Indice

Introduzione

L’analisi dei dati di durata è uno strumento utile in numerosi ambiti di studio e trova vaste possibilità di utilizzo anche nella ricerca economica. In particolare, nel contesto dell’analisi delle dinamiche del mercato del lavoro, per formulare e valutare politiche di intervento è di estrema importanza considerare la probabilità di abbandonare lo stato di disoccupazione in funzione della durata della permanenza stessa in tale stato e di altre caratteristiche socio-economiche o anagrafiche dell’individuo.

In tale contesto i dati sono tipicamente raccolti attraverso l’osservazione dello stato occupato da uno stesso individuo in successive waves di un’indagine di tipo panel.

Un caso di particolare interesse è il campionamento da stock dalla popolazione di disoccupati nella prima intervista, nella quale viene registrata la durata pregressa del periodo di disoccupazione. Le unità vengono ricontattate per una seconda rilevazione e vengono interrogate riguardo allo stato nel quale si trovano in tale momento.

Nel presente lavoro ipotizzeremo che tale informazione possa non essere riportata correttamente da ciascun individuo; lavori precedenti mostrano che non si tratta di una possibilità remota, ma tale errore di misura può costituire realisticamente un problema che investe porzioni non trascurabili dei dati a disposizione.

L'interesse principale riguarda gli effetti della misclassification sulla stima dei parametri di un modello di durata sotto l’assunzione che lo stato

riportato nella prima intervista, che tipicamente viene condotta in maniera più accurata, sia sempre corretto. Per valutare empiricamente tali effetti, vengono simulati dei campioni affetti da errori di misura: il risultato principale è che la presenza di errori porta a sottostimare gli effetti della durata della disoccupazione e delle altre covariate sulla probabilità di transitare all'occupazione.
Viene quindi proposto un Modello Gerarchico Bayesiano che prevede tra i parametri stimati lo stato realmente occupato da ciascun individuo e, di conseguenza, le proporzioni degli errori di misura; le distribuzioni a posteriori dei parametri vengono stimate tramite Monte Carlo Markov Chain dal software WinBUGS.
Il modello viene poi valutato sui campioni simulati affetti da errore di misura precedentemene usati: in particolare ne verrà valutata l'applicazione tramite gruppi di 100 campioni ciascuno, che differiscono per varie tipologie di dipendenza da durata e varie proporzioni di errore.

Nel primo capitolo viene descritto il modello in esame, il piano di campionamento considerato e la funzione di verosimiglianza; vengono inoltre effettuate stime tramite tale verosimiglianza su dati simulati.
Nel secondo capitolo, viene proposta una formulazione Bayesiana del modello in esame, fornita al software WinBUGS per ottenere una stima delle distribuzioni a posteriori dei parametri. Tale Modello Bayesiano viene utilizzato su dati simulati, in particolare su gruppi di 100 campioni caratterizzati da varie tipologie di dipendenza da durata.
Nel terzo capitolo, viene trattata la problematica dell'errore di misura nell'indicatore di censura; si applicano il metodo di stima di massima verosimiglianza ed il Modello Bayesiano a dati affetti da errori di misura in varie proporzioni e le stime ottenute sono confrontate con quelle mostrate

nei precedenti capitoli, al fine di valutare gli effetti dell'errore di misura sulle stime dei parametri.

Nel quarto capitolo, viene proposta una verosimiglianza corretta per dati affetti da errore di misura, ed il Modello Gerarchico Bayesiano che prevede la stima dello stato realmente occupato da ciascun individuo e delle proporzioni di unità erroneamente classificate. Tale modello viene testato sui gruppi di campioni affetti da errore di misura usati in precedenza ed i risultati confrontati con quelli ottenuti, tramite gli altri metodi presentati, nel capitolo 3.

Capitolo 1

Specificazione e stima di un modello di durata

1.1 L'analisi dei dati di durata

Nella loro accezione più generale, i dati di durata hanno origine in contesti longitudinali in cui sia possibile registrare la sequenza di stati occupati nel tempo da ciascun individuo e i tempi in cui sono avvenute le transizioni da uno stato all'altro. Gli stati devono essere in numero finito e chiaramente definiti, in modo che, in qualsiasi momento, sia possibile assegnare ciascun individuo ad uno e un unico stato. Uno degli scopi dell'analisi di questo tipo di dati è valutare la probabilità di transitare in un particolare stato in relazione alla storia pregressa dell'individuo e ad altre sue caratteristiche.
L'analisi dei dati di durata può essere perciò uno strumento utile in numerosi ambiti di studio e ricerca, ogni qualvolta sia possibile identificare un numero finito di stati dentro ai quali gli individui si trovano a transitare. Nell'ambito della ricerca sociale, ad esempio, possono venir identificati gli stati "celibe", "coniugato", "divorziato" ecc. se si vuole studiare la storia matrimoniale

degli individui; così come, in ambito medico, gli stati potranno riguardare l'essere affetti o meno da una particolare patologia.

Tale classe di modelli trova vaste possibilità di utilizzo anche nella ricerca economica. In particolare, nel contesto dell'analisi delle dinamiche del mercato del lavoro, è possibile occuparsi della durata della condizione di disoccupazione: considerare la probabilità di abbandonare lo stato di disoccupazione in funzione della durata della permanenza stessa in tale stato e di altre caratteristiche socio-economiche o anagrafiche dell'individuo, è di estrema importanza per formulare e valutare politiche di intervento sul mercato del lavoro.

Nel seguito, ci rifaremo a questo contesto ed a questa particolare applicazione, limitandoci per semplicità espositiva a trattare gli stati di "occupato" e disoccupato". I risultati ottenuti saranno però facilmente generalizzabili in qualsiasi altro ambito il modello verrà applicato. Per una trattazione più completa e generale si veda, ad esempio, Jenkins (2004) e la bibliografia ivi citata.

1.1.1 Modelli per dati di durata

I metodi per l'analisi dei dati di durata assumono differenti terminologie a seconda dell'ambito applicativo e del tipo di dati disponibili. Ad esempio, nella Event History Analisys viene registrata e studiata l'intera storia dell'individuo, ovvero lo stato occupato dallo stesso in ogni momento del tempo di osservazione, solitamente per analisi in campo socio-demografico (Allison, 1984). Modelli più semplici si focalizzano, invece, nell'analisi della permanenza in un unico stato, assumendo di registrare per ogni individuo una singola durata, ad esempio la lunghezza del suo periodo di tempo in stato di disoccupazione. Nel caso di modelli a rischi competitivi, le

destinazioni possono essere molteplici, ad esempio gli stati di occupazione e di Non Forza Lavoro.

Nel caso più semplice, vengono considerate transizioni da un singolo stato ad una singola destinazione. La durata della permanenza nello stato in esame è, per ciascun individuo, la realizzazione di una variabile aleatoria T: si parla di tempi discreti qualora tale variabile sia intrinsecamente discreta o se, pur continua, venga osservata soltanto ad intervalli di tempo; altrimenti si parla di tempi continui.

Nel contesto dei tempi continui, denotando f(t) ed F(t) rispettivamente funzione di densità e di ripartizione di T, viene definita funzione di sopravvivenza

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

Si definisca inoltre la funzione di rischio come

$$\theta(t) = \lim_{\Delta t \to 0} \frac{P(T > t + \Delta t)}{P(T > t)} = \frac{f(t)}{S(t)}.$$

La funzione di sopravvivenza è inoltre ricavabile a partire dalla funzione di rischio:

$$S(t) = \exp\left[- \int_0^t \theta(z) dz \right].$$

La variabile aleatoria T è univocamente definita a partire da una qualunque delle quattro funzioni appena definite.

La funzione di densità f(t) rappresenta la concentrazione della durata del periodo in esame, ad ogni possibile istante sull'asse temporale. La funzione di rischio rappresenta la medesima concentrazione ad ogni istante di tempo, ma ci si condiziona all'avvenuta sopravvivenza fino a quell'istante: tale funzione rappresenta quindi l'intensità delle transizioni ad ogni istante di tempo (Jenkins, 2004).

Dalla forma assunta dalla funzione di rischio è possibile valutare come la probabilità di transitare vari in funzione alla durata già trascorsa nello stato in esame: se la funzione è costante, ad esempio, la durata del periodo di tempo già trascorsa in uno stato non va ad influenzare la probabilità di transitare verso un nuovo stato. Se essa è crescente, al contrario, al crescere della propria permanenza in uno stato cresce anche la probabilità di uscirne. Qualora essa fosse decrescente, infine, sono gli individui da poco entrati nello stato in esame ad avere maggiore probabilità di transitare verso altri stati.

Sono molte le forme funzionali comunemente usate per la funzione di rischio. Qui citiamo solamente la funzione Weibull: $\theta(\alpha;t) = \alpha t^{\alpha-1}$. L'interpretazione di α, parametro di tale distribuzione, è naturale: infatti dalla forma della funzione di rischio è facile notare come in caso di $\alpha>1$ vi sia una dipendenza positiva tra la durata e la probabilità di transitare; tale dipendenza sarà negativa per $\alpha<1$. Infine, in caso di $\alpha=1$, non c'è dipendenza da durata: la funzione di rischio è costante e la distribuzione Weibull diventa assimilabile ad una distribuzione esponenziale, nota per la sua proprietà di assenza di memoria.

In molti casi, la probabilità di transitare viene valutata anche in relazione ad altre caratteristiche dell'individuo, che denoteremo come variabili X. In questo contesto, un'assunzione molto semplificativa è quella di rischi proporzionali:

$\theta(\alpha,\beta;t,X) = \theta_0(\alpha;t)\exp(\beta' X)$.

Tale assunzione comporta che le funzioni di rischio di individui con valori diversi nella covariata X siano tra loro proporzionali: il valore assunto dalla variabile X non va cioè ad influire sulla forma della funzione di rischio, e

quindi sul tipo di dipendenza da durata, ma soltanto sulla probabilità di transizione.

1.1.2 Piani di campionamento

Nello specificare correttamente un modello di durata è importante valutare anche il tipo di campionamento effettuato. In particolare distinguiamo tra campionamento da flusso o da stock. Nel primo caso gli individui vengono campionati dal flusso in entrata nello stato di disoccupazione, quindi ciascuna durata viene monitorata sin dall'inizio. Nel secondo caso si campiona dallo stock di individui che si trovano nello stato di interesse nel momento stesso del campionamento. Tale scelta, tuttavia, è all'origine del cosiddetto length-bias: la probabilità di essere inseriti nel campione aumenta all'aumentare della durata della propria disoccupazione (Salant, 1977), perciò il campione non rispecchierà la popolazione, ma tenderà a privilegiare individui da più tempo in stato di disoccupazione. Fra i possibili modi per evitare che ciò produca distorsioni nella stima del modello, il contributo alla verosimiglianza di ciascun individuo può essere condizionato alla durata già trascorsa nello stato in esame (Jenkins, 2004).

Un altro importante problema riguarda la finestra di osservazione, che raramente comprende tutti gli episodi nella loro completezza.

Si parla di censura a sinistra se non viene osservata la data di entrata nello stato: in questo caso non sarà di conseguenza nota neanche la durata totale.

In caso di censura a destra, la transizione non è ancora avvenuta in tutto il periodo d'osservazione; ancora, ciò non rende nota la durata totale del periodo. In questo caso il contributo alla verosimiglianza dato dall'individuo sarà S(t), con t il lasso di tempo trascorso dall'entrata nello stato alla fine del periodo di osservazione. Viene quindi costruita la variabile dicotomica δ,

con $\delta=0$ in caso di censura a destra e $\delta=1$ in caso di transizione, detta indicatore di censura.

Nel caso di campionamento da flusso la verosimiglianza diviene la seguente:

$$L(\alpha,\beta;t,x,\delta) = \prod_{i=1}^{n} \left([f(\alpha,\beta;t_i,x_i)]^{\delta_i} [S(\alpha,\beta;t_i,x_i)]^{1-\delta_i} \right) =$$

$$= \prod_{i=1}^{n} \left([\theta(\alpha,\beta;t_i,x_i)]^{\delta_i} [S(\alpha,\beta;t_i,x_i)] \right).$$

1.1.3 Stock sampling con follow-up

Nel contesto di analisi della durata della disoccupazione, nel quale ci muoviamo, i dati sono tipicamente raccolti attraverso l'osservazione dello stato occupato da uno stesso individuo in successive waves di un'indagine di tipo panel (Torelli e Paggiaro, 2002).

Nel presente lavoro, verrà preso in considerazione per semplicità il caso di indagini con due sole waves. Viene eseguito un campionamento da stock dalla popolazione di individui disoccupati. Gli individui campionati vengono intervistati una prima volta e viene registrata la durata del periodo che hanno già trascorso in stato di disoccupazione. Le stesse unità vengono ricontattate dopo un periodo di tempo k, tipicamente 3 mesi, e vengono interrogate riguardo allo stato nel quale si trovano in tale momento. In base alle risposte degli intervistati viene costruito l'indicatore di censura δ, con $\delta=0$ nel caso non sia ancora avvenuta una transizione allo stato di occupazione e $\delta=1$ in caso contrario. Questa particolare modalità di raccolta dati non permette di conoscere l'esatta durata della disoccupazione neanche per gli individui che non sono censurati a destra, ma soltanto che essa è compresa tra t_i e $t_i + k$, denotando con t_i la durata della permanenza nello

stato di disoccupazione dell'individuo i-esimo, pregressa alla prima intervista.

1.1.4 Funzione di verosimiglianza del modello

A partire del piano di osservazione definito in precedenza, e definiti α e β generici vettori di parametri, la funzione di verosimiglianza del modello può essere scritta come:

$$L(\alpha,\beta;x,t,\delta)=\prod_{i=1}^{n}\left(\left[1-\frac{S(t_i+k,x_i;\alpha,\beta)}{S(t_i,x_i;\alpha,\beta)}\right]^{\delta_i}\left[\frac{S(t_i+k,x_i;\alpha,\beta)}{S(t_i,x_i;\alpha,\beta)}\right]^{1-\delta_i}\right).$$

Da tale scrittura è possibile notare come le probabilità di transitare o non transitare vengano condizionate all'essere sopravvissuti in stato di disoccupazione fino al tempo t_i. Condizionare alle durate pregresse è una scelta obbligata dal campionamento da stock: tale condizionamento va cioè inserito nel modello per evitare il lenght-bias (Jenkins, 2004).

Sotto le assunzioni di rischi proporzionali e di forma Weibull per il rischio vale:

$$S(t,x)=\exp[-H_0(\alpha;t)\exp(x'\beta)],$$

con $H_0(\alpha;t)=\int_0^t\theta_0(\alpha;z)dz$ e $\theta_0(\alpha;t)=\alpha t^{\alpha-1}$, perciò:

$$\frac{S(t_i+k,x_i;\alpha,\beta)}{S(t_i,x_i;\alpha,\beta)}=\exp\left[\exp(x_i'\beta)\left(t_i^{\alpha}-(t_i+k)^{\alpha}\right)\right].$$

Dunque, la funzione di verosimiglianza del modello può venir scritta come:

$$L(\alpha,\beta;x,t,\delta)=\prod_{i=1}^{n}\left(P[\delta_i=1]\right)^{\delta_i}\left(1-P[\delta_i=1]\right)^{1-\delta_i}\quad(1),$$

con $P[\delta_i=1]=1-\exp\left[\exp(x_i'\beta)\left(t_i^{\alpha}-(t_i+k)^{\alpha}\right)\right]$.

1.2 Applicazione empirica con dati simulati

1.2.1 Simulazione dei dati con Stock Sampling

I campioni simulati per valutare i modelli trattati nel presente lavoro dovranno ricalcare ciò che avviene in caso di campionamento da stock per quanto riguarda le distribuzioni di probabilità delle variabili coinvolte, quali la variabile T, che rappresenta in questo caso la durata della disoccupazione pregressa al momento dell'intervista, le covariate X e l'indicatore di censura δ. Per una esaustiva trattazione teorica delle distribuzioni di probabilità delle variabili appena citate si veda Salant (1977).

Vengono inizialmente simulate le covariate X.

Assumendo che la probabilità di entrare nello stato di disoccupazione sia costante nel tempo antecedente la prima intervista, la funzione di densità della variabile T è proporzionale alla funzione di sopravvivenza della variabile durata totale D (Salant, 1977). Infatti, qualora un individuo entri nello stato di disocuppazione in un istante di tempo a distanza t dal momento dello studio, sarà osservabile con probabilità pari alla probabilità di rimanere nello stato di disoccupazione almeno per un periodo di lunghezza t, ovvero pari alla funzione di sopravvivenza calcolata in t.

Perciò, T è stata simulata tramite algoritmo di accettazione e rifiuto (Ripley, 1987), dalla distribuzione

$$f_T(t \mid x) \propto S_D(t \mid x) = \exp\left[-\exp(x'\beta)\left(t^{\alpha}\right)\right] \quad (2),$$

dove i valori x sono le covariate precedentemente simulate; il campionamento è stato in realtà effettuato da una distribuzione troncata rispetto alla precedente, in quanto i valori candidati sono stati simulati in

maniera uniforme tra 0 e una soglia massima di durata osservabile, che nel presente lavoro è stata fissata a 100 mesi. La soglia scelta è molto elevata, tale da rendere la distribuzione simulata decisamente simile a quella non troncata. Qualora, in ogni caso, non si desideri effettuare nessun troncamento, è possibile modificare l'algoritmo di accettazione e rifiuto, utilizzandone la versione per distribuzioni dal supporto non finito (Ripley, 1987): in questo caso va scelta una funzione di densità che, moltiplicata per un opportuno scalare, sovrasti $S_D(t \mid x)$ in ogni punto del suo supporto, quale ad esempio una variabile esponenziale con parametro opportunamente determinato.

L'indicatore di censura δ è una variabile di Bernoulli con valore del parametro dipendente dai valori di T ed X precedentemente simulati:

$$\delta \mid t, x \sim \text{Bern}\left(1 - \exp(x_i'\beta)\left(t_i^{\alpha} - (t_i + k)^{\alpha}\right)\right),$$

dove k è il tempo che intercorre tra le due interviste.

Un procedimento di simulazione nella pratica si è effettivamente utilizzato, in quanto più veloce, teoricamente del tutto equivalente rispetto a quello che è stato appena presentato.

Inizialmente per ogni unità sperimentale vengono simulate:

- le covariate X;
- la durata totale nello stato di disoccupazione, dalla distribuzione

$$D \mid t, x \sim \text{Weib}\left(\alpha, \frac{1}{\left(\exp(x'\beta)^{\frac{1}{\alpha}}\right)}\right);$$

- la variabile T rappresenta il lasso di tempo che intercorre tra il momento in cui l'individuo entra nello stato di disoccupazione ed il momento dell'intervista; in base a quanto precedentemente assunto, essa viene

quindi simulata in maniera uniforme tra 0 e la soglia massima di durata osservabile 100.

Ne consegue che l'individuo i-esimo venga tenuto nel campione soltanto qualora $d_i \geq t_i$, in quanto in caso contrario non sarebbe stato osservabile perché non più appartenente allo stock di disoccupati al momento dell'intervista. Questo procedimento è equivalente alla simulazione tramite algoritmo di accettazione e rifiuto dalla distribuzione $S_D(t \mid x)$; infatti per un individuo con $T=t_i$ la probabilità che la durata totale del periodo di disoccupazione sia maggiore di tale valore coincide con la funzione di sopravvivenza calcolata in t_i.

Infine, ad ogni unità mantenuta nel campione viene assegnato l'indicatore di censura nel seguente modo:

$\delta_i = 1$ se $t_i + k \geq d_i$,

$\delta_i = 0$ altrimenti.

Tale procedimento è equivalente alla simulazione da

$$\text{Bern}\left(1 - \exp(x_i'\beta)\left(t_i^{\alpha} - (t_i + k)^{\alpha}\right)\right).$$

Il parametro della variabile Bernoulli coincide, infatti, con la probabilità che la durata della disoccupazione non sia superiore a t_i +k, condizionata al fatto che essa sia maggiore di t_i, ovvero che l'individuo sia stato osservabile al momento dello studio.

1.2.2 Stima di massima verosimiglianza

Al fine di non dipendere eccessivamente dalle peculiarità di campioni di dimenensione ridotta, nel valutare i modelli in esame, si ricorre ad una congrua numerosità campionaria. Vengono difatti inizialmente simulati 3 campioni da 10.000 unità ciascuno, rispettivamente con dipendenza da

durata negativa, nulla e positiva: il logaritmo del parametro α assume rispettivamente valori –0,5, 0 e 0,25. Seguendo Torelli e Paggiaro (2002), i parametri β_0 sono stati fissati in modo tale che la durata media risulti non lontana da quella tipicamente osservata in Italia per la durata della disoccupazione. È stata inserita un'unica covariata X con distribuzione normale standard. Covariate discrete, in particolare dicotomiche, potrebbero rendere il parametro β di più difficile identificabilità, tuttavia nelle prove effettuate non sono emerse variazioni di rilievo rispetto al caso continuo.
La distanza tra le due successive interviste, k, è stata fissata a 3 mesi in analogia con ciò che succede in molti studi delle dinamiche del mercato del lavoro basati sulla rotazione del campione, in particolare nel contesto italiano.

Tabella 1.1. Modello Weibull a rischi proporzionali: stime di massima verosimiglianza su tre campioni simulati

parametro	vero valore	stima	se	Z value
		dipendenza negativa		
α	0,606	0,645	0,046	0,852
β_0	-2,500	-2,685	0,210	-0,881
β	1,000	1,093	0,066	1,412
		dipendenza nulla		
α	1,000	1,032	0,051	0,619
β_0	-4,000	-4,212	0,218	-0,971
β	1,000	1,084	0,062	1,347
		dipendenza positiva		
α	1,284	1,276	0,050	-0,156
β_0	-5,000	-4,918	0,207	0,395
β	1,000	1,024	0,053	0,450

Tramite la funzione di verosiglianza (1), si effettuano delle stime riportate nella Tabella 1.1 con i relativi errori standard. Inoltre, per verificare l'ipotesi di uguaglianza del parametro al suo vero valore, viene riportata

l'usuale statistica test $Z = \frac{\left(\hat{\vartheta} - \vartheta\right)}{se(\hat{\vartheta})}$, che si distribuisce asintoticamente come una normale standard per le proprietà degli stimatori di massima verosimiglianza.

Le stime dei parametri risultano complessivamente corrette e nessuna si distacca in maniera significativa dal vero valore del parametro.

Capitolo 2

Un approccio Bayesiano per l'analisi dei dati di durata

2.1 Modelli Bayesiani

La rappresentazione dell'incertezza riguardo ai parametri tramite distribuzioni di probabilità è il tratto distintivo dell'inferenza Bayesiana; il processo di apprendimento, nel contesto Bayesiano, consiste nell'aggiornamento delle opinioni iniziali riguardo al parametro ϑ (rappresentate dalla distribuzione $\pi(\vartheta)$) alla luce dei dati osservati: si ottiene una nuova distribuzione di probabilità per ϑ, $\pi(\vartheta \mid x)$, detta distribuzione a posteriori.

Questo procedimento viene effettuato tramite il teorema di Bayes, di cui riportiamo la versione generale,

$$\pi(\vartheta \mid x) = \frac{\pi(\vartheta) f(x \mid \vartheta)}{\int_{\Theta} \pi(\vartheta) f(x \mid \vartheta) d\vartheta},$$

con $f(x \mid \vartheta)$ funzione di densità di x dato ϑ.

In passato, affidarsi a metodi statistici di tipo Bayesiano era spesso scoraggiante a causa del peso computazionale richiesto dalla determinazione della densità a posteriori dei parametri, ad esempio per il calcolo dell'integrale a denominatore che essa richiede. Recentemente lo sviluppo di nuovi metodi di simulazione e tecniche computazionali ha rivoluzionato l'uso che viene fatto dell'inferenza Bayesiana nei più svariati campi di applicazione (Congdon, 2001).

Tale sviluppo permette ora di fruire maggiormente dei vari vantaggi offerti dalla statistica Bayesiana: offre una formalizzazione del processo di apprendimento che permette di ottenere dai dati un aggiornamento della conoscenza; non dipende da assunzioni di normalità asintotica, come spesso accade per i procedimenti di stima di massima verosimiglianza; fornisce come risultato del procedimento di stima l'intera distribuzione del parametro in esame e ciò permette di costruire intervalli di confidenza e test d'ipotesi in maniera più naturale di quanto accade nell'inferenza classica.

2.2 Markov Chain Monte Carlo

Nell'inferenza Bayesiana, soltanto in un numero ridotto di casi è possibile determinare analiticamente le distribuzioni a posteriori dei parametri. Il Markov Chain Monte Carlo è un metodo per simulare campioni da una generica distribuzione di probabilità; esso permette di simulare anche da distribuzioni multivariate e a meno di una costante di proporzionalità. Perciò si adatta alle esigenze della statistica Bayesiana, rendendo possibile simulare dalla distribuzione a posteriori, anche se i parametri sono numerosi e non è possibile determinare il valore dell'integrale posto a denominatore.

Il metodo si basa sulla costruzione di una catena markoviana che abbia come distribuzione limite la distribuzione desiderata, che in caso di applicazione dell'algoritmo in ambito bayesiano è $\pi(\vartheta \mid x)$. Gli elementi della serie simulata avranno dunque ciascuno distribuzione marginale $\pi(\vartheta \mid x)$; essi non saranno tuttavia indipendenti tra loro. Un'approssimazione delle quantità teoriche inerenti la distribuzione a posteriori dei parametri verrà poi fornita dalle corrispondenti statistiche empiriche della serie simulata.

Vi sono diverse varianti nell'algoritmo per costruire una catena ergodica con la distribuzione invariante desiderata f(x), dei quali nel seguito proporremo una breve sintesi; per una trattazione più completa si veda, ad esempio, Gilks, Richardson e Spiegelhalter (2006).

Con f(x) si intende la distribuzione di probabilità congiunta di un vettore aleatorio $x = (x_1,...,x_j,...,x_J)$. Tra gli algoritmi disponibili, il Gibbs Sampler richiede di conoscere, per ogni componente j del vettore, la distribuzione $f(x_j \mid x_{(j)})$, $x_{(j)} = (x_1,...,x_{j-1},x_{j+1},...,x_J)$, ovvero la sua distribuzione di probabilità condizionata alle componenti restanti.

Tale algoritmo, ad ogni passo della serie, simula ciclicamente ogni componente del vettore x dalle corrispondenti distribuzioni $f(x_j \mid x_{(j)})$. Non sempre, tuttavia, le funzioni di densità $f(x_j \mid x_{(j)})$ sono note o è possibile simulare in maniera agevole da esse.

In tali casi sono disponibili altri algoritmi, tra cui l'*adaptive rejection sampling*, lo *slice sampling*, l'*adaptive sampling*, il Metropolis-Hastings (Gamerman e Lopes, 2006).

Quest'ultimo algoritmo ad ogni passo della catena propone un nuovo valore simulandolo da una distribuzione di transizione arbitrariamente scelta

$h\left(x^* \mid x^{(t)}\right)$; viene accettato il valore proposto o tenuto il precedente con probabilità scelte in modo che la catena abbia distribuzione invariante f(x),

$$x^{(t+1)} = \begin{cases} x^* & \text{con probabilità } \alpha \\ x^{(t)} & \text{con probabilità } 1-\alpha \end{cases}$$

con $\alpha = \min\left\{1, f(x^*)h\left(x^{(t)} \mid x^*\right) / f\left(x^{(t)}\right)h\left(x^* \mid x^{(t)}\right)\right\}$.

Si dimostra che serie costruite tramite questi metodi, dopo un periodo iniziale necessario alla convergenza, detto burn-in, abbiano come distribuzione limite la distribuzione desiderata f(x) (Gamerman e Lopes, 2006).

2.3 Il software WinBUGS

WinBUGS è un software per l'analisi Bayesiana di complessi modelli statistici tramite metodo Markov Chain Monte Carlo (Spiegelhalter *et al.*, 2003). Il modello viene descritto attraverso uno specifico linguaggio testuale o, equivalentemente, attraverso un Directed Acyclic Graph (DAG). In questo caso, ogni quantità viene rappresentata con un nodo dal quale partono frecce verso i nodi da esso direttamente influenzati.

Il modello statistico si basa su assunzioni di indipendenza condizionata: si assume cioè che ciascun nodo, condizionatamente ai suoi nodi "genitori", sia indipendente da ogni altro nodo presente nel modello, eccetto i suoi nodi "discendenti" (Spiegelhalter *et al.*, 2003).

I nodi possono essere di tre tipi:

- costanti, quantità fissate e specificate in un apposito file di dati;

- stocastici, variabili aleatorie delle quali viene specificata la distribuzione. Possono essere osservate, nel qual caso si tratta di dati, o inosservate, nel qual caso di tratta di parametri;
- deterministici, funzioni non stocastiche di altri nodi.

I collegamenti fra due nodi possono indicare una dipendenza stocastica o non stocastica. Ogni DAG può essere rappresentato in maniera univoca tramite un linguaggio testuale, nel quale i primi collegamenti vengono indicati con ~, i secondi con <-.

L'assunzione di indipendenza condizionata di tali modelli rende la distribuzione congiunta facilmente fattorizzabile in termini di distribuzioni condizionate di ogni nodo rispetto ai propri nodi genitori; in altre parole, una volta fornite tutte le distribuzioni "figlio|genitore" il modello è completamente specificato e WinBUGS sceglie i metodi di simulazione direttamente sulla base delle relazioni espresse dalla struttura del modello.

Ove sia possibile, esso opta per un Gibbs Sampler; altrimenti sceglie l'adaptive rejection sampling in caso di densità log-concave. Vengono inoltre usati l'algoritmo Metropolis-Hastings, lo slice sampling o l'adaptive sampling (Congdon, 2003).

Il software sceglie, come periodo di burn-in della serie, la prima metà della serie; inoltre, è possibile effettuare un filtraggio, tenendo un valore ogni *n*: questo numero *n*, specificabile dall'utente, è detto *thinning interval*. È possibile, inoltre, specificare il numero di serie multiple che si vogliono simulare parallelamente ed i loro valori iniziali.

Nel presente lavoro verrà utilizzato il pacchetto "R2WinBUGS" del software R, che permette di gestire WinBUGS tramite R affinché esegua la stima di un modello, fornisca informazioni sulla convergenza della serie, e salvi quanto simulato in modo che sia facilmente accessibile da R per le ulteriori analisi (Sturtz, Ligges e Gelman, 2005).

2.3.1 Gli strumenti diagnostici del pacchetto "coda" di R

Sempre all'interno del software R, il pacchetto "coda" è finalizzato all'analisi di quanto prodotto tramite simulazioni Markov Chain Monte Carlo (Plummer *et al*., 2007). Tra i vari strumenti diagnostici proposti da tale pacchetto, andiamo ad elencarne alcuni, che verranno utilizzati nel seguito per l'analisi dell'output di WinBUGS.

La funzione "plot" propone alcuni strumenti grafici per una visualizzazione immediata delle distribuzioni a posteriori ottenute tramite simulazione: vengono visualizzate le serie simulate, tramite dei *plots*, uno per ogni parametro, che mostrano i valori simulati ad ogni iterazione; tali grafici permettono di avere intuitive ed immediate informazioni riguardo la convergenza della serie, la sua stazionarietà, l'entità dell'autocorrelazione. Inoltre, per ogni parametro, viene prodotta e visualizzata una stima della sua densità a posteriori a partire dalla serie simulata, tramite metodo Kernel (Silverman, 1986). Ciò permette, naturalmente, di avere un'immediata percezione della distribuzione a posteriori marginale per ogni parametro, e quindi delle stime a posteriori degli stessi.

Il metodo "summary" fornisce invece per la distribuzione a posteriori di ciascun parametro, varie statistiche riassuntive quali media, standard deviation e quantili; vengono prodotti per default i quantili 0,025, 0,25, 0,50, 0,75 e 0,975.

La funzione "autocorr" calcola l'autocorrelazione di ciascuna serie per i ritardi 0, 1, 5, 10, 50; nel caso di filtraggio della serie, i valori dei ritardi vengono presi relativamente al *thinning interval*. I valori simulati tramite MCMC sono, come detto in precedenza, tra loro dipendenti. Il valore dell'autocorrelazione offre una valutazione di quanto ogni nuovo valore

simulato dalla serie dipenda dai valori registrati nelle iterazioni precedenti: autocorrelazioni particolamente elevate suggeriscono che la serie, una volta assunto un determinato valore, rimarrà tendenzialmente per molte iterazioni intorno a tale valore, essendo tali iterazioni pesantemente influenzate dalle precedenti.

In tali condizioni di notevole lentezza della serie, anche raggiungere la convergenza può occupare un numero molto elevato di iterazioni: qualora si registrino alti valori nella tabella di autocorrelazione è dunque opportuno aumentare il numero di iterazioni previste; inoltre si può porre un *thinning interval* tale che la catena conservi la quasi totalità della sua informazione, ma occupi in memoria uno spazio notevolmente ridotto.

Tramite "crosscorr" viene calcolata la correlazione tra ciascuna coppia di serie, permettendo di valutare la dipendenza tra le stime dei parametri.

