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

Tassi di fecondità per età del Costa Rica. Un'analisi bayesiana.

Relatore Prof. Bruno Scarpa
Dipartimento di Scienze Statistiche

Laureando: Riccardo Griggio
Matricola N 1013835

Anno Accademico 2013/2014

Indice

Introduzione

I tassi di fecondità per età rappresentano il numero annuale di nascite da donne di una specifica età o fascia d'età per 1000 donne in quella fascia d'età. Questi tassi tornano utili come misura della struttura per età della fecondità, cioè la frequenza relativa della fecondità tra donne di differenti età tra gli anni riproduttivi, oppure come calcolo intermedio del tasso di fecondità totale.

In particolare, in questa tesi, si vogliono analizzare i tassi di fecondità per età del Costa Rica, nel periodo che va dall'anno 1972 al 2012. Esistono molti modelli per la loro stima, tra cui i più frequentemente usati sono la distribuzione di Hadwiger (Hadwiger, 1940; Gilje, 1969), la distribuzione Beta e la distribuzione Gamma di Hoem et al. (1981) e il modello spline quadratico di Schmertmann (2003), che verranno brevemente presentati nel primo capitolo. Nel nostro caso, però, si è voluto utilizzare un modello basato sulla distribuzione normale asimmetrica, che verrà presentata nel secondo capitolo, in modo che fosse flessibile e descrivesse sia forme simmetriche che asimmetriche. In generale i tassi di fecondità per età assumono una forma asimmetrica e unimodale, tuttavia in anni recenti e in alcune nazioni si sono osservate delle distribuzioni simmetriche simili alla normale. Per questo è utile usare un modello che includa sia distribuzioni simmetriche che asimmetriche.

Per stimare i parametri si è scelto un approccio bayesiano in modo da poter sfruttare i dati a disposizione per i diversi anni, avendo così la

possibilità di ricavare delle informazioni a priori dagli anni precedenti e aggiornare le stime per gli anni successivi.

Per analizzare i dati a nostra disposizione è stato utilizzato un primo modello proposto da Canale e Scarpa (2013), presentato nel terzo capitolo, in cui si assume che le nostre osservazioni sono continue e si distribuiscono come una normale asimmetrica. Viene assunta a priori la distribuzione normale-gamma inversa per i parametri di posizione e scala, mentre per il parametro di forma si assume che a priori si distribuisca come una normale asimmetrica. Per la stima di questi parametri, Canale e Scarpa (2013) propongono un algoritmo di Gibbs sampling ed introducono delle variabili latenti normali standard in modo da riuscire ad ottenere una coniugazione per i parametri di posizione e scala, invece per il parametro di forma la a posteriori appartiene alla classe di distribuzioni normale asimmetrica unificata.

In realtà, però, le nostre osservazioni non sono distribuite come una normale asimmetrica, visto che i tassi di fecondità per età vengono calcolati considerando gli intervalli di età, ma assumono la distribuzione di una multinomiale con i parametri ottenuti dalla densità della normale asimmetrica. Quindi si è deciso di provare una strada alternativa a quella già proposta. In questo nuovo modello, presentato nel quinto capitolo, per riuscire a coniugare la distribuzione multinomiale con quella normale asimmetrica vengono introdotte delle variabili latenti aventi distribuzione normale asimmetrica. In questo modo è possibile sfruttare alcuni dei risultati ottenuti da Canale e Scarpa (2013) e creare un nuovo algoritmo di Gibbs sampling.

Capitolo 1

Il tasso di fecondità per età

Il tasso di fecondità per età (in inglese Age-Specific Fertility Rate, ASFR) misura il numero annuale di nascite da donne di una specifica età o fascia d'età (di solito gruppi di 5 anni) per 1000 donne in quella fascia d'età ed è calcolato come

$$g_a = \frac{B_a}{E_a} 1000 \tag{1.1}$$

dove B_a è il numero di nascite da donne nel gruppo di età a in un dato anno o periodo di riferimento e E_a è il numero di donne nel gruppo di età a durante il periodo di riferimento specificato. I dati richiesti per il suo calcolo sono il numero di nascite in un dato anno o periodo di riferimento classificati per età della madre e il numero di donne in età riproduttiva (per esempio 15-44 o 15-49), in 1 anno o gruppi di 5 anni.

1.1 Il sistema statistico generatore dei dati

I dati sui tassi di fecondità per età possono essere ottenuti da tre sorgenti: dai sistemi di registrazione civile, dalle indagini a campione e dai censimenti. I sistemi di registrazione civile sono considerati la fonte migliore per le informazioni che ci interessano. Tuttavia, in alcuni paesi, in particolare quelli in regioni meno sviluppate, vi è la mancanza di

un sistema di registrazione civile o di un sistema di registrazione la cui copertura sia completa. Questi sistemi sono considerati completi se coprono il 90% o più di tutti i nati vivi presenti in un paese o in un'area. Nei paesi dove questi sistemi sono carenti, le indagini e i censimenti possono essere utilizzati per la stima dell'ASFR. Le stime dai censimenti sono derivate da domande sulle nascite durante uno specifico periodo precedente al censimento (di solito 12 mesi), mentre le stime dalle indagini possono essere derivate o da domande sulle nascite entro un determinato periodo precedente o dalla storia delle nascite parziale o completa. Confrontati con i dati dei registri civili completi, queste domande retrospettive tendono a produrre stime meno affidabili in quanto si basano sulla capacità degli individui di ricordare con precisione un evento che ha avuto luogo diversi mesi o anni prima. Inoltre, mentre i sistemi di registrazione civile tendono a generare stime annuali, la disponibilità di dati da indagini o censimenti dipende dall'esistenza di un'adeguata indagine o di adeguati programmi di censimento. I censimenti sono di norma condotti ogni 10 anni, mentre le indagini sono effettuate a differenti intervalli in diversi paesi. Nei paesi in via di sviluppo si svolgono, di solito, ogni 3-5 anni.

Generalmente, solo una fonte è fornita per anno per un paese. Quando più di una fonte è disponibile per lo stesso periodo, la preferenza è data alle stime basate sulle registrazioni civili. I dati dai sistemi di registrazione considerati come meno del 90% completi sono usati per i paesi dove fonti alternative sono o non disponibili o presentano problemi di comparabilità e quando questi dati possono fornire una valutazione sulle tendenze.

A seconda della fonte utilizzata vi sono delle limitazioni. Queste limitazioni sono:

- per i sistemi di registrazione civile, le stime sono soggette a limitazioni che dipendono dalla completezza delle registrazioni delle nascite. La comparabilità dei dati è inoltre influenzata dal trattamento dei bambini nati vivi ma morti prima della registrazione o nelle prime 24 ore di vita, dalla qualità delle informazioni riportate sull'età della madre, e dall'inserimento di nascite dei periodi precedenti. Le stime

della popolazione possono soffrire di limitazioni legate ad età errate e alla copertura delle età;

- per i dati da indagine e censimento, le principali limitazioni riguardano età errate, nascite omesse, data di nascita del bambino errata e, nel caso delle indagini, variabilità del campione.

1.2 Scopo dell'ASFR

Il tasso di fecondità per età ha due usi principali: uno come misura della struttura per età della fecondità, che è la frequenza relativa della fecondità tra donne di differenti età tra gli anni riproduttivi, e uno come calcolo intermedio per derivare il tasso di fecondità totale (in inglese Total Fertility Rate, TFR), cioè il numero medio di bambini che sarebbero nati vivi da una donna durante la sua vita se fosse passata attraverso gli anni della fecondità partorendo secondo l'attuale distribuzione dei tassi di fecondità per età.

Quando i dati sono ottenuti tramite censimento o tramite rilevamenti, si possono ottenere sia il numeratore che il denominatore del rapporto della formula (1.1).

Una semplice, anche se meno precisa, procedura per calcolare il denominatore dell'ASFR è di prendere la media del numero di donne in ogni fascia d'età durante il periodo di riferimento coperto da misurazione, cioè la media del numero di donne in ogni fascia d'età all'inizio e alla fine del periodo di riferimento.

I periodi di riferimento che comprendono più di un anno sono frequentemente utilizzati per calcolare gli ASFR dai dati dell'indagine; la logica è quella di diminuire la variabilità associata a numeri relativamente piccoli di nascite annuali, che si verificano per donne in gruppi di singoli anni o gruppi di 5 anni, e l'effetto distorcente della segnalazione di errori del periodo di riferimento.

Gli ASFR possono anche essere presenti per diversi gruppi di donne; per esempio, per donne attualmente sposate o in unione e per tutte

le donne in età riproduttiva. In società dove la fecondità è largamente confinata al matrimonio, gli ASFR per donne attualmente sposate o in unione forniranno più o meno la completa copertura di recente fecondità. Quando una grande quota della fecondità si verifica al di fuori di unioni riconosciute da queste società, tuttavia, la restrizione dell'ASFR a donne attualmente sposate comporterà una sottostima del livello di fecondità corrente.

L'ASFR è, inoltre, di particolare interesse in paesi, città o quartieri con interventi sulla riproduttività delle adolescenti con lo scopo di ridurre le gravidanze indesiderate.

1.3 Problemi derivanti dall'ASFR

A differenza del tasso di natalità, l'ASFR non è influenzato dalle differenze o dai cambiamenti nella composizione per età della popolazione e così è più utile nel confronto di differenti popolazioni o sotto-gruppi e nella misurazione di cambiamenti nel tempo. Tuttavia è influenzato dalle differenze o dai cambiamenti nel numero o nella percentuale di donne in età riproduttiva. Quindi, dei cambiamenti negli ASFR possono fornire informazioni fuorvianti riguardanti l'impatto dei programmi di pianificazione familiare sulla fecondità quando altri fattori che influenzano il rischio di gravidanza stanno cambiando (per esempio, per le fasce d'età 15-19 e 20-24, quando l'età al matrimonio sta crescendo rapidamente).

Per indirizzare questo problema, si possono calcolare gli ASFR solo per le donne che sono continuativamente sposate o in unione durante il periodo di riferimento della misura. La misura risultante è conosciuta come il tasso di fecondità per età nel matrimonio (in inglese marital age-specific fertility rate, MASFR). Tuttavia, per calcolare questa misura, sono richiesti i dati sulla durata del matrimonio o le storie matrimoniali. In pratica, i MASFR sono molto spesso approssimati calcolando gli ASFR per donne sposate o in unione nel periodo dell'indagine, anche se i valutatori dovrebbero riconoscere che questa figura approssima solamente il MASFR perché le

donne che sono sposate o in unione nel periodo dell'indagine possono non essere continuativamente sposate o in unione durante tutto il periodo di riferimento delle misurazioni (per esempio, per i 3-5 anni precedenti all'indagine).

1.4 Modelli per la stima dell'ASFR

La distribuzione dei tassi di fecondità per età comincia con un minimo all'inizio dell'età riproduttiva e poi cresce e raggiunge un massimo da qualche parte tra i 20 e i 30 anni (dipende dal paese in cui si trova la popolazione presa in esame). Poi decresce di nuovo fino a stabilizzarsi vicino ai 50 anni (per un esempio si veda la Figura 1.2). La grandezza del singolo tasso di fecondità per età è influenzata da differenze nelle pratiche matrimoniali e di gravidanza, presenza o assenza di controllo di fecondità e dalle normative riguardanti i divorzi e i nuovi matrimoni, ma la struttura generale rimane invariata negli anni e nei paesi. I paesi mostrano differenze riguardo alla velocità che ci impiega la curva a raggiungere il picco massimo e alla velocità che ci impiega a raggiungere la fine della campana di fecondità.

Seguendo Hoem et al. (1981), la curva di fecondità può essere scritta come

$$g(y; R, \theta_2, ..., \theta_r) = R \cdot h(y; \theta_2, ..., \theta_r), \tag{1.2}$$

dove $g(y; R, \theta_2, ..., \theta_r)$ è il tasso di fecondità per l'età y, $h(\cdot; \theta_2, ..., \theta_r)$ è la funzione di densità di probabilità sull'asse reale con $r - 1$ parametri, e R è l'r-esimo parametro che rappresenta il tasso di fecondità totale.

Sono state date molte specificazioni di $h(\cdot; \theta_2, ..., \theta_r)$ usando, per esempio, la distribuzione Hadwiger e le distribuzioni Beta e Gamma equivalenti alle curve di Pearson tipo I e III rispettivamente. In sequenza forniamo le formule matematiche per i modelli sopracitati e per la Spline quadratica, le quali sono i modelli più frequentemente usati nella letteratura demografica per stimare le curve di fecondità.

La funzione Hadwiger (Hadwiger, 1940; Gilje, 1969) è espressa da,

$$f(y) = \frac{ab}{c}\left(\frac{c}{y}\right)^{\frac{3}{2}} exp\left\{-b^2\left(\frac{c}{y} + \frac{y}{c} - 2\right)\right\},$$

dove y è l'età della madre al momento della nascita del bambino e a, b e c sono i tre parametri da stimare. Chandola et al. (1999), in contrasto con Hoem et al. (1981), sostengono che i parametri possono avere una interpretazione demografica come segue: il parametro a è associato alla fecondità totale, il parametro c è collegato all'età media alla nascita del figlio, il parametro b determina l'altezza della curva, mentre il termine $(ab)/c$ è collegato al tasso di fecondità per età massimo.

La funzione Gamma (Hoem et al., 1981) è data da,

$$f(y) = R\frac{1}{\Gamma(b)c^b}(y-d)^{b-1}exp\left\{-\left(\frac{y-d}{c}\right)\right\}, \qquad \text{per } y > d$$

dove, d rappresenta l'età di gravidanza più bassa, mentre il parametro R determina il livello di fecondità. I parametri b e c non hanno una interpretazione demografica diretta, ma Hoem et al. (1981) hanno riparametrizzato il modello sostituendo questi parametri con la moda m, la media μ e la varianza σ^2 della densità, dove $c = \mu - m$ e $b = (\mu - d)/c = \sigma^2/c^2$.

La funzione Beta di Hoem et al. (1981) è data dalla formula

$$f(y) = R\frac{\Gamma(A+B)}{\Gamma(A)\Gamma(B)}(\beta-\alpha)^{-(A+B-1)}(y-\alpha)^{A-1}(\beta-y)^{B-1}, \text{per } \alpha < y < \beta$$

I parametri sono collegati alla media ν e la varianza τ^2 attraverso le relazioni

$$B = \left\{\frac{(\nu-\alpha)(\beta-\nu)}{\tau^2} - 1\right\}\frac{\beta-\nu}{\beta-\alpha} \qquad \text{e} \qquad A = B\frac{\nu-\alpha}{\beta-\nu}.$$

Come Hoem et al. (1981) menzionano, i parametri α e β sono frequentemente interpretati come il limite di età inferiore e superiore di fecondità.

Schmertmann (2003) propose un modello alternativo per rappresentare il tasso di fecondità per età. Il modello proposto è dato da,

$$f(y) = \begin{cases} R \sum_{k=0}^{4} \theta_k (y - t_k)_+^2, & \alpha \leq y \leq \beta \\ 0, \qquad altrimenti \end{cases}$$

dove i nodi $t_0 < t_1 < ... < t_4$ ricadono nell'intervallo tra le età α e β, dove $t_0 = \alpha$ (la più bassa età di gravidanza) e $(y - t_k)_* \equiv MAX[0, y - t_k]$.

Come Schmertmann (2003) menziona, il modello spline quadratico è davvero utile per descrivere la forma di molte tabelle di fecondità ma richiede tredici parametri da stimare ed il loro significato è poco chiaro. Pertanto costruì un modello spline nel quale tre indici di età [α, P, H] determinano la forma della funzione $f(x)$, mentre il parametro R determina il livello di fecondità. I tre indici di età [α, P, H] sono, rispettivamente, l'età più giovane alla quale la fecondità cresce sopra lo zero, l'età alla quale la fecondità raggiunge il suo livello di picco e l'età più giovane oltre P alla quale la fecondità scende a metà del suo livello di picco. La riduzione del numero di parametri è ottenuto determinando i nodi di posizione dagli indici di età e imponendo restrizioni matematiche così che la funzione spline mimi le caratteristiche comuni degli ASFR.

Un'ulteriore distribuzione che può essere usata per specificare $h(\cdot; \theta_2, ..., \theta_r)$ è la distribuzione normale asimmetrica, la quale verrà trattata nel successivo capitolo.

1.5 I nostri dati

I dati che abbiamo a disposizione riguardano le nascite in Costa Rica dal 1972 al 2012 e le proiezioni della popolazione femminile, sempre del Costa Rica, dal 1950 al 2100. Per quanto riguarda i dati sulle proiezioni è stato preso in considerazione lo stesso periodo di quelli riguardanti le nascite, cioè dal 1972 al 2012. I dati sono divisi per età della madre nel caso del dataset sulle nascite e per età della donna nel caso del dataset sulle proieizioni della popolazione femminile. Questo ci ha permesso di definire una finestra di fecondità delle donne che è stata scelta dai 12 ai 55 anni. Le altre età sono state scartate perché considerate poco realistiche (per esempio, in alcuni anni erano presenti nascite all'età di 6 anni, probabilmente dovute a rilevazioni errate dei dati) o per mancanza di dati.

Nelle Tabelle 1.1 e 1.2 vengono riportate, per ogni anno, l'età media delle madri alla nascita dei figli e l'età media della popolazione femminile. Come si può notare l'età media delle madri si attesta sempre tra i 24 e i 26 anni, mentre l'età media della popolazione femminile è, all'inizio, intorno ai 27 anni e con il passare degli anni aumenta fino ad attestarsi sui 31 anni. Questo fenomeno può essere dovuto al fatto che le condizioni igeniche e sanitarie sono migliorate con il trascorrere degli anni, portando ad un inalzamento dell'aspettativa di vita della popolazione.

Anno	Età media delle madri	Età media delle donne
1972	25.89	27.35
1973	25.64	27.29
1974	25.32	27.24
1975	25.07	27.23
1976	24.93	27.23
1977	24.91	27.25
1978	25.01	27.28
1979	24.98	27.33
1980	24.94	27.42
1981	25.03	27.54
1982	25.15	27.67
1983	25.16	27.82
1984	25.29	27.98
1985	25.49	28.15
1986	25.60	28.32
1987	25.62	28.50
1988	25.68	28.67
1989	25.81	28.85
1990	25.95	28.99
1991	25.93	29.13

Tabella 1.1: Età media delle madri alla nascita dei figli ed età media della popolazione femminile del Costa Rica, 1972-1991.

Nella Figura 1.1 è rappresentato l'istogramma delle nascite per età della madre; a titolo d'esempio è stato considerato l'anno 1989. Si può esservare come la forma di questo istogramma segua, come era ovvio aspettarsi, quella spiegata nel Paragrafo 1.4 riguardo agli ASFR.

Grazie ai dataset a nostra disposizione sono stati calcolati gli ASFR, tramite la formula (1.1), e i TFR per ogni anno. Nella Figura 1.2 si può vedere la distribuzione dei tassi di fecondità per età del Costa Rica; anche in questo caso è stato considerato a titolo d'esempio l'anno 1989.

Anno	Età media delle madri	Età media delle donne
1992	25.99	29.26
1993	25.98	29.38
1994	25.87	29.48
1995	25.84	29.59
1996	25.78	29.71
1997	25.72	29.80
1998	25.58	29.85
1999	25.47	29.95
2000	25.36	30.09
2001	25.44	30.21
2002	25.41	30.29
2003	25.34	30.41
2004	25.33	30.55
2005	25.31	30.67
2006	25.24	30.80
2007	25.17	30.92
2008	25.16	31.03
2009	25.30	31.14
2010	25.41	31.26
2011	25.47	31.37
2012	25.50	31.48

Tabella 1.2: Età media delle madri alla nascita dei figli ed età media della popolazione femminile del Costa Rica, 1992-2012.

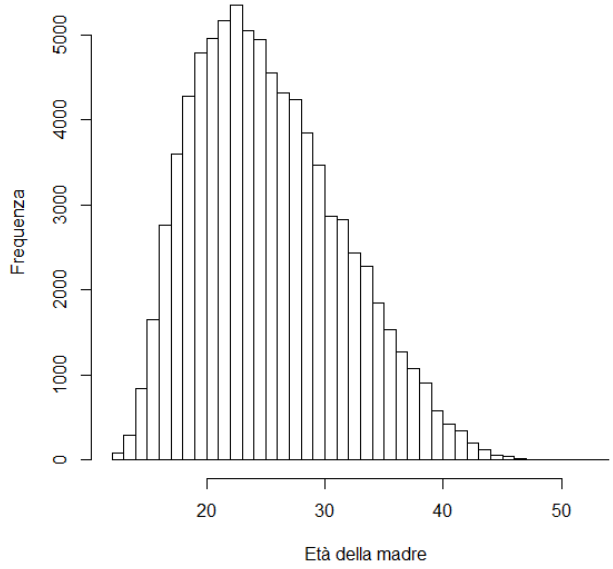

Figura 1.1: Istogramma delle nascite per età della madre nel Costa Rica, 1989.

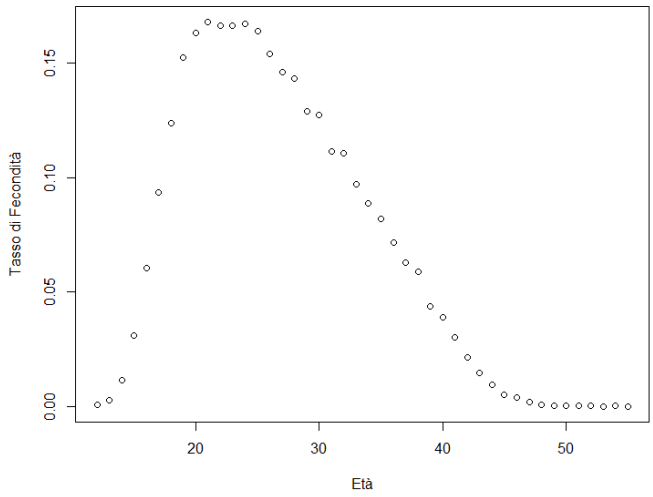

Figura 1.2: Tassi di fecondità per età del Costa Rica nel 1989.

In Figura 1.3 viene fatto un confronto tra le curve di fecondità del 1972, 1985, 1999, 2012 (scelti in modo da essere equidistanti gli uni dagli altri) mantenendo fissa la scala degli ASFR in modo da poter osservare il cambiamento di quest'ultimi a distanza di diversi anni. Come si può notare, gli ASFR, con il passare degli anni, tendono a diminuire vistosamente; a conferma di ciò vi è anche il grafico dell'andamento del TFR in Figura 1.4 che risulta, appunto, essere decrescente con il passare degli anni. Due possibili cause possono essere la nascita di un programma di controllo delle nascite o una maggiore istruzione e informazione della popolazione sulle malattie veneree, che ha portato, di conseguenza, ad un uso maggiore dei sistemi contraccettivi.

