Università degli studi di Milano–Bicocca
Scuola di Economia e Statistica

Corso di Laurea in
Scienze Statistiche ed Economiche

Criteri di selezione di un modello in ambito bayesiano

Relatore: Dott. Tommaso Rigon

Tesi di laurea di:
Giulia de Innocentiis
Matricola N. 865084

Anno Accademico 2022/2023

Alla mia nonna Rosa

Indice

Introduzione

Uno dei ruoli della statistica consiste nel derivare, a partire dalle osservazioni di un evento casuale, una stima sulla legge di probabilità che ha generato il fenomeno in questione. Le modalità e le assunzioni di base con cui svolgere determinato compito differenziano le diverse branche della statistica. In particolare si possono riconoscere due grandi scuole di pensiero differenti: la statistica frequentista e la statistica bayesiana. Nella statistica frequentista le informazioni che si utilizzano per fare inferenza sulla popolazione sono esclusivamente contenute nel campione di dati a disposizione. Nella statistica bayesiana invece vengono fatte delle assunzioni a priori riguardanti i parametri caratterizzanti il fenomeno in questione, precedentemente all'osservazione del campione; solo dopo quest'ultima, le ipotesi iniziali vengono aggiornate in base alle informazioni contenute nella determinazione osservata. I parametri vengono infatti considerati come variabili casuali con una propria distribuzione a priori, che aggiunge quindi una componente soggettiva nell'approccio bayesiano, non compresa nell'approccio frequentista.

In questa tesi viene fornita un'introduzione generale alla statistica bayesiana e ai metodi utilizzati nella fase di costruzione, inferenza e selezione di un modello di regressione in questo ambito.

La tesi è composta da quattro capitoli. Nel primo capitolo vengono presentate le fondamenta della statistica bayesiana, a partire dalla specificazione del teorema di Bayes e dal contesto di lavore base dell'analisi in questo ambito. Vengono quindi definiti i due elementi cardine di questo approccio: la distribuzione a priori e la distribuzione a posteriori dei parametri. Infine vengono illustrati i metodi e gli algoritmi principali utilizzati nella fase di inferenza bayesiana.

Nel secondo capitolo viene invece presentato il modello di regressione lineare multivariato in un'ottica bayesiana, le relative specificazioni delle distribuzioni a priori e a posteriori e l'utilizzo dell'algoritmo Gibbs Sampler specificatamente a questo modello.

Una volta definite le modalità di costruzione del modello, vengono illustrati alcuni degli strumenti necessari nella successiva fase di selezione del modello all'interno del terzo capitolo. In particolare viene esposto il concetto di capacità predittiva del modello e i criteri di informazione che ne derivano e, infine, il metodo della cross validation. In questo capitolo viene inoltre presentato un nuovo metodo di selezione proposto nel 2022 in Silva & Zanella (2022). Uno degli obiettivi di questa tesi è stato proprio lo studio di questo nuovo stimatore e delle sue proprietà e, soprattutto, la sua implementazione.

Nel quarto capitolo viene invece dimostrata la conoscenza acquisita in questo percorso tramite l'applicazione pratica di tutti i modelli e metodi illustrati tramite un'analisi di regressione polinomiale.

Capitolo 1

Statistica bayesiana e metodi Markov Chain Monte Carlo

In questo capitolo vengono illustrate le fondamenta della statistica bayesiana, ovvero le definizioni chiave necessarie per svolgere analisi in questo ambito e, naturalmente, il teorema di Bayes. Dopodiché vengono introdotti il principio dell'approssimazione di Monte Carlo e il concetto di catena di Markov e, successivamente, vengono descritti i principali metodi di campionamento *Markov Chain Monte Carlo*. Infine viene mostrata un'applicazione di quest'ultimi all'interno dell'analisi di sopravvivenza. La referenza principale utilizzata per questo capitolo è Hoff (2009).

Uno degli obiettivi della statistica consiste, partendo da un certo fenomeno di interesse, nel comprendere e stimare la sua distribuzione di probabilità. Nella statistica bayesiana tale scopo viene raggiunto utilizzando il teorema di Bayes, formalizzato nel XVIII secolo da Thomas Bayes. Alla conoscenza iniziale del fenomeno, espressa tramite la *distribuzione a priori* dei parametri della popolazione, vengono incorporate le informazioni ricavate dall'osservazione dello stesso, riassunti dal campione di osservazione; viene così ottenuta una conoscenza del fenomeno e delle sue caratteristiche più completa rispetto a quella iniziale, espressa tramite la *distribuzione a posteriori*.

Nonostante l'intuitività dei metodi bayesiani, spesso i calcoli che ne derivano risultato complessi. Per questo motivo vengono usati dei metodi di approssimazione basati sulla legge dei grandi numeri e sull'approssimazione di Monte Carlo.

1.1 Fondamenta della statistica bayesiana

Contesto di lavoro e teorema di Bayes

Seguendo la notazione di Hoff (2009), un modello statistico bayesiano è specificato da un modello parametrico $(\mathscr{Y}, f(y \mid \vartheta))$ e da una distribuzione a priori $(\Theta, p(\vartheta))$, dove

- Θ rappresenta lo **spazio parametrico**; esso include tutti i possibili valori del parametro ϑ che si crede caratterizzi il fenomeno di interesse;
- $p(\vartheta)$ rappresenta la **distribuzione a priori** del parametro; è la distribuzione di probabilità del parametro e rappresenta le assunzioni riguardo le caratteristiche della popolazione, prima di aver osservato i dati;
- $\mathscr{Y}$ rappresenta lo **spazio campionario**; l'insieme di tutti i possibili campioni estraibili dalla popolazione di interesse $\mathbf{Y}$, di cui verrà osservata solamente la singola determinazione $\boldsymbol{y} = (y_1, \dots y_n)$;
- $f(\boldsymbol{y} \mid \vartheta)$ è la ***likelihood*** o **funzione di verosimiglianza**; essa rappresenta la legge di distribuzione del campione $\boldsymbol{y}$ condizionatamente al parametro ϑ.

Nell'esplicitazione della distribuzione a priori, lo statistico esprime le proprie assunzioni riguardanti la popolazione studiata. Una volta estratto il campione dalla stessa, l'idea dell'inferenza bayesiana consiste nell'aggiornamento delle ipotesi fatte in partenza, tramite l'incorporazione delle informazioni contenute nel campione. Per fare ciò si costruisce la distribuzione di ϑ condizionata a y, denominata **distribuzione a posteriori**. Essa si ottiene tramite l'applicazione del teorema di Bayes:

$$p(\vartheta \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \vartheta)\, p(\vartheta)}{\int_\Theta p\left(\boldsymbol{y} \mid \tilde{\vartheta}\right) p\left(\tilde{\vartheta}\right)\, d\tilde{\vartheta}}. \tag{1.1}$$

Distribuzione a priori

Uno dei primi punti critici all'interno della statistica bayesiana è la specificazione della distribuzione a priori: essa può influenzare notevolmente la distribuzione a posteriori del parametro di interesse. Una caratteristica basilare che deve possedere è il comprendere tutti i valori di ϑ considerati plausibili secondo la percezione del fenomeno dallo statistico. Questo rivela chiaramente la soggettività

intrinseca di questo metodo: potenzialmente, ogni soggetto potrebbe fare assunzioni differenti riguardo all'evento considerato e definire quindi distribuzioni a priori differenti.

Si possono attuare diverse metodologie nella scelta della distribuzione da assegnare alla *prior*. Una prima metodologia è scegliere delle *prior* molto soggettive incorporando informazioni di esperti del fenomeno considerato. Nel caso opposto invece, quando non si ha nessuna informazione sul fenomeno considerato, si è soliti utilizzare delle *prior* poco o non informative, in modo da non influenzare eccessivamente la distribuzione a posteriori e lasciare che essa sia basata principalmente sulle informazioni contenute nel campione. Una metodologia usata per la sua intuitività è l'utilizzo delle distribuzioni a priori coniugate. In questo caso la distribuzione a priori e quella a posteriori fanno parte della stessa famiglia parametrica e il passaggio dalla distribuzione a priori a quella a posteriori consiste in un aggiornamento dei corrispondenti parametri. La principale ragione per cui questo metodo viene utilizzato è la semplicità dell'impianto matematico sottostante.

1.2 Metodi Markov Chain Monte Carlo

Uno dei principali problemi della statistica bayesiana è il calcolo della costante di normalizzazione all'interno della formula di Bayes (1.1), ovvero del seguente integrale:

$$\int_{\Theta} p\left(\boldsymbol{y}|\tilde{\vartheta}\right) p\left(\tilde{\vartheta}\right) \, d\tilde{\vartheta}.$$

Questo rende difficile e talvolta impossibile la determinazione esatta dei valori a posteriori dei parametri. Una possibile soluzione a questo problema è l'approssimazione degli stessi tramite il cosiddetto metodo Monte Carlo.

1.2.1 Approssimazione di Monte-Carlo

Il principio dell'approssimazione di Monte Carlo si basa sulla legge dei grandi numeri, per la quale è possibile approssimare il valore atteso relativo a un'intera popolazione con il valore atteso relativo a un campione estratto dalla stessa, purché il campione sia di ampiezza elevata.

Sia $\mathbf{Y} = (Y_1, \dots Y_n)$ un vettore di variabili aleatorie identicamente distribuite e $\boldsymbol{y} = (y_1, \dots y_n)$ la relativa determinazione. Sia ϑ il parametro di interesse che caratterizza la distribuzione di $\mathbf{Y}$.

Si ipotizzi che sia possibile ricavare un numero S di valori casuali, indipendenti ed identicamente distribuiti di ϑ dalla distribuzione a posteriori $p(\vartheta|\boldsymbol{y})$:

$$\vartheta^{(1)}, \ldots, \vartheta^{(n)} \sim \text{ i.i.d. } p(\vartheta|y_1, \ldots, y_n).$$

Allora, per il metodo Monte Carlo, la distribuzione empirica del campione $\vartheta^{(1)}, \ldots, \vartheta^{(n)}$ è in grado di approssimare la distribuzione $p(\vartheta|\boldsymbol{y})$. Al crescere del valore di S l'approssimazione diventa più precisa e centrata in corrispondenza dell'esatta distribuzione a posteriori. Per la legge dei grandi numeri si ha infatti che, se $\vartheta^{(1)}, \ldots, \vartheta^{(n)}$ sono campioni identicamente distribuiti da $p(\vartheta|y_1, \ldots, y_n)$ allora:

$$\frac{1}{S}\sum_{s=1}^{S} g\left(\vartheta^{(s)}\right) \to E\left[g(\vartheta|y_1, \ldots, y_n)\right] = \int_{\Theta} g(\vartheta)\, p(\vartheta|\boldsymbol{y})\, d\vartheta, \qquad \text{per } S \to \infty.$$

Per $S \to \infty$, questo implica che:

- la media campionaria Monte Carlo corrisponde approssimativamente, al crescere di S, al valore atteso reale della popolazione di interesse :

$$\bar{\vartheta} = \frac{\sum_{s=1}^{S} \vartheta^{(s)}}{S} \to E[\vartheta|y_1, \ldots, y_n];$$

- l'errore standard Monte Carlo converge alla deviazione standard reale:

$$\frac{\sum_{s=1}^{S}\left(\vartheta^{(s)} - \bar{\vartheta}\right)^2}{(S-1)} \to Var[\vartheta|y_1, \ldots, y_n].$$

Campionamento di importanza

Nelle casistiche in cui l'estrazione del campione della distribuzione a posteriori non è possibile si ricorre all'uso del **campionamento di importanza**. Esso introduce una distribuzione di probabilità alternativa $h(\vartheta)$, chiamata distribuzione di importanza, dalla quale i campioni vengono generati in modo più efficiente e semplice.

L'idea del campionamento di importanza consiste nello stimare il valore di una quantità utilizzando la media ponderata dei campioni generati dalla distribuzione di importanza. I campioni vengono pesati in base al rapporto tra la distribuzione di interesse e la distribuzione di importanza. In questo modo, i campioni più rari

nella distribuzione di interesse possono essere campionati più frequentemente dalla distribuzione di importanza, aumentando l'efficienza del metodo. Si avrà quindi che, per $\vartheta \sim h(\vartheta)$:

$$\int_{\Theta} \frac{g(\vartheta)\, p(\vartheta|y_1,\ldots,y_n)\, h(\vartheta)}{h(\vartheta)}\, d(\vartheta) = E\left[\frac{g(\vartheta|y_1,\ldots,y_n)}{h(\vartheta)}\right] \approx \frac{1}{S}\sum_{s=1}^{S} \frac{g(\vartheta|y_1,\ldots,y_n)}{h(\vartheta)}.$$

Nei casi in cui la distribuzione a posteriori è molto complessa, come nei casi che vengono analizzati in questa tesi, e difficile da calcolare, così come la costante di normalizzazione, è preferibile usare i metodi di campionamento MCMC piuttosto che il campionamento di importanza. Inoltre il campionamento di importanza prevede la scelta di una distribuzione di importanza che, se è molto distante dalla distribuzione a posteriori, può portare a problemi di alta varianza nel campionamento.

1.2.2 Catene di Markov e le relative proprietà

Fino ad ora sono stati presi in considerazione campioni indipendenti; tuttavia i metodi computazionali che più utilizzano l'approssimazione Monte-Carlo non utilizzano campioni indipendenti bensì campioni dipendenti e attribuibili a catene di Markov.

Definizione 1.1. Una sequenza $Y^{(0)}, Y^{(1)}, \ldots, Y^{(R)}$ di variabili casuali è detta una **catena di Markov** se la condizione

$$\mathbb{P}\left(Y^{(r+1)} \in A \,|\, y^{(0)},\ldots,y^{(r)}\right) = \mathbb{P}\left(Y^{(r+1)} \in A \,|\, y^{(r)}\right)$$

è verificata per ogni $r = 0,\ldots,R-1$ e per ogni insieme misurabile A.

In esse, lo stato delle catena al tempo n è indicato da Y_n mentre l'insieme $1,\ldots,n$ è l'insieme di tutti i possibili stati assumibili dalla catena. Dalla definizione si nota che in una catena di Markov l'elemento in posizione $(r+1)$-esima dipende unicamente dall'elemento in posizione r-esima. Nel caso di catene di Markov omogenee, la legge che definisce il passaggio dall'elemento r-esimo all'elemento $(r+1)$-esimo della sequenza viene chiamata *transition kernel* e definisce completamente la catena di Markov. Nei casi di catene continue, il *transition kernel* viene identificato tramite la legge di densità condizionata $k\left(y^{(r+1)}|y^{(r)}\right)$.

