

Università degli Studi di Padova  
Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in  
Scienze Statistiche



TESI DI LAUREA

**MODELLI DI REGRESSIONE DIRETTI E INDIRETTI PER IL  
RESIDUAL MEAN TIME IN DATI DI SOPRAVVIVENZA CON  
CENSURA A DESTRA**

Relatrice: Prof.ssa Giuliana Cortese  
Dipartimento di Scienze Statistiche

Laureando: Davide Serafin  
Matricola 1143234

Anno Accademico 2018/2019



# Indice

<b>Introduzione</b>	<b>5</b>
<b>1 Analisi di dati di durata: concetti base</b>	<b>9</b>
1.1 Funzioni base . . . . .	9
1.1.1 Funzione di sopravvivenza . . . . .	10
1.1.2 Funzione di rischio . . . . .	10
1.1.3 <i>Residual Mean Time Function</i> . . . . .	11
1.2 Stima delle funzioni base . . . . .	12
1.2.1 Stimatore di Kaplan-Meier . . . . .	13
1.2.2 Modello di Cox . . . . .	14
<b>2 <i>Inverse Probability Weighting (IPW)</i></b>	<b>17</b>
2.1 <i>Inverse Probability of Treatment Weighting (IPTW)</i> . . . . .	17
2.2 IPW in studi con dati mancanti . . . . .	19
2.3 <i>Inverse Probability of Censoring Weighting (IPCW)</i> . . . . .	21
<b>3 Stima del <i>Mean Time</i></b>	<b>25</b>
3.1 Stima per campioni omogenei . . . . .	26
3.2 Stima per campioni eterogenei . . . . .	28
3.3 Confronto tra gruppi con aggiustamento di covariate . . . . .	31
<b>4 Modelli di regressione per il <i>Residual Mean Time</i></b>	<b>33</b>
4.1 Modelli indiretti e diretti per il <i>Residual Mean Time</i> . . . . .	34
4.2 Una classe di modelli diretti per il <i>Residual Mean Time</i> . . . . .	37

4.2.1	Presentazione del modello . . . . .	37
4.2.2	Stima dei parametri . . . . .	38
4.2.3	Equivalenza tra lo stimatore IPCW e lo stimatore di Kaplan-Meier . . . . .	40
<b>5</b>	<b>Applicazione ai dati</b>	<b>45</b>
5.1	Presentazione del dataset . . . . .	45
5.1.1	<i>Endpoint</i> di sopravvivenza . . . . .	46
5.1.2	Variabili cliniche e di espressione genica . . . . .	50
5.2	Effetto marginale delle variabili sui tempi DFS . . . . .	52
5.2.1	Effetto delle variabili cliniche . . . . .	53
5.2.2	Effetto delle variabili di espressione genica . . . . .	64
5.2.3	Alcune considerazioni . . . . .	75
5.3	Stima univariata del <i>Residual Mean Time</i> . . . . .	77
5.4	Modelli di regressione . . . . .	81
5.4.1	Modello di Cox . . . . .	81
5.4.2	Modellazione del <i>Residual Mean Time</i> . . . . .	82
<b>6</b>	<b>Studi di simulazione</b>	<b>87</b>
<b>7</b>	<b>Conclusioni</b>	<b>93</b>
	<b>Bibliografia</b>	<b>97</b>
<b>A</b>	<b>Codice R - Applicazione ai dati</b>	<b>103</b>
<b>B</b>	<b>Codice R - Studi di simulazione</b>	<b>111</b>

# Introduzione

In uno studio clinico longitudinale l'*endpoint* primario per comparare più gruppi è spesso il tempo a uno specifico evento, come la morte o la recidiva di malattia. In questi casi, l'obiettivo principale è misurare la differenza tra le funzioni di sopravvivenza. È pratica comune assumere che il rapporto delle funzioni di azzardo sia costante nel tempo e utilizzare l'*hazard ratio* come parametro per quantificare la differenza tra gruppi. La stima dell'*hazard ratio* è ottenuta utilizzando il modello a rischi proporzionali di Cox. I valori dell'*hazard ratio* sono in genere difficili da interpretare, inoltre, quando l'assunzione di rischi proporzionali è violata, possono fornire una misura distorta.

Per questo motivo, si riscontra un interesse sempre maggiore in altre misure riassuntive globali basate sulla funzione di sopravvivenza che catturino il livello assoluto di rischio e siano più facili e dirette da interpretare rispetto all'*hazard ratio*. In questa tesi, si propone l'utilizzo del *residual mean time*, che rappresenta l'area al di sotto della funzione di sopravvivenza ed è interpretabile come l'aspettativa di vita residua di un individuo che è sopravvissuto fino a un certo istante  $t$ . Lo scopo della tesi è descrivere differenti metodi di stima per tale quantità, illustrando alcuni approcci di stima indiretta non parametrici e una classe di modelli di regressione diretti basati su equazioni di stima pesate con l'approccio *Inverse Probability of Censoring Weighting* (IPCW).

Il Capitolo 1 introduce le quantità e tecniche di stima sottostanti l'analisi di sopravvivenza utilizzate nel corso della tesi, con una prima descrizione della funzione d'interesse *residual mean time*. Il Capitolo 2 è dedicato all'approccio basato sull'*inverse probability weighting*: dopo aver descritto alcune delle note applicazio-

ni del metodo nel misurare l'effetto di un trattamento e nella gestione dei dati mancanti, si illustra l'applicazione nel contesto di analisi di sopravvivenza, con particolare enfasi sulla costruzione dei pesi che verranno utilizzati per la stima dei coefficienti di regressione dei modelli diretti proposti successivamente. Nel Capitolo 3 si propone uno stimatore non parametrico del *residual mean time*, con particolare attenzione alle complicazioni dovute alla presenza di censura a destra, che comportano l'utilizzo del *restricted residual mean time*. Inoltre, si descrive l'utilità di tale quantità nel misurare la differenza tra gruppi e i vantaggi ottenuti rispetto all'utilizzo di altre misure come l'*hazard ratio*.

In questo ambito, i modelli di regressione mirano a studiare l'associazione tra il *residual mean time* e le potenziali covariate. A tal fine, nel Capitolo 4 si introducono gli approcci studiati da Karrison (1987, 1997) e Zucker (1998), che propongono una stima indiretta del *restricted residual mean time* basata su modelli di regressione per la funzione di rischio come quello di Cox. Successivamente, si presenta una classe di modelli diretti per il *restricted residual mean time*, discutendo il procedimento di stima dei parametri e la costruzione dell'equazione di stima. In aggiunta, si dimostra l'equivalenza tra la stima del *restricted residual mean time* fornita dal modello e la stima ricavata tramite integrazione dello stimatore di Kaplan-Meier.

Nel Capitolo 5 si illustra l'applicazione dei metodi proposti, confrontando le stime del *restricted residual mean time* fornite dai modelli indiretti e diretti discussi nella tesi. L'obiettivo è valutare i vantaggi e le criticità di ciascun metodo, studiando per esempio l'effetto dei pesi IPCW nel modello diretto, assumendo censura indipendente o meno dalle covariate. Lo studio considerato è parte di un progetto clinico di ricerca in collaborazione con i dott.ri Rocco Cappellesso (Dipartimento di Medicina, Università di Padova), Tiziana Sanavia (*Department of Biomedical Informatics, Harvard Medical School*), Lorenzo Nicolè (Dipartimento di Medicina, Università di Padova). La malattia considerata è il melanoma cutaneo e lo scopo principale dello studio è valutare come i livelli di produzione dei geni recettori TAM e dei geni coinvolti nel processo di necroptosi influenzino le diverse risposte di sopravvivenza al tumore, in particolare rispetto agli eventi di morte e ricaduta della malattia.

Infine, nel Capitolo 6 si conducono alcuni studi di simulazione con lo scopo di valutare le proprietà inferenziali dei modelli di regressione diretti e indiretti per il *restricted residual mean time* presentati nella tesi.



# Capitolo 1

## Analisi di dati di durata: concetti base

Vengono introdotti i concetti sottostanti i metodi di analisi di sopravvivenza. Si presentano le funzioni base per il tempo continuo (si assume che l'evento d'interesse possa avvenire in qualsiasi istante di tempo) e alcuni degli usuali criteri di stima, che verranno richiamati nei capitoli successivi. Per la notazione utilizzata e eventuali approfondimenti si faccia riferimento a Klein e Moeschberger (2003).

### 1.1 Funzioni base

Si consideri il tempo trascorso da un soggetto in un determinato stato, che si conclude con il verificarsi di uno specifico evento e ha inizio dal momento in cui il soggetto è esposto al rischio di subire tale evento. Il tempo all'evento o tempo di sopravvivenza di un individuo, ossia la durata di permanenza in un determinato stato (definita come episodio), è la realizzazione di una variabile aleatoria continua  $T$  non negativa, con distribuzione tipicamente asimmetrica. La distribuzione di  $T$  è descritta dalle funzioni introdotte di seguito, che sono univocamente collegate tra di loro dalle relazioni esplicitate.

### 1.1.1 Funzione di sopravvivenza

Sia  $F(t)$  la funzione di ripartizione di  $T$  (probabilità di sperimentare l'evento fino al tempo  $t$ ) definita come

$$F(t) = Pr(T \leq t) = \int_0^t f(s)ds, \quad (1.1)$$

con  $f(t)$  funzione di densità di  $T$

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{Pr(t \leq T < t + \Delta t)}{\Delta t}, \quad (1.2)$$

funzione positiva che integra a uno. Dalla (1.2), si può mostrare che per un  $\Delta t > 0$  sufficientemente piccolo,  $f(t)\Delta t$  rappresenta la probabilità che l'evento si presenti tra  $t$  e  $t + \Delta t$ .

La funzione di sopravvivenza  $S(t)$  esprime la probabilità di sperimentare l'evento dopo un certo istante  $t$  e è il complemento della funzione di ripartizione

$$S(t) = 1 - F(t) = Pr(T > t) = \int_t^\infty f(s)ds. \quad (1.3)$$

Essa è monotona decrescente. In  $t = 0$  nessun soggetto ha sperimentato l'evento, quindi la probabilità di sopravvivenza è pari a 1. All'aumentare del tempo la funzione di sopravvivenza decresce in modo più o meno ripido a seconda della frequenza dell'evento. La distribuzione di  $T$  si dice propria se  $\lim_{t \rightarrow \infty} S(t) = 0$  (prima o poi l'evento viene sperimentato) e impropria se  $\lim_{t \rightarrow \infty} S(t) = g$  (esiste una probabilità  $g \neq 0$  di essere immuni all'evento). La funzione di sopravvivenza è utile per confrontare campioni con caratteristiche diverse.

In aggiunta, si può ottenere la relazione

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}. \quad (1.4)$$

### 1.1.2 Funzione di rischio

La funzione di rischio o di azzardo  $h(t)$  è definita come

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \quad (1.5)$$

Dalla (1.5) si deduce che per un  $\Delta t > 0$  sufficientemente piccolo,  $h(t)\Delta t$  rappresenta la probabilità che l'evento si presenti tra  $t$  e  $t + \Delta t$ , dato che si è sopravvissuti fino al tempo  $t$ . Più semplicemente,  $h(t)$  può essere interpretata come la propensione al cambiamento di stato nell'istante  $t$ , per chi è nella condizione di farlo. Come avviene per  $f(t)$ ,  $h(t)$  è una funzione positiva e non è una probabilità. Essa è utile per descrivere in che modo il rischio di sperimentare l'evento cambia nel tempo. La funzione di rischio, infatti, può assumere un andamento crescente (l'evento tende a verificarsi alla fine del periodo di osservazione: tipico di unità soggette a invecchiamento o usura), decrescente (l'evento tende a verificarsi all'inizio del periodo di osservazione: casi in cui vi è una selezione iniziale), costante, ma anche non monotono.

La funzione di rischio può essere riscritta come

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d \ln(S(t))}{dt}. \quad (1.6)$$

Una quantità connessa alla funzione di azzardo è la funzione di rischio cumulato  $H(t)$  definita da

$$H(t) = \int_0^t h(s) ds = -\ln(S(t)), \quad (1.7)$$

$$S(t) = \exp(-H(t)) = \exp\left(-\int_0^t h(s) ds\right). \quad (1.8)$$

Questa, quando non si ha a disposizione  $h(t)$ , può essere ottenuta tramite la funzione di sopravvivenza per avere un'idea del possibile andamento della funzione di rischio.

### 1.1.3 *Residual Mean Time Function*

Quando  $T$  ha funzione di probabilità  $f(t)$  e media finita, il *residual mean time* al tempo  $t$ , indicato con  $m(t)$ , è una funzione positiva definita da

$$m(t) = E(T - t \mid T > t) = \frac{\int_t^\infty (s - t)f(s)ds}{S(t)}. \quad (1.9)$$

Essa fornisce il tempo medio all'evento d'interesse, dato che l'evento non si è verificato fino all'istante  $t$ . Questa può ritenersi utile, per esempio, in trial clinici

dove si è interessati a conoscere per quanto tempo un paziente affetto da cancro al cervello ci si aspetti sopravvivere, dati i relativi fattori prognostici e il fatto che è ancora vivo in quell'istante (Sun *et al.* 2011).

Sfruttando la relazione (1.4), la regola di integrazione per parti e assumendo  $S(\infty) = 0$ , la (1.9) può essere riscritta come

$$\begin{aligned} E(T - t | T > t)S(t) &= \int_t^\infty (s - t)f(s)ds \\ &= \int_t^\infty -(s - t)dS(s) \\ &= -(s - t)S(s)|_t^\infty + \int_t^\infty S(s)ds \\ &= \int_t^\infty S(s)ds, \end{aligned}$$

che porta a

$$m(t) = E(T - t | T > t) = \frac{\int_t^\infty S(s)ds}{S(t)}. \quad (1.10)$$

Dall'espressione ricavata in (1.10), si evince che l'aspettativa di vita residua al tempo  $t$  corrisponde all'area al di sotto della curva di sopravvivenza da un certo istante  $t$ , divisa per il valore della funzione di sopravvivenza stessa in  $t$ . Dalla (1.10), è possibile ricavare il tempo medio all'evento (*mean time*)  $\mu = E(T)$  ponendo  $t = 0$ , che corrisponde all'area totale al di sotto della curva di sopravvivenza.

