UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA E INFORMATICA

Tesi di Laurea

UNO STUDIO BASATO SULLA VEROSIMIGLIANZA A COPPIE PER DATI BINARI DIPENDENTI

Relatore: Prof. ssa Laura Ventura

Correlatore: Dott. Cristiano Varin

Laureando: Paolo Girardi

Anno Accademico 2007 - 2008

Indice

Introduzione

Questa tesi ha come obiettivo lo studio della verosimiglianza a coppie e l'applicazione di questo strumento ad un caso studio. Questo approccio, alternativo alle consuete metodologie statistiche, non è ancora sufficientemente esplorato e coinvolge la nozione di verosimiglianza composita, che corrisponde ad una classe di pseudo-verosimiglianze, comprendente, tra l'altro, la verosimiglianza a coppie, la verosimiglianza di ordine m e la pseudo-verosimiglianza di Besag (1974). In generale, quando la funzione di verosimiglianza non può essere calcolata esplicitamente, si può definire una opportuna verosimiglianza composita facilmente calcolabile e utile per l'inferenza, senza l'ausilio di procedure di simulazione.

Grazie ad una collaborazione con Nicola Pisastu, ricercatore del CRS4 (Center for Advanced Studies, Research and Development in Sardinia), è stato analizzato un dataset genetico sulla Talassemia, malattia che si presenta in Sardegna con maggiore frequenza rispetto ad altre zone del Mediterraneo. Il dataset presenta una struttura di correlazione tra le osservazioni dettata principalmente dal fatto che i dati si riferiscono a persone con legami di parentela più o meno forte. L'idea è di modellare tramite la verosimiglianza a coppie tale struttura di dati, laddove la verosimiglianza classica non può fornire strumenti validi di inferenza a causa della mancanza dell'ipotesi di indipendenza tra le osservazioni.

Nel Capitolo 1, si fornisce una definizione della verosimiglianza composita in generale e, più in particolare, si definisce in maniera rigorosa il concetto alla base della verosimiglianza a coppie. Analogamente alla verosimiglianza genuina, si formalizzano le principali statistiche test collegate per la verifica di ipotesi sui parametri stimati. Inoltre, poiché il dataset si presenta con variabili risposta dicotomiche, per questo motivo nel paragrafo 1.4 si approfondisce il concetto di verosimiglianza a coppie per dati binari e le possibile alternative,

come le equazioni di stima generalizzate, delineando i pregi e i difetti delle due metodiche descritte.

Successivamente, nel Capitolo 2, si definiscono alcune tecniche utilizzate in letteratura per la costruzione di verosimiglianze a coppie e si mostrano i passaggi fondamentali per la sua costruzione con l'utilizzo del programma statistico R, un ambiente statistico per l'analisi di dati. Una copia di R può essere scaricata gratuitamente accedendo all'indirizzo web http://www.R-project.org/bin, in cui si trovano le versioni per i diversi sistemi operativi. In questa tesi le analisi sono state effettuate utilizzando la versione 2.5.1 per Windows.

Nel Capitolo 3 si procede con l'analisi del dataset sulla Talassemia, inizialmente dando importanza alla struttura delle osservazioni e all'imputazione dei dati mancanti in esso. Successivamente si è passati alla selezione delle variabili che risultavano discriminanti per la malattia, con la stima di semplici modelli predittivi binomiali per verificare se la selezione delle variabili era corretta, per la discriminazione degli affetti dalla malattia dai soggetti sani.

"'All models are wrong, but some models are useful'" è quanto afferma G. E. P. Box (1979), sottolineando che i modelli sono tutti sbagliati perché forniscono una visione semplificata della realtà, ma alcuni di essi sono utili in quanto, se ben specificati, riescono a cogliere le principali associazioni tra le variabili e a dare una chiave di lettura corretta sui fenomeni che riguardano l'analisi. Partendo da questo concetto, nel Capitolo 4 si procede alla stima di un modello di regressione con l'utilizzo della verosimiglianza a coppie. Considerando che il dataset ha osservazioni dipendenti tra persone con legami di parentela, si utilizza la verosimiglianza a coppie per tener conto della struttura di correlazione tra coppie di osservazioni, cosa che un modello lineare o lineare generalizzato non riuscirebbe a cogliere. Attraverso opportune tecniche computazionali si sono stimati gli errori delle stime, per poter commentare i risultati ottenuti dalla verosimiglianza a coppie e poter effettuare un confronto con il modello marginale presentato nel Capitolo 3.

Infine, nell'appendice si riportano i principali risultati delle analisi effettuate con l'ambiente statistico R e il codice dei programmi utilizzati per l'analisi dei dati con delle brevi spiegazioni sui passaggi effettuati.

In conclusione, il lavoro effettuato con questa tesi è stato interessante e stimolante in quanto si è effettuato lo studio e l'applicazione di una tecnica, quale la verosimiglianza a coppie, che non avevo incontrato nel mio percorso formativo. Tale metodologia statistica è in pieno

sviluppo, date le numerose recenti pubblicazioni in merito, e fa parte di un progetto di ricerca che coinvolge vari studiosi a livello nazionale e internazionale.

Capitolo 1
La Verosimiglianza a Coppie

1.1 Introduzione alla pseudo-verosimiglianza

In questo capitolo si focalizza l'attenzione sul concetto di pseudo-verosimiglianza e, in particolare, sulla verosimiglianza a coppie (*pairwise likelihood*). Questo tipo di pseudo-verosimiglianza risulta interessante e di grande utilità nei casi in cui si sia di fronte a dati in cui è presente una qualche struttura di dipendenza. Nell'analisi e modellazione di dati che hanno strutture di questo tipo, la scrittura della verosimiglianza propria può risultare complicata e difficile da esplicitare. Infatti, l'usuale assunzione, alla base della funzione di verosimiglianza, di indipendenza tra le osservazioni viene a mancare. Per questo motivo può essere utile ricorrere ad una verosimiglianza ridotta e strutturata in modo che, prendendo coppie (o triplette o, in generale, gruppi) di osservazioni, sia possibile modellare la dipendenza tra i dati e che le stesse coppie di osservazioni si possano presupporre indipendenti tra loro. In questa situazione, la funzione di "densità" usata nella verosimiglianza a coppie può essere una qualunque funzione di densità bivariata. Su questo argomento sono molti gli articoli presenti in letteratura, a significare che questa psuedo-verosimiglianza genera diffuso interesse, grazie anche alla crescente velocità computazionale degli elaboratori esistenti, che permette una più agevole stima dei modelli utilizzati. I primi tentativi di proporre una pseudo-verosimiglianza di questo tipo risalgono a Besag (1974) nel contesto della modellazione di dati spaziali, e alla verosimiglianza parziale di Cox (1975), introdotta per l'inferenza in modelli a rischio proporzionale. Successivi e più specifici usi della verosimiglianza a coppie sono attribuibili a Zeger (1988), Lindsay (1988), Liang (1992) e

Nott (1999), che hanno introdotto le equazioni di stima generalizzate (GEE) per l'inferenza in modelli bivariati, partendo dalle funzioni di distribuzione marginali e da una matrice, o funzione, di correlazione.

Per l'analisi di dati binari dipendenti, sui quali focalizzeremo l'attenzione in questa tesi, interessanti pubblicazioni sono Le Cessie e Van Houwelingen (1992), Kuk e Nott (1999) e Varin (2007). L'utilizzo della verosimiglianza a coppie per dati binari correlati risulta anche utile per analizzare la struttura di dipendenza all'interno dei dati stessi. Come suggerito da molti autori, le analisi di dati bivariati sono piuttosto sensibili rispetto alla specificazione della struttura di correlazione e la possibilità di stimare la correlazione, e di tenerne conto nella specificazione del modello, garantisce stime più precise dei parametri del modello.

In questo capitolo si presenta la verosimiglianza a coppie partendo dai primi spunti in letteratura su tale verosimiglianza. Nella parte centrale di questo capitolo si approfondirà questo tema portando esempi di studi effettuati con la verosimiglianza a coppie, in particolare nel caso di dati dicotomici: in questo caso si presenterà una trattazione completa della modellazione di dati binari, con particolare attenzione alle funzioni di densità usate in letteratura in questo caso. Alla fine di questo capitolo si presenteranno le eventuali alternative alla verosimiglianza a coppie, in particolare le GEE, evidenziandone tuttavia le carenze per dati dicotomici dipendenti.

1.2 La verosimiglianza a coppie

Nel caso in cui sia presente una struttura di correlazione nei dati, l'inferenza basata sulla verosimiglianza propria può diventare molto complessa, sia da specificare che da analizzare. In questo caso, può essere opportuno cercare di modificare la struttura della verosimiglianza propria sia per guadagnare in robustezza che per rimediare alla complessità della verosimiglianza completa. In questo paragrafo si considera una particolare pseudo-verosimiglianza, la verosimiglianza a coppie, utile quando vi è una complessa struttura di dipendenza nelle osservazioni.

Sia $Y=(Y_1,...,Y_n)$ un campione estratto da una variabile causale Y con densità $f(y;\theta)$, indicizzata da un parametro ignoto $\theta\in\Theta\subseteq R^m, m\geq 1$. Supponiamo che, a causa della dipendenza tra le componenti di *Y*, la densità $f(y;\theta)$ sia difficile da esplicitare, ma che sia possibile calcolare i contributi alla verosimiglianza per un qualche sotto insieme dei dati. Nasce da questa idea la verosimiglianza di Besag (1974), che ha portato alla formalizzazione della verosimiglianza composita (*composite likelihood)* di Lindsay (1988).

La verosimiglianza composita (CL) è definita come il prodotto ponderato delle verosimiglianze associate ai singoli eventi A_i , *i=1,...k,* ovvero

$$CL(\theta;y)=\prod_{i=1}^{k} f_{A_i}(y\in A_i;\theta)^{w_i}, \tag{1.1}$$

dove w_i sono i pesi assegnati ad ogni evento A_i e $f_{A_i}(.)$ la funzione di densità associata ad ogni singolo evento A_i . Nel caso più semplice la (1.1) origina la "*singlewise likelihood",* definita come:

$$L_{SL}(\theta;y)=\prod_{i=1}^{m} f_i(y_i;\theta)^{w_i}, \tag{1.2}$$

dove $f_i(.)$ è la funzione di densità scelta nella "*singlewise likelihood".* Invece, nel caso in cui si considerino coppie o più gruppi di osservazioni, nella (1.1) si utilizza una funzione di densità bidimensionale o multidimensionale e la verosimiglianza composita associata sarà

calcolata a coppie o a più dimensioni di dati. La verosimiglianza a coppie *(pairwise likelihood)* risulta così specificata

$$L_{PL}(\theta; y) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} f_{ij}(y_i, y_j; \theta)^{w_{ij}}, \tag{1.3}$$

dove $f_{ij}(.)$ è la funzione di densità per le coppie di osservazioni $y_i e\, y_j$. In particolare la funzione $f_{ij}(.)$ è una funzione di densità bivariata scelta, o costruita, opportunamente partendo dalle distribuzioni marginali di *Y***.** Una adeguata formalizzazione dei procedimenti che portano alla costruzione della verosimiglianza a coppie è discussa, ad esempio, in Cox e Reid (2004) e Varin (2007).

La stima del parametro θ avviene, in analogia con il caso della verosimiglianza completa, andando a massimizzare la (1.3) oppure uguagliando a zero la derivata prima della verosimiglianza a coppie, ossia la funzione *score* a coppie, e andando a determinare lo pseudo-stimatore di massima verosimiglianza, indicato con SMVC (Stimatore di Massima Verosimiglianza a Coppie, in letteratura *Maximum Composite Likelihood Estimator).* La (1.3) gode di molte proprietà di una verosimiglianza propria. In particolare:

- gli stimatori SMVC hanno distribuzione asintotica normale;
- gli stimatori SMVC sono invarianti, asintoticamente non distorti ed efficienti.

La funzione *score a coppie* è definita come

$$s(\theta) = s(\theta; y) = \frac{\partial}{\partial \theta} \log L_{PL}(\theta; y) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} s_{ij}(\theta; y_i, y_j), \tag{1.4}$$

dove $s_{ij}(\theta; y_i, y_j)$ è la derivata prima rispetto a θ dei singoli contributi alla log-verosimiglianza a coppie, ossia

$$s_{ij}(\theta; y_i, y_j) = \frac{\partial}{\partial \theta^T} \log f_{ij}(y_i, y_j, \theta;). \tag{1.5}$$

Poiché vi è l'ipotesi di indipendenza delle osservazioni se prese a coppie, sotto opportune condizioni di regolarità, lo SMCV risulta consistente e asintoticamente normale, ovvero

$$\hat{\theta}_{SMVC} \dot{\sim} N_m(\theta, G(\theta)^{-1}), \tag{1.6}$$

dove $G(\theta)$ è la matrice di informazione di Godambe (1960) associata alla verosimiglianza a coppie (1.3). L'espressione della $G(\theta)$ è

$$G(\theta) = H(\theta)^T J(\theta)^{-1} H(\theta), \tag{1.7}$$

con $H(\theta) = E[-\frac{\partial}{\partial\theta} s(\theta; Y)]$ e $J(\theta) = var(s(\theta; Y))$. La (1.7) è nota anche come matrice di informazione *sandwich.* Questa espressione per la matrice di varianze e covarianze è dovuta al fatto che, in questo caso, la seconda identità di Bartlett non risulta verificata, ossia $H(\theta) \neq J(\theta)$. Per costruzione, la $G(\theta)$ tiene conto della perdita di efficienza rispetto allo stimatore di massima verosimiglianza completo per θ. La stima di (1.7) è, in molti casi, complessa e la soluzione in forma esplicita può risultare complicata. Per questo motivo si ricorre a metodi *Bootstrap* o *Jack-kinfe* per stimare indirettamente la matrice di Godambe e ottenere di conseguenza gli *standard error* associati allo SMVC.

