare le stime come non distorte solo sotto alcune assunzioni.
L'imputazione viene fatta attraverso vari metodi che possono essere classificati in due gruppi:

Modellazione esplicita: la distribuzione predittiva viene modellata esplicitamente specificando formalmente un modello statistico, le ipotesi sono quindi esplicite. Questo gruppo include i metodi di imputazione tramite media, media condizionata, regressione e estrazione dalla distribuzione predittiva.

Modellazione implicita: si basa su algoritmi che assumono un modello statistico che non viene specificato. Le ipotesi sono quindi implicite, ma necessitano di essere controllate. Alcuni metodi appartenenti a questo gruppo sono: l'imputazione Hot deck, la sostituzione, il metodi Cold deck.

Per una descrizione dei vari metodi consultare Little & Rubin [13].

1.3.3 Procedure di ponderazione

Le procedure basate sull'utilizzo dei pesi sono una variante dell'analisi sui dati completi. Sono generalmente usate per compensare la non risposta totale e consistono nel modificare i pesi assegnati alle unità effettivamente osservate, in modo che risultino rappresentative anche di quelle non osservate. Abbiamo quindi una grande semplicità di applicazione. I pesi che vengono attribuiti sono inversamente proporzionali alla probabilità di osservazione (che deve essere stimata): questo può portare ad un aumento

della varianza. Infatti, i rispondenti con bassa probabilità di risposta avranno un peso particolarmente alto, assumendo una forte influenza sullo stimatore. Altro svantaggio di tale metodo è dato dalla difficoltà nel reperire le informazioni per costruire i pesi. Inoltre, non sempre le quantità pesate sono facilmente interpretabili. Per questi motivi la ponderazione si applica correttamene se le unità mancanti sono mediamente simili a quelle che hanno collaborato: con questo metodo si 'chiede' a coloro che hanno risposto di rappresentare anche il gruppo dei non-rispondenti.

1.3.4 Metodi basati sui modelli

Un modo efficiente per trattare il problema dei dati mancanti è quello di ipotizzare un modello parametrico sottostante i dati e stimare i parametri di tale modello attraverso i metodi di massima verosimiglianza. Si è visto nella sezione precedente che, se si assume che i dati siano di tipo MAR e che i parametri della funzione di densità dei dati (θ) siano distinti dai parametri del meccanismo che genera i dati mancanti (ϕ), si può fare inferenza sui parametri di interesse attraverso la funzione di verosimiglianza dei dati osservati, ignorando il meccanismo che genera i dati mancanti. Le procedure appartenenti a questo gruppo sono piuttosto flessibili e non necessitano di procedure 'ad hoc' per aggiustare le stime, oltre a disporre di stime asintotiche della varianza che tengono conto dell'incompletezza dei dati.
I metodi basati sui modelli non si pongono l'obiettivo di identificare un opportuno valore da assegnare al record con valori mancanti, ma piuttosto cercano di utilizzare tutta l'informazione disponibile per dare stime corrette dei parametri di interesse. Uno degli strumenti più noti a riguardo è l'algoritmo EM (Expectation Maximization) che consente di effettuare stime di massima verosimiglianza dei parametri di interesse su un dataset di dati incompleti, come se fossero completi. È questo strumento il nucleo di interesse di questa tesi e il prossimo capitolo cercherà di fornire un'attenta descrizione di tale algoritmo.

1.4 Conclusioni

La gestione dei dati mancanti è un problema che, sebbene rilevante, non sempre è affrontato adeguatamente. Molti ricercatori svolgono una semplice analisi dei dati completi assumendo una distribuzione MCAR dei dati mancanti, supportati anche dai software statistici che in molti casi implementano tale funzione. Il rischio di distorsione e di perdita di efficienza (oltre la notevole riduzione del set di dati, con la conseguente diminuzione del potere statistico in fase inferenziale) è quindi sottovalutato.
Per questo motivo si ritiene necessario sottolineare che è essenziale una riflessione sui dati mancanti, cercando di capire qual è il meccanismo generatore del caso preso in esame. Bisogna chiarire che non è essenziale procedere con dei metodi specifici per la trattazione dei dati mancanti, infatti, l'utilizzo di una strategia può essere un *'inopportuna complicazione'* come affermato da Fabbris [12] in riferimento all'uso del metodo di imputazione negli studi relazionali in ottica essenzialmente esplorativa. In ogni caso, la scelta di imputare o trattare in modo particolare i dati mancanti dipende molto dal contesto dell'analisi, in particolare, dalle ipotesi di partenza e dagli scopi prefissi.

Capitolo 2

L'algoritmo EM

L'algoritmo *Expetation-Maximization* (EM) è un algoritmo iterativo ampiamente utilizzato per calcolare le stime di massima verosimiglianza nel caso di dati incompleti. Questo strumento viene tendenzialmente usato quando la funzione di verosimiglianza assume forme particolarmente complicate e diventa necessario ricorrere a metodi numerici (ad esempio l'algoritmo di Newton-Raphson) che, però, possono essere molto onerosi a livello computazionale.
Il successo dell'algoritmo è dovuto alla semplicità di programmazione, al pregio di porre il problema di massimizzazione in termini statistici e alla sua generalità: infatti, le situazioni in cui può essere applicato comprendono non solo i casi evidenti di dati-incompleti, ma anche una grande varietà di situazioni in cui l'incompletezza dei dati non è così palese (variabili latenti, modelli log lineari...etc.).
Anche se i primi riferimenti ad un algoritmo simile risalgono al 1926, a presentarlo per la prima volta in modo completo, studiandone il comportamento e fornendo un ampio insieme di esempi, furono stati Dempster, Laird e Rubin nel 1977. Di seguito l'algoritmo EM verrà presentato in modo semplice e intuitivo, cercando comunque di utilizzare un'adeguata formalizzazione.

2.1 La logica dell'algoritmo EM

L'algoritmo EM formalizza un'idea elementare per trattare i dati mancanti che consiste nel:

1. Sostituire i valori mancanti con dei valori stimati;

2. Stimare i parametri;

3. Ri-stimare i dati mancanti, assumendo che le nuove stime dei parametri siano corrette;

4. Ri-stimare i parametri, ripetendo la procedura fino alla convergenza.

Ogni iterazione dell'algoritmo EM consiste in un passo E (Expectation step) ed in un passo M (Maximization step).

Il passo M è particolarmente semplice da descrivere: calcola le stime di massima verosimiglianza (SMV) di θ sui dati 'completati' (come se non fossero presenti dati mancanti). Quindi, il passo M sfrutta gli stessi metodi computazionali utilizzati per dati completi.

Il passo E trova i valori attesi condizionati dei 'dati mancanti', dati i valori osservati e le correnti stime dei parametri, quindi sostituisce i valori mancanti con questi attesi. 'Dati mancanti' è stato scritto tra apici in quanto l'algoritmo EM non sostituisce direttamente i valori mancanti con i valori attesi trovati al passo E, ma le funzioni di Y_{mis} che compaiono nella log-verosimiglianza dei dati completi $\ell(\theta|Y)$. È proprio per questo motivo che si ritiene che l'algortimo EM tratti il problema di dati mancanti a livello statistico e non semplicemente numerico.
Nostante l'algoritmo sia applicabile ad una vasta classe di modelli, è particolarmente utile quando i dati completi provengono da una famiglia esponenziale: in questa situazione il passo E si riduce al calcolo del valore atteso condizionato delle statistiche sufficienti per i dati completi e il passo M è, spesso, molto semplice a livello numerico.

2.2 Un esempio introduttivo

Per esemplificare la logica dell'algoritmo EM può essere proposto il seguente esempio, lo stesso utilizzato da Dempster, Laird e Rubin (1977) per introdurre l'algoritmo EM. Definiamo Y=(y_1, y_2, y_3, y_4) come la determinazione di una v.c. multinomiale con probabilità

$$\pi = (\pi_1, \pi_2, \pi_3, \pi_4) = \left(\frac{1}{2} - \frac{\theta}{2}, \frac{\theta}{4}, \frac{\theta}{4}, \frac{1}{2}\right) \tag{2.1}$$

L'obiettivo è trovare la SMV di θ.
Il vettore di dati osservati $Y_{obs} = (38, 34, 125)$ corrisponde alla osservazione della variabile di interesse Y=(y_1, y_2, y_3, y_4) con:

$$y_1 = 38 \tag{2.2}$$

$$y_2 = 34 \tag{2.3}$$

$$y_3 + y_4 = 125. \tag{2.4}$$

Quindi $Y_{obs} = (y_1, y_2, y_3 + y_4)$.
Si suppone che il vettore di dati osservati $Y_{obs} = (38, 34, 125)$ provenga da una variabile casuale con distribuzione multinomiale con probabilità di celle:

$$\pi = (\pi_1, \pi_2, \pi_3) = \left(\frac{1}{2} - \frac{\theta}{2}, \frac{\theta}{4}, \frac{\theta}{4} + \frac{1}{2}\right) \tag{2.5}$$

Il valore mancante si può quindi identificare come la parte di $y_3 + y_4$ corrispondente a y_3 (o y_4).
Se fosse stato osservato Y, la SMV di θ si sarebbe trovata massimizzando la funzione di verosimiglianza dei dati completi:

$$L(\theta|Y) = \frac{(y_1 + y_2 + y_3 + y_4)!}{y_1!y_2!y_3!y_4!} \cdot \pi_1^{y_1} \cdot \pi_2^{y_2} \cdot \pi_3^{y_3} \cdot \pi_4^{y_4}$$

Quindi la log-verosimiglianza per i dati completi è:

$$\ell(\theta|Y) \propto y_1 ln(1-\theta) + y_2 ln(\theta) + y_3 ln(\theta)$$

Risolvendo rispetto θ l'equazione di verosimiglianza

$$\frac{d}{d\theta} l(\theta|Y) = 0$$

Si ottiene la SMV nel caso dei dati completi:

$$\hat{\theta} = \frac{y_2 + y_3}{y_1 + y_2 + y_3} \tag{2.6}$$

Notando che la log-verosimiglianza è lineare in Y, calcolare il valore atteso di essa rispetto a Y, dato θ e Y_{obs}

$$E[Y_1 ln(1-\theta) + Y_2 ln(\theta) + Y_3 ln(\theta|Y_{obs}, \theta^{(t)})]$$

implica il calcolo del valore atteso di Y, dato θ e Y_{obs}. Nel caso di dati mancanti ciò comporta la sostituzione dei dati mancanti stessi con delle stime:

$$\begin{aligned} E[Y_1|\theta, Y_{obs}] &= 38 \\ E[Y_2|\theta, Y_{obs}] &= 34 \\ E[Y_3|\theta, Y_{obs}] &= 125(\theta/4)/(1/2 + \theta/4) \\ E[Y_4|\theta, Y_{obs}] &= 125(1/2)/(1/2 + \theta/4) \end{aligned}$$

Quindi alla t -esima iterazione, con $\theta^{(t)}$ stima corrente di θ, il passo E consiste nel calcolare

$$y_3^{(t)} = 125(\theta^{(t)}/4)/(1/2 + \theta^{(t)}/4) \tag{2.7}$$

Il passo M consiste nel trovare il massimo della funzione di log-verosimiglianza per dati completi, dato dalla (2.6), con $y_3 = y_3^{(t)}$

$$\theta^{(t+1)} = \frac{34 + y_3^{(t)}}{72 + y_3^{(t)}} \tag{2.8}$$

Iterando la (2.7) e la (2.8) si definisce l'algoritmo EM per questo problema. Nella tabella (2.1) vi sono le iterazioni dell'algoritmo, e si mostra la convergenza partendo da $\theta^{(0)} = 0.5$.

