Analisi della sopravvivenza tramite metodi bayesiani

Giulia de Innocentiis mat. 865084 Giorgia Faccanoni mat. 871869

24 Aprile 2025

Sommario

Nel presente articolo sono stati implementati modelli bayesiani per l'analisi di sopravvivenza di pazienti affette da tumore al seno. Per valutare l'effetto delle covariate sul tempo di sopravvivenza si sono stimati i modelli *Esponenziale*, *Weibull* e *Log Normale* ad effetti fissi e i modelli *Weibull* e *Log Normale* ad effetti random. Si sono ottenuti risultati migliori con il modello *Log Normale* a effetti random. Inoltre per interpolare al meglio la curva di sopravvivenza empirica è stato utilizzato il modello *Piecewise Exponential*.

1 Introduzione

L'analisi della sopravvivenza rappresenta un insieme di tecniche statistiche fondamentali per lo studio del tempo che intercorre tra il tempo iniziale, ad esempio la data di inizio di uno studio clinico, e il verificarsi di un evento di interesse, come la morte di un paziente o la ricorrenza di una malattia. Questo tipo di analisi trova largo impiego in ambito biomedico, epidemiologico e biologico. L'obiettivo di questa analisi è descrivere il comportamento del tempo all'evento nella popolazione in esame e quantificare l'impatto di fattori esplicativi sulla sopravvivenza. Tradizionalmente, questo ambito si è basato su approcci frequentisti, come il modello di *Cox* o le distribuzioni parametriche classiche, che offrono strumenti potenti ma presentano limitazioni nel trattamento della complessità strutturale dei dati. Negli ultimi decenni, l'approccio bayesiano ha guadagnato crescente interesse dato che, attraverso l'uso esplicito delle distribuzioni a priori e l'aggiornamento delle conoscenze pregresse tramite la distribuzione a posteriori, permette una modellazione più trasparente dell'incertezza e maggiore flessibilità nella specifica del modello. In quest'analisi sono stati presentati tre modelli parametrici bayesiani e un modello semiparametrico, mostrando le differenze tra di essi, interpretando l'effetto delle covariate sul tempo di accadimento dell'evento e infine confrontando le loro capacità predittive.

2 Materiali e metodi

2.1 Materiali

Il dataset considerato in questa analisi è il dataset *German Breast Cancer Data* (*gbcs*). Questo dataset è stato ricavato a partire da uno studio condotto dal *German Breast Cancer Group* in cui sono state reclutate 720 pazienti con cancro al seno con linfonodi primari positivi. Nello studio sono stati considerati due eventi di interesse: la recidiva del tumore e la morte della paziente, con relativi tempi di accadimento ed eventuali censure per ognuno dei due eventi. Nel dataset che si è utilizzato per l'analisi invece è stato considerato come evento unico la recidiva o la morte e quindi il tempo di sopravvivenza considerato è il tempo di sopravvivenza libero da recidiva. Inoltre nel dataset originale vi erano alcune osservazioni che presentavano dati mancanti che sono state rimosse; si è ottenuto quindi un dataset finale contenente 686 osservazioni e 11 variabili. Quest'ultime sono:

1. pid: codice identificativo della paziente;
2. age: età della paziente;
3. meno: variabile binaria che indica se la paziente è in menopausa;
4. size: dimensione del tumore in millimetri;
5. grade: variabile categorica che esprime il grado del tumore;
6. nodes: numero di linfonodi contenenti cellule tumorali;
7. pgr: recettori del progesterone misurate (fmol/l);

8. er: recettori degli estrogeni (fmol/l);
9. hormon: variabile binaria che indica se la paziente sta seguendo una terapia ormonale;
10. rfstime: tempo di sopravvivenza libero da recidiva espresso in giorni;
11. status: indica se la paziente è censurata o no: assume valore 0 nel caso in cui la paziente non abbia sviluppato l'evento, valore 1 quando si sviluppa l'evento.

2.2 Metodi

In quest'analisi sono stati utilizzati dei modelli bayesiani per l'analisi della sopravvivenza. L'analisi della sopravvivenza è un insieme di metodi statistici adatto allo studio dei tempi di accadimento di un evento. Per evento si intende un accadimento unico, definito in modo inequivocabile e di tipo binario. La variabile di interesse è il tempo di accadimento dell'evento, indicata da t.

Sia T la variabile casuale non negativa e continua rappresentante il tempo di sopravvivenza degli individui di una popolazione. Sia *f(t)* la funzione di densità di T e sia *F(t)* la funzione di ripartizione associata. La probabilità di sopravvivenza fino al tempo t di un individuo è data da :

$$S(t) = 1 - F(t) = P(T > t)$$

Si noti che la funzione di sopravvivenza è monotona decrescente con *S(0)*=1 e

$$S(\infty) = \lim_{t\to\infty} S(t) = 0$$

La funzione *hazard*, indicata con *h(t)*, indica il tasso istantaneo di fallimento, ovvero la probabilità di osservare l'evento al tempo t, ed è definita come:

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

La relazione tra la funzione di sopravvivenza e la funzione *hazard* è data :

$$h(t) = -\frac{d}{dt}\log(S(t)) \qquad S(t) = \exp\left(-\int_0^\infty h(u)du\right).$$

Dalle relazioni sopra indicate si può scrivere le funzione di densità dei tempi di sopravvivenza come

$$f(t) = h(t)\exp\left(-\int_0^\infty h(u)du\right).$$

Un problema che necessita particolare attenzione nell'analisi della sopravvivenza è la presenza di dati censurati e la loro implicazione nella specificazione del modello. I dati censurati si verificano quando non si osserva l'evento di interesse nel periodo di osservazione; questo può verificarsi per diversi motivi ma in questa analisi si sono considerati soltanto casi di censure non informative, che quindi non alterano la struttura dei modelli considerati.Per la tipologia di dati censurati considerati in questa analisi, ovvero dati censurati da destra, il tempo di sopravvivenza che si osserva è effettivamente il tempo di accadimento dell'evento, y_i, soltanto se esso è inferiore del tempo di censura c_i, ovvero soltanto se $y_i \leq c_i$. Altrimenti il tempo che si osserva è il tempo in cui si è avvenuta la censura del dato. Sotto questa rappresentazione è possibile riscrivere i dati come coppia di variabili (t_i, ν_i) dove:

$$t_i = \min(y_i, c_i) \qquad \nu_i = \begin{cases} 1 & \text{se } y_i \leq c_i \\ 0 & \text{se } y_i > c_i \end{cases}$$

Infine, utilizzando la precedente notazione, la funzione *likelihood* congiunta per un insieme di dati $D = (n, \mathbf{t}, \boldsymbol{\nu})$ in presenza di dati censurati è data da:

$$L(D) = \prod_{i=1}^{n} \{f(t_i)^{\nu_i} S(t_i)^{1-\nu_i}\}.$$

I modelli utilizzati sono stati:

1. Modello Esponenziale

   Il modello *Esponenziale* è uno dei principali modelli parametrici usati per l'analisi della sopravvivenza. Si suppone di avere n tempi di sopravvivenza indipendenti e identicamente distribuiti $\boldsymbol{t} = (t_1, ..., t_n)$, ognuno avente distribuzione *esponenziale* con parametro λ: $t_i \sim Exp(\lambda)$. Si definiscono gli indicatori di censura come $\nu_i = 0$ mentre $\nu_i = 1$ se il soggetto i-esimo ha sperimentato l'evento di interesse. La funzione di densità per ogni tempo di sopravvivenza è definita come:

   $$f(t_i) = \lambda exp\{-\lambda t_i\}$$

   e la funzione di sopravvivenza è data da:

   $$S(t_i|\lambda) = exp\{-\lambda t_i\}$$

   Dato $D = \{n, \mathbf{t}, \boldsymbol{\nu}\}$ il dataset contenente i dati osservati, la *likelihood* congiunta è:

   $$L(D|\lambda) = \prod_{i=1}^{n} \left(\lambda exp\{-\lambda t_i\}\right)^{\nu_i} \left(exp\{-\lambda t_i\}\right)^{1-\nu_i} = \lambda^{\sum_{i=1}^{n} \nu_i} + exp\left\{-\lambda \sum_{i=1}^{n} t_i\right\}$$

