Università degli Studi di Milano – Bicocca

Anno accademico 2009-2010

Dottorato di Ricerca in Statistica

XXII Ciclo

Tesi di Dottorato

Modelli Marginali Strutturali per lo studio dell'effetto causale di fattori di rischio in presenza di confondenti tempo dipendenti

Coordinatore: Prof. Giorgio Vittadini

Relatore: Prof. Giovanni Corrao

Candidata: Silvana A. Romio

Indice

Ringraziamenti

Il mio profondo ringraziamento alla Prof.ssa Rosaria Galanti della Division of Public Health Epidemiology, Department of Public Health Sciences, Karolinska Institutet che ha fornito i dati ed il supporto medico per l'analisi della corte BROMS.
Un particolare ringraziamento al Prof. Giovanni Corrao per la sua continua guida nella mia formazione.
Ringrazio anche il Prof. Rino Bellocco per i suoi suggerimenti e consigli.
Infine un particolare ringraziamento alla Dott.ssa Antonella Zambon, per avermi supportato in questi anni di dottorato e per la sua amicizia.

'se vuoi vedere le valli, sali in vetta a una motagna; se vuoi vedere la vetta di una montagna, sali su una nuvola; se invece aspiri a comprendere la nuvola chiudi gli occhi e pensa'
Kahlil Gibran

Introduzione

Uno degli obiettivi più importanti della ricerca epidemiologica è quello di analizzare la relazione tra uno o più fattori di rischio ed un evento sia esso malattia o morte. Negli ultimi anni, molti ricercatori hanno posto la loro attenzione sul significato della parola *'relazione'* al dilà delle questioni squisitamente matematiche e/o statistiche.
Già nel secolo $XVIII$ il filosofo David Hume aveva sollevato il problema dell'impossibilità di dimostrare un'ipotesi nell'ambito delle scienze empiriche. In Epidemiologia questo problema è particolarmente sentito anche per l'idea abbastanza diffusa in ambito scientifico che le dimostrazioni possono solo scaturire dagli esperimenti. Ma questo non sempre è vero, come spiegato da Rothman e Greenland [1]. Inoltre, negli studi epidemiologici si devono innanzitutto considerare tutti le possibili fonti di distorsione che possono inficiare l'associazione tra esposizione e outcome. Spesso, dunque, negli studi epidemiologici, l'obiettivo che può essere raggiunto è quello di testare relazioni 'non causali' tra l'esposizione e outcome, mentre l'ipotesi di causalità richiede assunzioni e dati che spesso non sono verificabili o disponibili.
Negli studi epidemiologici, spesso si presenta il problema della distorsione delle stime dei parametri d'interesse. Classicamente le due tipologie di errore sistematico che vengono considerate sono il confondimento e la distorsione da selezione. Anche se apparentemente il concetto di confondente può sembrare semplice, in realtà è estremamente complesso da formalizzare. Ciò è dimostrato dalla vasta letteratura riguardante tale concetto. Possiamo citare [2], [3], [4] tra molti altri. La definizione epidemiologica di confondente (ovvero un 'potenziale confondente C') è la seguente: una variabile che è legata sia all'esposizione X che all'outcome Y ma che non compare nel percorso causale della relazione esposizione-outcome (figura 1).

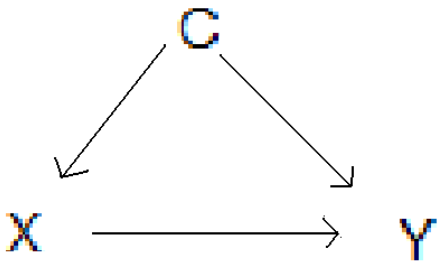


Figura 1: confondente

Pur essendo il confondimento molto noto e studiato in quasi tutti gli studi epidemiologici, spesso non è ben definito oppure risulta non sempre molto chiaro. In effetti, la definizione di confondente viene data in base al noto 'criterio associativo' [2]:

Criterio 1. *X e Y sono non confusi se per ogni variabile Z sulla quale non agisce X, Z risulta non associata a X oppure non associata all'outcome Y entro gli strati di X.*

Molte volte in effetti, viene chiamato 'confondimento' qualcosa che non lo è. Come viene chiaramente spiegato in [3] esistono essenzialmente tre interpretazioni del concetto di confondimento: la prima, considera il confondimento come una distorsione nella stima degli effetti causali, la seconda come l'equivalente della non 'collassabilità' (collapsability) mentre la terza ritiene che esiste confondimento quando non è possibile distinguere tra effetti principali e interazioni. La seconda interpretazione che è quella più utilizzata in ambito epidemiologico, può essere fuorviante come è stato mostrato attraverso numerosi esempi [3], [4]. Inoltre, la collassabilità dipende dalla misura di associazione utilizzata (cioè, può verificarsi la collassabilità utilizzando la differenza tra rischi ma non se si usa il rischio relativo oppure l'odds ratio).
Citando sempre Pearl [5], il concetto di confondimento è di natura causale e

di conseguenza tutti i metodi che si basano solo su tecniche di tipo statistico non riescono a controllare la distorsione provocata dal confondimento: 'non è possibile trasformare conoscenza statistica in conoscenza causale' spiegando questo fatto da un punto di vista probabilistico, dicendo che la funzione di distribuzione non può fornire alcuna informazione sui cambiamenti che su di essa possono verificarsi se le condizioni esterne subiscono variazioni [5]. Inoltre Pearl ha sottolineato l'importanza della differenza tra i due concetti, non intesa come contrapposizione ma per migliorare e precisare l'interpretazione delle misure utilizzate nelle analisi.
Dal punto di vista dell'analisi causale, si dice che esiste confondimento quando la misura di associazione non coincide con quella di effetto corrispondente, cioè quando ad esempio il rischio relativo non coincide con il rischio relativo causale.
Si pone quindi il problema di evidenziare quali siano i disegni e quali siano le ipotesi sulla base delle quali è possibile calcolare l'effetto causale oggetto di studio. Gli studi clinici controllati randomizzati sono nati con lo scopo di minimizzare l'influenza di errori sistematici nella misurazione dell'effetto di un fattore di rischio su di un outcome. Inoltre in questi studi le misure di associazione risultano essere uguali a quelle di effetto (cioè causali) come verrà mostrato nella sezione 1.2.
Negli studi osservazionali invece, la quantificazione dell'effetto causale risulta più complesso. In effetti, in questi studi spesso si presenta il problema dell'esistenza di una o più variabili che possono alterare o 'confondere' la relazione d'interesse in quanto lo sperimentatore non può in alcun modo intervenire sulle covariate osservate né sull'outcome. Si pone a questo punto il problema di identificare dei metodi che permettano di risolvere il problema del confondimento, alcuni dei quali sono discussi nella sezione 3.1. Il problema della distorsione da selezione viene trattato con più dettaglio nella sezione 2.2.
Lo studio delle differenze tra i concetti di causalità e associazione ha aperto la porta a un campo di ricerca trasversale che comprende filosofia, psicologia, matematica e statistica per citare solo alcune discipline. L'importanza della distinzione tra effetto causale ed associazione può essere essenziale al momento di prendere decisioni in materia di politiche sanitarie da seguire [6] in quanto queste decisioni spesso si basano sui risultati degli studi epidemiologici. Tuttavia, i modelli causali non sono ancora applicati di frequente

e solo negli ultimi anni vengono utilizzati in ambito epidemiologico [6], [7]. Ad oggi, gli studi che utilizzano questo tipo di modelli in ambito farmacoepidemiologico sono rari [6], [8]
L'obiettivo del presente lavoro è quello di studiare l'effetto causale di un fattore di rischio in presenza di confondenti tempo dipendenti e cioè una variabile che, condizionatamente alla storia di esposizione pregressa è un predittore dell'outcome e anche dell'esposizione successiva, applicando i metodi dell'inferenza causale a uno studio epidemiologico condotto per studiare un importante problema di sanità pubblica, ossia valutare se l'abitudine al fumo può essere considerato responsabile nella diminuzione dell'indice di massa corporea (body mass index - BMI) considerando come confondente tempo dipendente lo stesso BMI misurato al tempo precedente, utilizzando un modello marginale strutturale per misure ripetute avendo a disposizione i dati relativi ad una coorte di studenti svedesi.

Capitolo 1

Associazione e causalità

1.1 Concetti

Durante le ultime decadi dello scorso secolo diversi studi hanno messo in evidenza la necessità di rivedere i concetti di base dell'epidemiologia. Ad esempio, alcune indagini condotte per analizzare l'eventuale associazione tra assunzione di estrogeni ('pillola') e carcinoma dell'endometrio hanno prodotto risultati contraddittori [9]. Da qui la necessità di definire in modo più appropriato alcuni concetti quali le relazioni di associazione e causalità. Spesso questi due concetti vengono confusi anche se sono profondamente diversi. La più chiara ed intuitiva differenza tra associazione e causalità si basa sul tipo di relazione che le caratterizza: nel caso dell'associazione si tratta di una relazione senza una necessaria direzione specifica (*'undirected'*) e per tanto simmetrica, mentre la seconda è caratterizzata da una direzione specifica e quindi non simmetrica [9].
La necessità di formalizzare il concetto di causalità non è recente: da un punto di vista filosofico Hume tenta di fare ciò dicendo:

'We may define a cause to be an object, followed by another,...where, if the first object had not been, the second had never existed'

'possiamo definire una causa come un oggetto al quale segue un altro...dove se il primo non fosse stato, il secondo non sarebbe mai esistito' (Hume 1748, p.115).
Nella definizione precedente è implicito il concetto di *'controffattuale'* e cioè

cosa sarebbe accaduto (ovvero quale sarebbe stato l'effetto) se l'antecedente o causa fosse stato *'contrario ai fatti'*. Questa teoria che era nata nell'ambito della filosofia, viene presa in considerazione negli anni 90 da Rubin, Greenland, Robins e Pearl (per citare solo alcuni) per lo sviluppo di nuovi modelli che permettono di calcolare l'effetto causale di un fattore su di un outcome. In un primo approccio la teoria del contrafattuale può essere associata ad alcune interpretazioni in ambito frequentista o nei rischi competitivi [10].
Da un punto di vista squisitamente epidemiologico possiamo fare riferimento alle misure più familiari in questo ambito quali la differenza tra tassi, la differenza tra rischi (RD), il rapporto tra tassi e il rapporto tra rischi (RR). Le misure *di effetto*, ovvero quelle utilizzate per misurare l'eventuale relazione di causalità, confrontano cosa *accadrebbe in una* popolazione sotto due possibili ma distinte condizioni, delle quali al più una sola può verficarsi mentre le misure di *associazione* confrontano cosa *accade in due diverse* popolazioni (eventualmente la stessa popolazione considerata in due momenti diversi). Poichè in questo ultimo caso le due popolazioni possono essere osservate, è possibile misurare direttamente l'associazione mentre l'eventuale relazione di causalità no. Inoltre, quando la misura di associazione differisce dalla misura di effetto diciamo che esiste confondimento nella relazione.
Oltre alle differenze sopra riportate, è importante sottolineare il fatto che ogni volta che viene osservata una associazione tra due variabili si ricerca naturalmente l'eventuale relazione di causalità sottostante. Le strutture causali che possono spiegare o almeno contribuire alla relazione di associazione osservata tra due varibili d'interesse A e Y sono le seguenti:

- A può causare Y
- Y può causare A
- A e Y possono condividere una causa comune (confondente)
- *collider bias* (e.g. Berksonian bias) (cfr. figura 1.1 [11])

Una ulteriore precisazione merita l'ultima tipologia dell'elenco precedente. Se esiste una variabile Z sulla quale agiscono le variabili A e Y oppure che condivide delle cause comuni con queste variabili e in base alla quale viene fatta la selezione della polazione bersaglio, il condizionamento sulla variabile Z produrrà un collider bias. Nella figura 1.1 la variabile Z rappresenta il

collider. Un tipico esempio è quello degli studi caso controllo su base ospedaliera: ad esempio se si considera come esposizione d'interesse il tumore alla mammella e come outcome il meningioma allora i soggetti con entrambe le patologie verranno ospedalizzate per una o l'altra o per tutte e due. Questo podurrà una proporzione di soggetti ospedalizzati per entrambe le patologie superiore a quella corrispondenti ai soggetti ospedalizzati solo per meningioma [12].

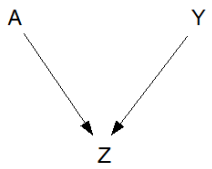


Figura 1.1: Berksonian bias

Queste strutture vengono analizzate qualitativamente attraverso i diagrammi causali che rappresentano le associazioni dedotte dalle osservazioni. Reciprocamente dalle relazioni causali tra le variabili possono essere dedotte le associazioni tra esse. Inoltre i diagrammi causali sono un ottimo strumento per individuare eventuali confondenti e altri tipologie di distorsione. Esiste una vasta letteratura su questo argomento [13].

1.2 Effetto causale in Epidemiologia

Supponiamo per fissare le idee di considerare una esposizione A dicotoma (ad esempio, l'abitudine a fumare) e un esito Y anch'esso dicotomo (ad esempio insorgenza di evento cardiovascolare). Formalmente, da un punto

di vista probabilistico, l'associazione tra A e Y può essere definita da

$$P[Y = 1|A = 1] \neq P[Y = 1|A = 0] \tag{1.1}$$

quindi la probabilità di sperimentare un evento cardiovascolare dipende dal fatto di essere fumatore.
Le misure a noi note quali differenza tra rischi (RD), rapporto tra rischi (RR) e odds ratio (OR) essenzialmente confrontano le due probabilità condizionate definite sopra e sono considerate misure di associazione e non causali.
Consideriamo adesso le due variabili aleatorie $Y_{a=1}$ e $Y_{a=0}$ che rappresentano gli esiti che sarebbero stati osservati se l'esposizione prendesse valori $a = 1$ e $a = 0$ rispettivamente sull'intera popolazione. Le variabili $Y_{a=1}$ e $Y_{a=0}$ vengono chiamate esiti potenziali (potential outcomes) ovvero esiti controfattuali in quanto (nel nostro esempio dicotomico) solo uno dei due è osservato per ogni soggetto (viene osservato cioè quello fattuale). Diremo che l'esposizione ha un effetto causale sull'outcome se e solo se per ogni soggetto

$$Y_{a=1} \neq Y_{a=0} \tag{1.2}$$

L'ipotesi nulla $Y_{a=1} = Y_{a=0}$ viene chiamata *sharp causal null hypothesis* (ipotesi causale nulla forte). Poichè l'esito non fattuali non è osservabile, cioè l'esito associato al trattamento non somministrato effettivamente (al livello di esposizione non realmente ricevuto) è un valore mancante, non è possibile individuare l'effetto causale sul singolo individuo. Per questo motivo è necessario definire l'effetto causale a livello di popolazione [14]:

Definizione 1. *Diciamo che l'esposizione A ha un effetto causale medio sull'esito Y nella popolazione se e solo se*

$$E[Y_a] \neq E[Y_{a'}] \tag{1.3}$$

per ogni coppia a, a' *con* $a \neq a'$.

Nell'esempio dicotomico sopracitato, la condizione della definizione diventa:

$$P[Y_{a=1} = 1] \neq P[Y_{a=0} = 1] \tag{1.4}$$

dove $P[Y_a = 1]$ rappresenta la proporzione di soggetti che avrebbero sviluppato l'esito (nell'esempio, diventano sovrappeso) se tutti i soggetti della popolazione avessero avuto un livello di esposizione a (ad esempio, se tutti i sogetti avessero fumato).
Nel caso generale l'ipotesi nulla causale sarà allora

$$H_0 : E[Y_a] = E[Y_{a'}] \, \forall a, a' \tag{1.5}$$

mentre nel caso dicotomico sarà

$$H_0 : P[Y_{a=1} = 1] = P[Y_{a=0} = 1] \tag{1.6}$$

Adesso è possibile definire le misure di effetto (ovvero causali) corrispondenti al RD, RR e OR in modo analogo a quelle di associazione. La differenza essenziale tra misure di associazione e misure di effetto è che le prime sono definite utilizzando le probabilità condizionali 1.1 (cioè su una parte dell'intera popolazione) mentre le seconde utilizzano probabilità marginali (cioè sull'intera popolazione). Ma il problema è sempre lo stesso: possiamo solo osservare gli esiti dei trattamenti (esposizioni) effettivamente osservati, non di quelli controfattuali!
Nel caso degli studi randomizzati tuttavia è possibile effettuare una stima consistente di tali misure. Infatti, negli studi randomizzati, il fatto che l'assegnazione dei trattamenti (esposizione) avvenga in modo casuale implica la *scambiabilità* dei soggetti rispetto ai gruppi: in effetti, ipotizzando per fissare le idee che l'esposizione è dicotoma (livelli zero e uno) e che l'assegnazione viene fatta in base al lancio di una moneta, è indifferente se il livello uno viene associato all'esito testa o croce, cioè l'assegnazione di ogni soggetto ad ogni singolo gruppo è casuale. Ciò implica

$$P[Y_a = 1|A = 1] = P[Y_a = 1|A = 0] = P[Y_a = 1] \tag{1.7}$$

ovvero $Y_a \coprod A \, \forall a$
Inoltre vale l'ipotesi di consistenza

$$Y_a = Y \tag{1.8}$$

se $A = a$ è il valore dell'esposizione effettivamente ricevuto dal soggetto.
Di conseguenza

$$P[Y_a = 1|A = a] = P[Y = 1|A = a] \tag{1.9}$$

Dalle (1.7) e (1.9) si evince facilmente che

$$\begin{aligned} P[Y_a = 1] = P[Y_a = 1|A = a] = P[Y = 1|A = a] \\ \Rightarrow P[Y_a = 1] = P[Y = 1|A = a] \end{aligned} \tag{1.10}$$

Ciò vuol dire che, sotto ipotesi di scambiabilità (assicurata negli esperimenti randomizzati), l'effetto causale coincide con l'associazione. Ovviamente tutto quello che è stato detto finora si basa sull'ipotesi che l'esperimento randomizzato non presenti perdite al follow-up, non-compliance o incertezza della cecità dello studio. Le perdite al follow-up implicano la non validità dell'ipotesi di scambiabilità in quanto i soggetti potrebbero abbandonare lo studio proprio in relazione all'esposizione a loro assegnata influenzando l'outcome osservato. La non compliance potrebbe comportare di analizzare i dati osservati secondo un approccio di intention to treat analysis o per protocol analysis. Nel primo caso si mantiene la scambiabilità ma si introduce una misclassificazione mentre nel secondo caso si perde l'assunto di scambiabilità ma non si introduce misclassificazione del trattamento. Di solito si preferisce la prima tipologia di analisi in quanto fornisce una misura di associazione non distorta se vale la ipotesi sharp causal per la vera esposizione (per protocol analysis). La mancanza di cecità implica che, pur essendo soddisfatta l'equazione (1.10), l'effetto causale è 'contaminato' dalle eventuali conseguenze dovute alla conoscenza del trattamento somministrato [14].
Per concludere osserviamo che molte volte non si hanno le informazioni sulla intera popolazione ma solo su un campione, quindi è auspicabile uno stimatore che goda delle tipiche proprietà statistiche. Per essere più specifici, essendo osservati solo alcuni soggetti (oppure se la randomizzazione vale solo in un sottoinsieme ma in modo tale che la differenza tra il gruppo degli esposti e dei non esposti diminuisca all'aumentare dell'ampiezza dei campioni considerati) possiamo affermare che $\hat{Pr}[Y = 1|A = a]$ è uno stimatore consistente di $\hat{Pr}[Y_a = 1]$ che a sua volta è uno stimatore consistente di $Pr[Y_a = 1]$.
Un'ultima ipotesi assunta tacitamente è degna di menzione: si assume la mancanza di interazione tra i soggetti nella definizione dell'effetto causale individuale (sharp null hypothesis)nel senso che l'esito relativo ad un soggetto non deve dipendere dal trattamento ricevuto da un altro soggeto. Questa ipotesi viene chiamata SUTVA (Stable Unit Treatment Value Assumption)

[15]. Ciò implica che l'effetto causale individuale non può essere calcolato in situazioni quali malattie contagiose o programmi educativi. In effetti, nel caso di un programma educativo per la prevenzione del fumo negli allievi delle scuole dove il trattamento è rappresentato dal seguire il programma (SI/NO) e l'esito è rappresentato dalla variabile dicotoma indicatrice (fumo SI/fumo NO), l'esito di un soggetto può essere influenzato dal trattameno di un altro soggetto in quanto è chiaro che l'esito di un soggetto non dipenderà soltanto dal proprio trattamento ma anche dal trattamento degli 'amici-compagni di classe' che possono influire sul suo outcome (interazione tra *amici*).Un altro esempio tipico in cui l'ipotesi SUTVA non viene verificata è negli studi riguardanti malattie contagiose: in effetti si supponga di voler studiare l'efficacia di un vaccino in una comunità di soggetti che interagiscono e si consideri come outcome lo stato di malattia dei soggetti (malato/non malato). Lo stato di malattia di un soggetto non dipenderà soltanto del fatto di essere vaccinato oppure no in quanto la sua probabilità di ammalarsi dipenderà anche dal fatto che gli altri soggetti della comunità a cui appartiene siano stati vaccinati oppure no.
Tuttavia, come fa notare Rubin [15] l'ipotesi SUTVA (come quella della non esistenza di confondenti non misurati) no può essere testata dai dati. É necessario quindi rivolgere particolere attenzione al disegno di analisi al fine di rendere tale ipotesi plausibile. Questa ipotesi viene discussa nell'analisi condotta nel presente lavoro.