1.3 Statistiche test collegate alla verosimiglianza a coppie

A partire dalla (1.6) si possono formulare delle statistiche test, come avviene usualmente con la verosimiglianza classica. Ad esempio, se siamo interessati ad effettuare un verifica di ipotesi su una componente γ di dimensione q del parametro θ, con $\theta = (\gamma, \tau)$, l'ipotesi nulla sarà $H_0: \gamma = \gamma_0$. Il corrispondente test alla Wald assume l'espressione

$$W_a = (\hat{\gamma}_{SMVC} - \gamma_0)^T G_{\gamma\gamma}(\hat{\theta}_{SMVC})(\hat{\gamma}_{SMVC} - \gamma_0), \tag{1.8}$$

dove $G_{\gamma\gamma}(\hat{\theta}_{SMVV})$ è la sotto-matrice di dimensione della matrice di Godambe, relativa a γ, calcolata con i parametri calcolati nel SMVC.
Il test score di verosimiglianza a coppie prende invece questa forma

$$W_s = s_\gamma(\theta_0)^T H_{\gamma\gamma}^{-1}(\theta_0) J_{\gamma\gamma}(\theta_0) H_{\gamma\gamma}^{-1}(\theta_0) s_\gamma(\theta_0), \tag{1.9}$$

dove $\theta_0 = (\gamma_0, \hat{\tau_{SMVC}}(\gamma_0))$, con $\hat{\tau_{SMVC}}(\gamma_0)$ stimatore parziale di massima verosimiglianza a coppie per τ con γ fissato, $s_\gamma(\theta)$ è la componente della funzione score di verosimiglianza relativa a γ e $H_{\gamma\gamma}(\theta_0)$ la sotto-matrice di $H(\theta)$ relativa al parametro γ e $J_{\gamma\gamma}(\theta_0)$ è la sotto-matrice di informazione stimata in θ_0. Come conseguenza della normalità asintotica dello SMCV, sia W_a che W_s hanno distribuzione asintotica nulla χ_q^2.
Un altro test che si può applicare, come nel caso della verosimiglianza, propria è la statistica log-rapporto di pseudo-verosimiglianza, data da

$$W_r = 2(\log L_{PL}(\hat{\theta}_{SMVC}; y) - \log L_{PL}(\theta_0; y)). \tag{1.10}$$

Tuttavia, in questo caso, la distribuzione asintotica nulla della statistica W_r non è quella usuale. Infatti, la distribuzione asintotica nulla di W_r è una combinazione lineare di χ_1^2 non indipendenti, per cui si ha

$$W_r \dot{\sim} \sum_i^q \lambda_i Z_i^2, \tag{1.11}$$

dove le Z_i^2 sono variabili casuali χ_1^2 e i λ_i sono gli autovalori della matrice $H_{\gamma\gamma}(\theta_0)^{-1} J_{\gamma\gamma}(\theta_0)$.
Esistono diverse proposte in letteratura per cercare di portarsi nella situazione di indipendenza e quindi usufruire di una statistica test della forma (1.10) con distribuzione nota. Ad esempio, Geys *et al.* (1999) propongono di dividere la statistica W_r per la media aritmetica dei valori degli autovalori. In questo modo si cerca di aggiustare la distribuzione asintotica per poter confrontare il valore osservato di W_r con il percentile opportuno di un χ_q^2. In alternativa,

Aerts e Claeskens (1999) e Bellio e Varin (2005) propongono un *Bootstrap* parametrico o una tecnica *Jack-knife* per calcolare in modo iterativo la distribuzione di W_r basandosi su simulazioni.

In letteratura, la (1.3) è stata utilizzata in diversi ambiti applicativi. In particolare, i principali contesti in cui è stata utilizzata la verosimiglianza a coppie sono:

- dati categoriali: molti autori hanno pubblicato lavori sulla verosimiglianza a coppie applicata a dati categoriali, come Leon (2005) e Renard (2004);
- dati di sopravvivenza: la verosimiglianza a coppie è utile per la stima di modelli di sopravvivenza multivariati in cui vengono osservati diversi eventi per ogni unità statistica; per approfondimenti si invita a Andersen (2004) e Zhao e Joe (2005);
- osservazioni longitudinali: la verosimiglianza a coppie è stata utilizzata in questo ambito da Molenberghs e Verbeke (2005);
- dati spaziali: è il più vasto ramo in cui la verosimiglianza a coppie ha avuto successo e utilizzo; si vedano i lavori di Nott e Ryden (1999) per l'analisi di immagini e Varin *et al.* (2005) per l'analisi di modelli lineari in dati spaziali;
- serie storiche: anche nelle serie storiche la verosimiglianza a coppie è stata utilizzata per l'inferenza in vari modelli, come per esempio in catene di Markov, come proposto da Ryden (1994).

Nel seguito di questo capitolo, si focalizza l'attenzione al contesto particolare della verosimiglianza a coppie applicata a dati binari. Inoltre, nei capitoli successivi, si discuterà un'applicazione a dei dati reali.

1.4 Dati binari

Il ricorso alla verosimiglianza a coppie per dati dipendenti con risposta dicotomica è molto utilizzato in letteratura, in quanto l'analisi dei dati binari risulta molto sensibile alla specificazione della struttura di correlazione tra le osservazioni.

Un primo studio, in questa direzione, è dovuta a Le Cessie e Van Houwelingen (1994) che hanno applicato la verosimiglianza a coppie per analizzare la mortalità infantile in uno studio *follow-up.* In questo particolare studio era di interesse analizzare la correlazione tra i parti gemellari, in quanto è ragionevole ipotizzare che le risposte siano correlate. In questo e in altri casi simili descritti in Varin (2007), si cerca una funzione bivariata sufficientemente "semplice" per i dati, da utilizzare nella (1.3). Si costruisce in questo modo una pseudo-verosimiglianza bivariata: la procedura più comune è quella in cui partendo dalle funzioni di densità marginale si costruisce una funzione congiunta tenendo conto che le due marginali sono dipendenti. Il metodo con cui si stimano i parametri può semplicemente essere basato su un algoritmo iterativo di massimizzazione e stima, metodo EM, oppure con il metodo di Newton-Raphson. La letteratura offre anche altre tecniche per tener conto della correlazione esistente tra coppie di dati. In molti casi (Zeger *et al.,* 1992) si ricorre all'uso di equazioni di stima generalizzate tra cui troviamo quelle di primo ordine (GEE1, Liang *et al.,* 1986) e di secondo ordine (GEE2, Zhao e Prentice, 1990). L'utilizzo di questo metodo per la stima dei parametri è computazionalmente semplice e di immediata comprensione. Sia le GEE che la SMVC forniscono stimatori consistenti e asintoticamente normali. Come descritto ampiamente da Molenberghs e Verbeke (2005), le equazioni di stima generalizzate GEE1 e GEE2 hanno caratteristiche molto simili alla funzione *score* di verosimiglianza.

Di seguito si andrà a esplicitare in modo rigoroso la verosimiglianza a coppie per dati dicotomici nel paragrafo 1.4.1, mentre un piccolo riassunto sui contenuti e proprietà delle GEE è presente nel paragrafo 1.4.2.

1.4.1 La verosimiglianza a coppie

Si consideri la seguente tipologia di dati binari con coppie di risposte del tipo (Y_i, Y_j) , con *i=1,...,n*, e *j=1,...,n*, con $j \neq i$, per un totale di $(n-1)^2$ coppie di osservazioni. Si assume che le osservazioni siano dipendenti all'interno della coppia, ma indipendenti fra differenti coppie di osservazioni. Le probabilità marginali $P(Y_i=1) e P(Y_j=1)$ sono indicate con $p_1 e p_2$, rispettivamente. La situazione è riassunta nella Tabella 1.1.

	$Y_j=1$	$Y_j=0$	
$Y_i=1$	p_{11}	p_{10}	p_1
$Y_i=0$	p_{01}	p_{00}	$1-p_1$
	p_2	$1-p_2$	1

Tabella 1.12 Tabella di probabilità della risposta (Y_i, Y_j) *in dati bivariati.*

Quando si è in presenza di una distribuzione univariata con variabile risposta binaria, il modello più utilizzato è il modello binomiale con legame funzione di legame probit.
Il GLM (modello lineare generalizzato) con funzione legame probit è una specificazione di un modello di regressione binaria che ha riscosso e riscuote una notevole popolarità (McCullagh e Nelder, 1989).
Sia *Y* una variabile dicotomica (ossia che assume soltanto i valori 0 e 1) e sia *X* una matrice di regressione. Il modello binomiale nel caso unidimensionale ipotizza che

$$Pr(Y=1 \mid X=x)=\Phi(\beta^T x), \tag{1.13}$$

dove *x* denota una riga della matrice di regressione X e $\Phi(.)$ è la funzione di ripartizione di una variabile casuale normale standard. Generalizzando la (1.13) per ogni risposta di *Y*, il vettore di parametri β viene usualmente stimato con il metodo della massima verosimiglianza. Partendo dalle funzioni di densità univariate per Y_i e Y_j , se queste variabili fossero indipendenti, la funzione di densità congiunta sarebbe il prodotto delle due marginali e di conseguenza la verosimiglianza composita risulterebbe come il prodotto delle

verosimiglianze marginali. Ma se le variabili sono dipendenti, si deve tener conto di questa dipendenza nel modello da specificare. In via preliminare si può assumere che la dipendenza tra due osservazioni possa essere modellata da un parametro ρ.
La funzione di verosimiglianza per questa tipologia di dati può quindi essere espressa come:

$$L(\beta, \gamma, \rho'; x) = \sum_{i=1}^{n} \sum_{j=1}^{n-1} \log \Phi_2(q_{1i} x_i^T \beta, q_{2j} x_j^T \gamma, \rho'), \tag{1.14}$$

dove

- $q_{1i} = \begin{Bmatrix} 1 \, se \, y_i \neq 0; \\ -1 \, altrimenti \end{Bmatrix}$,
- $q_{2j} = \begin{Bmatrix} 1 \, se \, y_j \neq 0; \\ -1 \, altrimenti \end{Bmatrix}$,
- β, γ sono parametri di regressione,
- $x_i e x_j$ sono le righe *i e j* della matrice di regressione X,
- $\rho' = q_1 q_2 \rho$,
- $Cov(\epsilon_i, \epsilon_j) = \rho$,

$\Phi_2(.)$ è la funzione di ripartizione di una normale bivariata e (ϵ_i, ϵ_j) sono gli errori tra i valori osservati e quelli predetti, utilizzando la stima di massima verosimiglianza a coppie di dati (SMVC) per la stima dei parametri del modello.

Per la specificazione del parametro ρ la letteratura offre vari approcci, ma i due che vengono principalmente usati sono:

1) **Correlazione tetracorica.**
   L'uso della correlazione tetracorica è stato introdotto da Pearson (1901) ed è basato sull'assunzione che per entrambe le variabili la distribuzione sia continua e normale. Infatti si suppone che le variabili risposta Y_i e Y_j, siano realizzazioni di una coppia di variabili latenti Z_i e Z_j con distribuzione normale standardizzata e correlazione ρ. In questo caso, il parametro ρ è chiamato correlazione tetracorica.

2) Odds ratio.

Un secondo metodo per analizzare l'associazione tra Y_i e Y_j è l'*odds ratio*. L'*odds ratio* indica il livello di associazione (negativa o positiva) ed è definito come

$$\rho = p_{00}\, p_{11} / p_{10}\, p_{01}\,. \tag{1.15}$$

Valori di $\rho>1$ denotano una associazione positiva, e valori di $\rho<1$ indicano una associazione negativa. Infine, il valore $\rho=1$ indica una situazione di indipendenza.

La misura di associazione proposta dall'*odds ratio* è migliore rispetto alla correlazione tetracorica, in quanto il suo valore varia tra 0 e infinito ed è di immediata interpretazione. Infatti, se il valore dell'*odds ratio* è superiore a 1 significa che vi è una associazione positiva tra le coppie di osservazioni; si ha un significato opposto se ha un valore inferiore a 1; mentre se il valore è pari a 1 significa che non vi è una sostanziale situazione di dipendenza tra i gruppi.

L'utilizzo della correlazione tetracorica o dell'*odds ratio* non è di fondamentale importanza nella (1.14), ma ci sono delle sostanziali differenze alla base della loro stima. La correlazione tetracorica, essendo compresa tra -1 e 1, ha delle difficoltà ad essere stimata e necessita molto spesso di una riparametrizzazione tra $-\infty$ e $+\infty$. Però uno dei vantaggi è che la correlazione tetracorica può essere estesa ad un arbitrario numero di osservazioni e di variabili sfruttando la distribuzione multinormale.

L'*odds ratio* è molto facile da calcolare, ma ci sono dei vincoli sulle caselle (vedi Tabella 1.12), in quanto si deve evitare di avere caselle senza osservazioni o con numerosità molto basse. Però l'*odds ratio* ha un dominio tra 0 e $+\infty$, per cui la sua stima è agevole con le routine di calcolo più comuni e non si rende necessaria una sua riparametrizzazione, ma soltanto una definizione della probabilità congiunta legata ad ogni cella della tabella.