   Per costruire un modello di tipo *Esponenziale* si introducono le covariate attraverso il parametro λ, nel seguente modo: $\lambda_i = exp(\boldsymbol{x_i^t \beta})$, dove il vettore $\boldsymbol{x_i}$ è il vettore delle covariate associato all'osservazione i-esima e $\boldsymbol{\beta}$ è il vettore dei coefficienti. Una volta definita questa riparametrizzazione è stata scelta come *prior* per i coefficienti $\boldsymbol{\beta}$ la distribuzione *Normale Multivariata*: $\boldsymbol{\beta} \sim N(\boldsymbol{\mu_0}, \Sigma_0)$. In questo caso la *posterior* derivante dal modello non è esprimibile in forma chiusa quindi si utilizzano tecniche di campionamento per poter stimare il modello. La *posterior* è data da:

   $$\pi(\boldsymbol{\beta}|D) = exp\left\{(\boldsymbol{x_i^t \beta})^{\sum_{i=1}^{n} \nu_i} - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu_0})^t \Sigma_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu_0}) - (\boldsymbol{x_i^t \beta}) \sum_{i=1}^{n} t_i\right\}$$

2. Modello Weibull

   Il modello *Weibull* è un modello parametrico molto utilizzato nell'analisi della sopravvivenza. Esso assume che la funzione *hazard* varia nel tempo, ovvero il rischio può aumentare o diminuire in base al parametro di scala della distribuzione *Weibull*. Si suppone che i tempi di sopravvivenza siano indipendenti e identicamente distribuiti come una *Weibull* con parametri (α, λ), ovvero $t_i \sim W(\alpha, \lambda)$ per ogni $i = (1, ..., n)$. La funzione di densità è definita come:

   $$f(t_i) = \frac{\alpha}{\lambda} \left(\frac{t_i}{\lambda}\right)^{(\alpha-1)} exp\left\{-\left(\frac{t_i}{\lambda}\right)^{\alpha}\right\}$$

   La funzione di sopravvivenza è data da:

   $$S(t_i|\alpha, \lambda) = exp\left\{-\left(\frac{t_i}{\lambda}\right)^{\alpha}\right\}$$

   Dato $D = \{n, \mathbf{t}, \boldsymbol{\nu}\}$ il dataset contenente i dati osservati, la *likelihood* congiunta è:

   $$L(D|\alpha, \lambda) = \prod_{i=1}^{n} \left(\frac{\alpha}{\lambda} \left(\frac{t_i}{\lambda}\right)^{(\alpha-1)} exp\left\{-\left(\frac{t_i}{\lambda}\right)^{\alpha}\right\}\right)^{\nu_i} \left(exp\left\{-\left(\frac{t_i}{\lambda}\right)^{\alpha}\right\}\right)^{1-\nu_i}$$

   Calcolando la *log-likelihood* dopo alcuni passaggi si ottiene la seguente funzione:

   $$l(D|\alpha, \lambda) = \sum_{i=1}^{n} \left\{v_i \left(log(\alpha) - \alpha(log(\lambda) + (\alpha - 1)log(t_i) - \left(\frac{t_i}{\lambda}\right)^{\alpha}\right) - (1 - v_i)\left(\frac{t_i}{\lambda}\right)^{\alpha}\right\}$$

   Per costruire un modello di sopravvivenza di tipo *Weibull* si introducono le covariate tramite il parametro λ, nel seguente modo: $\lambda_i = exp\{\mathbf{x_i^t \beta}\}$. Il vettore $\boldsymbol{x_i}$ è il vettore delle covariate associato all'osservazione i-esima e $\boldsymbol{\beta}$ è il vettore dei coefficienti. Quando entrambi i parametri α e λ sono ignoti non esiste una funzione *prior coniugata*. Una tipica specificazione è assumere che α si distribuisce come una *Gamma*: $\alpha \sim G(a_0, b_0)$ mentre i coefficienti $\boldsymbol{\beta}$ come una *Normale Multivariata*: $\boldsymbol{\beta} \sim N(\boldsymbol{\mu_0}, \Sigma_0)$. Poiché si assume che i parametri siano tra loro indipendenti, la *posterior congiunta* è data data da:

   $$\pi(\boldsymbol{\beta}, \alpha|D) \propto L(D|\boldsymbol{\beta}, \alpha)\pi(\boldsymbol{\beta}|\boldsymbol{\mu_0}, \Sigma_0)\pi(\alpha|a_0, b_0)$$

$$\pi(\boldsymbol{\beta},\alpha|D) \propto \alpha^{(a_0-1)} \exp\Bigg\{ - b_0\alpha - \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu_0})^t\Sigma_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu_0}) + \sum_{i=1}^{n}\left[v_i\Big(\log(\alpha) - \alpha\big(\log(\lambda)+(\alpha-1)\log(t_i)\big) - \left(\frac{t_i}{\lambda_i}\right)^{\alpha}\Big) - (1-v_i)\left(\frac{t_i}{\lambda_i}\right)^{\alpha}\right]\Bigg\}$$

Non essendo disponibile una forma chiusa per la posterior congiunta si è stimato il modello tramite il software STAN, ovvero tramite metodi di campionamento MCMC.

3. Modello Log Normale

Un ulteriore modello parametrico per l'analisi della sopravvivenza è il modello *Log Normale*. In questo modello si assume che i tempi di sopravvivenza si distribuiscano in modo indipendente come una distribuzione *log normale* con parametri μ, σ^2 , ovvero $t_i \sim LN(\mu,\sigma^2)$, e abbiano quindi la seguente funzione di densità:

$$f(t_i|\mu,\sigma) = (2\pi)^{-1/2}(t_i\sigma)^{-1}\exp\left\{-\frac{1}{2\sigma^2}(\log(t_i)-\mu)^2\right\}.$$

La funzione di sopravvivenza che ne deriva è data da :

$$S(t_i|\mu,\sigma) = 1 - \Phi\left(\frac{log(t_i)-\mu}{\sigma}\right).$$

Dato $D=\{n,\boldsymbol{t},\boldsymbol{\nu}\}$ il dataset contenente i dati osservati, la *likelihood* congiunta è data da

$$L(D|\mu,\sigma) = (2\pi\sigma^2)^{-\sum_{i=1}^{n}\nu_i/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\nu_i\left(\log(t_i)-\mu\right)^2)\right\}\prod_{i=1}^{n}t_i^{-\nu_i}\left(1-\Phi\left(\frac{log(t_i)-\mu}{\sigma}\right)\right)^{1-\nu_i}.$$

Per costruire un modello di sopravvivenza di tipo *Log Normale* si introducono le covariate tramite il parametro μ nel seguente modo: $\boldsymbol{\mu_i} = \boldsymbol{x_i^t\beta}$. Il vettore $\boldsymbol{x_i}$ è il vettore delle covariate associato all'osservazione i-esima e $\boldsymbol{\beta}$ è il vettore dei coefficienti. Inoltre, si è utilizzato invece che σ^2, la riparametrizzazione $\tau = \dfrac{1}{\sigma^2}$. Quando entrambi i parametri β e τ sono ignoti non è disponibile una funzione *prior coniugata*. La specificazione tipica che si è utilizzata per la distribuzione dei parametri è una *Normale Multivariata* condizionata a τ per i coefficienti $\boldsymbol{\beta}$: $\boldsymbol{\beta}|\tau \sim N_p\left(\mu_0,\tau^{-1}\Sigma_0\right)$, mentre per il parametro τ una distribuzione *Gamma*: $\tau \sim G(\alpha_0/2,\lambda_0/2)$. La *posterior congiunta* è data da :

$$\pi(\boldsymbol{\beta},\tau|D) \propto L(D|\boldsymbol{\beta},\tau)\pi(\boldsymbol{\beta}|\tau,\boldsymbol{\mu}_0,\Sigma_0)\pi(\tau,\alpha_0,\lambda_0)$$

$$\pi(\boldsymbol{\beta},\tau|D) \propto \tau^{\frac{a_0+d}{2}}\exp\left\{-\frac{\tau}{2}\left[\sum_{i=1}^{n}\nu_i(\log(t_i)-\boldsymbol{x_i^t\beta})^2 + (\boldsymbol{\beta}-\boldsymbol{\mu_0})'\Sigma_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu_0}) + \lambda_0\right]\right\} \times \prod_{i=1}^{n} y_i^{-\nu_i}\left(1-\Phi\left(\tau^{1/2}(\log(t_i)-\boldsymbol{x_i^t\beta})\right)\right)^{1-\nu_i}.$$

Non essendo disponibile una forma chiusa per la posterior congiunta si è stimato il modello tramite il software STAN, ovvero tramite metodi di campionamento MCMC.

