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

Statistica per le Tecnologie e le Scienze

STUDIO STATISTICO SU SOGGETTI AFFETTI DA DEMENZA SENILE: MODELLI DI SOPRAVVIVENZA PER DATI SULLA VITA RESIDUA

Relatore: prof.ssa Giuliana Cortese
Dipartimento di Scienze Statistiche

Laureanda: Rebecca Esegio
Matricola n. 2011389

Anno Accademico 2023/2024

Indice

Introduzione

I dati di sopravvivenza descrivono il periodo di tempo che intercorre tra un evento iniziale ed un evento finale. Negli studi epidemiologici, l'evento finale potrebbe coincidere con la ricaduta della malattia o il decesso causato da quest'ultima. Ad esempio, i dati sulla sopravvivenza vengono spesso utilizzati per descrivere la durata del tempo dall'esordio di una malattia alla morte o l'intervallo di tempo dal trattamento alla ricaduta. Spesso, in tali studi, non si rileva l'evento di interesse entro la finestra temporale dello studio per tutti i soggetti. Nel contesto medico, ciò può verificarsi perché è prevista la conclusione dello studio a causa di vincoli pratici, come limitazioni di budget, o perché alcuni pazienti vengono persi al follow-up, ovvero si perde il contatto con questi ultimi per una serie di possibili ragioni, solitamente intrinseche al soggetto stesso ma esterne allo studio. Di conseguenza, i dati di sopravvivenza sono caratterizzati dalla cosiddetta censura a destra. In questo caso è noto solamente che il tempo di attesa all'evento è maggiore del tempo di censura, ma non se ne conosce il valore preciso. Ad esempio, se uno studio termina prima che per un paziente si verifichi l'evento (decesso) a causa di una certa malattia, i ricercatori possono solo essere sicuri che questo paziente sopravvive fino alla fine dello studio, cioè la data in cui il soggetto viene censurato a destra, ma il tempo esatto di sopravvivenza rimane igno-

to. In questo elaborato, ci si riferisce alla censura a destra dovuta alla conclusione di uno studio come censura amministrativa per differenziarla dalla censura a destra per perdita al follow-up.

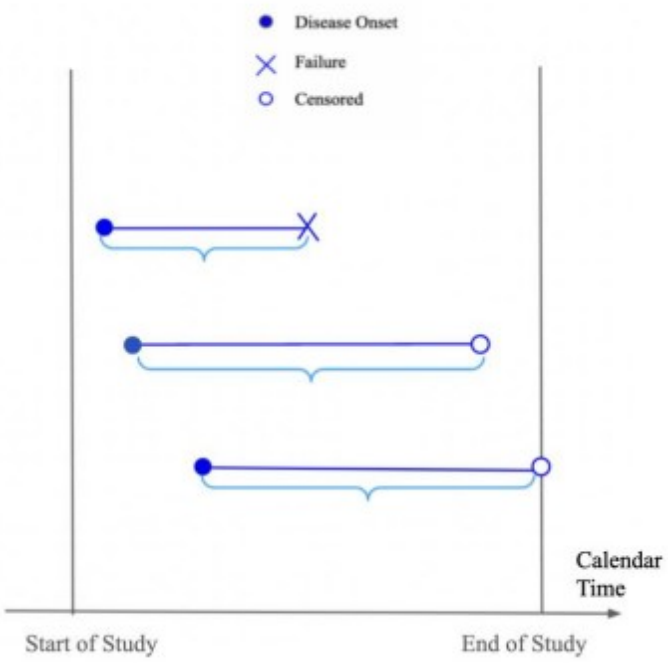

Figura 1: Studio di coorte incidente con tre ipotetici pazienti: l'evento viene osservato all'interno della finestra di studio per il primo soggetto, per il quale il tempo di sopravvivenza osservato coincide quindi con il tempo all'evento; per gli altri due soggetti si osserva censura a destra, dovuta alla perdita del paziente al follow-up (secondo soggetto) o al termine dello studio (terzo soggetto).

La *Figura 1* mostra uno schema di quello che viene definito uno studio di coorte incidente. Ad esempio, si suppone che l'interesse della ricerca sia il tempo di sopravvivenza legato ad una determinata malattia dalla diagnosi di quest'ultima. I pazienti reclutati in uno studio di coorte incidente non hanno

ancora ricevuto la diagnosi. Pertanto, vengono osservati sia il tempo di inizio che il tempo di censura dei soggetti. Tuttavia, sebbene gli studi di coorte incidenti forniscano dati precisi provenienti dalla popolazione target di interesse, sono spesso dispendiosi in termini di tempo e denaro.
A confronto, gli studi di coorte prevalenti sono più economici e pratici, il che li rende una pratica comune per la raccolta dei dati. Questi studi sono caratterizzati dall'identificazione e dalla recluta di individui che, all'inizio dello studio, hanno già ricevuto la diagnosi della malattia evento di interesse.

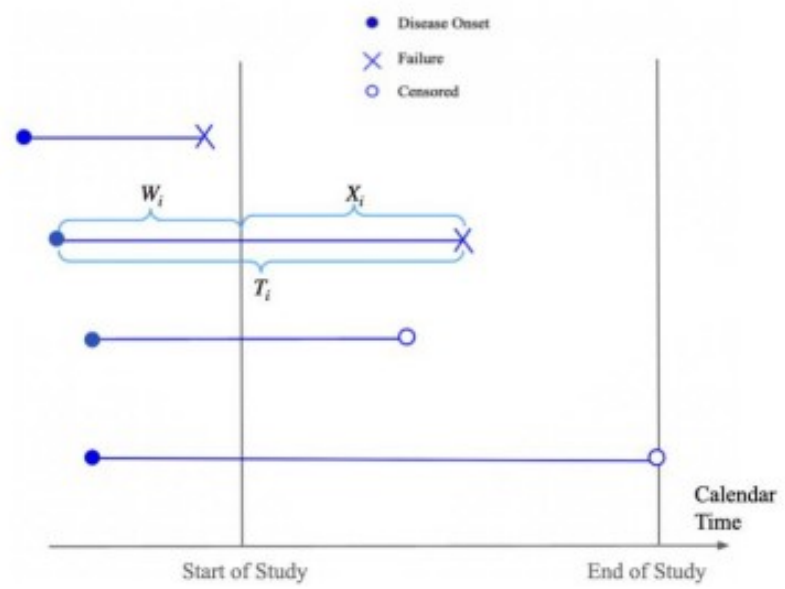

Figura 2: Studio di coorte prevalente con quattro ipotetici pazienti. Vengono reclutati all'interno dello studio il secondo, il terzo ed il quarto individuo, in quanto per il primo soggetto l'evento di interesse si verifica prima dell'inizio dell'esperimento. Solo il tempo di ricorrenza in avanti X_i può essere osservato direttamente per i soggetti reclutati. Per quanto riguarda il secondo individuo si osserva l'evento di interesse prima del termine dello studio, mentre per il terzo e quarto individuo si verifica una censura a destra.