Capitolo 2

Analisi qualitativa della causalità

2.1 DAGs

In questo paragrafo vengono date le nozioni di base dei grafici diretti aciclici (Direct Acyclic Graphs DAGs) [16] nonchè alcuni esempi del loro utilizzo negli studi epidemiologici. Vengono altresì forniti i criteri utilizzati per la determinazione dei confondenti.

2.1.1 Concetti di base

Nello studio delle relazioni causali si possono presentare diversi tipi di configurazioni che a volte possono essere molto complesse. É di enorme importanza quindi poter dare una rappresentazione grafica qualitativa di tali relazioni indipendentemente dalla valutazione quantitativa della loro intensità. Inoltre questi grafici devono contenere tutta l'informazione (cioè la conoscenza) posseduta da coloro che analizzano il problema. Si tratta quindi di uno strumento con un alto grado di soggettività. Ad esempio, la figura 2.1 rappresenta le relazioni tra diversi fattori di rischio (fumo(F), stress (ST), pressione sanguinea (PS)) ed insorgenza di evento cardiovascolare (ECV).

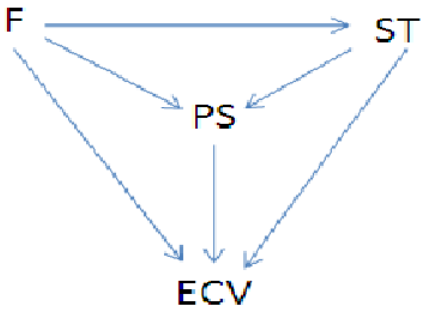


Figura 2.1: DAG

I DAG vengono utilizzati nello studio delle relazioni causali. Sono chiamati così perchè:

- le variabili presenti nel diagramma (nodi) possono essere collegate tra loro solo attraverso frecce (anzichè semplici linee di collegamento) (da cui il termine *diretti*);
- partendo da un nodo non è possibile ritornare sullo stesso qualunque sia il percorso rispettando l'orientamento delle frecce (da cui il termine *aciclici*).

Infine occorre osservare che i DAG non sono mere rappresentazioni grafiche delle relazioni causali tra le variabili considerate: le relazioni rappresentate nei DAG possono essere tradotte in termini di controfattuali attraverso equazioni strutturali non parametriche. Inoltre dai DAG è possibile capire quali sono le variabili che agiscono come confondenti e quali no, in base ad una specifica *teoria* dei DAG sviluppata da Judea Pearl [16].
É importante sottolineare che nei DAGs bisogna rappresentare tutte le possibili relazioni e che l'assenza di una freccia tra due variabili implica che si è *certi* dell'assenza di relazione tra le stesse.
Due dei concetti fondamentali della teoria dei DAG sono quello della *d-separazione* e del *back-door path* che verranno sviluppati nelle sezioni successive.

2.1.2 D-separazione

Prima di definire formalmente cosa si intende per d-separazione è necessario precisare la terminologia che verrà utilizzata. Le variabili vengono chiamate *nodi* del DAG. Un *cammino* tra due nodi è una sequenza di frecce che collegano i due nodi, indipendentemente dalla direzione delle frecce: ad esempio, in riferimento alla figura 2.1 un cammino è ECV-F-PS. Si dice che un nodo è un *collider* se esistono almeno due frecce che puntano verso di esso. Ad esempio, facendo sempre riferimento alla figura 2.1 PS è un collider nel cammino F-PS-ST. Se invece da un nodo partono due frecce (verso due nodi diversi) si dice che si ha un *fork*. Ad esempio nella figura 2.1 PS $\leftarrow$ F $\rightarrow$ ST costituisce un fork.
Se tra due nodi c'è una sola freccia, il nodo dal quale parte la freccia viene detto *genitore* mentre quello di arrivo viene detto *figlio*. Se partendo da un nodo A si arriva ad un nodo B seguendo il verso di una successione di frecce allora il nodo A viene detto ancestro mentre B viene detto *discendente*. Definiamo adesso il concetto di d-separazione:

Definizione 2. *Si dice che un cammino è d-separato (oppure bloccato o reso inattivo) da un insieme Λ di nodi se e solo se*

- *se il cammino è costituito da frecce successive, i nodi intermedi del cammino appartengono all'insime Λ*
- *se nel cammino c'è un fork, il nodo dal quale partono le due frecce appartiene all'insime Λ*
- *se nel cammino c'è un collider, né questo né un suo qualsiasi discendente deve appartenere all'insieme Λ*

Molte volte si usa la seguente notazione per rappresentare due insiemi X, Y di nodi d-separati da un terzo insieme di nodi Z, tutti facenti parte di un DAG G:

$$\left(X \coprod Y | Z\right)_G \tag{2.1}$$

Il concetto di d-separazione non è solo un concetto riguardante i DAGs ma è strettamente legato a quello di indipendenza condizionale. In effetti ogni DAG può essere visto come rappresentazione di un insieme di leggi di

probabilità congiunte dei nodi presenti nel DAG (si suppone che i vettori aleatori sono discreti, quindi assolutamente continui rispetto ad una misura di conteggio). Precisando:

Definizione 3. *Dato un DAG con nodi $V = \{V_1, \ldots, V_n\}$ e legge di probabilità congiunta Pr dei nodi, diremo che il DAG rappresenta Pr se e solo se*

$$Pr(V) = \prod_{i=1}^{n} Pr(V_i|PV_i) \tag{2.2}$$

dove PV_i rappresenta l'insieme dei genitori di V_i
La famiglia formata dal DAG e dall'insieme di leggi di probabilità che rappresenta viene chiamato Network Bayesiano.

Dalla definizione scaturisce che possono esistere diverse leggi di probabilità Pr associate allo stesso DAG. Questo fatto è il punto di forza dei DAG in quanto un singolo DAG rappresenta una classe di funzioni di probabilità. Uno dei risultati più importanti, che fu dimostrato da Verma e Pearl da una parte e nello stesso anno (1988) da Geiger è il seguente:

Teorema 1. *Dati tre insiemi disgiunti, la d-separazione in un DAG rispetto ad uno di essi equivale all'indipendenza condizionale rispetto allo stesso insieme per ogni legge rappresentata dal DAG.*

In formule, l'enunciato precedente può essere scrtto così:

$$\left(X \coprod Y|Z\right)_G \Leftrightarrow \left(X \coprod Y|Z\right)_{Pr} \tag{2.3}$$

per ogni Pr rappresentata da G.
É chiaro che se due insiemi disgiunti di nodi non sono d-separati da un altro insieme di nodi allora esiste almeno una distribuzione Pr rappresentata da G per la quale non si soddisfa la condizione di indipendenza condizionale. Ovviamente questa non indipendenza non necessariamente vale per tutte le distribuzioni rappresentate da G.
In virtù del teorema precedente, una volta individuati gli insiemi d-separati si hanno tutte le relazioni di indipendenza condizionale per tutte le leggi di probabilità rapresentate dal DAG. Siccome è molto semplice verificare la d-separazione utilizzando la definizione, risulta semplice la ricerca degli insiemi condizionatamente indipendenti.

Al fine di fornire una spiegazione euristica e per chiarire i concetti e risultati precedenti, si introducono alcuni esempi.
Esempio 1: nel DAG della figura 2.2

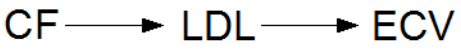


Figura 2.2: variabile intermedia

dove CF indica il consumo di formaggi (si/no), LDL avere il livello di trigliceridi alto (si/no) e ECV una malattia cardiovascolare (si/no) allora, condizionando sul nodo LDL (ad esempio, considerando i soggetti con alto livello di trigliceridi) farà si che CF e ECV siano indipendenti perchè il consumo di fomaggio agisce sul rischio di malattia cardiovascolare solo attraverso il livello di trigliceridi.

Esempio 2: nel DAG della figura 2.3

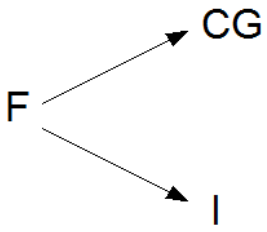


Figura 2.3: fork

dove F indica l'abitudine al fumo di sigaretta, CG il fatto di mangiare chewing-gum (si/no) e I soffrire d'insonnia, condizionando sul nodo F (ad esempio considerando i fumatori) il fatto di mangiare chewing-gum è indipendente dal fatto di soffrire d'insonnia (mentre prima del condizionamento, cioè nell'intera popolazione, le due variabili non risultano indipendenti).

Esempio 3: nel DAG della figura 2.4 dove TL indica il tipo di lavoro (manageriale/dipendente tra cui quello manageriale crea un elevato livello di stress), IA il fatto di vivere in una zona industriale ad alto inquinamento acustico (si/no) e ST il livello di stress. Ipotizzando che le uniche cause dello stress siano TL e IA, se si condiziona a ST cioè si considerano solo i soggetti che hanno un alto livello di stress, la distribuzione del lavoro manageriale è diversa nell'insieme dei soggetti che abitano in una zona ad alto inquinamento acustico da quella nell'insieme dei soggetti che abitano in una zona a basso inquinamento acustico. Di conseguenza, anche se originariamente le due variabili erano indipendenti, condizionando sul collider si crea una dipendenza inesistente (cioè fittizia).

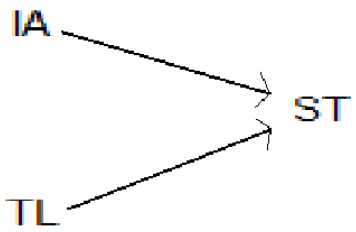


Figura 2.4: collider

Questi tre semplici esempi suggeriscono quanto possa essere pericoloso un condizionamento non ragionato delle variabili considerate nella analisi e quanto i DAG possano essere utili al momento di decidere su quali variabili condizionare. Nell'implementazione dei modelli statistici per l'analisi

delle relazioni quindi non basta aggiustare per tutte le variabili a disposizione come spesso accade nella pratica, in modo da *tutelarsi* da eventuali confondenti ma è necesario analizzare accuratamente ogni situazione.

2.1.3 Il criterio del back-door

Da quanto detto precedentemente, aggiustare per un collider può generare associazioni inesistenti nella realtà.
Un criterio grafico molto semplice che permette di individuare gli eventuali confondenti presenti in una relazione è quello del *backdoor path*. L'algoritmo può essere enunciato nel modo seguente [13]:

- eliminare ogni freccia tra l'esposizione e l'outcome d'interesse;
- controllare se nel nuovo grafo esiste un cammino non bloccato tra l'esposizione e l'outcome

Se tutti i cammini sono bloccati (cioè se tutti i cammini sono d-separati) allora non c'è confondimento. Essenzialmente, se tutti i cammini tra l'esposizione e l'outcome sono bloccati allora le due variabili risultano marginalmente indipendenti. Per chiarire il funzionamento dell'algoritmo vediamo due esempi. Nella figura 2.5

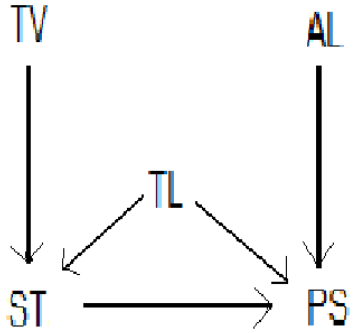


Figura 2.5: fork in un DAG

è rappresentato un DAG per lo studio della relazione tra stress (ST) e pressione sanguinea (PS) considerando anche i minuti trascorsi al giorno davanti alla televisione (TV), il consumo di alcool (AL) e il tipo di lavoro (TL). Tranne il cammino diretto tra ST e PS, l'unico cammino possibile è ST←TL→PS che risulta essere bloccato dall'insieme formato dalla variabile TL. Di conseguenza le variabili ST e PS sono indipendenti condizionatamente a TL (occorre quindi aggiustare per TL).
L'esempio della figura 2.6 rappresenta un classico esempio della letteratura [13].

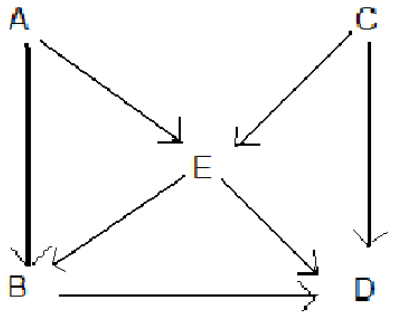


Figura 2.6: un esempio classico

In questo esempio, poichè la variabile E forma un fork nel cammino B←E→D e risulta essere un collider nel cammino B ←A→E←C→D. Di conseguenza, poichè è necessario condizionare su E (in virtù del fork) sarà neccessario condizionare anche per le variabili A e C che sono ancestri comuni di B e D cioè formano a loro volta due fork B←A→E→D e B←E←C→D.Di conseguenza l'insieme per il quale si dofrebbe condizionare è formato dai nodi A, E e C. Ma in realtà è possibile selezionare un insieme minimale di variabili per il quale condizionare. Tale insieme viene chiamato insieme *sufficiente* di variabili. Benchè formalmente non è un problema aggiustare per più variabili di quelle necessarie, molte volte accade di non avere tutte le informazioni sulle variabili coinvolte. Il fatto di poter ridurre

la quantità di variabili per le quali aggiustare risolve quindi un importante problema pratico.
Per decidere se un insieme di variabili (che non contiene discendenti né dell'esposizione né dell'outcome) è sufficiente per l'aggiustamento basta applicare il seguente algoritmo [13]:

- eliminare tutte le frecce che partono dalla variabile che rappresenta l'esposizione;
- aggiungere tutti gli archi generati dal controllo sulle variabili appartenenti all'insieme considerato
- se tutti i cammini non bloccati tra l'esposizione e l'outcome contengono almeno una variabile dell'insieme considerato allora l'insieme è sufficiente per il controllo.

Eventualmente l'insieme può essere anche vuoto (in tal caso non è necessario aggiustare per alcuna variabile).
Come esempio possiamo considerare ancora la figura 2.5. In questo caso l'insieme 'candidato' è quello formato dalla sola variabile TL. In effetti, dopo aver eliminato la freccia ST→PS e osservando che l'aggiustamento per TL non comporta l'aggiunta di alcuna freccia nel DAG, l'unico cammino che compare è ST←TL→PS che è bloccato (passa da) TL come detto sopra. Quindi l'insieme formato dal solo nodo TL è sufficiente.
Un altro esempio è quello fornito dalla figura 2.7.

In questo caso l'insieme vuoto risulta sufficiente poichè l'unico cammino 'backdoor ' risulta bloccato dall'insieme vuoto.
In fine, è possibile scegliere un insieme sufficiente minimale nel senso che nessun sottoinsieme di esso è sufficiente. Ad esempio nel caso della figura 2.6 un insieme sufficiente minimale è quello formato dai nodi A ed E. Ovviamente tale insieme minimale non necessariamente è unico. In effetti, anche l'insieme formato dai nodi C ed E risulta sufficiente minimale.

2.2 Distorsione

Nell'introduzione si è parlato del confondimento come di uno dei problemi che si presentano negli studi epidemiologici. In realtà il confondimento fa

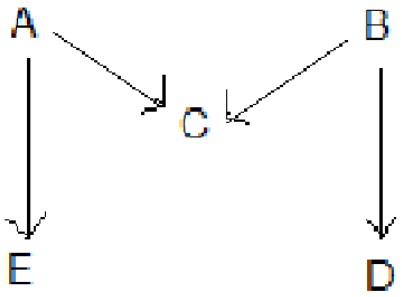


Figura 2.7: insiemi sufficienti minimali

parte della più ampia famiglia degli errori sistematici. Gli epidemiologi parlano in generale di *distorsione da selezione* in riferimento a un gran numero (a volte non ben definito) di distorsioni quali, ad esempio, l'effetto lavoratore sano, la distorsione da selezione negli studi caso-controllo, la perdita diferenziale al follow-up.
Così come in 1.1 sono state elencate le possibili strutture causali, in modo analogo possono essere rappresentate le diverse strutture di distorsione. Alcune possibili distorsione da selezione quali [17]:

- selezione non appropriata dei controlli in uno studio caso-controllo
- distorsione di Berkson
- perdita differenziale al follow up in studi longitudinali
- distorsione da dati mancanti
- distorsione da auto-selezione
- selezione da lavoratore sano

possono essere rappresentate schematicamente dalla figura 2.4. Considerando le variabili come indicato dalla figura 2.8

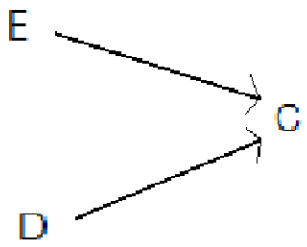


Figura 2.8: distorsione da selezione

nel caso in cui la variabile E rappresenta l'esposizione, D la malattia (outcome) e C è una variabile indicatrice (che assume il significato di selezione dalla coorte per uno studio caso-controllo)si avrà una rappresentazione della prima tipologia dell'elenco sopra riportato. Per la seconda (distorsione di Berkson) C rapresenta l'indicatore di ospedalizzazione; per la terza l'indicatore di censura, ecc.
Come si può capire chiaramente dagli esempi precedenti, applicando la definizione di d-separazione della sezione 2.1.2 è possibile decidere su quali variabili condizionare (ovvero aggiustare) al fine di controllare le eventuali distorsioni che possono presentarsi in uno studio.

