Università degli studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in

Scienze Statistiche

Inferenza bayesiana per dati ordinali multivariati: un caso di studio sulla parodontite in soggetti afroamericani

Relatore Prof. Antonio Canale
Dipartimento di Scienze Statistiche

Laureando Federico Garbin
Matricola 1105102

Anno Accademico 2016/2017

Indice

Introduzione

La parodontite è una delle forme più gravi di malattia del dente. Si può misurare la perdita di attacco clinico attraverso una sonda parodontale millimetrata: la misura che si ottiene è soggetta a naturale arrotondamento e risulta pertanto discreta e ordinata. Al crescere del valore vi è indicazione del peggioramento dello stato di salute del sito e del dente, nello specifico. Il processo che lo descrive è tuttavia di natura continua e latente, non direttamente osservabile. Verrà utilizzato un modello *probit* ordinale multivariato, con approccio di inferenza bayesiano, per spiegare la relazione tra il fenomeno latente sottostante e alcune covariate a livello di sito e soggetto, considerando una struttura di dipendenza spaziale tra i siti di rilevazione di tipo autoregressivo condizionato (CAR) nella matrice di varianze e covarianze. (Besag 1974)

Nel primo capitolo viene descritta la parodontite più nel dettaglio, assieme ai metodi di rilevazione. Si presenta quindi il dataset, definendo in particolare le variabili che interessano i soggetti e i siti. Viene rappresentata la struttura della dentatura con la numerazione dei siti e una griglia parodontale per definire il tipo di vicinanza tra di essi. Infine, verranno riportate brevi analisi descrittive sulle proporzioni empiriche relative alle cinque classi di gravità della malattia, per alcune covariate a livello di soggetto, stratificate per arcata superiore ed inferiore della dentatura.

Nel secondo capitolo si descrive il modello statistico attraverso una specificazione gerarchica dei parametri ed iperparametri coinvolti, la tecnica di *data augmentation* per l'introduzione delle variabili latenti, le distribuzioni a priori per i parametri di interesse, alcuni passaggi algebrici per ottenere le distribuzioni *full conditional* e l'algoritmo *Metropolis-within-Gibbs* per

l'aggiornamento dei parametri.

Nel terzo capitolo verranno presentati i risultati: alcune analisi sulla convergenza del modello per ogni singola componente, interpretazione delle stime dei parametri ottenute, intervalli di credibilità e alcune considerazioni finali.

Capitolo 1

La parodontite

1.1 Descrizione ed epidemiologia

I denti sono un organo duro e calcificato, il cui principale scopo è quello di svolgere la primissima fase della digestione, ovvero la masticazione. Data la particolare conformazione del cavo orale e l'attività che svolgono, tali organi non sono immuni dallo sviluppo di diversi tipi di patologie. La parodontite è un'infiammazione del parodonto, costituito da: osso alveolare, legamento parodontale, cemento radicolare e gengiva. Sebbene in fase iniziale essa risulti reversibile e curabile, una sua progressione porta ad un'incrementale perdita di attacco clinico (CAL - *clinical attachment loss*) che, nel caso peggiore, può portare ad una conseguente caduta del dente.

Il Prof.Tonetti, in un congresso della SIdP (*Società Italiana di Parodontologia e Implantologia*) («Quali strategie per arginare l'«epidemia» di parodontite?») riporta i dati relativi ad importanti studi epidemiologici svolti fino a quel momento, che collocano questa malattia nella sesta posizione tra le malattie più diffuse al mondo, con 743 milioni di individui affetti e un picco di incidenza attestato a 38 anni di età. Inoltre, nei paesi occidentali oltre il 40% della popolazione adulta è colpita dalla parodontite. Di questi individui, il 10-14% ne soffre in forma grave.

1.2 Valutazione parodontale approfondita

Si compone essenzialmente di quattro differenti esami:

1. Anamnesi medica e dento-parodontale
2. Esame obiettivo
3. Esami radiografici
4. Esami di laboratorio

L'anamnesi ha lo scopo di far emergere la presenza di una serie di elementi che influenzano l'insorgenza e la progressione della parodontite in termini di fattori o indicatori di rischio, tra i quali (come si vedrà nella Sezione 1.4) si può citare il fumo e la presenza di malattie croniche, come il diabete; oltre che ad altre importanti caratteristiche individuali su abitudini, stili di vita e familiarità per la malattia.

L'esame obiettivo si basa sull'ispezione ed osservazione dei denti. Di particolare interesse è il sondaggio parodontale, che permette di raccogliere molti parametri clinici. Si effettua tramite una sonda parodontale (Figura 1.1) che viene inserita fra dente e gengiva mantenendo una corretta angolazione ed una forza di 30 gr_f. Permette di misurare la profondità delle tasche gengivali e delle recessioni, di individuare il coinvolgimento delle forcazioni e rilevare eventuale sanguinamento al sondaggio. La sonda viene fatta scorrere lungo tutto il profilo di ogni dente e, usualmente, si registrano i dati relativi a sei siti: mesio-vestibolare, centro-vestibolare, disto-vestibolare, mesio-linguale, centro-linguale e disto-linguale.

Tra di essi, uno dei parametri di maggior importanza per la valutazione della gravità della malattia è rappresentato dalla perdita di attacco clinico (CAL), calcolato come distanza tra la giunzione amelo-cementizia (CEJ - *cementoenamel junction*) e la profondità massima di sondaggio (PD - *probe depth*) (Figura 1.2).

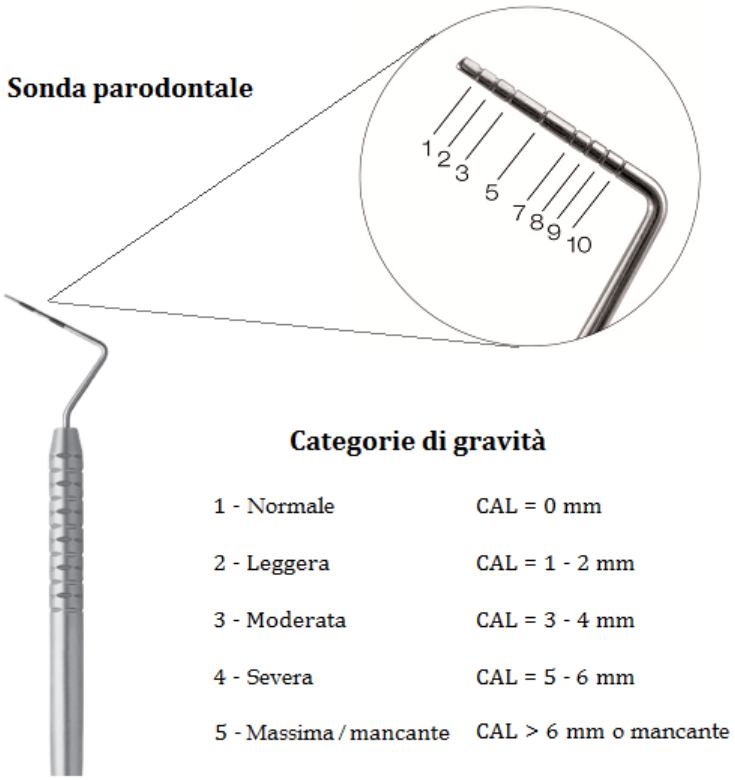

Figura 1.1: Sonda parodontale millimetrata per misurare la profondità della tasca gengivale. Fonte: Prodentis (*Graduated Periodontal Explorer - Williams Classic*)

1.3 Presentazione del dataset

I dati di cui si dispone per l'analisi sono stati gentilmente concessi dal Prof. Bandyopadhyay e riguardano uno studio clinico condotto dalla *Medical University of South Carolina* per determinare lo stato della parodontite di soggetti afro-americani di lingua creola *Gullah* affetti dal diabete di tipo 2, originariamente presentati in Fernandes et al. (2009).

Il campione è costituito da 288 soggetti, per i quali sono disponibili le misure di CAL per ognuno dei 28 denti, fatta eccezione per i 4 denti del giudizio. Sono state effettuate, per ogni individuo, le misurazioni tramite la sonda parodontale per ciascuno dei 6 siti relativi a ciascun dente. Nel caso il soggetto non presenti denti mancanti, sono disponibili quindi 168 rilevazioni

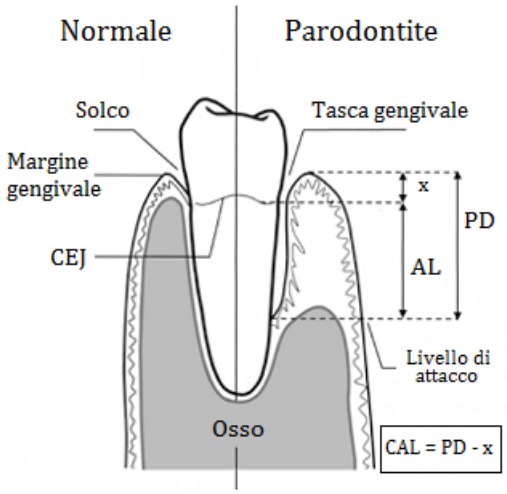

Figura 1.2: Perdita di attacco clinico: confronto tra soggetti normali e affetti da parodontite. Fonte: Arora et al. (2009)

complessive (Figura 1.3). Le prime 84 (1-84) vengono effettuate nella parte superiore, o mascella, le altre 84 (85-168) nella parte inferiore, o mandibola. Le misurazioni sono poi effettuate in due lati: interno (linguale), a diretto contatto con la lingua ed esterno (buccale), a diretto contatto con la bocca.

1.3.1 Covariate a livello di soggetto

Sono disponibili alcune variabili relative al soggetto, in particolare:

- Età
- Genere
- BMI (*Body Mass Index*) in kg/m^2
- Fumo
- HbA1c (emoglobina glicata) nel sistema $DCCT(\%)$

In particolare, la variabile fumo viene suddivisa in due categorie: "fumatori attuali/in passato" oppure "mai stato fumatore", mentre le variabili relative a BMI e al livello di emoglobina glicata vengono utilizzate mantenendo il valore numerico originale.

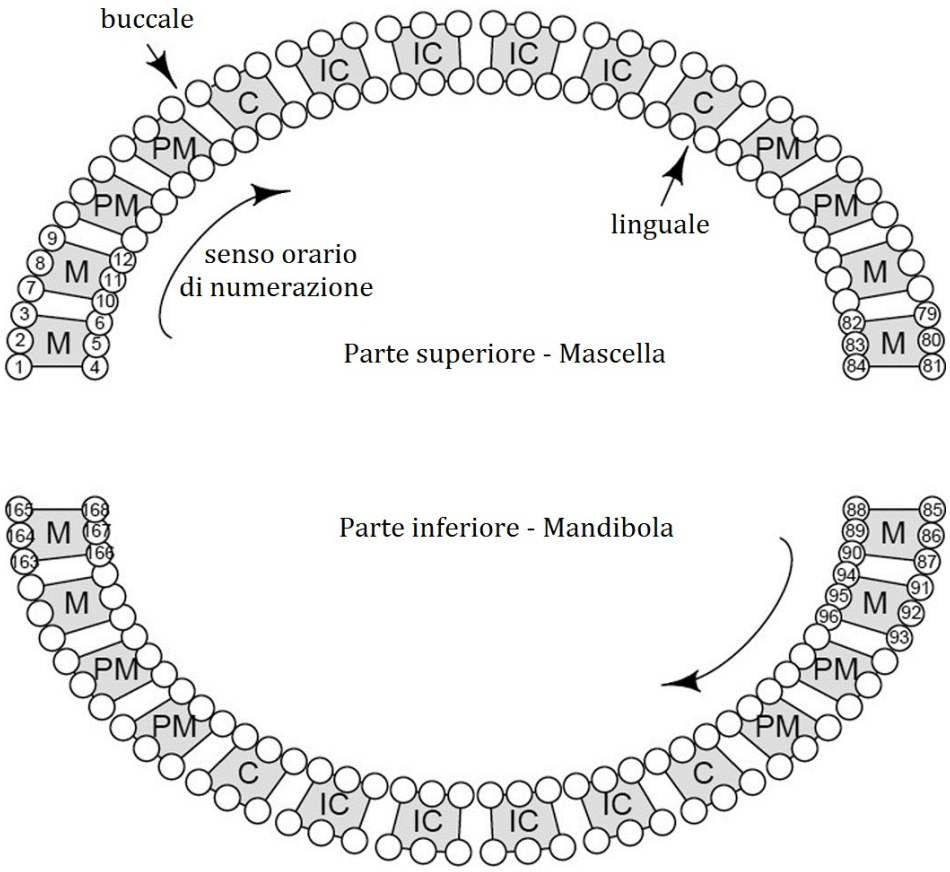

Figura 1.3: Siti di rilevazione per ogni tipo di dente (M=molari PM=premolari C=canini IC=incisivi). Fonte: Bandyopadhyay e Canale (2016)

1.3.2 Covariate a livello di sito

Di seguito vengono riportate le covariate a livello del sito:

- Sito esterno
- Indicatore parte inferiore (mandibola)/superiore (mascella)

L'indicatore relativo alla locazione del sito in un gap indica se il sito è mesio-buccale/disto-buccale/mesio-linguale/disto-linguale oppure se è posizionato al centro tra la parte mesiale e distale. Tralasciando queste nozioni, viene semplicemente utilizzata una variabile binaria che differenzia i siti centrali, in rosso nella Figura 1.4 dagli altri tipi di siti, in nero. E' necessario

definire il concetto di sito vicino. Vi sono tre tipi di distanza: distanza tra siti nello stesso lato (ad es. buccale) ad esclusione del sito centrale (*Tipo 1*), distanza tra siti di due denti diversi consecutivi posti nello stesso lato (*Tipo 2*) ed infine la distanza tra siti di lati opposti presenti nello stesso dente. Nell'immagine seguente vengono riportati i tipi di distanza sopra presentati (Figura 1.4).