1.4.2 Equazioni di stima generalizzate

Le equazioni di stima generalizzate proposte da Liang e Zeger (1986), sono state molto utilizzate negli ultimi anni. Nei GLM descritti da Nelder e Wedderburn (1972), l'associazione tra variabile risposta e variabili esplicative è data dalla funzione legame. Inoltre, i GLM assumono che le osservazioni siano indipendenti. Invece, nei modelli marginali introdotti da Zeger *et al.* (1985), l'interesse principale è modellare le marginali tenendo conto della struttura di dipendenza delle osservazioni: la correlazione tra le variabili risposta è stimata con un opportuno stimatore. L'obiettivo principale resta comunque analizzare gli effetti delle variabili esplicative sulla variabile risposta.

Equazioni Lineari Generalizzate di Primo ordine (GEE1)

Siano y_i i valori osservati sulla variabile risposta per *i=1, ..., n*. Per ogni y_i si dispone di un vettore di variabili esplicative x_i, il cui primo elemento è 1 per l'inclusione dell'intercetta nel modello considerato. Nei GLM le equazioni di stima dei parametri vengono ricavate partendo dalle equazioni di verosimiglianza e stimate per via iterativa. Questo procedimento rimane valido fino a quando le osservazioni non sono correlate, per cui la matrice di varianze-covarianze risulta diagonale. Per osservazioni correlate, la vera matrice di varianze-covarianze non ha una forma diagonale e per questo motivo Zeger *et al*. (1985) hanno proposto uno stimatore di tipo *sandwich,* più consistente per la stima della varianza, dato da

$$\hat{V}(\alpha) = A^{1/2}\hat{R}(\alpha)A^{1/2}, \tag{1.16}$$

dove $R(\alpha)$ è la matrice che definisce la struttura di dipendenza tra le osservazioni e α è un parametro ignoto di dipendenza che si deve stimare. Questa matrice viene comunemente detta *"working correlation matrix"*. Nella (1.16), *A* è una matrice diagonale ricavata direttamente dalla matrice di varianze-covarianze, e viene ottenuta calcolandone direttamente gli autovalori e inserendoli nella diagonale.

Il modo in cui viene costruita la matrice di correlazione $R(\alpha)$ rispecchia la natura di correlazione esistente tra i dati. Sono proposti vari metodi di stima di $R(\alpha)$; per approfondimenti si vedano Prentice (1988) e Liang e Zeger (1986). Una volta stimata la

varianza, l'equazione di stima generalizzata (GEE1) per i parametri di regressione β risulta

$$u(\beta)=\frac{1}{n}\sum_{i=1}^{n} D_i^T V_i^{-1}(y_i-\mu_i)=0, \tag{1.17}$$

dove $\mu_i=E(Y_i)$, D_i è il vettore delle derivate prime dell'equazione di stima rispetto a β e V_i è la matrice diagonale ricavata dalle matrice di varianze-covarianze precedentemente stimata. Sempre in Liang e Zeger (1986) sono presentati dei risultati di studi in cui si accerta che lo stimatore di β , basato sulla (1.17), risulta consistente e asintoticamente normale, sotto opportune condizioni di regolarità.

Equazioni Lineari Generalizzate di Secondo ordine (GEE2)

Le GEE2 sono molto simili alle GEE1, solo che si cerca la stima simultanea sia del parametro di correlazione α che del parametro di regressione β , ossia si risolve contemporaneamente il sistema di equazioni

$$u\begin{pmatrix}\alpha\\ \beta\end{pmatrix}=\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\frac{\delta\mu_i}{\delta\beta} & 0\\ 0 & \frac{\delta\sigma_i}{\delta\alpha}\end{pmatrix}\begin{pmatrix}V(y_i) & 0\\ 0 & V(z_i)\end{pmatrix}^{-1}\begin{pmatrix}y_i-\mu_i\\ z_i-\sigma_i\end{pmatrix}=0. \tag{1.18}$$

Nella (1.17), μ_i e σ_i sono i valori della media e della varianza delle osservazioni per riga, mentre y_i e z_i sono rispettivamente le osservazioni e la correlazione calcolata tra ogni coppia di osservazioni. Non ci addentriamo nei calcoli e trasformazioni di tale equazione di stima, ma per un approfondimento si rimanda a Zeger (1996). Come per le GEE1, anche con le GEE2 gli stimatori di β e α risultano consistenti e asintoticamente normali. L'uso delle GEE2 è consigliato se la struttura di correlazione all'interno dei dati è molto forte, il che induce ad una rapida convergenza dell'algoritmo e ad una migliore stima dei parametri.

1.5 Confronto tra la verosimiglianza a coppie e le GEE

La verosimiglianza a coppie è uno strumento molto flessibile in quanto non bisogna assumere nessuna formalizzazione sul modello utilizzato per la stima. Gli svantaggi a cui si vanno incontro con l'utilizzo della verosimiglianza a coppie sono soprattuto legati alla costruzione del modello bivariato da usare nel calcolo della verosimiglianza, in quanto molte volte succede che, costruendo la densità congiunta partendo dalle marginali, non si riesce a stimare in maniera opportuna la dipendenza tra le variabili e il modello porta pertanto a risultati non soddisfacenti. Un altro limite è a livello computazionale: per il calcolo della massima verosimiglianza nella verosimiglianza a coppie e la stima dei parametri di regressione, nonché di quello di correlazione, si usano algoritmi iterativi che vanno a massimizzare la verosimiglianza. Nel caso in cui però i parametri da stimare siano molti, ci si trova di fronte a tempi di calcolo elevati o a minimizzazioni locali e non globali. Questo problema si può in parte ridurre scegliendo modelli semplici e inizializzando l'algoritmo in vari punti in modo da andare a ricercare la soluzione ottima per la stima dei parametri.

Le GEE sono molto efficienti dal punto di vista computazionale, in quanto una volta definito il sistema di equazioni, la risoluzione e quindi la stima dei parametri avviene per via matriciale, senza andare incontro a massimizzazioni lunghe e laboriose. Inoltre, si è in grado di ricavare direttamente gli *standard error* delle stime, cosa che invece non avviene per le stime con la verosimiglianza a coppie, in quanto essi vengono ricavati con tecniche *Bootstrap* o *Jack-knife*.Il principale difetto delle GEE è che esse sono molto sensibili ad errate specificazioni del modello: per poter effettuare la stima dei parametri bisogna opportunamente specificare la matrice di correlazione, ma se la natura delle variabili è di difficile interpretazione, la *"working correlation matrix"* potrebbe portare a stime distorte sia nella correlazione che nei parametri di regressione.

Per questo motivo si tende a preferire la verosimiglianza a coppie quale strumento più flessibile e robusto, in grado di fornire conclusioni migliori in assenza di informazioni aggiuntive.

Nel prossimo capitolo andremo a verificare quali sono i passi che portano alla costruzione di una verosimiglianza a coppie, andando a mostrare quali sono le tecniche più usate e in che modo tale pseudo-verosimiglianza venga formalizzata e di poi implementata nel software statistico *R*.

Capitolo 2

Costruzione della verosimiglianza a coppie

Per la costruzione di una verosimiglianza a coppie della forma (1.3) si deve scegliere una opportuna funzione di densità bivariata in modo da tenere conto della struttura di dipendenza nei dati. A questo scopo si tende a utilizzare funzioni, tipo $f_{ij}(y_i, y_i, \theta)$ con $i,j=1,...,n$, semplici e modellabili su una vasta gamma di dati. I metodi principali per la costruzione di una verosimiglianza a coppie sono due: la prima tecnica è quella di partire da densità univariate e quindi di costruire la congiunta bivariata (cfr. Paragrafo 2.1), mentre la seconda è quella di partire da densità a due dimensioni e adattarla ai dati, per effettuare la stima dei parametri ricercati (cfr. Paragrafo 2.2). Nell'ultima parte di questo si discute, infine, su come creare una verosimiglianza a coppie per il dataset che sarà presentato nel Capitolo 3, partendo dalla teoria e applicandola alla costruzione di un programma in ambiente R.

2.1 La funzione Gauss-Copula

Cercare di modellare dati che sono correlati risulta complicato, in quanto la scelta, di una funzione che prova a tener conto della dipendenza, deve essere fatta con cura. Di conseguenza, è doveroso partire dalla ricerca di una funzione di probabilità che consideri la natura delle variabili in esame.

Di norma, si è facilmente in grado di specificare un modello univariato. Partendo da questo presupposto, un metodo che viene utilizzato per ricavare una funzione a due dimensioni partendo dalla funzione marginale è la funzione di "Gauss-Copula", le cui proprietà sono

ampiamente illustrate in Nelsen (1999). Il termine copula, che dai ricordi di analisi logica significa "parte di una proposizione che connette soggetto e predicato", è stato usato per la prima volta in Statistica nel 1959 da Abe Sklar, nel teorema che porta il suo nome, per designare la funzione che collega le distribuzioni marginali univariate a formare la relativa distribuzione multivariata.

La funzione Gauss-Copula, come tutte le copule, ha la proprietà che la distribuzione condizionata ed il valore atteso condizionato, sono esprimibili in termini della funzione copula. Nel seguito saranno descritti solo alcuni risultati fondamentali della teoria delle copule, rimandando per approfondimenti ai lavori di Genest e MacKay (1986), Joe (1997) e Nelsen (1999).

Una copula bivariata è una funzione $C: [0,1]\times[0,1] \rightarrow [0,1]$ le cui proprietà sono illustrate in Nelsen (1999). Per i nostri scopi, si richiama soltanto un teorema di Sklar (1959), che costituisce il risultato matematico fondamentale per l'applicazione delle copule a problemi di inferenza. Per il teorema di Sklar, date due variabili casuali X e Y con funzione di ripartizione congiunta F_{XY} e marginali F_X e F_Y , esiste una funzione copula C tale che, per ogni x e y con dominio su R, vi è:

$$F_{xy}(x, y)=C(F_X(x), F_Y(y)). \tag{2.1}$$

Se F_X e F_Y sono continue, allora la funzione C è unica; altrimenti C è unicamente determinata sul dominio Dom(F_X) x Dom(F_Y), dove Dom() denota il dominio delle marginali. In particolare, nella Gauss-Copula F_X e F_Y sono funzioni di ripartizione di una variabile normale, C è una funzione Copula opportunamente scelta e la funzione F_{XY} definita nella (2.1) è una funzione di ripartizione congiunta delle marginali F_X e F_Y.

Inoltre, si verifica che date due variabili casuali X e Y con distribuzione congiunta F_{XY}, marginali F_X e F_Y, e data una copula C, allora $U=F_X$ *e* $V=F_Y$ *s*ono variabili casuali uniformi standard con distribuzione congiunta $C(U,V)$. Generalmente i dati vengono modellati come illustrato nella Figura 2.2.

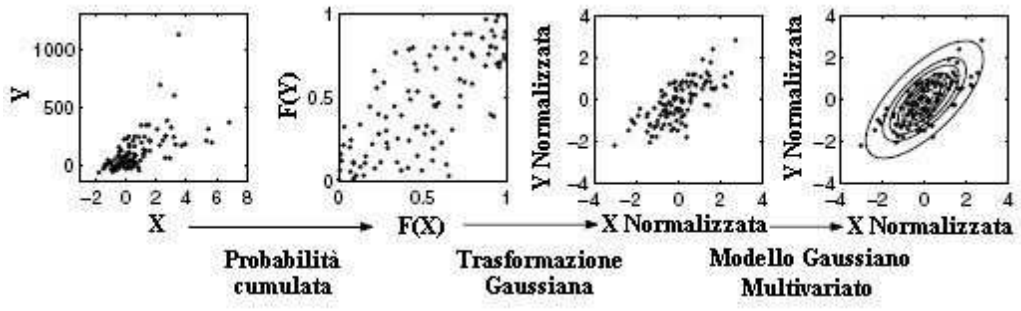

Figura 2.2 – Modellazione dei dati nel modello Gauss-Copula.

La funzione copula si presta bene nel caso in cui non è ben definito il parametro di interesse da andare a stimare o più semplicemente non è nota la natura delle variabili, per cui si cerca una funzione che comprenda una ampia classe di modelli. In tal modo non si va incontro ad una mal specificazione del modello che si andrà a stimare.

2.2 Verosimiglianza a coppie basata su una distribuzione bivariata nota

La costruzione di una verosimiglianza a coppie può avvenire partendo da una distribuzione di densità bivariata nota. Gli esempi in letteratura sono numerosi, in quanto è semplice partire da una distribuzione bivariata e andarla a modificare a seconda dei dati o dei parametri da stimare. Per alcune dimostrazioni ed esempi si possono ricordare le pubblicazioni di Dale (1986) e Le Cessie *et al*. (1994). Questo ultimo lavoro, già citato nel capitolo precedente, è risultato molto utile perché fornisce uno schema logico e i calcoli matematici alla base dello sviluppo della verosimiglianza a coppie. Lo studio effettuato da Le Cessie *et al.* (1994) riguarda lo studio di mortalità neonatale in parti gemellari, con lo scopo di modellare la probabilità di morte con una verosimiglianza a coppie che tenga conto della correlazione esistente tra i gemelli. Si tratta quindi di una analisi di dati binari correlati e gli autori usano una regressione logistica, andando a modellare l'associazione tra le risposte sia con la correlazione tetracorica sia con l'uso dell'*odds ratio*.