Le catene di Markov utilizzate ai fini di campionamento nei metodi Monte Carlo devono soddisfare le seguenti condizioni di regolarità, descritte qui in modo qualitativo:

- *irriducibilità*: la legge di transizione k permette lo spostamento in tutti i possibili stati della catena, qualsiasi sia il valori di partenza della stessa. Questo garantisce che una catena non si concentri esclusivamente su un sottospazio dello spazio campionario;
- *aperiodicità*: il suo ciclo di transizione non è deterministico. Se tutti gli stati della catena sono aperiodici, allora anche la catena nella sua interezza lo è;
- *Harris Recurrence*: la catena ritorna in un determinato sottospazio dello spazio parametrico un numero illimitato di volte.

La caratteristica più rilevante che queste catene devono possedere in ambito di campionamento è l'ammissione di una distribuzione di probabilità **stazionaria**. Ciò significa che, per $n \to \infty$ la distribuzione marginale dei suoi singoli elementi deve convergere a un'unica distribuzione $p(y)$, nonostante essi siano tra loro dipendenti. Negli algoritmi Markov Chain Monte Carlo tale distribuzione è la distribuzione da cui si vuole campionare ed è quindi solitamente la distribuzione a posteriori dei parametri. In questi metodi l'obiettivo è utilizzare l'approssimazione di Monte-Carlo nel seguente modo:

$$\int g(\boldsymbol{y})\, p(\boldsymbol{y})\, d\boldsymbol{y} \approx \frac{1}{R} \sum_{r=1}^{R} g\left(y^{(r)}\right),$$

dove $y^{(1)}, \ldots, y^{(r)}$ provengono da una catena di Markov con distribuzione stazionaria $p(\boldsymbol{y})$.

Per definire quali catene di Markov ammettono distribuzione stazionaria è necessario introdurre il concetto di **condizione di equilibrio**.

Definizione 1.2. Una catena di Markov $Y^{(1)}, \ldots, Y^{(R)}$ con *transition kernel* k soddisfa la condizione di equilibrio se esiste una funzione f tale che:

$$k(\boldsymbol{y}|\boldsymbol{y}^*) f(\boldsymbol{y}) = k(\boldsymbol{y}^*|\boldsymbol{y}) f(\boldsymbol{y}^*).$$

Una catena che soddisfa la condizione di equilibrio espressa dalla Definizione 1.2 per una funzione di densità p ammette come distribuzione stazionaria la stessa distribuzione p. Questa condizione è sufficiente ma non necessaria.

Una volta che si è stabilito l'ammissione di una legge di distribuzione stazionaria, bisogna assicurarsi che la catena di Markov converga alla stessa al crescere dell'ampiezza campionaria. La garanzia di convergenza alla stessa è fornita dal teorema Ergodico.

Definizione 1.3. (*Harris positiveness*) Una catena di Markov è detta **Harris positive** se è *Harris recurrent* e ammette una distribuzione di probabilità stazionaria.

Teorema 1.1 (Teorema Ergodico)**.** *Sia* $Y^{(1)}, \ldots, Y^{(R)}$ *una catena di Markov* Harris positive *con distribuzione stazionaria* p*. Sia* g *una funzione integrabile rispetto alla distribuzione* p*. Allora si ha convergenza quasi certa*

$$\frac{1}{R}\sum_{r=1}^{R} g\left(y^{(r)}\right) \longrightarrow \int g(\boldsymbol{y})\, p(\boldsymbol{y})\, d\boldsymbol{y}, \qquad \textit{per } R \to \infty.$$

Per generare una catena di Markov è quindi necessario generare il primo valore della catena $Y^0 \sim p_0$ e generare i successivi in base al *transition kernel* k.

L'importanza delle catene di Markov in ambito simulativo risiede nella possibilità di estrarre campioni casuali dalla distribuzione stazionaria della stessa catena. Infatti, per i risultati illustrati in precedenza, si ha che, per un'ampiezza elevata del campione, la legge di distribuzione dello stesso converge alla distribuzione stazionaria. Per questo, in corrispondenza di valore M abbastanza grande, $Y^{(M)}$ può essere considerato come un campione della distribuzione stazionaria cui si è interessati.

1.2.3 Algoritmo Metropolis-Hastings

Un primo algoritmo che combina l'utilizzo di una catena di Markov a un metodo di campionamento Monte Carlo è il Metropolis-Hastings. L'obiettivo dell'algoritmo è ottenere una catena di Markov che converga alla distribuzione di interesse. Per fare ciò l'idea di base è l'utilizzo di una distribuzione adeguata per proporre nuovi valori della catena, i quali verranno accettati all'interno della stessa con una certa probabilità. Per estrarre un campione da una generica distribuzione a posteriori $p(\vartheta|y_1, \ldots, y_n)$ è innanzitutto necessario fissare un valore iniziale della catena $\vartheta^{(0)}$. Dopodiché all'$r-esima$ iterazione dell'algoritmo i passi da svolgere sono i seguenti:

- generare ϑ^* dalla distribuzione di transizione $q\left(\vartheta^*|\vartheta^{(r-1)}\right)$;
- calcolare la probabilità di accettazione come:

$$\alpha\left(\vartheta^*, \vartheta^{(r-1)}\right) = \min\left\{1, \frac{p\left(\vartheta^*|y_1, \ldots, y_n\right)}{p\left(\vartheta^{(r-1)}|y_1, \ldots, y_n\right)} \frac{q\left(\vartheta^{(r-1)}|y_1, \ldots, y_n\right)}{q\left(\vartheta^*|y_1, \ldots, y_n\right)}\right\};$$

- aggiornare lo stato della catena come segue:

$$\vartheta^{(r)} = \begin{cases} \vartheta^* & \text{con probabilità } \alpha \\ \vartheta^{(r-1)} & \text{con probabilità 1-}\alpha \end{cases};$$

Il *transition kernel* della catena di Markov risultante è :

$$k\left(\vartheta^*|\vartheta\right) = q\left(\vartheta^*|\vartheta\right)\alpha(\vartheta^*, \vartheta) + \delta_\vartheta\left(\vartheta^*\right)\int q\left(s|\vartheta\right)\left(1 - \alpha(s|\vartheta)\right)\, ds$$

Tale kernel soddisfa la condizione di equilibrio. Di conseguenza la catena generata ammette come legge di distribuzione stazionaria $p\left(\vartheta|y_1 \ldots, y_n\right)$. Per un numero di iterazioni elevato, la sequenza di campioni prodotta dall'algoritmo rappresenterà un campione approssimato dalla distribuzione di interesse. È necessario ricordare che nella fase iniziale dell'algoritmo, denominata **burn-in**, i campioni generati non seguiranno la distribuzione stazionaria dato che affinché la catena raggiunga la convergenza è necessario un determinato numero di iterazioni.

Un elemento cruciale di questo metodo è la scelta della *proposal distribution*. Una scelta utile è l'utilizzo di una proposta simmetrica, dalla quale deriva una semplificazione della probabilità di accettazione α. In quest'ottica viene spesso utilizzato un *Random Walk* Gaussiano, ovvero

$$\left(\vartheta^*|\vartheta\right) \sim N\left(\vartheta, \sigma^2\right).$$

di cui un utilizzo verrà mostrato più avanti, nella Sezione 1.3.

1.2.4 Algoritmo Gibbs Sampler

L'algoritmo Gibbs Sampler è un ulteriore metodo Markov Chain Monte Carlo che viene utilizzato per generare campioni casuali da leggi di distribuzioni multivariate nei casi in cui il campionamento dei relativi parametri risulta

complessa se svolta congiuntamente. Il metodo qui illustrato prevede quindi la semplificazione dello stesso tramite il campionamento singolo delle componenti attraverso le *full conditional distributions*, ovvero la legge di distribuzione di ogni singolo parametro condizionata ai restanti. Sia $p(\vartheta|\boldsymbol{y})$ la distribuzione a posteriori di interesse con $\vartheta = (\vartheta_1 \ldots, \vartheta_L)$. Per ogni ϑ_l per $l = 1 \ldots, L$, la *full conditional distribution* è data da:

$$p(\vartheta_l|-) = p\left(\vartheta_l|\boldsymbol{y}, \vartheta_1 \ldots, \vartheta_{l-1}, \vartheta_{l+1}, \vartheta_L\right).$$

L'algoritmo in questione svolge iterativamente il campionamento delle *full conditional* e produce una sequenza di valori per i parametri la cui legge di distribuzione convergerà alla distribuzione multivariata di interesse. In particolare, dopo aver inizializzato $\vartheta^{(0)} = \left(\vartheta_1^{(0)} \ldots, \vartheta_L^{(0)}\right)$ i passi dell'algoritmo all'r-esima iterazione prevedono la generazione di:

$$\begin{aligned}
&\vartheta_1^{(r)} \sim p\left(\vartheta_1|\boldsymbol{y}, \vartheta_2^{(r-1)} \ldots, \vartheta_L^{(r-1)}\right);\\
&\vartheta_2^{(r)} \sim p\left(\vartheta_2|\boldsymbol{y}, \vartheta_1^{(r)}, \vartheta_3^{(r-1)} \ldots, \vartheta_L^{(r-1)}\right);\\
&\quad\vdots \qquad\qquad \vdots\\
&\vartheta_l^{(r)} \sim p\left(\vartheta_L|\boldsymbol{y}, \vartheta_1^{(r)} \ldots, \vartheta_{L-1}^{(r)}\right).
\end{aligned}$$

In questo caso la sequenza generata convergerà alla distribuzione a posteriori $p(\vartheta|\boldsymbol{y})$. Si noti che la sequenza generata è proprio una catena di Markov poiché il valore $\vartheta^{(r)}$ dipende dai valori $\vartheta^{(1)} \ldots, \vartheta^{(r-1)}$ attraverso il valore $\vartheta^{(r-1)}$.

Considerazioni generali

Per tutti gli algoritmi considerati, è necessario porre in evidenza che la stabilizzazione intorno alla distribuzione di interesse avviene qualsiasi sia l'inizializzazione scelta per il valore di partenza $\vartheta^{(0)}$. Tuttavia, affinché l'algoritmo raggiunga la convergenza sono necessarie un numero di iterazioni abbastanza elevato, motivo per cui i valori generati all'inizio dello stesso non dovranno essere considerati, poiché non distribuiti secondo la distribuzione di interesse. La fase dell'algoritmo che include i campioni menzionati viene detta **burn-in**.

Una volta generato il campione è necessaria una fase di diagnostica per verificare che esso sia stato generato correttamente. In tale fase è necessario verificare, anche attraverso metodi grafici, che la sequenza generata sia in effetti

stazionaria e che l'autocorrelazione tra i vari elementi della stessa sia compatibile con quella di una catena di Markov.

Gli algoritmi Gibbs Sampler e Metropolis-Hastings sono tra loro simili; il primo infatti è un caso particolare del secondo, nel quale le *full conditionals* sono le *proposal distribution* e la probabilità di accettazione α è pari a 1. Il Gibbs Sampler viene utilizzato esclusivamente per distribuzioni multivariate mentre il Metropolis-Hastings può essere utilizzato anche in distribuzioni univariate. Inoltre il Gibbs Sampler può essere utilizzato solo se le *full conditionals* sono ricavabili in forma chiusa; nei casi in cui questo non è possibile, come nel modello di regressione logistica, viene utilizzato il Metropolis-Hastings.

Per ulteriori approfondimenti sugli algoritmi trattati e altri possibili metodo di campionamento si vedano Robert & Casella (2010) e Robert & Casella (1999).

1.3 Utilizzo del Metropolis-Hastings nell'analisi di sopravvivenza

L'analisi di sopravvivenza è un insieme di metodi statistici adatto allo studio dei tempi di accadimento di un evento in ambito biomedico. Gli eventi considerati sono solitamente la diagnosi di una malattia oppure la morte. Gli studi vengono svolti seguendo un determinato numero di individui per un periodo prefissato e osservando per ognuno il tempo di accadimento dell'evento di interesse. E' possibile che l'evento di interesse non venga rilevato nel periodo di osservazione per diverse ragioni; in questo caso si parla di **dati censurati**.

Nel dataset *Stanford Heart Transplant*, disponibile nella documentazione standard di R ed analizzato a titolo esemplificativo nell'articolo Escobar & Meeker Jr (1992), sono disponibili i dati per un analisi di sopravvivenza. Il dataset contiene i dati relativi a 184 individui che hanno subito un trapianto di cuore, riportando per ognuno il tempo di sopravvivenza dopo l'aver subito il trapianto. L'evento di interesse in questo caso è la morte dell'individuo. Per ogni individuo viene osservata la coppia di valori $(\mathbf{t}, \mathbf{d})$ dove:

- **t** rappresenta il tempo al verificarsi dell'evento o il tempo disponibile all'ultima osservazione;

- **d** è l'indicatore dell'accadimento dell'evento:

$$\mathbf{d} = \begin{cases} 1 & \text{se l'evento viene osservato} \\ 0 & \text{se l'evento non viene osservato} \end{cases}$$

Si assume che i tempi di sopravvivenza siano variabili casuali indipendenti identicamente distribuite come una Weibull(γ, β). Le funzioni di **rischio** o ***hazard*** e di **sopravvivenza** nella distribuzione di Weibull(γ, β) sono rispettivamente:

$$h(t|\gamma, \beta) = \frac{\gamma}{\beta}\left(\frac{\mathbf{t}}{\beta}\right)^{\gamma-1}, \qquad S(t|\gamma, \beta) = \exp\left\{-\left(\frac{\mathbf{t}}{\beta}\right)^{\gamma}\right\}.$$

La funzione di densità viene poi ottenuta come:

$$f(\mathbf{t}|\gamma, \beta) = h(t|\gamma, \beta)\, S(t|\gamma, \beta).$$

Per stimare i parametri della distribuzione nel caso del dataset in questione si è utilizzato un approccio bayesiano. La funzione di verosimiglianza per questo modello parametrico per il vettore di parametri (γ, β) è data da:

$$p(\mathbf{t}, \mathbf{d}|\boldsymbol{\vartheta}) \propto \prod_{i=1}^{n} h(t_i|\gamma, \beta)^{d_i} S(t_i|\gamma, \beta) = \prod_{i:d_i=1} f(t_i|\gamma, \beta) \prod_{i:d_i=0} S(t_i|\gamma, \beta).$$

Dato che i parametri di interesse hanno valori strettamente positivi, si è deciso di utilizzare una riparametrizzazione degli stessi e considerare il vettore $\boldsymbol{\vartheta} = (\log\gamma, \log\beta)$. Per le distribuzioni a priori di essi sono state scelte due distribuzioni uguali:

$$\log(\gamma) \sim N(0, 100) \qquad \log(\beta) \sim N(0, 100).$$

Applicando il teorema di Bayes si può infine ricavare la distribuzione a posteriori dei parametri:

$$\begin{aligned} p(\boldsymbol{\vartheta}|\mathbf{t}, \mathbf{d}) &\propto p(\mathbf{t}, \mathbf{d}|\boldsymbol{\vartheta})\, p(\boldsymbol{\vartheta}) \\ &\propto \prod_{i:d_i=1} f(t_i|\gamma, \beta) \prod_{i:d_i=0} S(t_i|\gamma, \beta) \exp\left\{\left(-\frac{\gamma^2}{200}\right)\right\} \exp\left\{\left(-\frac{\beta^2}{200}\right)\right\} \end{aligned}$$

La distribuzione a posteriori dei parametri ha una forma molto complicata; per questo motivo essa viene approssimata tramite l'utilizzo dell'algoritmo *Random Walk* Metropolis-Hastings.