La matrice di adiacenza che definisce i siti vicini verrà presentata nel capitolo successivo. In particolare, verranno considerati vicini tutti i siti appartenenti allo stesso dente, considerando quindi solo le distanze di *Tipo 1* e *3*, includendo il sito centrale.

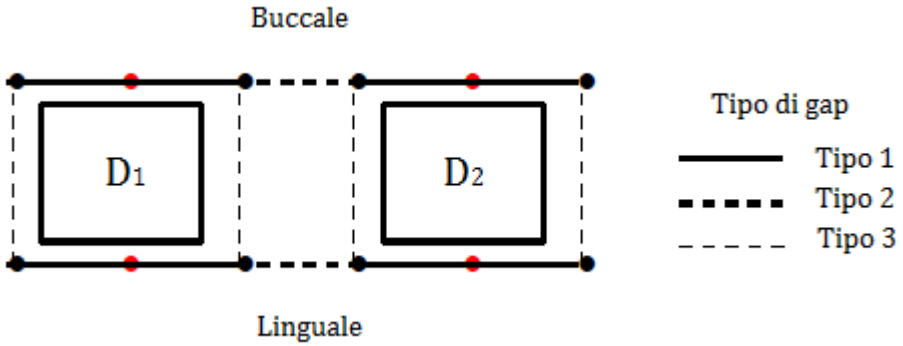

Figura 1.4: Griglia parodontale per due denti adiacenti situati nella stessa mascella. I punti rossi indicano i siti centrali, mentre i punti neri i siti non centrali, per i quali sono definiti tre tipi di distanze. Fonte: materiale aggiuntivo in Bandyopadhyay e Canale (2016)

1.4 Analisi descrittive

I soggetti inclusi nello studio sono 288, il 76% dei quali è di genere femminile. L'età media è di 55 anni e varia dai 26 agli 87 anni. Le numerosità più elevate si registrano tra i 50 e i 66 anni. Il 68% degli individui è obeso, mentre quelli in sovrappeso e normopeso sono rispettivamente il 22% e il 10%. Un solo soggetto è sottopeso. I fumatori costituiscono il 31% del campione, mentre il 59% dei soggetti ha livello alto/fuori controllo di emoglobina

glicata. Una successiva analisi viene svolta con lo scopo di evidenziare com'è distribuita la gravità della malattia (indicata con PD) nei vari siti: l'intera dentatura è stata suddivisa in quadranti (primo: siti 1-42, secondo: siti 43-84, terzo: 85-126, quarto: 127-168), sono stati evidenziati i siti appartenenti alla mascella e quelli relativi alla mandibola. La Figura 1.5 riporta la proporzione

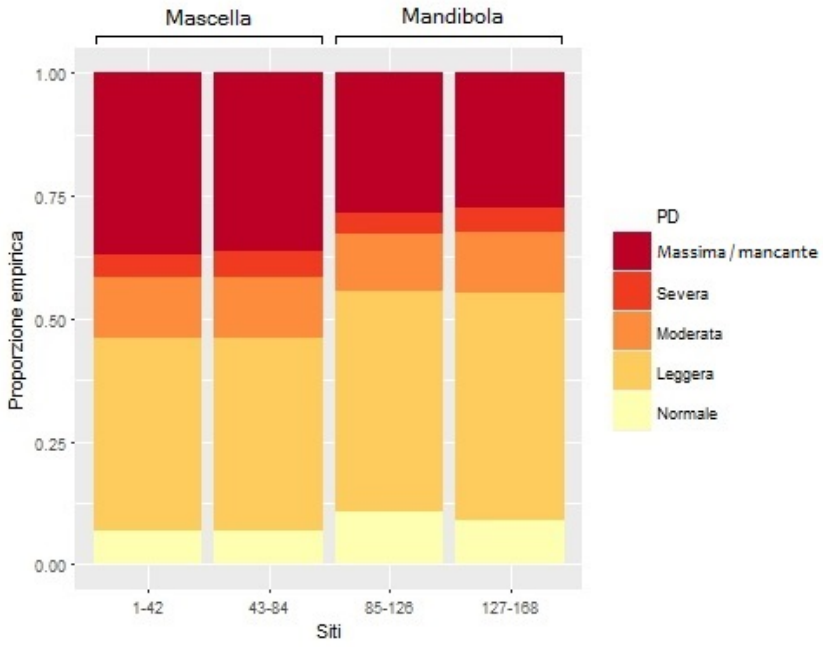

Figura 1.5: Proporzione empirica di gravità della parodontite suddivisa per quadranti della dentatura

empirica delle cinque categorie di gravità della parodontite. Notiamo come le proporzioni non siano omogenee se si confrontano mascella e mandibola, all'interno di tali suddivisioni invece le proporzioni sembrano essere simili. A dominare è la forma leggera di parodontite, la meno frequente è quella relativa al livello severo. Vi sono più denti mancanti nella parte superiore rispetto a quella inferiore della dentatura. Si è deciso in seguito di utilizzare un grafico analogo per valutare eventuali differenze di genere e per il livello di emoglobina glicata (HbA1c).

In Figura 1.6 si osserva come le proporzioni siano sistematicamente superiori per i maschi rispetto alle femmine. E' ancora evidente la non omogeneità tra mascella e mandibola per entrambi i generi. Gli uomini presentano un gra-

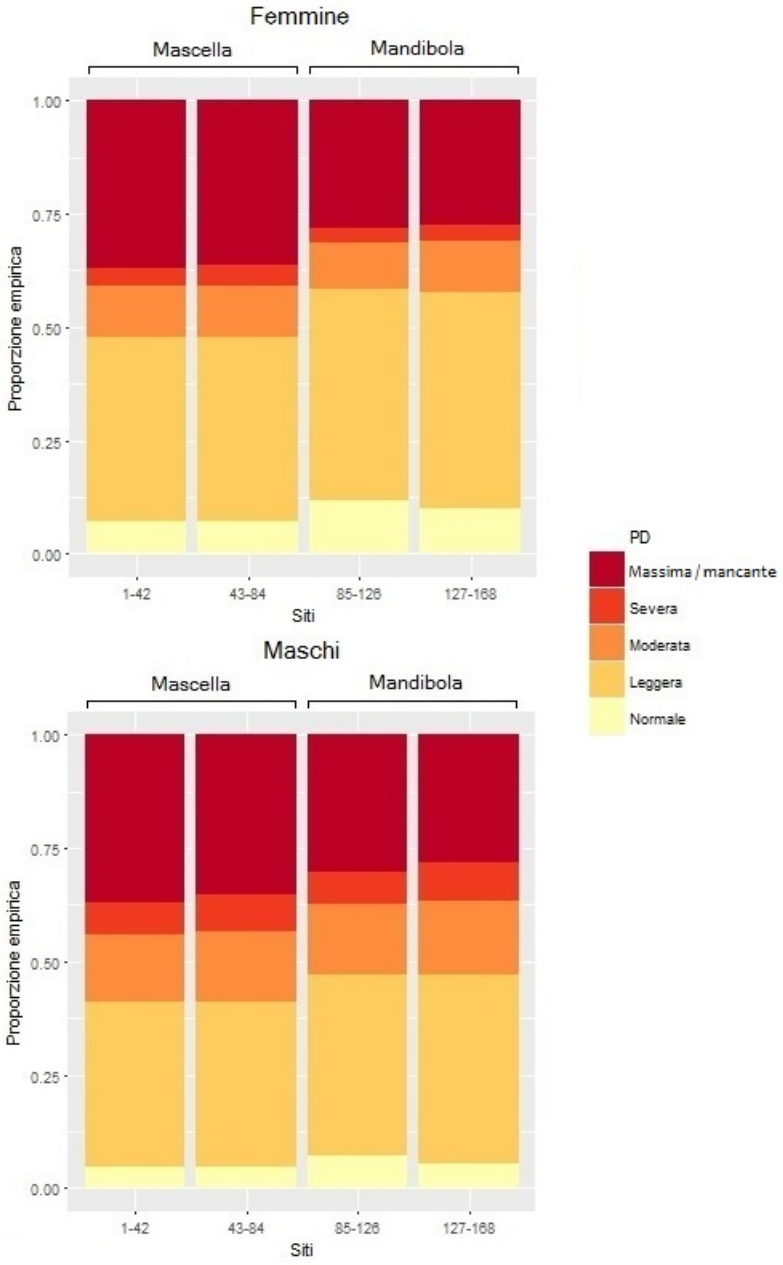

Figura 1.6: Proporzione empirica per genere di gravità della parodontite suddivisa per quadranti della dentatura

do moderato e severo della malattia superiore, in proporzione, rispetto alle donne. Le proporzioni relative al livello di emoglobina glicata sono diverse e non omogenee: per i soggetti il cui livello è fuori controllo, i casi di parodon-

tite severa e moderata sono superiori rispetto ai soggetti per i quali il livello è nella norma. (Figura 1.7).

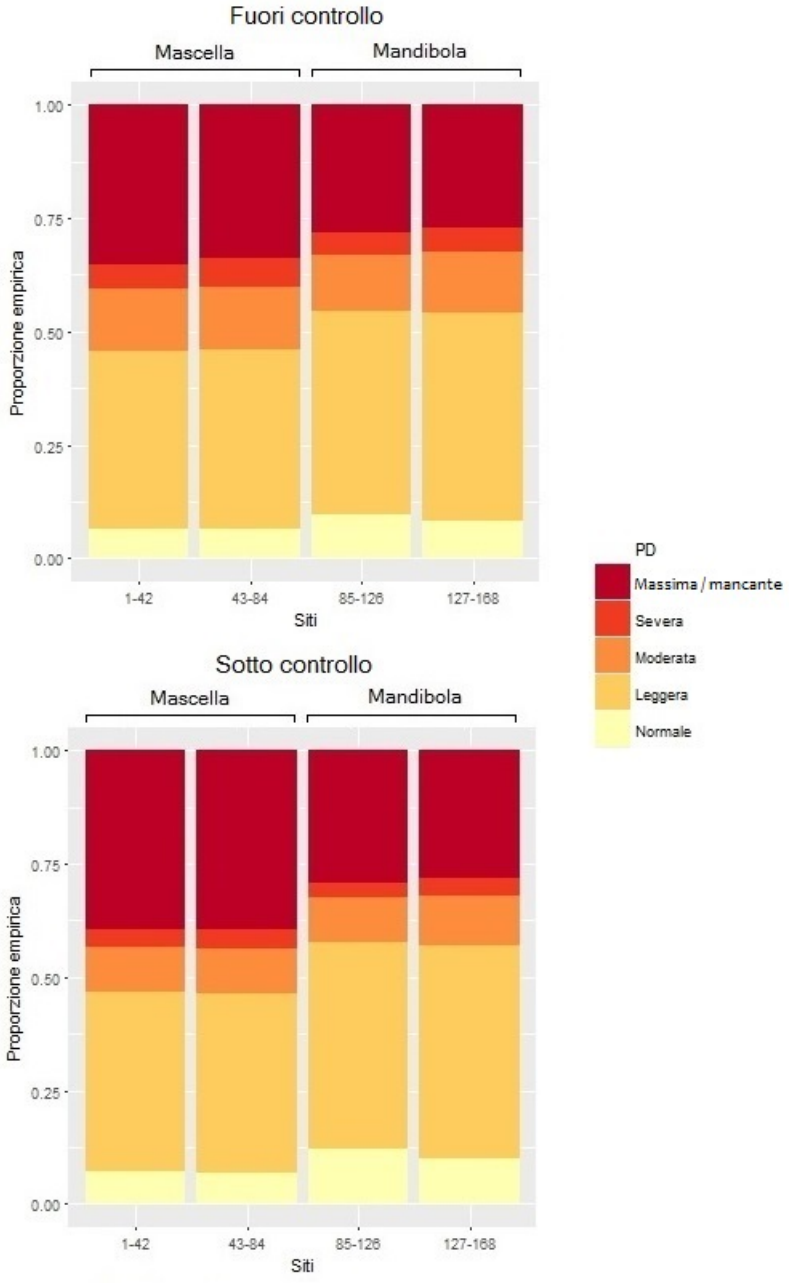

Figura 1.7: Proporzione empirica per livelli di emoglobina glicata di gravità della parodontite suddivisa per quadranti della dentatura

1.5 Obiettivo dell'analisi

Ci si pone l'obiettivo di definire la relazione tra la parodontite e le diverse covariate a livello di sito e soggetto, cercando di capire le caratteristiche per le quali il peggioramento della malattia è maggiore; in base alle misure di CAL rilevate sui soggetti. Si descriveranno anche le locazioni dei siti maggiormente esposti alla malattia, distinguendo quelli situati in posizione centrale o meno. Non vengono semplicemente sommate tra di loro tutte le $m = 168$ misure relative ad ogni soggetto considerando un modello univariato, poiché si ignorerebbe la struttura ordinale della gravità della malattia e il fenomeno latente che la spiega. Inoltre, si introduce una struttura di correlazione nella matrice di varianze e covarianze relativa alle misurazioni effettuate su ogni soggetto, secondo la quale siti adiacenti nello stesso dente presentano misure di CAL molto dipendenti tra di loro e meno correlate a quelle degli altri denti.