Usando la correlazione tetracorica si suppone che le due variabili risposta Y_i e Y_j, siano

realizzazioni di due variabili latenti Z_i e Z_j, che sono distribuite secondo una funzione normale bivariata con correlazione ρ. Di seguito, si suppone poi che la generica Y sia uguale a uno se $Z < g$,con g valore opportuno della distribuzione di ripartizione della normale.
La funzione di legame scelta è logistico e pertanto $p_i = (\exp(x_i\beta)/(1+\exp(x_i\beta)))$, con x_i i-esima riga della matrice di regressione X e β vettore dei parametri di regressione.
La probabilità congiunta p_{ij} prende la forma

$$p_{ij} = Pr(Y_i = 1, Y_j = 1) = \int_{-\infty}^{g}\int_{-\infty}^{g} \phi_2(t_i, t_j, \rho)\, dt_i\, dt_j, \qquad (2.3)$$

dove $\phi_2(.)$ indica la funzione di densità di una normale standard bivariata.
La stima dei parametri di regressione e associazione viene generalmente ottenuta con il metodo iterativo di Newton-Raphson. Il parametro di correlazione ρ varia tra -1 e 1, e può essere utile rimuovere questa limitazione andando a riparametrizzare ρ con $\rho' = \log((1+\rho)/(1-\rho))$.
Questo è un esempio di una funzione usata nella stima dei parametri in una verosimiglianza a coppie, ma generalmente si sceglie come funzione di partenza la funzione di densità più opportuna a seconda della tipologia di dati che si deve analizzare. Nel prossimo paragrafo andremo a costruire una verosimiglianza a coppie per i nostri dati andando a implementare un programma in linguaggio R per la stima dei parametri.

2.3 Costruzione della verosimiglianza a coppie per dati binari

Il primo punto che si deve affrontare per costruire una verosimiglianza a coppie è specificare cosa è di interesse andare a stimare con questa pseudo-verosimiglianza. Fondamentalmente, l'obiettivo principale consiste nello stimare i parametri di regressione, ma come già introdotto nel paragrafo precedente, ci si deve preoccupare di come trattare il parametro di correlazione.
In questo paragrafo ci si sofferma sull'analisi di dati binari, in cui per ogni realizzazione delle due variabili Y_i e Y_j si deve definire la probabilità di ciascun evento o realizzazione, come viene riportato nella Tabella 2.4. Come funzione di densità da cui partire si sceglie un modello binomiale bivariato con legame probit. La probabilità di ogni evento viene calcolata analogamente al caso univariato e è illustrata nella Tabelle 2.4.

Evento	Probabilità associata
(1,1)	$\int_{-x_i^T\beta}^{\infty}\int_{-x_j^T\beta}^{\infty}\phi(\mu,\nu;\alpha)\,d\mu\,d\nu$
(1,0)	$\int_{-x_i^T\beta}^{\infty}\int_{-\infty}^{-x_j^T\beta}\phi(\mu,\nu;\alpha)\,d\mu\,d\nu$
(0,1)	$\int_{-\infty}^{-x_i^T\beta}\int_{-x_j^T\beta}^{\infty}\phi(\mu,\nu;\alpha)\,d\mu\,d\nu$
(0,0)	$\int_{-\infty}^{-x_i^T\beta}\int_{-\infty}^{-x_j^T\beta}\phi(\mu,\nu;\alpha)\,d\mu\,d\nu$

Tabella 2.4 – Probabilità associata ad ogni evento della coppia Y_i e Y_j.

Nella Tabella 2.4, x_i e x_j indicano i due vettori della matrice di regressione X, $\Phi(.)$ è la funzione di densità della variabile normale standardizzata, mentre α è il parametro che indica la correlazione tra le coppie di variabili supponendo, per semplicità computazionale ed espositiva, che essa rimanga costante tra tutte le coppie del modello considerato. Come avviene nella verosimiglianza genuina, si calcola con questo metodo ogni contributo dato dalle coppie di osservazioni. Per questo motivo risulta utile lavorare sulla log-verosimiglianza a coppie, perché da la possibilità di sommare ogni singolo contributo alla verosimiglianza,

dato da ogni coppia. Si ottiene che la log-verosimiglianza a coppie per dati binari dipendenti, partendo da un modello binomiale bivariato con funzione legame probit, risulta

$$l_{PL}(\beta, \alpha; y_1, y_2) = \sum_{i=1}^{n} I(Y_{i,1} = \{0,1\}, Y_{i,2} = \{0,1\}) \log Prob(Y_{i,1} = \{0,1\}, Y_{i,2} = \{0,1\}), \quad (2.5)$$

dove, da come si può notare, vengono sommati il logaritmi delle probabilità delle coppie effettivamente osservate, in cui la variabile indicatrice *I(.)* definisce l'evento osservato, mentre le probabilità degli eventi sono quelle descritte nella Tabella 2.4.
Il programma, in linguaggio R, per il calcolo della verosimiglianza a coppie per dati binari è presentato nell'Appendice, al punto 3.
E' stato scritto un programma, utilizzando la funzione *optim(.)* di R, per calcolare le stime dei parametri di regressione e del parametro di correlazione. Il parametro α essendo una correlazione, non può assumere tutti i valori dello spazio reale, ma solo valori compresi tra -1 e 1. Si rende quindi necessaria una riparametrizzazione, per consentire a questo nuovo parametro di prendere qualsiasi valore nella retta reale, senza nessuna costrizione. La riparametrizzazione più usata è

$$\alpha^* = \log((\alpha - 1)/(1 - \alpha)). \quad (2.6)$$

Nel prossimo capitolo si presenta il dataset che si è analizzato ricorrendo alla verosimiglianza a coppie, andando a effettuare delle analisi esplorative preliminari e l'imputazione dei dati mancanti.

Capitolo 3

Un caso studio: la Talassemia in Sardegna

In questo capitolo si introducono le motivazioni che portano allo studio della popolazione sarda e, più in particolare, ci si concentra su dei dati che riguardano la talassemia, malattia ematica diffusa in molte aree della Sardegna. In collaborazione con il centro di ricerca CRS4, abbiamo analizzato un dataset di natura genetica sulla affezione da talassemia in un piccolo paese di quest'isola.

Lo studio preliminare avviene andando a verificare la struttura dei dati e poi andando a imputare, grazie a tecniche ad hoc, i dati mancanti presenti. Il capitolo si chiude con la stima di alcuni modelli binomiali basati, tuttavia, sull'ipotesi di indipendenza, per cercare di modellare la probabilità di malattia degli individui. Tali modelli marginali sono usati soltanto per la ricerca delle variabili più significative, che vengono impiegate successivamente, nel Capitolo 4, come base di partenza per l'analisi con l'utilizzo della verosimiglianza a coppie, schematizzata nel paragrafo 2.3.

3.1 Lo studio della popolazione sarda

In Sardegna è presente un'intensa attività di ricerca sulla genetica che coinvolge la popolazione, che è caratterizzata da un patrimonio genetico omogeneo, e che riguarda sostanzialmente due progetti:

- Progetto "ProgeNIA" – Studio della popolazione sarda, per la sua omogeneità, per lo studio dei tratti fenotipici legati all'invecchiamento, e a malattie complesse;
- Progetto "AKeA" - Studio dei marcatori della salute e della longevità dei Sardi.

Per eventuali approfondimenti in merito a questi progetti rimandiamo ai siti internet:

- http://it.wikipedia.org/wiki/Sardegna;
- http://www.crs4.it.

Studiare le malattie presenti in una popolazione, come quella della Sardegna, consiste nell'analisi dei geni caratteristici degli attuali abitanti dell'isola con due principali obiettivi. Il primo ha uno scopo prettamente biologico e antropologico ed è quello di ricostruire la storia naturale della popolazione. Essa consiste nella comprensione dell'entità, dei tempi e delle modalità della fondazione, unitamente alle successive o concomitanti dinamiche demografiche e evolutive. L'altro è, invece, applicativo ed ha la finalità di comprendere le cause genetiche di alcune patologie sfruttando alcune peculiarità della popolazione sarda, che la rendono di elezione per studi che prevedono l'utilizzo di isolati genetici. Entrambi gli obiettivi sono perseguiti attraverso lo studio molecolare di marcatori del DNA di individui della popolazione, mediante un approccio multidisciplinare che coinvolge biologi, medici, naturalisti, statistici, bioinformatici, biotecnologi, archeologi, antropologi e paleontologi.
L'interpretazione della variabilità genetica fa ritenere la popolazione sarda derivante da un gruppo di genti arrivate in Sardegna attraverso varie migrazioni nel Paleolitico superiore (14.000 anni fa).
L'antichità della fondazione, l'isolamento millenario e le difficili condizioni ambientali - ad esempio, la malaria - hanno generato nel tempo particolari caratteristiche antropologiche e genetiche. Per queste ragioni, i Sardi si differenziano non solo dagli altri europei, ma anche dai vari gruppi mediterranei.

I geni dei sardi si inquadrano perfettamente con le caratteristiche della popolazione europea con grosse differenze però in termini di:

- frequenze geniche (per lo più dovute a effetto del fondatore e deriva genetica casuale);
- presenza di sottotipi sardo-specifici (da imputarsi a mutazioni occorse nell'isola, dato il lungo tempo intercorso dalla fondazione ad oggi).

L'elevata variabilità genetica implica un numero rilevante di linee fondatrici. L'archeologia indica che la taglia effettiva della popolazione sarda sia stata molto importante relativamente alle altre aree geografiche coeve. Le grandi crisi demografiche medievali, infatti, non hanno potuto cancellare la struttura della popolazione, come è successo ad esempio nella vicina Corsica.
Le frequenze di alcuni geni sono state influenzate dalla presenza della selezione operata dal plasmodium, l'agente infettivo della malaria che, durante il suo ciclo vitale, parassitizza i globuli rossi del sangue dell'uomo attraverso le ghiandole salivari della zanzara.
La selezione ha agito aumentando la frequenza di geni che possono causare l'insorgenza di varie tipologie di talassemie, o del favismo, attraverso un processo noto come polimorfismo bilanciato.

3.2 Analisi del dataset “Talassemia” in formato pedigree

In questo paragrafo si descrive un particolare formato di dati, il formato Pedigree, molto in uso sia per la facilità di codificare i dati sia per l'utilizzo in molti programmi di analisi di dati. Per ulteriori informazioni in merito a questa tipologia di file e analisi dimostrative effettuate,si invita a visitare il sito internet del software statistico *Merlin*

http://www.sph.umich.edu/csg/abecasis/Merlin/tour/input_files.html,

uno dei più usati per analizzare file di questo tipo.

In breve, si può riassumere che il dataset in formato Pedigree ha i dati divisi con spazi bianchi, o tabulazioni, e le prime sei colonne sono così composte:

- Id della famiglia;
- Id della persona;
- Id del padre;
- Id della madre;
- Sesso (1=maschio; 2=femmina);
- Fenotipo,

dove gli “Id” sono dei numeri alfanumerico che identificano l'individuo e la sua famiglia nel dataset. Una limitazione del file di tipo Pedigree è che deve avere un fenotipo nella sesta colonna. Il fenotipo può essere una caratteristica quantitativa o una colonna di condizione di affezione di tipo qualitativo. Le altre colonne successive alla sesta, presenti nel dataset, si riferisco ai dati misurati dalle analisi genetiche sulle coppie di alleli.
In genetica, per allele si intende ogni forma vitale di DNA codificante per lo stesso gene: in altre parole, l'allele è responsabile della particolare modalità con cui si manifesta il carattere ereditario controllato da quel gene. In ogni individuo abbiamo un allele ereditato dal padre e uno dalla madre. Per cui ogni coppia di alleli deve essere presa insieme come replicazione dello stesso dato. Lo stato di affezione viene così codificato:

- 0 dato mancante;
- 1 Id non affetto o allele non attivo;
- 2 Id affetto o allele attivo.

Come si può capire la struttura dei dataset in formato Pedigree è semplice ed intuitivo. Nel prossimo paragrafo si presenta il lavoro di “pulizia” mirato all'imputazione di dati mancanti, purtroppo presenti nel dataset in questione, necessario per procedere nell'analisi dei dati.

3.3 Imputazioni di dati mancati e ricodifiche

Ad una prima analisi si è notato che nel dataset sono presenti molti dati mancanti, sia parzialmente che totalmente mancati, ossia intere righe di dati senza osservazioni.

La letteratura propone vari metodi per trattare dati incompleti, e l'applicazione dell'uno o dell'altro metodo dipende dal meccanismo generatore dei dati mancanti e dal tipo di analisi che si vuole condurre sul dataset completo. Nel dataset ci si trova di fronte sia ad unità statistiche con dati parzialmente mancanti, sia ad unità statistiche con dati totalmente incompleti. In questo ultimo caso, l'approccio è stato quello di eliminare gli individui con osservazioni totalmente mancanti, in quanto non si avevano a disposizione le caratteristiche e informazioni sufficienti per utilizzare qualche tecnica di imputazione per stimare l'intera riga mancante.

L'imputazione dei dati parzialmente mancanti è stata invece effettuata con una tecnica chiamata *hot deck* (Little *et al.,* 1987): questa tecnica sostituisce il valore mancante con un altro proveniente da un altro individuo, preso in maniera opportuna. L'aggettivo *hot* si riferisce al fatto che i valori imputati sono presi dall'indagine corrente, in contrapposizione al termine *cold* dell'imputazione *cold deck,* in cui i valori sono tratti da indagini precedenti. La definizione di *hot deck* che si presenta è abbastanza generale in quanto in letteratura, non è presente una definizione precisa comunemente accettata. Infatti, a questo proposito, bisogna sottolineare il fatto che non esiste una consolidata teoria in merito, ma l'applicazione di questa tecnica è dettata dal buon senso, più che da un rigoroso lavoro teorico. Il codice in linguaggio R utilizzato in questo dataset è consultabile nella Sezione 4 dell'Appendice.