Come *proposal distribution* è stato scelto un *Random Walk* Gaussiano centrato nel vettore di interesse:

$$\boldsymbol{\vartheta}^*|\boldsymbol{\vartheta} \sim \mathrm{N}_2\left(\boldsymbol{\vartheta}, 0.25^2 I_2\right)$$

Tramite l'algoritmo svolto con 50000 iterazioni è stato ottenuto un campione per entrambi i parametri di interesse; dopo aver verificato tramite strumenti grafici di diagnostica la bontà del campione ottenuti, sono stati calcolati gli indici a posteriori in Tabella 1.1.

Parametro	**Media**	**Deviazione standard**
γ	0.549	0.044
β	1247.973	226.605

Tabella 1.1: Valori dei parametri a posteriori

In Figura 1.1 e Figura 1.2 viene mostrata la media a posteriori con intervalli di credibilità al 95% per la funzione di sopravvivenza e la funzione di rischio.

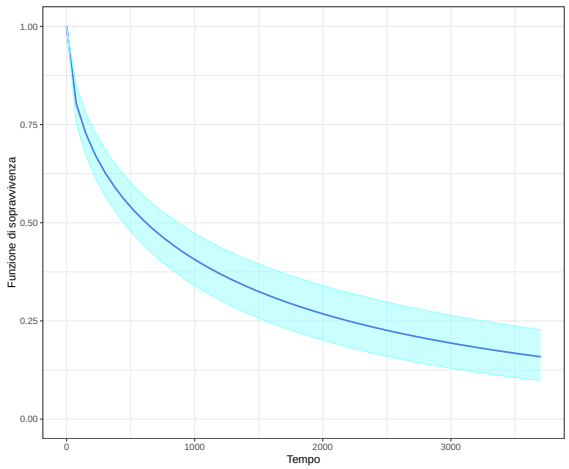

Figura 1.1: Funzione di sopravvivenza

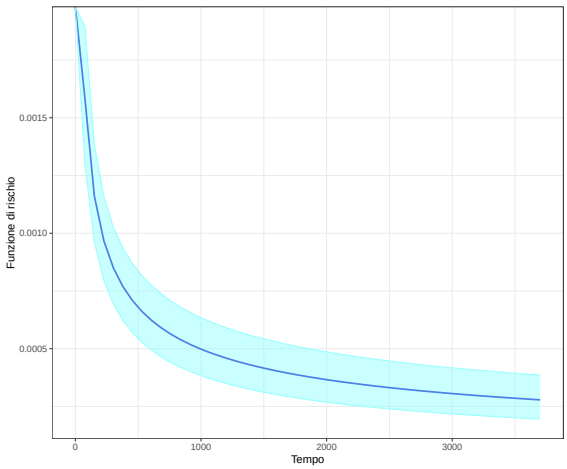

Figura 1.2: Funzione di rischio

Come si può notare, sia la funzione di sopravvivenza che la funzione di rischio hanno un andamento decrescente. La funzione di rischio tuttavia ha una decrescita più veloce e indica che, dopo aver subito il trapianto di cuore, il rischio di morire decresce di molto e velocemente. Inoltre, i valori che assume la funzione di rischio sono molto piccoli e vicino allo zero. Questo è molto positivo e sottolinea l'utilità e l'efficacia del trapianto. La funzione di sopravvivenza

invece mostra un andamento decrescente molto ripido per i primi giorni dopo il trapianto: questo può essere dovuto a casi in cui il trapianto non è andato a buon fine oppure casi in cui ci sono state complicazioni immediatamente successive all'operazioni. Dopo questo balzo iniziale, la decrescita ha un andamento più liscio. E' utile ricordare che nell'analisi sono stati considerati anche dati censurati per cui la decrescita della curva di sopravvivenza può essere dovuta anche alla presenza di dati censurati e non solo al verificarsi dell'evento.

Capitolo 2

Modello di regressione lineare

In questo capitolo viene illustrato il modello di regressione lineare in ambito bayesiano. Prima di tutto viene introdotta la distribuzione normale multivariata e le sue caratteristiche; dopodiché vengono presentati il modello parametrico e le distribuzione a priori dei parametri utilizzati in un modello di regressione bayesiano. Infine viene mostrato l'utilizzo dell'algoritmo Gibbs Sampler per l'approssimazione della distribuzione a posteriori dei parametri.

La regressione è un metodo utilizzato in statistica per studiare e indagare la relazione tra diversi fattori in un fenomeno di interesse. In particolare in ambito regressivo si è interessati a capire come una variabile di interesse, chiamata comunemente **variabile risposta** o **dipendente**, dipenda da altre variabili, chiamate **covariate** o **variabili indipendenti** e come essa vari al variare delle covariate. La relazione che lega le variabili può essere di varia natura; in questo capitolo verrà analizzato il modello di regressione basato su una relazione di tipo lineare.

2.1 Distribuzione normale multivariata

Un modello statistico è detto multivariato se, per ogni unità statistica, vengono effettuate diverse misurazioni riguardanti un numero $p > 1$ di variabili di interesse. Una delle leggi di distribuzione maggiormente utilizzata per la descrizione di un fenomeno multivariato è la generalizzazione della distribuzione normale univariata a p-dimensioni, ovvero la distribuzione **normale multivariata**.

Seguendo la notazione di Hoff (2009), un vettore di variabili casuali $\mathbf{Y}$ p-dimensionale è detto avere una distribuzione normale multivariata se la sua funzione di densità è data da:

$$p(\boldsymbol{y} \mid \boldsymbol{\vartheta}, \Sigma) = (2\pi)^{-p/2} \det(\Sigma)^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - \boldsymbol{\vartheta})^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\vartheta}) \right\}$$

dove

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} \qquad \boldsymbol{\vartheta} = \begin{pmatrix} \vartheta_1 \\ \vartheta_2 \\ \vdots \\ \vartheta_p \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \dots & \sigma_{1,p} \\ \sigma_{1,2} & \sigma_2^2 & \dots & \sigma_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,p} & \dots & \dots & \sigma_p^2 \end{pmatrix}.$$

Il supporto della variabile casuale $\mathbf{Y}$ è $\mathbb{R}^p$. La distribuzione è completamente caratterizzata dai parametri:

- $\boldsymbol{\vartheta}$ che rappresenta il **valore atteso** del vettore $\mathbf{Y}$; anch'esso ha supporto pari a $\mathbb{R}^p$;
- Σ che rappresenta la **matrice di varianza e covarianza** del vettore $\mathbf{Y}$. Gli elementi sulla diagonale rappresentano le varianze delle variabili e sono quindi elementi strettamente positivi.

2.2 Modello di regressione lineare in ambito bayesiano

In un modello di regressione l'oggetto di studio è la relazione tra una variabile casuale $\mathbf{Y}$ e un gruppo di covariate $\mathbf{x} = (x_1, x_2, \dots x_p)$. In particolare viene studiata la variazione della legge di distribuzione della variabile risposta in corrispondenza della variazione delle variabili indipendenti.

2.2.1 Modello parametrico

E' innanzitutto necessario definire un modello parametrico in grado di descrivere la relazione tra le variabili; formalmente questo si traduce nella costruzione della legge di distribuzione condizionata $p(y \mid \boldsymbol{x})$.

In un modello di regressione lineare normale la relazione tra il valore atteso della variabile $\mathbf{Y}$ e le covariate $\mathbf{x}$ è di tipo lineare per un insieme di parametri ed è formalizzabile come:

$$\mathbb{E}[Y \mid \mathbf{x}] = \beta_1 x_1 + \dots + \beta_p x_p = \boldsymbol{\beta}^T \mathbf{x}.$$

Utilizzando questa relazione è possibile specificare la legge di distribuzione della variabile risposta condizionata alle covariate e ai parametri come segue. Seguendo la notazione di Hoff (2009), sia $\mathbf{y} = (y_1, \dots, y_n)$ il vettore colonna n-dimensionale della variabile risposta e sia $\mathbf{X}$ la matrice $n \times p$ di covariate la cui i-esima riga corrisponde all'unità statistica $\mathbf{x}_i$; allora il **modello parametrico**

in una regressione normale è tale per cui $p\left(\mathbf{y} \mid \mathbf{X}, \beta, \sigma^2\right) \sim \mathcal{N}_p\left(\mathbf{X}\beta, \sigma^2\mathbf{I}_p\right)$. La sua funzione di densità è:

$$p\left(\mathbf{y} \mid \mathbf{X}, \beta, \sigma^2\right) \propto \left(\sigma^2\right)^{-p/2} \exp\left\{ -\frac{1}{2\sigma^2}\left(\boldsymbol{y} - \mathbf{X}\beta\right)^T\left(\boldsymbol{y} - \mathbf{X}\beta\right) \right\}$$

dove

$$\boldsymbol{X\beta} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} \beta_1 x_{1,1} + \ldots + \beta_p x_{1,p} \\ \vdots \\ \beta_1 x_{n,1} + \ldots + \beta_p x_{n,p} \end{pmatrix} = \begin{pmatrix} \mathbb{E}\left[Y_1 \mid \boldsymbol{\beta}, \mathbf{x}_1\right] \\ \vdots \\ \mathbb{E}\left[Y_n \mid \boldsymbol{\beta}, \mathbf{x}_n\right] \end{pmatrix}.$$

In un modello di regressione lineare la stima della variabile risposta è data da

$$Y_i = \boldsymbol{\beta}^T x_i + \varepsilon_i,$$

dove ε_i rappresentano gli **errori** commessi dal modello nella stima. Affinchè si tratti di un modello di regressione lineare normale essi devono essere tra loro indipendenti e identicamente distribuiti come:

$$\varepsilon_1, \ldots \varepsilon_n \sim \text{i.i.d. } \mathcal{N}\left(0, \sigma^2\right).$$

2.2.2 Distribuzioni a priori

La distribuzione del modello parametrico è caratterizzata dai parametri $\boldsymbol{\beta}$ e σ^2 e in un approccio di tipo bayesiano essi sono dotati di una legge di distribuzione propria.
Per la distribuzione a priori del parametro $\boldsymbol{\beta}$ si consideri la funzione di densità del modello parametrico come funzione di $\boldsymbol{\beta}$:

$$p\left(\mathbf{y} \mid \mathbf{X}, \beta, \sigma^2\right) \propto \exp\left\{ -\frac{1}{2\sigma^2}\left[\mathbf{y}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta - 2\beta^T\mathbf{X}^T\mathbf{y}\right] \right\}.$$

Il ruolo del parametro $\boldsymbol{\beta}$ sembra molto simile al ruolo del campione $\mathbf{y}$ e per questo motivo si suggerisce l'utilizzo di una distribuzione a priori coniugata e quindi l'utilizzo di una distribuzione normale multivariata come legge di distribuzione per $\boldsymbol{\beta}$.

Per quanto riguarda il parametro σ^2, riferito alla varianza della variabile risposta, è necessario, per ovvie ragioni, considerare variabili casuali con un supporto strettamente positivo. In ambito del modello di regressione lineare normale viene scelta come legge di distribuzione a priori per il parametro σ^2 la distribuzione gamma inversa.
Le distribuzioni a priori risultano quindi essere :

$$\boldsymbol{\beta} \sim \mathcal{N}_p\left(\beta_0, \Sigma_0\right) \qquad \sigma^2 \sim \text{inverse gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right).$$

Per entrambe le distribuzioni la scelta dei parametri che li caratterizzano, $\boldsymbol{\beta_0}$ e Σ_0 per il parametro $\boldsymbol{\beta}$ e ν_0 e σ_0^2 per il parametro σ^2, è da compiere coerentemente al fenomeno che si sta analizzando.

2.2.3 Distribuzioni a posteriori

In ambito bayesiano il passo successivo alla definizione del modello parametrico e della distribuzione a priori consiste nel ricavare la distribuzione a posteriori dei parametri. In un modello di regressione multipla normale il numero di parametri da stimare può essere potenzialmente molto elevato, motivo per cui la distribuzione a posteriori congiunta degli stessi diventa difficile da calcolare; per questa ragione si ricorre al calcolo di una sua approssimazione tramite gli algoritmi esposti nella Sezione 1.2. In particolare in questa sezione viene mostrato l'utilizzo dell'algoritmo Gibbs Sampler spiegato nella Sottosezione 1.2.4. La prima fase dell'algoritmo consiste nella derivazione delle *full conditional distributions* a partire dalla distribuzione congiunta dei parametri. Dato che le priori specificate sono tra loro indipendenti, la distribuzione congiunta sarà data da :

$$p\left(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \mathbf{X}\right) \propto p\left(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2\right) p\left(\boldsymbol{\beta}\right) p\left(\sigma^2\right).$$

Parametro β

Ricordando che la distribuzione a priori del parametro è $\beta \sim \mathcal{N}_p\left(\beta_0, \Sigma_0\right)$ e che la distribuzione del modello parametrico è $\left\{ \mathbf{y} \mid \mathbf{X}, \beta, \sigma^2 \right\} \sim \mathcal{N}_p(\mathbf{X}\beta, \sigma^2 \mathbf{I}_p)$, si può ottenere la *full conditional distribution* del parametro nel seguente modo:

$$
\begin{aligned}
p\left(\beta \mid \mathbf{y}, \mathbf{X}, \sigma^2\right) &\propto p\left(\mathbf{y}|\mathbf{X}, \beta, \sigma^2\right) \times p\left(\beta\right) \\
&\propto \exp\left\{ -\frac{1}{2\sigma^2}\left[\mathbf{y}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta - 2\beta^T\mathbf{X}^T\mathbf{y}\right] \right\} \cdot \\
&\qquad \exp\left\{ -\frac{1}{2}\left[\beta^T\Sigma_0^{-1}\beta + \beta_0^T\Sigma_0^{-1}\beta_0 - 2\beta^T\Sigma_0^{-1}\beta_0\right] \right\} \\
&\propto \exp\left\{ -\frac{1}{2\sigma^2}\left[\beta^T\mathbf{X}^T\mathbf{X}\beta - 2\beta^T\mathbf{X}^T\mathbf{y}\right] \right\} \cdot \\
&\qquad \exp\left\{ -\frac{1}{2}\left[\beta^T\Sigma_0^{-1}\beta - 2\beta^T\Sigma_0^{-1}\beta_0\right] \right\} \\
&= \exp\left\{ -\frac{1}{2}\left[\beta^T\left(\frac{\mathbf{X}^T\mathbf{X}}{\sigma^2} + \Sigma_0^{-1}\right)\beta - 2\beta^T\left(\frac{\mathbf{X}^T\mathbf{y}}{\sigma^2} + \Sigma_0^{-1}\beta_0\right)\right] \right\}.
\end{aligned}
$$

In quest'ultima espressione si riconosce il kernel di una densità normale multivariata caratterizzata da i seguenti indici di, rispettivamente, posizione e dispersione:

- $\mathbb{E}\left[\beta \mid \mathbf{y}, \mathbf{X}, \sigma^2\right] = \left(\frac{\mathbf{X}^T\mathbf{X}}{\sigma^2} + \Sigma_0^{-1}\right)^{-1}\left(\frac{\mathbf{X}^T\mathbf{y}}{\sigma^2} + \Sigma_0^{-1}\beta_0\right)$;
- $Var\left[\beta \mid \mathbf{y}, \mathbf{X}, \sigma^2\right] = \left(\frac{\mathbf{X}^T\mathbf{X}}{\sigma^2} + \Sigma_0^{-1}\right)^{-1}$.