Altri collegamenti utili con le quantità precedentemente discusse sono forniti dalle relazioni univoche

$$S(t) = \frac{m(0)}{m(t)} \exp\left(-\int_0^t \frac{ds}{m(s)}\right), \quad (1.11)$$

$$f(t) = \left(\frac{d}{dt}m(t) + 1\right) \left(\frac{m(0)}{m(t)^2}\right) \exp\left(-\int_0^t \frac{ds}{m(s)}\right), \quad (1.12)$$

$$h(t) = \frac{\frac{d}{dt}m(t) + 1}{m(t)}. \quad (1.13)$$

## 1.2 Stima delle funzioni base

Sia dato un campione di  $n$  soggetti ( $n$  episodi). Si considera il caso in cui ciascun episodio termina con il verificarsi dell'evento di interesse o con una censura a

destra, dovuta al periodo di osservazione considerato (alcuni individui non sperimentano l'evento prima della fine dello studio) o altre cause (per esempio l'uscita dal *follow-up* o il decesso del soggetto). Si assume che il meccanismo di censura sia indipendente dal processo di formazione dei tempi di sopravvivenza (censura non informativa).

### 1.2.1 Stimatore di Kaplan-Meier

Siano  $t_1 < t_2 < \dots < t_D$  i tempi in cui si verificano uno o più eventi, e  $E_i$  il numero di eventi che accadono in  $t_i$ . Sia  $R_i$  il numero di soggetti a rischio di sperimentare l'evento al tempo  $t_i$  (cioè il numero di soggetti che non hanno ancora sperimentato l'evento in  $t_i$  o che sperimentano l'evento in  $t_i$ ). Lo stimatore *Product-Limit* proposto da Kaplan e Meier (1958) fornisce una stima non parametrica della funzione di sopravvivenza

$$\widehat{S}_{KM}(t) = \begin{cases} 1 & \text{se } t < t_1 \\ \prod_{t_i \leq t} \left(1 - \frac{E_i}{R_i}\right) & \text{se } t_1 \leq t \end{cases} \quad (1.14)$$

Lo stimatore ha valore costante dopo  $t_D$  e viene rappresentato da una curva a gradini continua a destra, con ordinata il tempo di sopravvivenza stimato e ascissa i tempi all'evento. Questa vale 1 per  $t = 0$  (nessun soggetto ha sperimentato l'evento), decresce nel tempo, presenta salti negli istanti in cui l'evento è osservato. L'altezza dei gradini in ciascun  $t_i$  dipende dal numero di eventi osservati e dal numero di soggetti censurati. La formula di Greenwood (1.15) stima la varianza di  $\widehat{S}_{KM}(t)$ , e può essere utilizzata per dare una misura di precisione della stima di probabilità di sopravvivenza e il calcolo dell'intervallo di confidenza

$$\widehat{V}(\widehat{S}_{KM}(t)) = \widehat{S}_{KM}(t)^2 \sum_{t_i \leq t} \frac{E_i}{R_i(R_i - E_i)}. \quad (1.15)$$

Lo stimatore di Kaplan-Meier non permette di stimare direttamente la funzione di densità e di rischio. Tuttavia, è possibile ricavare informazioni indirette su quest'ultima o approssimarla tramite liscio, stimando la funzione di rischio cumulato attraverso la relazione  $H(t) = -\ln(S(t))$  e lo stimatore di Nelson-Aalen.

Nel caso di campioni eterogenei, è possibile utilizzare lo stimatore di Kaplan-Meier per confrontare le curve di sopravvivenza di due o più gruppi. A tale scopo, si può ricorrere a test statistici che verificano l'omogeneità delle curve, per evidenziare differenze statisticamente significative nelle distribuzioni degli episodi tra gruppi. Un approccio possibile è dato test dei ranghi logaritmici, un test non parametrico basato sui ranghi che saggia l'ipotesi nulla di uguaglianza delle funzioni di rischio di due o più gruppi.

### 1.2.2 Modello di Cox

Nell'analisi di dati di sopravvivenza può essere d'interesse studiare la relazione fra il tempo all'evento e una serie di variabili esplicative. Il modello di Cox (1972) è un modello di regressione semiparametrico che esprime il rischio in funzione del tempo e delle covariate. Dato un vettore di  $p$  covariate,  $X^T = (X_1, \dots, X_p)$  il modello si presenta nella forma

$$h(t | X) = h_0(t) \exp(\beta^T X) = h_0(t) \exp(\beta_1 X_1 + \dots + \beta_p X_p), \quad (1.16)$$

dove  $\beta = (\beta_1, \dots, \beta_p)^T$  è il vettore dei parametri e  $h_0(t)$  è la funzione rischio di base. Quest'ultima è una componente non parametrica, positiva, ignota, uguale per tutti i soggetti, dipende solo dal tempo e rappresenta la distribuzione della funzione azzardo per il gruppo di base ( $X = 0$ ).

Poiché la funzione di verosimiglianza dipende dalla funzione rischio di base, che non è specificata, e la distruzione dei tempi non è nota, è possibile derivare le stime dei parametri con il metodo della verosimiglianza parziale. Questa produce stime non distorte e asintoticamente normali, e si comporta come una verosimiglianza propria sotto molti aspetti.

Il modello di Cox è valido per covariate fisse nel tempo e rispetta l'assunzione di proporzionalità dei rischi. Infatti, dati due individui con covariate  $X^A$  e  $X^B$ , il rapporto tra i rispettivi rischi  $HR$  (*hazard ratio*)

$$HR = \frac{h(t | X^A)}{h(t | X^B)} = \frac{\exp(\beta_1 X_1^A + \dots + \beta_p X_p^A)}{\exp(\beta_1 X_1^B + \dots + \beta_p X_p^B)} \quad (1.17)$$

è costante al variare del tempo. Tuttavia, è un modello molto flessibile e con opportuni accorgimenti può essere esteso a situazioni in cui le covariate dipendono dal tempo e l'assunzione di rischi proporzionali è violata.

Ottenuta la stima dei coefficienti del modello, è possibile risalire alla funzione di rischio cumulato e alla funzione di sopravvivenza per un soggetto con vettore di covariate  $X$ . A tal proposito, lo stimatore di Breslow fornisce una stima della funzione di rischio cumulato per il gruppo base  $\widehat{H}_0(t)$ , che può essere utilizzata per calcolare la funzione di sopravvivenza associata  $\widehat{S}_0(t)$  tramite la (1.8). Da questa, la stima della funzione di rischio cumulato  $\widehat{H}(t | X)$  e di sopravvivenza  $\widehat{S}(t | X)$  per un individuo specifico sono generate da

$$\widehat{H}(t | X) = \widehat{H}_0(t) \exp(\widehat{\beta}_1 X_1 + \dots + \widehat{\beta}_p X_p), \quad (1.18)$$

$$\widehat{S}(t | X) = \widehat{S}_0(t)^{\exp(\widehat{\beta}_1 X_1 + \dots + \widehat{\beta}_p X_p)}. \quad (1.19)$$



## Capitolo 2

### *Inverse Probability Weighting (IPW)*

Gli usuali metodi impiegati nelle analisi statistiche considerano le osservazioni disponibili come ugualmente importanti. In alcune circostanze, tuttavia, risulta appropriato variare il peso assegnato a ciascuna unità statistica. In questo capitolo, viene introdotto il metodo dell'*Inverse Probability Weighting (IPW)*. La comprensione dell'IPW è di fondamentale importanza per generare le equazioni di stima dei modelli di sopravvivenza che verranno proposti nella tesi. Per capire la logica sottostante e le situazioni in cui è conveniente farne uso, vengono presentate alcune delle sue note applicazioni negli studi osservazionali per misurare l'effetto di un trattamento, negli studi con dati mancanti e nell'analisi di sopravvivenza per produrre stimatori che tengano conto del meccanismo di censura. In quest'ultima, si definisce lo stimatore di Kaplan-Meier come uno stimatore *Inverse Probability of Censoring Weighting*.

#### 2.1 *Inverse Probability of Treatment Weighting (IPTW)*

Per studiare l'effetto di un trattamento su un *outcome*, si consideri un campione di numerosità  $n$ , in cui la variabile indicatrice  $Z_i$  descrive il trattamento che ciascun soggetto  $i$  riceve ( $Z_i = 1$  per il gruppo trattato;  $Z_i = 0$  per il gruppo di controllo). Per ogni individuo, è possibile definire due *outcome* potenziali  $Y_{1i}$  e  $Y_{0i}$ . Questi rappresentano rispettivamente il valore della variabile risposta in caso

di assegnazione al primo trattamento (gruppo trattato) e al secondo trattamento (gruppo di controllo), indipendentemente dal trattamento effettivamente ricevuto. In questo modo,  $Y_{1i} - Y_{0i}$  costituisce l'effetto causale d'interesse per l' $i$ -mo soggetto, mentre  $E(Y_{1i} - Y_{0i})$  l'effetto causale medio del trattamento (*Average Treatment Effect*, ATE). Tuttavia, poiché ogni soggetto  $i$  viene assegnato esclusivamente a uno dei due gruppi di trattamento, si osserva un solo *outcome* per ogni individuo  $Y_i = Y_{0i} + (Y_{1i} - Y_{0i})Z_i$  e si incorre così nel problema fondamentale dell'inferenza causale (Angrist e Pischke 2009). Negli studi randomizzati, l'assegnazione casuale del trattamento comporta l'indipendenza tra trattamento e *outcome* potenziali  $\{Y_{1i}, Y_{0i}\} \perp\!\!\!\perp Z_i$  e permette così la stima dell'effetto medio confrontando il valore medio dell'*outcome* dei trattati con quello dei non trattati

$$\begin{aligned} E(Y_i | Z_i = 1) - E(Y_i | Z_i = 0) &= E(Y_{1i} | Z_i = 1) - E(Y_{0i} | Z_i = 0) \\ &= E(Y_{1i}) - E(Y_{0i}) = E(Y_{1i} - Y_{0i}). \end{aligned}$$

Negli studi osservazionali, spesso i soggetti non trattati differiscono sistematicamente dai soggetti trattati, e l'indipendenza tra trattamento e *outcome* potenziali viene a mancare. Dunque,  $\{Y_{1i}, Y_{0i}\} \not\perp\!\!\!\perp Z_i$  e di conseguenza  $E(Y_{1i} | Z_i = 1) \neq E(Y_{1i})$ ,  $E(Y_{0i} | Z_i = 0) \neq E(Y_{0i})$ . In questi casi, comparare i due gruppi di trattamento direttamente condurrebbe a una stima distorta dell'effetto (*selection bias*). Risulta quindi opportuno considerare un ulteriore set di variabili di controllo  $X_i$ , in modo da garantire che il trattamento sia indipendente dagli *outcome* potenziali, condizionatamente al vettore di variabili  $X_i$  (*Conditional Independence Assumption*, CIA):

$$\{Y_{1i}, Y_{0i}\} \perp\!\!\!\perp Z_i | X_i. \quad (2.1)$$

Si assume che le variabili che soddisfano la (2.1) siano note e osservate. Sotto la (2.1), la condizione di indipendenza è soddisfatta anche condizionatamente a una funzione scalare delle covariate, definita da Rosenbaum e Rubin (1983) come *propensity score*  $e_i = Pr(Z_i = 1 | X_i)$  (si assume  $0 < Pr(Z = 1 | X) < 1$ ). Lo scopo di tale tecnica è mimare le caratteristiche di uno studio randomizzato. Infatti, il *propensity score* ha le proprietà di un *balancing score*: nei soggetti con lo stesso *propensity score*, la distribuzione delle variabili di controllo osservate sarà

uguale tra trattati e non trattati. Il *propensity score* esprime la probabilità condizionata di assegnazione al trattamento e può essere utilizzato in quattro modi per rimuovere l'effetto di variabili confondenti nella stima dell'effetto del trattamento sull'*outcome* (Austin 2011): *propensity score matching*, *stratification on the propensity score*, *inverse probability of treatment weighting*, *covariate adjustment*.

L'*Inverse Probability of Treatment Weighting* (IPTW) è una tecnica che definisce un sistema di pesi utilizzando il *propensity score* e permette di stimare l'effetto medio di trattamento in studi osservazionali. Per ogni individuo  $i$ , viene dunque definito un peso  $w_i = \frac{Z_i}{e_i} + \frac{1-Z_i}{1-e_i}$ . La stima ATE è allora calcolabile come

$$\frac{1}{n} \sum_{i=1}^n \frac{Z_i Y_i}{e_i} - \frac{1}{n} \sum_{i=1}^n \frac{(1-Z_i) Y_i}{1-e_i}. \quad (2.2)$$

Negli studi osservazionali il vero *propensity score* non è noto e viene comunemente stimato con un modello di regressione logistica. In alternativa, sono disponibili altre tecniche di classificazione (Hastie *et al.* 2009) quali il *boosting* e il *bagging*, gli alberi di classificazione, le foreste casuali e le reti neurali. Nella costruzione del *propensity score* vengono generalmente incluse come variabili di controllo tutte le variabili a disposizione pre-trattamento, variabili che influenzano l'assegnazione del trattamento, variabili che influenzano l'*outcome* (potenziali confondenti), variabili che influenzano sia il trattamento che l'*outcome* (veri confondenti). Non sono invece da considerare variabili misurate dopo il trattamento, perché possono essere influenzate o modificate dal trattamento. Assunzioni chiave del metodo sono che la modellazione del *propensity score* sia correttamente specificata e non ci siano variabili confondenti non osservate. Sotto tali condizioni, l'IPTW permette una stima dell'effetto di tipo causale.

