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

Scienze Statistiche

Relazione Finale

Distribuzioni mistura per l'analisi della sopravvivenza in disegni clinici randomizzati a due stadi

Relatore: Prof.ssa Giuliana Cortese
Dipartimento di Scienze Statistiche

Laureando: Giovanna Ranzato
Matricola N. 1132446

Anno Accademico 2018/2019

Alla mia nonna,
che sarebbe stata tanto orgogliosa.

Indice

Introduzione

I disegni clinici randomizzati a due stadi stanno diventando sempre più comuni nell'analizzare il trattamento da seguire per certe malattie, specialmente quelle più "complesse". Essi consistono in una randomizzazione iniziale dei pazienti ad una terapia di primo stadio, detta terapia di induzione, e successivamente, in base alla risposta del soggetto e al suo eventuale consenso, in una seconda randomizzazione ad una terapia di secondo stadio, detta terapia di mantenimento. Lo scopo principale di questi disegni è quello di trovare la combinazione di terapia di induzione e terapia di mantenimento (*policy*) che massimizzi il tempo di sopravvivenza medio dei pazienti, al fine di somministrarla nella pratica ospedaliera quotidiana per il trattamento di alcune malattie.

In letteratura esistono però poche analisi di questo tipo; la maggior parte, infatti, consiste nel considerare i due stadi separatamente. Gli unici metodi di stima della distribuzione della sopravvivenza che considerino i due stadi contemporaneamente, al fine di trovarne la miglior combinazione, sono metodi non parametrici di recente studio: si tratta dei metodi LDT, WT e WRSE.
Considerato ciò, lo scopo di questa tesi è quello di studiare un metodo parametrico per analizzare la sopravvivenza dei pazienti sotto le varie *policies*. In particolare, implementeremo un modello mistura per trattare il tempo di sopravvivenza dei soggetti, variante di un modello già esistente che non tiene in considerazione la durata individuale della terapia di primo stadio per coloro che proseguono con il secondo.

Un esempio di disegno clinico randomizzato a due stadi è quello condotto dal *Cancer and Leukemia Group B* (CALGB), Protocollo numero 19808. Questo studio coinvolge pazienti al di sotto dei 60 anni malati di leucemia mieloide acuta (LMA). Durante il primo stadio del disegno essi ricevono casualmente una tipologia particolare di chemioterapia, tra due presenti; se alla fine di questo trattamento essi raggiungono una risposta completa alla malattia e se danno il proprio consenso, iniziano il secondo stadio e vengono randomizzati o ad una terapia immunoterapica o a nessun trattamento ulteriore (placebo). La somministrazione di un trattamento di secondo stadio è volta a debellare eventuali residui di malattia.

Utilizzeremo i dati provenienti da questo studio per trovare, tramite il metodo parametrico implementato, l'eventuale *policy* che, somministrata, porti al maggior tempo di sopravvivenza medio dei pazienti. Valuteremo inoltre la *performance* di questo metodo confrontandolo con alcune delle tecniche non parametriche menzionate.

Nel capitolo 1 di questo lavoro presentiamo gli aspetti teorici che sottostanno al metodo che proponiamo, sia nel caso di dati completi che censurati. Nel capitolo 2 ampliamo questi aspetti di base considerando tempi di terapia di primo stadio variabili per ogni individuo, arrivando così alla formulazione finale del nostro modello, in particolare nel caso di dati censurati. Il capitolo 3 è caratterizzato da una presentazione generale dei dati che verranno utilizzati, seguita da alcune prime analisi effettuate usando i metodi di analisi della sopravvivenza per disegni a due stadi già esistenti. Nel capitolo successivo eseguiamo le analisi vere e proprie basate sulle distribuzioni mistura, prima tramite simulazione dei dati e poi sui dati reali. Infine presentiamo le conclusioni tratte.

Le analisi sono state svolte utilizzando il *software RStudio*; il livello di significatività α è posto pari a 0.05 nel corso di tutto il lavoro presentato.

Capitolo 1

Metodo parametrico di analisi di sopravvivenza in disegni a due stadi: il modello mistura

1.1 Disegni clinici randomizzati a due stadi e relative tecniche di analisi

Malattie complesse come i tumori, l'AIDS o la depressione sono spesso trattate attraverso differenti combinazioni di terapie. In un disegno clinico a due stadi, infatti, i pazienti sono inizialmente randomizzati ad una prima terapia, detta terapia di induzione; in seguito, in base alla risposta della malattia al trattamento e al consenso del paziente, una seconda terapia, detta terapia di mantenimento, viene assegnata casualmente. Si tratta di una tipologia di studi adattivi, che consistono in una sequenza di trattamenti applicati in diversi stadi, in base alla storia individuale e alle risposte intermedie dei pazienti.

L'interesse di questi disegni è quello di analizzare l'effetto di diverse combinazioni di terapia di induzione e terapia di mantenimento, allo scopo di raggiungere il maggior beneficio in termini di sopravvivenza del paziente[1]. La maggior parte delle analisi precedentemente condotte su disegni di questo tipo non rispondeva a questa domanda, in quanto stimava separatamente le distribuzioni della sopravvivenza sotto la prima e la seconda terapia. La letteratura recente consiglia solo tre stimatori [4] che rispondano all'obiettivo di individuare la miglior combinazione in termini di allungamento medio della vita: lo stimatore LDT, studiato da Lunceford et al. (2002), stimatore semiparametrico della funzione di sopravvivenza,

[1]La sopravvivenza è definita come il tempo che intercorre dalla randomizzazione iniziale alla morte del paziente.

consistente ma non necessariamente efficiente; lo stimatore WT, proposto da Wahed e Tsiatis (2004), stimatore semiparametrico efficiente solo nel caso di dati non censurati; lo stimatore WRSE, studiato da Guo e Tsiatis (2005), stimatore non parametrico più efficiente dei precedenti. I primi due sono complessi e possono risultare difficili da implementare in pratica. Sembra inoltre esistere solo uno stimatore che consideri un approccio parametrico (Thall *et al.*, 2007) [17], in ambito bayesiano, ma il modello proposto soffre di sovraparametrizzazione.

Sulla base di queste considerazioni, lo scopo della tesi è quello di sviluppare un modello parametrico per analizzare dati da disegni randomizzati a due stadi, che possa essere facilmente implementato usando i *softwares* esistenti. In particolare, il modello proposto sarà un modello mistura, basato sugli aspetti teorici che adesso andiamo a presentare.

1.2 Il metodo: analisi di sopravvivenza per dati non censurati

Consideriamo un disegno clinico a due stadi in cui i pazienti sono inizialmente randomizzati ad una terapia di induzione tra A_1 e A_2; se essi rispettano i criteri di eleggibilità e se danno il proprio consenso, vengono casualmente assegnati alla terapia di mantenimento B_1 o B_2 [17]. L'obiettivo è perciò quello di confrontare le curve di sopravvivenza stimate sotto le varie combinazioni di terapie $A_jB_k(j,k = 1,2)$; la stima delle funzioni di sopravvivenza avrà un'interpretazione *intention-to-treat* [2].

Dal momento che i dati dei soggetti delle due terapie di induzione sono indipendenti, ci focalizziamo solo su coloro che che ricevono la terapia A_1; un discorso simile varrà per A_2. Per definire le distribuzioni di sopravvivenza per i trattamenti A_1B_1 e A_1B_2, sfruttiamo il contesto proposto da Lunceford et al. [8], basato su tempi di sopravvivenza controfattuali. Il tempo di sopravvivenza osservato per ogni paziente i risulta

$$T_i = (1 - R_i)T_{0i}^* + R_i\big[Z_iT_{1i}^* + (1 - Z_i)T_{2i}^*\big], \quad i = 1, 2, ..., n, \tag{1.1}$$

dove R_i indica il consenso/eleggibilità al trattamento successivo per il soggetto i ($R_i = 1$ se il paziente viene assegnato alla terapia di mantenimento, altrimenti $R_i = 0$) e T_{0i}^* è il tempo di sopravvivenza per il soggetto i se $R_i = 0$. Z_i, definito solo se $R_i = 1$, è un indicatore per l'assegnazione della terapia di mantenimento

[2] Considera facenti parte dell'esperimento tutti coloro che hanno iniziato il trattamento, a prescindere dal fatto che l'abbiano portato a compimento.

($Z_i = 1$ se il paziente i viene assegnato alla terapia B_1, 0 se assegnato a B_2); T^*_{1i} rappresenta quindi il tempo dalla randomizzazione iniziale alla morte se il soggetto i riceve la terapia di mantenimento B_1, equivalentemente T^*_{2i} per la terapia B_2. Le variabili $T^*_{ki}, k = 0, 1, 2$, sono tempi di sopravvivenza controfattuali (o potenziali), in quanto non possono essere osservati per tutti gli individui.

Si può dimostrare (si veda l'Appendice A.1) che la funzione di densità di T_i è una mistura[3]. Essa è infatti il risultato della combinazione delle funzioni di densità dei tempi di sopravvivenza potenziali: assunto $T^*_{ki} \sim f_k(\cdot;\theta_k), k = 0, 1, 2$, $R_i \sim Bernoulli(\pi_r)$ e $Z_i|R_i = 1 \sim Bernoulli(\pi_z)$, con $(\pi_r, \pi_z) = \pi$ e $(\theta_0, \theta_1, \theta_2) = \theta$, si ha che

$$f(t_i;\pi,\theta) = (1-\pi_r)f_0(t_i;\theta_0) + \pi_r\pi_z f_1(t_i;\theta_1) + \pi_r(1-\pi_z)f_2(t_i;\theta_2). \tag{1.2}$$

Si può notare come la somma dei pesi associati alle varie densità sia pari ad uno.

Come già menzionato, l'obiettivo è quello di stimare le funzioni di sopravvivenza sotto i trattamenti A_1B_1 e A_1B_2, al fine di analizzare quale combinazione massimizzi la probabilità di sopravvivere per i pazienti. A partire dalla funzione di densità calcolata, riusciamo a scrivere le funzioni di sopravvivenza $S_1(\cdot)$ e $S_2(\cdot)$, ossia le distribuzioni della sopravvivenza potenziale della popolazione se tutti i pazienti fossero stati rispettivamente assegnati ad A_1B_1 o ad A_1B_2, assumendo che qualche soggetto possa non sottoporsi ad una terapia di mantenimento. Le funzioni di sopravvivenza sono le seguenti:

$$\begin{aligned} S_1(t;\pi_r,\theta_0,\theta_1) &= 1-(1-\pi_r)F_0(t;\theta_0) - \pi_r F_1(t;\theta_1), \\ S_2(t;\pi_r,\theta_0,\theta_2) &= 1-(1-\pi_r)F_0(t;\theta_0) - \pi_r F_2(t;\theta_2), \end{aligned} \tag{1.3}$$

con $F_0(\cdot;\theta_0), F_1(\cdot;\theta_1)$ e $F_2(\cdot;\theta_2)$ funzioni di ripartizione del tempo di sopravvivenza associate rispettivamente all'assenza di una terapia di mantenimento, alla terapia di mantenimento B_1 e alla terapia di mantenimento B_2. Lo sviluppo delle formule (1.3) è riportato in Appendice A.2.

Per stimare $S_1(\cdot)$ e $S_2(\cdot)$ è necessario stimare i parametri π_r e θ. Per fare ciò dobbiamo scrivere la funzione di verosimiglianza associata ai vettori aleatori indipendenti ed identicamente distribuiti, osservabili per ogni soggetto, che chiamiamo D_i:

$$D_i = (R_i, R_i Z_i, T_i).$$

[3]Si definisce mistura una funzione di densità $f(x;\Theta) = \sum_{j=1}^{J} w_j f_j(x;\theta_j)$, con $\Theta = (w_1, ..., w_J, \theta_1, ..., \theta_J)$, $f_j(\cdot;\theta_j)$ funzione di densità dipendente dal parametro θ_j e w_j peso non negativo $\left(\sum_{j=1}^{J} w_j = 1\right)$ [2].

Si dimostra che la funzione di verosimiglianza (sviluppo in Appendice A.3) può essere fattorizzata in due componenti, la componente $L_1(\cdot)$ per il solo parametro π e la componente $L_2(\cdot)$ per il solo parametro θ, nel modo seguente:

$$L(\pi,\theta;d)=\prod_{i=1}^{n}L_i(\pi,\theta;d_i)=(1-\pi_r)^{\sum_i(1-r_i)}\ \pi_r^{\sum_i r_i}\ \pi_z^{\sum_i r_i z_i}\ (1-\pi_z)^{\sum_i r_i(1-z_i)} \times\prod_{i=1}^{n}\Big[f_0(t_i;\theta_0)^{1-r_i}f_1(t_i;\theta_1)^{r_i z_i}f_2(t_i;\theta_2)^{r_i(1-z_i)}\Big]. \tag{1.4}$$

Si nota infatti come $L(\pi,\theta;d)=L_1(\pi;r,rz)\cdot L_2(\theta;d)$; massimizzando separatamente $L_1(\cdot)$ e $L_2(\cdot)$ si riescono quindi ad ottenere gli stimatori di π e di θ. La stima di massima verosimiglianza di π, pari alle frequenze relative di r_i e di z_i, è data da

$$\hat{\pi}=(\hat{\pi}_r,\hat{\pi}_z)=\left(\frac{\sum_{i=1}^{n}r_i}{n},\frac{\sum_{i=1}^{n}r_i z_i}{\sum_{i=1}^{n}r_i}\right),$$

mentre la stima di θ dipende dalle assunzioni distributive che vengono imposte sui tempi di sopravvivenza potenziali $T^*_{ki}, k=0,1,2$.

1.2.1 Tempi di sopravvivenza esponenziali

Per semplicità assumiamo che i $T^*_{ki}, k=0,1,2$, siano variabili casuali esponenziali di media θ_k ($\theta_k>0$), per cui $f_k(t_i;\theta_k)=\theta_k^{-1}exp\Big(-\frac{t_i}{\theta_k}\Big)$ e $F_k(t_i;\theta_k)=1-exp\Big(-\frac{t_i}{\theta_k}\Big)$; il tasso (funzione di rischio) è $\lambda_k(t)=f_k(t)S_k(t)^{-1}=\theta_k^{-1}$ e i θ_k rappresentano perciò i parametri "valore medio" delle rispettive variabili casuali. Si avrà quindi, oltre alla stima di π già menzionata, che la stima di massima verosimiglianza di θ è pari a

$$\hat{\theta}=\Big(\hat{\theta}_0,\hat{\theta}_1,\hat{\theta}_2\Big)=\left(\frac{\sum_i(1-r_i)t_i}{\sum_i(1-r_i)},\frac{\sum_i r_i z_i t_i}{\sum_i r_i z_i},\frac{\sum_i r_i(1-z_i)t_i}{\sum_i r_i(1-z_i)}\right).$$

Le stime $\hat{\theta}_k$ coincidono con le medie aritmetiche dei tempi di sopravvivenza nei casi di assenza di terapia di mantenimento, presenza di terapia di mantenimento B_1 e presenza di terapia di mantenimento B_2.

Andando a sostituire in $S_1(\cdot)$ e in $S_2(\cdot)$ le stime ricavate, otteniamo le stime delle funzioni di sopravvivenza, le quali possono essere confrontate per individuare quale combinazione di terapie porti a raggiungere il maggior beneficio in termini

di allungamento della vita. In particolare si ha

$$\hat{S}_1(t) = S_1\big(t; \hat{\pi}_r, \hat{\theta}_0, \hat{\theta}_1\big) = (1-\hat{\pi}_r)exp\bigg(-\frac{t}{\hat{\theta}_0}\bigg) + \hat{\pi}_r exp\bigg(-\frac{t}{\hat{\theta}_1}\bigg),$$

$$\hat{S}_2(t) = S_2\big(t; \hat{\pi}_r, \hat{\theta}_0, \hat{\theta}_2\big) = (1-\hat{\pi}_r)exp\bigg(-\frac{t}{\hat{\theta}_0}\bigg) + \hat{\pi}_r exp\bigg(-\frac{t}{\hat{\theta}_2}\bigg).$$

1.3 Il metodo: analisi di sopravvivenza per dati censurati

Nell'ambito dell'analisi di dati di sopravvivenza spesso si incorre nel problema dei dati censurati, in particolare dei dati censurati a destra. Si parla di censura a destra quando la durata reale del tempo di sopravvivenza di un soggetto è superiore a quella osservata, ma non si sa di quanto.

Supponiamo di avere una variabile casuale C_i indicante il tempo che intercorre dalla randomizzazione iniziale del soggetto alla sua censura. Il tempo trascorso dalla randomizzazione all'evento osservato per ogni paziente sarà pari a

$$U_i = min(T_i, C_i), \quad i = 1, 2, ..., n,$$

con l'evento che può quindi essere l'evento morte se $T_i \leq C_i$ o l'evento censura se $C_i < T_i$. U_i rappresenterà perciò il tempo di sopravvivenza osservato per ogni individuo. Specifichiamo che la risposta intermedia alla terapia di induzione è solitamente valutata presto nello studio, in modo da poter osservare R_i per ogni soggetto; se qualche risposta non dovesse essere osservata, è consuetudine trattare questi pazienti come "non rispondenti", ossia associargli $R_i = 0$.
Per illustrare in modo più chiaro il concetto di censura a destra, riportiamo uno schema di esempio in Figura 1.1: il primo paziente decede durante il secondo stadio, il secondo è censurato, in quanto vive oltre il secondo stadio ma il tempo di sopravvivenza osservato coincide con la durata dello studio, il terzo vive oltre il primo stadio, ma non riceve una terapia di mantenimento, il quarto decede precocemente durante il primo stadio e verrà assunto come "non rispondente", il quinto viene censurato durante il secondo stadio (perchè ad esempio uscito dallo studio).

La funzione di densità del tempo di sopravvivenza osservato e le funzioni di sopravvivenza rimangono le stesse del paragrafo 1.2, ma in funzione di u. Per ricavare le stime dei parametri dobbiamo però scrivere la funzione di verosimiglianza associata alla nuova sequenza di vettori osservabili per ogni soggetto, che

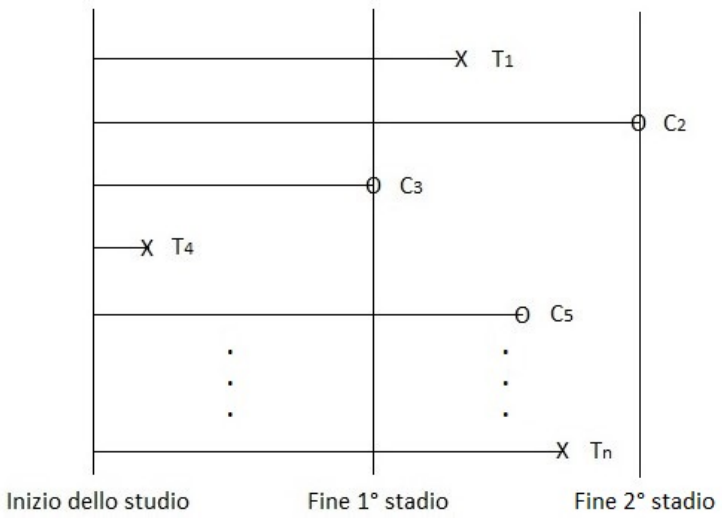


Figura 1.1: Dati censurati a destra.

chiamiamo sempre D_i:

$$D_i = (R_i, R_i Z_i, U_i, \Delta_i),$$

con $\Delta_i = I(T_i \leq C_i)$ indicatore "caso completo" ($\Delta_i = 1$ se non c'è censura, 0 altrimenti). Assumiamo che la censura sia completamente casuale, quindi indipendente dai dati. Nel costruire la funzione di verosimiglianza, dobbiamo considerare attentamente l'informazione che ogni osservazione ci fornisce [6]: un'osservazione corrispondente ad un evento esatto dà informazione riguardo la probabilità che l'evento accada in quel momento, che coincide approssimativamente con la funzione di densità $f(u_i)$; per un'osservazione censurata a destra sappiamo solo che l'evento è accaduto dopo un certo momento, quindi l'informazione è rappresentata dalla funzione di sopravvivenza $S(u_i)$. Partendo dalla (1.4), riusciamo ad adattare la funzione di verosimiglianza al caso di dati censurati a destra. L'espressione finale, il cui sviluppo è riportato in Appendice A.4, è la seguente[4]:

$$\begin{aligned} L(\pi, \theta; d) = \prod_{i=1}^{n} L_i(\pi, \theta; d_i) \propto \quad & (1-\pi_r)^{\sum_i (1-r_i)} \ \pi_r^{\sum_i r_i} \ \pi_z^{\sum_i r_i z_i} \ (1-\pi_z)^{\sum_i r_i (1-z_i)} \\ & \times \prod_{i=1}^{n} \Bigg\{ \Big[f_0(u_i;\theta_0)^{\delta_i} S_0(u_i;\theta_0)^{1-\delta_i} \Big]^{1-r_i} \\ & \times \Big[f_1(u_i;\theta_1)^{\delta_i} S_1(u_i;\theta_1)^{1-\delta_i} \Big]^{r_i z_i} \\ & \times \Big[f_2(u_i;\theta_2)^{\delta_i} S_2(u_i;\theta_2)^{1-\delta_i} \Big]^{r_i (1-z_i)} \Bigg\}. \end{aligned}$$

[4]La funzione di verosimiglianza $L(\cdot)$ è proporzionale all'espressione scritta, in quanto quest'ultima costituisce una funzione di verosimiglianza parziale.