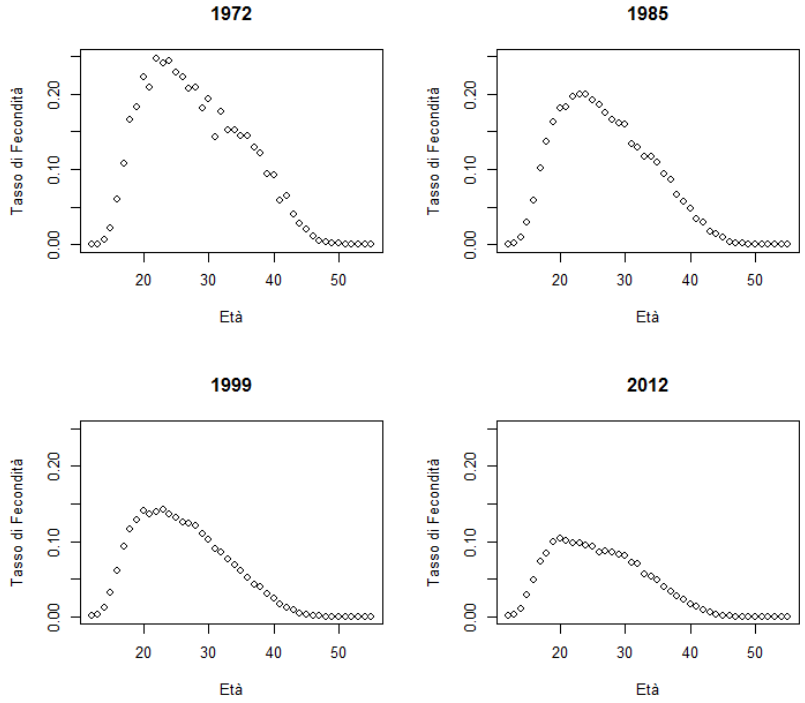

Figura 1.3: Tassi di fecondità per età degli anni 1972, 1985, 1999, 2012.

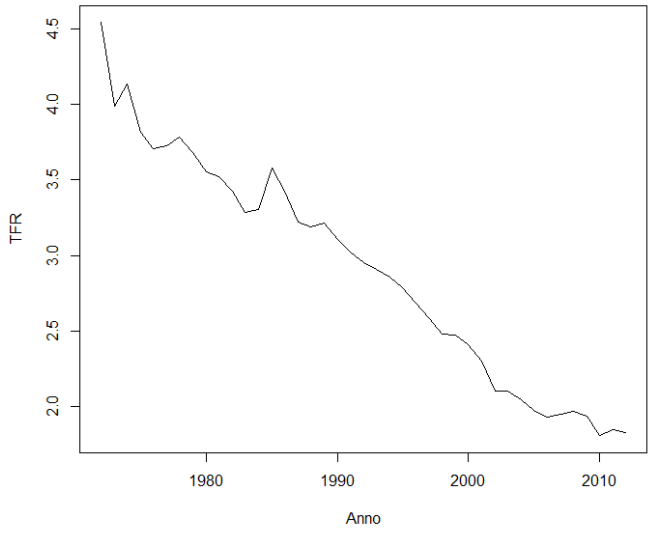

Figura 1.4: Tassi di fecondità totale del Costa Rica, 1972-2012.

Capitolo 2

La famiglia di distribuzioni normale asimmetrica

2.1 La distribuzione normale asimmetrica

2.1.1 Un utile lemma

Di seguito viene presentato un lemma centrale per il nostro sviluppo.

Lemma 2.1. *(Azzalini 1985) Sia f_0 una funzione di densità di probabilità in $\mathbb{R}^d$, sia $G_0(\cdot)$ una funzione di una distribuzione continua sull'asse reale, e sia $w(\cdot)$ una funzione a valore reali in $\mathbb{R}^d$, tale che*

$$f_0(-x) = f_0(x), w(-x) = -w(x), G_0(-y) = 1 - G_0(y) \tag{2.1}$$

per tutti i valori di $x \in \mathbb{R}^d$,$y \in \mathbb{R}$. Allora

$$f(x) = 2f_0(x)G_0\{w(x)\} \tag{2.2}$$

è una funzione di densità in $\mathbb{R}^d$.

Questo lemma ci permette di manipolare una densità "base" f_0 simmetrica attraverso una funzione di "perturbazione" $G_0\{w(x)\}$ per ottenere una nuova densità legittima f. Dato che c'è molta libertà per la scelta

degli ingredienti G_0 e w, allora l'insieme di distribuzioni che può essere ottenuto partendo da una data "base" f_0 è vasto. Il set di densità "perturbate" include sempre la densità "base", dato che $w(x) \equiv 0$ fornisce $f_0 = f$.

2.1.2 Definizione e prime proprietà

Azzalini (1985) sceglie, in (2.2), $f_0 = \phi$ come densità base e $G_0 = \Phi$, la funzione di densità e la funzione di distribuzione di una N(0,1), rispettivamente, e $w(x) = \alpha x$, per qualche valore reale α e definisce la funzione di densità di una distribuzione normale asimmetrica (skew-normal, SN)

$$\phi(x;\alpha) = 2\phi(x)\Phi(\alpha x) \qquad (-\infty < x < \infty), \tag{2.3}$$

la cui rappresentazione grafica è mostrata nella Figura 2.1 per alcune scelte di α. La funzione integranda di $\phi(x;\alpha)$ sarà definita come $\Phi(x;\alpha)$.

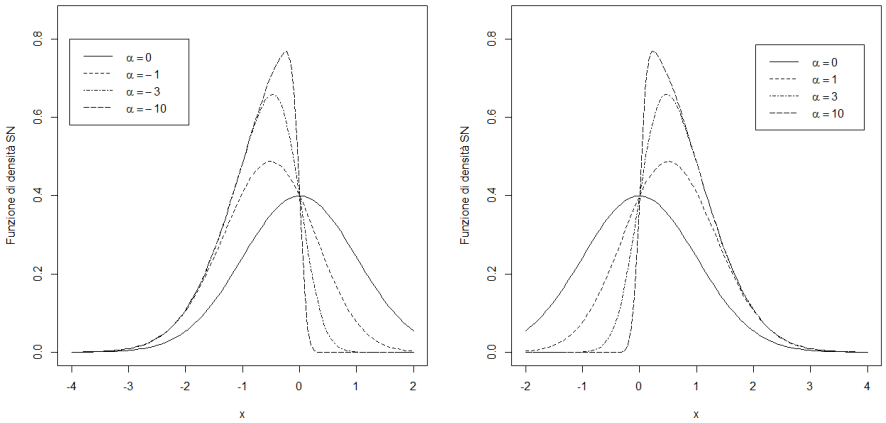

Figura 2.1: Funzione di densità normale asimmetrica quando $\alpha = 0, -1, -3, -10$ nel grafico di sinistra, e $\alpha = 0, 1, 3, 10$ nel grafico di destra.

Per l'ambito applicativo vengono introdotti i parametri di posizione e di scala. Se Z è una variabile casuale continua con funzione di densità

(2.3), allora la variabile

$$Y = \xi + \omega Z \qquad (\xi \in \mathbb{R}, \omega \in \mathbb{R}^+) \tag{2.4}$$

sarà chiamata variabile normale asimmetrica con parametro di posizione ξ, parametro di scala ω e parametro di forma α. La sua funzione di densità per $x \in \mathbb{R}$ è

$$\frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\frac{x-\xi}{\omega}\right) \equiv \frac{1}{\omega}\phi\left(\frac{x-\xi}{\omega};\alpha\right) \tag{2.5}$$

e possiamo scrivere

$$Y \sim SN(\xi, \omega^2, \alpha),$$

dove il quadrato di ω viene usato per analogia con la notazione $N(\mu, \sigma^2)$. Quando $\xi = 0$, $\omega = 1$ e torniamo alla densità (2.3), diciamo che la distribuzione è "normalizzata". Questo è il caso che considereremo molto frequentemente nel presente capitolo.

Riprendiamo alcune semplici proprietà da Azzalini (2014).

Proposizione 2.1. *Se Z indica una variabile casuale SN(0,1,α), avente funzione di densità $\phi(x;\alpha)$, sono vere le seguenti proprietà:*

(a) $\phi(x;0) = \phi(x)$ *per tutte le* x;

(b) $\phi(0;\alpha) = \phi(0)$ *per tutti gli* α;

(c) $-Z \sim SN(0,1,-\alpha)$, *equivalentemente* $\phi(-x;\alpha) = \phi(x;-\alpha)$ *per tutte le* x;

(d) $\lim_{\alpha\to\infty} \phi(x;\alpha) = 2\phi(x)I_{[0,\infty)}(x)$, *per tutte le* $x \neq 0$;

(e) $Z^2 \sim \chi^2_1$, *a prescindere da* α;

(f) *se* $Z' \sim SN(0,1,\alpha')$ *con* $\alpha' < \alpha$, *allora* $Z' <_{st} Z$.

La distribuzione limite (d) è chiamata distribuzione χ_1 o anche distribuzione semi-normale. L'affermazione (f) implica $E[Z'] < E[Z]$ e simili disuguaglianze tra i quantili di ogni livello.

La distribuzione normale standard è un elemento della famiglia delle densità normali asimmetriche, come indicato dalla proprietà (a). Per valori positivi di α otteniamo una distribuzione asimmetrica a destra, e per α negativo una distribuzione asimmetrica a sinistra. Un'altra importante connessione con la famiglia normale è la proprietà chi-quadrato (e).

2.1.3 La funzione generatrice dei momenti ed alcune implicazioni

Il seguente risultato sulla distribuzione normale è stato presentato ripetutamente in letteratura, con o senza dimostrazione; gli autori che hanno fornito una dimostrazione includono Ellison (1964) e Zacks (1981, pp. 53-54). Il risultato di Ellison è in sostanza più generale.

Lemma 2.2. *(Azzalini 1985) Se* $U \sim N(0,1)$ *allora*

$$\mathbb{E}\{\Phi(hU+k)\} = \Phi\left(\frac{k}{\sqrt{1+h^2}}\right), \qquad h,\, k \in \mathbb{R}. \tag{2.6}$$

Da questo risultato, la funzione generatrice dei momenti di Y è facilmente ottenuta, ed è

$$\begin{aligned} M(t) &= \mathbb{E}\{\exp(\xi t + \omega Z t)\} \\ &= 2\exp(\xi t + (1/2)\omega^2 t^2) \int_{\mathbb{R}} \phi(z-\omega t)\Phi(\alpha z)dz \\ &= 2\exp(\xi t + (1/2)\omega^2 t^2)\Phi(\delta\omega t) \end{aligned} \tag{2.7}$$

dove

$$\delta = \delta(\alpha) = \frac{\alpha}{\sqrt{1+\alpha^2}}, \qquad \delta \in (-1,1). \tag{2.8}$$

La moltiplicazione di (2.7) con la funzione generatrice dei momenti della distribuzione $N(\mu,\sigma^2)$, $\exp(\mu t + \sigma^2 t^2/2)$, è ancora una funzione del tipo (2.7). Dopo una semplice riduzione, otteniamo la seguente affermazione.

Proposizione 2.2. *(Chiogna 1998) Se* $Y_1 \sim SN(\xi, \omega^2, \alpha)$ *e* $Y_2 \sim N(\mu, \sigma^2)$ *sono variabili casuali indipendenti, allora*

$$Y_1 + Y_2 \sim SN(\xi + \mu, \omega^2 + \sigma^2, \tilde{\alpha}), \quad \tilde{\alpha} = \frac{\alpha}{\sqrt{1 + (1 + \alpha^2)\sigma^2/\omega^2}}. \tag{2.9}$$

Il parametro di forma $\tilde{\alpha}$ di $Y_1 + Y_2$ è più piccolo in valore assoluto del parametro di forma α di Y_1. Un altro aspetto da notare è che

$$\lim_{\alpha \to \pm\infty} \tilde{\alpha} = \pm\frac{\omega}{\sigma}. \tag{2.10}$$

Si consideri ora il caso in cui Y_1 e Y_2 sono entrambi "propri", cioè con un parametro di forma non nullo. Risulta che $Y_1 + Y_2$ non è del tipo SN, cioè la famiglia SN non è chiusa sotto convoluzione. Per la dimostrazione di questa proprietà si rimanda ad Azzalini (2014, pag. 27).

Una estensione del Lemma 2.2 per variabili SN può essere ottenuta come segue. Se $Z \sim SN(0, 1, \alpha)$ e $U \sim N(0, 1)$ sono variabili indipendenti, allora

$$\begin{aligned} \mathbb{E}\{\Phi(hZ + k)\} &= \mathbb{E}\{\mathbb{P}\{U \leq hz + k | Z = z\}\} \\ &= \mathbb{P}\{U - hZ \leq k\} \end{aligned} \tag{2.11}$$

e, utilizzando la Proposizione 2.2 per la distribuzione di $U - hZ$, arriviamo alla prima affermazione sottostante; la seconda è ottenuta in maniera simile.

Proposizione 2.3. *(Chiogna 1998) Se* $Z \sim SN(0, 1, \alpha)$ *e* $U \sim N(0, 1)$*, allora*

$$\mathbb{E}\{\Phi(hZ + k)\} = \Phi\left(\frac{k}{\sqrt{1 + h^2}}; -\frac{h\alpha}{\sqrt{1 + h^2 + \alpha^2}}\right), \tag{2.12}$$

$$\mathbb{E}\{\Phi(hU + k\alpha)\} = \Phi\left(\frac{k}{\sqrt{1 + h^2}}; \frac{\alpha}{\sqrt{1 + h^2(1 + \alpha^2)}}\right). \tag{2.13}$$

2.1.4 Rappresentazioni stocastiche

Una delle caratteristiche più attraenti della famiglia SN è che ammette una varietà di rappresentazioni stocastiche (si veda Azzalini, 2014). Queste sono utili per la generazione di numeri casuali, e in alcuni casi forniscono una motivazione per l'adozione della famiglia SN come modello stocastico per descrivere i dati.

Campionamento condizionato e selettivo

Una variabile $Z \sim SN(0, 1, \alpha)$ può essere ottenuta da una delle rappresentazioni

$$Z = \begin{cases} X_0 & se\, U < \alpha X_0, \\ -X_0 & altrimenti, \end{cases} \qquad Z = (X_0 | U < \alpha X_0), \tag{2.14}$$

dove X_0 e U sono variabili N(0, 1) indipendenti. Per lo scopo della generazione pseudo-casuale di numeri, la prima variante è chiaramente più efficiente, dato che non richiede nessun rifiuto, mentre l'ultima variante è più utile per le considerazioni teoriche.

Possiamo riesprimere questo costrutto introducendo la varibile normale bivariata (X_0, X_1) con le marginali standardizzate dove

$$X_1 = \frac{\alpha X_0 - U}{\sqrt{1 + \alpha^2}}$$

tale che $\text{cor}\{X_0, X_1\} = \delta(\alpha)$. Allora la rappresentazione (2.14) diventa

$$Z = \begin{cases} X_0 & se\, X_1 > 0 \\ -X_0 & altrimenti, \end{cases} \qquad Z = (X_0 | X_1 > 0). \tag{2.15}$$

Anche se la (2.15) è matematicamente equivalente alla costruzione precedente basata su (X_0, U), la seconda formulazione ha il vantaggio di avere una interpretazione invitante dal punto di vista della modellazione stocastica. In molti casi pratici, una variabile X_0' è osservata quando

un'altra variabile X_1', correlata con la prima, supera una certa soglia, portando ad una situazione di campionamento selettivo. Se questa soglia corrisponde al valore medio di X_1' e la normalità congiunta di (X_0', X_1') è valida, siamo effettivamente nel caso (2.15), fino ad un cambiamento non essenziale di posizione e di scala tra (X_0, X_1) e (X_0', X_1').

Rappresentazione additiva

Si consideri un valore arbitrario $\delta \in (-1, 1)$ e si usi la Proposizione 2.2 nel caso limite (2.10) con $\omega = |\delta|$, $\sigma = \sqrt{1-\delta^2}$ per ottenere la prossima affermazione. Se U_0, U_1 sono variabili indipendenti N(0,1), allora

$$Z = \sqrt{1-\delta^2}U_0 + \delta|U_1| \sim SN(0, 1, \alpha) \tag{2.16}$$

dove

$$\alpha = \alpha(\delta) = \frac{\delta}{\sqrt{1-\delta^2}} \tag{2.17}$$

Minimi e massimi

Si consideri una varibile normale bivariata (X, Y) con le marginali standardizzate e $\text{cor}\{X, Y\} = \rho$, allora

$$\begin{aligned} Z_1 &= \min\{X, Y\} \sim SN(0, 1, -\alpha), \\ Z_2 &= \max\{X, Y\} \sim SN(0, 1, \alpha). \end{aligned} \tag{2.18}$$

dove

$$\alpha = \sqrt{\frac{1-\rho}{1+\rho}}.$$

2.1.5 I momenti

Per calcolare i momenti di $Y \sim SN(\xi, \omega^2, \alpha)$, un metodo è tramite la funzione generatrice dei momenti (2.7) o, equivalentemente, tramite la

funzione generatrice dei cumulanti

$$K(t) = logM(t) = \xi t + (1/2)\omega^2 t^2 + \zeta_0(\delta\omega t) \tag{2.19}$$

dove

$$\zeta_0(x) = \log\{2\Phi(x)\}. \tag{2.20}$$

Possiamo inoltre fare uso delle derivate

$$\zeta_r(x) = \frac{d^r}{dx^r}\zeta_0(x) \qquad (r = 1, 2, ...) \tag{2.21}$$

le cui espressioni, per gli ordini più bassi, sono

$$\begin{aligned}
\zeta_1(x) &= \phi(x)/\Phi(x), \\
\zeta_2(x) &= -\zeta_1(x)\{x + \zeta_1(x)\} \\
&= -\zeta_1(x)^2 - x\zeta_1(x), \\
\zeta_3(x) &= -\zeta_2(x)\{x + \zeta_1(x)\} - \zeta_1(x)\{1 + \zeta_2(x)\} \\
&= 2\zeta_1(x)^3 + 3x\zeta_1(x)^2 + x^2\zeta_1(x) - \zeta_1(x), \\
\zeta_4(x) &= -\zeta_3(x)\{x + 2\zeta_1(x)\} - 2\zeta_2(x)\{1 + \zeta_2(x)\} \\
&= -6\zeta_1(x)^4 - 12x\zeta_1(x)^3 - 7x^2\zeta_1(x)^2 + 4\zeta_1(x)^2 \\
&\quad - x^3\zeta_1(x) + 3x\zeta_1(x),
\end{aligned} \tag{2.22}$$

dove $\zeta_1(x)$ coincide con il rapporto inverso di Mills, cioè il rapporto della funzione di densità di probabilità per la funzione di distribuzione cumulativa di una distribuzione, valutato a $-x$. Tutti gli $\zeta_r(x)$ per $r > 1$ possono essere scritti come funzioni di $\zeta_1(x)$ e potenze di x.

Usando la (2.22), le derivate $K(t)$ fino al quarto ordine sono immediate, si ottiene che

$$\mathbb{E}\{Y\} = \xi + \omega\mu_Z, \tag{2.23}$$

$$\text{var}\{Y\} = (\omega\sigma_Z)^2, \tag{2.24}$$

$$\mathbb{E}\{(Y - \mathbb{E}\{Y\})^3\} = \frac{1}{2}(4 - \pi)(\omega\mu_Z)^3, \tag{2.25}$$

$$\mathbb{E}\{Y - \mathbb{E}\{Y\})^4\} = 2(\pi - 3)(\omega\mu_Z)^4, \tag{2.26}$$

dove

$$\mu_Z = \mathbb{E}\{Z\} = b\delta, \qquad \sigma_Z^2 = \mathrm{var}\{Z\} = 1 - \mu_Z^2 = 1 - b^2\delta^2 \tag{2.27}$$

e

$$b = \zeta_1(0) = \sqrt{2/\pi}. \tag{2.28}$$

La standardizzazione del terzo e quarto cumulante produce le comuni misure di asimmetria e curtosi, cioè

$$\gamma_1\{Y\} = \gamma_1\{Z\} = \frac{4-\pi}{2}\frac{\mu_Z^3}{\sigma_Z^3}, \tag{2.29}$$

$$\gamma_2\{Y\} = \gamma_2\{Z\} = 2(\pi - 3)\frac{\mu_Z^4}{\sigma_Z^4}, \tag{2.30}$$

rispettivamente. Dallo schema di derivate $K(t)$ di ordine maggiore di due,

$$K^{(r)}(t) = (\delta\omega)^r\zeta_r(\omega\delta t), \qquad r > 2,$$

è visibile che l'r-esimo ordine della cumulante di Y è proporzionale a $(\delta\omega)^r$. Sfortunatamente, il calcolo esplicito del termine $\zeta_r(0)$ non sembra fattibile.

Siccome γ_1 e γ_2 sono spesso usati come misure dell'asimmetria e della curtosi in eccesso, rispettivamente, il loro comportamento e il loro intervallo numerico sono interessanti. Dalle espressioni sopra riportate, si vede che dipendono dai parametri solo tramite μ_Z/σ_Z, la quale a sua volta aumenta monotonicamente con α sino a $b/\sqrt{1-b^2}$. Quindi gli intervalli di γ_1 e γ_2 sono

$$(-\gamma_1^{max}, \gamma_1^{max}) \qquad [0, \gamma_2^{max}) \tag{2.31}$$

rispettivamente, dove

$$\gamma_1^{max} = \frac{\sqrt{2}(4-\pi)}{(\pi-2)^{3/2}} \approx 0.9953, \qquad \gamma_2^{max} = \frac{8(\pi-3)}{(\pi-2)^2} \approx 0.8692. \tag{2.32}$$

Questi intervalli non sono molto ampi, mostrando che la famiglia SN non fornisce un adeguato modello stocastico per i casi con alta asimmetria o curtosi. Inoltre, uno non può scegliere γ_1 indipendentemente da γ_2, dato che sono entrambi regolati da α.

2.2 La classe di distribuzione normale asimmetrica unificata

2.2.1 Definizione

Arellano-Valle e Azzalini (2006) introducono una estensione della distribuzione SN base, che servirà successivamente per la stima a posteriori del parametro di forma della nostra distribuzione.