4. Piecewise Exponential Model

Metodi semi-parametrici e non parametrici nell'analisi della sopravvivenza sono diventati sempre più popolari negli ultimi anni. Uno dei modelli più utilizzati in questo ambito è il modello semi-parametrico *Piecewise Constant Hazard Model*. L'idea del modello è quella di costruire una partizione dell'asse temporale in intervalli di ampiezza uguale, $0 < s_1 < s_2 < ... < s_J$ con $s_J < t_i$ per ogni $i = 1,...,n$, quindi si ottengono J intervalli: $(0,s_1],(s_1,s_2],...,(s_{j-1},s_j]$. Nel j-esimo intervallo si assume un *hazard* di base costante $h_0 = \lambda_j$ per ogni $t \in I_j = (s_{j-1},s_j]$. Dato $D=\{n,\boldsymbol{t},X,\boldsymbol{\nu}\}$ il dataset contenente i dati osservati dove $\nu_i = 1$ se l'osservazione i-esima ha sviluppato l'evento e il vettore $\lambda = (\lambda_1,...,\lambda_J)$, la funzione *likelihood* per gli n soggetti è data da:

$$L(D|\boldsymbol{\beta},\boldsymbol{\lambda}) = \prod_{i=1}^{n}\prod_{j=1}^{J}\left(\lambda_j exp\{\boldsymbol{x_i^T\beta}\}\right)^{\delta_{ij}\nu_i} exp\left\{-\delta_{ij}\left[\lambda_j(t_i-s_{j-1}) + \sum_{g=1}^{j-1}\lambda_g(s_g-s_{g-1})\right] exp\{\boldsymbol{x_i^T\beta}\}\right\}$$

dove $\delta_{ij} = 1$ se l'i-esima osservazione ha sviluppato l'evento nell'intervallo j-esimo. Questa funzione fa riferimento al modello *Piecewise Exponential Model*, esso è un modello semplice e molte volte è utilizzato come *benchmark* per confrontarlo con altri modelli semi o non parametrici più avanzati. Si può notare che se $J = 1$ allora $\lambda = \lambda_1$ e ci si riconduce al modello esponenziale. Una *prior* comune utilizzata per l'*hazard* di questo modello è la distribuzione *Gamma* indipendente, ovvero una distribuzione indipendente per ogni λ_j: $\lambda_j \sim G(a_0, b_0)$ per $j = 1, ..., J$, oppure si utilizza una *prior* dipendente (correlata). In questa analisi è stato implementato il modello sopra specificato utilizzando una funzione *prior Gamma* dipendente, derivante da un processo di *Gamma Markov Chain*, ovvero $\lambda_j \sim Gamma$ con valore atteso e varianza pari a:

$$E(\lambda_j|\lambda_{j-1}, ..., \lambda_1) = \lambda_{j-1}$$

$$Var(\lambda_j|\lambda_{j-1}, ..., \lambda_1) = \frac{(\lambda_{j-1})^2}{\alpha}$$

dove α è un iper-parametro scelto. Questo tipo di *prior* assume che ci sia dipendenza tra l'*hazard* di un intervallo rispetto all'intervallo successivo e in questo modo si introduce regolarità nel tempo. Il numero di intervalli è stato considerato pari a $J = (\frac{n}{log(n)})^{\frac{1}{1+2\gamma}}$, dove γ è il parametro di regolarità (*smoothness*) della vera funzione *hazard* che in assenza di informazione a priori è definito pari a $\gamma = \frac{1}{2}$.

3 Risultati

Data pre-processing

Innanzitutto si sono svolte alcune analisi descrittive dei dati. Si è notato che il dataset contiene alcuni valori anomali per la variabile *pgr*, poichè essa assume solitamente valori fino a circa 1000fmol/l, quindi sono state scartate 6 osservazioni, ottenendo un dataset di 680 osservazioni. Valori elevati di questa variabile rappresentano una buona risposta del tumore alla terapia ormonale. Si osserva, inoltre, che il numero di soggetti censurati è pari a 381, mentre i soggetti che hanno sviluppato l'evento sono 299. Il tempo mediano di sopravvivenza è pari circa a 1800 giorni. Nell'applicazione dei modelli le variabili continue sono state normalizzate. Tramite il grafico mostrato in Figura 1 si può notare la curva di sopravvivenza empirica stimata tramite lo stimatore Kaplan-Meier, che mostra, nel tempo, quante persone non hanno ancora sviluppato l'evento. Ogni segmento della curva rappresenta un soggetto censurato.

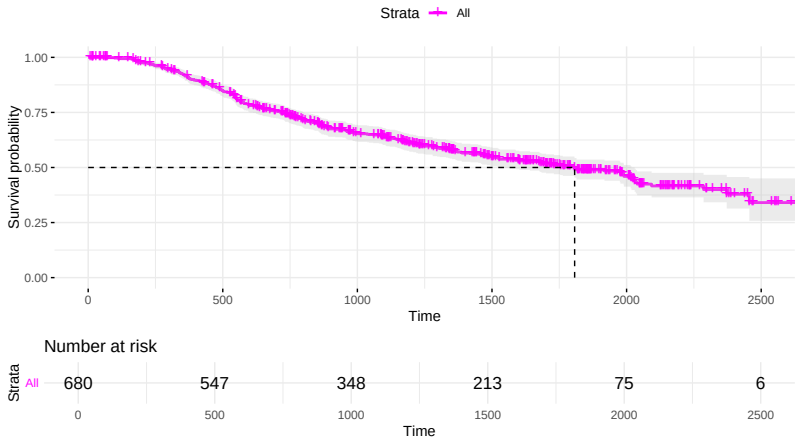

Figura 1: Curva di Kaplan-Meier

Modelli

Dopo aver concluso le procedure di pulizia del dato, sono stati stimati i tre modelli considerati, il modello *Esponenziale*, *Weibull* e *Log-Normale* tramite il software *STAN*. Come prima cosa sono stati implementati tutte e tre i modelli con tutte le covariate disponibili, ma dopo una procedura di *Covariate Selection* si è deciso di eliminare la variabile *er* poichè non significativa. I risultati di seguito fanno riferimento all'implementazione dei modelli senza la suddetta variabile.

1. Modello Esponenziale

   Come primo modello si deciso di sviluppare il modello *Esponenziale*. Come funzione *prior* sui coefficienti di regressione $\boldsymbol{\beta}$ è stata scelta una distribuzione *Normale Multivariata* con parametri: $\boldsymbol{\mu_0} = \mathbf{0}$ e Σ_0 una matrice diagonale con varianza pari a 10^4. Nel grafico in Figura 8 contenuta in Appendice 4 si mostrano le due traiettorie di numerosità 5000, con *burn-in* di 1000, ottenute dal campionamento dei coefficienti dalla loro distribuzione a posteriori, si nota che esse arrivano tutte a convergenza e seguono un andamento stazionario.

   Tramite strumenti di diagnostica, si è osservato che i coefficienti stimati sono ottimali e coerenti con l'analisi considerata. Tuttavia data la semplicità del modello e valori di *WAIC* e *LPML* meno soddisfacenti in confronto ai modelli *Weibull* e *Log-Normale*, si è deciso di non approfondire l'analisi con questo tipo di modello. Inoltre, c'è da ricordare che questo modello considera un *hazard* costante nel tempo, che però non riflette il comportamento dei dati.