Notiamo come anche in questo caso la verosimiglianza possa essere fattorizzata in due componenti, una in funzione del parametro π, $L_1(\pi; r, rz)$, e una in funzione del parametro θ, $L_2(\theta; d)$. Massimizzando separatamente queste due componenti, otteniamo per π_r e π_z le stesse stime del caso dei dati non censurati, mentre per ricavare le stime dei $\theta_k, k = 0, 1, 2$, dobbiamo imporre delle assunzioni distributive sui tempi di sopravvivenza potenziali.

1.3.1 Tempi di sopravvivenza esponenziali

Assumiamo che i tempi di sopravvivenza potenziali, $U^*_{ki}, k = 0, 1, 2$, siano variabili casuali esponenziali di media θ_k. Dopo aver calcolato le equazioni di verosimiglianza relative a $L_2(\theta; d)$ (sviluppo in Appendice A.5), otteniamo che la stima di massima verosimiglianza di θ è pari a

$$\hat{\theta} = \Big(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2\Big) = \left(\frac{\sum_i (1 - r_i)u_i}{\sum_i (1 - r_i)\delta_i}, \frac{\sum_i r_i z_i u_i}{\sum_i r_i z_i \delta_i}, \frac{\sum_i r_i (1 - z_i)u_i}{\sum_i r_i (1 - z_i)\delta_i}\right).$$

Notiamo come i $\hat{\theta}_k$ coincidano con le medie aritmetiche dei tempi di sopravvivenza nei casi di assenza di terapia di mantenimento, presenza di terapia di mantenimento B_1 e presenza di terapia di mantenimento B_2, considerando i soli soggetti con evento osservato ($\delta_i = 1$).

Andando a sostituire in $S_1(\cdot)$ e in $S_2(\cdot)$ le stime ricavate, otteniamo le stime delle funzioni di sopravvivenza, riportate di seguito:

$$\hat{S}_1(u) = S_1\big(u; \hat{\pi}_r, \hat{\theta}_0, \hat{\theta}_1\big) = (1 - \hat{\pi}_r) exp\left(-\frac{u}{\hat{\theta}_0}\right) + \hat{\pi}_r exp\left(-\frac{u}{\hat{\theta}_1}\right),$$

$$\hat{S}_2(u) = S_2\big(u; \hat{\pi}_r, \hat{\theta}_0, \hat{\theta}_2\big) = (1 - \hat{\pi}_r) exp\left(-\frac{u}{\hat{\theta}_0}\right) + \hat{\pi}_r exp\left(-\frac{u}{\hat{\theta}_2}\right).$$

Capitolo 2

Estensione del metodo: il modello mistura in disegni con durata di primo stadio variabile

2.1 Il tempo di sopravvivenza osservato

Nel capitolo 1 non abbiamo tenuto in considerazione il fatto che la durata del primo stadio del disegno clinico (ossia la durata della terapia di induzione), per coloro che proseguono con la seconda fase, potesse variare da soggetto a soggetto; T^*_{1i} e T^*_{2i} rappresentavano infatti i tempi di sopravvivenza potenziali per il paziente i che riceve rispettivamente le terapie di secondo stadio B_1 e B_2, a partire dalla randomizzazione iniziale nello studio. Ora invece consideriamo un disegno clinico randomizzato a due stadi in cui specifichiamo che la durata del primo stadio sia variabile per ogni individuo [4]; il tempo di sopravvivenza osservato per il paziente i risulta quindi modificato rispetto al capitolo precedente e in particolare si ha

$$T_i = (1 - R_i)T^*_{0i} + R_i\big[T^R_i + Z_i T^*_{1i} + (1 - Z_i)T^*_{2i}\big], \quad i = 1, 2, ..., n. \tag{2.1}$$

La variabile T^R_i rappresenta il tempo che intercorre dalla prima randomizzazione nello studio all'inizio del secondo stadio per il paziente i "rispondente", quindi la durata di tempo in cui questo individuo si sottopone alla terapia di induzione; T^*_{1i} e T^*_{2i} indicano i tempi di sopravvivenza potenziali del paziente i a partire dall'inizio del secondo stadio, ossia dall'inizio della terapia di mantenimento B_1 o B_2.

Con questa variazione la funzione di densità di T_i sarà diversa rispetto a quella calcolata nel capitolo 1 e di conseguenza anche le funzioni di sopravvivenza $S_1(\cdot)$ e $S_2(\cdot)$ e gli stimatori dei parametri. Tutto il lavoro presentato si baserà sull'utilizzo di questo modello, che in questo senso risulta essere una modifica del

già esistente modello parametrico presentato nel capitolo precedente. Eseguiamo quindi nuovamente le analisi, sia nel caso di dati completi che censurati.

2.2 Analisi di sopravvivenza per dati non censurati

La funzione di densità del tempo di sopravvivenza osservato rimane una mistura, combinazione però di funzioni di densità diverse rispetto a quelle della (1.2). Dato $T_i^R|R_i = 1 \sim f_R(\cdot;\theta_R)$, $T_{ki}^* \sim f_k(\cdot;\theta_k), k = 0,1,2$, $R_i \sim Bernoulli(\pi_r)$ e $Z_i|R_i = 1 \sim Bernoulli(\pi_z)$, con $(\pi_r,\pi_z) = \pi$ e $(\theta_R,\theta_0,\theta_1,\theta_2) = \theta$, si ha

$$f(t_i;\pi,\theta) = (1-\pi_r)f_0(t_i;\theta_0) + \pi_r\pi_z f_{R1}(t_i;\theta_R,\theta_1) + \pi_r(1-\pi_z)f_{R2}(t_i;\theta_R,\theta_2). \quad (2.2)$$

$f_{R1}(\cdot;\theta_R,\theta_1)$ rappresenta la funzione di densità della somma $T_i^R + T_{1i}^*$ (caso $Z_i = 1$), data dalla convoluzione[1] di $f_R(\cdot;\theta_R)$ con $f_1(\cdot;\theta_1)$, assunta l'indipendenza tra T_i^R e T_{1i}^*; stesso discorso per $f_{R2}(\cdot;\theta_R,\theta_2)$, funzione di densità di $T_i^R + T_{2i}^*$ (caso $Z_i = 0$). Facendo riferimento a $T_i^R + T_{1i}^*$, si definisce convoluzione di $f_R(\cdot)$ con $f_1(\cdot)$ la funzione definita nel seguente modo:

$$(f_R * f_1)(t_i) = \int_0^{t_i} f_R(j)f_1(t_i - j)dj = f_{R1}(t_i). \quad (2.3)$$

Per la dimostrazione della formula (2.2) si veda l'Appendice A.6.

Come nel capitolo precedente, a partire dalla (2.2) riusciamo a calcolare le funzioni di sopravvivenza $S_1(\cdot)$ e $S_2(\cdot)$, che riportiamo di seguito:

$$S_1(t;\pi_r,\theta_R,\theta_0,\theta_1) = 1 - (1-\pi_r)F_0(t;\theta_0) - \pi_r F_{R1}(t;\theta_R,\theta_1),$$
$$S_2(t;\pi_r,\theta_R,\theta_0,\theta_2) = 1 - (1-\pi_r)F_0(t;\theta_0) - \pi_r F_{R2}(t;\theta_R,\theta_2),$$

con $F_{R1}(\cdot;\theta_R,\theta_1)$ e $F_{R2}(\cdot;\theta_R,\theta_2)$ funzioni di ripartizione di $T_i^R + T_{1i}^*$ e di $T_i^R + T_{2i}^*$. Per stimare i parametri π_r e θ scriviamo la verosimiglianza in funzione del vettore di osservazioni $D_i = (R_i, R_iZ_i, T_i)$. L'espressione finale della funzione di verosimiglianza (stesso sviluppo della funzione di verosimiglianza del paragrafo 1.2) è la seguente:

$$L(\pi,\theta;d) = L_1(\pi;r,rz)\cdot L_2(\theta;d) = (1-\pi_r)^{\sum_i(1-r_i)}\ \pi_r^{\sum_i r_i}\ \pi_z^{\sum_i r_i z_i}\ (1-\pi_z)^{\sum_i r_i(1-z_i)}$$
$$\times \prod_{i=1}^{n}\Big[f_0(t_i;\theta_0)^{1-r_i} f_{R1}(t_i;\theta_R,\theta_1)^{r_i z_i} f_{R2}(t_i;\theta_R,\theta_2)^{r_i(1-z_i)}\Big].$$

[1] Date due variabili aleatorie continue e indipendenti X e Y, la funzione di densità di $X+Y$ viene detta la convoluzione delle funzioni di densità delle singole variabili [11].

Essendo $L_1(\cdot)$ sempre la stessa, lo stimatore di π coincide con quello del capitolo precedente; varia invece lo stimatore di θ, essendo $L_2(\cdot)$ ora costituita dal prodotto delle funzioni di densità di T^*_{0i}, $T^R_i + T^*_{1i}$ e $T^R_i + T^*_{2i}$. L'espressione di $\hat{\theta}$ dipenderà quindi dalle forme distributive $f_0(\cdot;\theta_0)$, $f_{R1}(\cdot;\theta_R,\theta_1)$ e $f_{R2}(\cdot;\theta_R,\theta_2)$.
Una volta ottenute le stime dei parametri, riusciamo a confrontare le curve di sopravvivenza stimate $\hat{S}_1(\cdot)$ e $\hat{S}_2(\cdot)$.

2.2.1 Tempi di sopravvivenza esponenziali

Assumiamo per semplicità che sia $T^*_{ki}, k = 0, 1, 2$, che T^R_i siano variabili casuali con distribuzione esponenziale di media rispettivamente θ_k e θ_R. $f_{R1}(\cdot;\theta_R,\theta_1)$ e $f_{R2}(\cdot;\theta_R,\theta_2)$ vengono quindi ottenute come convoluzione di variabili casuali esponenziali; si ricavano le seguenti funzioni di densità, la cui dimostrazione è riportata in Appendice A.7:

$$\begin{aligned} f_{R1}(t_i;\theta_R,\theta_1) &= \frac{e^{-t_i\theta_R^{-1}} - e^{-t_i\theta_1^{-1}}}{\theta_R - \theta_1}, \\ f_{R2}(t_i;\theta_R,\theta_2) &= \frac{e^{-t_i\theta_R^{-1}} - e^{-t_i\theta_2^{-1}}}{\theta_R - \theta_2}. \end{aligned} \tag{2.4}$$

Si può notare come la distribuzione della somma di due variabili esponenziali, di parametri differenti, abbia una forma chiusa.
Riscriviamo $L_2(\cdot)$ sostituendo le varie funzioni di densità:

$$L_2(\theta;d) = \prod_{i=1}^{n} \left[\left(\frac{e^{-t_i\theta_0^{-1}}}{\theta_0} \right)^{1-r_i} \left(\frac{e^{-t_i\theta_R^{-1}} - e^{-t_i\theta_1^{-1}}}{\theta_R - \theta_1} \right)^{r_i z_i} \left(\frac{e^{-t_i\theta_R^{-1}} - e^{-t_i\theta_2^{-1}}}{\theta_R - \theta_2} \right)^{r_i(1-z_i)} \right];$$

la si riesce a fattorizzare in due componenti: una contenente il parametro θ_0 (per il quale si ottiene, ovviamente, la stessa stima del paragrafo 1.2.1) e una contenente θ_R, θ_1 e θ_2. Per trovare le stime di questi ultimi ne abbiamo ricavato le rispettive equazioni di verosimiglianza; sfortunatamente queste non risultano risolvibili in forma esplicita, per cui bisognerà applicare un algoritmo di ottimizzazione numerica per derivare gli stimatori di questi parametri e quindi le stime delle funzioni di sopravvivenza.

2.3 Analisi di sopravvivenza per dati censurati

Come già spiegato nel paragrafo 1.3, nel caso di dati censurati a destra il tempo di sopravvivenza osservato per ogni paziente coincide con $U_i = min(T_i, C_i)$. La funzione di densità del tempo di sopravvivenza osservato e le funzioni di sopravvivenza rimangono le stesse del paragrafo 2.2, ma in funzione di u. La funzione di verosimiglianza, associata al vettore $D_i = (R_i, R_i Z_i, U_i, \Delta_i)$, con $\Delta_i = I(T_i \leq C_i)$, assume invece la seguente espressione (stesso sviluppo della funzione di verosimiglianza del paragrafo 1.3):

$$\begin{aligned} L(\pi, \theta; d) = L_1(\pi; r, rz) \cdot L_2(\theta; d) \propto \quad & (1-\pi_r)^{\sum_i (1-r_i)} \ \pi_r^{\sum_i r_i} \ \pi_z^{\sum_i r_i z_i} \ (1-\pi_z)^{\sum_i r_i (1-z_i)} \\ & \times \prod_{i=1}^{n} \Bigg\{ \Big[f_0(u_i; \theta_0)^{\delta_i} S_0(u_i; \theta_0)^{1-\delta_i} \Big]^{1-r_i} \\ & \times \Big[f_{R1}(u_i; \theta_R, \theta_1)^{\delta_i} S_{R1}(u_i; \theta_R, \theta_1)^{1-\delta_i} \Big]^{r_i z_i} \\ & \times \Big[f_{R2}(u_i; \theta_R, \theta_2)^{\delta_i} S_{R2}(u_i; \theta_R, \theta_2)^{1-\delta_i} \Big]^{r_i (1-z_i)} \Bigg\}. \end{aligned} \tag{2.5}$$

Come sempre si fattorizza $L(\cdot)$ nelle due componenti $L_1(\cdot)$ e $L_2(\cdot)$, dipendenti rispettivamente dai parametri $\pi = (\pi_r, \pi_z)$ e $\theta = (\theta_R, \theta_0, \theta_1, \theta_2)$; lo stimatore di massima verosimiglianza di π rimane lo stesso delle analisi precedenti, mentre lo stimatore di θ dipende dalle assunzioni distributive sui tempi di sopravvivenza.

2.3.1 Tempi di sopravvivenza esponenziali

Imponiamo che $U^*_{ki}, k = 0, 1, 2$, e U^R_i siano variabili casuali esponenziali rispettivamente di media θ_k e θ_R. Così come nel paragrafo 2.2.1, riusciamo ad esprimere $f_{R1}(\cdot; \theta_R, \theta_1)$ e $f_{R2}(\cdot; \theta_R, \theta_2)$ come convoluzione di variabili esponenziali. Ricaviamo le funzioni di sopravvivenza come complemento delle funzioni di ripartizione di $U^R_i + U^*_{1i}$ e $U^R_i + U^*_{2i}$, a loro volta ottenibili come integrale delle rispettive funzioni di densità. In particolare si ha

$$\begin{aligned} S_{R1}(u_i; \theta_R, \theta_1) = 1 - F_{R1}(u_i; \theta_R, \theta_1) &= \frac{\theta_R e^{-u_i \theta_R^{-1}} - \theta_1 e^{-u_i \theta_1^{-1}}}{\theta_R - \theta_1}, \\ S_{R2}(u_i; \theta_R, \theta_2) = 1 - F_{R2}(u_i; \theta_R, \theta_2) &= \frac{\theta_R e^{-u_i \theta_R^{-1}} - \theta_2 e^{-u_i \theta_2^{-1}}}{\theta_R - \theta_2}. \end{aligned} \tag{2.6}$$

Riusciamo allora a scrivere $L_2(\cdot)$, che si fattorizza nella componente costituita dal solo parametro θ_0 $\left(\hat{\theta}_0 = \frac{\sum_i (1-r_i) u_i}{\sum_i (1-r_i) \delta_i} \right)$ e nella componente contenente θ_R, θ_1 e θ_2; come nel caso dei dati non censurati, le equazioni di verosimiglianza per ricavare

$\hat{\theta}_R$, $\hat{\theta}_1$ e $\hat{\theta}_2$ non hanno soluzione esplicita e dovremo perciò applicare un algoritmo di ottimizzazione numerica.

2.3.2 Tempi di sopravvivenza Weibull

Essendo la funzione di verosimiglianza (2.5) quella a cui faremo successivamente riferimento per ricavare le stime dei parametri e quindi le stime delle curve di sopravvivenza, vediamo cosa si ottiene imponendo ai tempi di sopravvivenza una distribuzione un po' più complessa di quella esponenziale. Assumiamo che $U^*_{ki}, k = 0, 1, 2$, e U^R_i siano variabili casuali Weibull[2]: $U^*_{ki} \sim Weibull(\alpha_k, \lambda_k)$ e $U^R_i \sim Weibull(\alpha_R, \lambda_R)$.

Il problema che sorge in questo contesto riguarda l'espressione delle convoluzioni $f_{R1}(\cdot; \alpha_R, \alpha_1, \lambda_R, \lambda_1)$ e $f_{R2}(\cdot; \alpha_R, \alpha_2, \lambda_R, \lambda_2)$. Non si riesce infatti a ricavare in forma chiusa la convoluzione delle densità di due variabili Weibull tramite l'integrale (2.3); essa sarebbe ricavabile come approssimazione, si veda l'approssimazione di Leonard Johnson (1960) [13] o quella di Filho e Yacoub (2006) [12], o in forma esatta ma solo in riferimento a casi particolari [13] o tramite formule piuttosto complesse [19]. La funzione $L_2(\cdot)$ risulta quindi

$$
\begin{aligned}
L_2(\theta; d) \propto \quad & \prod_{i=1}^{n} \Bigg\{ \Bigg\{ \left[\frac{\alpha_0}{\lambda_0^{\alpha_0}} u_i^{\alpha_0 - 1} exp\Big(- \Big(\frac{u_i}{\lambda_0} \Big)^{\alpha_0} \Big) \right]^{\delta_i} \left[exp\Big(- \Big(\frac{u_i}{\lambda_0} \Big)^{\alpha_0} \Big) \right]^{1-\delta_i} \Bigg\}^{1-r_i} \\
& \times \Big[f_{R1}(u_i; \alpha_R, \alpha_1, \lambda_R, \lambda_1)^{\delta_i} S_{R1}(u_i; \alpha_R, \alpha_1, \lambda_R, \lambda_1)^{1-\delta_i} \Big]^{r_i z_i} \\
& \times \Big[f_{R2}(u_i; \alpha_R, \alpha_2, \lambda_R, \lambda_2)^{\delta_i} S_{R2}(u_i; \alpha_R, \alpha_2, \lambda_R, \lambda_2)^{1-\delta_i} \Big]^{r_i(1-z_i)} \Bigg\}.
\end{aligned}
$$

Per quanto riguarda la massimizzazione del fattore con α_0 e λ_0, solo la stima di massima verosimiglianza di λ_0 è ricavabile in forma chiusa $\left(\hat{\lambda}_0 = \left(\frac{\sum_i (1-r_i) u_i^{\hat{\alpha}_0}}{\sum_i (1-r_i)\delta_i} \right)^{\frac{1}{\hat{\alpha}_0}} \right)$; per ottenere la stima di α_0 bisognerà applicare un algoritmo di ottimizzazione numerica (l'equazione di verosimiglianza non è risolvibile in forma esplicita). Per derivare le stime dei restanti parametri bisognerà prima ricavare le convoluzioni e le rispettive funzioni di ripartizione tramite un *software* statistico e quindi applicare nuovamente un algoritmo di ottimizzazione.