Capitolo 3

Analisi quantitativa della causalità

3.1 Stima degli effetti causali in Epidemiologia

Gli studi osservazionali sono essenziali in ambito epidemiologico. In effetti, condurre studi randomizzati non sempre è possibile ad esempio per motivi di carattere etico.
Nel capitolo 1 è stato definito cosa si intende per effetto causale e cosa sono le misure di effetto. Negli studi osservazionali, e sotto opportune ipotesi, è possibile stimare tali misure. Il problema principale in questi studi è la mancanza di scambiabilità tra esposti e non esposti. In questi studi in effetti, potrebbe essere presente un eventuale fattore predittivo che può condizionare il livello dell'esposizione in studio. Riprendendo l'esempio della sezione 1.2 possiamo introdurre come fattore predittivo L la condizione di stress all'inizio dello studio. A questo punto però è possibile ipotizzare, in base alla conoscenza pregressa, una forma meno restrittiva della proprietà di scambiabilità ovvero la *scambiabilità condizionale*:

Definizione 4. *se A rappresenta l'esposizione, Y l'esito, Y_a l'esito controfattuale e L il fattore predittivo, diremo che vale l'ipotesi di scambiabilità*

condizionale se e solo se

$$\begin{aligned} P[Y_a = 1|A = 1, L = l] &= P[Y_a = 1|A = 0, L = l] \\ &= P[Y_a = 1|L = l] \end{aligned} \tag{3.1}$$

ovvero $Y_a \coprod A|L \ \forall a$.

É chiaro che la scambiabilità condizionale (indipendenza condizionale) non è equivalente a quella non condizionata a L (i.e. marginale) data da (1.7).
In termini epidemiologici il fattore predittivo L può essere considerato come un confondente (in questo caso misurato) non tempo dipendente. La possibilità di identificare l'effetto causale da uno studio osservazionale dipende dal fatto che il confondente L sia misurato. In presenza di confondenti non misurati non è possibile assumere l'indipendenza condizionale.
Sempre nell'ambito degli studi osservazionali oltre all'ipotesi di scambiabilità condizionale esistono altre due ipotesi che devono essere soddisfatte: la *consistenza* e la *condizione di positività*. La prima è già stata considerata nella sezione 1.2: $Y_a = Y$ se $A = a$ è il livello dell'esposizione effettivamente osservato. Questa ipotesi che a prima vista potrebbe sembrare banale, è sempre verificata negli studi randomizzati ma non necessariamente in quelli osservazionali. Essenzialmente, la validità di tale ipotesi riguarda la buona definizione del livello di esposizione: se tale livello non è specificato in modo dettagliato l'esito contrafattuale non sarebbe ben definito l'potesi di consistenza non sarebbe soddisfatta [18].
La condizione di positività è definita nel seguente modo:

$$P[A = a|L = l] > 0 \tag{3.2}$$

se $P[L = l] \neq 0$. Questa condizione assicura l'esistenza del rischio standar-

dizzato: in effetti tale rischio è dato da

$$\begin{aligned} R_{A=a} &= \sum_l Pr\left[Y=1|L=l,A=a\right] Pr\left[L=l\right] \\ &= \sum_l \frac{Pr\left[Y=1,L=l,A=a\right]}{Pr\left[L=l,A=a\right]} Pr\left[L=l\right] \\ &= \sum_l \frac{Pr\left[Y=1,L=l,A=a\right]}{Pr\left[A=a|L=l\right] Pr\left[L=l\right]} Pr\left[L=l\right] \\ &= \sum_l \frac{Pr\left[Y=1,L=l,A=a\right]}{Pr\left[A=a|L=l\right]} \end{aligned} \tag{3.3}$$

e per esistere il denominatore di ogni termine della somma deve essere diverso da zero.
A questo punto il problema diventa come calcolare una misura di effetto, ad esempio il RR causale. Per fare ciò esistono diversi metodi: la standardizzazione, il propensity score e l'inverse probability treatment weighting (IPTW).

3.1.1 Standardizzazione

Uno dei metodi classici per controllare il confondimento e quindi calcolare le stime corrette della variabile d'interesse è la *standardizzazione*. Essenzialmente, i diversi livelli del fattore predittivo L determinano gli strati all'interno dei quali i soggetti risultano omogenei rispetto a L. Nella *standardizzazione* i rischi vengono calcolati come media pesata dei rischi strato-specifici. Ne risulta che, sotto opportune ipotesi, il rischio relativo causale (che è una delle misure di effetto che possono essere considerate) coinciderà con il rischio relativo (i.e. la corrispondente misura di associazione). In

effetti, in presenza di scambiabilità condizionale avremo:

$$\begin{aligned} RR_{causal} &= \frac{Pr\left[Y_{a=1}=1\right]}{Pr\left[Y_{a=0}=1\right]} \\ &= \frac{\sum_l Pr\left[Y_{a=1}=1|L=l,A=1\right]Pr\left[L=l\right]}{\sum_l Pr\left[Y_{a=0}=1|L=l,A=0\right]Pr\left[L=l\right]} \\ &= \frac{\sum_l Pr\left[Y_{a=1}=1|L=l\right]Pr\left[L=l\right]}{\sum_l Pr\left[Y_{a=0}=1|L=l\right]Pr\left[L=l\right]} \\ &= \frac{\sum_l Pr\left[Y=1|L=l,A=1\right]Pr\left[L=l\right]}{\sum_l Pr\left[Y=1|L=l,A=0\right]Pr\left[L=l\right]} \end{aligned} \tag{3.4}$$

dove, in virtù di tale scambiabilità condizionale
$Pr\left[Y_a=1|L=l\right]=Pr\left[Y=1|L=l,A=a\right]$
Come si può vedere dalla (3.4) il RR_{causal} altro non è che il RR standardizzato usando la popolazione totale come riferimento.

3.1.2 Propensity Score

Abbiamo detto nell'introduzione che negli studi osservazionali, esiste confondimento quando la misura di effetto non coincide con la corrispondente misura di associazione. I metodi classici utilizzati in Epidemiologia mirano a controllare le variabili confondenti in modo tale da poter considerare la misura di asociazione come misura dell'effetto causale. Un metodo utilizzato per controllare la variabile confondente ormai noto in ambito epidemiologico è quello del *propensity score* [19]. In assenza di randomizzazione, il propensity score permette di stratificare i soggetti in insiemi in modo tale da produrre un bilanciamento per le covariate osservate [20]. Il razionale di questo metodo è il seguente: sia L l'insieme di variabili confondenti e si supponga che non ci siano confondenti non misurati. Il propensity score viene definito come la probabilità che un soggetto avente un particolare valore della covariata L (o del vettore di covariate L) sia trattato. In formule:

$$P\left(A=1|L=l\right) \tag{3.5}$$

La stratificazione mediante il propensity score fa si che i soggetti appartenenti a due categorie a confronto (trattati e non trattati) siano confrontabili all'interno di ogni strato definito dal propensity score. In altre parole, internamente ad ogni strato i soggetti trattati hanno la stessa distribuzione della

covariata L dei soggetti non trattati. Soggetti aventi lo stesso propensity score hanno lo stesso 'peso' nell'analisi stratificata.
Il concetto di propensity score nasce da quello più generale di funzione di bilanciamento. Si dice che una funzione $b(L)$ della covariata L è una funzione di bilanciamento se e solo se sodisfa la seguente condizione:

$$A \coprod L|b(L) \tag{3.6}$$

Il propensity score è una funzione di bilanciamento che gode delle seguenti proprietà [19], [20]:

- il propensity score bilancia le covariate osservate
- il propensity score può sostituire nell'aggiustamento le covariate da cui dipende, se queste sono sufficienti per il controllo del confondimento (cioè non si 'perde' nulla sostituendo le covariate considerate nell'aggiustamento con una loro sintesi appropriata)
- la stima del propensity score è migliore del suo vero valore in quanto quest'ultimo riesce a cogliere solo la distorsione sistematica proveniente dalle variabili confondenti mentre la stima del propensity score riesce a cogliere anche la componente dovuta all'errore casuale.

L'ultima proprietà è molto interessante in quanto ci si aspetterebbe di avere un comportamento migliore da parte del vero valore anzichè della sua stima [19]. Inoltre ha la proprietà di essere il meno fine. Ciò implica che il propensity score è la funzione di bilanciamento più semplice. Inoltre la stima del propensity score può essere calcolata utilizzando anche semplici modelli. Ad esempio il propensity score può essere stimato attraverso il seguente modello:

$$Pr(A_i = 1|L_i = l) = \frac{exp(l\beta)}{1 + exp(l\beta)} \tag{3.7}$$

cioè un modello logistico [21].
Fino a pochi ani fa il propensity score aveva la limitazione di non poter essere generalizzato al caso di esposizione tempo dipendente [22]. Recentemente, Bo Lu ha proposto un appaiamento per propensity score con covariate tempo dipendenti [23].

3.1.3 IPTW

Il metodo del *inverse probability treatment weighting* (IPTW) calcola il numeratore e il denominatore del RR causale costruendo una pseudo-popolazione che rappresenta gli outcome controfattuali. Essenzialmente questo metodo si basa sul ricalcolare le quantità (e pertanto le proporzioni) dei soggetti che potenzialmente sperimentano l'evento utilizzando le stesse proporzioni dell'esposizione 'fattuale' (anche per l'esposizione controfattuale). Ciò vuol dire che, considerando tutti i soggetti come non esposti (sia che il fattore prognostico sia presente oppure no) la proporzione di eventi è quella dei veri non esposti osservati.
Le numerosità della pseudo-popolazione possono essere calcolate moltiplicando le numerosità osservate per opportuni pesi (da qui il nome di inverse probability weighting). Questi pesi vengono calcolati come:

$$W = \frac{1}{f\left[A|L\right]} \tag{3.8}$$

Possiamo fare un esempio numerico molto semplice [24]. Si consideri una popolazione in cui diciotto soggetti vengono seguiti nel tempo. Di questi, dodici presentano uno stato di stress elevato e di questi otto fumano. Tra questi si osservano sei casi di evento cardiovascolare mentre tra i soggetti con elevato livello di stress non fumatori si osservano tre casi. Tra i sei soggetti il cui stato di stress non è elevato, tre fumano e di questi due soggetti sperimentano un evento cardiovascolare. Tra i restanti tre soggetti (non fumatori) si osservano due casi di evento cardiovascolare. I dati sono sintetizzati nella figura 3.1.
La figura 3.2 rappresenta la pseudo popolazione ottenuta se tutti i soggetti della popolazione originale fossero stati non fumatori (i.e. non trattati). Le numerosità dei rami corrispondenti a $Y = 0$ e $Y = 1$ rispecchiano le proporzioni originali. Analogamente si costruisce l'albero corrispondente alla pseudo popolazione ottenuta se tuti i soggetti fossero stati fumatori (i.e. trattati). Tale popolazione è rappresentata nella figura 3.3.
Unendo le pseudo popolazioni precedenti si ottiene la pseudo popolazione complessiva. Le numerosità di tale pseudo popolazione possono essere ottenute aplicando i pesi non stabilizzati definiti da (3.8). Nella figura 3.4 viene rappresentato il calcolo di tali numerosità utilizzando i pesi W.
Ad esempio, al primo insieme di soggetti aventi $A = 0$ e $L = 0$ corrisponde

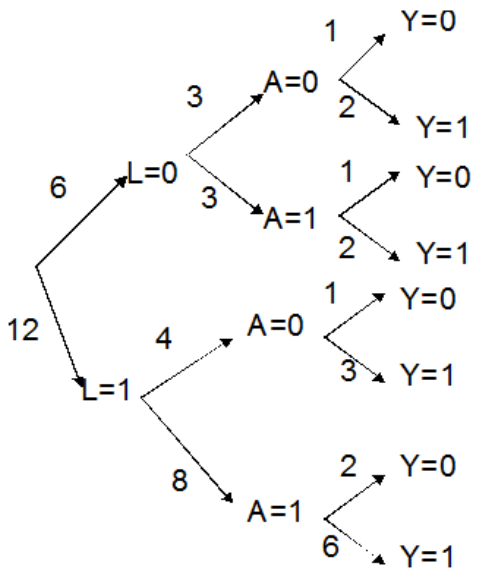


Figura 3.1: popolazione originale

un peso pari a

$$W = \frac{1}{f\left(A=0\right)|\left(L=0\right)} = \frac{1}{f\left(A=0, L=0\right)/f\left(L=0\right)} = \frac{f\left(L=0\right)}{f\left(A=0, L=0\right)} = \frac{6/18}{3/18} = 2 \tag{3.9}$$

e di conseguenza la numerosità corrispondente sarà pari a $2 * 1$ (peso calcolato*numerosità osservata). Usando i pesi W la numerosità della pseudo-popolazione aumenta (raddoppia). Se anziché considerare i pesi W consideriamo i *pesi stabilizzati* definiti da

$$SW = \frac{f\left(A\right)}{f\left(A|L\right)} \tag{3.10}$$

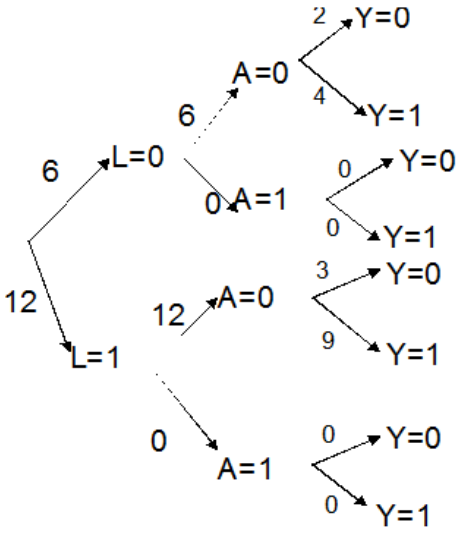


Figura 3.2: pseudo popolazione a=0

la numerosità della pseudo-popolazione rimane uguale a quella di partenza in quanto il numeratore rappresenta una distribuzione il cui integrale vale 1 (si potrebbe utilizzare quindi una qualsiasi funzione a somma o integrale 1). La figura 3.5 rappresenta i pesi stabilizzati. La scelta del numeratore nei pesi SW è indifferente ai fini della consistenza dello stimatore della misura di effetto. Cambia invece la variabilità dello stesso. É per questo motivo che in situazioni più complesse è preferibile utilizzare i pesi stabilizzati che rendono più efficiente lo stimatore.
Per poter utilizzare i due metodi sopra descritti (standardizzazione e IPTW) è necessario conoscere i dati sulla variabile confondente L e assumere che non ci sia confondimento non misurato. Inoltre i due metodi conducono alla stessa misura causale nei casi più semplici e cioè quando è possibile calcolare esplicitamente le probabilità utilizzate come pesi. Quando questo non è possibile a causa della complessità dei dati, è necessario stimare queste probabilità e quindi i risultati ottenuti dai due metodi possono differire.
Una ultima osservazione sui due metodi sopracitati. Tornando all'equazio-

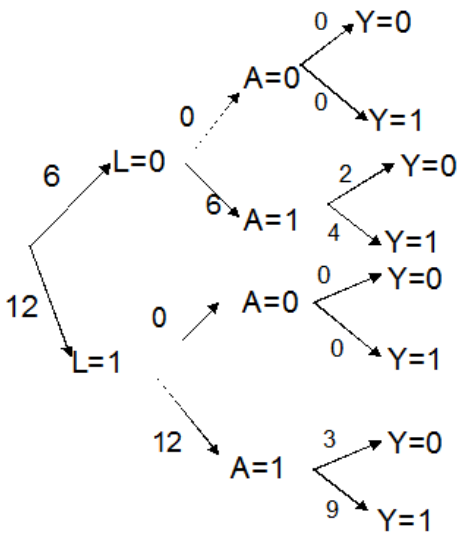


Figura 3.3: pseudo popolazione a=1

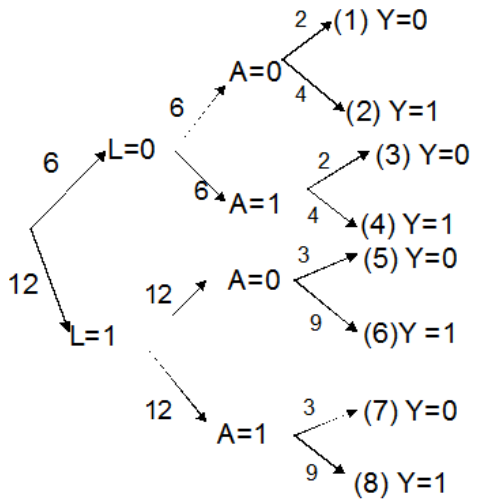
ne (3.4) e fissando le idee sul numeratore dell'ultimo rapporto, possiamo scrivere:

$$\sum_l Pr\left[Y=1|L=l,A=1\right]Pr\left[L=l\right]= \\ \sum_l Pr\left[Y=1,L=l,A=1\right]\frac{Pr\left[L=l\right]}{Pr\left[L=l,A=1\right]} \tag{3.11}$$

e l'ultima espressione può essere scritta come:

$$\sum_l Pr\left[Y=1,L=l,A=1\right]\frac{Pr\left[L=l\right]}{Pr\left[L=l\right]Pr\left[A=1|L=l\right]} \tag{3.12}$$

ovvero lo stimatore di Horvitz-Thompson. A questo punto è chiaro che l'ultima espressione corrisponde allo stimatore IPTW con pesi W (non stabilizzati).
Esiste una stretta relazione tra due dei metodi per l'aggiustamento del confondimento visti sopra, nella fattispecie la standardizzazione e l'IPTW. In



	W=1/f(A|L)
1	6/3=2
2	6/3=2
3	6/3=2
4	6/3=2
5	12/4=3
6	12/4=3
7	12/8=1.5
8	12/8=1.5

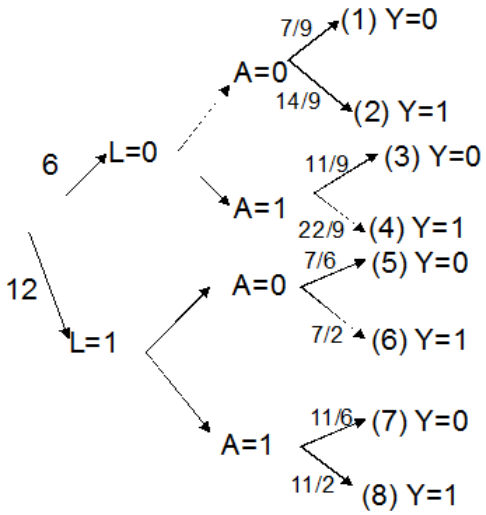
Figura 3.4: pesi W

effetti, se come nell'esempio presentato nel paragrafo 3.1.1 si utilizza la popolazone totale come popolazione obiettivo la misura d'interesse (cioè il rischio relativo standardizzato) può essere interpretato come il cambiamento proporzionale del rischio nella popolazione totale considerando tutti i soggetti esposti e tutti i soggetti non esposti ovvero considerando la pseudo-popolazione utilizzata nel metodo IPTW.
Infine, in presenza di un modificatore di effetto, poichè il metodo IPTW assegna i pesi considerando la distribuzione del modificatore di effetto nell'intera popolazione mentre il propensity score utilizza la distribuzione nella sotto-popolazione degli esposti [25]. Pertanto, in ambito farmacoepidemiologico è preferibile l'utilizzo del propensity score.