Per valutare la convengenza della catena di Markov, Geweke (1992) propone un test diagnostico basato sull'uguaglianza delle medie tra la prima e l'ultima parte della catena. Se la serie è stazionaria, la statistica di Geweke Z ha un distribuzione asintotica normale standard:

$$Z = \frac{\left(\bar{\theta}_a - \bar{\theta}_b\right)}{\left(V_a + V_b\right)^{\frac{1}{2}}},$$

dove $\bar{\theta}_a$ e $\bar{\theta}_b$ sono le medie del parametro rispettivamente per le prime n_a iterazioni e per le ultime n_b iterazioni, e V_a e V_b sono le loro rispettive varianze (Congdon, 2003). La funzione "geweke.diag" del pacchetto "coda" pone per default n_a uguale a 10% delle iterazioni ed n_b uguale al 50%. Tali valori sono naturalmente modificabili, ma va tenuto conto del fatto che il test viene effettuato sotto l'assunzione che le due parti della catena siano asintoticamente indipendenti; quindi $n_a + n_b$ deve essere mantenuto strettamente inferiore al totale delle iterazioni della catena.

Un ulteriore strumento per valutare la convergenza delle serie viene proposto da Gelman e Rubin (1992) ed implementato dal pacchetto "coda" tramite la funzione "gelman.diag". Questa diagnostica può essere calcolata soltanto su serie multiple simulate parallelamente, in quanto lo Scale Reduction Factor va a confrontare la variabilità dei parametri simulati tra le catene con la variabilità all'interno di ciascuna catena. Infatti, se i modelli sono particolarmente complessi o scarsamente identificabili, ci si aspetta una grande divergenza nel cammino delle differenti catene; la varianza di ogni singola catena sarà pertanto considerevolmente inferiore alla varianza tra le catene (Congdon, 2003). L'avvenuta convergenza verrà quindi diagnosticata quando le serie non avranno più memoria dei loro valori iniziali, scelti sovradispersi rispetto alla distribuzione a posteriori, ed i loro valori risulteranno tra loro indistiguibili.

Il metodo infatti si basa sul calcolo della seguente quantità:

$$\hat{\sigma}^2 = (n-1)\frac{W}{n} + \frac{B}{n},$$

dove n è il numero delle iterazioni, B/n la varianza empirica tra le catene e W la media delle varianza empirica di ciascuna catena. La statistica usata come diagnostica della convergenza è:

$$R = \sqrt{\frac{(d+3)\hat{V}}{(d+1)W}},$$

$$\hat{V} = \hat{\sigma}^2 + \frac{B}{nm},$$

dove d sono i gradi di libertà stimati tramite il metodo dei momenti.

Valori sostanzialmente maggiori di 1 indicano mancanza di convergenza. La funzione consente anche il calcolo del limite superiore dell'intervallo di confidenza di R; tuttavia è necessaria cautela nel valutare tale limite in

quanto il suo calcolo si basa sull'assunzione che la variabile in esame sia normale.

Brooks e Gelman (1997), inoltre, propongono una versione multivariata della diagnostica: anche questa viene calcolata tramite la funzione "gelman.diag" di R, e permette di valutare la convergenza delle serie relative a ciascun parametro del modello nel loro complesso.

Infine, la funzione "HPDinteval" costruisce, a partire dalle serie simulate, un intervallo High Posterior Density (HPD) per ciascun parametro con livello di credibilità scelto dall'utente: a partire dalla $\pi(\vartheta \mid x)$, cioè, la funzione costruisce un intervallo di credibilità includendovi i valori di ϑ a cui corrisponde una densità a posteriori più elevata.

2.4 Specificazione del modello di durata in ambito Bayesiano

Il problema di analisi di dati di durata, ampiamente descritto nel capitolo precedente, può venir affrontato anche attraverso metodi d'inferenza di tipo Bayesiano. Il modello Bayesiano risultante viene definito dalla distribuzione a priori $\pi(\vartheta)$ dei parametri del modello, $\vartheta = (\alpha, \beta)$, e dalla verosimiglianza $f(x \mid \vartheta)$, che coincide con la verosimiglianza (1) ottenuta nel paragrafo 1.1.4.

2.4.1 Scelta delle distribuzioni a priori

Per la distribuzione a priori $\pi(\vartheta)$ è opportuno effettuare delle scelte molto accurate, in modo tale che le distribuzioni di probabilità rispecchino

effettivamente eventuali conoscenze pregresse sui parametri e siano funzionali ad un corretto processo di stima dei parametri.

Adottare distribuzioni a priori eccessivamente concentrate su alcuni punti dello spazio parametrico può indurre le stime a posteriori dei parametri a convergere proprio su tali punti, indipendentemente dalla vera distribuzione del parametro. Al contrario, scegliere distribuzioni scarsamente informative, con varianze eccessive, può andare a discapito dell'identificabilità del modello e consentire al processo di stima di scegliere valori appartenenti allo spazio parametrico ma decisamente poco ragionevoli per il modello e la popolazione in esame.

Le distribuzioni dei parametri vengono ipotizzate tra loro indipendenti, perciò la distribuzione a priori $\pi(\vartheta)$ è il prodotto delle distribuzioni a priori di ogni singola componente del vettore.

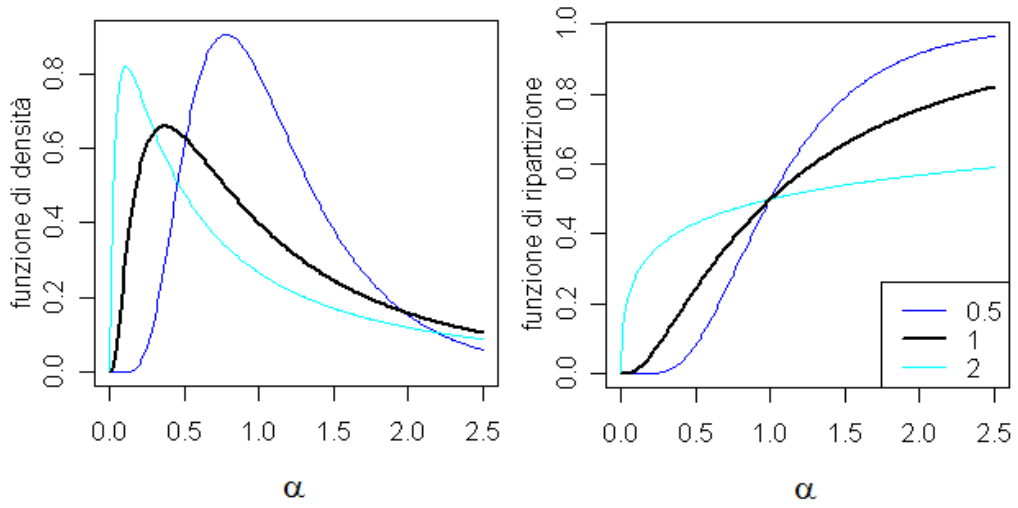
Per ogni parametro vengono prese in considerazione alcune distribuzioni tra quelle con supporto pari al suo spazio campionario. In ciascuna di tali famiglie di distribuzioni vengono individuati gli elementi più opportuni per rappresentare le opinioni a priori sul parametro e ottenere stime corrette; le distribuzioni proposte vengono poi effettivamente usate in processi di stima, al fine di escludere quelle che eventualmente presentano comportamenti anomali.

Il parametro α definisce la dipendenza tra la durata nello stato di disoccupazione e la probabilità di uscire da tale stato; ricordiamo che valori minori di 1 indicano una dipendenza negativa; valori maggiori di 1 indicano dipendenza positiva; il valore spartiacque corrisponde all'assenza di dipendenza e quindi ad una funzione di rischio costante. Tra le distribuzioni che abbiano come supporto il semiasse positivo, vengono proposte la lognormale e la Gamma. Si decide inoltre di fissare a 1 un indice di posizione opportunamente scelto.

Quanto alla prima distribuzione proposta, è dunque opportuno che μ, la media della variabile normale, abbia valore 0, in modo che la sua trasformata esponenziale abbia mediana 1. Il parametro σ^2, varianza della variabile normale, può essere scelto in base all'informatività che si vuole dare alla distribuzione a priori.

Nella Figura 2.1 sono confrontate le funzioni di ripartizione della distribuzione con vari valori di σ^2.

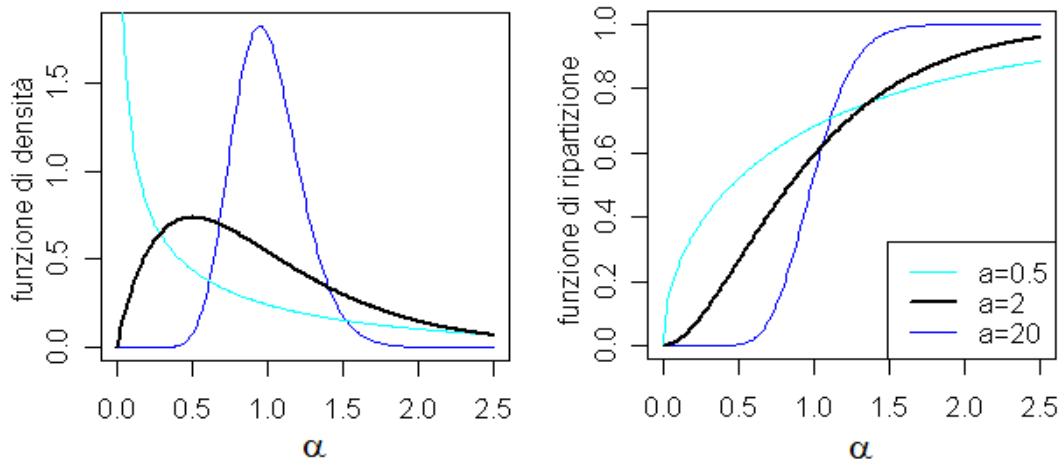
Figura 2.1. Funzioni di densità e ripartizione di una variabile lognormale, con vari valori per la varianza sulla scala logaritmica (riportati in legenda)



Le variabili aleatorie Gamma, con $f(x) = \left(\frac{s^a}{\Gamma(a)} \right) x^{a-1} e^{-sx}$, hanno valore atteso pari ad $\frac{a}{s}$ e varianza pari ad $\frac{a}{s^2}$. Inoltre la moda è pari a 0 qualora a sia minore di 1, e $\frac{a-1}{s}$ altrimenti.

Si decide dunque di fissare $a = s$ in modo da ottenere una distribuzione che abbia media 1. Valori di a vicini ad 1, renderebbero la moda uguale o vicina a 0, ma al crescere di a decresce la varianza. Mostriamo quanto appena esposto nella Figura 2.2 che confronta vari valori di a, con $a = s$.

Figura 2.2. Funzioni di densità e ripartizione di una variabile Gamma, con vari valori per il parametro a (riportati in legenda), e s=a



Si sono effettuate analisi preliminari con entrambe le distribuzioni, la prima con varianza pari a 1, la seconda con a=2. Le due distribuzioni fanno pervenire a risultati del tutto equivalenti. Nel seguito verrà usata soltanto la distribuzione a priori $\Gamma(2,2)$, in quanto essa assegna minor probabilità a valori di particolarmente grandi e poco ragionevoli nel contesto in esame (si può facilmente vedere dal valore della funzione di ripartizione delle due distribuzioni nel punto 2,5, dai grafici 2.1 e 2.2).

Per i parametri β è opportuno adottare distribuzioni definite sull'intero asse reale e simmetriche. Vengono quindi scelte delle distribuzioni normali tra loro indipendenti; la media viene fissata a 0, tranne per l'intercetta β_0 il cui valore atteso viene posto a -4, in modo da ottenere valori medi delle durate coerenti con quelli osservati nella realtà. La scelta della varianza, come di consueto, va effettuata in base a considerazioni sull'informatività delle a priori.

Scelte diverse riguardo a tali varianze, ad esempio 4 o 1, non sembrano produrre alcuna differenza significativa nella stima dei parametri. Nel seguito, verrà posta varianza 4 per β_0 e 1 per i generici parametri β.

2.4.2 Specificazione del modello

La distribuzione a posteriori dei parametri $\pi(\vartheta|x) \propto \pi(\vartheta) f(x|\vartheta)$ a causa della complessità del modello, non è analiticamente determinabile. Di conseguenza, verranno usati i metodi Markov chain Monte Carlo, attraverso il software WinBUGS.

Il modello statistico Bayesiano deve essere specificato al software WinBUGS graficamente o nel corrispondente linguaggio testuale; entrambe le specificazioni verranno riportate in appendice A. Indipendentemente dal linguaggio utilizzato, comunque, il modello va specificato attraverso distribuzioni di probabilità condizionate. Esso, inoltre, si basa su assunzioni di indipendenza condizionata: si assume che ciascun nodo, condizionatamente ai suoi nodi "genitori", sia indipendente da ogni altro nodo presente nel modello, eccetto i suoi nodi "discendenti".

La variabile δ_i è dunque, per ogni unità i, la realizzazione di una variabile Bernoulli con probabilità P_i, funzione deterministica dei parametri e delle variabili t_i e x_i, che nel modello sono costanti; i vettori δ_i, t_i e x_i vengono forniti infatti al programma come dati.

Le probabilità P_i vengono calcolate come $1 - \exp\left(\exp(x_t'\beta)\left(t_i^{\alpha} - (t_i + k)^{\alpha}\right)\right)$, dati gli assunti di rischi proporzionali e distribuzione di tipo Weibull, con α, β_0 e β parametri; essi sono infatti nodi stocastici con distribuzioni a priori precedentemente specificate.

Di conseguenza, il modello viene specificato attraverso le seguenti distribuzioni di probabilità:

$\alpha \sim \Gamma(2,2)$,

$\beta_0 \sim N(-4,4)$,

$$\beta \sim N(0,1),$$

$$\delta_i \sim \text{Bern}\left(1 - \exp\left(\exp(x_i'\beta)\left(t_i^{\alpha} - (t_i + k)^{\alpha}\right)\right)\right).$$

2.5 Stima del Modello Bayesiano sui campioni simulati

Come brevemente esposto nel paragrafo 2.1, nell'ambito della statistica Bayesiana il parametro ϑ è aleatorio, a differenza della statistica classica nella quale ϑ è un valore fissato. Tuttavia, nel seguito ci si riferirà al "vero valore" del parametro, intendendo con tale termine il valore del parametro con il quale sono stati simulati i dati.

Il modello Bayesiano è stato applicato, tramite il software WinBUGS, sugli stessi tre campioni su cui erano state effettuate stime di massima verosimiglianza (paragrafo 1.2.2). Sono state simulate tre catene parallele di 20.000 iterazioni ciascuna. Per valutare la sensibilità del metodo di stima ai valori iniziali, le tre catene sono state fatte partire dai quantili rispettivamente 0,025, 0,50 e 0,975 di una catena, simulata precedentemente, che consta di 5000 iterazioni (Congdon, 2003). Come previsto per default da WinBUGS, la prima metà delle iterazioni, ovvero le prime 10.000, funge da periodo di burn-in; inoltre, è stato scelto un *thinning interval* di 5.

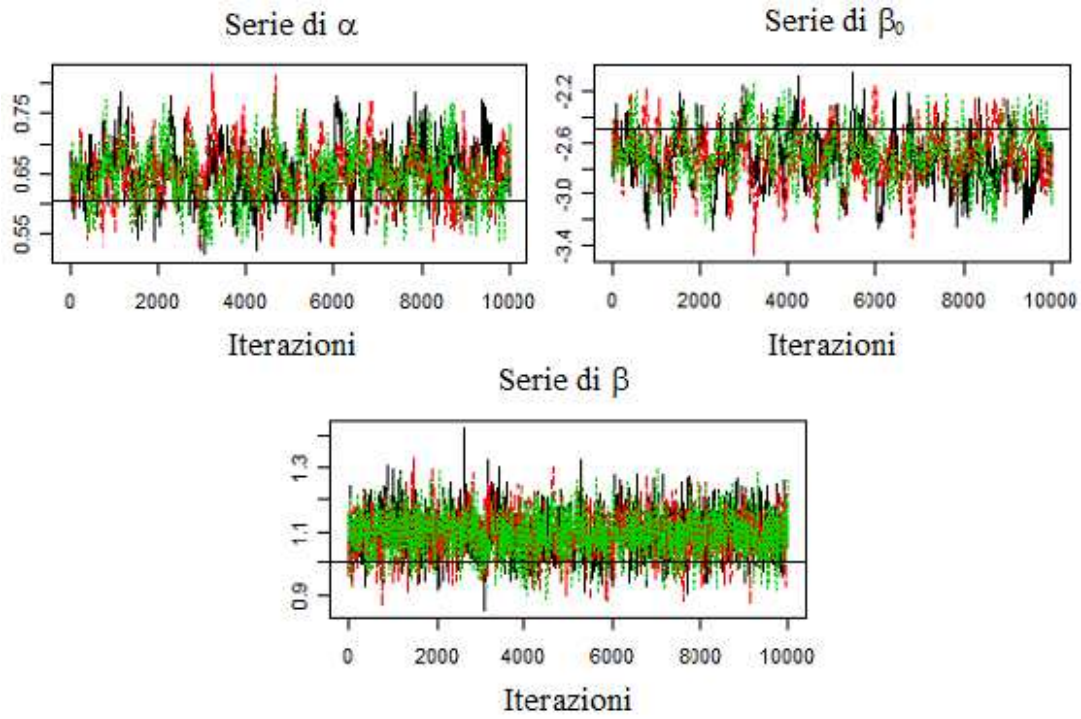
2.5.1 Problematiche di convergenza ed autocorrelazione

Attraverso il pacchetto "coda" sono state valutate varie caratteristiche delle serie ottenute quali la convergenza, la stazionarietà, l'autocorrelazione, nonché la variabilità e la correttezza delle stime.

La Figura 2.3 mostra, per ciascun parametro, le tre serie simulate, rappresentate tramite diversi colori ed i "veri valori" tramite una retta nera, limitatamente al campione relativo alla dipendenza da durata negativa; gli altri campioni, difatti, presentano un analogo comportamento relativamente alla convergenza e ad alla stazionarietà, quindi i risultati che presenteremo e commenteremo di seguito sono ad essi estendibili.

Si noti come le serie dei parametri α e β_0 presentino generalmente una maggior lentezza nelle loro oscillazioni rispetto alle serie dei parametri β; il grafico sembrerebbe suggerire che le serie necessitino di numerose iterazioni per spostarsi all'interno dello spazio parametrico: dunque ci aspettiamo autocorrelazioni elevate per i parametri α e β_0.

Figura 2.3. Weibull con dipendenza da durata negativa: tre catene MCMC parallele per la simulazione delle distribuzioni a posteriori dei parametri α, β e β_0



Tuttavia, per ogni parametro, le tre serie parallele simulate non divergono su regioni diverse dello spazio campionario, ma variano all'interno di un

comune intervallo di valori; questa indipendenza rispetto ai punti iniziali, seppur molto distanti tra loro, suggerisce che sia stata raggiunta la convergenza delle serie.

Si noti inoltre come in alcuni punti le serie di α e β_0 sembrino speculari tra loro; ad esempio a picchi positivi dell'una corrispondono picchi negativi nell'altra. Tale fenomeno indica una forte dipendenza negativa tra le due serie. Le Tabelle 2.1 e 2.2 mostrano le autocorrelazioni e crosscorrelazioni delle serie.

I valori presenti in tali Tabelle confermano quanto ipotizzato tramite l'osservazione dei grafici precedentemente riportati: le serie dei parametri α e β_0 presentano autocorrelazioni a ritardo 5 decisamente alte, prossime ad 1.

Tabella 2.1. Weibull con dipendenzada durata negativa: autocorrelazioni a vari ritardi delle catene MCMC

Ritardo	α	β_0	β
5	0,847	0,897	0,270
25	0,578	0,604	0,101
50	0,351	0,368	0,049
250	-0,009	-0,007	0,016

Tabella 2.2. Weibull con dipendenza da durata negativa: crosscorrelazioni tra le catene MCMC

	α	β_0	β
α	1	-0,963	0,330
β_0		1	-0,398
β			1

Ciò indica una forte dipendenza di ciascun valore simulato dal precedente valore della serie. Tali autocorrelazioni continuano ad assumere valori positivi e non prossimi allo 0 anche dopo un numero considerevole di ritardi, come 25 e 50. Le autocorrelazioni si annullano, comunque, dopo 250 ritardi: il contrario avrebbe indicato un comportamento patologico o comunque

problematico della catena, tale da compromettere la convergenza della serie. Anche le serie relative al parametro β presentano delle autocorrelazioni positive, ma di modesta entità.

Passando alla matrice delle crosscorrelazioni, immediatamente si nota il valore della correlazione tra le serie di α e β_0: esso è prossimo a –1 ed indica una dipendenza negativa molto forte tra le due serie, come era stato precedentemente ipotizzato dall'andamento tra loro speculare. Le serie, in altre parole, sono tra loro pressoché proporzionali ed i valori assunti dall'una sono fortemente influenzati dai valori assunti dall'altra. La serie del parametro β è positivamente correlata con quella del parametro α, e negativamente con quella del parametro β_0, ma i valori di tali correlazioni non sono particolarmente elevati.

Crosscorrelazione e autocorrelazioni per α e β_0 sono due fenomeni tra loro legati e sono un segnale della difficile identificabilità dei due parametri. È questa una caratteristica non specifica del campione in esame o del tipo di dipendenza da durata, ma intrinseca del modello Weibull; gli indici esaminati assumono valori estremamente simili in tutti i campioni presi in considerazione. I due parametri concorrono infatti congiuntamente alla determinazione del valore medio della durata; essi sono quindi tra loro legati, in quanto devono assumero valori tali da riprodurre la media della durata osservata.

Esaminiamo ora le diagnostiche per la convergenza di Geweke e di Gelman e Rubin, descritti nel paragrafo 2.3.1, riportati nelle Tabelle 2.3 e 2.4.

Tabella 2.3. Weibull con dipendenza da durata negativa: statistica di Geweke su ciascuna catena parallela

Serie	α	β_0	β
1	-1,454	1,400	0,794
2	-1,597	1,641	-1,041
3	-0,631	0,615	-0,599

Tabella 2.4. Weibull con dipendenza da durata negativa: statistica di Gelman e Rubin e limite superiore del relativo intervallo di confidenza a livello 0,95

	α	β_0	β
Stima puntuale	1,01	1,01	1,00
Quantile 0,975	1,02	1,03	1,01

psrf multivariato: 1,01

La statistica test Z di Geweke viene proposta per le tre serie parallele di ogni parametro; per nessuna di esse viene rifiutata, a livello di significatività 0,05, l'ipotesi di stazionarietà. Anche la statistica R di Gelman e Rubin assume valori prossimi ad 1 in ciascun parametro e nella versione multivariata: le due diagnostiche sembrano quindi indicare che le serie MCMC abbiano raggiunto la convergenza.

2.5.2 Distribuzioni a posteriori stimate

Appurata dunque la convergenza delle serie, è possibile esaminare le stime delle distribuzioni a posteriori ottenute tramite simulazione. Per ogni campione, verranno mostrate nelle Figure 2.4, 2.5 e 2.6 le stime delle densità a posteriori dei tre parametri, ottenute dalle serie simulate tramite metodo Kernel. Inoltre, verranno riportati media, standard deviation, quartili e quantili 0,025 e 0,975 delle serie simulate nella Tabella 2.6.

Il modello sembra avere un comportamento globalmente buono, in quanto le distribuzioni a posteriori si scostano nettamente dalle distribuzioni a priori, nella direzione dei "veri valori" dei parametri.

Si noti inoltre come generalmente le distribuzioni a posteriori siano approssimativamente simmetriche e normali; ulteriore conferma se ne ha dalla notevole vicinanza tra i media e mediana di ciascuna distribuzione.

Figura 2.4 Weibull con dipendenza da durata negativa: stime delle distribuzioni a posteriori dei parametri α, β e β_0

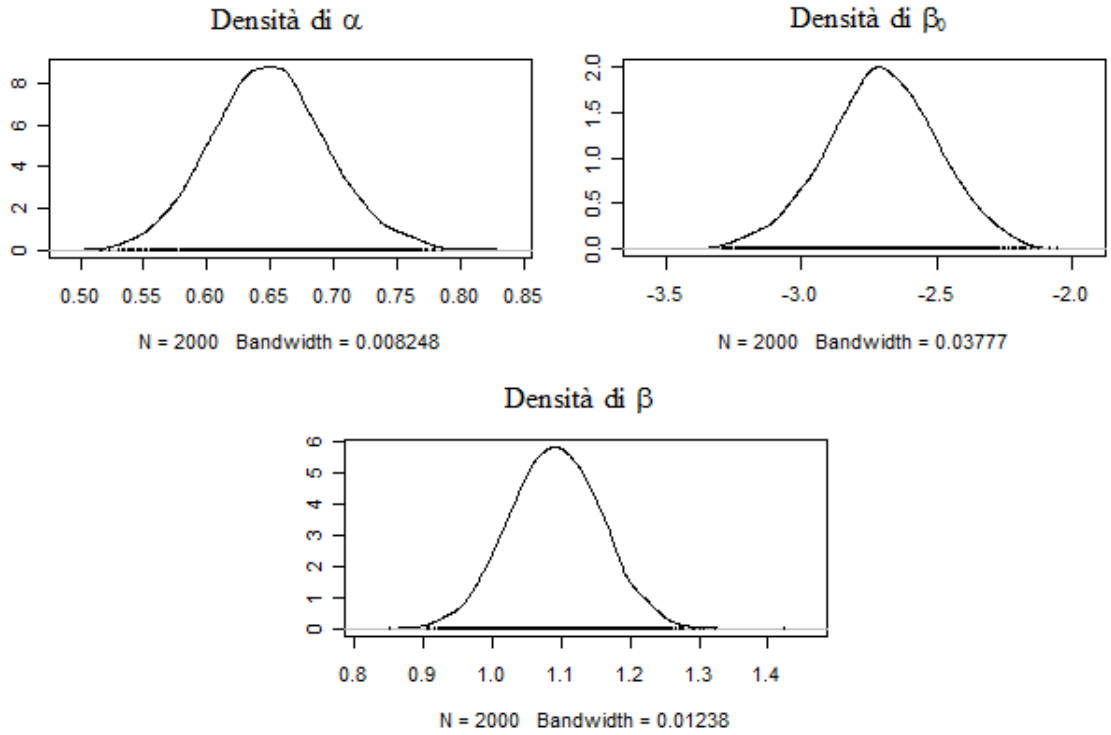


Figura 2.5 Weibull con dipendenza da durata nulla: stime delle distribuzioni a posteriori dei parametri α, β e β_0

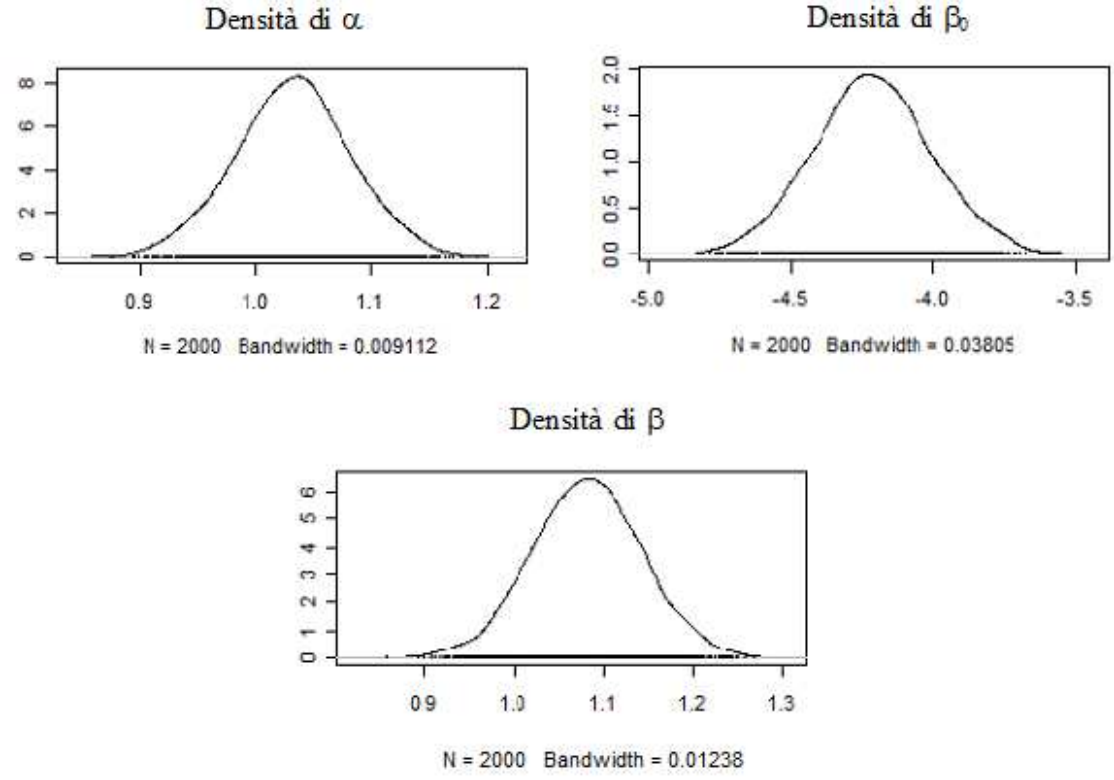


Figura 2.6 Weibull con dipendenza da durata positiva: stime delle distribuzioni a posteriori dei parametri α, β e β_0

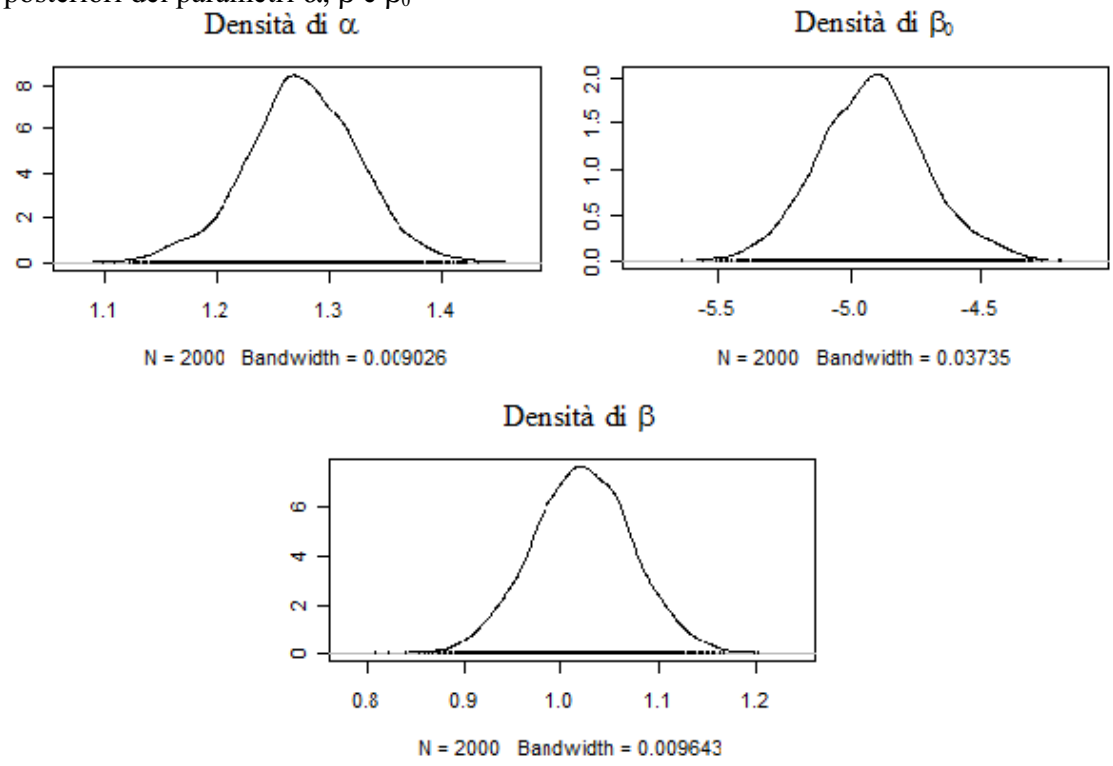


Tabella 2.5. Modello Weibull a rischi proporzionali: stime di massima verosimiglianza (1) su tre campioni simulati

parametro	vero valore	stima	se	Z value
dipendenza negativa				
α	0,606	0,645	0,046	0,852
β_0	-2,500	-2,685	0,210	-0,881
β	1,000	1,093	0,066	1,412
dipendenza nulla				
α	1,000	1,032	0,051	0,619
β_0	-4,000	-4,212	0,218	-0,971
β	1,000	1,084	0,062	1,347
dipendenza positiva				
α	1,284	1,276	0,050	-0,156
β_0	-5,000	-4,918	0,207	0,395
β	1,000	1,024	0,053	0,450

Tabella 2.6. Modello Weibull con diverse tipolodie di dipendenza da durata: indici di sintesi delle distribuzioni a posteriori stimate sui tre campioni in esame

parametro	vero valore	Media	SD	2,5%	25%	50%	75%	97,5%
dipendenza negativa								
α	0,606	0,649	0,045	0,563	0,619	0,648	0,678	0,742
β_0	-2,500	-2,704	0,205	-3,125	-2,837	-2,702	-2,565	-2,307
β	1,000	1,092	0,067	0,962	1,047	1,092	1,137	1,224
dipendenza nulla								
α	1,000	1,032	0,049	0,934	0,999	1,032	1,065	1,129
β_0	-4,000	-4,215	0,207	-4,623	-4,353	-4,217	-4,079	-3,800
β	1,000	1,082	0,061	0,964	1,042	1,083	1,123	1,202
dipendenza positiva								
α	1,284	1,275	0,050	1,169	1,244	1,275	1,309	1,373
β_0	-5,000	-4,915	0,206	-5,322	-5,055	-4,915	-4,786	-4,478
β	1,000	1,021	0,052	0,920	0,986	1,022	1,056	1,123

Scegliendo la media della distribuzione a priori come stima puntuale dei parametri, si verifica che tali stime sono pressoché coincidenti con quelle di massima verosimiglianza ottenute sugli stessi campioni (paragrafo 1.2.2); la notevole numerosità campionaria, infatti, fa sì che sulla distribuzione a posteriori la funzione di verosimiglianza abbia un'influenza decisamente più marcata rispetto alla distribuzione a priori. Si vedrà comunque come questo avvenga anche con dimensioni più limitate.