Si conclude quindi che, a posteriori,

$$
\boldsymbol{\beta} \sim \mathcal{N}_p\left(\left(\frac{\mathbf{X'X}}{\sigma^2} + \Sigma_0^{-1}\right)^{-1}\left(\frac{\mathbf{X'y}}{\sigma^2} + \Sigma_0^{-1}\beta_0\right), \left(\frac{\mathbf{X'X}}{\sigma^2} + \Sigma_0^{-1}\right)^{-1}\right).
$$

Parametro σ^2

Si ricordi che la distribuzione a priori del parametro è $\sigma^2 \sim$ inverse-gamma $\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$ e che la distribuzione del modello parametrico è $\left\{ \mathbf{y} \mid \mathbf{X}, \beta, \sigma^2 \right\} \sim \mathcal{N}_p(\mathbf{X}\beta, \sigma^2\mathbf{I}_p)$. Per ottenere la distribuzione a posteriori, per facilità matematica, si considera l'inverso del parametro σ^{-2} che si distribuisce come $\sigma^{-2} \sim$ gamma $\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$.

La *full conditional distribution* del parametro σ^{-2} è quindi ottenibile come:

$$\begin{aligned} p\left(\sigma^{-2} \mid \mathbf{y}, \mathbf{X}, \beta\right) &\propto p\left(\mathbf{y}|\mathbf{X}, \beta, \sigma^{-2}\right) \times p\left(\sigma^{-2}\right) \\ &\propto \sigma^{-2n} \exp\left\{ -\frac{1}{2}\sigma^{-2}\left[\mathbf{y}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta - 2\beta^T\mathbf{X}^T\mathbf{y}\right] \right\} \cdot \\ &\quad \sigma^{-2\left(\frac{\nu_0}{2}-1\right)} \exp\left\{ -\left(\frac{\nu_0\sigma_0^2}{2}\right)\sigma^{-2} \right\} \\ &= \sigma^{-2\left(\frac{\nu_0}{2}+n-1\right)} \exp\left\{ \sigma^{-2}\left[\frac{\mathbf{y}^T\mathbf{y}}{2} + \frac{\beta^T\mathbf{X}^T\mathbf{X}\beta}{2} - \beta^T\mathbf{X}^T\mathbf{y} + \frac{\nu_0\sigma_0^2}{2}\right] \right\}. \end{aligned}$$

In quest'ultima espressione si riconosce il kernel di una distribuzione gamma con i seguenti parametri:

- **parametro di forma** pari a $(\nu 0 + n)/2$;
- **parametro di scala** pari a $\frac{1}{2}\left[\mathbf{y'y} + \beta'\mathbf{X'X}\beta - 2\beta'\mathbf{X'y} + \nu_0\sigma_0^2\right]$.

Si conclude quindi che, a posteriori,

$$\sigma^2 \sim \text{inverse gamma}\left(\frac{(\nu 0 + n)}{2}, \frac{\left[\mathbf{y'y} + \beta'\mathbf{X'X}\beta - 2\beta'\mathbf{X'y} + \nu_0\sigma_0^2\right]}{2}\right).$$

Utilizzo del Gibbs Sampler

Una volta ottenute le *full conditional distributions* dei parametri è possibile approssimare la distribuzione a posteriori congiunta tramite l'utilizzo dell'algoritmo Gibbs Sampler.

Dati i valori $\left\{ \boldsymbol{\beta}^{(s)}, \sigma^{2(s)} \right\}$ al passo r-esimo dell'algoritmo vengono generati i valori al passo (r+1)-esimo

- campionando $\boldsymbol{\beta}^{(s+1)} \sim p\left(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \sigma^{2(s)}\right)$;
- campionando $\sigma^{2(s+1)} \sim p\left(\sigma^2 \mid \mathbf{y}, \mathbf{X}, \boldsymbol{\beta}^{(s+1)}\right)$.

Svolgendo un numero elevato di iterazioni si ottiene un sequenza di valori $\left\{ \boldsymbol{\beta}, \sigma^2 \right\}$ in grado di approssimare la distribuzione a posteriori dei parametri.

Un'applicazione del modello di regressione lineare tramite l'utilizzo del Gibbs Sampler verrà mostrato nel Capitolo 4.

Capitolo 3

Selezione del modello

In questo capitolo vengono illustrati alcuni metodi che permettono la selezione del modello migliore in un'analisi di regressione. In primo luogo viene considerato come metodo di confronto tra i modelli la loro capacità predittiva e vengono quindi presentati diversi approcci per la sua valutazione. In particolare vengono analizzati i criteri basati sull'informazione e il metodo della *leave-one-out cross validation*. Infine viene presentato un metodo proposto recentemente in Silva & Zanella (2022) in grado di risolvere alcune problematiche presentate dagli approcci illustrati in precedenza.

Una delle sfide che si deve affrontare in ambito di regressione è la selezione del modello che meglio si adatta ai dati disponibili. Questo consiste nella scelta del forma del modello e nella selezione delle variabili con maggiore capacità esplicativa rispetto alla variabile risposta. Quando si analizzano dati reali questo è molto complicato poiché si possono avere diverse variabili a disposizione e prima di valutare il modello non si ha nessuna evidenza di quali siano effettivamente in grado di fornire informazioni utili alla stima della variabile dipendente. Sono state quindi costruite diverse metodologie e criteri in grado di valutare e confrontare i modelli costruiti in modo di essere in grado di selezionare il modello migliore, e quindi le variabili maggiormente informative.

La referenza principale utilizzata per questo capitolo è Gelman et al. (2013).

3.1 Capacità predittiva del modello

Un modello può essere valutato sotto diversi aspetti. L'aspetto che viene preso in considerazione in questo capitolo, e su cui vengono costruiti i criteri che vengono qui presentati, è la capacità predittiva del modello. Per **capacità predittiva** di un modello si intende l'accuratezza con cui un modello, tramite i parametri stimati,

riesce a stimare la variabile risposta sulla base delle sue covariate. Il modo ideale per stimarla sarebbe utilizzando un nuovo campione, proveniente dalla stessa distribuzione da cui sono stati generati i dati, ma che non è stato utilizzato nella fase di costruzione del modello. Ciò che si vorrebbe fare è quindi stimare la variabile risposta del nuovo campione utilizzando i parametri del modello costruiti sulla base di tutte le osservazioni disponibili in precedenza e i valori delle covariate del nuovo campione. Dalla notazione di Gelman et al. (2013), sia f la reale distribuzione dei dati, y il campione osservato e $\tilde{y}$ il nuovo campione proveniente sempre da f. Allora la misura di capacità predittiva del modello per un singola osservazione $\tilde{y}_i$, utilizzando la scala logaritmica, è data da

$$\log p(\tilde{y}_i \mid \boldsymbol{\vartheta}) = \log \mathbb{E}_{\vartheta}[p(\tilde{y}_i \mid \boldsymbol{\vartheta})] = \log \int p(\tilde{y}_i \mid \boldsymbol{\vartheta})\, p(\boldsymbol{\vartheta} \mid y)\, d\vartheta,$$

dove ϑ proviene dalla distribuzione a posteriori $p(\boldsymbol{\vartheta} \mid y)$. L'espressione qui sopra rappresenta la densità predittiva del campione $\tilde{y}_i$ indotta dalla distribuzione a posteriori di ϑ. Considerando però che, al momento della valutazione del modello, il campione nuovo $\tilde{y}$ è sconosciuto, si costruisce la densità predittiva del modello tramite il loro valore atteso. La funzione in grado di fare questo è chiamata *expected out-of-sample log predictive density* (ELPD) ed è data da

$$\text{ELPD} = \mathbb{E}_f(\log p_{\text{post}}(\tilde{y}_i)) = \int \log p_{\text{post}}(\tilde{y}_i) f(\tilde{y}_i)\, d\tilde{y}. \tag{3.1}$$

Se si considera l'intero campione nuovo $\tilde{y}$ invece che un singolo elemento di esso, la misura della capacità predittiva è data dalla somma dei contributi dei singoli elementi; la funzione in grado di fare questo è l'ELPPD $= \sum_{i=1}^{n} \mathbb{E}_f(\log p_{\text{post}}(\tilde{y}_i))$.

Il problema fondamentale di questo è che, oltre al fatto che il nuovo campione non è disponibile al momento della valutazione del modello, anche la distribuzione reale dei dati f non è conosciuta e quindi calcolare la funzione (3.1) non è possibile. Per questo si ricorre spesso a una sua approssimazione, ottenibile tramite diversi metodi.

Un primo metodo intuitivo è la valutazione della capacità predittiva tramite i dati disponibili. Questo significa calcolare la funzione (3.1) considerando come nuovo il campione su cui è stato costruito il modello. Questo può essere fatto utilizzando la funzione *log pointwise predictive density* (LPPD), definita come ù:

$$\text{LPPD} = \log \prod_{i=1}^{n} p_{\text{post}}(y_i) = \sum_{i=1}^{n} \log \int p(y_i \mid \vartheta)\, p(\vartheta \mid y)\, d\vartheta. \tag{3.2}$$

In pratica essa viene calcolata come:

$$\text{computed LPPD} = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p\left(y_i \mid \vartheta^s\right) \right).$$

L'evidente limite di questo modello è che si sta calcolando una misura su dei dati che non solo non sono nuovi, ma che sono anche gli stessi su cui è stato stimato il modello. Per questo motivo la funzione calcolata in questo modo produrrà una sovrastima di quella che è la reale capacità predittiva del modello.

3.2 Criteri di informazione

Il secondo metodo che viene analizzato per l'approssimazione della funzione (3.1) è l'utilizzo dei **criteri di informazione**. Essi sono così chiamati poiché valutano e misurano la quantità di informazione contenuta nel campione. Questi criteri nascono con l'obiettivo di risolvere il problema di sovrastima che si ha quando si valuta il modello tramite la funzione (3.2). Infatti tutti i criteri valutano la capacità predittiva tramite l'aggiunta a una funzione simile alla LPPD di un elemento in grado di correggere quest'errore. I criteri di informazione sono valutati sulla scala della devianza e sono tutti della forma:

$$\text{IC} = -2 \log p\left(y \mid \hat{\vartheta}\right) + \text{correzione}(p).$$

3.2.1 *Akaike information criterion*

Uno dei criteri di informazione più diffusi è l'*Akaike information criterion* (AIC). Esso viene utilizzato soprattutto in ambito frequentista e, in effetti, non contiene elementi bayesiani al suo interno. La sua formulazione è

$$\text{AIC} = -2 \log p\left(y \mid \hat{\vartheta}_{mle}\right) + 2k.$$

La stima della funzione di densità a posteriori dei dati avviene utilizzando una stima puntuale dei parametri e non l'intera distribuzione a posteriori degli stessi. La stima puntuale che viene utilizzata è la stima di massima verosimiglianza che viene usata nella fase di inferenza in ambito frequentista. Inoltre la correzione per la sovrastima viene fatta per il numero k dei parametri stimati dal modello.

3.2.2 *Deviance information criterion*

Un altro criterio di informazione, simile all'AIC, è il *Deviance information criterion* (DIC). Esso può essere visto come la versione bayesiana dell'AIC perchè presenta la stessa logica di quest'ultimo ma con due modifiche di stampo puramente bayesiano. La sua formulazione è data da

$$\mathrm{DIC} = -2 \log p\left(y \mid \hat{\vartheta}_{\mathrm{Bayes}}\right) + 2p_{\mathrm{DIC}}.$$

La prima modifica che si può notare è nella stima dei parametri: essa è ancora una stima puntuale ma invece di essere la stima di massima verosimiglianza è pari alla media a posteriori dei parametri, ovvero $\hat{\vartheta}_{\mathrm{Bayes}} = \mathbb{E}(\vartheta \mid y)$. La seconda modifica risiede nel numero dei parametri per cui viene corretta la funzione. In ambito bayesiano il numero effettivo dei parametri tiene conto non solo del numero di parametri espliciti nel modello, ma anche dell'incertezza o dell'informazione che il modello fornisce su tali parametri ed è una misura della complessità del modello. Per questo motivo è un numero che necessita di essere stimato. Nella formula del DIC il numero dei parametri è rappresentato da p_{DIC} che è pari a

$$p_{\mathrm{DIC}} = 2\left(\log p\left(y \mid \hat{\vartheta}_{\mathrm{Bayes}}\right) - \mathbb{E}_{\mathrm{post}}\left(\log p\left(y \mid \vartheta\right)\right)\right).$$

Il valore atteso nella formula sovrastante è calcolabile come

$$\mathbb{E}_{\mathrm{post}}\left(\log p\left(y \mid \vartheta\right)\right) = \frac{1}{S}\sum_{i=1}^{S} \log p\left(y \mid \vartheta^s\right),$$

dove i valori ϑ^s sono campioni della distribuzione a posteriori dei parametri.

3.2.3 *Watanabe-Akaike information criterion*

Un criterio che invece prevede un approccio completamente bayesiano è il *Watanabe-Akaike information criterion* (WAIC). Esso, oltre a stimare il numero effettivo dei parametri, non utilizza una stima puntuale per ϑ ma sfrutta tutta la distribuzione a posteriori, e in particolare tutti i campioni disponibili estratti da essa. La funzione che viene utilizzata per l'approssimazione della funzione (3.1) è proprio la funzione LPPD.