Una volta ottenute le stime di tutti i parametri, è possibile ricavare le stime

[2]La distribuzione Weibull, $X \sim Weibull(\alpha, \lambda)$, di parametro di forma $\alpha > 0$ e parametro di scala $\lambda > 0$ ha funzione di densità $f(x; \alpha, \lambda) = \frac{\alpha}{\lambda^\alpha} x^{\alpha-1} exp\big(- (\frac{x}{\lambda})^\alpha \big)$, con supporto $\mathbb{R}^+$. La funzione di ripartizione è $F(x; \alpha, \lambda) = 1 - exp\big(- (\frac{x}{\lambda})^\alpha \big)$.

delle funzioni di sopravvivenza, utilizzando le formule

$$\hat{S}_1(u) = 1 - (1 - \hat{\pi}_r)F_0(u; \hat{\alpha}_0, \hat{\lambda}_0) - \hat{\pi}_r F_{R1}(u; \hat{\alpha}_R, \hat{\lambda}_R, \hat{\alpha}_1, \hat{\lambda}_1),$$

$$\hat{S}_2(u) = 1 - (1 - \hat{\pi}_r)F_0(u; \hat{\alpha}_0, \hat{\lambda}_0) - \hat{\pi}_r F_{R2}(u; \hat{\alpha}_R, \hat{\lambda}_R, \hat{\alpha}_2, \hat{\lambda}_2).$$

Capitolo 3

Esempio di disegno randomizzato a due stadi

3.1 I dati: disegno clinico randomizzato a due stadi in pazienti con leucemia mieloide acuta

I dati che analizzeremo provengono dallo studio CALGB (*Cancer and Leukemia Group B*) 19808 [7]. Lo scopo di questo studio è quello di sottoporre i pazienti al di sotto dei 60 anni malati di leucemia mieloide acuta (LMA) ad una terapia composta da due stadi, al fine di valutarne l'effetto sulla loro sopravvivenza. I soggetti malati di LMA, infatti, nonostante il raggiungimento di una risposta completa (CR) a seguito di chemioterapia, possono ospitare ancora residui minimi di malattia. Un successivo trattamento immunoterapico, in questo caso l'interleuchina-2 ricombinante (IL-2), può essere efficace se somministrato in presenza di residui di malattia.

Nello studio in questione i pazienti vengono randomizzati ad una tipologia di chemioterapia, ADE (citarabina, daunorubicina ed etoposide) o ADEP (citarabina, daunorubicina, etoposide e valspodar); se CR viene raggiunta essi possono essere randomizzati, previo consenso, alla somministrazione sottocutanea di IL-2, per 90 giorni, o a nessun trattamento ulteriore (Placebo). Utilizzando la notazione dei capitoli precedenti, i trattamenti di primo stadio A_1 e A_2 sono quindi rispettivamente rappresentati da ADEP ed ADE, mentre quelli di secondo stadio B_1 e B_2 da IL-2 e Placebo.

Obiettivo primario dello studio è quello di confrontare la sopravvivenza dei pazienti sotto le diverse linee di trattamento, al fine di stabilire quale combinazione di terapia di induzione e terapia di mantenimento massimizzi la probabilità di sopravvivere per questi soggetti. Implementeremo perciò le stime delle funzioni

di sopravvivenza ricavate nel capitolo precedente e le confronteremo con le stime ottenute tramite alcuni metodi non parametrici, con lo scopo di verificare le *performances* delle varie tecniche nei disegni randomizzati a due stadi.

3.2 Analisi preliminari e analisi di sopravvivenza non parametriche

Il *dataset* originario è composto da 302 pazienti che vengono randomizzati ad una terapia di primo stadio, ADEP o ADE (A_1 o A_2), di durata variabile; di questi, solo 292 intraprendono effettivamente il trattamento. Si tratta di 155 maschi e 137 femmine, tutti al di sotto dei 60 anni (si va da un'età minima di 18 anni al momento della registrazione nello studio, fino ad un'età massima di 59 anni). Un totale di 223 pazienti riesce a raggiungere una risposta completa alla terapia di induzione ed è quindi candidabile a proseguire con il secondo stadio. Tra questi soggetti, solo 92 ricevono una terapia di mantenimento ($R_i = 1$), di cui 46 vengono randomizzati al trattamento immunoterapico IL-2 (B_1) e 46 al trattamento con Placebo (B_2); un elevato numero di pazienti non intraprende quindi una terapia di mantenimento, a causa di svariati motivi come l'interruzione volontaria del *follow-up*, la comparsa di un'altra malattia o lo sviluppo di tossicità in seguito alla somministrazione del trattamento di primo stadio.

Tra i 292 pazienti nello studio, 119 decedono durante il periodo di terapia, 173 sopravvivono o escono dallo studio e sono quindi censurati. Il loro tempo di sopravvivenza (U_i) oscilla tra un minimo di un giorno e un massimo di 4762 giorni (13 anni), secondo l'andamento riportato in Figura 3.1. Si può osservare come la distribuzione dei giorni di sopravvivenza, trattandosi di variabile temporale, sia asimmetrica a destra e presenti un andamento tipico di una mistura. La variabile risulta infatti avere due andamenti: un primo andamento decrescente fino a 3500 giorni di sopravvivenza (9 anni), con un picco iniziale di pazienti che decedono o escono dallo studio durante il primo anno e mezzo a partire dalla registrazione nel protocollo, e un secondo andamento decrescente dopo i 3500 giorni. Dai capitoli precedenti sappiamo che la funzione di densità del tempo di sopravvivenza osservato è calcolabile come mistura delle funzioni di densità dei tempi di sopravvivenza dei pazienti che non ricevono una terapia di mantenimento ("non rispondenti", $R_i = 0$) e di quelli che invece proseguono con il secondo stadio ("rispondenti", $R_i = 1$). Analizziamo quindi separatamente le distribuzioni dei giorni di sopravvivenza in questi due gruppi (Figura 3.2) e mostriamo come la densità del tempo di sopravvivenza globale sia una combinazione delle densità

del tempo di sopravvivenza di "rispondenti" e di "non rispondenti" (Figura 3.3). Effettivamente la distribuzione dei giorni di sopravvivenza dei "non rispondenti" ha un picco tra 0 e 500 giorni e quella dei "rispondenti" tra 3500 e 4000, fenomeno che si riscontra nell'istogramma globale di Figura 3.1. Osservando allo stesso tempo le densità in Figura 3.3 (stimate imponendo ampiezza di banda pari a 500 e nucleo gaussiano), è chiaro come la densità globale (in rosso) sia il risultato della combinazione delle singole densità dei soggetti con $R_i = 0$ e di quelli con $R_i = 1$ (in verde).

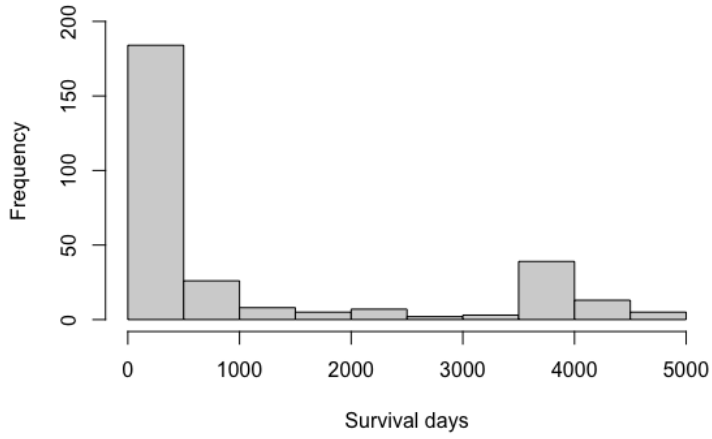
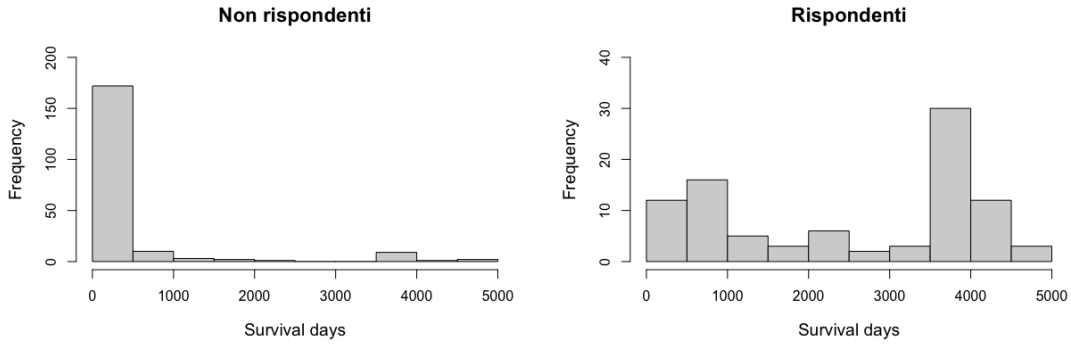


Figura 3.1: Distribuzione dei giorni di sopravvivenza dei pazienti nello studio.



(a) (b)

Figura 3.2: Distribuzione dei giorni di sopravvivenza dei pazienti "non rispondenti" (a) e dei pazienti "rispondenti" (b).

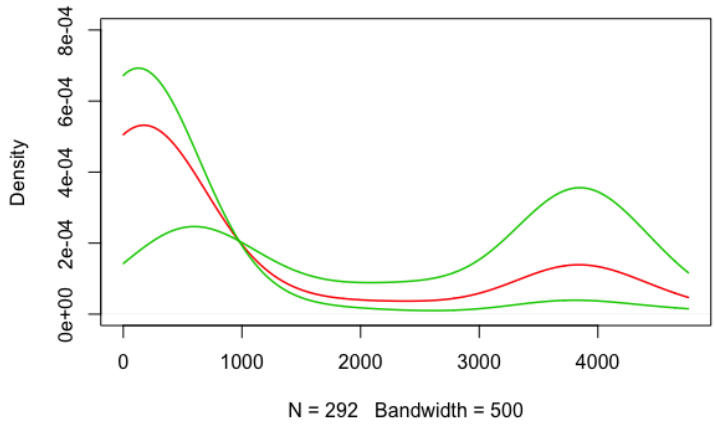


Figura 3.3: Densità dei giorni di sopravvivenza di "non rispondenti" e "rispondenti" in verde, densità globale dei giorni di sopravvivenza in rosso.

Un errore comune nei disegni clinici randomizzati a due stadi, come già menzionato, è quello di analizzare la sopravvivenza separatamente nelle due fasi di induzione e mantenimento; nell'articolo di Kolitz [7], relativo al *dataset* che stiamo considerando, l'*endpoint* primario è infatti rappresentato da DFS[1] (*Disease Free Survival*) calcolato a partire dalla data della randomizzazione al trattamento di secondo stadio e non dall'inizio dello studio. Come analisi preliminare decidiamo appositamente di riportare la distribuzione dei giorni di sopravvivenza a partire dall'inizio della terapia di secondo stadio, senza tenere conto del trattamento precedente (Figura 3.4). La distribuzione dei pazienti censurati (sopravvissuti o usciti dallo studio) e non censurati (deceduti) è simile nei due gruppi di terapia di secondo stadio, per cui potremmo affermare che la sopravvivenza mediana sia maggiore per coloro che ricevono il trattamento immunoterapico rispetto a coloro cui viene somministrato il placebo. Eseguendo però un *t-test* si accetta l'ipotesi nulla di uguaglianza delle medie nei due campioni (p-*value*=0.074), per cui si afferma che la sopravvivenza di coloro che ricevono IL-2 e quella di coloro che non ricevono trattamento ulteriore è mediamente la stessa.

Inseguendo invece lo scopo di questa tesi, ossia quello di ricavare uno stimatore della funzione di sopravvivenza che porti a trovare la miglior combinazione di trattamenti in termini di allungamento medio della vita, abbiamo menzionato nel paragrafo 1.1 alcuni stimatori non parametrici che rispondono a questa domanda. Tramite il pacchetto DTR (*Dynamic Treatment Regimes*) di *RStudio* riusciamo

[1]DFS rappresenta la sopravvivenza libera da malattia, ossia l'intervallo di tempo dalla randomizzazione al trattamento alla recidiva di malattia o morte qualsiasi sopraggiunga prima.

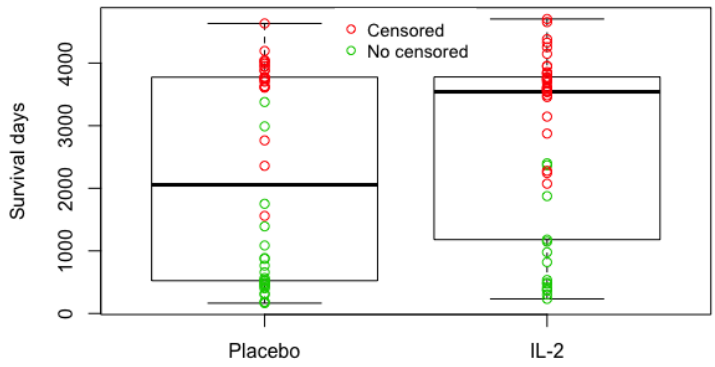


Figura 3.4: Distribuzione dei giorni di sopravvivenza per terapia di mantenimento, a partire dall'inizio del secondo stadio.

ad implementare due di questi stimatori [15]: lo stimatore LDT e lo stimatore WRSE.

LDT è uno stimatore semiparametrico calcolato a partire dalla (1.1) nel caso di dati censurati e che sfrutta il metodo *inverse-weighting*; si tratta di uno stimatore consistente ma non necessariamente efficiente. In particolare, $\hat{S}_{1LDT}(\cdot)$ viene ottenuto pesando ogni osservazione per una quantità $Q_{1i} = 1 - R_i + \pi_z^{-1} R_i Z_i$, in modo tale da assegnare peso π_z^{-1} ai pazienti che intraprendono la *policy* A_1B_1, peso uno a coloro che si fermano al primo stadio A_1 e peso zero a coloro che intraprendono A_1B_2; un ragionamento analogo vale per $\hat{S}_{2LDT}(\cdot)$ $\big($con $Q_{2i} = 1 - R_i + (1 - \pi_z)^{-1} R_i (1 - Z_i)\big)$. Riportiamo in Figura 3.5 le stime delle funzioni di sopravvivenza per le varie linee di trattamento ottenute applicando LDT; sottolineiamo che ogni curva viene quindi stimata considerando il contributo sia di "rispondenti" che di "non rispondenti" a ciascuna terapia. Per verificare l'uguaglianza delle curve tra loro, sia globalmente che a coppie, eseguiamo un test di Wald puntuale (unico disponibile nel pacchetto per gli stimatori LDT); essendo le stime di A_1B_1 e A_1B_2 correlate (così come quelle di A_2B_1 e A_2B_2), dal momento che entrambe usano l'informazione dei pazienti "non rispondenti", il test utilizza la matrice di varianza e covarianza non diagonale tra le stime [8]. In Tabella 3.1 riportiamo solamente i *p-values* risultati significativi ai vari istanti temporali considerati (ci spostiamo di anno in anno all'interno del *range* di riferimento). A discapito di quanto potrebbe sembrare dal grafico, risulta esserci una differenza tra le quattro curve di sopravvivenza (al limite del livello di significatività α) a sei mesi a partire dall'inizio dello studio ($Time = 0.5$); questo perchè l'intervallo

di confidenza di ciascuna curva è più preciso nei primi tempi dello studio (come si può vedere in Figura 3.5), dal momento che gli *standard-errors* delle stime sono più piccoli. Seguendo il metodo LDT, potremmo perciò affermare che a sei mesi

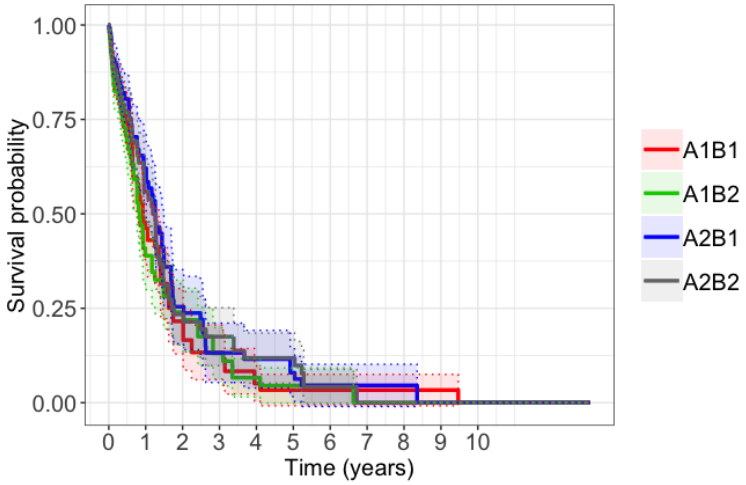


Figura 3.5: Stime delle funzioni di sopravvivenza per le varie *policies* di trattamento, ottenute tramite il metodo LDT ($A1$=ADEP, $A2$=ADE, $B1$=IL-2, $B2$=Placebo).

Policies	$Time = 0.5$	$Time = 4.5$
$A_1B_1 = A_1B_2 = A_2B_1 = A_2B_2$	0.047	
$A_1B_1 = A_2B_2$		0.034
$A_1B_2 = A_2B_1$	0.020	

Tabella 3.1: *p-values* risultati significativi nel testare l'uguaglianza tra le curve di Figura 3.5, a vari istanti temporali.

dall'inizio dello studio abbiano più probabilità di sopravvivere ad LMA i pazienti che ricevono ADE+IL-2 (A_2B_1) rispetto ad ADEP+Placebo (A_1B_2), mentre a quattro anni e mezzo dall'inizio coloro cui viene somministrato ADE+Placebo (A_2B_2) rispetto ad ADEP+IL-2 (A_1B_1); nei restanti istanti temporali seguire una linea di trattamento piuttosto che un'altra non comporta differenze in termini di sopravvivenza.