Capitolo 2

Modello statistico ed inferenza bayesiana

2.1 Modello *probit* ordinale multivariato

Il modello *probit* ordinale multivariato è una generalizzazione del modello *probit* ordinale univariato. Si indichi con $i = (1, ..., n)$ l'insieme degli individui, $n = 288$, con $j = (1, ..., m)$ l'insieme delle rilevazioni di CAL relative all'i - esimo individuo, $m = 168$. Nel caso univariato, l'indice j viene posto uguale ad uno, ad ogni individuo è associata quindi una sola rilevazione. Nel caso multivariato, invece, la matrice delle rilevazioni complessive $\mathbf{Y}$ è di dimensione $m \times n$: y_{ij} è relativa alla j-esima rilevazione riferita all'i-esimo individuo.

La variabile risposta y_{ij} può assumere uno tra i possibili valori ordinati, o categorie, k_w, dove w rappresenta il numero di diverse categorie diverse per ogni rilevazione. In questo caso specifico, si pone $k_w = k$. La variabile risposta y_{ij} è relativa allo stato di salute dei siti per ciascun dente. Per ogni sito l'intervallo dei valori di CAL misurati varia da 0 a 16 mm. E' comune adottare una discretizzazione della variabile risposta in K categorie ordinate in cui la prima categoria indica un ottimo stato di salute del sito, mentre l'ultima indica un pessimo stato di salute del sito oppure un dente mancante, che si presume sia dovuto alla gravità della malattia e non a cause accidentali. Si segue la discretizzazione descritta in Figura 1.1. L'ordinamento delle categorie è indotto quindi da una variabile latente continua sottostante che

può essere interpretata come il peggioramento dello stato di salute del sito. Inoltre, non si impone che le categorie siano delimitate da valori di soglia equispaziati tra di loro; poiché, ad esempio, la differenza reale di gravità che sussiste tra la categoria 2 e 3 non è detto sia simile a quella presente tra la categoria 3 e 4. Questo perché la conoscenza dello stato di salute è limitata all'ordine delle categorie e non all'esatta progressione di quest'ultimo, che è un fenomeno di natura continua e latente, non tipicamente costante.

Si definisce Y_i^* una variabile latente normale m-variata per l'i-esimo individuo. Nel modello *probit* univariato, la variabile latente presenta dei valori di soglia entro i quali è compresa, che corrispondono a diversi valori di y_i (Johnson e Albert 1999). Estendendo al caso multivariato, ogni realizzazione y_i si ottiene dalla relativa variabile latente Y_i^* nel modo seguente: si definisce $A_i = (\alpha_1, ..., \alpha_{k-1})$ l'insieme dei possibili valori di soglia, ogni rilevazione avente K possibili categorie avrà quindi $K-1$ soglie. Si pone $\alpha_0 = -\infty$ e $\alpha_K = \infty$ per ciascun soggetto, per motivi di identificabilità del modello. Sempre per lo stesso motivo, è opportuno fissare almeno un valore di soglia, ad esempio $\alpha_1 = 0$. Si osserverà il valore $y_{ij} = k$ se $\alpha_{k-1} < Y_{ij}^* < \alpha_k$. La verosimiglianza per la variabile risposta si ottiene integrando nello spazio multidimensionale delle variabili latenti vincolate:

$$P(y_{ij} = k|X_i, Z_j, \beta, \gamma, \mathbf{\Sigma}) = \int_{A_{iT}} \dots \int_{A_{i1}} \phi_T(Y_i^*|X_i, Z_j, \beta, \gamma, \mathbf{\Sigma}) dY_1^* \dots dY_T^* \tag{2.1}$$

Vengono indicati con $i = 1, ..., n$ gli individui e con $j = 1, ..., T$ le rilevazioni, $\boldsymbol{Y}_i$ è quindi un vettore T-dimensionale e ciascun elemento può assumere valori in $k = 1, ..., K-1$, A_{ik} è il relativo intervallo $(\alpha_{i,k-1}, \alpha_{i,k})$, β e γ sono i coefficienti di regressione, $\mathbf{\Sigma}$ rappresenta invece la matrice di varianze e covarianze che contiene le dipendenze spaziali tra i siti di rilevazione, X_i e Z_j sono le covariate e $\phi_T(Y_i^*|X_i, Z_j, \beta, \gamma, \mathbf{\Sigma})$ la funzione di densità della distribuzione normale multivariata.

Nell'analisi qui presentata si assume che le soglie siano le stesse per tutti gli individui, poiché il processo latente sottostante di interesse, ovvero il peggioramento della salute del sito, è sempre il medesimo per tutti gli individui. Dato la natura gerarchica dei dati, nella sezione successiva verrà specificata questa struttura in modo più dettagliato.

2.2 Specificazione gerarchica

Il modello da cui si prende spunto è quello definito in Bandyopadhyay e Canale (2016), dal quale si adotta la seguente specificazione: la matrice delle covariate a livello dell'individuo è $\boldsymbol{X}_{n\times p}$ ($p = 5$), dove l'i-esima riga $\boldsymbol{x}_i$ rappresenta un vettore di p covariate per il soggetto i, mentre la matrice delle covariate a livello di sito è indicata con $\boldsymbol{Z}_{m\times q}$ ($q = 2$), dove la j-esima riga $\boldsymbol{z}_j$ rappresenta un vettore di q covariate per il sito j. Un modello tipicamente utilizzato in un contesto di variabile risposta a più categorie, e quello multinomiale, definito per tutte le m rilevazioni di un soggetto, $\boldsymbol{Y}_i$, come $Y_{ij} \overset{ind}{\sim} Mult(1, \pi_{0ij}, ..., \pi_{Kij})$, dove $\sum_{k=0}^{K-1} \pi_{kij} = 1$.

Nel caso in esame viene violato l'assunto di indipendenza tra le rilevazioni, assunzione presente nel modello multinomiale, poiché esiste una non trascurabile forma di correlazione spaziale tra i siti di rilevazione, fondamentale da considerare.

Si assume una distribuzione multivariata relativa a tutte le m rilevazioni fatte su un soggetto: questa distribuzione dipende dal valore atteso μ_i, $i = 1, ..., n$, da una matrice di varianze e covarianze $\boldsymbol{\Sigma}$ (nella quale la dipendenza spaziale tra i siti di rilevazione è regolata dal parametro di lisciamento ρ) e dalle soglie α_k, $k = 1, ..., K-1$. Il valore atteso μ_i dipende, a sua volta, dalle covariate $\boldsymbol{\beta}$, a livello del soggetto e $\boldsymbol{\gamma}$, a livello del sito. In Figura 2.1 viene riportato un grafico relativo al modello presentato in precedenza, in cui si indicano con b_0, b_1, g_0, g_1, r_0 e r_1 gli iper-parametri fissati. Si definisce il seguente modello, relativamente ad ogni soggetto:

$$y_i = k \quad \Leftrightarrow \quad \alpha_{k-1} < Y_i^* < \alpha_k \tag{2.2}$$

$$Y_i^* = \mu_i + \varepsilon_i \tag{2.3}$$

$$\mu_i = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} \tag{2.4}$$

$$\varepsilon_i \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Sigma}) \tag{2.5}$$

Il termine di errore ha distribuzione normale m-variata di media nulla e matrice di varianze e covarianze $\boldsymbol{\Sigma} = \boldsymbol{D} - \rho\boldsymbol{W}$. Si definisca l'insieme dei siti di rilevazione per ogni soggetto come $\mathbf{s} = (s_1, ..., s_m)$. La matrice $\boldsymbol{\Sigma}$ è la differenza tra una matrice diagonale $\boldsymbol{D}$ in cui il j-esimo elemento rappresenta

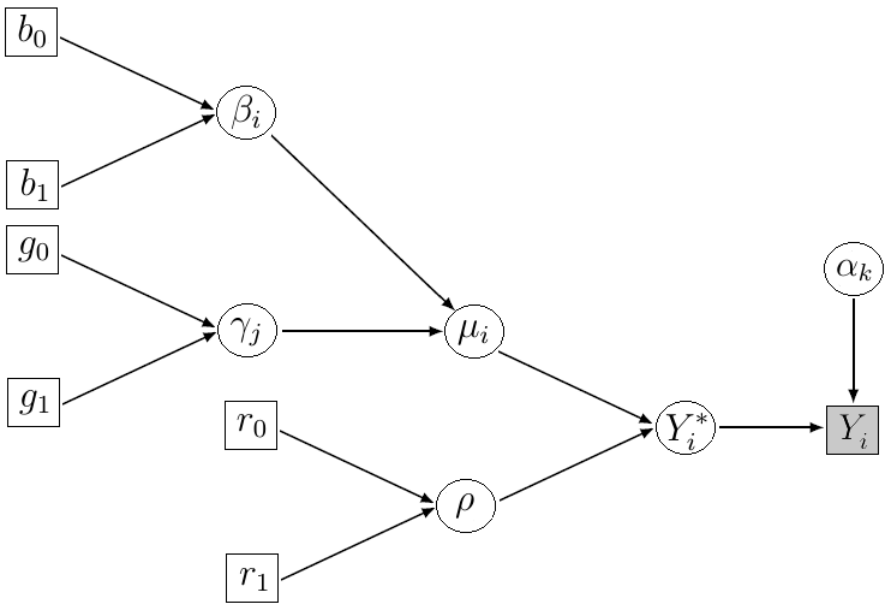

Figura 2.1: Struttura gerarchica bayesiana del modello

il numero di siti vicini alla locazione s_j, $\boldsymbol{W}$ è la matrice di adiacenza in cui $w_{jj'} = 1$ se s_j è un vicino di $s_{j'}$, altrimenti $w_{jj'} = 0$; infine ρ è un parametro di lisciamento che regola il grado di dipendenza spaziale. In particolare, vi sarà alta dipendenza tra siti vicini nello stesso dente e bassa dipendenza tra siti più lontani e in denti diversi.

2.2.1 *Data augmentation* e verosimiglianza aumentata

La funzione di verosimiglianza del modello *probit* ordinale multivariato, condizionatamente ai valori osservati, è la seguente:

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \rho | \boldsymbol{X}, \boldsymbol{Z}, Y_i) = \prod_{i=1}^{n} \left[\Pr\left(Y_{i1} = y_{i1}, ..., Y_{im} = y_{im} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \rho\right) \right] \quad (2.6)$$

Essa, nel caso del modello *probit* ordinale multivariato, è difficile da trattare. Inoltre, definire un algoritmo di campionamento di *Gibbs*, che comporta il calcolo delle distribuzioni condizionate, mantenendo la funzione di verosimiglianza *probit* risulta poco pratico. Albert e Chib (1993) utilizzano un approccio simile nel caso univariato, che verrà adattato al caso multivariato in esame. Le variabili latenti vengono trattate come valori mancanti, si uti-

lizza la tecnica del *data augmentation* per condurre l'inferenza. È un metodo introdotto da Tanner e Wong (1987), principalmente per motivi di carattere computazionale. Questa tecnica si basa su un algoritmo iterativo: si supponga che Y siano i dati osservati e θ l'ignoto parametro di interesse. Se l'interesse è quello di simulare dalla distribuzione $f(Y|\theta)$, l'idea è di introdurre la variabile latente Y^*, in modo che sia semplice simulare dalla distribuzione congiunta $f(Y, Y^*|\theta)$. La distribuzione di partenza può essere ottenuta marginalizzando rispetto alla variabile latente:

$$f(Y|\theta) = \int f(Y, Y^*|\theta)\, dY^* \tag{2.7}$$

L'algoritmo itera tra un passo di imputazione, nel quale vengono simulate le variabili latenti e un passo di stima a posteriori, fino a raggiungere la convergenza. I campioni dell'ignoto parametro θ possono essere utilizzati in fase di inferenza. Nel seguente algoritmo $\boldsymbol{\theta}$ indica il vettore di parametri $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \rho)$.