I quantili 0,025 e 0,975 possono essere visti come estremi di un intervallo di credibilità con probabilità 0,95 per ciascun parametro. Alternativamente, è possibile costruire un intervallo tramite approssimazione normale, utilizzando media e deviazione standard delle serie. Gli intervalli costruiti attraverso questi due metodi sono comunque estremamente simili, ad ulteriore conferma dell'approssimativa normalità delle distribuzioni a posteriori. Inoltre, è possibile costruire degli intervalli HPD ma essi, sempre a causa della simmetria e dell'approssimativa normalità delle distribuzioni, risulterebbero assolutamente equivalenti a quelli costruiti tramite quantili. In ciascun campione esaminato, i "valori veri" dei parametri appartengono agli

intervalli di credibilità, indipendentemente da come questi siano stati costruiti.
Si noti attraverso l'ampiezza degli intervalli ed i valori delle deviazioni standard, come la variabilità delle stime sia piuttosto ridotta; tale variabilità è, in tutti gli esempi mostrati, analoga a quella delle stime ottenute tramite metodo di massima verosimiglianza.

2.6 Stime ripetute su insiemi di 100 campioni

I campioni sui quali è appena stato applicato il modello Bayesiano sono particolarmente numerosi, constando di 10.000 unità, tuttavia qualche caratteristica rilevata nelle stime appena riportate potrebbe dipendere dallo specifico campione in esame.
Per avere una valutazione più generale e più obiettiva del modello in esame, si è ricorsi alla simulazione di un grande numero di campioni: questo lavoro di simulazione fa parte di un disegno sperimentale più vasto, che prevede anche la valutazione di altri modelli. In questo paragrafo, tuttavia, verrà descritto il metodo ed i risultati sono riportati limitatamente al modello tuttora in esame.
Sono stati simulati 100 campioni, di numerosità pari a 1000 ciascuno, con una sola covariata X con distribuzione normale e tempo k tra le due successive interviste pari a 3 mesi; per i metodi usati per la simulazione si veda il paragrafo 1.2.1. Su ciascun campione simulato viene applicato il modello in esame tramite WinBUGS producendo per ciascun parametro una catena da 20.000 iterazioni; per il burn-in è stata utilizzata la prima metà delle iterazioni; il thinning interval è stato posto a 5.

Ogni serie consta quindi di 2000 valori, dei quali sono stati calcolati media e quartili; per ogni parametro è stato inoltre creato un intervallo di credibilità HPD a livello 0,8 ed è stata riportata una variabile dicotomica che indica l'effettiva appartenenza del "vero valore" del parametro a tale intervallo. Per ciascuno dei 100 campioni, inoltre, sono state valutate la crosscorrelazione e le autocorrelazioni a ritardo 5 dei parametri α e β_0, al fine di valutare la situazione di problematica identificabilità dei due parametri emersa dall'esame dei precedenti campioni.
Questo procedimento è stato ripetuto per tre insiemi di parametri relativi a tipologie di dipendenza da durata differenti: tali insiemi di parametri sono gli stessi scelti per le precedenti analisi.

2.6.1 Stima dei parametri

Una prima valutazione della distribuzione delle stime attraverso i campioni viene effettuata tramite un metodo grafico come il box-plot ed usando come stima puntuale dei parametri il valore medio della relativa distribuzione a posteriori simulata. Si riportano nella Figura 2.7 quelli relativi alla situazione di dipendenza da durata negativa.
Tali distribuzioni appaiono approssimativamente simmetriche e centrate intorno al vero valore del parametro; vari elementi, quali la distanza interquartile, la lunghezza dei "baffi" del box-plot, nonché la presenza di singoli campioni nei quali le stime si rilevano decisamente anomale, mostrano un'elevata variabilità nelle distribuzioni.
Delle quantità rilevate su ciascun campione vengono calcolate media e deviazione standard, riportate nella Tabella 2.7.

Figura 2.7. Boxplot relativi a stime puntuali (media campionaria della distribuzione a posteriori) dei parametri α(=0,606), β_0(=-2,5) e β(=1) per 100 campioni

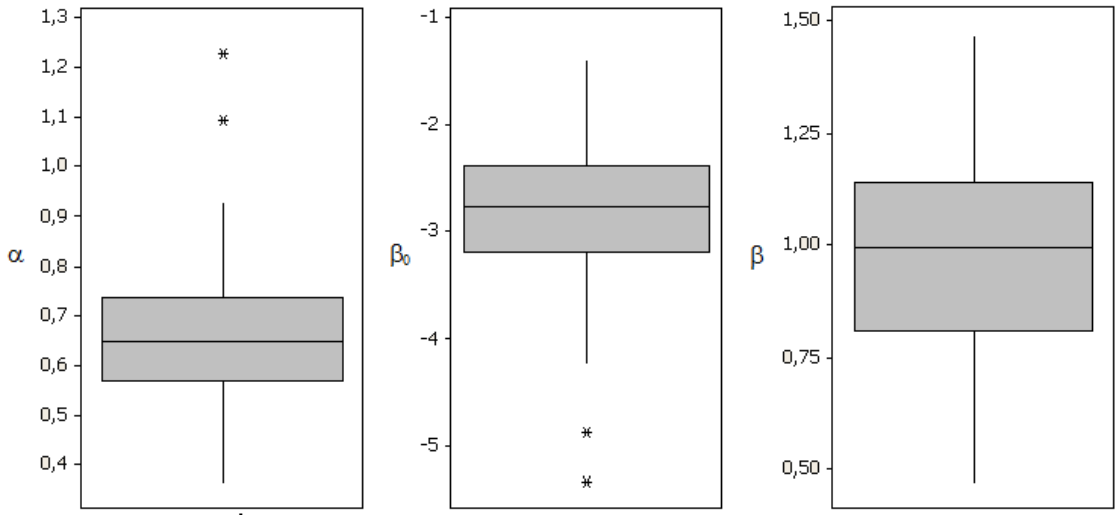


Tabella 2.7. Media e standard deviation di indicatori delle distribuzioni a posteriori per gruppi di 100 campioni con diversa tipologia di dipendenza da durata

	α		β_0		β	
	media	**sd**	**media**	**sd**	**media**	**sd**
			dipendenza da durata negativa			
vero valore	*0,606*		*-2,500*		*1,000*	
media	0,668	0,146	-2,838	0,673	0,991	0,219
I quartile	0,573	0,141	-3,273	0,688	0,845	0,215
mediana	0,666	0,146	-2,839	0,670	0,990	0,219
III quartile	0,761	0,152	-2,405	0,654	1,136	0,223
copertura	0,790		0,780		0,770	
			dipendenza da durata nulla			
vero valore	*1,000*		*-4,000*		*1,000*	
media	1,032	0,129	-4,168	0,554	0,940	0,180
I quartile	0,931	0,123	-4,587	0,582	0,814	0,178
mediana	1,030	0,129	-4,160	0,552	0,939	0,180
III quartile	1,131	0,134	-3,740	0,528	1,064	0,183
copertura	0,840		0,820		0,800	
			dipendenza da durata positiva			
vero valore	*1,284*		*-5,000*		*1,000*	
media	1,266	0,141	-4,982	0,599	0,990	0,176
I quartile	1,161	0,137	-5,405	0,625	0,873	0,172
mediana	1,262	0,141	-4,969	0,596	0,989	0,176
III quartile	1,368	0,146	-4,547	0,574	1,107	0,180
copertura	0,780		0,740		0,790	

Come rilevato dalla Figura 2.7, le stime appaiono in media corrette, tuttavia la loro variabilità tra i campioni è notevole, come si può notare dalle standard deviation della media, per ogni parametro. La variabilità delle distribuzioni a posteriori all'interno di ciascun campione, non sembra in media particolarmente elevata, come si evince osservando i quartili. Quanto agli intervalli HPD si riporta la media della variabile dicotomica indicante l'appartenenza del "vero valore" del parametro al relativo intervallo, che quindi corrisponde alla proporzione di campioni in cui ciò avviene. L'intervallo viene costruito a livello 0,8, e le proporzioni risultano difatti sempre molto vicine a tale valore.

2.6.2 Autocorrelazione delle serie

Le autocorrelazioni a ritardo 5 di α e β_0, riportate nella Tabella 2.9, sono molto alte, prossime all'unità, ma hanno anche una variabilità estremamente ridotta; ciò conferma questa caratteristica come propria del modello in esame e costante al variare dei campioni in esame. La stessa osservazione può essere fatta riguardo la forte dipendenza negativa tra le due serie di parametri.

Tali caratteristiche sembrano accentuarsi leggermente al crescere del parametro α.

Tabella 2.8. Medie e sd dei valori di autocorrelazione a ritardo 5 e crosscorrelazione delle serie di α e β_0, su gruppi di 100 campioni con varie tipologie di dipendenza da durata

	negativa		nulla		positiva	
	media	sd	media	sd	media	sd
autocorrelazione α	0,841	0,021	0,873	0,019	0,891	0,016
autocorrelazione β_0	0,891	0,014	0,911	0,014	0,924	0,011
crosscorrelazione α β_0	-0,960	0,006	-0,969	0,005	-0,974	0,004

Per captare eventuali comportamenti problematici della catena è importante considerare anche le autocorrelazioni a ritardi maggiori di 5; è stato dunque costruito, per ogni campione, un indice come numero di autocorrelazioni che superano le seguenti soglie: 0,70 per le autocorrelazioni a ritardo 25, 0,55 per le autocorrelazioni a ritardo 50 e 0,20 per le autocorrelazioni a ritardo 250.

Si rileva che nel caso di dipendenza negativa nessuna autocorrelazione supera i limiti di cui sopra; in caso di dipendenza nulla i campioni nei quali il valore dell'indice si stacca da 0 sono poche unità; nel caso dipendenza positiva in molti campioni tale indice assume valori diversi da 0, anche se generalmente non molto elevati. Ciò conferma che, al variare della tipologia di dipendenza da durata, vi sono maggiori difficoltà nell'identificabilità dei singoli parametri e quindi maggior lentezza nella simulazione delle serie; non si rileva, tuttavia, alcuna sistematica relazione tra la presenza di autocorrelazioni elevate e la correttezza delle stime. Come precedentemente rilevato, comunque, è consigliabile tenere conto di eventuali eccessive autocorrelazioni nelle serie al fine di scegliere opportunamente il numero di iterazioni ed il *thinning interval.*

Capitolo 3

Gli effetti dell'errore di misura

3.1 Gli errori di misura nell'indicatore di censura

Nel contesto dell'analisi delle dinamiche del mercato del lavoro, molti studi vengono effettuati tramite successive waves di rilevazioni sugli stessi individui, a distanze di tempo regolari e predeterminate. Qualora il campionamento effettuato sia da stock, al momento della prima intervista tutti gli individui sono in stato di disoccupazione. In ciascuna rilevazione successiva, viene richiesto all'intervistato di specificare quale stato stia occupando. Nel particolare modello che verrà preso in esame nel presente lavoro, composto di due sole waves di rilevazione, tale informazione corrisponde all'indicatore di censura δ.

Ipotizziamo che tale informazione possa non essere riportata correttamente da ciascun individuo. Ciò può essere causato da molteplici fattori, anche in base alla tipologia di intervista effettuata: ad esempio, può dipendere da semplici errori di compilazione da parte dell'intervistato o di trascrizione da parte dell'intervistatore. Oppure, in molti casi, l'individuo intervistato potrebbe non conoscere precisamente la definizione che viene data di ciascun stato (Istat, 2004). Ad esempio, nel contesto della Rilevazione

Continua delle Forze Lavoro (RCFL) condotta dall'Istat, un individuo risulta occupato se ha più di 15 anni e nella settimana di riferimento dichiara, in alternativa, di:

- aver svolto almeno un'ora di lavoro in un'attività che preveda un corrispettivo monetario o in natura;
- aver svolto almeno un'ora di lavoro non retribuito nella ditta di un familiare nella quale collabora;
- essere assente dal lavoro con opportune condizioni.

Un individuo è classificato come disoccupato (o in cerca di occupazione) se ha un'età compresa tra i 15 e i 74 anni, non è occupato e dichiara di:

- aver effettuato almeno un'azione attiva di ricerca di lavoro nei trenta giorni precedenti all'intervista ed essere disponibile a lavorare entro due settimane successive all'intervista;
- iniziare un lavoro entro tre mesi dalla data dell'intervista ma essere comunque disponibile a lavorare entro le due settimane all'intervista, qualora fosse possibile anticipare l'inizio del lavoro (Istat, 2004).

I rimanenti intervistati sono inseriti nella categoria dei Non Forza Lavoro, che comprendono studenti, pensionati, casalinghe ma anche potenziali lavoratori che non cercano un'occupazione, ad esempio perchè scoraggiati.

Le definizioni possono non coincidere con la nozione comune delle parole "occupato", "disoccupato" e ciò può indurre i rispondenti a riportare erroneamente il loro stato.

Nel contesto della RCFL, inoltre, la prima rilevazione avviene tramite tecnica CAPI (intervista faccia a faccia), mentre le successive tramite tecnica CATI (intervista telefonica), quindi possono essere condotte in maniera meno accurata.

Poterba e Summers (1984) hanno provato, nel caso dell'indagine statunitense Current Population Survey (CPS), a valutare la proporzione di

unità erroneamente classificate negli stati di occupazione, disoccupazione e Non Forza Lavoro. Ciò veniva effettuato sulla base di una "Reinterview Survey": parte degli intervistati della CPS veniva ricontattata dopo circa una settimana dall'intervista e veniva invitata a descrivere le attività compiute in questa settimana. Le attività che il rispondente affermava di aver effettuato sono quindi state confrontate con la risposta registrata nella CPS circa il proprio stato: si è evidenziata una proporzione non trascurabile di risposte incoerenti.

Poterba e Summers rilevano che lo 0,16% degli individui occupati sono classificati come disoccupati; viceversa, il 3% degli individui disoccupati ha risposto di essere occupato. Se si considera anche la categoria dei NFL tra i possibili stati, si notano proporzioni di errori piuttosto alte, probabilmente perché questo stato può venire facilmente confuso con quello di disoccupazione: si nota che la proporzione di disoccupati classificati come NFL supera il 10%; nella sottopopolazione femminile, tale proporzione raggiunge il 13%.

Lo studio di Poterba e Summers (1984) non ha, naturalmente, una valenza universale, in quanto esso riguarda un'indagine in particolare; tuttavia, esso mostra che la misclassification non è una possibilità remota, ma costituisce realisticamente un problema che investe porzioni non trascurabili dei dati a disposizione.

In un contesto semplificato, nel quale vengono considerati soltanto gli stati di occupazione e disoccupazione, per descrivere l'entità dell'errore di misura, definiamo i parametri:

$$\alpha_0 = P(\delta = 1 | A = 0)$$

$$\alpha_1 = P(\delta = 0 | A = 1),$$

dove A è una variabile dicotomica che indica il vero stato assunto dall'individuo (con A=1 in caso di transito allo stato di occupazione),

mentre la variabile δ indica lo stato riportato nell'intervista ed il suo valore 1 corrisponde al fatto che si riporti lo stato di occupazione. I due parametri in altre parole corrispondono alla probabilità di rispondere in maniera errata, rispettivamente per individui rimasti nello stato di disoccupazione o transitati all'occupazione.

In caso di pattern di misclassification simmetrica, si ipotizza che la probabilità di rispondere correttamente sia la stessa nei due stati, ovvero si assume l'uguaglianza dei due parametri α_0 e α_1. L'errore di misura potrà dipendere anche da una qualche caratteristica degli individui: i parametri α_0 e α_1 non sono costanti all'interno della popolazione, ma cambiano in base al valore assunto da una qualche variabile Z.

3.2 Indicazioni dalla letteratura sugli effetti dell'errore di misura

Gli effetti dell'errore di classificazione dell'indicatore di censura non sono stati molto trattati nel campo della letteratura. Nell'ambito dell'analisi dei dati di durata, infatti, ha ricevuto maggior interesse l'errore di misura nella durata T (e.g., Holt, McDonald e Skinner, 1991; Meier, Richardson e Hughes, 2003) e soprattutto nelle esplicative X (Carroll *et al.*, 2006).

Considerando il problema in maniera più generale, vari studi sono stati comunque condotti sull'effetto dell'errata classificazione in modelli che presentano analogie col modello d'interesse. Sempre nell'ambito del mercato del lavoro Skinner e Torelli (1993) e Skinner (1998) prendono in considerazione i modelli per la stima di flussi. Da tali studi emerge che

l'errata classificazione dello stato di appartenenza può portare a gravi distorsioni nella stima dei parametri.

Hausman, Abrevaya e Scott-Morton (1998) si occupano di modelli lineari generalizzati in cui la variabile dipendente è dicotomica e ipotizzano che venga erroneamente misurata. Viene proposta la formulazione classica tramite variabile latente $y_i^* = x_i^t\beta + \varepsilon_i$, con F funzione di ripartizione di ε nota. Il pattern di misclassification è definito dai parametri α_0 e α_1, come descritto nel precedente paragrafo. Gli autori mostrano che, in questo contesto, le stime dei parametri risultano inconsistenti; la funzione di ripartizione F può essere scelta arbitrariamente, quindi le conclusioni dello studio riguardano anche i classici metodi di stima logit e probit.

In questo contesto è possibile identificare tutti i parametri, compresi quelli che riguardano l'errore di misura, alla sola condizione che $\alpha_0 + \alpha_1 < 1$; tale condizione è decisamente non restrittiva, in quanto se essa non fosse rispettata la qualità dei dati sarebbe tale da impedire ogni sensata analisi successiva. Hausman *et al.* (1998) mostrano che la matrice di informazione di Fisher del modello non è diagonale a blocchi, negando quindi l'indipendenza tra le stime dei parametri β e di quelli che riguardano l'errore di misura. Inoltre, vengono proposte formule analitiche per le derivate parziali di $\hat{\beta}$ rispetto ai parametri α_0 e α_1, calcolate in $\alpha_0 = \alpha_1 = 0$: tramite tali funzioni è possibile ottenere un'approssimazione numerica dell'effetto dell'errore sulla stima.

Il modello di analisi di dati di durata qui trattato, tuttavia, non può essere ricondotto nella classe di modelli considerata da Hausman *et al.* (1998): la probabilità di transitare, $1 - \exp\left(\exp(x_t'\beta)\left(t_i^{\alpha} - (t_i + k)^{\alpha}\right)\right)$, non può essere

scritta in funzione di una combinazione lineare delle covariate con dei parametri β.

Nel caso di campionamento da flusso ciò sarebbe possibile, a patto di considerare covariate del modello $\left(x_i^t, \ln(t_i)\right)$ ed un vettore di parametri (β,α); sarebbe infatti possibile scrivere:

$$P(A_i = 1) = 1 - \exp(-\exp(x_i'\beta + \alpha \ln(t_i))),$$

con A una variabile dicotomica che indica se l'individuo è realmente transitato allo stato di occupazione. Questo caso, si inserisce nella formulazione proposta da Hausman *et al.* (1998), dove la funzione di legame è cloglog ed ε si distribuisce come una variabile Laplace (Florens, Fougère e Mouchart, 2007), quindi sarebbe possibile usare gli strumenti proposti dagli autori per valutare analiticamente gli effetti dell'errata classificazione.

Nel caso in esame, a causa del condizionamento dettato dallo stock sampling, l'ultimo addendo nella somma risulterebbe $\ln\left((t_i + k)^\alpha - t_i^\alpha\right)$, rendendo quindi impossibili ulteriori semplificazioni che permettano di scrivere $P(A_i = 1)$ in funzione di una qualche combinazione lineare di covariate e parametri. Si potrebbe effettuare un tentativo di approssimazione di $(t_i + k)^\alpha$, tramite polinomi di Taylor di primo grado, in $t_i^\alpha + k\alpha (t_i + k)^{\alpha - 1}$; l'espressione $\ln\left((t_i + k)^\alpha - t_i^\alpha\right)$ può essere quindi semplificata come $(\alpha - 1)\ln(t_i + k) + \ln(k) + \ln(\alpha)$; com'è noto, tuttavia, l'approssimazione tramite polinomi di Taylor non è sempre accurata e richiede valori particolarmente piccoli di k, quindi non si rivela appropriata. In altre parole, i risultati di Hausman *et al.* (1998) sarebbero validi anche per il modello in esame solamente se il parametro α, che lega la probabilità di

transitare alla durata della disoccupazione, fosse già noto e ci si limitasse alla stima dei parametri β.

Torelli e Paggiaro (2002) affrontano l'esame degli effetti dell'errata classificazione in maniera empirica, attraverso il metodo Monte Carlo. Gli autori assumono che nella prima intervista lo stato di disoccupazione venga riportato sempre correttamente, mentre le risposte della successiva rilevazione possano essere scorrette. Per la simulazione dei dati, nelle scelte della numerosità campionaria, della distanza tra interviste successive e della tipologia delle covariate e del valore dei relativi parametri, viene ricalcato ciò che tipicamente accade negli studi delle dinamiche del mercato del lavoro basati sulla rotazione del campione. Vengono simulati campioni con α_0 e α_1 pari a circa 0,05 e 0,03 rispettivamente, per dipendenze da durata negativa, nulla o positiva: i risultati mostrano, in ciascun caso, distorsioni verso il valore 1 nella stima del parametro α, mentre i parametri β subiscono una forte attenuazione. Tali effetti diventano più macroscopici all'aumentare dell'errore di misura. I campioni simulati nello studio, rispecchiando quanto tipicamente avviene nella realtà, contengono una proporzione di individui effettivamente transitati decisamente ridotta rispetto a quella degli individui rimasti in stato di disoccupazione. Per tale motivo, viene mostrato come il valore di α_0 abbia più peso di α_1 nel determinare distorsioni nelle stime.

3.3 Stime di massima verosimiglianza su dati affetti da errore di misura

Per valutare empiricamente gli effetti della misclassification sulla stima dei parametri del modello, vengono simulati dei dati affetti da errori di misura.

Lavoriamo sotto l'assunzione che nella prima intervista lo stato venga riportato sempre correttamente, e che quindi tutti gli individui inseriti nel campione siano realmente disoccupati. L'ipotesi non è implausibile, in quanto tale intervista viene tipicamente effettuata con maggiore accuratezza; ad esempio, nel contesto della RCFL, la prima rilevazione viene effettuata con tecnica CAPI (intervista faccia a faccia), mentre le successive tramite tecnica CATI (intervista telefonica). Inoltre è stata fatta la stessa assunzione nei lavori di Poterba e Summers (1995) e di Torelli e Paggiaro (2002).

Paggiaro e Torelli (2004) hanno invece esaminato uno scenario nel quale si ipotizza che nella prima rilevazione qualche occupato possa essere stato classificato come disoccupato; vengono quindi simulati dei dati, assumendo che tali unità classificate erroneamente abbiano comportamenti leggermente differenti nella durata del periodo di ricerca dell'occupazione, quali ad esempio una maggiore probabilità di transizione, assenza di dipendenza da durata o di effetti della covariata X. In tutti i casi appena elencati si osservano attenuazioni nelle stime dei parametri, ma esse, anche in caso di proporzioni di errore non realistiche, sono decisamente meno rilevanti di quelle causate da errata classificazione nell'indicatore di censura.

A partire da tali assunzioni, i tre 3 campioni da 10.000 unità utilizzati nei paragrafi 1.2.2 e 2.5 vengono ripresi e viene creata una nuova variabile δ_i che indica lo stato riportato dall'individuo; lo stato reale viene indicato con A_i. Qualora non siano presenti errori di misura, δ_i ed A_i coincidono per ogni individuo presente nel campione.

Nel presente contesto, invece, la variabile δ_i viene costruita a partire dalla variabile A_i, in questo modo:

$$\delta_i | A_i = 1 \sim \text{Bern}(1 - \alpha_1),$$

$$\delta_i | A_i = 0 \sim \text{Bern}(\alpha_0),$$

in base alle definizioni dei parametri α_1 e α_0 enunciate nel paragrafo 3.1.

Per ciascuno dei 3 campioni in esame ne vengono creati due, con diversi pattern di misclassification: in entrambi i casi essa è simmetrica, nel primo i parametri α_0 e α_1 assumono valore 0,02, mentre nel secondo 0,07.

Su tali campioni vengono poi effettuate stime di massima verosimiglianza. Sulla base di quanto ottenuto da Torelli e Paggiaro (2002), ci aspettiamo distorsioni nelle stime più macroscopiche nel secondo gruppo di campioni. Come nel paragrafo 1.2.2, in Tabella 3.1 vengono riportati i valori delle stime, le loro deviazioni standard ed il valore assunto dalla statistica test Z per il confronto con il vero valore del campione. Tale valore sarà affiancato da un asterisco qualora il test porti ad un rifiuto al livello del 5%.

Tabella 3.1. Weibull con varie tipologie di dipendenza da durata: stime di massima verosimiglianza (1) su campioni affetti da errore in varie proporzioni

		$\alpha_0 = \alpha_1 = 0{,}02$			$\alpha_0 = \alpha_1 = 0{,}07$		
parametro	**vero valore**	**stima**	**se**	**Z value**	**stima**	**se**	**Z value**
				dipendenza negativa			
α	0,606	0,737	0,039	3,343 *	0,818	0,030	7,065 *
β_0	-2,500	-2,641	0,174	-0,811	-2,414	0,133	0,647
β	1,000	0,744	0,054	-4,723 *	0,315	0,039	-17,535 *
				dipendenza nulla			
α	1,000	1,006	0,043	0,144	1,010	0,032	0,326
β_0	-4,000	-3,723	0,183	1,516	-3,164	0,137	6,081 *
β	1,000	0,726	0,051	-5,323 *	0,388	0,037	-16,371 *
				dipendenza positiva			
α	1,284	1,199	0,043	-1,976 *	1,108	0,033	-5,333 *
β_0	-5,000	-4,285	0,174	4,109 *	-3,477	0,135	11,281 *
β	1,000	0,735	0,045	-5,888 *	0,421	0,035	-16,542 *

Si noti come il parametro β sia generalmente e fortemente sottostimato, soprattutto in quei campioni in cui l'entità dell'errore di misura è maggiore. In ogni campione in esame, l'ipotesi che il parametro β sia uguale al suo vero valore 1 può essere sempre rifiutata, e con forte evidenza sperimentale.

Qualora β avesse valore negativo, esso subirebbe una sovrastima; si può quindi parlare di attenuazione per tale parametro.

Le stime del parametro α subiscono una distorsione verso il valore 1, facendo quindi emergere in maniera ridotta la dipendenza tra durata e probabilità di transizione. Anche per questo parametro si tende a rifiutare l'uguaglianza al suo vero valore corrispondente, con l'ovvia eccezione del caso di dipendenza nulla. Infine, si noti la sovrastima del parametro β_0: tale fenomeno non si verifica in caso di dipendenza negativa, ma è sempre più evidente al crescere del parametro α.

Per avere una valutazione più generale degli effetti della misclassification sulla stima dei parametri che non dipenda dai singoli campioni presentati, è stato usato il metodo Monte Carlo con 100 repliche, i cui risultati sono stati riportati diffusamente nell'appendice B.

Da tali simulazioni emerge, tra l'altro come il valore di α_0 influisca maggiormente sull'entità delle distorsioni di quello di α_1: quest'ultimo, infatti, produce una distorsione tanto più forte quanto più numerosi sono i transitati; naturalmente, vale la considerazione opposta per α_0. Nei campioni in esame la proporzione dei transitati è, come negli analoghi campioni reali, piuttosto bassa (il 15%) e ciò spiega l'asimmetria negli effetti dei due tipi di errata classificazione.

3.4 Il Modello Bayesiano su dati affetti da errore di misura

Anche il Modello Bayesiano viene ora applicato a dati affetti da errori di misura, per valutare i suoi effetti sulle distribuzioni a posteriori stimate. I

campioni utilizzati sono i medesimi sui quali sono state effettuate stime di massima verosimiglianza.

3.4.1 Errori di classificazione al 2%

Vengono inizialmente esaminati i risultati relativi ai campioni con proporzione di risposte errate 2%. Di seguito, si riportano nella Figura 3.1 le serie simulate raffrontate con i "veri valori" dei parametri e nelle Tabelle 3.2 e 3.3 i valori delle loro autocorrelazioni e crosscorrelazioni al fine di valutare eventuali mutamenti su convergenza e stazionarietà delle serie dovuti alla presenza di dati errati. Come nel paragrafo 2.5, la Figura 3.1 e le Tabelle 3.2 e 3.3 sono relative alla situazione di dipendenza da durata negativa ma i commenti a riguardo sono estendibili agli altri casi.

Figura 3.1. Weibull con dipendenza da durata negativa: tre catene MCMC parallele per la simulazione delle distribuzioni a posteriori dei parametri α, β e β_0 da un campione affetto da errori al 2%

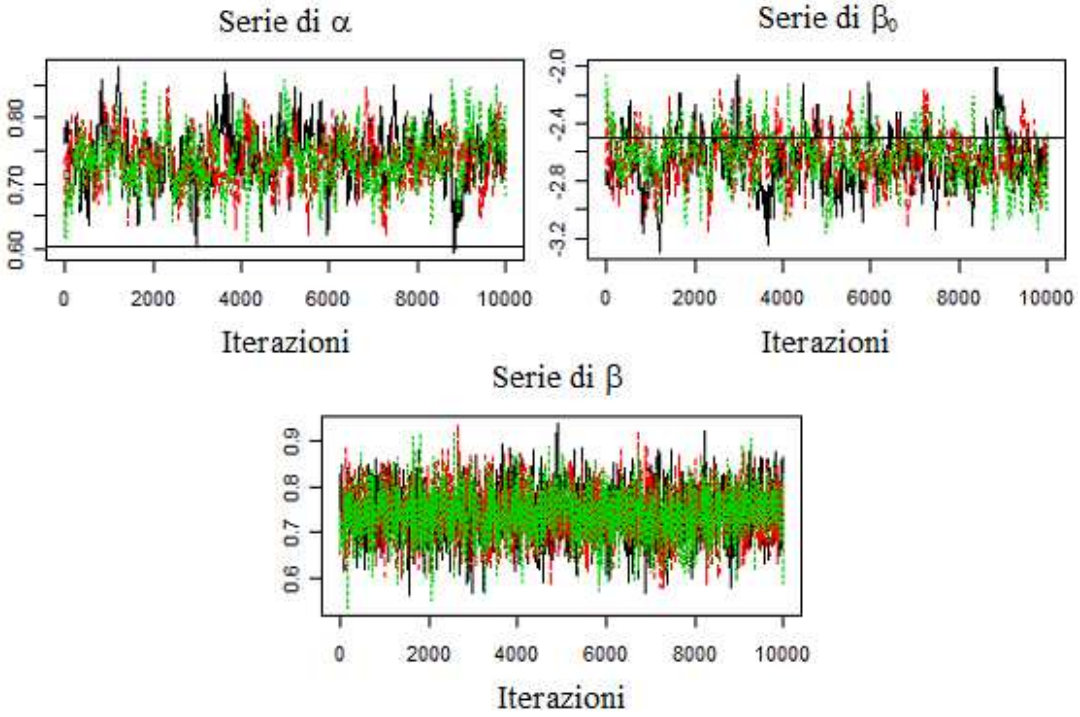


Tabella 3.2. Weibull con dipendenza da durata negativa: autocorrelazioni a vari ritardi delle catene MCMC; dati affetti da errori di classificazione al 2%

Serie	α	β_0	β
5	0,868	0,903	0,198
25	0,567	0,587	0,070
50	0,344	0,349	0,030
250	0,002	0,009	0,009

Tabella 3.3. Weibulle con dipendenza da durata negativa: crosscorrelazioni tra le catene MCMC; dati affetti da errori di classificazione al 2%

	α	β_0	β
α	1	-0,970	0,319
β_0		1	-0,308
β			1

Da tutti gli indicatori riportati, si nota come la situazione sia del tutto assimilabile a quella riportata nel paragrafo 2.5 e riferita ai dati senza errori. Anche in questo caso, infatti, si rileva per α e β_0 la forte correlazione negativa, gli alti valori di autocorrelazione e la lenta variazione delle loro serie. La difficile identificabilità dei due parametri è, difatti, una caratteristica strutturale del modello, che non dipende dai dati sui quali viene stimato.

Esaminiamo ora le distribuzioni a posteriori stimate: le Figure 3.2, 3.3 e 3.4 mostrano le densità di ciascun parametro stimate tramite metodo Kernel. Inoltre indici di sintesi delle distribuzioni a posteriori stimate vengono riportate nella Tabella 3.5; essa è affiancata dalla Tabella 3.4, che ripropone per confronto le stime ottenute tramite massima verosimiglianza.

Si noti come la forma delle distribuzioni a posteriori sia ancora approssimativamente normale; inoltre si osserva dai valori delle deviazioni standard che la variabilità di tali distribuzioni è del tutto analoga a quella riscontrata nelle stime effettuate con modello analogo ma su dati non affetti da errore di misura.

