Università degli Studi di Firenze

Dipartimento di Statistica G. Parenti

Dottorato di Ricerca in Statistica Applicata

XXIII ciclo - SECS-S/01

Modelli per dati longitudinali Gaussiani con osservazioni censurate

Antonietta Chianca

Tutor: **Annibale Biggeri**

Co-Tutor: **Bianca De Stavola**

Co-Tutor: **Dorothea Nitsch**

Coordinatore: **Fabio Corradi**

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

Molti studi epidemiologici longitudinali consistono nel seguire nel tempo una coorte di soggetti per rilevarne periodicamente informazioni sullo stato di salute e anche dati di sopravvivenza. Tali studi possono presentare problemi di dati mancanti o osservazioni censurate, ad esempio in seguito al decesso. In questa tesi vengono trattati modelli statistici applicabili a dati longitudinali con osservazioni censurate, nel caso in cui l'uscita dallo studio può essere ipotizzata non casuale, parlando di dropout o censoring informativo. Tali modelli sono detti joint models per dati longitudinali e di sopravvivenza. Nel caso di missing non casuale c'è dipendenza tra il processo di risposta e quello dei dati mancanti. Non tener conto di questa dipendenza può portare a stime distorte. I joint models studiati e applicati in questa tesi possono essere fatti rientrare in una classe ampia di modelli detti shared parameter models, in cui un modello per misure di risposta longitudinale è linkato con un modello che descrive il meccanismo dei dati mancanti, attraverso un set di effetti casuali che sono condivisi tra i due processi. In tal caso si produce un'indipendenza condizionata tra i due processi, date le covariate, portando a semplificazioni nelle stime. I joint models sono stati applicati ai dati derivanti dallo Studio Osservazionale Longitudinale ULSAM, condotto in Svezia, e per il quale sono stati arruolati tutti gli uomini viventi nell'Uppsala Country nati tra il 1920 e il 1924. Lo scopo dello studio è di analizzare l'effetto nel tempo dello

stile di vita (attività fisica, fumo, obesità,...) sul biomarker GFR (stimato a partire dai valori della cistatina), che è un indicatore della funzionalità renale.

I dati del biomarker sono stati misurati in tre occasioni sulla coorte di uomini osservati. La prima visita è stata effettuata all'età di circa 70 anni. Alla terza visita, avvenuta quando gli uomini avevano circa 82 anni, la mortalità rispetto alla coorte iniziale era abbastanza rilevante, arrivando al 41%. Si è ipotizzato che i soggetti deceduti producano un dropout informativo o missing non a caso (MNAR). Secondo tale definizione il processo dei dati mancanti dipende dai valori non osservati della variabile risposta. Nel nostro caso può essere considerato ragionevole che la sopravvivenza degli uomini usciti dallo studio in seguito al decesso dipenda dal biomarker GFR, ovvero dalla stato della propria funzionalità renale. Per tener conto del supposto missing data informativo dovuto al decesso, sono stati applicati joint models dei dati longitudinali e di sopravvivenza teorizzati da Henderson, Diggle e Dobson (2000), secondo un'impostazione Bayesiana proposta da Guo e Carlin (2004). Per l'applicazione ai dati ULSAM, sono state effettuate elaborazioni adattando al caso specifico programmi scritti in Winbugs. I risultati del joint model sono stati confrontati con quelli derivanti dalle analisi separate, sia secondo l'impostazione classica frequentista che bayesiana, ovvero un modello lineare ad effetti casuali per i dati longitudinali e un modello di Weibull per i dati di sopravvivenza.

Capitolo 1

Metodologia classica

1.1 Introduzione

In questo capitolo riportiamo le modellizzazioni classiche che sono utilizzate generalmente per le analisi di dati longitudinali con outcome continuo e per i dati di sopravvivenza. Per l'analisi dei dati longitudinali sono presentati i modelli ad effetti casuali a due stadi cosi come proposti da Laird e Ware (1982). Tali modelli sono stati diffusamente considerati nella letteratura successiva per la definizione di modelli congiunti o *joint models* per dati longitudinali e di sopravvivenza, in particolare per situazioni di drop-out informativo, come si vedrà di seguito. Per quanto riguarda l'analisi di sopravvivenza sono brevemente riportati i modelli di hazard proporzionale di Cox e Weibull. Sarà inoltre riportata l'impostazione bayesiana dei modelli considerati.

1.2 Modelli longitudinali ad effetti casuali

Molti studi longitudinali sono realizzati per investigare il cambiamento nel tempo di una caratteristica che è misurata ripetutamente per ogni partecipante allo studio (Laird and Ware (1982)). Nel nostro caso, a partire dallo studio longitudinale ULSAM, la caratteristica d'interesse misurata ripetutamente è una stima del Glomerular filtration rate (GFR), indicatore della funzionalità renale. Le misure multiple sono prese per ogni individuo; tali misure possono variare tra gli individui per il numero e il tempo di osservazione, producendo un dataset detto non bilanciato. In un modello a due stadi la distribuzione di probabilità per le misure multiple ha la stessa forma per ogni individuo, mentre i parametri di queste distribuzioni variano tra gli individui. Le distribuzioni di questi parametri nella popolazione, detti effetti casuali o 'random effects', costituiscono il secondo stadio del modello. I modelli a due stadi non richiedono che i dati longitudinali siano bilanciati. Laird e Ware definiscono una famiglia di modelli per misure seriali che includono modelli di crescita, detti *growth models*, e modelli per misure ripetute come casi particolari.

Sia y_{ij} la variabile risposta e $\boldsymbol{x}_{ij}$ il vettore di variabili esplicative osservate al tempo t_{ij} per il soggetto i, con $i = 1, ..., m$ e $j = 1, ..., n_i$. Sia $\boldsymbol{y}_i$ il vettore normale multivariato delle risposte per l'i-simo individuo, con $E(\boldsymbol{y}_i) = \boldsymbol{\mu}_i$ vettore delle medie di dimensione $(n_i \times 1)$, e $\boldsymbol{\Sigma}$ matrice di varianza e covarianza di dimensione $(n_i \times n_i)$. Il modello proposto da Laird e Ware è il seguente:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i \tag{1.1}$$

dove $\boldsymbol{\alpha}$ è il vettore $(p \times 1)$ dei parametri non noti della popolazione (effetti fissi), $\boldsymbol{X}_i$ è la matrice $(n_i \times p)$ del disegno nota che collega $\boldsymbol{\alpha}$ alle misure longitudinali $\boldsymbol{y}_i$, $\boldsymbol{b}_i$ è il vettore $(k \times 1)$ di effetti casuali individuali non osservabili, $\boldsymbol{Z}_i$ è una matrice del disegno $(n_i \times k)$ nota che collega $\boldsymbol{b}_i$ a $\boldsymbol{y}_i$, e $\boldsymbol{e}_i \sim N(\mathbf{0}, \boldsymbol{R}_i)$ è il vettore dei residui indipendenti e normalmente distribuiti, dove $\boldsymbol{R}_i$ è una matrice di varianza e covarianza $(n_i \times n_i)$ definita positiva. Nel primo stadio $\boldsymbol{\alpha}$ e $\boldsymbol{b}_i$ sono considerati effetti fissi. Nel secondo stadio i parametri $\boldsymbol{\alpha}$ sono assunti come effetti fissi, mentre i parametri $\boldsymbol{b}_i \sim N(\mathbf{0}, \boldsymbol{D})$ sono assunti normalmente distribuiti con media nulla e matrice di varianza e covarianza $\mathbf{D}$ di dimensione $(k \times k)$.

Marginalmente il vettore $\boldsymbol{y}_i$ si distribuisce normalmente con media $\boldsymbol{X}_i\boldsymbol{\alpha}$ e matrice di varianza e covarianza $\boldsymbol{R}_i + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'$. Per ottenere le stime di massima verosimiglianza o di massima verosimiglianza ristretta dei parametri non noti possono essere utilizzati metodi iterativi come il Newthon Raphson e l'algoritmo EM (Dempster et al., 1977; Dempster et al., 1981; Harville, 1977; Laird, 1982).

1.3 Modelli di sopravvivenza

L'analisi di sopravvivenza, o analisi dei dati del tempo all'evento, si riferisce a metodi statistici per i quali la variabile risposta di interesse è, per il singolo individuo, il tempo intercorso dall'inizio del follow-up fino al verificarsi dell'evento, detto anche fallimento. In studi epidemiologici l'inizio dell'osservazione può essere la nascita o l'inizio dell'esposizione ad alcuni fattori di rischio, mentre l'evento d'interesse può essere il decesso, l'incidenza di una malattia, il ricovero.

L'evento di interesse può non essere osservato per ogni individuo, ad esempio perchè lo studio è chiuso prima che l'evento sia sperimentato (in tal caso si conosce solo che il tempo all'evento è più grande di quello osservato), o perchè alcuni soggetti possono non volere, o non potere per varie ragioni, continuare a partecipare allo studio e fornire le informazioni (soggetti detti 'dropouts'). In questi casi si hanno informazioni mancanti considerate *right censored*. In genere si assume che il censoring sia non informativo.

1.3.1 Concetti matematici di base

Si indichi con T una variabile casuale continua non-negativa che denota il tempo di sopravvivenza, ovvero il tempo trascorso fino all'occorrenza dell'evento di interesse. Si indichi con $f(t)$ la funzione di densità di probabilità di T e con $F(t)$ la corrispondente funzione di distribuzione cumulata. La funzione di sopravvivenza, che fornisce la probabilità che un individuo sopravviva dopo il tempo t, è definita come

$$S(t) = 1 - F(t) = P(T > t) = \int_t^\infty f(u)du \tag{1.2}$$

La funzione $S(t)$ ha una serie di proprietà: (a) è monotona non-crescente; (b) $S(0) = 1$, ovvero all'inizio dello studio si assume che tutti gli individui siano vivi; (c) $S(\infty) = \lim_{n\to\infty} S(t) = 0$, ovvero se lo studio procedesse all'infinito non ci sarebbe nessun sopravvissuto. La funzione di densità è data da

$$f(t) = \lim_{\Delta t\to 0} P(t \leq T < t + \Delta t) \tag{1.3}$$

Un'altra importante funzione usata nella descrizione dei tempi di sopravvivenza è l'hazard function, o tasso istantaneo di fallimento al tempo t, che

è definito come

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{1.4}$$

Si noti che $\lambda(t)$ non è una probabilità condizionata di fallimento, ma un tasso condizionato di fallimento, e che quindi il limite superiore della funzione non è 1. Comunque $\lambda(t)\Delta t$ approssima una probabilità di fallimento nell'intervallo $(t, t + \Delta t]$, dato che si è sopravvissuti al tempo t. Le funzioni $S(t)$ e $f(t)$ possono essere espresse in funzione di $\lambda(t)$. Dall'equazione (1.4) abbiamo che

$$\lambda(t) = \frac{S'(t)}{S(t)} = -\frac{d}{dt} log\left\{S(t)\right\} \tag{1.5}$$

per cui segue che

$$S(t) = \exp\left(-\int_0^t \lambda(u)du\right) = \exp\left(-\Lambda(t)\right) \tag{1.6}$$

dove $\Lambda(t) = \int_0^t \lambda(u)du$ è detta hazard function cumulativa. Combinando la (1.4) e la (1.6) si ha

$$f(t) = \lambda(t)S(t) = \lambda(t) \exp\left(-\int_0^t \lambda(u)du\right) \tag{1.7}$$

1.3.2 Modello di Cox

Negli studi epidemiologici può essere di particolare interesse la relazione tra i tempi di sopravvivenza dei soggetti e una serie di covariate osservate durante lo studio. Per questo esistono modelli che consentono di incorporare le informazioni su tali covariate. Il modello di sopravvivenza più diffuso fu proposto da Cox (1972).

Dato un vettore delle covariate $\boldsymbol{x}$ e un vettore dei coefficienti di regressione $\boldsymbol{\beta}$, sia $\lambda_0(t)$ la *baseline hazard function* al tempo t. L'hazard al tempo t per un individuo i è dato da

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t) \exp\left(\boldsymbol{x}'\boldsymbol{\beta}\right)$$

La principale assunzione alla base del modello di Cox è che l'*hazard ratio* per due individui sia costante nel tempo, per cui si parla di proportional hazard model. Se il numero di individui è n, i dati di sopravvivenza osservati sono dati da n coppie (T_i^*, δ_i), dove $T_i^* = min(T_i, C_i)$, essendo C_i il tempo di right censoring osservato per l'i-simo soggetto, mentre δ_i è una variabile indicatrice che assume valore 1 nel caso in cui si è verificato l'evento, ovvero $T_i \leq C_i$, e 0 altrimenti.

Assumendo che i tempi T_i siano *iid*, la funzione di verosimiglianza può essere scritta come

$$L(\beta, \lambda_0(t)|D) \propto \prod_{i=1}^{n} [\lambda_0(T_i^*) \exp(\boldsymbol{x}_i' \boldsymbol{\beta})]^{\delta_i} \exp\left(- \int_0^{T_i^*} \lambda_0(u) \exp(\boldsymbol{x}_i' \boldsymbol{\beta}) du \right) \tag{1.8}$$

dove $D = (n, \boldsymbol{T}^*, \boldsymbol{X}, \boldsymbol{\delta})$, con $\boldsymbol{T}^*$, $\boldsymbol{X}$, $\boldsymbol{\delta}$ vettori di n elementi. Per la stima dei parametri β si può risolvere rispetto a β la verosimiglianza parziale, che non necessita della stima del baseline hazard function $\lambda_0(t)$. La verosimiglianza parziale è data da

$$PL(\beta|D) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{x}_{(i)}' \boldsymbol{\beta})}{\sum_{k \in R_i} \exp(\boldsymbol{x}_k' \boldsymbol{\beta})} \right)^{\delta_i} \tag{1.9}$$

dove R_i rappresentano gli esposti al rischio. Le osservazioni censurate contribuiscono solo al denominatore.

1.3.3 Modello di Weibull

Il modello di Cox è un modello semiparametrico in quanto la forma di $\lambda_0(t)$ non è specificata. Possono essere considerati anche modelli parametrici come il modello di Weibull, per il quale i tempi di sopravvivenza sono assunti seguire la distribuzione Weibull, con parametro di scala λ e parametro di pendenza γ. Il baseline hazard function si può scrivere come

$$\lambda_0(t) = \lambda \gamma t^{\gamma - 1} \tag{1.10}$$

e l'hazard function è data da

$$\lambda(t|\boldsymbol{x}) = \lambda\gamma t^{\gamma-1}\exp\left(\boldsymbol{x}'\boldsymbol{\beta}\right)$$

1.4 Approccio bayesiano

Sia $\boldsymbol{\theta}$ il vettore di parametri non noti, e $\mathbf{y} = (y_1, ..., y_n)$ i dati osservati, la cui distribuzione di probabilità è $f\left(\mathbf{y}\,|\theta\right)$. L'approccio bayesiano assume che $\boldsymbol{\theta}$ sia a sua volta una variabile casuale con una certa distribuzione a priori $\pi(\boldsymbol{\theta})$ (basata sulle nostre conoscenze su $\boldsymbol{\theta}$), che, combinata con la verosimiglianza attraverso il teorema di Bayes, porta alla distribuzione a posteriori di $\boldsymbol{\theta}$ data da

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y},\theta)}{p(\mathbf{y})} = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{\int f(\mathbf{y}|\theta)\pi(\theta)d\theta} \tag{1.11}$$

Il denominatore della (1.11), detta anche densità marginale dei dati $\mathbf{y}$, non dipende dai parametri $\boldsymbol{\theta}$, rappresentando solo una costante di normalizzazione, così possiamo scrivere

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\theta)\pi(\theta) \tag{1.12}$$

Se non abbiamo informazioni a priori sui parametri $\boldsymbol{\theta}$ o si vuole fare inferenza esclusivamente a partire dai dati, in genere viene specificata una a priori non informativa (chiamata anche *flat* o *vague*). Esempi di tali distribuzioni a priori sono la uniforme o la normale con una grande varianza. A volte possono essere specificate delle a priori dette improprie, tali che non richiedono che l'integrale della distribuzione sia pari ad 1. Non è però possibile ottenere sempre a posteriori proprie. Un esempio è il caso di random effects models per dati longitudinali, nel quale la dimensione dello spazio parametrico aumenta con l'ampiezza campionaria. In tali modelli, le informazioni sui dati possono essere insufficienti per identificare tutti i parametri, per cui alcune

delle distribuzioni a priori sui parametri individuali devono essere informative.

Il modello bayesiano considerato nell'equazione 1.11 è anche detto gerarchico; esso può essere visto come costituito da due stadi, uno per $f(\mathbf{y}|\boldsymbol{\theta})$, la verosimiglianza delle osservazioni $\mathbf{y}$ dati i parametri $\boldsymbol{\theta}$, e uno per $\pi(\boldsymbol{\theta}|\boldsymbol{\eta})$, la distribuzione a priori dei parametri $\boldsymbol{\theta}$ del modello dato il vettore degli iperparametri $\boldsymbol{\eta}$. Si può quantificare l'incertezza riguardo gli iperparametri $\boldsymbol{\eta}$ ponendo un secondo-stadio di distribuzioni a priori, cosiddette *hyperprior*, $h(\boldsymbol{\eta})$. In tal caso la distribuzione a posteriori per $\boldsymbol{\theta}$, ottenuta marginalizzando anche su $\boldsymbol{\eta}$, ed è data da:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{\int p(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\eta}) d\eta}{\int\int p(\mathbf{y}, \boldsymbol{u}, \boldsymbol{\eta}) d\boldsymbol{\eta} d\boldsymbol{u}} = \frac{\int f(\mathbf{y}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}|\boldsymbol{\eta}) h(\boldsymbol{\eta}) d\boldsymbol{\eta}}{\int\int f(\mathbf{y}|\boldsymbol{u}) \pi(\boldsymbol{u}|\boldsymbol{\eta}) h(\boldsymbol{\eta}) d\boldsymbol{\eta} d\boldsymbol{u}} \tag{1.13}$$

Per vantaggi computazionali si può scegliere la a priori per $\pi(\boldsymbol{\theta}|\boldsymbol{\eta})$ appartenente alla famiglia di distribuzioni che è coniugata con la verosimiglianza $f(\mathbf{y}|\theta)$. In tal caso si ottiene una distribuzione a posteriori $p(\boldsymbol{\theta}|\mathbf{y})$ appartenente alla stessa famiglia della a priori.

Considerando, come esempio, la distribuzione Gaussiana $f(\mathbf{y}|\boldsymbol{\theta}) \sim N(\mathbf{y}|\mu, \sigma^2)$, in cui entrambi i parametri $\boldsymbol{\theta} = (\mu, \sigma^2)$ sono assunti non noti. Se guardiamo alla verosimiglianza f come funzione solo di μ, la a priori coniugata è ancora data da una distribuzione normale, per cui possiamo assumere che $\pi_1(\mu) \sim N(\theta|\mu_1, \tau^2)$. Invece, guardando ad f come ad una funzione solo di σ, si ottiene una espressione proporzionale al reciproco di una Gamma, per cui $\pi_2(\sigma^2) \sim IG(\sigma^2|a, b)$. Assumendo l'indipendenza di μ e σ^2 si ottiene $\pi(\boldsymbol{\theta}) = \pi_1(\mu)\pi_2(\sigma^2)$. La possibilità di usare a priori coniugate che producono a posteriori in forma chiusa le rende particolarmente importanti nell'analisi bayesiana. Anche se una singola a priori coniugata può risultare inadeguata nel riflettere le conoscenze a priori, in taluni casi (ad esempio multimodalità,

code pesanti, ecc.), può essere più flessibile utilizzare una mistura finita di a priori coniugate, che consente comunque dei calcoli della a posteriori semplificati. Una mistura di a priori coniugate porta ad una mistura di a posteriori coniugate.