## 2.2 IPW in studi con dati mancanti

Le analisi di dati sono tipicamente complicate dalla presenza di dati mancanti nei dataset a disposizione. Le unità statistiche si distinguono così in complete, se le variabili associate sono completamente osservate, e incomplete, se una o più

variabili contengono valori mancanti. Gli approcci più comuni prevedono la sostituzione dei dati mancanti con tecniche di imputazione, o il restringimento dell'analisi alle sole osservazioni complete. Nel caso in cui gli individui con casi completi differiscono sistematicamente da quelli con casi incompleti, quest'ultimo approccio comporta risultati distorti nelle analisi. L'IPW costituisce un metodo alternativo per ridurre tale distorsione (Seaman e White 2011).

Si supponga che il ricercatore sia interessato ad adattare un modello di regressione lineare generalizzato (*analysis model*) a un campione di numerosità  $n$ , dove  $Y_i$  e  $X_i$  sono rispettivamente la variabile risposta e il vettore di covariate per l'individuo  $i$ . La stima del vettore di parametri  $\theta$  del modello viene fornita dalla funzione punteggio  $U_i(\theta)$ , tramite  $\sum_{i=1}^n U_i(\theta) = 0$ . Restringendosi ai soli casi completi, viene introdotta la variabile indicatrice  $R_i$ , con valore  $R_i = 1$  se sia  $X_i$  che  $Y_i$  sono completamente osservate e  $R_i = 0$  in caso contrario. La stima di  $\theta$  risolve allora l'equazione  $\sum_{i=1}^n R_i U_i(\theta) = 0$ . Mediante l'IPW, a ogni caso completo viene assegnato un peso  $w_i$ , pari al reciproco della probabilità di essere un caso completo:

$$\sum_{i=1}^n R_i w_i U_i(\theta) = 0. \quad (2.3)$$

In questo modo, individui con una bassa probabilità di essere casi completi vengono pesati di più e hanno una maggiore influenza sulla stima dei parametri dell'*analysis model*. La stima dei pesi  $w_i$  segue la stessa logica della costruzione del *propensity score*, introdotto nel paragrafo precedente. Usualmente, viene stimato un modello logistico che ha come *outcome* la variabile  $R$  (*missingness model*), per prevedere la probabilità di essere un caso completo. Sia dato un ulteriore vettore di variabili aggiuntive  $Z$  osservate, utilizzate solamente al fine di costruire il sistema di pesi e quindi non incluse nell'*analysis model*. Il set di predittori  $H$  del *missingness model* viene scelto tra le variabili in  $X$ ,  $Y$  e  $Z$ , in modo che la condizione

$$Pr(R = 1 \mid X, Y, H) = Pr(R = 1 \mid H) \quad (2.4)$$

sia soddisfatta. In accordo con la costruzione del *propensity score*, è necessario che il *missingness model* sia correttamente specificato. Può capitare, per esempio, che

ad alcuni individui siano assegnati pesi molto grandi perché le probabilità stimate sono incorrette. Tale situazione è spesso dovuta a una scorretta specificazione del *missingness model*. È quindi opportuno stabilire una relazione adeguata tra i predittori  $H$  e la probabilità di dati mancanti, considerando eventualmente interazioni e trasformazioni delle covariate. Sotto le assunzioni di corretta specificazione del *missingness model* e di credibilità dell'equazione (2.4), l'IPW garantisce la rimozione della distorsione dei parametri nell'*analysis model*.

Il metodo appena descritto può essere adattato anche a indagini in cui è di interesse studiare delle caratteristiche rare nella popolazione (Mansournia e Altman 2016). In un campione rappresentativo della popolazione, in cui ogni unità ha la stessa probabilità di essere estratta, soggetti con caratteristiche peculiari saranno presenti in scarse quantità. Per aumentarne il numero, questi individui vengono sovracampionati incrementando la loro probabilità di selezione nel campione. Per tale motivo, il campione non risulta più rappresentativo della popolazione di riferimento e di conseguenza le stime d'interesse risultano distorte. Stime pesate tramite IPW correggono il *selection bias* generato: ogni individuo del campione viene pesato con il reciproco della probabilità di campionamento.

## 2.3 *Inverse Probability of Censoring Weighting (IPCW)*

Nel contesto di analisi di sopravvivenza, l'*Inverse Probability of Censoring Weighting* (IPCW) è stato proposto da Robins e Rotnitzky (1992) per produrre stimatori che tengano conto del meccanismo di censura a destra. Seguendo la logica del paragrafo precedente, la presenza di censura può essere pensata come un problema di dati mancanti. Lo scopo risulta così correggere per le osservazioni censurate dando dei pesi aggiuntivi ai soggetti non censurati, in modo da stimare classi di modelli di regressione come se la censura fosse assente, utilizzando equazioni di stima pesate del tipo (2.3).

Siano  $T$  e  $C$  variabili casuali continue che descrivono rispettivamente il tempo all'evento e il tempo alla censura a destra, e  $X^T = (X_1, \dots, X_p)$  un vettore di  $p$  covariate. Si assuma che  $T$  sia indipendente da  $C$ , date le covariate  $X$ . Per  $i =$

$1, \dots, n$ , siano  $(T_i, C_i, X_i)$  repliche indipendenti di  $(T, C, X)$  e si supponga vengano osservate  $(Y_i, \delta_i, X_i)$ , con  $Y_i = \min(T_i, C_i)$  e  $\delta_i = I(T_i \leq C_i)$ . In particolare,  $Y_i = C_i$  quando  $\delta_i = 0$  (episodio  $i$ -mo censurato) e  $Y_i = T_i$  quando  $\delta_i = 1$  (episodio  $i$ -mo non censurato). Si consideri la funzione di sopravvivenza di  $C$ , indicata come

$$G(t | X) = Pr(C > t | X), \quad (2.5)$$

che può dipendere o meno dal vettore di covariate  $X$ . Essa costituisce la probabilità condizionata di rimanere non censurati fino al tempo  $t$ . L'IPCW assegna un peso ai soggetti non censurati, definito dal reciproco della funzione di sopravvivenza di  $C$ . Nella pratica, la (2.5) viene stimata dai dati basandosi su  $(Y_i, 1 - \delta_i, X_i; i = 1, \dots, n)$ : gli individui che sperimentano l'evento d'interesse vengono considerati come osservazioni censurate e viceversa (Willems 2014). Per fare questo, è necessario stabilire se il meccanismo di censura sia indipendente dal vettore di covariate  $X$  o dipendente.

Quando il meccanismo di censura è indipendente dalle covariate, la funzione di sopravvivenza per il tempo alla censura  $G(t | X) = G(t)$  può essere stimata tramite lo stimatore di Kaplan-Meier (Sun *et al.* 2011). Si rimanda alla (2.8) per la formulazione. In situazioni di censura dipendente, la stima di  $G(t | X)$  deve tener conto dell'informazione data dalle covariate. Risulta quindi necessario modellare il tempo alla censura  $C$  con il vettore di covariate  $X$  a disposizione (Sun e Zhang 2009). Nel caso in cui ci sia un numero limitato di variabili discrete e la numerosità  $n$  del campione sia sufficientemente elevata, la funzione di sopravvivenza di  $C$  può essere calcolata tramite la versione stratificata dello stimatore di Kaplan-Meier. Questa è una tecnica di stima semplice e diretta da applicare per la costruzione di pesi dipendenti dalle covariate. Altrimenti, in presenza di covariate continue o numerosità moderata, è opportuno stimare modelli di regressione correttamente specificati per la distribuzione dei tempi alla censura. A tal proposito, si può modellare la funzione di rischio  $h_C(t | X)$  di  $C$  tramite un modello di Cox

$$h_C(t | X) = h_{C_0}(t) \exp(\beta_C^T X), \quad (2.6)$$

dove  $h_{C_0}(t)$  è la funzione rischio di base e  $\beta_C$  il vettore dei parametri. Da questo, attraverso lo stimatore di Breslow, viene stimata la funzione di sopravvivenza as-

sociata (paragrafo 1.2.2). Il modello di Cox non è detto sia adeguato per descrivere il meccanismo di censura e altri approcci potrebbero condurre a soluzioni migliori. Il modello di regressione additivo di Aalen (1989) e i modelli a tempi accelerati (Klein e Moeschberger 2003) costituiscono altri possibili stimatori di  $G(t | X)$ . In ogni caso, prima di procedere con la costruzione di pesi dipendenti dalle covariate, conviene indagare sulla effettiva dipendenza di  $C$  dalle covariate, e validare le assunzioni dei modelli utilizzati per stimare  $G(t | X)$ , per esempio, testando la proporzionalità degli azzardi se si è scelto il modello di Cox.

Un risultato rilevante nell'utilizzo dell'*Inverse Probability Weighting* in ambito di sopravvivenza è fornito da Satten e Datta (2001). Questi hanno dimostrato che lo stimatore di Kaplan-Meier è esprimibile come uno stimatore IPCW. Si assuma che  $T$  e  $C$  siano indipendenti. Per  $j = 1, \dots, J$ , siano  $\gamma_j$  i tempi ordinati alla censura o all'evento,  $n_j$  il numero di persone che subiscono l'evento al tempo  $\gamma_j$  e  $m_j$  il numero di episodi censurati al tempo  $\gamma_j$ . Sia  $R(t) = \sum_{j=1}^J (n_j + m_j) I(\gamma_j \geq t)$  il numero di persone a rischio di sperimentare l'evento all'istante  $t$ . Lo stimatore di Kaplan-Meier della (1.14) può essere riscritto come

$$\hat{S}_{KM}(t) = \prod_{j|\gamma_j \leq t} \left(1 - \frac{n_j}{R(\gamma_j)}\right). \quad (2.7)$$

Lo stimatore di Kaplan-Meier della funzione di sopravvivenza per il tempo alla censura è allora definito da

$$\hat{G}_{KM}(t) = \prod_{j|\gamma_j \leq t} \left(1 - \frac{m_j}{R(\gamma_j)}\right). \quad (2.8)$$

In assenza di censura, la funzione di ripartizione di  $T$ , ossia  $F(t)$ , potrebbe essere stimata empiricamente dai dati come media di termini indipendenti e identicamente distribuiti

$$F^*(t) = \frac{1}{n} \sum_{i=1}^n I(T_i \leq t). \quad (2.9)$$

Il metodo IPCW permette di stimare la funzione di ripartizione di  $T$  in presenza di censura. Come sopra, si considera la media di termini indipendenti e identicamente distribuiti, ma ci si restringe ai soli casi in cui la censura è assente ( $\delta_i = 1$ ) e si

pesano le unità per il reciproco della funzione di sopravvivenza di  $C$ . La stima di quest'ultima è fornita dalla (2.8). Si ottiene

$$\widehat{F}_{IPCW}(t) = \frac{1}{n} \sum_{i=1}^n \frac{I(Y_i \leq t)\delta_i}{\widehat{G}_{KM}(Y_{i-})}. \quad (2.10)$$

Nella (2.10), poiché  $I(Y_i \leq t)\delta_i = I(T_i \leq t)\delta_i$ , il termine  $I(T_i \leq t)$  è interscambiabile con  $I(Y_i \leq t)$ . Da questa, la stimatore IPCW della funzione di sopravvivenza è definito da

$$\widehat{S}_{IPCW}(t) = 1 - \widehat{F}_{IPCW}(t) = 1 - \frac{1}{n} \sum_{i=1}^n \frac{I(Y_i \leq t)\delta_i}{\widehat{G}_{KM}(Y_{i-})}. \quad (2.11)$$

Nello studio citato, Satten e Datta provano l'uguaglianza tra  $\widehat{S}_{IPCW}(t)$  e  $\widehat{S}_{KM}(t)$ . Si noti che  $\widehat{F}_{IPCW}(t)$  e  $\widehat{F}_{KM}(t) = 1 - \widehat{S}_{KM}(t)$  sono entrambe curve a gradini con continuità a destra e possibili salti in  $\gamma_j$ . La dimostrazione si basa sul fatto che se, per un fissato tempo  $\gamma_j$ , le quantità  $\widehat{F}_{IPCW}(t)$  e  $\widehat{F}_{KM}(t)$  presentano un salto di pari grandezza, allora  $\widehat{F}_{IPCW}(t) = \widehat{F}_{KM}(t)$  e ciò comporta  $\widehat{S}_{IPCW}(t) = \widehat{S}_{KM}(t)$ .

# Capitolo 3

## Stima del *Mean Time*

Si riprendano il contesto e le notazioni utilizzate nel paragrafo 2.3. Siano  $T$  e  $C$  variabili casuali continue che descrivono rispettivamente il tempo all'evento e il tempo alla censura a destra. Si assuma che il meccanismo di censura sia non informativo. Per  $i = 1, \dots, n$ , siano  $(T_i, C_i)$  repliche indipendenti di  $(T, C)$  e si supponga vengano osservate  $(Y_i, \delta_i)$ , con  $Y_i = \min(T_i, C_i)$  e  $\delta_i = I(T_i \leq C_i)$ . In particolare,  $Y_i = C_i$  quando  $\delta_i = 0$  (episodio  $i$ -mo censurato) e  $Y_i = T_i$  quando  $\delta_i = 1$  (episodio  $i$ -mo non censurato).

Il tempo medio all'evento (*mean time*) è stato definito nel paragrafo 1.1.3 come  $\mu = E(T)$ , e corrisponde al *residual mean time* calcolato all'istante  $t = 0$ , ossia  $m(0)$ . In letteratura, tale quantità prende spesso il nome di tempo di sopravvivenza medio (*mean survival time*). Si presenta la stima del *mean time* per campioni omogenei, soffermandosi sulle problematiche causate dalla presenza di censura a destra negli studi. Questa conduce, infatti, all'utilizzo del *restricted mean time*. Successivamente, si dimostra come tale quantità sia utile nel comparare la sopravvivenza tra gruppi diversi in alternativa al comune *hazard ratio*. Infine, si discutono metodi che permettono il confronto tra gruppi aggiustato con delle covariate aggiuntive.