Siano V_0 e V_1 variabili indipendenti tali che

$$V_{0\gamma} \sim LTN_m(-\gamma; 0, \Gamma), \quad V_1 \sim N_d(0, \Psi), \tag{2.33}$$

dove Γ e Ψ sono matrici di correlazione, e la notazione $LTN_m(c; \mu, \Sigma)$ indica una variabile normale multivariata con componenti troncati sotto c, e si consideri la trasformazione

$$Y = \xi + \omega\{B_0 V_{0\gamma} + B_1 V_1\}, \tag{2.34}$$

dove $B_0 = \Delta\Gamma^{-1}$ e B_1 è una matrice $d \times d$ tale che

$$B_1 \Psi B_1^T = \bar{\Omega} - \Delta\Gamma^{-1}\Delta^T \tag{2.35}$$

dove i termini Γ, Δ, $\bar{\Omega}$ sono recuperati da una appropriata partizione della matrice di correlazione

$$\Omega^* = \begin{pmatrix} \Gamma & \Delta^T \\ \Delta & \bar{\Omega} \end{pmatrix}.$$

La funzione di densità di Y diventa

$$f(y) = \phi_d(y - \xi; \omega\bar{\Omega}\omega)\frac{\Phi_m(\gamma + \Delta^T\bar{\Omega}^{-1}\omega^{-1}(y-\xi); \Gamma - \Delta^T\bar{\Omega}^{-1}\Delta)}{\Phi_m(\gamma; \Gamma)} \quad (2.36)$$

per $y \in \mathbb{R}^d$, dove $\Phi_d(\cdot; \Sigma)$ è la funzione di distribuzione di una normale d-variata con matrice di varianze e coviarianze Σ e ω è una matrice diagonale $d \times d$. Questa espressione è chiamata densità normale asimmetrica unificata (unified skew- normal, per semplicità di pronuncia si adotta l'acronimo SUN), quindi scriviamo $Y \sim SUN_{m,d}(\xi, \gamma, \omega, \bar{\Omega}, \Delta, \Gamma)$.

Un caso particolare è presentato nel seguente Lemma 2.3. Questo lemma ci sarà utile per un algoritmo di simulazione per il prelievo di osservazioni dalla distribuzione a posteriori (3.4), che verrà presentata nel Paragrafo 3.1.

Lemma 2.3. *(Canale e Scarpa 2013) Siano $V_0 \sim LTN_q(-\gamma; 0, \Gamma)$, $V_1 \sim N(0,1)$ con V_0 indipendente da V_1 e la notazione $LTN_d(\tau; \mu, \Sigma)$ indica una distribuzione normale d-variata con media μ e matrice di varianze e covarianze Σ troncata a sinistra di τ. Se*

$$Y = \xi + \omega(\Delta\Gamma^{-1}V_0 + \sqrt{1 - \Delta^T\Gamma^{-1}\Delta}V_1),$$

allora $Y \sim SUN_{1,q}(\xi, \gamma, \omega, 1, \Delta, \Gamma)$.

È evidente che la simulazione dal modello qui sopra può essere fatta facilmente affidandosi ad un efficiente algoritmo di campionamento per distribuzioni normali troncate. È necessaria, inoltre, l'inversa della matrice Γ di dimensione $n \times n$ e quindi il costo computazionale totale dipende molto dal calcolo di Γ^{-1}. L'onere computazionale cresce con la dimensione del campione n. Per eseguire una generica inversione di matrice, è ben noto che sono richieste $O(n^3)$ operazioni. Data la particolare espressione per Γ, è disponibile una forma chiusa per la sua inversione. Usando la formula di Sherman-Morrison (p.e., Golub e Van Loan, 1989, p.50), possiamo

scrivere

$$\begin{aligned}\Gamma^{-1} &= (I - D(\Delta)^2 + \Delta\Delta^T) \\ &= diag\{1/(1-\delta_i^2)\} - \frac{1}{1+\sum_{i=1}^n \delta_i^2(1-\delta_i^2)^{-1}} diag\{1/(1-\delta_i^2)\}\Delta\Delta^T \times \\ &\quad diag\{1/(1-\delta_i^2)\} \\ &= diag\{1/(1-\delta_i^2)\} - \frac{1}{1+\sum_{i=1}^n \delta_i^2(1-\delta_i^2)^{-1}} \tilde{\Delta},\end{aligned}$$

dove $\tilde{\Delta}$ è una matrice $n \times n$ con elementi $\tilde{\delta}_{ij} = \delta_i\delta_j(1-\delta_i^2)^{-1}(1-\delta_j^2)^{-1}$.

2.2.2 Alcune proprietà della SUN

Nel seguente paragrafo non vengono discusse nel dettaglio le proprietà formali della famiglia SUN, ma ne vengono soltanto fornite delle brevi descrizioni tratte da Arellano-Valle e Azzalini (2006).

Prendendo in considerazione che la funzione generatrice dei momenti di $V_{0\gamma}$ valutata in s è

$$\exp((1/2)s^T\Gamma s)\frac{\Phi_m(\gamma+\Gamma s;\Gamma)}{\Phi_m(\gamma;\Gamma)}, \quad (s \in \mathbb{R}^m),$$

segue che, sotto la condizione (2.35), la funzione generatrice dei momenti di (2.36) è

$$M(t) = \exp(\xi^T t + (1/2)t^T\Omega t)\frac{\Phi_m(\gamma+\Delta^T\omega t;\Gamma)}{\Phi_m(\gamma;\Gamma)} \quad (t \in \mathbb{R}^d). \qquad (2.37)$$

I momenti e i cumulanti possono essere ottenuti direttamente da (2.37) o da un adeguato adattamento delle espressioni date da Gupta et al. (2004). Il calcolo dei momenti è più semplice quando $\Gamma = diag(\tau_1^2, ..., \tau_m^2)$, dato che la funzione generatrice dei cumulanti si riduce a

$$K(t) = logM(t) = \xi^T t + (1/2)t^T\Omega t + \sum_{j=1}^m log\Phi(\tau_j^{-1}\gamma_j + \tau_j^{-1}\delta_{\cdot j}^T\omega t) - \log\Phi(\gamma;\Gamma),$$

dove $\delta_{\cdot 1}, ..., \delta_{\cdot m}$ sono le colonne di Δ. Da questa espressione otteniamo

$$\mathbb{E}\{Y\} = K'(0) = \xi + \sum_{j=1}^{m} \zeta_1(\tau_j^{-1}\gamma_j)\tau_j^{-1}\omega\delta_{\cdot j}$$

e

$$var\{Y\} = K''(0) = \Omega + \sum_{j=1}^{m} \zeta_2(\tau_j^{-1}\gamma_j)\tau_j^{-2}\omega\delta_{\cdot j}\delta_{\cdot j}^T\omega,$$

dove $\zeta_r(x)$ è la r-esima derivata di $\zeta_0(x) = \log\{2\Phi(x)\}$.

È facile da vedere che, dalla costruzione del Paragrafo 2.2.1, la famiglia SUN è chiusa sotto marginalizzazione e sotto condizionamento.

Consideriamo la distribuzione di una forma quadratica $Q(Z) = Z^T AZ$, dove $Z \sim SUN_{d,m}(0, \gamma, 1_d, \Omega^*)$ e A è una matrice simmetrica $d \times d$ di rango p. Si può dimostrare che la funzione generatrice dei momenti di $Q(Z)$ è

$$M_Q(t) = |I_d - 2tA\bar{\Omega}|^{-1/2} \frac{\Phi_m(\gamma; \Gamma + 2t\Delta^T(I_d - 2tA\bar{\Omega})^{-1}A\Delta)}{\Phi_m(\gamma; \Gamma)}.$$

Un importante caso particolare si ha quando $A = \bar{\Omega}^{-1}$. Per ottenere $Q(Z) \sim \chi^2_d$ in maniera analoga al caso di Z con distribuzione normale o SN, abbiamo bisogno di due condizioni: (a) che $|I_d - 2tA\bar{\Omega}|^{-1/2} = (1-2t)^{-d/2}$, che rimane vera se $S = \bar{\Omega}^{-1}$; (b) che la frazione in $M_Q(t)$ sia pari ad 1. Quest'ultima condizione è soddisfatta quando $\gamma = 0$ se $m = 1$, e quando $A\Delta = 0$ se $m \geq 1$.

Capitolo 3

Il modello per la stima bayesiana dei parametri

In questo capitolo verrà presentato un metodo bayesiano per la stima dei parametri della distribuzione normale asimmetrica. È stato scelto un approccio bayesiano perché, a nostra disposizione, abbiamo i dati per diversi anni, che ci permettono di ricavare delle informazioni a priori dall'anno $t-1$ utili per la stima dei parametri dell'anno t. Un'ulteriore motivazione che ha portato a questa scelta è che la stima di α pone dei problemi intrinsechi. Si assuma che siano noti $\xi = 0$ e $\omega = 1$. In questo caso, la funzione di verosimiglianza per α è solo il prodotto di n funzioni di ripartizione di una normale standard. Se inoltre assumiamo che tutte le osservazioni siano positive (o negative), allora la verosimiglianza è monotona crescente (o decrescente), il che ci porta ad una stima di massima verosimiglianza di +(-) ∞. In aggiunta, sia con osservazioni positive che negative, la verosimiglianza profilo per α ha sempre un punto stazionario in zero e la funzione di verosimiglianza può essere piuttosto piatta.

Non esiste una forma esplicita per la stima bayesiana della distribuzione normale asimmetrica, per questo si ricorrerà a metodi computazionali basati sul metodo MCMC (Markov Chain Monte Carlo).

3.1 Verosimiglianza e specificazioni a priori

Per prima cosa, in questo paragrafo, viene ipotizzato che ξ e ω siano noti in modo da riuscire ad ottenere la distribuzione condizionata di α. Successivamente, nel Paragrafo 3.3, verrà presentato l'algoritmo di Gibbs sampling proposto da Canale e Scarpa (2013) con tutti i parametri ignoti. Quindi assumiamo che ξ, ω siano noti e, senza perdita di generalità, che $\xi = 0$ e $\omega = 1$. La verosimiglianza del modello (2.5) per un campione i.i.d. $y = (y_1, ..., y_n)$ di dimenzione n è

$$f(y; \alpha) = \prod_{i=1}^{n} 2\phi(y_i)\Phi(\alpha y_i). \tag{3.1}$$

Canale e Scarpa (2013) hanno proposto due distribuzioni a priori informative per il parametro scalare di forma α. La prima è semplicemente una normale e può essere scelta con l'obiettivo di centrare la priori su una particolare ipotesi di α. La seconda proposta, invece, è una distribuzione normale asimmetrica. Quest'ultima proposta, che viene presentata di seguito, sarà quella che verrà utilizzata nell'algoritmo di Gibbs sampling.

A priori normale asimmetrica per α

Assumiamo a priori che il parametro α sia distribuito come una normale asimmetrica, cioè,

$$\alpha \sim \pi(\alpha), \pi(\alpha) = \frac{2}{\psi_0}\phi\left(\frac{\alpha - \alpha_0}{\psi_0}\right)\Phi\left(\lambda_0\frac{\alpha - \alpha_0}{\psi_0}\right), \tag{3.2}$$

dove α_0 e ψ_0 sono rispettivamente gli iperparametri di posizione e scala e λ_0 è l'iperparametro di forma che riflette il nostro pensiero sulla direzione

dell'asimmetria. In questo caso la distribuzione a posteriori diventa

$$\begin{aligned}\pi(\alpha; y) &\propto \phi\left(\frac{\alpha-\alpha_0}{\psi_0}\right)\Phi\left(\lambda_0\frac{\alpha-\alpha_0}{\psi_0}\right)\prod_{i=1}^{n}\Phi(\alpha y_i) \\ &\propto \phi\left(\frac{\alpha-\alpha_0}{\psi_0}\right)\Phi_{n+1}\left(\begin{bmatrix} y\alpha_0 \\ 0 \end{bmatrix} + \begin{bmatrix} y \\ \lambda_0/\psi_0 \end{bmatrix}(\alpha-\alpha_0); I_{n+1}\right), \quad (3.3)\end{aligned}$$

dove I_d è la matrice identità di dimensione d.

La funzione di densità nell'equazione (3.3) appartiene anch'essa alla classe di distribuzioni SUN, più precisamente

$$\alpha|y \sim SUN_{1,n+1}(\alpha_0, \gamma, \psi_0, 1, \Delta, \Gamma) \quad (3.4)$$

dove $\Delta = [\delta_i]_{i=1,\dots,n+1}$ con $\delta_i = \psi_0 z_i(\psi_0^2 z_i^2 + 1)^{-1/2}$ e $z = (y^T, \lambda_0\psi_0^{-1})^T$, $\gamma = (\Delta_{1:n}\alpha_0\psi_0^{-1}, 0)$, $\Gamma = I - D(\Delta)^2 + \Delta\Delta^T$, e dove $D(V)$ è una matrice diagonale, i cui elementi coincidono con quelli del vettore V.

Un interessante caso, da un punto di vista pratico, si ottiene considerando $\alpha_0 = 0$. Questa scelta per l'iperparametro è equivalente ad avere un'informazione a priori grezza solo sul lato dell'asimmetria della distribuzione dei dati: infatti, assumendo valori positivi o negativi per l'iperparametro di forma λ_0, si mette maggior massa a priori sul semi-asse positivo o negativo.

La media e varianza a posteriori, nel caso sopra esposto, diventano

$$\begin{aligned}&\mathbb{E}\{\alpha; y\} = \zeta_1(0_n; \tilde{\Gamma}), \\ &\text{var}\{\alpha; y\} = \psi_0^2 + \zeta_2(0_n; \tilde{\Gamma}),\end{aligned}$$

dove 0_n è un vettore $n \times 1$ di zeri, $\zeta_k(x; \Sigma)$ è la k-esima derivata di $\log(2\Phi_n(x, \Sigma))$ con $x \in \mathbb{R}^n$, e la $\tilde{\Gamma}$ è una matrice semidefinita positiva con $1/(\delta_i^2)$ sulla diagonale e 1 sugli elementi fuori dalla diagonale ottenuta come $\tilde{\Gamma} = D(\Delta)^{-1}\Gamma_i D(\Delta)^{-1}$.

3.2 Elicitazione a priori

Come già detto all'inizio di questo capitolo, nel nostro caso abbiamo a disposizione i dati per diversi anni e quindi possiamo ricavare delle utili informazioni per elicitare le a priori. Gli anni contigui avranno una distribuzione dei dati simile tra di loro, quindi le informazioni degli anni precedenti possono fornirci già dei dettagli utili su come sarà la distribuzione negli anni successivi. In particolare si possono usare le stime dei parametri ottenute nell'ultimo anno per le a priori dell'anno successivo. Tuttavia, per quanto riguarda il parametro di forma non c'è una coniugazione tra la a priori e la a posteriori. Infatti la a priori è una distribuzione SN mentre la a posteriori è appartenente alla classe di distribuzioni SUN. Per questo motivo dobbiamo seguire una strada alternativa, dove useremo le relazioni che legano i momenti della distribuzione SN con i suoi parametri. Richiamiamo, quindi, le formule enunaciate nel Paragrafo 2.1.5

$$\begin{aligned}
&\mathbb{E}\{Y\} = \xi + \omega\sqrt{2/\pi}\delta \\
&\text{var}\{Y\} = \omega^2(1-(2/\pi)\delta^2) \\
&\gamma_1\{Y\} = \frac{4-\pi}{2}\left[\frac{\mu_Z^2}{\sigma_Z^2}\right]^{3/2} \\
&\gamma_2\{Y\} = 2(\pi-3)\left[\frac{\mu_Z^2}{\sigma_Z^2}\right]^{2}
\end{aligned} \tag{3.5}$$

dove $\delta = \alpha/\sqrt{1+\alpha^2}$, $\mu_Z = \sqrt{2/\pi}\delta$ e $\sigma_Z^2 = 1-(2/\pi)\delta^2$. In questo modo, ad esempio, è sufficiente invertire la terza equazione della (3.5) per ottenere il parametro α per la nostra a priori della distribuzione SN dell'anno t usando i momenti della distribuzione SUN ottenuta a posteriori nell'anno $t-1$.

Una ulteriore considerazione va fatta per il segno dell'asimmetria, che può essere facilmente incorporato usando π in (3.2) centrata in zero. In questa espressione, un valore positivo (negativo) di λ_0 porta ad una distribuzione a priori asimmetrica che assegna bassa probabilità all'asimmetria negativa (positiva). Quindi alti valori di λ_0 danno bassa probabilità di

avere valori negativi di α e viceversa.

3.3 Calcolo a posteriori

Per fare inferenza sul vettore completo dei parametri specifichiamo una distribuzione normale gamma inversa per i parametri di posizione e scala e la distribuzione a priori descritta nel Paragrafo 3.1 per il parametro di forma. Nello specifico assumiamo che la distribuzione a priori per l'intero vettore dei parametri del modello (3.1) sia

$$\pi(\xi, \omega, \alpha) = N(\xi; \xi_0, k\omega^2) \times Ga(\omega^{-2}; a, b) \times \pi(\alpha; \theta_0) \tag{3.6}$$

dove $\pi(\alpha; \theta_0)$ ha un adeguato vettore dei parametri θ_0.

Seguendo Bayes e Branco (2007), Arellano-Valle et al. (2009) e Gancho et al. (2011) e basandoci sulla rappresentazione stocastica della distribuzione normale asimmetrica introdotta da Azzalini (1986), la quale è un particolare caso del Lemma 2.3, introduciamo delle variabili latenti normali standard $\eta_1, ..., \eta_n$. Condizionatamente a queste variabili latenti, possiamo considerare la generica i-esima osservazione come normalmente distribuita con media $\xi + \delta|\eta_i|$ e varianza $(1 - \delta^2)\omega^2$. Grazie a questa interpretazione otteniamo una coniugazione per i parametri di posizione e scala. Quest'ultimo argomento ha permesso a Canale e Scarpa (2013) di costruire un efficiente algoritmo di Gibbs sampling, il quale si aggiorna attraverso i seguenti passi:

- Aggiornamento di η_i dalla sua distribuzione a posteriori condizionata

$$\eta_i \sim TN_0(\delta(y_i - \xi), \omega^2(1 - \delta^2))$$

  dove δ è $\alpha/\sqrt{\alpha^2 + 1}$ e $TN_\tau(\mu, \sigma^2)$ è una normale troncata sotto τ con media μ e varianza σ^2.

- Campionare (ξ, ω) da

$$N(\hat{\mu}, \hat{\kappa}\omega^2) InvGam(a + (n + 1)/2, b + \hat{b})$$

dove

$$\hat{\mu} = \frac{\kappa \sum_{i=1}^{n}(y_i - \delta\eta_i) + (1-\delta^2)\xi_0}{n\kappa + (1-\delta^2)}$$

$$\hat{\kappa} = \frac{\kappa(1-\delta^2)}{n\kappa + (1-\delta^2)}$$

$$\hat{b} = \frac{1}{2(1-\delta^2)}\left\{\delta^2 \sum_{i=1}^{n}\eta_i^2 - 2\delta\sum_{i=1}^{n}\eta_i(y_i-\xi) + \sum_{i=1}^{n}(y_i-\xi)^2 + \frac{1-\delta^2}{\kappa}(\xi-\xi_0)^2\right\}.$$

- Campionare α da

$$\alpha \sim \pi(\alpha|y^*)$$

dove $y^* = (y_i - \xi)/\omega$ per $i = 1, ..., n$, e $\pi(\alpha|y)$ è la distribuzione a posteriori ottenuta nel Paragrafo 3.1.

Capitolo 4

I risultati del modello

4.1 La preparazione dei dati

Nel modello presentato nel Capitolo 3, le osservazioni che entrano nella verosimiglianza sono l'età della madre al momento della nascita, quindi, partendo dal dateset delle nascite, è stato creato un vettore dei dati per ogni anno con tante osservazioni contententi una specifica età a seconda di quante nascite ci sono state da madri con quella età in quel determinato anno, ad esempio se nel dataset delle nascite nell'anno 1980 all'età di 20 anni ci sono 10 nascite allora nel vettore del 1980 ci saranno 10 osservazioni di valore 20. Visto che, in ogni anno, l'ordine di grandezza delle osservazioni sull'età della madre al momento della nascita si aggira intorno alle decina di migliaia, è stato preso in considerazione un campione casuale di 1000 osservazioni per ogni anno.

4.2 I parametri da inserire nell'algoritmo

Per l'algoritmo di stima, oltre al numero di iterazioni e ad un valore iniziale per i parametri ξ, ω^2 e α, abbiamo bisogno di un valore per gli iperparametri ξ_0, κ, a, b, α_0, ψ_0 e λ_0 della distribuzione a priori (3.6). Nel primo anno (1972) si è cercato di essere il meno informativi possibile, quindi di avere una varianza ampia, di conseguenza sono stati scelti i

seguenti valori per gli iperparametri: $\xi_0 = 18$, $\kappa = 100$, $a = 3$, $b = 7000$, $\alpha_0 = 0$, $\psi_0 = 100$ e $\lambda_0 = 1$. A λ_0 è stato assegnato un valore basso, pari ad 1, perché si è voluto essere poco informativi anche per quanto riguarda l'asimmetria di α, così da dare probabilità $1/2$ ad α di essere asimmetrico a destra e $1/2$ di essere asimmetrico a sinistra. Invece α_0 è stato posto pari a 0 per le considerazioni fatte nell'ultima parte del Paragrafo 3.2. Infine il numero di iterazioni dell'algoritmo è stato scelto uguale a 10000.

Per quanto riguarda gli anni successivi al primo si è tenuto conto di alcune informazioni che vengono fornite dai dati dell'anno precedente. Quindi per i valori iniziali dei parametri si è scelto di inserire le stime a posteriori dell'anno precedente, così come per ξ_0 dove si è scelta la stima a posteriori di ξ dell'anno prima ed, infine, il valore di λ_0 è stato ricavato dall'indice di asimmetria γ_1 delle stime di α dell'anno precedente, estrapolando α dalla terza equazione in (3.5), che nel nostro caso corrisponde a λ_0. Invece i parametri κ, a, b, α_0 e ψ_0 sono mantenuti costanti per ogni anno, perché, anche se le distribuzioni vengono centrate sui valori dell'anno prima, vogliamo mantenere una certa variabilità per permettere alle stime di cambiare di anno in anno e non vincolarci alle stime ottenute nell'anno precedente.

4.3 Le stime

Nel seguente paragrafo verranno presentati i risultati ottenuti grazie all'algoritmo di Gibbs sampling enunciato nel Paragrafo 3.3. Tutti i risultati sono al netto del burn-in che è stato scelto pari a 1000 perché, dopo aver osservato i grafici per i vari parametri, sembra che dopo 1000 iterazioni le catene arrivino a convergenza.

Visto che questo algoritmo comporta un grande carico computazionale e ci sono 41 anni da analizzare, dopo il raggiungimento del burn-in l'algortimo è stato fatto girare in parallelo su più processori caricando su ogni processore le stime ottenute nell'ultima iterazione in modo che proseguisse da dove era rimasto. In questo modo il carico computazionale

viene ripartito sui vari processori e si impiega meno tempo per analizzare tutti i dati.