Algoritmo 1 *Data augmentation*

Iterazione t-esima

1. Simulare $Y^* \sim f(Y^*|\boldsymbol{\theta}, Y) \propto f(Y, Y^*|\boldsymbol{\theta})$

2. Simulare ogni $\theta_t \sim f(\theta_t|Y^*, Y) \propto f(Y, Y^*|\theta_t) f(\theta_t)$

La verosimiglianza aumentata risulta quindi essere:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \rho, Y_i^*|\boldsymbol{X}, \boldsymbol{Z}, Y_i) = &\prod_{i=1}^{n}\prod_{j=1}^{m} \mathbb{1}(y_{ij} = k)\mathbb{1}(\alpha_{k-1} < Y_{ij}^* < \alpha_k) \\ &\Pr\left[Y_{ij}^* \in (\alpha_{k-1}, \alpha_k)\right]\end{aligned} \tag{2.8}$$

2.3 Distribuzioni a priori

Sono state definite le seguenti distribuzioni a priori per i parametri di interesse: $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$, sono i coefficienti associati alle variabili relative al soggetto, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$, sono i coefficienti associati alle variabili relative al sito e ρ è il parametro di lisciamento. Le a priori sono così definite:

$$\begin{aligned} \boldsymbol{\beta}' &\sim \mathcal{N}_p(\mathbf{0}, 10\boldsymbol{I}_p) \\ \boldsymbol{\gamma}' &\sim \mathcal{N}_q(\mathbf{0}, 10\boldsymbol{I}_q) \\ \rho &\sim \mathcal{U}(0.65, 1) \end{aligned} \tag{2.9}$$

Per i parametri relativi alle soglie $\boldsymbol{\alpha} = (\alpha_2, ..., \alpha_4)$ si è deciso di specificare una distribuzione non informativa. Le distribuzioni a priori dei parametri $\boldsymbol{\gamma}$ e $\boldsymbol{\beta}$ sono non informative, centrate sul valore nullo e con varianza pari a 10. Il parametro di lisciamento, che può oscillare nel *range* $[-1, 1]$, è supposto avere una distribuzione a priori in un intervallo abbastanza stretto, più questo valore si avvicina ad 1 e più la dipendenza è forte.

2.4 Distribuzione a posteriori

La distribuzione a posteriori aumentata, introducendo le variabili latenti, è data dalla seguente espressione:

$$\begin{aligned} \pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \rho | \boldsymbol{X}, \boldsymbol{Z}, Y_i^*) = &\prod_{i=1}^{n} \Pr\left(Y_{i1} = y_{i1}, ..., Y_{im} = y_{im} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \rho\right) \times \\ &\times \pi(\boldsymbol{\alpha})\pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma})\pi(\rho) \end{aligned} \tag{2.10}$$

La prima parte è relativa alla funzione di verosimiglianza, la seconda parte è relativa alle distribuzioni a priori dei parametri di interesse. Sostituendo le funzioni generali con le distribuzioni a priori e la funzione di verosimiglianza aumentata specificata in precedenza, si ottiene la seguente distribuzione a

posteriori:

$$\begin{aligned}\pi(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\rho|\boldsymbol{X},\boldsymbol{Z},Y_i^*) &= \prod_{i=1}^{n}\prod_{j=1}^{m}\mathbb{1}(y_{ij}=k)\mathbb{1}(\alpha_{k-1}<Y_{ij}^*<\alpha_k)\times \\ &\times \Pr\left[Y_{ij}^*\in(\alpha_{k-1},\alpha_k)\right]\frac{1}{\sqrt{|20\pi\boldsymbol{I}_p|}}\exp\left\{-\frac{1}{2}\boldsymbol{\beta}'\left(\frac{1}{10}\boldsymbol{I}_p\right)\boldsymbol{\beta}\right\}\times \\ &\times \frac{1}{\sqrt{|20\pi\boldsymbol{I}_q|}}\exp\left\{-\frac{1}{2}\boldsymbol{\gamma}'\left(\frac{1}{10}\boldsymbol{I}_q\right)\times\boldsymbol{\gamma}\right\}\mathbb{1}(0.65<\rho<1)\times \\ &\times \mathbb{1}(\alpha_{k-1}<\alpha_k<\alpha_{k+1})\end{aligned} \tag{2.11}$$

2.5 Algoritmo *Metropolis-within-Gibbs*

Verranno di seguito riportate le distribuzioni *full conditional* per i parametri di interesse, mentre nella sezione successiva si descriverà il *Gibbs sampling* nel quale è stato introdotto un passo di *Metropolis-Hastings* per simulare il parametro di lisciamento ρ, dato che la sua distribuzione non è in forma chiusa.

2.5.1 Distribuzioni *full conditional*

1. Variabili latenti Y^*

Il primo passo consiste nel simulare le variabili latenti Y_i^* da una variabile normale multivariata troncata, condizionata ai valori dei parametri $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\rho$, di media μ_i e matrice di precisione $\boldsymbol{\Omega}=\boldsymbol{\Sigma}^{-1}$. La generazione viene effettuata attraverso un *Gibbs sampling* che campiona le osservazioni dalle distribuzioni univariate condizionate, anch'esse normali troncate. Quindi, tutti i campioni vengono accettati ad eccezione dei primi valori, che vengono scartati come periodo di *burn-in*. Per un approfondimento si veda l'algoritmo in (Kotecha e Djuric 1999). La distribuzione a posteriori aumentata risulta pertanto:

$$\begin{aligned} Y_i^* &\sim \mathcal{N}_m(\mu_i, \boldsymbol{\Sigma})\mathbb{1}_{A_i} \\ A_i &= \bigotimes_{j=1}^{m} \Big[\alpha_{y_{ij-1}}, \alpha_{y_{ij}}\Big] \end{aligned} \tag{2.12}$$

2. Parametri di regressione $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$

Le distribuzioni condizionate per i parametri di regressione si ricavano secondo gli stessi passaggi algebrici, differiscono solo nelle dimensioni: nel caso dei coefficienti $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ la *full conditional* sarà una normale p-variata, mentre per $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ sarà una normale q-variata.

$$\begin{aligned} \pi(\boldsymbol{\beta}|Y_i^*, \boldsymbol{X}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \rho) &\propto \\ &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{|20\pi \boldsymbol{I}_p|}} \exp\Big\{ -\frac{1}{2}\boldsymbol{\beta}'\Big(\frac{1}{10}\boldsymbol{I}_p\Big)\boldsymbol{\beta}\Big\} \times \\ &\quad \times \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\Big\{ -\frac{1}{2}\sum_{i=1}^{n} \big(Y_i^* - \mu_i\big)'\boldsymbol{\Sigma}^{-1}\big(Y_i^* - \mu_i\big)\Big\} \\ &\propto \exp\Big\{ -\frac{1}{2}\boldsymbol{\beta}'\Big(\frac{1}{10}\boldsymbol{I}_p\Big)\boldsymbol{\beta}\Big\} - 2\boldsymbol{\beta}'\boldsymbol{X}'\big(\boldsymbol{Y}'^* - \boldsymbol{\mu}\big) + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\ &\propto \exp\Big\{ -\frac{1}{2}\Big[\boldsymbol{\beta}'\Big(\frac{1}{10}\boldsymbol{I}_p + \boldsymbol{X}'\boldsymbol{X}\Big)\boldsymbol{\beta}\Big] + \boldsymbol{\beta}'\boldsymbol{X}'\big(\boldsymbol{Y}'^* - \boldsymbol{\mu}\big)\Big\} \end{aligned} \tag{2.13}$$

Si può riconoscere il nucleo di una densità normale multivariata di media μ_β e varianza V_β così definite:

$$\begin{aligned} V_\beta &= \Big(m\boldsymbol{X}'\boldsymbol{X} + \frac{1}{10}\boldsymbol{I}_p\Big)^{-1} \\ \mu_\beta &= V_\beta\Big(\sum_{j=1}^{m} \boldsymbol{X}'Y_j'^*\Big) \end{aligned} \tag{2.14}$$

Si è definito con $Y_j'^* = y_{ij}'^* - \boldsymbol{Z}\boldsymbol{\gamma}$ il vettore delle variabili latenti a cui è stato sottratto il prodotto delle covariate per i coefficienti relativi ai siti, per ognuno degli $n = 288$ soggetti, dato che tale distribuzione non dipende da $\boldsymbol{\gamma}$.

Analogamente si calcola la distribuzione *full conditional* per i coefficienti $\boldsymbol{\gamma}$.

$$\begin{aligned}
\pi(\boldsymbol{\gamma}|Y_i^*, \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \rho) \propto \\
&\propto \prod_{i=1}^{n} \frac{1}{\sqrt{|20\pi \boldsymbol{I}_q|}} \exp\left\{ -\frac{1}{2}\boldsymbol{\gamma}'\left(\frac{1}{10}\boldsymbol{I}_q\right)\boldsymbol{\gamma}\right\} \times \\
&\quad \times \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n}\left(Y_i^* - \mu_i\right)'\boldsymbol{\Sigma}^{-1}\left(Y_i^* - \mu_i\right)\right\} \\
&\propto \exp\left\{ -\frac{1}{2}\boldsymbol{\gamma}'\left(\frac{1}{10}\boldsymbol{I}_q\right)\boldsymbol{\gamma}\right\} - 2\boldsymbol{\gamma}'\boldsymbol{Z}'\left(\boldsymbol{Y}'^* - \boldsymbol{\mu}\right) + \boldsymbol{\gamma}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\gamma} \\
&\propto \exp\left\{ -\frac{1}{2}\left[\boldsymbol{\gamma}'\left(\frac{1}{10}\boldsymbol{I}_q + \boldsymbol{Z}'\boldsymbol{Z}\right)\boldsymbol{\gamma}\right] + \boldsymbol{\gamma}'\boldsymbol{Z}'\left(\boldsymbol{Y}'^* - \boldsymbol{\mu}\right)\right\}
\end{aligned} \tag{2.15}$$

Si può riconoscere il nucleo di una densità normale multivariata di media μ_γ e varianza V_γ così definite:

$$\begin{aligned}
V_\gamma &= \left(n\boldsymbol{Z}'\boldsymbol{Z} + \frac{1}{10}\boldsymbol{I}_q\right)^{-1} \\
\mu_\gamma &= V_\gamma\left(\sum_{i=1}^{n}\boldsymbol{Z}'Y_j'^*\right)
\end{aligned} \tag{2.16}$$

Si è definito con $Y_i'^* = y_{ij}'^* - \boldsymbol{X}\boldsymbol{\beta}$ il vettore delle variabili latenti a cui è stato sottratto il prodotto delle covariate per i coefficienti relativi ai soggetti, per ognuno degli $m = 168$ siti; dato che questa distribuzione non dipende da $\boldsymbol{\beta}$. Le distribuzioni *full conditional* sono così definite:

$$\pi(\boldsymbol{\beta}|Y_i^*, \boldsymbol{X}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \rho) \sim \mathcal{N}_p(\mu_\beta, V_\beta) \tag{2.17}$$

$$\pi(\boldsymbol{\gamma}|Y_i^*, \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \rho) \sim \mathcal{N}_q(\mu_\gamma, V_\gamma) \tag{2.18}$$

3. Parametri relativi alle soglie $\alpha_2, \alpha_3, \alpha_4$

Come già detto in precedenza, il dominio delle variabili latenti è l'insieme dei numeri reali, quindi vengono poste $\alpha_0 = -\infty$ e $\alpha_K = \infty$. Inoltre, per rendere i parametri del modello identificabili, è opportuno fissare un'altra soglia tra le rimanenti: si è optato per definire $\alpha_1 = 0$. Restano quindi da ottenere le distribuzioni *full conditional* per i parametri $\alpha_2, \alpha_3, \alpha_4$. Si indica

con k il generico valore della soglia k-esima α_k, $k = 2, 3, 4$, la distribuzione *full conditional* ad essa relativa è di seguito riportata:

$$\begin{aligned}&\pi(\alpha_k|Y_i^*, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \alpha_{l,\{l\neq k\}}, \rho) \propto \\ &\propto \prod_{i=1}^{n}\prod_{j=1}^{m} \mathbb{1}(y_{ij} = k)\mathbb{1}(\alpha_{k-1} < Y_{ij,k}^* < \alpha_k)\, \mathbb{1}(y_{ij} = k+1)\times \\ &\quad \times \mathbb{1}(\alpha_k < Y_{ij,k+1}^* < \alpha_{k+1})\end{aligned} \tag{2.19}$$

Nel caso univariato, Albert e Chib (1993) definiscono una distribuzione uniforme compresa tra il più grande valore simulato della variabile latente per un soggetto che fa parte della categoria precedente alla soglia di interesse e il più piccolo valore simulato della variabile latente per un soggetto appartenente alla categoria successiva alla soglia di interesse.