Figura 3.2. Weibull con dipendenza da durata negativa: stime delle distribuzioni a posteriori dei parametri α, β e β_0 su dati affetti da errori di classificazione al 2%

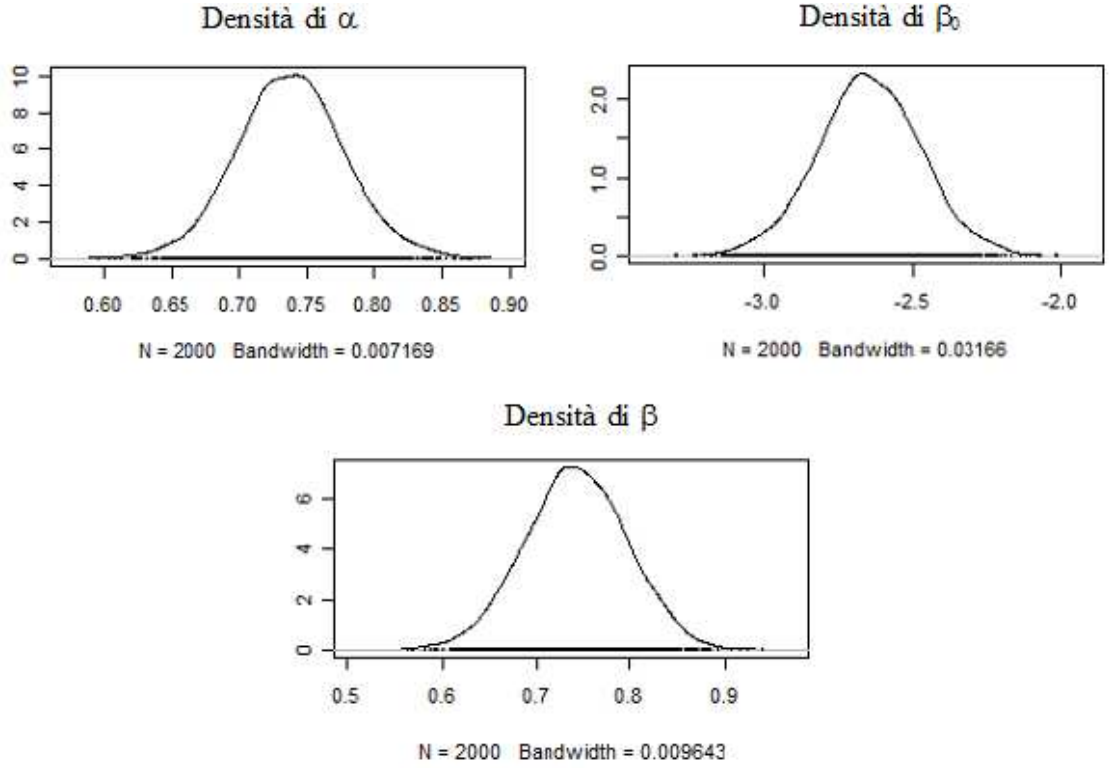


Figura 3.3. Weibull con dipendenza da durata nulla: stime delle distribuzioni a posteriori dei parametri α, β e β_0 su dati affetti da errori di classificazione al 2%

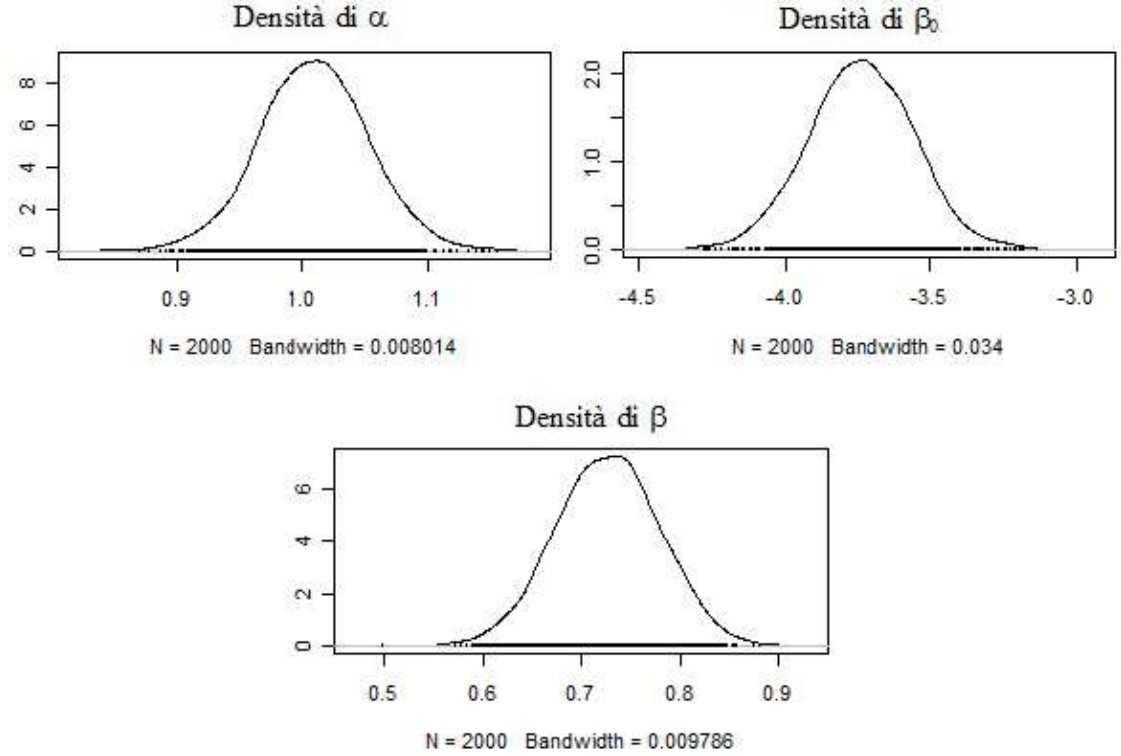


Figura 3.4. Weibull con dipendenza da durata positiva: stime delle distribuzioni a posteriori dei parametri α, β e β_0 su dati affetti da errori di classificazione al 2%

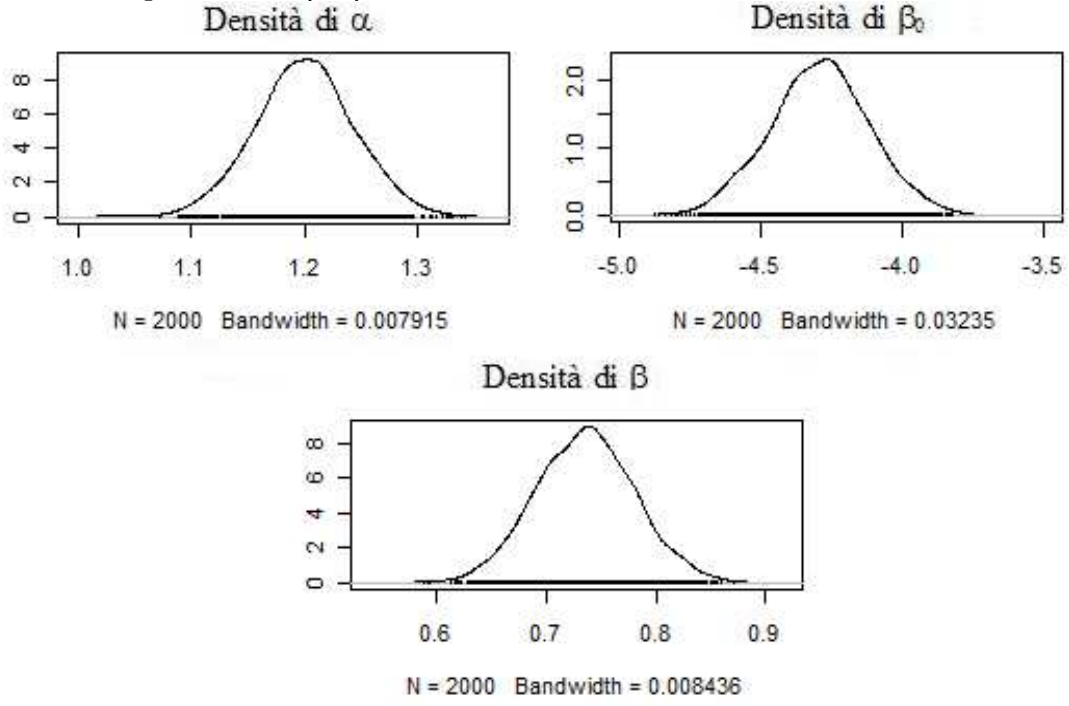


Tabella 3.4. Weibull con varie tipologie di dipendenza da durata: stime di massima verosimiglianza (1) su dati affetti da errore al 2%

		$\alpha_0 = \alpha_1 = 0{,}02$		
parametro	**vero valore**	**stima**	**se**	**Z value**
		dipendenza negativa		
α	0,606	0,737	0,039	3,343 *
β_0	-2,500	-2,641	0,174	-0,811
β	1,000	0,744	0,054	-4,723 *
		dipendenza nulla		
α	1,000	1,006	0,043	0,144
β_0	-4,000	-3,723	0,183	1,516
β	1,000	0,726	0,051	-5,323 *
		dipendenza positiva		
α	1,284	1,199	0,043	-1,976 *
β_0	-5,000	-4,285	0,174	4,109 *
β	1,000	0,735	0,045	-5,888 *

Tabella 3.5. Indici di sintesi delle distribuzioni a posteriori stimate sui tre campioni in esame, affetti da errore di classificazione al 2%

parametro	vero valore	Media	SD	2,5%	25%	50%	75%	97,5%
dipendenza da durata negativa								
α *	0,606	0,738	0,039	0,659	0,712	0,738	0,764	0,819
β_0	-2,500	-2,649	0,175	-3,003	-2,762	-2,652	-2,534	-2,291
β *	1,000	0,742	0,054	0,634	0,706	0,742	0,778	0,847
dipendenza da durata nulla								
α	1,000	1,009	0,043	0,923	0,980	1,009	1,038	1,094
β_0	-4,000	-3,737	0,183	-4,098	-3,860	-3,736	-3,614	-3,375
β *	1,000	0,726	0,053	0,623	0,690	0,726	0,762	0,828
dipendenza da durata positiva								
α *	1,284	1,200	0,044	1,115	1,172	1,200	1,229	1,286
β_0 *	-5,000	-4,293	0,179	-4,644	-4,408	-4,290	-4,175	-3,942
β *	1,000	0,735	0,045	0,646	0,705	0,736	0,765	0,826

Soltanto i valori centrali delle distribuzioni a posteriori sembrano venir modificati per effetto della misclassification; i loro spostamenti avvengono nelle stesse direzioni di quelli avvenuti nel medesimo contesto per le stime di massima verosimiglianza: usando come stime puntuali dei parametri le medie delle distribuzioni a posteriori, si nota che le stime del parametro α vengono distorte verso il valore 1, il parametro β è fortemente attenuato, il parametro β_0 subisce una sovrastima tanto più forte al crescere di α. Tali stime sono estremamente simili a quelle ottenute sugli stessi campioni tramite stima di massima verosimiglianza.

Si costruiscono infine intervalli di credibilità a livello 0,95, attraverso i quantili o l'approssimazione normale; essi risultano molto simili tra loro. Nella Tabella 3.5 sono stati segnalati tramite un asterisco i parametri il cui "vero valore" non appartiene ai suddetti intervalli; questi parametri sono gli stessi per cui, in ambito di massima verosimiglianza, si rifiutava il test di uguaglianza ai loro veri valori.

3.4.2 Errore di classificazione al 7%

Si riportano ora le distribuzioni a posteriori dei parametri stimate sui campioni nei quali la proporzione di risposte errate raggiunge il 7%. Quanto alle problematiche di convergenza e stazionarietà della serie, esse non subiscono variazioni di rilievo rispetto al caso di dati senza errori o con proporzione di errori 2%. Si rimanda perciò ai paragrafi 2.5 o 3.4.1.

Le Figure 3.5, 3.6 e 3.7 mostrano le densità stimate per ciascun paragrafo, mentre gli indici di sintesi relativi alle distribuzioni a posteriori sono riportato in Tabella 3.7.

Figura 3.5. Weibull con dipendenza da durata negativa: stime delle distribuzioni a posteriori dei parametri α, β e β_0 su dati affetti da errori di classificazione al 7%

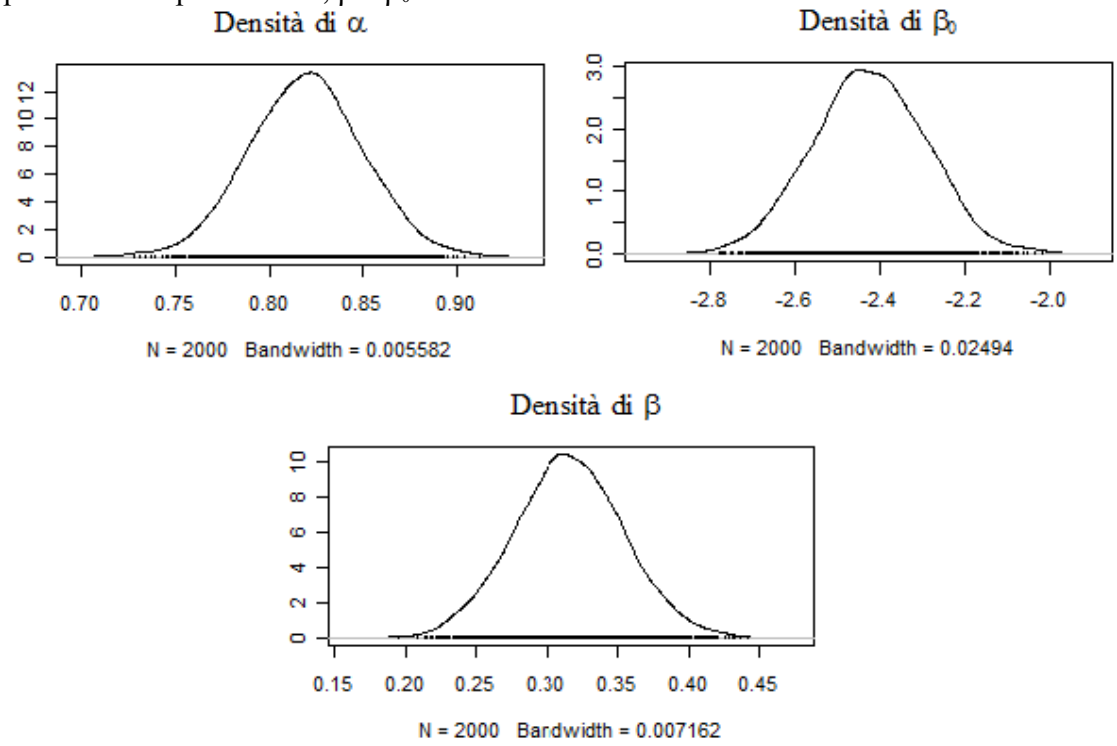


Figura 3.6. Weibull con dipendenza da durata nulla: stime delle distribuzioni a posteriori dei parametri α, β e β_0 su dati affetti da errori di classificazione al 7%

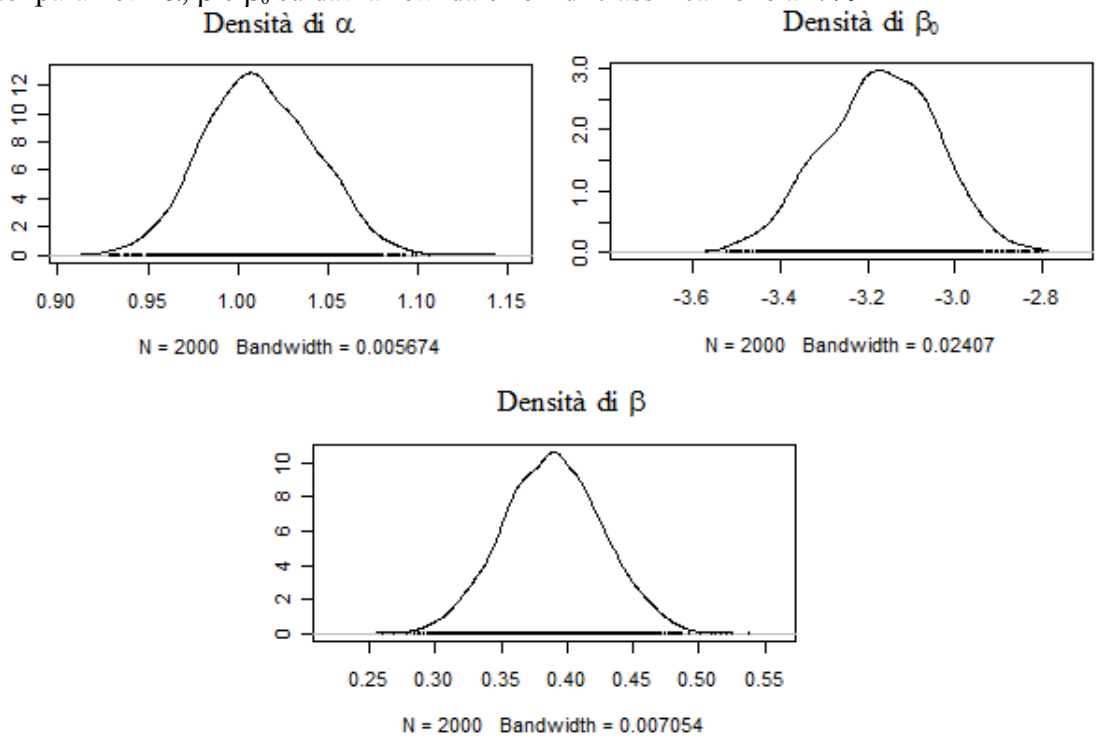


Figura 3.7. Weibull con dipendenza da durata positiva: stime delle distribuzioni a posteriori dei parametri α, β e β_0 su dati affetti da errori di classificazione al 7%

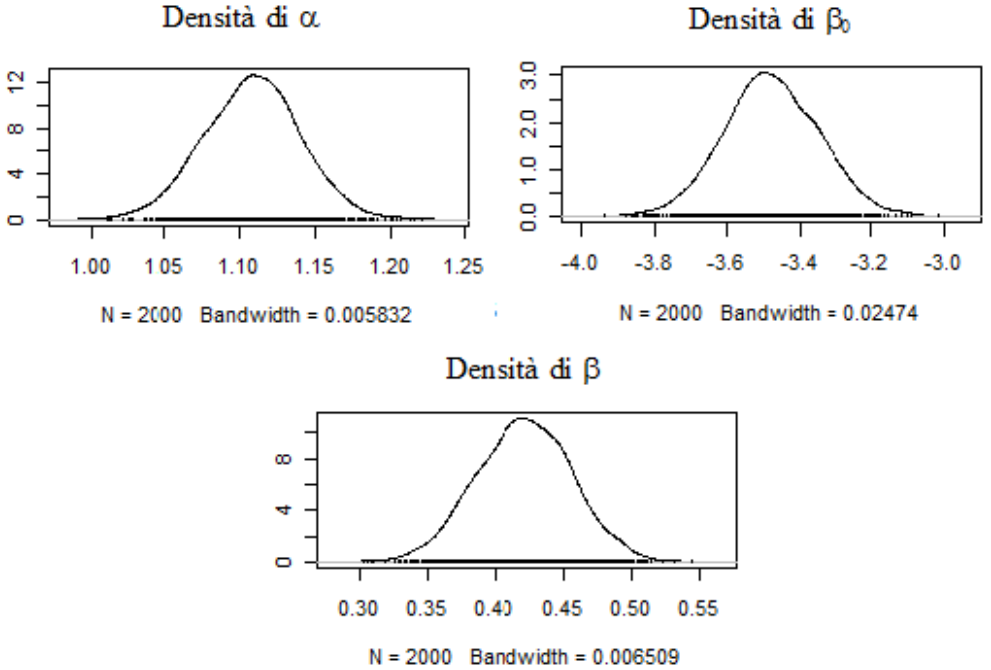


Tabella 3.6. Weibull con varie tipologie di dipendenza da durata: stime di massima verosimiglianza (1) su dati affetti da errore al 7%

		$\alpha_0 = \alpha_1 = 0{,}07$		
parametro	**vero valore**	**stima**	**se**	**Z value**
		dipendenza negativa		
α	0,606	0,818	0,030	7,065 *
β_0	-2,500	-2,414	0,133	0,647
β	1,000	0,315	0,039	-17,535 *
		dipendenza nulla		
α	1,000	1,010	0,032	0,326
β_0	-4,000	-3,164	0,137	6,081 *
β	1,000	0,388	0,037	-16,371 *
		dipendenza positiva		
α	1,284	1,108	0,033	-5,333 *
β_0	-5,000	-3,477	0,135	11,281 *
β	1,000	0,421	0,035	-16,542 *

Tabella 3.7. Indici di sintesi delle distribuzioni a posteriori stimate sui tre campioni in esame, affetti da errore di classificazione al 7%

parametro	**vero valore**	**Media**	**SD**	**2,5%**	**25%**	**50%**	**75%**	**97,5%**
			dipendenza da durata negativa					
α *	0,606	0,820	0,030	0,761	0,799	0,820	0,839	0,879
β_0	-2,500	-2,424	0,134	-2,683	-2,513	-2,426	-2,333	-2,161
β *	1,000	0,315	0,038	0,239	0,290	0,315	0,341	0,392
			dipendenza da durata nulla					
α	1,000	1,012	0,031	0,954	0,990	1,011	1,033	1,072
β_0 *	-4,000	-3,172	0,129	-3,429	-3,259	-3,167	-3,079	-2,930
β *	1,000	0,389	0,038	0,316	0,363	0,389	0,414	0,465
			dipendenza da durata positiva					
α *	1,284	1,108	0,032	1,043	1,087	1,109	1,129	1,170
β_0 *	-5,000	-3,477	0,133	-3,735	-3,565	-3,481	-3,386	-3,212
β *	1,000	0,421	0,035	0,351	0,397	0,421	0,445	0,489

Rispetto ai campioni precedentemente esaminati, con errori di misura assenti o presenti in misura minore, le distribuzioni a posteriori mantengono la forma e la variabilità ma subiscono delle vistose traslazioni. Esse avvengono nelle direzioni precedentemente elencate, ma sono di maggior entità: si noti, ad esempio, che le distribuzioni a posteriori di β sono concentrate intorno a valori decisamente minori di 1 come 0,3.

Anche in questo caso, inoltre, vengono segnalati in Tabella 3.7 tramite un asterisco i parametri il cui "vero valore" non appartenga all'intervallo di credibilità costituito dai quantili 0,025 e 0,975 della distribuzione a posteriori; essi coincidono con quelli per cui la statistica Z portava a rifiutare il test di uguaglianza ai loro veri valori in ambito di massima verosimiglianza, riportati sempre tramite asterisco in Tabella 3.6.

3.5 Stime ripetute su insiemi di 100 campioni

Anche in questo caso si vuole valutare il processo di stima in maniera più generale, senza dipendere dagli specifici campioni in esame; per fare ciò, ci inseriamo nel più ampio disegno sperimentale cui si è fatto riferimento nel paragrafo 2.6. In tale contesto, sono stati simulati un considerevole numero di campioni, per la precisione 100, di 1000 unità; ad ognuno di essi veniva poi applicato il modello Bayesiano e gli esiti di tale processo di stima sono stati registrati tramite l'insieme di indici descritto nel paragrafo 2.6.

Il procedimento era stato ripetuto su tre insiemi di parametri, che corrispondono rispettivamente a situazioni di dipendenza da durata negativa, nulla e positiva.

In ciascun campione utilizzato in tale contesto, viene ora simulata la variabile δ_i, che indica lo stato riportato dall'individuo a partire dalla variabile A_i; le modalità di tali simulazioni vengono chiarite nel paragrafo 3.3. Analogamente a quanto fatto per i singoli 3 campioni in esame, vengono prodotte due variabili δ_i, corrispondenti a due diversi pattern di misclassification, entrambi simmetrici con proporzioni di errori rispettivamente 2% e 7%. Su ciascuno dei campioni così ottenuti viene applicato il modello Bayesiano; le modalità con le quali è stato eseguito il

processo di stima e gli indici di sintesi calcolati sono gli stessi del precedente contesto, nel quale non erano presenti errori di misura.

Tabella 3.8. Media e standard deviation di indicatori delle distribuzioni a posteriori in gruppi di 100 campioni con dipendenza da durata negativa affetti da errore in varie proporzioni

	α		β_0		β	
vero valore	*0,606*		*-2,500*		*1,000*	
	media	**sd**	**media**	**sd**	**media**	**sd**
			$\alpha_0 = \alpha_1 = 0{,}02$			
media	0,765	0,128	-2,795	0,583	0,639	0,221
I quartile	0,683	0,125	-3,157	0,600	0,522	0,219
mediana	0,762	0,128	-2,790	0,579	0,638	0,221
III quartile	0,844	0,133	-2,430	0,567	0,755	0,224
copertura	0,520		0,670		0,200	
			$\alpha_0 = \alpha_1 = 0{,}07$			
media	0,865	0,095	-2,632	0,425	0,321	0,145
I quartile	0,800	0,092	-2,914	0,439	0,236	0,144
mediana	0,863	0,095	-2,626	0,424	0,320	0,145
III quartile	0,928	0,098	-2,346	0,411	0,404	0,146
copertura	0,080		0,760		0,000	

Tabella 3.9. Media e standard deviation di indicatori delle distribuzioni a posteriori in gruppi di 100 campioni con dipendenza da durata nulla affetti da errore in varie proporzioni

	α		β_0		β	
vero valore	*1,000*		*-4,000*		*1,000*	
	media	**sd**	**media**	**sd**	**media**	**sd**
			$\alpha_0 = \alpha_1 = 0{,}02$			
media	1,016	0,112	-3,776	0,495	0,705	0,190
I quartile	0,929	0,109	-4,137	0,520	0,598	0,185
mediana	1,013	0,112	-3,766	0,494	0,705	0,189
III quartile	1,100	0,116	-3,408	0,471	0,811	0,195
copertura	0,860		0,820		0,350	
			$\alpha_0 = \alpha_1 = 0{,}07$			
media	1,008	0,098	-3,168	0,431	0,368	0,125
I quartile	0,942	0,095	-3,447	0,445	0,289	0,123
mediana	1,008	0,098	-3,167	0,431	0,367	0,125
III quartile	1,074	0,101	-2,884	0,419	0,447	0,127
copertura	0,820		0,270		0,000	

Tabella 3.10. Media e standard deviation di indicatori delle distribuzioni a posteriori in gruppi di 100 campioni con dipendenza da durata positiva affetti da errore in varie proporzioni

	α		β_0		β	
vero valore	*1,284*		*-5,000*		*1,000*	
	media	**sd**	**media**	**sd**	**media**	**sd**
			$\alpha_0 = \alpha_1 = 0{,}02$			
media	1,188	0,115	-4,326	0,484	0,708	0,146
I quartile	1,099	0,113	-4,683	0,501	0,609	0,144
mediana	1,186	0,116	-4,320	0,488	0,707	0,146
III quartile	1,274	0,119	-3,959	0,473	0,806	0,148
copertura	0,660		0,460		0,280	
			$\alpha_0 = \alpha_1 = 0{,}07$			
media	1,102	0,097	-3,496	0,407	0,393	0,119
I quartile	1,033	0,093	-3,776	0,423	0,316	0,118
mediana	1,101	0,096	-3,492	0,405	0,392	0,119
III quartile	1,170	0,100	-3,210	0,390	0,469	0,120
copertura	0,250		0,030		0,000	

Per ogni tipologia di dipendenza da durata, vengono riportati i risultati ottenuti con le due proporzioni di errore di misura: le Tabelle 3.8, 3.9, 3.10 contengono le medie e le standard deviation degli indici rilevati sui gruppi di 100 campioni. Le distorsioni nelle stime dei parametri appaiono invariate per direzione ed entità da quelle rilevate dall'esame dei singoli campioni. La variabilità delle stime tra i campioni rimane sostenuta, ma sembra leggermente ridotta rispetto a quella osservata in assenza di errori di misura; naturalmente parte di tale riduzione è dovuta all'attenuazione dei parametri, ma il cambiamento di scala, in quanto lieve, non può giustificare da solo tale diminuzione nella variabilità delle stime.

La proporzione di campioni nei quali il "vero valore" del parametro appartiene all'intervallo HPD si distacca ora dallo 0,8, tranne nei casi particolari in cui non avvengono distorsioni; tale proporzione si abbassa particolarmente nel caso del parametro β ed arriva a 0 nei campioni con maggiori proporzioni di errate classificazioni.

Tabella 3.11. Medie e sd dei valori di autocorrelazione a ritardo 5 e crosscorrelazione delle serie di α e β_0 in gruppi di 100 campioni con varie tipologie di dipendenza da durata e affetti da errore in varie proporzioni

	negativa		**nulla**		**positiva**	
	media	**se**	**media**	**se**	**media**	**se**
			$\alpha_0 = \alpha_1 = 0{,}02$			
autocorrelazione α	0,841	0,021	0,873	0,019	0,891	0,016
autocorrelazione β_0	0,891	0,014	0,911	0,014	0,924	0,011
crosscorrelazione $\alpha\,\beta_0$	-0,960	0,006	-0,969	0,005	-0,974	0,004
			$\alpha_0 = \alpha_1 = 0{,}07$			
autocorrelazione α	0,858	0,017	0,875	0,017	0,890	0,016
autocorrelazione β_0	0,900	0,013	0,910	0,014	0,921	0,011
crosscorrelazione $\alpha\,\beta_0$	-0,965	0,004	-0,971	0,004	-0,974	0,004

Nella Tabella 3.11 sono stati riportati i valori di autocorrelazione e crosscorrelazione per i parametri α e β_0; tali quantità non subiscono alcuna variazione di rilievo, sia per valore centrale che variabilità, rispetto alle corrispondenti registrate in assenza di errori di misura, riportate nel paragrafo 2.6.2.

Già dall'analisi dei singoli campioni, era emerso come la presenza degli errori di misura avesse influenza sui valori centrali delle distribuzioni a posteriori stimate, ma non sulle problematiche di convergenza, stazionarietà ed autocorrelazione delle serie.

L'indice costruito a partire dalle autocorrelazioni a ritardi maggiori di 5 ha un comportamento del tutto analogo a quello riportato nel paragrafo 2.6.2 per il caso di assenza di errori di misura.

Capitolo 4

Un Modello Gerarchico Bayesiano per dati affetti da errore di misura

4.1 Verosimiglianza corretta per dati affetti da errori di misura

Come è stato ampiamente mostrato precedentemente, la stima di massima verosimiglianza produce stime non corrette se effettuata su dati affetti da errore di misura nell'indicatore di censura.

La verosimiglianza (1) usata per tali stime nel paragrafo 1.2.2 non tiene infatti conto del potenziale errore di misura nei valori δ_i.

La seguente verosimiglianza risulta invece corretta per dati affetti da errori di misura:

$$L(\alpha,\beta,\alpha_0,\alpha_1;x,t,\delta) = \prod_{i=1}^{n} \left(P[\delta_i = 1]\right)^{\delta_i} \left(1 - P[\delta_i = 1]\right)^{1-\delta_i} \quad (3),$$

$$\text{con } P[\delta_i = 1] = (1-\alpha_1)\left(1 - \frac{S(t_i + k, x_i;\alpha,\beta)}{S(t_i, x_i;\alpha,\beta)}\right) + \alpha_0 \frac{S(t_i + k, x_i;\alpha,\beta)}{S(t_i, x_i;\alpha,\beta)} =$$

$$= (1-\alpha_1) + (\alpha_0 + \alpha_1 - 1)\exp\left(\exp(x_i'\beta)\left(t_i^{\alpha} - (t_i + k)^{\alpha}\right)\right).$$

Si noti come (1) coincida con (3) se α_0 e α_1 sono uguali a 0. Nel caso in cui α_0 e α_1 siano note, tale verosimiglianza fornisce stime consistenti. Il metodo di stima viene proposto da Poterba e Summers (1995); Torelli e Paggiaro (2002) mostrano, attraverso il metodo Monte Carlo, di aver ottenuto stime corrette dei parametri.

I campioni da 10.000 unità esaminati precedentemente vengono ripresi e su quelli affetti da errori di misura (si rimanda al paragrafo 3.3) vengono effettuate stime di massima verosimiglianza, usando la verosimiglianza (3) con α_0, α_1 pari ai valori usati nella simulazione del campione.

Tabella 4.1. Weibull con varie tipologie di dipendenza da duarat: stime di massima verosimiglianza (3) su campioni affetti da errore in varie proporzioni; stime effettuate tramite la verosimiglianza (1) nella colonna "ver.(1)"

		$\alpha_0 = \alpha_1 = 0{,}02$				$\alpha_0 = \alpha_1 = 0{,}07$			
par.	**vero**	**ver.(1)**	**stima**	**se**	**Z value**	**ver.(1)**	**stima**	**se**	**Z value**
				dipendenza negativa					
α	0,606	0,737 *	0,675	0,056	1,240	0,818 *	0,605	0,074	-0,010
β_0	-2,500	-2,641	-2,862	0,253	-1,432	-2,414	-2,539	0,330	-0,117
β	1,000	0,744 *	1,149	0,086	1,727	0,315 *	0,987	0,115	-0,108
				dipendenza nulla					
α	1,000	1,006	1,014	0,062	0,223	1,010	0,995	0,079	-0,062
β_0	-4,000	-3,723	-4,176	0,266	-0,662	-3,164 *	-4,040	0,338	-0,117
β	1,000	0,726 *	1,080	0,081	0,990	0,388 *	1,084	0,103	0,814
				dipendenza positiva					
α	1,284	1,199 *	1,230	0,060	-0,900	1,108 *	1,357	0,085	0,859
β_0	-5,000	-4,285 *	-4,994	0,254	0,024	-3,477 *	-5,278	0,366	-0,760
β	1,000	0,735 *	1,028	0,065	0,431	0,421 *	1,070	0,091	0,769

La Tabella 4.1 riporta i valori delle stime, dei loro errori standard, e del test Z; tali valori vengono affiancati a quelli delle stime ottenute sugli stessi campioni con la verosimiglianza (1). Vengono evidenziate tramite asterisco le stime che risultano significativamente diverse dal vero valore del parametro. Si noti come le stime ottenute risultino decisamente più vicine ai veri valori dei parametri, rispetto a quelle ottenute tramite verosimiglianza

(1); difatti, in nessun caso viene rifiutato il test di uguaglianza con i veri valori, a differenza di quello che avviene per le stime precedenti.