Di seguito vengono riportati i grafici dei valori stimati dei parametri. Sono riportati a titolo di esempio solo gli anni 1987 e 2004, ma tutte le considerazioni fatte valgono anche per gli altri anni. Questa tipologia di grafici ci fornisce delle informazioni utili per valutare la bontà delle stime ottenute. Infatti le Figure 4.1 e 4.2 mostrano che le stime non presentano un andamento particolare, quindi le catene sono già arrivate a convergenza. Queste indicazioni ci portano a concludere che le stime prodotte dall'algoritmo sono valide.

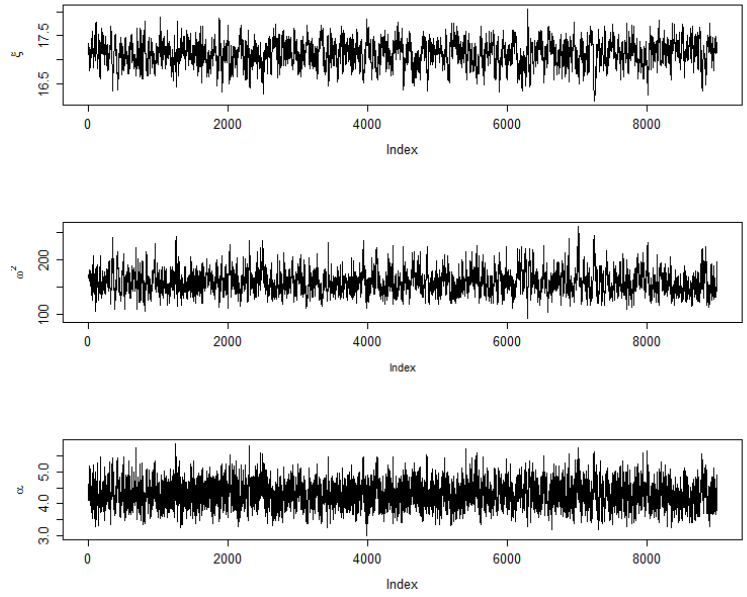

Figura 4.1: Valori stimati dei parametri ξ, ω^2 e α dell'anno 1987.

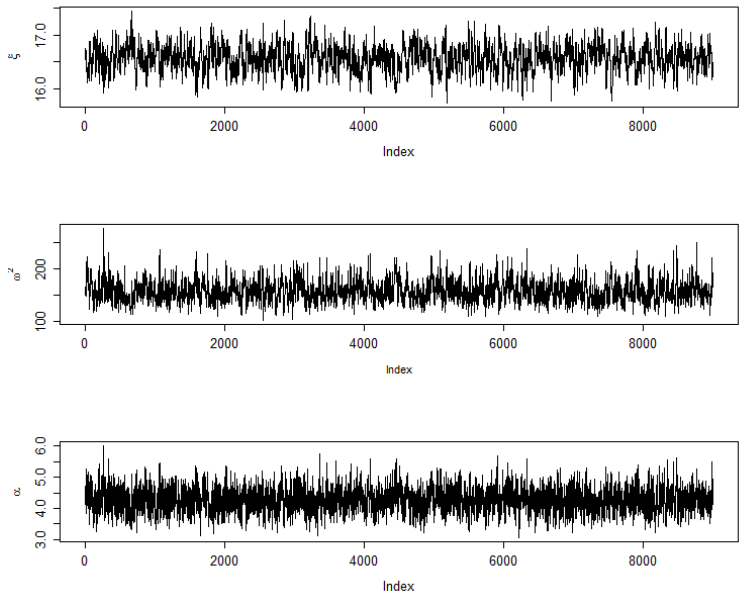

Figura 4.2: Valori stimati dei parametri ξ, ω^2 e α dell'anno 2004.

Nelle Figure 4.3 e 4.4 viene mostrato come si adattano le curve di fecondità ottenute ai dati reali. Come si può notare la curva non riesce a seguire perfettamente l'andamento dei dati ma ci si avvicina. Però, non in tutti i casi le curve stimate hanno una forma simile a quella della distribuzione dei tassi di fecondità. Per degli esempi si vedano le Appendici A e B, dove sono raffigurate le curve di fecondità stimate per ogni anno.

Un altro aspetto interessate da osservare è come cambia la distribuzione dei parametri a priori e a posteriori. Nelle Figure 4.5 e 4.6 si può notare come la distribuzione a priori sia molto piatta, mentre a posteriori sia molto più concentrata.

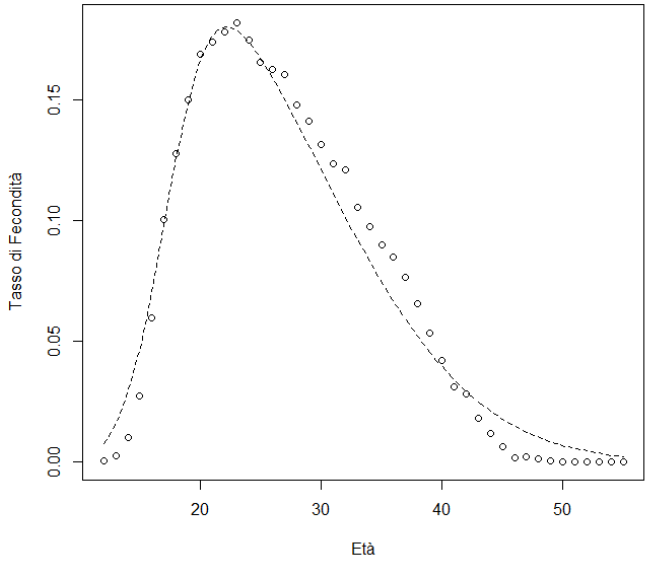

Figura 4.3: Curva di fecondità stimata per l'anno 1987.

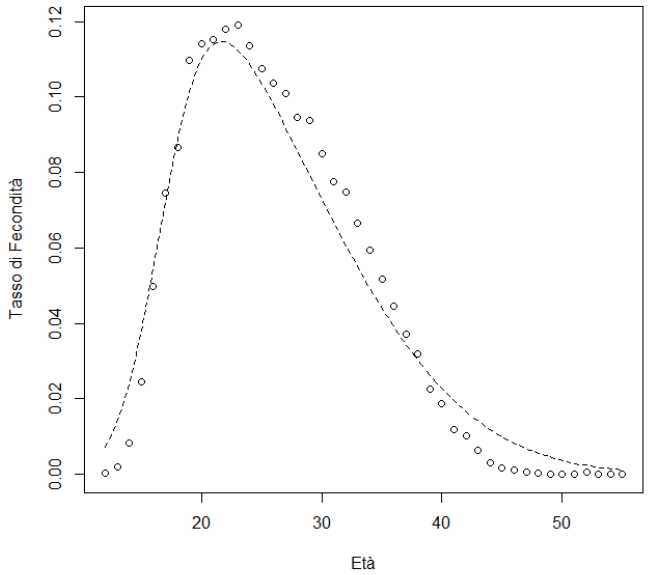

Figura 4.4: Curva di fecondità stimata per l'anno 2004.

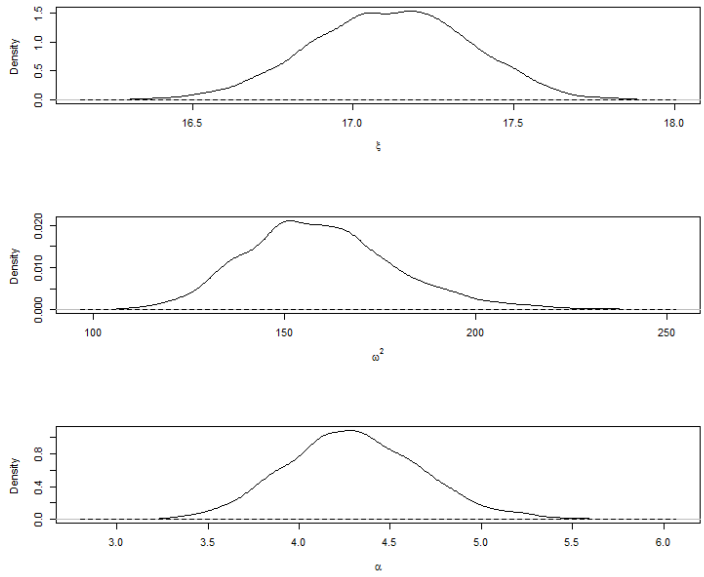

Figura 4.5: Densità a posteriori (linea continua) e a priori (linea tratteggiata) dei parametri ξ, ω^2 e α dell'anno 1987.

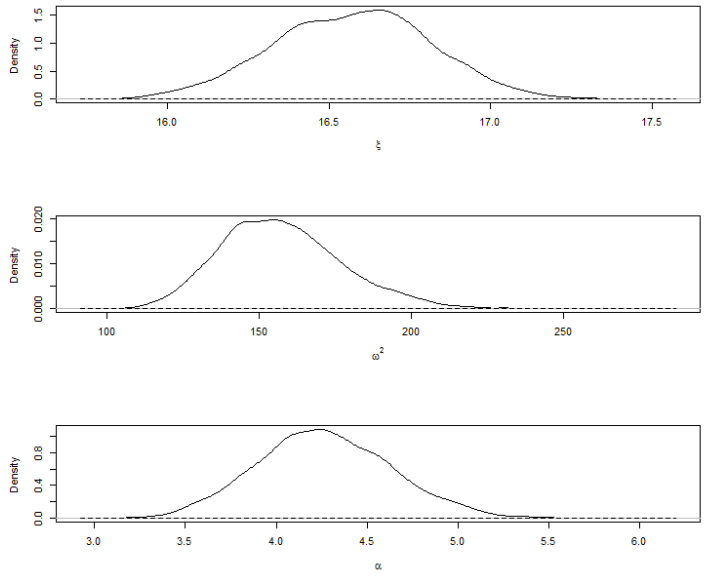

Figura 4.6: Densità a posteriori (linea continua) e a priori (linea tratteggiata) dei parametri ξ, ω^2 e α dell'anno 2004.

Nelle Tabelle 4.1 e 4.2 sono presenti le stime a posteriori dei parametri e la relativa deviazione standard per ogni anno. Come si può notare il valore delle stime non cambia molto al variare degli anni; l'unico parametro che tende a variare un po' di più è ω^2 che presenta un minimo nell'anno 1975 con il valore di 147.8644 e un massimo nell'anno 1994 con 181.5823. Inoltre la deviazione standard è molto piccola per quanto riguarda i parametri ξ e α, mentre è decisamente più grande per ω^2.

Anno	ξ		ω^2		α	
	stima	dev. std.	stima	dev. std.	stima	dev. std.
1972	17.1388	0.24578	156.6401	19.27682	4.1645	0.35421
1973	17.1661	0.25114	157.7434	20.32790	4.1608	0.36796
1974	16.8080	0.23973	153.6669	20.09700	4.2980	0.36808
1975	16.6434	0.24073	147.8644	18.96986	4.2750	0.36408
1976	16.6800	0.25887	158.2510	21.41215	4.3843	0.38848
1977	16.4177	0.24174	148.4466	20.41715	4.4927	0.39069
1978	16.7894	0.24667	158.0739	20.34263	4.2711	0.37118
1979	16.5532	0.24640	161.0131	20.88833	4.3778	0.38275
1980	16.8488	0.24201	162.1815	21.18904	4.3782	0.37236
1981	16.5899	0.26322	159.3603	20.13850	4.1954	0.36155
1982	17.2515	0.26277	162.6507	22.32749	4.3223	0.38605
1983	16.7389	0.25028	162.2176	21.05092	4.5014	0.38522
1984	16.9631	0.24091	153.7903	20.23141	4.4418	0.38423
1985	16.9539	0.27519	168.9478	22.24102	4.2502	0.37617
1986	17.0449	0.29901	177.9705	25.10053	4.3143	0.39580
1987	17.1152	0.24856	159.5909	20.01180	4.3026	0.37624
1988	17.1111	0.27234	162.9908	22.11004	4.3245	0.39371
1989	16.8621	0.27024	170.5852	22.74192	4.2383	0.38171
1990	17.1159	0.28020	176.8802	23.48907	4.2968	0.38568
1991	17.0312	0.26932	173.0241	22.51098	4.2518	0.38185

Tabella 4.1: Stime posteriori e deviazione standard dei parametri ξ, ω^2 e α, 1972-1991.

Anno	ξ		ω^2		α	
	stima	dev. std.	stima	dev. std.	stima	dev. std.
1992	16.8246	0.26750	169.9404	22.06971	4.1405	0.36218
1993	16.9392	0.29751	181.0620	24.74300	4.1298	0.38346
1994	16.8304	0.31103	181.5823	25.51401	4.1113	0.38833
1995	16.8038	0.29201	173.1512	22.13252	4.0426	0.37704
1996	16.9500	0.26515	171.3133	22.23660	4.1849	0.37303
1997	16.8485	0.26999	175.9101	23.31056	4.1718	0.38212
1998	16.5905	0.26914	168.2768	22.65927	4.2188	0.38399
1999	16.4845	0.27820	173.0367	22.06550	4.1023	0.36566
2000	16.2632	0.26991	165.0106	22.26195	4.2281	0.38287
2001	16.7112	0.25516	166.9170	20.61437	4.1152	0.36208
2002	16.8502	0.25146	164.1647	20.51448	4.1796	0.35986
2003	16.6215	0.26280	167.0426	20.82887	4.1607	0.35987
2004	16.5725	0.24218	157.3300	20.18550	4.2675	0.37341
2005	16.7406	0.27249	167.2203	21.98415	4.2255	0.38096
2006	16.5840	0.24316	154.0451	19.71649	4.2907	0.36903
2007	16.5965	0.24593	164.2123	20.92225	4.3150	0.38233
2008	16.0947	0.24169	161.5232	19.56277	4.2741	0.36667
2009	16.5294	0.25708	170.9143	22.25468	4.2637	0.37607
2010	16.4699	0.27554	173.2821	24.42371	4.3136	0.39419
2011	16.5516	0.27118	169.7708	22.80741	4.2909	0.38750
2012	16.6018	0.26697	167.7714	23.02731	4.3225	0.38931

Tabella 4.2: Stime a posteriori e deviazione standard dei parametri ξ, ω^2 e α, 1992-2012.

Grazie alle stime ottenute dal modello si possono ricavare la media, la varianza e l'indice di asimmetria a posteriori utilizzando le prime tre equazioni in (3.5). Le medie per ogni anno di questi valori vengono rappresentate nelle Figure 4.7, 4.8 e 4.9 insieme alle bande di credibilità con livello di confidenza del 95%. Nella Figura 4.7 si può notare che la media a posteriori all'inizio decresce leggermente fino al 1977, per poi crescere fino a stabilizzarsi intorno al 1986 ed infine descrescere di nuovo dal 1997 con un piccolo aumento negli ultimi 4 anni. Invece nella Figura 4.8 si osserva che la varianza a posteriori ha all'inzio un andamento crescente fino al 1994 , poi decresce fino al 2006 ed infine ricrescere leggermente. Infine nella Figura 4.9 si vede che l'indice di asimmetria a posteriori ha un adamento costante nei primi anni fino al 1990, per poi diminuire leggermente fino al 1995 e dopo aumentare nuovamente fino agli ultimi anni.

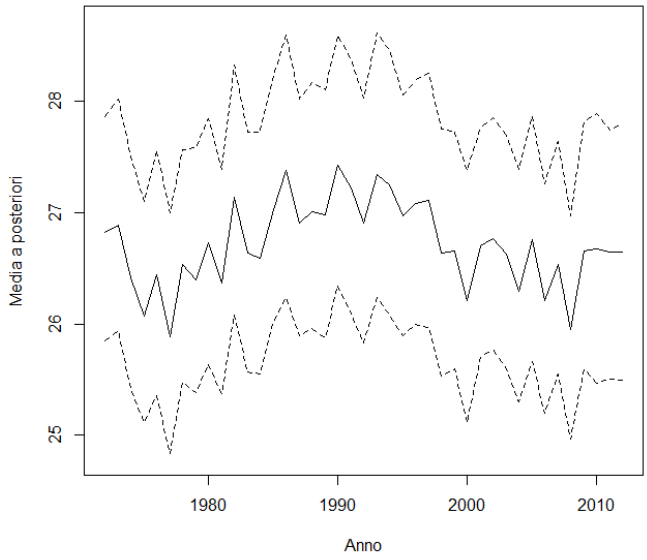

Figura 4.7: Medie a posteriori con bande di credibilità al 95%.

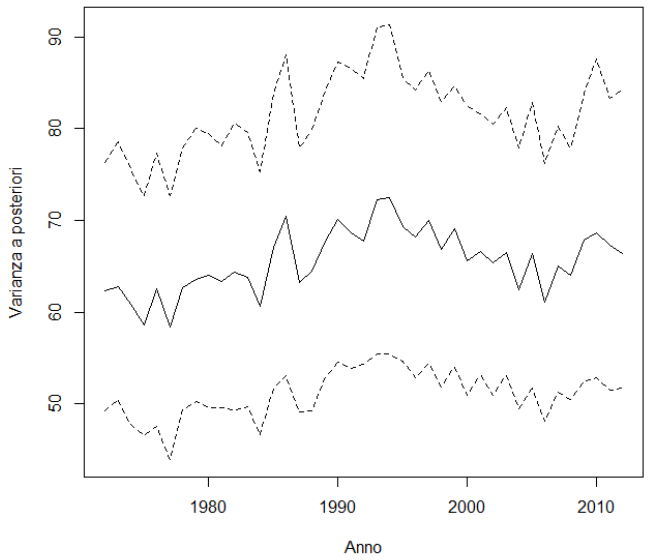

Figura 4.8: Varianze a posteriori con bande di credibilità al 95%.

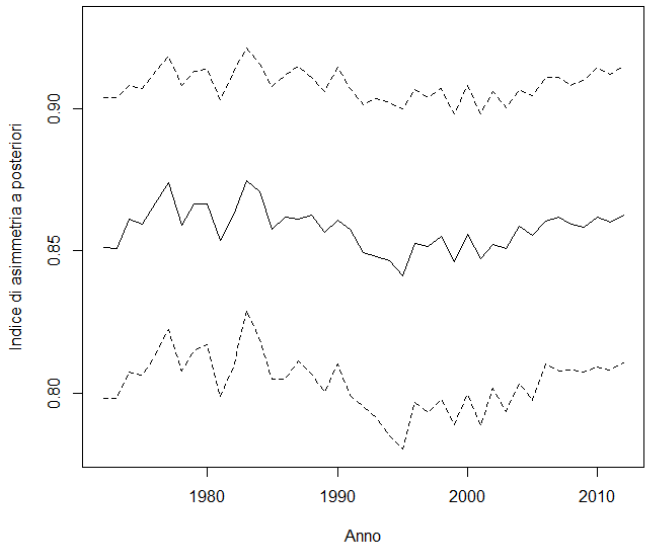

Figura 4.9: Indici di asimmetria a posteriori con bande di credibilità al 95%.

Nella Figura 4.10 sono rappresentate le curve di fecondità stimate dal 1972 al 2012 ad intervalli di 4 anni. In tutti gli anni la curva ha un andamento molto simile, con un valore del tasso di fecondità per età sopra lo zero intorno ai 13 anni, un picco tra i 20 e i 26 anni e nuovamente un valore del tasso di fecondità vicino allo zero dopo i 50 anni.

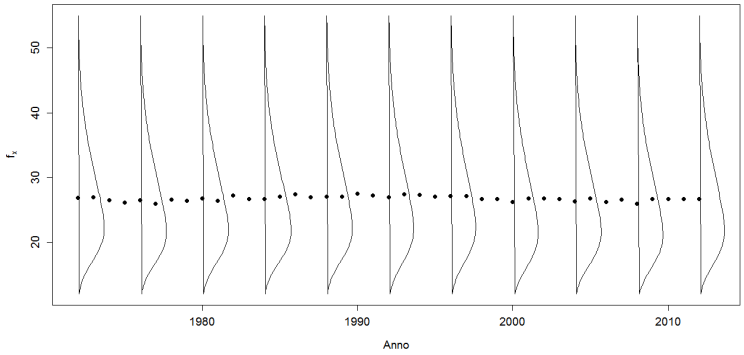

Figura 4.10: Curve di fecondità del Costa Rica, 1972-2012. I punti indicano la media a posteriori per ogni anno.

Capitolo 5

La nostra proposta

Nei capitoli precedenti assumiamo che le osservazioni $y_1, ..., y_n$, dove n è il numero di osservazioni, sono una variabile continua che misura l'età della madre alla nascita del figlio. Tuttavia, noi non osserviamo questa quantità, quello che veramente osserviamo è il fatto che un figlio nasce in una certa classe di età della madre (nel nostro caso queste classi di età sono di un anno). Per questo possiamo pensare che le nostre osservazioni siano dei vettori del tipo $\mathbf{x}_i = [x_{i,12}, ..., x_{i,j}, ..., x_{,55}]$ in cui ciascun elemento assume valore 1 se in quell'anno la donna considerata ha avuto un figlio nella j-esima classe di età e 0 altrimenti, cioè $x_{i,j} = 1$ se $y_i = j$ e 0 altrimenti. Questa variabile si distribuisce come una multinomiale $MN(m, p_1, ..., p_j)$, dove $m = 1$ e p_j è uguale a

$$p_j = \mathbb{P}\{y_i = j\} = \int_{j-1/2}^{j+1/2} f_{SN}(t; \xi, \omega^2, \alpha)dt$$

dove f_{SN} indica la funzione di densità di una variabile normale asimmetrica, y_i è l'età della madre i-esima al momento della nascita del figlio e $j = 12, ..., 55$, cioè j può assumere come valore una tra le età della finestra di fecondità considerata.

Nel seguente capitolo verrà proposto un metodo per la stima dei parametri tramite Gibbs sampling come nel Capitolo 3, ma con l'aggiunta di variabili latenti che permettano di coniugare la distribuzione multinomiale

delle nostre osservazioni con la distribuzione SN, in modo che sia possibile stimare i parametri necessari per la stima delle curve di fecondità.

Visto che i parametri della nostra multinomiale dipendono dai parametri della SN e che grazie alla variabili latenti che introdurremo riusciremo ad ottenere una coniugazione con la distribuzione SN, le considerazioni del Capitolo 3 sono valide anche in questo caso.