2. Modello Weibull

   Un altro modello sviluppato è stato il modello *Weibull*. In questo caso i tempi di sopravvivenza si distribuiscono come una *Weibull* e la funzione *hazard* varia nel tempo in base al parametro di scala α. Come funzione a priori per il parametro α è stata scelta una *Gamma* con parametri $a_0 = 2$ e $b_0 = 2$, mentre per i parametri $\boldsymbol{\beta}$ una *Normale Multivariata* con media $\boldsymbol{\mu_0} = \mathbf{0}$ e matrice di varianza e covarianza Σ_0 diagonale con varianza pari a 10^4. Nel grafico in Figura 9, contenuta in Appendice 4, si mostrano le traiettorie ottenute utilizzando due catene di numerosità 5000 e *burn-in* di 1000. Si può notare che le traiettorie seguono un andamento stazionario, per cui l'approssimazione della distribuzione è attendibile.

   I coefficienti stimati sono stati valutati tramite l'*Effective-Sample Size* in cui in tutti i casi si ottengono valori che mostrano un andamento della catena stazionario e omogeneo. Inoltre, in Figura 10 e in Figura 11 (in Appendice 4), si nota che i coefficienti $\boldsymbol{\beta}$ stimanti assumono a posteriori una curva di densità simile a quella di una Gaussiana. Per questo modello è stato ottenuto un valore di *LPML* pari a -2940.6 e un *WAIC* pari a 5881.2.

   Nella seguente Tabella 1 vengono illustrate le medie a posteriori dei coefficienti, del parametro di scala α e i relativi intervalli di credibilità, calcolati tramite metodo *Highest Posterior Density*. In questo tipo di modello la funzione *hazard* è definita come: $h(t_i) = \alpha \left(\frac{1}{exp\{\boldsymbol{x_i^t\beta}\}}\right)^{\alpha} t_i^{\alpha-1}$; si può notare quindi che il parametro di scala α, quando assume valori maggiori di 1, provoca un aumento dell'*hazard* all'aumentare del tempo. Nel caso analizzato il parametro di scala assume un valore pari a 0.3977, quindi l'*hazard* diminuisce all'aumentare del tempo. Questo potrebbe essere spiegato dal fatto che la terapia ormonale considerata potrebbe fare effetto solo dopo un periodo di tempo, per questo motivo all'inizio dello studio il rischio sembra essere più elevato. Il coefficiente β_1, associato alla variabile *age*, mostra come all'aumentare di un anno l'età delle pazienti il rischio di accadimento aumenta, mentre la sopravvivenza diminuisce, questo perché il valore del coefficiente è negativo. Anche il coefficiente associato alla variabile *size* è negativo e questo sta a significare che, un aumento del valore della variabile a parità di tutto il resto, porta ad un aumento del rischio e ad una diminuzione della sopravvivenza. Questo vale anche per il coefficiente legato alla variabile *nodes*. Per quanto riguarda le variabili *meno* e *hormon*, invece, i coefficienti a loro associati hanno segno positivo, questo sta a significare che una donna in menopausa o una donna che ha ricevuto un trattamento ormonale ha una sopravvivenza più elevata e un rischio minore, a parità di tutto il resto, rispetto a una donna non in menopausa e che non ha svolto una terapia ormonale. Infine alla variabile *grade* è associato segno positivo e un valore della media a posteriori elevato, che sta a significare che al crescere della gravità del tumore la sopravvivenza della paziente aumenta, che rappresenta un risultato poco coerente.

Coefficiente	Media a posteriori	HPD al 95%
age	-0.9084	(-1.32, -0.51)
size	-0.3360	(-0.59, -0.09)
nodes	-0.5598	(-0.76, -0.37)
pgr	1.4515	(1.04, 1.88)
meno	2.2778	(1.44, 3.02)
grade	3.5874	(3.34, 3.84)
hormon	1.6363	(1.03, 2.22)
alpha	0.3977	(0.36, 0.43)

Tabella 1: Medie a posteriori e IC dei parametri nel modello *Weibull*

Nel modello *Weibull* è difficile interpretare di quanto aumenta o diminuisce la sopravvivenza o il rischio in termini percentuali a causa del parametro di scala α. Per questo motivo si è scelto di interpretare i coefficienti anche in modo grafico attraverso la rappresentazione delle curve di sopravvivenza stimate. In Figura 2 vengono rappresentate due curve di sopravvivenza distinte per le pazienti che hanno ricevuto una terapia ormonale rispetto a quelle che non l'hanno seguita, mantenendo valori costanti, pari alla mediana, per il resto delle covariate. Si può notare una forte differenza tra le due curve, concludendo che la terapia ormonale è un fattore che incide molto sulla sopravvivenza di soggetti affetti da cancro al seno.

Come si può notare in Figura 3, anche l'età è un fattore che incide molto sulla sopravvivenza, infatti si osserva una discrepanza tra la curva di una donna di trent'anni e una donna di sessant'anni, stimata a parità del resto delle variabili esplicative.

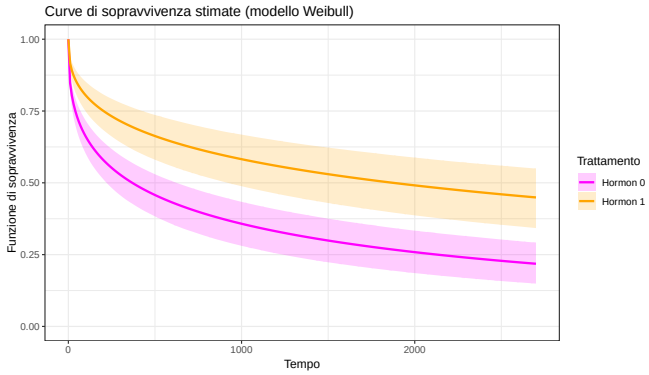

Figura 2: Curve di sopravvivenza per variabile *hormon* nel modello *Weibull*

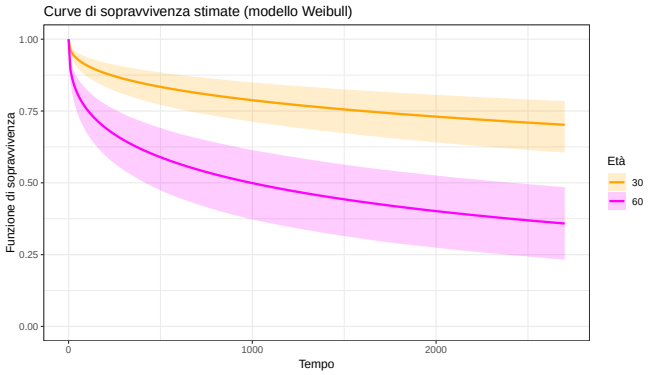

Figura 3: Curve di sopravvivenza per variabile *età* nel modello *Weibull*

3. Modello Log Normale

Un ulteriore modello sviluppato per questa analisi è stato il modello *Log Normale*, in cui si suppone che i tempi di sopravvivenza seguano una distribuzione *Log Normale*. Si è utilizzata come specificazione del modello la tipica mostrata in precedenza e si è scelta per il parametro τ una *Gamma* con parametri $\alpha_0 = 2$ e $\lambda_0 = 2$, mentre per i coefficienti di regressione $\boldsymbol{\beta}$ una *Normale Multivariata* centrata in $\boldsymbol{\mu_0} = 0$ e matrice di varianza e covarianza Σ_0 diagonale con varianza pari a 10^4. Nei grafici in Figura 12 contenuta in Appendice 4 sono mostrate le traiettorie ottenute utilizzando due catene da 5000 iterazioni e un periodo di *burn-in* pari a 1000. Le traiettorie mostrano un andamento regolare e stazionario, che porta quindi a una approssimazione attendibile della posterior congiunta di interesse. Questo viene confermato anche tramite l'*Effective Sample Size* del campione abbastanza elevata e i grafici in Figura 13 e in Figura 14 contenute in Appendice 4, che mostrano curve di densità con andamenti simili a curve di densità gaussiane. Per questo modello è stato ottenuto un valore di *LPML* pari a -2870.3 e un *WAIC* pari a 5740.6.