L'*hot deck* si basa sull'individuazione del valore da imputare

$$\hat{y}_{is} = \hat{y}_i + \hat{e}_i , \tag{3.1}$$

dove $\hat{y}_{is}$ è il valore imputato mediante tecnica stocastica, determinato aggiungendo un residuo $\hat{e}_i$ al valore imputato con tecnica deterministica, e $\hat{e}_i$ è tale che il suo valore medio è nullo, e pertanto $E(\hat{y}_{is}) = \hat{y}_i$.

Questi procedimenti sono molto usati nella pratica campionaria e possono assumere schemi molto elaborati per selezionare le unità di imputazione. La caratteristica comune a tutte le procedure *hot deck* è quella di selezionare un donatore che abbia caratteristiche simili a quelle del non rispondente, allo scopo di ridurre la distorsione causata dalla non risposta. In questo caso il donatore viene scelto in modo da risultare più simile all'individuo con il dato mancante. La tecnica usata è quella di calcolare la matrice delle distanze tra i vari individui e di cercare i soggetti possibili donatori tra quelli con la distanza minore. Nel dataset "Talassemia" si è potuto notare che, in molti casi, l'individuo donatore era un genitore dell'individuo con dato mancante, questo perché, come si può immaginare, i geni tra persone strettamente imparentate sono simili e la distanza tra i due profili risulta piccola.

Bisogna tener presente che l'effetto delle procedure *hot deck* è quello di produrre un aumento della varianza dello stimatore della media, rispetto all'imputazione deterministica. Infatti, la duplicazione delle osservazioni secondo un meccanismo casuale introduce un ulteriore elemento di variabilità nel processo di stima, per cui la varianza della stime calcolate su valori imputati con criterio stocastico risulta sistematicamente maggiore a quella basata sull'imputazione di valori medi. Inoltre, un ulteriore svantaggio risulta dal fatto che l'imputazione deterministica distorce la distribuzione della variabile di interesse attenuando la varianza del carattere tra gli elementi del campione: l'assegnazione del valore medio dei rispondenti a ciascun valore mancante (complessivamente o all'interno di ciascuna classe di aggiustamento) aumenta artificialmente la densità delle osservazioni in corrispondenza del valore medio, mentre la scelta casuale di un valore da imputare conserva le caratteristiche della distribuzione, nell'ipotesi in cui i valori imputati si distribuiscano in maniera analoga a quelli osservati.

Successivamente all'imputazione dei dati mancanti, nell'analisi del dataset "Talassemia" si è resa necessaria la ricodifica di ogni coppia di alleli, in modo da semplificare l'informazione in essi contenuta, infatti, se gli alleli di una coppia hanno valori diversi risulta intuitivo capire che è lo stesso sia se ci troviamo di fronte alla coppia (1,2) che alla coppia (2,1), in quanto un allele proviene dalla madre e uno dal padre. Per questo motivo le possibili combinazioni della coppia di alleli sono entrambi uno, entrambi due o diversi.

Nella Tabella 3.2 viene riportato uno schema esemplificativo della ricodificazione effettuata.

Allele 1	Allele 2	Ricodifica
1	1	UNO
1	2	DIV
2	1	DIV
2	2	DUE

Tabella 3.2 – Ricodifica degli alleli

Alla fine, dopo avere eseguito l'imputazione dei dati mancanti e la ricodifica delle variabili delle coppie di alleli, si è ottenuto un dataset formato da 554 persone e da 1408 variabili. A questo punto, il nostro obiettivo è ricondurci ad una situazione con un minor numero di variabili, in quanto 1408 variabili sono troppe per poter fare ulteriori analisi o stimare uno eventuale modello. Questo è il principale problema che si analizzerà e che si cercherà di risolvere nel prossimo Paragrafo.

3.4 Analisi preliminare del dataset

Con il dataset pulito e ricodificato, si è in presenza di 252 maschi e 302 persone di sesso femminile, mentre la presenza della talassemia, riscontrata in 88 persone, si distribuisce in modo indipendente tra i 2 sessi, in quanto il sesso non è un fattore di rischio per la malattia in esame, come mostrato nella Tabella 3.3.

	Non malati	Malati	
Maschi	212	40	252
Femmine	254	48	302
	466	88	554

Tabella 3.3 – Distribuzione della malattia nei 2 sessi.

Anche il test Chi-quadro di Pearson applicato alla Tabella 3.3 ci suggerisce che il sesso non è un fattore di rischio per la malattia (*p-value* = 0.0121).

Con complessivamente 1408 variabili, non si è in grado di effettuare le analisi descrittive classiche e quindi, in questa fase si desidera cercare di usare qualche tecnica per trovare quegli alleli che siano correlati con la presenza della malattia nell'individuo. Per poter studiare un modello computazionalmente semplice, in particolare, si cerca un numero accettabile di regressori, ad esempio 20/30 alleli, che discriminano i malati da quelli non affetti dalla patologia. Complessivamente si hanno solo 554 osservazioni e per avere stime significative dobbiamo limitare il numero delle variabili, anche per poter dare un'interpretazione all'eventuale modello che si potrà stimare con esse.

Per fare questa selezione tra gli alleli, si possono usare molte tecniche, ma le caratteristiche del dataset hanno imposto alcune scelte su come effettuare l'analisi, in quanto si è in presenza di più variabili rispetto al numero di osservazioni e quindi qualsiasi tecnica di regressione non otterrebbe risultati soddisfacenti dal punto di vista teorico, e neanche da quello computazionale.

Come si può immaginare, non si possono studiare situazioni di dipendenza tra le variabili, e quindi si è cercato di quantificare quanto ogni variabile discriminasse tra malati e non malati. Una possibile soluzione poteva essere quella di studiare le tabelle di contingenza tra la variabile fenotipo con ogni variabile del dataset, ma nelle celle di frequenza si trovavano molto spesso numerosità piccole o addiritura uguali a zero. Pertanto non si poteva continuare con questa analisi, essendo il test di Pearson associato alla tabella non corretto.

Si è allora stimato un modello binomiale con una funzione di legame probit e con variabile risposta presenza/assenza della malattia, e si è usato come regressore ogni allele presente nel dataset. Così si sono stimati 1402 modelli binomiali, con l'obiettivo di verificare quali alleli risultassero significativi e comportavano, di conseguenza, una riduzione in devianza per ogni modello marginale stimato.

Come riferimento sulla bontà dei modelli stimati si può utilizzare l'indice AIC (*Akaike Information Criterion*, Akaike, 1973) o l'indice BIC *(Bayesian Information* Criterion, Schwarz, 1978). L'indice AIC è dato da

$$\mathrm{AIC} = 2\mathrm{k} - 2\log(\hat{L}), \tag{3.4}$$

dove *k* è il numero di parametri del modello stimato e $\hat{L}$ è la verosimiglianza massimizzata del modello stesso. L'indice BIC è, invece, dato da:

$$\mathrm{BIC} = k\log(n) - 2\log(\hat{L}), \tag{3.5}$$

dove *n* è il numero delle osservazioni prese in esame per la stima del modello. L'indice BIC effettua una maggiore selezione rispetto all'AIC, in quanto penalizza un modello con un numero elevato di parametri. Nel nostro caso abbiamo una numerosità campionaria costante e l'unico elemento di differenza tra i modelli è il numero di livelli presenti nel fattore nel modello. Infatti, ci sono alleli che presentano tutti e tre i livelli del fattore e altri che ne hanno soltanto due. Per questo motivo si è preso come riferimento il valore ottenuto calcolando l'indice BIC. Viene riportato nella Figura 3.6 un grafico di dispersione dei valori del BIC dei modelli stimati per ogni variabile del dataset in questione.

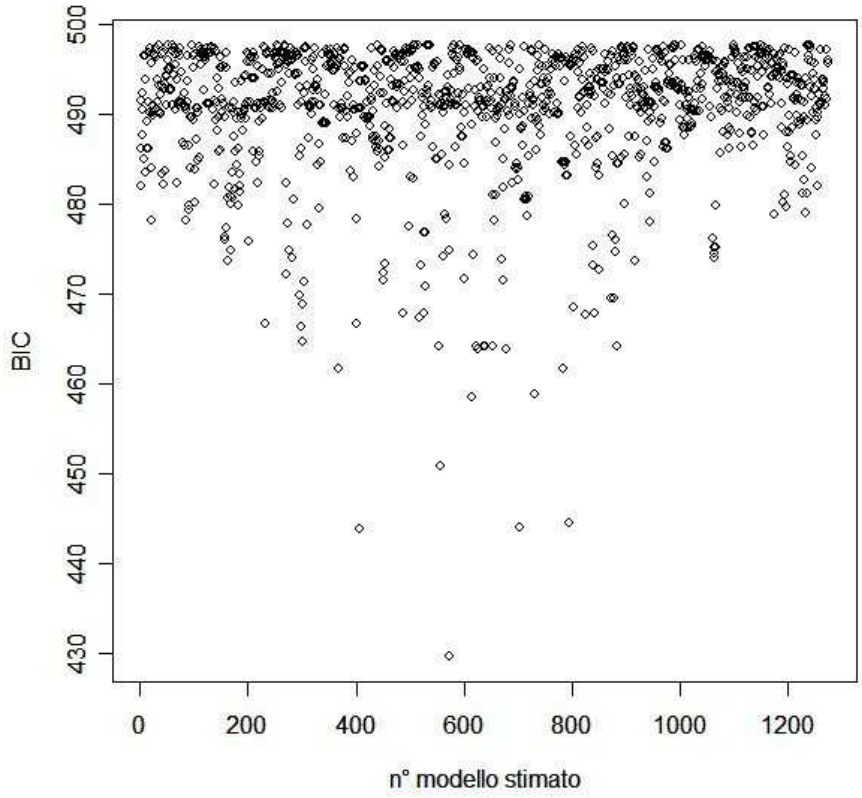

Figura 3.6 – Valori di BIC ottenuti dal modello.

Come si può vedere dalla Figura 3.6, ci sono degli alleli che comportano una maggiore riduzione del BIC e molti altri alleli che comportano una minore riduzione in devianza. Per questo motivo è opportuno scegliere un valore di soglia per selezionare gli alleli migliori con i quali proseguire l'analisi e l'eventuale stima di un modello predittivo.
Dal grafico si può controllare i valori del BIC ottenuti in alcuni modelli e si sceglie di impostare il limite di selezione delle variabili, al valore di 480, a cui corrisponde la selezione di 80 variabili che risultano marginalmente significative.
Bisogna sottolineare che questo approccio esclude completamente la dipendenza latente tra le variabili. Infatti, essendo le risposte correlate, gli errori standard associati alle stime vengono sovrastimati, in virtù del fatto che non viene utilizzata la matrice di Godambe per il calcolo degli *standard error.* Nel paragrafo successivo si valuteranno quali variabili selezionare per la stima di un modello binomiale predittivo.

3.5 Costruzione di un modello predittivo

Inizialmente si procede con la stima di un modello binomiale con legame probit con tutte le variabili significative trovate al punto precedente. Con procedura *stepwise,* si sono eliminate le variabili che comportano una minore riduzione in devianza confrontando sempre i due modelli annidati con un'analisi della devianza (*anova*). La metodologia di selezione del modello, chiamata *backward*, comporta la stima successiva di modelli annidati, togliendo una variabile alla volta. La scelta della variabile da eliminare viene valutata con l'indice BIC. Una volta ultimata questa procedura si sono ottenuti 22 alleli significativi. Il modello stimato sembra avere una buona riduzione della devianza rispetto al modello con la sola intercetta e una discreta capacità di previsione. I risultati ottenuti per la stima di questo modello con l'ausilio del software statistico R sono riportati nella Sezione 1 dell'Appendice.

Previsione	Non malati effettivi	Malati effettivi
Non malati previsti	453	12
Malati previsti	13	76

Tabella 3.7 – Tabella di confusione per il modello completo.

La tabella di confusione (Azzalini, 2004) riportata in Tabella 3.7 fornisce una indicazione della buona capacità di prevedere la malattia del modello stimato.

La percentuale di falsi positivi è di 2.58%, mentre i falsi negativi sono il 15,78%.

Un ulteriore metodo per analizzare la capacità predittiva di un modello binomiale è l'utilizzo della curva ROC (*Receiver Operating Characteristic* o *Relative Operating Characteristic*). Si tratta di una metodologia sviluppata per la prima volta durante la II Guerra mondiale per l'analisi delle immagini radar e lo studio del rapporto segnale/disturbo. Essa venne ben presto applicata in altri campi della tecnica e, a partire dagli anni '70, anche in campo medico (Lusted, 1971), inizialmente allo scopo di quantificare l'attendibilità dei responsi di immagini radiografiche interpretate da operatori diversi (Goodenough *et al.*, 1974, Hanley e McNeil, 1982). In tempi più recenti, l'utilizzo delle curve ROC si è fatto relativamente comune per la valutazione non solo delle immagini, ma anche dei più svariati test sia nel settore medico (con particolare riguardo alla valutazione dei test clinici di laboratorio) (Erdrich, 1981, Henderson, 1993), che, in minor misura, in quello veterinario (Greineret *et al.*, 2000).