5.1 Le variabili latenti

Introduciamo delle variabili latenti continue $v_1, ..., v_n$ aventi distribuzione $SN(\xi, \omega^2, \alpha)$, e osserviamo y_i, dove $y_i = j$ se $j - 1/2 \leq v_i < j + 1/2$. Sia (3.6) la distribuzione a priori dei parametri ξ, ω^2 e α, come nel caso dei capitoli precedenti, allora la distribuzione a posteriori coniugata di $(\xi, \omega^2, \alpha, V)$, dove $V = [v_1, ..., v_n]$, diventa

$$\begin{aligned}\pi(\xi, \omega^2, \alpha, V|Y) \propto & f_N(\xi, \xi_0, \kappa\omega^2) f_{InvGa}(\omega^2; a, b) f_{SN}(\alpha; \alpha_0, \psi_0^2, \lambda_0) \times \\ & \prod_{i=1}^{n} f_{SN}(v_i; \xi, \omega^2, \alpha) \times \left\{ \sum_{j=12}^{55} 1(y_i = j) \times \right. \\ & \left. 1(j - 1/2 \leq v_i < j + 1/2) \right\}. \end{aligned} \tag{5.1}$$

dove $Y = [y_1, ..., y_n]$, $f_N(\cdot; \mu, \sigma^2)$ è la funzione di densità di una normale di parametri μ e σ^2, $f_{InvGa}(\cdot; c, d)$ è la funzione di densità di una gamma inversa di parametri c e d, $f_{SN}(\cdot; \zeta, \rho^2, \tau)$ è la funzione di densità di una normale asimmetrica di parametri ζ, ρ^2 e τ, e $1(X \in A)$ è la funzione indicatrice che è uguale ad 1 se la variabile casuale X è contenuta in A.

Dall'equazione (5.1) è facile ottenere le distribuzioni a posteriori condizionate per i parametri ξ, ω^2 e α. Per trovare tali distribuzioni è sufficiente evidenziare i fattori che dipendono dai parametri che ci interessano,

scartando quelli che coinvolgono gli altri parametri. Le distribuzioni sono

$$\pi(\xi, \omega^2|\alpha, V, Y) \propto f_N(\xi; \xi_0, \kappa\omega^2) f_{InvGa}(\omega^2; a, b) \times \prod_{i=1}^{n} f_{SN}(v_i; \xi, \omega^2, \alpha); \tag{5.2}$$

$$\pi(\alpha|\xi, \omega^2, V, Y) \propto f_{SN}(\alpha; \alpha_0, \psi_0^2, \lambda_0) \prod_{i=1}^{n} f_{SN}(v_i; \xi, \omega^2, \alpha). \tag{5.3}$$

È altrettanto semplice ricavare le distribuzioni a posterirori delle variabili casuali latenti v_i, che, per la singola variabile latente, risulta essere

$$\pi(v_i|\xi, \omega^2, \alpha, y_i = j) \propto f_{SN}(v_i; \xi, \omega^2, \alpha) \times \left\{ \sum_{j=12}^{55} 1(y_i = j) 1(j - 1/2 \leq v_i < j + 1/2) \right\} \tag{5.4}$$

cioè $v_i|\xi, \omega, \alpha, y_i = j \sim SN(\xi, \omega^2, \alpha)$ troncata a sinistra di $j - 1/2$ e a destra di $j + 1/2$.

Quindi la funzione di densità a posteriori della variabile latente v_i diventa

$$\pi(v_i|\xi, \omega^2, \alpha, y_i = j) = \begin{cases} C f_{SN}(v_i; \xi, \omega^2, \alpha) & se\, j - 1/2 \leq v_i < j + 1/2 \\ 0 & altrimenti \end{cases} \tag{5.5}$$

dove $C = 1/\int_{j-1/2}^{j+1/2} f_{SN}(t; \xi, \omega^2, \alpha) dt$ è la costante di normalizzazione, che rende la funzione di densità f_{SN} una densità propria nel supporto della variabile v_i.

5.2 Il nuovo algoritmo

Come detto all'inizio di questo capitolo dovremo ricondurci ai risultati ottenuti nel Capitolo 3, quindi, anche in questo caso, introduciamo delle variabili latenti normali standard $\eta_1, ..., \eta_n$ e, condizionatamente a queste

variabili latenti, possiamo considerare la generica i-esima variabile v_i normalmente distribuita con media $\xi+\delta|\eta_i|$ e varianza $(1-\delta^2)\omega^2$, dove $\delta = \alpha/\sqrt{\alpha^2+1}$. In questo modo si ottiene una coniugazione per i parametri di posizione e scala.

Nel primo passo del seguente algoritmo bisogna generare dei valori dalla densità (5.5). Per fare ciò è stato creato un algoritmo di accettazione-rifiuto utilizzando come distribuzione per le proposte una uniforme nell'intervallo $[j-1/2, j+1/2]$.

In questo modo è stato possibile implementare un algoritmo per la stima dei parametri ξ, ω e α, che si aggiorna attraverso i seguenti passi:

(a) Aggiornamento di v_i dalla distribuzione

$$v_i|\xi,\omega^2,\alpha,y_i=j \sim TSN_{j-1/2,j+1/2}(\xi,\omega^2,\alpha)$$

dove $TSN_{\epsilon-1/2,\epsilon+1/2}(\nu,\iota,\theta)$ è una normale asimmetrica troncata a sinistra di $\epsilon-1/2$ e a destra di $\epsilon+1/2$.

(b) Aggiornamento di η_i dalla sua distribuzione a posteriori condizionata

$$\eta_i \sim TN_0(\delta(v_i-\xi),\omega^2(1-\delta^2))$$

dove δ è $\alpha/\sqrt{\alpha^2+1}$ e $TN_\tau(\mu,\sigma^2)$ è una normale troncata sotto τ con media μ e varianza σ^2.

(c) Campionare (ξ,ω) da

$$N(\hat{\mu},\hat{\kappa}\omega^2)InvGam(a+(n+1)/2, b+\hat{b})$$

dove

$$\hat{\mu} = \frac{\kappa\sum_{i=1}^n(v_i-\delta\eta_i)+(1-\delta^2)\xi_0}{n\kappa+(1-\delta^2)}$$

$$\hat{\kappa} = \frac{\kappa(1-\delta^2)}{n\kappa+(1-\delta^2)}$$

$$\hat{b} = \frac{1}{2(1-\delta^2)}\left\{\delta^2\sum_{i=1}^n\eta_i^2 - 2\delta\sum_{i=1}^n\eta_i(v_i-\xi)+\sum_{i=1}^n(v_i-\xi)^2+\frac{1-\delta^2}{\kappa}(\xi-\xi_0)^2\right\}.$$

(d) Campionare α da

$$\alpha \sim \pi(\alpha|v^*)$$

dove $v^* = (v_i - \xi)/\omega$ per $i = 1, ..., n$, e $\pi(\alpha|v)$ è la distribuzione a posteriori ottenuta nel Paragrafo 3.1.

Per il codice R dell'algoritmo si veda l'Appendice E.

Capitolo 6

I risultati del nuovo modello

Le considerazioni fatte nel Paragrafo 4.2 valgono anche in questo caso, quindi non verranno riprese in questo capitolo. Verranno invece presentati i risultati ottenuti con il nuovo algoritmo proposto nel Paragrafo 5.2.Come nel caso precedente, tutti i risultati sono al netto del burn-in, che è pari 1000 perchè dopo 1000 iterazioni le catene sembrano arrivare a convergenza.

Con il nuovo algoritmo il carico computazionale è aumentato, dato che bisogna aggiornare le nuove variabili latenti $v_1, ..., v_n$ ad ogni iterazione. Quindi, come nel caso precedente, dopo il raggiungimento del burn-in l'algoritmo viene fatto girare in parallelo su più processori, caricando su ogni processore le stime ottenute nell'ultima iterazione, cioè la millesima, in modo che il ciclo presegua da dove era rimasto. In questo modo il calcolo delle stime avviene più velocemente. Diversamente dal precedente modello, il numero di iterazioni dell'algoritmo è stato scelto pari a 5000 perché ritenute sufficienti per ottenere delle buone stime e per rendere ulteriormente più veloci i calcoli.

Nelle Figure 6.1 e 6.2 vengono riportati i grafici dei valori stimati dei parametri. Anche in questo caso, come nel Paragrafo 4.3, vengono presi a titolo d'esempio gli anni 1987 e 2004 ma le considerazioni che vengono fatte valgono anche per gli altri anni.

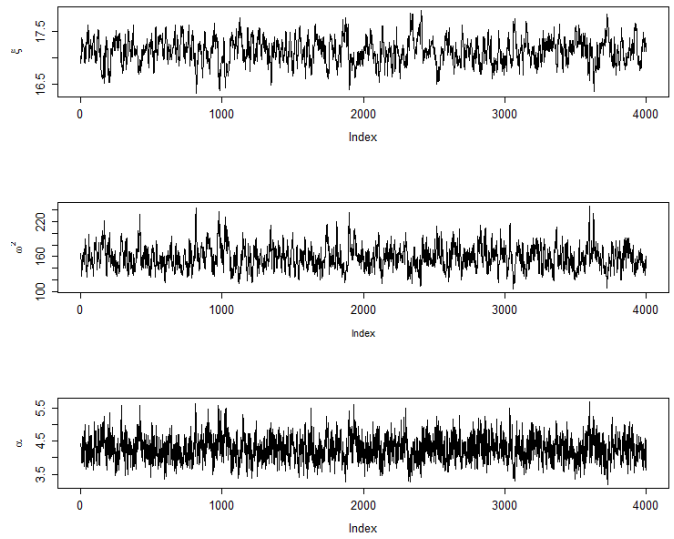

Figura 6.1: Valori stimati dei parametri ξ, ω^2 e α dell'anno 1987.

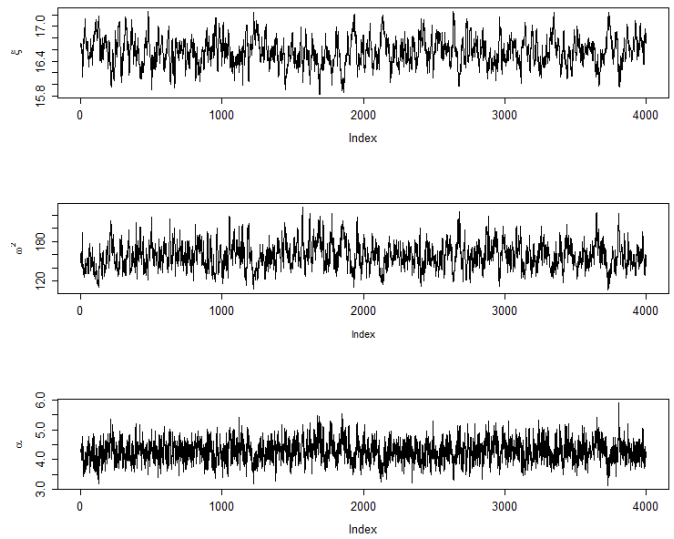

Figura 6.2: Valori stimati dei parametri ξ, ω^2 e α dell'anno 2004.

Da questi grafici si osserva che le stime non sembrano avere andamenti particolari con il passare delle iterazioni, quindi l'algoritmo produce delle stime valide.

Nelle Figure 6.3 e 6.4 sono rappresentate le curve di fecondità stimate insieme ai tassi di fecondità per età empirici. Le curve non seguono perfettamente l'andamento dei dati, ma sembrano comunque addattarsi abbastanza bene. Tuttavia, anche con questo modello, vi sono alcuni casi in cui le curve di fecondità non hanno una forma simile a quella della distribuzione dei tassi di fecondità per età. Per degli esempi si vedano le Appendici C e D, dove sono presenti i grafici di tutte le curve di fecondità ottenute con il nuovo modello.

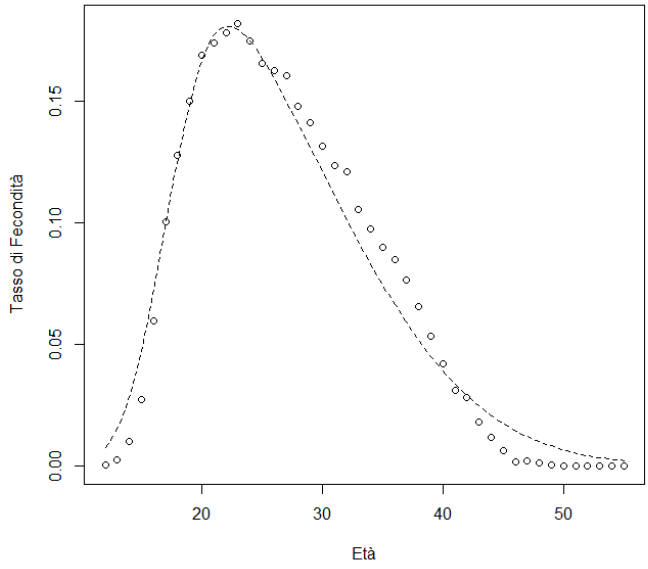

Figura 6.3: Curva di fecondità stimata per l'anno 1987.

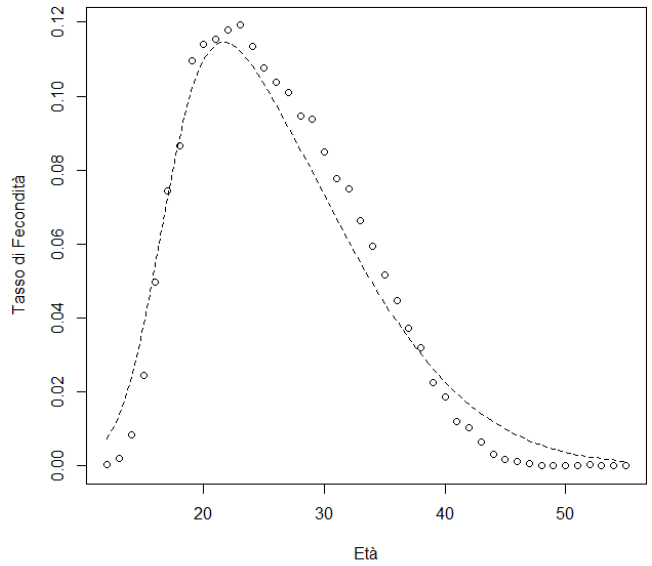

Figura 6.4: Curva di fecondità stimata per l'anno 2004.

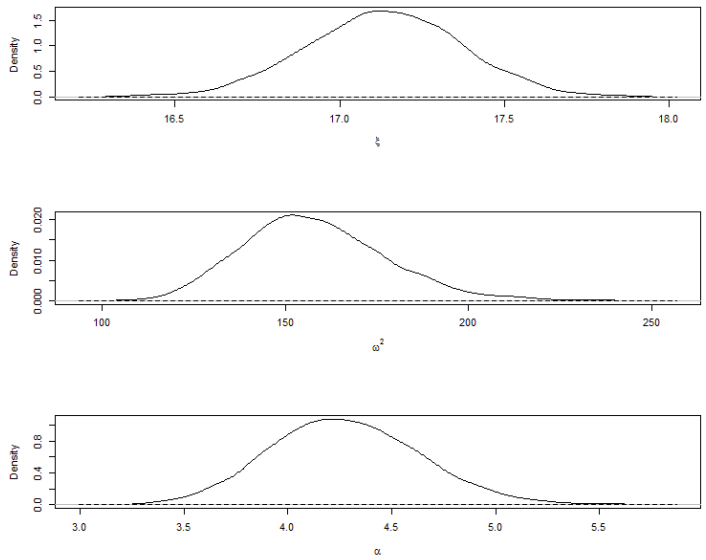

Figura 6.5: Densità a posteriori (linea continua) e a priori (linea tratteggiata) dei parametri ξ, ω^2 e α dell'anno 1987.

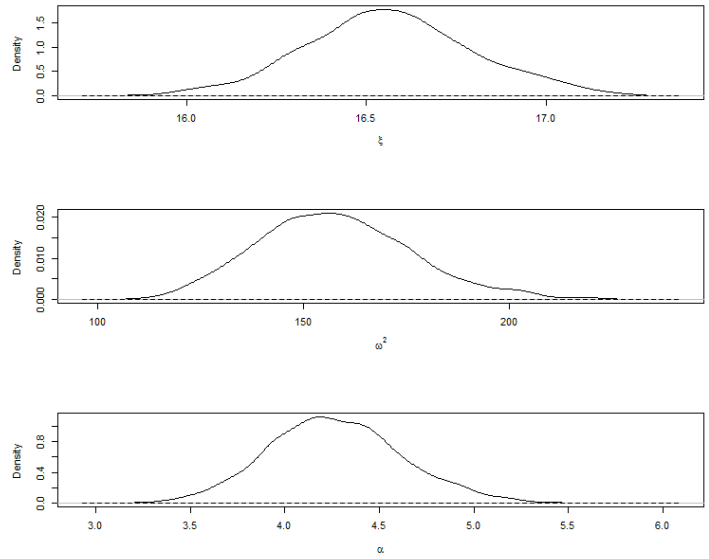

Figura 6.6: Densità a posteriori (linea continua) e a priori (linea tratteggiata) dei parametri ξ, ω^2 e α dell'anno 2004.

Nelle Figure 6.5 e 6.6 si possono vedere le distribuzioni a priori e a posteriori dei parametri. Come nel Paragrafo 4.3 la distribuzione a priori risulta molto piatta, mentre quella a posteriori molto più concentrata.

Nelle Tabelle 6.1 e 6.2 sono presenti le stime dei parametri e la relativa deviazione standard per ogni anno. Con il passare degli anni le stime di ξ e α non cambiano molto, mentre ω^2 varia di più con un minimo di 148.1737 nell'anno 1975 ed un massimo di 184.0182 nell'anno 1994. Anche in questo caso, la deviazione standard è decisamente bassa per i parametri ξ e α, mentre per ω^2 risulta più grande.

Anno	ξ		ω^2		α	
	stima	dev. std.	stima	dev. std.	stima	dev. std.
1972	17.1369	0.24949	157.1073	19.88499	4.1718	0.36266
1973	17.1697	0.24421	157.5011	18.97583	4.1515	0.34744
1974	16.7908	0.25403	155.9983	20.87516	4.3230	0.37760
1975	16.6402	0.23531	148.1737	18.90241	4.2823	0.36373
1976	16.6737	0.24987	157.2417	20.24378	4.3628	0.37974
1977	16.3969	0.24210	150.2170	20.11523	4.5102	0.38357
1978	16.7693	0.24778	157.8886	20.02364	4.2710	0.36932
1979	16.5650	0.28029	160.6856	21.36136	4.3725	0.38791
1980	16.8731	0.24376	160.5241	20.70756	4.3615	0.37563
1981	16.6106	0.23667	158.1546	19.05331	4.1872	0.35260
1982	17.2245	0.26275	163.9919	22.15966	4.3427	0.39059
1983	16.7612	0.25514	162.8037	22.46169	4.5158	0.39482
1984	16.9765	0.23309	153.2553	19.48250	4.4417	0.37845
1985	16.9639	0.26798	168.1072	22.40598	4.2355	0.37242
1986	17.0514	0.28962	179.0053	24.27546	4.3308	0.39298
1987	17.1345	0.23692	158.1544	19.73565	4.2798	0.36442
1988	17.1327	0.25854	161.5544	21.25682	4.3083	0.38517
1989	16.8520	0.28542	172.2167	24.38318	4.2567	0.40074
1990	17.1495	0.27758	174.4970	23.16338	4.2676	0.38634
1991	17.0250	0.27772	173.3359	22.84665	4.2474	0.37832

Tabella 6.1: Stime posteriori e deviazione standard dei parametri ξ, ω^2 e α, 1972-1991.

Anno	ξ		ω^2		α	
	stima	dev. std.	stima	dev. std.	stima	dev. std.
1992	16.8801	0.26582	167.3256	21.07130	4.1064	0.36622
1993	16.9548	0.27964	180.1791	23.38533	4.1066	0.37086
1994	16.7952	0.30241	184.0182	25.61624	4.1486	0.39700
1995	16.7721	0.28977	175.8214	22.90717	4.0875	0.36937
1996	16.9488	0.26530	170.3574	22.16978	4.1787	0.37148
1997	16.8528	0.27582	173.9971	22.22728	4.1498	0.36100
1998	16.6218	0.26659	167.3787	21.46665	4.2074	0.36950
1999	16.5262	0.26573	170.7966	20.63461	4.0781	0.35273
2000	16.2821	0.25252	163.4126	20.25318	4.2013	0.36376
2001	16.7112	0.27447	166.5154	21.88263	4.1076	0.37309
2002	16.8637	0.26822	164.3248	21.10235	4.1838	0.36775
2003	16.6318	0.24214	167.3934	20.50123	4.1715	0.35645
2004	16.5642	0.23728	157.8245	19.00888	4.2691	0.35812
2005	16.7094	0.26464	169.4063	21.58109	4.2518	0.36990
2006	16.5885	0.24674	153.8975	19.60008	4.3002	0.37052
2007	16.6207	0.25688	163.1818	20.87449	4.3003	0.38170
2008	16.0992	0.26150	161.7257	21.35989	4.2690	0.38885
2009	16.4987	0.29636	173.1447	23.27262	4.2964	0.38390
2010	16.4622	0.26636	174.8582	23.11179	4.3323	0.38557
2011	16.5429	0.26629	169.8497	22.23201	4.2922	0.37681
2012	16.5971	0.26154	167.0904	22.06257	4.3136	0.38710

Tabella 6.2: Stime a posteriori e deviazione standard dei parametri ξ, ω^2 e α, 1992-2012.

Con le stime ottenute dall'algoritmo si possono ricavare le medie, le varianze e gli indici di asimmetria a posteriori, le cui medie per ogni anno vengono rappresentate nelle successive figure insieme alle bande di credibilità con livello di confidenza del 95%. La media a posteriori, rappresentata nella Figura 6.7, decresce leggermente fino al 1977, poi da qui cresce fino a stabilizzarsi intorno al 1986 e dal 1997 riporta un andamento decrescente, con un piccolo aumento negli ultimi 4 anni. Invece la varianza a posteriori (Figura 6.8) riporta un andamento crescente fino al 1994 per poi decrescere fino al 2006 ed avere un piccolo inalzamento negli ultimi 6 anni. Infine l'andamento dell'indice di asimmetria a posteriori (Figura 6.9) rimane più o meno costante dal 1972 al 1991, poi diminuisce leggermente l'anno successivo e dal 2001 ricomincia a crescere fino al 2012.