### 3.1 Stima per campioni omogenei

Nel paragrafo 1.1.3, è stato mostrato che il tempo medio all'evento  $\mu$  corrisponde all'area al di sotto della corrispondente funzione di sopravvivenza:

$$\mu = m(0) = \int_0^{\infty} S(t)dt. \quad (3.1)$$

In presenza di censura, la media campionaria dei tempi all'evento  $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$  sottostima il vero valore del tempo medio (Karrison 1997). Nei loro studi, Kaplan e Maier (1958) mostrano che un naturale stimatore per  $\mu$  si ottiene sostituendo  $S(t)$  con  $\hat{S}(t)$ , ottenuto tramite lo stimatore *Product-Limit* della funzione di sopravvivenza. Dunque,

$$\hat{\mu} = \int_0^{\infty} \hat{S}(t)dt. \quad (3.2)$$

Lo stimatore *Product-Limit* è una funzione a scalini e quindi nella (3.2) non c'è alcuna funzione matematica da integrare. Tuttavia, integrare lo stimatore *Product-Limit* equivale a dividere l'area al di sotto della curva in rettangoli, calcolare l'area di ciascun rettangolo e sommarle (Sheldon 2013):

$$\hat{\mu} = \sum_{i=1}^D \hat{S}(t_{i-1})(t_i - t_{i-1}), \quad (3.3)$$

dove  $t_0 = 0$  e  $t_D$  è l'ultimo tempo osservato.

A causa della inevitabile presenza di censura a destra nel campione, la coda della funzione di sopravvivenza è spesso difficile da stimare correttamente. Infatti, per come è costruito, lo stimatore *Product-Limit* è ben definito solo fino al più grande tempo osservato nello studio  $t_{\max}$ . Quando in  $t_{\max}$  si osserva un evento, la stima della curva di sopravvivenza da  $t_{\max}$  in poi vale 0 e lo stimatore (3.2) è non distorto. Al contrario, se in  $t_{\max}$  si osserva una censura, la stima della funzione di sopravvivenza da  $t_{\max}$  in poi non è definita perché non è noto quando l'osservazione associata subirà l'evento. In tal caso, poiché  $\hat{S}(t)$  non è determinata ovunque,  $\hat{\mu}$  è indefinita. Una possibile soluzione è trattare l'ultima osservazione dello studio (*Last Observation Time*, LOT) come se avesse sperimentato l'evento, anche se effettivamente censurata (Efron 1967). Questa comporta la convergenza

di  $\widehat{S}(t)$  a 0 dopo  $t_{\max}$  e permette di stimare il tempo medio restringendosi all'intervallo  $[0, t_{\max}]$ . In alternativa, si può anche assumere che l'ultimo tempo all'evento sia l'ultimo tempo osservato (*Last Event Time*, LET), trascurando le censure successive. Irwin (1949) suggerisce di considerare il tempo medio restringendosi a un intervallo prefissato  $[0, \tau]$ , dove  $\tau$  viene scelto dal ricercatore come il più grande tempo possibile di sopravvivenza nello studio. Tale quantità prende il nome di *restricted mean time* ed è interpretabile come l'aspettativa di vita fino all'orizzonte temporale  $\tau$ :

$$\mu(\tau) = E(\min(T, \tau)) = \int_0^\tau S(t)dt, \quad (3.4)$$

per qualsiasi tempo  $\tau > 0$ . Esso rappresenta la media di  $T$  per tutti gli individui dello studio seguiti fino al tempo  $\tau$  e corrisponde all'area al di sotto della associata funzione di sopravvivenza fino all'istante  $\tau$ . Si noti che l'area al di sopra della curva di sopravvivenza fino all'istante  $\tau$  è invece detta *restricted mean time lost* e corrisponde a  $\tau - \mu(\tau)$  (Zhao *et al.* 2016). In altri termini,  $\mu$  è il limite di  $\mu(\tau)$  per  $\tau \rightarrow \infty$ . Poiché  $\int_0^{t_2} S(t)dt > \int_0^{t_1} S(t)dt$  per  $t_2 > t_1$ , il *mean time* eccede il *restricted mean time* per qualsiasi  $\tau$ . Dunque,  $\mu(\tau)$  è una funzione monotona crescente di  $\tau$  (Royston e Parmar 2011). Il *restricted mean time* è stimato in modo consistente dall'area al di sotto della curva di sopravvivenza di Kaplan-Maier fino all'istante  $\tau$ :

$$\widehat{\mu}(\tau) = \int_0^\tau \widehat{S}(t)dt, \quad (3.5)$$

dove  $\tau$  è il più grande tempo osservato o è prestabilito dal ricercatore. Utilizzando la notazione del paragrafo 1.2.1, la varianza dello stimatore (3.5) è

$$\widehat{V}(\widehat{\mu}(\tau)) = \sum_{i=1}^D \left( \int_{t_i}^\tau \widehat{S}(t)dt \right)^2 \frac{E_i}{R_i(R_i - E_i)}, \quad (3.6)$$

(Klein e Moeschberger 2003) e l'intervallo di confidenza a livello  $1 - \alpha$  per il tempo medio è dunque dato da

$$\widehat{\mu}(\tau) \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{V}(\widehat{\mu}(\tau))}. \quad (3.7)$$

## 3.2 Stima per campioni eterogenei

Spesso, i dati di uno studio clinico sono eterogenei per il disegno sperimentale (se per esempio si vuole confrontare due diversi tipi di trattamento) o per le caratteristiche delle unità statistiche (i soggetti studiati possono avere aspetti clinici e biologici diversi). In questi casi, è d'interesse confrontare due o più gruppi. Negli studi longitudinali, in cui l'*endpoint* primario è il tempo a uno specifico evento, le sottostanti funzioni di sopravvivenza e di rischio forniscono una misura del comportamento dell'evento d'interesse nel tempo. Tuttavia, a differenza della funzione di sopravvivenza, valori assoluti della funzione di rischio sono difficili da interpretare clinicamente.

Nel comparare più gruppi, il confronto delle curve di sopravvivenza in un singolo punto non è un modo di procedere corretto, in quanto la scelta del punto risulta arbitraria. Un possibile approccio è considerare il test dei ranghi logaritmici. La differenza tra due gruppi è comunemente quantificata con l'*hazard ratio* ( $HR$ ) stimato tramite il modello di Cox, assumendo che il rapporto tra le funzioni di azzardo dei due gruppi sia costante al variare del tempo. L'ipotesi nulla  $HR = 1$  (uguaglianza dei due gruppi) viene accertata con il test del rapporto di verosimiglianza sul parametro associato nel modello di Cox ( $\beta = 0$ ). Tale test è asintoticamente equivalente al test dei ranghi logaritmici.

Ciò nonostante, la violazione dell'assunzione di proporzionalità mette in dubbio l'interpretazione di un singolo *hazard ratio*, in quanto questo ora varia nel tempo. Inoltre, in assenza di proporzionalità, il test dei ranghi logaritmici, seppur non necessita di assunzioni sulla forma delle sottostanti funzioni di distribuzione, perde potenza (per esempio quando le curve di sopravvivenza dei gruppi si incrociano). Nel modello di Cox, la suddetta assunzione può essere testata con procedure di *goodness-of-fit* basate sui residui di Schoenfeld e sui residui cumulati di martingala (Cortese *et al.* 2010). La stima non parametrica dell'*hazard ratio* come funzione del tempo proposta da Gilbert *et al.* (2002) può fornire, invece, un'utile ispezione visiva dell'andamento dell'*hazard ratio* nel tempo.

Considerato ciò, Uno *et al.* (2014), Royston e Parmar (2011) propongono la

differenza (o il rapporto) dei *restricted mean time* nei due gruppi  $(\mu_0(\tau), \mu_1(\tau))$  come alternativa robusta nel misurare la differenza tra due curve di sopravvivenza

$$\Delta\mu(\tau) = \mu_1(\tau) - \mu_0(\tau). \quad (3.8)$$

La differenza (o il rapporto) tra le mediane o le funzioni di sopravvivenza in un fissato punto  $t$  costituiscono altre misure robuste, ma presentano alcuni svantaggi. La mediana ( $Me$ ) viene calcolata tramite lo stimatore di Kaplan-Meier ( $\widehat{Me} = \inf\{t : \widehat{S}(t) \leq 0.5\}$ ), dunque potrebbe non essere identificata empiricamente quando la frequenza dell'evento è relativamente bassa o il tempo di *follow-up* è breve. La stima della funzione di sopravvivenza in un fissato punto  $t$ , invece, è espressa dal valore dello stimatore di Kaplan-Meier in  $t$ , e la differenza tra gruppi è vincolata dalla specifica scelta di  $t$ . Al contrario, la (3.8) è in grado di catturare il profilo globale della differenza tra gruppi con una singola misura riassuntiva. Infatti, attraverso l'integrazione della funzione di sopravvivenza, ciascun *restricted mean time* prende in considerazione l'intera funzione di sopravvivenza fino a  $\tau$ . Inoltre, grazie alla stima di  $\mu_1(\tau)$  e  $\mu_0(\tau)$ , la differenza dei *restricted mean time* permette migliori valutazioni a livello clinico sul contrasto tra gruppi rispetto al convenzionale *hazard ratio*. La differenza dei *restricted mean time* viene considerata una quantità primaria quando l'assunzione di proporzionalità è violata, ma può anche essere riportata come misura secondaria nei casi in cui l'assunzione di proporzionalità è soddisfatta.

Uno dei modi di procedere per ottenere la stima dei *restricted mean time* per ogni gruppo è specificare un modello di Cox avente come covariata la variabile  $Z$  identificativa del gruppo. Per esempio, nel caso del confronto tra due trattamenti,  $Z$  è una variabile indicatrice ( $Z = 0$  per il gruppo di controllo;  $Z = 1$  per il gruppo trattato) e il modello è esprimibile come

$$h(t | Z) = h_0(t) \exp(\beta_1 Z), \quad (3.9)$$

dove  $h_0(t)$  è la funzione di rischio del gruppo di controllo e  $\beta_1$  il logaritmo dell'*hazard ratio* (gruppo trattato contro gruppo di controllo). Tramite lo stimatore di Breslow, è possibile ricavare la stima della funzione di sopravvivenza del gruppo di

controllo  $\widehat{S}_0(t)$  e del gruppo trattato  $\widehat{S}_1(t) = \widehat{S}_0(t)^{\exp(\widehat{\beta}_1)}$ , come descritto nel paragrafo 1.2.2. Tali quantità vengono integrate numericamente su  $[0, \tau]$  per ottenere il *restricted mean time* del gruppo di controllo

$$\widehat{\mu}_0(\tau) = \int_0^\tau \widehat{S}_0(t) dt \quad (3.10)$$

e del gruppo trattato

$$\widehat{\mu}_1(\tau) = \int_0^\tau \widehat{S}_1(t) dt. \quad (3.11)$$

La procedura descritta, tuttavia, produce risultati distorti quando l'assunzione di proporzionalità del modello di Cox non è rispettata. Per questo motivo, è preferibile ricavare  $\widehat{S}_0(t)$  e  $\widehat{S}_1(t)$  singolarmente con lo stimatore di Kaplan-Meier, che equivale ad utilizzare funzioni di sopravvivenza stimate con un modello di Cox stratificato per la variabile trattamento (Kalbfleisch e Prentice 2002). Si noti che tali procedure possono essere estese anche al calcolo dei *restricted mean time* ( $\widehat{\mu}_g(\tau)$ ) per più di due gruppi ( $g = 0, 1, \dots$ ). In aggiunta, i *restricted mean time* possono essere stimati con il metodo di Andersen *et al.* (2004) delle *pseudo-observations* e con i *flexible parametric survival model* proposti da Royston e Parmar (2002). Si vedano i relativi articoli per ulteriori approfondimenti.

Ottenute  $\widehat{\mu}_1(\tau)$  e  $\widehat{\mu}_0(\tau)$ , la stima della differenza tra i due gruppi è allora

$$\widehat{\Delta\mu}(\tau) = \widehat{\mu}_1(\tau) - \widehat{\mu}_0(\tau), \quad (3.12)$$

che ha *standard error*

$$se(\widehat{\Delta\mu}(\tau)) = \sqrt{V(\widehat{\mu}_1(\tau)) + V(\widehat{\mu}_0(\tau))}. \quad (3.13)$$

Ciò permette di costruire il relativo intervallo di confidenza riferendosi alle tabelle della distribuzione normale standard

$$\widehat{\Delta\mu}(\tau) \pm z_{1-\frac{\alpha}{2}} se(\widehat{\Delta\mu}(\tau)), \quad (3.14)$$

e di saggiare l'ipotesi nulla di uguaglianza dei gruppi, tramite

$$z = \frac{\widehat{\Delta\mu}(\tau)}{se(\widehat{\Delta\mu}(\tau))}. \quad (3.15)$$

### 3.3 Confronto tra gruppi con aggiustamento di covariate

Nel paragrafo 3.2, è stata presentata una procedura basata sul *restricted mean time* per confrontare la sopravvivenza in gruppi diversi. Karrison (1987, 1997) estende il metodo per permettere di incorporare un vettore di covariate  $X$  nell'analisi. I *restricted mean time* di ciascun gruppo possono essere stimati con modelli a rischi proporzionali. A tal fine, nel paragrafo 3.2 è stato specificato un modello di Cox (3.9) avente come variabile esplicativa l'identificativo del gruppo. Tale modello permette l'inclusione di un set aggiuntivo di covariate, ma richiede comunque che i coefficienti associati al gruppo, oltre a quelli delle covariate aggiuntive, siano costanti nel tempo. Karrison (1987, 1997) propone una generalizzazione del modello di Cox, con l'obiettivo di evitare l'assunzione sottostante di proporzionalità degli azzardi. Si ricorre a un modello esponenziale a tratti, composto da funzioni di rischio di base costanti a tratti separate per ciascun gruppo. Le covariate hanno un effetto moltiplicativo sulla funzione di rischio che è lo stesso per entrambi i gruppi. Simbolicamente,

$$h_g(t | X) = h_{gv} \exp(\beta^T X) \quad t \in I_v, \quad (3.16)$$

dove  $g$  indica il gruppo ( $g = 0, 1$ ) e  $\{I_v\}$  è una serie di intervalli in  $[0, \tau]$ . In questo modo, l'assunzione di proporzionalità è fatta rispetto all'effetto delle covariate e non rispetto al gruppo di appartenenza. Zucker (1998) descrive una procedura semplificata per implementare l'approccio di Karrison. Per evitare l'assunzione di proporzionalità dell'effetto del gruppo, Zucker propone un modello di Cox stratificato per gruppo:

$$h_g(t | X) = h_{0g}(t) \exp(\beta^T X), \quad (3.17)$$

dove la funzione di rischio di base  $h_{0g}(t)$  è non specificata. Il suddetto modello può essere stimato con la verosimiglianza parziale costruendo separate funzioni di verosimiglianza parziale per ciascun gruppo, moltiplicando tra loro le funzioni e scegliendo poi i valori dei coefficienti che massimizzano tale funzione (Kalbfleisch e Prentice 2002). A differenza del modello esponenziale a tratti di Karrison (3.16),

il modello stratificato di Cox ha il vantaggio di non richiedere la divisione dell'asse temporale in sottointervalli.