Questo problema di stima si può risolvere anche utilizzando l'algoritmo di Newton-Raphson: partendo da un valore $\theta^{(0)} = 0.5$ i valori di θ per le prime due iterazioni sono $\theta^{(1)} = 0.63636363$ e $\theta^{(2)} = 0.62696867$. Confrontando questi valori con le iterazioni dell'algoritmo EM della tabella (2.1) si nota che, partendo dallo stesso valore iniziale, dopo solo due iterazioni l'algoritmo di Newton-Raphson è già abbastanza vicino al valore della SMV mentre con l'algoritmo EM dobbiamo aspettare la quinta iterazione. Questo esempio mette in luce uno dei problemi dell'algoritmo EM, vale a dire la sua lentezza di convergenza.

Tabella 2.1: Stima del parametro θ, iterazioni dell'algoritmo EM

t	$\theta^{(t)}$	$\theta^{(t)}$- $\hat{\theta}$	$(\theta^{(t+1)}$- $\hat{\theta})/(\theta^{(t)} - \hat{\theta})$
0	0.500000000	0.126821498	0.1465
1	0.608247423	0.018574075	0.1346
2	0.624321051	0.002500447	0.1330
3	0.626488879	0.000332619	0.1328
4	0.626777323	0.000044176	0.1328
5	0.626815632	0.000005866	0.1328
6	0.626820719	0.000000779	-
7	0.626821395	0.000000104	-
8	0.626821484	0.000000014	-

2.3 Formalizzazione dell'algoritmo EM

Nel paragrafo 1.2.2 abbiamo visto che la distribuzione dei dati completi può essere fattorizzata in

$$f(Y_{mis}|Y_{obs}, \theta)f(Y_{obs}|\theta).$$

La log-verosimiglianza diviene quindi

$$\ell(\theta|Y) = \ell(\theta|Y_{obs}, Y_{mis}) = \ell(\theta|Y_{obs}) + \ln[f(Y_{mis}|Y_{obs}, \theta)] \tag{2.9}$$

dove $\ell(\theta|Y)$ è la verosimiglianza dei dati completi, $\ell(\theta|Y_{obs})$ è la verosimiglianza dei dati osservati, mentre l'ultimo termine $ln[f(Y_{mis}|Y_{obs}, \theta)]$ è il logaritmo della funzione di densità dei missing, dati i valori osservati e θ.
L'obiettivo è stimare θ massimizzando la verosimiglianza dei dati osservati rispetto a θ, per Y_{obs} fissato. Infatti, quando vi sono dati mancanti la funzione di verosimiglianza che viene massimizzata è quella dei dati osservati. Però, come è già stato detto, questa procedura può essere molto laboriosa. Si cerca così di semplificare il problema, 'accontentandosi' di calcolare il valore atteso della log-verosimiglianza rispetto alla distribuzione condizionata di Y_{mis} dato Y_{obs} e θ. Scriviamo la (2.9) nel seguente modo:

$$\ell(\theta|Y_{obs}) = \ell(\theta|Y_{obs}, Y_{mis}) - ln[f(Y_{mis}|Y_{obs}, \theta)] \tag{2.10}$$

Il valore atteso della (2.10) rispetto alla distribuzione di Y_{mis} dato Y_{obs} e θ è:

$$E[\ell(\theta|Y_{obs})] = Q(\theta|\theta^{(t)}) - H(\theta|\theta^{(t)})$$

dove

$$Q(\theta|\theta^{(t)}) = \int [\ell(\theta|Y_{obs}, Y_{mis})] f(Y_{mis}|Y_{obs}, \theta^{(t)}) dY_{mis} \tag{2.11}$$

e

$$H(\theta|\theta^{(t)}) = \int ln[f(Y_{mis}|Y_{obs}, \theta)] f(Y_{mis}|Y_{obs}, \theta^{(t)}) dY_{mis}. \tag{2.12}$$

considerando che $E[\ell(\theta|Y_{obs})]$ rispetto a Y_{mis} è $\ell(\theta|Y_{obs})$ si ha:

$$\ell(\theta|Y_{obs}) = Q(\theta|\theta^{(t)}) - H(\theta|\theta^{(t)}) \tag{2.13}$$

Se $\theta^{(t)}$ è SMV per $\ell(\theta|Y_{obs})$, anche $H(\theta|\theta^{(t)})$ è massimizzato quando $\theta = \theta^{(t)}$ in quanto, per la disuguaglianza di Jensen, si ha che $H(\theta|\theta^{(t)}) \leq H(\theta^{(t)}|\theta^{(t)})$. Di conseguenza anche $Q(\theta|\theta^{(t)})$ è massimizzato quando $\theta = \theta^{(t)}$.
Da qui abbiamo l'algoritmo EM: poichè $\ell(\theta|Y_{obs})$ è difficile da massimizzare si preferisce massimizzare il valore atteso condizionato della log-verosimiglianza per dati completi $Q(\theta|\theta^{(t)})$. La massimizzazione di $Q(\theta|\theta^{(t)})$ assicura, per quanto detto sopra, la massimizzazione anche di $\ell(\theta|Y_{obs})$. Partendo da una stima iniziale $\theta^{(0)}$ e detta $\theta^{(t)}$ la stima corrente di θ, i due passi dell'algoritmo EM sono:

Passo E : calcola il valore atteso rispetto alla distribuzione di Y_{mis} della log-verosimiglianza per dati completi, dato Y_{obs} e θ:

$$Q(\theta|\theta^{(t)}) = \int [\ell(\theta|Y_{obs}, Y_{mis})] f(Y_{mis}|Y_{obs}, \theta^{(t)}) dY_{mis} \tag{2.14}$$

Passo M : calcola $\theta^{(t+1)}$ massimizzando $Q(\theta|\theta^{(t)})$:

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)}) \forall \theta \tag{2.15}$$

Grazie al lavoro di Dempster, Laird e Rubin (1977) si ha la sicurezza che ad ogni iterazione di un algoritmo EM la log-verosimiglianza è non decrescente. Infatti, si consideri una sequenza di iterazioni $\theta^{(0)}, \theta^{(1)} ... \theta^{(t)} ...$ dove $\theta^{(t+1)} = M(\theta^{(t)})$ per qualche funzione $M(\cdot)$. Si riporta di seguito l'enunciato del teorema che afferma che ogni iterazione aumenta o lascia invariata la verosimiglianza (Little & Rubin [13]).

Teorema 1 *Ad ogni iterazione un algoritmo EM aumenta la log-verosimiglianza* $\ell(\theta|Y_{obs})$ *cioè* $\ell(\theta^{(t+1)}|Y_{obs}) \geq \ell(\theta^{(t)}|Y_{obs})$. *Vale l'uguaglianza se e solo se* $Q(\theta^{(t+1)}|\theta^{(t)}) = Q(\theta^{(t)}|\theta^{(t)})$.

Inoltre, Dempster, Laird e Rubin (1977) dimostrano che, se $\theta^{(t)}$ converge, allora converge ad un punto stazionario. Questo avviene solo sotto alcune restrizioni applicate alla funzione di densità (quali l'appartenenza ad una famiglia esponenziale regolare e $\ell(\theta|Y_{obs})$ limitata), ma non dovrebbe stupire in quanto nessun algoritmo iterativo assicura la convergenza ad un punto stazionario. È necessario specificare tuttavia che, quando la log-verosimiglianza ha diversi punti stazionari, la convergenza dell'algoritmo EM dipende dalla scelta del valore iniziale. Per questo motivo si raccomanda di prevedere diverse iterazioni dell'algoritmo da più punti iniziali. In ogni caso, nella maggioranza dei problemi pratici rilevanti, si è visto che l'algoritmo EM converge quasi sempre ad un massimo locale.

2.4 L'algoritmo EM per famiglie esponenziali

L'algortimo EM assume una forma particolarmente semplice quando i dati completi Y hanno una distribuzione appartenente alla famiglia esponenziale regolare, vale a dire se è esprimibile in

$$f(Y, \theta) = \exp[s(Y)d(\theta) + c(Y) - b(\theta)]$$

dove θ è parametro ignoto, $c(\cdot)$ e $d(\cdot)$ sono funzioni note la cui scelta individua una particolare distribuzione.
La forma appena vista si può scrivere alternativamente come

$$f(Y, \theta) = t(Y) \exp[s(Y)d(\theta)] \frac{1}{a(Y)} \tag{2.16}$$

con $t(Y) = \exp[c(Y)]$ e $a(\theta) = \exp[b(\theta)]$. $t(\cdot)$, $b(\cdot)$ e $a(\cdot)$ funzioni note.
$s(Y)$ è un vettore di statistiche sufficienti per dati completi (1 x d),
θ è un vettore di parametri (1 x d).

Per definizione, le statistiche sufficienti portano tutte le informazioni riguardanti θ. Riportando l'espressione usata da Piccolo [4] si ha che, *tutte le informazioni riguardanti* θ *- che pure esistevano nel campione casuale - vengono integralmente* trasferite *nello*

stimatore sufficiente.
Per questo motivo il passo E per questa distribuzione si può ridurre alla stima della statistica sufficiente per dati completi $s(y)$ con:

$$s^{(t+1)} = E[s(Y)|Y_{obs}, \theta^{(t)}]$$

Il passo M determina la nuova stima di $\theta^{(t+1)}$ di θ risolvendo le equazioni di verosimiglianza

$$E[s(Y)|\theta] = s^{(t+1)}$$

che sono semplicemente le equazioni di verosimiglianza per dati completi con $s(y)$ sostituito da $s^{(t+1)}$.

2.5 Pregi e difetti dell'algoritmo EM

Si è visto che l'algoritmo EM ha diversi pregi, ma anche alcuni diffetti. Riassumendo si possono elencare tra i pregi:

- Non prevede il calcolo e l'inversione di matrici di informazione (usate nell'algoritmo Newton-Raphson);
- È facile da costruire poichè il passo E e il passo M sono basati su calcoli compiuti sui dati completi;
- È di facile implementazione;
- Pone il problema di massimizzazione della funzione di verosimiglianza in presenza di dati mancanti in termini statistici: il passo E completa i dati mentre il passo M calcola di stima di massima verosimiglianza sui dati completi;
- Ad ogni iterazione aumenta la log-verosimiglianza. Inoltre, nella maggioranza dei problemi pratici converge ad un massimo locale.

Tra i diffetti si possono invece riportare:

- È conveniente solo quando il passo E può essere calcolato direttamente, per questo viene usato frequentemente con variabili appartenti alla famiglia esponenziale;

- Il tasso di convergenza può essere molto lento, soprattutto se vi sono molti dati mancanti;
- La convergenza ad una SMV (cioè ad un massimo globale) non è sempre garantita;
- Argomento non trattato in questa tesi, ma che si ritiene giusto riportare: l'algoritmo EM non fornisce automaticamente gli errori standard delle stime. Bisogna infatti ricorrere o all'algoritmo SEM o calcolarli con il metodo di Louis.