3.2 Modelli per lo studio degli effetti causali

3.2.1 Esposizione puntuale non tempo dipendente

Supponiamo per fissare le idee una esposizione A dicotoma (0/1) e una variabile dicotoma Y per l'esito oggetto di studio (outcome) considerato



	SW=f(A)/f(A|L)
1	7/18 *6/3=7/9
2	7/18 *6/3=7/9
3	11/18 *6/3=11/9
4	11/18 *6/3=11/9
5	7/18 *12/4=7/6
6	7/18 *12/4=7/6
7	11/18 *12/8=11/12
8	11/18 *12/8=11/12

Figura 3.5: pesi stabilizzati SW

alla fine del follow-up. La misura di associazione (ad esempio il rischio relativo RR)può essere calcolata dal seguente modello [22]:

$$logP\left(Y=1|A=a\right)=\theta_0+\theta_1*a \tag{3.13}$$

e di conseguenza

$$RR=\frac{exp\left(\theta_0+\theta_1\right)}{exp\left(\theta_0\right)}=exp\left(\theta_1\right) \tag{3.14}$$

cioè $exp\left(\theta_1\right)$ è la stima del RR grezzo. Il corrispondente modello marginale è:

$$logP\left(Y_a=1\right)=\alpha_0+\alpha_1*a \tag{3.15}$$

In questo caso avremo

$$RR=\frac{exp\left(\alpha_0+\alpha_1\right)}{exp\left(\alpha_0\right)}=exp\left(\alpha_1\right) \tag{3.16}$$

essendo adesso $exp\left(\alpha_1\right)$ il RR causale. I coefficienti α_1 e θ_1 coincideranno in assenza di confondimento.

Il modello (3.15) è *marginale* in quanto modella la probabilità marginale dell'esito contrafattuale ed è *saturo* in quanto la quantità di parametri incognti che compaiono a destra nell'equazione coincide con quella di sinistra (i.e. la quantità di controfattuali). In presenza di confondimento (ma ipotizzando sempre l'assenza di confondimento non misurato), la stima dei parametri del modello causale viene fatta attraverso lo stesso tipo di modello ma 'pesato': in effetti, la pseudo-popolazione Y_a viene mimata inserendo nell'analisi che utilizza il modello (3.13), i pesi IPTW calcolati come i reciproci delle probabilità che vengono ricavate attraverso il modello [22]

$$logitPr\left(A_0 = 1|L_0 = l_0\right) = \beta_0 + \beta_1 * l_0 \tag{3.17}$$

Essendo l'esposizione dicotoma, questo modello è sufficiente per stimare tutti i pesi necessari in quanto l'altra probabilità $Pr\left(A_0 = 1|L_0 = l_0\right)$ altro non è che il complemento a 1 della probabilità calcolata tramite la (3.17).
Anche nel caso di esposizione non dicotoma (esposizione ordinale oppure continua) è possibile stimare i pesi utilizzando opportuni modelli per il calcolo delle probabilità che li definiscono.
Anche in questo contesto è possibile fare un confronto tra i modelli marginali strutturali e la standardizzazione.

3.2.2 Esposizione tempo dipendente

Nella sezione precedente è stata analizzata una situazione molto semplice. In effetti, la variabile di esposizione (A) era fissa così come il confondente (L). Supponiamo che sia l'esposizione che il confondente siano tempo dipendenti ed inoltre che quest'ultimo sia influenzato dall'esposizione pregressa. Come prima supponiamo che l'outcome di interesse sia misurato alla fine del follow-up. La situazione più semplice in cui le uniche variabili che intervengono sono il confondente, l'esposizione e l'outcome, potrebbe essere rappresentata dal DAG della figura 3.6. Il confondente $L\left(t\right)$ è sottointeso che sia misurato prima dell'esposizione $L\left(t\right)$ [22]. Nel caso in cui l'esposizione sia tempo dipendente come nella figura 3.6, Robins sottolinea che la variabile di esposizione è la *storia* di esposizione del soggetto. Se la quantità di livelli dell'esposizione A è pari a n e la quantità di tempi osservati è k, esistono n^k possibili combinazioni di esposizione (di cui una osservata). Siccome ad ogni particolare *combinazione* di trattamento è associato

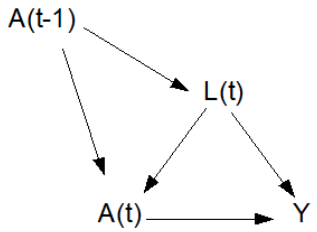


Figura 3.6: confondente tempo dipendente

un particolare outcome controfattuale, si avranno $n^k - 1$ possibili outcomes controfattuali e un outcome osservato. Questo fatto rappresenta un problema dal punto di vista dell'impostazione del modello da considerare e della stima dei parametri che misurano l'effetto causale: in effetti, come impostare un modello che dipenda dalla storia di trattamento e come calcolare le probabilità che intervengono nel calcolo dei pesi?
Per quanto riguarda il modello e considerando l'esposizione dicotoma, Robins [22] propone di considerare il seguente:

$$logP\left(Y_a = 1\right) = \beta_0 + \beta_1 * cum\left(\bar{a}\right) \tag{3.18}$$

dove $cum\left(\bar{a}\right)$ rappresenta la dose cumulata di esposizione durante tutto il follow-up. Come prima, la stima dei parametri del modello può essere ottenuta mediante il modello logistico pesato

$$logP\left(Y = 1|\bar{A} = \bar{a}\right) = b_0 + b_1 * cum\left(\bar{a}\right) \tag{3.19}$$

con i pesi calcolati come reciproci del prodotto cumulato delle probabilità $Pr\left(A_k = a_k|\bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_k = \bar{l}_k\right)$ per ogni soggetto cioè:

$$W = \frac{1}{\prod_{k=0}^{K} Pr\left(A_k = a_k|\bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_k = \bar{l}_k\right)} \tag{3.20}$$

per i pesi non stabilizzati; in modo analogo vengono calcolati i pesi stabilizzati visti in 3.1.3

$$SW = \frac{\prod_{k=0}^{K} Pr\left(A_k = a_k | \bar{A}_{k-1} = \bar{a}_{k-1}\right)}{\prod_{k=0}^{K} Pr\left(A_k = a_k | \bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_k = \bar{l}_k\right)} \tag{3.21}$$

Come si può osservare, il modello precedente coincide essenzialmente con quello della sezione 3.2.1 nel caso in cui la varibile di esposizione anzichè essere dicotoma è politomica. Chiaramente i pesi vengono calcolati in modo diverso in quanto nel caso in cui l'esposizione è tempo dipendente, i pesi devono essere calcolati considerando questa dipendenza dell'esposizione dal tempo.
Dal punto di vista applicativo, il calcolo dei pesi può creare alcuni problemi in quanto adesso sono necessarie più di due probabilità (una complementare dall'altra) per calcolare tutti i pesi?. In effetti, se $A(k) = 0$ indica l'assenza di trattamento e $A(k) = 1$ indica che il soggetto è trattato a tempo k, in generale si possono avere le situazioni della tabella 3.1.

Tabella 3.1: configurazione del trattamento a due tempi successivi

$A(k-1)$	$A(k)$
0	0
0	1
1	0
1	1

Si dovebbe quindi calcolare la probabilità di essere in trattamento per ogni istante temporale considerato e per ogni configurazione della storia di trattamento pregressa. Ma ciò non è possibile con un unico modello. Se invece si assume che una volta che il soggetto inizia il trattamento non lo sospende, allora l'unica probabilità che è necessario calcolare è quella corrispondente alla prima riga

$$logitP\left[A(k) = 0 | A(k-1) = 0, \bar{L}(k)\right] \tag{3.22}$$

in quanto quella corrispondente alla seconda riga è il suo complemento a 1, la configurazioni della terza riga della tabella non è possibile e e quella

della ultima riga è pari a 1. In questo caso i pesi possono essere calcolati utilizzando un modello loistico. Si assume quindi che l'esposizione non può essere intermittente nel calcolo dei pesi cioè la successione di zeri e uno che rappresentano la storia di esposizione di un soggetto deve essere:

$$(00\ldots011\ldots1)$$

Questo assunto, nato da una esigenza di tipo applicativo, può essere interpretato come un *'intention to continue treatment analysis'* [26]. In termini formali, Robins [27] dimostra che in assenza di mispecificazione del modello e di confondenti non misurati, il modello pesato è quello che permette di calcolare l'effetto causale in modo non distorto. Considerando sempre il modello che permette di studiare la relazione di associazione

$$E\left[Y|\bar{A}\right] = \gamma_1 + \gamma_2 cum\left(\bar{A}\right) \tag{3.23}$$

se vale l'ipotesi di *esogeneità causale* definita da

$$Y_{\bar{a}} \coprod A\left(k\right)|\bar{A}\left(k-1\right) \tag{3.24}$$

ovvero la probabilità di essere trattato al tempo k condizionatamente alla storia di trattamento pregressa e ai confondenti misurati e non misurati, dipende solo dalla storia passata di tattamento. Allora il parametro γ_2 non è interpretabile solo come parametro di associazione ma anche come parametro causale.
Una ipotesi più debole che può essere testata dai dati è quella di *esogeneità statistica*:

$$\bar{L}\left(k\right) \coprod A\left(k\right)|\bar{A}\left(k-1\right) \tag{3.25}$$

che è chiaramente implicata dall'ipotesi precedente di esogeneità causale. L'implicazione reciproca non è vera e non è possibile dai dati verificare quando vale la esogeneità causale.
L'ipotesi di esogeneità statistica equivale alla seguente

$$P\left[A\left(k\right)|\bar{A}\left(k-1\right)=0,\bar{L}\left(k\right)\right] = P\left(\left(A\left(k\right)|\bar{A}\left(k-1\right)\right)\right) \tag{3.26}$$

ovvero

$$\frac{P\left[A\left(k\right)|\bar{A}\left(k-1\right)=0,\bar{L}\left(k\right),\bar{U}\left(k\right)\right]}{P\left(\left(A\left(k\right)|\bar{A}\left(k-1\right)\right)\right)} = 1 \tag{3.27}$$

I pesi stabilizzati definiti in 3.1.3 sono quindi una misura dello scostamento dalla situazione di esogeneità statistica. Di conseguenza, nella pseudopopolazione creata utilizzando i pesi, vale la esogeneità statistica.
Il risultato fondamentale dimostrato da Robins è che in assenza di confondenti non misurati e di mispecificazione del modello utilizzato, si ha che:

- l'esogeneità statistica implica l'esogeneità causale
- lo stimatore ottenuto dal modello pesato (considerando il metodo IPTW) risulta uno stimatore non distorto e che converge in probabilità allo stimatore del modello marginale per l'outcome controfattuale $Y_{\bar{a}}$
- in presenza di esogeneità statistica gli stimatori ottenuti dal modello di associazione (come stimatori dei minimi quadrati ordinari) e dal modello marginale strutturale pesato coincidono.

Formalmente quanto sopra viene assicurato dal seguente teorema:

Teorema 2. *Se la condizione definita da* (3.23) *è vera, allora* $E[Y_{\bar{a}}]$ *è l'unica funzione di* $c(\bar{a})$ *che soddisfa l'equazione*

$$E\left[q\left(\bar{A}\right)\left(Y - c\left(\bar{A}\right)\right)/SW\right] = 0 \tag{3.28}$$

per ogni funzione $q\left(\bar{A}\right)$ *dove SW sono i pesi stabilizzati definiti nella sezione 3.1.3.*

Dimostrazione. Per la dimostrazione si rimanda a [27] □

Ricapitolando possiamo rappresentare quanto detto sopra con lo schema della figura 3.7
I modelli marginali strutturali possono includere anche covariate al *baseline* nonchè l'interazione di queste con il trattamento d'interesse. Anche i pesi vengono stimati aggiungendo al modello le covariate al *baseline*. I modelli marginali strutturali non possono però essere utilizzati per studiare l'interazione con una covariata tempo dipendente.

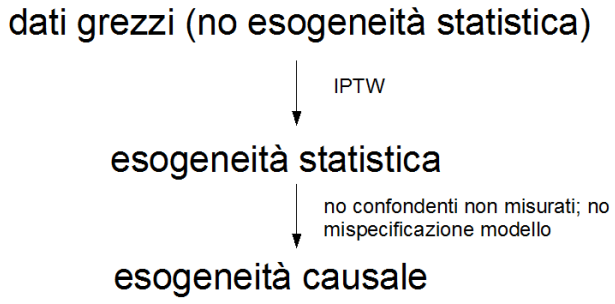


Figura 3.7:

3.2.3 Misure ripetute

Negli studi longitudinali, per i quali vengono raccolte le informazioni relative all'outcome e alla covariata d'interesse (e.g. trattamento ovvero fattore di rischio) in tempi successivi (misure ripetute) l'interesse è quello di studiare se e come la covariata influisce sul cambiamento della media marginale dell'outcome. Il modello che solitamente viene utilizzato in questa situazione è quello che utilizza le *generalized estimating equations* (GEEs) proposto da Zeger e Liang [28],[29]. Essenzialmente il modello permette di considerare la struttura di correlazione esistente all'interno dell'insieme di osservazioni del singolo soggetto.
Anche in questo caso, in presenza di confondenti tempo dipendenti, la stima dei parametri ha una interpretazione di tipo associativo, non causale. Per ottenere uno stimatore che possa essere interpretato da un punto di vista causale è neccessario considerare lo stesso tipo di modello GEE ma *pesato*, utilizzado per ogni soggetto i pesi calcolati con il metodo IPTW visto nella sezione 3.1.3 indicato in [6].
Dal punto di vista applicativo, molti programmi statistici permetono di stimare i parametri del suddetto modello considerando anche i pesi. Ad esempio utilizzando il software *Statistical Analysis System* (SAS) è possibile utilizzare la procedura *genmod* con la specificazione *weight* (o equivalentemente *scwgt*) per la variabile utilizzata per i pesi (stabilizzati oppure no).
Chiaramente i modelli marginali strutturali non rappresentano la soluzione a tutti i problemi. Infatti possiedono alcune limitazioni dovute alle ipotesi sottostanti il modello quali la presenza di confondenti non misurati, la cor-

rettezza del modello stesso nonchè quella del modello utilizzato per il calcolo dei pesi, la consistenza (come definita in 1.2) e la condizione di positività. Tuttavia si presentano come uno strumento valido per l'analisi dell'effetto causale.

3.2.4 Variabili strumentali

Il problema del confondimento non misurato può essere affrontato tramite l'utilizzo di variabili strumentali. Questa metodologia è largamente utilizzata in econometria. Una variabile strumentale è una variabile che predice l'esposizione ed è associata all'outcome solo attraverso l'esposizione (quindi non può essere associata all'outcome né direttamente né attraverso percorsi che siano misurati o no) [30]. Nella figura 3.8 viene rappresentato un DAG con una variabile strumentale.

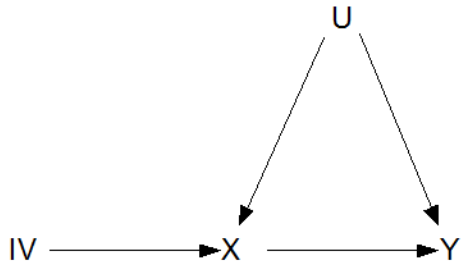


Figura 3.8: variabile strumentale

La variabile X rappresenta l'esposizione, Y l'outcome, U il confondente non misurato e IV la variabile strumentale.
Un modello per la stima degli effetti del trattamento X sull'outcome Y utilizzando la variable strumentale IV può essere il sistema seguente [31]

$$\begin{aligned} Y &= \alpha + \beta X + e \\ X &= \gamma + \delta IV + f \end{aligned} \tag{3.29}$$

dove e ed f rappresentano le componenti di errore. Se l'associazione e l'outcome sono dicotome è possibile calcolare lo stimatore 'variabile strumentale'

$\hat{\beta}_{IV}$ come rapporto tra l'associazione della variabile strumentale e l'outcome e la relazione tra la variabile strumentale e l'esposizione, ovvero:

$$\hat{\beta}_{IV} = \frac{P\left(Y = 1|Z = 1\right) - P\left(Y = 1|Z = 0\right)}{P\left(X = 1|Z = 1\right) - P\left(X = 1|Z = 0\right)} \tag{3.30}$$

É opportuno osservare che in questo caso $P\left(Y = 1|Z = 1\right) = E\left[Y|\, Z = 1\right]$ (analogamente gli altri termini). Quindi il numeratore rappresenta la differenza tra i rischi per l'outcome (RD_Y) mentre il denominatore rappresenta la differenza tra i rischi per l'esposizione (RD_X).
Tuttavia, le variabili strumentali non rappresentano la panacea in materia di controllo dei confondenti non misurati. Il primo problema che si presenta in ambito biomedico è quello di trovare una variabile strumentale per ogni problema analizzato. Anche se in letteratura sono presenti alcuni esempi di utilizzo di tali variabili ([30], [32], [33]) l'utlizzo delle variabili strumentali può introdurre distorsione in particolare se tale variabile è debolmente associata all'esposizione e non risulta appropriato nel caso di esposizioni tempo dipendenti [34]. Un altro approccio è quello di analizzare il problema dell'eventuale presenza di confondimento non misurato attraverso un'analisi della sensitività come proposto da Brumback et al. in [35].

Capitolo 4

Modelli marginali nello studio BROMS

4.1 Introduzione

Gli effetti negativi del fumo di sigaretta sulla salute ed il carico sulla spesa sanitaria associati a questa abitudine sono ormai di pubblico dominio da molti anni [36]. Le conseguenze avverse si presentano in particolare negli adolescenti in quanto, come viene riportato in molti lavori, il fumo di sigaretta provoca una diminuzione dell'appetito aumentando la possibilità di un ritardo della crescita [37]. Un altro fatto noto è che, sia tra gli adolescenti che tra gli adulti, i fumatori hanno una altezza ed un peso inferiore ai non fumatori [38], [39], [40]. Le motivazioni biologiche di questi effetti del fumo di sigaretta sono legate ad una riduzione dell'efficienza dell'utilizzo di energia [41], [42] e ad un aumento del tasso metabolico e della termogenesi [43] [44]. Un recente studio [45] ha mostrato che, nelle giovani adolescenti, la persistenza dell'abitudine al fumo di sigarette è associato ad uno sviluppo ritardato del peso, dell'altezza ed del BMI e che l'inizio di questa abitudine è associato ad una riduzione dell'aumento del peso e del BMI ma non dell'altezza. Risultati simili sono stati ottenuti in uno studio successivo [46].
Due sono le relazioni che si vogliono indagare in merito all'abitudine al fumo negli adolescenti: primo, il sovappeso aumenta il rischio di iniziare a fumare tabacco? Secondo, il fumo di sigaretta decrementa la media del BMI?. Il primo problema è già stato affrontato [47]: i risultati dello studio sugerisco-

no che il sovrappeso nelle adolescenti femmine aumenta il rischio di iniziare a fumare tabacco. Lo studio del secondo quesito invece è più complesso per la presenza di un confondente tempo dopendente, nella fattispecie lo stesso BMI : in effetti il valore del BMI pregresso è un predittore del BMI attuale ma anche dell'attuale condizione di essere fumatore. Inoltre, come detto sopra, diversi studi suggeriscono un effetto negativo del consumo di tabacco pregresso sul BMI attuale [45], [46].
L'obiettivo della presente analisi è quello di indagare l'eventuale effetto causale dell'abitudine al fumo sulla media del BMI utilizzando i dati (longitudinali) di una coorte di adolescenti svedesi.