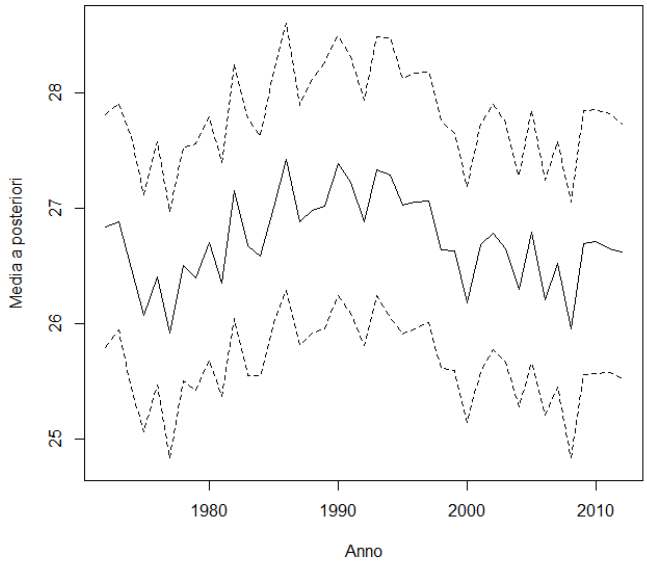

Figura 6.7: Medie a posteriori con bande di credibilità al 95%.

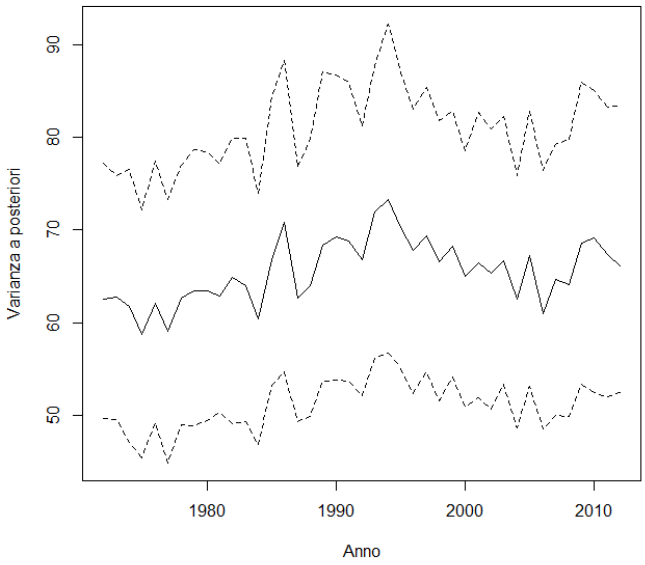

Figura 6.8: Varianze a posteriori con bande di credibilità al 95%.

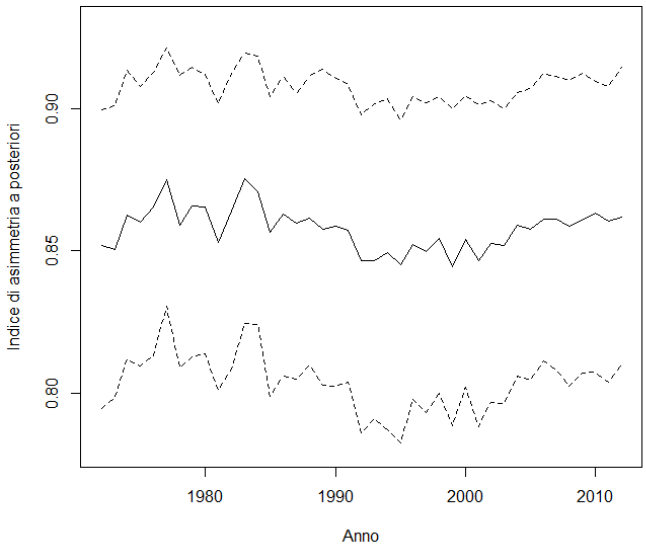

Figura 6.9: Indici di asimmetria a posteriori con bande di credibilità al 95%.

Infine sono rappresentate in Figura 6.10 le curve di fecondità stimate dal 1972 al 2012 ad intervalli di 4 anni. Anche qui, come nel Capitolo 4, le curve sono molto simili. Il tasso di fecondità assume un valore sopra lo zero intorno ai 13 anni, con un picco tra i 20 e i 26 anni e ritorna verso lo zero dopo i 50 anni.

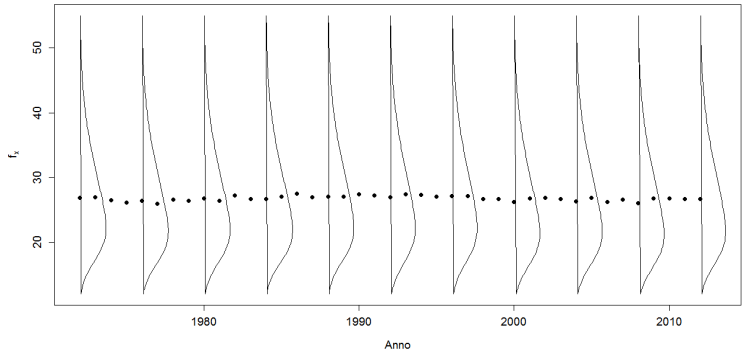

Figura 6.10: Curve di fecondità del Costa Rica, 1972-2012. I punti indicano la media a posteriori per ogni anno.

Capitolo 7

Conclusioni

In questa tesi è stato presentato uno studio sui tassi di fecondità per età del Costa Rica dal 1972 al 2012. Vista la forma della distribuzione di questi tassi, si è pensato che la distribuzione normale asimmetrica fosse una buona scelta per la stima delle curve di fecondità. In particolare, sono state intraprese due vie per raggiungere questo obiettivo. Nella prima si assume che le osservazioni che entrano nella verosimiglianza sono l'età della madre al momento della nascita del figlio, che si distribuiscono come una SN. In questo modo, stimando i parametri da questa distribuzione, si ottengono i parametri per la distribuzione degli ASFR. Nella seconda i dati che osservo non sono l'età della madre alla nascita del figlio, ma il fatto che ci sia stata una nascita in una determinata fascia d'età e si distribuiscono come una multinomiale $MN(m, p_1, ..., p_j)$, con $m = 1$ e $p_j = \mathbb{P}\{y_i = j\}$, dove quest'ultimo è uguale all'integrale da $j-1/2$ a $j+1/2$ della funzione di densità di una SN, dove $j = 12, ..., 55$. Quindi vengono introdotte n variabili latenti continue $v_1, ..., v_n$, dove n è il numero di osservazioni, che si distribuiscono come delle SN, valide solo nell'intervallo da $j - 1/2$ a $j + 1/2$ condizionatamente al fatto che y_i sia uguale a j. In questo modo si ottiene una coniugazione con la distribuzione SN che ci permette di stimare i parametri per le curve di fecondità. In tutti e due i casi si è utilizzato un approccio bayesiano perché avendo a disposizione i dati per diversi anni si ha la possibilità di ricavare delle informazioni a

priori dall'anno $t-1$ per l'anno t.

Dai Capitoli 4 e 6 si evince che con tutti e due i metodi si ottengono delle stime valide, con una variabilità contenuta per quanto riguarda i parametri ξ e α, mentre per il parametro ω^2 è decisamente più grande. In tutti e due i casi, le curve di fecondità stimate non seguono perfettamente l'andamento dei dati ma, di solito, non vi si discostano di molto. Tuttavia ci sono degli anni in cui le curve non si adattano molto bene alla distribuzione degli ASFR.

I due metodi non presentano dei risultati molto diversi l'uno dall'altro, anzi sono sostanzialmente uguali. Infatti confrontando le stime presenti nelle Tabelle 4.1 e 4.2 con quelle presenti nelle Tabelle 6.1 e 6.2 si può notare che sono pressoché identiche per gli stessi anni. È possibile, anche, osservare questo fenomeno confrontando i grafici delle curve di fecondità stimate con i due metodi nello stesso anno. Infatti nelle Figure 7.1 e 7.2, che riguardano gli anni presi ad esempio nei Capitoli 4 e 6, si vede che le curve hanno lo stesso andamento.

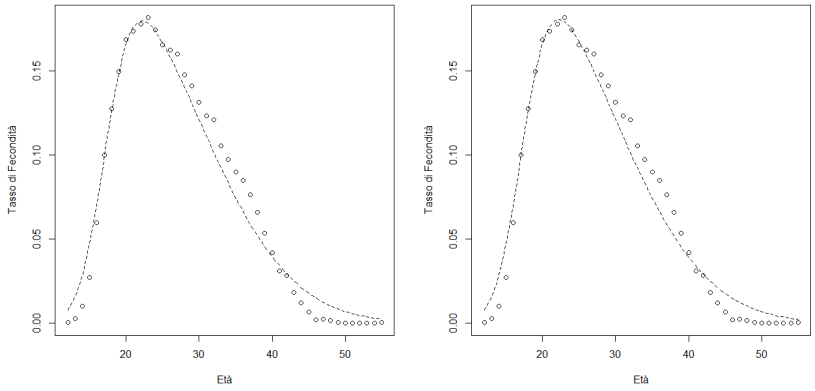

Figura 7.1: Curva di fecondità del 1987 stimata con il primo modello (sinistra) e il secondo modello (destra).

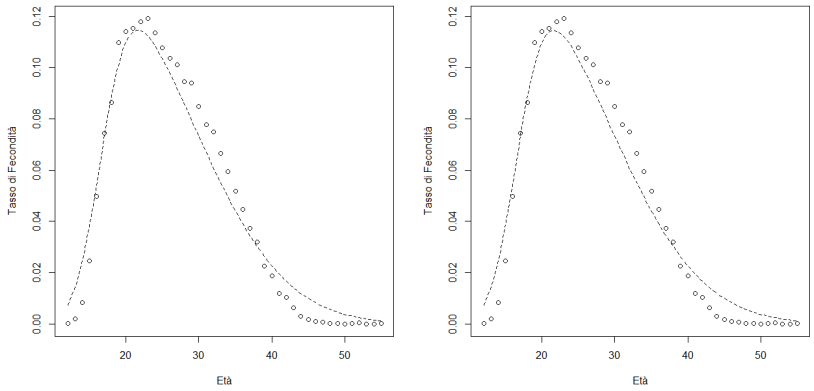

Figura 7.2: Curva di fecondità del 1987 stimata con il primo modello (sinistra) e il secondo modello (destra).

Un problema che affligge tutti e due i metodi riguarda gli algoritmi basati sul Gibbs sampling, che sono computazionalmente pesanti e di conseguenza i calcoli sono lenti, anche se implementati utilizzando algoritmi paralleli. Per migliorare questo aspetto ed accorciare i tempi d'attesa, si può pensare di implementare le funzioni usate per le simulazioni nel linguaggio di programmazione C piuttosto che in R, il quale generalmente è poco prestante sui cicli.

In conclusione il secondo modello sarebbe preferibile rispetto a quello proposto da Canale e Scarpa (2013) perché viene presa in considerazione la vera distribuzione delle osservazioni. Tuttavia, in questo caso specifico, le stime che si ottengono dai due metodi sono sostanzialmente uguali, quindi può risultare preferibile usare il primo modello essendo più semplice rispetto al secondo.

Bibliografia

[1] Albert, J. H., Chib, S., 1993. Bayesian Analysis of Binary and Polychotomous Response Data. Journal of the American Statistical Association, 88, 442: 669-679.

[2] Arellano-Valle, R. B., Azzalini, A., 2006. On the unification of families of skew-normal distributions. Scandinavian Journal of Statistics, 33, 561-574.

[3] Arellano-Valle, R. B., Genton, M. G., Loschi, R.H., 2009. Shape mixture of multivariate skew-normal distributions. Journal of Multivariate Analysis, 100, 91-101.

[4] Azzalini, A., 1985. A class of distributions which includes the normal ones. Scandinavian Journal of Statistics, 12, 171-178.

[5] Azzalini, A., 1986. Further results on a class of distributions which includes the normal ones. Statistica, 46, 199-208.

[6] Azzalini, A., 2005. The Skew-normal Distribution and Related Multivariate Families. Scandinavian Journal of Statistics, 32, 159-188.

[7] Azzalini, A., Capitanio, A., 2014. The Skew-Normal and Related Families. Institute of Mathematical Statistics Monographs, Cambridge University Press, Cambridge.

[8] Bayes, C., Branco, M., 2007. Bayesian inference for the skewness parameter of the scalar skew-normal distribution. Brazilian Journal of Probability and Statistics, 21, 141-163.

[9] Canale, A., Scarpa, B., 2013. Informative Bayesian inference for the skew-normal distribution.

[10] Chandola, T., Coleman, D. A., Horns, R. W., 1999. Recent European fertility patterns: fitting curves to "distorted" distributions. Population Studies, 53, 3: 317-329.

[11] Chiogna, M., 1998. Some results on the scalar skew-normal distribution. Journal of the Italian Statistical Society, 7, 1-13.

[12] Ellison, B. E., 1964. Two theorems for inferences about the normal distribution with applications in acceptance sampling. Journal of the American Statistical Association, 59, 89-95.

[13] Gancho, V. G., Dey, D., Lachos, V. H., Andrade, M. G., 2011. Bayesian nonlinear regression models with scale mixtures of skew-normal distributions: Estimation and case influence diagnostics. Computational Statistics and Data Analysis, 55, 588-602.

[14] Gilje, E., 1969. Fitting curves to age-specific fertility rates: some examples. Statistical Review of the Swedish National Central Bureau of Statistics III, 7:118-134.

[15] Golub, G. H., Van Loan, C. F., 1989. Matrix Computations, 2nd Edition. Johns Hopkins University Press, Baltimore.

[16] Hadwiger, H., 1940. Eine analytische reprodutions-funktion fur biologische Gesamtheiten. Skandinavisk Aktuarietidskrift, 23, 101-113.

[17] Hoem, J. M., Madsen, D., Nielsen, J. L., Ohlsen, E., Hansen, H. O., Rennermalm, B., 1981. Experiments in modelling recent Danish fertility curves. Demography, 18: 231-244

[18] Mazzucco, S., Scarpa, B., 2011. Fitting age-specific fertility rates by a skew-symmetric probability density function. Working Paper Series, 10, Department of Statistical Sciences, University of Padua.

[19] Mazzuco, S. and Scarpa, B., 2014. Fitting age-specific fertility rates by a flexible generalized skew normal probability density function. Journal of the Royal Statistical Society: Series A (Statistics in Society).

[20] Neal, P., Kypraios, T., 2013. Exact Bayesian inference via data augmentation. Statistics and Computing, Springer US.

[21] Peristera, P., Kostaki, A., 2007. Modeling fertility in modern populations. Demographic Research, 16, 6: 141-194.

[22] Schmertmann, C. P., 2003. A system of model fertility schedules with graphically intuitive parameters. Demographic Research, 9, 5: 82-110.

[23] Zacks, S., 1981. Parametric Statistical Inference. Oxford: Pergamon Press.

[24] Measure Evaluation PRH: Age specific fertility rates. http://www.cpc.unc.edu/measure/prh/rh_indicators/specific/fertility/age-specific-fertility-rates

[25] United Nations: Age-specific fertility rate. http://www.un.org/esa/population/publications/WFR2009_Web/Data/Meta_Data/ASFR.pdf

[26] Age-Specific Fertility Rates and the Total Fertility Rate. https://www.k4health.org/sites/default/files/Total%20Fertility%20Rate%20and%20Age-%Specific%20Fertility%20Rate.pdf

Appendice A

Curve di fecondità stimate con il primo modello (1972-1993)

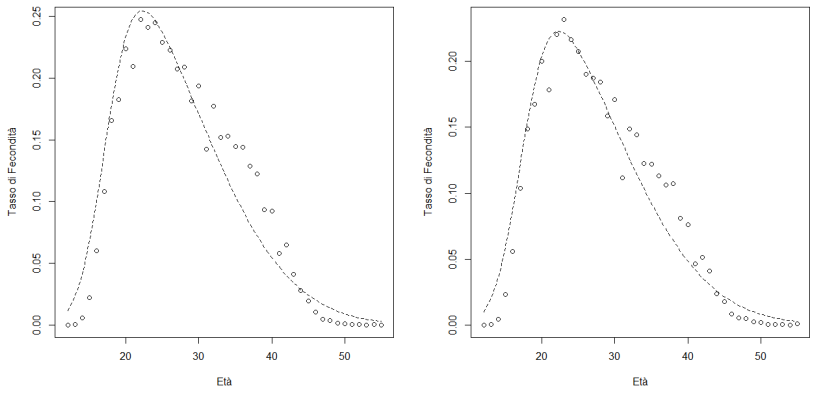

Figura A.1: Curva di fecondità del 1972 (sinistra) e del 1973 (destra).

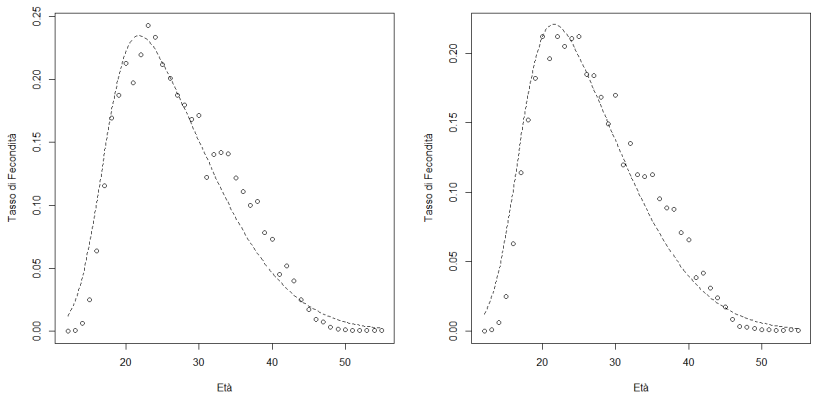

Figura A.2: Curva di fecondità del 1974 (sinistra) e del 1975 (destra).

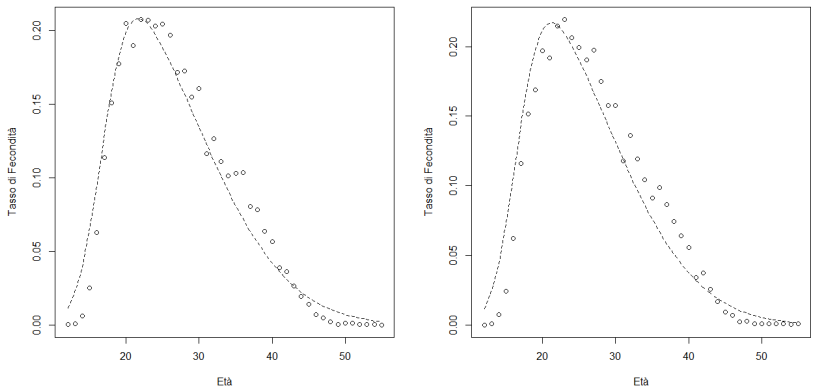

Figura A.3: Curva di fecondità del 1976 (sinistra) e del 1977 (destra).

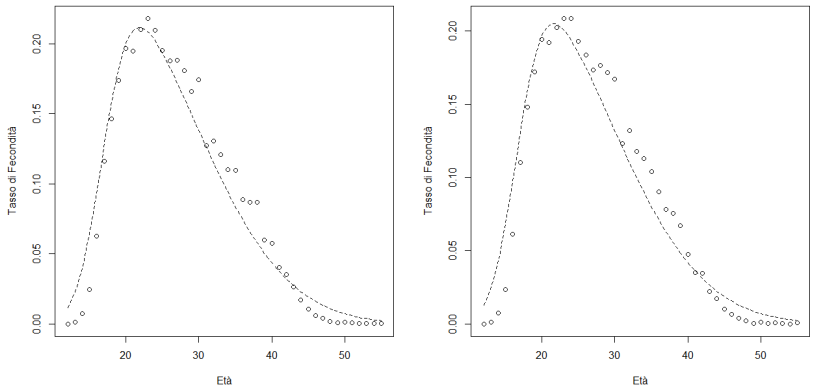

Figura A.4: Curva di fecondità del 1978 (sinistra) e del 1979 (destra).

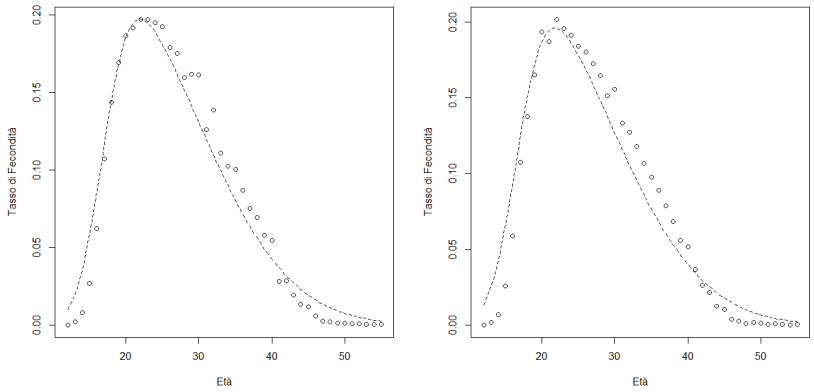

Figura A.5: Curva di fecondità del 1980 (sinistra) e del 1981 (destra).

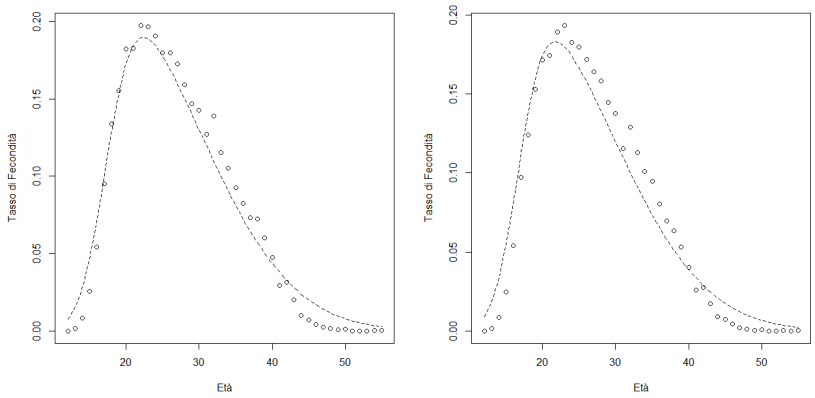

Figura A.6: Curva di fecondità del 1982 (sinistra) e del 1983 (destra).

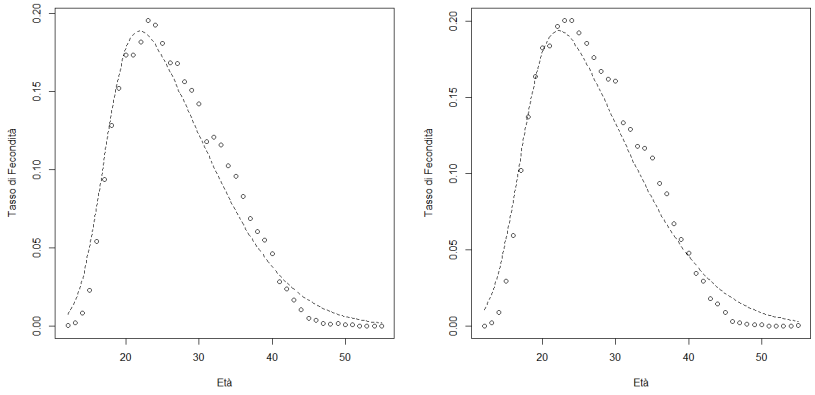

Figura A.7: Curva di fecondità del 1984 (sinistra) e del 1985 (destra).