2.6 L'algoritmo EM nei modelli lineari

2.6.1 Variabile risposta con valori missing

Un ulteriore utilizzo dell'algoritmo EM riguarda la sua applicazione ai modelli lineari generalizzati quando vi sono dei dati mancanti nella variabile risposta. Si assume che i dati siano mancanti *at random* (MAR) e le covariate completamente osservate. Se i dati sono MAR significa che la non risposta a Y (variabile risposta) dipende completamente dai valori osservati delle altre variabili, le covariate $X_1, ..., X_p$. Per spiegare in dettaglio questa modalità, aggiungiamo le seguenti notazioni:

- $Y_i \sim EF(b(\theta_i))$ con $b'(\theta_i) = \mu_i$;
- Y_i è legata alle covariate attraverso il predittore lineare η_i dove $\eta_i = x_i^T \beta$;
- la funzione legame $g(\mu_i) = \eta_i$ è determinata dal tipo di distribuzione.

L'obiettivo è stimare β.
Con l'algoritmo EM si può stimare il valore atteso della log-verosimiglianza per θ con $\theta = \theta^{(t)}$ (passo E) e utilizzare questo valore atteso per calcolare la regressione di Y con $X_1, ..., X_p$ covariate (passo M), si ottengono così dei nuovi valori per le y_i mancanti da cui si può ricavare la succesiva stima di θ, $\theta^{(t+1)}$, e quindi ritonare al passo M. Naturalmente, per quanto detto sopra, si possono usare le statistiche sufficienti al posto della log-verosimiglianza.

2.6.2 Covariate con valori missing

In questa sezione, l'algoritmo EM viene utilizzato per stimare i parametri in un GLM (modello lineare generalizzato) che presenta dei missing tra la covariate. Il metodo che verrà riportato si chiama 'metodo dei pesi' ed è stato ideato da Ibrahim [11].
Si assume che i dati siano mancanti *at random* (MAR) e che la variabile risposta (Y) sia completamente osservata. Se le covariate, $X_1, ..., X_p$ sono MAR, significa che la non risposta dipende completamente dai valori osservati della variabile risposta e delle altre covariate.
Poniamo, allora, $X = (X_1, ..., X_p)$ la matrice di covariate proveniente da una distribuzione discreta con parametri $\Gamma = (\Gamma 1, ..., \Gamma r)$. Assumiamo inoltre che $Y|X$ provenga da una famiglia esponenziale con parametri (α, ϕ), dove α e ϕ sono distinti da Γ.
Il parametro di interesse è α (cioè i coefficienti di regressione). Sottolineiamo che la distribuzione congiunta (Y, X) è data dalla distribuzione condizionata $Y|X$ e dalla distribuzione marginale di X. Prendendo come y_i e $x_i = (x_1, ..., x_p)$ la i -esima riga corrispondente alla i -esima osservazione, assumiamo che per le n osservazioni le $y_i|x_i$ sono indipendenti e le x_i siano *iid*. Prendiamo ora $\theta = (\alpha, \phi, \gamma)$ e scriviamo $x_i = (x_{obs,i}, x_{mis,i})$ dove $x_{obs,i}$ e $x_{mis,i}$ indicano rispettivamente i valori osservati e quelli mancanti di x_i.

Il passo E dell'algoritmo EM è dato da

$$\begin{aligned} Q_i(\theta|\theta^{(t)}) &= E(\ell(\theta; x_i, y_i)|x_{obs,i}, y_i, \theta^{(t)}) = \\ &= \sum_{x_{mis,i}} p(x_{mis,i}|x_{obs,i}, y_i, \theta^{(t)})\ell(\theta; x_i, y_i) \end{aligned} \tag{2.17}$$

Dove $\theta^{(t)}$ è la stima corrente di θ, $p(x_{mis,i}|x_{obs,i}, y_i, \theta^{(t)})$ è la distribuzione condizionata dei valori mancanti dati i valori osservati e la stima corrente di θ, e la sommatoria è su tutti i possibili valori di $x_{mis,i}$.
Grazie al teorema di Bayes $p(x_{mis,i}|x_{obs,i}, y_i, \theta^{(t)})$ si può scrivere come:

$$p(x_{mis,i}|y_i, x_{obs,i}, \theta^{(t)}) = \frac{p(y_i|x_{mis,i}, x_{obs,i}, \theta^{(t)})p(x_i|\theta^{(t)})}{\sum_{x_{mis,i}} p(y_i|x_i, \theta^{(t)})p(x_i|\theta^{(t)})} \tag{2.18}$$

Notiamo che la (2.18) è ora 'costruita' da probabilità note in quanto la $Y|X$ viene da una distribuzione esponenziale con parametri (α, ϕ) e la X viene da una distribuzione discreta con parametro γ. Partendo quindi da una stima di $\theta^{(t)} = (\alpha^{(t)}, \phi^{(t)}, \gamma^{(t)})$ abbiamo tutti

i componenti che servono per calcolare la (2.18). La sommatoria è calcolata su $x_{mis,i}$, questo significa che si deve calcolare su tutti i possibili valori che può assumere il dato mancante: nel caso di una binomiale con n=1 estrazioni, i valori possibili per la nostra variabile saranno 0 e 1.
Riassumendo, il passo E dell'algoritmo può essere scritto come:

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= \sum_{i=1}^{n} Q_i(\theta|\theta^{(t)}) = \\ &= \sum_{i=1}^{n} \sum_{x_{mis,i}} w_{i,(t)} \ell(\theta; x_i, y_i) \end{aligned} \tag{2.19}$$

dove i $w_{i,(t)} = p(x_{mis,i}|x_{obs,i}, y_i, \theta^{(t)})$ si possono esprimere come nella (2.18) e sono i pesi corrispettivi alle osservazioni mancanti (quindi per le osservazioni non mancanti sono pari a 1).

Il passo M diventa quindi la massimizzazione della (2.19), che corrisponde al calcolare la SMV per dati completi (dove i valori mancanti sono stati sostituiti dai valori pesati calcolati sulla base dei dati osservarti).

2.7 Conclusioni

La presentazione teorica dell'algoritmo è ora conclusa. Sono moltissimi gli aspetti che non sono stati trattati ed è giusto accennare al fatto che l'algoritmo EM ha moltissime versioni, tra cui le più note sono: l'algoritmo SEM (Supplemented EM), che fornisce la matrice di varianza e covarianza delle stime, l'algoritmo ECM (Expectation/Conditional Maximization) che semplifica il passo M quando la massimizzazione non è diretta, l'algoritmo MCEM (Monte Carlo EM) che cerca di valutare numericamente il passo E quando questo è difficile da calcolare.
Molti sono stati i tentativi in letteratura per rendere più rapido l'algoritmo EM, si veda per esempio Louis (1982), Horng (1987).
Negli ultimi anni le estensioni dell'algoritmo si sono moltiplicate, tanto che è difficile elencare tutte le novità: ogni estensione cerca di rimediare agli svantaggi dell'algoritmo esistenti. Per un riferimento completo alle diverse versioni si veda Mclachlan, Geoffrey [6]

Si ricorda che la presentazione fatta in questa tesi non ha l'obiettivo di dare un'esaustiva descrizione dell'algoritmo, ma vuole solamente fornire una presentazione a livello intuitivo, cercando comunque di dare una minima formalizzazione complessiva.
Nel prossimo capitolo si presenterà un'applicazione pratica dell'algoritmo su dati reali allo scopo di chiarire il funzionamento operativo di questo strumento, che risulta essere uno dei più usati in presenza di dati mancanti.

Capitolo 3

Binge Drinking

É aumentato tra i giovani di 11-18 anni il consumo di bevande alcoliche. La notizia emerge dal focus elaborato dall'ISS secondo cui nel corso degli ultimi anni a partire dal 1998 sono aumentate per entrambi i sessi le prevalenze di consumatori: se nel 1998 l'abitudine al bere caratterizzava il 18.2% dei maschi e il 12% delle femmine tra i 14 e i 18 anni, nel 2003 le percentuali sono salite rispettivamente al 25% e al 19% (Istituto Superiore Sanità [9]).
Con l'espressione binge drinking si fa riferimento all'abitudine di consumare quantità eccessive (convenzionalmente 6 o più bicchieri di bevande alcoliche, anche diverse) in una singola occasione. Questo comportamento è presente prevalentemente nei paesi del nord Europa, ma si è fortemente radicato anche nel nostro Paese, in particolar modo nella fascia giovanile della popolazione, 'contaminando' anche le generazioni di adulti e anziani, prevalentemente di sesso maschile (Istituto Superiore Sanità [9]). Proprio per questo motivo a decorrere dall'anno 2003, nell'indagine Multiscopo sulle famiglie (ISTAT) sono state introdotte domande relative all'assunzione di alcol e sul fenomeno del binge drinking nella sezione 'Bevande'; per la prima volta tale sezione è stata estesa anche alla popolazione di 11-13 anni (in precedenza si partiva dai 14 anni di età).
In Italia la modalità di consumo degli alcolici sta infatti sostanzialmente cambiando: in passato il suo consumo era tendenzialmente moderato e si trattava principalmente di vino, assunto prevalentemente durante i pasti (bere vino per accompagnare i pasti o in occasioni particolari, infatti, fa parte della storia e cultura del nostro paese) (ISTAT

[7]); negli ultimi anni, invece, l'assunzione di alcol si è estesa a situazioni differenti da quelle tradizionalemnte conviviali: è sempre più frequente bere fuori dall'orario dei pasti e, soprattutto, allo scopo di ubriacarsi.