Le medie a posteriori dei parametri e i relativi intervalli di credibilità, calcolati tramite metodo *Highest Posterior Density*, sono mostrati in Tabella 2.

Coefficiente	Media a posteriori	HPD al 95%
age	-0.7551	(-1.09, -0.41)
size	-0.2408	(-0.47, 0.00)
nodes	-0.6902	(-0.94, -0.47)
pgr	0.9884	(0.71, 1.28)
meno	2.0505	(1.36, 2.74)
grade	3.0001	(2.78, 3.21)
hormon	1.3628	(0.85, 1.87)
tau	0.1450	(0.12, 0.17)

Tabella 2: Medie a posteriori e IC dei parametri nel modello *Log Normale*

Per questa tipologia di modello, è possibile interpretare l'effetto dei coefficienti in modo diretto sul logaritmo del tempo di sopravvivenza. Tuttavia, volendo confrontare i risultati con gli altri modelli stimati, si è

utilizzata la versione esponenziata dei coefficienti e il suo impatto sul tempo mediano di sopravvivenza. In particolare si ha che, a parità di tutte le altre covariate, l'effetto delle singole è il seguente:

(a) *Age*: Aumentando di un anno l'età di una paziente, il tempo mediano di sopravvivenza diminuisce del 53.00% ;

(b) *Size*: Aumentando di una unità la dimensione del tumore, il tempo mediano di sopravvivenza diminuisce del 21.4% ;

(c) *Nodes*: Aumentando di una unità il numero dei linfonodi in cui si trova la presenza di cellule tumorali, il tempo mediano di sopravvivenza diminuisce del 49.5%;

(d) *Pgr*: Aumentando di una unità la quantità di recettore di progesterone rilevati, il tempo mediano di sopravvivenza aumenta del 168.6%;

(e) *Meno*: Passando da una donna non in menopausa a una in menopausa, il tempo mediano di sopravvivenza aumenta del 677.17%;

(f) *Grade*: Aumentando di un'unità il grado del tumore, il tempo mediano di sopravvivenza aumenta del 1908.7%;

(g) *Hormon*:. Passando da una donna che non ha seguito la terapia ormonale a una che l'ha seguita, il tempo mediano di sopravvivenza aumenta del 290.07%.

Similmente al modello Weibull mostrato in precedenza, il coefficiente associato alla variabile *grade* presenta dei valori non coerenti con il significato della variabile. Le variabili più impattanti risultano essere le variabili *Meno* e *Hormon* e per evidenziare il loro effetto sul tempo di sopravvivenza atteso dei pazienti si sono svolte delle rappresentazioni grafiche di alcune curve di sopravvivenza. In Figura 4 si possono notare due curve di sopravvivenza per pazienti non in menopausa e con tumore di grado 2 differenziate per la terapia ormonale seguita o meno, mantenendo i valori delle altre covariate fissi nei valori mediani. Si può notare subito come per coloro che sono state sottoposte alla terapia ormonale la sopravvivenza attesa è visibilmente superiore.

In Figura 5 vengono presentate due curve di sopravvivenza per pazienti con tumore di grado 2 sottoposte a terapia ormonale differenziate per lo stato di menopausa, mantenendo i valori delle altre covariate fissi nei valori mediani. Si può notare come la curva di sopravvivenza attesa per coloro che sono in menopausa è visibilmente superiore.

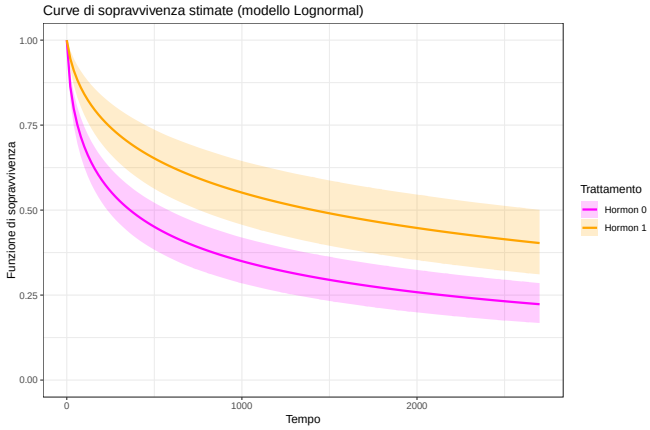

Figura 4: Curve di sopravvivenza per variabile *hormon* nel modello *Log Normale*

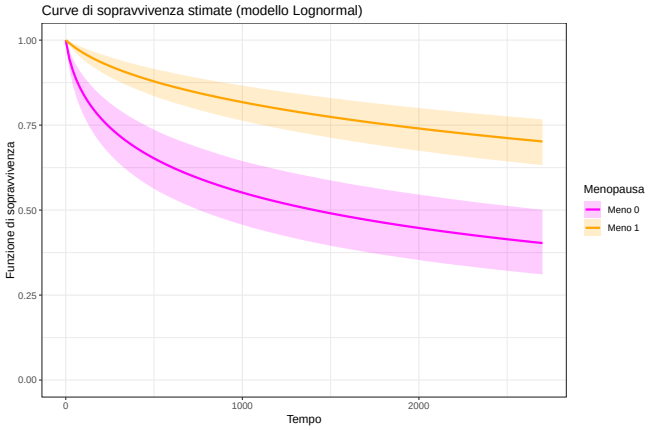

Figura 5: Curve di sopravvivenza per variabile *meno* nel modello *Log Normale*

4. Modelli con effetti random

Un evidente problema dei modelli con effetti fissi è l'impatto troppo elevato della variabile *grade*, la cui interpretazione in entrambi i casi mostra che all'aumentare del grado del tumore, la sopravvivenza aumenta, il che non è coerente intuitivamente. Essendo una variabile categorica che assume 3 valori, si è deciso di stimare nuovamente i modelli *Log Normale* e *Weibull* utilizzando la variabile *grade* come effetto random; in particolare considerando quindi un'intercetta random per ognuno dei tre livelli di *grade*. Per farlo si è considerato un vettore di parametri $\boldsymbol{u}$ di lunghezza 3, in cui ogni elemento rappresenta l'effetto random associato a una categoria della variabile *grade*. Per questi parametri è stata utilizzata come prior una *Normale Multivariata* centrata nel vettore nullo e con matrice di varianza e covarianza diagonale, percui

per ogni elemento del vettore $\boldsymbol{u}$ si ha che $u_j \sim N(0, 10^4), j = 1, 2, 3$. Gli effetti random nei due modelli di regressione sono stati introdotti tramite il predittore lineare. Per il modello *Weibull* il predittore lineare è diventato $\lambda_{ij} = exp\{\mathbf{x_i^t}\boldsymbol{\beta} + grade_i * u_j\}$. Per il modello *Log Normale* invece il predittore lineare è diventato $\mu_{ij} = x_i^t\boldsymbol{\beta} + grade_i * u_j$. In Tabella 3 sono mostrati le medie a posteriori dei parametri e gli intervalli di credibilità calcolati tramite *Highest Posterior Density* per entrambi i modelli.