1.4.1 Modelli longitudinali ad effetti casuali

In questa sezione riportiamo i modelli longitudinali lineari ad effetti casuali secondo l'impostazione bayesiana, utilizzando la simbologia proposta da Carlin (Carlin B. P., Louis T. A. (2008)).
Nel caso di dati longitudinali con misure ripetute, si indichi con Y_{ij} la j-sima misura ripetuta dell'i-simo individuo nello studio, $j = 1, ..., s_i$ e $i = 1, ..., n$. Un modello ad effetti casuali misto è indicato con:

$$Y_i = \mathbf{X}_i\alpha + \mathbf{W}_i\beta_i + \epsilon_i \tag{1.14}$$

dove $\mathbf{X}_i$ è una matrice delle covariate $s_i \times p$, $\mathbf{W}_i$ è una matrice del disegno $s_i \times q$, α è un vettore dei parametri $p \times 1$ e β_i è un vettore degli effetti casuali soggetto-specifici $q \times 1$, che si assume normalmente distribuito con vettore delle medie $\mathbf{0}$ e matrice di varianza e covarianza $\mathbf{V}$. I coefficienti β_i dovrebbero catturare ogni effetto medio soggetto-specifico, rendendo le componenti ϵ_i non più correlate. Si assume dunque che, dato β_i, le componenti di ϵ_i siano indipendenti, con $\sum = \sigma^2 \cdot \mathbf{I}_{s_i}$
Il modello (1.14) è applicabile anche nel caso in cui non tutti gli individui hanno lo stesso numero di osservazioni, nel caso di dati mancanti ad esempio a causa della mancata partecipazione alle visite previste dallo studio. Inoltre può essere rimossa l'assunzione di una comune varianza dei soggetti σ^2. Anche l'assunzione di un errore Gaussiano può essere sostituita con un'altra densità simmetrica ma a code più pesanti, come una t di Student. E' anche

possibile considerare covariate i cui valori cambiano nel tempo. In una impostazione bayesiana Carlin (Carlin B. P., Louis T. A. (2008)) propone una specificazione del modello individuando distribuzioni a priori per α, σ^2 e $\mathbf{V}$. Alcuni autori bayesiani considerano la distribuzione degli effetti casuali β_i come parte della a priori ("primo stadio" o parte strutturale), mentre la distribuzione di $\mathbf{V}$ forma un'altra parte ("secondo stadio" o parte soggettiva). Queste due parti insieme sono dette *hierarchical prior*.

Consideriamo come esempio un modello molto semplice, del tipo

$$Y_{ij} \sim N(\alpha_i + \beta_i x_{ij}, \sigma^2), j = 1, ..., s_i, i = 1, ..., n \tag{1.15}$$

Tale modello è detto ad effetti casuali e considera la stima di differenti intercette e pendenze per ogni individuo i.
Si assume inoltre che il vettore dei parametri delle intercette e pendenze derivino da una stessa popolazione con distribuzione normale,

$$\boldsymbol{\theta}_i \equiv \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim N\left(\boldsymbol{\theta}_0 \equiv \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}, \sum\right), i = 1, ..., n \tag{1.16}$$

Nell'analisi bayesiana abbiamo una struttura gerarchica, in cui dobbiamo specificare una distribuzione a priori per σ^2 al secondo stadio e una distribuzione a priori al terzo stadio per $\boldsymbol{\theta}_0$ e $\sum$.
Per semplicità si possono scegliere forme coniugate per ognuna di queste distribuzioni, per cui

$$\sigma^2 \sim IG(a, b), \tag{1.17}$$

$$\boldsymbol{\theta}_0 \sim N(\boldsymbol{\eta}, C), \tag{1.18}$$

$$\Sigma^{-1} \sim W((\rho R)^{-1}, \rho) \tag{1.19}$$

dove W indica una distribuzione *Wishart*, una generalizzazione multivariata della distribuzione Gamma. In questo esempio R è una matrice 2x2 e $\rho \geq 2$ è un parametro che rappresenta i gradi di libertà.

In generale $V \sim W(D, n)$ è una matrice $p \times p$ simmetrica e definita positiva, con D una matrice simmetrica e definita positiva e $n > 0$, parametro di pendenza o gradi di libertà, posto che $n \geq p$.
La funzione di densità di V è proporzionale a

$$p(V|n, D) \propto \frac{|V|^{(n-p-1)/2}}{|D|^{n/2}} exp\left[-\frac{1}{2}tr(D^{-1}V)\right] \tag{1.20}$$

Per questa distribuzione si ha che $E(V_{ij}) = nD_{ij}$, $Var(V_{ij}) = n(D_{ij}^2 + D_{ii}D_{jj})$, e $Cov(V_{ij}, V_{kl}) = n(D_{ik}D_{jl} + D_{il}D_{jk})$.
Secondo la parametrizzazione scelta nella (1.19) si ha $E(\Sigma^{-1}) = R^{-1}$, per cui R^{-1} è la a priori della precisione attesa dei parametri θ_i, mentre R è la a priori della varianza attesa. Inoltre, $Var(\Sigma_{ij})$ è decrescente con ρ, per cui piccoli valori di ρ corrispondono a distribuzioni delle a priori vaghe.

Gli iperparametri del modello sono a, b, $\boldsymbol{\eta}$, C, ρ e R, tutti assunti noti, mentre il numero totale di parametri non noto nel modello è α_i per il numero di individui n, β_i per il numero di individui n, α_0, β_0, σ^2 e le tre componenti della matrice di varianza e covarianza Σ. La scelta di a priori coniugate rende possibile ottenere facilmente stime dei parametri tramite un Gibbs sampler. La distribuzione a posteriori di $\boldsymbol{\theta}_i$ è data da

$$\boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\theta}_0, \Sigma^{-1}, \sigma^2 \sim N(D_i(\sigma^{-2}X_i^T\mathbf{y}_i + \Sigma^{-1}\boldsymbol{\theta}_0), D_i) \tag{1.21}$$

dove $D_i^{-1} = \sigma^{-2}X_i^T X_i + \Sigma^{-1}$.
Analogamente, la *full conditional* per $\boldsymbol{\theta}_0$ è

$$\boldsymbol{\theta}_0|\mathbf{y}, \{\boldsymbol{\theta}_i\}, \Sigma^{-1}, \sigma^2 \sim N(V(n\Sigma^{-1}\bar{\boldsymbol{\theta}} + C^{-1}\boldsymbol{\eta}), V), \tag{1.22}$$

dove $V = (n\Sigma^{-1} + C^{-1})^{-1}$ e $\bar{\boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\theta}_i$

Per il nostro esempio la distribuzione a priori Wishart (1.20) è proporzionale a

$$\left|\Sigma^{-1}\right|^{(\rho-3)/2} exp\left[-\frac{1}{2}tr(\rho R\Sigma^{-1})\right] \tag{1.23}$$

Combinando la verosimiglianza normale per gli effetti casuali (1.16) con questa distribuzione a priori si ottiene le seguente distribuzione Wishart aggiornata

$$\Sigma^{-1}|\mathbf{y}, \{\boldsymbol{\theta}_i\}, \boldsymbol{\theta}_0, \sigma^2 \sim W\left(\left[\sum_{i=1}^{n}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_0)(\boldsymbol{\theta}_i - \boldsymbol{\theta}_0)^T + \rho R\right]^{-1}, n + \rho\right)$$

Infine, la *full conditional* per σ^2 è la distribuzione Inverse Gamma aggiornata

$$\sigma^2|\mathbf{y}, \{\boldsymbol{\theta}_i\}, \boldsymbol{\theta}_0, \Sigma \sim IG\left(\frac{s}{2} + a, \left[\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - X_i\boldsymbol{\theta}_i)^T(\mathbf{y}_i - X_i\boldsymbol{\theta}_i) + b^{-1}\right]^{-1}\right)$$

dove $s = \sum_{i=1}^{n} s_i$. Così, la proprietà condizionante del Gibbs sampler consente una forma chiusa delle distribuzioni *full conditional* per ogni parametro nel modello. Riscrivendo il modello (1.15), centrando la covariata sulla media

$$Y_{ij} \sim N(\alpha_i + \beta_i(x_{ij} - \overline{x}), \sigma^2), j = 1, ..., s_i, i = 1, ..., n \tag{1.24}$$

è possibile supporre che le a priori di α_i e β_i siano indipendenti, per cui si può porre $\sum = Diag(\sigma_\alpha^2, \sigma_\beta^2)$ e replicare la a priori Wishard (1.19) con il prodotto di a priori Inverse Gamma, $IV(a_\alpha, b_\alpha)$ e $IV(a_\beta, b_\beta)$. Per i restanti parametri vengono scelte a priori vaghe, vale a dire $C^{-1} = 0$ (che fa scomparire i

parametri $\boldsymbol{\eta}$ dalle distribuzioni condizionate), $a = a_\alpha = a_\beta = \epsilon$ e $b = b_\alpha = b_\beta = 1/\epsilon$, dove $\epsilon = 0.001$.

1.4.2 Modelli di sopravvivenza

Supponiamo di avere uno studio con i individui, con $i = 1, ..., n$, per i quali si hanno $j = 1, ..., s_i$ misure ripetute. Indichiamo con t_{ij} il tempo al decesso o al censoring, con x_{ij} una covariata e con δ_{ij} un indicatore del decesso o meno (1 deceduto, 0 vivo). Secondo l'ipotesi di hazard proporzionale, un individuo con valore della covariata x_{ij} ha un hazard proporzionale all'hazard al baseline $\lambda_0(t)$, al passare del tempo t. Se assumiamo un modello Weibull per l'hazard al baseline tale che $\lambda_0(t) = \rho_i t_{ij}^{\rho_i - 1}$, l'hazard per il soggetto i-simo può essere scritto come

$$\begin{aligned} \lambda(t_{ij}; x_{ij}) &= \lambda_0(t_{ij})\omega_i \exp(\beta_0 + \beta_1 x_{ij}) \\ &= \rho_i t_{ij}^{\rho_i - 1} \exp(\beta_0 + \beta_1 x_{ij} + W_i) \end{aligned} \tag{1.25}$$

dove $\rho_i > 0$, $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ e $W_i = log\omega_i$ è un termine *frailty* unità-specifico. Mentre W_i cattura differenze tra le occasioni di misura dei soggetti, ρ_i rappresenta differenti hazards al baseline; essi possono o crescere ($\rho_i > 1$) o decrescere ($\rho_i < 1$). Possiamo assumere che gli effetti casuali seguano una distribuzione normale, in particolare possiamo porre $W_i \sim N(0, 1/\tau)$ e $\rho_i \sim Gamma(\alpha, \alpha)$.
Inoltre se indichiamo

$$\mu_{ij} = \exp(\beta_0 + \beta_1 x_{ij} + W_i) \tag{1.26}$$

allora $t_{ij} \sim Weibull(\rho_i, \mu_{ij})$. Questa scrittura è usata nella definizione di modelli di sopravvivenza di Weibull in Winbugs.

Capitolo 2

Modelli per dati longitudinali con missing data informativo

2.1 Introduzione

In questo capitolo, dopo aver introdotto il concetto di missing data nell'ambito degli studi longitudinali e dei tipi di modelli sviluppati in letteratura per questo tipo di problema statistico, si è focalizzata l'attenzione su uno di questi tipi di modelli, gli shared parameter models. L'uso di shared (random) parameter models per dati longitudinali è un approccio che tiene conto di un processo di dati mancanti non casuale (NMAR), detto anche missing data informativo. In tale ambito, un modello per misure di risposta longitudinale è linkato con un modello che descrive il meccanismo dei dati mancanti, attraverso un set di effetti casuali che sono condivisi tra i due processi. Nel caso di missing non casuale c'è dipendenza tra il processo di risposta e quello dei dati mancanti. Non tener conto di questa dipendenza può portare a stime distorte. Nello specifico si è trattato il caso in cui il processo di dati mancanti sia un processo di sopravvivenza.

2.2 Missing data informativo nei dati longitudinali

Rubin (1976) e Little e Rubin (1987) formalizzano il meccanismo dei dati mancanti. Tale meccanismo è detto completamente a caso (MCAR) se la probabilità di missing è indipendente sia dagli outcomes osservati che da quelli non osservati, missing a caso (MAR) se la probabilità di missing dipendente solo dagli outcomes osservati, e non missing a caso (NMAR) se la probabilità di missing dipende dagli outcomes non osservati e/o da quelli osservati. Diggle e Kenward (1994) definiscono dropout informativo quello che induce al MNAR. Nel caso di MAR si può ignorare il meccanismo dei dati mancanti in quanto l'inferenza basata sulla verosimiglianza porta a stime non distorte. Si parla di meccanismo dei dati mancanti ignorabile. Le ragioni che portano ad avere dati mancanti in uno studio possono essere varie, per cui nella realtà è difficile classificare un processo come MAR o MNAR. La letteratura sviluppata nell'ambito dei missing data ha portato allo sviluppo di complesse modellizzazioni che si basano su assunzioni non testabili circa la relazione tra le misure osservate e il meccanismo dei dati mancanti, per cui in tale contesto particolare importanza hanno assunto le analisi di sensitività.

In generale per trattare dati longitudinali in presenza di valori mancanti sono stati proposti vari modelli: (a) *selection models*; (b) *pattern-mixture models*; (c) *shared parameter models*, che differiscono per il tipo di fattorizzazione usata. Seguiamo la terminologia di Rubin, in cui Y_{ij} rappresenta la misura per l'i-simo soggetto ($i = 1, ..., N$), all'occasione di misura j ($j = 1, ..., n_i$). Definiamo, per ogni occasione j,

$$R_{ij} = \begin{cases} 1 & Y_{ij} \text{ è osservato} \\ 0 & \text{altrimenti} \end{cases}$$

Consideriamo il vettore $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{in_i})$ partizionato in due sottovettori $\boldsymbol{Y}_i^0$ dei valori osservati, e $\boldsymbol{Y}_i^m$ dei valori mancanti. Il vettore $\boldsymbol{Y}_i$ rappresenta dunque il vettore completo, costituito anche dai possibili valori mancanti, mentre $\boldsymbol{R}_i$ è l'indicatore o processo di dati mancanti. La distribuzione di interesse è quella costituita dai dati completi e dall'indicatore dei dati mancanti $(\boldsymbol{Y}_i, \boldsymbol{R}_i)$. Indicando X_i e W_i le matrici delle covariate per il processo di misura e dei dati mancanti, e con $\boldsymbol{\theta}$ e $\boldsymbol{\psi}$ i rispettivi vettori dei parametri, la distribuzione dei dati completa è data da

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) \tag{2.1}$$

Il *selection model* è basato sulla seguente fattorizzazione

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i | X_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i | \boldsymbol{y}_i, W_i, \boldsymbol{\psi}) \tag{2.2}$$

ovvero dal prodotto della densità marginale del processo di misura e della densità del processo dei dati mancanti, condizionato all'outcome.
Alternativamente si può considerare una famiglia di modelli detta *pattern-mixture models* (Little 1993), che è basata sulla seguente fattorizzazione

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i | \boldsymbol{r}_i, X_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i | W_i, \boldsymbol{\psi}) \tag{2.3}$$

La terza famiglia, nella quale possiamo far rientrare i joint models applicati in questa tesi, è chiamata *shared-parameter models*, per la quale

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{b}_i) = f(\boldsymbol{y}_i | \boldsymbol{r}_i, X_i, \boldsymbol{\theta}, \boldsymbol{b}_i) f(\boldsymbol{r}_i | W_i, \boldsymbol{\psi}, \boldsymbol{b}_i) \tag{2.4}$$

In tali modelli viene inserito un vettore di effetti casuali (o latenti) unità-specifici $\boldsymbol{b}_i$, *shared* o condivisi tra entrambe le componenti della distribuzione congiunta. In tal caso $\boldsymbol{Y}_i$ e $\boldsymbol{R}_i$ sono mutualmente indipendenti, dato l'effetto casuale $\boldsymbol{b}_i$ e le covariate. Nell'ambito di tali modelli, a partire dalla fine degli

anni '80, si è sviluppato un interesse nei metodi di modellizzazione congiunta dei dati longitudinali e di quelli relativi ai tempi di sopravvivenza (Wu e Carroll (1988), Pawitan e Self (1993), De Gruttola e Tu (1994), Faucett e Thomas (1996), Hogan e Laird (1997), Wulfsohn e Tsiatis (1997), Henderson et al. (2000)).

2.3 Shared parameter models

Il primi autori ad introdurre shared parameters models furono Wu e Carroll (1988), i quali modellarono il processo di risposta con un modello lineare ad effetti casuali, collegato ad un processo di censoring, in cui i coefficienti casuali del modello lineare erano considerati come covariate di un modello probit. Reviews di tali modelli sono presenti in Little (1995), Hogan e Laird (1997b) e Vonesh, Greene e Schlucher (2006).

2.3.1 Formulazione del modello

Sia $\boldsymbol{y}_i = (\boldsymbol{y}_i^o, \boldsymbol{y}_i^m)$ il vettore di outcomes longitudinali completo per l'i-simo soggetto, con $\boldsymbol{y}_i^o$ che rappresenta i dati osservati e $\boldsymbol{y}_i^m$ i dati mancanti non osservati. Sia inoltre $\boldsymbol{b}_i \sim N(0, \Sigma)$ il vettore di effetti casuali per il soggetto i-simo che può essere *shared* tra il processo dei dati longitudinali e quello dei dati mancanti. Si indichi inoltre con T_i il tempo all'evento. Questa simbologia differisce della teoria classica sui missing data, in cui il processo di dati mancanti è indicato con $\boldsymbol{R}_i$, dove R_{ij} è un indicatore di presenza di informazione o meno.