L'intero tempo di sopravvivenza viene indicato come T_i ed è composto dal tempo di ricorrenza all'indietro W_i (*backward recurrence time*) ed il tempo di ricorrenza in avanti X_i, risultante da: $T_i - W_i$ (*forward recurrence time*).
Nella letteratura correlata, alcuni autori si riferiscono al *backward recurrence time* come alla durata corrente e al *forward recurrence time* come alla *vita residua*. Poiché il tempo di ricorrenza all'indietro si verifica prima dell'inizio dello studio, è possibile osservare solo il tempo di ricorrenza in avanti. Indicato come C_i il tempo di censura, il tempo effettivamente osservato dall'inizio dello studio fino alla censura è $X_i = min(C_i, T_i - W_i)$, che è troncato a sinistra e possibilmente censurato a destra. Il termine *troncamento a sinistra* si riferisce ai pazienti (ad esempio, il primo individuo nella *Figura 2*) per i quali si verifica l'evento prima ancora dell'inizio dello studio e, a causa dello schema adottato nel campionamento, non entrano a far parte del campione. Questo introduce un'ulteriore aspetto di uno studio di coorte prevalente: i soggetti che osserviamo non costituiscono un campione casuale della popolazione target, ma piuttosto tendono ad avere tempi di sopravvivenza più lunghi. Un soggetto è osservabile in uno studio di coorte prevalente solo se sopravvive abbastanza a lungo per entrare a far parte dello studio stesso.

Poiché tutti i soggetti reclutati negli studi di coorte prevalenti hanno già sperimentato l'evento iniziale, tali studi richiedono meno tempo e risorse finanziarie rispetto a quelli incidenti. Idealmente, per poter risalire alla data di inizio dell'evento si potrebbe intervistare i soggetti stessi; in pratica, tuttavia, i tentativi di ricordare quando potrebbe essersi verificato l'esordio possono essere inaffidabili.

Questa difficoltà pratica dei dati troncati a sinistra ottenuti da uno studio di coorte prevalente fornisce la motivazione per utilizzare solo i tempi di vita residua X_i, eventualmente censurati, al fine di stimare la distribuzione di sopravvivenza dei soggetti nello studio. Per stimare quindi la curva di sopravvivenza di interesse di T_i si fa ricorso alla relazione di quest'ultima con la funzione di densità del *forward recurrence time*.

Capitolo 1

Metodi per l'analisi

In uno studio di coorte prevalente con follow-up, un approccio per rimuovere qualsiasi potenziale influenza dall'incertezza nella misurazione delle date di insorgenza reali è attraverso l'utilizzo delle sole vite residue. Poiché quest'ultime vengono misurate da una data di screening ben definita (giorno di prevalenza) fino al fallimento/censura, queste durate temporali osservate sono essenzialmente prive di errori.
Indipendentemente dai dati a disposizione, per stimare la funzione di sopravvivenza, si può adottare un approccio parametrico o non parametrico. Utilizzando i dati sulla vita residua, si presenta lo stimatore non parametrico di massima verosimiglianza (NPMLE) e si propone uno stimatore "*stacked*", risultante dalla combinazione lineare di singoli stimatori di funzioni di sopravvivenza parametrici e non, con pesi "*stacking*" ottimali ottenuti minimizzando la funzione di punteggio Brier.

1.1 Funzioni probabilistiche fondamentali

La variabile aleatoria T, definita come tempo di attesa all'evento, è una variabile: non negativa, continua con supporto $[0, \infty)$ ed ha distribuzione di probabilità che può essere equivalentemente specificata tramite la sua funzione di sopravvivenza, di densità o di rischio istantaneo (*funzione hazard*).

1.1.1 Funzione di sopravvivenza

La funzione di sopravvivenza è definita come la probabilità che un evento non si verifichi prima di un certo istante di tempo t. Essendo T la variabile casuale che rappresenta il tempo di attesa all'evento di interesse, allora la funzione di sopravvivenza, indicata comunemente con S(t), è definita come:

$$S(t) = P(T > t) \tag{1.1}$$

Si tratta di una funzione non crescente, tale che

$$S(0) = 1$$

e

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

Definiamo inoltre il valore atteso della variabile aleatoria T, indicato con E[T] = μ, che rappresenta la media dei tempi di attesa:

$$E[T] = \int_0^\infty t f(t)\, dt$$

1.1.2 Funzione di densità

Per quanto riguarda la funzione di densità per la stessa variabile T, essa viene definita come:

$$f(t) = dF(t)/dt \tag{1.2}$$

Quest'ultima rappresenta la densità della probabilità che un evento si verifichi in un determinato intervallo di tempo $[t,t+dt]$, dato che il soggetto è sopravvissuto fino a t e si ha che:

$$f(t) \geq 0$$

per ogni x e

$$\int_0^\infty f(t)\,dt = 1$$

Essa si lega alla funzione di sopravvivenza in quanto:

$$f(t) = -dS(t)/dt$$

e

$$S(t) = \int_0^\infty f(t)\,dt$$

1.1.3 Funzione di rischio instantaneo o *hazard*

La funzione di rischio o *hazard* è definita come la probabilità istantanea condizionata che l'evento si verifichi nell'istante di tempo t, dato che un individuo è sopravvissuto fino a tale istante ed è definita come:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{1.3}$$

La funzione di rischio si ottiene anche come rapporto tra densità e funzione di sopravvivenza, ossia:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{dlogS(t)}{dt}$$