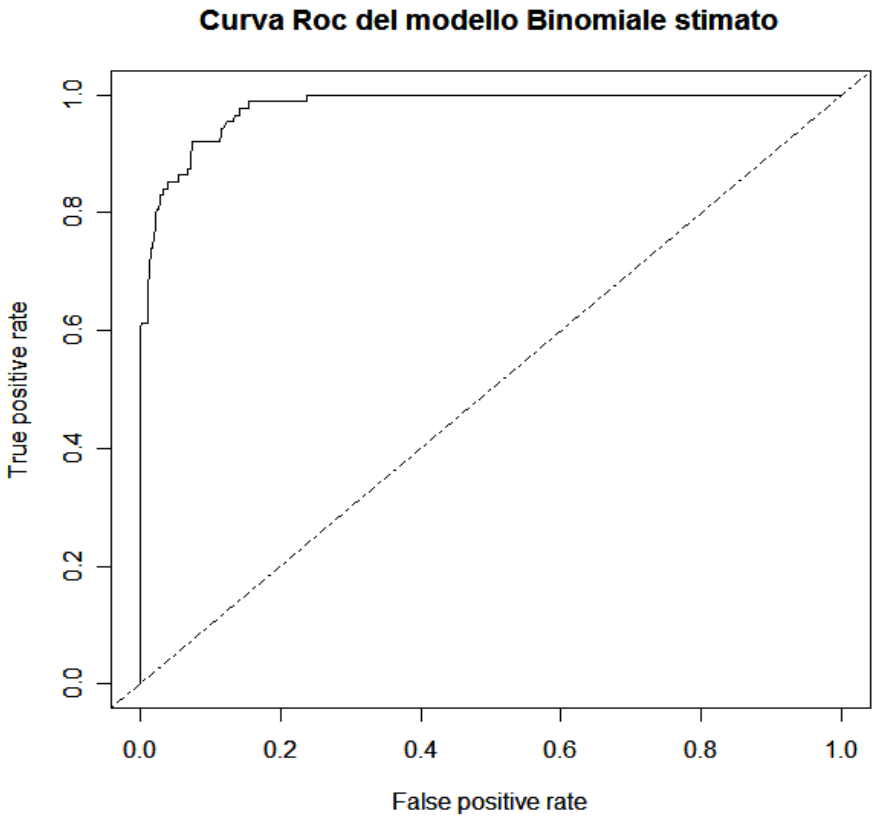

Figura 3.8 – Curva ROC per il modello completo.

L'analisi ROC viene effettuata attraverso lo studio della funzione che – in un test quantitativo – lega la probabilità di ottenere un risultato vero-positivo nella classe dei malati-veri (ossia la "sensibilità") alla probabilità di ottenere un risultato falso-positivo nella classe dei non-malati (ossia uno meno la "specificità"). In altre parole, vengono studiati i rapporti fra i falsi positivi e i falsi negativi. Come si può vedere dalla curva Roc del modello binomiale stimato nella Figura 3.8, esso apporta un significativo miglioramento rispetto alla situazione casuale rappresentata dalla linea diagonale.

Tutte le analisi compiute per stimare il modello binomiale sono state effettuate non considerando la dipendenza tra le variabili. Quindi chiaramente esso non è ben specificato, e come tale, le stime posso distorte e gli *standard error* sottostimati.

La selezione delle variabili è avvenuto scegliendo come valore di soglia 0,10 nel test di significatività marginale. Nel prossimo paragrafo si stima un modello ridotto, abbassando tale soglia di inclusione delle variabili a 0,001 e continuando con una procedura stepwise per selezionare, in questo modo, un piccolo numero di variabili.

3.6 Stima di un modello ridotto

Si è tentato di semplificare ulteriormente il modello stimato nel paragrafo precedente, prendendo ancora come riferimento l'indice BIC e le analisi *anova* sui fattori. Diminuire il numero di variabili nel modello è utile sia ai fini interpretativi, sia per avere pochi parametri da stimare e quindi abbreviare il tempo computazionale della loro stima. Infatti, come si andrà a vedere nel capitolo successivo, l'algoritmo di stima utilizzato è oneroso in termini di tempo e pertanto si deve cercare di ridurre al massimo in numero di variabili e, in tal modo, abbreviare il tempo per la loro stima.

Il modello ridotto è riportato nella Sezione 2 dell'Appendice, e presenta sette variabili. Rispetto al modello precedente è aumentata la devianza residua e gli indici AIC e BIC, ma l'aumento è stato contenuto considerando che sono state eliminate più della metà delle variabili. Infatti, per la selezione delle variabili, si è fissato un valore di significatività nel modello pari a 0.001, quindi molto basso. La capacità di previsione del modello resta buona, come si può verificare dalla tabella di confusione riportata in 3.9.

Previsione	Non malati effettivi	Malati effettivi
Non malati previsti	452	28
Malati previsti	14	60

Tabella 3.9 – Tabella di confusione per il modello completo

La percentuale di falsi positivi è di 3.04%, mentre i falsi negativi sono il 31.82%.
Dunque il modello riesce a prevedere bene i non malati, ma il potere discriminante tra le persone effettivamente affette da talassemia è considerevolmente più basso rispetto al modello precedente, in quanto ben 16 soggetti ammalati sono collocati tra le persone non affette da malattia. Però il fatto di avere meno variabili in questo modello consente una migliore interpretazione e una maggiore flessibilità.
Nel capitolo successivo si analizzerà il dataset prendendo coppie di dati studiando in particolare modo i parenti di primo e secondo grado. Successivamente si andrà a stimare di un modello di regressione ricorrendo alla verosimiglianza a coppie.

Capitolo 4

Stima con la verosimiglianza a coppie

Lo scopo del Capitolo 4 è quello di utilizzare la verosimiglianza a coppie per il calcolo di un modello di regressione. Infatti, il modello del capitolo precedente, anche se ha prodotto risultati rilevanti, non è ben specificato, in quanto non si tiene in considerazione la dipendenza tra le osservazioni. In quest'ultima fase si descrivono quali passi sono stati compiuti per la stima dei parametri di regressione andando a calcolare con opportune tecniche i valori degli *standard error* associati alle stime, facendo un confronto con quanto ottenuto nel capitolo precedente.

4.1 Analisi preliminare sulle coppie

Dopo aver terminato l'analisi esplorativa, si è potuto notare che ci sono diverse coppie di alleli che discriminano in modo forte tra malati e non malati. Però i modelli calcolati nel capitolo precedente non tengono conto della dipendenza tra le risposte e di conseguenza i modelli, oltre a non essere correttamente specificati, contengono stime distorte dei parametri e gli *standard error* vengono in genere sovrastimati. Come ampiamente discusso nel Capitolo 2, si cerca quindi di prendere le osservazioni a coppie, o più semplicemente a gruppi, in modo che questi risultano indipendenti tra loro e in modo da poter stimare un parametro di dipendenza all'interno dei gruppi stessi. Del dataset in questione siamo al corrente della struttura parentale e, quindi, possiamo andare a verificare che tipo di dipendenza sussiste tra soggetti appartenenti alla stessa famiglia. Nei dati sono presenti 492 coppie di osservazioni con parentela di primo grado, quindi legami madre/padre e figlio, e 775 coppie di osservazioni di

secondo grado in cui teniamo conto di tutte le parentele in cui ci sono 2 salti generazionali, quindi per esempio nonno con nipote, fratello e sorella e zio e nipote. Inizialmente sono state calcolate ben 2540 coppie di parenti di secondo grado, ma la maggior parte di queste appartenevano a osservazioni con dati totalmente mancanti e quindi sono state escluse dall'analisi.

Come si può verificare dalla Tabella 4.1, sembra sussistere una certo tipo di associazione nella presenza della malattia tra parenti di primo grado: infatti, facendo un semplice test di indipendenza di Pearson per tabelle di contingenza, si ottiene un valore della statistica test pari a 31.3357, la cui significatività associata, nella distribuzione di riferimento χ^2_1 , è pari a zero. Anche il valore dell'*odds ratio*, calcolato con le numerosità delle celle nella Tabella 4.1, ha un valore pari a 4.33, quindi, essendo di molto superiore a 1, esprime una forte associazione positiva tra la presenza della malattia nei figli e nei genitori.

	Genitore sano	Genitore malato
Figlio/a sano/a	354	60
Figlio/a malato/a	45	33

Tabella 4.1 – Distribuzione della malattia tra genitori e figli.

Nel caso in cui si considerano coppie formate da parenti di secondo grado, si ha una distribuzione della malattia tra le osservazioni come quella presente nella Tabella 4.2.

	Parente 2° sano	Parente 2° malato
Nipote o Fratello sano	535	102
Nipote o Fratello malato	98	41

Tabella 4.2 – Distribuzione della malattia tra parenti di secondo grado.

Ad una prima visione anche qui vi è una struttura di associazione positiva tra i parenti di secondo grado. Sia il test di Pearson (*p-value* pari a 0.001) che l'odds ratio (2.19) testimoniano che anche in questo caso vi è una associazione, seppur più debole, tra i parenti di secondo grado sulla presenza/assenza della malattia.

4.2 Stima di modelli utilizzando la verosimiglianza a coppie

Considerando solo le coppie appartenenti alla parentela di primo grado, si procede alla stima un modello di regressione utilizzando la verosimiglianza a coppie descritta nel Paragrafo 2.3. Le variabili di regressione sono quelle selezionate nel modello ridotto del Paragrafo 3.6, e quindi il programma implementato in R deve stimare 14 parametri, cioè 12 parametri di regressione, l'intercetta e il parametro di dipendenza. Il programma utilizzato è presentato nella sezione 3 dell'Appendice. In prima battuta si decide di inizializzare la funzione *optim(.)* da diversi punti per verificare come si comporta il processo di stima. La stima presenta delle difficoltà computazionali in quanto la funzione di verosimiglianza è irregolare e c'è la presenza di massimi locali. Per questo motivo si decide di far partire il processo di stima partendo dai valori dei parametri ottenuti dal modello marginale, vedi sezione 2 dell'Appendice, e con valore di correlazione standard pari a 0.63, valore normalizzato dell'*odds ratio* nella Tabella 4.1. La funzione *optim(.)* di R usata per la stima restituisce i valori dei parametri, il valore della funzione massimizzata e indica il numero di iterazioni eseguite dall'algoritmo in modo che la stima converga. Inoltre la funzione *optim(.)* può utilizzare diversi algoritmi specializzati alla massimizzazione; per ulteriori informazioni ed esempi pratici rimandiamo all'esauriente *help* in linea di R.

Stabiliti i dati iniziali con cui partire con la massimizzazione della log-verosimiglianza a coppie, si procede con l'algoritmo di ottimizzazione in modo da ricercare i nuovi valori dei parametri del modello, utilizzando in questo caso, solo le 492 coppie di osservazioni dei parenti di primo grado della Tabella 4.1.

	Modello Glm Marginale	Modello *Pairwise* di 1° grado
Intercetta	5.4983	5.5681
Allele 177-uno	-1.1417	-0.9697
Allele 177-due	1.9467	2.0109
Allele 248-uno	-3.8891	-3.8323
Allele 303-uno	0.8287	0.6660
Allele 303-due	0.2492	0.2967
Allele 329-uno	0.0097	0.0384
Allele 329-due	1.1816	0.6027
Allele 501-uno	0.8382	1.0827
Allele 501-due	-1.4111	-1.1797
Allele 639-due	-2.4016	-2.4428
Allele 679-uno	0.8138	0.8615
Allele 679-due	-1.0092	-1.2468
Correlazione	---	0.9314

Tabella 4.3 – Stima del modello marginale e utilizzando la verosimiglianza a coppie.

Nella Tabella 4.3 sono confrontati i valori dei parametri del modello marginale con quello proposto dalla verosimiglianza a coppie. Non ci sono forti differenze tra i valori osservati. La funzione *optim(.)* di R restituisce il valore di massima verosimiglianza -224,12 e la correlazione tetracorica è pari a 0.9314 che, se normalizzato tra -1 e 1, risulta pari a 0.4347, quindi discretamente positiva. Ora si può formulare un ulteriore modello per tenere conto della parentela introducendo anche le coppie di secondo grado. Il parametro di correlazione nel modello viene così espresso

$$\alpha^* = \alpha_1 + z\,\alpha_2, \tag{4.4}$$

dove α_1 è la correlazione tra le coppie di primo grado, z è una variabile indicatrice che vale 1 nel caso di coppie di secondo grado e 0 altrimenti, mentre α_2 indica la variazione di correlazione tra i due gradi di parentela.

Il programma in R usato differisce leggermente dal precedente ed è riportato nella sezione 5 dell'Appendice.

	Modello Glm Marginale	**Modello Pairwise di 2° grado**
Intercetta	5.4983	5.5440
Allele 177-uno	-1.1417	-1.1376
Allele 177-due	1.9467	1.9697
Allele 248-uno	-3.8891	-3.8858
Allele 303-uno	0.8287	0.9058
Allele 303-due	0.2492	0.2652
Allele 329-uno	0.0097	0.0261
Allele 329-due	1.1816	1.2145
Allele 501-uno	0.8382	0.8865
Allele 501-due	-1.4111	-1.3529
Allele 639-due	-2.4016	-2.3662
Allele 679-uno	0.8138	0.9322
Allele 679-due	-1.0092	-1.0047
Correlazione 1	---	0.7752
Correlazione 2	---	-0.1678

Tabella 4.5 – Stima del modello marginale e utilizzando la verosimiglianza a coppie considerando le osservazioni di primo e secondo grado di parentela.