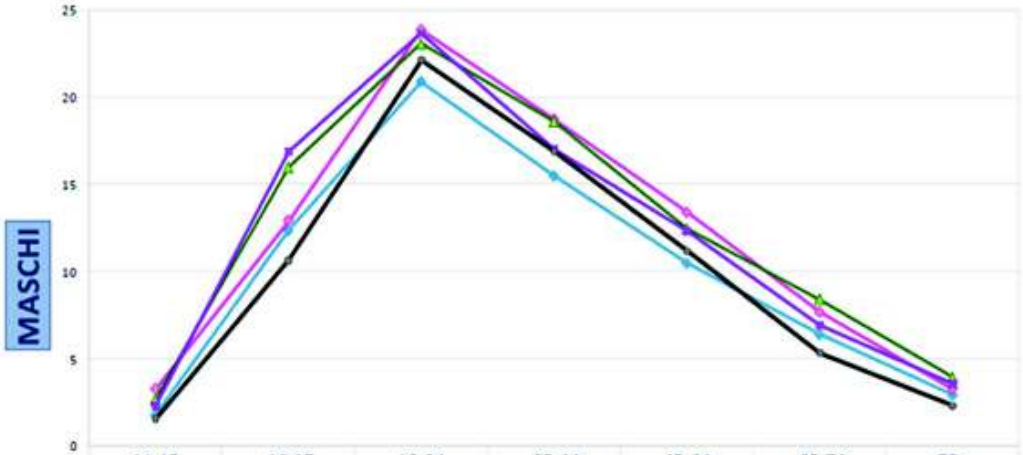
3.1 Binge drinking: fenomeno in crescita

Il bere per ubriacarsi si sta estendendo a fasce sempre più ampie di popolazione, in particolar modo ai giovanissimi (Arcidiacono, Caianiello [3]).
Come riporta l'Istituto Nazionale della Sanità [9], la percentuale dei binge drinkers è più elevata tra gli uomi che tra le donne, per tutte le fascie d'età ad eccezione di quella al di sotto del limite legale (16 anni), in cui non si registrano differenze sostanziali tra le percentuali.
Il fenomeno, inoltre, cresce all'aumentare dell'età e raggiunge i valori più elevati per entrambi i sessi a 18-24 anni, successivamente la percentuale torna in diminuzione.
Le figure (3.1) e (3.2) mostrano efficacemente tutti questi elementi. Il picco per i maschi si registra in corrispondenza alla classe di età 18-24; per le ragazze, al contrario, si ottiene intorno ai 16-17 anni sino al 2006, solo di recente si è spostato intorno ai 18-24 anni, quasi a suggerire che le binge drinkers delle generazioni precedenti abbiano mantenuto tali abitudini facendo 'traslare' la curva nel tempo.
A livello nazionale, molti sono stati i piani d'azione a tutela di questa fascia di popolazione che è considerata quella più a rischio per le problematiche correlate all'alcol. É pertanto importante cercare di migliorare le conoscenze riguardo ad un'abitudine relativamente nuova per il nostro Paese e distante dalle abitudini mediterranee che traevano ispirazione dalla moderazione e dal consumo di alcolici ai pasti.

3.2 Una prima analisi: creazione del modello logistico

Come già accenato, al fine di poter analizzare il fenomeno del binge drinking nel contesto italiano, a partire dall'indagine Multiscopo ISTAT dell'anno 2003, sono state introdotte delle nuove domande circa l'assunzione di alcol. In particolare, la domanda

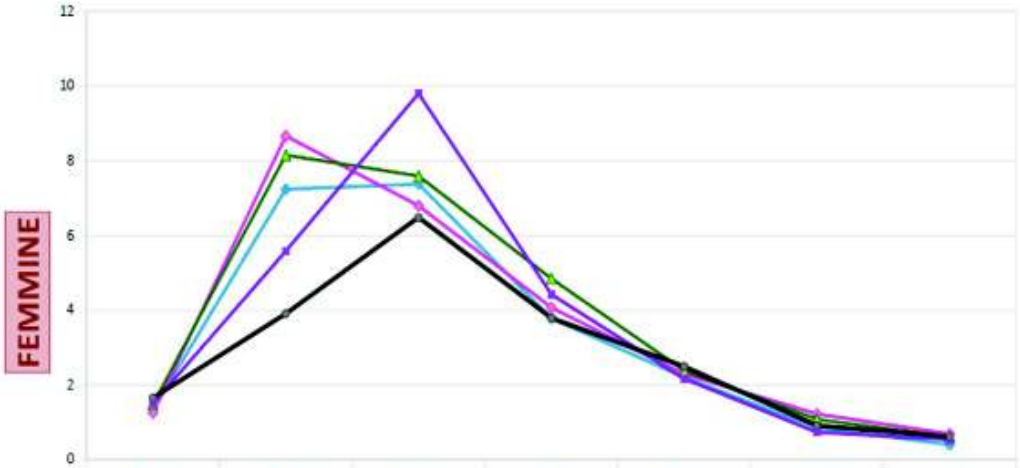
Figura 3.1: Individui che hanno sperimentato il *binge drinking*, frequenze percentuali per cento maschi della stessa fascia d'età. Anno 2003 - 2008. Italia



	11-15	16-17	18-24	25-44	45-64	65-74	75+
2003	1,7	12,3	20,9	15,5	10,5	6,4	2,9
2005	3,2	12,9	23,9	18,7	13,4	7,6	3,2
2006	2,6	16,0	23,1	18,6	12,4	8,4	3,9
2007	2,3	16,9	23,6	17,0	12,3	6,9	3,5
2008	1,5	10,6	22,1	16,9	11,2	5,3	2,3

Fonte: Elaborazioni Osservatorio Nazionale Alcol CNESPS e WHO CC Research on Alcohol su dati Istat Indagine "Multiscopo sulle Famiglie-Aspetti della vita Quotidiana". Anno 2009

Figura 3.2: Individui che hanno sperimentato il *binge drinking*, frequenze percentuali per cento femmine della stessa fascia d'età. Anno 2003 - 2008. Italia



	11-15	16-17	18-24	25-44	45-64	65-74	75+
2003	1,3	7,2	7,4	3,8	2,2	0,9	0,4
2005	1,3	8,7	6,8	4,1	2,3	1,2	0,7
2006	1,5	8,2	7,6	4,8	2,4	1,1	0,6
2007	1,5	5,6	9,8	4,4	2,2	0,7	0,5
2008	1,7	3,9	6,5	3,8	2,5	0,9	0,6

Fonte: Elaborazioni Osservatorio Nazionale Alcol CNESPS e WHO CC Research on Alcohol su dati Istat Indagine "Multiscopo sulle Famiglie-Aspetti della vita Quotidiana". Anno 2009

Figura 3.3: Domanda relativa all'assunzione di alcol, Indagine multiscopo sulle Famiglie: Aspetti della vita Quotidiana. ISTAT, 2005.

relativa al binge drinking interroga i rispondenti sul consumo di almeno 6 bicchieri di bevande alcoliche in un'unica occasione negli ultimi 12 mesi, vedi immagine (3.3).

Per l'analisi riportata di seguito, utilizzeremo i dati dell'Indagine Multiscopo sulle Famiglie -Aspetti della vita Quotidiana- dell'anno 2005 (ISTAT).
Per questa prima analisi ignoriamo la presenza di dati mancanti nella variabile *binge drinking*, questi verrano presi in considerazione nel successivo capitolo. Avremo quindi modo di vedere i differenti risultati che si possono ottenere ignorando o meno i missing. Le analisi presentate di seguito non tengono quindi conto della presenza di dati mancanti nella variabile *binge drinking*, né si preoccupano delle distorisioni che tale mancanza può causare. Anche altre variabili che verrano di seguito utilizzate presentano dei valori mancanti, ma si preferisce trattarle come completamente osservate, imputando i valori mancanti nelle variabili quantitative e creando modalità separate per i valori mancanti nelle variabili nominali e ordinali. Questo perchè i missing presenti nelle altre variabili sono poco numerosi e non si ritiene che influiscano particolarmente sul calcolo delle stime di interesse.
L'obiettivo è quello di determinare i fattori di rischio per il binge drinking. A riguardo sono state prese in considerazione le seguenti variabili:

- Età: come presentato nel paragrafo precedente, il fenomeno binge drinking è particolarmente diverso in base alla fascia di età degli intervistati. Nella seguente analisi si è interessati a studiare tale fenomeno per i giovanissimi, dagli 11 ai 17 anni;

- Sesso: anche il genere comporta un approccio tendenzialmente diverso al fenomeno in questione, e per questo motivo non può essere ignorato;

- Ripartizione geografica: esiste una sostanziale differenza nell'assunzione di alcol da parte dei giovani in base alla ripartizione geografica, in particolare il binge drinking è più diffuso nell´Italia Settentrionale (Arcidiacono, Caianiello [3]);

- Abitudine al fumo: l'alcol e il fumo si ritengono due comportamenti a rischio fortemente associati, tendenzialmente *'chi eccede nel consumo di alcol spesso associa anche altri comportamenti a rischio, uno di questi è l'abitudine al fumo'.* (ISTAT [7]);

- Consumo di alcolici da parte dei genitori: il fatto che i genitori siano consumatori di alcolici, anche se in modo molto moderato, potrebbe aumentare la probabilità che i giovani riescano ad avvicinarsi alle bevande alcoliche, essendo presenti già all'interno dell'abitazione (Arcidiacono, Caianiello [3]);

- Abitudine al fumo dei genitori: per quanto detto nei due punti precedenti, anche l'abitudine al fumo dei genitori non può essere ignorata;

- Titolo di studio dei genitori: tendenzialmente, l'abitudine a bere alcolici quotidianamente decresce all'aumentare del titolo di studio (ISTAT [8]), e questo potrebbe influire sull'assunzione degli alcolici da parte dei figli per quanto detto nei punti precedenti;

- Attività fisica continuativa: si può ritenere che praticare attività sportiva in modo continuativo sia legato ad uno stile di vita più salutare, e quindi ad un minor uso di bevande alcoliche;

- Benessere economico generale della famiglia: analogamente, una situazione di benessere elevato dovrebbe corrispondere ad uno stile di vita più salutare e, dunque, ad un minor uso di bevande alcoliche.

Nella tabella (3.1) sono descritte le variabili per tipologia e modalità di risposta.

La selezione sopra riportata delle variabili si basa su un'analisi di contenuto, ma è al-

trettanto importante che la selezione si basi anche su analisi di tipo statistico, in quanto, in previsione di uno studio di tipo multivariato, bisogna 'scremare' le variabili esplicative che sono potenziali determinanti della variabile dipendente (Fabbris [12]). Per le variabili dicotomiche, in questo caso *sesso*, *abitudine fumo* e *consumo alcolici dei genitori*, è necessario calcolare la sensibilità e la specificità: tutte le variabili hanno riportato una specificità non inferiore a 0,7 e una sensibilità non inferiore a 0.09. Verranno quindi utilizzate nelle analisi succesive.
La variabiale quantitativa discreta *età* è stata analizzata applicando un modello di regressione logistica avente come predittori solo l'intercetta e la variabile in questione. L'analisi ha portato al rifiuto dell'ipotesi nulla di indipendenza, con un livello di significatività della variabile inferiore a 0.001. Per la variabile quantitativa continua *benessere famiglia* si è utilizzato lo stesso procedimento, ma in questo caso l'analisi ha portato ad una accettazione dell'ipotesi nulla (p-value superiore a 0.5), abbiamo comunque deciso di non abbandonare tale variabile, credendo che in un'ottica multivariata potrebbe essere significativa in relazione ad altre variabili.
Tutte le variabili rimanenti sono di tipo ordinale (*attività fisica*, *titolo studio genitori*, *abitudine fumo genitori*) o nominale (*ripartizione geografica*) e per valutare la loro significatività individuale è stato utilizzato il coefficiente χ^2 di Pearson. Solamente le variabili *ripartizione geografica* e *abitudine fumo genitori* sono risultate significativamente correlate con la variabile *binge drinking* (p-value inferiore a 0.0001). In ogni caso, non essendo il numero di variabili elevato, anche *attività fisica* e *titolo studio genitori* verrano inserite nelle analisi successive, in quanto può essere possibile che diventino significative in interazione con altre. Cerchiamo ora di valutare il contributo complessivo delle singole variabili al fenomeno del binge drinking, passando alla selezione delle variabili in un'ottica multivariata: applichiamo dunque il metodo di selezione *stepwise* per la regressione logistica.
Il modello stimato è il seguente:

$$y = età + sesso + ripartizione\ geografica + abitudine\ fumo + consumo\ alcolici\ genitori + età{*}sesso$$

dove *età*sesso* è l'interazione tra la variabile *età* e la variabile *sesso*. Vengono presentati in tabella (3.2) le stime dei parametri del modello creato.