4.2 Metodi e strumenti

La coorte BROMS

Lo studio di coorte BROMS (acronimo svedese per 'Children Smoking and their Enviroment in Stockholm region') è stato approvato dal Comitato Etico del Karolinska Institutet del Huddinge University Hospital, come precisato in [47]. La coorte è costituita da 3020 adolescenti svedesi delle scuole dell'area urbana di Stoccolma (Svezia) che hanno partecipato a una indagine annuale sul consumo di tabacco condotta nelle scuole di appartenenza a partire dal Gennaio 1998. La prima indagine è stata condotta durante il quinto grado della scuola obbligatoria (l'età media degli studenti è risultata pari a 11.6 anni) mentre le successive sono state condotte per i seguenti quattro anni della scuola dell'obbligo (una volta all'anno e sempre nel mese di Gennaio) fino all'età di 15 anni. Successivamente altre due indagini sono state condotte alle età di 17 e 18 anni. La partecipazione allo studio è stata alta in particolare tra gli studenti aventi genitori con un alto grado d'istruzione, con una minor probabilità di iniziare a fumare entro la fine del follow-up e con percezione di essere sovrappeso all'età di 15 anni.
Le caratteristiche della coorte e del disegno sono maggiormente dettagliati in altri studi [47], [48], [49].
La coorte a disposizione per il presente studio è quella utilizzata in [47] formata dai 2922 soggetti che si sono dichiarati non fumatori all'inizio del follow-up.

Definizione dell'esposizione

L'esposizione d'interesse in questo studio è rappresentata dalla quantità dichiarata dagli studenti di tempo di fumo di cigaretta è indicata come *cumsmo*. I valori di questa variabile sono compresi tra zero e sei. L'esposizione è tempo dipendente e la sua definizione è stata calcolata a partire da una variabile di tipo dicotomo, che rappresenta la risposta degli studenti e che assume valore 0 se il soggetto dichiara di non fumare oppure di fumare meno di una volta al mese, uno in tutti gli altri casi. La definizione di questa covariata dicotomica rappresenta la soglia generalmente accettata per l'uso di tabacco tra li adolescenti giovani.

Definizione dell'outcome

Di interese è il BMI (rapporto tra il peso misurato in kilogrammi e il quadrato della altezza misurata in metri, per ogni singolo soggetto) ed è di tipo continuo. La rilevazione e registrazione dei dati è stata condotta da operatori della scuola fino all'ottavo anno scolastico compreso mentre per gli anni successivi sono state considerate le informazioni rilevate dalle dichiarazioni degli studenti ai quali erano state date indicazioni specifiche su come eseguire le misure del peso e della altezza (le rilevazioni dovevano essere effettuate senza scarpe e senza i vestiti). Solo per l'ottava classe si hanno sia le informazioni raccolte dagli scolari sia quelle raccolte dall'operatore sanitario. Questo ha permesso di verificare l'attendibilità delle informazioni riportate dagli studenti attraverso il calcolo del coefficiente di concordanza k che è risultato pari a 0.89 [47].

Covariate

Le covariate disponibili al basale sono il genere, l'abitudine al fumo dei genitori (considerata come covariata indicatrice dell'abitudine al fumo di almeno uno dei due genitori), la condizione di sovrappeso e lo status socioeconomico della famiglia, quest'ultima definita in base al grado di scolarizzazione della madre dello studente (in mancanza di questa informazione viene utilizzata l'informazione del padre). In assenza di entrambi i dati, viene utilizzato il livello educativo richiesto dalla tipologia d'impiego del genitore. La covariata utilizzata è di tipo categorico a tre livelli: 'scuola dell'obbligo', 'scuola

superiore' oppure 'college'.
La covariata relativa al sovrappeso al basale merita una discussione più approfondita. Come riportato in [47] le soglie del BMI considerate nella definizione di 'sovrappeso' e 'obesità' nei bambini e negli adolescenti è leggermente diversa dalle soglie considerate nelle definizioni degli stessi concetti negli adulti [50]. Come è noto, le soglie del BMI per gli adulti sono 25 e 30 rispettivamente per sovrappeso e obesità. Nei bambini e negli adolescenti invece è necessario considerare il costante cambiamento del corpo (in termini di altezza e massa corporea). Nel lavoro di Cole et al. precedentemente citato [50] vengono presentate le definizioni di sovrappeso e obesità basate sulla proposta dell'International Obesity Task Force (IOTF) che nel 1997 ha stabilito un indice relativo a queste due condizioni nei bambini e negli adolescenti.
In questo lavoro (come in [47]) vengono utilizzate le soglie derivate dal metodo LMS di Cole [50].

Dati mancanti

Uno dei problemi fondamentali che presenta l'analisi di dati longitudinali è l'eventuale presenza di dati mancanti. La mancanza d'informazione può comportare la distorsione delle stime calcolate [51]. Da qui l'importanza dello studio e applicazione di metodi che permettono calcolare le stime dei parametri d'interesse anche in presenza di valori mancanti. Una tassonomia delle diverse configurazioni che possono assumere i dati mancanti ormai accettata universalmente è quella data da Rubin [52]. Questa classificazione è basata sul meccanismo che può produrre i dati mancanti (la così detta *missingness*): missing completely at random (MCAR) quando non ci sono differenze sistematiche tra i valori mancanti e quelli osservati, missing at random (MAR) quando ogni differenza sistematica tra i valori mancanti e quelli osservati può essere spiegata attraverso i dati osservati, oppure missing not at random (MNAR) se le differenze sistematiche tra valori mancanti e valori osservati non possono esere spiegate dai soli valori osservati [53]. Esistono numerosi lavori nei quali vengono confrontati i diversi metodi che permettono di analizzare insiemi di dati contenenti valori mancanti tra cui [54],[55],[56].
Formalmente [52], se Y è la variabile d'interesse per la quale sono presenti dati mancanti e R è la variabile indicatrice che assume valore 1 quando Y

è mancante e 0 altrimenti e se Y_{com} rappresenta l'insieme di tutti i valori Y (mancanti e non) diremo che la configurazione dei dati mancanti è MCAR se

$$P\left(R|Y_{com}\right) = P\left(R\right) \tag{4.1}$$

cioè la *missingness* non dipende dalla variabile d'interesse. L'insieme Y_{com} può essere partizionato in due sottoinsiemi: Y_{obs} e Y_{mis} che racchiudono rispettivamente i dati osservati e quelli mancanti. La configurazione MAR può essere definita dalla seguente condizione:

$$P\left(R|Y_{com}\right) = P\left(R|Y_{obs}\right) \tag{4.2}$$

cioè la *missingness* non dipende dai dati mancanti, mentre sarà MNAR se dipende dai dati mancanti Y_{mis}. Il problema principale nella trattazione dei dati mancanti è l'imposibilità di testare le ipotesi relative alla tipologia di dati mancanti utilizzando i soli dati a disposizione. In effetti, le tipologie precedentemente descritte riguardano il processo che genera tali dati mancanti e che esula dal controllo dell'analista. Per questo motivo è importante analizzare le eventuali differenze tra i soggetti aventi valori mancanti e non in base ad altre caratteristiche (covariate) che possono eventualmente intervenire nel processo di generazione dei valori mancanti. Comunque l'ipotesi MAR è quella che viene accettata nella maggior parte delle applicazioni non solo perchè di semplice implementazione attraverso la maggior parte dei software statistici utilizzati, ma perchè è stato dimostrato [54],[57] che in molte situazioni tale assunzione, anche se considerata erroneamente, non origina errori gravi nelle stime nè nei corrispondenti errori standard.
La configurazione dei dati mancanti può essere *monotona* oppure no. Nel primo caso i valori mancanti presentano una struttura ben determinata: è possibile considerare un ordinamento delle variabili presenti nel dataset in modo tale che se compare un dato mancante in una di esse allora tutte le variabili succesive nell'ordinamento sono mancanti. Ciò implica che i valori mancanti non si possono presentare in forma intermittente. Un esempi classico è quello del drop-out negli studi longitudinali. Nella coorte oggetto di studio invece, la configurazione dei dati mancanti non è di tipo monotono. Inoltre la configurazione può essere di tipo univariato se i valori mancanti si presentano su di una sola variabile (oppure su un insieme di variabili che presentano dei dati mancanti contemporaneamente) mentre sono disponibili i valori di altre covariate in modo completo.

Esistono diversi metodi per trattare i dati mancanti. Qui verranno solo citati i più utilizzati. Per una trattazione esaustiva ed esauriente di tali metodi si rimanda a [54].
I metodi per la trattazione dei dati mancanti possono essere classificati in base alle procedure utilizzate. Uno dei primi metodi utilizzati per la trattazione dei dati mancanti è quello basato sulla eliminazione delle osservazioni aventi almeno un dato mancante (*case deletion*). Ovviamente questo metodo implica la perdita di numerose informazioni. Inoltre le stime ottenute applicando questo metodo possono essere fortemente distorte e imprecise. In alternativa sono stati proposti alcuni metodi che prevedono l'imputazione singola del dato mancante. Alcuni esempi sono il '*last value carried forward*' (LVCF) che imputa il dato mancante al tempo $t+1$ 'trascinando' l'ultima informazione a disposizione al tempo t, oppure l'interpolazione lineare che imputa il dato mancante al tempo $t+1$ applicando un modello di regressione lineare che utilizza i dati a tempi t e $t+2$.
Altri metodi invece si basano sulla funzione di massima verosimiglianza. Questi metodi risultano più appropriati sotto ipotesi MCAR oppure MAR mentre, sotto ipotesi MNAR, possono produrre stime distorte. Inoltre questi metodi richiedono insiemi di dati di numerosità elevata per poter assicurare la validità delle proprietà asintotiche degli stimatori di massima verosimiglianza utilizzati. L'algoritmo senz'altro più noto per la determinazione delle stime di massima verosimiglianza in presenza di dati mancanti è l'algoritmo EM ('expectation-maximization') proposto da Dempster et al (1997) [58].
In alternativa ai metodi che si basano sull'imputazione singola dei dati mancanti esiste un metodo basato sull'imputazione multipla. Questo metodo è stato proposto da Rubin (1987) [59] ed attualmente è implementato in molti programmi statistici di ampio uso. Questo metodo necessita la specificazione della distibuzione dei dati mancanti Y_{mis} condizionata ai dati osservati Y_{obs} e al parametro θ, a differenza dei metodi basati sulla massima verosimiglianza che hanno bisogno della specificazione della distribuzione congiunta $P(Y_{obs}, Y_{mis}; \theta)$. Quindi i dati mancanti, per ogni soggetto, sono imputati dai propri valori osservati. La logica del metodo è la seguente: per ogni valore mancante vengono imputati un certo numero m (almeno 5) di valori anzichè un valore singolo. In questo modo si creano più insiemi di dati completi su ciascuno dei quali viene eseguita l'analisi (ad esempio una analisi

della sopravvivenza, un'analisi logistica, ecc.). Le stime ottenute dall'analisi fatta su ogni dataset imputato vengono 'sintetizzate' in una unica stima (una sorta di stima 'pooled') $\hat{\theta}_p$ con una corrispondente variabilità. Questa misura di sintesi è la media aritmetica delle stime ottenute cioè

$$\hat{\theta}_p = \frac{1}{m}\sum_{j=1}^{m}\hat{\theta}j \tag{4.3}$$

mentre la varianza totale della stima di sintesi $\hat{\theta}_p$ è data da [59]

$$V_{tot} = \frac{\sum_{j=1}^{m} Var_j}{m} + \left(1+\frac{1}{m}\right)\frac{\sum_{j=1}^{m}\left(\hat{\theta}_j - \hat{\theta}_p\right)^2}{m-1} \tag{4.4}$$

dove $\hat{\theta}_j$ è la stima ottenuta dall'analisi eseguita sul j-esimo dataset imputato mentre Var_j è la sua varianza che quindi si pesenta come la somma di una varianza 'entro' imputazione e una varianza 'tra' imputazioni. Inoltre, lo stimatore

$$\frac{\hat{\theta_p} - \theta}{\sqrt{V_{tot}}} \tag{4.5}$$

si distribuisce asintoticamente con una t-Student con v_m gradi di libertà espressi da

$$v_m = (m-1)\left[\frac{V_{tot}}{\left(1+\frac{1}{m}\right)\frac{\sum_{j=1}^{n}\left(\hat{\theta}_j - \bar{\theta}\right)^2}{m-1}}\right]^2 \tag{4.6}$$

Rubin [59] dimostra inoltre che l'efficienza della stima ottenuta può essere calcolata attraverso la seguente espressione:

$$eff = \frac{1}{1+\frac{\lambda}{m}} \tag{4.7}$$

dove m rappresenta la quantità di imputazioni e λ rappresenta il tasso di informazione mancante inteso come l'aumento relativo della varianza dovuto ai valori mancanti. Questa quantità può essere stimata come $\frac{\tau}{1+\tau}$ dove τ è dato da

$$\tau = \left(1+\frac{1}{m}\right)\frac{\left(1+\frac{1}{m}\right)\frac{\sum_{j=1}^{m}\left(\hat{\theta}_j - \hat{\theta}_p\right)^2}{m-1}}{\frac{\sum_{j=1}^{m} Var_j}{m}} \tag{4.8}$$

Notare che l'efficienza è un numero reale compreso tra zero e uno. Inoltre, quando m tende a infinito, la efficienza tende a uno, mentre se λ tende a infinito, l'efficienza tende a zero avendo quindi delle proprietà matematiche coerenti con il concetto di efficienza che rappresenta.
Emerge chiaramente l'analogia con le stime 'pooled' e le corrispondenti varianze utilizzate negli studi di meta analisi.
Nel caso della coorte in studio le covariate con dati mancanti sono la covariata esposizione relativa all'abitudine al fumo e la variabile di esposizione BMI dalla sesta classe in poi. La percentuale di valori mancanti nelle covariate abitudine al fumo dei genitori e status socioeconomico è molto bassa (1% circa) mentre l'informazione sul genere è completa.
Nelle figure 4.1 e 4.2 sono rappresentate le distribuzioni dei dati mancanti nell'outcome BMI e della variabile esposizione per ciascun anno a partire dalla classe sesta, stratificati in base alla categoria di sovrappeso al baseline (i.e. SI/NO) e per genere.

Come detto precedentemente non è possibile testare l'ipotesi relativa alla configurazione di tipo MAR per i dati mancanti utilizzando i soli dati a disposizione. Tuttavia, attraverso alcune analisi effettuate su tali dati, è possibile sostenere la plausibilità di tale configurazione. Ad esempio, nella coorte BROMS la distribuzione del BMI nella classe t non presenta differenze significative se condizionata al fatto che il BMI è un dato mancante alla classe $t+1$ tranne che per la classe ottava e dodicesima (p-value pari a 0.002 e 0.0004 rispettivamente). Inoltre esistono differenze significative tra le proporzioni di dati mancanti tra ragazze e ragazzi nelle classi nona e tredicesima (p-value pari a 0.0002 in entrambi i casi), con una percentuale di dati mancanti superiore nei maschi rispetto alle femmine. Questo risultato è atteso in quanto è noto che le ragazze sono più collaborative. Non ci sono invece differenze significative tra le proporzioni di dati mancanti nei fumatori e nei non fumatori. Infine, non esiste alcuna differenza significativa nelle percentuali di drop-out stratificando per genere oppure per sovrappeso al baseline.
Alla luce di tutte queste considerazioni è possibile ritenere plausibile una configurazione di tipo MAR per i dati mancanti nella variabile BMI e nella variabile esposizione relativa al consumo di tabacco.
Si è scelto d'imputare i dati mancanti mediante il metodo di imputazione multipla (Multiple Imputation (MI)). Nelle analisi è stato utilizzato il sot-

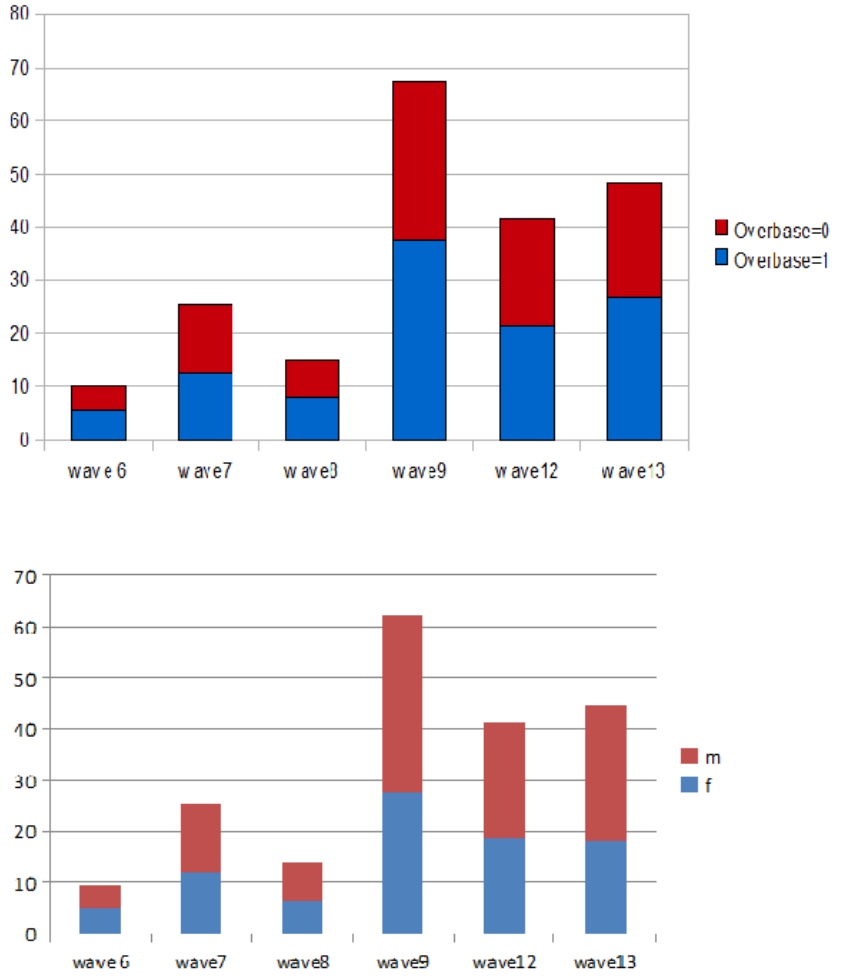


Figura 4.1: distribuzione dei dati mancanti per il BMI per categorie di sovrappeso al basale (grafico in alto) e per genere (grafico in basso)