Per ogni gruppo, si assuma che le osservazioni siano indipendenti e identicamente distribuite. Sia  $n_g$  la numerosità di ciascun gruppo ( $g = 0, 1$ ) e  $n = n_0 + n_1$  la numerosità del campione. La funzione di rischio cumulato di base per il gruppo  $g$  specifico,  $H_{0g}(t) = \int_0^t h_{0g}(s)ds$ , è stimata con la procedura di Breslow ( $\widehat{H}_{0g}(t)$ ). La funzione di sopravvivenza di base per il gruppo  $g$  è stimata con  $\widehat{S}_{0g}(t) = \exp(-\widehat{H}_{0g}(t))$ . La funzione di sopravvivenza per il gruppo  $g$  condizionata a specifici valori delle covariate, è allora stimabile come  $\widehat{S}_g(t | X) = \exp(-e^{\widehat{\beta}^T X} \widehat{H}_{0g}(t))$ . Questa viene integrata numericamente per stimare il *restricted mean time* del gruppo  $g$  per fissati valori delle covariate  $X$ :

$$\widehat{\mu}_g(\tau | X) = \int_0^\tau \widehat{S}_g(t | X) dt. \quad (3.18)$$

Infine, un confronto tra i due gruppi aggiustato è dato da

$$\widehat{\Delta\mu}(\tau | X) = \widehat{\mu}_1(\tau | X) - \widehat{\mu}_0(\tau | X). \quad (3.19)$$

L'approccio descritto permette la stima e il confronto dei *restricted mean time* aggiustati per un qualsiasi fissato valore di  $X$ . Tuttavia, confronti in livelli diversi del vettore di covariate conducono a differenti stime dell'effetto. Per ottenere un singolo e generale confronto aggiustato, è possibile calcolare la curva di sopravvivenza "media" di ciascun gruppo, utilizzando l'intera distribuzione empirica delle covariate di entrambi i gruppi:

$$\widehat{S}_{.g}(t) = \frac{1}{n} \sum_{h=0}^1 \sum_{j=1}^{n_h} \widehat{S}_g(t | X_{hj}). \quad (3.20)$$

Di conseguenza, la stima del corrispondente *restricted mean time* è

$$\widehat{\mu}_{.g}(\tau) = \int_0^\tau \widehat{S}_{.g}(t) dt, \quad (3.21)$$

e la stima dell'effetto è

$$\widehat{\Delta\mu}_{.}(\tau) = \widehat{\mu}_{.1}(\tau) - \widehat{\mu}_{.0}(\tau). \quad (3.22)$$

Per ulteriori approfondimenti e le procedure di inferenza si rimanda al relativo articolo (Zucker 1998).

## Capitolo 4

# Modelli di regressione per il *Residual Mean Time*

Negli studi di coorte e osservazionali, il campione è prevalentemente in forma eterogenea e un problema che frequentemente ricorre nell'analisi di dati di sopravvivenza è aggiustare la funzione di sopravvivenza tenendo conto dell'informazione fornita da variabili esplicative. Si consideri il tempo all'evento  $T > 0$ , variabile già discussa nei capitoli precedenti, e un vettore di covariate  $X^T = (X_1, \dots, X_p)$ . È di interesse stabilire la relazione tra il tempo all'evento e una o più variabili esplicative. Per esempio, può essere utile identificare i fattori di rischio di una malattia, controllando per confondenti, o comparare le funzioni di sopravvivenza per due o più trattamenti, al fine di determinare la prognosi di un paziente con determinate caratteristiche.

I modelli per la funzione di rischio costituiscono il principale approccio per l'analisi di regressione per dati di sopravvivenza con censura a destra. Nella prima parte del capitolo si mostra come questi possano essere utilizzati per stimare indirettamente il *residual mean time*. In alternativa, si introducono alcuni modelli di regressione che hanno il vantaggio di fornire un'interpretazione diretta dell'effetto delle covariate sul *residual mean time*. Tra questi, si presenta una classe di modelli per il *restricted residual mean time*, basata su equazioni di stima pesate tramite

IPCW per modellare il meccanismo di censura a destra. In aggiunta, si dimostra che in presenza di variabili esplicative qualitative la stima IPCW del *restricted residual mean time* è equivalente a uno stimatore costruito con lo stimatore di Kaplan-Meier.

## 4.1 Modelli indiretti e diretti per il *Residual Mean Time*

I modelli di regressione per dati di sopravvivenza sono frequentemente specificati tramite la funzione di rischio. Tra questi, i *multiplicative hazard rate model* esprimono la funzione di rischio condizionata come prodotto tra la funzione di rischio di base  $h_0(t)$  e una funzione non negativa delle covariate  $c(\beta^T X)$ :

$$h(t | X) = h_0(t)c(\beta^T X). \quad (4.1)$$

Nei *multiplicative hazard rate model*, l'*hazard ratio* di due individui con vettore di covariate  $X^A$  e  $X^B$ ,  $HR = \frac{h(t|X^A)}{h(t|X^B)} = \frac{c(\beta^T X^A)}{c(\beta^T X^B)}$ , è costante nel tempo. La funzione di sopravvivenza condizionata può essere espressa tramite la funzione di sopravvivenza del gruppo di base  $S_0(t)$ :

$$S(t | X) = S_0(t)c(\beta^T X). \quad (4.2)$$

Nella maggior parte delle applicazioni viene utilizzato il modello di Cox con  $c(\beta^T X) = \exp(\beta^T X)$ , in cui la funzione di rischio di base  $h_0(t)$  è trattata in modo non parametrico ed è una funzione arbitraria non negativa. Un'altra modellazione possibile della funzione di rischio è data dalla famiglia degli *additive hazard rate model*

$$h(t | X) = h_0(t) + \exp(\beta^T X), \quad (4.3)$$

di cui fanno parte il modello semiparametrico di Lin e Ying (1994) e il modello non parametrico di Aalen (1989) (quando  $\beta = \beta(t)$ , ossia i coefficienti associati alle covariate variano nel tempo).

Nei modelli per la funzione di rischio citati, l'effetto di un trattamento o delle covariate è dato dal rapporto o dalla differenza delle funzioni di rischio. Nei classici modelli di regressione lineare con variabile risposta quantitativa l'effetto delle

covariate è fornito dalla differenza di valori medi. Così come avviene nei modelli lineari, anche negli studi in cui la variabile risposta è il tempo di sopravvivenza, potrebbe essere utile esprimere l'effetto delle covariate sulla scala del tempo medio di sopravvivenza (Andersen *et al.* 2004). Inoltre, come approfondito nel Capitolo 3, va evidenziato che il tempo medio di sopravvivenza (*mean time*) è una quantità che presenta numerosi vantaggi nel misurare l'effetto delle covariate o di un trattamento rispetto all'*hazard ratio*.

A tal proposito, il modello di Cox fornisce una stima indiretta del *mean time* e più in generale del *residual mean time*. Infatti, specificare un modello per la funzione di rischio del tipo (4.1), come il modello di Cox, comporta la modellazione della funzione di sopravvivenza per via (4.2), e di conseguenza del *mean time* tramite

$$E(T | X) = \int_0^{\infty} S(t | X) dt, \quad (4.4)$$

e del *residual mean time* tramite

$$E(T - t | T \geq t, X) = \frac{\int_t^{\infty} S(s | X) ds}{S(t | X)}. \quad (4.5)$$

Come evidenziato nel paragrafo 3.1, a causa della presenza di censura a destra, tipicamente non si ha molta informazione sulla coda della funzione di sopravvivenza e quindi il tempo medio di sopravvivenza risulta difficile da stimare correttamente. Dunque, è preferibile considerare il *restricted mean time*

$$\mu(\tau | X) = E(\min(T, \tau) | X) = \int_0^{\tau} S(t | X) dt, \quad (4.6)$$

per  $\tau > 0$ , e il *restricted residual mean time*

$$E(\min(T, \tau) - t | T \geq t, X) = \frac{\int_t^{\tau} S(s | X) ds}{S(t | X)}, \quad (4.7)$$

per  $0 \leq t \leq \tau$ . Come riportato da Tian *et al.* (2013), in un modello di Cox la stima della funzione di sopravvivenza è fornita tramite lo stimatore di Breslow e questa permette la stima del *restricted mean time* e del *restricted residual mean time* per via (4.6) e (4.7). Per la stima del *restricted mean time*, per esempio, si calcolerà

$$\hat{\mu}(\tau | X) = \int_0^{\tau} \exp(-\hat{H}_0(t) e^{\hat{\beta}^T X}) dt, \quad (4.8)$$

dove  $\widehat{\beta}$  e  $\widehat{H}_0(t)$  sono rispettivamente la stima di massima verosimiglianza parziale di  $\beta$  e lo stimatore di Breslow della funzione di rischio cumulato di base,  $H_0(t)$ .

In aggiunta, nel Capitolo 3, sono stati presentati altri studi riguardanti l'analisi di regressione del *restricted mean time*, estendibili al *restricted residual mean time* con le relazioni appena considerate. Karrison (1987, 1997) descrive un modello a effetti proporzionali del tipo (4.1) con funzione di base costante a tratti, e utilizza il suddetto modello per  $\mu(\tau | X)$ , attraverso la relazione (4.6). Zucker (1998) generalizza tale approccio basandosi su un modello di Cox stratificato.

I modelli di regressione citati forniscono una previsione del *restricted mean time* e del *restricted residual mean time*. Tuttavia, queste quantità non sono modellate direttamente con le covariate, ma per mezzo della funzione di azzardo, e dunque l'influenza delle covariate sulle quantità d'interesse non è descritta esplicitamente da dei coefficienti di regressione. A tale scopo, per valutare l'effetto delle covariate sul *residual mean time*, indicato con  $m(t) = E(T - t | T > t)$ , Maguluri e Zhang (1994) propongono un modello a effetti proporzionali

$$m(t | X) = m_0(t) \exp(\beta^T X), \quad (4.9)$$

dove  $\beta$  è il vettore dei parametri del modello e  $m_0(t)$  una funzione *residual mean time* del gruppo base ( $X = 0$ ) non nota. Nel relativo articolo, i parametri vengono stimati con dati completi (cioè in assenza di censura). Chen *et al.* (2005), Chen e Cheng (2005) riprendono il modello e presentano procedure di stima che considerano la presenza di censura a destra. Successivamente, Chen e Cheng (2006), Chen (2007) presentano un modello additivo

$$m(t | X) = m_0(t) + \beta^T X \quad (4.10)$$

e ne discutono vari metodi di stima in presenza e assenza di censura. Infine, Sun e Zhang (2009), Sun *et al.* (2011) studiano una classe di modelli trasformati nella forma

$$m(t | X) = g(m_0(t) + \beta^T X), \quad (4.11)$$

dove  $g(\cdot)$  è una funzione legame non negativa, due volte differenziabile e strettamente crescente. Si noti che per  $g(x) = \exp(x)$  e  $g(x) = x$  il modello si riduce al

(4.9) e al (4.10). La stima dei parametri di regressione, proposta nei relativi articoli, avviene con equazioni di stima generalizzate che utilizzano la tecnica *Inverse Probability of Censoring Weighting*, permettendo così di modellare la censura a destra.

## 4.2 Una classe di modelli diretti per il *Residual Mean Time*

Siano  $T$  e  $C$  variabili casuali continue che descrivono rispettivamente il tempo all'evento e il tempo alla censura a destra, e  $X^T = (X_1, \dots, X_p)$  un vettore di  $p$  covariate. Si assuma che  $T$  sia indipendente da  $C$ , date le covariate  $X$ . Per  $i = 1, \dots, n$ , siano  $(T_i, C_i, X_i)$  repliche indipendenti di  $(T, C, X)$  e si supponga vengano osservate  $(Y_i, \delta_i, X_i)$ , con  $Y_i = \min(T_i, C_i)$  e  $\delta_i = I(T_i \leq C_i)$ . In particolare,  $Y_i = C_i$  quando  $\delta_i = 0$  (episodio  $i$ -mo censurato) e  $Y_i = T_i$  quando  $\delta_i = 1$  (episodio  $i$ -mo non censurato). Per il momento, si assuma anche che  $C$  sia indipendente dalle covariate  $X$  e che  $G(\tau) > 0$  per ogni punto di restrizione fissato  $\tau$ , dove  $G(s)$  rappresenta la funzione di sopravvivenza di  $C$ .

### 4.2.1 Presentazione del modello

In questo capitolo, si considera una classe di modelli di regressione diretti per il *restricted residual mean time* (Cortese *et al.* 2017), nella forma

$$E(\min(T, \tau) - t \mid T \geq t) = h(\tilde{X}^T \beta), \quad (4.12)$$

per  $0 \leq t \leq \tau$ , dove  $\tilde{X}^T = (1, X^T)$  è un vettore di dimensione  $(p + 1)$  e  $\beta$  il vettore di parametri di dimensione  $(p + 1)$ . Il primo elemento di  $\beta$  rappresenta il coefficiente associato al gruppo base. La funzione legame  $h(\cdot)$  è specificata a priori dal ricercatore e si assume sia non negativa, strettamente crescente e due volte differenziabile.