Inoltre integrando rispetto a t, si ha:

$$S(t) = exp[-\int_0^t h(s)\, ds]$$

Si definisce, in aggiunta, la funzione rischio cumulato che fornisce una stima del rischio cumulato di un evento nel tempo e può essere specificata matematicamente come l'integrale della funzione di rischio istantaneo nel tempo:

$$H(t) = \int_0^t h(s)\, ds \tag{1.4}$$

Essa si lega alla funzione di sopravvivenza tramite l'equazione:

$$H(t) = -log[S(t)]$$

1.2 Assunzioni del modello

Il metodo utilizzato si basa sul rapporto tra la funzione di densità del *forward recurrence time* $(T_i - W_i)$ e la distribuzione di sopravvivenza (o target), S.
Questa relazione si basa su due assunzioni:

- Censura casuale (e non informativa): si presenta nel momento in cui ogni soggetto ha un tempo di censura stocasticamente indipendente dal momento della suo decesso. Il tempo di sopravvivenza osservato è il minimo tra il momento della censura e quello della morte e la distribuzione dei tempi di censura non dipende dai parametri di interesse della verosimiglianza.
  Tramite questa assunzione si può affermare che la conoscenza del tempo censura non fornisce alcuna informazione circa il fatto che l'evento possa verificarsi in maniera imminente. Per un dato censurato si sa solo che il tempo di sopravvivenza è maggiore del tempo di censura osservato.

- Processo di incidenza stazionaria: si presuppone che il tasso di insorgenza della malattia nella popolazione sia costante. Come conseguenza dell'assunzione di stazionarietà, la distribuzione dei dati che presentano troncamento è uniforme e le distribuzioni del *forward recurrence time* e del *backward recurrence time* sono identiche.

1.3 Relazione tra la distribuzione della vita residua e la distribuzione del tempo di sopravvivenza

Supponiamo di denotare la funzione di densità del *forward recurrence time*, ovvero il tempo che intercorre dall'inizio dello studio alla censura, con f_{fwrd}.
Quindi, sotto le ipotesi avanzate precedentemente, è noto il Teorema 1 (*A Comparison of Stacking Methods to Estimate Survival Using Residual Lifetime Data, 2022*) sulla relazione

tra la densità f_{fwrd} e la curva di sopravvivenza target S, di T_i, dove μ= E[T]:

$$f_{fwrd}(t) = \frac{S(t)}{\mu} \tag{1.5}$$

La seguente dimostrazione utilizza la distribuzione distorta dalla lunghezza di T, $\tilde{\mathrm{T}}$. L'uso di $\tilde{\mathrm{T}}$ riflette che i dati di sopravvivenza raccolti negli studi di coorte prevalenti sono distorti dalla lunghezza, cioè, è più probabile che gli individui con un tempo di sopravvivenza più lungo vengano reclutati nello studio.
Indichiamo con $\tilde{\mathrm{T}} = \tilde{\mathrm{W}} + \tilde{\mathrm{X}}$ la distribuzione distorta dalla lunghezza di T. La probabilità di osservare $\tilde{\mathrm{T}}$ è proporzionale a T_i . Con tassi di incidenza stazionari, W segue una distribuzione uniforme f_W, con una corrispondente funzione di densità cumulativa F_W. Dato che T ha una funzione di densità f_T ed un tempo di sopravvivenza medio μ= E[T]:

$$f_{T,W}(t, w) = \frac{f(t)}{\int_0^\infty S(u)\, du} \tag{1.6}$$

Essendo l'integrale al denominatore uguale a μ, segue che:

$$f_{T,W}(t, w) = \frac{f(t)}{\mu} \tag{1.7}$$

$$f_W(w) = \frac{S(w)}{\mu} \tag{1.8}$$

Sotto l'assunzione di stazionarietà i *forward e backward recurrence times* presentano la stessa distribuzione:

$$f_T(t) = \frac{f(t)F_W(t)}{\int_0^\infty f(w)S(u)du} \tag{1.9}$$

Inoltre in virtù della stessa ipotesi, si ha $F_W(t) = tf_W(t)$, che conduce all'equazione:

$$f_T(t) = \frac{tf(t)}{\mu} \tag{1.10}$$

Derivando la distribuzione della *forward recurrence time* f_{fwrd}, si ottiene:

$$f_{fwrd}(x) = \frac{S(x)}{\mu} \tag{1.11}$$

,

la quale implica:

$$f_{fwrd}(0) = \frac{S(0)}{\mu} = \frac{1}{\mu} \tag{1.12}$$

Pertanto, dall'equazione (1.5), $f_{fwrd}(t) = S(t)f_{fwrd}(0)$, lo stimatore naturale di S risulta:

$$\hat{S}(t) = \frac{\hat{f}_{fwrd}(t)}{\hat{f}_{fwrd}(0)} \tag{1.13}$$

L'equazione (1.5) viene utilizzata per costruire le funzioni di verosimilgianza utilizzate nelle stime parametriche, mentre

l'equazione (1.13) viene utilizzata nelle stime non parametriche.

1.4 Stimatore Parametrico

Si utilizza lo stimatore di massima verosimiglianza per modelli parametrici. Nello specifico, sia θ il vettore dei parametri della distribuzione parametrica assunta. Lo stimatore di massima verosimiglianza $\hat{\theta}$ è ottenuto massimizzando la funzione di verosimiglianza:

$$\mathrm{L}(\theta) = \prod_{i=1}^{n} f_{fwrd}^{\delta_i}(x_1) S_{fwrd}^{1-\delta_i}(x_i) \tag{1.14}$$

Sostituendo l'equazione (1.5) si può riscrivere la funzione di verosimiglianza:

$$\mathrm{L}(\theta) = \prod_{i=1}^{n} \frac{S(X_i;\theta)^{\delta_i} \int_{z>x_i} S(z;\theta)^{1-\delta_i}}{\mu(\theta)} \tag{1.15}$$

Uno stimatore parametrico $\hat{S}$ per la funzione di sopravvivenza S risulta quindi essere:

$$\hat{S}(t) = S(t;\hat{\theta}) \tag{1.16}$$

1.5 Stimatore Non Parametrico

Considerando che i modelli parametrici possono avere ipotesi non specificate, si può supporre che uno stimatore non parametrico fornisca un'opzione più solida. Poiché si utilizzano

dati realtivi al *forward recurrence time*, un approccio standard non parametrico per i dati censurati a destra (ovvero, la curva di Kaplan-Meier applicata al *forward recurrence time*) ignorerebbe la struttura speciale suggerita dall'equazione (1.5). In particolare, tale equazione implica che la densità del *forward recurrence time* è una funzione non crescente, fatto che può essere sfruttato nella procedura di stima.