Nel caso multivariato si ottiene questa distribuzione secondo un ragionamento analogo: ogni individuo contribuisce con un solo termine alla distribuzione a posteriori, ossia quando tutte le sue funzioni indicatrici sono pari ad 1. Se consideriamo quindi una particolare soglia α_k, tutti i termini che non coinvolgono α_k possono essere rimossi come costanti di proporzionalità, ottenendo la distribuzione (2.19). La prima funzione indicatrice è ripetuta per tutte le latenti comprese nell'intervallo (α_{k-1}, α_k), mentre la seconda funzione indicatrice è ripetuta per tutte le latenti comprese tra (α_k, α_{k+1}). Poichè il prodotto di funzioni indicatrici è non nullo se e solo se tutte le funzioni indicatrici assumono valore 1, possono essere omesse: se $\alpha_k > Y_{i,k-1}^*, \forall\, Y_{i,k-1}^*$ allora $\alpha_k > \max(Y_{i,k-1}^*)$. Analogamente, se $\alpha_k < Y_{i,k}^*, \forall\, Y_{i,k}^*$ allora $\alpha_k < \min(Y_{i,k}^*)$. La forma della distribuzione in questo intervallo è costante, si tratta quindi di una distribuzione uniforme. Poiché le distribuzioni *full conditional* per le soglie in una dimensione non dipendono dalle soglie nelle altre dimensioni, possiamo simulare una soglia alla volta indipendentemente dalle altre. Affinché venga mantenuto l'ordinamento crescente delle soglie $(\alpha_0 < \alpha_1 < \ldots < \alpha_{k-1} < \alpha_K)$, si definisce la distribuzione come segue:

$$\begin{aligned}&\pi(\alpha_k|Y^*, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \alpha_{l,\{l\neq k\}}, \rho) \propto \mathcal{U}\left[\, \max(\max(Y_{i,k}^* : Y_{i,k} = k), \alpha_{k-1}),\right. \\ &\left.\min(\min(Y_{i,k}^* : Y_{i,k} = k+1), \alpha_{k+1})\,\right]\end{aligned} \tag{2.20}$$

Tuttavia, nonostante sia immediato simulare da questa distribuzione, la differenza tra i due estremi dell'uniforme risulta essere piccola, perciò i valori simulati saranno molto vicini tra di loro. Il seguente argomento è stato affrontato in Lynch (2007), dove si evidenziano problemi di convergenza e di *mixing*, soprattutto per numerosità campionarie elevate. Inoltre, i valori simulati risultano essere molto autocorrelati tra di loro, con correlazioni significative anche a ritardi elevati. Una soluzione viene presentata in Cowles (1996), in cui la simulazione dalla distribuzione uniforme viene rimpiazzata da un passo di *Metropolis-Hastings*. Vengono proposti dei candidati valori di soglia nell'intero intervallo delle soglie adiacenti a quella di interesse da una distribuzione normale troncata, utilizzando il criterio di accettazione-rifiuto per decidere se accettare il valore proposto. Il miglioramento è dovuto al fatto che, sebbene non tutti i valori ottenuti vengano accettati, le soglie effettuano spostamenti maggiori; portando ad un abbassamento dell'autocorrelazione tra ritardi successivi, vi è un marcato miglioramento del *mixing* ed infine ogni componente arriva più rapidamente convergenza. Tuttavia, si è preferito adottare l'approccio di Albert e Chib (1993), poiché la distribuzione da cui si simula è disponibile in forma chiusa e tutti i valori proposti vengono sempre accettati.

4. Parametro di lisciamento ρ

Il parametro che regola il grado di dipendenza spaziale non presenta una distribuzione in forma chiusa:

$$\begin{aligned}\pi(\rho|Y^*, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \propto & \prod_{i=1}^{n} \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (Y_i^* - \mu_i)' \boldsymbol{\Sigma}^{-1} (Y_i^* - \mu_i) \right\} \times \\ & \times \mathbb{1}(0.65 < \rho < 1) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{|2\pi(\boldsymbol{D} - \rho\boldsymbol{W})|}} \times \\ & \times \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (Y_i^* - \mu_i)' (\boldsymbol{D} - \rho\boldsymbol{W})^{-1} (Y_i^* - \mu_i) \right\} \times \\ & \times \mathbb{1}(0.65 < \rho < 1)\end{aligned} \tag{2.21}$$

Si rende quindi necessario un passo di *Metropolis-Hastings*, un algoritmo di tipo MCMC particolarmente utile quando i parametri non presentano

distribuzioni in forma chiusa (Ntzoufras 2009). Una catena di *Markov* con passeggiata casuale è tale che $\rho^{(c+1)} = \rho^{(c)} + \varepsilon_t$, con ε_t indipendente da $\rho^{(c)}$. Propongo quindi un valore $\rho^{(c)}$ utilizzando come *proposal* un *random walk* uniforme. Una scelta comune è quella di porre

$$\rho^{(c)}|\rho^{(k-1)} \sim \mathcal{U}\left(\rho^{(k-1)} - \varepsilon_t, \rho^{(k-1)} + \varepsilon_t\right) \tag{2.22}$$

Il valore proposto viene accettato con probabilità

$$\alpha = [\min(1, t)] \tag{2.23}$$

In genere, si preferisce optare per l'algoritmo con passeggiata casuale uniforme, dato che permette di esplorare lo spazio degli stati attraverso una conoscenza locale della distribuzione. E' opportuno regolare opportunamente il valore di ε_t in modo da rendere efficace l'algoritmo. Si definisce t nel seguente modo, dove con c si indica il valore proposto mentre con $k-1$ il valore del parametro ρ relativo al passo precedente:

$$\begin{aligned}
t &= \frac{f(Y^*|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \rho^{(c)})}{f(Y^*|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \rho^{(k-1)})} \frac{\pi(\rho^{(c)})}{\pi(\rho^{(k-1)})} \\
&= \frac{\frac{1}{\sqrt{|2\pi(\boldsymbol{D} - \rho^{(c)}\boldsymbol{W})|}} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n}(Y_i^* - \mu_i)'(\boldsymbol{D} - \rho^{(c)}\boldsymbol{W})^{-1}(Y_i^* - \mu_i)\right\}}{\frac{1}{\sqrt{|2\pi(\boldsymbol{D} - \rho^{(k-1)}\boldsymbol{W})|}} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n}(Y_i^* - \mu_i)'(\boldsymbol{D} - \rho^{(k-1)}\boldsymbol{W})^{-1}(Y_i^* - \mu_i)\right\}} \\
&= \frac{|(\boldsymbol{D} - \rho^{(k-1)}\boldsymbol{W})|}{|(\boldsymbol{D} - \rho^{(c)}\boldsymbol{W})|} \frac{\exp\left\{ -\frac{1}{2}\sum_{i=1}^{n}(Y_i^* - \mu_i)'(\boldsymbol{D} - \rho^{(c)}\boldsymbol{W})^{-1}(Y_i^* - \mu_i)\right\}}{\exp\left\{ -\frac{1}{2}\sum_{i=1}^{n}(Y_i^* - \mu_i)'(\boldsymbol{D} - \rho^{(k-1)}\boldsymbol{W})^{-1}(Y_i^* - \mu_i)\right\}} \\
&= \frac{|\boldsymbol{\Sigma}^{(k-1)}|}{|\boldsymbol{\Sigma}^{(c)}|} \frac{\exp\left\{ -\frac{1}{2}\sum_{i=1}^{n}(Y_i^* - \mu_i)'\boldsymbol{\Sigma}^{-1(c)}(Y_i^* - \mu_i)\right\}}{\exp\left\{ -\frac{1}{2}\sum_{i=1}^{n}(Y_i^* - \mu_i)'\boldsymbol{\Sigma}^{-1(k-1)}(Y_i^* - \mu_i)\right\}}
\end{aligned} \tag{2.24}$$

L'algoritmo viene di seguito riportato:

Algoritmo 2 *Metropolis-within-Gibbs sampling*

1. Inizializzazione dei parametri $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\rho}$

   Iterazione t-esima

2. Simulare $Y^*_{t+1}|\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t, \rho_t$ dalla (2.12)
3. Simulare $\boldsymbol{\alpha}_{t+1}|Y^*_{t+1}, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t, \rho_t$ dalla (2.20)
4. Simulare $\boldsymbol{\beta}_{t+1}|Y^*_{t+1}, \boldsymbol{\alpha}_{t+1}, \boldsymbol{\gamma}_t, \rho_t$ dalla (2.17)
5. Simulare $\boldsymbol{\gamma}_{t+1}|Y^*_{t+1}, \boldsymbol{\alpha}_{t+1}, \boldsymbol{\beta}_{t+1}, \rho_t$ dalla (2.18)
6. Simulare ρ_{t+1} utilizzando un passo di *Metropolis-Hastings* dalla (2.24)
7. Reiterare dal passo 2

Capitolo 3

Applicazione ai dati e risultati

Nel seguente capitolo si applica il modello descritto nel Capitolo 2 ai dati. Si campionano $n = 125$ soggetti, che corrispondono ad un totale di 21000 osservazioni. Per l'inferenza a posteriori si sono effettuate 20000 iterazioni dell'algoritmo *Metropolis-within-Gibbs sampler*, di cui le prime 4000 sono state scartate come periodo di *burn-in*. L'algoritmo è computazionalmente oneroso e il tempo impiegato per l'intera esecuzione è stato di 17 ore, utilizzando una *workstation* fissa dotata di un processore AMD *A8-6500 APU with Radeon*$^{\text{TM}}$ *HD Graphics 3.50GHz* con 8Gb di memoria *RAM*. Il codice è stato scritto tramite il *software* R Core Team (2015) e viene riportato in Appendice A. Per la scelta degli iper-parametri delle distribuzioni a priori si è utilizzata la formulazione della sezione 2.3.

3.1 Convergenza del modello ed interpretazione dei risultati

Per analizzare se il modello raggiunge effettivamente la convergenza, si sono ispezionati i *trace plots* relativi ai parametri di soglia $\boldsymbol{\alpha}$, ai coefficienti di regressione $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ ed, infine, al parametro di lisciamento ρ che regola il grado di dipendenza spaziale tra i siti.

3.1.1 Covariate relative al soggetto

Come si può notare dalla Figura 3.1, tutti e tre i relativi *trace plots* dei parametri età, genere e BMI dell'individuo presentano un buon *mixing*, segno che le distribuzioni sono giunte a convergenza ed in maniera piuttosto rapida. Stesso discorso si può fare per i parametri relativi alla variabile indicatrice fumatore e per il livello di emoglobina glicata (Figura 3.2).

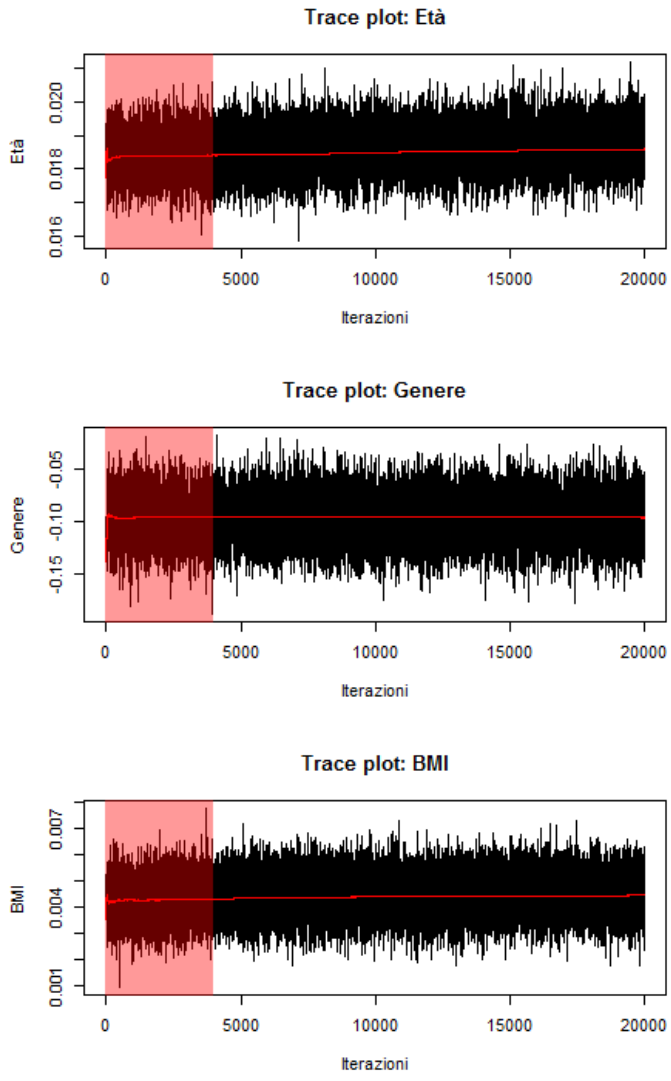

Figura 3.1: *Trace plots* per i coefficienti di regressione relativi alle covariate del soggetto: età, genere e BMI

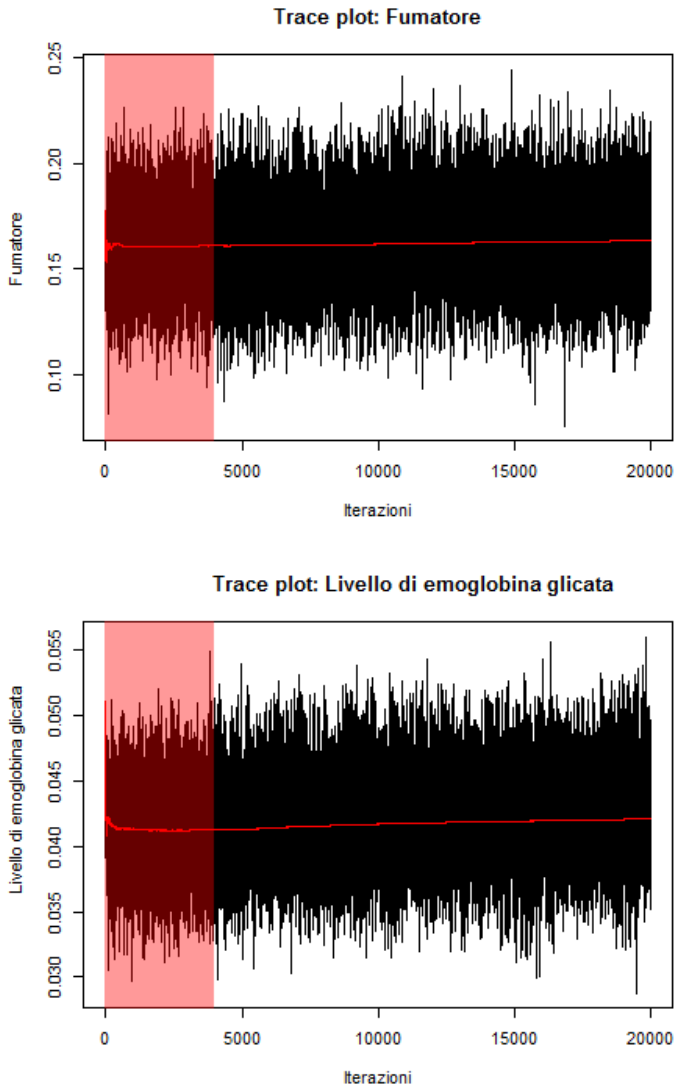

Figura 3.2: *Trace plots* per i coefficienti di regressione relativi alle covariate del soggetto: fumatore e livello emoglobina glicata

I grafici seguenti riportano invece le funzioni di autocorrelazione relative ai *trace plots* sopra presentati. Le caratteristiche rispecchiano quello attese: già dopo pochi ritardi l'autocorrelazione rientra negli intervalli di confidenza (Figura 3.3). L'andamento della funzione di autocorrelazione è analogo anche per le altre due covariate a livello di soggetto: fumatore e livello di emoglobina glicata (Figura 3.4).