WRSE (*Weighted Risk Set Estimator*) è uno stimatore non parametrico calcolato a partire dalla (2.1) nel caso di dati censurati; si tratta di uno stimatore maggiormente intuitivo ed efficiente rispetto a LDT. A differenza di LDT, questo stimatore assegna ai soggetti pesi che sono tempo-dipendenti; in particolare, con riferimento ad A_1B_1, si ha:

$$W_{1i}(u) = 1 - R_i(u) + \pi_z^{-1} R_i(u) Z_i,$$

con $R_i(u) = R_i I\big(T_i^R \leq u\big)$. Al paziente i verrà quindi assegnato peso pari a π_z^{-1} se all'istante u risponde ad A_1 $\big(R_i(u) = 1\big)$ e riceve B_1, peso pari a uno se all'istante u non è ancora stata osservata una risposta ad A_1 $\big(R_i(u) = 0\big)$ o se il paziente si è fermato ad A_1 e peso zero se all'istante u risponde ad A_1 e riceve B_2. In Figura 3.6 possiamo osservare le stime delle curve di sopravvivenza per le varie combinazioni $A_j B_k, j, k = 1, 2$, ottenute tramite questo metodo. Come prima cosa, notiamo la differenza di andamento di queste curve rispetto a quelle di Figura 3.5: la probabilità di sopravvivere alla leucemia si differenzia maggiormente tra le *policies* e non si annulla dopo i 10 anni dall'inizio dello studio. A discapito

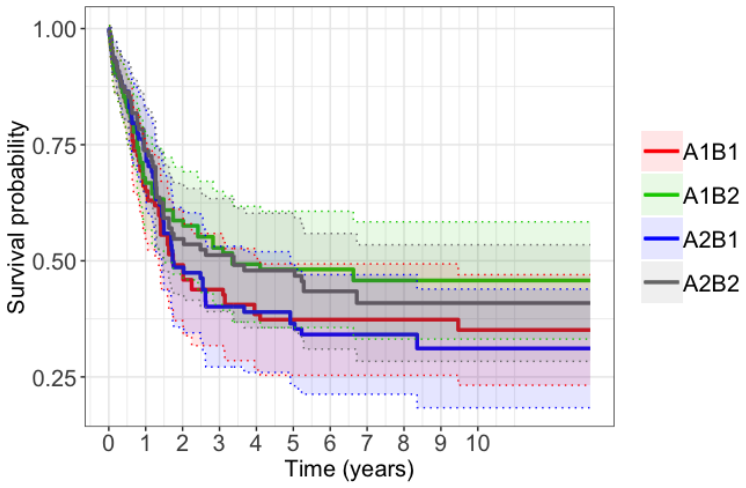


Figura 3.6: Stime delle funzioni di sopravvivenza per le varie *policies* di trattamento, ottenute tramite il metodo WRSE ($A1$=ADEP, $A2$=ADE, $B1$=IL-2, $B2$=Placebo).

di quanto potrebbe sembrare dal grafico, conducendo il test di Wald puntuale si accetta sempre l'ipotesi nulla di uguaglianza della sopravvivenza sotto le diverse linee di trattamento (*p-values* > 0.05), sia globalmente che a coppie; ciò è dovuto ad una maggiore imprecisione delle stime rispetto a LDT (*standard-errors* più elevati), per cui notiamo intervalli di confidenza più ampi nel grafico. Anche conducendo un *logrank test* pesato (test non parametrico disponibile nel pacchetto per lo stimatore WRSE), si accetta sempre l'ipotesi nulla di uguaglianza delle varie curve di sopravvivenza stimate. Seguendo il metodo WRSE, si potrebbe quindi affermare che la probabilità di sopravvivere per i pazienti rimanga la stessa indipendentemente da quale combinazione di trattamenti essi intraprendano.

Un altro stimatore non parametrico della funzione di sopravvivenza, adattato ai disegni a due stadi, che possiamo utilizzare è rappresentato dallo stimatore di Kaplan-Meier. Si tratta del metodo più "semplice", tanto da valergli il nome di stimatore NAIVE; tende però a sovrastimare il contributo dei "non rispondenti" alla distribuzione della sopravvivenza ed è perciò uno stimatore distorto. Osserviamo

le curve di sopravvivenza per le varie *policies* stimate tramite questo metodo (Figura 3.7). Per testare l'ipotesi di uguaglianza di queste curve tra loro, conduciamo un *logrank test* a coppie (codice R in Appendice B.1): in tutti i confronti si accetta l'ipotesi nulla di uguaglianza (p-*values* > 0.05). Stimando la sopravvivenza con il metodo di Kaplan-Meier, si potrebbe perciò affermare che la probabilità di sopravvivenza alla malattia non cambi in base alla linea di trattamento intrapresa.

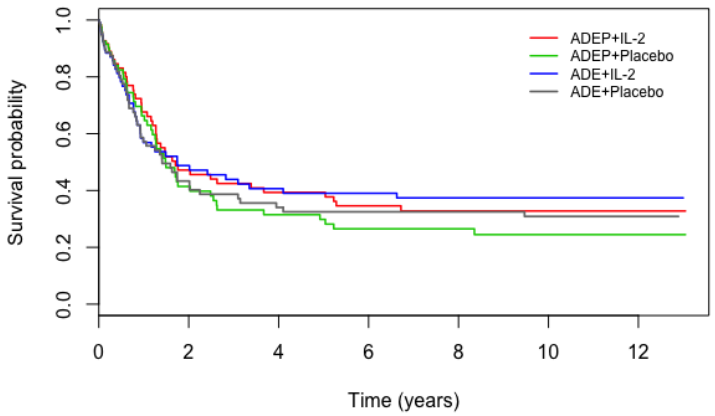


Figura 3.7: Stime delle funzioni di sopravvivenza per le varie *policies* di trattamento, ottenute tramite il metodo di Kaplan-Meier.

Nel capitolo seguente analizziamo i risultati che si ottengono stimando la sopravvivenza per le varie *policies* di trattamento tramite il metodo parametrico descritto in questa tesi.

Capitolo 4

Distribuzioni mistura per l'analisi di sopravvivenza

4.1 Studio di simulazione

Nel corso di questo paragrafo analizzeremo la *performance* del metodo parametrico proposto, confrontandone i risultati con quelli dei metodi non parametrici visti, tramite simulazione dei dati. Le variabili casuali sono rappresentate dai tempi di sopravvivenza potenziali ($T^*_{ki}, k = 0, 1, 2$), dalla durata della terapia di primo stadio per i soggetti che proseguono con il secondo (T^R_i) e dalle variabili relative al tempo di censura (C_i), al tipo di terapia di induzione assegnato (X_i), all'eleggibilità al trattamento di secondo stadio (R_i) e alla tipologia di terapia di mantenimento (Z_i).

Le simulazioni sono costruite in modo tale da imitare il contesto dell'applicazione sui dati del capitolo 3. Stabiliamo perciò una numerosità campionaria pari a 300 osservazioni e generiamo le variabili

$$R_i \sim Bernoulli(0.3) \quad \text{e} \quad Z_i \sim Bernoulli(0.5),$$

per cui avremo un tasso di risposta alla terapia di induzione (π_r) pari al 30% e un tasso di pazienti assegnati a B_1 (π_z) pari al 50%. Imponiamo inoltre che

$$X_i \sim Bernoulli(0.5) \quad \text{e} \quad C_i \sim Unif(0, \gamma).$$

Avremo così che il tempo di censura è distribuito uniformemente tra zero e γ anni (γ scelto in modo da avere circa il 30% dei soggetti censurati) e che il tasso di individui assegnati ad A_1 piuttosto che ad A_2 coincide con il 50% dei pazienti. Nel corso del paragrafo effettueremo però un confronto solamente tra coloro che

ricevono terapia di induzione A_1, sapendo che le stesse analisi valgono per coloro che ricevono A_2, essendo i soggetti nei due gruppi indipendenti.
Per quanto riguarda i tempi di sopravvivenza potenziali e la durata della terapia di primo stadio, possiamo assumerne una qualsiasi forma distributiva continua con supporto positivo; tra le più comuni per i tempi di sopravvivenza ritroviamo le distribuzioni esponenziale, Weibull, Gompertz, Gamma e log-normale [10]. L'ideale sarebbe stato proseguire il paragrafo differenziandolo per almeno due tipologie di distribuzione, ad esempio esponenziale e Weibull, analizzate teoricamente nel capitolo 2. Ricavare, per ogni *dataset* simulato, le convoluzioni e le rispettive funzioni di sopravvivenza tramite la funzione `integrate()` di *RStudio* risultava però di difficile implementazione, per cui si è deciso di concentrarsi solamente sul caso esponenziale.
Una volta effettuata la scelta distributiva, il tempo di sopravvivenza osservato viene calcolato utilizzando la seguente formula (paragrafo 2.1):

$$T_i = (1 - R_i)T_{0i}^* + R_i\big[T_i^R + Z_iT_{1i}^* + (1 - Z_i)T_{2i}^*\big], \quad i = 1, 2, ..., n.$$

La variabile U_i, relativa al tempo di sopravvivenza di ciascun paziente nel caso di dati censurati e rappresentante la nostra "variabile risposta", viene ricavata come valore minimo tra T_i e C_i. Ciascun soggetto avrà così associata una variabile δ_i, pari a uno se ne avviene il decesso o a zero se ne avviene la censura.

Sono state svolte 1000 simulazioni Monte Carlo; le stime derivanti da ognuna delle 1000 simulazioni vengono aggregate in un unico risultato e confrontate tra di loro.

4.1.1 Tempi di sopravvivenza esponenziali

Assumiamo le seguenti distribuzioni per la durata individuale della terapia di induzione e per i tempi di sopravvivenza potenziali

$$T_i^R \sim Exp(5), \quad T_{0i}^* \sim Exp(1), \quad T_{1i}^* \sim Exp(0.125), \quad T_{2i}^* \sim Exp(0.167),$$

e poniamo $\gamma = 7$. Con i tassi imposti, otteniamo che i valori assunti dal vettore di parametri θ, medie di ciascun tempo di sopravvivenza non censurato, sono $\theta_R = 5^{-1} = 0.2$, $\theta_0 = 1$, $\theta_1 = 0.125^{-1} = 8$ e $\theta_2 = 0.167^{-1} = 6$. Per stimare le curve di sopravvivenza $S_1(\cdot)$ e $S_2(\cdot)$ tramite la tecnica parametrica proposta, dovremo ricavare le stime degli elementi di θ, oltre a quella del parametro π_r della distribuzione di R_i; i valori assunti dai parametri di interesse sono quelli in Tabella 4.1.

π_r	θ_R	θ_0	θ_1	θ_2
0.3	0.2	1	8	6

Tabella 4.1: Valori assunti dai parametri di interesse in ogni *dataset* simulato.

In questo specifico contesto il tempo di sopravvivenza osservato T_i viene ricavato come combinazione di variabili casuali esponenziali, più precisamente come mistura delle tre variabili T^*_{0i}, $T^R_i + T^*_{1i}$ e $T^R_i + T^*_{2i}$. Le funzioni di densità delle due "variabili somma" sono ottenute come convoluzione di funzioni di densità esponenziali, la cui forma è chiusa $\big($vedi formule (2.4)$\big)$. In Figura 4.1 possiamo infatti osservare come la densità di T_i (in rosso) risulti essere una combinazione delle densità dei vari tempi potenziali (in verde); notiamo la forma esponenziale della densità di T^*_{0i} e le forme non note delle densità di $T^R_i + T^*_{1i}$ e di $T^R_i + T^*_{2i}$. La figura è stata ottenuta imponendo ampiezza di banda pari a 1 e nucleo gaussiano e prendendo come riferimento un qualsiasi *dataset* simulato.

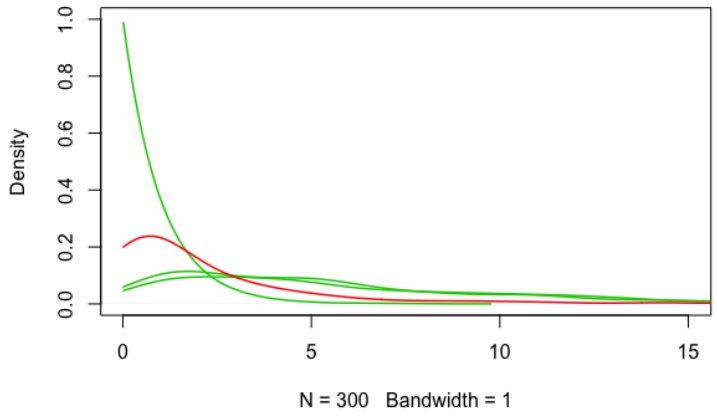


Figura 4.1: Densità degli anni di sopravvivenza dei "non rispondenti" e dei "rispondenti" assegnati a B_1 e a B_2 in verde, densità globale degli anni di sopravvivenza in rosso.

Innanzitutto stimiamo attraverso il metodo parametrico proposto le funzioni di sopravvivenza e i relativi *standard errors* in corrispondenza di due precisi istanti temporali scelti come riferimento: a un anno e a 5 anni dall'inizio dello studio. Per fare ciò dobbiamo prima implementare tramite un algoritmo numerico le stime di θ_R, θ_1 e θ_2 (non disponibili in forma esplicita, come visto nel paragrafo 2.3.1), per

ognuna delle 1000 simulazioni. Il contributo alla funzione di log-verosimiglianza è

$$l_2(\theta_R, \theta_1, \theta_2) \propto \sum_{i=1}^{n} \left[\log \left(\frac{e^{-\frac{u_i}{\theta_R}} - e^{-\frac{u_i}{\theta_1}}}{\theta_R - \theta_1} \right)^{\delta_i r_i z_i} + \log \left(\frac{\theta_R e^{-\frac{u_i}{\theta_R}} - \theta_1 e^{-\frac{u_i}{\theta_1}}}{\theta_R - \theta_1} \right)^{(1-\delta_i) r_i z_i} + \right.$$
$$\left. + \log \left(\frac{e^{-\frac{u_i}{\theta_R}} - e^{-\frac{u_i}{\theta_2}}}{\theta_R - \theta_2} \right)^{\delta_i r_i (1-z_i)} + \log \left(\frac{\theta_R e^{-\frac{u_i}{\theta_R}} - \theta_2 e^{-\frac{u_i}{\theta_2}}}{\theta_R - \theta_2} \right)^{(1-\delta_i) r_i (1-z_i)} \right], \tag{4.1}$$

che, se ottimizzato, permette di ricavare le stime dei parametri richiesti, in quanto non presenta un andamento monotono. In Figura 4.2 possiamo osservare $l_2(\theta_R, \theta_1, \theta_2)$, al variare dei valori di θ_R, θ_1 e θ_2, per un qualsiasi *dataset* simulato e possiamo effettivamente notare la presenza di un unico punto di massimo in ognuno dei tre casi. Le stime vengono ottimizzate tramite l'algoritmo "L-BFGS-B", un metodo quasi-Newton che consente di corredare i parametri con limite

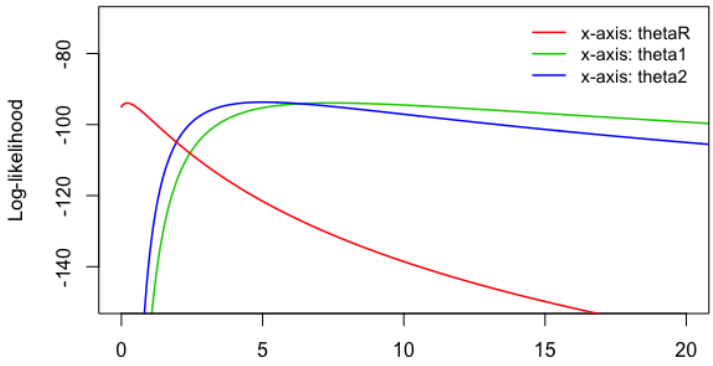


Figura 4.2: Andamento della funzione di log-verosimiglianza fattorizzata al variare dei valori assunti da θ_R, θ_1 e θ_2 (asse delle ascisse).

inferiore e/o superiore del rispettivo spazio parametrico (nel nostro caso equivalente a $\mathbb{R}^+$). Riportiamo in Tabella 4.2 le medie delle stime ottenute per ogni parametro, con il relativo *standard error* medio tra parentesi; l'algoritmo è stato completato correttamente in 994 simulazioni su 1000. Nella stessa tabella riportiamo anche le medie delle stime di π_r e θ_0, disponibili invece in forma chiusa $\left(\hat{\pi}_r = \frac{\sum_i r_i}{300}, \hat{\theta}_0 = \frac{\sum_i (1-r_i) u_i}{\sum_i (1-r_i) \delta_i}\right)$, e i loro *standard errors*. Possiamo vedere come le stime ottenute si avvicinino ai veri valori dei parametri, in particolar modo quelle di π_r, θ_0 e θ_1; notiamo inoltre la presenza di *standard errors* più elevati relativi agli stimatori di θ_1 e θ_2.

$\hat{\pi}_r$	$\hat{\theta}_R$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$
0.299 (0.037)	0.639 (0.634)	0.998 (0.106)	8.043 (3.999)	5.812 (2.638)

Tabella 4.2: Medie delle stime dei parametri di interesse, *standard errors* tra parentesi.

Riusciamo allora a fornire una stima media della probabilità di sopravvivenza a un anno (u_1) e a 5 anni (u_2) dall'inizio dello studio, per i soggetti che seguono le linee di trattamento A_1B_1 e A_1B_2, sostituendo le stime ottenute:

$$\hat{S}_1(u) = S_1(u; \hat{\pi}_r, \hat{\theta}_R, \hat{\theta}_0, \hat{\theta}_1) = 1 - (1 - \hat{\pi}_r)F_0(u; \hat{\theta}_0) - \hat{\pi}_r F_{R1}(u; \hat{\theta}_R, \hat{\theta}_1),$$
$$\hat{S}_2(u) = S_2(u; \hat{\pi}_r, \hat{\theta}_R, \hat{\theta}_0, \hat{\theta}_2) = 1 - (1 - \hat{\pi}_r)F_0(u; \hat{\theta}_0) - \hat{\pi}_r F_{R2}(u; \hat{\theta}_R, \hat{\theta}_2),$$

con $u \in \{u_1, u_2\}$, $F_{R1}(\cdot)$ e $F_{R2}(\cdot)$ funzioni di ripartizione disponibili in forma chiusa $\big($vedi formule (2.6)$\big)$. Riportiamo in Tabella 4.3 le stime ottenute ed i corrispondenti *standard errors* tra parentesi. Per il calcolo degli *standard errors* abbiamo applicato il metodo delta multivariato [18]: a partire dalle distribuzioni degli stimatori di massima verosimiglianza $\hat{\theta}_R$ e $\hat{\theta}_k, k = 0, 1, 2$ (si è deciso di mantenere $\hat{\pi}_r$ costante, in modo da non complicare troppo l'applicazione del metodo), abbiamo calcolato le distribuzioni delle funzioni di sopravvivenza stimate e quindi ne abbiamo ricavato gli *standard errors* (dimostrazione in Appendice A.8).

t	$\hat{S}_1(t)$	$\hat{S}_2(t)$
1	0.535 (0.031)	0.527 (0.033)
5	0.172 (0.047)	0.143 (0.049)

Tabella 4.3: Stime di $S_1(\cdot)$ e $S_2(\cdot)$ a un anno e a 5 anni dall'inizio dello studio (t), *standard errors* tra parentesi.

Abbiamo così che la probabilità di sopravvivenza alla malattia per i soggetti che intraprendono la *policy* A_1B_1 è pari al 53.5% ad un anno dall'inizio dello studio e al 17.2% a 5 anni; per i soggetti assegnati ad A_1B_2 coincide invece con il 52.7% ad un anno dall'inizio dello studio e con il 14.3% a 5 anni. Il codice R con cui abbiamo ottenuto questi risultati è riportato in Appendice B.2.