Per la correzione della sovrastima, il numero dei parametri può essere stimato come segue:

$$p_{WAIC} = 2\sum_{i=1}^{n} \left(\log \mathbb{E}_{post}\left(p\left(y_i \mid \vartheta\right)\right) - \mathbb{E}_{post}\left(\log p\left(y_i \mid \vartheta\right)\right)\right).$$

La formulazione del criterio è data quindi da:

$$\text{WAIC} = -2\left(\text{LPPD} - p_{WAIC}\right) = -2\left(\sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p\left(y_i \mid \vartheta^s\right)\right) - p_{WAIC}\right)$$

dove i valori ϑ^s sono i valori estratti dalla distribuzione a posteriori tramite un numero S di estrazioni.

Considerazioni generali

Un primo elemento da sottolineare è che il numero effettivo dei parametri calcolato tramite p_{DIC} e p_{WAIC} è diverso dal numero dei parametri reale quando le distribuzioni a priori dei parametri sono informative. Nel caso di distribuzioni a priori non informative e modelli semplici, come il modello lineare, è molto probabile che il numero stimato di parametri sia molto simile a quello reale. Nel caso di modelli invece con strutture gerarchiche o priori informative questo non succede.

Invece che le metriche p_{DIC} e p_{WAIC} introdotte in precedenza può essere utilizzato per entrambi i criteri anche un altro metodo per la stima del numero dei parametri: l'utilizzo della varianza della densità logaritmica a posteriori del modello. Per quanto riguarda il DIC, questo viene preferito rispetto alla metrica p_{DIC} quando quest'ultima fornisce risultati negativi. Nel calcolo del WAIC viene preferito spesso questo metodo rispetto alla metrica p_{WAIC} poichè fornisce risultati più stabili.

In un ambito di selezione bayesiano, tra questi tre criteri, quello con le caratteristiche più desiderabili è il WAIC, proprio per il suo approccio completamente bayesiano e l'utilizzo della distribuzioni a posteriori nella sua interezza, invece che delle stime puntuali.

3.3 Cross-validation

Una terzo approccio per la stima della funzione (3.1) è la *Cross-validation* (CV). Questo metodo prevede innanzitutto la divisione del dataset a disposizione in una parte di *training* e una parte di *test*. L'idea di base consiste nell'implementare il modello con i dati del *training* e quindi costruire la distribuzione a posteriori dei parametri solo con questi dati; in seguito viene valutato il modello considerando come campione nuovo i dati contenuti nel *test*. È intuibile che, maggiore è la numerosità della parte di *training*, maggiore è l'accuratezza del modello.

Ipotizzando che la distribuzione a posteriori sia rappresentata da S campioni estratti dalla stessa tramite simulazione, la densità predittiva, in scala logaritmica, per la parte di *test* è data da

$$\log p_{training}(y_{test}) = \log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_{test} \mid \vartheta)\right), \qquad \vartheta \sim p(\vartheta \mid y_{training}).$$

3.3.1 *Leave-one-out cross validation*

Il caso limite di questo approccio è il *Leave-one-out cross validation* (LOO-CV), in cui il dataset viene diviso in tante partizioni quante sono le osservazioni da cui è composto. Date n osservazioni, si avrà quindi che l'i-esima osservazione compone il *test set* mentre il dataset composto dalle n-1 restanti osservazioni è considerato come *training set*. L'inferenza del modello viene quindi fatta n volte con ogni volta una partizione di *training* e *test* differenti. Per n volte viene quindi stimata la quantità di interesse $p(y_i \mid y_{-i})$ e, ipotizzando che la distribuzione a posteriori sia rappresentata da S campioni estratti dalla stessa tramite simulazione, una stima della funzione (3.1) si ottiene come

$$\mathrm{LPPD}_{loo-cv} = \sum_{i=1}^{n} \log p_{post(-i)}(y_i), \text{calcolata come } \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p\left(y_i \mid \vartheta^{is}\right)\right).$$

Un primo svantaggio di questo approccio è che, idealmente, le partizioni in cui viene suddiviso il dataset dovrebbero essere tra loro indipendenti. Come è evidente nel caso limite qui presentato, ogni volta il *training* su cui viene stimato il modello cambia soltanto per una singola osservazione; questo comporta la quasi totale dipendenza tra le partizioni. Un secondo svantaggio, il principale, risiede nel costo computazionale dell'implementazione di tale metodo. È infatti

necessario fare inferenza sul modello, estrarre un campione di numerosità ampia dalla distribuzione a posteriori, n volte. Questo richiede una grande quantità di tempo, anche per valori di n non elevati.

Una possibile soluzione è l'utilizzo della *K-fold cross-validation* in cui il dataset viene suddiviso in un numero k di partizioni, minore del numero di osservazioni n. In questo modo la correlazione tra i diversi training diminuisce, così come il costo computazionale, anche se non di molto.

Una soluzione più rilevante per ridurre i tempi computazionali è l'utilizzo di metodi che riescano a calcolare l'approssimazione della quantità di interesse $p(y_i \mid y_{-i})$ senza dovere ogni volta stimare nuovamente il modello.

3.3.2 *Importance sampling*

Un'approssimazione della *leave-one-out cross validation* è possibile tramite l'utilizzo del campionamento di importanza. Secondo questo metodo è possibile approssimare la quantità di interesse $p(y_i \mid y_{-i})$ se le n osservazioni sono condizionatamente indipendenti tra loro e utilizzando S campioni ϑ dalla distribuzione a posteriori $p(\vartheta \mid y)$, tramite l'uso degli *importance ratios* definiti come:

$$r_i^s = \frac{1}{p(y_i \mid \vartheta^s)} \propto \frac{p(\vartheta^s \mid y_{-i})}{p(\vartheta^s \mid y)}.$$

La quantità di interesse $p(y_i \mid y_{-i})$ viene quindi stimata come:

$$p(y_i \mid y_{-i}) \approx \frac{\sum_{s=1}^{S} r_i^s p(y_i \mid \vartheta^s)}{\sum_{s=1}^{S} r_i^s} \approx \frac{1}{\frac{1}{S}\sum_{s=1}^{S} \frac{1}{p(y_i \mid \vartheta^s)}}. \tag{3.3}$$

La funzione (3.1) viene approssimata, tramite l'*importance sampling leave-one-out cross validation* (IS-LOO):

$$\mathrm{ELPD}_{\mathrm{IS-loo}} = \sum_{i=1}^{n} \log\left(\frac{\sum_{s=1}^{S} r_i^s p(y_i \mid \vartheta^s)}{\sum_{s=1}^{S} r_i^s}\right).$$

Il principale vantaggio dell'uso del campionamento di importanza risiede nella sostanziale riduzione del costo computazione dato che prevede un unico campionamento dalla distribuzione a posteriori. Nonostante questa caratteristica allettante, non viene quasi mai utilizzato questo metodo a causa di un suo grave svantaggio: gli *importance ratios* possono avere alta, addirittura infinita, varianza e questo provoca instabilità nello stimatore.

3.3.3 *Pareto Smoothed importance sampling*

Una versione dell'*IS-LOO* presentata in Vehtari et al. (2017) proposta per risolvere il problema dell'alta varianza degli *importance ratios* è il *Pareto Smoothed importance sampling leave one out cross validation* (PSIS-LOO).

L'idea di base di questo metodo è l'utilizzo della distribuzione di Pareto per "smussare" i valori più estremi della coda di destra degli *importance ratios*. In particolare il metodo qui presentato prevede i seguenti passi:

- modellare il 20% degli *importance ratios* più grandi con la distribuzione di Pareto;
- stabilizzare gli *importance ratios* sostituendo gli M ratios più elevati con i valori attesi della distribuzione di Pareto costruita, dove M è il numero di simulazioni usate per la costruzione della stessa. I nuovi *ratios* saranno quindi
$$\tilde{w}_i^s = F^{-1}\left(\frac{z-0.5}{M}\right), \quad z = 1, \ldots, M$$
dove F^{-1} è l'inversa della funzione di ripartizione della distribuzione di Pareto, s indicizza le simulazioni fatte e i indicizza l'unità statistica;
- troncare ogni *ratio* $\tilde{w}_i^s$ per garantire la stima con varianza finita e ottenere infine i pesi
$$w_i^s = S^{3/4}\bar{w}_i$$
dove $\bar{w}_i$ è la media dei pesi $\tilde{w}_i^s$ per tutti i valori di S.

Una volta ottenuti i nuovi pesi, l'approssimazione della funzione (3.1) è data da:

$$\mathrm{ELPD}_{\mathrm{PSIS-LOO}} = \sum_{i=1}^{n} \log\left(\frac{\sum_{s=1}^{S} w_i^s p\left(y_i \mid \vartheta^s\right)}{\sum_{s=1}^{S} w_i^s}\right).$$

La diagnostica per la stabilità della stima ottenuta viene svolta analizzando il parametro di forma $\hat{k}$ della distribuzione di Pareto. Per valori di k inferiori a 0.5, la varianza degli *importance ratios* è finita, vale quindi il teorema centrale del limite e le stime convergono. Per valori compresi tra 0.5 e 1, la varianza degli *importance ratios* può essere infinita ma la media esiste, per cui il teorema centrale del limite continua a valere e le stime convergono, anche se lentamente. Per valori superiori a 1 invece la varianza e la media degli *importance ratios* non esistono e quindi non si hanno stime affidabili. Empiricamente si è osservato che fino a

valori di k pari a 0.7 non si hanno problemi. Anche il PSIS-LOO presenta quindi della problematiche.

3.4 Stimatore Mixture di Silva e Zanella

In questa sezione verrà presentato un nuovo stimatore per la quantità $p(y_i \mid y_{-i})$ proposto nell'articolo Silva & Zanella (2022) da Giacomo Zanella e Luca Silva nel settembre 2022.

Lo stimatore in questione è stato proposto con un obiettivo preciso: risolvere i problemi presentati dalla *leave-one-out cross validation* e dagli altri stimatori di $p(y_i \mid y_{-i})$ presentati in precedenza in contesti ad alta dimensionalità. Essi presentano alta varianza e quindi risultati instabili e inaffidabili quando il numero di covariate nel modello è molto più elevato del numero delle osservazioni. Con il metodo proposto invece viene garantita una varianza finita dello stimatore e quindi una prestazione migliore in contesti multidimensionali. Inoltre il metodo proposto ha lo stesso costo computazionale previsto dagli altri metodi perché prevede anch'esso un unico campionamento dei parametri.

Definizione dello stimatore

Lo stimatore proposto è una misura della probabilità predittiva LOO $p(y_i \mid y_{-i})$. La prima differenza con gli altri stimatori illustrati è il fatto che i valori ϑ dei parametri considerati non provengono dalla classica distribuzione a posteriori $p(\vartheta \mid y)$ bensì da una nuova distribuzione q_{mix} data da:

$$q_{mix}(\vartheta, I) = \frac{p(\vartheta)\, p(y_{-I} \mid \vartheta))}{\sum_{j=1}^{n} p(y_{-j})} \qquad (\vartheta, I) \in \{1, \dots n\}. \tag{3.4}$$

La distribuzione è una distribuzione congiunta tra ϑ e I, che formalmente è una variabile casuale su $\{1, \dots, n\}$.

Si noti che la definizione di q_{mix} in questo modo è tale per cui vale l'uguaglianza $q_{mix}(\vartheta \mid I = i) = p(\vartheta \mid y_{-i})$ e inoltre permette di ricavare $p(y_i \mid y_{-i})$ come il seguente valore atteso:

$$p(y_i \mid y_{-i}) = \mathbb{E}_{(\vartheta, I) \sim q_{mix}} [p(y_i \mid \vartheta \mid I = i)].$$

Lo stimatore proposto si ricava con la seguente procedura:

- si ottiene un campione di numerosità S di $\vartheta_1, \vartheta_2, \ldots, \vartheta_S$ dalla distribuzione marginale di ϑ $q_{mix}(\vartheta)$ ricavata dalla distribuzione congiunta $q_{mix}(\vartheta, I)$:

$$q_{mix}(\vartheta) = \frac{\sum_{j=1}^{n} p(\vartheta)\, p\left(y_{-j} \mid \vartheta\right)}{\sum_{j=1}^{n} p\left(y_{-j}\right)} \propto p(\vartheta \mid y) \left(\sum_{j=1}^{n} p\left(y_j \mid \vartheta\right)^{-1} \right);$$

- per ogni osservazione $i \in \{1, \ldots, n\}$, si estraggono campioni pesati da $p(\vartheta \mid y_{-i})$ assegnando a ogni campione in $\{\vartheta_s\}_{s=1,\ldots,S}$ il peso

$$w_i^{(mix)}(\vartheta) = q_{mix}(I = i \mid \vartheta) = \frac{p(y_i \mid \vartheta)^{-1}}{\sum_{j=1}^{n} p\left(y_j \mid \vartheta\right)^{-1}},$$

  che è la probabilità condizionata di $I = i$ dato ϑ secondo la distribuzione congiunta $q_{mix}(\vartheta, I)$;

- infine per ogni osservazione $i \in \{1, \ldots, n\}$ si stima la probabilità $p(y_i \mid y_{-i})$ attraverso il nuovo stimatore

$$\hat{\mu}_i^{(mix)} = \frac{\sum_{s=1}^{S} p(y_i \mid \vartheta_s)\, w_i^{(mix)}(\vartheta_s)}{\sum_{s=1}^{S} w_i^{(mix)}(\vartheta_s)}. \tag{3.5}$$

Considerazioni sullo stimatore

Nell'articolo Silva & Zanella (2022) vengono sottolineate due peculiarità importanti dello stimatore. Prima di tutto viene mostrato come la distribuzione in (3.4) possa essere interpretata coma una mistura delle distribuzioni a posteriori LOO, formulandola quindi come $q_{mix} = \sum_{i=1}^{n} \pi_i p(\vartheta \mid y_{-i})$ con i pesi della mistura pari a $\pi_i = p(y_{-i}) \left(\sum_{j=1}^{n} p\left(y_{-j}\right)\right)^{-1}$, sotto le condizioni $\sum_{i=1}^{n} \pi_i = 1$ e $\pi_i \geqslant 0$. L'algoritmo si occupa di stimare effettivamente le quantità π_i per ogni componente della mistura. Sfruttando la riscrittura

$$\pi_i = \frac{\sum_j p\left(y_j \mid y_{-j}\right)}{p\left(y_i \mid y_{-i}\right)},$$

si può infatti esprimere la quantità di interesse come

$$p\left(y_i \mid y_{-i}\right) = \frac{\sum_j p\left(y_j \mid y_{-j}\right)}{\pi_i}.$$

Dalla stima di quest'ultima tramite la mistura ne deriva il miglioramento delle prestazioni dello stimatore in (3.5) rispetto allo stimatore (3.3) dato che i pesi della mistura sono tipicamente più facili da stimare e che sono per costruzione di valore massimo pari a 1, una caratteristica importante per la robustezza dello stimatore. Successivamente, nella terza sezione, viene dimostrato come la varianza asintotica dello stimatore, definita come:

$$AV_i^{(mix)} = \lim_{S\to\infty} S \operatorname{var}\left(\frac{\hat{\mu}_i^{(mix)}}{\mu_i}\right)$$

sia finita nel caso dello stimatore proposto, risultato fondamentale per l'obiettivo iniziale degli autori.