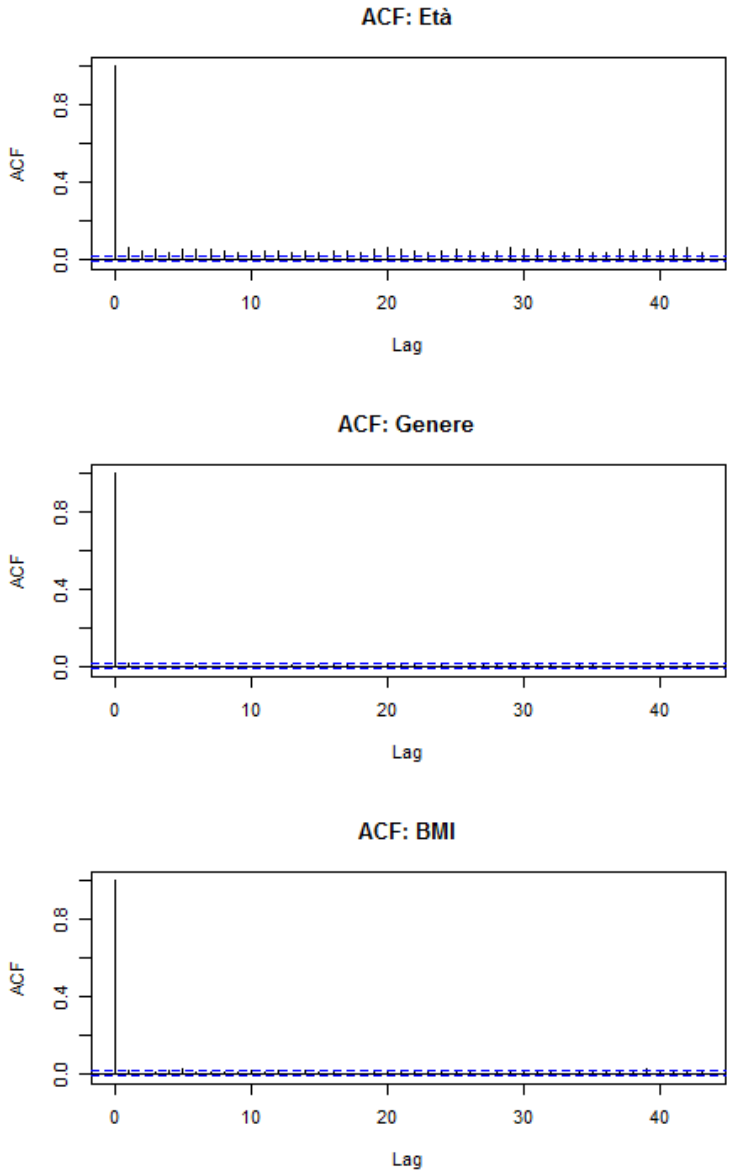

Figura 3.3: Funzione di autocorrelazione per i coefficienti di regressione relativi alle covariate del soggetto: età, genere e BMI

Si riporta, infine, l'oggetto più interessante dell'inferenza: le distribuzioni a posteriori relative ai cinque parametri del soggetto. Già dai grafici si ha un'indicazione del verso dell'associazione delle covariate con l'aumentare della gravità della parodontite.

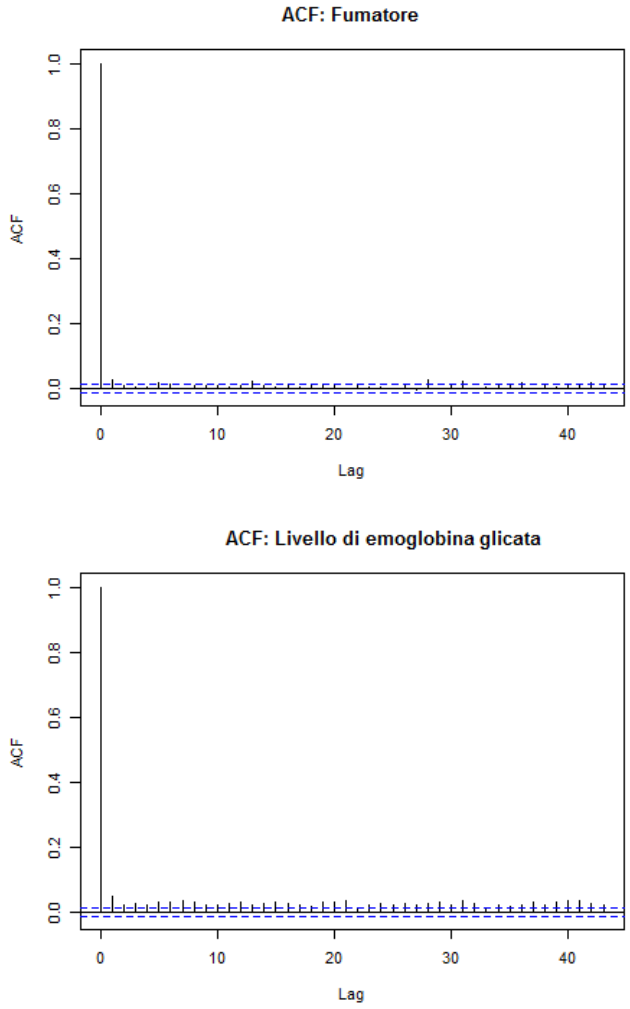

Figura 3.4: Funzione di autocorrelazione per i coefficienti di regressione relativi alle covariate del soggetto: fumatore e livello di emoglobina glicata

Nel primo grafico della Figura 3.5 si può notare come la parodontite si aggravi con l'avanzare dell'età. A soffrirne maggiormente sono i soggetti di sesso maschile, a conferma di quanto visto con le prime analisi descrittive. Inoltre, vi è un'associazione positiva tra parodontite e BMI.

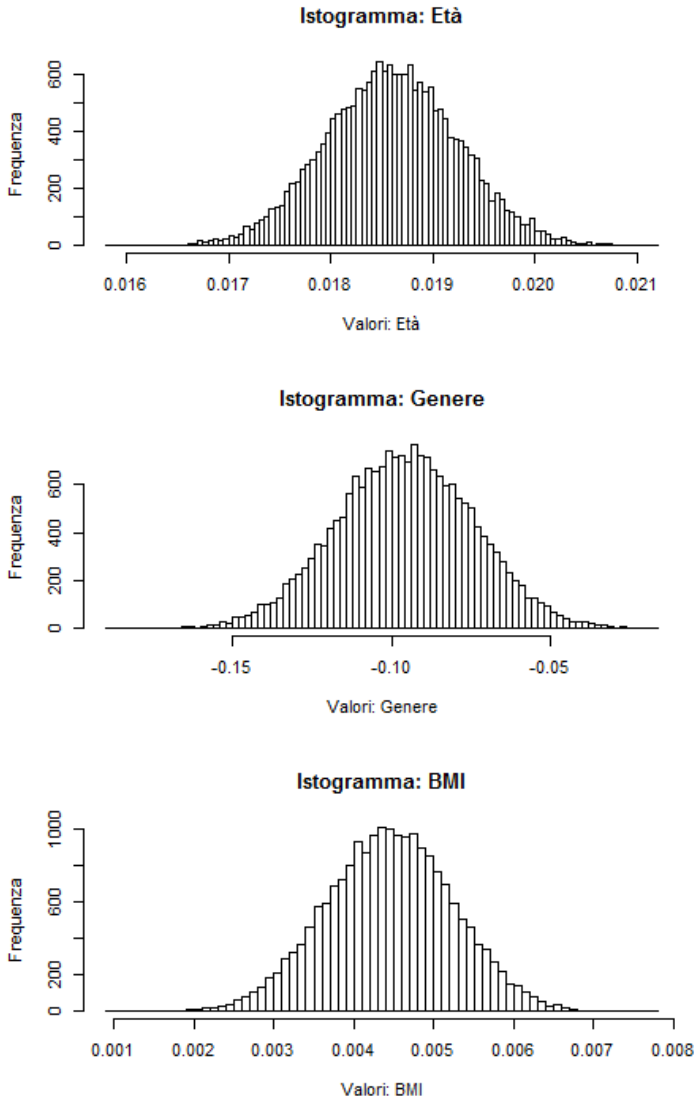

Figura 3.5: Distribuzioni a posteriori per i coefficienti di regressione relativi alle covariate del soggetto: età, genere e BMI

In Figura 3.6 si riportano le distribuzioni a posteriori per altre due covariate: fumatore e livello di emoglobina glicata. Vi è un'associazione positiva tra fumo e parodontite. Un alto livello di emoglobina glicata nel sangue comporta un deterioramento della salute del parodonto: i soggetti in esame soffrono tutti di questa malattia che è quindi direttamente collegata alla salute

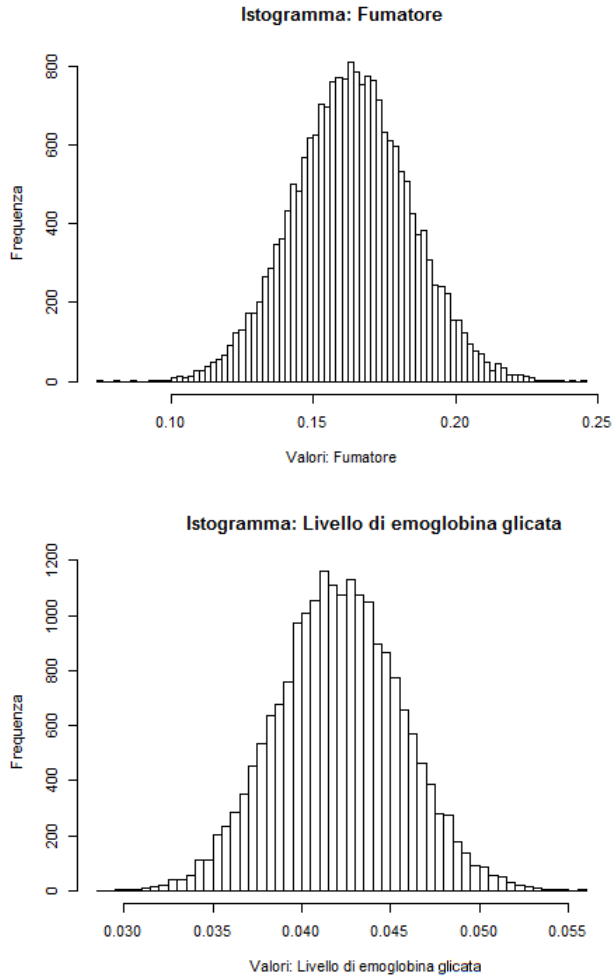

Figura 3.6: Distribuzioni a posteriori per i coefficienti di regressione relativi alle covariate del soggetto: fumatore e livello di emoglobina glicata

dei loro denti. Il fenomeno è noto in letteratura (Engebretson et al. 2013), uno studio in cui si evidenzia il fatto che il peggioramento sia superiore nei soggetti diabetici rispetto agli individui sani.

3.1.2 Covariate relative al sito

In merito alle covariate disponibili per il sito di rilevazione, come si può osservare dalla (Figura 3.7), si è raggiunta la convergenza per ciascuna delle due componenti.