Tuttavia la reale entità dell'errore di misura è conosciuta nel caso di dati simulati, ma ovviamente non nel caso di campioni reali. Per i parametri α_0 e α_1 viene allora usata una loro stima, fornita da altre fonti di dati o altri studi effettuati su popolazioni comparabili con quella in esame. Tuttavia tale metodo non risulta consigliabile, in quanto nel caso i parametri α_0 e α_1 inseriti nella verosimiglianza si discostino da quelli reali, si otterranno distorsioni nella stima dei parametri α e β. Inoltre, tramite tale metodo, viene sottostimata la deviazione standard delle stime ottenute, in quanto non viene presa in considerazione la varianza legata alla stima dei parametri α_1 e α_0 (Hausman *et al.*, 1998).

Paggiaro e Torelli (2004) utilizzano l'algoritmo EM per ottenere una massimizzazione della verosimiglianza (3), non solo nei parametri α e β ma anche in α_0 e α_1. Lo studio riporta che, probabilmente a causa della difficile identificabilità di alcuni parametri, l'algoritmo converge spesso a valori sul confine dello spazio parametrico o non ragionevoli.

4.2 Modelli Gerarchici Bayesiani

Si parla di Modelli Gerarchici Bayesiani riferendosi ad una strategia di costruzione del modello nel quale le quantità non osservate (quali ad esempio parametri, dati mancanti o affetti da errore di misura, effetti casuali ecc.) sono organizzate in un numero discreto di livelli, legati tra loro da funzioni deterministiche o relazioni stocastiche che descrivono le caratteristiche insite nei dati (Richardson e Best, 2003). L'applicabilità di

tale classe di modelli ha subito un netto miglioramento con lo sviluppo degli algoritmi computazionali, in particolare del metodo Markov Chain Monte Carlo.

Tali modelli vengono spesso usati, ad esempio, in ambito multilevel, qualora sia presente una gerarchia anche nei dati. Vengono generalmente usati anche nel caso in cui si vogliano esaminare dati relativi ad unità sperimentali tra loro simili, quali ad esempio piccole aree geografiche; tali dati si assumono realizzazioni da una famiglia di distribuzioni casuali comune, i cui parametri sono a loro volta realizzazioni di una variabile aleatoria; i parametri legati a tale variabile sono detti *iperparametri*.

Ad esempio, in un modello Poisson-Gamma, il primo livello consta di conteggi osservati, realizzazioni di variabili di tipo Poisson, tra loro condizionalmente indipendenti date le medie delle distribuzioni dalle quali sono stati generati. Nel secondo livello del modello, le medie stesse sono realizzazioni di una variabile aleatoria Gamma; infine, i parametri di tale variabile Gamma occupano il terzo livello del modello.

$$x_i \sim P(\theta_i),$$

$$\theta_i \sim \Gamma(\alpha, \beta),$$

$$(\alpha, \beta) \sim \pi(\vartheta),$$

con $\pi(\vartheta)$ distribuzione a priori per α e β il cui parametro ϑ non è aleatorio, ma fissato.

Esempi di applicazione di questo tipo di modelli possono essere tassi di mortalità in varie aree geografiche o proporzione di successi negli esami in diversi scuole. Queste procedure sono spesso usate per effettuare uno *smoothing* verso la media di una variabile aleatoria osservata in molteplici unità (Congdon, 2003).

Si basano su un'implicita assunzione: tutte le unità sono tra loro intercambiabili, o comunque abbastanza simili da giustificare l'assunzione di una comune densità.
I Modelli Gerarchici Bayesiani, al di là di questi esempi, sono usati ovunque sia possibile rappresentare il problema a vari livelli di astrazione e sia possibile descriverlo sulla base di distribuzioni di probabilità ognuna delle quali condizionata al proprio livello superiore.

4.3 Un Modello Gerarchico Bayesiano per dati affetti da errore di misura

L'algoritmo EM, utilizzato per ottenere una massimizzazione della verosimiglianza (3), converge spesso a valori sul confine dello spazio parametrico o non ragionevoli, probabilmente a causa della difficile identificabilità di alcuni parametri (Paggiaro e Torelli, 2004). Inquadrare il modello in abito Bayesiano potrebbe essere decisivo per la stima dell'intero insieme di parametri, in quanto la scelta di opportune distribuzioni a priori può aiutare il procedimento di stima a non convergere su punti dello spazio parametrico poco sensati per il modello e la popolazione in esame.

4.3.1 Scelta delle distribuzioni a priori

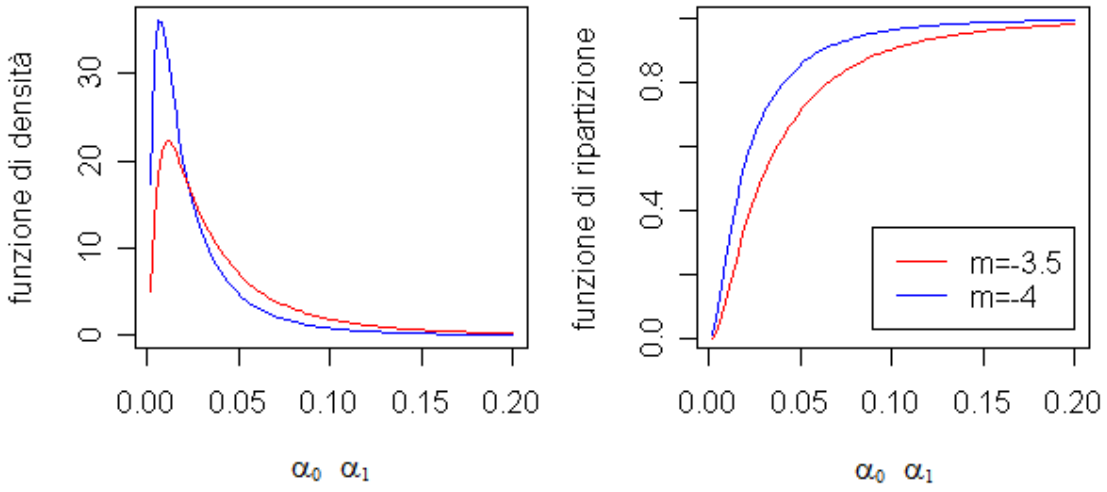
Per $\vartheta = (\alpha, \beta_0, \beta, \alpha_0, \alpha_1)$ viene fissata distribuzione a priori $\pi(\vartheta)$ pari al prodotto delle distribuzioni a priori di ogni singola componente del vettore, dove per la scelta delle distribuzioni dei primi tre elementi del vettore si rimanda al paragrafo 2.4.1.

Per i parametri α_1 e α_0 è necessario scegliere distribuzioni continue con supporto (0,1), che si assumono tra loro indipendenti; viene proposta la variabile aleatoria Beta oppure distribuzione Normale sulla trasformazione logit del parametro.

Per valutare quali elementi prendere in considerazione all'interno di tali famiglie di distribuzioni si rammenti che è opportuno utilizzare una distribuzione a priori concentrata su valori prossimi allo 0.

Se si assume che la trasformata logit del parametro sia normale, è opportuno fissarne la media *m* tra –4 e –3,5, mentre per la varianza si consigliano valori non eccessivamente elevati, ad esempio 1; la distribuzione viene mostrata nella Figura 4.1 con varie parametrizzazioni.

Figura 4.1. Funzioni di densità e ripartizione di una variabile aleatoria la cui trasformata logit si distribuisce come una normale di media *m*, per vari valori del parametro *m*



La distribuzione Beta, con $f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$, ha valore atteso pari a $\frac{a}{a+b}$ e varianza pari a $\frac{ab}{(a+b)^2(a+b+1)}$. Se si vuole fissare la media su valori bassi, vicini allo 0, è dunque necessario fissare oppotunamente la proporzione tra i parametri a e b; ad esempio, fissare una

proporzione pari 1 a 19 determina una media 0,05, mentre fissare una proporzione 1 a 15 determina una media 0,063. Si presti attenzione che al decrescere della proporzione, e quindi della media, tende a decrescere anche la varianza, rischiando di creare a priori molto concentrate su un singolo valore e quindi troppo informative.

Oltre alla media, è necessario decidere la forma della distribuzione, attraverso il valore del parametro a (b è ad esso vincolato). Per a uguale a 1 si ottiene una distribuzione con moda sullo 0. Per a immediatamente maggiore di 1 si ottiene una distribuzione il cui supporto non contiene lo 0, ma con moda molto vicina a tale valore. Al crescere di a la moda si sposta sempre più verso il valore medio della distribuzione e la densità assume forme più arrotondate; tuttavia anche la varianza decresce, con le conseguenze appena accennate. Le Figure 4.2 e 4.3 mostrano quanto appena accennato.

La distribuzione Normale sulla trasformazione logit, provata con vari valori medi (quali -3,5, -4) e varie varianze (quali 0,5, 1), ha prodotto talvolta distribuzioni a posteriori molto diffuse, che estendevano il loro supporto anche su valori poco sensati soprattutto per quanto riguarda α_1.

Figura 4.2. Funzioni di densità e ripartizione di una variabile aleatoria Beta, per vari valori del parametro *a*, con *b*=15*a* (media fissata a 0,063)

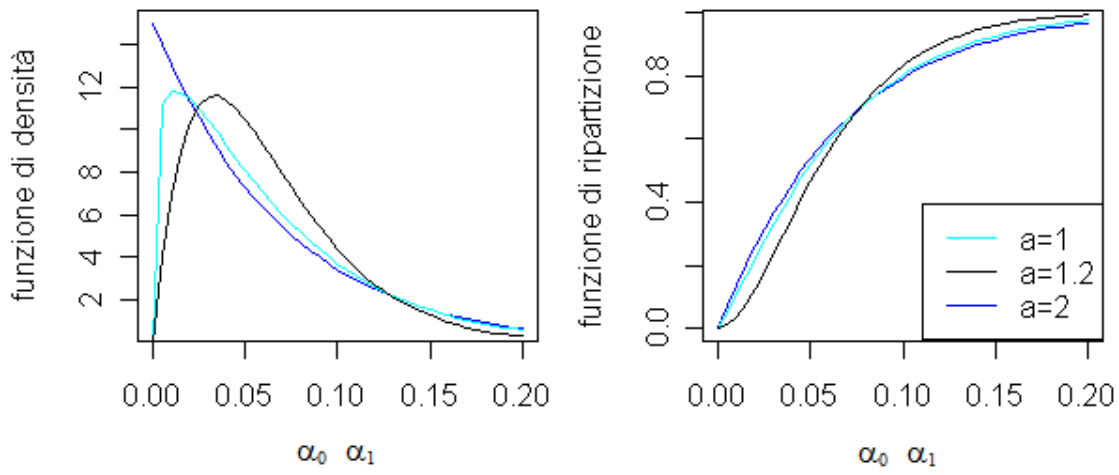
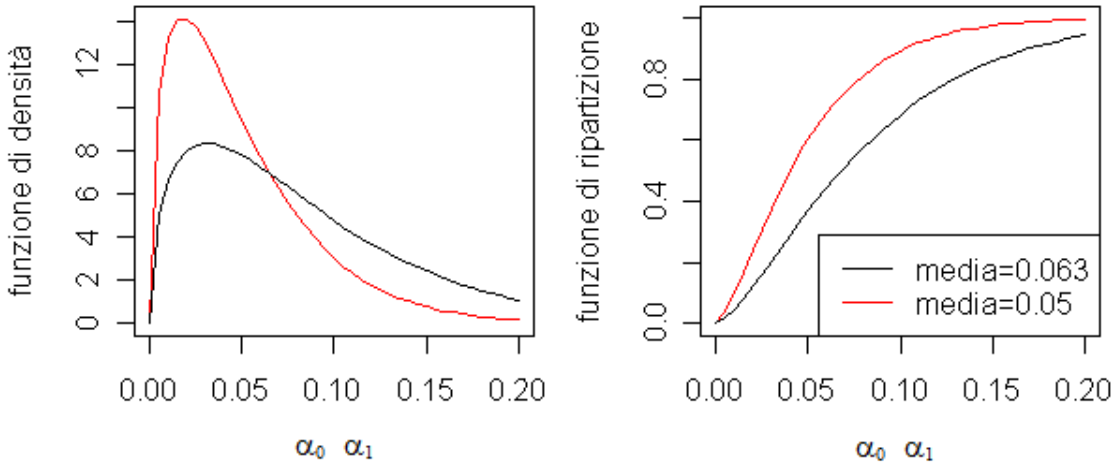


Figura 4.3. Funzioni di densità e ripartizione di una variabile aleatoria Beta, per vari valori della media, con il parametro *a* fissato a 1,5



La distribuzione Normale sulla trasformazione logit, provata con vari valori medi (quali -3.5, -4) e varie varianze (quali 0,5, 1), ha prodotto talvolta distribuzioni a posteriori molto diffuse, che estendevano il loro supporto anche su valori poco sensati per i parametri soprattutto per quanto riguarda α_1.

Quanto alla distribuzione Beta è opportuno porre attenzione alla scelta della sua forma, in quanto l'uso di distribuzioni con ridotta varianza, come B(2,38), portava a distribuzioni a posteriori stimate dei parametri α_0 e α_1 estremamente attratte dalla media della distribuzione 0,05, indipendentemente dai loro "veri valori".

Confrontando le distribuzioni B(1, 15) e B(1,5, 16,5) entrambe hanno mostrato un comportamento globalmente buono. La prima sembrava presentare migliori prestazioni nel caso di basse proporzioni di errore, al contrario della seconda. Il ricercatore potrà tenere conto nella scelta delle a priori delle sue conoscenze sull'ordine di grandezza delle proporzioni di errore nella popolazione in esame; non per fissare distribuzioni concentrate sui presunti valori di α_0 e α_1, ma per scegliere distribuzioni più performanti

nel range di tali presunti valori, seppur con un comportamento globalmente accettabile.

4.3.2 Specificazione del modello

La verosimiglianza corretta $f(x|\vartheta)$ coincide con la (3) presentata nel paragrafo 4.1. La distribuzione a posteriori dei parametri $\pi(\vartheta|x) \propto \pi(\vartheta) f(x|\vartheta)$ a causa della complessità del modello, non è analiticamente determinabile. Di conseguenza, verranno usati i metodi Markov Chain Monte Carlo, attraverso il software WinBUGS, graficamente o nel corrispondente linguaggio testuale; entrambe le specificazioni sono riportate nell'appendice A.

In ogni caso, il modello verrà espresso dividendo le quantità ignote in più livelli e specificando la distribuzione di ciscun livello condizionata a quello superiore.

In questo contesto nel quale si ipotizza la presenza di errori di misura, la variabile δ_i, ovvero lo stato riportato dall'individuo, non coincide più con A_i, variabile indicante il reale transito dell'individuo; δ_i è perciò osservabile, a differenza di A_i.

δ_i è una variabile aleatoria di Bernoulli con probabilità che dipendono, in maniera non stocastica, dai parametri α_1 e α_0 e dalla variabile A_i:

$$P[\delta_i = 1] = (1 - \alpha_1)A_i + \alpha_0(1 - A_i).$$

Delle distribuzioni aleatorie di α_1 e α_0 è stato discusso nel paragrafo precedente; i nodi A_i sono anch'essi stocastici e, non essendo quantità osservate, sono anch'essi dei parametri. Le variabili A_i costituiscono dunque un livello intermedio di parametri; sono generate dal medesimo modello probabilistico che nel modello senza errori era all'origine della variabile δ_i

(paragrafo 2.4.2). Le variabili A_i si distribuiscono cioè come variabili Bernoulli con probabilità P_i, funzione deterministica dei parametri α, β_0 e β, e delle variabili t_i e x_i che vengono fornite al programma come dati.

Tale modello è dunque specificato dalle seguenti distribuzioni di probabilità:

$$\alpha \sim \Gamma(2,2),$$

$$\beta_0 \sim N(-4,4),$$

$$\beta \sim N(0,1),$$

$$A_i \sim Bern\left(1 - \exp\left(\exp(x_i'\beta)\left(t_i^{\alpha} - (t_i + k)^{\alpha}\right)\right)\right),$$

$$\alpha_0 \sim Beta(1{,}5, 22{,}5),$$

$$\alpha_1 \sim Beta(1{,}5, 22{,}5),$$

$$\delta_i \sim Bern\left((1-\alpha_1)A_i + \alpha_0(1 - A_i)\right);$$

ed è perciò ascrivibile tra i Modelli Gerarchici Bayesiani.

Nonostante le distribuzioni a priori scelte per α_1 e α_0 assegnino ridottissima probabilità ai valori dello spazio parametrico non prossimi al valore 0, in alcuni campioni α_1 veniva stimato con valori molto alti, prossimi ad 1: una sorta di "label switching" (Redner e Walker, 1984), quindi, tale da indurre il modello ad una confusione tra gli individui che rispondono in maniera veritiera e quelli che rispondono in maniera errata. Dovrebbe essere particolarmente difficile per le serie raggiungere valori di questo tipo e mantenervisi per le successive iterazioni; rimanere in particolari regioni dello spazio parametrico, anche se improbabili, potrebbe essere conseguenza dell'eventuale uso dell'algortimo Gibbs Sampler da parte di WinBUGS, nel quale le simulazioni di ogni passo della serie si basano sul condizionamento ai valori assunti nel passo precedente.

Stime di proporzioni d'errore eccezionalmente alte porta naturalmente anche a stimare valori anomali per gli altri parametri: la dipendenza da durata

viene stimata positiva con valori di α di molto superiori all'unità, e l'effetto delle covariate di segno opposto a quello reale. Di conseguenza, qualora le stime siano affette da tale problematica, l'utente può avvedersene agevolmente e ripetere il processo di stima. Tuttavia, è possibile evitare che il fenomeno si verifichi semplicemente imponendo dei limiti in fase di stima ai valori simulati per i parametri α_1 e α_0; il software WinBUGS permette di specificare agevolmente questo tipo di vincoli ed essi equivalgono ad imporre ai due parametri delle densità a priori uguali a 0 fuori dall'intervallo di valori all'interno del quale si vuole consentire che essi varino. Nel nostro caso, si sceglie di impedire ai due parametri di superare il valore 0,25: livelli di errata classificazione superiori risultano poco ragionevoli e tali da rendere i dati inadatti per qualsiasi ulteriore analisi.

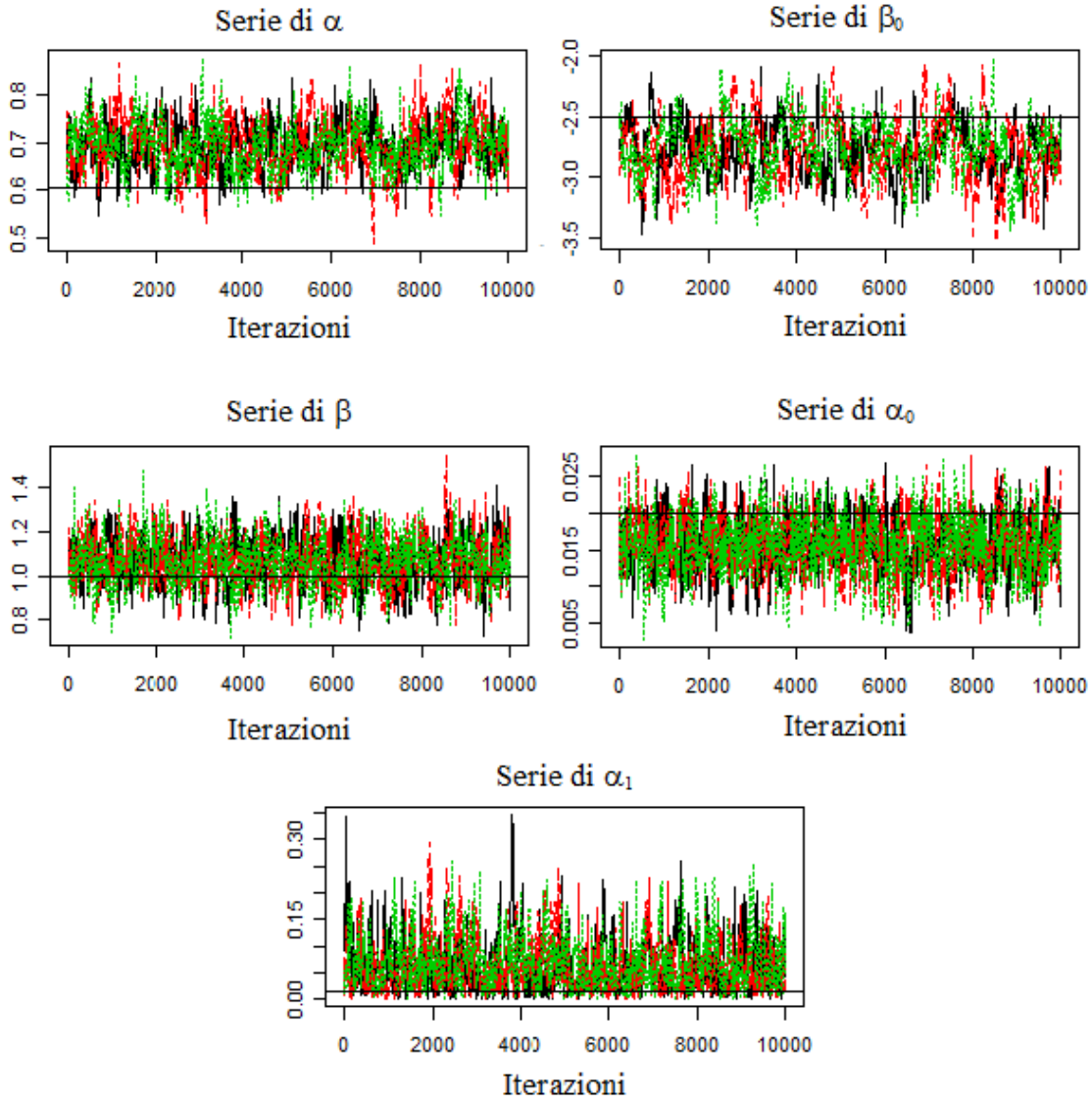
4.4 Stime effettuate tramite Modello Gerarchico Bayesiano

Il modello proposto viene applicato sui campioni affetti da errore di misura precedentemente utilizzati; sono state simulate tre catene parallele di 20.000 iterazioni ciascuna. Le tre catene sono state fatte partite dai quantili rispettivamente 0,025, 0,50 e 0,975 di una catena, simulata precedentemente, che consta di 5000 iterazioni. Le prime 10.000 interazioni fungono da periodo di burn-in; il *thinning interval* è di 5. La stima di tale modello è particolarmente impegnativa computazionalmente, anche perché tra i parametri da stimare sono presenti le variabili A_i, che sono in numero pari agli individui del campione.

4.4.1 Problematiche di convergenza ed autocorrelazione

Si valutano innanzitutto come di consueto, la convergenza, la stazionarietà e l'autocorrelazione delle serie usate per la stima delle distribuzioni a posteriori dei parametri.

Figura 4.4. Weibull con dipendenza da durata negativa: tre catene MCMC parallele per la stima del Modello Gerarchico Bayesiano; dati affetti da errore di misura al 2%



La Figura 4.4 e le Tabelle 4.2 e 4.3 sono relativi al campione con proporzione di errore 2% e dipendenza da durata negativa.

La Figura 4.4 mostra, per ciascun parametro, le tre catene MCMC parallele simulate, rappresentate con diversi colori; la retta di colore nero è posta all'altezza del "vero valore" del parametro.

Le serie relative alla stima dei parametri α_0 e α_1 hanno un comportamento globalmente buono, quanto ad autocorrelazione, che suggerisce sia stata raggiunta la convergenza; emerge tuttavia l'eccessiva variabilità della catena relativa al parametro α_1, se si osserva la scala del corrispondente grafico: tale catena infatti assume spesso anche valori poco ragionevoli per il parametro in esame.

Tabella 4.2. Weibull con dipendenza da durata negativa: autocorrelazioni delle catene MCMC per la stima del Modello Gerarchico Bayesiano; dati affetti da errori al 2%.

Ritardi	α	β_0	β	α_0	α_1
5	0,850	0,925	0,703	0,681	0,778
25	0,515	0,675	0,291	0,225	0,325
50	0,306	0,455	0,107	0,058	0,146
250	0,032	0,011	-0,028	-0,001	-0,048

Tabella 4.3. Weibull con dipendenza da durata negativa: crosscorrelazioni tra le catene MCMC per la stima del Modello Gerarchico Bayesiano; dati affetti da errori al 2%.

	α	β_0	β	α_0	α_1
α	1	-0,827	0,044	-0,259	-0,038
β_0		1	-0,402	-0,186	0,261
β			1	0,680	0,103
α_0				1	0,058
α_1					1

Le serie relative ai parametri α e β_0 mantengono le loro problematiche, quali elevata autocorrelazione ed identificabilità difficoltosa. Le autocorrelazioni a ritardo 5 per i parametri β, α_0 e α_1 sono abbastanza elevate e hanno valori

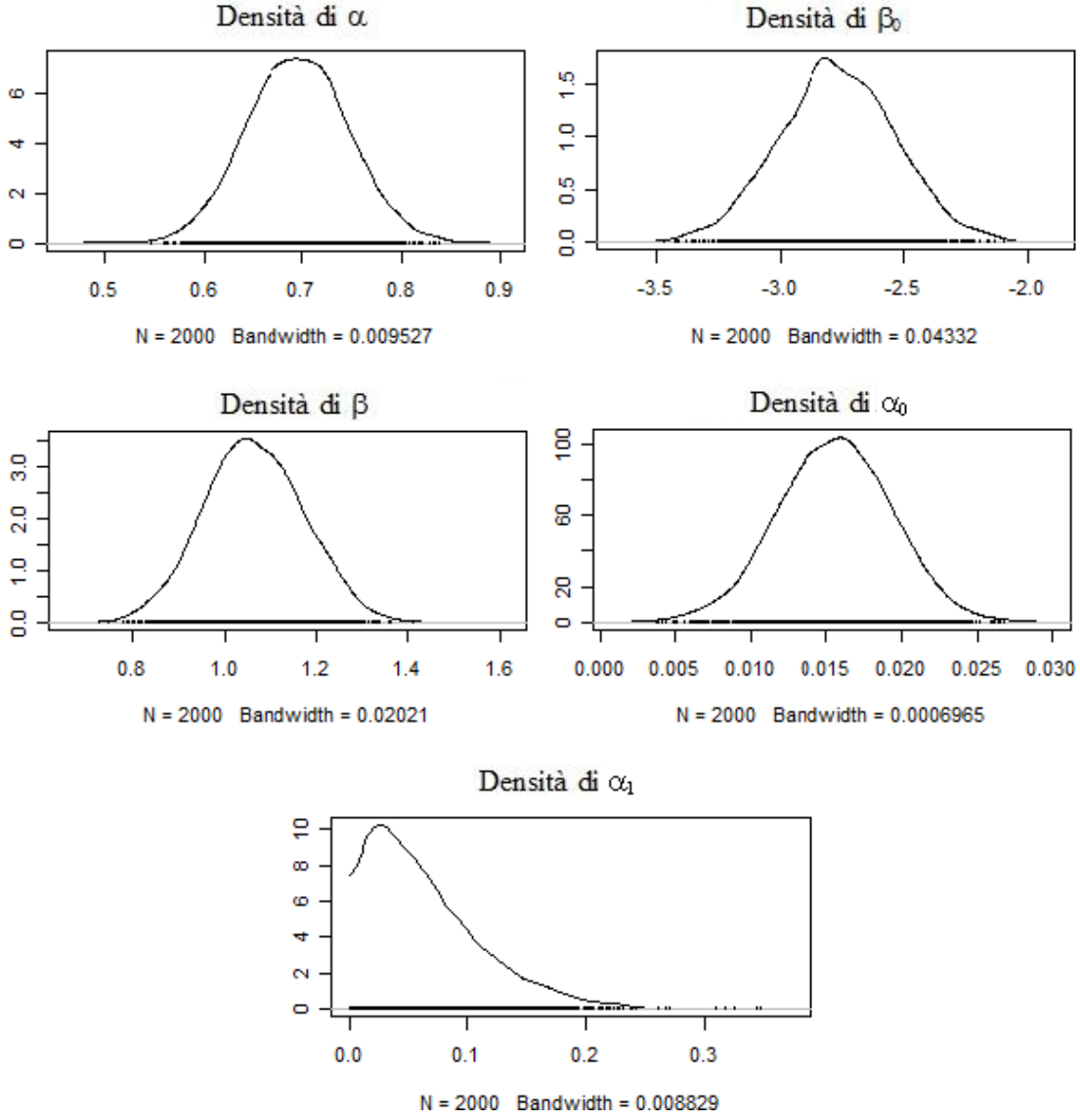
intorno allo 0,7, tuttavia ai ritardi 25 e 50 esse diventano trascurabili. Le autocorrelazioni della prima coppia di parametri, invece, assumono valori prossimi all'unità e paragonabili a quelli osservati nei processi di stima precedentemente effettuati tramite Modello Bayesiano. Permane la forte dipendenza negativa tra le serie relative ad α e β_0, anche se in misura leggermente minore. Tra le altre catene non emergono crosscorrelazioni particolarmente significative, ad eccezione della coppia di parametri β e α_0; difatti, maggiore sarà la proporzione di errori, maggiore sarà l'entità dell'attenuazione che essi provocano sulle stime di β e, quindi, la correzione di tale attenuazione che il modello di trova ad operare: la correlazione non è comunque tale da rendere problematica la convergenza delle serie.
Esaminando gli altri campioni esaminati si nota che al crescere del parametro α emergono maggiormente i segnali legati a difficoltà nell'identificazione, quali le elevate autocorrelazioni anche a ritardi molto ampi, la forte dipendenza tra coppie di parametri, valori significativi nelle diagnostiche di Geweke e Gelman.

4.4.2 Distribuzioni a posteriori stimate

La Figura 4.5 è relativa alle distribuzioni a posteriori stimate dal modello ed i grafici sono realizzati a partire dalle serie simulate tramite metodo Kernel.
La stima della distribuzione a posteriori per il parametro α_0 ha prodotto un esito molto positivo, in quanto tale distribuzione si discosta nettamente dalla relativa a priori, ha una variabilità molto ridotta ed è centrata su un valore molto vicino a quello reale del parametro.
Lo stesso non si può dire della distribuzione a posteriori di α_1: come emerso dai grafici relativi alle serie, infatti, essa ha una variabilità decisamente

molto ampia ed il suo supporto si estende su regioni dello spazio campionario poco realistiche. Inoltre, il suo valore centrale rimane 0,06, il medesimo della distribuzione a priori.

Figura 4.5. Weibull con dipendenza da durata negativa: distribuzioni a posteriori dei parametri stimate tramite Modello Gerarchico Bayesiano; dati affetti da errori al 2%



Tale parametro non appare dunque identificato e ciò avviene probabilmente a causa dello scarso impatto che esso ha sulla stima dei parametri (paragrafo 3.3): tale impatto è limitato rispetto a quello di α_0 per la ridotta proporzione

di transitati presenti nel campione in esame. Le difficoltà nell'identificazione di α_1 non appaiono dunque come strutturali del modello, ma dipendenti dalla proporzione di transitati e, di conseguenza, dall'ambito di applicazione del modello e dalla specifica popolazione in esame.

Proprio per lo scarso impatto del parametro, le difficoltà nell'identificazione di α_1 non sembrano compromettere l'identificazione e la stima delle distribuzioni a posteriori nei restanti parametri. Esse si discostano nettamente dalle corrispondenti a priori, e lo fanno muovendosi nella giusta direzione: le stime appaiono pressoché corrette e la loro variabilità non eccessiva. Nella Tabella 4.4 riportiamo le stime di massima verosimiglianza (1) ottenute sui medesimi dati (paragrafo 3.3), per confrontarle con quelle ottenute tramite il modello Gerarchico Bayesiano, contenute nella medesima Tabella: delle distribuzioni a posteriori ottenute verranno mostrati media, standard deviation, quartili e quantili 0,025 e 0,975.