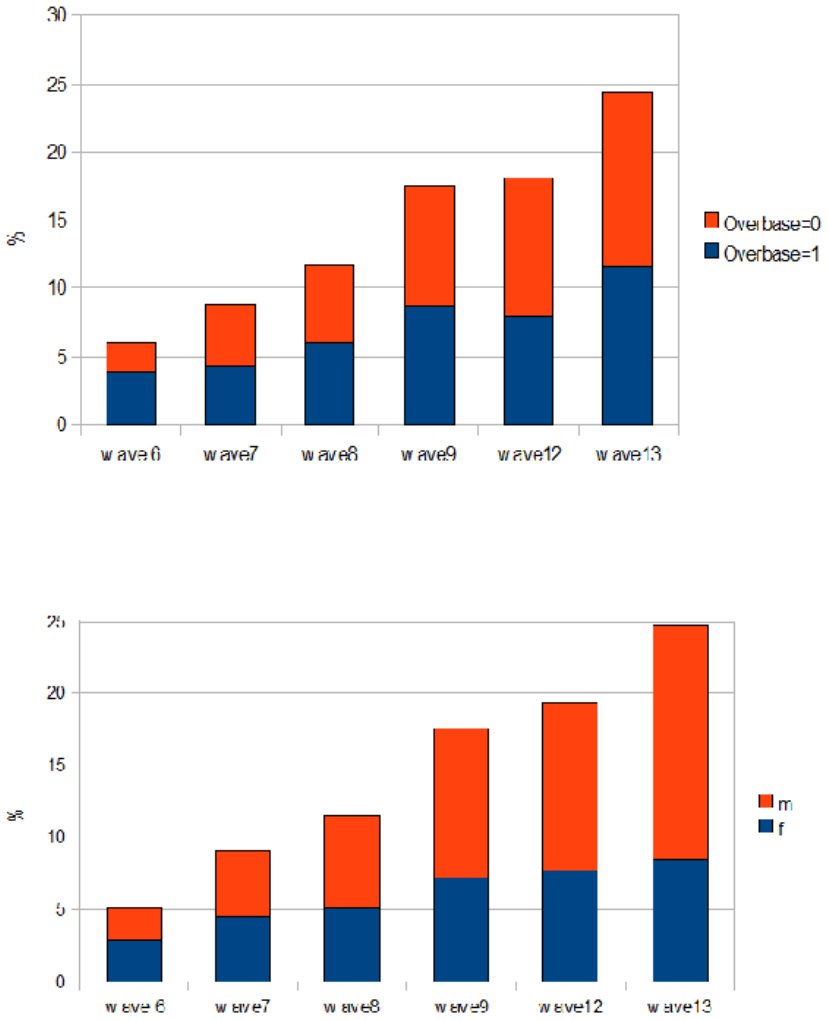


Figura 4.2: distribuzione dei dati mancanti per l'esposizione al fumo per categorie di sovrappeso al basale (grafico in alto) e per genere (grafico in basso)

fware SAS (versione 9.1) [60] che permette di effettuare tale imputazione attraverso due procedure: la procedura MI che realizza l'imputazione dei dati mancanti creando i diversi dataset imputati e successivamente la procedura MIANALYZE che sintetizza le stime ottenute dall'analisi eseguita sui singoli dataset imputati seguendo la procedura di Rubin citata precedentemente. Anche se teoricamente è sufficiente imputare cinque dataset per applicare il metodo, in generale è consigliabile imputare almeno venti insiemi per migiorare la performance del metodo [53].
La procedura MI consente di scegliere tra tre metodi di imputazione: regressione, propensity score e Markov Chain Monte Carlo (MCMC). I primi due metodi possono essere applicati solo se la configurazione dei dati mancanti è di tipo monotono. Nel metodo basato sulla regressione i valori imputati vengono generati utilizzando successive regressioni lineari per la variabile oggetto d'imputazione. Il razionale del metodo basato sul propensity score è quello di raggruppare le osservazioni in base al loro propensity score, per ogni variabile con dati mancanti. Il propensity score viene calcolato come la probabilità che una variabile abbia un valore mancante condizionatamente ai valori assunti dalle precedenti variabili nell'ordinamento considerato.
Quando i dati mancanti non sono di tipo monotono ma si può ritenere che la distribuzione delle variabili da imputare sia di tipo normale multivariato è necessario utilizzare il metodo MCMC per l'imputazione. Essenzialmente il metodo MCMC utilizza procedure d'inferenza Bayesiana per la determinazione di una catena di Markov che converge in distribuzione. I dati mancanti vengono imputati dalla distribuzione limite della catena. I valori iniziali utilizzati dal processo che genera la catena di Markov sono ottenuti applicando l'algoritmo EM che a sua volta (se non viene specificato nulla) utilizza come valori iniziali per la media e la varianza quelli calcolati sui dati completi. Inoltre è possibile scegliere se utilizzare una unica catena di Markov per tutte le imputazioni (*single chain*) oppure utilizzare più catene per le diverse imputazioni (*multiple chain*) [61], [62].
Poichè la configurazione dei dati mancanti nella coorte BROMS non è di tipo monotono è stato utilizzato il metodo MCMC sia con catena singola che multipla. Sono stati imputati i valori dei dati mancanti per le variabili relative al BMI e all'abitudine al fumo di sigaretta costruendo 20 insiemi di dati utilizzando la suddetta PROC MI di SAS (versione 9.1) [60]. Per quanto riguarda le variabili dicotome come l'abitudine al fumo, Schafer (1997)

[63] sugerisce di trattarle nel processo d'imputazione come continue e quindi arrotondare a zero oppure a uno (vale a dire, se il valore imputato è minore di 0.5 viene arrotondato a 0, viceversa viene arrotondato a 1). Horton et al. nel 2003 [64] hanno analizzato la distorsione delle stime sia nel caso dell'arrotondamento di valori imputati di variabili binarie sia nel caso in cui tali valori imputati vengano lasciati con i valori ottenuti dall'imputazione cioè senza alcun arrotondamento. Dal loro studio emerge che le stime calcolate utilizzando i valori imputati non arrotondati non sono distorte, contrariamento a quanto accade nel caso dell'arrotondamento. Un ulteriore confronto tra la distorsione delle stime prodotte quando vengono imputati i valori mancanti di variabili dicotome è stato proposto da Ake [65].
Nella coorte BROMS le variabili utilizzate nel modello per l'imputazione sono tutte le covariate predittive del BMI e del consumo di sigarette (quindi tutte le covariate al basale comprese BMI e indicatrice del consumo di sigarette). Questa scelta si basa sul fatto che il metodo dell'imputazione multipla MI produce stime non distorte se nel modello utilizzato per l'imputazioni viene considerata una quantità sufficiente di covariate predittive dei valori mancanti [53].

Analisi statistica

Come specificato sopra per studiare l'effetto causale del fumo di tabacco sulla media del BMI è necessario utilizzare un modello strutturale marginale per misure ripetute [6]. In presenza di confondenti tempo dipendenti il DAG associato può essere estremamente complesso. Nel caso del problema di cui ci occupiamo è possibile rappresentare il DAG associato in modo molto schematico come una concatenazione di 'cellule' ciascuna delle quali può essere rappresentata come in figura 4.3.
In questo DAG sono rappresentate le relazioni dirette ed indirette tra abitudine al fumo pregresso e presente nonchè quelle tra BMI pregresso e attuale. Il primo modello considerato è il seguente:

$$E\left[BMI_{\bar{a}}\left(t+1\right)|V\right] = \beta_0 + \beta_1 * cumsmo\left(t\right) + \beta_2 * t + \beta_3^{'} * V \qquad (4.9)$$

dove $\mathrm{BMI}_{\bar{a}(t+1)}$ rappresenta la variabile controfattuale associata al BMI al tempo $t+1$ relativamente alla storia di trattamento rappresentata da $\bar{a}$, *cumsmo* è la covariata tempo dipendente che rappresenta gli anni totali di fumo (i.e. la quantità totale di anni in cui è stata dichiarata l'abitudine

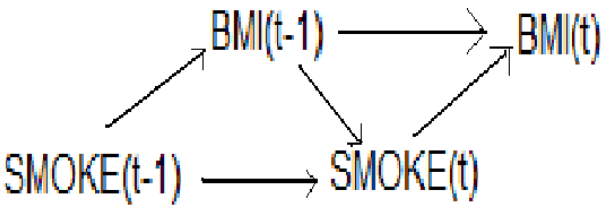


Figura 4.3: DAG BROMS

al fumo di tabacco) per ogni anno di rilevazione, t rappresenta la classe scolastica (considerata come variabile categorica ordinale) e V rappresenta il vettore delle covariate considerate al basale. Anche il parametro β_3 è un vettore. Poichè le covariate del vettore V sono considerate al basale (quindi non tempo dipendenti), è possibile anche studiare le interazioni tra queste e l'esposizione *cumsmo* [22].
Per quanto riguarda il vettore di covariate V abbiamo considerato il genere, l'abitudine al fumo dei genitori, la condizione sovrappeso all'inizio dello studio e lo status socio-economico (queste covariate sono già state definite precedentemente). Nei modelli marginali strutturali, l'inclusione dei fattori di rischio al basale nei modelli per il trattamento è di fondamentale importanza per ottenere stime consistenti dei parametri come osservato da Delaney et al. (2002)[8] e mostrato da Lefebvre et al. (2008), [66] con uno studio di simulazione attraverso il quale viene analizzata la distorsione e variabilità degli stimatori ottenuti con il metodo IPTW in diversi scenari. Questo studio sugerisce che se nel modello per il calcolo dei pesi si considerano tutte le covariate associate alla variabile di risposta, si ottiene un notevole miglioramento dello stimatore ottenuto.
Il modello (4.9) consente di studiare l'effetto causale dell'esposizione d'interesse (in questo caso gli anni cumulati di fumo dichiarato) entro ogni categoria delle covariate considerate al basale. Ricordiamo che nel modello marginale strutturale è necessario che la rilevazione del valore al tempo t del confondente tempo dipendente sia precedente all'assegnazione dell'esposizione corrispondente allo stesso tempo. Nel caso della coorte BROMS la rilevazione dell'abitudine al fumo si riferisce al fumo abituale di più di una cigaretta la mese. Questa esposizione è naturalmente riferita ad un tempo precedente quello della rilevazione del BMI. Di conseguenza, per il calcolo dei pesi, è stato considerato il valore del BMI nella classe precedente.In ogni tempo il vettore delle misurazioni del BMI è stato traslato (in avanti) rispetto all'indicazione della classe scolastica. Nella sezione 3.2.2 l'equazione (3.21) esprime i pesi stabilizzati per il trattamento (esposizione). In presenza di dati censurati è possibile aggiustare questi pesi tenendo conto della censura [6],[67]. Infatti è possibile considerare questa come un ulteriore fattore d'interesse attraverso un indicatore C tempo dipendente che assume valore pari a zero al tempo t se il soggetto è ancora sotto osservazione 'dopo' il tempo t, uno altrimenti. Sotto ipotesi equivalente a quella

dell'assenza di confondenti non misurati e cioè:

$$Y_{\bar{a}}(t+1) \coprod C(t+1) | \bar{C}(t) \equiv 0, \bar{L}(t), \bar{A}(t) \tag{4.10}$$

è possibile calcolare i pesi da utilizzare nella stima dei parametri del modello marginale strutturale come prodotto dei pesi calcolati attraverso (3.21) moltiplicati per i pesi calcolati attraverso la (4.11)

$$SW_{cens}(j) = \frac{\prod_{k=0}^{j} Pr\left(C(k+1) = 0 | \bar{C}(k) = 0, \bar{A}(k), V\right)}{\prod_{k=0}^{j} Pr\left(C(k+1) = 0 | \bar{C}(k) = 0, \bar{A}(k), \bar{L}_k\right)} \tag{4.11}$$

Anche questi pesi devono essere stimati opportunamente come i pesi SW.
Una delle ipotesi alla base dei modelli marginali strutturali è quella dell'assenza di confondenti non misurati. Purtroppo non è possibile testare questa ipotesi dai soli dati.
Una delle obiezioni che può essere sollevata riguarda il fatto che in realtà ci potrebbero essere dei confondenti non misurati come ad esempio le abitudini alimentari e l'attività fisica. Questa informazione è stata rilevata solo per la classe nona e quindi non è stato possibile inserirla nell'analisi complessiva. Tuttavia, è stata condotta un'analisi considerando solo le informazioni dalla classe nona in poi, considerando oltre alle variabili già incluse precedentemente anche le informazioni relative alle abitudini alimentari e all'attività fisica nell'insieme delle covariate al baseline. Per quanto riguarda le abitudini alimentari è stata considerata una variabile categorica a tre livelli che indica la tipologia di dieta in base alla quantità di giorni alla settimana in cui lo studente consuma frutta fresca, verdura cruda, verdura cotta, patatine fritte, snacks, bibite gasate, dolci. L'attività fisica è rapresentata da una variabile categorica a sei livelli, ciascuno dei quali definito dalla numerosità di ore alla settimana dedicate a tale attività. I risultati ottenuti da questa analisi sono simili a quelli ottenuti dall'analisi complessiva.
Tuttavia è sempre possibile la presenza di confondenti non misurati come ad esempio i fattori di tipo genetico. A questo punto l'unica possibilità è quella di effettuare una analisi di sensibilità come indicato nella sezione 3.2.3.
Un altro limite alla validità del modello proposto è quello relativo all'ipotesi SUTVA presentata nella sezione 1.2. Come detto in precedenza, secondo questa ipotesi l'outcome controfattuale di un soggetto non deve dipendere

dall'esposizione di un altro sogetto ma solo dalla propria esposizione. Ricordiamo che esempi classici dove tale ipotesi non si soddisfa è negli studi riguardanti programmi educativi e malattie contagiose. Al contrario nello studio sulla coorte BROMS l'ipotesi SUTVA può essere ritenuta valida in quanto l'outcome d'interesse è il BMI del singolo studente e l'abitudine al fumo di un compagno di classe o amico dello studente potrebbe influenzare il suo comportamento ma probabilmente cambierebb anche la sua abitudine al fumo.
In linea con quanto suggerito da Hernán in [6] l'osservazione di uno studente è stata censurata alla classe t se dalla classe $t+1$ in poi risultavano mancanti sia l'informazione relativa all'abitudine al fumo di sigaretta sia quella relativa al BMI, considerando questa censura non informativa.
Infine, sono stati esclusi dalla coorte tre soggetti per i quali era disponibile soltanto l'informazione al baseline.

4.2.1 Risultati

La coorte BROMS era composta da 2919 soggetti che hanno accumulato 13613 anni-persona. Nella tabella 4.1 vengono riportate le numerosità e le percentuali relative alle caratteristiche al baseline della coorte BROMS classificata per genere. Lo status socio economico viene indicato come basso, medio o alto in corrispondenza alle categorie definite precedentemente. Nella tabella 4.2 sono riportate le numerosità e le percentuali delle osservazioni censurate nelle diverse classi (per un totale di 505 censure corrispondenti al 17.3%) degli studenti). L'andamento delle medie del BMI per tutte le classi considerate e stratificate per genere sono rappresentate in figura 4.4. I casi incidenti di fumo sono specificati nella tabella 4.3 I dati dello studio sono stati analizzati sia con il modello classico per misure ripetute (*generalized estimating equations* GEE) sia con il modello marginale strutturale per misure ripetute. In entrambi i modelli, oltre alle covariate al basale, è stata inclusa l'interazione della variabile d'interesse *cumsmo* con la covariata relativa al genere. Questa interazione risponde al fatto che, come già indicato nella sezione 4.1, l'associazione tra abitudine al fumo e BMI è stata osservata principalmente nelle donne. Tutte le analisi sono state condotte utilizzando la procedura GENMOD di SAS (versione 9.1) che permette di applicare un modello di tipo GEE per misure ripetute modellando la struttura di correlazione all'interno di ogni singolo soggetto attraverso la

Tabella 4.1: caratteristiche al basale della coorte BROMS

	femmine N (%)	maschi N (%)
media BMI (sd)	18.50 (2.96)	18.50 (2.83)
Sovrappeso	233 (16.1)	283 (19.2)
Genitori fumatori	563 (39.3)	547 (37.4)
Status socioeconomico		
Basso	170 (11.9)	171 (11.7)
Medio	531 (37.1)	575 (39.5)
Alto	728 (51.0)	712 (48.8)
Numerosità	1444	1475

Tabella 4.2: Distribuzione nel tempo delle censure

	Studenti censurati N (%)
sesta	14 (2.8)
settima	44 (8.6)
ottava	76 (15.1)
nona	123 (24.4)
dodicesima	248 (49.1)
totale	505 (100.0)

opzione REPEATED. Attraverso l'opzione TYPE è possibile determinare la struttura di correlazione. Nello studio in esame è stata scelta la struttura di correlazione di tipo $AR(1)$, in modo tale da tener conto della distanza tra due osservazioni successive $Y(j)$ e $Y(k)$ in quanto riferita alle misure ripetute (e successive nel tempo) del BMI. In particolare,

$$corr\left[Y(j), Y(k)\right] = \rho^{|j-k|} \tag{4.12}$$

Inoltre è possibile stimare i parametri del modello strutturale marginale semplicemente introducendo nel modello i pesi IPTW soggetto-specifici attraverso l'opzione WEIGHT (o equivalentemente SCWGT).

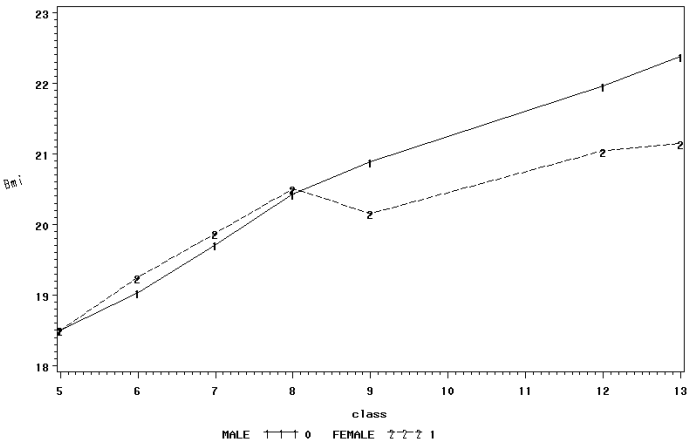


Figura 4.4: Andamento del BMI medio nel tempo, stratificato per genere

Tabella 4.3: casi incidenti di fumo nella coorte BROMS

femmine N (%)	maschi N (%)	totale N (%)
572 (39.7)	424 (28.8)	997 (37.2)
1444	1475	2919

Per il calcolo dei pesi SW e W sono stati considerati due modelli: il primo (nella tabella, *Modello marginale strutturale 1*), nel quale si assume l'ipotesi che i soggetti che iniziano a fumare continuano a farlo come spegato in 3.2.2 mentre nel secondo (nella tabella, *Modello marginale strutturale 2*) i pesi vengono calcolati attraverso un modello logistico che tiene conto del trattamento al tempo precedente. I risultati relativi ai dati originali (cioè non imputati) per i tre modelli (marginale classico e i due modelli marginali strutturali considerati) sono riportati in tabella 4.4, 4.5 e 4.6.