Per valutare la *performance* di questo metodo, lo confrontiamo con le tecniche non parametriche LDT, WRSE e Kaplan-Meier, studiando le proprietà degli stimatori tramite le seguenti quantità empiriche: distorsione (*bias*), precisione (*standard error*, S.E.) e probabilità di copertura (CP). Andiamo perciò a misurare quanto la stima di ciascuna funzione di sopravvivenza si discosti in media dal vero valore, quanto precise siano queste stime e quanto la probabilità di copertura dell'intervallo di confidenza di Wald si avvicini al valore nominale del 95%. I risultati

sono riportati in Tabella 4.4. Nella colonna "CP%" precisiamo anche, tra parentesi, qual è la percentuale di valori che non è compresa nei rispettivi intervalli di confidenza, presente nella coda sinistra e nella coda destra.

t	Policy	$\hat{S}(t)$	Bias	S.E.	CP%
Metodo parametrico					
1	A_1B_1	0.535	0.006	0.031	89.5 (7.6 sx, 2.9 dx)
	A_1B_2	0.527	0.006	0.033	91.6 (6.3 sx, 2.1 dx)
5	A_1B_1	0.172	0.003	0.047	91.0 (5.9 sx, 3.1 dx)
	A_1B_2	0.143	0.003	0.049	93.5 (3.1 sx, 3.4 dx)
Metodo LDT					
1	A_1B_1	0.461	-0.068	0.047	64.2 (0.7 sx, 35.1 dx)
	A_1B_2	0.454	-0.066	0.046	63.2 (0.4 sx, 36.4 dx)
5	A_1B_1	0.037	-0.132	0.018	18.9 (0.2 sx, 80.9 dx)
	A_1B_2	0.037	-0.103	0.017	23.4 (0.8 sx, 75.8 dx)
Metodo WRSE					
1	A_1B_1	0.522	-0.007	0.047	94.2 (2.3 sx, 3.5 dx)
	A_1B_2	0.531	0.010	0.046	94.8 (4.5 sx, 0.7 dx)
5	A_1B_1	0.147	-0.022	0.047	88.1 (0.4 sx, 11.5 dx)
	A_1B_2	0.174	0.035	0.050	91.1 (7.6 sx, 1.3 dx)
Metodo Kaplan-Meier					
1	A_1B_1	0.527	-0.002	0.060	94.3 (3.3 sx, 2.4 dx)
	A_1B_2	0.520	0.0003	0.060	94.3 (2.4 sx, 3.3 dx)
5	A_1B_1	0.168	-0.001	0.056	92.9 (1.8 sx, 5.3 dx)
	A_1B_2	0.137	-0.003	0.052	92.4 (1.5 sx, 5.9 dx)

Tabella 4.4: Risultati delle simulazioni basate su 1000 campioni Monte Carlo, ognuno di dimensione pari a 300 unità, per i vari metodi di analisi della sopravvivenza.

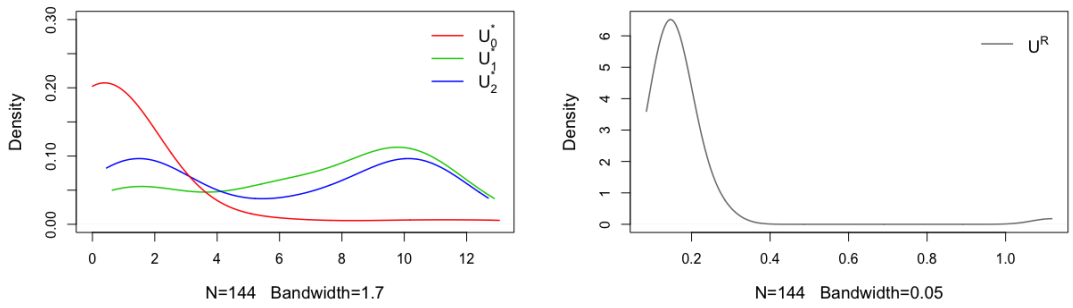
Confrontando tra loro i vari metodi, vediamo come le stime delle funzioni di sopravvivenza ottenute in modo parametrico siano complessivamente più precise di quelle ricavate tramite i metodi WRSE e Kaplan-Meier (*standard errors* leggermente più bassi). Si tratta di un risultato atteso, essendo generalmente i metodi parametrici più precisi rispetto a quelli non parametrici. Sono inoltre, insieme a quelle ricavate tramite Kaplan-Meier, le stime meno distorte, con valori di *bias* tendenti allo zero. Quest'ultimo risultato ci sorprende, in quanto le stime di Kaplan-Meier dovrebbero risultare distorte, come spiegato nel paragrafo 3.2; ciò può essere dipeso da una serie di fattori come gli istanti temporali considerati o il numero di simulazioni effettuate. Le probabilità di copertura per gli intervalli di Wald al 95% delle funzioni di sopravvivenza si avvicinano al valore nominale in tutti i casi (88-94%), eccetto per il metodo LDT, per il quale assumono valori molto bassi. Inoltre, la percentuale di funzioni non compresa nei rispettivi intervalli di confidenza non è

equamente distribuita a destra e a sinistra degli intervalli; in particolare, per quanto riguarda il metodo parametrico si nota un'asimmetria verso sinistra. Tutte le tecniche portano quindi a stime asimmetriche delle funzioni di sopravvivenza.

In conclusione, possiamo affermare che il metodo implementato sembri essere una buona tecnica per stimare la sopravvivenza in disegni randomizzati a due stadi; ricordiamo che questi risultati sono però vincolati dalle forme distributive scelte per le variabili simulate. Non si notano differenze particolari con i metodi di WRSE e Kaplan-Meier, i quali funzionano altrettanto bene. L'unico metodo sconsigliato risulta essere LDT, visti i valori bassi di CP e una forte asimmetria sulla coda destra della distribuzione delle funzioni di sopravvivenza; questa tecnica porta infatti a sottostimare la probabilità di sopravvivere alla malattia (a 5 anni risulterebbe solamente del 3.7%).

4.2 Applicazione ai dati

Vogliamo ora applicare il metodo parametrico proposto al *dataset* relativo allo studio CALGB 19808, analizzarne i risultati e confrontarli con quelli ottenuti tramite i metodi non parametrici. Si confrontano tra loro solo i pazienti che ricevono ADEP come terapia di induzione (144 soggetti), dal momento che le stesse analisi verrebbero svolte per coloro che vengono assegnati ad ADE (148 soggetti).



(a) (b)

Figura 4.3: Densità dei tempi di sopravvivenza potenziali (in anni) (a) e della durata della terapia di induzione (in anni) per i pazienti "rispondenti" (b).

Per stimare le funzioni di sopravvivenza risulta necessario effettuare delle assunzioni distributive sui tempi di sopravvivenza potenziali e sulla durata della terapia di induzione per chi prosegue con la fase di mantenimento; vedremo poi quali di queste portino ad un miglior adattamento del modello ai dati. Prima di fare ciò,

in Figura 4.3 osserviamo le distribuzioni empiriche delle variabili $U^*_{ki}, k = 0, 1, 2$, e U^R_i (stimate imponendo nucleo gaussiano e ampiezza di banda indicata nei grafici).

4.2.1 Assunzione esponenziale sui tempi di sopravvivenza

Supponiamo per semplicità che le variabili $U^*_{ki}, k = 0, 1, 2$, e U^R_i siano tutte casualmente generate da distribuzioni esponenziali, rispettivamente $Exp(\lambda_k)$ e $Exp(\lambda_R)$, di parametri $\theta_k = \lambda_k^{-1}$ e $\theta_R = \lambda_R^{-1}$ non noti.

Per derivare le stime $\hat{S}_1(\cdot)$ e $\hat{S}_2(\cdot)$, dobbiamo prima ottenere le stime dei parametri π_r, θ_R e θ_k. Conosciamo l'espressione delle stime di π_r e θ_0, mentre per ricavare quelle di θ_R, θ_1 e θ_2 massimizziamo la componente $l_2(\cdot)$ della funzione di log-verosimiglianza, con espressione data dalla formula (4.1), tramite un algoritmo di ottimizzazione numerica. L'algoritmo scelto è "L-BFGS-B"; forniti dei valori arbitrari di partenza per i tre parametri e raggiunta la convergenza, l'algoritmo ne produce le stime. Riportiamo in Tabella 4.5 le stime finali dei vari parametri di interesse, corredate dei rispettivi *standard errors* tra parentesi.

$\hat{\pi}_r$	$\hat{\theta}_R$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$
0.313 (0.039)	0.280 (0.240)	2.658 (0.420)	20.842 (7.404)	10.979 (3.341)

Tabella 4.5: Stime dei parametri di interesse, *standard errors* tra parentesi.

Riusciamo così a stimare le funzioni di sopravvivenza associate alle *policies* ADEP+IL-2 $\left(\hat{S}_1(\cdot)\right)$ e ADEP+Placebo $\left(\hat{S}_2(\cdot)\right)$, con i rispettivi intervalli di confidenza. In Figura 4.4 ne osserviamo l'andamento: la probabilità di sopravvivere per gli individui assegnati ad ADEP+IL-2 è sistematicamente più alta, per qualsiasi istante temporale dopo i due anni, della probabilità di sopravvivenza di coloro che ricevono ADE+Placebo; gli intervalli di confidenza delle due curve si sovrappongono. Per confrontarle, conduciamo un test del rapporto di verosimiglianza per testare l'ipotesi nulla $H_0 : \theta_1 = \theta_2$; questi parametri rappresentano infatti l'unico elemento non in comune tra le due funzioni di sopravvivenza, per cui un'eventuale accettazione di H_0 comporterebbe l'uguaglianza delle due curve. La statistica test è pari a 3.69, con *p-value* associato uguale a 0.055. Non abbiamo perciò abbastanza evidenza per accettare l'ipotesi nulla e per poter di conseguenza affermare che la probabilità di sopravvivere alla malattia sia la stessa indipendentemente dalla *policy* intrapresa.

Per valutare la bontà di adattamento di questo modello, confrontiamo i risultati ottenuti con quelli dei metodi WRSE e Kaplan-Meier (paragrafo 3.2), presi come riferimento; tralasciamo il metodo LDT, avendo visto dalle simulazioni es-

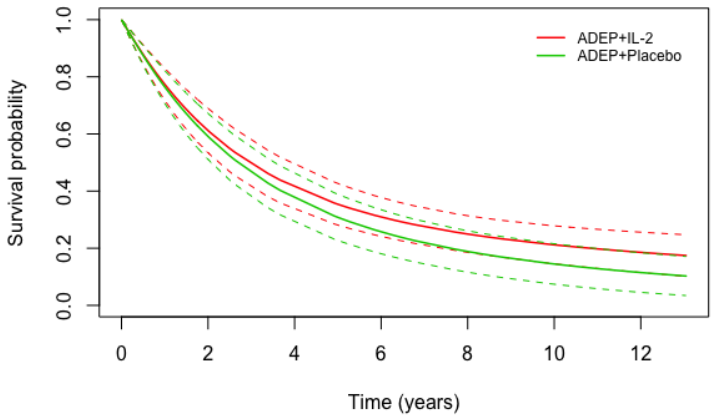
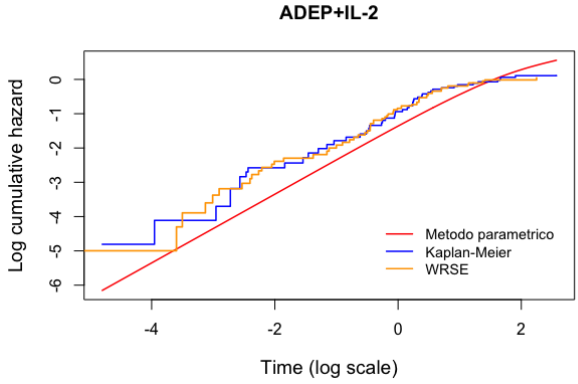


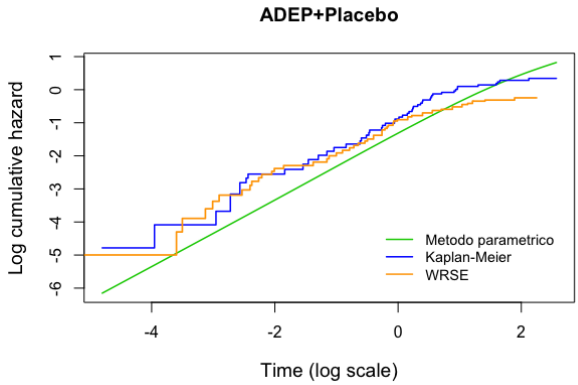
Figura 4.4: Stime delle funzioni di sopravvivenza e relativi intervalli di confidenza (linee tratteggiate) per le *policies* di trattamento ADEP+IL-2 e ADEP+Placebo, ottenuti tramite il metodo parametrico.

sere quello che funziona peggio. Decidiamo di confrontare, anzichè le curve di sopravvivenza, il logaritmo delle funzioni di rischio cumulativo stimate dai vari metodi, per un paragone di maggior impatto visivo grazie alla relazione di quasi linearità con il logaritmo del tempo di sopravvivenza. La funzione di rischio cumulativo, espressa come $\Lambda(t) = -\log S(t)$, rappresenta il rischio cumulato di non sopravvivere alla malattia nell'intervallo $(0, t]$. Da una valutazione grafica delle Figure 4.5a e 4.5b, notiamo come $\Lambda(t)$ stimata con il metodo parametrico non si sovrapponga alle funzioni stimate con le tecniche non parametriche, per entrambe le combinazioni di trattamenti; in particolare il modello esponenziale tende a sottostimare il rischio cumulativo di non sopravvivere alla leucemia.

L'assunzione esponenziale per i tempi di sopravvivenza potenziali e per la durata della terapia di induzione non sembra perciò portare ad un modello che si adatti particolarmente bene ai dati in esame. L'AIC (*Akaike Information Criteria*) associato a questo modello è inoltre pari a 306.05.



(a)



(b)

Figura 4.5: Confronto tra le funzioni di rischio cumulativo stimate con il metodo parametrico e con i metodi WRSE e Kaplan-Meier, per le *policies* ADEP+IL-2 (a) ADEP+Placebo (b).

4.2.2 Assunzioni miste sui tempi di sopravvivenza

Conduciamo ora un'analisi esplorativa per capire quale forma distributiva si avvicini maggiormente alla distribuzione empirica delle variabili $U^*_{ki}, k = 0, 1, 2$, e U^R_i, al fine di stimare un modello che abbia un adattamento migliore ai dati in esame. Per fare ciò analizziamo quale distribuzione di ciascuna variabile porti alla stima della funzione di rischio cumulativo che più si avvicini a quella stimata tramite Kaplan-Meier, metodo preso come riferimento. Le distribuzioni che decidiamo di considerare sono l'esponenziale, la Weibull e la log-normale. Riportiamo i grafici per ciascuna variabile nelle Figure 4.6, 4.7a, 4.7b e 4.8 (codice R di esempio in Appendice B.3). Da un'analisi di Figura 4.6, deduciamo che le funzioni di rischio cumulativo che più si avvicinano a quella prodotta dal metodo di Kaplan-Meier sono quelle stimate tramite la distribuzione Weibull e la distribuzione log-normale; optiamo per la distribuzione Weibull per il semplice fatto che per tempi molto

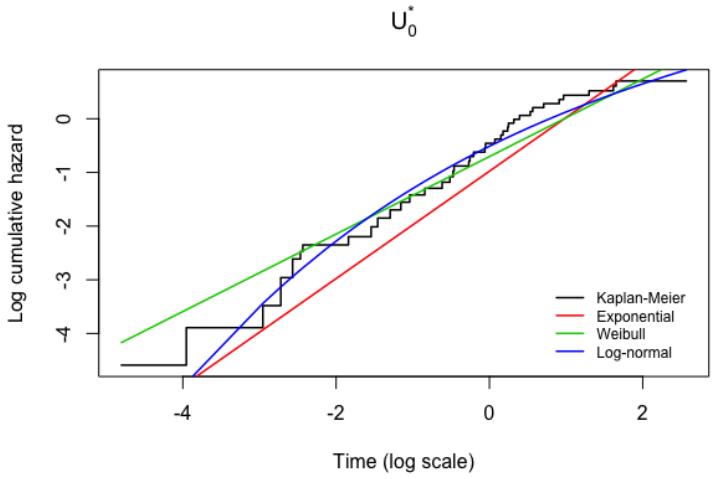
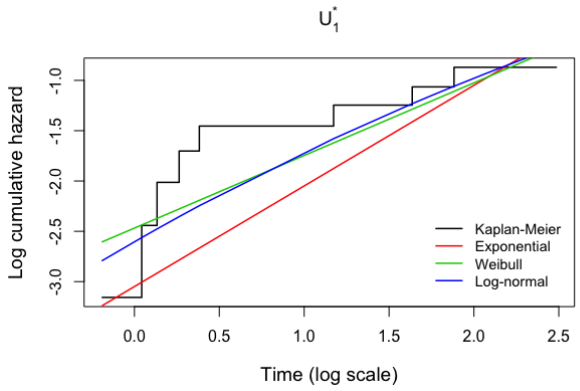


Figura 4.6: Confronto tra le funzioni di rischio cumulativo stimate imponendo varie distribuzioni a U_{0i}^* con quella stimata tramite il metodo di Kaplan-Meier.

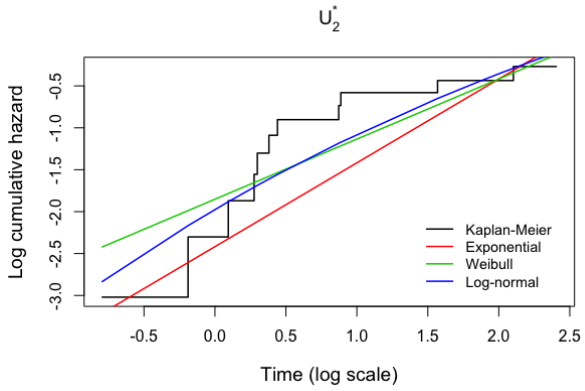
bassi non tende a sottostimare il rischio di non sopravvivere alla malattia. Assumiamo perciò che U_{0i}^*, tempo di sopravvivenza dei pazienti "non rispondenti", sia realizzazione casuale di una variabile Weibull, $Weibull(\alpha_0, \lambda_0)$, di parametri α_0 e λ_0 ignoti. Dalle Figure 4.7a, 4.7b e 4.8 notiamo invece come la funzione di rischio cumulativo stimata tramite Kaplan-Meier si allontani da una distribuzione lineare; risulta quindi più difficile valutare l'adattamento delle funzioni stimate tramite le varie assunzioni distributive. Da un'analisi grafica che perciò lascia qualche dubbio, ipotizziamo che la distribuzione più adatta per le variabili U_{1i}^*, U_{2i}^* e U_i^R sia la log-normale[1]; queste variabili rappresentano rispettivamente il tempo di sopravvivenza, dall'inizio del secondo stadio, dei soggetti assegnati a IL-2, il tempo di sopravvivenza, dall'inizio del secondo stadio, dei soggetti assegnati a Placebo e la loro durata della terapia di primo stadio. Assumiamo quindi che queste variabili siano realizzazioni casuali di variabili log-normali, rispettivamente $logN(\mu_1, \sigma_1), logN(\mu_2, \sigma_2)$ e $logN(\mu_R, \sigma_R)$, di parametri ovviamente non conosciuti.

Per stimare le funzioni di sopravvivenza associate alle *policies* ADEP+IL-2 e ADEP+Placebo, dobbiamo prima ricavare le stime dei nove parametri di interesse. Massimizzando la funzione di verosimiglianza (2.5), calcoliamo le stime di π_r e di λ_0 (forma esplicita nel paragrafo 2.3.2); per ottenere la stima di α_0 applichiamo l'algoritmo di ottimizzazione "L-BFGS-B". Analogamente avviene per ottenere le

[1]La distribuzione log-normale, $X \sim logN(\mu, \sigma)$, di parametri $\mu \in \mathbb{R}$ e $\sigma \geq 0$ e supporto $\mathbb{R}_0^+$, è tale che il suo logaritmo segua la distribuzione normale $N(\mu, \sigma^2)$. I parametri μ e σ rappresentano perciò la media e la deviazione standard del logaritmo della variabile X.



(a)



(b)

Figura 4.7: Confronto tra le funzioni di rischio cumulativo stimate imponendo varie distribuzioni a U^*_{1i} (a) e U^*_{2i} (b) con quella stimata tramite il metodo di Kaplan-Meier.

stime dei parametri associati alle distribuzioni log-normali: la convoluzione delle densità di due variabili log-normali non è calcolabile in forma chiusa tramite l'integrale (2.3) [3], per cui ci serviamo del calcolo numerico tramite R per ricavarla; successivamente troviamo $\hat{\mu}_R, \hat{\sigma}_R, \hat{\mu}_1, \hat{\sigma}_1, \hat{\mu}_2$ e $\hat{\sigma}_2$ utilizzando l'algoritmo di ottimizzazione numerica. Per il calcolo delle funzioni di sopravvivenza delle "variabili somma", complemento delle rispettive funzioni di ripartizione, si è utilizzata la seguente formula (esempio relativo a $U^R_i + U^*_{1i}$) [14] :

$$S_{R1}(u_i) = 1 - F_{R1}(u_i) = 1 - \int_0^{u_i} f_R(j) F_1(u_i - j) dj.$$

Le stime di massima verosimiglianza finali sono riportate in Tabella 4.6, con i rispettivi *standard errors* tra parentesi. Osserviamo la presenza di *standard errors* più elevati per le stime di μ_1, σ_1, μ_2 e σ_2.