La distribuzione congiunta di $\boldsymbol{y}_i$, T_i e $\boldsymbol{b}_i$ può essere scritta, in una notazione

semplificata, che non esplicita i condizionamenti sulle covariate X_i, come

$$f(\boldsymbol{y}_i, \boldsymbol{T}_i, \boldsymbol{b}_i) = f(\boldsymbol{y}_i|\boldsymbol{T}_i, \boldsymbol{b}_i) f(\boldsymbol{T}_i|\boldsymbol{b}_i) f(\boldsymbol{b}_i) \tag{2.5}$$

Si assume che, condizionando sugli effetti casuali $\boldsymbol{b}_i$, le risposte non dipendano dallo stato di missing, dunque la quantità $f(\boldsymbol{y}_i|\boldsymbol{T}_i, \boldsymbol{b}_i)$ diventa $f(\boldsymbol{y}_i|\boldsymbol{b}_i)$. Inoltre si assume che gli elementi di $\boldsymbol{y}_i$ siano condizionatamente indipendenti dato $\boldsymbol{b}_i$, per cui la densità del vettore delle risposte condizionato sugli effetti casuali, $f(\boldsymbol{y}_i|\boldsymbol{b}_i)$, può essere decomposto nel prodotto delle densità per i valori osservati e non osservati di $\boldsymbol{y}_i$

$$f(\boldsymbol{y}_i|\boldsymbol{b}_i) = f(\boldsymbol{y}_i^o|\boldsymbol{b}_i) f(\boldsymbol{y}_i^m|\boldsymbol{b}_i) \tag{2.6}$$

Lo *shared parameter model* per modellare congiuntamente $(\boldsymbol{y}_i, \boldsymbol{T}_i)$ può essere scritto come

$$\begin{aligned} f(\boldsymbol{y}_i, \boldsymbol{T}_i) &= \int_b f(\boldsymbol{y}_i, \boldsymbol{T}_i|\boldsymbol{b}_i) f(\boldsymbol{b}_i) d\boldsymbol{b}_i & (2.7) \\ &= \int_b f(\boldsymbol{y}_i|\boldsymbol{b}_i) f(\boldsymbol{T}_i|\boldsymbol{b}_i) f(\boldsymbol{b}_i) d\boldsymbol{b}_i & (2.8) \end{aligned}$$

avendo ipotizzato che $\boldsymbol{y}_i|\boldsymbol{b}$ e $\boldsymbol{T}_i|\boldsymbol{b}$ siano condizionatamente indipendenti dato $\boldsymbol{b}_i$.

2.3.2 Shared parameter model per missing data non ignorabile o informativo

Assumendo che le y_{ij} siano condizionatamente indipendenti dato $\boldsymbol{b}_i$, lo *shared parameter model* può essere scritto come

$$\begin{aligned} f(\boldsymbol{y}_i^o, \boldsymbol{T}_i) &= \int_{y^m} \int_b f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{T}_i|\boldsymbol{b}_i) f(\boldsymbol{b}_i) d\boldsymbol{b}_i d\boldsymbol{y}_i^m \\ &= \int_{y^m} \int_b f(\boldsymbol{y}_i^o|\boldsymbol{b}_i) f(\boldsymbol{y}_i^m|\boldsymbol{b}_i) f(\boldsymbol{T}_i|\boldsymbol{b}_i) f(\boldsymbol{b}_i) d\boldsymbol{b}_i d\boldsymbol{y}_i^m \\ &= \int_b f(\boldsymbol{y}_i^o|\boldsymbol{b}_i) f(\boldsymbol{T}_i|\boldsymbol{b}_i) f(\boldsymbol{b}_i) \left\{ \int_{y^m} f(\boldsymbol{y}_i^m|\boldsymbol{b}) d\boldsymbol{y}_i^m \right\} d\boldsymbol{b}_i \\ &= \int_b f(\boldsymbol{y}_i^o|\boldsymbol{b}_i) f(\boldsymbol{T}_i|\boldsymbol{b}_i) f(\boldsymbol{b}_i) d\boldsymbol{b}_i \end{aligned} \tag{2.9}$$

Secondo la classificazione fatta da Little (1995) questi modelli ricadono nella classe dei *random-effects dependent selection models*. Tali modelli possono assumere varie forme a seconda dell'outcome dei dati longitudinali, se continuo o discreto, e della funzione che descrive il meccanismo dei dati mancanti. In questa tesi in particolare si farà riferimento a densità congiunte ottenute nel caso di missing monotono (ad esempio quando i pazienti escono dallo studio senza potervi rientrare) e misurato su un tempo continuo. Una reviews di tali modelli è contenuta in Tsiatis e Davidian (2004).

Vediamo perchè lo *shared parameter model* incorpora un meccanismo di dati mancanti NMAR.

Per definizione il missing not at random, NMAR, si ha quando la probabilità di *missingness* dipende dai dati non osservati $\boldsymbol{y}^m$. Lo *shared parameter model* induce correlazione marginale tra $\boldsymbol{y}$ e $\boldsymbol{T}$ attraverso la comune dipendenza da

$\boldsymbol{b}$. La densità congiunta di $\boldsymbol{T}_i$ dato $\boldsymbol{y}_i = (\boldsymbol{y}_i^o, \boldsymbol{y}_i^m)$ è

$$\begin{aligned} f(\boldsymbol{T}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m) &= \frac{\int_b f(\boldsymbol{T}_i, \boldsymbol{y}_i^o, \boldsymbol{y}_i^m|\boldsymbol{b}_i) f(\boldsymbol{b}_i) d\boldsymbol{b}_i}{\int_b f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m|\boldsymbol{b}_i) f(\boldsymbol{b}_i) d\boldsymbol{b}_i} \\ &= \frac{\int_b f(\boldsymbol{T}_i|\boldsymbol{b}_i) f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m|\boldsymbol{b}_i) f(\boldsymbol{b}_i) d\boldsymbol{b}_i}{\int_b f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m|\boldsymbol{b}_i) f(\boldsymbol{b}_i) d\boldsymbol{b}_i} \\ &= \int_b f(\boldsymbol{T}_i|\boldsymbol{b}_i) f(\boldsymbol{b}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m) d\boldsymbol{b}_i \end{aligned}$$

In generale la distribuzione condizionata di $\boldsymbol{T}_i|\boldsymbol{y}_i$ dipende da $\boldsymbol{y}_i^m$ attraverso la distribuzione a posteriori di $\boldsymbol{b}_i$, poichè $f(\boldsymbol{b}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m)$, che può essere vista come una distribuzione a posteriori *empirical Bayes*, dipende da $\boldsymbol{y}_i^m$.
Ci sono vari approcci per la stima dei parametri della distribuzione congiunta attraverso la massimizzazione della funzione di verosimiglianza $L = \prod_i f(\boldsymbol{y}_i^o, \boldsymbol{T}_i)$, dove la f è data dalla (2.9). Poichè la massimizzazione della verosimiglianza prevede l'integrazione sulla distribuzione degli effetti casuali, questa può essere impegnativa dal punto di vista computazionale. Un'alternativa alla valutazione diretta dell'integrale possono essere l'approccio EM Monte Carlo (McCulloch, 1997) o le approssimazioni di Laplace (Gao, 2004). Se gli effetti casuali sono pochi, solo uno o due, possono essere utilizzate tecniche di integrazione numerica come la Gaussian quadrature o adaptive Gaussian quadrature. In genere la massimizzazione della verosimiglianza prevede la scrittura di software per la particolare applicazione.

In generale per descrivere i dati longitudinali si usa un modello lineare ad effetti misti, in cui le misure sono rilevate a determinati intervalli di tempo o occasioni. Tali modelli sono abbastanza flessibili in quanto non richiedono che i dati $\boldsymbol{y}_i$ siano osservati allo stesso set di occasioni o abbiano le stesse dimensioni. Per quanto riguarda il meccanismo dei dati mancanti consideriamo il caso particolare in cui esso sia dato da un processo di sopravvivenza, che è possibile modellare in quanto si dispone di informazioni sull'eventuale

verificarsi dell'evento di interesse (decesso, malattia o altro) e della durata intercorsa prima del verificarsi dell'evento o fino alla fine dello studio (in caso di right censoring).

Follman e Wu (1995) descrivono shared parameter models per studi longitudinali con MNAR o informativo, usando una approccio generale tramite modelli lineari generalizzati

$$\begin{aligned} \eta_y(E[\boldsymbol{y}_i|\boldsymbol{b}_i]) &= \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i \\ \eta_T(E[\boldsymbol{T}_i|\boldsymbol{b}_i]) &= \boldsymbol{W}_i\gamma_1 + \boldsymbol{b}_i\gamma_2 \end{aligned}$$

dove η_y e η_T sono funzioni link monotone, e $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$ e $\boldsymbol{W}_i$ sono matrici delle covariate fisse.

In sintesi uno shared parameter model è ottenuto specificando due sottomodelli:

- un modello condizionato per i dati longitudinali $f(\boldsymbol{y}_i|\boldsymbol{b}_i)$, in cui le $\boldsymbol{b}_i$ rappresentano effetti casuali soggetto-specifici
- un modello condizionato per i dati di sopravvivenza $f(\boldsymbol{T}_i|\boldsymbol{b}_i)$ in cui le $\boldsymbol{b}_i$ rappresentano covariate soggetto specifiche.

Ora, se T_i è il tempo all'evento, noi osserviamo $T_i^* = min(T_i, C_i)$ dove C_i è il tempo di censoring, e δ_i, variabile indicatrice che è uguale ad 1 se $T_i^* = T_i$ e 0 se $T_i^* = C_i$. La verosimiglianza per lo shared parameter model è

$$L(\boldsymbol{\theta}_y, \boldsymbol{\theta}_{T|y}) = \prod_{i=1}^{N} \int_b f(\boldsymbol{y}_i|\boldsymbol{b}_i; \beta, \sigma^2) f(\boldsymbol{b}_i; \Sigma) f(\boldsymbol{T}_i^*, \delta_i|\boldsymbol{b}_i; \boldsymbol{\theta}_{T|y}) d\boldsymbol{b}i$$

dove

$$f(\boldsymbol{T}_i^*, \delta_i|\boldsymbol{b}_i; \boldsymbol{\theta}_{T|y}) = \left[f_{T|y}(T_i^*|\boldsymbol{b}_i; \boldsymbol{\theta}_{T|y})\right]^{\delta_i} \left[1 - F_{T|y}(T_i^*|\boldsymbol{b}_i; \boldsymbol{\theta}_y)\right]^{1-\delta_i}$$

Capitolo 3

Joint models per dati longitudinali e di sopravvivenza

3.1 Introduzione

Molti studi epidemiologici consistono nel seguire nel tempo una coorte di soggetti per esaminare la relazione tra una o più variabili esplicative e il rischio di sviluppare una malattia o di morire. Le misure delle variabili esplicative sono spesso rilevate a periodici intervalli di tempo e generano un dataset costituito da misure ripetute di covariate tempo-dipendenti, misure singole di covariate fissate, e il tempo di sviluppo della malattie o del decesso o il censoring per ogni soggetto (Faucett e Thomas (1996)). In generale, l'interesse epidemiologico è su due aspetti di questi dati: a) la relazione tra un outcome misurato ripetutamente e il tempo e/o altre covariate, ovvero come si modifica nel tempo una variabile (es. biomarker) e quali sono i fattori di rischio che possono influenzare tale variazione; b) la relazione tra la covariata tempo dipendente e la probabilità di sviluppare una malattia o il decesso del soggetto. Entrambi questi problemi possono essere considerati in un'unica

analisi modellando un outcome continuo sul tempo e simultaneamente collegando l'outcome al rischio di malattia o morte, a seconda dell'interesse dello studio.
In tale ambito a partire dalla fine degli anni '80 è stato sviluppato l'approccio joint modeling, che utilizza i dati in maniera efficiente, riducendo la distorsione dovuta al missing data informativo, problema spesso presente nei dati longitudinali, poiché produce stime più accurate della forza della associazione tra l'outcome del modello longitudinale e il rischio di malattia o morte.

I joint models presentati in questo capitolo sono modelli per l'analisi dei dati che provengono da studi longitudinali nei quali ogni soggetto produce una sequenza di misure a pre-specificati tempi nel follow-up e tempi di sopravvivenza.

3.2 Formulazione del modello

In presenza di missing data informativo, ovvero quando il processo di missing data è dipendente dal tasso di cambiamento individuale dell'outcome considerato nel modello longitudinale, le stime sono distorte, per cui si deve tener conto nella modellizzazione del processo di missing data. Un processo di missing data informativo può essere ad esempio causato dalla morte dei partecipanti allo studio. Nella realtà possono essere presenti altri processi di missing data non informativi, e indipendenti dal processo primario informativo.
Wu e Carroll (1988) furono i primi ad introdurre i joint models per tener conto del missing data informativo negli studi longitudinali, adottando un modello lineare ad effetti misti per i dati longitudinali e un modello probit

per modellare la probabilità del verificarsi dell'evento di interesse (esempio decesso).
Il modello lineare ad effetti misti può essere scritto come:

$$Y_{ij} = \mathbf{X}_i(t_{ij})\beta + Z_i(t_{ij})\mathbf{b}_i + e_{ij} \tag{3.1}$$

dove si assume generalmente che $Z_i(t)' = (1, t)$, ovvero un modello ad intercetta e pendenza casuale. Nel modello le quantità e_{ij} e $\mathbf{b}_i$ rappresentano rispettivamente le misure d'errore e gli effetti casuali, entrambi mutualmente indipendenti, Gaussiani, e a media nulla. Wulfsohn e Tsiatis (1997) proposero, come modello dell'hazard, un modello di Cox

$$\lambda_i(t) = \lambda_0(t)exp\left\{\alpha\mathbf{Z}_i(t)\mathbf{b}_i\right\} \tag{3.2}$$

dove $\lambda_0(t)$ è una funzione non parametrica. Secondo questo modello il parametro α determina l'associazione tra il modello longitudinale e il tasso di dropout, e l'eventuale effetto casuale (frailty), nel modello di sopravvivenza.

3.3 Joint model nella impostazione bayesiana

Faucett e Thomas (1996) propongono un'impostazione bayesiana di un joint model, veramente simile al modello a due stadi proposto da Tsiatis (1995), utilizzando il metodo Marcov Chain Monte Carlo del Gibbs sampling per stimare la distribuzione a posteriori dei parametri non noti del modello, dato i dati osservati. Il modello proposto è valido anche per dati non equamente spaziati (unbalanced), con differenti numeri di osservazioni per soggetto (missing data) e censoring dei tempi di sopravvivenza. Faucett e Thomas (1996), attraverso studi di simulazione in cui confrontano le stime ottenute

dal joint model con quelle derivanti dalle analisi separate ottenute usando metodi standard, mostrano che, modellando congiuntamente i dati longitudinali e di sopravvivenza, si riduce la distorsione dei parametri dovute a missing data informativo (nel modello longitudinale) e a covariate misurate con errore (nel modello di sopravvivenza).

Modello congiunto e assunzioni

Si presuppone che per ogni soggetto $i = 1, ..., I$ si abbiano le seguenti informazioni: (i) $j = 1, ..., J_i$ misure di una covariata continua tempo dipendente z_{ij}, possibilmente misurata con errore; (ii) tempi di osservazione di queste misure t_{ij}; (iii) tempi di ingresso nello studio e_i; (iv) indicatore dello stato di malattia o decesso d_i; (v) tempo intercorso dal baseline all'evento malattia (o decesso) o dell'intero follow-up in caso di censoring s_i. Il modello non richiede che il numero di osservazioni o gli intervalli di tempo tra le osservazioni siano uniformi tra i soggetti. Gli autori specificano il modello congiunto in termini di due sottomodelli: (i) il *covariate tracking model*, modello longitudinale che descrive la relazione tra un outcome misurato ripetutamente e il tempo, ed eventualmente altre covariate ; (ii) il *disease risk model*, modello di sopravvivenza, che descrive la relazione tra il rischio di malattia o decesso e la covariata tempo-dipendente, outcome del modello longitudinale.

Covariate Tracking Model

Gli autori assumono un classico *measurement error model* per la covariata osservata. Il modello, proposto in maniera semplificata con il solo predittore tempo, ma che è possibile generalizzare introducendo altre variabili esplicative, è

$$z_{ij} = x_i(t_{ij}) + \epsilon_{ij} \tag{3.3}$$

dove $x_i(t_{ij})$ è il valore della vera covariata, non osservata, al tempo t_{ij}, e ϵ_{ij} sono termini d'errore assunti indipendenti e normalmente distribuiti a media nulla.
Si è assunto, per il valore della vera covariata, un modello ad effetti casuali con intercetta e pendenza soggetto-specifica

$$x_i(t) = \alpha_i + \beta_i(t) \tag{3.4}$$

Si è assunto inoltre che gli effetti casuali abbiano una distribuzione normale bivariata con media μ_α e μ_β e matrice di varianza e covarianza

$$\Sigma = \begin{pmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{pmatrix}$$

dove $\sigma_{\alpha\beta} = \rho\sigma_\alpha\sigma_\beta$ e ρ è la correlazione tra l'intercetta casuale α_i e la pendenza casuale β_i.

Disease Risk Model
Per la parte del modello relativa ai dati di sopravvivenza gli autori assumono un proportional hazard model con rischio che dipende log-linearmente dal vero valore della covariata longitudinale. L'hazard per il soggetto i al tempo t è

$$\lambda_i(t) = \lambda_0(t)exp(\gamma x_i(t)) \tag{3.5}$$

dove $\lambda_0(t)$ è l'hazard al baseline. Nello specifico gli autori assumono un modello per l'hazard al baseline, costante per sottointervalli di tempo, partizionando la scala del tempo in $k = 1, ..., K$ intervalli, con $\lambda_0(t) = \lambda_k$.
Gli autori stimano la distribuzione congiunta a posteriori di tutti i parametri del modello non noti

$$\left[\{\alpha_i, \beta_i\}, \mu_\alpha, \mu_\beta, \Sigma, \sigma_\epsilon^2, \lambda_0(t), \gamma | \{z_{ij}\}, \{t_{ij}\}, \{s_i\}, \{d_i\}, \{e_i\}\right] \tag{3.6}$$

usando un *Gibbs sampling*, un metodo Monte Carlo che consente di generare campioni dalla distribuzione a posteriori congiunta dei parametri non noti del modello, condizionatamente solo ai dati osservati. Il metodo prevede di campionare iterativamente dalla distribuzione full conditional di ogni parametro, dati i valori correnti assegnati a tutti gli altri parametri, e i dati osservati. Faucett e Thomas utilizzano delle *flat prior* per i parametri μ_α, μ_β,γ, $|\Sigma|^{-3/2}$ per Σ, $1/\sigma_\epsilon^2$ per σ_ϵ^2 e $1/\lambda_k$ per λ_k. La procedura usata per generare i campioni da ogni distribuzione è la seguente:

Effetti Casuali: α_i *e* β_i*, con* $i = 1, ..., I$

La distribuzione a posteriori per ogni effetto casuale (per esempio α_i) all'(*m+1*)-sima iterazione, è proporzionale al prodotto delle tre distribuzioni condizionali

$$\left[\alpha_i|\beta_i^{(m)}, \mu_\alpha^{(m)}, \mu_\beta^{(m)}, \Sigma^{(m)}, \sigma_\epsilon^{2(m)}, \lambda_0(t)^{(m)}, \gamma^{(m)}, \{z_{ij}\}, s_i, d_i\right] \tag{3.7}$$

$$\propto \left[\{z_{ij}\}|\alpha_i, \beta_i^{(m)}, \sigma_\epsilon^{2(m)}\right] \times \left[\alpha_i|\beta_i^{(m)}, \mu_\alpha^{(m)}, \mu_\beta^{(m)}, \Sigma^{(m)}\right] \times \left[s_i, d_i|\alpha_i, \beta_i^{(m)}, \lambda_0(t)^{(m)}, \gamma^{(m)}\right] \tag{3.8}$$

per $i = 1, ..., I$, mentre l'apice *(m)* indica l'm-simo valore campionato per ogni parametro. I primi due termini sono distribuzioni normali rispettivamente proporzionali a

$$\exp\left\{-\sum_{j=1}^{J_i}\left[z_{ij} - (\alpha_i + \beta_i t_{ij})\right]^2 / \left(2\sigma_\epsilon^2\right)\right\} \tag{3.9}$$

e

$$\exp\left\{-\left[\alpha_i - \left(\mu_\alpha + \frac{\sigma_{\alpha\beta}}{\sigma_\beta^2}\left(\beta_i - \mu_\beta\right)\right)\right]^2 / \left[2\sigma_\alpha^2\left(1-\rho^2\right)\right]\right\} \tag{3.10}$$

mentre l'ultimo termine è la full likelihood dei parametri di sopravvivenza basati sui valori assunti veri, non osservati, della covariata

$$\left\{\lambda_0(s_i)\exp\left[\gamma\left(\alpha_i + \beta_i s_i\right)\right]\right\}^{d_i} \exp\left\{-\int_{e_i}^{s_i}\lambda_0(t)\exp\left[\gamma\left(\alpha_i + \beta_i t\right)\right]\mathrm{d}t\right\} \tag{3.11}$$

Poichè la forma della distribuzione dei parametri è abbastanza complicata, gli autori usano una procedura rejection sampling per generare i campioni dalla distribuzione a posteriori esatta.