Si noti inoltre che lo stimatore prevede un solo campionamento dalla distribuzione $q_{mix}(\vartheta)$ per cui oltre ad avere caratteristiche più attraenti degli altri stimatori illustrati, non prevede aggiuntivi costi computazionali.

Implementazione

Per quanto riguarda l'implementazione pratica bisogna innanzitutto campionare i valori di ϑ della nuova distribuzione q_{MIX} utilizzando, come di consueto, la scala logaritmica:

$$\log q_{mix}(\vartheta) = \log p(\vartheta) + \sum_{i=1}^{n} \log p(y_i \mid \vartheta) + LSE\left(\{-\log p(y_i \mid \vartheta)\}_{i=1,\dots,n}\right) + const,$$

dove la funzione *LSE* è data da $LSE(\boldsymbol{x}) = \log\left(\sum_{i=1}^{n} \exp(x_i)\right)$. Una volta ottenuto il campione $\vartheta_1, \vartheta_2, \dots, \vartheta_S$ si calcolano gli stimatori per ogni unità statistica come segue:

- costruire la matrice $n \times S$ dei termini di log-verosimiglianza $l_{is} = \log p(y_i \mid \vartheta_s)$ per $i = 1, \dots, n$ e $s = 1, \dots, S$;
- costruire la matrice $n \times S$ di pesi (in scala logaritmica) definiti come $\tilde{w}_{is} = \log\left(w_i^{(mix)}(\vartheta_s)\right)$ e calcolati tramite l'uguaglianza $\tilde{w}_{is} = -l_{is} - \tilde{z}_s$ dove $\tilde{z}_s = LSE\left(-\{l_{is}\}_{i=1,\dots,n}\right)$, per $i = 1, \dots, n$ e $s = 1, \dots, S$;
- calcolare il logaritmo dello stimatore in (3.5) come $\log \hat{\mu}_i^{(mix)} = \tilde{z} - LSE\left(\{\tilde{w}_{is}\}_{s=1,\dots,S}\right)$ dove $\tilde{z} = LSE\left(\{z_s\}_{s=1,\dots,S}\right)$.

Il codice per l'implementazione qua descritta, insieme a quella degli altri criteri illustrati in questo capitolo, è mostrato nell'Appendice 4.2

Capitolo 4

Applicazione al dataset cholesterol

L'obiettivo di questa applicazione è utilizzare gli algoritmi e i criteri studiati in questa tesi per identificare, all'interno di una regressione polinomiale, il modello che meglio si adatta ai dati, ovvero il grado del polinomio che meglio ne permetta la spiegazione. L'applicazione è svolta sul dataset *cholesterol* sulla base dell'analisi sviluppata in Efron & Hastie (2013).

Il dataset comprende i dati relativi a uno studio della durata di 7 anni svolto su 164 pazienti sottoposti alla somministrazione della **Cholestyramine**, un farmaco per l'abbassamento del colesterolo. Lo studio è stato svolto con lo scopo di verificare se il farmaco in questione contribuisse alla diminuzione del colesterolo negli individui. Le due variabili considerate sono state:

- *cholesterol decrease*: differenza tra il livello del colesterolo a inizio dello studio e il livello a fine dello studio;
- *compliance*: dose del farmaco somministrata, standardizzata.

I modelli costruiti studiano la variazione della variabile risposta, la diminuzione del colesterolo, in funzione della dose del farmaco ricevuta.

In Figura 4.1 viene mostrato l'andamento dalla variabile risposta in funzione della covariata. Come si può notare la relazione tra le variabili non è di tipo lineare per cui non è possibile utilizzare una regressione lineare standard. Nell'analisi svolta si è provato quindi ad applicare una regressione polinomiale, utilizzando modelli per diversi gradi del polinomio, dal primo al sesto. Attraverso l'utilizzo dei criteri illustrati nel Capitolo 3 si è scelto come modello ottimale per l'adattamento ai dati, la regressione polinomiale di grado terzo, la cui retta stimata viene mostrata nella Figura 4.1.

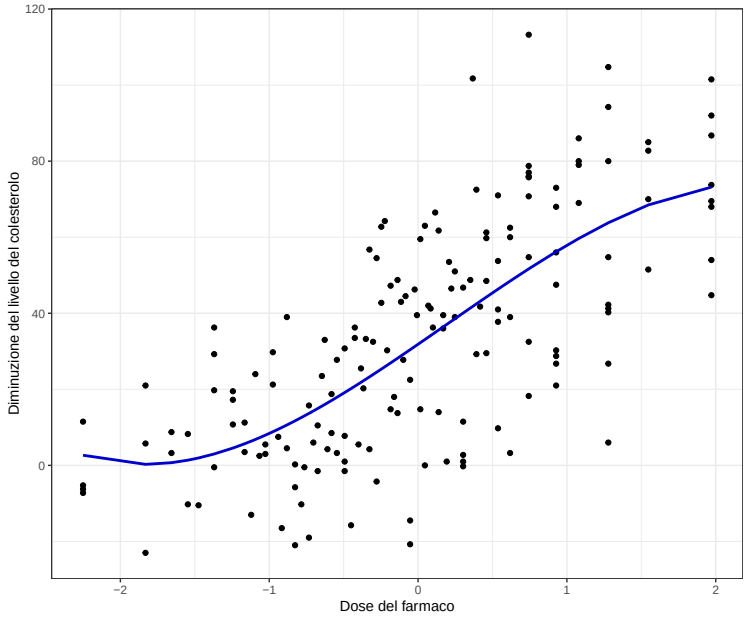

Figura 4.1: Regressione polinomiale di grado terzo stimata

4.1 Costruzione dei modelli

Per svolgere l'analisi di regressione polinomiale sono stati considerati 6 modelli, ognuno per ogni grado del polinomio, dal primo al sesto. Per ogni modello l'inferenza sui parametri è stata fatta utilizzando un approccio bayesiano, ovvero attraverso lo studio della distribuzione a posteriori dei parametri. Essa è stata approssimata tramite l'utilizzo degli algoritmi Gibbs Sampler e Metropolis-Hastings. Inoltre, per ogni modello è stata stimata anche la distribuzione q_{MIX} formalizzata in Silva & Zanella (2022) per poter calcolare il nuovo stimatore proposto nello stesso articolo. Di seguito vengono mostrati i risultati ottenuti nell'inferenza a posteriori soltanto per il modello polinomiale di grado terzo, in quanto rivelatosi il migliore alla fine dell'analisi.

Derivazione delle distribuzioni a posteriori

Per prima cosa sono state stimate le distribuzioni a posteriori dei parametri tramite l'utilizzo del Gibbs Sampler. Nella regressione polinomiale la stima dei parametri avviene nello stesso modo della regressione lineare multivariata per cui le *full conditionals distributions* utilizzate hanno la forma descritta nel Capitolo 2.

In Figura 4.2 vengono mostrati i grafici delle distribuzioni a posteriori per i coefficienti di regressione stimati dall'algoritmo con 100000 iterazioni. Affinché le stime convergano alla reale distribuzione a posteriori dei parametri, il *trace plot* della sequenza dei valori stimati deve presentare un andamento stazionario mentre il grafico della densità della distribuzione deve avere un andamento simile a quello di una distribuzione normale. Per tutte le variabili considerate, i grafici mostrati presentano in effetti un andamento stazionario per cui l'approssimazione della distribuzione è attendibile. Anche la varianza, di cui non viene mostrato il grafico, presenta una distribuzione approssimata stazionaria per cui anche la sua stima risulta affidabile.

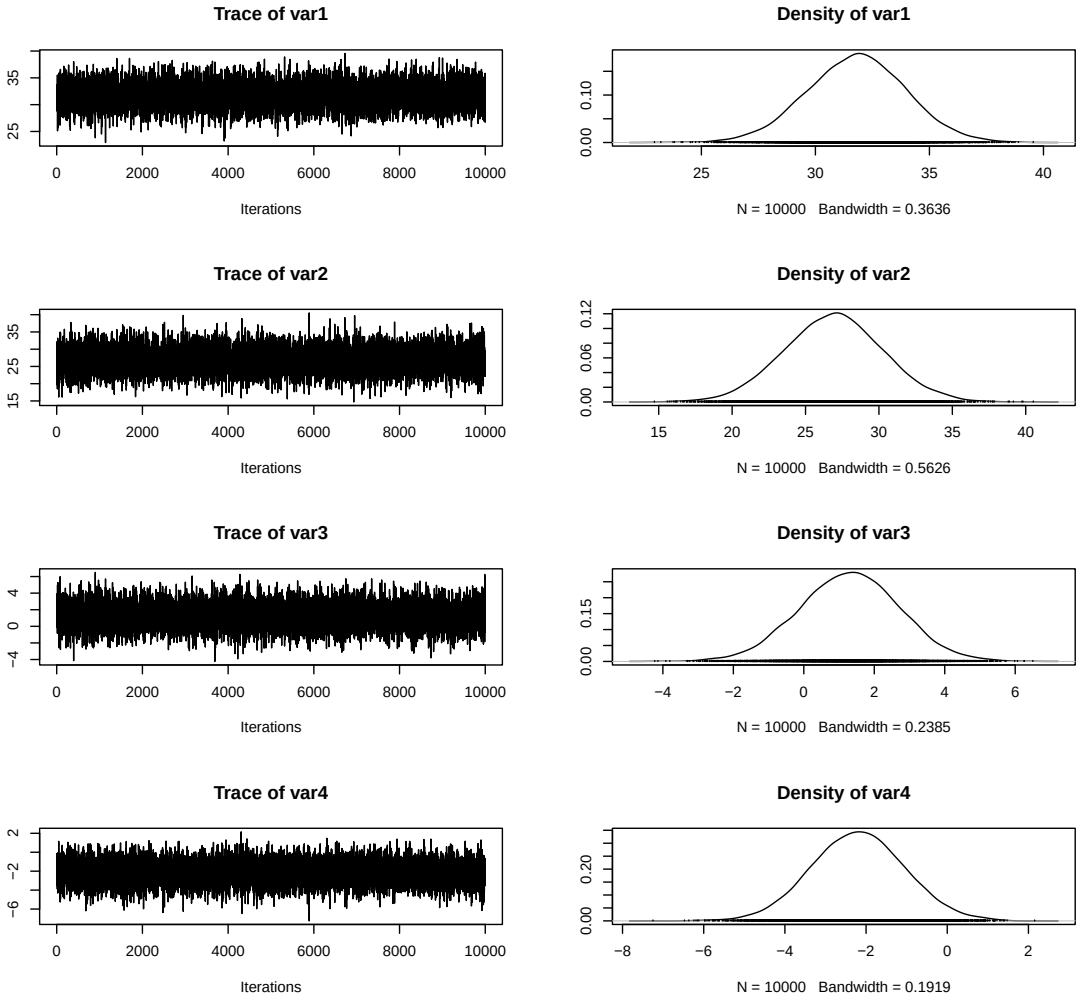

Figura 4.2: Trace plot e densità dei 4 coefficienti di regressione

I risultati ottenuti a posteriori tramite questo metodo di calcolo sono mostrati nella Tabella 4.1. Questa mostra la media a posteriori, lo *standard error* e

l'intervallo di credibilità al 95% dei parametri stimati attraverso il Gibbs Sampler.

	Media	**Standard Error**	**Intervallo di credibilità al** 95%
Intercetta	31.808	2.189	[27.523, 36.106]
x	26.89	3.370	[20.297, 33.557]
$\mathbf{x}^2$	1.290	1.432	[-1.51, 4.117]
$\mathbf{x}^3$	-2.173	1.152	[-4.449, 0.088]
$\boldsymbol{\sigma}^2$	486.134	55.453	[390.16, 604.307]

Tabella 4.1: Indici a posteriori dei parametri per il Gibbs Sampler

I coefficienti di regressione e la varianza sono stati inoltre stimati tramite l'utilizzo dell'algoritmo Metropolis-Hastings, utilizzando come *proposal distribution* una distribuzione normale multivariata e svolgendo anche in questo caso 100000 iterazioni. Anche in questo caso i *trace plot* e i grafici della densità mostrano un andamento stazionario e simile a quello nella Figura 4.2; la convergenza alla distribuzione di interesse sembra quindi essere garantita.

I risultati a posteriori sono presentati nella Tabella 4.2, la quale mostra la media a posteriori, lo *standard error* e l'intervallo di credibilità al 95% .

	Media	**Standard Error**	**Intervallo di credibilità al** 95%
Intercetta	31.852	2.179	[27.568, 36.095]
x	26.918	3.367	[20.331, 33.562]
$\mathbf{x}^2$	1.284	1.432	[-1.486, 4.116]
$\mathbf{x}^3$	-2.183	1.149	[-4.458, 0.068]
$\boldsymbol{\sigma}^2$	482.61	55.076	[386.542, 602.638]

Tabella 4.2: Indici a posteriori dei parametri per il Metropolis-Hastings

I risultati ottenuti tramite i due metodi di campionamento sono tra loro molto simili e gli intervalli di credibilità e lo *standard error* confermano un andamento standard delle stime, che possono essere ritenute affidabili in entrambi i casi.

Infine, tramite l'utilizzo del Metropolis-Hastings è stata derivata anche la distribuzione q_{MIX} per la costruzione dello stimatore proposto in Silva & Zanella (2022).

4.2 Confronto tra i modelli

Una volta costruiti, i modelli si sono confrontati utilizzando i criteri illustrati nel Capitolo 3. Per la costruzione dei modelli si è scelto di considerare l'approssimazione della distribuzione a posteriori fornita dal Metropolis-Hastings; sulla base di questa è stata valutata la capacità predittiva per ogni modello. In Tabella 4.3 vengono mostrati i risultati ottenuti dall'applicazione delle diverse metodologie per ogni grado del polinomio considerato.

Criteri di selezione	**Grado del polinomio**					
	1°grado	2°grado	3°grado	4°grado	5°grado	6°grado
AIC	1484.91	1485.5	1483.90	1485.84	1486.26	1488.26
DIC	1484.9	1485.44	1483.93	1485.83	1486.35	1487.88
WAIC	1484.42	1484.76	1482.89	1484.59	1484.89	1486.27
PSIS-LOO	1484.5	1485.0	1483.2	1485.1	1485.6	1487.1
Stimatore $\mu_i^{(\mathbf{mix})}$	1484.65	1485.01	1483.06	1485.19	1485.25	1487.47

Tabella 4.3: Confronto tra i vari gradi del polinomio

Per scegliere il modello migliore è necessario individuare quello che presenta i valori dei criteri inferiori. Come si può notare i risultati forniti da tutti i criteri sono tra loro molto simili e differiscono per poco. Tuttavia, il modello della regressione polinomiale di grado terzo presenta dei valori leggermente inferiori per tutte le metriche calcolate e risulta quindi essere il grado ottimale del polinomio per l'adattamento ai dati considerati.