Riusciamo allora ad ottenere le funzioni di sopravvivenza stimate $\hat{S}_1(\cdot)$ e $\hat{S}_2(\cdot)$ (il

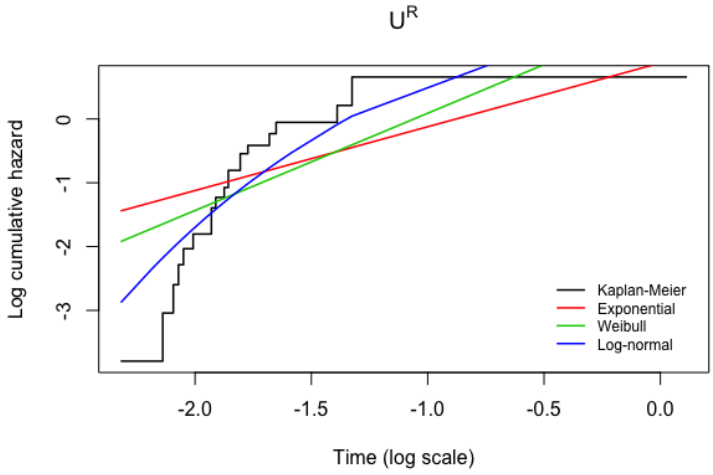


Figura 4.8: Confronto tra le funzioni di rischio cumulativo stimate imponendo varie distribuzioni a U_i^R con quella stimata tramite il metodo di Kaplan-Meier.

$\hat{\pi}_r$	$\hat{\alpha}_0$	$\hat{\lambda}_0$	$\hat{\mu}_R$	$\hat{\sigma}_R$	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\mu}_2$	$\hat{\sigma}_2$
0.313	0.713	2.665	0.157	0.363	4.558	5.723	1.832	5.416
(0.039)	(0.074)	(0.591)	(0.156)	(0.118)	(2.268)	(3.146)	(1.813)	(3.583)

Tabella 4.6: Stime dei parametri di interesse, *standard errors* tra parentesi.

codice R usato è riportato in Appendice B.4), espresse dalle formule

$$\hat{S}_1(u) = 1 - (1 - \hat{\pi}_r)\left[1 - exp\left(- \left(\frac{u}{\hat{\lambda}_0}\right)^{\hat{\alpha}_0}\right)\right] - \hat{\pi}_r F_{R1}(u; \hat{\mu}_R, \hat{\sigma}_R, \hat{\mu}_1, \hat{\sigma}_1),$$

$$\hat{S}_2(u) = 1 - (1 - \hat{\pi}_r)\left[1 - exp\left(- \left(\frac{u}{\hat{\lambda}_0}\right)^{\hat{\alpha}_0}\right)\right] - \hat{\pi}_r F_{R2}(u; \hat{\mu}_R, \hat{\sigma}_R, \hat{\mu}_2, \hat{\sigma}_2);$$

le funzioni di ripartizione di $U_i^R + U_{1i}^*$ e $U_i^R + U_{2i}^*$, somme di variabili log-normali, non sono ottenibili in forma chiusa. Le due curve stimate sono riportate in Figura 4.9. Dal grafico, con un andamento molto simile a quello del modello esponenziale, sembrerebbe che la probabilità di sopravvivere alla leucemia sia maggiore per coloro che intraprendono la linea di trattamento ADEP+IL-2. Conduciamo allora due test del rapporto di verosimiglianza per testare rispettivamente l'ipotesi nulla $H_{01}: \sigma_1 = \sigma_2$, con μ_1 e μ_2 diversi, e l'ipotesi nulla $H_{02}: \mu_1 = \mu_2$, con σ_1 e σ_2 diversi. Entrambe le statistiche test si distribuiscono come un chi quadrato con un grado di libertà. Il *p-value* ottenuto per verificare H_{01} è uguale a 0.902, mentre quello per verificare H_{02} a 0.050. Possiamo quindi affermare che le deviazioni standard associate alle due curve stimate possono essere assunte uguali, mentre

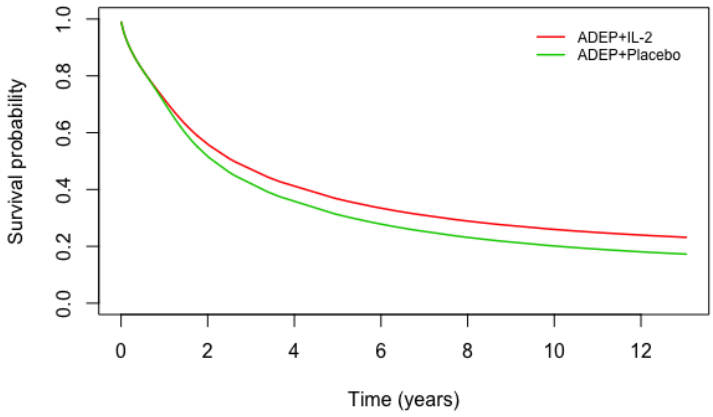
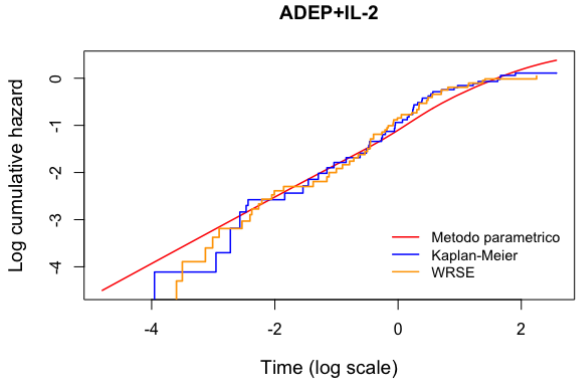


Figura 4.9: Stime delle funzioni di sopravvivenza per le *policies* di trattamento ADEP+IL-2 e ADEP+Placebo, ottenute tramite il metodo parametrico.

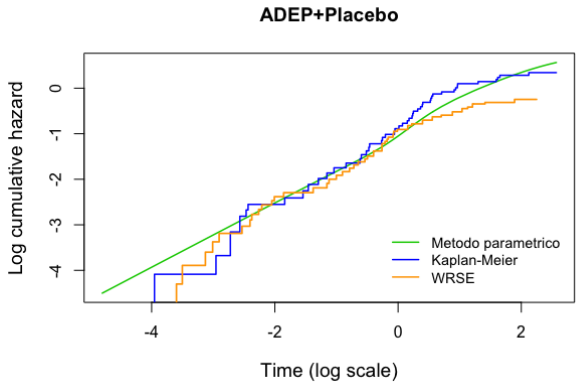
non c'è evidenza per dire altrettanto sulle medie. Conduciamo allora un altro test del rapporto di verosimiglianza per testare $H_0 : \mu_1 = \mu_2$ sotto l'assunzione che σ_1 e σ_2 siano uguali. La statistica test è pari a 3.32, con *p-value* 0.068. Possiamo perciò accettare H_0 e assumere che le due curve di sopravvivenza siano uguali: condurre una linea di trattamento piuttosto che un'altra non comporta differenze in termini di allungamento medio della vita.

Per verificare la bontà di adattamento del modello stimato, confrontiamo il logaritmo della funzione di rischio cumulativo con quello ottenuto tramite i metodi non parametrici WRSE e Kaplan-Meier, così come abbiamo fatto per il caso esponenziale. In Figura 4.10a osserviamo il confronto per la linea di trattamento ADEP+IL-2: l'adattamento ai dati sembra essere buono, dal momento che la funzione di rischio cumulativo tende a sovrapporsi a quelle prese come riferimento. Per la *policy* ADEP+Placebo (Figura 4.10b), la funzione stimata si allinea maggiormente a quella del metodo di Kaplan-Meier piuttosto che a quella di WRSE; in ogni caso l'adattamento risulta migliore rispetto a quello del precedente modello esponenziale.

Le specifiche assunzioni fatte sui tempi di sopravvivenza potenziali e sulla durata individuale della terapia di induzione (distribuzione Weibull per U_{0i}^*, distribuzione log-normale per U_i^R, U_{1i}^* e U_{2i}^*) sembrano perciò portare ad un adattamento del modello ai dati migliore rispetto a quello del modello esponenziale, come si può vedere confrontando i grafici delle funzioni di rischio cumulativo dei due casi. L'AIC associato a questo modello è pari a 283.60; essendo questo un indice di bontà di adattamento ai dati ed essendo la regola quella di preferire il modello



(a)



(b)

Figura 4.10: Confronto tra le funzioni di rischio cumulativo stimate con il metodo parametrico e con i metodi WRSE e Kaplan-Meier, per le *policies* ADEP+IL-2 (a) e ADEP+Placebo (b).

con AIC più basso, concludiamo quindi che il modello appena stimato è migliore rispetto a quello del paragrafo 4.2.1. Il risultato è perciò in linea con quanto si sperava di ottenere. Di certo l'adattamento del modello ai dati potrebbe essere ulteriormente migliorato, in particolare in corrispondenza di valori bassi del tempo di sopravvivenza, scegliendo delle distribuzioni parametriche diverse e più flessibili.

Conclusioni

Nel corso di questo lavoro è stato implementato un metodo parametrico per studiare la sopravvivenza di pazienti trattati con disegni clinici randomizzati a due stadi. Si tratta di un modello mistura, combinazione delle funzioni di densità dei tempi di sopravvivenza dei soggetti "non rispondenti", "rispondenti" assegnati alla terapia di secondo stadio B_1 e "rispondenti" assegnati alla terapia di secondo stadio B_2. Il modello è una trasformazione di un modello già esistente (Wahed, 2010), il quale non tiene in considerazione la durata variabile del trattamento di primo stadio per i pazienti che proseguono con la fase di mantenimento. Si basa sulla teoria dei tempi di sopravvivenza controfattuali (potenziali) espressa da Lunceford et al. (2002), secondo cui il tempo di sopravvivenza osservato per ogni soggetto può essere calcolato come combinazione dei rispettivi tempi controfattuali; da questo concetto ne deriva quello di modello mistura. In base alle assunzioni distributive sui tempi di sopravvivenza potenziali, cambierà perciò la distribuzione della mistura e di conseguenza la bontà di adattamento del modello ai dati in esame. Per questo motivo, un lavoro esplorativo importante consiste nello scegliere la distribuzione che meglio si adatti alla densità empirica di ciascun tempo di sopravvivenza e che produca stime il più possibile in linea con quelle dei metodi non parametrici per disegni a due stadi presi come riferimento. I risultati ottenuti nei dati relativi ad uno studio clinico in pazienti con leucemia mieloide acuta confermano ciò: l'adattamento del modello stimato imponendo ad ogni tempo di sopravvivenza la distribuzione più adatta, scelta dopo un'attenta analisi grafica, è risultato migliore rispetto a quello stimato assumendo per semplicità la distribuzione esponenziale per ciascun tempo di sopravvivenza. Il modello con assunzioni miste risulta infatti avere AIC più basso e produce stime che si avvicinano maggiormente a quelle prodotte dalle tecniche prese come riferimento.

In questa tesi si è deciso di studiare un metodo parametrico in quanto tutte le principali tecniche di analisi di sopravvivenza per disegni randomizzati a due stadi sono non parametriche. Si è voluto quindi valutare la *performance* di un metodo che può essere più intuitivo e facile da implementare in pratica rispetto ad uno non parametrico, anche se a vantaggio di quest'ultimo ne rimane la maggior

flessibilità. Abbiamo perciò confrontato il modello mistura con due tecniche non parametriche prese come riferimento: i metodi WRSE e Kaplan-Meier. Ne è emerso che il metodo implementato gode di buone proprietà, molto simili a quelle dei metodi di riferimento, per cui potrebbe esserne consigliato l'utilizzo come tecnica di stima parametrica della sopravvivenza in disegni a due stadi. In particolare, esso produce stimatori delle funzioni di sopravvivenza non distorti e precisi. Risulta però, come già detto, vincolato dalle assunzioni distributive scelte per i tempi di sopravvivenza.
I risultati ottenuti nel *dataset* analizzato sono in linea con quanto prodotto dalle tecniche WRSE e Kaplan-Meier: facendo riferimento al modello con adattamento ai dati migliore, non risultano esserci differenze in termini di sopravvivenza tra coloro che seguono la *policy* ADEP+IL-2 (A_1B_1) e coloro che seguono la *policy* ADEP+Placebo (A_1B_2). Ricevere l'immunoterapia IL-2 piuttosto che il placebo, a seguito del completamento del trattamento di primo stadio, non comporta un allungamento medio della sopravvivenza; IL-2 non risulta perciò essere una terapia di mantenimento efficace. L'analisi per coloro che ricevono la terapia di induzione ADE (A_2) può essere fatta separatamente e allo stesso modo, essendo i dati dei due gruppi di primo stadio indipendenti. Secondo i metodi non parametrici applicati, non c'è differenza nella probabilità di sopravvivere alla malattia neppure tra le linee di trattamento ADE+IL-2 (A_2B_1) e ADE+Placebo (A_2B_2) e nemmeno complessivamente confrontando tra loro le quattro *policies*.

Analisi più approfondite della tecnica implementata potrebbero riguardare la scelta di altre assunzioni distributive per i tempi di sopravvivenza oltre a quelle da noi considerate (esponenziale, Weibull e log-normale). Utilizzando il pacchetto `flexsurv()` di *RStudio* possono essere prese in considerazione anche le distribuzioni Gompertz, Gamma e Gamma Generalizzata, le quali potrebbero portare ad un adattamento del modello ai dati ancora migliore rispetto a quello qui ottenuto. Un problema che sussiste riguarda la forma non chiusa della maggior parte delle convoluzioni, ossia delle funzioni di densità della somma di variabili, qui rappresentate dalla durata della terapia di primo stadio per i pazienti "rispondenti" e dal loro tempo di sopravvivenza a partire dal secondo stadio. Esse possono essere ricavate tramite calcolo numerico utilizzando un *software* statistico, anche se non sempre ciò può risultare automatico; si potrebbero anche usare dei pacchetti specifici, adatti a ricavare la funzione di densità della somma di due variabili.
Inoltre, per valutare in maniera maggiormente accurata la *performance* del modello proposto, converrebbe effettuare un numero maggiore di simulazioni Monte Carlo (almeno 2000), tenendo presente l'onere computazionale che ne deriva, e considerare più istanti temporali diversi per il confronto tra le funzioni di

sopravvivenza stimate.

Appendice A

Dimostrazione formule

A.1 Funzione di densità di T_i

Per facilitare la dimostrazione si scrivono le distribuzioni condizionate del tempo dati R_i e Z_i come funzioni di probabilità, pur sapendo che ciò è inappropriato essendo il tempo una variabile casuale continua.

$$\begin{aligned} f(t_i;\pi,\theta) &= \sum_{r_i=0,1} Pr(T_i = t_i | R_i = r_i) Pr(R_i = r_i) \\ &= Pr(R_i = 0) Pr(T_{0i}^* = t_i) + Pr(R_i = 1) Pr\Big(\{Z_i T_{1i}^* + (1 - Z_i) T_{2i}^*\} = t_i\Big) \\ &= (1 - \pi_r) f_0(t_i;\theta_0) + \pi_r \sum_{z_i=0,1} Pr\Big(\{Z_i T_{1i}^* + (1 - Z_i) T_{2i}^*\} = t_i | Z_i = z_i\Big) Pr(Z_i = z_i) \\ &= (1 - \pi_r) f_0(t_i;\theta_0) + \pi_r \pi_z f_1(t_i;\theta_1) + \pi_r (1 - \pi_z) f_2(t_i;\theta_2) \end{aligned}$$

A.2 Funzioni di sopravvivenza

$$\begin{aligned} S_1(t;\pi_r,\theta_0,\theta_1) = 1 - F_1(t;\pi_r,\theta_0,\theta_1) &= 1 - \int_0^t f_1(j;\pi_r,\theta_0,\theta_1) dj \\ &= 1 - \int_0^t \Big[(1 - \pi_r) f_0(j;\theta_0) + \pi_r f_1(j;\theta_1)\Big] dj \\ &= 1 - (1 - \pi_r) F_0(t;\theta_0) - \pi_r F_1(t;\theta_1), \end{aligned}$$

con $f_1(\cdot;\pi_r,\theta_0,\theta_1)$ e $F_1(\cdot;\pi_r,\theta_0,\theta_1)$ rispettivamente funzione di densità e funzione di ripartizione della variabile *outcome* potenziale T_1, se tutti i pazienti fossero quindi stati assegnati ad A_1B_1.
La dimostrazione è analoga per $S_2(\cdot;\pi_r,\theta_0,\theta_2)$.

A.3 Funzione di verosimiglianza per dati non censurati

Per il contributo $i-esimo$ alla funzione di verosimiglianza, vale che:

$$
\begin{aligned}
L_i(\pi,\theta;d_i) = f(d_i;\pi,\theta) &= Pr(R_i = r_i, R_iZ_i = r_iz_i, T_i = t_i) \\
&= \begin{cases} (1-\pi_r)f_0(t_i;\theta_0) & \text{se } d_i = (0,0,t_i) \quad (\text{ossia } R_i = 0) \\ \pi_r\pi_z f_1(t_i;\theta_1) & \text{se } d_i = (1,1,t_i) \quad (\text{ossia } R_i = 1, Z_i = 1) \\ \pi_r(1-\pi_z)f_2(t_i;\theta_2) & \text{se } d_i = (1,0,t_i) \quad (\text{ossia } R_i = 1, Z_i = 0) \end{cases} \\
&= (1-\pi_r)^{1-r_i}(\pi_r\pi_z)^{r_iz_i}\Big[\pi_r(1-\pi_z)\Big]^{r_i(1-z_i)} f_0(t_i;\theta_0)^{1-r_i} f_1(t_i;\theta_1)^{r_iz_i} f_2(t_i;\theta_2)^{r_i(1-z_i)}
\end{aligned}
$$

Perciò si ottiene:

$$
\begin{aligned}
L(\pi,\theta;d) = \prod_{i=1}^{n} L_i(\pi,\theta;d_i) =& (1-\pi_r)^{\sum_i(1-r_i)}\ \pi_r^{\sum_i r_i}\ \pi_z^{\sum_i r_iz_i}\ (1-\pi_z)^{\sum_i r_i(1-z_i)} \\
&\times \prod_{i=1}^{n}\Big[f_0(t_i;\theta_0)^{1-r_i} f_1(t_i;\theta_1)^{r_iz_i} f_2(t_i;\theta_2)^{r_i(1-z_i)}\Big] \\
=& L_1(\pi;r,rz)\cdot L_2(\theta;d)
\end{aligned}
$$

A.4 Funzione di verosimiglianza per dati censurati

Probabilità di osservare un evento:
$Pr(U_i = u_i, \Delta_i = 1) = Pr(U_i = u_i|\Delta_i = 1)Pr(\Delta_i = 1) =$
$= f_T(t_i;\theta)Pr(T_i \le C_i) = f_T(t_i;\theta)Pr(C_i \ge T_i) = f_T(t_i;\theta)\big(1 - F_C(t_i;\gamma)\big) =$
$= f_T(t_i;\theta)S_C(t_i;\gamma),$
con $F_C(t_i;\gamma)$ e $S_C(t_i;\gamma)$ funzione di ripartizione e funzione di sopravvivenza della variabile C_i nel punto t_i, assunto che C_i abbia una determinata distribuzione $f_C(\cdot;\gamma)$; $f_T(\cdot;\theta)$ funzione di densità di T_i.