Tabella 4.4. Weibull con dipendenza da durata negativa: indici di sintesi delle distribuzioni a posteriori stimate tramite Modello Gerarchico Bayesiano e stime di massima verosimiglianza (1) su dati affetti da errori al 2%

		MV		Modello bayesiano						
par.	vero	stima	se	mean	sd	2,5%	25%	50%	75%	97,5%
α	0,606	0,737	0,039	0,695	0,051	0,595	0,660	0,695	0,729	0,796
β_0	-2,500	-2,641	0,174	-2,771	0,235	-3,230	-2,925	-2,774	-2,613	-2,313
β	1,000	0,744	0,054	1,065	0,109	0,856	0,990	1,061	1,138	1,278
α_0	0,020			0,016	0,004	0,008	0,013	0,016	0,018	0,023
α_1	0,020			0,065	0,049	0,005	0,028	0,054	0,091	0,185

Si noti come il modello Gerachico Bayesiano riesca a riportare le stime dei parametri verso il loro vero valore; in questo caso fa eccezione il parametro β_0, la cui distorsione subita a causa della misclassification non appare comunque significativa. Le deviazioni standard delle distribuzioni a posteriori appaiono leggermente maggiori rispetto agli standard error delle stime di massima verosimiglianza; nei precedenti esempi, qualora sugli

stessi campioni venivano applicate stime di massima verosimiglianza e Modello Bayesiano, gli standard error dell'una e le standard deviation dell'altro erano pressochè coincidenti.

Ora invece si registra questa maggiore variabilità nel Modello Gerarchico Bayesiano, a causa della sua maggiore complessità e della concomitante stima dei parametri α_0, α_1 e A_i.

Un intervallo di credibilità con significatività 0,95 per ciascun parametro può avere come estremi i quantili 0,025 e 0,975. Alternativamente, poiché dai grafici precedenti emerge l'approssimativa normalità delle distribuzioni a posteriori, ad esclusione del parametro α_1, è possibile costruire gli intervalli tramite approssimazione normale, utilizzando media e deviazione standard delle serie.

Gli intervalli costruiti attraverso questi metodi sono estremamente simili, ad ulteriore conferma dell'approssimativa normalità delle distribuzioni a posteriori, sempre escludendo il parametro α_1. In ciascun campione esaminato, i "valori veri" dei parametri appartengono agli intervalli di credibilità, indipendentemente da come questi siano stati costruiti.

Si mostrano nelle Figure 4.6 e 4.7 e nelle Tabelle 4.5 e 4.6 i risultati del modello sui campioni relativi a dipendenza da durata nulla e positiva.

Come nel caso della dipendenza da durata negativa, emerge positivamente la capacità del modello di stimare distribuzioni a posteriori per α_0 particolarmente concentrate e collocate intorno al suo "valore esatto"; riguardo al parametro α_1 i risultati rimangono non accettabili. Anche per i restanti parametri si ottengono distribuzioni a posteriori che si spostano nella corretta direzione rispetto alle loro a priori e soprattutto rispetto alle stime ottenute tramite massima verosimiglianza e non sono eccessivamente disperse.

Figura 4.6. Weibull con dipendenza da durata nulla: distribuzioni a posteriori dei parametri stimate tramite Modello Gerarchico Bayesiano; dati affetti da errori al 2%

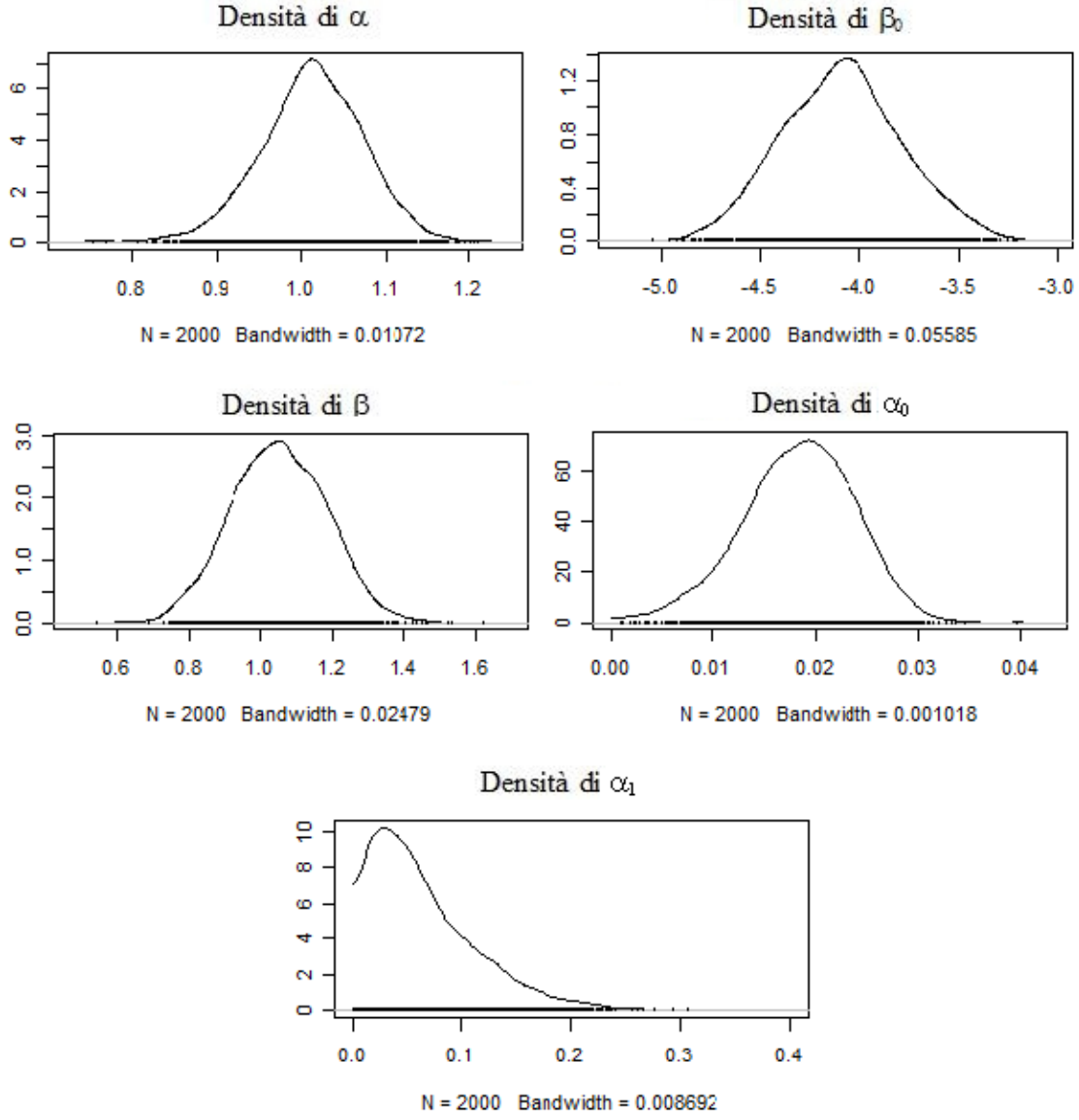


Tabella 4.5. Weibull con dipendenza da durata nulla: indici di sintesi delle distribuzioni a posteriori stimate tramite Modello Gerarchico Bayesiano e stime di massima verosimiglianza (1) su dati affetti da errori al 2%

		MV		Modello bayesiano						
par.	vero	stima	se	mean	sd	2,5%	25%	50%	75%	97,5%
α	1,000	1,006	0,043	1,014	0,059	0,891	0,977	1,015	1,054	1,127
β_0	-4,000	-3,723	0,183	-4,092	0,300	-4,663	-4,302	-4,092	-3,893	-3,487
β	1,000	0,726	0,051	1,050	0,133	0,793	0,956	1,050	1,144	1,306
α_0	0,020			0,018	0,005	0,006	0,015	0,019	0,022	0,028
α_1	0,020			0,066	0,050	0,005	0,029	0,054	0,091	0,192

Figura 4.7. Weibull con dipendenza da durata positiva: distribuzioni a posteriori dei parametri stimate tramite Modello Gerarchico Bayesiano; dati affetti da errori al 2%

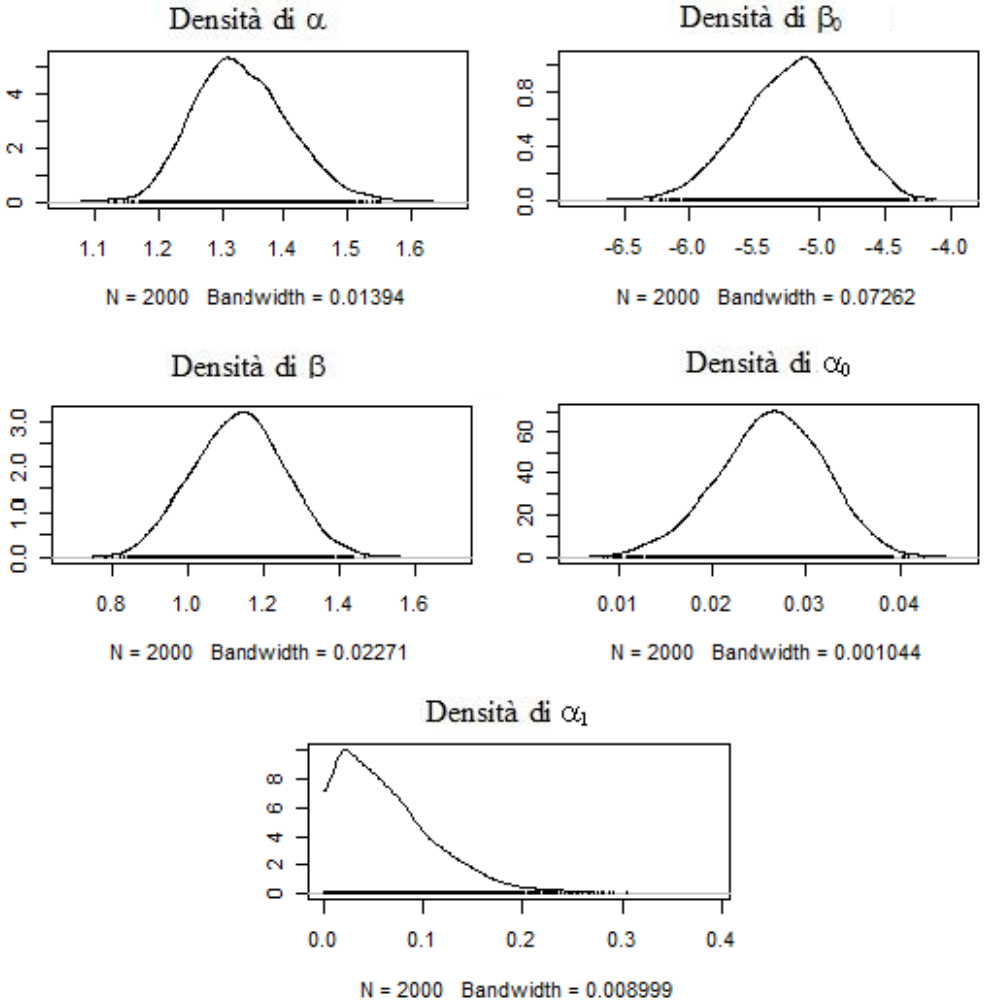


Tabella 4.6. Weibull con dipendenza da durata positiva: indici di sintesi delle distribuzioni a posteriori stimate tramite Modello Gerarchico Bayesiano e stime di massima verosimiglianza (1) su dati affetti da errori al 2%

		MV		Modello bayesiano						
par.	vero	stima	se	mean	sd	2,5%	25%	50%	75%	97,5%
α	1,284	1,236	0,043	1,332	0,075	1,200	1,279	1,327	1,381	1,487
β_0	-5,000	-4,465	0,179	-5,222	0,390	-6,016	-5,479	-5,198	-4,952	-4,502
β	1,000	0,808	0,046	1,135	0,122	0,899	1,050	1,136	1,217	1,373
α_0	0,020			0,026	0,006	0,015	0,023	0,026	0,030	0,037
α_1	0,020			0,067	0,051	0,005	0,027	0,055	0,092	0,194

Le distribuzioni a posteriori, sempre con l'eccezione di α_1, hanno distribuzione approssimativamente normale; possono quindi venir costruiti

intervalli di credibilità nelle due modalità, tra loro pressoché coincidenti, e si osserva che ad ognuno di essi appartiene il "vero valore" del parametro.

Il modello è stato applicato in maniera analoga anche sui campioni con proporzione di risposte errate 7%. Nelle Figure 4.8, 4.9 e 4.10 e nelle Tabelle 4.7, 4.8 e 4.9 viene riportato quanto ottenuto per i tre campioni nelle tre diverse tipologie di dipendenza da durata.

Nei tre campioni proposti, il primo parametro di misclassification α_0 è ancora una volta stato stimato efficacemente: il modello appare dunque capace di stimare distribuzioni a posteriori per tale parametro che si distaccano nettamente dalla distribuzione a priori in entrambe le direzioni. La distribuzione a posteriori ottenuta, in entrambi i casi, ha una ridotta variabilità e forma approssimativamente normale.

Le distribuzioni a posteriori di α_1, al contrario, sono del tutto coincidenti a quelle ottenute nei campioni precedenti, nei quali il parametro assumeva valore 0,02: la variabilità rimane eccessivamente ampia, e la media coincide ancora con quella della distribuzione a priori per il parametro.

Le stime dei restanti parametri sono globalmente molto buone, in quanto si sono spostate nella direzione opposta rispetto alle distorsioni che avvenivano nel caso di stima di massima verosimiglianza. La loro variabilità è leggermente aumentata rispetto ai corrispondenti campioni con proporzione di errore più ridotta, ma rimane comunque contenuta. Ancora, costruendo degli intervalli di credibilità a livello 0,95 tutti i "veri parametri" vi appartengono, a differenza di quanto accadeva con la stima di massima verosimiglianza.

Figura 4.8. Weibull con dipendenza da durata negativa: distribuzioni a posteriori dei parametri stimate tramite Modello Gerarchico Bayesiano; dati affetti da errori al 7%

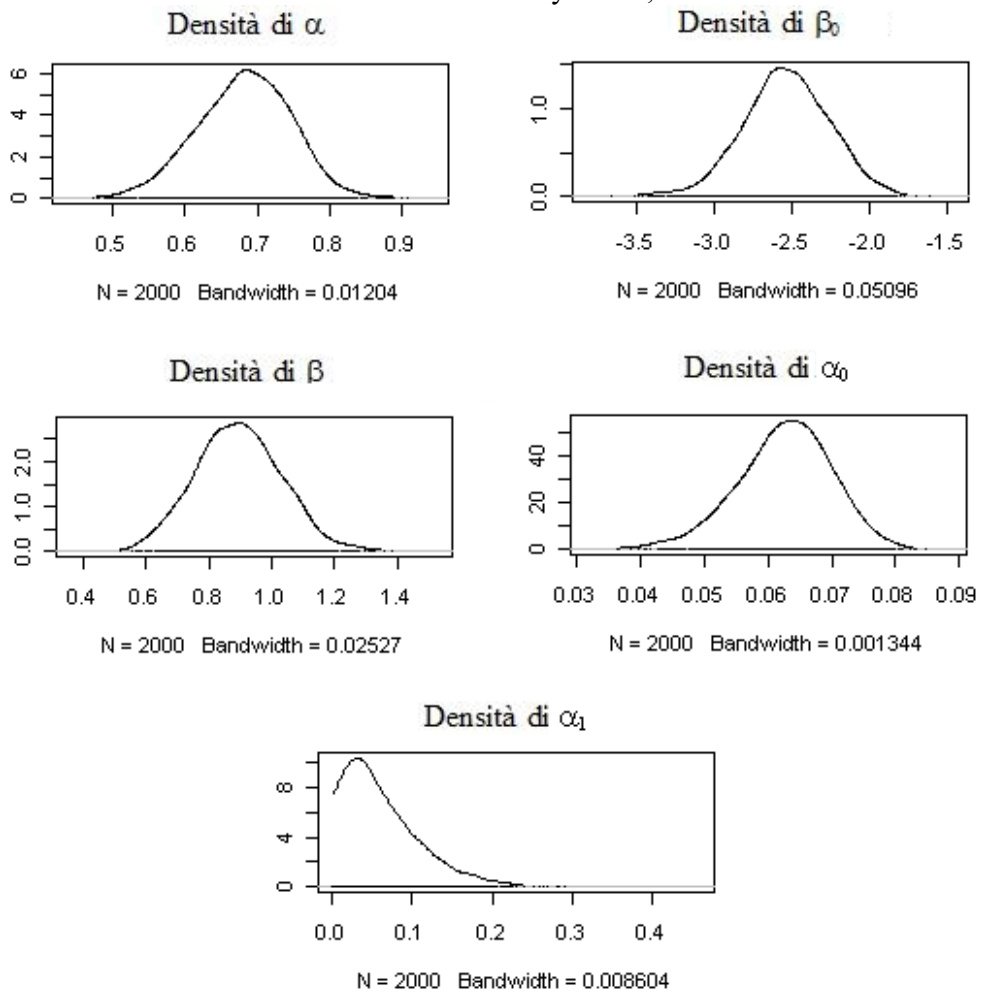


Tabella 4.7. Weibull con dipendenza da durata negativa: indici di sintesi delle distribuzioni a posteriori stimate tramite Modello Gerarchico Bayesiano e stime di massima verosimiglianza (1) su dati affetti da errori al 7%

		MV		Modello bayesiano						
par.	vero	stima	se	mean	sd	2,5%	25%	50%	75%	97,5%
α	0,606	0,818	0,030	0,682	0,065	0,551	0,639	0,685	0,727	0,803
β_0	-2,500	-2,414	0,133	-2,535	0,278	-3,095	-2,713	-2,536	-2,346	-2,004
β	1,000	0,315	0,039	0,895	0,136	0,641	0,803	0,893	0,985	1,171
α_0	0,070			0,063	0,007	0,046	0,058	0,063	0,068	0,076
α_1	0,070			0,065	0,050	0,005	0,028	0,053	0,090	0,187

Figura 4.9. Weibull con dipendenza da durata nulla: distribuzioni a posteriori dei parametri stimate tramite Modello Gerarchico Bayesiano; dati affetti da errori al 7%

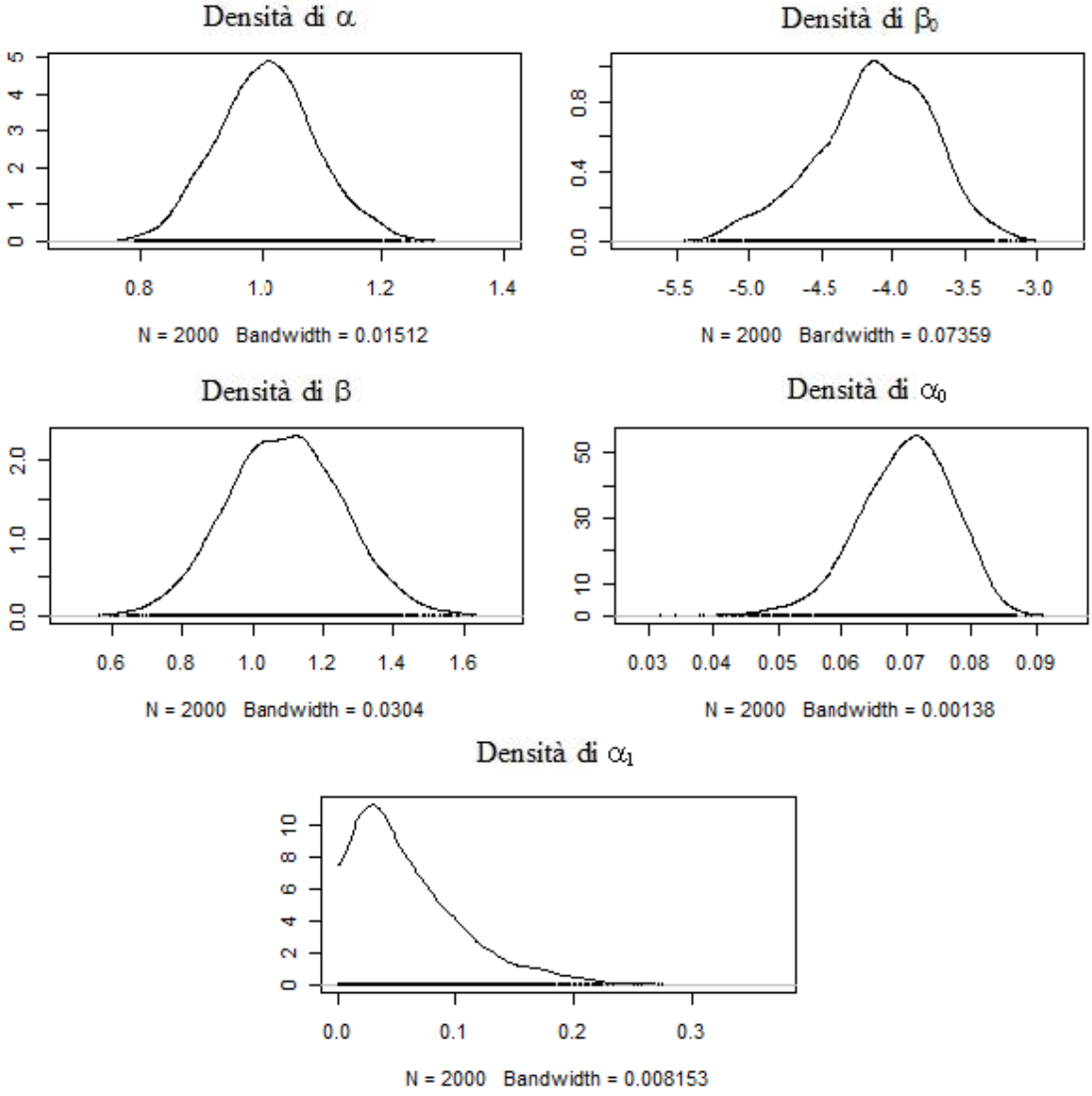


Tabella 4.8. Weibull con dipendenza da durata nulla: indici di sintesi delle distribuzioni a posteriori stimate tramite Modello Gerarchico Bayesiano e stime di massima verosimiglianza (1) su dati affetti da errori al 7%

		MV		Modello bayesiano						
par.	**vero**	**stima**	**se**	**mean**	**sd**	**2,5%**	**25%**	**50%**	**75%**	**97,5%**
α	1,000	1,010	0,032	1,009	0,083	0,853	0,953	1,008	1,062	1,181
β_0	-4,000	-3,164	0,137	-4,131	0,403	5,020	-4,372	-4,105	-3,842	-3,414
β	1,000	0,388	0,037	1,092	0,163	0,777	0,979	1,093	1,204	1,417
α_0	0,070			0,070	0,007	0,054	0,065	0,070	0,075	0,083
α_1	0,070			0,062	0,048	0,005	0,026	0,050	0,085	0,184

Figura 4.10. Weibull con dipendenza da durata positiva: distribuzioni a posteriori dei parametri stimate tramite Modello Gerarchico Bayesiano; dati affetti da errori al 7%

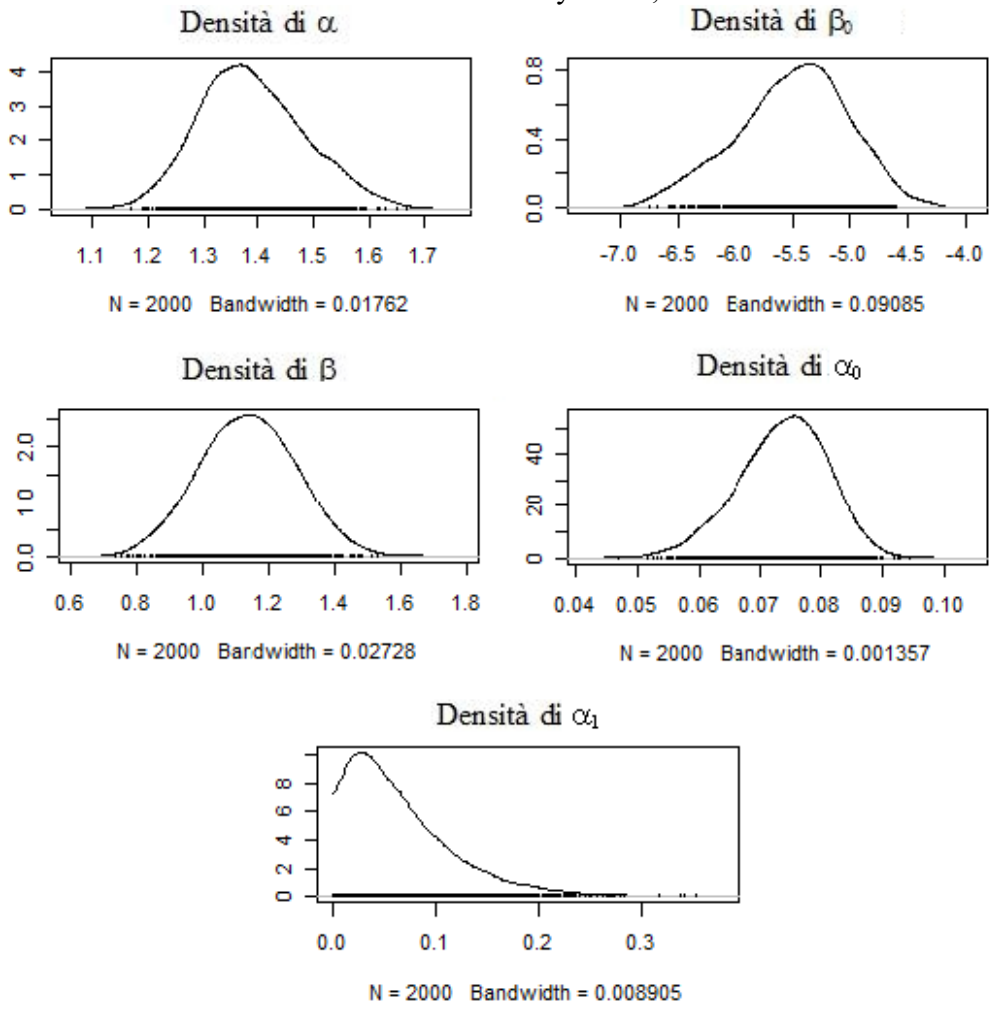


Tabella 4.9. Weibull con dipendenza da durata positiva: indici di sintesi delle distribuzioni a posteriori stimate tramite Modello Gerarchico Bayesiano e stime di massima verosimiglianza (1) su dati affetti da errori al 7%

		MV		Modello bayesiano						
par.	vero	stima	se	mean	sd	2,5%	25%	50%	75%	97,5%
α	1,284	1,156	0,033	1,388	0,095	1,219	1,321	1,381	1,450	1,588
β_0	-5,000	-3,674	0,138	-5,507	0,489	-6,541	-5,820	-5,468	-5,166	-4,646
β	1,000	0,481	0,036	1,137	0,147	0,852	1,035	1,137	1,238	1,424
α_0	0,070			0,074	0,007	0,058	0,069	0,074	0,079	0,087
α_1	0,070			0,067	0,052	0,005	0,028	0,054	0,092	0,203

4.5 Stime ripetute su insiemi di 100 campioni

Da quanto emerge dall'applicazione del Modello Gerarchico Bayesiano sui sei campioni in esame, l'efficacia del modello nel produrre stime generalmente accettabili, nonostante la presenza degli errori e la difficile identificabiltà di alcuni parametri, è da ritenersi globalmente soddisfacente. Tuttavia, è necessario chiedersi se eventuali inesattezze nelle stime siano da imputare alla particolarità dei dati in esame, o al metodo di stima stesso o, ad esempio, se esso fornisca risultati accettabili su qualsiasi campione.

Si opta quindi per un disegno sperimentale che prevede la simulazione di un numero elevato di campioni e, per ciascuno di essi, la valutazione di vari modelli sotto vari livelli di misclassification. Alcuni risultati emersi da tale procedimento sono stati già discussi nei precedenti capitoli. Si riassumono qui i principali risultati.

Sono stati simulati 100 campioni, di numerosità pari a 1000 ciascuno, con una sola covariata X con distribuzione normale, tempo k tra le due successive interviste pari a 3 mesi, non affetti da errore di misura; per i metodi usati per la simulazione si veda il paragrafo 1.2.1. Su ogni campione simulato sono state stimate le distribuzioni a posteriori dei parametri α, β_0 e β tramite il modello Bayesiano descritto nel paragrafo 2.4; tramite WinBUGS è stata simulata una catena da 20.000 iterazioni; per il burn-in sono state utilizzate le prime 10.000 delle iterazioni ed il *thinning interval* è stato posto a 5, di conseguenza si sono ottenute catene di lunghezza 2000.

Di ogni serie simulata sono stati calcolati media e quartili; per ogni parametro è stato inoltre creato un intervallo di credibilità HPD a livello 0,8 ed è stata riportata una variabile dicotomica che indica l'effettiva appartenenza del “vero valore” del parametro a tale intervallo. Sono state

inoltre calcolate la crosscorrelazione e le autocorrelazioni a ritardo 5 dei parametri α e β_0. I risultati di tale procedimento sono stati riportati nel paragrafo 2.6.

Per ciascuno dei campioni in esame, è stata in un secondo momento simulata la variabile δ_i, con proporzioni di errata classificazione pari a 0,02 e 0,07 per entrambi le sottopopolazioni dei transitati e non. Sui campioni così ottenuti è stato applicato il medesimo modello Bayesiano e sono state effettuate le medesime analisi appena riportate; i loro risultati sono stati pesentati nel paragrafo 3.6.

Prendendo inizialmente in considerazione soltanto le stime puntuali di ogni parametro ottenute come media della distribuzione a posteriori stimata, il Modello Bayesiano ha prodotto stime piuttosto disperse all'interno dei gruppi di 100 campioni, indipendentemente dalla presenza e dall'entità dell'errore di misura. I valori centrali delle distribuzioni a posteriori si distanziano dal vero valore del parametro al crescere della proporzione di errore; le direzioni nelle quali vengono distorte le stime vengono trattate nel paragrafo 3.4.

La difficile indentificabilità dei parametri α e β_0 viene riscontrata in tutti i tipi di campioni cui viene applicato il modello: le autocorrelazioni a ritardo 5 e la crosscorrelazione dei due parametri sono estremamente vicine rispettivamente a 1 e -1, con ridottissima variabilità.

Inoltre, è stato costruito per ogni campione un indice che tenga conto delle autocorrelazioni a ritardo maggiore di 5, per valutare se la catena abbia avuto comportamenti problematici nella convergenza; per le modalità di costruzione di tale indice si rimanda al paragrafo 2.6.2. Si riscontrano autocorrelazione a ritardi elevati crescenti al crescere del parametro α; questa tendenza è indipendente dall'eventuale presenza di errori di misura nei dati.

Sugli stessi campioni sono poi state effettuate stime tramite il Modello Gerarchico Bayesiano: nel seguito si presentano e analizzano autocorrelazioni e crosscorrelazione per α e β_0, medie e quartili delle distribuzioni a posteriori e proporzioni dell'appartenenza del "vero parametro" all'intervallo HPD a livello 0,80.

4.5.1 Autocorrelazione delle serie

Si riportano nella Tabella 4.10, per i due pattern di misclassification proposti, la media e la standard deviation, su ciascun insieme di campioni, dei valori di autocorrelazione a ritardo 5 e crosscorrelazione per i parametri α e β_0.

Tabella 4.10. Medie e sd di autocorrelazione e crosscorrelazione delle serie di α e β_0, su gruppi di 100 campioni con varie tipologie di dipendenza da durata e affetti da errore di misura in varie proporzioni

	negativa		nulla		positiva	
	media	sd	media	sd	media	sd
			$\alpha_0 = \alpha_1 = 0{,}02$			
autocorrelazione α	0,823	0,042	0,856	0,042	0,889	0,029
autocorrelazione β_0	0,919	0,017	0,943	0,019	0,956	0,014
crosscorrelazione $\alpha\,\beta_0$	-0,815	0,079	-0,823	0,105	-0,882	0,075
			$\alpha_0 = \alpha_1 = 0{,}07$			
autocorrelazione α	0,829	0,058	0,857	0,051	0,881	0,037
autocorrelazione β_0	0,933	0,025	0,947	0,025	0,958	0,019
crosscorrelazione $\alpha\,\beta_0$	-0,682	0,133	-0,705	0,146	-0,731	0,174

Le problematiche legate ai due parametri permangono anche nel Modello Gerarchico Bayesiano: le autocorrelazioni del parametro α, tuttavia, appaiono leggermente attenuate ripetto a quelle registrate nel precedente modello; anche la correlazione negativa tra le serie corrispondenti ai due parametri, pur rimanendo forte non raggiunge le immediate vicinanze del valore -1; aumenta, inoltre, la variabilità di tutti gli indici proposti, pur

rimanendo molto ridotta. I valori di sintesi sembrano confermare quanto precedentemente ipotizzato: l'identificabilità dei due parametri tende a diventare più problematica al variare della tipologia di dipendenza da durata, in particolare al crescere del parametro α. Al crescere dell'errore di misura, invece, si riscontra una leggera diminuzione nella correlazione negativa tra le serie dei due parametri.

Nel caso del Modello Gerachico Bayesiano si riscontrano valori dell'indice riguardante le autocorrelazioni a ritardi maggiori di 5 più alti rispetto alle applicazioni del Modello Bayesiano senza errori sul medesimo gruppo di campioni: è ragionevole infatti supporre che la maggior complessità del modello richieda uno sforzo computazionale maggiore.

Si nota, ancora una volta, la tendenza dell'indice ad aumentare al variare della tipologia di dipendenza da durata. Inoltre, problematiche di elevata autocorrelazione tendono a diventare più rilevanti anche al crescere della proporzione dell'errore di misura nei dati.

Queste indicazioni possono essere usate dall'utente per la scelta del numero di iterazioni e del *thin interval* più opportuni per la stima del modello sui dati in esame: non essendoci problemi di convergenza, per ottenere dalla distribuzione a posteriori dei parametri un campione approssimativamente indipendente è sufficiente estrarre valori lontani tra loro nelle serie generate dal metodo MCMC.