Medie degli effetti casuali: μ_α e μ_β

La distribuzione a posteriori, ad esempio di μ_α (e in maniera simile per μ_β), è normale con media $\overline{\alpha} - \left(\overline{\beta} - \mu_\beta\right) \sigma_{\alpha\beta}/\sigma_\beta^2$ e varianza $\sigma_\alpha^2 \left(1 - \rho^2\right)/I$, dove $\overline{\alpha}$ e $\overline{\beta}$ sono le medie tra i soggetti dei valori degli effetti casuali campionati correntemente.

Effetti casuali della matrice di Varianza e Covarianza: Σ

Per generare la matrice di varianza e covarianza Σ gli autori seguono la procedura di Zeger e Karim (1991). Prima generano una distribuzione Wishart standardizzata con $I-1$ gradi di libertà, $\mathbf{W}$, e poi calcolano $\Sigma = (\mathbf{H}'\mathbf{W}\mathbf{H})^{-1}$ dove $\mathbf{H}$ è

$$\left(-\sum_{i=1}^{I} \mathbf{B}_i \mathbf{B}_i'\right)^{-1} = \mathbf{H}'\mathbf{H} \tag{3.12}$$

e $\mathbf{B}_i$ è il vettore $\left[\left(\alpha_i - \mu_\alpha\right), \left(\beta_i - \mu_\beta\right)\right]'$.

Varianza dell'errore: σ_ϵ^2

Con una a priori proporzionale a $1/\sigma_\epsilon^2$, la distribuzione a posteriori per la varianza dell'errore è un Inverse-Gamma con parametri:

$$N/2 = \sum_{i=1}^{I} J_i/2$$

e

$$SS/2 = \sum_{i=1}^{I} \sum_{j=1}^{J_i} \left[z_{ij} - \left(\alpha_i + \beta_i t_{ij}\right)\right]^2/2 \tag{3.13}$$

Per generare il campione dalla distribuzione a posteriori di σ_ϵ^2 si può calcolare

SS/χ^2_N.

Baseline Hazard: λ_k, $k = 1, ..., K$

La distribuzione a posteriori del baseline hazard λ_k ha una distribuzione Gamma con parametri

$$D_k = \sum_{i=1}^{I} d_i I_{\{t_{k-1} < s_i \leq t_k\}}$$

e

$$Y_k = \sum_{i=1}^{I} I_{\{s_i > t_{k-1}\}}^{min(s_i, t_k)} \exp\left[\gamma\left(\alpha_i + \beta_i t\right)\right] \mathtt{d}t \tag{3.14}$$

dove $I_{\{t_{k-1} < s_i \leq t_k\}}$ è un indicatore che l'i-simo soggetto sviluppi la malattia, mentre $I_{\{s_i > t_{k-1}\}}$ è un indicatore che indica se il soggetto è stato sotto osservazione nel k-simo periodo.

Parametro del rischio di malattia

La distribuzione a posteriori di γ è proporzionale alla verosimiglianza data nell'equazione (3.11).

Gli autori usano un rejection sampling tradizionale per generare i campioni da questa distribuzione (Ripley B. (1987)).

3.4 Generalizzazione dei joint models di Henderson, Diggle e Dobson (2000)

La letteratura sviluppata sui modelli congiunti è stata basata prevalentemente su applicazioni specifiche per specifici dataset, per cui gli obiettivi di investigazione statistici e scientifici dipendevano dalle applicazioni di interesse e potevano focalizzarsi sul processo longitudinale (per risolvere un

eventuale drop-out informativo), sul processo di sopravvivenza (per introdurre una covariata tempo-dipendente misurata con errore) o dando uguale importanza ai due processi longitudinale e di sopravvivenza. Nel loro articolo Henderson, Diggle e Dobson (2000) considerano di eguale interesse le componenti longitudinali e di sopravvivenza del modello congiunto proposto, sebbene la distinzione è più concettuale in quanto la metodologia si applica alla stessa maniera negli altri due casi.
Henderson, Diggle e Dobson (2000) sviluppano una metodologia molto flessibile, formulando una classe di modelli per il comportamento congiunto di dati longitudinali e di sopravvivenza, che includevano una serie di modelli proposti in precedenza da altri autori. Tale classe di modelli, in assenza di associazione tra processo longitudinale e di sopravvivenza, dovrebbe ricondurre agli stessi risultati ottenuti considerando delle analisi separate sui dati, ovvero un modello gaussiano lineare con errori correlati (modelli lineari ad effetti misti) e un modello di hazard proporzionale semi-parametrico o un modello parametrico con o senza frailty.
Gli autori postulano un processo Gaussiano bivariato latente

$$W(t) = \{W_1(t), W_2(t)\}$$

ed assumono che i due processi di misura e di evento siano condizionatamente indipendenti dato $W(t)$ e le covariate.

Il modello generale e la verosimiglianza associata
La notazione usata dai due autori non è quella standard adottata nel caso di analisi longitudinali o di sopravvivenza. Si considerino m soggetti seguiti sull' intervallo di tempo $[0, \tau)$.
Per ogni soggetto i si hanno una serie di misure quantitative $\{y_{ij} : j = 1, ..., n_i\}$ al tempo $\{t_{ij} : j = 1, ..., n_i\}$, insieme alla realizzazione di un processo di conto

$\{N_i(u) : 0 \leq u \leq \tau\}$ per gli eventi (malattia o morte) e di un processo zero-uno $\{H_i(u) : 0 \leq u \leq \tau\}$, che indica se il soggetto è a rischio di vivere l'evento al tempo u. Il processo di conto ha salti ai tempi $\{u_{ij} : j = 1, ..., N_i(\tau)\}$, con $N_i(\tau)$ che non può essere maggiore di uno per i dati di sopravvivenza. Gli autori assumono che il tempo in cui vengono rilevate le misure t_{ij} è non-informativo, cioè indipendente dai processi di misura e di conto. Si assume inoltre che il censoring del tempo di sopravvivenza, che può esserci in seguito alla fine dello studio, sia non-informativo, e che i dati provenienti da soggetti diversi siano generati da processi indipendenti.

Formulazione del modello

Gli autori linkano i due sottomodelli, che chiamano *measurement* e *intensity*, attraverso un processo Gaussiano bivariato latente a media nulla $W(t) = \{W_1(t), W_2(t)\}$. Tali sottomodelli sono:

(a) *sottomodello measurement*

Data la sequenza di misure $y_{i1}, y_{i2}, \ldots$, ai tempi $t_{i1}, t_{i2}, \ldots$, il processo di misura è dato da

$$Y_{ij} = \mu_i(t_{ij}) + W_{1i}(t_{ij}) + Z_{ij} \tag{3.15}$$

dove $\mu_i(t_{ij})$ rappresenta la risposta media e $Z_{ij} \sim N(0, \sigma_z^2)$ sono errori mutualmente indipendenti. Si assume che $\mu_i(t)$ siano descritti da un modello lineare

$$\mu_i(t) = x_{1i}(t)'\beta_1$$

nel quale i vettori $x_{1i}(t)$ rappresentano variabili esplicative possibilmente tempo-dipendenti, con β_1 i corrispondenti coefficienti di regressione.

(b) *sottomodello intensity*

Il processo d'intensità dell'evento al tempo t è dato da un modello moltiplicativo semi-parametrico

$$\lambda_i(t) = H_i(t)\alpha_0(t)\exp\{x_{2i}(t)'\beta_2 + W_{2i}(t)\} \tag{3.16}$$

con la forma di $\alpha_0(t)$ non specificata. I vettori $x_{2i}(t)$ possono, ma non necessariamente, avere elementi in comune con $x_{1i}(t)$.

I processi per i dati longitudinale Y e per i dati di sopravvivenza N sono condizionatamente indipendenti, date le covariate X e i due processi latenti W_1 e W_2. Il link tra W_1 e W_2 è detto associazione latente. Tale associazione può essere assente. La combinazione delle equazioni (3.15) e (3.16) abbraccia un ampio range di modelli specifici proposti per le analisi separate per misure longitudinali continue (esempio Laird e Ware (1982)) e per outcome di sopravvivenza. In particolare, si può generalizzare $W_{1i}(t)$ come

$$W_{1i}(t) = d_{1i}(t)'U_{1i} + V_{1i}(t) \tag{3.17}$$

dove $d_{1i}(t)$ indica il vettore di variabili esplicative, $U_{1i} \sim MNV(0, \Sigma_1)$ è il corrispondente vettore di effetti casuali e $V_{1i}(t)$ è un processo Gaussiano stazionario a media nulla, varianza σ^2_{v1} e funzione di correlazione $r_1(u) = \mathtt{cov}\{V_{1i}(t), V_{1i}(t-u)/\sigma_1^2\}$. Il processo $W_{2i}(t)$ è specificato in maniera simile a $W_{1i}(t)$.

La funzione di verosimiglianza

Sia θ il vettore dei parametri non noti presenti nel modello congiunto. Henderson, Diggle e Dobson (2000) definiscono $\mathcal{W}_{2i}$ il path completo di W_{2i} sull'intervallo $[0, \tau)$, e $\mathcal{W}_2$ l'insieme di questi path su tutti i soggetti. Condizionatamente a $\mathcal{W}_2$, i dati di sopravvivenza sono indipendenti dai dati longitudinali, per cui si può scrivere la funzione di verosimiglianza $L = L(\theta, Y, N)$

come il prodotto della verosimiglianza della distribuzione marginale delle misure osservate Y per la distribuzione condizionale degli eventi, N, dati i valori osservati Y, ovvero

$$L = L_Y \times L_{N|Y} = L_Y(\theta, Y) \times E_{\mathcal{W}_2|Y}\left\{L_{N|W_2}\left(\theta, N|\mathcal{W}_2\right)\right\} \tag{3.18}$$

dove $L_Y(\theta, Y)$ è la distribuzione marginale normale multivariata della Y, mentre $L_{N|W_2}\left(\theta, N|\mathcal{W}_2\right)$ è la verosimiglianza condizionale per i dati di sopravvivenza. Indicando con $A_0(t) = \int_0^t \alpha_0(u)du$ l'intensità cumulativa al baseline, si può scrivere

$$L_{N|W_2}\left(\theta, N|\mathcal{W}_2\right) = \prod_i \left(\left[\prod_t \left[\exp\left\{x_{2i}(t)'\beta_2 + W_{2i}(t)\right\}\alpha_0(t)\right]^{\Delta N_i(t)} \right] \right.$$
$$\left. \times \exp\left[- \int_0^\tau H_i(t) \exp\left\{x_{2i}(t)'\beta_2 + W_{2i}(t)\right\} dA_0(t) \right] \right)$$

Per la parte longitudinale gli autori assumono un modello lineare ad effetti casuali di Laird e Ware (1982), come già fatto nella precedente letteratura (Tsiatis et al. (1995), Faucett e Thomas (1996) e Wulfsohn e Tsiatis (1997)),

$$W_1(t) = U_1 + U_2(t) \tag{3.19}$$

dove (U_1, U_2) è una distribuzione bivariata gaussiana a media nulla e varianze σ_1^2 e σ_2^2.

Per la parte di sopravvivenza gli autori propongono come specificazione

$$W_2(t) = \gamma_1 U_1 + \gamma_2 U_2 + \gamma_3(U_1 + U_2(t)) + U_3 \tag{3.20}$$

dove $U_3 \sim N(0, \sigma_3^2)$ è indipendente da (U_1, U_2) e rappresenta un frailty, mentre γ_1, γ_2 e γ_3 misurano l'associazione introdotta rispettivamente attraverso l'intercetta, la pendenza, e il valore corrente W_1.

Il metodo di stima adottato dagli autori è un' estenzione dell'*algoritmo EM* di

massimizzazione della verosimiglianza proposto da Wulfsohn e Tsiatis (1997). Secondo tale metodo nel *E-step* si ottengono le log-verosimiglianze attese per i dati completi, condizionati ai dati osservati e alle stime correnti dei paramentri, e nell'*M-step* si massimizza la log-verosimiglianza attesa per ottenere nuove stime dei parametri. Si procede fino a convergenza.

3.5 Ulteriori sviluppi

Xu e Zeger (2001) estendono il modello di Faucett e Thomas (1996). Essi propongono un modello a processi latenti più flessibile rispetto a quelli proposti in precedenza in quanto applicabile a dati di misure ripetute continue, di conteggio o categoriche. Il modello include, oltre agli effetti casuali, un processo correlato serialmente per permettere una flessibile modellizzazione della autocorrelazione nel processo della variabile longitudinale misurata ripetutamente.

Wang e Taylor (2001) propongono un'estensione del modello di Faucett e Thomas (1996) che prevede un modello longitudinale per dati continui che incorpora, oltre alle covariate, un'intercetta casuale e una misura d'errore, un processo stocastico *integrated Ornstein-Uhlenbeck* (IOU), che rappresenta una famiglia di strutture di covarianza. Gli autori costruiscono una traiettoria lineare usando intercette casuali e una pendenza fissa. Il processo IOU porta la traiettoria a variare intorno ad una linea parametrica stimata. Includendo il processo stocastico, la componente longitudinale del modello fornisce una struttura della traiettoria individuale dell'outcome (biomarker) più flessibile rispetto a quella derivante da un modello ad effetti casuali standard, anche se uno svantaggio può essere la complessità del modello.

Brown e Ibrahim (2003) propongono un joint model che rilassa le assunzioni distribuzionali del modello longitudinale usando come a priori sui parametri del modello longitudinale il processo di Dirichlet. In tal modo la distribuzione a posteriori dei parametri longitudinali è libera da vincoli parametrici, portando a stime più robuste.

La letteratura statistica sui joint model si è anche sviluppata estendendosi al caso di misure ripetute multiple e outcome dei tempi di sopravvivenza multipli (Huang et a. (2001), Chi e Ibrahim (2006), Deslandes E., Chevret S. (2010)).

Capitolo 4

Applicazione

4.1 Introduzione

In questa tesi sono stati applicati joint models per dati longitudinali e di sopravvivenza alla coorte ULSAM al fine di analizzare l'effetto nel tempo dello stile di vita (attività fisica, fumo, obesità,...) sul biomarker GFR, che è un indicatore della funzionalità renale. I dati del biomarker, o meglio della cistatina, in base alla quale sono stati ricavati i valori stimati del GFR, sono misurati ripetutamente a distanza di vari anni sulla coorte di uomini. Tali dati non sono completi ma presentano valori mancanti, poichè non tutti gli uomini erano presenti alle visite successive rispetto a quella considerata come baseline, effettuata quando gli individui avevano circa 70 anni.
La mancata presenza alle visite successive, se dovuta a cause diverse dal decesso, si assume produca nei dati un missing data casuale (MAR). Nel caso in cui, invece, si è verificato il decesso, i dati mancati sono considerati possibilmente non casuali (NMAR o informativi).
Per tener conto del supposto missing data informativo dovuto al decesso, sono stati applicati i joint models teorizzati da Henderson, Diggle e Dob-

son (2000), secondo un'impostazione Bayesiana proposta da Guo e Carlin (2004).

Per l'applicazione ai dati ULSAM sono state effettuate delle elaborazioni adattando al caso specifico programmi scritti in Winbugs, software dedicato alle analisi bayesiane. I risultati del joint model sono stati confrontati con quelli derivanti dalle analisi separate, longitudinale e di sopravvivenza, applicate sia secondo l'impostazione frequentista che bayesiana.

4.2 Joint model bayesiano di Guo e Carlin

L'idea chiave in Henderson et al. (2000) è di connettere i processi longitudinali e di sopravvivenza con un processo Gaussiano bivariato latente. I dati longitudinali e di sopravvivenza sono assunti indipendenti, dato il processo latente e le covariate considerate nei modelli.

Guo e Carlin (2004) implementarono la versione frequentista dell'approccio proposto da Henderson et al. (2000) utilizzando il software SAS. Il programma, abbastanza complicato, prevedeva solo la stima puntuale e gli errori standard asintotici dei parametri. Questo motivò gli autori a ricercare una soluzione alternativa a quella classica attraverso un approccio Bayesiano, che permetta una stima a posteriori esatta utilizzando metodi Marcov chain Monte Carlo (MCMC).

Un problema nell'analisi bayesiana può essere la scelta delle distribuzioni a priori che dovrebbero essere sufficientemente non informative, ed il confronto tra i modelli, che può essere effettuato tramite il DIC criterion (Spiegelhalter et al. 2002).

4.2.1 Approccio classico

Supponiamo di avere m soggetti seguiti nell'intervallo di tempo $[0, \tau)$. Per ogni soggetto si dispone di un insieme di misure longitudinali quantitative $\{y_{ij}, j = 1, ..., n_i\}$ ai tempi $\{s_{ij}, j = 1, ..., n_i\}$. Tali misure possono essere parzialmente *missing*. Si hanno inoltre dei tempi di sopravvivenza t_i ad un certo endpoint, con possibilità di misure censurate. Brevemente riprendiamo l'approccio proposto da Henderson et al. (2000) per le analisi separate ed il joint model.

Modelli dei dati longitudinali

Consideriamo modelli misti longitudinali, con effetti fissi e casuali, come in Laird e Ware (1982).
Data una successione di misure per l'i-simo soggetto $y_{i1}, y_{i2}, ..., y_{in_i}$, ripetute ai tempi $s_{i1}, s_{i2}, ..., s_{in_i}$, un modello lineare ad effetti casuali o misti per dati continui gaussiani è modellato come

$$y_{ij} = \mu_i(s_{ij}) + W_{1i}(s_{ij}) + \epsilon_{ij} \tag{4.1}$$

dove $\mu_i(s) = \mathtt{x}_{1i}^T(s)\beta_1$ è la risposta media, $W_{1i}(s) = \mathtt{d}_{1i}^T(s)U_i$ incorpora effetti casuali soggetto-specifici, e $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ è una sequenza di misure di errore mutualmente indipendenti. Il vettore $\mathtt{x}_{1i}(s)$ può essere costituito anche da covariate time-varying, mentre i β_1 sono i corrispondenti coefficienti di regressione.