Si nota che in particolare, AIC e DIC forniscono risultati che differiscono per cifre molto piccole e questo rispecchia la loro specificazione teorica. Essi infatti differiscono per il numero dei parametri e per la stima puntuale dei parametri che viene utilizzata. Tuttavia, in questo caso, sono state scelte delle distribuzioni a priori caratterizzate da indici di posizione e dispersione ottenuti tramite il metodo dei minimi quadrati. Quest'ultimo viene utilizzato nell'inferenza dei parametri in ambito frequentista e, nel caso della regressione lineare, esso coincide con il metodo della massima verosimiglianza. Da questo deriva che le stime di massima verosimiglianza utilizzate nell'AIC e le stime a posteriori, ricavate a partire da distribuzioni a priori caratterizzate da questi valori specifici, sono tra loro simili. Questo porta alla conseguente similitudine tra AIC e DIC. Se fossero state scelte delle distribuzioni a priori differenti o maggiormente informative probabilmente questa somiglianza non si sarebbe ottenuta.

E' utile inoltre sottolineare che lo stimatore $\mu_i^{(mix)}$ presenta dei valori conformi agli altri stimatori; questo dimostra che esso è uno stimatore affidabile nella previsione della capacità predittiva del modello e può essere utilizzato alternativamente agli altri. In questo caso non si nota il miglioramento delle prestazioni rispetto agli altri poiché l'analisi svolta è molto semplice e nessuno dei criteri utilizzati presenta delle problematiche. In analisi ad alta dimensionalità, con un numero di covariate molto superiore al numero di osservazioni, gli altri metodi avrebbero presentato dei problemi, ovvero instabilità e inaffidabilità numerica. In questi casi utilizzare il nuovo stimatore avrebbe portato a risultati più soddisfacenti in quanto sarebbe stato quello in grado di fornire risultati più affidabili. In questa analisi ci si è limitati soltanto a verificarne il funzionamento e la coerenza con le altre metodologie ma in futuro può essere utilizzato come una valida alternativa ai metodi classici. Inoltre, lo stimatore $\mu_i^{(mix)}$ e lo stimatore PSIS-LOO dovrebbero, almeno in questo semplice caso, coincidere; tuttavia questo non avviene per delle leggerissime discrepanze dovute all'errore Monte Carlo.

Interpretazione del modello

Il modello migliore risulta quindi essere il modello con il grado terzo del polinomio. In questo modello, utilizzando come stime puntuali le medie a posteriori ottenute tramite Metropolis-Hastings, la variabile risposta sarà ottenuta come:

$$\mathbb{E}[Y_i] = 31.852 + 26.918 * x_i + 1.284 * x_i^2 - 2.183 * x_i^3.$$

Avendo a disposizione la completa distribuzione a posteriori dei parametri, è possibile costruire la distribuzione a posteriori dei dati predittiva. Infatti, conoscendo i valori della media e varianza a posteriori che caratterizzano la distribuzione, è possibile, tramite simulazione, generare dei dati dal modello parametrico. In questo modo si avrà una distribuzione a posteriori predittiva dal modello. Nella Figura 4.3 viene mostrata la rappresentazione grafica della media a posteriori predittiva, con l'intervallo di credibilità dei valori al 95%.

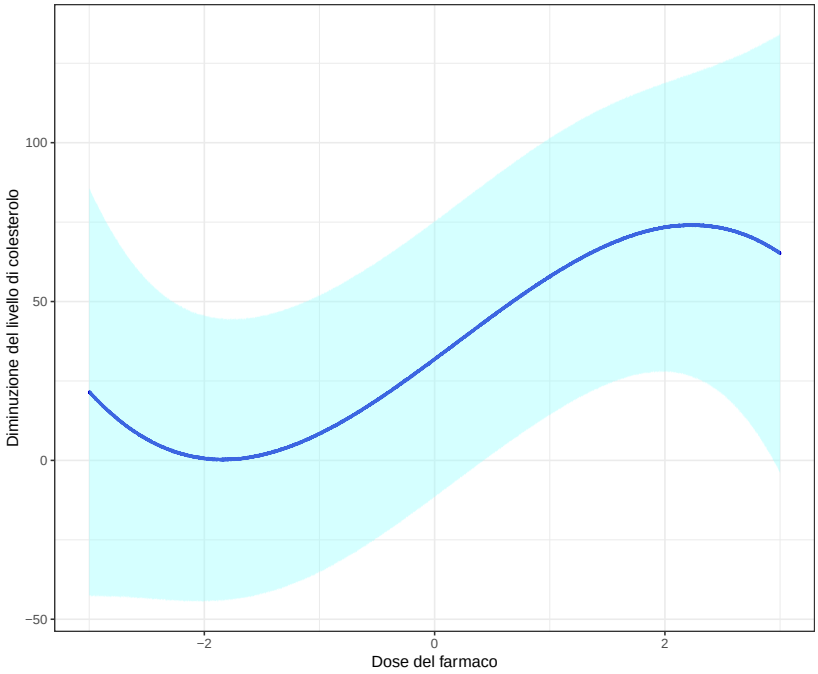

Figura 4.3: Distribuzione predittiva a posteriori

La distribuzione a posteriori predittiva è, in questo caso, un utile supporto ai medici e ai somministratori del farmaco **Cholestyramine**. Infatti tramite essa è possibile prevedere, in corrispondenza della dose del farmaco somministrata, quale sarà la diminuzione del livello di colesterolo che il paziente osserverà. Ricordando che i valori della dose del farmaco sono stati standardizzati, il valore 0 della dose del farmaco standardizzata rappresenta la somministrazione del valore medio della dose. Per un soggetto che ha assunto la dose media del farmaco la diminuzione del livello del colesterolo sarà, in media, pari a 31.852. Per qualsiasi valore di dose assunta del farmaco, viene prevista in media una diminuzione o un non aumento del livello del colesterolo. Tuttavia, per valori della dose somministrati molto bassi, gli intervalli di credibilità includono valori negativi della diminuzione del colesterolo; ciò significa che per valori bassi della è possibile che il livello di colesterolo aumenti. In particolare questo avviene per valori inferiori al livello medio della dose. Per valori leggermente superiori alla media (da valori superiori a 0.5 in termini di dose standardizzata) si osserva che gli intervalli di credibilità non includono valori negativi della diminuzione del livello di colesterolo; ciò significa, che superato questo limite di dose, si osserverà, con un livello di confidenza del 95%, una diminuzione del livello di colesterolo.