Probabilità di osservare una censura:
$Pr(U_i = u_i, \Delta_i = 0) = Pr(U_i = u_i|\Delta_i = 0)Pr(\Delta_i = 0) =$
$= f_C(c_i;\gamma)Pr(T_i > C_i) = f_C(c_i;\gamma)\big(1 - F_T(c_i;\theta)\big) = f_C(c_i;\gamma)S_T(c_i;\theta).$

La funzione di verosimiglianza in funzione del parametro θ è quindi pari a

$$
\begin{aligned}
L(\theta; u, \delta) &= \prod_{i \in E} \Big[f_T(t_i;\theta) S_C(t_i;\gamma) \Big] \prod_{i \in C} \Big[f_C(c_i;\gamma) S_T(c_i;\theta) \Big] \\
&= \prod_{i=1}^{n} \Big[f_T(t_i;\theta)^{\delta_i} S_C(t_i;\gamma)^{1-\delta_i} f_C(c_i;\gamma)^{\delta_i} S_T(c_i;\theta)^{1-\delta_i} \Big] \\
&\propto \prod_{i=1}^{n} \Big[f_T(t_i;\theta)^{\delta_i} S_T(c_i;\theta)^{1-\delta_i} \Big] = \prod_{i=1}^{n} \Big[f(u_i;\theta)^{\delta_i} S(u_i;\theta)^{1-\delta_i} \Big],
\end{aligned}
$$

con E l'insieme dei soggetti per i quali si è osservato l'evento e C l'insieme delle osservazioni censurate. Nel calcolo di $L(\pi, \theta; d)$ viene considerata solo l'espressione a cui la verosimiglianza in funzione di θ è proporzionale; in questo senso abbiamo a che fare con una funzione di verosimiglianza parziale.

A.5 Equazioni di verosimiglianza relative a $L_2(\theta; d)$

Assunto $U^*_{ki} \sim Exp(\lambda_k), k = 0, 1, 2$, con $\lambda_k = \theta_k^{-1}$, abbiamo che

$$
\begin{aligned}
L_2(\theta; d) \propto \quad & \theta_0^{-\sum_i \delta_i (1-r_i)} \;\; \theta_1^{-\sum_i \delta_i r_i z_i} \;\; \theta_2^{-\sum_i \delta_i r_i (1-z_i)} \\
& \times exp\Bigg\{ -\sum_{i=1}^{n} \Big[\theta_0^{-1} u_i (1-r_i) + \theta_1^{-1} u_i r_i z_i + \theta_2^{-1} u_i r_i (1-z_i) \Big] \Bigg\}.
\end{aligned}
$$

Ricaviamo l'equazione di verosimiglianza relativamente al parametro θ_0; lo stesso ragionamento vale per θ_1 e θ_2. Dal momento che si riesce a separare $L_2(\cdot)$ in tre fattori, ognuno funzione di un parametro θ_k, calcoliamo la derivata prima rispetto a θ_0 della componente contenente solo il parametro stesso, ottenendo la seguente equazione di verosimiglianza:

$$
-\theta_0^{-1} \sum_{i=1}^{n} \delta_i (1 - r_i) + \theta_0^{-2} \sum_{i=1}^{n} u_i (1 - r_i) = 0
$$

Da qui si ricava l'espressione per lo stimatore di massima verosimiglianza di θ_0 ottenuta nel paragrafo 1.3.1.

A.6 Funzione di densità di T_i (durata di primo stadio variabile)

$$
\begin{aligned}
f(t_i;\pi,\theta) &= \sum_{r_i=0,1} Pr(T_i = t_i | R_i = r_i) Pr(R_i = r_i) \\
&= (1-\pi_r) f_0(t_i;\theta_0) + \pi_r \sum_{z_i=0,1} Pr\Big(\{T_i^R + Z_i T_{1i}^* + (1-Z_i) T_{2i}^*\} = t_i | Z_i = z_i\Big) \\
&\quad \times Pr(Z_i = z_i) \\
&= (1-\pi_r) f_0(t_i;\theta_0) + \pi_r \pi_z \int_0^{t_i} f_R(j;\theta_R) f_1(t_i - j;\theta_1) dj + \\
&\quad + \pi_r (1-\pi_z) \int_0^{t_i} f_R(j;\theta_R) f_2(t_i - j;\theta_2) dj \\
&= (1-\pi_r) f_0(t_i;\theta_0) + \pi_r \pi_z f_{R1}(t_i;\theta_R,\theta_1) + \pi_r (1-\pi_z) f_{R2}(t_i;\theta_R,\theta_2)
\end{aligned}
$$

La funzione di densità di $T_i^R + T_{1i}^*$, $f_{R1}(\cdot)$, e quella di $T_i^R + T_{2i}^*$, $f_{R2}(\cdot)$, sono rispettivamente date dalla convoluzione di $f_R(\cdot)$ con $f_1(\cdot)$ e di $f_R(\cdot)$ con $f_2(\cdot)$.

A.7 Funzioni di densità di $T_i^R + T_{1i}^*$ e $T_i^R + T_{2i}^*$

Assunto $T_{ki}^* \sim Exp(\lambda_k), k = 0, 1, 2$, con $\lambda_k = \theta_k^{-1}$, e $T_i^R \sim Exp(\lambda_R)$, con $\lambda_R = \theta_R^{-1}$, abbiamo che

$$
\begin{aligned}
f_{R1}(t_i;\theta_R,\theta_1) &= \int_0^{t_i} f_R(j;\theta_R) f_1(t_i - j;\theta_1) dj = \int_0^{t_i} \theta_R^{-1} exp\Big(-\frac{j}{\theta_R}\Big) \theta_1^{-1} exp\Big(-\frac{t_i - j}{\theta_1}\Big) dj \\
&= \theta_R^{-1}\theta_1^{-1} exp\Big(-\frac{t_i}{\theta_1}\Big) \int_0^{t_i} exp\Big(\frac{\theta_R - \theta_1}{\theta_R\theta_1} j\Big) dj \\
&= \theta_R^{-1}\theta_1^{-1} exp\Big(-\frac{t_i}{\theta_1}\Big) \left[\frac{\theta_R\theta_1}{\theta_R - \theta_1} exp\Big(\frac{\theta_R - \theta_1}{\theta_R\theta_1} j\Big) \Big|_0^{t_i} \right] = \frac{e^{-t_i\theta_R^{-1}} - e^{-t_i\theta_1^{-1}}}{\theta_R - \theta_1}.
\end{aligned}
$$

Analogamente si ottiene $f_{R2}(t_i;\theta_R,\theta_2) = \frac{exp\left(-t_i\theta_R^{-1}\right) - exp\left(-t_i\theta_2^{-1}\right)}{\theta_R - \theta_2}$.

A.8 *Standard errors* funzioni di sopravvivenza: metodo delta multivariato

Metodo delta multivariato. *Si supponga che* $Y_n = (Y_{n1}, ..., Y_{nk})$ *sia una sequenza di vettori casuali tali che*

$$\sqrt{n}(Y_n - \mu) \dot{\sim} N(0, \Sigma).$$

Sia $g : \mathbb{R}^k \to \mathbb{R}$ *e sia*

$$\nabla g(y) = \begin{pmatrix} \frac{\partial g}{\partial y_1} \\ \vdots \\ \frac{\partial g}{\partial y_k} \end{pmatrix}.$$

Sia ∇_μ *equivalente a* $\nabla g(y)$ *calcolato in* $y = \mu$ *e si assuma che gli elementi di* ∇_μ *siano diversi da zero. Allora*

$$\sqrt{n}\big(g(Y_n) - g(\mu)\big) \dot{\sim} N\big(0, \nabla_\mu^T \Sigma \nabla_\mu\big).$$

Per le proprietà degli stimatori di massima verosimiglianza si ha che

$$\begin{bmatrix} \hat{\theta}_R \\ \hat{\theta}_0 \\ \hat{\theta}_1 \end{bmatrix} \dot{\sim} N_3 \left(\begin{bmatrix} \theta_R \\ \theta_0 \\ \theta_1 \end{bmatrix}, \begin{bmatrix} \sigma_R^2 & \sigma_{R0} & 0 \\ \sigma_{0R} & \sigma_0^2 & 0 \\ 0 & 0 & \sigma_1^2 \end{bmatrix} \right);$$

chiamiamo Σ la matrice di varianza e covarianza. È necessario applicare il metodo delta multivariato per calcolare la distribuzione, e quindi lo *standard error*, di $g\big(\hat{\theta}_R, \hat{\theta}_0, \hat{\theta}_1\big) = S_1\big(u_i; \hat{\theta}_R, \hat{\theta}_0, \hat{\theta}_1\big) = 1 - (1 - \hat{\pi}_r)F_0\big(u_i; \hat{\theta}_0\big) - \hat{\pi}_r F_{R1}\big(u_i; \hat{\theta}_R, \hat{\theta}_1\big)$ (manteniamo $\hat{\pi}_r = 0.3$). Dopo aver calcolato

$$\nabla g(\theta_R, \theta_0, \theta_1) = \nabla = \begin{pmatrix} \frac{\partial g}{\partial \theta_R} \\ \frac{\partial g}{\partial \theta_0} \\ \frac{\partial g}{\partial \theta_1} \end{pmatrix},$$

otteniamo che $\hat{S}_1(u_i) \sim N\big(S_1(u_i), \sigma^2_{S_1}\big)$, con $\sigma^2_{S_1} = \nabla^T \Sigma \nabla$. In particolare, avremo che lo *standard error* di $\hat{S}_1(u_i)$ è uguale a

$$\begin{aligned} \hat{\sigma}_{S_1} = \Bigg(& \hat{\pi}_r^2 \hat{\sigma}_R^2 \cdot \frac{\left[e^{-u_i \hat{\theta}_R^{-1}}\big(\hat{\theta}_1 - u_i + \hat{\theta}_1 \hat{\theta}_R^{-1} u_i\big) - \hat{\theta}_1 e^{-u_i \hat{\theta}_1^{-1}}\right]^2}{\big(\hat{\theta}_R - \hat{\theta}_1\big)^4} - 2\hat{\pi}_r(\hat{\pi}_r - 1)\hat{\sigma}_{R0} \cdot \frac{u_i e^{-u_i \hat{\theta}_0^{-1}}}{\hat{\theta}_0^2} \\ & \times \frac{e^{-u_i \hat{\theta}_R^{-1}}\big(\hat{\theta}_1 - u_i + \hat{\theta}_1 \hat{\theta}_R^{-1} u_i\big) - \hat{\theta}_1 e^{-u_i \hat{\theta}_1^{-1}}}{\big(\hat{\theta}_R - \hat{\theta}_1\big)^2} + (\hat{\pi}_r - 1)^2 \hat{\sigma}_0^2 \cdot \frac{u_i^2 e^{-2u_i \hat{\theta}_0^{-1}}}{\hat{\theta}_0^4} + \hat{\pi}_r^2 \hat{\sigma}_1^2 \\ & \times \frac{\left[e^{-u_i \hat{\theta}_1^{-1}}\big(\hat{\theta}_R - u_i + \hat{\theta}_R \hat{\theta}_1^{-1} u_i\big) - \hat{\theta}_R e^{-u_i \hat{\theta}_R^{-1}}\right]^2}{\big(\hat{\theta}_R - \hat{\theta}_1\big)^4} \Bigg)^{\frac{1}{2}}. \end{aligned}$$

Gli stessi sviluppi sono eseguiti per calcolare lo *standard error* di $\hat{S}_2(u_i)$, sostituendo θ_2 al posto di θ_1.

Appendice B

Codici *R*

B.1 Costruzione *logrank test*

Per confrontare le curve di sopravvivenza stimate con il metodo di Kaplan-Meier risulta necessario costruire "a mano" il *logrank test*, dal momento che quello automatico che si ottiene con il comando `survdiff()` terrebbe conto solo dei pazienti "rispondenti" nella stima di ciascuna curva. Riportiamo il codice *R* per il confronto tra le *policies* ADEP+IL-2 (A_1B_1) e ADEP+Placebo (A_1B_2); codici analoghi si usano per il confronto a coppie tra le altre *policies*.

```
library(base)

j=union(unique(surv1$time),unique(surv2$time)) #tempi distinti
#degli eventi osservati in ciascun gruppo (surv1: stima curva
#A1B1, surv2: stima curva A1B2)
N_1j=NULL
for (i in 1:101){ #vettore n.risk di surv1 ha 101 elementi
  N_1j[i]=surv1$n.risk[i]
}
for (i in 102:122){ #vettore j ha 122 elementi
  N_1j[i]=surv1$n.risk[max(which(j[i]>surv1$time))]
}
N_1j=sort(N_1j,decreasing=T) #numero soggetti a rischio al tempo j
#nel gruppo A1B1
O_1j=NULL
for (i in 1:101){
  O_1j[i]=surv1$n.event[i]
}
```

```
for (i in 102:122){
  O_1j[i]=0
}
O_1j=O_1j[order(j,decreasing=F)] #numero eventi osservati al tempo
#j nel gruppo A1B1

j=union(unique(surv2$time),unique(surv1$time))
N_2j=NULL
for (i in 1:99){
  N_2j[i]=surv2$n.risk[i]
}
for (i in 100:122){
  N_2j[i]=surv2$n.risk[max(which(j[i]>surv2$time))]
}
N_2j=sort(N_2j,decreasing=T) #numero soggetti a rischio al tempo j
#nel gruppo A1B2
O_2j=NULL
for (i in 1:99){
  O_2j[i]=surv2$n.event[i]
 }
for (i in 100:122){
  O_2j[i]=0
}
O_2j=O_2j[order(j,decreasing=F)] #numero eventi osservati al tempo
#j nel gruppo A1B2

N_j=N_1j+N_2j
O_j=O_1j+O_2j
E_1j=(O_j*N_1j)/N_j #valore atteso di O_1j
V_j=(O_j*(N_1j/N_j)*(1-N_1j/N_j)*(N_j-O_j))/(N_j-1) #varianza di
#O_1j
t_12=sum((O_1j-E_1j))/sqrt(sum(V_j)) #logrank test
p_12=2*(1-pnorm(abs(t_12))) #p-value=0.391: accetto H0
```

B.2 Stime funzioni di sopravvivenza (studio di simulazione)

Riportiamo il codice R utilizzato per applicare il metodo parametrico di stima delle funzioni di sopravvivenza alle 1000 simulazioni Monte Carlo effettua-