Per la stima vengono utilizzate le 492 coppie di primo grado e le 775 coppie di secondo grado, aumentando così la numerosità campionaria. Anche in questo modello, come nel precedente, le stime dei parametri restano molto vicine a quelle ottenute con il modello marginale. La correlazione resta positiva sia tra i parenti di primo grado che nel secondo grado, anche se nel secondo grado questa scende leggermente come quanto verificato nella Tabella 4.2. Ulteriori analisi su questi risultati non possono essere effettuate se non calcolando gli *standard error* delle stime, obiettivo del prossimo Paragrafo.

4.3 Creazione dei gruppi e stima degli errori standard

Una volta terminato il processo di stima dei parametri di regressione, si deve andare alla ricerca del valore degli *standard error* in modo da poter verificare se i valori stimati per i parametri di regressione risultano significativi. Ricavare la matrice di Godambe in modo diretto risulta in questa circostanza complicato; quindi si cerca una sua stima per via indiretta utilizzando il metodo *Jack-knife*. Il metodo *Jack-knife* in campioni unidimensionali di lunghezza *n* prevede la creazione di *n* gruppi togliendo una osservazione alla volta, in modo da effettuare la stima di un valore tramite una funzione di stima *s(.)* replicata per *n* volte. Con *n* stime viene poi calcolata la varianza dello stimatore. Lo stimatore è espresso come

$$\hat{jk}(x)=\frac{1}{n}\sum_{i=1}^{n} s(x_i), \tag{4.6}$$

dove x_i è il vettore *x* privata della *i-esima* osservazione e *s(.)* è la funzione di stima.

La principale problematica che si incontra in questo caso riguarda il fatto che non si può applicare questo metodo senza trascurare il fatto che le coppie non sono tra loro indipendenti, in quanto più coppie possono appartenere alla stessa famiglia. L'unica soluzione che possiamo applicare è la creazione di gruppi pseudo indipendenti con una numerosità abbastanza omogenea. Per maggiori approfondimenti su questa metodologia si rimanda a Bienias *et al*. (2002). La maggior informazione presente nel dataset per la definizione dei gruppi + la parentela e quindi si cerca di aggregare le unità in gruppi, in modo da andare alla ricerca di coppie di osservazioni tra loro imparentate. Si procede in questo algoritmo fino a coppie di osservazioni distanti fino al terzo grado di legame familiare. Si trovano con questo metodo 81 gruppi di coppie con distribuzione non omogenea, in quanto solo 7 gruppi hanno una numerosità tale da superare le 20 unità. Su suggerimento di un ricercatore del centro CRS4, si procede a unire gruppi di coppie con numero identificatore vicino (Id), in quanto anche questo dato definisce osservazioni tra loro vicine a livello familiare. Selezionando e unendo manualmente le osservazioni, si riescono ad ottenere 19 gruppi con distribuzione omogenea. Successivamente, si procede con la stima dei parametri togliendo un gruppo alla volta. In questo modo si stimano 19 modelli e quindi si ricavano 19 stime dei parametri.

Il calcolo della varianza e degli *standard error* è semplice e i risultati ottenuti sono riportati nella Tabella 4.7.

	Modello Glm Marginale	**Standard error**	**Modello Pairwise**	**Standard error**
Intercetta	5.4983	0.6760	5.5681	0.5076
Allele 177-uno	-1.1417	0.2092	-0.9697	0.1037
Allele 177-due	1.9467	0.4753	2.0109	0.2121
Allele 248-uno	-3.8891	0.5229	-3.8323	0.4194
Allele 303-uno	0.8287	0.2433	0.6660	0.0817
Allele 303-due	0.2492	0.2319	0.2967	0.0775
Allele 329-uno	0.0097	0.2202	0.0384	0.1140
Allele 329-due	1.1816	0.2916	0.6027	0.1460
Allele 501-uno	0.8382	0.2280	1.0827	0.1115
Allele 501-due	-1.4111	0.3875	-1.1797	0.2282
Allele 639-due	-2.4016	0.3053	-2.4428	0.1471
Allele 679-uno	0.8138	0.4610	0.8615	0.1847
Allele 679-due	-1.0092	0.1988	-1.2468	0.1248
Correlazione	---	---	0.9314	0.2608

Tabella 4.7 – Confronto nella stima dei parametri e degli standard error tra modello marginale e quello dello ottenuto con la verosimiglianza a coppie.

Come si può notare le stime degli *standard error* nel modello a coppie sono tutte inferiori rispetto ai valori ottenuti nel modello marginale, e quindi le stime dei parametri risultano più significative. Inoltre la correlazione è significativa, in quanto abbiamo un valore nella statistica test pari a 3.5713 (*p-value* 0.00136), quindi la correlazione stimata è fortemente significativa. Ora possiamo valutare la capacità di previsione del modello con i parametri stimati e riportiamo nella Tabella 4.6 la matrice di confusione in modo da verificare come il modello di regressione, con i parametri calcolati con la verosimiglianza a coppie, discrimina i malati dai non affetti dalla malattia.

Previsione	Non malati effettivi	Malati effettivi
Non malati previsti	448	30
Malati previsti	18	58

Tabella 4.8 – Tabella di confusione per il modello utilizzando la verosimiglianza a coppie.

La percentuale di falsi positivi è di 3.86%, mentre i falsi negativi sono il 34,09%. La capacità di previsione peggiora leggermente se confrontiamo tali dati con il modello marginale, ma tale confronto è fuori luogo in quanto, alla luce della significatività della correlazione, il modello marginale non è specificato correttamente.
Per potere verificare che le stime sono più precise, si utilizza il test sulla nullità del parametro e come distribuzione del test si sfrutta la normalità asintotica degli stimatori massima verosimiglianza.

	P-value modello marginale	P-value Verosimiglianza a coppie
Intercetta	4.17e-16	5.92e-27
Allele 177-uno	4.85e-08	8.20e-20
Allele 177-due	4.21e-05	2.41e-20
Allele 248-uno	1.02e-13	5.90e-19
Allele 303-uno	0.000659	2.96e-15
Allele 303-due	0.282669	0.000523
Allele 329-uno	0.964863	0.7538796
Allele 329-due	5.08e-05	1.59e-04
Allele 501-uno	0.000238	2.67e-21
Allele 501-due	0.000271	1.25e-06
Allele 639-due	3.63e-15	1.04e-60
Allele 679-uno	0.077495	1.50e-05
Allele 679-due	3.85e-07	1.69e-22

Tabella 4.9 – Confronto di significatività nella stima dei parametri.

La Tabella 4.9 ci rappresenta un confronto tra i valori di significatività della statistica-test sulla nullità del parametro, tra modello marginale e modello stimato con la verosimiglianza a coppie. Con quest'ultimo metodo, le stime risultano tutte più precise, a testimonianza che riusciamo a cogliere meglio la struttura latente nei dati. La variazione risulta molto marcata, e c'è perfino un parametro, quello in riferimento all'allele 303-due, che passa da non significativo a fortemente significativo utilizzando come soglia il valore di 0.05.
L'allele 639-due è quello più rilevante all'interno del modello, dato concordante con quanto verificato presso il gruppo di ricerca CRS4 in Sardegna. Nel prossimo Paragrafo andremo a definire le principali problematiche incontrate nel percorso di stima con la verosimiglianza a coppie, soffermando l'attenzione nella critica della metodologia computazionale utilizzata.

4.4 Problemi computazionali e di stima

Le principali difficoltà nell'utilizzo della verosimiglianza a coppie sono due. La prima riguarda il tempo di calcolo computazionale dei programmi realizzati: per esempio, nel software creato per la stima dei parametri, descritto in sezione 3 dell'Appendice, si utilizza il ciclo *for(.)* che nel programma R risulta molto lento, basti pensare che una stima dei parametri del modello più semplice, presentato nel Paragrafo 4.2, ci si impiega circa 30 minuti, lavorando con un computer dotato di un buon livello di capacità di calcolo. I tempi di stima risultano elevati anche nel calcolo del modello a correlazione variabile e soprattuto nella stima degli *standard error*, in cui la stima ha impiegato circa 10 ore. Per quest'ultimo motivo non sono stati presentati gli *standard error* del modello con correlazione di primo e secondo grado.

La prima soluzione è quella di sostituire il ciclo *for(.)* con la funzione *apply(.)*; in questo modo si sono ridotti i tempi di stima, ma non in modo soddisfacente. Nella Tabella 4.10 vengono riportati i tempi computazionali per il calcolo delle principali stime calcolate nel Capitolo 4.

Una possibile soluzione alla problematica è quella di esportare il codice in formato C e in questo caso si segnala il sito internet http://www.hipersoft.rice.edu/rcc/, dove un gruppo di ricercatori sta creando un software per compilare i programmi in R direttamente in C senza bisogno di modificare il codice e senza l'utilizzo di un compilatore esterno.

Processo di stima	Parametri stimati	Tempo di calcolo
Modello completo marginale	43	0.43 sec
Modello ridotto marginale	13	0.23 sec
Modello con SMVC 1° grado	14	circa 25 minuti
Jack-knife per il calcolo degli s.e. in SMVC 1°	14	circa 9 ore
Modello con SMVC 2° grado	15	circa 55 minuti

Tabella 4.10 – Tempi di calcolo per i principali processi di stima.

La natura dei dati ha inciso profondamente sulla struttura della verosimiglianza che si deve massimizzare. Le variabili di regressione sono dicotomiche e, come si può immaginare. la log-verosimiglianza si presenta irregolare. Questo problema comporta la presenza di massimi e minimi locali locali e gli algoritmi utilizzati dalla funzione *optim(.)* di R non riescono a convergere in modo soddisfacente. Molta importanza viene data dai valori iniziali scelti nel programma di inizializzazione. La miglior soluzione è impostare diversi valori nei parametri di partenza e verificare di volta in volta il valore di verosimiglianza ottenuta. Purtroppo, questo procedimento con i mezzi a disposizione non è realizzabile, in quanto, come visto sopra, una singola massimizzazione impiega troppo tempo dal punto di vista computazionale.

Capitolo 5

Conclusioni

A chiusura di tutto di questo lavoro, sono emersi diversi elementi che meritano di essere verificati attentamente. L'obiettivo di questa tesi è una breve rassegna su quanto è stato pubblicato sulla verosimiglianza a coppie, l'approfondimento e l'uso di questa tecnica per lo studio di dati binari dipendenti. I dati binari, sono per natura, difficili da analizzare specie se ci troviamo di fronte a variabili di regressione di tipo fattoriale. L'utilizzo della verosimiglianza a coppie ha consentito una stima del livello di correlazione dello stato di affezione alla malattia tra parenti di primo e di secondo grado. Calcolando gli *standard error,* nel modello con i familiari di primo grado, si possono trarre delle conclusioni positive:

- l'uso della verosimiglianza a coppie nella stima di un modello di regressione ha comportato una migliore precisione nella stima di quasi tutti parametri stimati rispetto al modello marginale. L'incremento di significatività risulta marcato;
- la correlazione stimata, anche se di livello discreto, risulta essere fortemente significativa e quindi viene colta, almeno parzialmente, la struttura di dipendenza latente, uno dei principali obiettivi di questa analisi;
- nel modello con la stima della correlazione di primo e secondo grado viene confermato che la dipendenza tra lo stato di malattia tra parenti cala al crescere del grado di parentela, però di questo modello non abbiamo gli *standard error* e quindi non possiamo verificare, in un eventuale test, se le stime sono rilevanti.

Per poter effettuare le analisi presenti nel Capitolo 4 abbiamo dovuto effettuare alcune scelte che, tenendo conto dei risultati ottenuti, non sono trascurabili:

- il formato Pedigree del file, ha una struttura che non si integra agevolmente con i software di statistica più utilizzati e per il calcolo delle coppie dipendenti ci siamo fermati al secondo grado, trascurando correlazioni di ordine superiore, a causa, soprattutto, dei dati incompleti, presenti nel dataset, e della complessità computazionale del calcolo dei legami di parentela tra osservazioni;
- la scelta delle variabili da includere nel modello utilizzato per la stima a coppie nel Capitolo 4 è stata effettuata senza tener conto della dipendenza; un modo migliore e, certamente più corretto per procedere, è quello di introdurre l'analisi con la verosimiglianza a coppie anche in via preliminare ed effettuare una selezione *forward*, nello stesso modo di un modello lineare semplice;
- a causa dell'elevato tempo di stima non si è tenuto conto delle possibili interazioni tra le variabili scelte nel modello iniziale, interazioni, tuttavia, presenti quando si incontrano dati di natura genetica;
- una singola stima dei parametri con la verosimiglianza a coppie ha un tempo di calcolo troppo oneroso. A causa di questo motivo non siamo in grado di supportare ulteriori analisi dopo il primo livello di parentela.

Come specificato nel Paragrafo 4.4, gli attuali algoritmi di ottimizzazione, non offrono risultati confortanti in termini di tempo e di convergenza. Tuttavia occorre sottolineare che il modello risulta ben specificato e pronto per verificare la presenza di altri alleli significativi, oltre a quelli presenti nel modello ridotto di partenza. La verosimiglianza a coppie si è dimostrata una tecnica flessibile e concreta per cercare di modellare dati dipendenti, senza specificare una struttura rigida sui parametri.

Appendice

Programmi e Risultati in R

Tutti i comandi e i risultati presenti in questa parte sono derivati dall'ambiente di lavoro del programma statistico R. Per una sintassi più completa dei comandi utilizzati si rimanda all'aiuto in linea di R e a Masarotto G. e Iacus S. M. (2003).