Codice R

```
################
###CAPITOLO 1###
################

#Analisi della sopravvivenza

#Caricamento dei dati e delle librerie necessarie
library(coda)
library(ggplot2)

library(survival)
dati<-stanford2
str(dati)
t<-dati$time
d<-dati$status

#Funzione di log-verosimiglianza
loglik<- function(t,d,gamma, beta){
  log_hazard<-sum(d*((gamma-1)*log(t)+log(gamma)-gamma*log(beta)))
  log_survival<- sum(-(t/beta)^gamma)
  log_hazard+log_survival
}

#Specificazione delle distribuzioni a priori

log_prior<-function(theta){
  sum(dnorm(theta, mean=0, sd=sqrt(100)))
}
```

```
#Specificazione della distribuzione a posteriori

log_posterior<-function(t, d,theta){
  loglik(t,d,exp(theta[1]), exp(theta[2]))+log_prior(theta)
}

#Algoritmo M-H per i valori a posteriori dei parametri

#Specificazione dei valori iniziali dei parametri
sigma2<-diag(x=c(0.25^2, 0.25^2), ncol=2)
theta<-c(0,0)

MH<-function(theta,sigma2, t,d,M, burn_in){

  log_p<-log_posterior(t,d,theta)
  th.matrix<-matrix(NA,nrow=M, ncol=2)

  for (i in 1:(M+burn_in)){

    prop<-rnorm(2, mean=theta, sd=0.25)  #campiono i valori dei
                                          parametri dalla proposal
                                          distribution
    log_prop<-log_posterior(t,d,prop)

    accept_rate<-min(1, exp(log_prop-log_p)) #calcolo del
                                             tasso di accettazione

    if (runif(1)<accept_rate){ #aggiornamento dei parametri
      theta<-prop
      log_p<-log_prop
    }

    if (i>burn_in){
      th.matrix[i-burn_in,]<-theta #escludo i valori che
                                   #fanno parte del periodo
```

```
                                  #burn_in
    }
  }
  th.matrix
}


stima<-as.mcmc(exp(MH(c(0,0),sigma2,t,d,50000,5000)))
stima_gamma<-stima[,1]
stima_beta<-stima[,2]
plot(stima) #grafici per la diagnostica

effectiveSize(stima) #ampiezza campionaria
1-rejectionRate(stima) #tasso di accettazione
autocorr(stima)

#Grafici delle due funzioni

#Stima della funzione di sopravvivenza
range(t) #0.5-3700
grid <- seq(0, 3700, length = 50)

S_mean <- numeric(length(grid))
S_upper <- numeric(length(grid))
S_lower <- numeric(length(grid))

for (i in 1:length(grid)) {
  S_mean[i] <- mean(pweibull(grid[i], shape = stima_gamma,
                             stima_beta, lower.tail = FALSE))
  S_lower[i] <- quantile(pweibull(grid[i], shape = stima_gamma,
                            stima_beta, lower.tail = FALSE), 0.025)
  S_upper[i] <- quantile(pweibull(grid[i], shape = stima_gamma,
                            stima_beta, lower.tail = FALSE), 0.975)
}

data_plot <- data.frame(Time = grid, Mean = S_mean, Upper = S_upper,
```

```
Lower = S_lower)

ggplot(data = data_plot, aes(x = Time, y = Mean, ymin = Lower,
  ymax = Upper)) +
  geom_line(col="blue") +
  theme_bw() +
  ylab("Funzione di sopravvivenza") +
  ylim(c(0, 1)) +
  geom_ribbon(alpha = 0.5, col="aquamarine", fill="aquamarine")

#Stima della funzione hazard
grid <- seq(0, 3700, length = 50)

H_mean <- numeric(length(grid))
H_upper <- numeric(length(grid))
H_lower <- numeric(length(grid))

for (i in 1:length(grid)) {
  H_mean[i] <- mean(((stima_gamma/stima_beta)*
                       (grid[i]/stima_beta)^(stima_gamma-1)))
  H_lower[i] <- quantile(((stima_gamma/stima_beta)*
                            (grid[i]/stima_beta)^(stima_gamma-1)), 0.025)
  H_upper[i] <- quantile(((stima_gamma/stima_beta)*
                            (grid[i]/stima_beta)^(stima_gamma-1)), 0.975)
}

data_plot <- data.frame(Time = grid, Mean = H_mean, Upper = H_upper,
Lower = H_lower)

ggplot(data = data_plot, aes(x = Time, y = Mean, ymin = Lower,
  ymax = Upper)) +
  geom_line(col="blue") +
  theme_bw() +
  ylab("Funzione Hazard")+
  geom_ribbon(alpha = 0.5, col="aquamarine", fill="aquamarine")
```

```
################
###CAPITOLO 3###
################
#Per la costruzione dei parametri è necessario avere la distribuzione
#a posteriori dei parametri

#Algoritmo Gibbs Sampler per la distribuzione a posteriori
# di un modello di regressione lineare multivariata
Gibbs_lin <- function(M, Y, X, beta_0, Sigma_0, v0, sigma2_0) {
  n <- nrow(X)
  out <- matrix(NA, nrow = M, ncol = length(beta_0) + 1)
  sigma2 <- as.numeric(var(Y))

  for (i in 1:M) {
    # campiono beta
    sigma.n <- solve((t(X) %*% X) / sigma2 + solve(Sigma_0))
    mean.n <- sigma.n %*% ((t(X) %*% Y) / sigma2 + solve(Sigma_0) %*% beta_0)

    beta <- rmvnorm(1, mean = mean.n, sigma = sigma.n)

    # campiono sigma_2
    SSR <- ((t(Y) %*% Y) + beta %*% t(X) %*% X %*% t(beta)
    - 2 * beta %*% t(X) %*% Y)
    sigma2 <- 1 / rgamma(1, shape = (n + v0) / 2,
    rate = (v0 * sigma2_0 + SSR) / 2)

    out[i, ] <- c(beta, sigma2)
  }
  out
}

#Algoritmo Metropolis-Hastings per la distribuzione a posteriori
# di un modello di regressione lineare multivariata

#Distribuzione a priori
logprior <- function(beta, sigma, beta_0, sigma_0, v0, sigma2_0) {
```

```
  dmvnorm(c(beta), mean = c(beta_0), sigma = sigma_0, log = TRUE) +
 dinvgamma(sigma^2, shape = v0 / 2, scale = v0 * sigma2_0 / 2, log = T)
}

#Funzione di log-verosimiglianza
log_lik <- function(n,Y,X,beta,sigma) {
  mu<- X %*% beta
  sum(dnorm(Y, mean=mu, sd=sigma, log=T))
}

#Distribuzione a posteriori
logpost <- function(n, Y, X, beta, varianza, beta_0, sigma_0,
v0, sigma2_0) {
  if(varianza <= 0) return(-Inf)
  log_lik(n,Y,X,beta,varianza) +
  logprior(beta, varianza, beta_0, sigma_0, v0, sigma2_0)
}

LSE <- function(x) {
  x_max <- max(x)
  x_max + log(sum(exp(x - x_max)))
}

#Distribuzione q_mix
logpost_qmix <- function(n, Y, X, beta, varianza, beta_0, sigma_0,
v0, sigma2_0) {
  if(varianza <= 0) return(-Inf)
  log_lik(n,Y,X,beta,varianza) +
  logprior(beta, varianza, beta_0, sigma_0, v0, sigma2_0) +
  LSE(-log_lik_s(Y, varianza, X, beta))
}

#Algoritmo M-H per la distribuzione a posteriori classica
MH <- function(beta, sigma2, n, Y, X, M, burn_in, beta_0, Sigma_0,
v0, sigma2_0) {
```

```
  log_p <- logpost(n, Y, X, beta, sqrt(sigma2), beta_0, Sigma_0,
  v0, sigma2_0)

  th.matrix <- matrix(NA, nrow = M, ncol = length(beta) + 1)

  for (i in 1:(M + burn_in)) {
    theta <- rmvnorm(1, c(beta, sigma2), sigmaP)

    beta_prop <- theta[1:ncol(X)]
    sigma_prop2 <- theta[ncol(X)+1]

    log_prop <- logpost(n, Y, X, beta_prop, sqrt(sigma_prop2), beta_0, Sigma_0,
    v0, sigma2_0)
    accept_rate <- min(1, exp(log_prop - log_p))
    if (runif(1) < accept_rate) {
      beta <- beta_prop
      sigma2 <- sigma_prop2
      log_p <- log_prop
    }
    if (i > burn_in) {
      th.matrix[i - burn_in, ] <- c(beta, sigma2)
    }
  }
  th.matrix
}

#Algoritmo M-H per la distribuzione q_mix
MH_qmix <- function(beta, sigma2, n, Y, X, M, burn_in, beta_0, Sigma_0,
v0, sigma2_0) {

  log_p <- logpost_qmix(n, Y, X, beta, sqrt(sigma2), beta_0, Sigma_0,
  v0, sigma2_0)

  th.matrix <- matrix(NA, nrow = M, ncol = length(beta)+1)

  for (i in 1:(M + burn_in)) {
```

```
    theta <- rmvnorm(1, c(beta, sigma2), sigmaP)

    beta_prop <- theta[1:ncol(X)]
    sigma_prop2 <- theta[ncol(X)+1]

    log_prop <- logpost_qmix(n, Y, X, beta_prop, sqrt(sigma_prop2), beta_0,
    Sigma_0, v0, sigma2_0)
    accept_rate <- min(1, exp(log_prop - log_p))
    if (runif(1) < accept_rate) {
      beta <- beta_prop
      sigma2 <- sigma_prop2

      log_p <- log_prop
    }
    if (i > burn_in) {
      th.matrix[i - burn_in, ] <- c(beta,sigma2)
    }
  }
  th.matrix
}

#Criteri di informazione

AIC<-function(n,Y,X,beta,sigma){
  -2*log_lik(n,Y, X, beta, sigma)+2*(length(beta)+1)

}

DIC<-function(fit, n, Y,X){

  coeff_bayes<-colMeans(fit)[1:dim(fit)[2]-1]
  sigma_bayes<-sqrt(colMeans(fit)[dim(fit)[2]])
  log_p_bayes<-log_lik(n,Y, X, coeff_bayes, sigma_bayes)

  somma<-as.numeric(0)
```

```
  for (i in 1:dim(fit)[1]){
    add<-log_lik(n, Y, X, fit[i,1:dim(fit)[2]-1],sqrt(fit[i,dim(fit)[2]]))
    somma<-somma+add
  }

  valore<-somma/dim(fit)[1]

  p_dic<-2*(log_p_bayes -valore)

  DIC<- -2*log_p_bayes+2*p_dic
  DIC
}

#WAIC

lppd<-function(fit, X, Y){

  somma_j<-as.numeric(0)

  for (i in 1:dim(fit)[1]){
    beta<-fit[i,1:dim(fit)[2]-1]
    sigma<-fit[i,dim(fit)[2]]
    mu<-X%*%beta
    add_i<-dnorm(Y,mu, sqrt(sigma))
    somma_j<-somma_j+add_i
  }

  add_j<-log(somma_j/dim(fit)[1])
  add_j
}


media<-function(fit, X, Y){

  somma_j<-as.numeric(0)
```

```
  for (i in 1:dim(fit)[1]){

    beta<-fit[i,1:dim(fit)[2]-1]
    sigma<-fit[i,dim(fit)[2]]
    mu<-X%*%beta

    add_i<-dnorm(Y,mu, sqrt(sigma), log=T)
    somma_j<-somma_j+add_i
  }
  media_j<-somma_j/dim(fit)[1]
  media_j
}


p_waic<-function(fit, X, Y){
  2*sum(lppd(fit, X, Y)-media(fit,X,Y))
}


WAIC<-function(fit, X, Y){
  p<-p_waic(fit, X, Y)
  lppd_tot<-sum(lppd(fit, X, Y))
  -2*(lppd_tot-p)
}

#Leave-One-Out cross validation

#EXACT LOO-CV

lppd_loocv<-function(fit, X, Y){

  somma_j<-as.numeric(0)

  for (i in 1:dim(fit)[1]){
    beta<-fit[i,1:9]
    sigma2<-fit[i,10]
```

```
    mu<-X%*%beta
    add_i<-dnorm(Y,mu, sqrt(sigma2))
    somma_j<-somma_j+add_i
  }

  add_j<-log(somma_j/dim(fit)[1])
  add_j
}

somma_i<-as.numeric(0)
for (i in 1:nrow(X)){
  X_i<-X[-i,]
  Y_i<-Y[-i]

  Sigma_0<-solve((t(X_i)%*%X_i)/nrow(X_i)*as.numeric(var(Y_i)))
  beta_0<-solve(t(X_i)%*%X_i)%*%t(X_i)%*%Y_i
  v0<-1
  sigma2_0<-(t(Y_i)%*%Y_i-2%*%t(beta_0)%*%t(X_i)%*%Y_i+t(beta_0)
  %*%t(X_i)%*%X_i%*%beta_0)/(nrow(X_i)-length(beta_0))

  fit_i<-as.mcmc(Gibbs_lin(M,Y_i,X_i,beta_0, Sigma_0, v0, sigma2_0))
  add_i<-lppd_loocv(fit_i, X[i,],Y[i])
  add_i

}

#PSIS-LOO

library(loo)
M<-10000
loglik_matrix<-matrix(rep(0,n*M),ncol=n, nrow=M)
for (i in 1:dim(fit)[1]){
  beta<-fit[i,1:dim(fit)[2]-1]
  sigma2<-fit[i,dim(fit)[2]]
  mu<-X%*%beta
  loglik_matrix[i,]<-dnorm(Y,mu, sqrt(sigma2), log=T)
```

```
}
loo(loglik_matrix)

#Stimatore Mixture proposto da Zanella e Silva

S <- 10000
theta_matrix <- theta_qmix

#Matrice di log-verosimiglianza
loglik_matrix <- matrix(rep(0, n * S), nrow = n, ncol = S)
for (i in 1:dim(theta_matrix)[1]) {
  beta <- theta_matrix[i,1:dim(fit)[2] -1]
  sigma2<-theta_matrix[i,dim(fit)[2] ]
  mu <- X%*%beta
  loglik_matrix[,i] <- dnorm(Y, mu, sqrt(sigma2), log = T)
}

#Matrice dei pesi
wi_s <- matrix(rep(0, n * S), ncol = S, nrow = n)
z_s <- vector()

for (s in 1:S) {
  z_s[s] <- LSE(-loglik_matrix[, s])
}
z_s

for (s in 1:S) {
  for (i in 1:n) {
    wi_s[i, s] <- -loglik_matrix[i, s] - z_s[s]
  }
}

#Stimatore
second <- c()
for (i in 1:n) {
```

```
  second[i] <- LSE(wi_s[i, ])
}
ztilde <- LSE(-z_s)
lppd_qmix<-log_mu_mix <- ztilde - second
sum(lppd_qmix)*-2

################
###CAPITOLO 4###
################

library(ggplot2)
library(coda)
library(mvtnorm)
library(invgamma)

#Caricamento del dataset
chol<-read.table("cholesterol.txt", header=T)

#Costruzione dei dataset per svolgere le
#sei regressioni

library(tidyverse)
chol<-chol%>%as_tibble()

dati1<-chol
dati2<-chol%>%
  mutate(x2=compliance^2)
dati3<-chol%>%
  mutate(x2=compliance^2, x3=compliance^3)
dati4<-chol%>%
  mutate(x2=compliance^2, x3=compliance^3, x4=compliance^4)
dati5<-chol%>%
  mutate(x2=compliance^2, x3=compliance^3, x4=compliance^4,
  x5=compliance^5)
dati6<-chol%>%
  mutate(x2=compliance^2, x3=compliance^3, x4=compliance^4,
```

```
  x5=compliance^5,x6=compliance^6)

n <- nrow(chol)

#Esempio per il polinomio di grado terzo

Y<-dati3$cholesterol.decrease
X<-model.matrix(cholesterol.decrease~., data=dati3)
n<-nrow(X)
Sigma_0 <- solve((t(X) %*% X) / (nrow(X) * as.numeric(var(Y))))
beta_0 <- solve(t(X) %*% X) %*% t(X) %*% Y
v0 <- 1
sigma2_0 <- (t(Y) %*% Y - 2 %*% t(beta_0) %*% t(X) %*% Y + t(beta_0)
%*% t(X) %*% X %*% beta_0) / (nrow(X) - length(beta_0))



mod3<-lm(cholesterol.decrease~., data=dati3)
sum3<-summary(mod3)

gib3<-as.mcmc(Gibbs_lin(100000, Y, X, beta_0, Sigma_0, v0, sigma2_0))
plot(gib3)
cov(gib3)

sigmaP <- vcov(mod3)
z <- c(rep(0, ncol(X)))
sigmaP <- rbind(sigmaP, z)
sigmaP <- cbind(sigmaP, c(rep(0, ncol(X)),3058.3655077))
sigmaP <- 2.38^2 * sigmaP / (ncol(X)+1)

beta <- beta_0
sigma2 <- sum3$sigma^2

theta_post3 <- as.mcmc(MH(beta,sigma2, n, Y, X, 100000, 5000, beta_0,
Sigma_0,v0, sigma2_0))
acf(theta_post3)
plot(theta_post3)
```

```
colMeans(theta_post3)

beta_mle<-coef(mod3)
sigma_mle<-sum3$sigma

AIC_1(n,Y,X,beta_mle, sigma_mle )

DIC(theta_post3,n,Y,X)

WAIC(theta_post3,X,Y)

#PSIS-LOO

M<-100000
loglik_matrix<-matrix(rep(0,n*M),ncol=n, nrow=M)
fit<-theta_post3
for (i in 1:dim(fit)[1]){
  beta<-fit[i,1:dim(fit)[2]-1]
  sigma2<-fit[i,dim(fit)[2]]
  media<-X%*%beta
  loglik_matrix[i,]<-dnorm(Y,media, sqrt(sigma2), log=T)

}
loo(loglik_matrix)

theta_qmix3<-as.mcmc(MH_qmix(beta, sigma2, n, Y, X, 100000, 5000,
beta_0, Sigma_0, v0, sigma2_0))
plot(theta_qmix3)

#Stimatore Mixture
S <- 100000
theta_matrix <- theta_qmix3
loglik_matrix <- matrix(rep(0, n * S), nrow = n, ncol = S)
for (i in 1:dim(theta_matrix)[1]) {
  beta <- theta_matrix[i,1:dim(theta_matrix)[2] -1]
```

```
  sigma2<-theta_matrix[i,dim(theta_matrix)[2] ]
  media <- X%*%beta
  loglik_matrix[,i] <- dnorm(Y, media, sqrt(sigma2), log = T)
}

wi_s <- matrix(rep(0, n * S), ncol = S, nrow = n)
z_s <- vector()

for (s in 1:S) {
  z_s[s] <- LSE(-loglik_matrix[, s])
}

for (s in 1:S) {
  for (i in 1:n) {
    wi_s[i, s] <- -loglik_matrix[i, s] - z_s[s]
  }
}

second <- c()
for (i in 1:n) {
  second[i] <- LSE(wi_s[i, ])
}
ztilde <- LSE(-z_s)
lppd_qmix<-log_mu_mix <- ztilde - second
sum(lppd_qmix)*-2

#Costruzione del grafico tramite i valori a posteriori

beta<-colMeans(theta_post3)[-5]
scarto<-sqrt(colMeans(theta_post3)[5])
y_stimati<-beta[1]+beta[2]*chol$compliance+beta[3]*chol$compliance^2
+ beta[4]*chol$compliance^3

ggplot(chol, aes(x=compliance, y=cholesterol.decrease)) +
  geom_point( col="black") +
   geom_line(aes(x=compliance, y=y_stimati), col="blue3", lwd=1)+
```

```
  labs(x="Dose del farmaco",
       y="Diminuzione del livello del colesterolo")+
  theme_bw()

#Distribuzione predittiva
grid<-seq(-3,3, by=0.001)
beta<-theta_post3[,-5]
scarto<-sqrt(theta_post3[,5])

F_mean <- numeric(length(grid))
F_upper <- numeric(length(grid))
F_lower <- numeric(length(grid))

for (i in 1:length(grid)) {
  F_mean[i] <- mean(rnorm(nrow(beta), mean=beta[,1]+beta[,2]*grid[i]
    +beta[,3]*grid[i]^2+beta[,4]*grid[i]^3, sd=scarto), lower.tail = FALSE)
  F_lower[i] <- quantile(rnorm(nrow(beta), mean=beta[,1]+beta[,2]*grid[i]
  + beta[,3]*grid[i]^2+beta[,4]*grid[i]^3, sd=scarto), lower.tail = FALSE, 0.025)
  F_upper[i] <-quantile(rnorm(nrow(beta), mean=beta[,1]+beta[,2]*grid[i]
  + beta[,3]*grid[i]^2+beta[,4]*grid[i]^3, sd=scarto), lower.tail = FALSE, 0.975)
}

data_plot <- data.frame(Dose = grid, Mean = F_mean, Upper = F_upper,
Lower = F_lower)

ggplot(data = data_plot, aes(x = Dose, y = Mean, ymin = Lower, ymax = Upper)) +
  geom_line(col="blue3", lwd=1) +
  theme_bw() +
  xlab("Dose del farmaco")+
  ylab("Diminuzione del livello di colesterolo") +
  geom_ribbon(alpha = 0.4, fill="darkslategray1")
```

Ringraziamenti

Per prima cosa vorrei ringraziare il professor Rigon per la disponibilità e il supporto che mi ha dato sia durante il percorso di stage che durante il percorso di tesi.

Ringrazio la mia famiglia, in particolare la mia mamma, per avermi sempre incoraggiato durante i miei percorsi di studio.

Ringrazio i miei compagni e amici di università per aver reso questi tre anni molto più belli e divertenti di quanto mi sarei mai aspettata. In particolare, ringrazio Giorgia per avermi aiutato nella preparazione di tutti gli esami e della tesi e per essermi stata accanto nei momenti più belli, ma soprattutto in quelli più neri.

Ringrazio le mie amiche di sempre per avermi sempre supportato e sopportato (e per ascoltare sempre tutte le mie lamentele!).

Ringrazio Stefano per essermi sempre stato vicino in questi ultimi anni, quando più ne avevo bisogno.

Bibliografia

Efron, B. & Hastie, T. (2013). Computer age statistical inference.

Escobar, L. A. & Meeker Jr, W. Q. (1992). Assessing influence in regression analysis with censored data. *Biometrics* , 507–528.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*, vol. 580. Springer.

Robert, C. P. & Casella, G. (1999). *Monte Carlo statistical methods*, vol. 2. Springer.

Robert, C. P. & Casella, G. (2010). *Introducing monte carlo methods with r*, vol. 18. Springer.

Silva, L. & Zanella, G. (2022). Robust leave-one-out cross-validation for high-dimensional bayesian models. *arXiv preprint arXiv:2209.09190* .

Vehtari, A., Gelman, A. & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing* **27**, 1413–1432.