Modelli dei dati di sopravvivenza

Per modellare i dati di sopravvivenza si possono utilizzare modelli parametrici, come l'esponenziale o la Weibull, o semiparametrici come il modello ad

hazard proporzionale di Cox.

Modello di Weibull

Assumiamo che il tempo di sopravvivenza per l'i-simo soggetto segua una distribuzione di Weibull $t_i \sim \text{Weibull}(r, \mu_i(t))$, dove

$$log(\mu_i(t)) = \mathtt{x}_{2i}^T(t)\beta_2 + W_{2i}(t) \tag{4.2}$$

e $r > 0$. Il vettore $\mathtt{x}_{2i}(t)$ può essere costituito anche da covariate time-varying, mentre i β_2 sono i corrispondenti coefficienti di regressione. Alcune o tutte le variabili contenute nel modello di sopravvivenza possono essere le stesse presenti nel modello longitudinale. $W_{2i}(t)$ include effetti delle covariate soggetto-specifiche e un'intercetta (detta *frailty*). L'hazard al tempo t è dato da

$$\lambda_i(t) = rt^{r-1}\mu_i(t) = rt^{r-1}exp\left(\mathtt{x}_{2i}^T(t)\beta_2 + W_{2i}(t)\right) \tag{4.3}$$

che è monotono in t, decrescendo per $r < 1$ e crescendo per $r > 1$, mentre è costante per $r = 1$. In quest'ultimo caso la Weibull si riconduce ad una distribuzione esponenziale, con hazard costante nel tempo.

Modello di Cox

Un modello dell'hazard proporzionale semiparametrico di Cox è del tipo

$$\lambda_i(t) = \lambda_0(t)exp\left(\mathtt{x}_{2i}^T(t)\beta_2 + W_{2i}(t)\right) \tag{4.4}$$

in cui la hazard function al baseline $\lambda_0(t)$ può assumere una qualsiasi forma (si veda Cox e Oakes, 1984).

4.2.2 Joint Model

Il joint model proposto da Henderson et al. (2000) prevede di modellare congiuntamente i due processi, longitudinale e di sopravvivenza, tramite un

processo Gaussiano bivariato latente a media zero su $(W_{1i}, W_{2i})^T$. Quando l'associazione tra i due processi esiste, l'inferenza prodotta dal joint model dovrebbe essere meno distorta e più efficiente. Il modello congiunto è specificato nel modo seguente

$$W_{1i}(s) = U_{1i} + U_{2i}s \tag{4.5}$$

che rappresenta un modello longitudinale con intercetta e pendenza casuale. Questa espressione non deve necessariamente essere lineare. Inoltre

$$W_{2i}(t) = \gamma_1 U_{1i} + \gamma_2 U_{2i} + \gamma_3(U_{1i} + U_{2i}(t)) + U_{3i} \tag{4.6}$$

dove i parametri γ_1, γ_2 e γ_3 misurano l'associazione tra i due sottomodelli indotta dalla intercetta casuale, dalla pendenza casuale e dal valore longitudinale previsto al tempo dell'evento $W_{1i}(t)$. La coppia di variabili latenti U_{1i} e U_{2i} ha distribuzione bivariata Gaussiana $(U_{1i}, U_{2i})^T \sim N(0, \sum)$, mentre $U_{3i} \sim N(0, \sigma_3^2)$ rappresenta termini frailty indipendenti da $(U_{1i}, U_{2i})^T$.

4.3 Analisi dei dati ULSAM

4.3.1 Descrizione dei dati

I dati usati per l'applicazione provengono da uno studio longitudinale osservazionale condotto in Svezia, l'Uppsala Longitudinal Study of Adult Men (ULSAM).

Lo studio è iniziato nel 1970 e ha previsto l'arruolamento di tutti gli uomini viventi nell' Uppsala Country nati tra il 1920 e il 1924. Al primo arruolamento parteciparono allo studio 2322 uomini di circa 50 anni. 1221 uomini furono di nuovo esaminati dopo 20 anni, all'età di circa 70 anni, e ulteriori esami furono fatti alla coorte all'età di circa 77 e 82 anni. È stato somministrato un questionario che includeva informazioni demografiche e su abitudini

di vita, ed effettuate misurazioni (pressione del sangue, peso, altezza,...) e analisi di laboratorio su campioni di sangue e urine. Oltre ai dati longitudinali sono stati ricavati i dati di sopravvivenza, avvalendosi dell'archivio di mortalità locale. Maggiori dettagli sullo studio longitudinale ULSAM si possono trovare sul sito web: http://www.pubcare.uu.se/ULSAM/

4.3.2 Lo studio

Motivazione dello studio da un punto di vista epidemiologico

Dalla letteratura sappiamo che la compromissione della funzione renale è altamente prevalente negli anziani ed è un fattore di rischio per malattie cardiovascolari e mortalità. L'interesse prevalente dello studio è sui fattori di rischio, collegati allo stile di vita, del declino nel tempo della funzionalità renale, misurata dal tasso di filtrazione glomerulare o *Glomerular Filtration Rate* (GFR), in età adulta nella popolazione generale. Lo studio è stato effettuato sulla coorte di uomini ULSAM.

Variabili collegate al declino della funzione renale

I predittori del declino della funzione renale sono stati tratti dalla letteratura. Il principale predittore del GFR è l'età.

Una review dei fattori di rischio o markers collegati alla malattia cronica dei reni, o *Chronic Kidney Disease*, nella popolazione generale è riportata su Lancet (El Nahas A. M., Bello A. K. (2005)). In essa i fattori sono classificati in iniziali e di progressione:

Initiation factor: Obesità, ipertensione, diabete, fumo, iperlipidemia e albuminuria;

Progression factor: Fattori non modificabili includono razza, età, sesso e fat-

tori genetici. Altri fattori sono ipertensione sistemica, iperlipidemia, obesità, albuminuria, fumo.
Altra variabile di interesse è l'attività fisica (Robinson-Cohen C., Katz R., Mozaffarian D. et al. (2009)). Alti livelli di attività fisica sono associati ad un basso rischio di rapido declino della funzione renale tra gli anziani.
L'albuminuria o *urine-albumin*, condizione per la quale nelle urine c'è un anormale quantità di proteine, è anche un predittore di mortalità, insieme al GFR, nella popolazione generale (Astor B. C., Hallan S. I., Miller E. R. , Yeung, E. and Coresh J. (2008)).

Assunzioni alla base dello studio
Alla base dell'analisi che presenteremo in seguito c'è l'ipotesi che il processo dei dati mancanti generato dal decesso sia o possa essere informativo. In tal caso ci si attende che le persone che muoiono prima abbiano un livello di GFR più basso, quindi abbiano una maggiore probabilità di avere una compromissione renale. Dai dati non si può conoscere qual'è il tipo di meccanismo di dati mancanti, se a caso o informativo, per cui ci si deve avvalere di conoscenze esterne. La letteratura scientifica ha messo in relazione la funzionalità renale con la mortalità (generale o cardiovascolare), mostrando che una compromissione della funzione renale può avere effetti negativi sulla sopravvivenza.
Nella figura 4.1 sono riportate le curve di sopravvivenza empiriche di Kaplan-Meier per quartili di eGFR, dove il GFR è stimato in base alla cistatina su dati della coorte ULSAM, con baseline a 70 anni. La sopravvivenza è peggiore per coloro che hanno bassi livelli di GFR, in particolare un GFR stimato ricadente nel primo quartile (11-52 ml/min/1.73 m), avvalorando l'ipotesi che gli uomini che sono usciti prima dallo studio per decesso potessero avere un

livello di GFR più basso, portando a stime del modello classico longitudinale, lineare ad effetti casuali, distorte perchè basate su un campione 'selezionato' di uomini con uno stato di salute migliore.

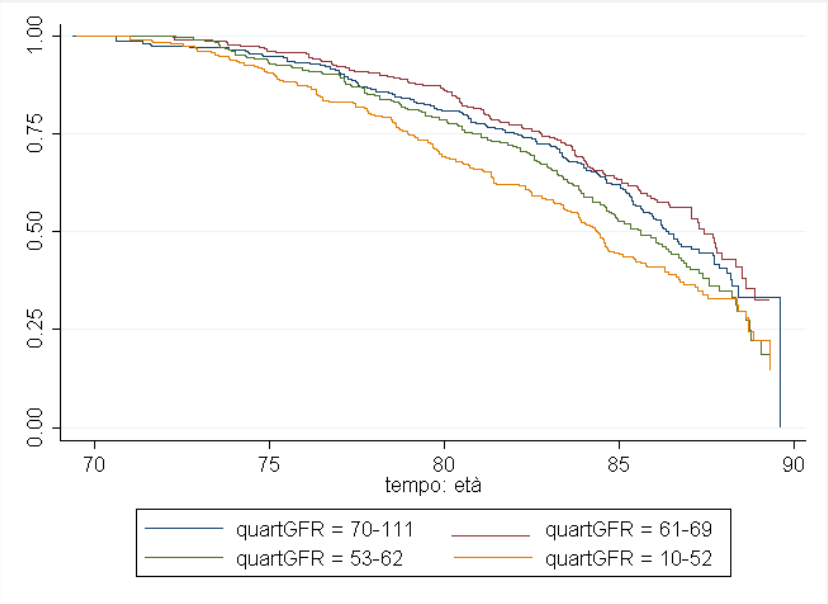

Figura 4.1: Andamento dei dati di sopravvinenza della coorte ULSAM: stime di Kaplain-Meier per quartili di GFR stimato in base alla cistatina

Missing data informativo nei dati longitudinali

Per lo studio abbiamo fissato il baseline a 70 anni. Parte degli uomini che parteciparono allo studio a 70 anni non hanno potuto parteciparvi nelle rilevazioni successive, a 77 o 82 anni, perché deceduti. Altri non si sono presentati alle visite per motivi a noi non noti, seppur invitati, e qualcuno si è trasferito fuori dall'Uppsala Country.

Il biomarker eGFR (valore stimato in base alla cistatina), collegato allo stato di salute del paziente, presenta dunque dati mancanti dovuti principalmente alla mortalità, oltre che ad altre cause. I valori del eGFR per età alla visita

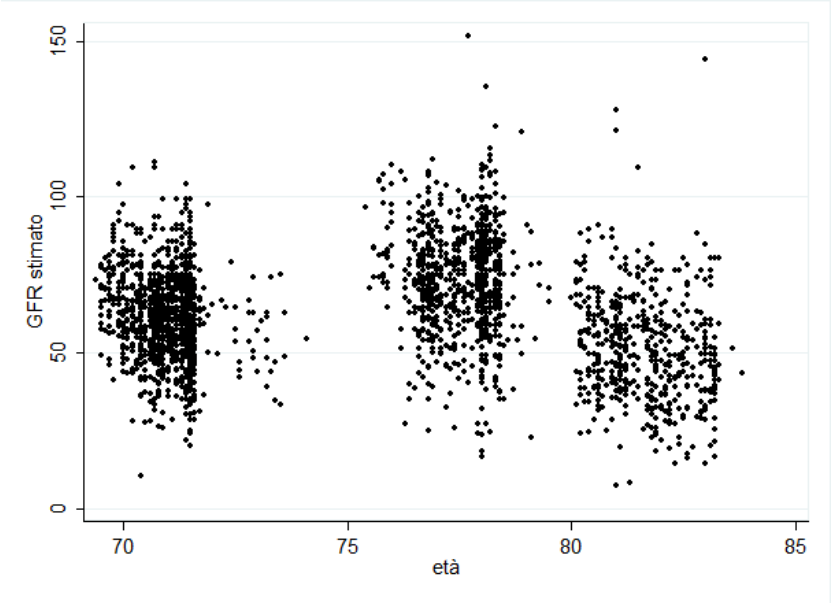

Figura 4.2: GFR stimato in base alla cistatina per età alla visita nella coorte di uomini ULSAM

sono rappresentati nella figura 4.2.

Si assume, per l'analisi statistica, che il missing per uscita dallo studio dovuto al decesso sia informativo, mentre il missing per altri motivi sia missing at random (MAR).
Disponiamo di tre misure ripetute del biomarker eGFR, che consentono di applicare un'analisi longitudinale con outcome continuo (modello lineare ad effetti misti). Per tutti gli uomini partecipanti allo studio si hanno inoltre informazioni sui tempi del decesso, che consentono di applicare modelli di sopravvivenza (modello di Weibull, modello di Cox).

4.3.3 Problemi del dataset

Lo studio presenta alcuni problemi.

In primo luogo il basso numero di misure ripetute del eGFR, che sono solo tre. Ciò non consente di avere una chiara identificazione dell'andamento, che richiederebbe la misura di più valori nel tempo. Inoltre si genera un gran numero di dati mancanti al passare del tempo, dovuto anche alla anzianità della coorte. L'ampiezza campionaria per il baseline, che è a 70 anni, e per le altre due rilevazioni, a circa 77 e 82 anni, è rispettivamente di 1221, 838 e 526. C'è un rapido incremento dei missing data dovuto principalmente al decesso, ma anche a non partecipazione alla visita per altre cause. Fino alla terza visita, nella quale gli uomini avevano un'età vicina agli 82 anni, sono deceduti 504 uomini (41.3 %). Alla terza visita inoltre 191 uomini (15.6 %) non si sono presentati per altri motivi.

Un' altra caratteristica di questi dati è che essi non sono perfettamente bilanciati; ciò significa che non intercorre sempre lo stesso tempo tra rilevazioni successive dei dati sui singoli individui. Nell'analisi longitudinale dunque non è identificabile una particolare struttura che descriva la correlazione tra le misure nel tempo.

Inoltre, come è possibile vedere nel grafico 4.3, l'andamento del eGFR non è lineare, come ci si attendeva, ma assume più una forma quadratica. Per verificare se ci fossero stati errori di laboratorio sono state effettuate delle rianalisi, ricavando la misura della cistatina con un altro metodo di laboratorio (*Gentian method*), su un campione di 50 unità di sangue per la seconda visita ed altre 50 unità per la terza visita. I dati campionati sono risultati simili a quelli originari (*Dade method*), per cui questi ultimi sono stati utilizzati per l'analisi statistica.

C'è da tener conto inoltre che il valore del GFR è stimato in base alla misura

della cistatina, per cui si potrebbero verificare errori di misura con impatto sulle stime, in particolare per i valori del GFR più elevati (Stevens et al. (2009)).

4.3.4 I modelli

L'outcome scelto per l'analisi longitudinale è l'eGFR, preso come variabile continua gaussiana, mentre le covariate sono variabili collegate allo stile di vita (attività fisica, obesità, fumo, ipertensione). Altra covariata considerata nell'analisi è l'albuminuria. Non abbiamo inserito nel modello invece le variabili che rappresentano i lipidi (colesterolo HDL, LDL e trigliceridi) per la loro interazione con attività fisica. Il modello di sopravvivenza conterrà essenzialmente le stesse variabili contenute nel modello longitudinale.

Outcome del modello longitudinale

Come biomarker della funzione renale abbiamo usato il GFR stimato. È consuetudine nella clinica effettuare una stima di tale valore applicando formule che esplicitano la relazione del vero GFR con alcune proteine, la cistatina o la creatinina. Il GFR stimato è stato calcolato a partire dalla *serum cystatin C*, utilizzando la formula di Larsson:

$$eGFR = 77.239 * CystatinC^{(-1.2623)}$$

Nello studio si è scelto di usare la cistatina poiché nel dataset si ha il dato per le tre misure ripetute, a 70, 77 e 82 anni, a differenza della creatinina, per la quale non si dispone del dato dopo i 70 anni. Inoltre alcuni studi supportano l'utilizzo della cistatina rispetto alla creatinina per la stima del GFR nel caso di coorti di anziani (Shlipak M. G., Katz R. Kestenbaum B. et al. (2009)).

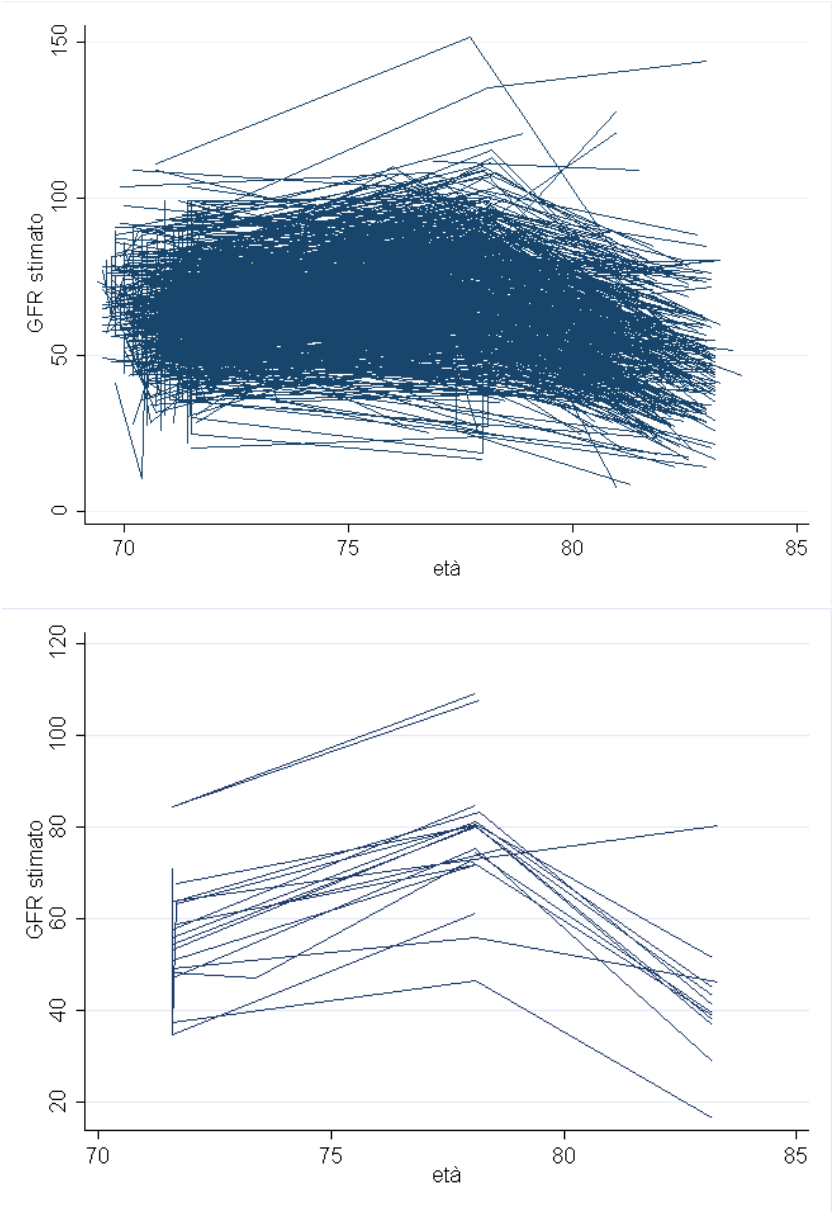

Figura 4.3: Andamento dei dati longitudinali: GFR stimato dalla cistatina per tutti gli individui della coorte ULSAM inclusi nello studio (fig. in alto) e per una selezione di essi (fig. in basso)

In ambito clinico si assume l'esistenza nell'uomo di una malattia cronica dei reni (*Cronic Kidney Disease*) se si ha un GFR più basso di 60 ml/min/1.73 m.