4.5.2 Distribuzioni a posteriori stimate

Nel seguito (Figure 4.11-20 e Tabelle 4.11-16) per ciascuna delle tipologie di dipendenza da durata esaminate, vengono riportati:

- per ciascun parametro, boxplot delle stime puntuali sui 100 campioni in esame, ottenute come medie della distribuzioni a posteriori simulate. Per

avere una visione più immediata delle prestazioni del modello, il "vero valore" del parametro è stato rappresentato nei grafici tramite una retta rossa. (Non sono stati riportati ovunque i grafici relativi al parametro α_1, in quanto essi, indipendentemente dal valore del parametro e dalla dipendenza da durata, mostrano una distribuzione simmetrica e molto concentrata intorno al valore 0,06 che corrisponde al valore medio della distribuzione a priori.)

- nel caso della dipendenza da durata negativa, i boxplot relativi ad α e β vengono affiancati dagli analoghi rappresentanti le stime puntuali effettuate tramite Modello Bayesiano su dati rispettivamente non affetti e affetti da errore di misura; inoltre i boxplot relativi ad α_0 e α_1 sono stati affiancati e riprodotti sulla stessa scala, al fine di visualizzare meglio le differenze nella prestazione del modello per le stime dei due parametri
- tabelle contenenti media e deviazione standard di vari indicatori registrati per ciascun campione, quali medie e quartili delle distribuzioni a posteriori stimate tramite Modello Gerarchico Bayesiano. Le tabelle contengono inoltre la proporzione di campioni nei quali il "vero valore" del parametro risultava appartenere al corrispondente intervallo HPD.
- per confronto si riportano gli analoghi risultati ottenuti sui medesimi campioni tramite Modello Bayesiano senza stima degli errori.

Tutte le Figure e le Tabelle vengono proposte prima per i campioni con proporzione d'errore 0,02, e di seguito per i campioni con proporzione d'errore 0,07.

Figura 4.11 Weibull con dipendenza da durata negativa: stime puntuali (media campionaria della distribuzione a posteriori) di α tramite Modello Bayesiano (A) e, su dati affetti da errore al 2%, Modello Bayesiano (B) e Modello Gerarchico Bayesiano (C)

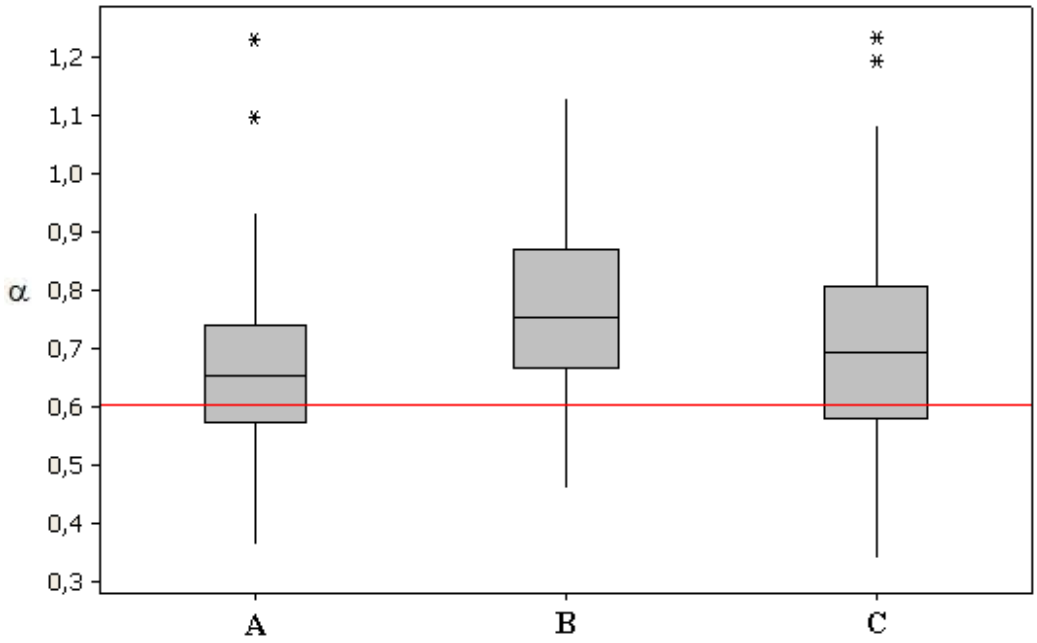


Figura 4.12 Weibull con dipendenza da durata negativa: stime puntuali (media campionaria della distribuzione a posteriori) di β tramite Modello Bayesiano (A) e, su dati affetti da errore al 2%, Modello Bayesiano (B) e Modello Gerarchico Bayesiano (C)

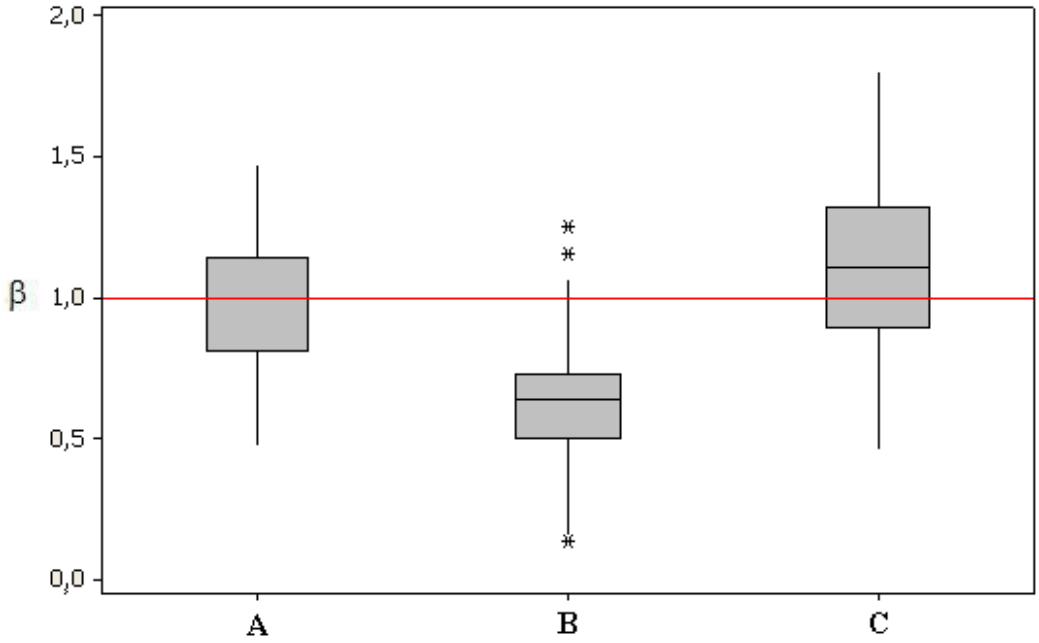


Figura 4.13 Weibull con dipendenza da durata negativa: stime puntuali (media campionaria della distribuzione a posteriori) di α_0 e α_1 tramite Modello Gerarchico Bayesiano su dati affetti da errore al 2%

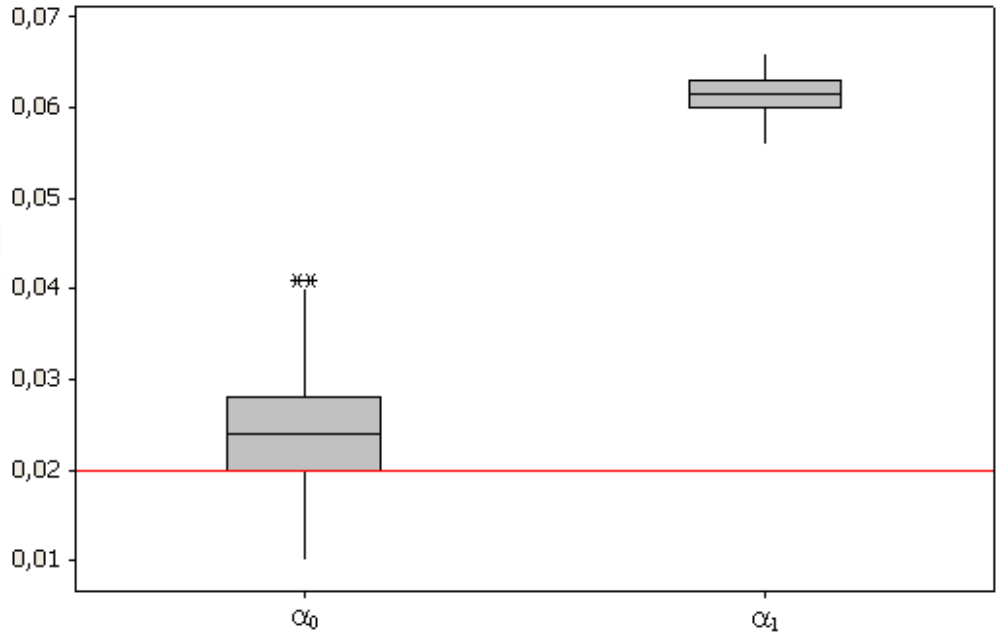


Figura 4.14 Weibull con dipendenza da durata negativa: stime puntuali (media campionaria della distribuzione a posteriori) di α tramite Modello Bayesiano (A) e, su dati affetti da errore al 7%, Modello Bayesiano (B) e Modello Gerarchico Bayesiano (C)

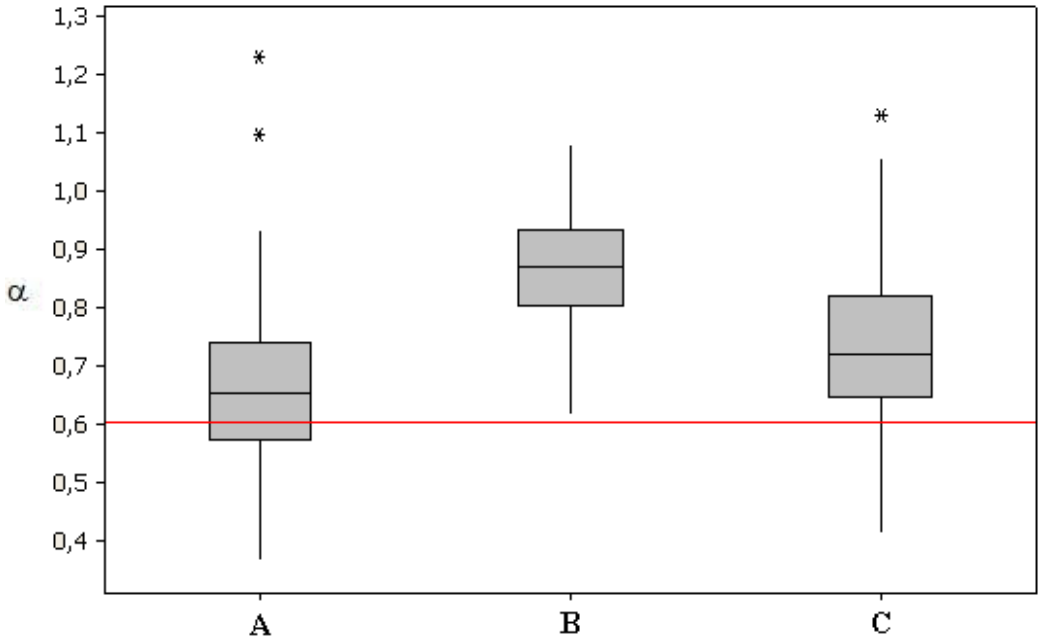


Figura 4.15 Weibull con dipendenza da durata negativa: stime puntuali (media campionaria della distribuzione a posteriori) di α tramite Modello Bayesiano (A) e, su dati affetti da errore al 2%, Modello Bayesiano (B) e Modello Gerarchico Bayesiano (C)

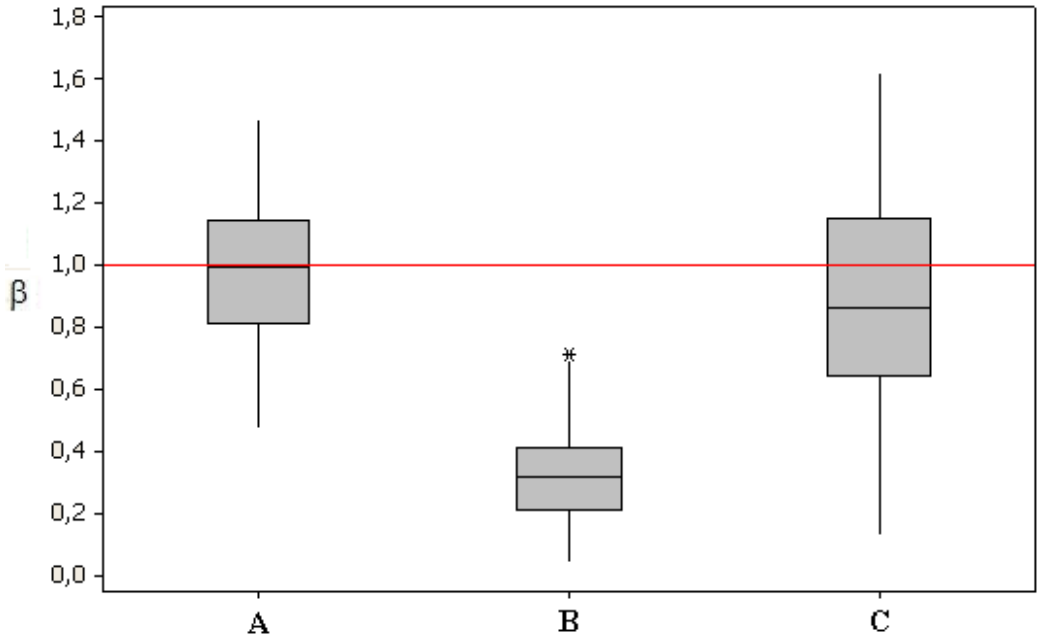


Figura 4.16 Weibull con dipendenza da durata negativa: stime puntuali (media campionaria della distribuzione a posteriori) di α_0 e α_1 tramite Modello Gerarchico Bayesiano su dati affetti da errore al 7%

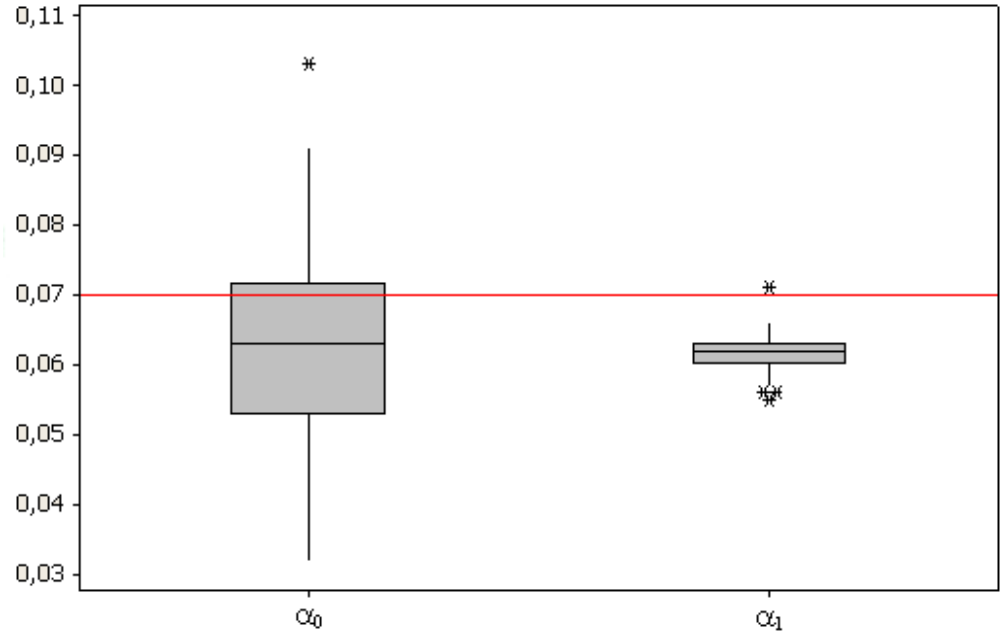


Tabella 4.11 Weibull con dipendenza da durata negativa: media e sd di indici di sintesi delle distribuzioni a posteriori stimate tramite Modello Bayesiano e Modello Gerarchico Bayesiano su dati affetti da errore al 2%

		α		β_0		β					
		media	**sd**	**media**	**sd**	**media**	**sd**				
Modello Bayesiano	media	0,765	0,128	-2,795	0,583	0,639	0,221				
	I quartile	0,683	0,125	-3,157	0,600	0,522	0,219				
	mediana	0,762	0,128	-2,790	0,579	0,638	0,221				
	III quartile	0,844	0,133	-2,430	0,567	0,755	0,224				
	copertura	0,520		0,670		0,200					
	vero valore	*0,606*		*-2,500*		*1,000*		*0,020*		*0,020*	
		α		β_0		β		α_0		α_1	
		media	**sd**	**media**	**sd**	**media**	**sd**	**media**	**sd**	**media**	**sd**
MGB	media	0,698	0,165	-3,202	0,780	1,112	0,288	0,024	0,007	0,062	0,002
	I quartile	0,572	0,163	-3,745	0,837	0,841	0,271	0,017	0,007	0,026	0,001
	mediana	0,697	0,168	-3,147	0,760	1,082	0,291	0,024	0,007	0,050	0,002
	III quartile	0,822	0,173	-2,599	0,718	1,351	0,322	0,031	0,008	0,085	0,003
	copertura	0,800		0,750		0,910		0,870		1,000	

Tabella 4.12 Weibull con dipendenza da durata negativa: media e sd di indici di sintesi delle distribuzioni a posteriori stimate tramite Modello Bayesiano e Modello Gerarchico Bayesiano su dati affetti da errore al 7%

		α		β_0		β					
		media	**sd**	**media**	**sd**	**media**	**sd**				
Modello Bayesiano	media	0,865	0,095	-2,632	0,425	0,321	0,145				
	I quartile	0,800	0,092	-2,914	0,439	0,236	0,144				
	mediana	0,863	0,095	-2,626	0,424	0,320	0,145				
	III quartile	0,928	0,098	-2,346	0,411	0,404	0,146				
	copertura	0,080		0,760		0,000					
	vero valore	*0,606*		*-2,500*		*1,000*		*0,070*		*0,070*	
		α		β_0		β		α_0		α_1	
		media	**sd**	**media**	**sd**	**media**	**sd**	**media**	**sd**	**media**	**sd**
MGB	media	0,733	0,145	-3,263	0,823	0,890	0,333	0,063	0,013	0,062	0,002
	I quartile	0,591	0,150	-3,847	0,955	0,581	0,324	0,050	0,015	0,026	0,001
	mediana	0,737	0,149	-3,143	0,786	0,853	0,341	0,065	0,014	0,050	0,002
	III quartile	0,874	0,153	-2,558	0,690	1,166	0,363	0,077	0,013	0,085	0,003
	copertura	0,820		0,810		0,860		0,850		1,000	

Figura 4.17 Weibull con dipendenza da durata nulla: stime puntuali (media campionaria della distribuzione a posteriori) tramite Modello Gerarchico Bayesiano su dati affetti da errore al 2%

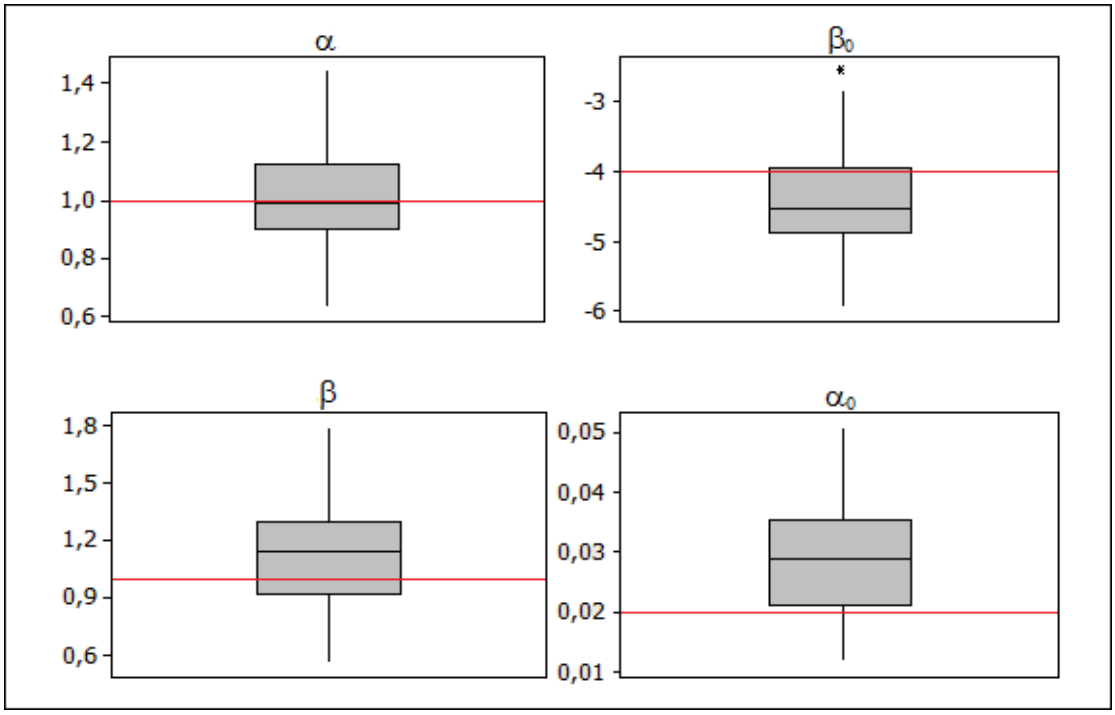


Figura 4.18 Weibull con dipendenza da durata nulla: stime puntuali (media campionaria della distribuzione a posteriori) tramite Modello Gerarchico Bayesiano su dati affetti da errore al 7%

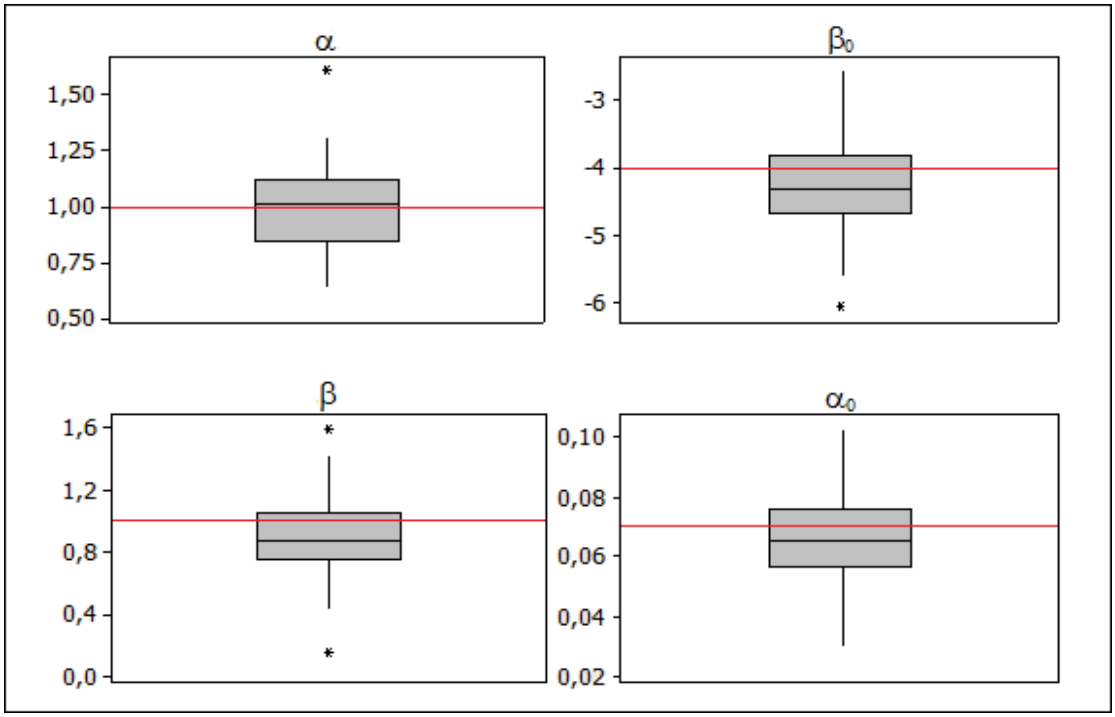


Tabella 4.13 Weibull con dipendenza da durata nulla: media e sd di indici di sintesi delle distribuzioni a posteriori stimate tramite Modello Bayesiano e Modello Gerarchico Bayesiano su dati affetti da errore al 2%

		α		β_0		β					
		media	**sd**	**media**	**sd**	**media**	**sd**				
Modello Bayesiano	media	1,016	0,112	-3,776	0,495	0,705	0,190				
	I quartile	0,929	0,109	-4,137	0,520	0,598	0,185				
	mediana	1,013	0,112	-3,766	0,494	0,705	0,189				
	III quartile	1,100	0,116	-3,408	0,471	0,811	0,195				
	copertura	0,860		0,820		0,350					
	vero valore	*1,000*		*-4,000*		*1,000*		*0,020*		*0,020*	
		α		β_0		β		α_0		α_1	
		media	**sd**	**media**	**sd**	**media**	**sd**	**media**	**sd**	**media**	**sd**
MGB	media	1,005	0,166	-4,407	0,663	1,123	0,246	0,029	0,008	0,062	0,001
	I quartile	0,865	0,176	-4,983	0,725	0,873	0,229	0,021	0,009	0,025	0,001
	mediana	1,004	0,163	-4,323	0,662	1,096	0,245	0,028	0,011	0,050	0,001
	III quartile	1,146	0,163	-3,746	0,615	1,341	0,273	0,037	0,011	0,087	0,003
	copertura	0,840		0,886		0,935		0,739		1,000	

Tabella 4.14 Weibull con dipendenza da durata nulla: media e sd di indici di sintesi delle distribuzioni a posteriori stimate tramite Modello Bayesiano e Modello Gerarchico Bayesiano su dati affetti da errore al 7%

		α		β_0		β					
		media	**sd**	**media**	**sd**	**media**	**sd**				
Modello Bayesiano	media	1,008	0,098	-3,168	0,431	0,368	0,125				
	I quartile	0,942	0,095	-3,447	0,445	0,289	0,123				
	mediana	1,008	0,098	-3,167	0,431	0,367	0,125				
	III quartile	1,074	0,101	-2,884	0,419	0,447	0,127				
	copertura	0,820		0,270		0,000					
	vero valore	*1,000*		*-4,000*		*1,000*		*0,070*		*0,070*	
		α		β_0		β		α_0		α_1	
		media	**sd**	**media**	**sd**	**media**	**sd**	**media**	**sd**	**media**	**sd**
MGB	media	0,985	0,175	-4,243	0,698	0,906	0,255	0,066	0,014	0,062	0,002
	I quartile	0,836	0,190	-4,893	0,806	0,614	0,250	0,053	0,016	0,026	0,001
	mediana	0,990	0,172	-4,104	0,703	0,867	0,262	0,067	0,016	0,052	0,002
	III quartile	1,138	0,171	-3,463	0,637	1,168	0,283	0,082	0,016	0,086	0,003
	copertura	0,897		0,907		0,938		0,856		1,000	

Figura 4.19 Weibull con dipendenza da durata positiva: stime puntuali (media campionaria della distribuzione a posteriori) tramite Modello Gerarchico Bayesiano su dati affetti da errore al 2%

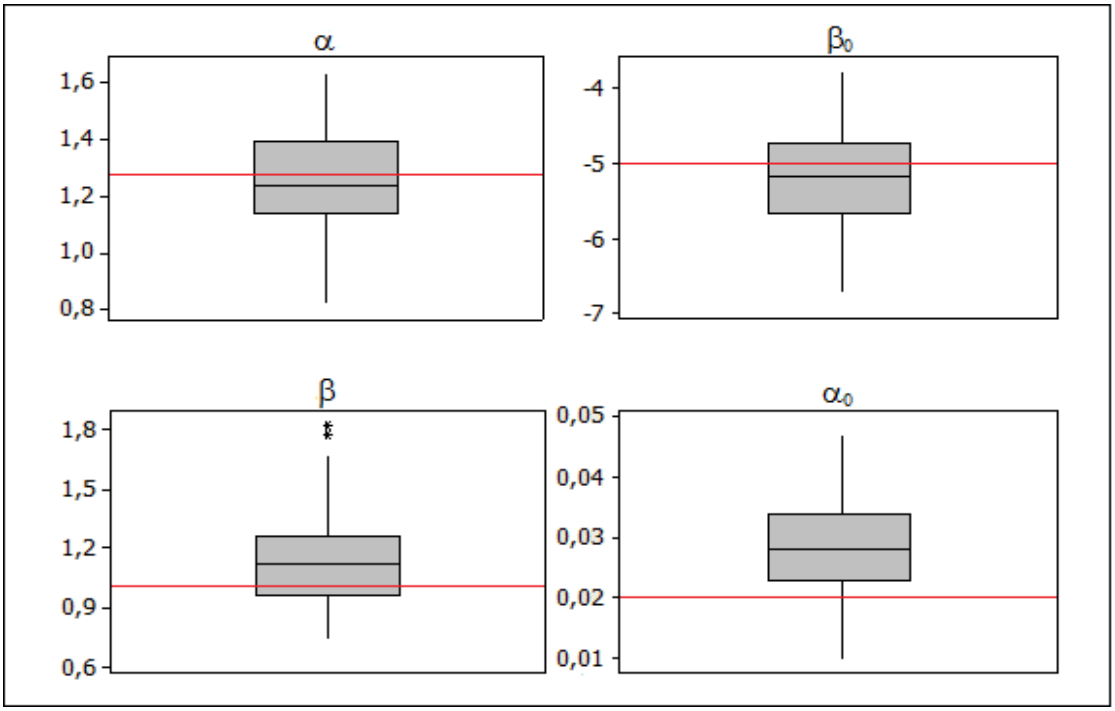


Figura 4.20 Weibull con dipendenza da durata positiva: stime puntuali (media campionaria della distribuzione a posteriori) tramite Modello Gerarchico Bayesiano su dati affetti da errore al 7%

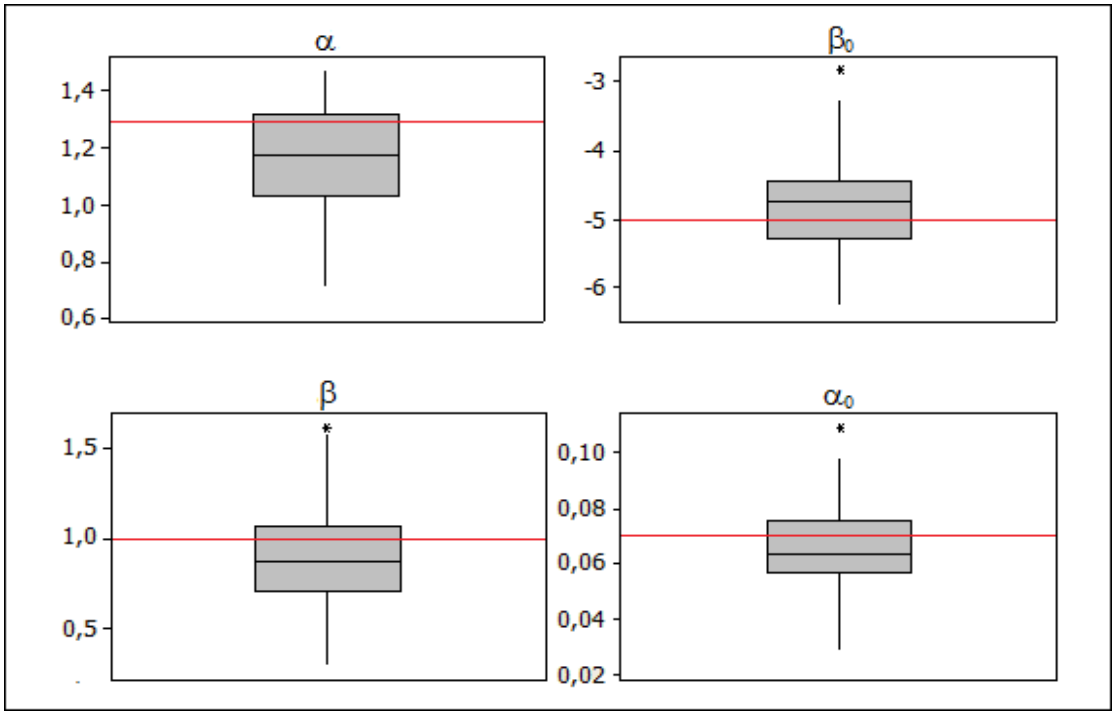


Tabella 4.15 Weibull con dipendenza da durata positiva: media e sd di indici di sintesi delle distribuzioni a posteriori stimate tramite Modello Bayesiano e Modello Gerarchico Bayesiano su dati affetti da errore al 2%

		α		β_0		β					
		media	**sd**	**media**	**sd**	**media**	**sd**				
Modello Bayesiano	media	1,188	0,115	-4,326	0,484	0,708	0,146				
	I quartile	1,099	0,113	-4,683	0,501	0,609	0,144				
	mediana	1,186	0,116	-4,320	0,488	0,707	0,146				
	III quartile	1,274	0,119	-3,959	0,473	0,806	0,148				
	copertura	0,660		0,460		0,280					
	vero valore	*1,284*		*-5,000*		*1,000*		*0,020*		*0,020*	
		α		β_0		β		α_0		α_1	
		media	**sd**	**media**	**sd**	**media**	**sd**	**media**	**sd**	**media**	**sd**
MGB	media	1,267	0,174	-5,214	0,655	1,136	0,234	0,029	0,008	0,063	0,002
	I quartile	1,127	0,173	-5,794	0,720	0,923	0,214	0,020	0,008	0,027	0,001
	mediana	1,261	0,172	-5,136	0,669	1,113	0,235	0,028	0,009	0,051	0,002
	III quartile	1,402	0,177	-4,554	0,620	1,325	0,261	0,037	0,009	0,087	0,003
	copertura	0,880		0,900		0,890		0,820		1,000	

Tabella 4.16 Weibull con dipendenza da durata positiva: media e sd di indici di sintesi delle distribuzioni a posteriori stimate tramite Modello Bayesiano e Modello Gerarchico Bayesiano su dati affetti da errore al 7%

		α		β_0		β					
		media	**sd**	**media**	**sd**	**media**	**sd**				
Modello Bayesiano	media	1,102	0,097	-3,496	0,407	0,393	0,119				
	I quartile	1,033	0,093	-3,776	0,423	0,316	0,118				
	mediana	1,101	0,096	-3,492	0,405	0,392	0,119				
	III quartile	1,170	0,100	-3,210	0,390	0,469	0,120				
	copertura	0,250		0,030		0,000					
	vero valore	*1,284*		*-5,000*		*1,000*		*0,070*		*0,070*	
		α		β_0		β		α_0		α_1	
		media	**sd**	**media**	**sd**	**media**	**sd**	**media**	**sd**	**media**	**sd**
MGB	media	1,168	0,174	-4,773	0,659	0,914	0,269	0,067	0,015	0,064	0,011
	I quartile	1,015	0,180	-5,444	0,759	0,652	0,252	0,051	0,017	0,026	0,001
	mediana	1,166	0,167	-4,652	0,683	0,887	0,277	0,069	0,017	0,051	0,002
	III quartile	1,318	0,171	-4,001	0,611	1,155	0,298	0,084	0,016	0,087	0,004
	copertura	0,850		0,900		0,880		0,860		1,000	

Il parametro α_0 viene stimato in maniera globalmente buona e le variabilità di tali stime puntuali all'interno dell'insieme dei campioni in esame è

decisamente ridotta; la distanza interquartile all'interno di ciascuna distribuzione a posteriori stimata è inoltre in media molto piccola.