1 Output di R di un modello Binomiale per la discriminazione dei malati. Alpha=0.10

glm(formula = I(dati[, 6] - 1) ~ ., family = binomial(probit), data = dati[, ok])

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.814e+00	-1.203e-01	-1.354e-02	-5.401e-05	2.912e+00

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	7.35537	1.75416	4.193	2.75e-05	***
X177DUE	2.59657	0.78528	3.307	0.000945	***
X177UNO	-1.70036	0.41321	-4.115	3.87e-05	***
X248UNO	-4.41464	1.10611	-3.991	6.58e-05	***
X293DUE	-0.90730	0.58280	-1.557	0.119522	
X293UNO	-1.71893	0.55372	-3.104	0.001907	**
X303DUE	1.40181	0.51601	2.717	0.006595	**
X303UNO	1.55344	0.48346	3.213	0.001313	**
X329DUE	2.08654	0.46692	4.469	7.87e-06	***
X329UNO	-0.82796	0.40393	-2.050	0.040389	*
X404DUE	0.63587	0.37805	1.682	0.092574	.
X404UNO	-0.88162	0.56111	-1.571	0.116131	
X453DUE	-1.04001	0.48309	-2.153	0.031333	*
X453UNO	-2.06065	266.40633	-0.008	0.993828	
X501DUE	-1.43417	0.52799	-2.716	0.006602	**
X501UNO	1.88289	0.47956	3.926	8.63e-05	***
X539DUE	-0.65121	0.37838	-1.721	0.085248	.
X539UNO	-0.21111	0.45054	-0.469	0.639370	
X605DUE	0.75879	0.39775	1.908	0.056430	.
X605UNO	-1.58854	0.55841	-2.845	0.004445	**
X609DUE	0.81106	0.51545	1.574	0.115602	

X609UNO -1.76739 0.47662 -3.708 0.000209 ***
X633DUE 1.27560 0.67947 1.877 0.060471 .
X633UNO 0.03539 0.35085 0.101 0.919664
X639DUE -1.79263 0.63509 -2.823 0.004763 **
X679DUE -2.30726 1.24271 -1.857 0.063364 .
X679UNO 2.70906 1.51893 1.784 0.074499 .
X688DUE 0.73360 1.29710 0.566 0.571689
X688UNO -4.71682 1.32778 -3.552 0.000382 ***
X749DUE 0.30811 0.73402 0.420 0.674661
X749UNO 1.24793 0.59627 2.093 0.036360 *
X750DUE 1.18648 0.47825 2.481 0.013107 *
X750UNO 0.72999 0.48760 1.497 0.134367
X810DUE -2.68426 1.27664 -2.103 0.035502 *
X810UNO -0.92278 0.54684 -1.687 0.091509 .
X923DUE -1.15240 0.51904 -2.220 0.026403 *
X923UNO 0.41924 0.36124 1.161 0.245818
X933DUE 1.72926 0.58658 2.948 0.003198 **
X933UNO -0.16780 0.35793 -0.469 0.639212
X971DUE -0.54232 0.52155 -1.040 0.298421
X971UNO -0.80254 0.33601 -2.388 0.016918 *
X1006DUE 0.82908 0.47667 1.739 0.081979 .
X1006UNO 0.45308 0.36452 1.243 0.213892
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 485.03 on 553 degrees of freedom
Residual deviance: 142.12 on 511 degrees of freedom
AIC: 228.12

Number of Fisher Scoring iterations: 14

2 Output di R di un modello Binomiale per la discriminazione dei malati. Alpha=0.005

```
glm(formula = I(dati[, 6] - 1) ~ ., family = binomial(probit),
    data = dati[, ok3])

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.07550  -0.34204  -0.19097  -0.05702   3.56627

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.4983     0.6760   8.134 4.17e-16 ***
X177DUE       1.9467     0.4753   4.096 4.21e-05 ***
X177UNO      -1.1417     0.2092  -5.457 4.85e-08 ***
X248UNO      -3.8891     0.5229  -7.438 1.02e-13 ***
X303DUE       0.2492     0.2319   1.074 0.282669
X303UNO       0.8287     0.2433   3.406 0.000659 ***
X329DUE       1.1816     0.2916   4.052 5.08e-05 ***
X329UNO       0.0097     0.2202   0.044 0.964863
X501DUE      -1.4111     0.3875  -3.642 0.000271 ***
X501UNO       0.8382     0.2280   3.675 0.000238 ***
X639DUE      -2.4016     0.3053  -7.867 3.63e-15 ***
X679DUE      -1.0092     0.1988  -5.076 3.85e-07 ***
X679UNO       0.8138     0.4610   1.765 0.077495 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 485.03  on 553  degrees of freedom
Residual deviance: 245.43  on 541  degrees of freedom
AIC: 271.43

Number of Fisher Scoring iterations: 7
```

3 Programma per il calcolo della verosimiglianza a coppie in ambiente R con correlazione costante

```
#funzione per il calcolo della densità della funzione pairwise
bprob<-function(y,alpha2){
-I((y[1]==1)&(y[2]==1))*log(pmvnorm(mean=c(0,0),lower=c(-y[3],-
y[4]),upper=c(Inf,Inf),sigma=matrix(c(1,alpha,alpha,1),2,2)))
-I((y[1]==0) & (y[2]==1))*log(pmvnorm(mean=c(0,0),lower=c(-Inf,-y[4]),upper=c(-
y[3],Inf),sigma=matrix(c(1,alpha,alpha,1),2,2)))
-I((y[1]==1) & (y[2]==0))*log(pmvnorm(mean=c(0,0),lower=c(-y[3],-Inf),upper=c(Inf,-
y[4]),sigma=matrix(c(1,alpha,alpha,1),2,2)))
-I((y[1]==0) & (y[2]==0))*log(pmvnorm(mean=c(0,0),lower=c(-Inf,-Inf),upper=c(-y[3],-
y[4]),sigma=matrix(c(1,alpha,alpha,1),2,2)))
  }

#funzione per il calcolo della verosimiglianza pairwise
pairwise<-function(theta,y,x)
{
  int <- theta[1]
  beta1 <- theta[2:13]
  alpha1 <- theta[14]
  alpha2<-(exp(alpha1)-1)/(exp(alpha1)+1)
  means <- matrix(0, nrow=492, ncol=2)
  means[,1]<-int+as.matrix(x[,1:12])%*%beta1
  means[,2]<-int+as.matrix(x[,13:24])%*%beta1
  sum(apply(cbind(y,means),1,bprob,alpha2=alpha2))
}

# massimizzazione della verosimiglianza del punto precedente con la funzione optim
ris<-optim(c(rep(0,13),0.5),pairwise,y=y,x=x)
```

4 Imputazione dati mancanti con tecnica hot-deck

```
#imputazione dati mancanti tramite hot deck
hotdeck<-function(dati){
n<-dim(dati)[1]
m<-dim(dati)[2]
dist<-dist(dati[,7:m])
dist2<-as.matrix(dist)
n<-dim(dati)[1]
for(j in 7:m){
for(i in 1:n){
if(is.na(dati[i,j])){
dist2[i,i]<-100000000000
min<-min(dist2[i,])
pos1<-posizione(min,dist2[i,])
dati[i,j]<-dati[pos1,j]
if(is.na(dati[i,j])){
min2<-min(dist2[i,-pos1])
pos2<-posizione(min2,dist2[i,])
dati[i,j]<-dati[pos2,j]
if(is.na(dati[i,j])){
min3<-min(dist2[i,-c(pos1,pos2)])
pos3<-posizione(min3,dist2[i,])
dati[i,j]<-dati[pos3,j]
if(is.na(dati[i,j])){
min4<-min(dist2[i,-c(pos1,pos2,pos3)])
pos4<-posizione(min4,dist2[i,])
dati[i,j]<-dati[pos4,j]
}}}}}}
dati }
```

5 Programma per il calcolo della verosimiglianza a coppie in ambiente R con correlazione variabile

```
pairwise2<-function(theta,y,x)
{
 int <- theta[1]
 beta1 <- theta[2:13]
 alpha1 <- theta[14]
 alpha12 <- theta[15]
 alpha2[1]<-(exp(alpha1)-1)/(exp(alpha1)+1)
 alpha2[2]<-(exp(alpha12)-1)/(exp(alpha12)+1)
 means <- matrix(0, nrow=1267, ncol=2)
 means[,1]<-int+as.matrix(x[,1:12])%*%beta1
 means[,2]<-int+as.matrix(x[,13:24])%*%beta1
 tt<-cbind(y,means)
 sum(apply(tt,1,bprob,alpha2=alpha2))
}
bprob<-function(y,alpha2){
alphat<-alpha2[1]+y[3]*alpha2[2]
-I((y[1]==1)&(y[2]==1))*log(pmvnorm(mean=c(0,0),lower=c(-y[4],-
y[5]),upper=c(Inf,Inf),sigma=matrix(c(1,alphat,alphat,1),2,2)))
-I((y[1]==0)&(y[2]==1))*log(pmvnorm(mean=c(0,0),lower=c(-Inf,-y[5]),upper=c(-
y[4],Inf),sigma=matrix(c(1,alphat,alphat,1),2,2)))
-I((y[1]==1)&(y[2]==0))*log(pmvnorm(mean=c(0,0),lower=c(-y[4],-Inf),upper=c(Inf,-
y[5]),sigma=matrix(c(1,alphat,alphat,1),2,2)))
-I((y[1]==0)&(y[2]==0))*log(pmvnorm(mean=c(0,0),lower=c(-Inf,-Inf),upper=c(-y[4],-
y[5]),sigma=matrix(c(1,alphat,alphat,1),2,2)))
}

optim(c(ini,1.0,0),pairwise2,y=yt2,x=xt2)
```

Bibliografia

Agresti, A. (2002). *Categorical Data Analysis*, Second Edition. Wiley, New York.

Azzalini, A. (1983), Maximun likelihood of order m for stochastic processes. Biometrika 70, 367-381.

Bellio, R. e Varin, C. (2005) A pairwise likelihood approach to generalized linear models with crossed random effetcs. Statistical modelling 5, 217-227.

Besag, J. (1974). Spatial Interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society B 34, 192-236..

Bienias J. L., Kott P. e Evans D. A. (2002). A*pplication of the delete-a-group jackknife variance estimator to analyses of data from a complex longitudinal survey.* Proceedings of Statistics Canada Symposium, Chicago.

Cessie, S. L. e Houwelingen, J. v. (1994). Logistic regression for correlated binary data. Applied Statistics 43, 95-108.

Cox, D.R. e Reid, N. (2004). A note on the pseudolikelihood constructed form marginal densities. Biometrika 91, 729-737.

De Leon, A.R. (2005). Pairwise likelihood approach to grouped continuos model and its extension. Statistics and Probability Letters 75, 49-57.

Geys H., Molenberghs, G. e Lipsitz, S.R. (1998) A note on the comparison of pseudolikelihood and generalized estimating equations for marginal odds ratio model. Journal of Statistical Computation and Simulation 62, 45-72.

Geys H., Molenberghs, G. e Ryan , L.M. (1997) Pseudolikelihood inference for clustered binay data. Communications in Statistics: Theory and Methods 26, 2743-2767.

Kuk, A.Y.C. e Nott. D.J. (2000). A pairwise likelihood approach to analyzing correlated binary data. Statistics and Probability Letters 47, 329-335.

Lindsay, B.G. (1988). Composite likelihood methods. Contemporaney mathematics 80, 221-240.

Liang, K. Y. e Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. Biometrika 73, 12-22

Molenberghs, G. e Verbeke, G. (2005). *Models for discrete longitudinal data*. Spinger-Verlag, New York.

Nott, D.J. e Ryden T. (1999). Pairwise likelihood methods for inference in image models. Biomerika 86, 661-676.

Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation. Biometrics 47, 823-883.

Renard, D., Molenberghs, G. e Geys, H. (2004) A pairwise likelihood approach to estimation in mutilevel probit models. Computational Statistics and Data Analysis 44,649-667.

Varin, C., Høst, G. e Skare, Ø. (2005). Pairwise inference in spatial generalized mixed models. Compuatational Statistics and Data Analysis 49, 1173-1191.

Varin, C. e Vidoni, P. (2005). A note on composite likelihood inference and model selection. Biometrika 92, 519-528.

Zhao, Y. e Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. The Canadian Journal of Statistics 77, 642-648.

Ziegler, A. e Kastner, C. (1996). The generalized estimating equations in the past ten year: an overview and a biomedical application. Informatik, Epidemiologie und Biometrie in Medizin und Biologie 386, 24.

Ringraziamenti

Alla fine di questo percorso di studi desidero ringraziare le persone che mi sono state vicine e che hanno permesso il raggiungimento di questo importate risultato.
Desidero ringraziare la Prof.ssa Ventura per avermi concesso, dopo l'esperienza della tesi nella laurea triennale, questa importante opportunità di studio, considerando la disponibilità e la pazienza dimostrata in qualsiasi momento. Inoltre ringrazio il Dott. Varin per l'importante aiuto fornito durante alcune particolari fasi della tesi.
Ringrazio i miei genitori, mia sorella Michela e tutti i miei familiari che hanno creduto in me e che mi hanno sempre sostenuto in tutti questi anni.
Un caloroso saluto ai miei compagni di corso, in particolare Fabio, Elisa, Michele e Stefano, per aver condiviso i momenti di affanno e di spensieratezza della vita universitaria.
Una dedica particolare a tutti gli amici, in particolare ad Alessia che mi è stata vicina nella stesura di questo lavoro, e a Riccardo per essere sempre stato presente nonostante gli impegni lavorativi.
Un abbraccio a tutti i componenti del Gruppo di Musici e Sbandieratori del mio paese per le bellissime giornate trascorse in vostra compagnia.