Gli studiosi Denby e Vardi hanno proposto un algoritmo per determinare lo stimatore non parametrico di massima verosimiglianza (NPMLE) di una funzione di densità non crescente quando si utilizzano dati potenzialmente censurati a destra (*A Comparison of Stacking Methods to Estimate Survival Using Residual Lifetime Data, 2022*). Quando l'osservazione più grande viene censurata, la funzione di verosimiglianza nella (1.14) ha solo un valore di massimo relativo, ma nessun massimo assoluto. Pertanto, Denby e Vardi hanno proposto la stima della massima verosimiglianza (MLE) della densità limitata a M, dove tutta la massa di probabilità rimanente è collocata in un tempo estremamente grande, ossia M.
Denotiamo con D_m l'insieme tutte le funzioni di densità continue a sinistra non crescenti con supporto (0, M]. L'MLE limitata a M è:

$$max_{g \in D_M} = L(g|data) \tag{1.17}$$

Tuttavia, questo NPMLE con limite M presenta un bias locale vicino a zero. Si considera che questo bias si presenta anche in assenza di censura, dove le proprietà asintotiche del NPMLE sotto il vincolo di densità decrescente sono state sta-

bilite. Questo porta ad ipotizzare che questo fenomeno possa essere il risultato di un diverso tasso di convergenza vicino a zero.
L'equazione (1.13) rivela che lo stimatore proposto dipende da $\hat{f}_{fwrd}(0)$, ed è quindi necessario correggere questo bias locale, motivo per cui viene utilizzato lo stimatore di Denby-Vardi corretto.

1.6 Procedure di impilamento

Supponendo che ci siano m modelli presi in considerazione nell'analisi, si calcolano prima i corrispondenti stimatori delle funzioni di sopravvivenza, utilizzando la procedura di massima verosimiglianza per i modelli parametrici e lo stimatore di Denby-Vardi nel caso non parametrico.
Si propone quindi uno stimatore per bilanciare i pro e i contro dei diversi modelli combinandoli linearmente, ottenendo:

$$\hat{S}_{fwrd}(x) = \sum_{k=1}^{m} a_k \hat{S}_{k,fwrd}(x)$$

I pesi per la combinazione lineare sono ottenuti tramite l'algoritmo proposto da Wey, Connet e Rudser (*A Comparison of Stacking Methods to Estimate Survival Using Residual Lifetime Data, 2022*), in cui vengono minimizzati gli errori quadratici della funzione di sopravvivenza, misurati dal punteggio Brier. Si descrive in seguito questa procedura.

1.6.1 Il punteggio Brier

L'errore quadratico per le funzioni di sopravvivenza in un dato momento t viene misurato dal *punteggio Brier*. In assenza di censura si ha:

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}(Z_i(t) - \hat{S}(t))^2$$

dove $Z_i(t) = 1(t_i > t)$ e t_i è il tempo dell'evento per l'i-esima osservazione.
Per valutare la prestazione dello stimatore al tempo t, se l'evento dell'i-esima osservazione non si verifica entro l'istante t, allora la quantità $1 - \hat{S}(t) = \hat{F}(t)$ contribuisce al punteggio Brier. In questo caso, viene penalizzata la più piccola probabilità di sopravvivenza stimata al tempo t, $\hat{S}(t)$. Al contrario, se l'evento dell'osservazione i-esima si è verificato entro il tempo t, viene penalizzato un valore di $\hat{S}(t)$ maggiore
.
Tuttavia, poiché l'evento potrebbe non essersi osservato prima di t, vengono utilizzati i *pesi di probabilità inversa di censura* (IPCW) per tenere conto della probabilità che un'osservazione venga censurata.

Sia t_i il tempo dell'evento per l'osservazione i-esima, c_i il tempo di censura, G la funzione di sopravvivenza della distribuzione del tempo di censura, $T_i(t) = min(t_i, t)$ e $\Delta_i(t) = 1_{T_i<c_i}$, si può definire:

$$IPCW - IPCW(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i(t)}{G(T_i(t))}(Z_i(t) - \hat{S}(t))^2$$

dove il peso di un'osservazione non censurata dipende dal fatto che l'evento si verifichi entro il tempo t, le osservazioni censurate con $c_i > t$ contribuiscono a IPCW-IPCW(t) e le osservazioni censurate con $c_i < t$ contribuiscono solo indirettamente, attraverso la stima della distribuzione del tempo di censura.

Per ottimizzare i pesi $\hat{a}_1, ..., \hat{a}_m$ per combinare gli m modelli, il punteggio Brier dell'IPCW viene minimizzato su una serie di punti temporali, sotto i vincoli $\hat{a}_k > 0$ e $\sum_{k=1}^{m} a_k = 1$. Ovvero:

$$\hat{a} = argmin_\alpha \sum_{r=1}^{s}\sum_{i=1}^{n}\frac{\Delta_i(t)}{G(T_i(t))}(Z_i(t) - \sum_{k=1}^{m} a_k \hat{S}_k^{-i}(t))^2$$

dove $a = (a_1, a_2, ..., a_m)$ e $\hat{S}_k^{-i}$ è la sopravvivenza stimata del modello k tralasciando l'iesima osservazione, risultante dalla convalida incrociata. In pratica, viene utilizzata la convalida incrociata a 5 per facilitare l'efficienza computazionale.