	Modello Log-normale		**Modello Weibull**	
Coefficiente	Media a posteriori	HPD al 95%	Media a posteriori	HPD al 95%
age	0.12264	(-0.01, 0.27)	0.06442	(-0.07, 0.19)
size	-0.09036	(-0.17, 0.00)	-0.08063	(-0.16, 0.00)
nodes	-0.27597	(-0.36, -0.19)	-0.20976	(-0.27, -0.15)
pgr	0.21123	(0.10, 0.31)	0.23937	(0.12, 0.36)
meno	-0.25273	(-0.56, 0.02)	-0.20307	(-0.46, 0.07)
hormon	0.30796	(0.12, 0.50)	0.27310	(0.09, 0.46)
grade 1	7.92285	(7.54, 8.28)	8.21008	(7.80, 8.58)
grade 2	7.42167	(7.21, 7.64)	7.71160	(7.51, 7.91)
grade 3	7.27227	(7.02, 7.52)	7.59957	(7.36, 7.84)
tau	1.02863	(0.85, 1.22)		
alpha			1.35641	(1.23, 1.50)

Tabella 3: Medie a posteriori e IC dei parametri nel modello random *Log Normale* e *Weibull*

Per quanto riguarda il modello *Log Normale* si nota che gli effetti random stimati portano ad avere risultati più coerenti. Se si considerano tutte le covariate nulle, una paziente con tumore di grado 1 ha un tempo di sopravvivenza mediano stimato pari a 2750 giorni, una paziente con tumore di grado 2 ha un tempo di sopravvivenza mediano stimato pari a 1669 giorni, e infine una paziente con tumore di grado 3 ha un tempo di sopravvivenza mediano stimato pari a 1436 giorni. L'effetto quindi di *grade* ha un valore molto più sensato perchè al crescere della gravità del tumore i giorni di sopravvivenza attesi diminuiscono. Per quanto riguarda le altre covariate, le variabili *age* e *meno* perdono la loro significatività, come si può notare dagli intervalli di credibilità; le altre covariate invece rimangono coerenti con il modello *Log Normale* a effetti fissi ma hanno un effetto molto meno marcato e in alcuni casi più sensato. Ad esempio per la variabile *hormon* si ha, che passando da una donna che non ha seguito la terapia ormonale a una che l'ha seguita, il tempo mediano di sopravvivenza aumenta del 34.98% invece che del 270%. Per questo modello è stato ottenuto un valore di *LPML* pari a -2569.2 e un *WAIC* pari a 5138.4.

In Figura 6 vengono mostrate tre curve di sopravvivenza differenti per il modello *Log Normale* per gli effetti random stimati per le tre categorie della variabile *grade*. Le curve mostrate sono rappresentative di una paziente con valori delle covariate fisse nei valori mediani, non in menopausa e con terapia ormonale. Si può notare come sia più elevata la curva per il livello del tumore meno grave (*Grade 1*).

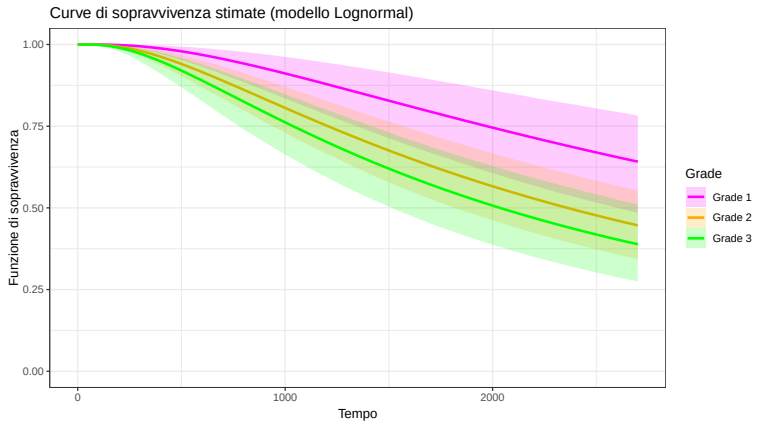

Figura 6: Curve di sopravvivenza per variabile *grade* nel modello *Log Normale random*

Per quanto riguarda il modello *Weibull* l'effetto dei coefficienti di regressione stimati e dei coefficienti random è simile al modello *Log Normale*. Anche in questo caso si ha un miglioramento rispetto al modello a effetti

fissi, inoltre il parametro di scala α assume un valore superiore a 1, il che è più sensato per questo tipo di analisi. Questo significa che all'aumentare del tempo la probabilità di sviluppare l'evento aumenta. Per questo modello è stato ottenuto un valore di *LPML* pari a -2591.2 e un *WAIC* pari a 5182.4.

5. Modello Piecewise Exponential

   Infine è stato considerato un modello semi-parametrico *Piecewise Exponential*, un modello esponenziale a tratti in cui a ogni intervallo del tempo λ_j, per $j = 1, ...J$, è stata assegnata una distribuzione a priori *Gamma* dipendente, ovvero con media λ_{j-1} e varianza $\frac{(\lambda_{j-1})^2}{\alpha}$. Per implementarlo è stato utilizzato il pacchetto del software R *BayesSurvival*. In questo caso, come si può vedere in Figura 7, è stata considerata soltanto la curva di sopravvivenza stimata dal modello in relazione alla curva di *Kaplan-Meier*, ovvero la stima della curva di sopravvivenza empirica. Si può notare che il modello interpola in modo ottimale la curva empirica, e quindi una specificazione dell'*hazard* variabile nel tempo permette un migliore adattamento del modello ai dati.

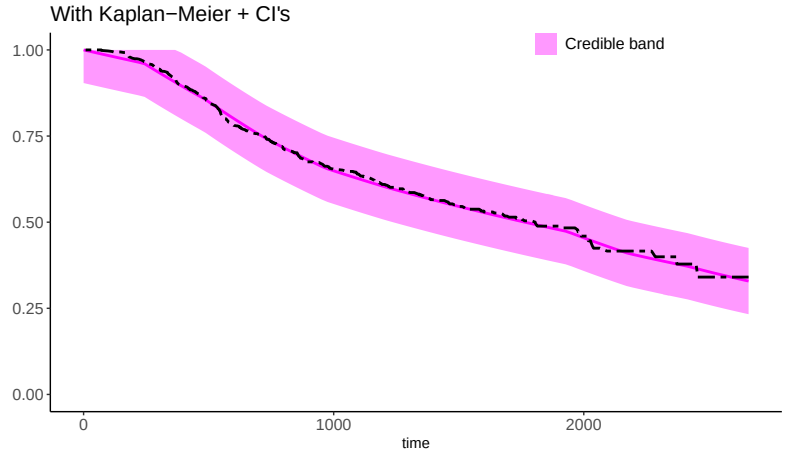

Figura 7: Curve di sopravvivenza di *Kaplan-Meier* contro la curva stimata

4 Conclusione

In questo progetto sono stati presentati diversi modelli bayesiani per l'analisi della sopravvivenza che sono stati implementati tramite il software *STAN* e applicati al dataset *Breast Cancer*. Tutti i modelli considerati hanno ottenuto risultati soddisfacenti e coerenti tra di loro. Sia nel caso dei modelli a effetti fissi che nel caso a effetti random il modello *Log-Normale* presenta un valore del criterio *WAIC* minore rispetto agli altri modelli. Il modello migliore selezionato è il modello *Log Normale* con effetti random perché permette di risolvere problematiche specifiche per il problema analizzato producendo dei coefficienti più adatti ai dati considerati. Si potrebbe estendere e arricchire quest'analisi utilizzando ulteriori metodologie, in particolare sviluppando il modello di *Cox* in ambito bayesiano oppure modelli bayesiani per l'analisi della sopravvivenza non parametrici.