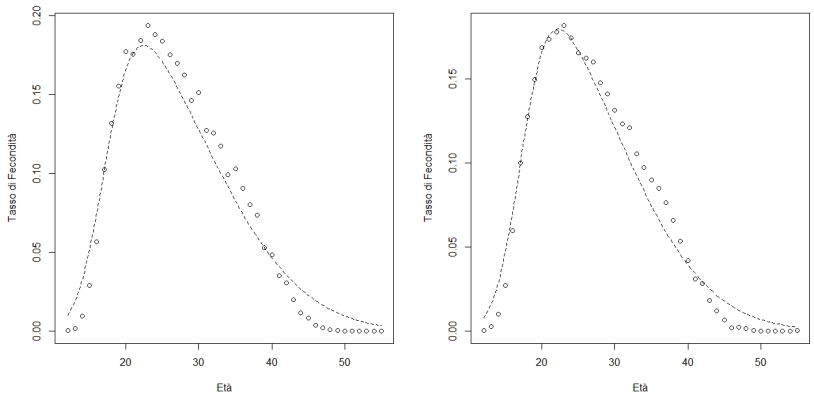

Figura A.8: Curva di fecondità del 1986 (sinistra) e del 1987 (destra).

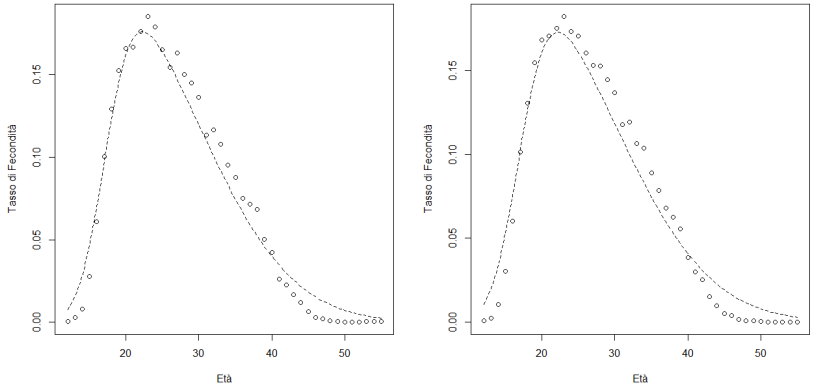

Figura A.9: Curva di fecondità del 1988 (sinistra) e del 1989 (destra).

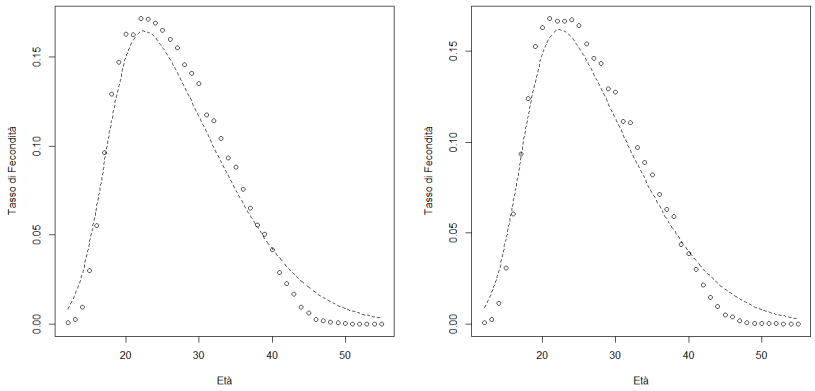

Figura A.10: Curva di fecondità del 1990 (sinistra) e del 1991 (destra).

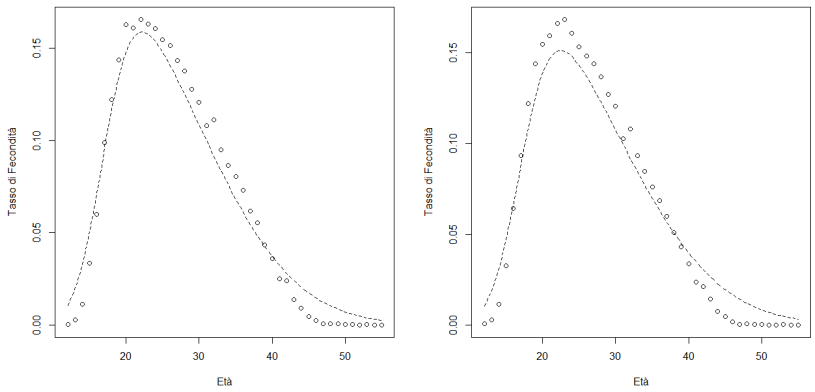

Figura A.11: Curva di fecondità del 1992 (sinistra) e del 1993 (destra).

Appendice B

Curve di fecondità stimate con il primo modello (1994-2012)

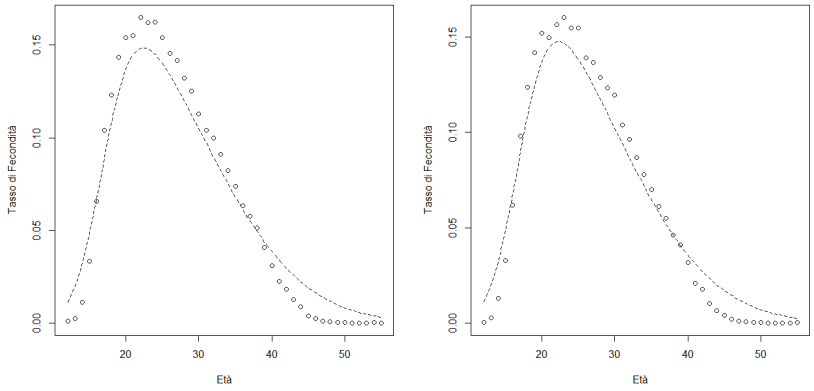

Figura B.1: Curva di fecondità del 1994 (sinistra) e del 1995 (destra).

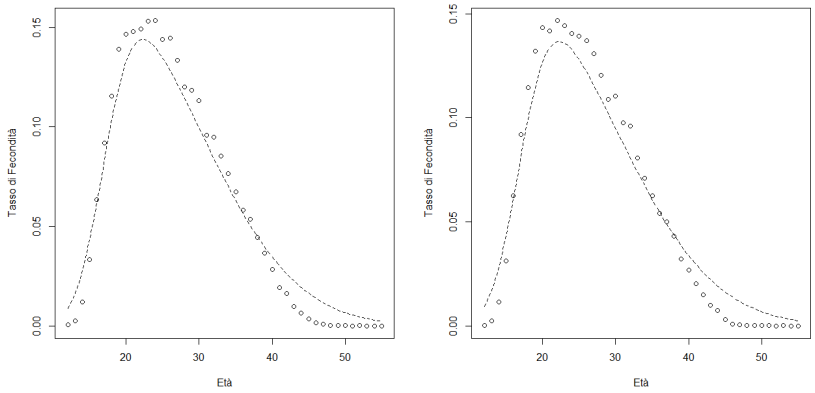

Figura B.2: Curva di fecondità del 1996 (sinistra) e del 1997 (destra).

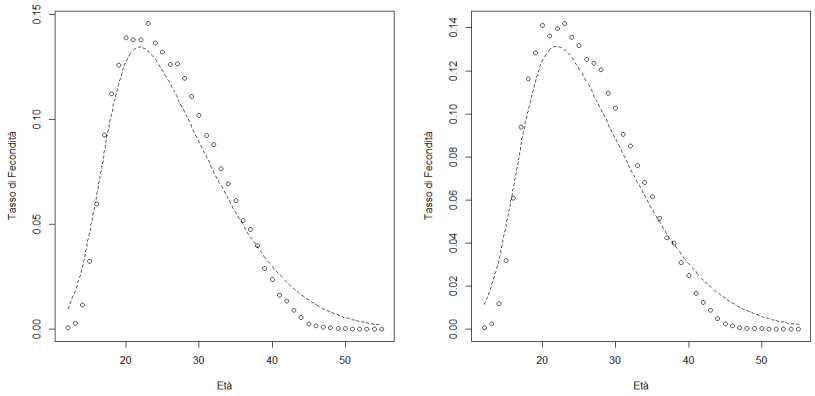

Figura B.3: Curva di fecondità del 1998 (sinistra) e del 1999 (destra).

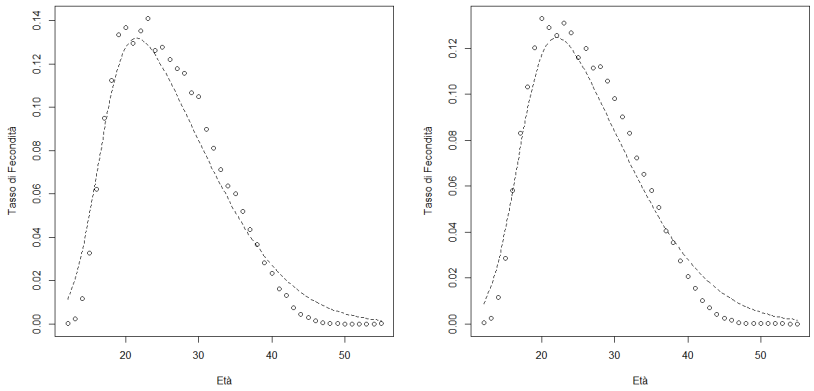

Figura B.4: Curva di fecondità del 2000 (sinistra) e del 2001 (destra).

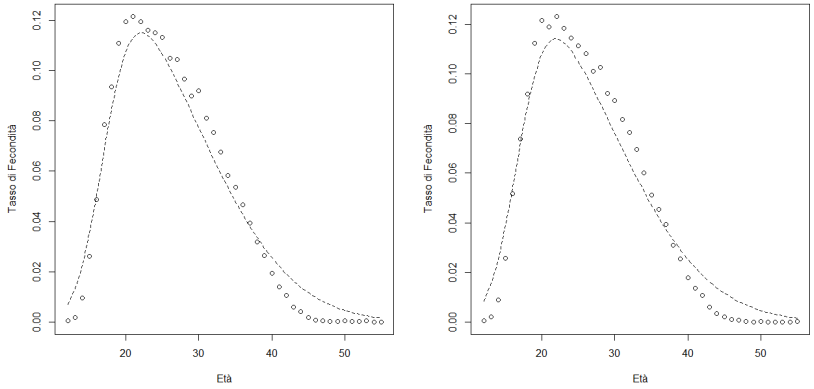

Figura B.5: Curva di fecondità del 2002 (sinistra) e del 2003 (destra).

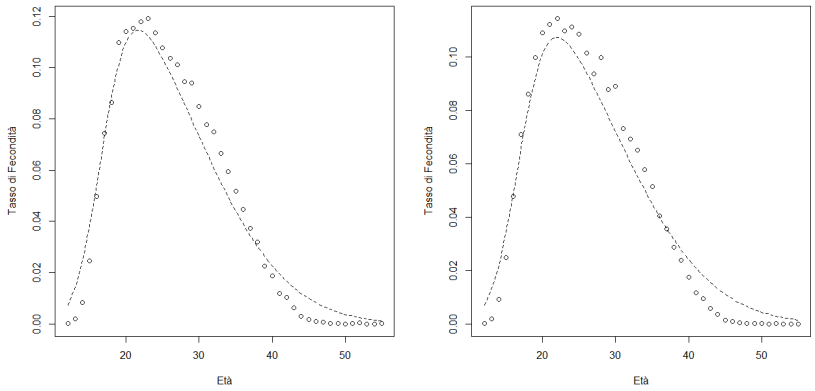

Figura B.6: Curva di fecondità del 2004 (sinistra) e del 2005 (destra).

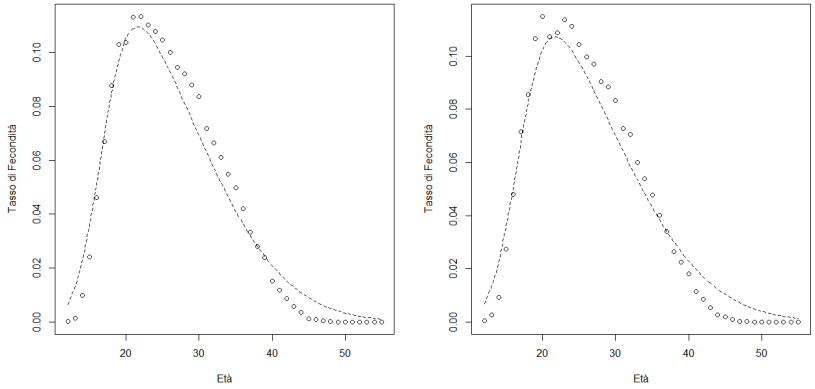

Figura B.7: Curva di fecondità del 2006 (sinistra) e del 2007 (destra).

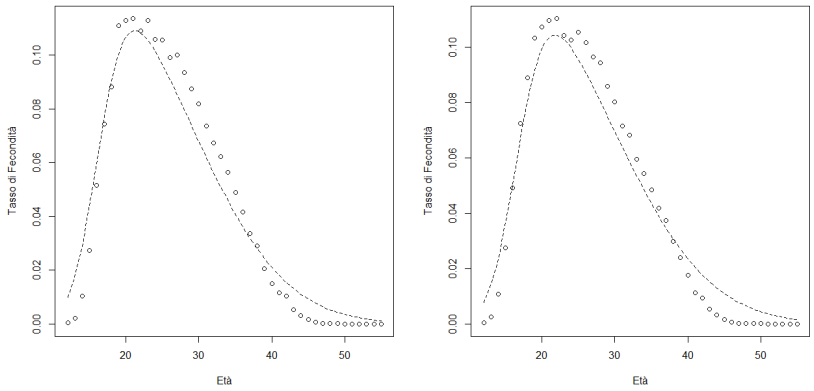

Figura B.8: Curva di fecondità del 2008 (sinistra) e del 2009 (destra).

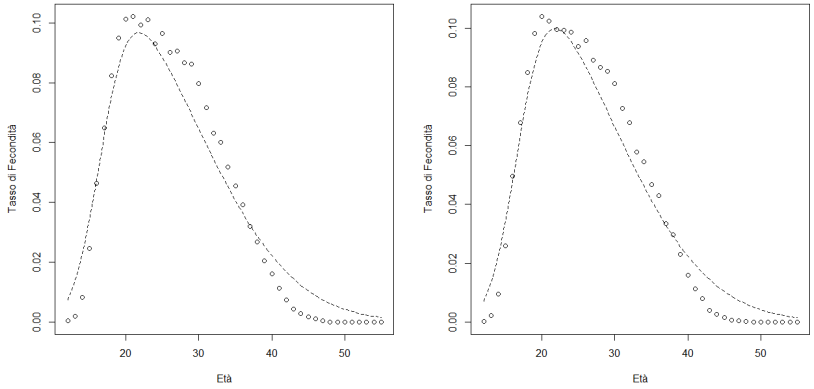

Figura B.9: Curva di fecondità del 2010 (sinistra) e del 2011 (destra).

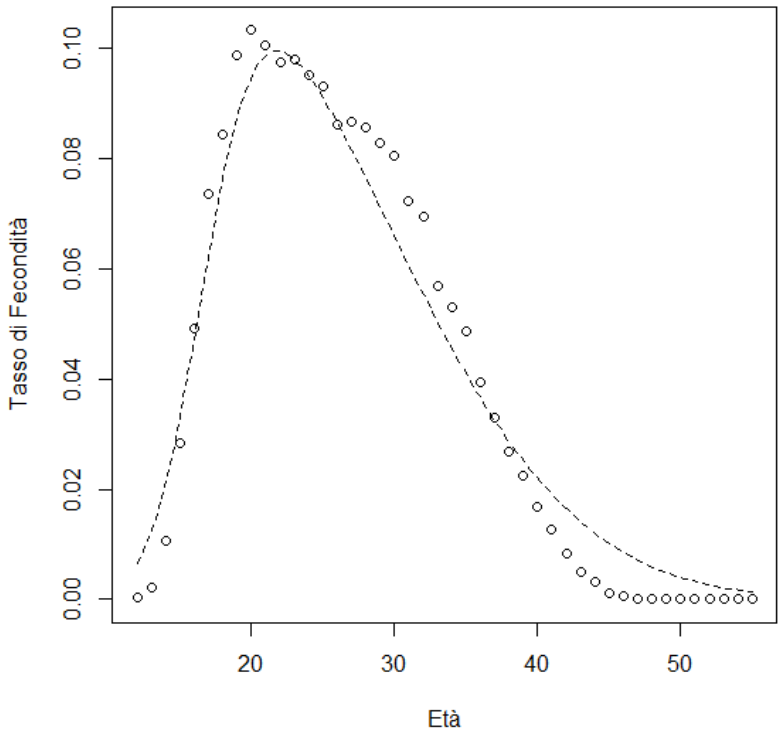

Figura B.10: Curva di fecondità del 2012.

Appendice C

Curve di fecondità stimate con il secondo modello (1972-1993)

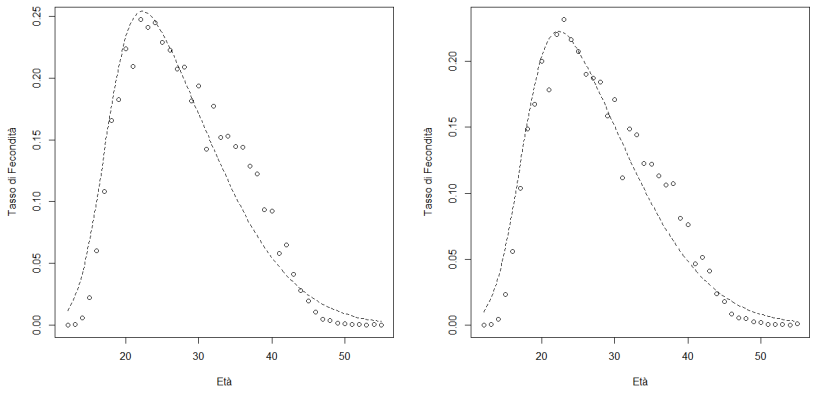

Figura C.1: Curva di fecondità del 1972 (sinistra) e del 1973 (destra).

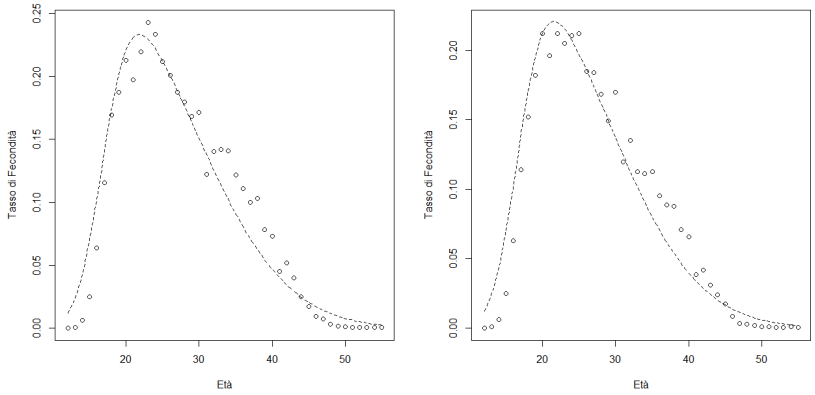

Figura C.2: Curva di fecondità del 1974 (sinistra) e del 1975 (destra).

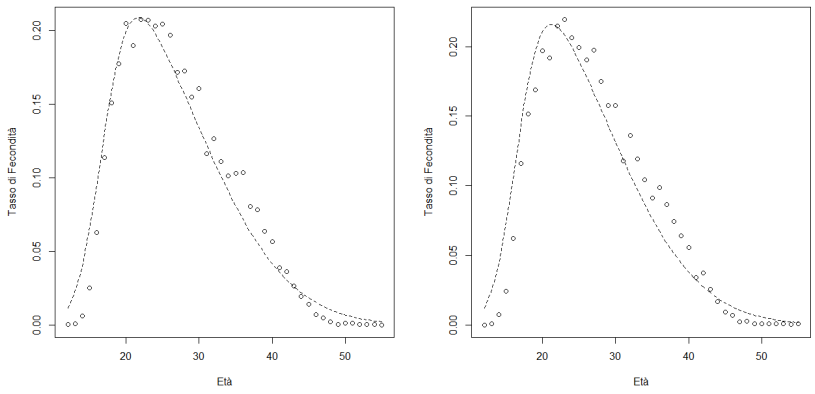

Figura C.3: Curva di fecondità del 1976 (sinistra) e del 1977 (destra).

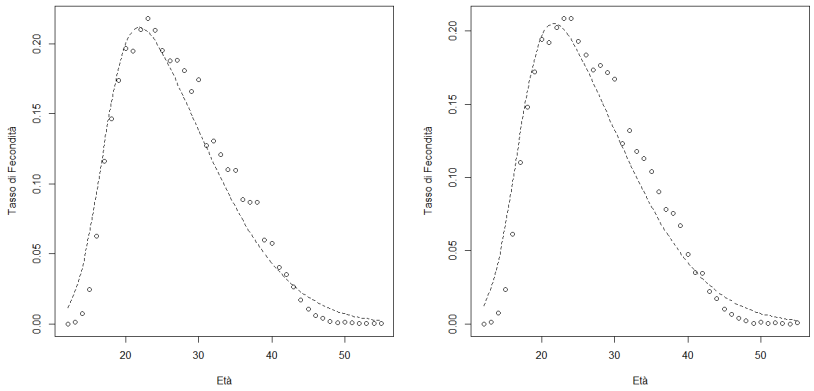

Figura C.4: Curva di fecondità del 1978 (sinistra) e del 1979 (destra).

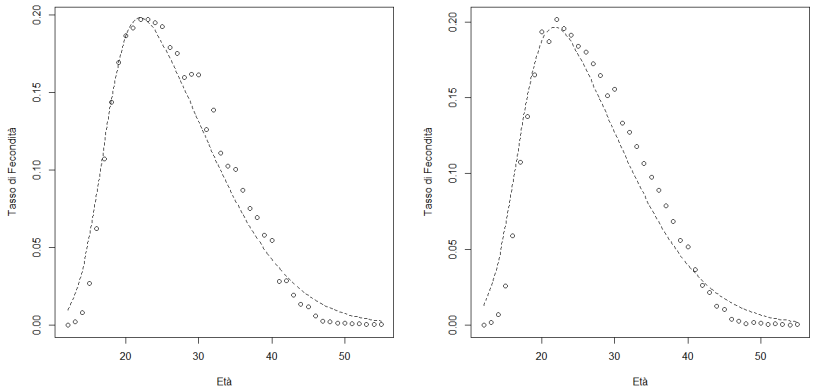

Figura C.5: Curva di fecondità del 1980 (sinistra) e del 1981 (destra).