Si può dimostrare che esiste un insieme di pesi, tale che il modello "*stacked*" ha un rendimento almeno pari a quello del miglior candidato disponibile in termini di errori quadratici medi tra gli m modelli presi in considerazione.
Infine, lo stimatore di sopravvivenza "*stacked*" del *forward*

reccurence time ottenuto in questo modo garantisce che la corrispondente funzione di densità abbia una densità non crescente, poiché una funzione di densità è la derivata negativa di una curva di sopravvivenza:

$$f_{fwrd}(x) = -\frac{d}{dx} S_{fwrd}(x) = \sum_{k=1}^{m} a_k f_{k,fwrd}(x)$$

ottenendo così una combinazione lineare di funzioni di densità non crescenti per i dati temporali del *forward reccurence time*.
Si può quindi ottenere uno stimatore *"stacked"* per la funzione di sopravvivenza S nel seguente modo, tenendo conto di m diversi modelli, che nel caso preso in esame successivamente saranno 5 (NPMLE, *Weibull, Log-Logistic, Log-Normale; Gamma*):

$$\hat{S}(t) = \sum_{k=1}^{m} \hat{a}_k \hat{S}_k(t)$$

Capitolo 2

Studio di simulazione

Utilizzando i dati di vita residua simulati censurati a destra, si valuta lo stimatore di impilamento *("stacked")*. Si esaminano le prestazioni dei singoli stimatori parametrici e non parametrici rispetto allo stimatore "*stacked*", quando i dati sulla vita residua sono soggetti a proporzioni crescenti di censura amministrativa. L'obiettivo è quello di valutare il vantaggio crescente, al diminuire del follow-up, dell'utilizzo di uno stimatore di "*stacking*" sia con il modello di massima verosimiglianza non parametrica corretto (NPMLE) che con le funzioni di sopravvivenza parametriche.

Per simulare una serie di dati sulla vita residua censurati a destra, si genera prima una data di inizio, O, da una distribuzione Uniforme con supporto (0, 50). Successivamente si genera un tempo di fallimento, T, da una distribuzione Weibull con parametri di forma e scala entrambi pari a 2. Con la metodologia proposta, non si considera l'inclusione di covariate all'interno del modello. Si campionano le coppie di tempi di insorgenza e fallimento (O_i , T_i), per cui $T_i >$ (50 - O_i), fino ad un campione di dimensione n. Casi con $T_i <$ (50 – O_i) vengono filtrati per rappresentare gli individui non

osservabili in uno studio di coorte prevalente, il cui evento di interesse si verifica prima dell'inizio dello studio. I tempi campionati, per i=1, 2, ..., n , sono stati quindi censurati a destra da una costante C, corrispondente alla censura amministrativa. Per generare un set di dati con tasso di censura amministrativa, q, il tempo di censura C^* risulta pari a 1 - q^{th} quantile di X (*forward recurrence time*).

Nella serie di simulazioni, si ipotizza che i tempi di censura siano distribuiti secondo una distribuzione Weibull e che i tempi di vita residua siano censurati amministrativamente spostando le date di fine studio per risultare, rispettivamente, al 10%, 40%, o 60% di censura.
Per ciascuna percentuale di censura, si adattano i modelli *NPMLE, Weibull, Log-Logistic, Log-Normale* e *Gamma* corretti.
Utilizzando tutti e cinque i sottomodelli, si determinano i pesi "*stacking*" ottimali e si calcolano gli errori DISSE utilizzati per confrontare la funzione di sopravvivenza stimata dal modello $\hat{S}(t_j)$ con la funzione di sopravvivenza empirica basata sui dati $S_0(t_j)$ per i vari modelli, quando stimati separatamente e quando combinati in un modello "*stacked*".

$$DISSE = \sum_{j=1}^{k} (t_j - t_{j-1})(\hat{S}(t_j) - S_0(t_j))^2$$

Il DISSE è la versione discretizzata dell'integrale dato da: $\int_0^\infty (\hat{S}(t) - S(t))^2 \, dt$. Esso viene calcolato nell'intervallo di tempo (0-10) ed il limite superiore del supporto viene fissato a 50 poiché le funzioni di sopravvivenza sottostanti nelle simu-

lazioni Weibull presentano una probabilità trascurabile oltre quel punto.
Si considera inoltre un secondo modello "*stacked*" che include l'NPMLE corretto e tutti i modelli parametrici tranne il Weibull (ovvero il vero modello di generazione dei dati). Si utilizzano campioni di dimensione 125 (ovvero 125 vite residue osservate) in 100 esecuzioni di simulazione.

Si riportano, di seguito, nella *Tabella 2.1* i pesi medi per il modello "*stacked*" che include tutti i sottomodelli e per uno che include tutti i sottomodelli tranne il Weibull (modello di generazione dei dati).

Tabella 2.1: Pesi medi per il modello stacked che include tutti i sottomodelli(prima riga per ogni modello indicato) e per uno che include tutti i sottomodelli tranne il Weibull (seconda riga per ogni modello indicato).

Proporzione di censura amministrativa			
Modelli	10%	40%	60%
NPMLE	0.0034928	0.0000214	0.0000293
	0.0033900	0.0000437	0.0000314
Weibull	0.8385529	0.8058903	0.7368124
	NA	NA	NA
Log-Logistic	0.0299094	0.0270967	0.0346384
	0.0615676	0.1485210	0.1280891
Log-Normale	0.0236943	0.0162095	0.0300002
	0.0441654	0.037526	0.0500009
Gamma	0.1043503	0.1508033	0.1985489
	0.8908768	0.7469221	0.8219107

I pesi medi per un modello "*stacked*" combinano le previsioni dei modelli di base assegnando loro dei pesi, i quali vengono solitamente conferiti in base alle prestazioni dei modelli stes-

si. Utilizzando il modello "*stacked*", si nota che quasi tutto il peso viene spostato dal modello NPMLE corretto alla distribuzione di Weibull, la quale presenta pesi maggiori per ogni percentuale di censura amministrativa.
Mentre, quando il modello Weibull è escluso dal modello "*stacked*", la maggior parte del peso è stato spostato al modello Gamma, con pesi medi nettamente superiori rispetto agli altri modelli.

Si riportano ora gli errori DISSE medi in *Tabella 2.2* per i cinque modelli adattati (*NPMLE, Weibull, Log-Logistic, Log-Normale* e *Gamma*), per un modello *stacked* che include tutti i sottomodelli ed uno che include tutti i sottomodelli tranne il modello Weibull (modello di generazione dei dati).

Tabella 2.2: DISSE medi per i cinque modelli adattati, il modello stacked che include tutti i sottomodelli e per uno che include tutti i sottomodelli delli tranne il Weibull (modello di generazione dei dati).