Appendice

Grafici di diagnostica

Modello Esponenziale

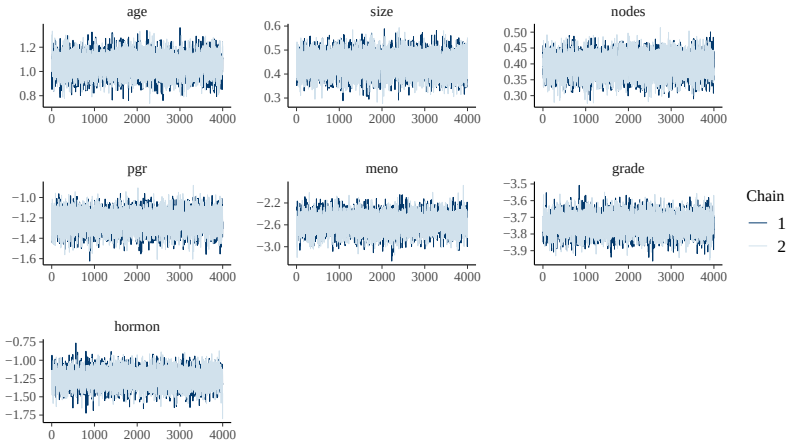

Figura 8: Traiettorie coefficienti $\boldsymbol{\beta}$ modello *Esponenziale*

Modello Weibull

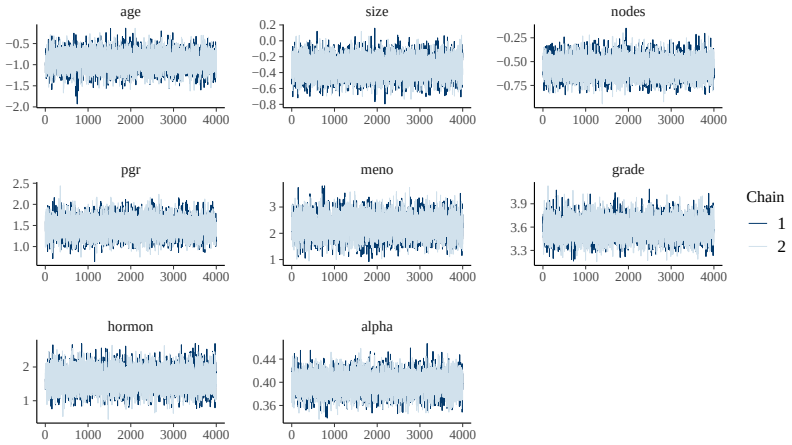

Figura 9: Traiettorie coefficienti $\boldsymbol{\beta}$ e α modello *Weibull*

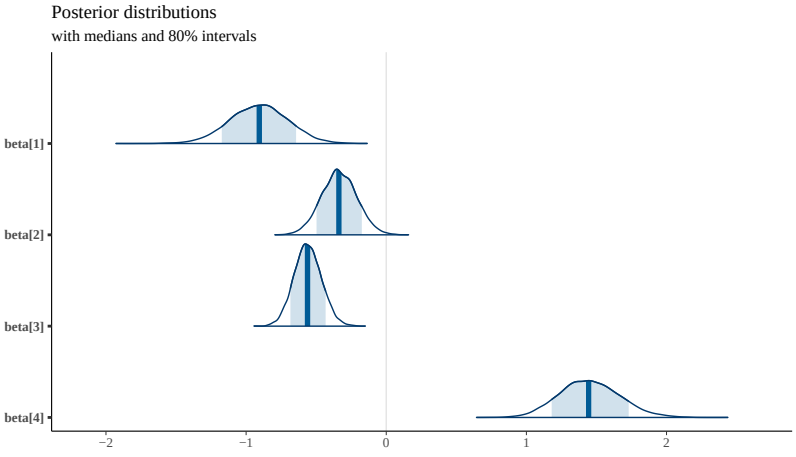

Figura 10: Densità a posteriori coefficienti $\boldsymbol{\beta_{1:4}}$

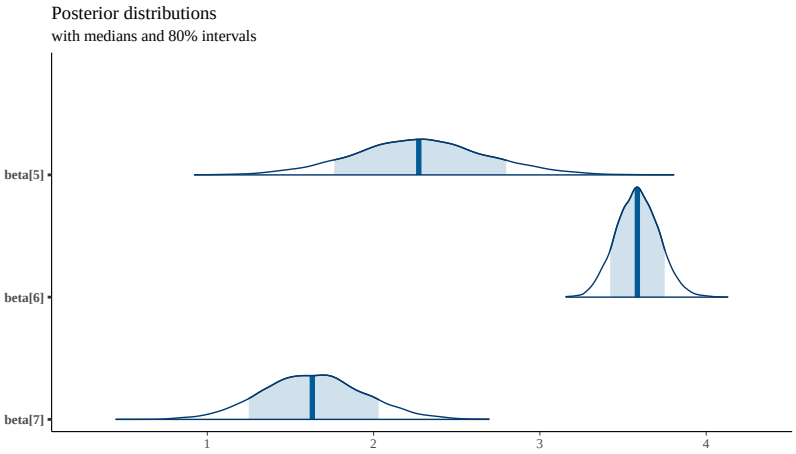

Figura 11: Densità a posteriori coefficienti $\boldsymbol{\beta_{5:7}}$

Modello Log Normale

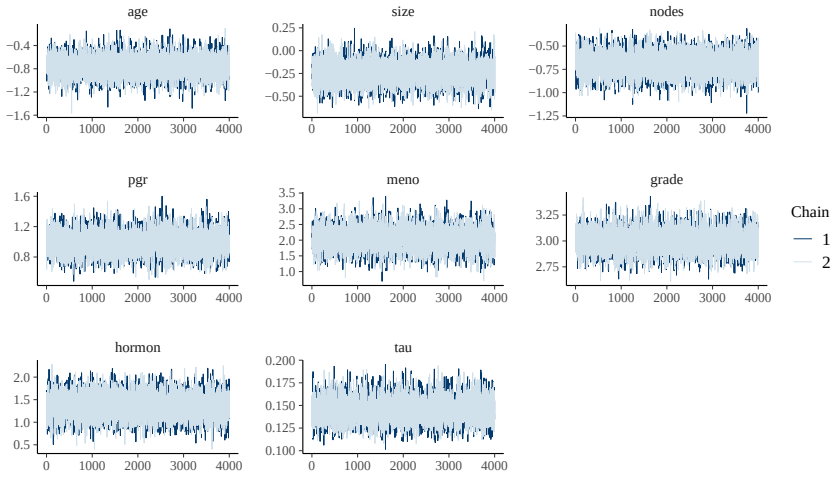

Figura 12: Traiettorie coefficienti $\boldsymbol{\beta}$ e τ modello *Log Normale*

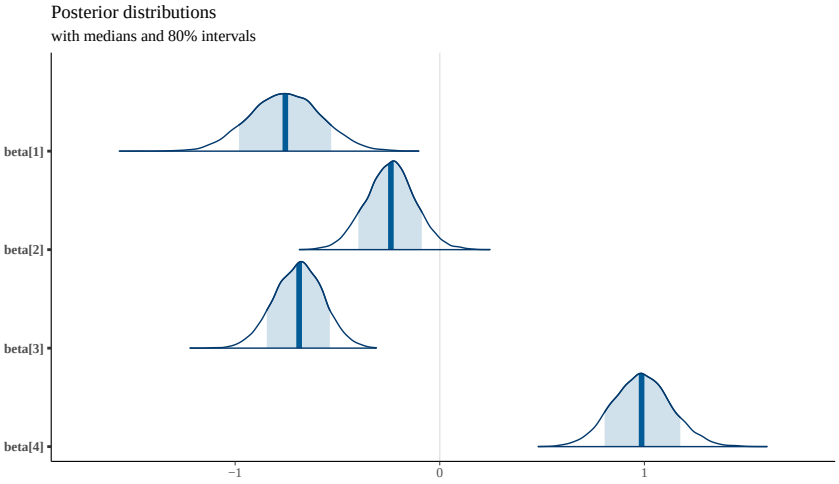

Figura 13: Curve di densità per i coefficienti $\boldsymbol{\beta_{1:4}}$ del modello *Log Normale*

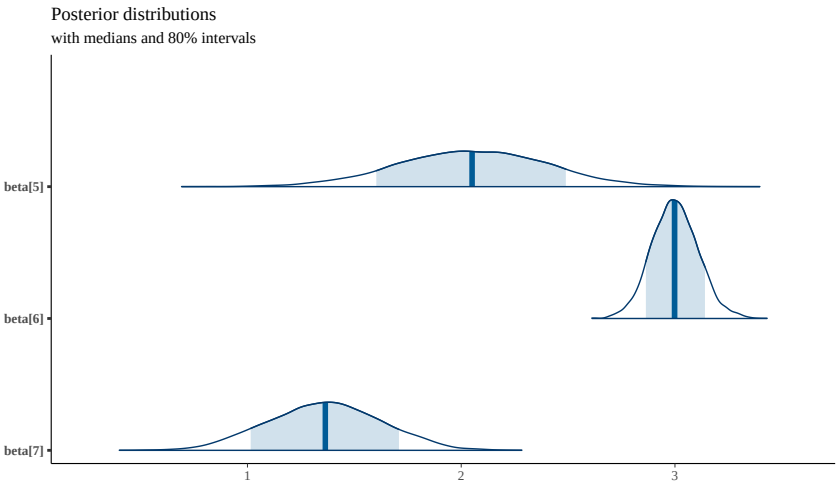

Figura 14: Curve di densità per i coefficienti $\boldsymbol{\beta_{5:7}}$ del modello *Log Normale*