Il modello ottenuto si ritiene sensato sia dal punto di vista statistico che logico, non si rende quindi neccesaria la forzatura di variabili non accettare dalla procedura stepwise.

3.3 Discussione e conclusioni

Da questa prima analisi potrebbero sorgere alcuni dubbi sulle stime ottenute, soprattutto se confrontate con la letteratura precedente. Per esempio sembra che l'essere di sesso femminile aumenti di circa 5 volte la probabilità di sperimentare il binge drinking ($e^{1.56} = 4.7$), in contrasto con quanto detto nel paragrafo (3.1). Invece, a conferma di quanto detto nel paragrafo (3.1), vediamo che esiste un interazione negativa (inibizione) tra la variabile sesso e la variabile età: l'età sembra agire agire in modo diverso a seconda del sesso. Tuttavia, la stima dei coefficienti di regressione rispetto la variabile *ripartizione geografica* sembrano essere troppo poco 'incisive': il fenomeno del binge drinking, in letteratura è considerato quasi 'invisibile' nel Sud Italia (Arcidiacono, Caianiello [3]), ma l'odds ratio, pur confermano tale tendenza riporta un effetto inibitorio: $e^{-0.40} = 0.68$, ma non di così grande portata come ci saremmo aspettati.
In seguito a questa analisi potrebbero quindi sorgere alcuni dubbi sulla validità delle stime ottenute: forse, il metodo utilizzato non è il più adatto e *forse* è il caso di prendere in considerazione la presenza dei dati mancanti.
Nel prossimo capitolo verrà vagliata questa possibilità, cercando di analizzare il fenomeno del binge drinking in maniera più adeguata. Torneremo quindi sulle stime dei coefficienti di regressione e sugli odds ratio per comprendere come, ed in che misura, la presenza di dati mancanti abbia probabilmente distorto la reale comprensione del fenomeno.

Tabella 3.1: Descrizione e codifica delle variabili utilizzate nella procedura stepwise

Variabile	**Tipologia**	**Descrizione**
Età	variabile quantitativa discreta	numero di anni compiuti
Benessere famiglia	variabile quantitativa continua	livello benessere valutato da 0 a 1
Sesso	variabile dicotomica	1=maschio 2=femmina
Ripartizione geografica	variabile nominale	1=Nord 2=Centro 3=Sud/Isole
Fumo	variabile nominale	0=Non fumatore 1=Fumatore 2=Dato mancante
Attività fisica	variabile ordinale	0=No 1=Raramente 2=1 o più volte a settimana 3=1 o più volte al mese
Consumo alcolici genitori	variabile nominale	0=Nessun genitore 1=Almeno un genitore 2=Dati mancanti
Abitudine fumo genitori	variabile ordinale	0=Nessun genitore 1=Almeno un ex fumatore 2=Almeno un fumatore 3=Dati mancanti

Tabella 3.2: Stime dei coefficienti del modello logistico con variabile risposta 'sperimentazione del fenomeno *binge drinking*'

Parametro	Stima coefficiente	Standard Error	p-value
Intercetta	-8.14	0.66	$< .0001$
Età	0.43	0.10	$< .0001$
Sesso 2	1.56	0.96	0.1023
Rip. geografica[1]*: Centro Italia*	-0.32	0.19	0.0829
Rip. geografica[1]*: Sud Italia/Isole*	-0.40	0.14	0.0040
Abitudine fumo[2]*: Fumatore*	1.50	0.17	$< .0001$
Abitudine fumo[2]*: Dati mancanti*	3.08	0.19	$< .0001$
Alcol genitori[3]*: Almeno un genitore*	0.19	0.15	0.2007
Alcol genitori[3]*: Dati mancanti*	0.59	0.20	0.0036
*Età*sesso*	-0.15	0.07	0.0228

[1] la modalità di riferimento è il 'Nord Italia'.

[2] la modalità di riferimento è il 'Non fumantore'.

[3] la modalità di riferimento è 'Nessun genitore'.

Capitolo 4

... e i dati mancanti?

Nel capitolo precedente abbiamo analizzato il fenomeno del binge drinking e abbiamo ottenuto dei risultati non privi di senso, ma ci hanno fatto sorgere dei piccoli dubbi a riguardo: alcuni effetti sembravano troppo accentuati mentre altri erano poco 'incisivi'. È quindi necessaria una rivalutazione del procedimento usato per capire cosa non abbiamo colto o guardato con abbastanza attenzione. Molto probabilmente la presenza di alcuni dati mancanti ha avuto un ruolo decisivo sul livello di distorsione avuto, è quindi necessario perdere in considerazione questa possibilità. Analizziamo quindi la variabile *binge brinking* per capire come sono distribuiti i missing: riassumendo le risposte ottenute si ha che 3291 individui hanno di non aver vissuto il fenomeno in questione, mentre solo 171 dichiarano di averlo sperimentato. Notiamo che sono 3752 i soggetti presenti nel set di dati, ma 290 (circa 8%) non hanno fornito risposta. Se questi 290 valori mancanti sono sostanzialmente diversi da quelli osservati siamo in una situazione di dati mancnati 'non ignorabili', e questo potrebbe aver distorto i risultati ottenuti. L'ipotesi di una differenza tra i dati osservati e quelli non osservati non è così improbabile visto l'argomento in questione: come già precisato, le domande che riguardano il comportamento relativo all'uso di alcolici sono considerate sensibili dagli intervistati e quindi più soggette a mancate risposte, soprattutto se la risposta potrebbe compromettere 'la propria immagine sociale' in quanto prova di un cattivo comportamento. Sempre riportando le parole di Bosco [1]

> 'Un ruolo decisivo è giocato quindi da quelle conoscenze, credenze e convinzioni che si riferiscono alla norma socialmente condivisa, a quella rappresen-

tazione interiorizzata di ciò che è 'giusto' nella società in cui viviamo.'

Grazie alla procedura ideata da Ibrahim e Lipsitz (che utilizza il 'metodo dei pesi' introdotto nel Capitolo 2), sarà possibile correggere questa situazione e ottenere delle stime migliori di quelle avute nel capitolo precedente. Bisogna sottolineare, come nel capitolo precedente, che le variabili indipendenti usate nel modello creato, non sono tutte completamente osservate ma, si preferisce trattare come completamente osservate, imputando i valori mancanti nelle variabili quantitative o creando modalità separate per i valori mancanti nelle variabili ordinali o nominali. Questo perché, come già detto nel Capitolo 3, i missing presenti nelle covariate sono poco numerosi e non si ritiene che influiscano particolarmente sul calcolo delle stime di interesse. In ogni caso, se si volesse estendere il procedimento a tutte le covariate sarebbe necessario uno sforzo solamente computazionale, ma non concettuale: una volta compreso il meccanismo, esso può essere facilmente esteso a tutte le variabili desiderate. Si è perciò deciso di trattare come completamente osservate le variabili dipendenti.

4.1 Modello e notazione

Abbiamo stabilito che la non risposta alla domanda in questione dipende da diversi fattori, sia riguardanti l'individuo stesso, sia legati all'aver vissuto o meno l'evento in questione. Siamo quindi in un caso di dati mancanti 'non ignorabili' e diventa necessario specificare un meccanismo generatore di dati mancanti, che cercherà di spiegare la probabilità di non risposta. A questo scopo, la variabile z che varrà 1 se la variabile risposta è mancante, 0 se è osservata. La sua distruzione dipenderà da alcune variabili esplicativa, le stesse che spiegano la probabilità di sperimentare o meno il fenomeno del binge drinking, oltre alla variabile *binge drinking* stessa. In questo modo la variabile *binge drinking* di nostro interesse verrà trattata come variabile esplicativa con dati mancanti nel modello m .
Per facilità esplicativa, in seguito la variabile *binge drinking* sarà chiamata y.

Essendo in un caso di dati 'non ignorabili', la completezza dei dati del modello consiste nella distribuzione congiunta della variabile risposta y e l'indicatore di dati mancanti

m. Dal momento che le variabili esplicative x sono considerate pienamente osservate, queste sono trattate come fisse.
Supponiamo che $y_1, ..., y_n$ siano osservazioni indipendenti, in cui ogni y_i ha una distribuzione binomiale con $d_i = 1$ dimensione del campione e p_i probabilità di successo, $i = 1, ..., n$. Inoltre, prendiamo $x_i^T = (x_{i1}, ..., x_{ip})$ indicante il p-vettore delle variabili esplicative per l'osservazione i -esima, e $\beta^T = (\beta_1, ..., \beta_p)$ indicante il corrispondente p-vettore dei coefficienti di regressione. Viene inserito un 1 nel vettore x_i^T se viene utilizzata l'intercetta. La verosimiglianza per dati completi $(y_i|x_i, \beta)$ è data da:

$$f(y_i|x_i, \beta) = \exp(y_i \beta x_i^T - \log(1 + \exp(\beta x_i^T))). \tag{4.1}$$

Avendo specificato

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

per $i = 1, ..., n$. Specifichiamo un modello di regressione logistica per m_i. Prendiamo $z_i = (x_i, y_i)$ e sia α un $(p + 1)$-vettore di coefficienti di regressione per r_i. Abbiamo:

$$f(m_i|z_i, \alpha) = \exp(m_i \alpha z_i^T - \log(1 + \exp(\alpha z_i^T))), \tag{4.2}$$

Vediamo che in (4.2) abbiamo che la probabilità di y_i mancante (m_i =1) dipende, sia dalla risposta stessa della variabile y_i , sia dal vettore x_i di variabili esplicative. Infatti, notiamo che, se $\alpha_{p+1} = 0$, allora $f(r_i|z_i, \alpha)$ non dipende da y_i , quindi i dati mancanti sono distribuiti in modo casuale e il meccanismo di dati mancanti è ignorabile. Ma, se $\alpha_{p+1} \neq 0$, allora il meccanismo di dati mancati dipende da y_i e non è ignorabile. In conclusione, la verosimiglianza per dati completi è ottenuta da:

$$f(m, y|x, \beta, \alpha) = \prod_{i=1}^{n} f(y_i|x_i, \beta) f(m_i|z_i, \alpha) \tag{4.3}$$

4.2 Stima dei coefficienti di regressione

La verosimiglianza per dati completi in (4.3) tratta in sostanza la y_i come variabile esplicativa mancante nel modello di $(m_i|z_i, \alpha)$. In questo modo, grazie ad Ibrahim (1990), le stime di massima verosimiglianza (α, β) possono essere ottenute grazie all'algoritmo

EM con il 'metodo dei pesi'.