	Proporzione di censura amministrativa		
Modelli	10%	40%	60%
NPMLE	0.1076062	0.3733248	0.4378117
Weibull	0.0259302	0.08215753	0.0938027
Log-Logistic	0.0362179	0.07589733	0.0892650
Log-Normale	0.0334567	0.06605067	0.0733694
Gamma	0.0291386	0.07192574	0.0772642
Modello "Stacked" (tutti i modelli compresi)	0.0278312	0.0834681	0.0947039
Modello "Stacked" (senza Weibull)	0.0385673	0.0910352	0.1012375

Dalla *Tabella 2.2* si osserva che, come ci si aspetterebbe, i valori degli errori DISSE crescono con l'aumentare della percentuale di censura amministrativa, (sia per i singoli modelli sia per i due modelli "*stacked*") in quanto quest'ultima intro-

duce un tipo di mancanza di dati che influenza la precisione delle stime. Quando si presenta una censura amministrativa significativa nei dati, il modello può riscontrare delle difficoltà nel fare previsioni accurate per gli elementi censurati e per tutto il range di dati, influenzando in tal modo gli errori DISSE.

All'aumentare della percentuale di censura amministrativa si osserva che, sebbene gli errori DISSE dei modelli individuali e "*stacked*" aumentino tutti, il DISSE medio per il modello non parametrico di massima verosimiglianza (NPMLE) aumenta ad un ritmo molto più rapido rispetto a quello dei modelli parametrici individuali e dei due modelli "*stacked*". Questo accade poiché la censura amministrativa riduce il periodo di follow-up e la gamma dei dati osservati sulla durata di vita residua, influenzando così in modo più grave lo stimatore non parametrico di massima verosimiglianza. L'NPMLE si esibisce male poiché la censura amministrativa aumenta la sua incapacità per prevedere qualsiasi probabilità di sopravvivenza potenzialmente diversa da zero dopo la fine dello studio, cioè dopo la censura amministrativa.

Inizialmente l'obiettivo era quello di migliorare l'NPMLE corretto quando si stima la funzione di sopravvivenza utilizzando solo le vite residue osservate da uno studio di coorte prevalente con follow up. Ciò attraverso l'introduzione di modelli parametrici in uno stimatore "*stacked*", pur mantenendo l'NPMLE corretto, ipotizzando che i modelli parametrici attenuassero una carenza dell'NPMLE corretto: il fallimento dello stimatore nel catturare il comportamento della coda della funzione di sopravvivenza, in particolare quando il periodo di

follow-up è breve. Tuttavia, si è mostrato che, confrontando gli stimatori, utilizzando essenzialmente i loro DISSE medi, l'NPMLE corretto non funziona in maniera soddisfacente né da solo né come membro dello stack all'aumento della percentuale di censura amministrativa.

Capitolo 3

Applicazione

Si procede quindi con l'applicazione dello stimatore "*stacked*" proposto, utilizzandolo per stimare la sopravvivenza servendosi dei dati raccolti nel centro di Channing House.
Channing House è un centro per anziani di Palo Alto, in California, i cui dati sono stati raccolti tra l'apertura della clinica nel 1964 ed il 1° luglio 1975, per i cui individui è stata registrata l'età al momento dell'ingresso e al momento dell'uscita dalla casa di cura o del decesso.

Il dataset è presente all'interno del pacchetto *boot* del software *R* [*RStudio*] e nell'analisi vengono cosiderate le variabili:

- *entry*: variabile relativa all'età del residente (in mesi) al momento dell'ingresso nel centro.
- *exit*: variabile relativa all'età del residente (in mesi) al momento del decesso, della partenza dal centro o del 1° luglio 1975, a seconda di quale evento si sia verificato per primo.
- *time*: variabile relativa al periodo di tempo (in mesi) che il residente ha trascorso a Channing House.

- *status*: variabile indicatore dell'evento. Vale 1 se il residente è morto a Channing House, mentre se vale 0 indica che il soggetto ha lasciato la casa prima del 1° luglio 1975 o che era ancora vivo e viveva nel centro in quella data.

Questi dati presentano censura a destra e sono caratterizzati da un troncamento a sinistra in quanto i residenti sono entrati a Channing House ad età diverse, la loro vita non è stata osservata prima del loro ingresso e non vengono, dunque, rilevate le osservazioni su individui il cui decesso è antecedente all'inizio dello studio. Pertanto, solo gli individui che hanno vissuto abbastanza a lungo da entrare nel centro di pensionamento saranno inclusi nei dati.
Nell'analisi i tempi in mesi relativi alle variabili dell'età e del tempo vengono convertiti in anni.

3.0.1 Verifica dell'ipotesi di stazionarietà

Si attesta inizialmente la validità dell'ipotesi di stazionarietà, precedentemente delineata nel primo capitolo, in quanto assunzione necessaria per la succesiva costruzione dei modelli.
In letteratura sono stati proposti due approcci per verificare l'ipotesi di stazionarietà: una valutazione grafica ed una procedura analitica.
In particolare, il professore e studioso M. Asgharian ha dimostrato che l'ipotesi di stazionarietà può essere verificata graficamente confrontando gli stimatori di Kaplan-Meier basati sui *backward* e *forward recurrence times* (Statisticsin-Medicine, M.Asgharian, D. B. Wolfson, e X. Zhang (2006)). Un'ampia discrepanza di quest'ultimi indica che l'ipotesi di stazionarietà non è valida.
Gli studiosi V. Addona e D. Wolfson hanno proposto un test

analitico per verificare tale ipotesi (LifetimeDataAnalysis,V. Addona e D. B. Wolfson (2006)). Essi hanno dimostrato che constatare l'ipotesi di stazionarietà equivale a constatare che le distribuzioni del *backward reccurence time* e *forward reccurence time* sono le stesse. Entrambi gli approcci proposti vengono implementati mediante l'utilizzo delle apposite funzioni presenti nel pacchetto *CoxPhLb* del software R.

Risulta quindi possibile esplorare graficamente l'ipotesi di stazionarietà utilizzando i vettori corrispondenti al *backward recurrence time*, al *forward recurrence time* e la variabile indicatrice dell'evento, generando un grafico con le due curve di Kaplan-Meier corrispondenti, che viene riportato di seguito.