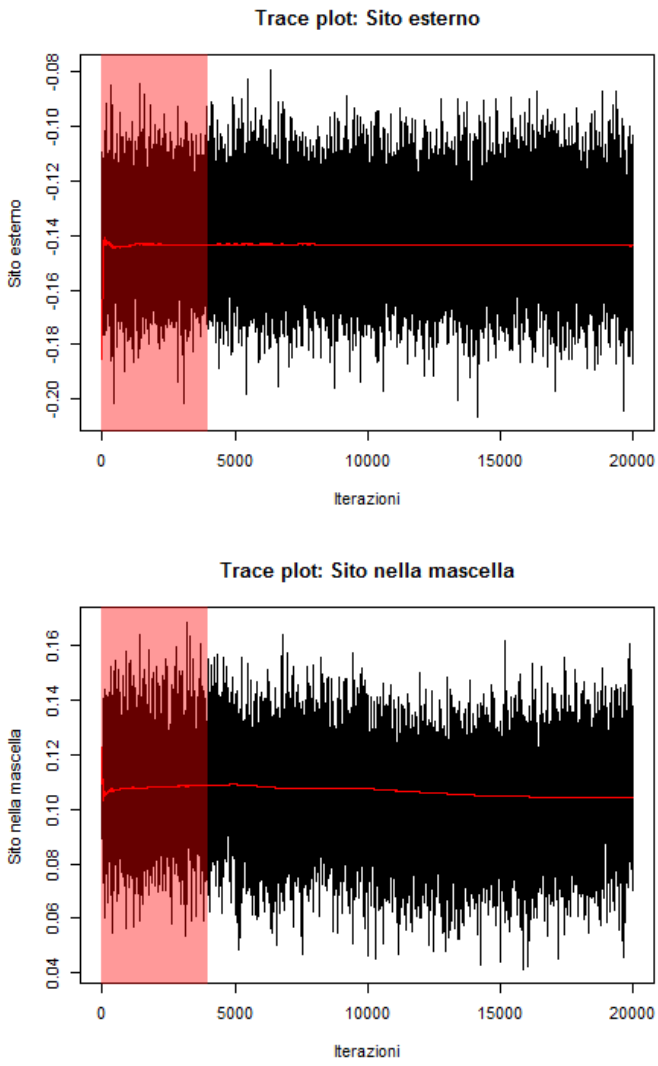

Figura 3.7: *Trace plots* per i coefficienti di regressione relativi alle covariate del sito: sito esterno e sito nella mascella

Le funzioni di autocorrelazione, riportate di seguito, evidenziano come per la variabile relativa al sito locato nella mascella, l'autocorrelazione oscilla

al di fuori in prossimità dell'estremo superiore dell'intervallo di confidenza. Tuttavia, in termini di convergenza, non si evidenziano particolari problemi.

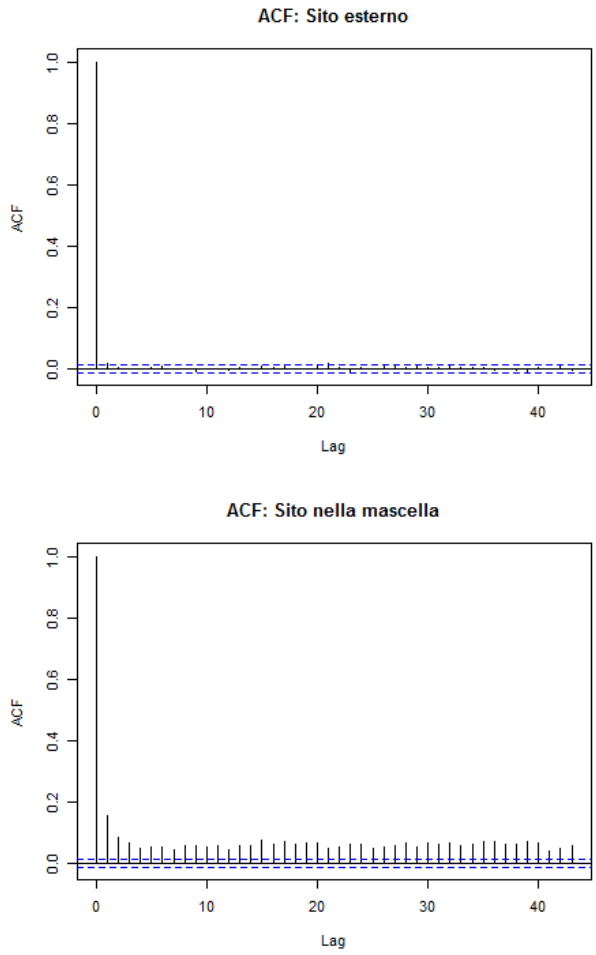

Figura 3.8: Funzioni di autocorrelazione per i coefficienti di regressione relativi alle covariate del sito: sito esterno e sito nella mascella

Le distribuzioni a posteriori, in Figura 3.9, portano alle seguenti considerazioni: i siti locati nella parte centrale del dente, dal lato della lingua e della bocca, sono negativamente associati con la progressione della malattia. Quelli situati nella parte superiore presentano uno stato di salute inferiore e sono più soggetti a deterioramento.

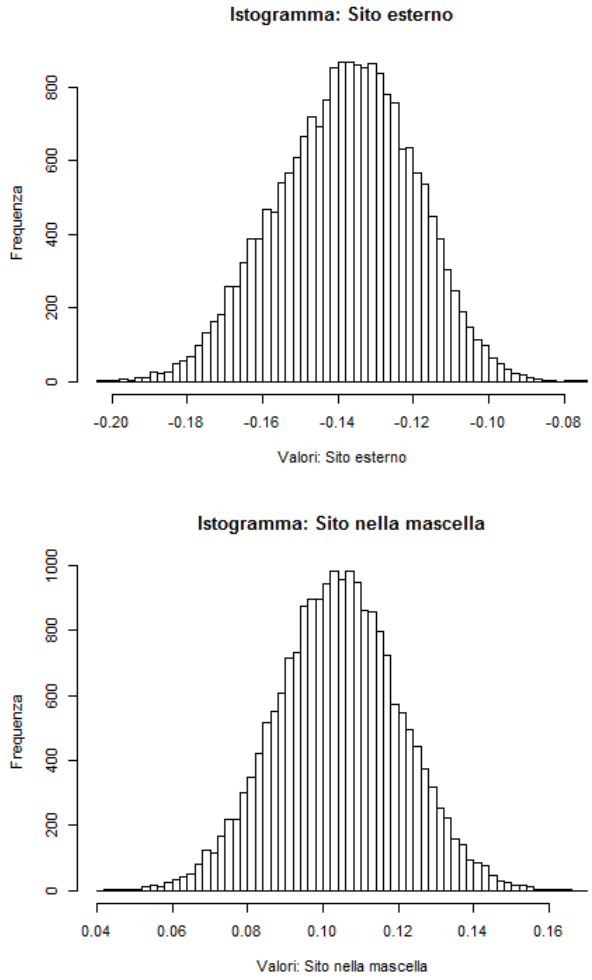

Figura 3.9: Distribuzioni a posteriori per i coefficienti di regressione relativi alle covariate del sito: sito esterno e sito nella mascella

3.1.3 Parametro di lisciamento

Il parametro di lisciamento ρ giunge a convergenza, seppure presenti un'autocorrelazione significativa anche oltre il quarantesimo ritardo. Tuttavia la decrescita è di carattere esponenziale e ci si aspetta rientri nei relativi intervalli di confidenza nei successivi ritardi. L'indicazione è comunque chiara: la stima di questo parametro è prossima al limite superiore del suo dominio, perciò la dipendenza tra i siti nello stesso dente è molto alta, attorno allo 0.99.

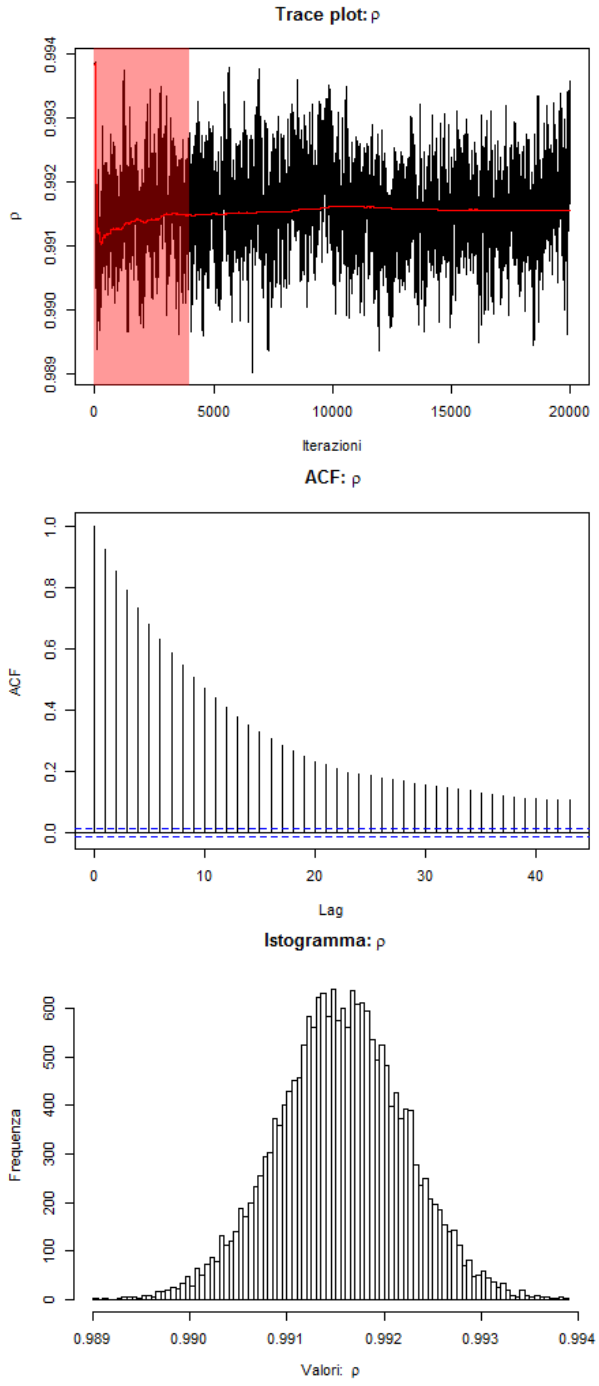

Figura 3.10: *Trace plot*, funzione di autocorrelazione e distribuzione a posteriori del parametro di lisciamento

3.2 Media a posteriori ed intervalli di credibilità

La Tabella 3.1 riporta le stime dei parametri e i corrispettivi intervalli di credibilità al 95%. Si può notare come il valore nullo non sia mai contenuto, per nessun parametro, nel relativo intervallo.

Livello	**Variabili**	**2.5 %**	**Media**	**97.5 %**
Soggetto	Età	0.01731	0.01860	0.01985
	Genere	-0.13718	-0.09583	-0.05259
	BMI	0.00288	0.00445	0.00597
	Fumatore	0.12343	0.16348	0.20359
	Livello di emoglobina glicata	0.03518	0.04209	0.04890
Sito	Sito esterno	-0.17323	-0.13807	-0.10418
	Sito nella mascella	0.07208	0.10417	0.13722
	Parametro di lisciamento	0.99030	0.99155	0.99287

Tabella 3.1: Media a posteriori e intervalli di credibilità HPD al 0.95

Conclusioni

Questo lavoro di tesi è incentrato sullo studio della parodontite in relazione ad alcune covariate disponibili per i soggetti e i siti di rilevazione della misura relativa alla perdita di attacco clinico. La modellazione proposta, attraverso un modello *probit* ordinale multivariato con approccio di inferenza bayesiano, considera la particolare struttura gerarchica di tipo soggetto/siti di rilevazione secondo la quale le osservazioni sono suddivise. Il processo latente sottostante potrebbe essere diverso al variare delle categorie di gravità e non ben descritto da una semplice discretizzazione della variabile risposta in classi equispaziate. Inoltre, ci si pone l'obiettivo di modellare la particolare struttura di dipendenza spaziale indotta dalla matrice di adiacenza definita per i siti, che si è dimostrata essere elevata poiché la stima del parametro di lisciamento è prossima all'estremo superiore del suo dominio.

Le indicazioni che derivano dalle stime dei parametri sono le seguenti: la malattia tende ad aggravarsi in soggetti fumatori, in particolare nei maschi, con un livello di emoglobina glicata fuori dai limiti normali e un elevato BMI. Essa si presenta in forme sempre più gravi con l'aumentare dell'età. Relativamente alle locazioni di rilevazione, i siti centrali tendono a manifestare un superiore deterioramento rispetto agli altri, lo stesso si verifica per i siti dell'arcata superiore della dentatura.

Gli sviluppi futuri di questa tesi sono molteplici: in questo lavoro si è supposto che i valori delle misurazioni mancanti, il cui significato è la perdita del dente e delle sue relative sei rilevazioni, appartengano alla fascia più grave della malattia, la quale ha provocato un grave deterioramento con conseguente distaccamento del dente. Tuttavia, in certe situazioni può risultare un'approssimazione un po' forzata e causare la perdita di un'importante for-

ma di eterogeneità tra i soggetti: sarebbe quindi opportuno considerare un ulteriore tipo di correlazione tra soggetti, a livello dell'intero dente, in quanto soggetti con molti denti mancanti potrebbero avere caratteristiche diverse dal comportamento medio degli altri soggetti. In secondo luogo, potrebbe essere interessante utilizzare differenti specificazioni della matrice di adiacenza: indicare come locazioni vicine i siti del dente, dello stesso tipo, nelle differenti arcate; oppure definire una struttura di vicinanza tra i quadranti dell'intera dentatura.

Un'importante lavoro che tratta questo tipo di dati è stato svolto, tramite un approccio bayesiano non parametrico, da Reich et al. (2013).