Per esplicitare il termine dell'intercetta nel predittore lineare, la (4.12) può essere riscritta come

$$E(\min(T, \tau) - t \mid T \geq t) = h(\alpha + X^T \beta). \quad (4.13)$$

La (4.13) definisce una ricca famiglia di modelli grazie alla funzione legame  $h(\alpha, \beta) = h(\alpha + X^T \beta)$ . Tale scelta può essere basata sui dati osservati e sull'interpretazione dei parametri di regressione che si desidera. Ad esempio, per avere un effetto delle covariate simile a quello degli *additive hazard model* e dei *multiplicative hazard model* (paragrafo 4.1), si possono considerare rispettivamente  $h(x, y) = x + y$  e  $h(x, y) = x \exp(y)$ . La prima scelta della funzione legame comporta un modello additivo

$$E(\min(T, \tau) - t \mid T \geq t) = \alpha + X^T \beta, \quad (4.14)$$

dove il singolo parametro  $\beta_k$  indica la variazione della aspettativa di vita residua fino all'istante  $\tau$  per un incremento unitario della covariata  $X_k$ , a parità delle altre variabili esplicative. La seconda scelta della funzione legame conduce a un modello moltiplicativo

$$E(\min(T, \tau) - t \mid T \geq t) = \tilde{\alpha} \exp(X^T \beta), \quad (4.15)$$

dove  $\tilde{\alpha} = \exp(\alpha)$  e il termine  $\exp(\beta_k)$  indica la variazione relativa della aspettativa di vita residua fino all'istante  $\tau$  per un incremento unitario della covariata  $X_k$ , a parità delle altre variabili esplicative.

Si noti che la (4.13) permette di modellare il *restricted mean time* ponendo  $t = 0$ , che corrisponde al *restricted residual mean time* sull'intera durata dello studio.

## 4.2.2 Stima dei parametri

Per costruire l'equazione di stima del modello (4.12), si consideri la tecnica IPCW (paragrafo 2.3), che permette di tener conto della presenza di censura. Per  $t < \tau$ , si definisca

$$M_i = M_i(\beta) = \frac{I(\min(T_i, \tau) \leq C_i)}{G(\min(Y_i, \tau) -)} I(Y_i \geq t) (\min(Y_i, \tau) - t - h(\tilde{X}_i^T \beta)). \quad (4.16)$$

La (4.16) fornisce le basi per la costruzione dell'equazione di stima e si dimostra essere una funzione a media zero. Tale proprietà garantisce la non distorsione dell'equazione di stima proposta.

Sia  $\Delta_i^\tau = I(\min(T_i, \tau) \leq C_i)$  e  $U_i(Y_i) = I(Y_i \geq t)(\min(Y_i, \tau) - t - h(\tilde{X}_i^T \beta))$ . Si consideri

$$E_{T,C}(M_i) = E_T \left( E_C \left( \frac{\Delta_i^\tau U_i(Y_i)}{G(\min(Y_i, \tau) -)} \mid T_i, X_i \right) \mid T_i \geq t, X_i \right). \quad (4.17)$$

Si noti che  $\Delta_i^\tau U_i(Y_i) = \Delta_i^\tau U_i(T_i)$ , dunque è possibile sostituire  $Y_i$  con  $T_i$  per la presenza dell'indicatore  $\Delta_i^\tau$  e si ottiene

$$\begin{aligned} E_{T,C}(M_i) &= E_T \left\{ E_C \left[ \frac{\Delta_i^\tau U_i(T_i)}{G(\min(T_i, \tau) -)} \mid T_i, X_i \right] \mid T_i \geq t, X_i \right\} \\ &= E_T \left\{ E_C[\Delta_i^\tau \mid T_i, X_i] \frac{U_i(T_i)}{G(\min(T_i, \tau) -)} \mid T_i \geq t, X_i \right\} \\ &= E_T \left\{ Pr(C_i \geq \min(T_i, \tau) \mid T_i, X_i) \frac{U_i(T_i)}{G(\min(T_i, \tau) -)} \mid T_i \geq t, X_i \right\} \\ &= E_T \left\{ G(\min(T_i, \tau) -) \frac{U_i(T_i)}{G(\min(T_i, \tau) -)} \mid T_i \geq t, X_i \right\} \\ &= E_T \{ U_i(T_i) \mid T_i \geq t, X_i \} \\ &= E_T \{ I(T_i \geq t)(\min(T_i, \tau) - t - h(\tilde{X}^T \beta)) \mid T_i \geq t, X_i \} \\ &= E_T \{ \min(T_i, \tau) - t - h(\tilde{X}^T \beta) \mid T_i \geq t, X_i \} \\ &= E_T \{ \min(T_i, \tau) - t \mid T_i \geq t, X_i \} - E_T \{ \min(T_i, \tau) - t \mid T_i \geq t, X_i \} \\ &= 0. \end{aligned} \quad (4.18)$$

Le quantità  $M_i$  dell'espressione (4.16) dipendono dai pesi IPCW,  $1/G(s)$ , e quindi da una funzione di probabilità non nota  $G(s)$ . Tale funzione deve essere stimata dai dati in modo da poter risolvere l'equazione di stima IPCW proposta di seguito. Si denoti con  $\widehat{M}_i(\beta)$  la controparte empirica di  $M_i(\beta)$ , dove  $G(s)$  è sostituita con la stima associata  $\widehat{G}(s)$ . La funzione di sopravvivenza per il tempo alla censura viene stimata dai dati osservati basandosi su  $(Y_i, 1 - \delta_i, X_i; i = 1, \dots, n)$ . Nel caso in cui  $C$  è indipendente dalle covariate, è possibile ricorrere allo stimatore di Kaplan-Meier. In caso contrario, in presenza di censura dipendente, per stimare  $G(s)$  è necessario tener conto dell'informazione fornita dalle covariate e dunque

è opportuno ricorrere alla versione stratificata dello stimatore di Kaplan-Meier o a modelli correttamente specificati che stimino la funzione di sopravvivenza per i tempi alla censura, come il modello semiparametrico di Cox. Per ulteriori approfondimenti, si rimanda alla costruzione dei pesi dell'IPCW nel paragrafo 2.3.

Utilizzando la tecnica IPCW (Robins e Rotnitzky 1992), l'equazione di stima pesata che si considera è allora

$$s(\beta) = \sum_{i=1}^n Dh(\tilde{X}_i^T \beta) \tilde{X}_i \widehat{M}_i(\beta) = 0, \quad (4.19)$$

dove  $Dh(\eta_i) = \partial h(\eta_i) / \partial \eta_i$ . Risolvendo l'equazione (4.19) si ottengono le stime dei coefficienti  $\beta$  del modello.

Sia  $\widehat{\beta}$  la radice unica della (4.19), allora per la proprietà di non distorsione dell'equazione di stima segue che lo stimatore  $\widehat{\beta}$  è consistente e  $\sqrt{n}(\widehat{\beta} - \beta)$  è asintoticamente normale con media 0 e matrice di covarianza stimata in modo consistente dallo stimatore *sandwich* (Davison 2003, Liang e Zeger 1986, Pace e Salvan 2001)

$$\widehat{V}(\widehat{\beta}) = I(\widehat{\beta})^{-1} \left[ \sum_{i=1}^n s_i(\widehat{\beta}) s_i^T(\widehat{\beta}) \right] I(\widehat{\beta})^{-1}, \quad (4.20)$$

con

$$s_i(\beta) = Dh(\tilde{X}_i^T \beta) \tilde{X}_i \widehat{M}_i(\beta) \quad (4.21)$$

e

$$I(\beta) = \sum_{i=1}^n \frac{I(\min(T_i, \tau) \leq C_i)}{\widehat{G}(\min(Y_i, \tau)-)} I(Y_i \geq t) Dh(\tilde{X}_i^T \beta)^2 \tilde{X}_i \tilde{X}_i^T. \quad (4.22)$$

### 4.2.3 Equivalenza tra lo stimatore IPCW e lo stimatore di Kaplan-Meier

Nel caso di una singola covariata di tipo qualitativo, lo stimatore IPCW del paragrafo precedente è equivalente a uno stimatore costruito con lo stimatore di Kaplan-Meier. Per semplicità, si consideri il caso in cui vi sia un unico gruppo. Sia  $E(\min(T, \tau) - t \mid T \geq t) = h(t)$ , si dimostra che vale la relazione

$$\widehat{h}(t) = \frac{\int_t^\tau \widehat{S}_{KM}(s) ds}{\widehat{S}_{KM}(t)}. \quad (4.23)$$

Nel paragrafo 2.3 lo stimatore di Kaplan-Meier, indicato con  $\widehat{S}_{KM}(t)$ , è stato espresso come stimatore IPCW

$$\widehat{S}_{KM}(t) = 1 - \frac{1}{n} \sum_{i=1}^n \frac{I(T_i \leq C_i)}{\widehat{G}(Y_i-)} I(Y_i \leq t), \quad (4.24)$$

dove  $\widehat{G}(t)$  è lo stimatore di Kaplan-Meier della funzione di sopravvivenza  $G(t)$  del tempo alla censura. Nella espressione (4.24), grazie alla presenza dell'indicatore di censura  $I(T_i \leq C_i)$ , è possibile sostituire  $Y_i$  con  $T_i$  e per  $t \leq \tau$  lo stimatore è riscrivibile come

$$\begin{aligned} \widehat{S}_{KM}(t) &= 1 - \frac{1}{n} \sum_{i=1}^n \frac{I(T_i \leq C_i)}{\widehat{G}(T_i-)} I(T_i \leq t) \\ &= 1 - \frac{1}{n} \sum_{i=1}^n \frac{I(\min(T_i, \tau) \leq C_i)}{\widehat{G}(\min(T_i, \tau)-)} I(T_i \leq t). \end{aligned} \quad (4.25)$$

Per  $\widehat{G}(\tau) > 0$ , segue che

$$\frac{1}{n} \sum_{i=1}^n \frac{I(\min(T_i, \tau) \leq C_i)}{\widehat{G}(\min(T_i, \tau)-)} = 1. \quad (4.26)$$

Per dimostrare la relazione (4.26), si divida il campione in due gruppi di soggetti, il primo con  $Y_i \leq \tau$  e il secondo con  $Y_i > \tau$ . L'espressione a sinistra della (4.26) ha numeratore  $I(\min(T_i, \tau) \leq C_i)I(Y_i \leq \tau) = I(T_i \leq C_i)I(Y_i \leq \tau)$  per il primo gruppo, e numeratore  $I(\min(T_i, \tau) \leq C_i)I(Y_i > \tau) = I(Y_i > \tau)$  per il secondo gruppo. Allora è possibile scrivere

$$\frac{1}{n} \sum_{i=1}^n \frac{I(\min(T_i, \tau) \leq C_i)}{\widehat{G}(\min(T_i, \tau)-)} = \frac{1}{n} \sum_{i=1}^n \frac{I(T_i \leq C_i)I(Y_i \leq \tau)}{\widehat{G}(\min(T_i, \tau)-)} + \frac{1}{n} \sum_{i=1}^n \frac{I(Y_i > \tau)}{\widehat{G}(\tau)}, \quad (4.27)$$

in cui la prima sommatoria è esprimibile tramite lo stimatore (4.25) all'istante  $\tau$ , ossia

$$\frac{1}{n} \sum_{i=1}^n \frac{I(T_i \leq C_i)I(Y_i \leq \tau)}{\widehat{G}(\min(T_i, \tau)-)} = 1 - \widehat{S}_{KM}(\tau), \quad (4.28)$$

e

$$\frac{1}{n} \sum_{i=1}^n \frac{I(Y_i > \tau)}{\widehat{G}(\tau)} = \widehat{S}_{KM}(\tau), \quad (4.29)$$

dato che  $\sum_{i=1}^n I(Y_i > \tau)$  è il numero di soggetti a rischio di sperimentare l'evento all'istante  $\tau$ , ossia  $R(\tau)$ , e  $\widehat{G}(\tau)\widehat{S}_{KM}(\tau) = R(\tau)/n$  (Satten e Datta 2001). Di conseguenza

$$\frac{1}{n} \sum_{i=1}^n \frac{I(\min(T_i, \tau) \leq C_i)}{\widehat{G}(\min(T_i, \tau)-)} = (1 - \widehat{S}_{KM}(\tau)) + \widehat{S}_{KM}(\tau) = 1. \quad (4.30)$$

Grazie all'equazione (4.30), risulta che

$$\frac{1}{n} \sum_{i=1}^n \frac{I(\min(T_i, \tau) \leq C_i)}{\widehat{G}(\min(T_i, \tau)-)} (I(T_i > t) + I(T_i \leq t)) = 1, \quad (4.31)$$

e da questa si ottiene una rappresentazione alternativa della (4.25):

$$\widehat{S}_{KM}(t) = \frac{1}{n} \sum_{i=1}^n \frac{I(\min(T_i, \tau) \leq C_i)}{\widehat{G}(\min(T_i, \tau)-)} I(T_i > t). \quad (4.32)$$

Integrando su  $[0, \tau]$  la (4.32), si ricava

$$\begin{aligned} \int_0^\tau \widehat{S}_{KM}(s) ds &= \int_0^\tau \frac{1}{n} \sum_{i=1}^n \frac{I(\min(T_i, \tau) \leq C_i)}{\widehat{G}(\min(T_i, \tau)-)} I(T_i > s) ds \\ &= \frac{1}{n} \sum_{i=1}^n \frac{I(\min(T_i, \tau) \leq C_i)}{\widehat{G}(\min(T_i, \tau)-)} \int_0^\tau I(T_i > s) ds \\ &= \frac{1}{n} \sum_{i=1}^n \frac{I(\min(T_i, \tau) \leq C_i)}{\widehat{G}(\min(T_i, \tau)-)} \min(T_i, \tau), \end{aligned} \quad (4.33)$$

dove  $\int_0^\tau I(T_i > s) ds = \min(T_i, \tau)$ , perché

$$\int_0^\tau I(T_i > s) ds = \begin{cases} \tau & \text{se } \tau \leq T_i \\ T_i & \text{se } T_i \leq \tau \end{cases}. \quad (4.34)$$

Lo stimatore (4.33) è esattamente lo stimatore IPCW per  $h(0) = E(\min(T, \tau))$ , che si ricava risolvendo l'equazione di stima del modello quando  $t = 0$ . Poiché  $\int_0^\tau \widehat{S}_{KM}(s) ds$  è uno stimatore consistente di  $E(\min(T, \tau))$  (paragrafo 3.1), anche lo stimatore IPCW è uno stimatore consistente di  $E(\min(T, \tau))$ .