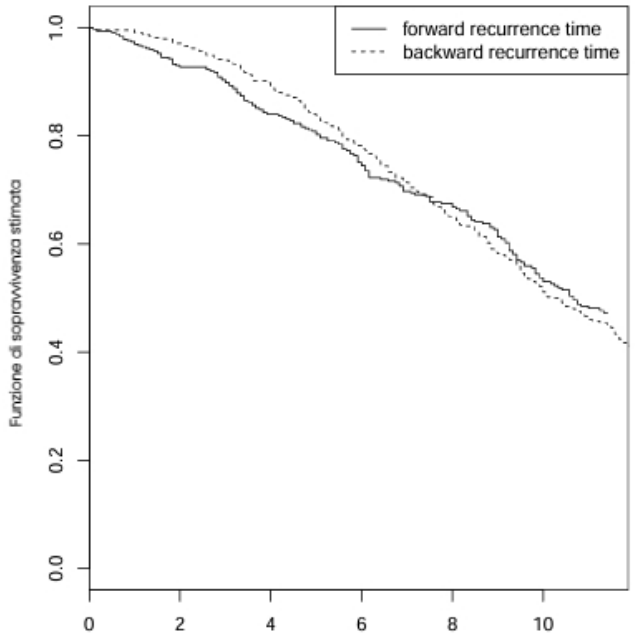

Figura 3.1: Verifica grafica dell'ipotesi di stazionarietà per i dati di Channing House.

Dal grafico risultante nella *Figura 3.1* risulta evidente che l'ipotesi di stazionarietà viene soddisfatta per i dati presi in analisi, in quanto le due curve non mostrano differenze significative tra di loro.
Si esamina ulteriormente l'ipotesi di stazionarietà conducendone il test analitico. Sulla base dei risultati, possiamo concludere che la stazionarietà del processo di incidenza è ragionevole per i dati presi in esame con un *p-value* pari a 0.794 (livello di significatività fissato $\alpha = 0.05$).

3.0.2 Verifica del modello "*stacked*"

Lo stimatore "*stacked*" proposto include 5 modelli: NPMLE, Weibull, Log-Normale, Log-Logistic e Gamma.

Nella *Figura 3.2* sono riportate le curve di sopravvivenza stimate per il modello "*stacked*", raffigurate in rosso, e per il modello non parametrico di massima verosimiglianza (NPMLE), raffigurate in nero, con i relativi limiti di confidenza puntuali al 95%.
Dal grafico, si nota che la stima per il modello NPMLE cattura solo in minima parte la forma della stima per il modello "*stacked*" e non risulta in grado di delineare il comportamento della funzione di sopravvivenza per i dati in analisi. Al contrario, il modello "*stacked*", prendendo in cosiderazione non solo il modello NPMLE ma anche altri quattro modelli parametrici, cattura interamente il comportamento della curva di sopravvivenza.

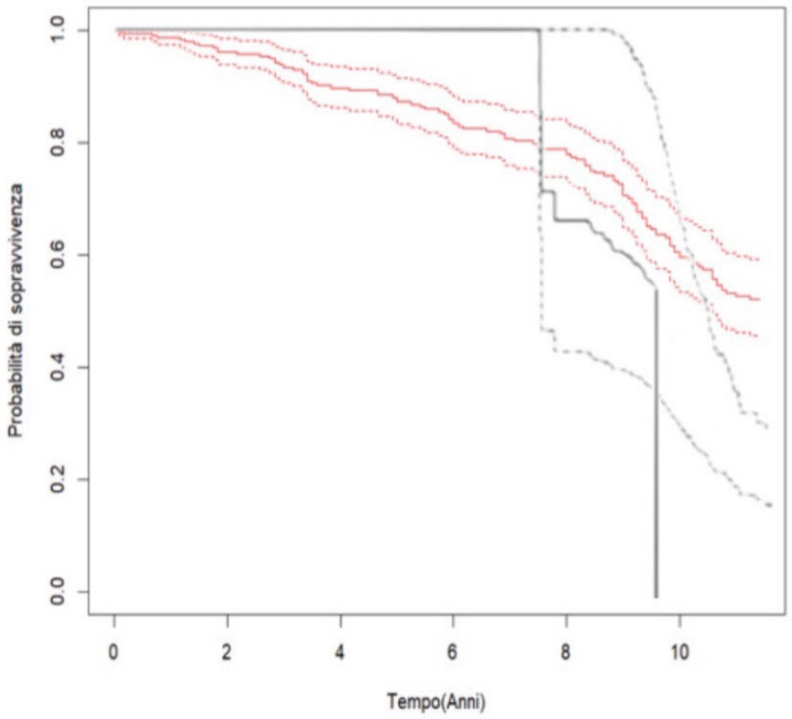

Figura 3.2: Grafico che presenta la funzione di sopravvivenza per il modello di massima verosimiglianza non parametrico con limiti di confidenza puntuali al 95% (nero) insieme alle stime della funzione di sopravvivenza per il modello "*stacked*" con limiti di confidenza puntuali bootstrappati al 95% (rosso).

Si riportano nella *Tabella 3.1* un elenco dei pesi dei 5 singoli sottomodelli facenti parte del modello "*stacked*", per ottenere un'ulteriore valutazione delle prestazioni del modello stesso.

Pesi individuali dei sottomodelli				
NPMLE	Weibull	Log-Logistic	Log-Normale	Gamma
$1.486\text{x}10^{-5}$	$2.128\text{x}10^{-9}$	$7.158\text{x}10^{-9}$	$9.978\text{x}10^{-7}$	0.0022295

Tabella 3.1: Tabella dei pesi per i 5 sottomodelli

I pesi riportati nella *Tabella 3.1* riflettono la capacità del modello "*stacked*" di spostare l'assegnazione del peso ad un

modello che, inizialmente, potrebbe non essere stato preso in considerazione da solo, come il modello Gamma.

In questo capitolo si voleva valutare lo stimatore "*stacked*" proposto, a confronto dello stimatore non parametrico di massima verosimiglianza (NPMLE), applicandolo al dataset Channing, la cui natura dei dati motiva l'utilizzo di stimatori basati sul *forward recurrence time*.
In seguito ad una verifica grafica ed analitica dell'ipotesi di stazionarietà, necessaria per procedere successivamente nelle analisi, si sono confrontate le curve di sopravvivenza stimate dal modello non parametrico NPMLE e dal modello "*stacked*", composto da cinque sottomodelli (NPMLE, *Weibull, Log-Logistic, Log-Normale e Gamma).* Da questo paragone, risulta evidente che l'adattamento risulti migliore per il modello "*stacked*", il quale riesce a stimare adeguatamente la sopravvivenza servendosi unicamente del *forward recurrence time*, a differenza del modello NPMLE, il quale descresce velocemente fino a zero. Inoltre, riportando i pesi dei sottomodelli presi in considerazione, che possono essere interpretati come indicatori dell'affidabilità o della capacità predittiva del modello all'interno dello stimatore "*stacked*", si osserva che il peso viene assegnato al modello *Gamma*, il quale inizialmente non sarebbe stato preso in considerazione singolarmente nell'analisi in questione.