I risultati mostrano chiaramente come in tutti e tre i modelli sono il sovrappeso e l'interazione tra gli anni dichiarati di abitudine al fumo e genere le covariate che risultano significativamente associate ad una diminuzione

Tabella 4.4: Risultati del modello marginale classico

modello marginale classico	β	SE	$p - value$
anni di fumo	0.046	0.060	0.448
genere			
femmine	0.022	0.079	0.776
maschi	0	-	-
fumo dei genitori			
fumatori	0.114	0.082	0.166
non fumatori	0	-	-
status socioeconomico			
basso	0	-	-
medio	0.078	0.136	0.568
alto	0.028	0.132	0.829
sovrappeso			
si	5.063	0.131	$< .0001$
no	0	-	-
fumo cumulato*genere	-0.246	0.069	0.0004

Tabella 4.5: Risultati del modello marginale strutturale 1

	β	SE	$p - value$
anni di fumo	0.053	0.073	0.464
genere			
femmine	−0.039	0.086	0.650
maschi	0	-	-
fumo dei genitori			
fumatori	0.055	0.089	0.532
non fumatori	0	-	-
status socioeconomico			
basso	0	-	-
medio	0.085	0.141	0.548
alto	0.018	0.138	0.895
sovrappeso			
si	5.092	0.145	$< .0001$
no	0	-	-
fumo cumulato*genere	−0.322	0.083	$< .0001$

Tabella 4.6: Risultati del modello marginale strutturale 2

	β	SE	$p - value$
anni di fumo	0.047	0.072	0.5141
genere			
femmine	-0.041	0.086	0.6365
maschi	0	-	-
fumo dei genitori			
fumatori	0.058	0.089	0.509
non fumatori	0	-	-
status socioeconomico			
basso	0	-	-
medio	0.085	0.140	0.542
alto	0.021	0.138	0.878
sovrappeso			
si	5.089	0.144	$< .0001$
no	0	-	-
interazione fumo-genere	-0.317	0.082	0.0001

del BMI. Invece non risulta significativo l'effetto degli anni di fumo sul BMI. Si osservano leggere differenze tra le stime puntuali dei modelli marginale classico e strutturali mentre non ci sono differenze per quanto riguarda la significatività. Gli errori standard risultanti sono tendenzialmente superiori nei nei modelli marginali strutturali rispetto al modello marginale classico. Una analisi analoga è stata condotta sui dati completi (cioè sui soggetti che hanno tutte le informazioni e quindi senza dati mancanti) e sui dati imputati utilizzando, come detto precedentemente, l'imputazione multipla con il metodo MCMC. Per quanto riguarda l'imputazione multipla nelle figure 4.5 a 4.10 è riportata la diagnostica relativamente alla non indipendenza tra gli insiemi di dati imputati per i valori del BMI.

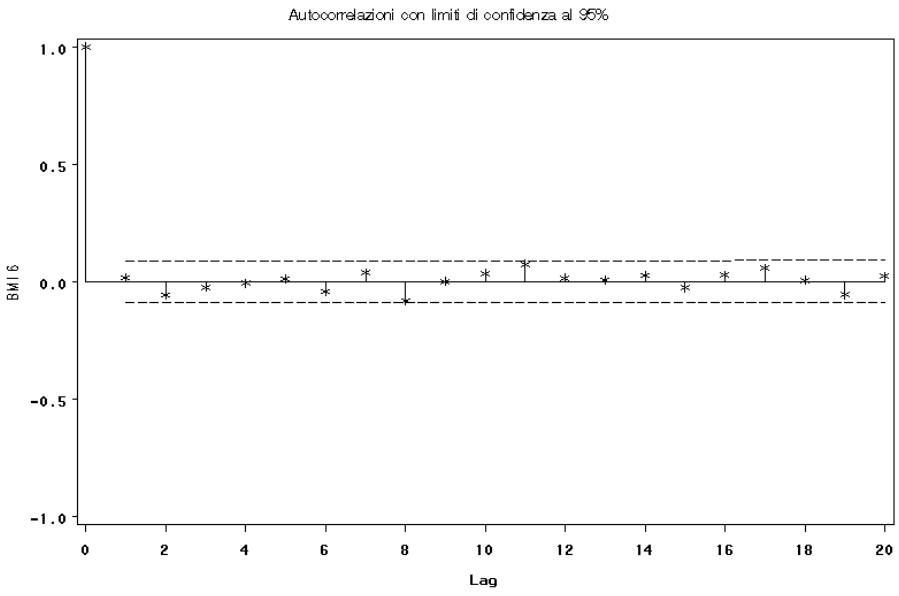


Figura 4.5: autocorrelazione BMI6

I risultati ottenuti sia dall'analisi sui dati completi (*completers*) sia sui valori imputati sono molto simili a quelli ottenuti sui dati originali (vengono riportati solo i risultati per questi ultimi).
Le distribuzioni dei pesi IPTW utilizzati per il calcolo delle stime dei parametri dei due modelli marginali strutturali $SW_{esposizione}$ e $W_{esposizione}$ per

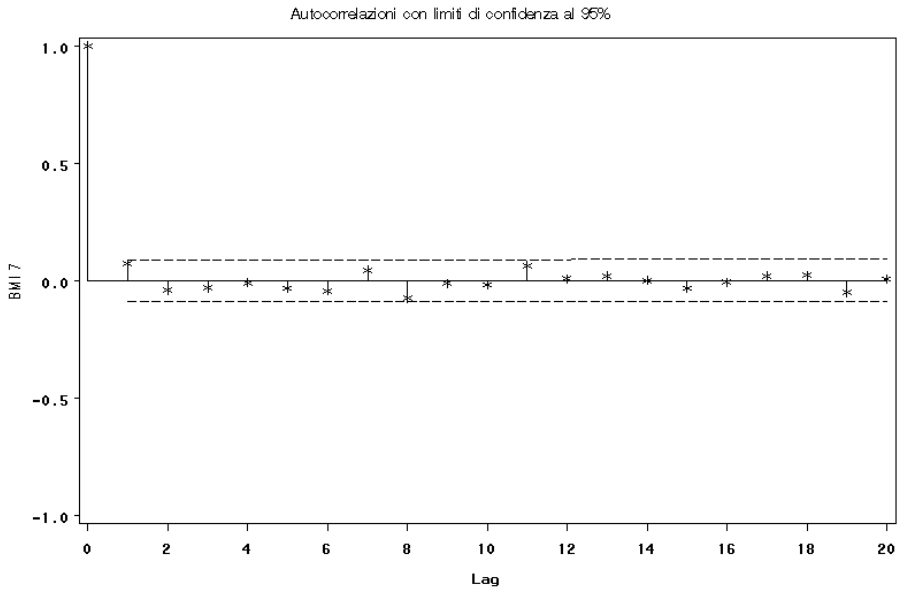


Figura 4.6: autocorrelazione BMI7

l'esposizione nonchè i pesi SW_{cen} e W_{cen} calcolati per l'indicatore del censoring per le classi seconda e dodicesima vengono riportate nelle tabelle 4.7 e 4.8. Inoltre nelle figure 4.11 e 4.12 sono rappresentati i box-plot dei pesi relativi all'esposizione per il modello marginale strutturale 1 mentre nelle figure 4.13 e 4.14 sono rappresentati i box-plot dei pesi relativi all'esposizione per il modello marginale strutturale 2. Come si può vedere chiaramente da questi grafici, i pesi stabilizzati hanno una distribuzione meno variabile dei pesi non stabilizzati.

4.2.2 Discussione

Per studiare l'effetto causale del fumo di sigaretta sul BMI è stato condotto uno studio di coorte costituita da adolescenti svedesi (coorte BROMS) utilizzando due tipologie di modelli marginali per misure ripetute: il modello classico (MMC) e il modello marginale strutturale(MSM) [6]. L'elevata numerosità della coorte BROMS e la accuratezza e tipologia dei dati raccolti

Tabella 4.7: Distribuzione dei pesi stabilizzati e non dell'esposizione e la censura per il modello marginale strutturale 1 e per le classi sesta e dodicesima

	media	mediana	1^{mo} p	99^{mo} p
classe sesta				
$SW_{esposizione}$	0.9998	0.9999	0.9973	1.0023
$W_{esposizione}$	1.9587	1.0123	1.0069	51.7563
SW_{cen}	1.0000	0.9999	0.9991	1.0009
W_{cen}	1.0035	1.0032	1.0012	1.0082
classe dodicesima				
$SW_{esposizione}$	0.9999	0.9934	0.9299	1.1163
$SW_{esposizione}$	6.7782	1.5687	1.2347	68.3884
SW^{cen}	0.9878	0.9895	0.9418	1.0259
W_{cen}	1.1732	1.1568	1.0535	1.5257

Tabella 4.8: Distribuzione dei pesi stabilizzati e non dell'esposizione e la censura per il modello marginale strutturale 2 e per le classi sesta e dodicesima

	media	mediana	1^{mo} p	99^{mo} p
classe sesta				
$SW_{esposizione}$	0.9998	0.9999	0.9980	1.0016
$W_{esposizione}$	1.9583	1.0123	1.0000	50.6877
SW_{cen}	1.0000	0.9999	0.9991	1.0009
W_{cen}	1.0035	1.0032	1.0012	1.0082
classe dodicesima				
$SW_{esposizione}$	0.9998	0.9983	0.9512	1.0786
$SW_{esposizione}$	47.3276	1.4794	1.1603	517.8872
SW^{cen}	0.9878	0.9895	0.9418	1.0259
W_{cen}	1.1732	1.1568	1.0535	1.5257

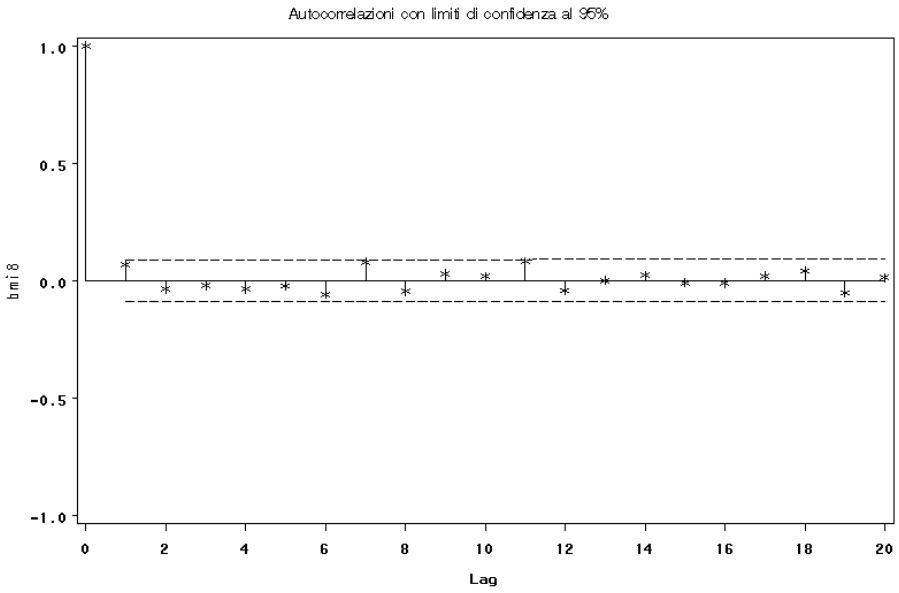


Figura 4.7: autocorrelazione BMI8

sui soggetti appartenenti alla coorte la rendono particolarente adatta allo studio di fenomeni dinamici comportamentali caratteristici dell'adolescenza. Infatti questa coorte è una tra le più numerose presenti in letteratura. Individuare il processo che lega l'abitudine al fumo alle sue principali conseguenze in età adolescenziale, rappresenta un target importante per definire le adeguate strategie di sanità pubblica. Il principali risultato di questa analisi è l'evidenza che l'effetto dell'abitudine al fumo di sigaretta è modificato dal genere. Si evidenzia come l'effetto causale cumulato del fumo di sigaretta sulla riduzione del BMI è significativo solo nelle donne: infatti, la stima del parametro relativo all'interazione tra l'esposizione al fumo e genere è pari a -0.322 (p-value < 0.001) mentre la stima del parametro relativo al consumo cumulato di sigarette nei maschi è non significativo e pari a 0.053 (p-value pari a 0.464). Questo risultato è consistente con quanto riportato in studi precedenti [45].
Tuttavia, essendo questo uno studio osservazionale, la sua capacità di misurare l'effetto causale del fumo sul BMI è discutibile. Infatti, le stime

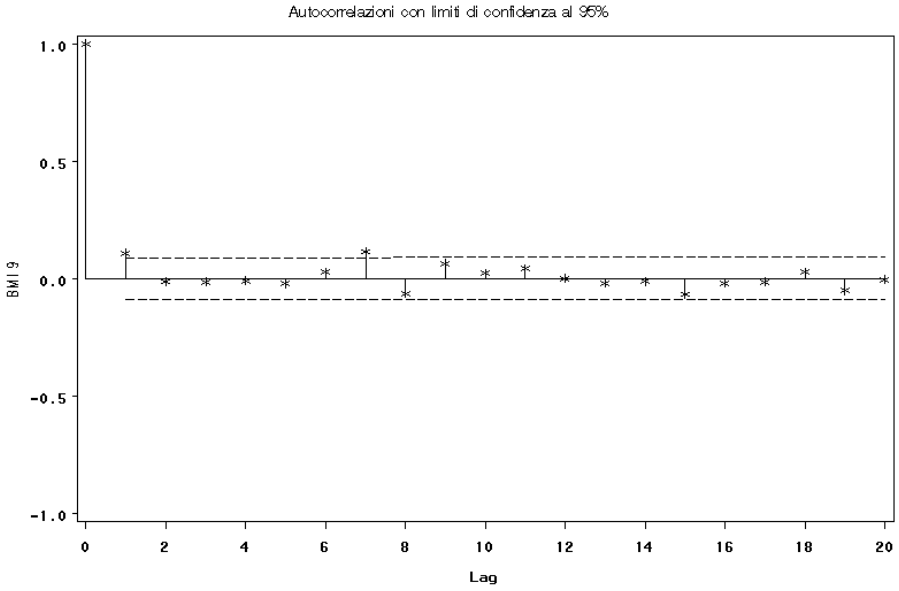


Figura 4.8: autocorrelazione BMI9

ottenute potrebbero essere affette da *selection bias* ma questo è poco probabile in quanto la coorte è stata reclutata attraverso un campionamento casuale effettuato in tutte le scuole dell'area di Stoccolma (per un totale di novantasei scuole effettivamente partecipanti), assegnando ad ogni scuola un peso proporzionale alla dimensione della scuola stessa. Un'altra forma di errore sistematico che potrebbe aver inficiato le stime ottenute è quello dovuto alla misclassificazione dell'esposizione e/o dell'outcome. Se tale errore esiste, tuttavia, non ci sono motivi per ritenerla di tipo differenziale e quindi l'eventuale effetto atteso sarebbe quello di un mascheramento dell'effetto dell'esposizione. Comunque è poco probabile che ci sia un problema di questo tipo in quanto le misurazioni delle variabili sono state eseguite da personale specializzato fino alla classe ottava, mentre da questa classe in poi le misure, pur essendo riportate dagli studenti, sono comunque risultate attendibili. L'attendibilità di queste misure è stato studiato in [47] attraverso l'indice di concordanza *kappa*. Infine, la forma di errore sistematico più rilevante potrebbe essere quella del confondimento e in particolare

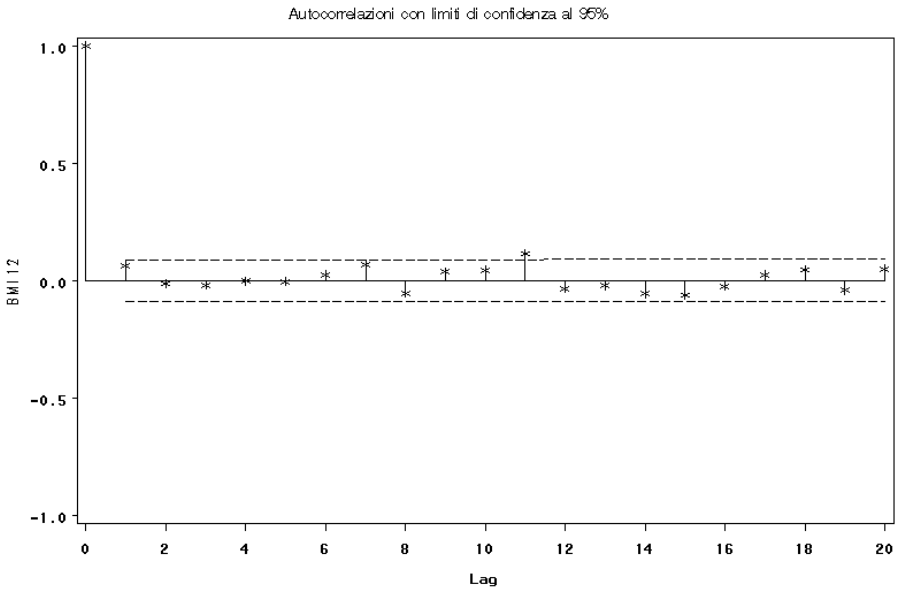


Figura 4.9: autocorrelazione BMI12

quello dovuto a confondenti tempo dipendenti. Quest'ultimo problema in particolare è stato affrontato attraversol'applicazione dei MSM.
I risultati ottenuti applicando gli MMC e gli MSM sono sovrapponibili. Questo fatto potrebbe essere spiegato da una debole intensità del legame tra il confondente tempo dipendente e l'esposizione.
Inoltre, l'applicabiità di questi modelli dipende da alcuni fattori. Innanzitutto l'assenza di mispecificazione dei modelli utilizzati per il calcolo dei pesi IPTW (stabilizzati o no). Per affrontare questo problema sono stati considerati due possibili modelli per il calcolo di tali pesi che hanno portato a stime dei parameri tra loro molto simili. Inoltre, potrebbero non essere soddisfatte le ipotesi sottostanti il modello quali l'ipotesi SUTVA, della consistenza, della positività e soprattutto dell'assenza di confondenti non misurati. Per quanto riguarda le prime tre ipotesi, in questo studio non sembrano emergere evidenze di una violazione delle stesse. Rimane tuttavia l'ipotesi di assenza di confondimento non misurato. Non possiamo escludere la presenza di tale confondimento nello stimare la relazione og-

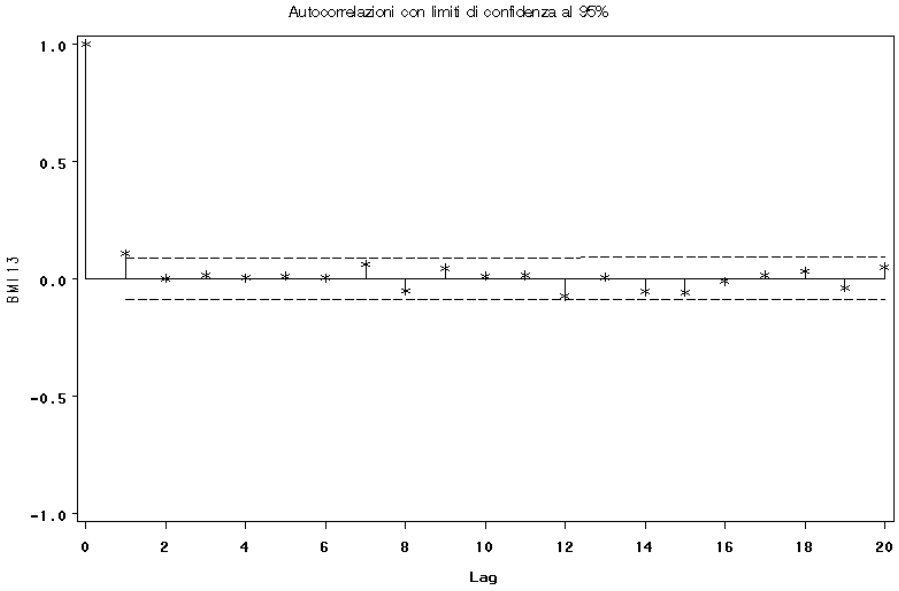


Figura 4.10: autocorrelazione BMI13

getto d'indagine data la ridotta disponibilità di alcune variabili relative alle abitudini alimentari, attività fisica e consumo di alcool, per ogni anno di rilevazione, e i frequenti cambiamenti in questi comportamenti e abitudini durante l'adolescenza. Sebbene Brumback et al. (2004) [35] e Robins et al. (1999) [68] suggeriscono di utilizzare una analisi della sensibilità per la rilevazione della mancata validità dell'ipotesi di confondimento non misurato, in questo lavoro si è preferito quantificare direttamente quanto la mancanza di queste informazioni influenzano le stime ottenute. A tal fine sono state utilizzate le informazioni relative alle abitudini alimentari ed all'attività fisica disponibili nella classe nona (unica informazione in possesso). Sono stati applicati i modelli utilizzati in questo lavoro alle informazioni a partire da questa classe, ottenendo dei risultati molto simili a quelli ottenuti non considerando queste variabili.
Infine, i modelli marginali strutturali non permettono di analizzare l'eventuale interazione tra l'esposizione e una covariata tempo dipendente. In questi casi è necessario implemetare altri tipi di modelli come i modelli