Dalla (2.17) si ha che il contributo individuale dell'i -esima osservazione nella log-verosimiglianza è dato da:

$$E[\ell(\alpha,\beta|x_i,y_i,r_i)] = \begin{cases} \sum_{y_i=0}^{d_i} \ell(\alpha,\beta|x_i,y_i,r_i) f(y_i|m_i,x_i,\alpha,\beta) & \text{se } y_i \text{ è mancante} \\ \ell(\alpha,\beta|x_i,y_i,m_i) & \text{se } y_i \text{ è osservata} \end{cases} \tag{4.4}$$

Dove $\ell(\alpha,\beta|x_i,y_i,m_i) = \log(f(m_i,y_i|x_i,\beta,\alpha))$ è la funzione di log-verosimiglianza per i dati completi per l'osservazione i -esima e $f(y_i|r_i,x_i,\alpha,\beta)$ è la distribuzione condizionata dei dati mancanti in base ai dati osservati. Bisogna sottolineare che nella (4.4) la sommatoria è sulle y_i se queste sono mancanti, ma non se sono osservate. Come si è mostrato nel paragrafo (2.6.2) il passo E della (4.4) assume la forma di una log-verosimiglianza ponderata sui dati completi, dove $f(y_i|m_i,x_i,\alpha,\beta)$ sono i pesi. Notiamo che questi pesi possono essere espressi come:

$$f(y_i|m_i,x_i,\alpha,\beta) = \frac{f(y_i|x_i,\beta)f(m_i|z_i,\alpha)}{\sum_{y_i=0}^{d_i} f(y_i|x_i,\beta)f(m_i|z_i,\alpha)} \tag{4.5}$$

sempre per quanto detto nel paragrafo (2.6.2) e per il teorema di Bayes.

In questo modo i pesi sono calcolati interamente usando i modelli per y_i e m_i calcolati sui dati completi.

Guardiamo con attenzione alla sommatoria riportata al denominatore della funzione dei pesi: il limite inferiore è 0 ($y_i = 0$) mentre il limite superiore è d_i. Nell'esempio in questione y_i può essere o uguale a 0, oppure uguale a 1. Quindi il numero di prove d_i è 1. Riassumendo, la sommatoria viene calcolata sostituendo a y_i prima il valore 0 e poi il valore 1. Inoltre, dato che i pesi sono utilizzati nella (4.4) solamente se la y_i è mancante, avremmo che m_i è sempre pari a 1 nella funzione dei pesi. I pesi espressi nella (4.5) saranno quindi pari a:

$$\frac{f(y_i|x_i,\beta)f(m_i=1|z_i,\alpha)}{f(y_i=1|x_i,\beta)f(m_i=1|z_i,\alpha)+f(y_i=0|x_i,\beta)f(m_i=1|z_i,\alpha)}$$

Utilizzando la (4.4) possiamo scrivere il passo E per tutte le n osservazioni alla $(t+1)$-esima iterazione come

$$Q(\alpha,\beta|\alpha^t,\beta^t) = \sum_{i=1}^{n}\sum_{y_i}^{d_i} w_{iy_i,(t)}\ell(\alpha,\beta|x_i,y_i,m_i), \tag{4.6}$$

dove

$$w_{iy_i,(t)} = f(y_i|m_i, x_i, \alpha, \beta) = \frac{f(y_i|x_i, \beta)f(m_i|z_i, \alpha)}{\sum_{y_i=0}^{d_i} f(y_i|x_i, \beta)f(m_i|z_i, \alpha)} \tag{4.7}$$

Se y_i è osservata, segue dalla (4.4) che $w_{iy_i,(t)} = 1$. Analizziamo ancora le sommatorie riportate nella (4.6). La prima sommatoria viene calcolata su tutti i rispondenti presenti nel dataset, sia su quelli per i quali abbiamo l'osservazione della variabile, sia su quelli per i quali l'osservazione è mancante. Questo perchè stiamo scrivendo il passo E per tutte le osservazioni.
La seconda sommatoria ha come limite inferiore 0 ($y_i = 0$) mentre il limite superiore è d_i che nell'esempio in questione abbiamo posto pari a 1. Questo significa che, nel caso la y_i sia mancante, bisogna includere nella verosimiglianza sia il caso che la y_i sia uguale a zero, sia il caso in cui la y_i sia pari a 1. Quindi, prendendo la i -esima osservazione mancante, avremmo che rientrerà nella verosimiglianza 'due volte': una volta 'ipotizzando' il suo valore pari a 1 (pesando per la probabilità che sia pari a 1):

$$\ell(\alpha, \beta|x_i, y_i = 1, m_i = 1)\frac{f(y_i = 1|x_i, \beta)f(m_i = 1|z_i, \alpha)}{\sum_{y_i=0}^{d_i} f(y_i|x_i, \beta)f(m_i|z_i, \alpha)},$$

e una volta 'ipotizzando' il suo valore pari a 0 (pesando per la probabilità che sia pari a 0):

$$\ell(\alpha, \beta|x_i, y_i = 0, m_i = 1)\frac{f(y_i = 0|x_i, \beta)f(m_i = 1|z_i, \alpha)}{\sum_{y_i=0}^{d_i} f(y_i|x_i, \beta)f(m_i|z_i, \alpha)}$$

Il passo M massimizza la funzione in (4.6), che equivale a calcolare la stima di massima verosimiglianza pesata. Ciò comporta la stima di una regressione logistica sia per $(y_i|x_i, \beta)$ e $(r_i|z_i, \alpha)$ nella quale ogni osservazione mancante è sostituita da una coppia (in quanto sono 2 i possibili valori per la y_i) di osservazioni pesate. Abbiamo quindi che la (4.6) trasforma il problema in una stima sui dati completi pesati e rende la stima molto semplice da calcolare con un qualsiasi software statistico.

4.3 Procedura operativa

Riassumiamo ora i passaggi necessari per costruire l'algoritmo:

1. Selezionare un insieme x di variabili esplicative per la variabile risposta y;
2. Creare l'insieme z di variabili esplicative per la variabile m , composto dall'insieme x unito alla variabile y , $z = (x, y)$?;
3. Calcolare i coefficienti di regressione α e β che verranno utilizzati come le stime del passo 0: $\alpha^{(0)}$ e $\beta^{(0)}$;
4. Costruire una funzione che calcola i pesi $w_{iy_i,(t)}$ dati i coefficienti di regressione $\alpha^{(t)}$ e $\beta^{(t)}$;
5. Pesare i valori presenti nel dataset con i pesi costruiti al punto 4;
6. Ricalcolare le stime dei coefficienti di regressione $\alpha^{(t+1)}$ e $\beta^{(t+1)}$ sui nuovi valori pesati;
7. Con le nuove stime ricalcolare i pesi (punto 4) e ripartire dal punto 5;
8. Continuare fino a quando non vengono soddisfatti i criteri di convergenza.

Implementiamo ora, attraverso l'uso del software R i vari passi elencati utilizzando i dati in possesso:

Passo 1: Come esposto nel Capitolo 3 le variabili che possono descrivere la variabile risposta y (sperimentazione dell'evento specificato) sono:

$$y = età + sesso + ripartizione\ geografica + abitudine\ fumo + consumo\ alcolici\ genitori + età{*}sesso$$

Dove *età*sesso* è l'interazione tra la variabile *età* e la variabile *sesso*. Nella tabella (4.1) è presentata una descrizione dettagliata delle variabili.

Passo 2: Il vettore z sarà composto da:

Tabella 4.1: Descrizione e codifica delle variabili utilizzate nel modello logistico che spiega il fenomeno *binge drinking*

Variabile	**Tipologia**	**Descrizione**
Età	variabile quantitativa continua	numero di anni compiuti
Sesso	variabile dicotomica	1=maschio 2=femmina
Ripartizione geografica	variabile nominale	1=Nord 2=Centro 3=Sud/Isole
Abitudine fumo	variabile nominale	0=Non fumatore 1=Fumatore 2=Dato mancante
Consumo alcolici genitori	variabile nominale	0=Nessun genitore 1=Almeno un genitore 2=Dato mancante

*z = (y, età, sesso, ripartizione geografica, abitudine fumo, consumo alcolici genitori, età*sesso*

Passo 3: Le stime di $\beta^{(0)}$ e $\alpha^{(0)}$ sono riportate nella tabella (4.2);

Passo 4, 5, 6 e 7: Vengono pesati i valori e ricalcolate le nuove stime di $\alpha^{(t+1)}$ e $\beta^{(t+1)}$, vengono riportati nelle tabelle (4.3) e (4.4) i valori ottenuti ad ogni iterazione dell'algoritmo;

Passo 8: Nel seguente esempio facciamo procedere l'algoritmo fino a quando la log-verosimiglianza calcolata nella i -esima iterazione dell'algoritmo non consegue un aumento considerevole rispetto all'iterazione precedente. Per 'aumento considerevole' si intende maggiore di 0.1.
Questo avviene alla 19-esima iterazione come mostrato nella tabella (4.5).

Tabella 4.2: Stime dei coefficienti di regressione β e α al passo 0 dell'algoritmo EM

Parametro	$\beta^{(0)}$	$\alpha^{(0)}$
Intercetta	-8.14	-1.17
y	-	3.42
Età	0.43	-0.20
Sesso 2	1.56	-0.04
Rip. geografica[1]*: Centro Italia*	-0.32	0.41
Rip. geografica[1]*: Sud Italia/Isole*	-0.40	0.54
Abitudine fumo[2]*: Fumatore*	1.50	-0.82
Abitudine fumo[2]*: Dati mancanti*	3.08	5.14
Alcol genitori[3]*: Almeno un genitore*	0.19	-0.16
Alcol genitori[3]*: Dati mancanti*	0.59	0.50
*Sesso*età*	-0.15	-0.02

[1] la modalità di riferimento è il 'Nord Italia'.

[2] la modalità di riferimento è il 'Non fumantore'.

[3] la modalità di riferimento è 'Nessun genitore'.

4.4 Discussione

Per prima cosa bisogna sottolineare che qualsiasi commento che verrà di seguito fatto si baserà solo sui valori assunti dai coefficienti di regressione α e β e non tratterà in nessun modo la loro significatività.
Questo non perchè si ritenga irrilevante la significatività delle stime (anzi, è indubbia la loro importanza e essenzialità), ma perchè, l'algoritmo EM non fornisce automaticamente la matrice di varianza-covarianza asintotica delle stime di massima verosimiglianza, ma necessita di procedimenti particolari: alcuni autori, ad esempio Louis (1982), hanno proposto metodi per calcolare tale matrice, ma ogni metodo è specifico del problema considerato e generalmente comporta calcoli che possono essere complessi. In ogni caso si è deciso di non trattare questo argomento nella presente tesi e, pur essendo consapevoli del limite fissato, trattiamo come significative tutte le stime ottenute.