Anche la variabilità delle stime del parametro α_1 all'interno dell'insieme di campioni è decisamente ridotta; tuttavia i valori centrali non sono i reali valori del parametro ma corrispondono entrambi a 0,06, la media della distribuzione a priori. Inoltre, come suggeriscono le medie dei quartili, la variabilità della distribuzione a posteriori stimata per tale parametro è, all'interno di ciascun campione, in media particolarmente alta. Ciò conferma che il parametro ha problemi di identificazione, con i tassi di transizione osservati nei campioni trattati.

Quanto ai restanti parametri, le stime puntuali sono in media approssimativamente corrette; in particolare, confrontando i risultati dei due modelli proposti, si apprezza la capacità del Modello Gerarchico Bayesiano di correggere, almeno in media, le distorsioni provocate dall'errore di misura sul modello precedentemente esaminato.

In media il Modello Gerachico Bayesiano fornisce dunque delle stime accettabili, tuttavia la variabilità di tali stime sull'insieme di campioni in esame è decisamente notevole. Tale caratteristica era stata già rilevata sul Modello Bayesiano precedentemente esaminato, ma nel contesto del Modello Gerarchico Bayesiano essa appare ancora più accentuata.

La proporzione di campioni i cui corrispondenti intervalli HPD contengono il “vero valore” del parametro oscilla intorno a 0,8, livello di probabilità degli intervalli. Fa eccezione il parametro α_1, la cui proporzione assume sempre valore 1, a causa della notevole ampiezza degli intervalli stimati per tale parametro.

4.6 Analisi di sensibilità

Il modello è stato valutato anche in condizioni leggermente differenti, al fine di testarne la sensibilità. Anche campioni con pattern di misclassification asimmetrici sono stati simulati: le stime dei parametri α_1 ed α_0 e, di conseguenza, quelle dei restanti parametri non ne risultano variate; le serie relative ai due parametri, infatti, sono tra loro pressoché indipendenti, come si può notare dalla matrice di crosscorrelazione riportata nella Tabella 4.3; inoltre, la distribuzione a posteriori stimata di α_1 non sembra dipendere dall'effettivo "valore" assunto dal parametro, a causa della difficile identificabilità del parametro in caso di tassi di transizione ridotti.

La distanza tra le due interviste successive, k, va ad influenzare direttamente il numero di individui transitati allo stato di occupazione, quindi ciò dovrebbe influire sul processo di stima, in quanto immaginiamo che l'identificabilità del modello, ed in particolare del parametro α_1, possa migliorare se la proporzione di transitati si facesse meno esigua.

Sono dunque stati fatti dei tentativi con k uguale a 6. Dai risultati non sono emerse differenze significative nella correttezza e nella variabilità delle stime.

La proporzione dei transitati può, in realtà, essere modificata anche scegliendo opportunamente il parametro β_0: seguire questa strada, tuttavia, può far incappare in forme molto particolari delle funzioni di rischio $\theta(t)$ e di densità f(t) della variabile T, durata del priodo di disoccupazione; forme molto particolari della distribuzione della durata ne rendono di difficile stimabilità i parametri, rendendo il processo di stima più difficoltoso e maggiormente soggetto a distorsioni.

Inoltre, la covariata X nei campioni riportati è continua e, più precisamente, normale; non si è rilevata alcuna importante differenza nei risultati, scegliendo come covariata una variabile di tipo dicotomico.
Infine, sono state valutate le prestazioni del modello con varie distribuzioni a priori per i parametri: si vedano a riguardo i paragrafi 2.4.1 e 4.3.1.

Conclusioni

Dal lavoro di simulazione emerge come la presenza di errori di misura nell'indicatore di censura, anche se in proporzioni ridotte, produca forti distorsioni nella stima dei parametri. In particolare, sia in caso di dipendenza da durata negativa che positiva, la stima viene distorta verso il caso di dipendenza nulla.

Inoltre, i parametri β, che rappresentano l'effetto delle covariate sulla probabilità di transitare, vengono distorti verso lo 0.

La proporzione di errore di misura influisce, naturalmente, sull'entità delle distorsioni; tuttavia, a causa di ridotti tassi di transizione nei campioni in esame, risulta decisamente più influente la proporzione di disoccupati che affermano di essere transitati, rispetto al tipo opposto di errore .

La verosimiglianza corretta per dati affetti da errore fornisce stime consistenti, tuttavia presuppone l'esatta conoscenza delle proporzioni di errore.

Il Modello Gerarchico Bayesiano non richiede invece tale tipo di conoscenza, ma le proporzioni di errore e le reali transizioni per ciascun individuo vengono stimate dal modello stesso.

Tale modello fornisce risultati globalmente positivi ed in media le stime puntuali Bayesiane coincidono con i “veri valori” dei parametri, intesi come parametri con i quali sono stati simulati i dati. Le distorsioni riscontrate nelle stime effettuate tramite massima verosimiglianza quindi scompaiono o

comunque appaiono decisamente attenuate nelle stime effettuate tramite Modello Gerachico Bayesiano.

Anche la proporzione di disoccupati che riportano erroneamente il loro stato viene stimata con elevata precisione dal modello.

Si riscontrano, comunque, le seguenti problematiche che impongono all'utente delle cautele in fase di stima e interpretazione dei risultati:

- la variabilità delle stime, all'interno dei gruppi di campioni esaminati, è piuttosto elevata;
- la proporzione di coloro tra i transitati che riportano erroneamente il loro stato è difficilmente identificabile; tale caratteristica non è tuttavia strutturale del modello, ma dipende dalla proporzione di transitati nel campioni e, di conseguenza, dell'ambito di applicazione del modello e dalla specifica popolazione in esame;
- l'identificabilità dei parametri α e β_0 risulta problematica, a causa della forte dipendenza che li lega, che si manifesta con un'elevata correlazione negativa tra le serie relative ai due parametri;
- in alcuni casi si riscontrano autocorrelazioni molto alte, anche a ritardi elevati; ciò avviene soprattutto in particolari casi di dipendenza da durata ed entità di errore di misura: l'utente sulla base di tali informazioni potrà effettuare scelte opportune in ambito di stima delle serie.

Viste le premesse riguardo la realistica presenza di errore di misura nei dati e le positive performances del modello, si rileva sicuramente di interesse studiarne ulteriormente il comportamento ed, eventualmente, migliorarne le prestazioni.

Infine, il modello può essere facilmente adattato e sperimentato in contesti differenti. Ad esempio, il modello è agevolmente estendibile al caso di modelli a rischi competitivi; sicuramente interessante sarà anche esaminare il caso in cui la probabilità di riportare erroneamente il proprio stato non sia

costante per tutti gli individui, ma dipenda da una qualche loro caratteristica, quali ad esempio età e livello di istruzione.

Bibliografia

Allison, P. (1984) *Event History Analysis*, Sage, Newbury Park CA.

Bassi, F., Torelli, N. e Trivellato, U. (1998) "Data and modelling strategies in estimating labour force gross flows affected by classification errors", *Survey Methodology*, 24, 109-122.

Brooks, S.P. e Gelman, A. (1997) "General methods for monitoring convergence of iterative simulations", *Journal of Computational and Graphical Statistics*, **7**, 434-455.

Carroll, J., Ruppert, D., Stefanski, L.A. e Crainiceanu C.M. (2006) *Measurement error in non linear models: a modern perspective*, Boca Raton: Chapman & Hall/CRC.

Congdon, P. (2001) *Bayesian Statistical Modelling*, Chichester: Wiley.

Congdon, P. (2003) *Applied Bayesian Modelling*, New York: Wiley.

Florens, J., Fougère, D. e Mouchart, M. (2007) *Duration Models and Point Processes*, IZA Discussion Paper No. 2971,
disponibile in SSRN: <http://ssrn.com/abstract=1011133>.

Gamerman, D. e Lopes, H.F. (2006) *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*, Boca Raton: Chapman&Hall/CRC.

Gelman, A. e Rubin, D.B. (1992) "Inference from iterative simulation using multiple sequences", *Statistical Science*, **7**, 457-511.

Geweke, J. (1992) "Evaluating the accuracy of sampling-based approaches to calculating posterior moments" in *Bayesian Statistics 4*, a cura di J.M. Bernardo, J.M. Berger, A.P. Dawid e A.F.M. Smith, Oxford: Clarendon Press.

Gilks, W.R., Richardson, S. e Spiegelhalter, D.J. (1996) *Markov chain Monte Carlo in practice*, London: Chapman&Hall.

Hausman, J.A., Abrevaya J. e Scott-Morton, F.M. (1998) "Misclassification of the Dependent Variable in a Discrete-Response Setting", *Journal of Econometrics*, 87, 239-269.

Holt D., McDonald, J. e Skinner, C.J. (1991) "The effect of measurement error on event history analysis" in *Measurement Error in Surveys*, a cura di P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz e S. Sudman, New York: Wiley, 665-686.

Istat (2004), *La nuova rilevazione sulle Forze Lavoro. Contenuti, metodologie, organizzazione*, Roma: Istat.

Jenkins, S.P. (2004) *Survival Analysis*, Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK, disponibile in:
<www.iser.essex.ac.uk/teaching/degree/stephenj/ec968lnotesv6.pdf>.

Meier, A.S., Richardson, B.A. e Hughes J.P. (2003) "Discrete Proportional Hazard Models for Mismeasured Outcomes", *Biometrics*, 59, 947-954.

Paggiaro, A. e Torelli, N. (2004) "The effect of classification errors in survival data analysis", *Statistical Methods & Applications*, 13, 213-225.

Plummer, M., Best, N., Cowles, K. e Vines, K. (2007) *coda: Output analysis and diagnostics for MCMC*, R package version 0.12-1.

Poterba, J.M. e Summers, L.H. (1984) "Response Variation in the CPS: caveats for Unemployment Analysts", *Monthly Labor Review*, 37-43.

Poterba, J.M. e Summers, L.H. (1995) "Unemployment Benefits and Labor Market Transitions: a Multinomial Logit Model with Errors in Classification", *Review of Economics and Statistics*, 77, 207-216.

Redner, R.A. e Walker, H.F. (1984) "Mixture densities, maximum likelihood and the EM algorithm", *SIAM Review*, 26:195-239.

Richardson, S. e Best, N. (2003) "Bayesian hierarchical models in ecological studies of health-environment effects", *EnvironMetrics*, 14, 129-147.

Ripley, B.D. (1987) *Stochastic Simulation*, New York: Wiley.

Salant, S.W. (1977) "Search theory and duration data: A theory of sorts", *Quarterly Journal of Economics*, 91:39-57.

Silverman, B.W. (1986) *Density Estimation,* London: Chapman and Hall.

Skinner, C.J. (1998) "Logistic Modelling of Longitudinal Survey data with Measurement Error", *Statistica Sinica*, 8,1045-1058.

Skinner, C.J. e Torelli, N. (1993) "Measurement Errors and the estimation of Gross Flows from Longitudinal Economic Data", *Statistica*, 3, 391- 405.

Spiegelhalter, D., Thomas, A., Best, N. e Lunn, D. (2003) *WinBUGS User Manual*, disponibile in:
< http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf>.

Sturtz, S., Ligges, U. e Gelman, A. (2005) "R2WinBUGS: A Package for Running WinBUGS from R", *Journal of Statistical Software*, 12(3), 1-16.

Torelli N. e Paggiaro A. (2002) "Estimating transition models with misclassification" in *IASS Topics, 53rd Session of the International Statistical Institute, Proceedings, Invited Papers*, IASS, 391-410

Appendice A

Specificazione dei Modelli Bayesiani nel linguaggio WinBUGS

Nel presente lavoro sono stati utilizzati vari modelli di tipo Bayesiano: le distribuzioni a posteriori sono state stimate tramite metodo Markov Chain Monte Carlo usando il software WinBUGS. Vengono ora descritte le modalità di specificazione dei modell Bayesiani nel linguaggio del software, mentre per le descrizioni dei due modelli si rimanda ai paragrafi 2.4.2 e 4.3.2. Per una trattazione più completa del linguaggio del software WinBUGS ed in generale delle sue modalità di utilizzo si veda Spiegelhalter *et al.* (2003).

Il primo Modello Bayesiano di cui ci occupiamo è definito dalle seguenti distribuzioni di probabilità:

$\alpha \sim \Gamma(2,2)$,

$\beta_0 \sim N(-4,4)$,

$\beta \sim N(0,1)$,

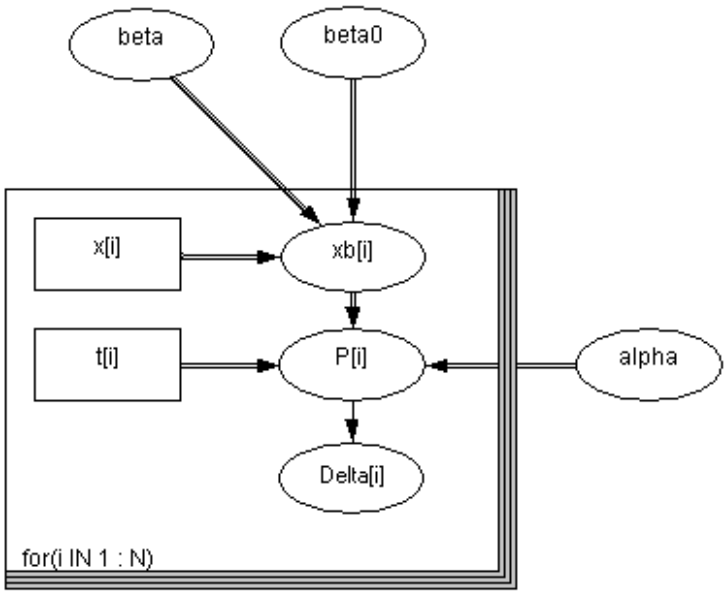
$\delta_i \sim \text{Bern}\left(1 - \exp\left(\exp(x_i'\beta)\left(t_i^{\alpha} - (t_i + k)^{\alpha}\right)\right)\right)$.

Si rimanda al paragrafo 2.4.2 per la descrizione di tale modello.

Di seguito, tali specificazioni vengono espresse attraverso il linguaggio testuale di WinBUGS ed il grafico corrispondente nella Figura A.1:

```
model
{
for(i in 1:N)
{xb[i] <- beta0 + beta*x[i]
P[i] <-1-exp(exp(xb[i])*(pow(t[i],alpha)-pow(t[i+k,alpha)))
Delta[i] ~ dbern(P[i])
}
alpha ~ dgamma(2,0.5)
beta0 ~ dnorm(-4,0.25)
b ~ dnorm(0,1)
}
```

Figura A.1. Specificazione del Modello Bayesiano, presentato nel paragrafo 2.4.2, tramite linguaggio grafico del software WinBUGS



Ogni quantità è rappresentata tramite un nodo, di forma rettangolare se costante, e ovale se stocastica o funzione deterministica di altre quantità. Le frecce doppie rappresentano una dipendenza non stocastica, a differenza di quelle singole. Si noti che parti ripetute, indicizzate tramite *i* che varia da 1 a N, vengono rappresentate tramite il riquadro rettangolare che le contiene.

Per quanto riguarda il linguaggio testuale si ricorda che i collegamenti stocastici tra due nodi vengono rappresentati con ~, mentre i collegamenti non stocastici con <-. Per una più vasta trattazione del linguaggio del software WinBUGS si rimanda al paragrafo 2.3.
Le distribuzioni di probabilità vengono espresse nel linguaggio testuale in maniera del tutto naturale: si segnala che inserire nella specificazione nodi e collegamenti di tipo deterministico non è strettamente necessario, ma può essere conveniente inserirli per facilitare le notazioni (Spiegelhalter, 2003); inoltre si ponga attenzione alla particolarità della sintassi WinBUGS che vuole come secondo parametro della distribuzione normale, non la varianza della distribuzione, ma il suo inverso.

Il modello descritto nel paragrafo 4.3.2 è definito dalle seguenti distribuzioni di probabilità:

$$\alpha \sim \Gamma(2,2),$$

$$\beta_0 \sim N(-4,4),$$

$$\beta \sim N(0,1),$$

$$A_i \sim Bern\left(1 - \exp\left(\exp(x_i'\beta)\left(t_i^{\alpha} - (t_i + k)^{\alpha}\right)\right)\right),$$

$$\alpha_0 \sim Beta(1{,}5, 22{,}5),$$

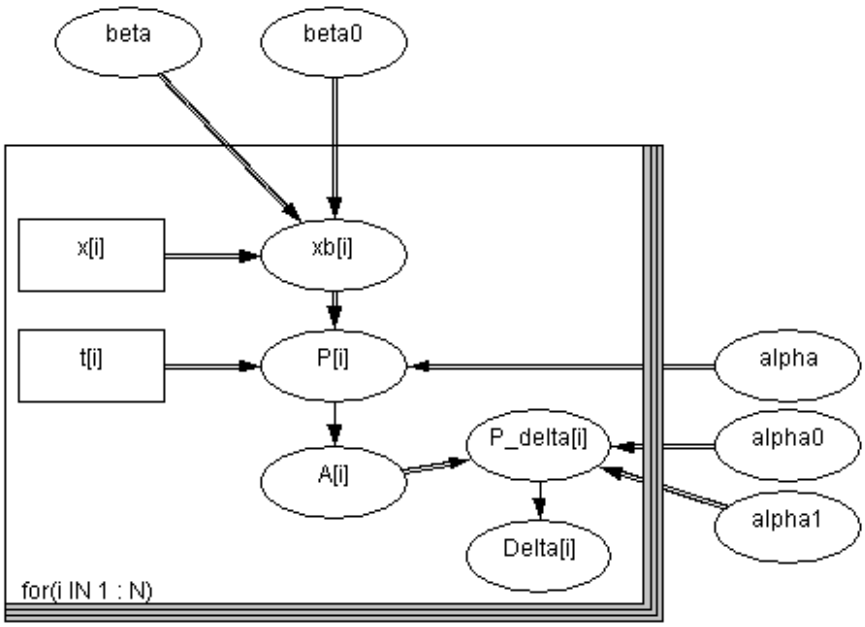
$$\alpha_1 \sim Beta(1{,}5, 22{,}5),$$

$$\delta_i \sim Bern\left((1-\alpha_1)A_i + \alpha_0(1 - A_i)\right);$$

esso è perciò un Modello Gerarchico Bayesiano, in quanto le quantità ignote da stimare nel modello sono divise in due livelli, di cui l'inferiore formato dai parametri A_i.
Riportiamo entrambe le possibili specificazioni per il software WinBUGS, nell linguaggio testuale e nel linguaggio grafico nella Figura A.2:

```
model
{
for(i in 1:N)
{xb[i] <- beta0 + beta*x[i]
P[i] <-1-exp(exp(xb[i])*(pow(t[i],alpha)-pow(t[i+k,alpha)))
A[i] ~ dbern(P[i])
P_Delta[i] <- (1- alpha1)*A[i] + alpha0*(1-A[i])
Delta[i] ~ dbern(P_Delta[i])
}
alpha ~ dgamma(2,0.5)
beta0 ~ dnorm(-4,0.25)
b ~ dnorm(0,1)
alpha0 ~ dbeta(1.5,22.5)I(0,0.25)
alpha1 ~ dbeta(1.5,22.5)I(0,0.25)
}
```

Figura A.2. Specificazione del Modello Gerarchico Bayesiano, presentato nel paragrafo 4.3.2, tramite linguaggio grafico del software WinBUGS



Si noti che in questo secondo modello le frecce singole, che rappresentano dipendenze stocastiche, sono due come due sono appunto i livelli nei quali sono organizzati i parametri del modello; ricordiamo che i collegamenti di tipo deterministico non sono necessari per la specificazione del modello, e

quindi la loro presenza non deve confondere sul numero di livelli del modello.

Anche in questo caso, la trascrizione delle distribuzioni di probabilità che definiscono il modello nel linguaggio WinBUGS è del tutto naturale; si noti la sintassi per specificare i particolari limiti imposti ai parametri α_0 e α_1: la lettera I accanto alle distribuzioni a priori dei due parametri è seguita da parentesi contenenti l'intervallo nel quale si vuole limitare la stima dei parametri in questione; qualora sia specificato soltanto uno dei due estremi dell'intervallo, non si pone alcuna limitazione dal lato opposto. Si ricorda che questi due vincoli sono opzionali e possono venir usati per evitare problemi di *label switching* (si veda il paragrafo 4.3.2) o comunque di stime di valori poco ragionevoli per le proporzioni di errore.

Appendice B

Ulteriori evidenze dall'analisi di dati affetti da errore di misura

Per avere una valutazione più generale degli effetti della misclassification sulla stima dei parametri che non dipenda dai singoli campioni presentati, si ricorre al metodo Monte Carlo con 100 repliche. In riferimento a Torelli e Paggiaro (2002), i campioni sono stati simulati in modo che ricalcassero quelli realmente utilizzati nell'ambito di analisi della durata della disoccupazione in studi basati sulla rotazione del campione: la dimensione campionaria è stata fissata a 1000; la tipologia delle covariate e il valore dei relativi parametri sono stati scelti in base a quanto tipicamente osservato nella RCFL italiana per la popolazione dei disoccupati. Le probabilità di errata classificazione sono state fissate a valori che non risultano lontani a quelli stimati in precedenti lavori (Poterba e Summers, 1995; Bassi, Torelli e Trivellato, 1995).

Nella Tabella B.1, si riportano i risultati di una prima simulazione in cui viene fatta variare la dipendenza dalla durata, attraverso il parametro α il cui logaritmo assume rispettivamente i valori -0,5, 0 e 0,5.

Tabella B.1. Metodo Monte Carlo con 100 campioni sui quali sono state effettuate stime di massima verosimiglianza (1); dati affetti da errore con α_0=0,05, α_1= 0,01; covariate: sesso (1=M), età, coniugato(1=sì), educazione (1=alto livello)

parametri	vero	media	sd	vero	media	sd	vero	media	sd
α	0,606	0,774	0,074	1,000	0,997	0,090	1,649	1,397	0,079
intercetta	-2,500	-1,454	0,256	-4,000	-2,281	0,424	-6,000	-3,484	0,354
sesso	1,000	0,659	0,138	1,000	0,683	0,171	1,000	0,741	0,150
età	-1,000	-0,599	0,391	-1,000	-0,641	0,513	-1,000	-0,738	0,433
coniugato	0,000	0,039	0,206	0,000	0,018	0,213	0,000	0,008	0,187
educazione	1,000	0,683	0,227	1,000	0,677	0,232	1,000	0,727	0,192

Si noti la distorsione delle stime del parametro α verso il valore 1 e l'attenuazione dei parametri β, presente ed approssimativamente di medesima intensità in tutti i pattern di dipendenza da durata esaminati. I risultati si inquadrano quindi in quanto ottenuto nello studio di Torelli e Paggiaro (2002).

Tabella B.2. Metodo Monte Carlo con 100 campioni sui quali sono state effettuate stime di massima verosimiglianza (1); dati affetti da errore di misura in varie proporzioni; covariate: sesso (1=M), età, coniugato(1=sì), educazione (1=alto livello)

		Parametri	vero	α_1					
			valore	0,01		0,05		0,1	
				media	sd	media	sd	media	sd
α_0	0,01	α	0,606	0,681	0,070	0,698	0,086	0,700	0,074
		intercetta	-2,500	-1,605	0,417	-1,720	0,460	-1,804	0,407
		sesso	1,000	0,904	0,259	0,901	0,246	0,905	0,236
		età	-1,000	-0,866	0,654	-0,950	0,619	-0,987	0,751
		coniugato	0,000	0,001	0,276	0,021	0,277	0,057	0,262
		educazione	1,000	0,881	0,285	0,949	0,318	0,902	0,286
	0,05	α	0,606	0,774	0,074	0,770	0,082	0,791	0,074
		intercetta	-2,500	-1,454	0,256	-1,421	0,400	-1,578	0,366
		sesso	1,000	0,659	0,138	0,607	0,224	0,597	0,210
		età	-1,000	-0,599	0,391	-0,621	0,526	-0,679	0,472
		coniugato	0,000	0,039	0,206	-0,006	0,204	0,030	0,227
		educazione	1,000	0,683	0,227	0,647	0,277	0,671	0,250
	0,1	α	0,606	0,836	0,060	0,842	0,063	0,854	0,068
		intercetta	-2,500	-1,207	0,314	-1,283	0,307	-1,350	0,330
		sesso	1,000	0,452	0,162	0,463	0,176	0,431	0,171
		età	-1,000	-0,536	0,413	-0,415	0,408	-0,385	0,399
		coniugato	0,000	-0,004	0,182	0,007	0,176	0,017	0,191
		educazione	1,000	0,544	0,207	0,552	0,218	0,465	0,226

Si esamina poi come gli effetti sulle stime dei parametri varino rispetto al pattern di misclassification, rappresentato dai parametri α_0 e α_1 (Tabella B.2). Non riportiamo i risultati ottenuti per i tre valori del parametro α, in quanto l'entità delle distorsioni non subisce variazioni di rilievo al variare della dipendenza da durata; viene riportato il caso di dipendenza da durata negativa, in quanto essa viene tipicamente osservata in studi di questo tipo. Come si può vedere esaminando la diagonale della Tabella B.2, le distorsioni nelle stime crescono con le proporzioni di risposte errate. Si noti l'asimmetria della Tabella, confrontando ad esempio i casi α_0=0,01, α_1= 0,05 e α_0=0,05, α_1= 0,01: α_0 produce cioè un effetto maggiore sulla stima dei parametri di quello prodotto da α_1, tanto che all'interno di ciascuna riga della Tabella l'entità delle distorsioni è pressoché uniforme. Infatti, α_1 produce una distorsione tanto più forte quanto maggiore è il numero dei transitati e naturalmente, vale la considerazione opposta per α_0; nei campioni in esame la proporzione dei transitati è, infatti, del 15% circa. Si osservi inoltre il primo riquadro della Tabella B.2 (α_0=0,01, α_1=0,01) in cui si riporta la situazione di una minima quota di errata classificazione nei dati, ovvero di soli 10 individui in media in ciascun campione che riportano il loro stato non correttamente. La conseguente distorsione nelle stime è comunque di entità non trascurabile: si prenda ad esempio β_2, che subisce un'attenuazione da -1 a -0,86.

Nel simulare i campioni appena utilizzati è stato assunto che la probabilità di rispondere erroneamente fosse la medesima per ogni individuo del campione. Ciò potrebbe non risultare vero, in quanto non è implausibile che la predisposizione a dare risposte corrette possa dipendere da qualche caratteristica dell'unità sperimentale; ad esempio, l'età dell'individuo, il suo

livello d'istruzione o una variabile che indica se l'intervista è stata effettuata sull'individuo stesso o un suo familiare.

Abbiamo esaminato su dati simulati l'effetto dell'errore di misura in alcuni casi particolari di misclassification dipendente da altre variabili; abbiamo esaminato queste problematiche senza poi proporre opportuni modelli per la corretta stima dei parametri, ma soltanto come spunto per eventuali ulteriori analisi nell'ambito: sarà importante infatti considerare che gli effetti della misclassification possono essere estremamente variabili per pattern ed entità delle distorsioni in base ad associazioni della misclassification con altre variabili.

Assumiamo che le probabilità di riportare erroneamente il proprio stato dipendano dal valore assunto da una qualche variabile Z, ed introduciamo le funzioni:

$$\alpha_0(Z) = P(\delta = 1 | A = 0; Z),$$

$$\alpha_1(Z) = P(\delta = 0 | A = 1; Z).$$

Abbiamo esaminato il caso in cui la predisposizione dell'individuo a dare risposte corrette dipenda da una qualche sua caratteristica, inserita nel modello anche come covariata, ad esempio il sesso o l'età.

Dunque i campioni sono stati simulati con un'unica covariata X con distribuzione normale standard e l'errata classificazione viene definita dai parametri α_0 e α_1:

$$\alpha_0(x_i) = \alpha_1(x_i) = \frac{\exp(\gamma_0 + \gamma_1 x_i)}{1 + \exp(\gamma_0 + \gamma_1 x_i)},$$

dove il parametro γ_0 è stato fissato a –4; per valori nulli di x α_0 e α_1 valgono 0,018.

Analogamente al precedente esempio, vengono riportati soltanto i risultati ottenuti in caso di dipendenza negativa nella Tabella B.3.

Tabella B.3. Metodo Monte Carlo con 100 campioni sui quali sono state effettuate stime di massima verosimiglianza (1); dati affetti da errore di misura associato alla variabile X tramite il parametro γ_0

Parametri	vero	γ_0=-4 γ_1=1		γ_0=-4 γ_1=-1	
		media	sd	media	sd
α	0,606	0,764	0,118	0,766	0,088
intercetta	-2,500	-1,798	0,500	-1,537	0,451
x	1,000	0,984	0,181	-0,073	0,206
α	0,606	0,765	0,093	0,754	0,113
intercetta	-2,500	-1,556	0,401	-1,772	0,479
x	-1,000	0,106	0,186	-0,965	0,172

È interessante notare come, in alcuni casi, ad esempio qualora la covariata X sia positivamente associata alla probabalità di errore ed il corrispondente parametro β sia positivo, la distorsione del parametro β è decisamente ridotta, tanto che la stima è vicinissima ad 1.

Se cambia il segno del parametro o dell'associazione tra X ed errore, al contrario, l'attenuazione è così accentuata che il parametro stimato non è significativamente diverso da 0.

Inoltre abbiamo ipotizzato che la probabilità di rispondere correttamente possa essere influenzata dalla durata di permanenza nello stato di disoccupazione pregressa alla prima intervista; viene dunque proposto il seguente pattern di errata classificazione:

$$\alpha_0(t_i) = \alpha_1(t_i) = \frac{\exp(\gamma_0 + \gamma_1 t_i)}{1 + \exp(\gamma_0 + \gamma_1 t_i)},$$

con γ_0 pari a –4, analogamente al caso precedente; γ_1 assume valori tali che i parametri α_0 e α_1 variano tra 0,06 e 0,02 circa.

Dalla Tabella B.4 si noti come, nel caso di dipendenza da durata positiva (α =exp(0,5)), la distorsione di α sia maggiore per γ_1=-0,01, quando cioè ad alte durate corrisponda maggiore rischio di transito, ma anche minore rischio di misclassification. L'analogo, ma in maniera opposta, avviene per dipendenza da durata negativa.

Tabella B.4. Metodo Monte Carlo con 100 campioni sui quali sono state effettuate stime di massima verosimiglianza (1); dati affetti da errore di misura associato alla variabile T tramite il parametro γ_0.

Parametri	vero	γ_0=-4 γ_1=0,01		γ_0=-3 γ_1=-0,01	
		media	sd	media	sd
		Dipendenza negativa			
α	0,606	0,789	0,083	0,709	0,082
intercetta	-2,500	-1,745	0,370	-1,327	0,403
x	1,000	0,727	0,206	0,659	0,223
		Dipendenza nulla			
α	1,000	1,057	0,089	0,960	0,085
intercetta	-4,000	-2,761	0,415	-2,260	0,382
x	1,000	0,797	0,228	0,691	0,214
		Dipendenza positiva			
α	1,649	1,480	0,117	1,361	0,086
intercetta	-6,000	-4,035	0,478	-3,441	0,364
x	1,000	0,825	0,206	0,699	0,187