Dataset dell'analisi longitudinale

L'analisi è applicata agli uomini inclusi nella coorte ULSAM. Il baseline è a 70 anni. Dei 1221 uomini esaminati di circa 70 anni (tra il 1991 e il 1995), 28 sono stati esclusi perchè mancanti della misura della cistatina. I soggetti al baseline, per i quali si dispone della misura del eGFR, sono dunque 1193. Le misure sono state replicate per gli uomini intorno ai 77 anni (rilevate tra il 1998 ed il 2001) e agli 82 anni (rilevate tra il 2003 e il 2005). Dall'analisi sono stati eliminati 14 uomini non presenti al baseline, ma presenti alla rilevazione successiva. Dal dataset finale sono stati inoltre eliminati i record nei quali si aveva almeno un valore mancante nelle covariate, per cui l'analisi finale è stata effettuata su 1022 uomini. Ricordiamo che i dati non sono perfettamente bilanciati, con valori mancanti dell'eGFR alle rilevazioni successive rispetto al baseline.

Per l'analisi longitudinale classica applichiamo un modello lineare ad effetti misti che implicitamente assume che il processo di missingness è casuale (MAR), cioè non collegato a valori dell'outcome non misurati, mentre nel sottomodello longitudinale dell'analisi congiunta ipotizziamo un processo di dati mancanti, dovuto al decesso, informativo.

I fattori di rischio di cui si vuole valutare l'associazione con l'eGFR sono:

- età in occasione delle tre rilevazioni, considerata come time varying,
- l'attività fisica,
- l'obesità, espressa dalla misura della circonferenza della vita,
- l'ipertensione, espressa dall'essere sottoposto o meno a trattamento farmacologico per l'ipertensione,

- il fumo,
- l'albumina nelle urine.

Il principale scopo dell'analisi è di studiare l'associazione nel tempo dei sopraelencati fattori di rischio collegati allo stile di vita, considerati al baseline, e il declino dell'eGFR.

Includiamo, dunque, nel modello le seguenti covariate:

Occ77c, età prevista in occasione delle visite (70, 77 o 82 anni), variabile centrata su 77,
Att-fisica, attività fisica (1 sedentaria, 2 moderata, 3 regolare, 4 atletica),
Circ-Vita, circonferenza della vita (cm),
Ipertensione, variabile dummy per aver ricevuto un trattamento farmacologico per l'ipertensione (1=si, 0=no),
Fumo, variabile dummy per l'essere fumatore (1=si, 0=no),
Log-album, logaritmo u-albumin excr. rate ($\mu g/min$).

Presentiamo di seguito i risultati delle analisi separate, longitudinale e di sopravvivenza, secondo l'impostazione classica frequentista e le analisi separate e congiunte (joint model) secondo l'impostazione bayesiana.

4.3.5 Risultati usando modelli classici

Le analisi separate, longitudinale e di sopravvivenza, sono state condotte utilizzando STATA.
Sia y_{ij} il livello della j-sima misura di eGFR sull'i-simo uomo dello studio longitudinale, con $j = 1, ..., m_i$ e $i = 1, ..., n$. Il modello lineare ad effetti

casuali è

$$y_{ij} = \beta_{11} + \beta_{12}s_{ij} + \beta_{13}s_{ij}^2 + \beta_{14}Att - fisica_i + \beta_{15}Circ - vita_i + \beta_{16}Ipertensione_i + \beta_{17}Fumo_i + \beta_{18}l - album_i + \beta_{19}l - album_i \times s_{ij} + W_{1i}(s_{ij}) + \epsilon_{ij} \quad (4.7)$$

Specifichiamo l'outcome $y_{ij} = eGFR_{ij}$ come la misura dell'eGFR, $s_{ij} = Occ77c_{ij}$ che rappresenta approssimativamente l'età in occasione delle tre visite (valori fissati a 70, 77 e 82, centrati su 77), e $s_{ij}^2 = occ77c_{ij}^2$, per tener conto dell'andamento quadratico dei dati. In tal caso $W_{1i}(s_{ij}) = U_{1i} + U_{2i}s_{ij}$ rappresenta l'effetto casuale, dove $(U_{1i}, U_{2i})^T \sim N(\mathbf{0}, \mathbf{\Sigma})$. In questo modello si assume che varino nel tempo sia l'intercetta che la pendenza, dunque differenti soggetti hanno differente eGFR al baseline, e differente declino dell'eGFR nel tempo.
Specificato ciò, il modello può essere riscritto nel seguente modo

$$eGFR_{ij} = \beta_{11} + \beta_{12}occ77c_{ij} + \beta_{13}occ77c_{ij}^2 + \beta_{14}Att - fisica_i + \beta_{15}Circ - vita_i + \beta_{16}Ipertensione_i + \beta_{17}Fumo_i + \beta_{18}l - album_i + \beta_{19}l - album_i \times occ77c_{ij} + U_{1i} + U_{2i}occ77c_{ij} + \epsilon_{ij}$$

I risultati dell'analisi longitudinale secondo l'inferenza classica sono mostrati nella tabella 4.1. Si è utilizzato il comando *xtmixed* in STATA, specificando il metodo di stima *ml* di massima verosimiglianza, e una matrice di varianza e covarianza Σ non strutturata *cov(un)*.

Il coefficiente di regressione medio stimato per la variabile *occ77c* (-1.749), CI (-1.984, -1.513) è significativo, per cui passando da una visita (avendo centrato su 77) a quella successiva vi è un andamento decrescente, per la componente lineare, del GFR stimato. Il valore per il termine quadratico

Tabella 4.1: Analisi classica: modello ad effetti misti longitudinale

Parametri	Stima puntuale	95% CI
Intercetta (β_{11})	83.708	$(74.272, 93.144)$
occ77c (β_{12})	-1.749	$(-1.984, -1.513)$
occ77c2 (β_{13})	-0.543	$(-0.573, -0.513)$
Att-fisica (β_{14})	1.201	$(-0.058, 2.459)$
Circ-vita (β_{15})	-0.077	$(-0.167, 0.012)$
Ipertensione (β_{16})	-4.957	$(-6.701, -3.213)$
Fumo (β_{17})	-3.651	$(-5.678, -1.624)$
log-album (β_{18})	-2.129	$(-2.996, -1.261)$
log-album*occ77c (β_{19})	-0.240	$(-0.340, -0.139)$

occ77c2 (-0.543), CI(-0.573 -0.513) è anch'esso significativo. Predittori del declino del GFR nel tempo, significativi al 95%, sono l'ipertensione (-4.957), il fumo (-3.651) e un alto livello di albumina nelle urine *l-album* (-2.129). Tale effetto cresce con l'avanzare dell'età *l-albumin*occ77c*, dato che l'interazione è significativa. L'attività fisica e la circonferenza della vita non sono significative al 95%, ma lo sono al 90%. L'attività fisica *Att-fisica* (1.201) CI(-0.058, 2.459) è l'unica variabile nel modello che ha un effetto positivo sul GFR.

Per l'analisi di sopravvivenza consideriamo un modello parametrico di Weibull. Tutte le covariate sono considerate al baseline, per cui si può scrivere l'equazione per il log-relative hazard, senza effetti casuali, come

$$log(\mu_i) = \beta_{21} + \beta_{22} Att - fisica_i + \beta_{23} Circ - vita_i + \beta_{24} Ipertensione_i + \beta_{25} Fumo_i + \beta_{26} l - album_i + \beta_{27} eta' - visita_i \quad (4.8)$$

dove *eta'-visita* è l'età alla visita al baseline.

I coefficienti stimati sono riportati nella tabella 4.2. L'attività fisica è signi-

Tabella 4.2: Analisi classica: modello di sopravvivenza di Weibull

Parametri	Hazard Ratio	coefficienti	95% CI
Intercetta (β_{21})	-	-62.737	$(-73.931, -51.542)$
Att-fisica (β_{22})	0.777	-0.252	$(-0.382, -0.123)$
Circ-vita (β_{23})	1.007	0.007	$(-0.003, 0.016)$
Ipertensione (β_{24})	1.436	0.362	$(0.191, 0.533)$
Fumo (β_{25})	1.454	0.374	$(0.185, 0.564)$
log-album (β_{26})	1.146	0.136	$(0.070, 0.202)$
età-visita (β_{27})	0.930	-0.072	$(-0.215, 0.070)$

ficativamente protettiva, per cui soggetti che fanno attività fisica hanno un tasso di sopravvivenza migliore rispetto a quelli che adottano uno stile di vita sedentario. Persone ipertese, che fumano e con un alto livello di albumina nelle urine invece hanno un tasso di sopravvivenza peggiore rispetto rispettivamente ai non ipertesi, non fumatori e con un basso livello di albumina. Le altre due variabili inserite nel modello, circonferenza della vita e età alla visita non sono significative. Il modello di Cox, qui non riportato, dà risultati veramente molto simili.

4.3.6 Risultati usando Joint Models

Scelta delle a priori

Il modello congiunto proposto di seguito segue l'impostazione di Guo e Carlin (2004), i quali considerano una versione pienamente bayesiana implementata tramite il metodo MCMC, utilizzando il software Winbugs.

Nel modello bayesiano vengono adottate a priori non-informative o vaghe,

in modo da permettere il confronto con i risultati dell'analisi classica. Le a priori considerate sono 'proprie', ma con valori degli iperparametri scelti in modo tale da avere il minimo impatto sui dati. In particolare per il sottomodello longitudinale sono considerate normali multivariate per il vettore degli effetti principali $\beta_1 = (\beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{15}, \beta_{16}, \beta_{17}, \beta_{18}, \beta_{19})^T$ e inverse gamma per la varianza dell'errore σ^2_ϵ, entrambi con una bassa precisione e di conseguenza alta varianza. Si è ragionato in maniera simile per definire le a priori del sottomodello di sopravvivenza. Sono scelte normali vaghe per i coefficienti $\beta_2 = (\beta_{21}, \beta_{22}, \beta_{23}, \beta_{24}, \beta_{25}, \beta_{26}, \beta_{27})$ e inverse gamma per la variabilità dell'eventuale termine di frailty σ^2_3. Per i parametri degli effetti casuali comuni ad entrambi i modelli, ovvero per la variabilità, si è assunto un inverse Wishart (essenzialmente una gamma multivariata; vedi 1.19). Per i parametri di associazione γ_1 e γ_2 sono scelte a priori normali anch'esse abbastanza vaghe. Per selezionare il modello si può utilizzare il DIC (Deviance Information Criterion; Spiengelhalter et al., 2002), una generalizzazione del AIC (Akaike Information Criterion), scegliendo il modello con più piccolo DIC totale.

Il modello congiunto che abbiamo considerato nell'applicazione introduce un'intercetta e pendenza casuale nel sottomodello longitudinale. Sotto il modello congiunto la sopravvivenza degli uomini appartenenti alla coorte ULSAM è collegata a due caratteristiche che giudano il pattern longitudinale, ovvero il livello iniziale del GFR e il relativo tasso di decremento. Questo è ragionevole in quanto un alto livello nel GFR rappresenta un migliore stato di salute; ci dovremmo attendere che gli individui con un GFR basso o con un più rapido declino abbiano una peggiore sopravvivenza.

Risultati

I risultati del joint model sono basati su tre catene parallele MCMC di 60000 iterazioni ognuna, che seguono 10000 iterazioni 'burn-in'. Vogliamo confrontare i risultati ottenuti sotto i modelli separati classici, con quelli dei modelli separati bayesiani e del joint model bayesiano. Per quanto riguarda l'analisi separata bayesiana, il modello longitudinale assume la forma (4.7), mentre quello di sopravvivenza la (4.8). Nel joint model invece il sottomodello longitudinale assume la stessa forma (4.7), mentre in quello di sopravvivenza si aggiunge la componente W_{2i}. Il modello diventa

$$log(\mu_i) = \beta_{21} + \beta_{22}Att - fisica_i + \beta_{23}Circ - vita_i + \beta_{24}Ipertensione_i + \beta_{25}Fumo_i + \beta_{26}l - album_i + \beta_{27}eta' - visita_i + W_{2i}(t) \quad (4.9)$$

dove $W_{2i}(t) = W_{2i} = \gamma_1 U_{1i} + \gamma_2 U_{2i}$.

Con tale modello congiunto si può testare se il differente livello del GFR iniziale γ_1 e il relativo tasso di decremento γ_2 hanno un effetto sulla sopravvivenza degli uomini della coorte ULSAM.

Tutte le variabili non dicotomiche sono state centrate intorno alla rispettiva media, per tener conto della correlazione tra i coefficienti. Le stime a posteriori dei coefficienti di regressione $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$, e dei loro intervalli di credibilità al 95%, per le analisi separate bayesiane sono riportate nella tabella (4.3), mentre le stime per il joint model sono riportate nella tabella (4.4).

I risultati delle analisi bayesiane, separate e congiunte, sono abbastanza simili.

Una differenza tra le stime dei coefficienti del modello classico e di quello bayesiano longitudinale, sia separato che congiunto, è nell'intercetta e nella

Tabella 4.3: Analisi bayesiana separata: modello ad effetti misti longitudinale e modello di sopravvivenza di Weibull

Parametri	Media a posteriori - mod. longitudinale	95% CI
Intercetta (β_{11})	74.83	$(73.48, 76.19)$
occ77c (β_{12})	-2.232	$(-2.39, -2.074)$
occ77c2 (β_{13})	-0.542	$(-0.570, -0.514)$
Att-fisica (β_{14})	1.297	$(0.035, 2.563)$
Circ-vita (β_{15})	-0.084	$(-0.174, 0.006)$
Ipertensione (β_{16})	-4.717	$(-6.468, -2.978)$
fumo (β_{17})	-3.385	$(-5.421 - 1.346)$
log-album (β_{18})	-2.149	$(-3.025, -1.269)$
log-album*occ77c (β_{19})	-0.243	$(-0.354, -0.131)$
sigma$_{11}$(Σ_{11})	151.7	$(132.4, 172.9)$
sigma$_{22}$(Σ_{22})	1.253	$(1.033, 1.502)$
cor (ρ)	0.426	$(0.328, 0.516)$
sigma$^2_\epsilon$	81.21	$(72.88, 90.33)$
Parametri	Media a posteriori - mod. Weibull	95% CI
Intercetta (β_{21})	-6.449	$(-6.918, -6.004)$
Att-fisica (β_{22})	-0.244	$(-0.373, -0.115)$
Circ-vita (β_{23})	0.007	$(-0.003, 0.016)$
Ipertensione (β_{24})	0.341	$(0.170, 0.513)$
Fumo (β_{25})	0.359	$(0.168, 0.548)$
log-album (β_{26})	0.133	$(0.067, 0.198)$
età-visita (β_{27})	0.127	$(-0.011, 0.265)$

Tabella 4.4: Analisi bayesiana congiunta

Parametri	Media a posteriori - joint model	95% CI
Intercetta (β_{11})	74.28	$(72.9, 75.66)$
occ77c(β_{12})	-2.357	$(-2.531, -2.185)$
occ77c2(β_{13})	-0.548	$(-0.577, -0.520)$
Att-fisica(β_{14})	1.371	$(0.089, 2.653)$
Circ-vita (β_{15})	-0.085	$(-0.176, 0.007)$
Ipertensione (β_{16})	-4.762	$(-6.520, -2.997)$
Fumo (β_{17})	-3.492	$(-5.529, -1.455)$
log-album (β_{18})	-2.319	$(-3.200, -1.441)$
log-album*occ77c (β_{19})	-0.268	$(-0.381, -0.158)$
sigma$_{11}$(Σ_{11})	156.2	$(136.2, 178.2)$
sigma$_{22}$(Σ_{22})	1.259	$(1.038, 1.512)$
cor (ρ)	0.442	$(0.344, 0.532)$
sigma$^2_\epsilon$	80.91	$(72.64, 90.01)$
Intercetta (β_{21})	-6.608	$(-7.118, -6.125)$
Att-fisica (β_{22})	-0.253	$(-0.385, -0.119)$
Circ-vita (β_{23})	0.005	$(-0.004, 0.015)$
Ipertensione (β_{24})	0.367	$(0.190, 0.544)$
Fumo (β_{25})	0.394	$(0.192, 0.590)$
log-album (β_{26})	0.140	$(0.071, 0.208)$
età-visita (β_{27})	0.088	$(-0.055, 0.230)$
γ_1	-0.020	$(-0.029, -0.011)$
γ_2	-0.133	$(-0.320, 0.047)$

variabile relativa all'età all'occasione delle visite. In particolare, rispetto all'analisi separata classica, il modello bayesiano congiunto fornisce una stima della media a posteriori dell'intercetta, e quindi un GFR medio al baseline, più bassa (β_{11}, 74.28 vs. 83.71), e una stima della media a posteriori della variabile occasione alla visita, e quindi una componente lineare del declino del GFR all'avanzare dell'età, più accentuata (occ77c = β_{12} (-2.36), IC(-2.53, -2.19) vs. (-1.75) IC(-1.98, 1.51)). Nell'analisi bayesiana l'attività fisica β_{14} è significativa al livello del 5% (nel senso bayesiano; il 95% degli intervalli di credibilità contiene lo 0), mentre non lo era nel modello classico. L'effetto positivo dell'attività fisica è più rilevante nel joint model rispetto al modello separato bayesiano (1.37 IC(0.09, 2.65) vs. 1.30 IC(0.04 2.56)).
Sia nel modello congiunto che separato bayesiano, le altre variabili che hanno un effetto negativo sul GFR sono l' ipertensione (-4.72) IC(-6.47, -2.98), il fumo (-3.39) IC(-5.42, -1.35) e la log-albumina nelle urine (-2.15) IC(-3.03, -1.27).
Anche nei modelli di sopravvivenza le stime dei parametri sono simili, con un'effetto negativo sulla sopravvivenza dell'ipertensione, fumo e albumina, ed effetto positivo dell'attività fisica.
La stima a posteriori del parametro di associazione γ_1 nell'analisi congiunta è negativa e 'significativamente' diversa da zero, fornendo evidenza di associazione tra i due sottomodelli, indicando che i valori iniziali del GFR sono negativamente associati con l'hazard di morte. Il parametro di associazione γ_2, che indica l'associazione con la pendenza, invece non è 'significativo'.

Riportiamo nelle pagine successive dei grafici con diagnostiche dei modelli congiunti e separati bayesiani, in particolare:
- nella figura (4.4) e (4.5) riportiamo le statistiche di Gelman Rubin per ver-

ificare la convergenza per i parametri stimati a partire dai tre differenti set di valori iniziali, rispettivamente per il modello bayesiano separato longitudinale e di sopravvivenza.
- nelle figure (4.6), (4.7) e (4.8) riportiamo le tracce e le densità a posteriori stimate.
- nelle figure (4.9) e (4.10) riportiamo le statistiche di Gelman e Rubin per investigare la convergenza dei parametri del joint model.
Infine nella figura (4.11) sono rappresentati i valori osservati e i valori previsti secondo il joint model per alcuni individui della coorte ULSAM.