Il metodo presentato ha l'obiettivo di stimare i coefficienti di una regressione logistica quando la variabile risposta presenta dei valori mancanti e il meccanismo di dati mancanti

Tabella 4.3: Stima dei coefficienti di regressione α per ogni iterazione dell'algoritmo EM

i	*Int.*	*y*	*Età*	*Sesso: F*	*Rip.geo. Centro*	*Rip.geo. Sud/Isole*	*Fumo Fumat.*	*Fumo D.M.*[1]	*Alcol Un gen.*	*Alcol D.M.*[1]	*Sesso* età*
0	-4.49	1.03	0.05	1.25	0.21	0.30	-0.04	4.97	0.00	0.71	-0.11
1	-4.70	0.75	0.07	1.37	0.20	0.28	0.03	5.06	0.01	0.73	-0.12
2	-4.81	0.63	0.08	1.41	0.19	0.27	0.06	5.12	0.01	0.74	-0.13
...	...	...	...	...	...	...	...	...	...	...	...
17	-5.22	-0.01	0.11	1.55	0.15	0.24	0.17	5.25	0.04	0.77	-0.14
18	-5.22	-0.02	0.11	1.55	0.15	0.24	0.17	5.25	0.04	0.77	-0.14
19	-5.22	-0.03	0.12	1.55	0.15	0.24	0.17	5.25	0.04	0.77	-0.14

[1] Dati Mancanti

è 'non ignorabile'. L'idea chiave del metodo è che la risposta mancante può essere trattata come variabile dipendente mancante nel modello della m e questo ci permette di applicare il 'metodo dei pesi' di Ibrahim.

Sotto l'ipotesi che tutte le stime siano significative ci rendiamo conto di essere in presenza di un meccanismo generatore di dati mancanti NMAR, in quanto la probabilità di non risposta è legata sia ai valori osservati, sia ai valori non osservati della variabile y. In questo caso, una analisi fatta sui soli dati osservati avrebbe portato sicuramente a stime distorte.
Trattando le stime dell'algoritmo EM come corrette e significative, vediamo che una stima della distorsione relativa è definita come $(\hat{\beta_{EM}} - \hat{\beta_{CC}})/\hat{\beta_{EM}}$, dove $\hat{\beta_{EM}}$ e $\hat{\beta_{CC}}$ denotano rispettivamente, le stime basate sull'algoritmo EM e sui casi completi. Per una completa visione delle distorsioni relative si veda la tabella (4.6).

L'errore relativo stimato è superiore al 7% in valore assoluto per tutti i coefficienti, arrivando a toccare anche il 167,8% nel parametro *Abitudine fumo 2* (modalità mancanti della variabile *abitudine fumo*). Bisogna sottolineare che la stima sui dati completi equivale alla stima di massima verosimiglianza sotto l'ipotesi di dati mancanti 'ignorabili' (MAR e MCAR), ottenuta nel Capitolo 3. In questo caso un ipotesi di dati mancanti 'ignorabili' ha portato a distorsioni sulle stime di grandi dimensioni.

Tabella 4.4: Stima dei coefficienti di regressione β per ogni iterazione dell'algoritmo EM

i	*Int.*	*Età*	*Sesso: F*	*Rip.geo. Centro*	*Rip.geo. Sud/isole*	*Fumo Fumatore*	*Fumo D.M.*[1]	*Alcol Un gen.*	*Alcol D.M.*[1]	*Sesso* età*
0	-11.44	0.64	1.52	-0.60	-0.80	1.66	3.61	0.36	0.75	-0.15
1	-12.02	0.67	1.49	-0.643	-0.85	1.64	3.37	0.37	0.77	-0.14
2	-12.09	0.68	1.44	-0.64	-0.86	1.63	3.07	0.37	0.78	-0.14
...	...	...	...	...	...	...	...	...	...	...
17	-11.72	0.66	1.04	-0.63	-0.85	1.62	1.20	0.34	0.76	-0.12
18	-11.72	0.66	1.04	-0.63	-0.85	1.621	1.17	0.34	0.75	-0.12
19	-11.72	0.66	1.03	-0.63	-0.85	1.62	1.15	0.34	0.75	-0.12

[1] Dati Mancanti

Ma come si sono modificate le stime dopo l'applicazione dell'algoritmo EM? Vediamo che la semplice analisi sui dati completi aveva portato a stimare una maggior effetto del sesso: l'essere femmina (*sesso*) sembrava far aumentare di quasi 5 volte ($e^{1.56} = 4.7$) la probabilità di sperimentare l'evento in questione, mentre un analisi sui dati 'completati' dall'algoritmo EM mostra che questo aumento è solo pari a 2.80 ($e^{1.03} = 2.80$), più coerente con ciò che viene riportato nella letteratura precedente.
Anche la stima del rischio relativo alla *ripartizione geografica* di appartenenza è ora più in linea con la letteratura, vediamo che l'essere meridionale *ripartizione geografica= 3* resta prottetivo nei confronti del binge drinking, aumentando però di intensità: ignorando i dati mancanti otteniamo una stima del rischio relativo pari a $e^{-0.40} = 0.68$, successivamente l'applicazione dell'algoritmo EM risulta pari a $e^{-0.85} = 0.43$.
La mancata risposta alla domanda relativa all'abitudine sul fumare avrebbe aumentato la probabilità di sperimentare l'evento di 21.8 volte ($e^{3.08} = 21.8$), mentre un'analisi più accurata tramite l'algoritmo EM fa diminuire questo aumento a 3.2 volte ($e^{1.15} = 3.2$).
L'aumento di un anno di età sembrava far aumentare la probabilità di bere 6 o più bicchierini di super alcolici in un'unica occasione di 1.5 volte ($e^{0.43} = 1.5$), mentre, dopo l'applicazione dell'algoritmo scopriamo un aumento di circa 2 volte ($e^{0.66} = 1.9$).
Vediamo come, se si fosse fatta l'analisi solo sui dati osservati, si sarebbe data maggiore (o minore) rilevanza agli effetti di alcune variabili rispetto a quella dovuta.
Riguardo alla probabilità di non-risposta, vediamo come sono cambiati i coefficienti di

Tabella 4.5: Differenza tra i valori delle logverosimiglianze per ogni iterazione dell'algoritmo

i	$\ell^{(i)} - \ell^{(i-1)}$	i	$\ell^{(i)} - \ell^{(i-1)}$
0	-	11	0.661
1	-	12	0.511
2	5.432	13	0.516
3	4.344	14	0.368
4	3.211	15	0.290
5	2.470	16	0.230
6	2.087	17	0.182
7	1.464	18	0.144
8	1.687	19	0.115
9	1.117	20	0.091
10	0.857		

regressione prima e dopo l'applicazione dell'algoritmo EM: un evidente cambiamento riguarda la variabile *sesso*, l'essere di sesso femminile sembrava far diminuire la probabilità di non risposta, mentre dopo l'applicazione dell'algoritmo, vediamo che questa aumenta di quasi 5 volte. La mancata risposta alla domanda relativa all'abitudine al fumo non sembra avere un influenza diversa prima o dopo l'applicazione dell'algoritmo, rimane comunque alta (il coefficiente di regressione relativo è pari a 5.25), ma la cosa non dovrebbe stupire molto essendo entrambe le domande 'sensibili' e riguardanti ambiti simili, vedi Bosco [1].

4.5 Conclusioni

Con questo capitolo si è conclusa la presente trattazione dell'algoritmo EM.
Riassumendo si sono volute evidenziare la capacità dell'algoritmo EM di diminuire il livello di distorsione ottenibile qualora non si tenga conto del meccanismo generatore dei dati mancanti. Abbiamo visto che esiste il rischio è di dare troppa (o troppa poca) importanza ad alcuni effetti, e grazie all'algoritmo EM questo pericolo può essere evitato.
Per chiarezza e completezza è nel seguito riportato il programma R utilizzato per la

Tabella 4.6: Stime dei coefficienti di regressione β con dati completi e dopo l'applicazione dell'algoritmo EM. Calcolo della distorsione relativa.

Parametro	$\hat{\beta}_{CC}$	$\hat{\beta}_{EM}$	$\frac{(\hat{\beta}_{EM}-\hat{\beta}_{CC})}{\hat{\beta}_{EM}}$
Intercetta	-8,14	-11,72	30,5 %
Età	0,43	0,66	34,8 %
Sesso 2	1,56	1,03	-51,5 %
Rip. geografica[1]*: Centro Italia*	-0,32	-0,63	49,2 %
Rip. geografica[1]*: Sud Italia/Isole*	-0,4	-0,85	52,9 %
Abitudine fumo[2]*: Fumatore*	1,5	1,62	7,4 %
Abitudine fumo[2]*: Dati Mancanti*	3,08	1,15	-167,8 %
Alcol genitori[3]*: Almeno un genitore*	0,19	0,34	44,1 %
Alcol genitori[3]*: Dati mancanti*	0,59	0,75	21,3 %
*Età*sesso*	-0,15	-0,12	-25,0 %

[1] la modalità di riferimento è il 'Nord Italia'.

[2] la modalità di riferimento è il 'Non fumantore'.

[3] la modalità di riferimento è 'Nessun genitore'.

creazione e implementazione dell'algoritmo EM (vedi figura (4.1)), ma è importante precisare che non esiste un programma 'standard', bensì è necessario 'costruire' l'algoritmo EM *ad hoc*.

Riguardo al fenomeno del binge drinking i risultati ottenuti dopo l'applicazione dell'algoritmo EM sembrano essere in linea con quanto riportato in letteratura, questo significa che in questo contesto non si potevano ignorare i dati mancanti, in quanto sostanzialmente diversi da quelli osservati. Si sottolinea quindi l'importanza di valutare la presenza di dati mancanti e rendere trasparenti tutte le decisoni prese a riguardo, ricordando che le basi per una buona ricerca scientifica si trovano non solo nella validità dei risultati ottenuti, ma soprattutto nella riproducibilità dei processi utilizzati e nella chiarezza d'intenti.

Figura 4.1: Programma R: creazione algoritmo EM

```
EM<-function(y,predl1,pred22,max.iter,tol,mem){
      #Creazione della funzione che calcola i pesi, dati predl=x*beta e pred2=y*alpha:
            pesil<-function(predl,pred2){
            wl<-((exp(predl-log(l+exp(predl))+(pred2)-log(l+exp(pred2))))/
                              (exp(predl-log(l+exp(predl))+pred2-log(l+exp(pred2)))
                        +exp(-log(l+exp(predl))+pred2-log(l+exp(pred2)))))
            w0<-((exp(-log(l+exp(predl))+(pred2)-log(l+exp(pred2))))/
                              (exp(predl-log(l+exp(predl))+pred2-log(l+exp(pred2)))
                              +exp(-log(l+exp(predl))+pred2-log(l+exp(pred2)))))
                        newpesi=ifelse(r==l,ifelse(y==0, w0, wl),l)
                        return(newpesi)}
      #Inizio dell'algoritmo
         for(iter in 0:max.iter){
            if(iter==0) {predl=predll} else{predl=predA}
            if(iter==0) {pred2=pred22} else{pred2=predB}
      #Calcolo dei pesi per ogni iter-esima iterazione
            pesi<-pesil(predl,pred2)
      #Calcolo i modelli sui valori pesati:
           A<-glm(y~eta+sessol+ripl+fumo+consumo_modl+eta:sesso,
            weights=pesi, family=binomial)
            predA<-predict(A,dati_nuovil)
           B<-glm(r ~y+ eta+sessol+ripl+fumo+consumo_modl+eta:sesso,
            weights=pesi, family = binomial)
            predB<-predict(B,dati_nuovil)
      #Selezione dei coefficienti di regressione:
            beta=A$coefficients
            alpha=B$coefficients
      #Calcolo la log-verosimiglianza di tutte le n osservazioni:
            loglike=logLik(A)+logLik(B)
            mem[iter]=loglike
            diflike=ifelse(iter>l,mem[iter]-mem[iter-l],l)
      #Specificazione criteri di convergenza (tol=0.1):
            if(diflike<tol) break
            if(diflike<0) stop("Verosimiglianza non aumenta")
            cat("         ", iter, "              ", loglike,"     ",beta,"\n")}
      list(betal, alphal, iter)}
```

Conclusioni

L'obiettivo che la presente tesi si prefiggeva era di studiare il fenomeno del *binge drinking* e dei suoi fattori determinanti in presenza di alcune osservazioni mancanti.