In generale, risolvendo l'equazione di stima del modello per  $h(t)$

$$\sum_{i=1}^n \frac{I(\min(T_i, \tau) \leq C_i)}{\widehat{G}(\min(T_i, \tau)-)} I(T_i > t) (\min(T_i, \tau) - t - h(t)) = 0, \quad (4.35)$$

si ottiene

$$\widehat{h}(t) = \frac{\sum_{i=1}^n \frac{I(\min(T_i, \tau) \leq C_i)}{\widehat{G}(\min(T_i, \tau)-)} (\min(T_i, \tau) - t) I(T_i > t)}{\sum_{i=1}^n \frac{I(\min(T_i, \tau) \leq C_i)}{\widehat{G}(\min(T_i, \tau)-)} I(T_i > t)}. \quad (4.36)$$

Questo è esattamente lo stimatore IPCW per  $E(\min(T, \tau) - t \mid T \geq t)$ .

Infine, si noti che integrando su  $[t, \tau]$  la (4.32) si ricava

$$\begin{aligned} \int_t^\tau \widehat{S}_{KM}(s) ds &= \int_t^\tau \frac{1}{n} \sum_{i=1}^n \frac{I(\min(T_i, \tau) \leq C_i)}{\widehat{G}(\min(T_i, \tau)-)} I(T_i > s) ds \\ &= \frac{1}{n} \sum_{i=1}^n \frac{I(\min(T_i, \tau) \leq C_i)}{\widehat{G}(\min(T_i, \tau)-)} \int_t^\tau I(T_i > s) ds \\ &= \frac{1}{n} \sum_{i=1}^n \frac{I(\min(T_i, \tau) \leq C_i)}{\widehat{G}(\min(T_i, \tau)-)} (\min(T_i, \tau) - t) I(T_i > t), \end{aligned} \quad (4.37)$$

dove  $\int_t^\tau I(T_i > s) ds = (\min(T_i, \tau) - t) I(T_i > t)$ , perché per  $t < \tau$

$$\int_t^\tau I(T_i > s) ds = \begin{cases} \tau - t & \text{se } t < \tau \leq T_i \\ T_i - t & \text{se } t < T_i \leq \tau \\ 0 & \text{se } T_i \leq t < \tau \end{cases}. \quad (4.38)$$

Dunque, la (4.36) equivale a

$$\widehat{h}(t) = \frac{\int_t^\tau \widehat{S}_{KM}(s) ds}{\widehat{S}_{KM}(t)}. \quad (4.39)$$

Si noti che quando l'assunzione di indipendenza tra  $C$  e la covariata qualitativa è rilassata e dunque  $G(s)$  è stimata, per esempio, tramite lo stimatore di Kaplan-Meier in versione stratificata per la censura, la relazione (4.39) è ancora valida per ogni strato della covariata.



# Capitolo 5

## Applicazione ai dati

Utilizzando i dati scaricati dal *Genomic Data Commons* (GDC), si illustra l'applicazione delle tecniche trattate nei capitoli precedenti. Le analisi vengono svolte tramite il software R (ver. 3.6.1, [www.r-project.org](http://www.r-project.org)). I modelli di regressione di Cox e i modelli diretti additivi trattati nella tesi sono implementati utilizzando i pacchetti R `survival` e `timereg`, mentre i grafici delle curve di sopravvivenza sono ottenuti con il pacchetto `survminer`. Il codice R relativo è disponibile in Appendice A.

### 5.1 Presentazione del dataset

I dati sono stati scaricati dal *Genomic Data Commons* (GDC), *Data Portal* del *Cancer Genome Atlas* (TCGA). Questa è una piattaforma del *National Cancer Institute* (NCI) che permette a ricercatori medici e bioinformatici di cercare e scaricare dati sui tumori da analizzare. Il TCGA è un consorzio internazionale che costituisce la più grande raccolta di dati per l'analisi genomica dei tumori. Esso contiene più di 10,000 campioni di pazienti estendendosi a 33 tipi di tumore, mettendo a disposizione dati sul DNA a quelli di espressione genica e di proteine.

Nello studio proposto il tipo di neoplasia considerata è il melanoma. Questo è un tumore, spesso molto aggressivo, che origina nella cute o più raramente negli

occhi o nelle mucose. Si sviluppa dai melanociti, che sono le cellule che producono la melanina, il pigmento che conferisce alla cute la sua colorazione. L'obiettivo dell'analisi è studiare come i livelli di produzione di specifici geni cambiano in pazienti con diverse risposte di sopravvivenza al tumore.

Il dataset originario si compone di 295 unità statistiche e 266 variabili. Tra queste, 26 variabili sono di natura clinica (età del paziente, stadio del tumore, ...) e 240 variabili sono relative alle espressioni geniche. Il dataset è stato pulito considerando le variabili cliniche di maggiore interesse ai fini dell'analisi ed eliminando alcune variabili di poca informazione a causa di un'elevata presenza di valori mancanti. Inoltre, le modalità di alcune variabili qualitative sono state ridefinite e raggruppate per avere una più alta numerosità nelle classi. Per la selezione delle variabili di espressione genica, è di interesse tener conto dei geni coinvolti nel processo di necroptosi (Ripk1, Ripk3, Mlkl, Casp8) e dei geni recettori TAM (Tyro3, Axl, Mertk) coinvolti nel processo di riconoscimento delle cellule apoptotiche. Il dataset finale si compone di 236 unità statistiche e 19 variabili, presentate nei paragrafi successivi. I pazienti considerati vengono osservati dal momento dell'asportazione del tumore tramite intervento chirurgico, e non sono sottoposti ad alcuna terapia adiuvante nel periodo di osservazione.

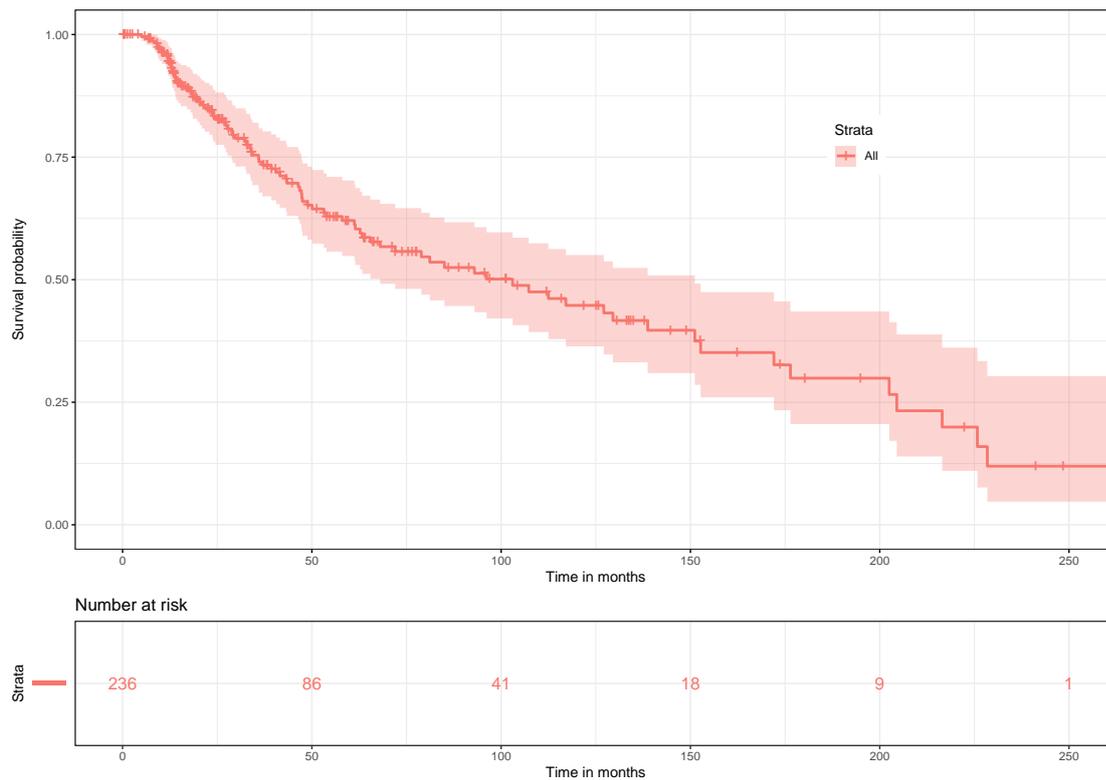
### 5.1.1 *Endpoint* di sopravvivenza

Per ogni paziente, si ha a disposizione di due endpoint da monitorare per la analisi di sopravvivenza: l'*Overall Survival* (OS) e la *Disease-free Survival* (DFS).

L'*Overall Survival* (OS) rappresenta il *gold-standard* tra gli *endpoint* primari ed è definita come l'intervallo di tempo dalla diagnosi alla morte del paziente per qualsiasi causa. La OS è descritta dalla variabile continua OS\_MONTHS che esprime il tempo in mesi di un paziente dalla asportazione del tumore alla morte per qualsiasi causa o all'uscita dallo studio. Di seguito, si riportano i quartili, la media e lo scarto quadratico medio della variabile,

Min	1° Qu.	Mediana	Media	3° Qu.	Max	<i>sd</i>
0.33	14.37	23.28	53.66	69.82	268.53	55.25

dove il più grande tempo di osservazione è 268.53 mesi (22.38 anni). La variabile dicotomica `OS_STATUS` esprime lo stato del paziente alla fine del periodo di osservazione, morto (`=1`) o ancora in vita (`=0`). Il decesso costituisce l'evento di interesse e viene sperimentato da 92 individui (0.39%), mentre i rimanenti 144 (0.61%) sono soggetti a censura a destra. La funzione di sopravvivenza e il relativo intervallo di confidenza al 95% vengono stimati con lo stimatore di Kaplan-Meier (Figura 5.1). Come si osserva dal grafico, le censure e gli eventi si verificano soprattutto per



**Figura 5.1:** Stima della funzione di sopravvivenza OS.

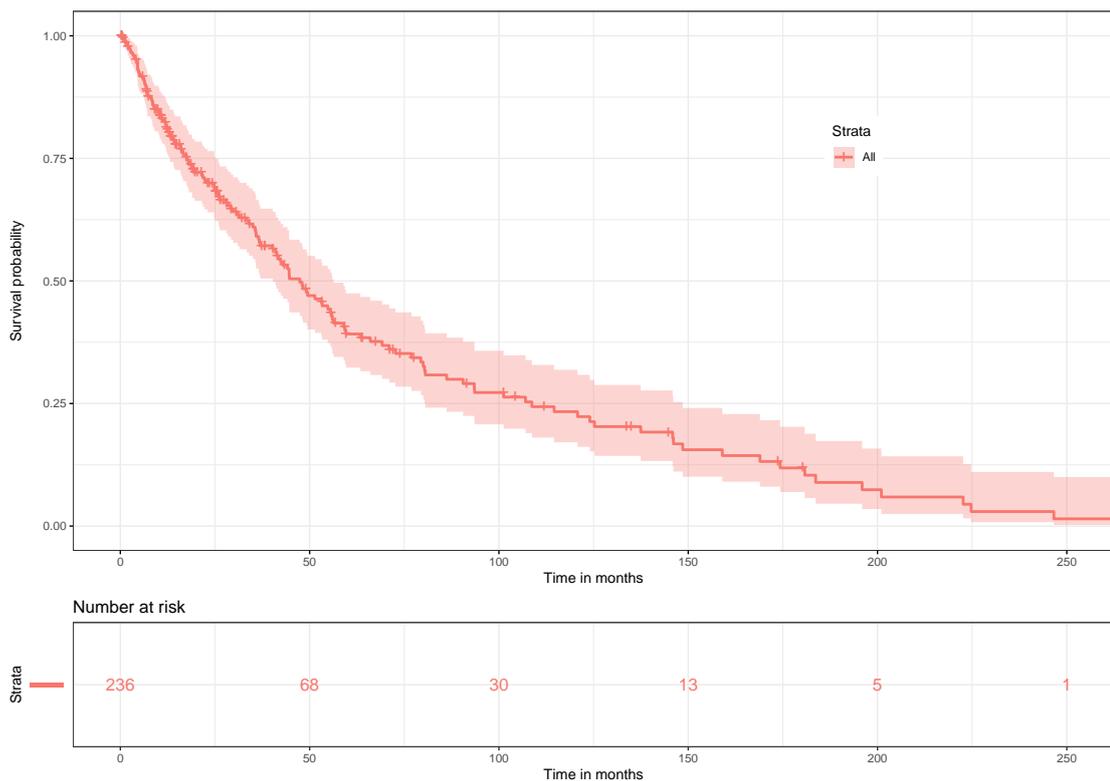
tempi relativamente brevi. Ciò comporta una stima più precisa nel primo periodo, con intervalli di confidenza più ampi nell'ultimo periodo di osservazione. Il tempo di sopravvivenza mediano è di 103.02 (95% CI: 68.04; 151.15) mesi.

La *Disease-free Survival* (DFS) è un *endpoint* surrogato che richiede campioni di numerosità inferiori e tempi di *follow-up* più brevi dell'OS. La DFS è descritta dalla variabile continua `DFS_MONTHS` che esprime il tempo in mesi di un paziente

dalla asportazione del tumore alla recidiva di malattia o all'uscita dallo studio. Di seguito, si riportano i quartili, la media e lo scarto quadratico medio della variabile,

Min	1° Qu.	Mediana	Media	3° Qu.	Max	<i>sd</i>
0.33	11.08	25.56	44.36	56.09	266.95	50.48

dove il più grande tempo di osservazione è 266.95 mesi (22.25 anni). La variabile dicotomica *DFS\_STATUS* esprime lo stato del paziente alla fine del periodo di osservazione, recidivo (=1) o non recidivo (=0). La recidiva di malattia costituisce l'evento di interesse e viene sperimentato da 147 individui (0.62%), mentre i rimanenti 89 (0.38%) sono soggetti a censura a destra. La funzione di sopravvivenza e il relativo intervallo di confidenza al 95% vengono stimati con lo stimatore di Kaplan-Meier (Figura 5.2). Come si osserva anche per l'OS, le censure e gli eventi si verificano



**Figura 5.2:** Stima della funzione di sopravvivenza DFS.

soprattutto per tempi relativamente brevi. Tuttavia, a differenza dell'OS, si ha un maggior numero di eventi che di censure. Il tempo di sopravvivenza mediano è di 47.40 (95% CI: 40.24, 56.34) mesi.