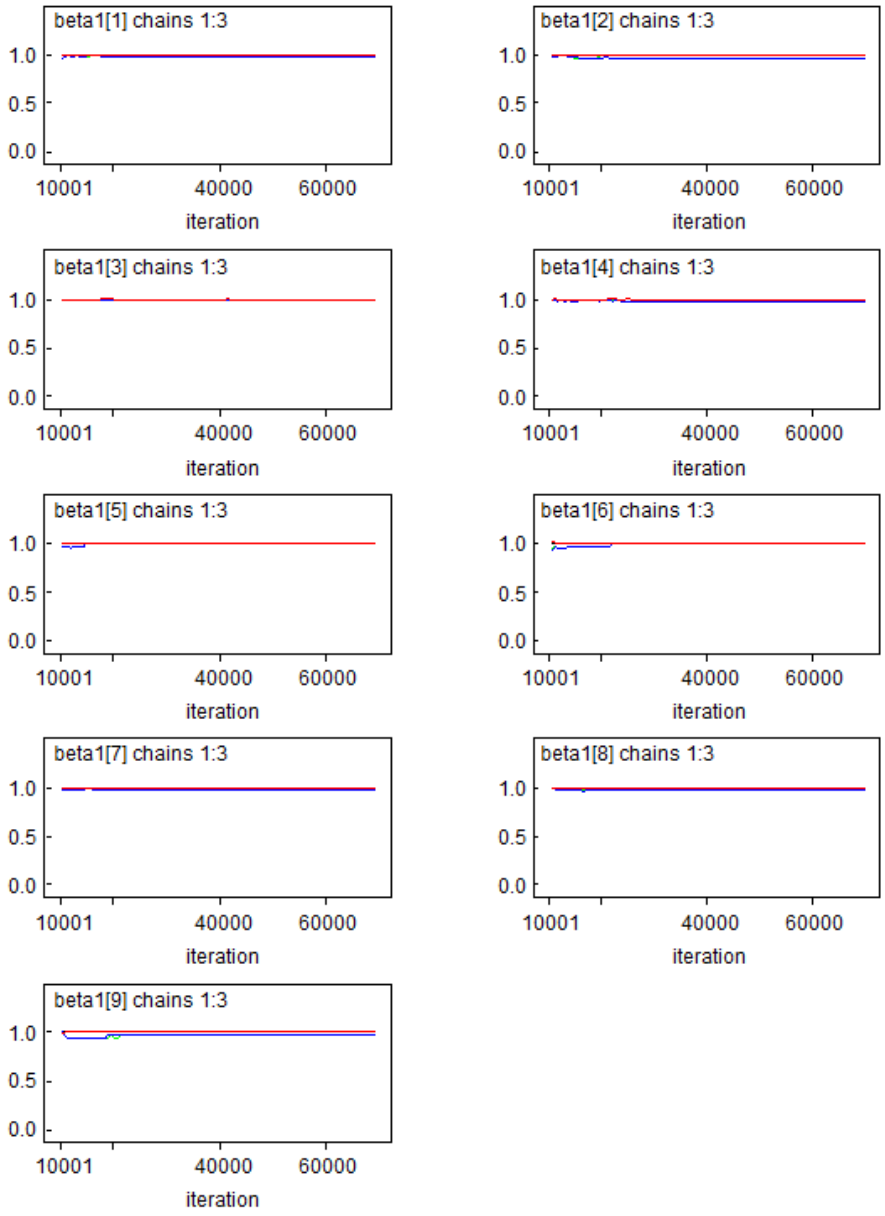

Figura 4.4: Convergenza di Gelman Rubin: modello longitudinale ad effetti casuali bayesiano separato

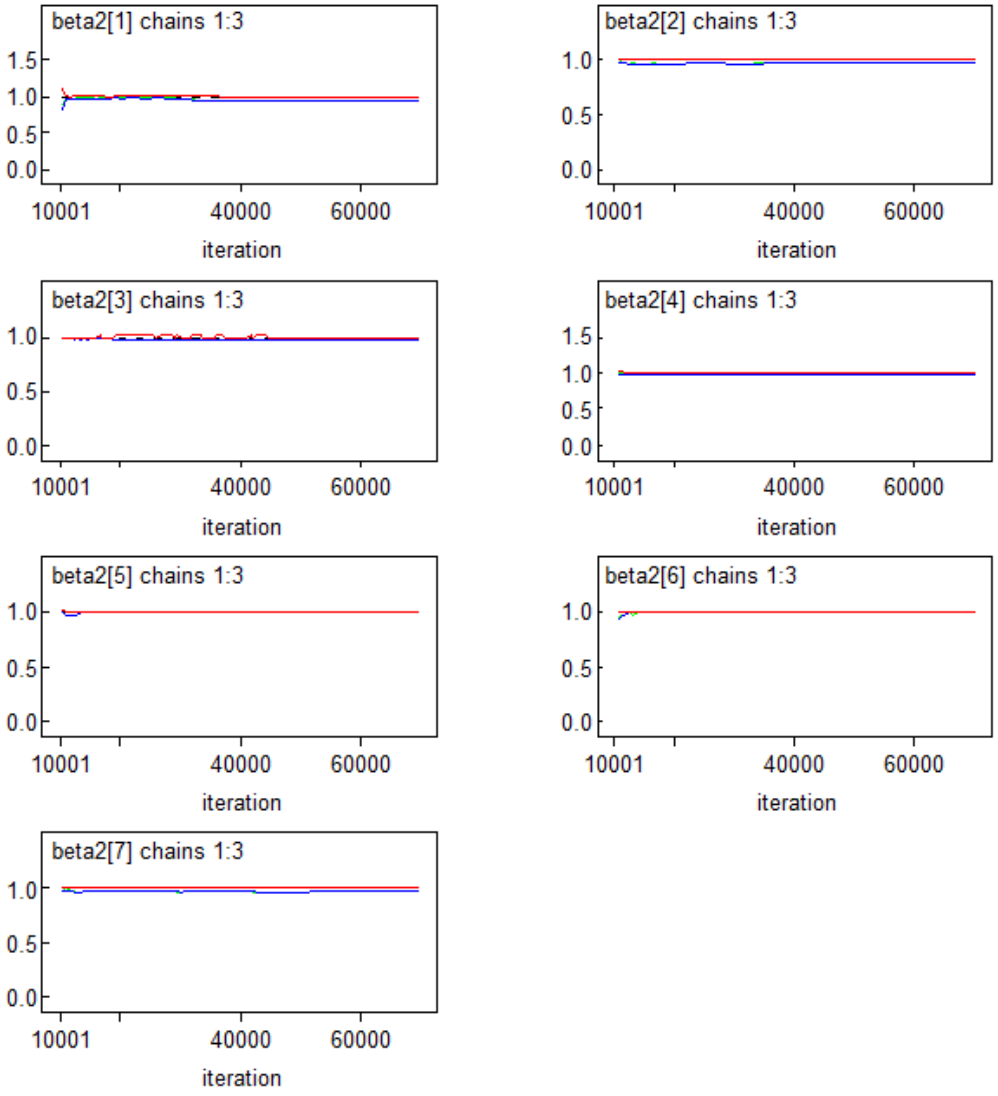

Figura 4.5: Convergenza di Gelman Rubin: modello di sopravvivenza di Weibull bayesiano separato

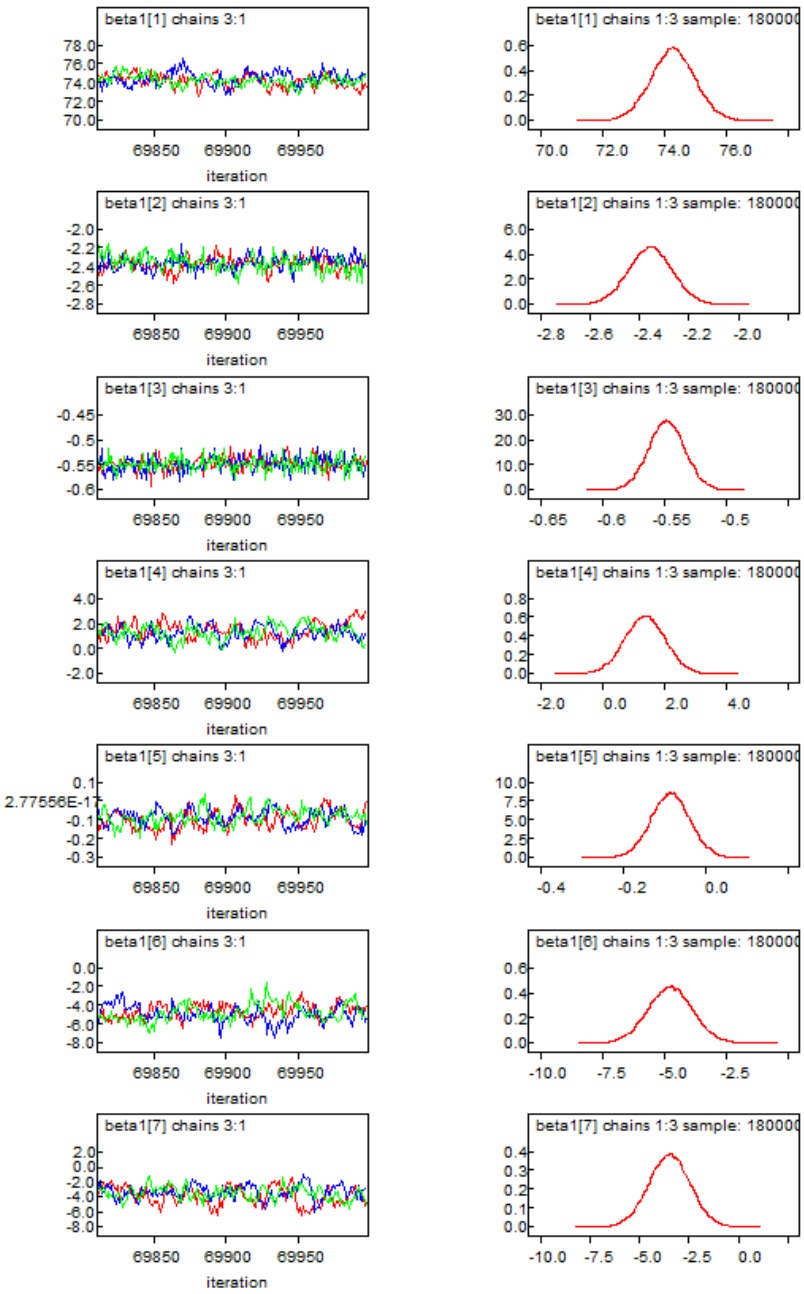

Figura 4.6: Traccia e stima della densità di Kernel per il joint model (a)

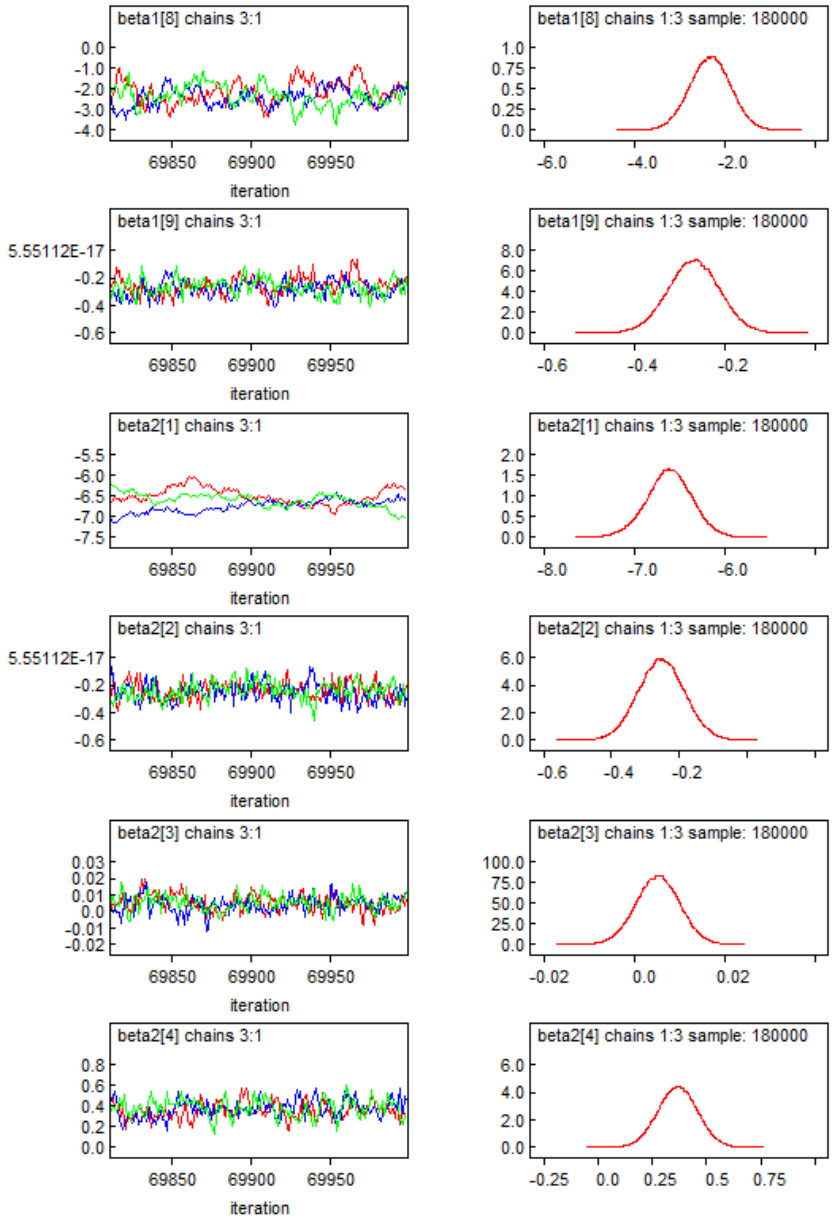

Figura 4.7: Traccia e stima della densità di Kernel per il joint model (b)

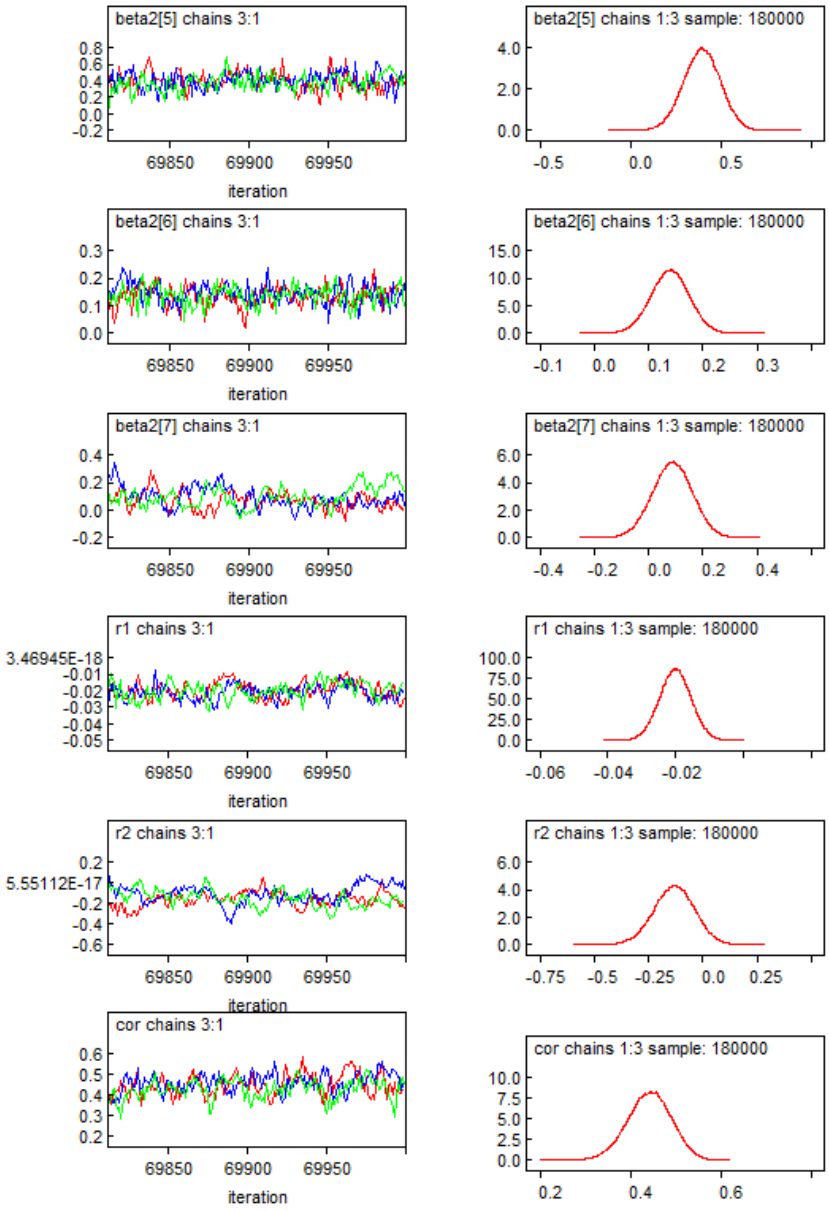

Figura 4.8: Traccia e stima della densità di Kernel per il joint model (c)

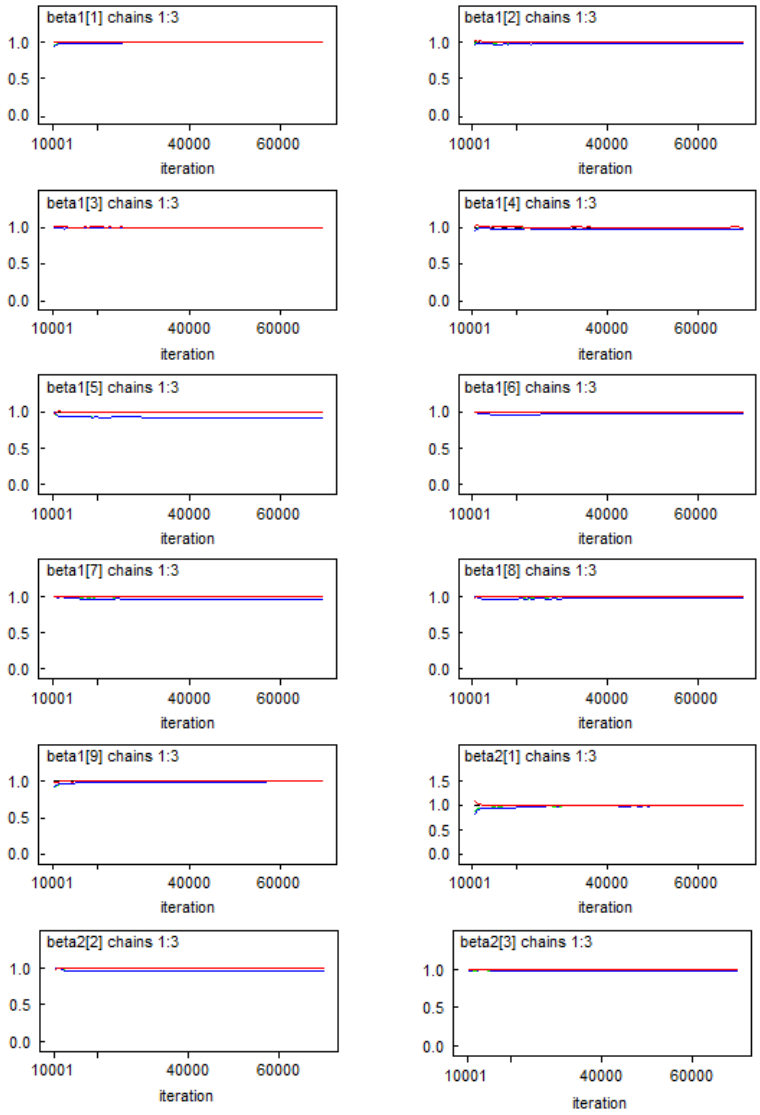

Figura 4.9: Convergenza di Gelman Rubin dei parametri del joint model (a)