Modello Weibull con effetti random

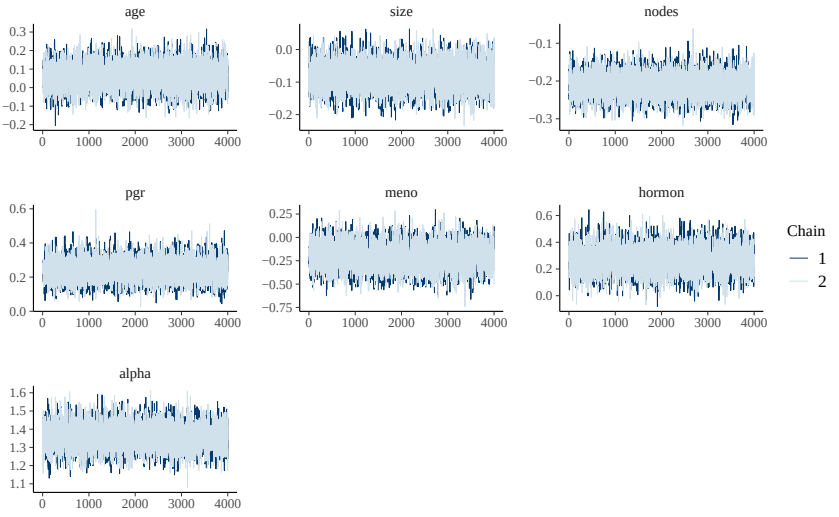

Figura 15: Traiettorie coefficienti $\boldsymbol{\beta}$ e α modello *Weibull con random*

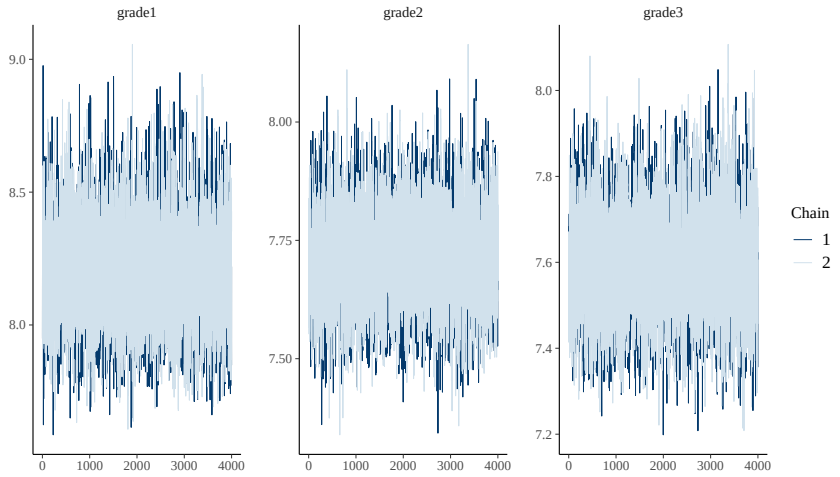

Figura 16: Traiettorie coefficienti $\boldsymbol{u}$ modello *Weibull con random*

Modello Log Normale con effetti random

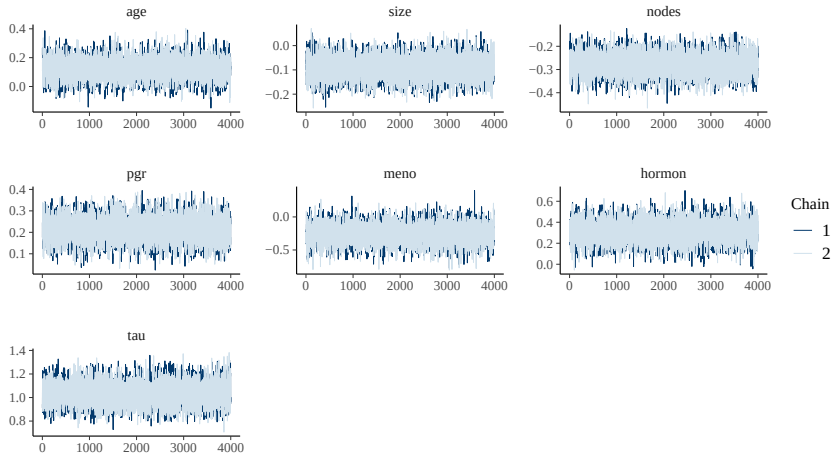

Figura 17: Traiettorie coefficienti $\boldsymbol{\beta}$ e τ modello *Log Normale con random*

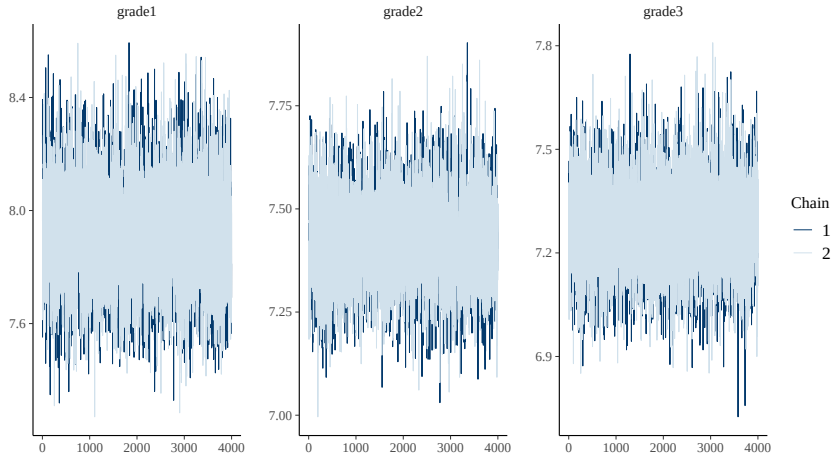

Figura 18: Traiettorie coefficienti $\boldsymbol{u}$ modello *Log Normale con random*

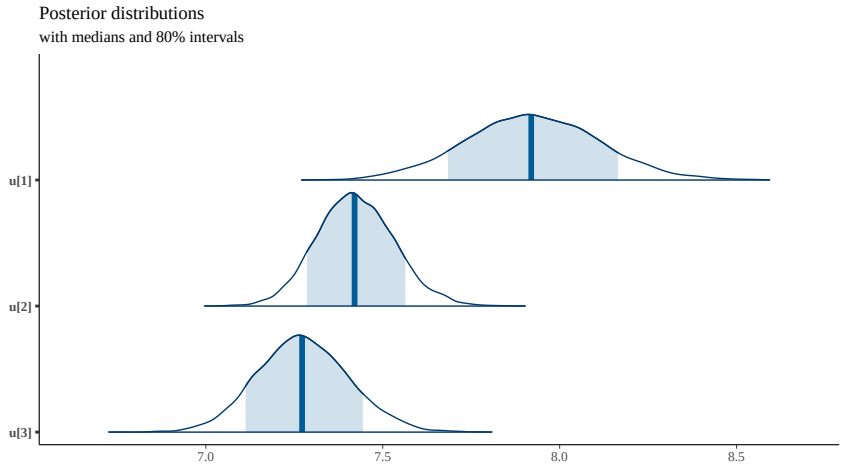

Figura 19: Curve di densità per i coefficienti $\boldsymbol{u}$ del modello *Log Normale con random*

Riferimenti bibliografici

[1] (2013) J. Ibrahim. and M. Chen and D. Sinha, *Bayesian Survival Analysis*, Springer Science & Business Media

[2] *Breast Cancer Data*
`https://www.kaggle.com/datasets/utkarshx27/breast-cancer-dataset-used-royston-and-altman/data`

[3] *Survival models in STAN*
`https://mc-stan.org/docs/stan-users-guide/survival.html`

[4] *Package 'BayesSurvival'*
`https://cran.r-project.org/web/packages/BayesSurvival/BayesSurvival.pdf`