È stato dato particolare risalto al problema derivante dall'incompletezza delle osservazioni, ritenendo che, se ignorato, avrebbe portato a distorsioni abbastanza forti nelle stime dei fattori determinanti il *binge drinking*. La nostra supposizione sembra aver trovato conferma: le stime dei coefficienti hanno subito variazioni di rilievo in seguito al trattamento dei dati mancanti, avvicinandosi così, inoltre, ai risultati presenti in letteratura.

I risultati ottenuti ci permettono inoltre di dare risalto al secondo obiettivo di questa tesi: evidenziare l'importanza di una riflessione relativa ai dati mancanti ogniqualvolta ci si appresti ad intraprendere uno studio quantitativo. Solitamente si definisce la statistica come la 'tecnologia' necessaria per trasformare dati ed informazioni elementari in nuove conoscenze, in ipotesi di decisione, in previsioni, e, quindi, in soluzioni di problemi concreti. É quindi necessario che i dati e le informazioni elementari ai quali si applica la tecnologia statistica, siano corretti e completi, in caso contrario, si può incorrere nel rischio di generare distorsioni nelle nuove conoscenze prodotte. È infatti sempre opportuno ricordare che, come afferma una frase particolarmente celebre tra gli analisti, *'garbage in, garbage out'*. Diviene quindi d'obbligo domandarsi, ogni volta che ci si appresta a compiere delle analisi: 'Qual è la qualità dei dati con i quali sto lavorando?', 'Posso migliorarli in qualche modo? Se sì, come? Ne "vale la pena"?'. In risposta a queste domande sono stati sviluppati i primi due capitoli di questa tesi.

Tornando allo studio del *binge drinking*, i risultati ottenuti, in seguito alla correzione delle stime avvenuta mediante l'algoritmo EM, sembrano mostrare che, tra i giovani di

età compresa tra gli 11 ed i 17 anni, le femmine sono tendenzialmente più propense a sperimentare il fenomeno rispetto ai coetanei di sesso maschile. Inoltre, il *binge drinking* sembra essere molto più diffuso nelle regioni del Nord Italia, mentre risulta pressoché assente al Sud. Infine, anche il consumo di alcolici da parte dei genitori e l'abitudine al fumo si sono rivelati fattori di rischio di una certa rilevanza.

Tali risultati, in linea con la letteratura precedente, riconfermano nuovamente l'esistenza di un fenomeno, piuttosto recente nel nostro paese, ma che sta sensibilmente cambiando alcuni aspetti dello stile di vita dei giovanissimi.

Siamo comunque consapevoli che sono molti gli aspetti, sia di ordine sociale che tecnico, che necessiterebbero di un maggiore approfondimento: dal punto di vista sociologico, avrebbe una notevole rilevanza uno studio più contestualizzato e che dia importanza a quelle che possono essere le influenze esterne a cui sono soggetti i giovani (pubblicità, film,...); dal punto di vista tecnico, non è stata presa in considerazione la significatività delle stime ottenute e questo potrebbe aver portato ad impiegare nel modello variabili non così fortemente correlate col fenomeno.

Consapevoli di questi limiti e di altre possibili mancanze, si spera di aver comunque fornito una trattazione coerente e abbastanza completa relativamente al fenomeno dei dati mancanti, da un lato, e un'introduzione insieme tecnica e qualitativa sul fenomeno del *binge drinking* dall'altro. Questo, con l'auspicio di aver inoltre creato un lavoro piacevole per i pochi che lo leggeranno.

Bibliografia

[1] Andrea Bosco, *Come si costruisce un questionario*, prima edizione, Le Bussole.

[2] Bosio C.A. (1997) 'Grazie no! : il fenomeno dei non rispondenti', *Politica e Sondaggio*, a cura di P. Ceri, Rosenberg & Sellier, Torino.

[3] Caterina Arcidiacono, Elisabetta Caianiello, 'Nuovi stili di consumo alcolico negli adolescenti italiani: allarme sociale al Nord e fenomeno invisibile al Sud', *Statistica & Società*, anno V n. 1-2-3.

[4] Domenico Piccolo, *Statistica*, seconda edizione, il Mulino.

[5] E. Scafato, S. Ghirini, L. Galluzzo, C. Gandin, S. Martire e R. Russo - L'alcol e i giovani: un'analisi dei fattori determinanti l'abuso. Centro Collaboratore WHO per la Ricerca e la Promozione della Salute su Alcol e Problematiche Alcolcorrelate - Osservatorio Nazionale Alcol. CNESPS. Istituto Superiore di Sanità (ISS), Roma

[6] Geoffrey J. McLachlan, Thriyambakam Krishnan , *The EM algorithm and extensions*, seconda edizione, Wiley-Interscience.

[7] ISTAT. L'uso e l'abuso di alcol in Italia, 20 aprile 2006

[8] ISTAT. L'uso e l'abuso di alcol in Italia, 22 aprile 2010

[9] Istituto Superiore Sanità, Osservatorio Nazionale Alcol, CNESPS, *BINGE DRINKING: un'abitudine consolidata nel tempo tra i giovani*, Roma, 29 aprile 2009.

[10] Joseph G. Ibrahim and Stuart R. Lipsitz (1996), 'Parameter Estimation from incomplete Data in Binomial Regression When the Missing Data Mechanism is Nonignorable', *Biometrics*, Iternational Biometric Society, Vol.52, No.3, pp. 1071-1078.

[11] Joseph G. Ibrahim (1990), 'Incomplete Data in Generalized Linear Models', *Journal of the American Statistical Association*, American Statistical Association, Vol.85, No. 411, pp 765-769.

[12] Luigi Fabbris, *Statistica multivariata*, prima edizione, McGraw-Hill.

[13] Roderick J.A. Little & Donald B. Rubin (2002), *Statistical analysis with missing data*, seconda edizione, Wiley Interscience.

Ringraziamenti

Giunta al termine di questo lavoro desidero ringraziare ed esprimere la mia riconoscenza nei confronti di tutte le persone che, in modi diversi, mi sono state vicine e hanno permesso e incoraggiato sia i miei studi che la realizzazione e stesura di questa tesi. I miei più sentiti ringraziamenti vanno a:

Dott.Stefano Mazzuco: per la continua disponibilità e prontezza nei charimenti e suggerimenti, per la rilettura critica di tutti i capitoli della tesi e per avermi spinto a trattare un argomento che non ritenevo alla mia portata, ma che mi ha dato grande soddisfazione.

Alla mia famiglia che mi ha sempre supportato e sopportato. Grazie a mamma e papà per la continua presenza e fiducia in me: grazie per l'appoggio nei momenti in cui la voglia di studiare era poca, le soluzioni a tanti problemi e il ridimensionamento di problemi che in realtà erano inutili, per l'interesse per i miei studi, i sacrifici di tempo ed economici che avete dovuto fare, ma soprattutto grazie per la fiducia che mi avete dato, per il fatto che eravate convinti che avrei potuto farcela, per i pasticini dopo ogni esame e le telefonate dopo ogni voto. Grazie per tutto, per le possibilità che mi avete dato nel corso di tutta mia vita e per quelle che ancora mi darete. La vostra continua presenza è veramente importante.

Grazie ai miei fratelloni, Matteo e Lisa: anche voi siete sempre stati presenti, mi avete fatto ridere (e arrabbiare), mi avete permesso di staccare la testa dai libri e mi avete insegnato che nella vita c'è molto altro di bello, grazie per non avermi mai fatto sentire la 'secchiona' esagerata di casa, per aver sopportato i miei stati d'umore peggiori.

A Diego: per essermi stato vicino nei momenti in cui veramente era difficile farlo, per aver sopportato i miei sfoghi, per avermi ripetuto i motivi importanti per continuare gli studi nei momenti in cui io non ne trovavo.

Grazie per avermi ascoltato quando parlavo di un mondo che non è il tuo e per essere riscito a capirmo quando non mi capivo più da sola.

Il tuo è stato un appoggio importante, che mi ha permesso di arrivare a questa prima tappa dei miei studi con maggior serenità (pensa se non c'eri come sarei presa!).

A Silvia: non so quando mi sarei laureata senza il tuo aiuto! Grazie per le ripetizioni, le spiegazioni continue, le soluzioni agli esercizi, la rilettura e la correzione di ogni pagina che ho scritto. Grazie per l'ospitalità, per le ore di chiacchere, per gli sfoghi e per i silenzi, per i pranzi e le cene (salmone incluso), per le ore di telefonate negli orari più strani. Ma al di là di queste cose pratiche (che sono però importanti), ti voglio ringraziare per la profonda amicizia che è nata, per il fatto che ti sforzi di insegnarmi che è importante anche il 'colore del grano' (anche se ti ho insegnato io la parte più rurale!), che si può essere un'amica importante anche se si vive a km di distanza. Grazie per aver sopportato il mio italiano e aver imparato il dialetto, per i consigli preziosi ma soprattutto per essere sempre disponibile ad ascoltarmi, anche quando magari avresti molto altro da fare! Sono veramente felice dell'amicizia che è nata.

A Licia, Ilary e Carlotta: per i pranzi in compagnia, le feste e i numerosi dolci! Per aver reso la mia vita a Padova più allegra e piacevole. Grazie a Licia per le ore passate insieme per i progetti e per le battaglie nei confronti di Silvia (e difficile farle cambiare idea), grazie per il tuo ottimismo e la tua carica di energia e voglia di nuovo. Grazia a Ilary per i consigli e suggerimenti, per le chiaccherate su quello che succede in facoltà. Grazie a Carlotta per l'energia che riesce a trasmettere, per la sua solarità e compagnia.

Nadia, Roberto e Manu: non scrivo qualcosa per ogn'uno di voi altrimenti avrei bisogno di troppo spazio. Grazie per i momenti di svago, per le pizze in compagnia, per

avermi fatto divertire e per il vostro impegno nel viziarmi!
Un grazie particolare a Nadia, per tutti gli anni di amicizia e sostegno.

Grazie anche a tutti i miei colleghi in piscina, per le sostituzioni e la comprensione che mi hanno dimostrato, soprattutto in questo ultimo periodo. Grazie a tutti i bambini dei miei corsi, per avermi migliorato tante giornate che dopo ore di studio non erano state piacevoli.
Grazie alla mia famiglia allargata (Nonni, Zii e cugini), tanti dei vostri insegnamenti mi hanno aiutato ad arrivare fino a qui.
Grazie a tutti coloro che mi hanno aiutata, che mi sono stati vicini e mi hanno fatto apprezzare il percorso che ho fatto.

Daniela