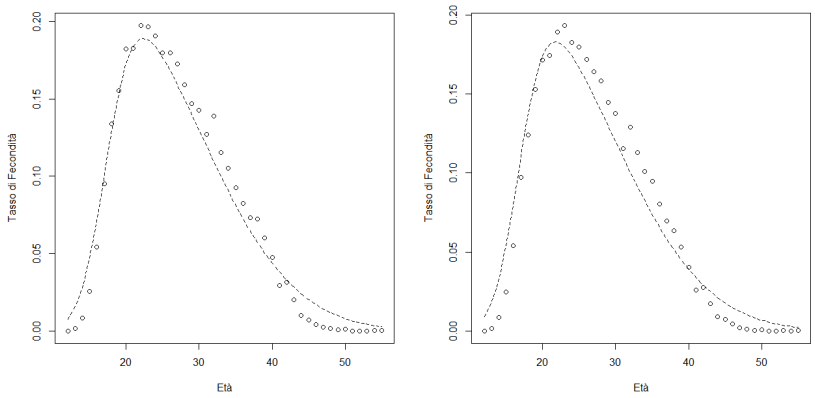

Figura C.6: Curva di fecondità del 1982 (sinistra) e del 1983 (destra).

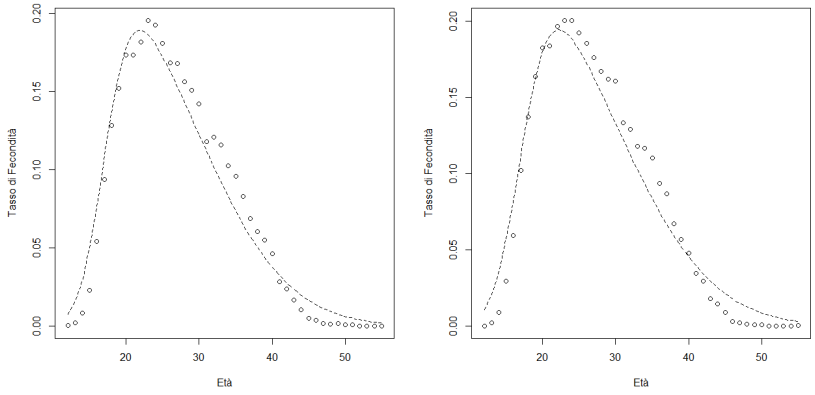

Figura C.7: Curva di fecondità del 1984 (sinistra) e del 1985 (destra).

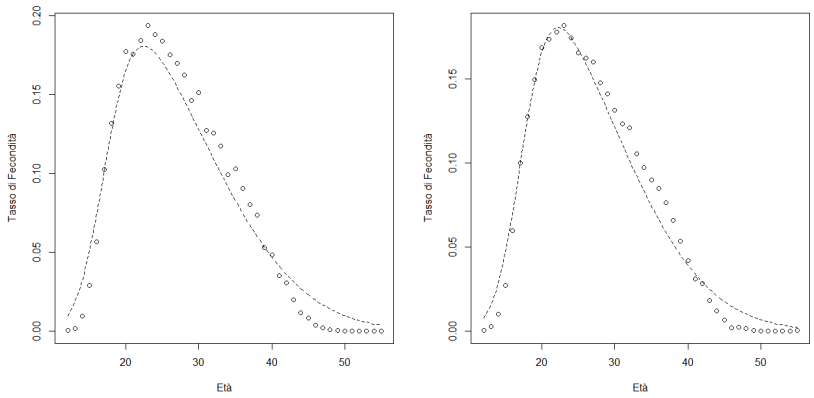

Figura C.8: Curva di fecondità del 1986 (sinistra) e del 1987 (destra).

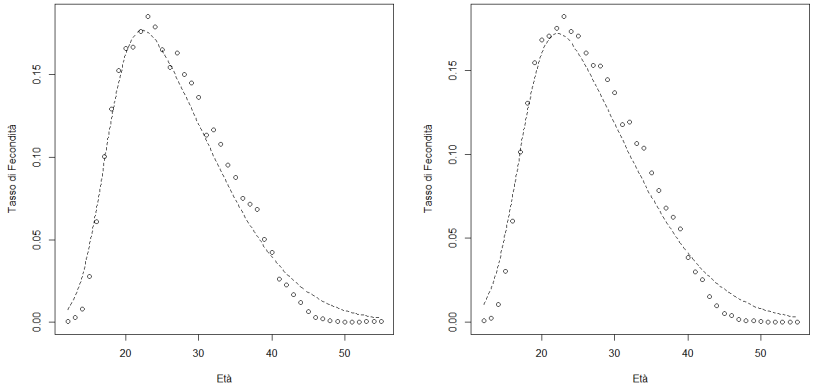

Figura C.9: Curva di fecondità del 1988 (sinistra) e del 1989 (destra).

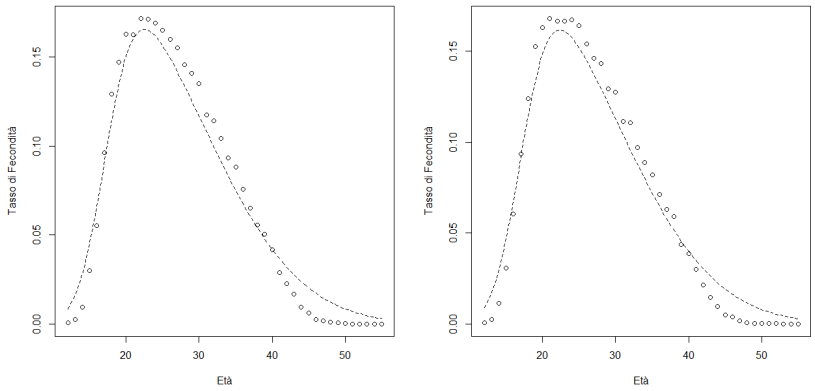

Figura C.10: Curva di fecondità del 1990 (sinistra) e del 1991 (destra).

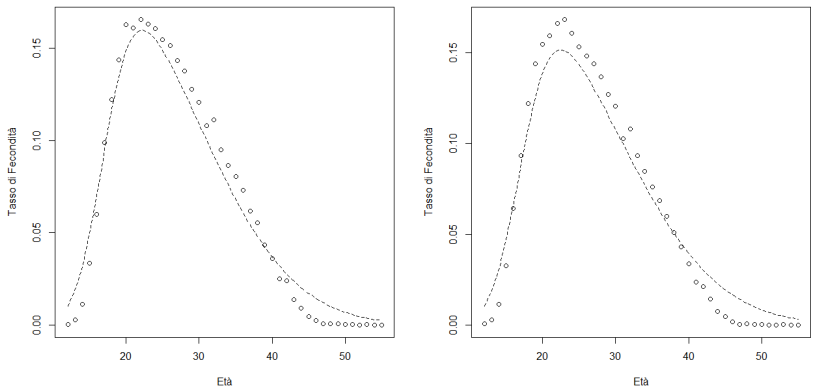

Figura C.11: Curva di fecondità del 1992 (sinistra) e del 1993 (destra).

Appendice D

Curve di fecondità stimate con il secondo modello (1994-2012)

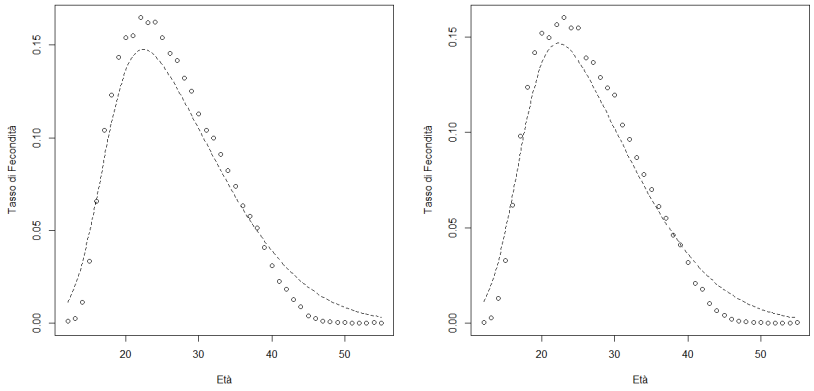

Figura D.1: Curva di fecondità del 1994 (sinistra) e del 1995 (destra).

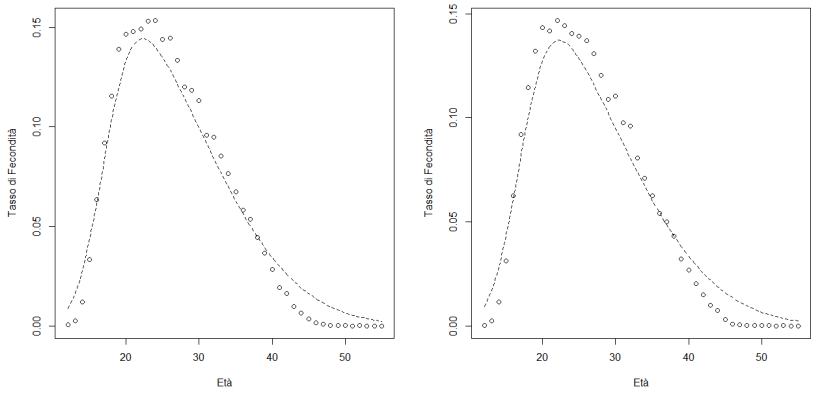

Figura D.2: Curva di fecondità del 1996 (sinistra) e del 1997 (destra).

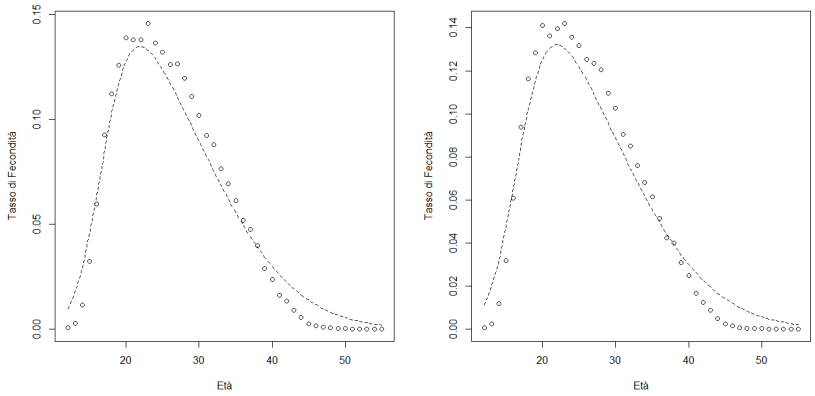

Figura D.3: Curva di fecondità del 1998 (sinistra) e del 1999 (destra).

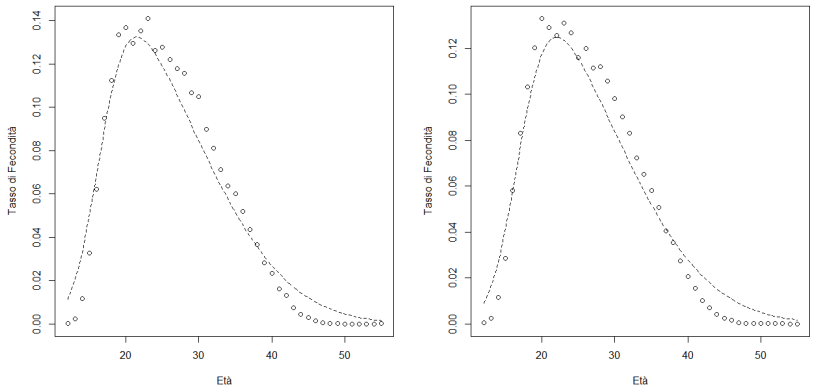

Figura D.4: Curva di fecondità del 2000 (sinistra) e del 2001 (destra).

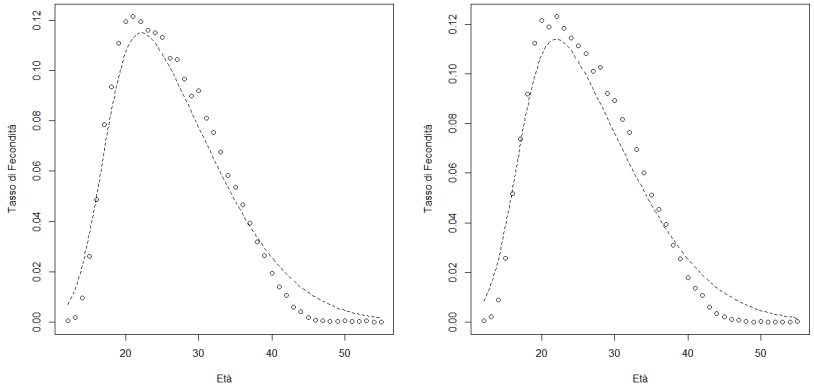

Figura D.5: Curva di fecondità del 2002 (sinistra) e del 2003 (destra).

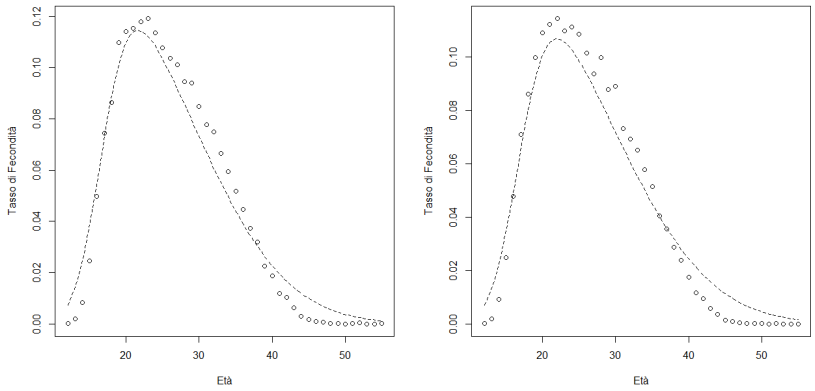

Figura D.6: Curva di fecondità del 2004 (sinistra) e del 2005 (destra).

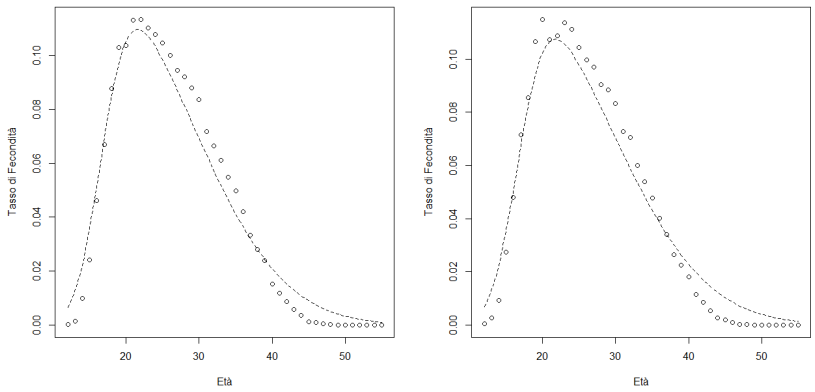

Figura D.7: Curva di fecondità del 2006 (sinistra) e del 2007 (destra).

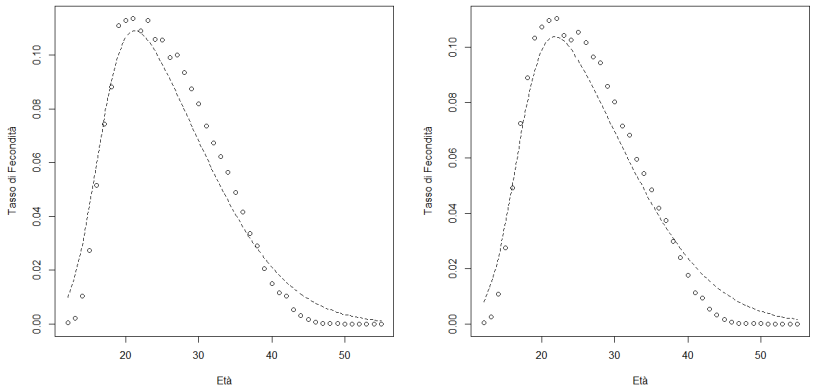

Figura D.8: Curva di fecondità del 2008 (sinistra) e del 2009 (destra).

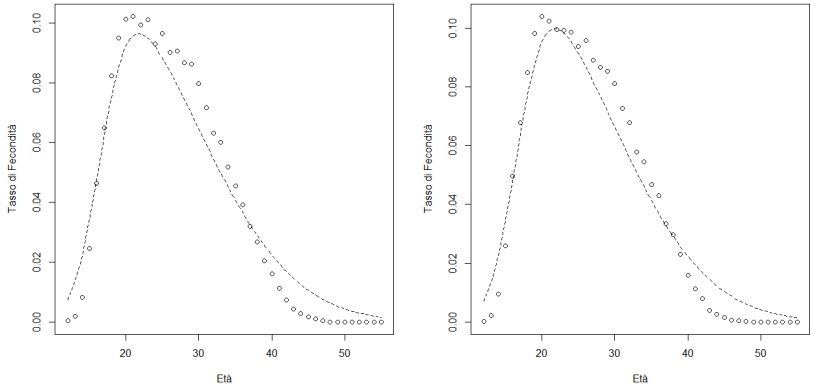

Figura D.9: Curva di fecondità del 2010 (sinistra) e del 2011 (destra).

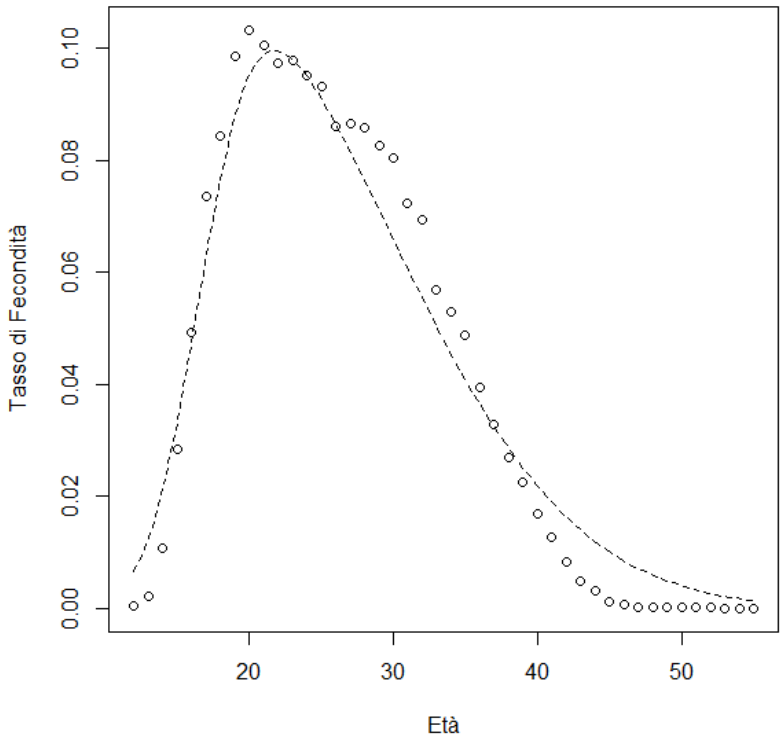

Figura D.10: Curva di fecondità del 2012.

Appendice E

Codice R dell'algoritmo del secondo modello

```
########################################
#funzione SN troncata
########################################

f <- function(x,xi,omega,alpha,j)
{
(dsn(x,xi,omega,alpha)/(psn(j+(1/2),xi,omega,alpha)-
psn(j-(1/2),xi,omega,alpha)))*I(x>=(j-(1/2)) & x<=(j+(1/2)))
}

########################################
# Accettazione-rifiuto con proposta uniforme
########################################

rf.AR1 <- function(n,xi,omega,alpha,j)
{
out<-c()
accepted<-c()
```

```
b<- -nlminb(10,function(x) -f(x,xi,omega,alpha,j)/
dunif(x,j-(1/2),j+(1/2)),lower=j-(1/2),upper=j+(1/2))$objective

while(length(out)<n)
{
xs<-runif(n,j-(1/2),j+(1/2))
u<-runif(n)
acc <- u <f(xs,xi,omega,alpha,j)/b/dunif(xs,j-(1/2),j+(1/2))
out<-c(out,xs[acc])
accepted<-c(accepted,acc)
}
list(values=out[1:n],acceptance=mean(accepted))
}

#######################################
#Algoritmo di Gibbs sampling
#######################################

gibbs <- function(N,y,alpha,xi,omega,a,b,xi0,k,alpha0,phi0,lambda0){



  n<- length(y)
  eta <- rep(NA,n)
  delta.cap<- matrix(0,nrow=n+1,ncol=1)


  V<-rep(0,n)

   A <- rep(NA,N)
B <- rep(NA,N)
C <- rep(NA,N)
```

```
for (i in 1:N){

for(j in 1:n){
J<-y[j]
V[j]<-rf.AR1(1,xi,sqrt(omega),alpha,J)$values
   }

delta <- alpha/sqrt((alpha^2) + 1)

for(r in 1:n){
eta[r] <- rtruncnorm(1,mean=delta*(V[r]-xi),sd=sqrt(omega*
(1-delta^2)),a=0)
}

mu.hat <- (k*sum(V-(delta*eta))+((1-delta^2)*xi0))/
((n*k)+(1-delta^2))

k.hat <- (k*(1-delta^2))/((n*k)+(1-delta^2))

b.hat <- (1/(2*(1-(delta^2))))*(((delta^2)*sum(eta^2))-
(2*delta*sum(eta*(W-xi)))+sum((W-xi)^2)+(((1-(delta^2))/k)*((xi-xi0)^2)))

xi <- rnorm(1,mu.hat,k.hat*omega)
    omega <- 1/rgamma(1,a+((n+1)/2),b+b.hat)


y.star <- (V-xi)/sqrt(omega)

z<-c(y.star,lambda0/phi0)


     delta.cap<- (phi0*z)/(sqrt((phi0^2*z^2)+1))
```

```
gamma <- c(delta.cap[1:n]*(alpha0/phi0),0)

alpha <- rsun(1, alpha0, gamma, phi0, delta.cap, algo = "gibbs")$sample

A[i] <- alpha
B[i] <- xi
C[i] <- omega


}


abc <- cbind(A,B,C)
write.table(abc,file="datiaug.txt")
list(alpha=A,xi=B,omega=C,W=W)
}
```

Tesi scritta utilizzando LaTeX.