Capitolo 4

Conclusioni

In questo elaborato si conduce uno studio su dati di sopravvivenza, i quali descrivono il periodo di tempo trascorso tra un evento iniziale ed un evento finale di interesse, come la ricaduta di una malattia o il decesso dovuto a quest'ultima. Spesso, l'evento di interesse non viene osservato prima del termine dello studio per tutti i soggetti, a causa di vincoli pratici o in quanto si perde il contatto con alcuni individui per una serie di possibili motivi, intrinsechi al soggetto stesso ma esterni alla ragione dello studio. Di conseguenza, i dati di sopravvivenza sono caratterizzati dalla censura a destra.
Nel caso in esame i dati sono contraddistinti inoltre da troncamento a sinistra, in riferimento ai pazienti per i quali si verifica l'evento prima ancora dell'inizio dello studio e pertanto non entrano a far parte del campione.
Quindi, tutti i soggetti in esame hanno già sperimentato l'evento iniziale, la cui data precisa di inizio è tuttavia sconosciuta. Questa difficoltà pratica dei dati troncati a sinistra, ottenuti da uno studio di coorte prevalente, fornisce la motivazione per utilizzare solo i tempi di vita residua al fine di stimare la distribuzione di sopravvivenza dei soggetti nello studio.

Questo studio confronta, utilizzando i dati sulle vite residue, l'adattamento dello stimatore non parametrico di massima verosimiglianza (NPMLE) e di uno stimatore "*stacked*", che comprende modelli parametrici e non, con pesi di "*stacking*" ottimali ottenuti minimizzando la funzione di punteggio Brier.

Nel Capitolo 2, utilizzando dati di vita residua simulati con censura a destra, si sono esaminate le prestazioni dei singoli stimatori parametrici e non parametrici rispetto allo stimatore "*stacked*", composto da cinque sottomodelli (NPMLE, Weibull, Log-Logistic, Log-Normale e Gamma), quando i dati sulla vita residua sono soggetti a proporzioni crescenti di censura amministrativa.
L'obiettivo era valutare il vantaggio crescente, al diminuire del follow-up, dell'utilizzo di uno stimatore di "*stacking*", sia con il modello NPMLE corretto, che con le funzioni di sopravvivenza parametriche.
Confrontando gli stimatori, utilizzando i loro errori DISSE medi, si è mostrato che l'NPMLE corretto non funziona in maniera soddisfacente né da solo né come membro dello stimatore "*stacked*", all'aumento della percentuale di censura amministrativa. Questo accade in quanto al crescere della percentuale di censura, aumenta l'incapacità del modello di prevedere qualsiasi probabilità di sopravvivenza potenzialmente diversa da zero dopo la fine dello studio, cioè dopo la censura amministrativa.

Nel Capitolo 3 si è valutato lo stimatore "*stacked*" proposto, in confronto allo stimatore non parametrico di massima verosimiglianza (NPMLE) applicandolo al dataset Channing,

la cui natura dei dati motiva l'utilizzo di stimatori basati sul *forward recurrence time.* Si è verificata innanzitutto l'ipotesi di stazionarietà graficamente, confrontando le curve del *backward e forward recurrence time* e mediante il relativo test analitico, il cui risultato ha portato all'accetazione dell'ipotesi. Si sono confrontate le curve di sopravvivenza stimate dal modello non parametrico NPMLE e dal modello "*stacked*", dal cui paragone è emerso che l'adattamento risulta migliore per il modello "*stacked*", in quanto riesce a stimare adeguatamente la sopravvivenza servendosi unicamente del *forward recurrence time*, a differenza del modello NPMLE. Inoltre, riportando i pesi dei sottomodelli presi in considerazione, si è osservato che il peso viene assegnato al modello Gamma, il quale inizialmente non sarebbe stato preso in considerazione singolarmente nell'analisi in questione.

Inizialmente, lo scopo era di migliorare il modello NPMLE corretto quando si stimava la funzione di sopravvivenza utilizzando solamente le vite residue, osservate in uno studio di coorte prevalente con follow-up. L'obiettivo era quello di introdurre dei modelli prametrici in uno stimatore "*stacked*", pur mantenendo il modello NPMLE, ipotizzando che i modelli parametrici avrebbero attenuato l'incapacità del modello NPMLE di catturare il comportamento della funzione di sopravvivenza. Questo scopo viene raggiunto nelle analisi, sia nella simulazione che servendosi di dati reali di vita residua, in quanto è risultato che lo stimatore "*stacked*" supera lo stimatore NPMLE in termini di adattamento, consentendo stime accurate per le distribuzioni di sopravvivenza, nel caso in cui si abbiano dati riguardanti le vite residue dei soggetti nello studio.

Bibliografia

M.Asgharian, D. B. Wolfson, e X. Zhang (2006). *Checking stationarity of the incidence rate using prevalent cohort survival data. Statistics in Medicine.*

V. Addona e D. B. Wolfson (2006). A formal test for the stationarity of the incidence rate using data from a prevalent cohort study with follow-up.

Sitografia

McVittie J. H., Wolfson D. B., Vittorio Addona V. e Li Z. (2022). *Stacked survival models for residual lifetime data.*
https://bmcmedresmethodol.biomedcentral.com

Li Z. (2022). *A Comparison of Stacking Methods to Estimate Survival Using Residual Lifetime Data from Prevalent Cohort Studies.*
https://digitalcommons.macalester.edu/

Hyun Lee C., Zhou H., Ning J., Diane D. L. e Yu Shen Y. (2020). *CoxPhLb: An R Package for Analyzing Length Biased Data under Cox Model.*
https://www.ncbi.nlm.nih.gov/

Keiding N., Jason P. , Oluf H., Slama R. (2011). *Accelerated failure time regression for backward recurrence times and current durations.*
https://www.sciencedirect.com/