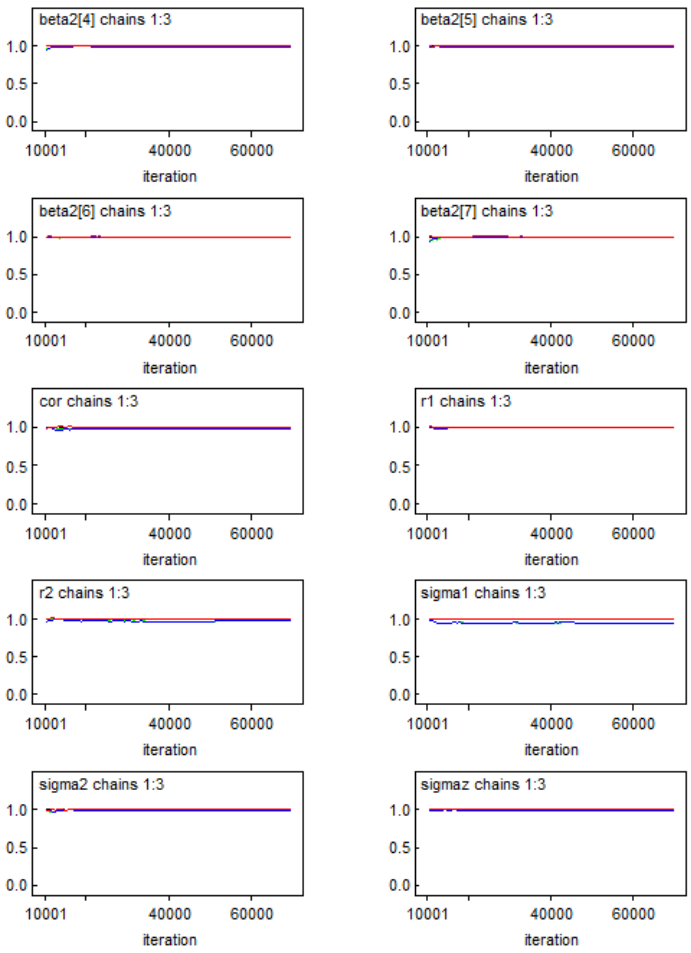

Figura 4.10: Convergenza di Gelman Rubin dei parametri del joint model (b)

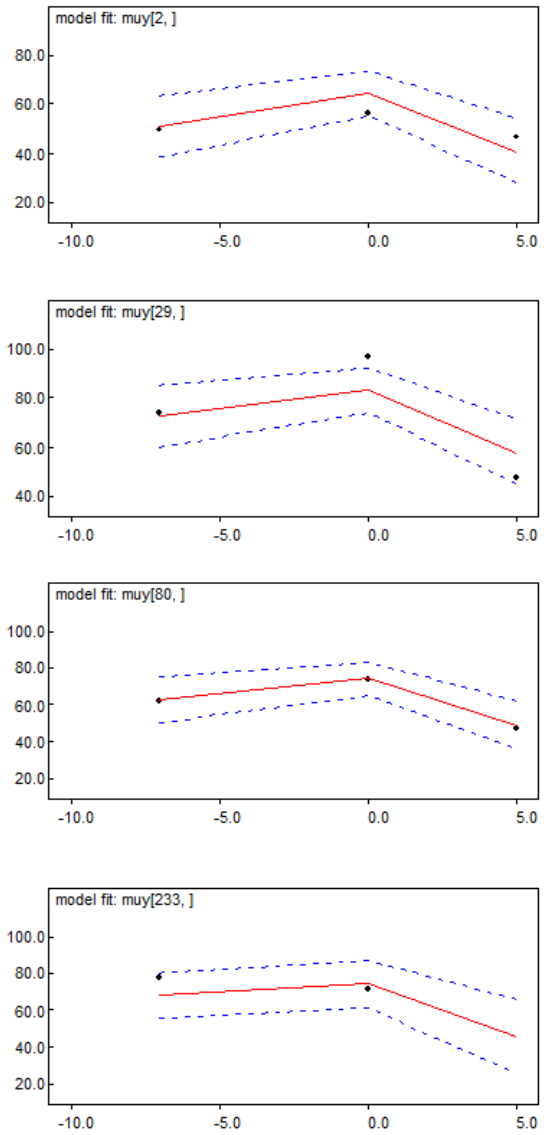

Figura 4.11: Valori osservati e previsti secondo il joint model per alcuni individui della coorte ULSAM

Conclusioni

La letteratura scientifica sui joint models si è sviluppata molto negli ultimi anni nell'ambito della ricerca biomedica, in particolar modo per l'analisi di dati provenienti da trial clinici, per verificare l'effetto di diverse terapie, o, anche se in maniera minore, per l'analisi di dati su studi osservazionali longitudinali; per entrambi questi tipi di studi si hanno misure ripetute su un insieme di variabili e problemi di osservazioni censurate, che possono portare a casi di missing data o dropout informativo.

Le metodologie sviluppate in letteratura, di interesse in questa tesi, sono basate sull'integrazione dei dati longitudinali e di sopravvivenza, non più analizzati in maniera separata, per tener conto della possibile dipendenza tra specifici biomarker e dati sui tempi all'evento, es. malattia o morte.

Vista la complessità computazionale di tali metodi, essi non sono ancora diventati routinari nella ricerca biomedica, seppure gli sviluppi metodologici siano stati svariati. In questa tesi abbiamo applicato i joint models secondo un'impostazione bayesiana proposta da Guo e Carlin (2004), utilizzando il software Winbugs.

Abbiamo applicato i joint models ad una coorte di anziani dello studio osservazionale longitudinale ULSAM, per studiarne la funzianalità renale in relazione ad alcune variabili collegate essenzialmente allo stile di vita.

Alcune caratteristiche dei dati, ovvero il basso numero di misure ripetute del

biomarker GFR, l'alta percentuale di dati mancanti e l'andamento non lineare del GFR all'avanzare dell'età, rendono difficile una chiara identificazione del declino del GFR e quindi della funzionalità renale nel tempo.
Dall'applicazione dei joint models è risultata una stima del coefficiente di associazione tra il modello longitudinale e di sopravvivenza indotta dalla intercetta casuale negativa, per cui ad bassi valori del GFR al baseline è associata una maggiore mortalità. Ciò rispecchia quanto già noto dalla letteratura secondo cui il GFR è un predittore della mortalità, generale e cardiovascolare. L'approccio tramite joint modeling utilizza i dati in maniera efficiente, riducendo la distorsione dovuta al missing data informativo, poiché produce stime più accurate della forza della associazione tra l'outcome del modello longitudinale e il rischio di malattia o morte. Oltretutto tali modelli forniscono stime valide anche nel caso di missing a caso.

Struttura della tesi

Nel primo capitolo abbiamo presentato le modellizzazioni classiche utilizzare generalmente per le analisi di dati longitudinali con outcome continuo e per i dati di sopravvivenza. Per l'analisi dei dati longitudinali sono stati considerati i modelli ad effetti casuali a due stadi cosi come proposti da Laird e Ware (1982). Per quanto riguarda l'analisi di sopravvivenza si è posta l'attenzione sui modelli di hazard proporzionale di Weibull. E' stata inoltre riportata l'impostazione bayesiana di tali modelli. Nel secondo capitolo, dopo aver introdotto il concetto di missing data non casuale (NMAR) o informativo nell'ambito degli studi longitudinali, e presentati i tipi di modelli sviluppati in letteratura per questo tipo di problema statistico, si è focalizzata l'attenzione sugli shared parameter models, classe di modelli generale entro la quale rientrano i joint models per dati longitudinali e di sopravvivenza. Nel

terzo capitolo è stata presentata una review dei joint models. Particolare attenzione è stata posta sul modello di Faucett e Thomas (1996), tra i primi a proporre l'impostazione bayesiana. A partire dal loro modello sono state sviluppate svariate estensioni. E' stato poi presentato il modello di Henderson et al. (2000), approccio generale che ingloba una serie di joint models sviluppati in precedenza; esso segue l'impostazione frequentista. Nel quarto capitolo infine sono riportati i risultati dell'applicazione dei joint models, secondo l'impostazione bayesiana di Guo e Carlin (2004), ai dati della coorte ULSAM, ed effettuati i confronti tra le stime derivanti dall'analisi separata classica frequentista, l'analisi bayesiana separata e il joint model.

Bibliografia

Astor B. C., Hallan S. I., Miller E. R. , Yeung, E. and Coresh J. (2008). Glomerular Filtration Rate, albuminuria, and risk of cardiovascular and all cause mortality in the US population. *American Journal of Epidemiology*, **167**, 1226-1234.

Berzuini C. and Larizza C. (1996). A unified approach for modeling longitudinal and failure time data, with application in medical monitoring. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, **18**, N.2, 109-123.

Brown E.R. and Ibrahim J.G. (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*, **59**, 221-228.

Carlin B. and Luois (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman Hall.

Carlin B. P., Louis T. A. (2008). *Bayesian Methods for Data Analysis*, Third Edition. Chapman & Hall.

Collett D. (2003) *Modelling Survival Data in Medical Research*, Second Edi-

tion. Chapman & Hall.

Chi YY, Ibrahim JG. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, **62**, 432-445.

Cox D. (1972). Regression models and life tables. *Journal of the Royal Statistical Society*, Series B, **34**, 187-220.

Cox D. and Oakes D. (1984). *Analysis of survival data*. Chapman and Hall: London.

DeGruttola V. and Tu X. (1994). Modelling progression of cd4-lymphocyte count and its relationship to survival time. *Biometrics*, **50**, 1003-1014.

Dempster A., Laird N., and Rubin D. (1977). Maximum likelihood from incom- plete data via the em algorithm. *Journal of the Royal Statistical Society*, Series B, **39**, (1) 1-38.

Deslandes E., Chevret S. (2010). Joint modeling of multivariate longitudinal data and the dropout process in a competing risk setting: application to ICU data. *BMC Medical Research Methodology*, **10**:69.
Diggle P. (1988). An approach to the analysis of repeated measurements. *Biometrics*, **44**, 959-971.

Diggle P., kenward M.G. (1994). Informative dropout in longitudinal data analysis. *Applied Statistics*, **43**, 49-94.

Diggle P., Heagerty P., Liang K., and Zeger S. (2002). *Analysis of Longitudinal Data*. Oxford University Press.

Dobson A. and Henderson R. (2003). Diagnostics for joint longitudinal and dropout time mod- elling. *Biometrics*, **59**, 741-751.

El Nahas A. M., Bello A. K. (2005). Chronic kidney disease: the global challenge. *Lancet*, **365**, 331-340.

Faucett C. J. and Thomas D. C. (1996), Simultaneously modeling censored survival data and repeatedly measured covariates: A Gibbs sampling approach, *Statistics in Medicine*, **15**, 1663-1685.

Finkelstein D.M. and Schoenfeld D.A. (1999). Combining mortality and longitudinal measures in clinical trials. *Statistics in Medicine*, **18**, 1341-1354.

Fitzmaurice G.M., Laird N.M., and Ware J.H. (2004). *Applied Longitudinal Analysis*. Hoboken, NJ: Wiley.

Fitzmaurice G., Davidian M., Verbeke G., Molenberghs G. (2009) *Longitudinal Data Analysis*. Chapman & Hall.

Follmamm D. and Wu M. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics*, **51**, 151-168.

Gao S. (2004). A shared random effect parameter approach for longitudinal dementia data with non ignorabole missing datas. *Statistics in Medicine*, **23**,

211-219.

Gao S. and Thiebaut R. (2009). Mixed-effect Models for Truncated Longitudinal Outcomes with Nonignorable Missing Data, *Journal of Data Science*, **7**, 13-25.

Guo Xu, Carlin B.P.(2004), Separate and joint modelling of longitudinal and event time data using standard computer packages, *American Statistical Association*, **58**, 16-24

Ha I. D., Park T. and Lee Y. (2003). Joint modelling of repeated measures and survival time data. *Biometrical Journal*, **45**, 647-658.

Harville D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320-340.

Henderson R., Diggle P., and Dobson, A. (2000), Joint modeling of longitudinal measurements and event time data *Biostatistics*, **4**, 465-480.

Hogan J. W. and Laird N. (1997). Model-based approaches to analysing incom- plete longitudinal and failure time data. *Statistics in medicine*, **16**, 259-272.

Hogan J. and Laird N. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine*, **16**, 239-257.

Huang W. Z., Zeger S. L., Anthony J. C., Garrett E. (2001). Latent variable model for joint analysis of multiple repeated measures and bivariate event times. *Journal of the American Statistical Association*, **96**, 906-914.

Ibrahim J. G., Chen M. H., and Sinha D. (2004). Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine studies. *Statistica Sinica*, **14**, 847-867.

Kaplan L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-481.

Li L., Hu B., and Greene T. (2009) A Semiparametric Joint Model for Longitudinal and Survival Data with Application to Hemodialysis Study. *Biometrics*, **65**, 737-745.

Lin HQ, McCulloch CE, Mayne ST. (2002). Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. *Statistics in Medicine*, **21**, 2369-2382.

Laird N. and Ware J. (1982). Random effects in longitudinal data. *Biometrics*, **38**, 963-974.

Larsen K. (2004). Joint analysis of time-to-event and multiple binary indicators of latent classes. *Biometrics*, **60**, 85-92.

Little R. (1995). Modeling the dropout mechanism in repeated measures studies. *Journal of the American Statistical Association*, **90**, 1112-1121.

Little R. and Rubin D. (2002). *Statistical Analysis with Missing Data*, 2nd edition. New York: Wiley.

Liu M. and Ying Z. (2007). Joint analysis of longitudinal data with informative right censoring. *Biometrics*, **63**, 363-371.

Maaravi, Y., Bursztyn, M., Hammerman-Rozenberg,R. and Stessman, J. (2007) Glomerular filtration rate estimation and mortality in an elderly population. *Q J Med*, **100**, 441-449.

Marubini E. and Valsecchi M. (2004). *Analysing Survival Data from Clinical Trials and Observational Studies*. Wiley: New York.

McCullagh P. and Nelder J. (1989). *Generalized Linear Model*. Chapman & Hall Ltd.

Molenberghs G. and Kenward M. (2007). *Missing Data in Clinical Studies*. New York: Wiley.

Nitsch D., Dietrich D. F., von Eckardstein A. et al. (2006), Prevalence of renal impairment and its association with cardiovascular risk factors in a general population: results of the Swiss SAPALDIA study. *Nephrol Dial Transplant*, **21**, 935-944.

Pawitan Y. and Self S. (1993). Modelling disease marker processes in AIDS. *Journal of the American Statisical Association*, **88**, 719-726.

Prentice R. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, **69**, 331-342.

Rabe-Hesketh S., Skrondal A. (2005). *Multilevel and Longitudinal Modeling Using Stata*, Second Edition. Stata Press, Texas.

Rifkin D. E. , Katz R. et al. (2010). Albuminuria, impaired kidney function and cardiovascular outcomes or mortality in the elderly. *Nephrol Dial Transplant*, **25**, 1560-1567.

Rizopoulos D., Verbeke G., and Molenberghs G. (2008). Shared parameter models under random effects misspecification. *Biometrika*, **95**, 63-74.

Robins J.M., Rotnitzky A. and Zhao L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**, 106-121.

Robinson-Cohen C., Katz R., Mozaffarian D. et al. (2009) Physical Activity and Rapid Decline in Kidney Function Among Older Adults *Arch Intern Med.*. **169**(22), 2116-2123.

Rubin D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581-592.

Schluchter M. (1992). Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine*, **11**, 1861-1870.

Schluchter M. D., Greene T., and Beck G. J. (2001). Analysis of change in the presence of informative censoring: Application to a longitudinal clinical trial of progressive renal disease. *Statistics in Medicine*, **20**, 989-1007.

Shlipak M. G., Katz R., Kestenbaum B. et al. (2009) Rate of kidney function decline in older adult: a comparison using creatinine and cystatin C. *American Journal of Nefrology*, **30**, 171-178.

Song X., Davidian M., and Tsiatis A. (2002). A semiparametric likelihood approach to joint modelling of longitudinal and time to event data. *Biometrics*, **58**, 742-753.

Song X., Davidian M., and Tsiatis A. A. (2002). An estimator for the proportional hazards model with multiple longitudinal covariates measured with errors. **Biostatistics**, **3**, 511-528.

Spiegelhalter D., Best N., Carlin B., and van der Linde A. (2002). Bayesian measure of model complexity and fit. *Journal of the Royal Statistical Society*, Series B, **64**, 3, 583-639.

Stevens L. A. , Schmid C. H. , Greene T. et al. (2009) Factors other than glomerular filtration rate affect serum cystatin C levels. *Kidney International.* **75**, 652-660.

Tsiatis A. and Davidian M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, **14**, 809-834.

Tsiatis A., DeGruttola V., and Wulfsohn M. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, **90**, 27-37.

Verbeke G. and Molenberghs G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

Vonesh E. F., Greene T. and Schluchter M. D.(2006). Shared parameter models for the joint analysis of longitudinal data with event times. *Statistics in Medicine*, **25**, 143-163.

Wang Y. and Taylor M. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of American Statistical Association*, **455**, 895-905.

Wu L. (2002). A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to AIDS studies. *Journal of American Statistical Association*, **97**, 955-964.

Wu M. C. and Bailey K. (1988). Analysing changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medicine*, **7**, 337-346.

Wu M. and Carroll R. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, **44**, 175-188.

Wulfsohn M. S. and Tsiatis A. A. (1997), A joint model for survival and longitudinal data measured with error, *Biometrics*, **53**, 33–339.

Xu J. and Zeger S. L. (2001), Joint analysis of longitudinal data comprising repeated measures and times to events, *Applied Statistics*, **50**, 375-387.

Zeng D., Cai J. (2005). Simultaneous modelling of survival and longitudinal data with an application to repeated quality of life measures. *Lifetime Data Analysis*, **11**, 151-174.