Appendice A

Codice R

Codice A.1: Librerie

```
library(mvtnorm)
library(tmvtnorm)
library(Matrix)
library(MASS)
library(TeachingDemos)
```

Codice A.2: Alcuni parametri utili

```
# D : Matrice diagonale, numero di siti vicini per ogni dente
# W : Matrice di adiacenza, tutti i siti appartenenti allo stesso dente
#     sono vicini

m <- 168
n <- 100
D <- SM
W <- ADJ2

# Creazione valori unici e dataset ridotto (n=125 soggetti)

sequenza <- seq(1,n*m, by=m)
nrighe <- rep(rep(1:n), each=m)
mydata_rid <- mydata2[1:length(nrighe),]
```

Codice A.3: Distribuzioni full conditionals

```
# 1) Full conditional per BETA (5-variata)

rf.beta.cond <- function(data, y_star){
  X <- as.matrix(data[sequenza, 4:8]) # covariate del sito
  Z <- as.matrix(data[1:m, 9:10])     # covariate del soggetto
  zgamma <- as.numeric(Z %*% gamma)
  yj <- sweep(y_star, 1, zgamma, "-")
  yj <- t(X) %*% apply(yj, 2, sum)
  varianza <- solve(m * t(X) %*% X + 1/10 * diag(5))
  media <- varianza %*% yj
  rmvnorm(1, media, varianza)
}

# 2) Full conditional per GAMMA (Bivariata)

rf.gamma.cond <- function(data, y_star){
  X <- as.matrix(data[sequenza, 4:8])
  Z <- as.matrix(data[1:m, 9:10])
  xbeta <- as.numeric(X %*% beta)
  yi <- sweep(y_star, 2, xbeta, "-")
  yi <- t(Z) %*% apply(yi, 1, sum)
  varianza <- solve(n * t(Z) %*% Z + 1/10 * diag(2))
  media <- varianza %*% yi
  rmvnorm(1, media, varianza)
}

# 3) Full conditionals per ALPHA

rf.alpha.cond <- function(data, y_star, alphaold){
  alphanew <- rep(NA,6)
  alphanew[1] <- -Inf
  alphanew[2] <- 0
  alphanew[6] <- Inf
```

```
  y <- as.vector(as.matrix(CALmatrix[,1:n]))
  y_star <- as.vector(y_star)

  alphanew[3] <- runif(1, max(max(y_star[y==2]), alphaold[2]),
                  min(min(y_star[y==3]), alphaold[4]))  # a2
  alphanew[4] <- runif(1, max(max(y_star[y==3]), alphaold[3]),
                  min(min(y_star[y==4]), alphaold[5]))  # a3
  alphanew[5] <- runif(1, max(max(y_star[y==4]), alphaold[4]),
                  min(min(y_star[y==5]), alphaold[6]))  # a4
  alphanew
}

# 4) Generazione delle variabili latenti dalla distribuzione normale
#     troncata multivariata

rmvnorm_trunc <- function(data, beta, gamma, rho, alpha){

  H <- solve(D - rho * W)     # Matrice di precisione

  y_star <- matrix(NA, nrow=m, ncol=n)
  alpha <- sort(alpha)
  for(i in unique(data$idsub)){
    X <- data[data$idsub==i, 4:8]
    Z <- data[data$idsub==i, 9:10]
    y <- CALmatrix[,i]
    mui <- as.numeric(as.matrix(X) %*% beta + as.matrix(Z) %*% gamma)
    a <- cbind(alpha[y], alpha[y+1])
    y_star[,i] <- rtmvnorm(5, mean=mui, lower=a[,1], upper=a[,2],
                  algorithm="gibbs", H=H, burn.in=4)[5,]
  }
  y_star
}
```

Codice A.4: Algoritmo Metropolis-within-Gibbs sampler

```
# Densita' della normale multivariata per il Metropolis step
```

```
ldmvnorm <- function(y_star, beta, gamma, sigma, data, alpha) {
  dens <- rep(NA, ncol(y_star))
  sigma <- solve(sigma) # matrice di varianze e covarianze
  for(i in 1:ncol(y_star)){
    X <- data[data$idsub==i, 4:8]
    Z <- data[data$idsub==i, 9:10]
    y <- CALmatrix[,i]
    mui <- as.matrix(X) %*% beta + as.matrix(Z) %*% gamma
    mui <- as.numeric(mui)
    dens[i] <- dmvnorm(x = y_star[,i], mean = mui, sigma = sigma, log=T)
    }
    sum(dens)
}

# 5) Algoritmo Metropolis-within-Gibbs

f.gibbs <- function(nsim, start, data, eps){
  out <- list(alpha = matrix(NA, nsim, 6), beta = matrix(NA, nsim, 5),
          gamma = matrix(NA, nsim, 2), rho = matrix(NA, nsim, 1))
  x <- start
  accepted <- numeric(1)
  for(i in 1:nsim){
        # Update Y
        y_star <- rmvnorm_trunc(data, as.numeric(x$beta), as.numeric(x
                  $gamma), as.numeric(x$rho), as.numeric(x$alpha))
        pre_alpha <- x$alpha
        # Update ALPHA
        x$alpha <- rf.alpha.cond(data, y_star, as.numeric(pre_alpha))
        # Update BETA
        x$beta <- rf.beta.cond(data, y_star)
        # Update GAMMA
        x$gamma <- rf.gamma.cond(data, y_star)
        xs <- x

        # Metropolis step per update rho con passeggiata casuale uniforme
```

```
        continue=T
        while(continue){                        #rho deve stare in [-1,1]
          rhonew <- x$rho + runif(1,-eps,eps)
          if(rhonew>-1 & rhonew<1) continue=F
        }
        rhonew <- x$rho + runif(1,-eps, eps)   #rho(c)
        xs$rho <- rhonew
        sigmanew <- D - rhonew * W
        rhoold <- x$rho                         #rho(k-1)
        sigmaold <- D - rhoold * W
        mui <- as.matrix(data[,4:8]) %*% as.numeric(x$beta) +
                as.matrix(data[,9:10]) %*% as.numeric(x$gamma)
        alpha_value <- min(1, exp(ldmvnorm_trunc(y_star, as.numeric(xs$
                       beta), as.numeric(xs$gamma), sigmanew, data, as.
                       numeric(xs$alpha)) - ldmvnorm_trunc(y_star, as.
                       numeric(xs$beta), as.numeric(xs$gamma), sigmaold,
                       data, as.numeric(xs$alpha))))
        if(runif(1) < alpha_value){
          accepted <- accepted + 1
          x$rho <- xs$rho
        }
        out$alpha[i,] <- x$alpha
        out$beta[i,] <- x$beta
        out$gamma[i,] <- x$gamma
        out$rho[i,] <- x$rho
        }
        list(values = out, accepted = accepted/nsim)
}
```

Codice A.5: Inizializzazione dei parametri

```
# Modello logit cumulato per inizializzazione dei parametri di regressione

yclass <- factor(mydata_rid$CALclass)
```

```
fit.polr <- polr(yclass ~ age + female + BMI + smoker + HbA1c_lev + gap +
             maxilla, data=mydata_rid, Hess = T)
summary(fit.polr)

nsim <- 20000
burn.in <- 4000
beta <- fit.polr$coefficients[1:5]
gamma <- fit.polr$coefficients[6:7]
rho <- .991
alpha <- c(-Inf, 0, 0.7111, 1.1555, 1.355, Inf)
start <- list(alpha = alpha, beta = beta, gamma = gamma, rho = rho)

ptm <- proc.time()
output <- f.gibbs(nsim, start, mydata_rid, eps=.00052)
proc.time()-ptm
```

Codice A.6: Media a posteriori e intervalli di credibilita' al 95%

```
emp.hpd <- function (x, conf = 0.95){
  conf <- min(conf, 1 - conf)
  n <- length(x)
  nn <- round(n * conf)
  x <- sort(x)
  xx <- x[(n - nn + 1):n] - x[1:nn]
  m <- min(xx)
  nnn <- which(xx == m)[1]
  return(c(x[nnn], x[n - nn + nnn]))
}

# Medie a posteriori e intervalli di credibilita' HPD al 95%

# Eta', Genere, BMI, Fumatore, Livello di emoglobina glicata

mean(output$values$beta[-burn.in])
emp.hpd(output$values$beta[-burn.in])
```

```
# Sito esterno, Sito nella mascella

mean(output$values$gamma[-burn.in])
emp.hpd(output$values$gamma[-burn.in])

# Parametro di lisciamento

mean(output$values$rho[-burn.in])
emp.hpd(output$values$rho[-burn.in])
```

Ringraziamenti

Nella conclusione del percorso di studi universitari si viene a creare una situazione molto particolare: da un lato l'emozione e la gioia, nonché la consapevolezza, di esserci riusciti e potersi affacciare al complesso mondo lavorativo; dall'altro la mente non può che ritornare alle innumerevoli esperienze di vita vissute ed amicizie instaurate in un ambiente ideale e completo come lo è sempre stata questa facoltà, che resteranno impresse a lungo termine e difficilmente verranno dimenticate o perse.

Un sentito ringraziamento va al Prof. Canale, per aver sempre dimostrato massima disponibilità e dedizione nei miei confronti in tutto il periodo di tesi, risultando in ogni occasione attento alle mie richieste e mancanze.

Non basterebbero assolutamente queste poche righe per nominare tutte le persone che hanno reso questi ultimi due anni speciali, sia all'interno del mondo universitario che all'esterno.

Ai miei genitori va riconosciuto l'importante merito di avermi sostenuto in qualsiasi momento, impegnandosi affinché si venissero a creare le condizioni e la tranquillità nel potermi concentrare a tempo pieno negli studi, non facendomi mai mancare nulla ed assecondando sempre le mie esigenze.

Alla mia splendida ragazza, nonché collega, amica e compagna di studio Ilaria, con la quale ho condiviso tutti questi cinque intensi e avventurosi anni, perché abbiamo sempre saputo superare insieme qualsiasi ostacolo, creando un legame difficile da immaginare: grazie, di cuore, ora siamo grandi.

A tutti gli amici e ragazzi di Padova, con i quali c'è sempre stato il modo di vivere ogni singolo giorno, un giorno particolare; tra risate e momenti produttivi, un gruppo coeso che terrò sempre a me stretto. Un grazie sincero ad Alessandro, Giulia e Ciro; con i quali mi sono confrontato e interfacciato

in questo ultimo intenso periodo di tesi.

Si ringrazia vivamente, infine, il Center for Oral Health Research (COHR) presso la Medical University of South Carolina (MUSC) per aver fornito i dati ed il contesto di questa tesi, in particolare il Prof. Dipankar Bandyopadhyay del Dipartimento di Biostatistica della Virginia Commonwealth University (VCU).

Bibliografia

Albert, J.H. e S. Chib (1993). «Bayesian Analysis of Binary and Polychotomous Response Data». In: *Journal of the American Statistical Association* 88.422, pp. 669–679.

Arora, M. et al. (2009). «Association of Environmental Cadmium Exposure with Periodontal Disease in U.S. Adults». In: *Environmental Health Perspectives* 117 (5).

Bandyopadhyay, D. e A. Canale (2016). «Non-parametric spatial models for clustered ordered periodontal data». In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 65.4, pp. 619–640.

Besag, J. (1974). «Spatial Interaction and the Statistical Analysis of Lattice Systems». In: *Journal of the Royal Statistical Society Series B (Methodological)* 36 (2).

Cowles, M.K. (1996). «Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models». In: *Statistics and Computing* 6 (2).

«Quali strategie per arginare l'«epidemia» di parodontite?» In: *Italian Dental Journal* Anno X Numero 2. A cura di Griffin Editore, p. 3. ISSN: 1970-7428.

Engebretson, S.P. et al. (2013). «The effect of nonsurgical periodontal therapy on hemoglobin a1c levels in persons with type 2 diabetes and chronic periodontitis: A randomized clinical trial». In: *JAMA* 310.23, pp. 2523–2532.

Fernandes, J.K. et al. (2009). «Periodontal Disease Status in Gullah African Americans With Type 2 Diabetes Living in South Carolina». In: *Journal of Periodontology* 80.7, pp. 1062–1068.

Johnson, V.E. e J.H. Albert (1999). *Ordinal Data Modeling*. Statistics for Social and Behavioral Sciences. New York: Springer Verlag.

Kotecha, J.H. e P.M. Djuric (1999). «Gibbs Sampling Approach for Generation of Truncated Multivariate Gaussian Random Variables». In: *Proceedings of the Acoustics, Speech, and Signal Processing, 1999. On 1999 IEEE International Conference - Volume 03*. ICASSP '99. Washington, DC, USA: IEEE Computer Society, pp. 1757–1760. ISBN: 0-7803-5041-3.

Lynch, S.M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. 1ª ed. Statistics for social and behavioral sciences. Springer, pp. 291–294.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS / Ioannis Ntzoufras*. Wiley Hoboken, N.J.

Prodentis. *Graduated Periodontal Explorer - Williams Classic*. URL: http://www.prodentis.com/p-2122443963418804437-Sonde-Williams-parodontale-graduee-Classic.html.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Reich, J.B. et al. (2013). «A Nonparametric Spatial Model for Periodontal Data With Nonrandom Missingness». In: *Journal of the American Statistical Association*.

Tanner, M.A. e W.H. Wong (1987). «The Calculation of Posterior Distributions by Data Augmentation». In: *Journal of the American Statistical Association* 82.398, pp. 528–540.