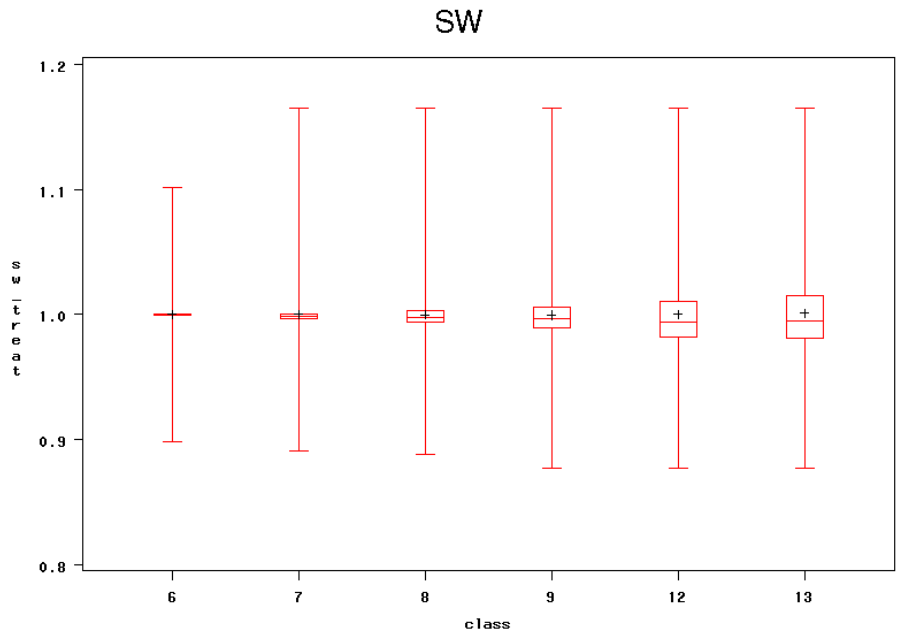


Figura 4.11: Modello marginale strutturale 1: pesi stabilizzati per il trattamento

strutturali innestati (*structural nested models*, SNM) [69].
Anche se questi modelli sono teoricamente adatti a misurare l'effetto causale di un'esposizione in presenza di confondenti tempo dipendenti, la loro complessità di applicazione e le ipotesi sottostanti il modello richiedono una particolare attenzione da parte dell'utilizzatore. É comunque auspicabile che nei prossimi anni queste metodologie diventino uno strumento standard di analisi.

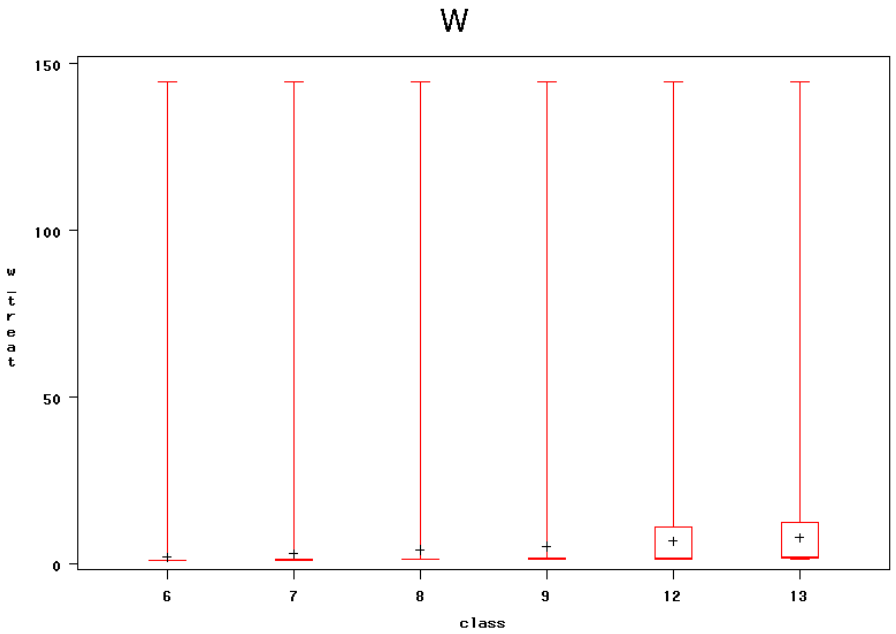


Figura 4.12: Modello marginale strutturale 1: pesi non stabilizzati per il tratamento

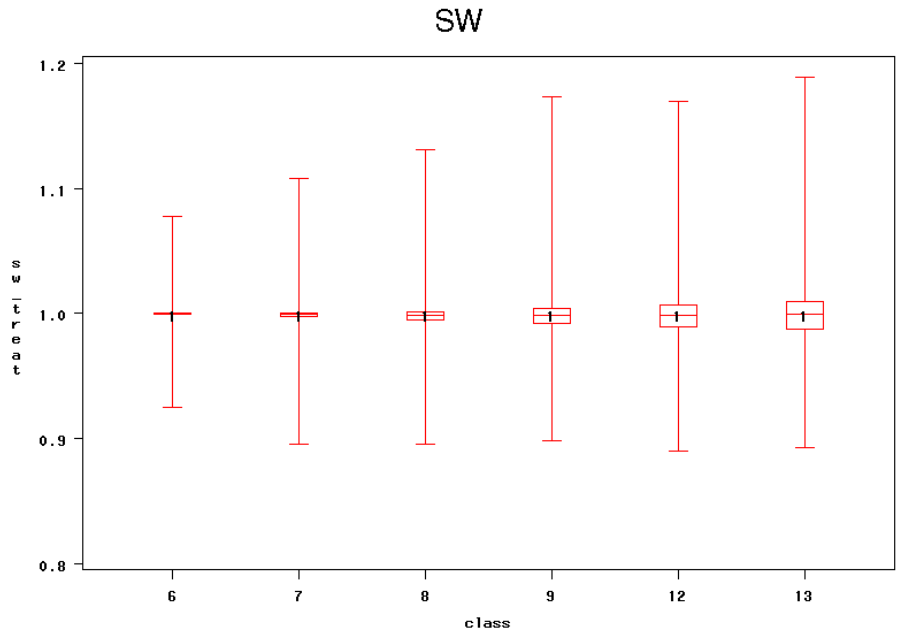


Figura 4.13: Modello marginale strutturale 2: pesi stabilizzati per il trattamento

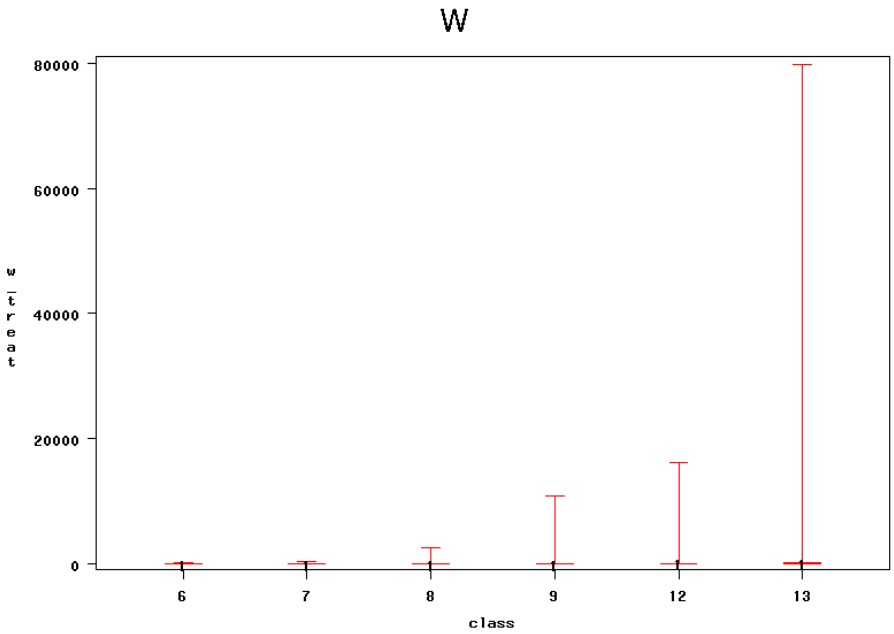


Figura 4.14: Modello marginale strutturale 2: pesi non stabilizzati per il tratamento

Bibliografia

[1] Rothman K, Greenland S. *Causation and causal inference in Epidemiology*,American Journal of Pubblic Health 2005;95:144:150.

[2] Pearl J. *Why there is no statistical test for confounding, why many think there is, and why they arealmost right*, Cognitive Systems Laboratory, Computer Science Department. University of California 1998;Technical Report R-256.

[3] Greenland S, Robins JM, Pearl J.*Confounding and Collapsability in Causal Inference*, Statistical Science , 1999;14:29-46.

[4] Miettinen O, Cook EF. *Confounding: Essence and Detection*, American Journal of Epidemiology, 1981; 114:593-603.

[5] Pearl J. *Causal Inference in the Health Sciences: A Conceptual Introduction*, Health Services and Outcomes Research Methodology 2002;2:189-220.

[6] Hernán MA, Brumback BA, Robins JM.*Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures*,Statistics in Medicine 2002;21:1689-1709.

[7] Bodnar L M, Davidian M, Siega-Riz AM, Tsiatis AA.*Marginal Structural Models for Analyzing Causal Effects of Time-Dependent Treatments: An Application in Perinatal Epidemiology*,Statistics in Medicine 2002;21:1689-1709.

[8] Delaney JAC, Daskalopoulou SS, Suissa S.*Traditional versus Marginal Structural Models to Estimate the Effectiveness of β-*

blocker Use on Mortality After Myocardial Infarction,Statistics in Medicine 2002;21:1689-1709.

[9] Rothman K, Greenland S, Lash T. *Modern Epidemiology*, Lippincot Williams & Wilkins, 3th edition, 2008.

[10] Greenland S. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, Wiley & Sons, 2004.

[11] Greenland S, Pearl J. *Causal Diagrams*, Computer Science Department. University of California 2006;Technical Report R-332.

[12] Schwartzbaum et al. *Berkson's bias reviewed*, European Journal of Epidemiology 2003;18:1109-1112.

[13] Greenland S, Pearl J, Robins JM. *Causal Diagrams for epidemiologic research*, Epidemiology 1999; 10:37-47.

[14] Hernán MA. *A definition of causal effect for epidemiological research*,J. Epidemiol Community Health 2004;58:265-271.

[15] Rao CR, Miller JP, Rao DC. *Epidemiology and Medical Statistics*.

[16] Pearl J. *Causal Diagrams for Empirical Research*, Biometrika 1995;82:669-710.

[17] Hernán MA, Hernánandez-Díaz S, Robins JM. *A structural approach to selection bias*,Epidemiology 2004;15:615-625.

[18] Hernán MA, Taubman SL. *Does obesity shorten life? The importance of well-defined interventions to answer causal questions*,J. Epidemiol Community Health 2006;60:578-586.

[19] Rosenbaum PR, Rubin DB. *The Central Role of the Propensity Score in Observational Studies for Causal Effects*,Biometrika 1983;70:41-55.

[20] Joffe MM, Rosenbaum PR. *Invited commentary: Propensity Scores*,American Journal of Epidemiology 1999; 4:327-333.

[21] Hirano K, Imbens G. *Estimation of Causal Effects using propensity score weighting: an application to data on ight heart catheterization*,Health Service and Outcomes Research Methodology 2001;2:259-278.

[22] Robins JM,Hernán MA, Brumback BA. *Marginal Structural Models and Causal Inference in Epidemiology*,Epidemiology 2000;11:550-560.

[23] Bo Lu. *Propensity score matching with time-dependent covariates*,Biometrics 2005;61:721-728.

[24] Hernán MA, Robins JM. *Estimating causal effect for epidemiological data*,J. Epidemiol Community Health 2006;60:578-586.

[25] Sturmer T, Rothman KJ, Glynn RJ. *Insights into different results from different causal contrasts in the presence of effect-measure modification*,Pharmacoepidemiol Drug Saf 2006;15:698-709.

[26] Fewell Z, Hernán MA, Wolfe F, Tilling K, Choi H, Sterne HA C.*Controlling for time-dependent confounding using marginal structural models*,The Stata Journal 2004;4:402-420.

[27] Robins JM. *Association, Causation and marginal structural models*,Synthese 1999;121:151-179.

[28] Zeger SL, Lang KY. *Longitudinal data analysis for discrete and continuous outcomes*,Biometics 1986; 42:121-130.

[29] Lang KY, Zeger SL. *Longitudinal data analysis using generalized linear models*,Biometrika 1986; 73:13-22.

[30] Rassen JA et al. *Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships*, Journal of Clinical Epidemiology 2009; article in press.

[31] Martens EP et al. *Instrumental variables.Applications and limitations*, Epidemiology 2006;17:260-267.

[32] Brookhart MA et al. *Evaluating short-termdrug effect using physician-specific precribing preference as an instrumental variable*, Epidemiology 2006;17:268-275.

[33] Henneman TA et al. *Estimating causal parameters in marginal structural models with unmeasured confounders using instrumental variables* , Working paper series University of California, Berkeley 2002;Paper 104.

[34] Hernán MA, Robins JM. *Instruments for causal inference. An epidemiologist's dream*,Epidemiology 2006;17:360-372.

[35] Brumback BA, Hern´àn MA, Haneuse SJPA, Robins JM. *Sensitivity analysis for unmeasured confounding assuming a marginal structural model for repeated measured*, Statistics in Medicine 2004;23:749-767.

[36] *Preventing tobacco use among young people: a report of the Surgeon General*, Atlanta, GA: US. Department of Health and Human Services, Office on Smoking and Health, 1994.

[37] Camp DE, Klesges RC, Relyea G. *The relationship between body weight concerns and adolescent smoking*,Health Psychol 1993;1:24-32.

[38] Klesges RC, Ward KD, Ray JW, et al. *The prospective relationships between smoking and weight in a young, biracial cohort: the Coronary Artery Risk Development in Young Adults Study*,J Consult Clin Psychol 1998;66:987-93.

[39] Lall KB, Singhi S, Gurnani M,et al. *Somatotype, physical growth and sexual maturation in young male smokers*,J Epidemiol Commun Health 1980;34:295-8.

[40] Klesges RC, Meyers AW. *Smoking, body weight and their effects on smoking behavior: a comprehensive review of the literature*,Psychological Bulletin 1989; 106:204-230.

[41] Wagner-Sradar SA, Levine SA, Morley JE, et al.*Effects of cigarette smoke and nicotine on feeding and energy* Psysiol Behav 1984;32:389-95.

[42] Cryer PE, Haymond MW, Santiago JV, Shah SD. *Norepinephrine and epinephrine release and adrenergic mediation of smoking associated with hemodynamic and metabolic events*,N England J Med 1976;295:573-7.

[43] Newsholme EA. *A possible metabolic basis for the control of body weight*,N England J Med 1980; 302:400-5.

[44] Perkins KA, Epstein LH, Marks BL, et al. *The effect of nicotine on energy expenditure during light physical activity*,N England J Med 1989;320:898-903.

[45] Stice E, Martìnez EE. *Cigarette smoking prospectively prediscts retarded physical growth among female adolescents*,Journal of Adolescent Health 2005;37:363-370.

[46] Fidler JA, West R et al. *Does smoking in adolescence affect body mass index, waist or height? Findings from a longitudinal study*, Adiction 2007;102:1493-1501.

[47] Caria MP et al. *Overweight and perception of overweight as predictors of smokeless tobacco use and of cigarette smoking in cohort of Swedish adolescents*,Adiction 2007;102:1493-1501.

[48] Galanti MR, Rosendahl I et al. *Early gender differences in adolescent tobacco use-the experience of a Swedish cohort*, Scand J Public Health 2003;29:314-17.

[49] Post A, Galanti MR, Gilljam H. *School and family participation in a longitudinal study of tobacco use: some methodological notes*, Eur J Public Health 2003;13:75-6.

[50] Cole TJ et al. *Establishing a standard definition for child overweight and obesity worldwide: international survey*, BMJ 2000;320:1240-1243.

[51] Fairclough DL, Peterson HE, Chang V. *Why are missing quality of life data a problem in clinical trials of cancer therapy*, Stat Med 1998;17:667-677.

[52] Rubin D B. *Inference and missing data*, Biometrika 1976;63:581-592.

[53] Sterne JAC et al. *Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls*, BMJ 2009;338:b2393.

[54] Schafer JL, Graham JW. *Missing data: our view of the state of art*, Psychological Methods 2002;7:147-177.

[55] Engels JM, Diehr P. *Imputation of missing longitudinal data: a comparison of methods*, Journal of Clinical Epidemiology 2003;56:968-976.

[56] Donders ART et al. *Review: A gentle introduction to imputation of missing values*, Journal of Clinical Epidemiology 2006;59:1087-1091.

[57] Collins LM, Schaffer JL, Kam CM.*A comparison of inclusive and restrictive strategies in modern missing-data procedures*, Psychological Methods 2001;6:330-351.

[58] Dempster AP, Laird NM, Rubin DB. *Maximum likelihood estimation from incomplete data via EM algorithm (with discussion)*, Journal of the Royal Statistical Society, Series B 1977;39:1-38.

[59] Rubin DB. *Multiple imputation for non-response in surveys*, New Work:Wiley 1987.

[60] SAS/STAT Institute *Software version 9.1.*Cary (NC): SAS Institute, Inc.

[61] Yang C Yuan. *Multiple imputation for missing data: concepts and new development (Version 9.0)*, SAS Institute Inc. 1987.

[62] Molenberghs G, Verbeke G. *Mdels for discrete longitudinal data*, New York: Springer 2005.

[63] Schafer J L. *Analysis of incomplete multivariate data*, London: Chapman and Hall 1997.

[64] Horton NJ, Lipsitz SR, Parzen M. *A potential for bias when rounding in Multiple Imputation*,American Statistician 2003;57:229-232.

[65] Ake CF *Rounding after multiple imputation with non-binary categorical covariates*, Paper 112-30,SUGI 30.

[66] Lefebvre G, Delaney JAC, Platt RW. *Impact of mis-specification of the treatment model on estimates from a marginal structural model*, Statistics in Medicine 2008;27:3629-3642.

[67] Hernán MA, Brumback BA, Robins JM. *Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men*, Epidemiology 2000;11:561-570.

[68] Robins JM, Rotnitzky A, Scharfstein DO *Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal models*, Halloran E, Berry D eds. Statistical Methods in Epidemiology: The Environment and Clinical Trials. New York: Springer Verlag 1999; 1-92.

[69] Robins J M. *Structural nested failure times models*In: Andersen P K, Keiding N, section eds. Survival Analysis. In Armitage P, Colton T eds. The Encyclopedia of Biostatistics. Chichester, UK:John Wiley and Sons, 1998; 4372-4389.