te (set.seed(123)).

```
library(numDeriv)

loglikelihood=function(par,data){ #funzione di log-verosimiglianza
                                  #fattorizzata
    thetaR=par[1]
    theta1=par[2]
    theta2=par[3]
    U=data$U
    R=data$R
    Z=data$Z
    delta=data$delta
    val.loglikelihood=sum(delta*R*Z*log((exp(-U/thetaR)-exp(-U/
                      theta1))/(thetaR*exp(-U/thetaR)-theta1*
                      exp(-U/theta1))))+sum(R*Z*log((thetaR*exp(
                      -U/thetaR)-theta1*exp(-U/theta1))/(thetaR-
                      theta1)))+sum(delta*R*(1-Z)*log((exp(-U/
                      thetaR)-exp(-U/theta2))/(thetaR*exp(-U/
                      thetaR)-theta2*exp(-U/theta2))))+sum(
                      R*(1-Z)*log((thetaR*exp(-U/thetaR)-theta2*
                      exp(-U/theta2))/(thetaR-theta2)))
    return(val.loglikelihood)
}

par.stime=function(data,opt){ #stime parametri
  R=data$R
  U=data$U
  delta=data$delta
  pi_r.hat=sum(R)/nrow(data)
  theta0.hat=sum((1-R)*U)/sum((1-R)*delta)
  thetaR.hat=opt[1]
  theta1.hat=opt[2]
  theta2.hat=opt[3]
  return(c(pi_r.hat,thetaR.hat,theta0.hat,theta1.hat,theta2.hat))
}

par.se=function(data,hess,pi_r.hat,theta0.hat){ #s.e. parametri
  R=data$R
  U=data$U
  delta=data$delta
  info.fisher=solve(-hess)
```

```
  se_thetaR.hat=sqrt(diag(info.fisher))[1]
  se_theta1.hat=sqrt(diag(info.fisher))[2]
  se_theta2.hat=sqrt(diag(info.fisher))[3]
  se_pi_r.hat=sqrt((sum(1-R)/(1-pi_r.hat)^2+sum(R)/pi_r.hat^2)^
              (-1))
  se_theta0.hat=sqrt((2*sum(U*(1-R))/theta0.hat^3-sum((1-R)*
                delta)/theta0.hat^2)^(-1))
  return(c(se_pi_r.hat,se_thetaR.hat,se_theta0.hat,se_theta1.hat,
         se_theta2.hat))
}

surv.stime=function(par.hat,data){ #stime funz. di sopravvivenza
  U=data$U
  pi_r.hat=par.hat[1]
  thetaR.hat=par.hat[2]
  theta0.hat=par.hat[3]
  theta1.hat=par.hat[4]
  theta2.hat=par.hat[5]
  t1=max(U[U<=1])
  t2=max(U[U<=5])
  SURV11.1=1-(1-pi_r.hat)*(1-exp(-t1/theta0.hat))-pi_r.hat*((
           thetaR.hat*(1-exp(-t1/thetaR.hat))-theta1.hat*(1-exp(
           -t1/theta1.hat)))/(thetaR.hat-theta1.hat))
  SURV12.1=1-(1-pi_r.hat)*(1-exp(-t1/theta0.hat))-pi_r.hat*((
           thetaR.hat*(1-exp(-t1/thetaR.hat))-theta2.hat*(1-exp(
           -t1/theta2.hat)))/(thetaR.hat-theta2.hat))
  SURV11.2=1-(1-pi_r.hat)*(1-exp(-t2/theta0.hat))-pi_r.hat*((
           thetaR.hat*(1-exp(-t2/thetaR.hat))-theta1.hat*(1-exp(
           -t2/theta1.hat)))/(thetaR.hat-theta1.hat))
  SURV12.2=1-(1-pi_r.hat)*(1-exp(-t2/theta0.hat))-pi_r.hat*((
           thetaR.hat*(1-exp(-t2/thetaR.hat))-theta2.hat*(1-exp(
           -t2/theta2.hat)))/(thetaR.hat-theta2.hat))
  return(c(SURV11.1,SURV12.1,SURV11.2,SURV12.2))
}

loglikelihood2=function(par,data){ #funzione di log-verosim.
                                   #fattorizzata l2
  thetaR=par[1]
  theta0=par[2]
  theta1=par[3]
  theta2=par[4]
  U=data$U
```

```
  R=data$R
  Z=data$Z
  delta=data$delta
  val.loglikelihood2=sum(delta*R*Z*log((exp(-U/thetaR)-exp(-U/
                     theta1))/(thetaR*exp(-U/thetaR)-theta1*exp(-
                     U/theta1))))+sum(R*Z*log((thetaR*exp(-U/
                     thetaR)-theta1*exp(-U/theta1))/(thetaR-
                     theta1)))+sum(delta*R*(1-Z)*log((exp(-U/
                     thetaR)-exp(-U/theta2))/(thetaR*exp(-U/
                     thetaR)-theta2*exp(-U/theta2))))+sum(R*(1-Z)*
                     log((thetaR*exp(-U/thetaR)-theta2*exp(-U/
                     theta2))/(thetaR-theta2)))+sum(log(theta0^(
                     -delta*(1-R))*exp(-U/theta0*(1-R))))
  return(val.loglikelihood2)
}

covariance=function(hessian){ #covarianza tra thetaR e theta0
  info.fisher2=solve(-hessian)
  sigmaR0=info.fisher2[1,2]
  return(sigmaR0)
}

surv.se=function(par,se,data,sigmaR0){ #s.e. funzioni di
                                       #sopravvivenza
  U=data$U
  pi_r=par[1]
  thetaR=par[2]
  theta0=par[3]
  theta1=par[4]
  theta2=par[5]
  t1=max(U[U<=1])
  t2=max(U[U<=5])
  #applicazione metodo delta multivariato:
  se_SURV11.1=sqrt((pi_r-1)^2*se[3]^2*exp(-2*t1/theta0)*t1^2/
              theta0^4-2*(pi_r-1)*pi_r*sigmaR0*exp(-t1/theta0)*t1/
              theta0^2*(exp(-t1/thetaR)*(-t1+theta1+theta1/thetaR*
              t1)-theta1*exp(-t1/theta1))/(thetaR-theta1)^2+
              pi_r^2*se[2]^2*(exp(-t1/thetaR)*(-t1+theta1+theta1/
              thetaR*t1)-theta1*exp(-t1/theta1))^2/(thetaR-theta1)
              ^4+pi_r^2*se[4]^2*(exp(-t1/theta1)*(-t1+thetaR+
              thetaR/theta1*t1)-thetaR*exp(-t1/thetaR))^2/(thetaR-
              theta1)^4)
```

```
  se_SURV12.1=sqrt((pi_r-1)^2*se[3]^2*exp(-2*t1/theta0)t1^2/
              theta0^4-2*(pi_r-1)*pi_r*sigmaR0*exp(-t1/theta0)*t1/
              theta0^2*(exp(-t1/thetaR)*(-t1+theta2+theta2/thetaR*
              t1)-theta2*exp(-t1/theta2))/(thetaR-theta2)^2+
              pi_r^2*se[2]^2*(exp(-t1/thetaR)*(-t1+theta2+theta2/
              thetaR*t1)-theta2*exp(-t1/theta2))^2/(thetaR-theta2)
              ^4+pi_r^2*se[5]^2*(exp(-t1/theta2)*(-t1+thetaR+
              thetaR/theta2*t1)-thetaR*exp(-t1/thetaR))^2/(thetaR-
              theta2)^4)
  se_SURV11.2=sqrt((pi_r-1)^2*se[3]^2*exp(-2*t2/theta0)t2^2/
              theta0^4-2*(pi_r-1)*pi_r*sigmaR0*exp(-t2/theta0)*t2/
              theta0^2*(exp(-t2/thetaR)*(-t2+theta1+theta1/thetaR*
              t2)-theta1*exp(-t2/theta1))/(thetaR-theta1)^2+
              pi_r^2*se[2]^2*(exp(-t2/thetaR)*(-t2+theta1+theta1/
              thetaR*t2)-theta1*exp(-t2/theta1))^2/(thetaR-theta1)
              ^4+pi_r^2*se[4]^2/*(exp(-t2/theta1)*(-t2+thetaR+
              thetaR/theta1*t2)-thetaR*exp(-t2/thetaR))^2/(thetaR-
              theta1)^4)
  se_SURV12.2=sqrt((pi_r-1)^2*se[3]^2*exp(-2*t2/theta0)*t2^2/
              theta0^4-2*(pi_r-1)*pi_r*sigmaR0*exp(-t2/theta0)*t2/
              theta0^2*(exp(-t2/thetaR)*(-t2+theta2+theta2/thetaR*
              t2)-theta2*exp(-t2/theta2))/(thetaR-theta2)^2+
              pi_r^2*se[2]^2*(exp(-t2/thetaR)*(-t2+theta2+theta2/
              thetaR*t2)-theta2*exp(-t2/theta2))^2/(thetaR-theta2)
              ^4+pi_r^2*se[5]^2*(exp(-t2/theta2)*(-t2+thetaR+
              thetaR/theta2*t2)-thetaR*exp(-t2/thetaR))^2/(thetaR-
              theta2)^4)
  return(c(se_SURV11.1,se_SURV12.1,se_SURV11.2,se_SURV12.2))
}

conv=NULL
par_stime=matrix(NA,nrow=1000,ncol=5)
par_se=matrix(NA,nrow=1000,ncol=5)
stime11=matrix(NA,nrow=1000,ncol=2)
colnames(stime11)=c("t=1","t=5")
stime12=matrix(NA,nrow=1000,ncol=2)
colnames(stime12)=c("t=1","t=5")
se11=matrix(NA,nrow=1000,ncol=2)
colnames(se11)=c("t=1","t=5")
se12=matrix(NA,nrow=1000,ncol=2)
colnames(se12)=c("t=1","t=5")
```

```
for (j in c(1:516,518:603,605:1000)){ #errore a j=517 e j=604
  dataset=cbind(X[,j],R[,j],Z[,j],T0[,j],TR[,j],T1[,j],T2[,j],
          C[,j],T[,j],U[,j],delta[,j])
  dataset=as.data.frame(dataset)
  colnames(dataset)=c("X","R","Z","T0","TR","T1","T2","C","T","U",
                      "delta")
  datasetPAR=dataset[X[,j]==1,]

  #parametri:
  optimization=optim(par=c(thetaR,theta1,theta2),fn=loglikelihood,
               data=datasetPAR,hessian=FALSE,control=list(trace=
               TRUE,fnscale=-1),method="L-BFGS-B",lower=rep
               (0.0001,3),upper=rep(Inf,3))
               #ottengo stime ottimizzate di thetaR,theta1,theta2
  conv[j]=optimization$convergence

  par_stime[j,]=par.stime(datasetPAR,optimization$par)
                #stime parametri

  hess=hessian(func=loglikelihood,x=c(par_stime[j,2],par_stime[j,
       4],par_stime[j,5]),data=datasetPAR)
  par_se[j,]=par.se(datasetPAR,hess,par_stime[j,1],par_stime[j,3])
             #s.e. parametri

  #funzioni di sopravvivenza:
  r1stime=surv.stime(par_stime[j,],datasetPAR)
  stime11[j,]=cbind(r1stime[c(1,3)]) #stime A1B1
  stime12[j,]=cbind(r1stime[c(2,4)]) #stime A1B2

  optimization2=optim(par=c(thetaR,theta0,theta1,theta2),fn=
                loglikelihood2,data=datasetPAR,hessian=TRUE,
                control=list(trace=TRUE,fnscale=-1),method=
                "L-BFGS-B",lower=rep(0.01,4),upper=rep(Inf,4))
  sigmaR0.hat=covariance(optimization2$hessian)
  r1se=surv.se(par_stime[j,],par_se[j,],datasetPAR,sigmaR0.hat)
  se11[j,]=cbind(r1se[c(1,3)]) #s.e. A1B1
  se12[j,]=cbind(r1se[c(2,4)]) #s.e. A1B2
}

table(conv) #994 convergenza "0", 4 convergenza "52", 2 NA
```

```
#combino i risultati:
mean_stime11=apply(stime11,2,mean,na.rm=TRUE)
mean_stime12=apply(stime12,2,mean,na.rm=TRUE)
mean_se11=apply(se11,2,mean,na.rm=TRUE)
mean_se12=apply(se12,2,mean,na.rm=TRUE)
```

B.3 Costruzione grafici per assunzioni distributive tempi di sopravvivenza

Riportiamo il codice usato per costruire il grafico relativo alla variabile U_{0i}^*; codici analoghi sono stati usati per le altre variabili. L'utilizzo del comando `survreg()` ha reso necessaria una riparametrizzazione dei parametri associati a ciascuna distribuzione [16].

```
dati.adep=dati[dati$trt_induct=="ADEP",] #solo A1

index=which(dati.adep$rec_intens==0) #solo R=0

library(survival)

surv.KM=survfit(Surv(dati.adep$survivaldays[index],
        dati.adep$status[index])~1,data=dati.adep,
        type="kaplan-meier",se.fit=F) #Kaplan-Meier
H.KM=-log(surv.KM$surv)
plot(log(surv.KM$time),log(H.KM),type="s",lwd=1.5,
     ylab="Log cumulative hazard",xlab="Time (log scale)",
     main=expression('U'[0]^'*'))

surv.exp=survreg(Surv(dati.adep$survivaldays[index],
         dati.adep$status[index])~1,data=dati.adep,
         dist="exponential") #assunzione esponenziale
theta0.hat=exp(surv.exp$coefficients)
H.exp=-log(exp(-(dati.adep$survivaldays[index])/theta0.hat))
lines(sort(log(dati.adep$survivaldays[index])),sort(log(H.exp)),
      lwd=1.5,col=2)

surv.wei=survreg(Surv(dati.adep$survivaldays[index],
         dati.adep$status[index])~1,data=dati.adep,dist="weibull")
         #assunzione Weibull
alpha0.hat=1/surv.wei$scale
```

```
lambda0.hat=exp(surv.wei$coefficients)
H.wei=-log(exp(-((dati.adep$survivaldays[index])/lambda0.hat)^
      alpha0.hat))
lines(sort(log(dati.adep$survivaldays[index])),sort(log(H.wei)),
      lwd=1.5,col=3)

surv.lnorm=survreg(Surv(dati.adep$survivaldays[index],
           dati.adep$status[index])~1,data=dati.adep,
           dist="lognormal") #assunzione log-normale
sigma0.hat=surv.lnorm$scale
mu0.hat=surv.lnorm$coefficients
H.lnorm=-log(1-pnorm(log(dati.adep$survivaldays[index]),
        mean=mu0.hat,sd=sigma0.hat))
lines(sort(log(dati.adep$survivaldays[index])),sort(log(H.lnorm)),
      lwd=1.5,col=4)

legend(0.7,-3,legend=c("Kaplan-Meier","Exponential","Weibull",
       "Log-normal"),col=c(1,2,3,4),lwd=1.5,cex=0.75,box.lty=0,
       lty=1.5)
```

B.4 Stime funzioni di sopravvivenza (assunzioni miste tempi di sopravvivenza)

Grazie alla funzione `integrate()` si è riusciti a calcolare le convoluzioni $f_{R1}(\cdot)$ e $f_{R2}(\cdot)$ e le rispettive funzioni di ripartizione. Di seguito il codice R completo con cui abbiamo ottenuto le stime finali $\hat{S}_1(\cdot)$ e $\hat{S}_2(\cdot)$.

```
dati.adep=dati[dati$trt_induct=="ADEP",] #solo A1
dati.adep$trt_immuno=as.numeric(dati.adep$trt_immuno)-1
index=which(is.na(dati.adep$trt_immuno)==TRUE)
dati.adep$trt_immuno[index]=0 #impongo valore qualsiasi ai dati
                              #mancanti, tanto Z e' definito solo
                              #per R=1

#ricavo le convoluzioni fR1 e fR2:
f.R=function(x,muR,sigmaR){
  dlnorm(x,meanlog=muR,sdlog=sigmaR)
}
f.1=function(y,mu1,sigma1){
  dlnorm(y,meanlog=mu1,sdlog=sigma1)
```

```
}
f.R1=function(z,muR,sigmaR,mu1,sigma1){
  integrate(function(x,z){
    f.R(x,muR,sigmaR)*f.1(z-x,mu1,sigma1)
  },lower=0.0001,upper=z,z=z,stop.on.error=FALSE)$value
}
fR1=Vectorize(f.R1) #convoluzione

f.2=function(y,mu2,sigma2){
  dlnorm(y,meanlog=mu2,sdlog=sigma2)
}
f.R2=function(z,muR,sigmaR,mu2,sigma2){
  integrate(function(x,z){
    f.R(x,muR,sigmaR)*f.2(z-x,mu2,sigma2)
  },lower=0.0001,upper=z,z=z,stop.on.error=FALSE)$value
}
fR2=Vectorize(f.R2) #convoluzione

#ricavo le funzioni di ripartizione delle convoluzioni, FR1 e FR2:
F.1=function(y,mu1,sigma1){
  plnorm(y,meanlog=mu1,sdlog=sigma1)
}
F.R1=function(z,muR,sigmaR,mu1,sigma1){
  integrate(function(x,z){
    f.R(x,muR,sigmaR)*F.1(z-x,mu1,sigma1)
  },lower=0.0001,upper=z,z=z,stop.on.error=FALSE)$value
}
FR1=Vectorize(F.R1) #funzione di ripartizione

F.2=function(y,mu2,sigma2){
  plnorm(y,meanlog=mu2,sdlog=sigma2)
}
F.R2=function(z,muR,sigmaR,mu2,sigma2){
  integrate(function(x,z){
    f.R(x,muR,sigmaR)*F.2(z-x,mu2,sigma2)
  },lower=0.0001,upper=z,z=z,stop.on.error=FALSE)$value
}
FR2=Vectorize(F.R2) #funzione di ripartizione


#per ricavare le stime dei parametri delle distribuzioni
#log-normali:
```

```
loglikelihood=function(par,data){ #funzione di log-verosimiglianza
                                  #fattorizzata
  muR=par[1]
  sigmaR=par[2]
  mu1=par[3]
  sigma1=par[4]
  mu2=par[5]
  sigma2=par[6]
  U=data$survivaldays
  R=data$rec_intens
  Z=data$trt_immuno
  delta=data$status
  fR1=fR1(U,muR,sigmaR,mu1,sigma1)
  fR2=fR2(U,muR,sigmaR,mu2,sigma2)
  FR1=FR1(U,muR,sigmaR,mu1,sigma1)
  FR2=FR2(U,muR,sigmaR,mu2,sigma2)
  val.loglikelihood=log(prod((fR1^delta*(1-FR1)^(1-delta))^(R*Z)*
                    (fR2^delta*(1-FR2)^(1-delta))^(R*(1-Z))))
  return(val.loglikelihood)
}
optimization=optim(par=c(-1.9,0.4,1.7,0.9,1.4,1.1),fn=
             loglikelihood,data=dati.adep,hessian=TRUE,control=
             list(trace=TRUE,fnscale=-1),method="L-BFGS-B",lower=
             rep(c(-Inf,0.0001),3),upper=rep(Inf,6))
optimization$par #stime ottimizzate di muR,sigmaR,mu1,sigma1,mu2,
                 #sigma2

#per ricavare la stima di alpha0 (Weibull):
loglikelihood2=function(par,data){ #funz. di log-verosimiglianza
                                   #fattorizzata
  alpha0=par
  U=data$survivaldays
  R=data$rec_intens
  delta=data$status
  val.loglikelihood2=log(alpha0)*sum(delta*(1-R))-alpha0*log(3)*
                     sum(delta*(1-R))+(alpha0-1)*sum(delta*(1-R)*
                     log(U))-sum((1-R)*(U/3)^alpha0)
  return(val.loglikelihood2)
}
optimization2=optim(par=1,fn=loglikelihood2,data=dati.adep,
              hessian=TRUE,control=list(trace=TRUE,fnscale=-1),
              method="L-BFGS-B",lower=0.0001,upper=Inf)
```

```
optimization2$par #stima ottimizzata di alpha0

#stime parametri:
muR.hat=optimization$par[1]
sigmaR.hat=optimization$par[2]
mu1.hat=optimization$par[3]
sigma1.hat=optimization$par[4]
mu2.hat=optimization$par[5]
sigma2.hat=optimization$par[6]
alpha0.hat=optimization2$par
lambda0.hat=(sum(dati.adep$survivaldays^alpha0.hat*(1-dati.adep$
            rec_intens))/sum(dati.adep$status*(1-dati.adep$
            rec_intens)))^(1/alpha0.hat)
pi_r.hat=sum(dati.adep$rec_intens)/nrow(dati.adep)

#stime funzioni di sopravvivenza:
SURV11=1-(1-pi_r.hat)*(1-exp(-(dati.adep$survivaldays/lambda0.hat)
       ^alpha0.hat))-pi_r.hat*FR1(dati.adep$survivaldays,muR.hat,
       sigmaR.hat,mu1.hat,sigma1.hat)
SURV12=1-(1-pi_r.hat)*(1-exp(-(dati.adep$survivaldays/lambda0.hat)
       ^alpha0.hat))-pi_r.hat*FR2(dati.adep$survivaldays,muR.hat,
       sigmaR.hat,mu2.hat,sigma2.hat)

plot(sort(dati.adep$survivaldays),sort(SURV11,decreasing=TRUE),
     ylab="Survival␣probability",xlab="Time␣(years)",col=2,
     lwd=1.5,type="l",ylim=c(0,1))
lines(sort(dati.adep$survivaldays),sort(SURV12,decreasing=TRUE),
      col=3,lwd=1.5,type="l")
legend(9.3,1,legend=c("ADEP+IL-2","ADEP+Placebo"),col=c(2,3),
       lty=1,cex=0.75,box.lty=0,lwd=1.5)
```

Bibliografia

[1] Azzalini, A., and Scarpa, B. *Data Analysis and Data Mining: An introduction.* OUP USA, 2012.

[2] Costa Filho, I. G. *Mixture models for the analysis of gene expression: Integration of multiple experiments and cluster validation.* PhD thesis, 2008.

[3] Furman, E., Hackmann, D., and Kuznetsov, A. On log-normal convolutions: An analytical-numerical method with applications to economic capital determination. *SSRN Electronic Journal* (2017).

[4] Guo, X., and Tsiatis, A. A weighted risk set estimator for survival distributions in two-stage randomization designs with censored survival data. *The International Journal of Biostatistics 1*, 1 (2005).

[5] Jackson, C. H. flexsurv: A platform for parametric survival modeling in R. *Journal of Statistical Software 70* (2016).

[6] Klein, J. P., and Moeschberger, M. L. *Survival analysis: Techniques for censored and truncated data.* Springer Science & Business Media, 2006.

[7] Kolitz, J. E., et al. Recombinant interleukin-2 in patients aged younger than 60 years with acute myeloid leukemia in first complete remission: Results from Cancer and Leukemia Group B 19808. *Cancer 120*, 7 (2014), 1010–1017.

[8] Lunceford, J. K., Davidian, M., and Tsiatis, A. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics 58*, 1 (2002).

[9] National Cancer Institute. `https://www.cancer.gov/publications/dictionaries`. Accessed: 2018-11-7.

[10] Rodríguez, G. Parametric survival models. Tech. rep., Princeton: Princeton University, 2010.

[11] Ross, S. M. *Calcolo delle probabilità.* Maggioli Editore, 2013.

[12] Samimi, H., and Akbari, M. An accurate approximation to 3-parameter Weibull sum distribution. *International Research Journal of Applied and Basic Sciences 4*, 6 (2013), 1524–1529.

[13] Schlenker, G. J. Methods for calculating the probability distribution of sums of independent random variables. Tech. rep., Army Armament, Munitions and Chemical Command. Systems Analysis Office. Rock Island, Illinois, 1986.

[14] StatLect. `https://www.statlect.com/fundamentals-of-probability/sums-of-independent-random-variables`. Accessed: 2019-02-11.

[15] Tang, X., and Melguizo, M. Package "DTR". *R Top Doc* (2015).

[16] Therneau, T. M., and Lumley, T. Package "survival". *R Top Doc 128* (2015).

[17] Wahed, A. S. Inference for two-stage adaptive treatment strategies using mixture distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 59*, 1 (2010).

[18] Wasserman, L. *All of statistics: A concise course in statistical inference*. Springer Science & Business Media, 2013.

[19] Yilmaz, F., and Alouini, M.-S. Sum of Weibull variates and performance of diversity systems. In *Proceedings of the 2009 International Conference on Wireless Communications and Mobile Computing: Connecting the World Wirelessly* (2009), ACM, pp. 247–252.

Ringraziamenti

Il primo grazie va alla Prof.ssa Cortese, per la costanza e la disponibilità dimostratemi in tutti i mesi di lavoro assieme. Grazie inoltre per i consigli e gli importanti aiuti "extrascolastici", l'ho apprezzato molto.

Il ringraziamento più importante va alla mia famiglia. Grazie a mia mamma per essere il mio punto di riferimento quotidiano, a mio papà per avermi spronato nei momenti più duri e a mia sorella Alessandra per essere la mia compagna di risate e spensieratezza.

Infine grazie a Margherita, Michela e Sara, non so come sarebbero stati questi anni senza di voi.