Confrontando l'OS con la DFS, si osserva che tra i pazienti che non hanno avuto recidiva di malattia nel periodo di osservazione, nessuno ha subito l'evento di morte. Tra i pazienti che hanno avuto recidiva di malattia, 92 (0.63%) sono morti e 55 (0.37%) sono risultati ancora vivi alla fine del periodo di osservazione. Nella Tabella 5.1 si riportano gli indici di sintesi della distribuzione della variabile `DFS_MONTHS`, con i relativi boxplot in Figura 5.3, per i pazienti che hanno sperimentato l'evento di recidiva di malattia (gruppo A, 147 osservazioni), i pazienti che non hanno sperimentato l'evento di recidiva (gruppo B, 89 osservazioni), i pazienti che hanno sperimentato l'evento di recidiva e l'evento di morte (gruppo C, 92 osservazioni), i pazienti che hanno sperimentato l'evento di recidiva e non l'evento di morte (gruppo D, 55 osservazioni). All'interno di ciascun gruppo, si nota che la media è sempre maggiore della mediana a causa della forte asimmetria a destra dei tempi. Confrontando i gruppi A e B, i pazienti che hanno sperimentato l'evento di recidiva di malattia hanno tempi maggiori in termini di media e mediana rispetto ai pazienti che non hanno sperimentato la recidiva. Tale risultato inverosimile è dovuto al fatto che gran parte delle censure avvengono in periodi relativamente brevi e queste hanno dunque un ruolo rilevante nella comparazione dei due gruppi.

Gruppo	Min	1° Qu.	Mediana	Media	3° Qu.	Max	<i>sd</i>
A	0.59	10.58	29.83	49.53	59.46	266.95	56.05
B	0.33	12.16	23.09	35.83	43.33	180.26	38.40
C	1.28	11.98	27.12	50.39	59.40	266.95	58.50
D	0.59	7.13	33.38	48.08	61.19	246.58	52.19

**Tabella 5.1:** Quartili, media e scarto quadratico medio di `DFS_MONTHS` per i gruppi A, B, C, D.

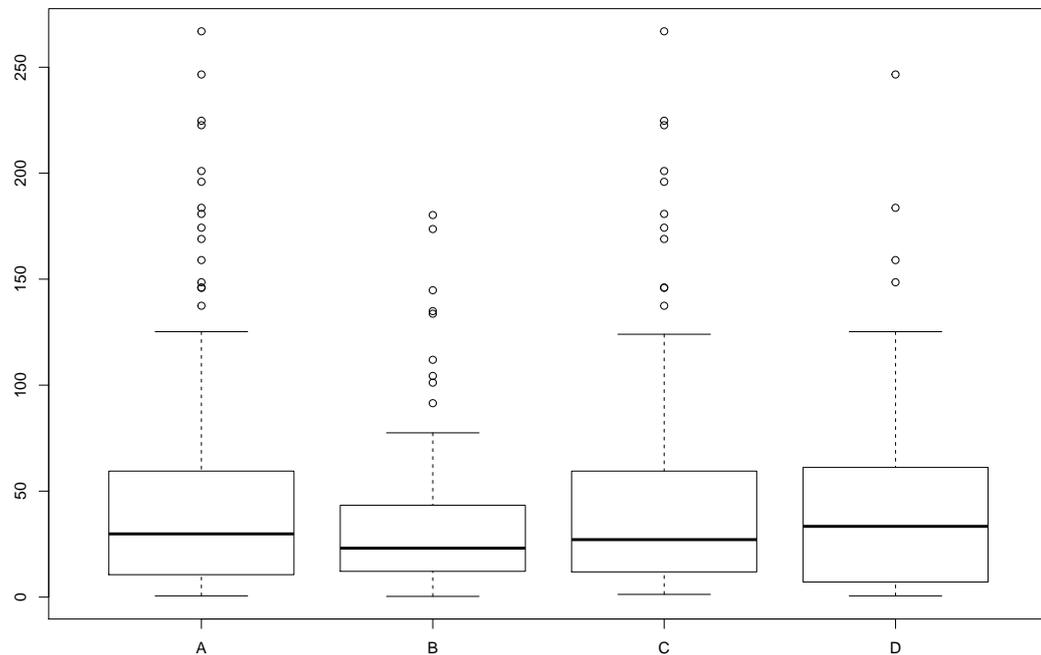


Figura 5.3: Boxplot di DFS\_MONTHS per i gruppi A, B, C, D.

### 5.1.2 Variabili cliniche e di espressione genica

Nella Tabella 5.2 si riportano gli indici di sintesi delle variabili continue considerate nell'analisi.

L'espressione dei geni Ripk1, Ripk3, Mkl, Casp8, Tyro3, Axl, Mertk è descritta rispettivamente dalle variabili continue positive **RIPK1**, **RIPK3**, **MLKL**, **CASP8**, **TYR03**, **AXL**, **MERTK**. Le distribuzioni delle variabili di espressione genica presentano tutte una asimmetria a destra. In media, l'espressione di Tyro3 è la più alta ( $2025.80 \pm 1392.27$ ), mentre l'espressione di Ripk3 è la più bassa ( $86.72 \pm 92.45$ ). Inoltre, per ogni paziente si considera la possibile presenza di mutazione puntiforme del gene Braf identificata dalla variabile dicotomica **BRAF\_MUT**. La mutazione del gene si verifica nel 51.27% dei pazienti. Da letteratura, Braf è il gene più mutato nel tumore del melanoma, in particolare la mutazione V600E è la più dif-

Variabile	Min	1° Qu.	Mediana	Media	3° Qu.	Max	<i>sd</i>
RIPK1	503.90	1018.30	1216.20	1261.80	1440.20	2471.30	372.05
RIPK3	2.86	23.71	58.98	86.72	123.85	578.71	92.45
MLKL	12.14	79.27	151.57	197.77	253.53	1606.56	183.56
CASP8	51.60	293.40	443.90	487.30	620.20	1752.60	267.57
TYR03	167.30	1108.90	1636.30	2025.80	2559.60	9439.90	1392.27
AXL	25.17	259.64	540.37	1240.18	1135.23	15623.81	2148.53
MERTK	6.55	92.17	234.06	380.93	437.87	4361.02	518.36
AGE	18.00	48.00	60.00	58.94	72.00	90.00	15.88
FRACTION_GENOME_ALTERED	0.00	0.16	0.27	0.30	0.40	0.91	0.20
PURITY	0.12	0.64	0.76	0.72	0.84	0.96	0.18
CONSENSUS_SCORE_CD8	2.00	126.90	243.20	232.60	339.00	462.50	125.91

**Tabella 5.2:** Quartili, media e scarto quadratico medio delle variabili continue.

fusa (Flaherty *et al.* 2010). La percentuale di mutazioni del genoma è descritta da `FRACTION_GENOME_ALTERED`, variabile continua con valori compresi tra 0 e 1.

I tumori non sono semplicemente masse di cellule maligne, ma ecosistemi complessi composti da diversi tipi di cellule. In ciascun paziente, l'insieme di cellule del campione di tessuto prelevato include anche tipi di cellule non cancerose come i fibroblasti, le cellule immunitarie, le cellule endoteliali e le cellule epiteliali. Considerato ciò, la variabile continua `PURITY` rappresenta la *tumor purity*, ossia una stima della proporzione di cellule tumorali nell'insieme di cellule del campione di tessuto (Aran *et al.* 2015), e assume valori compresi tra 0 e 1. La variabile continua `CONSENSUS_SCORE_CD8`, invece, quantifica la proporzione di cellule T CD8+ nel campione di tessuto e fornisce una stima dell'infiltrato immunitario nel tumore. La stima si basa su algoritmi bioinformatici di deconvoluzione dei dati di sequenziamento dell'RNA (Finotello e Trajanoski 2018). In particolare, più bassi sono i valori assunti dalla variabile, maggiore è l'infiltrato immunitario.

Per quanto riguarda le informazioni sul melanoma, si tiene in considerazione la variabile dicotomica `TUMOR_SITE` che identifica se il tumore ha avuto origine nel busto o sugli arti (=Trunk/Extremities) o in altre zone del corpo (=Other).

Per la maggior parte dei pazienti (83.05%) il tumore è situato sul busto o sugli arti. Per la stadiazione del melanoma si considera la classificazione TNM (Balch *et al.* 2001), che quantifica l'estensione del tumore primitivo (T), l'estensione di metastasi nei linfonodi regionali (N) e la presenza o assenza di metastasi a distanza (M). Le informazioni ottenute sulla misura dei livelli T, N e M vengono combinate in stadi che forniscono una descrizione precisa dell'estensione anatomica della malattia. Lo stadio del tumore è contenuto nella variabile qualitativa `AJCC_PATHOLOGIC_TUMOR_STAGE`: la maggior parte dei pazienti (43.22%) si colloca negli stadi III o IV (=III/IV), il 36.86% nello stadio II (=II) e il 19.91% negli stadi 0 o I (=0/I).

La variabile dicotomica `SEX` identifica il sesso del paziente, il 63.98% dei pazienti è di sesso maschile. L'età del paziente è descritta dalla variabile continua `AGE`, i pazienti hanno in media 58.94 ( $\pm 15.88$ ) anni.

## 5.2 Effetto marginale delle variabili sui tempi DFS

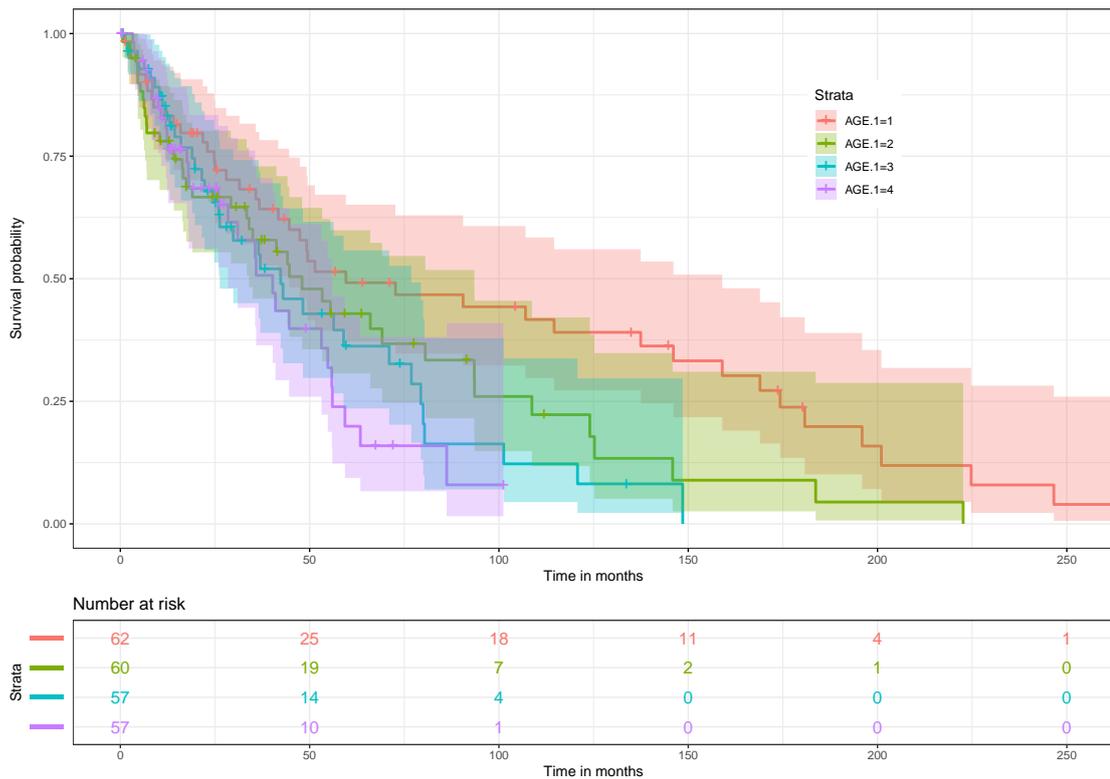
Per ogni variabile, si misura l'effetto marginale sui tempi di sopravvivenza liberi dalla malattia (DFS). L'evento di recidiva presenta una frequenza maggiore dell'evento di morte, pertanto, ai fini della tesi si decide di studiare l'effetto delle variabili sui tempi DFS, che presentano una percentuale minore di censura rispetto all'*endpoint* OS. Per quantificare e testare la significatività dell'effetto, si considerano dei modelli di Cox con una sola covariata e il test dei ranghi logaritmici. Quest'ultimo assegna pesi uguali agli eventi che si verificano nel *follow-up* e perde potenza quando l'*hazard ratio* dei gruppi non è assunto proporzionale. In tale situazione, si ricorre al test Tarone-Ware, al test Peto-Peto e al test Peto-Peto modificato che sono più potenti del test dei ranghi logaritmici (Karadeniz e Ercan 2017) e sono costruiti sulla base di pesi che assegnano maggiore importanza agli eventi che si verificano in tempi brevi. Per verificare l'assunzione di proporzionalità dei rischi dei modelli di Cox, si ricorre ai metodi grafici e verifiche d'ipotesi basati sui residui di Schoenfeld scalati e sui residui cumulati di martingala. I residui di Schoenfeld scalati sono definiti per ciascuna variabile e per ciascun tempo

all'evento (Grambsch e Therneau 1994). Dato  $\beta_k(t) = \beta_k + \theta_k g_k(t)$ , dove  $\beta_k$  è il coefficiente associato alla variabile esplicativa  $k$  e  $g_k(\cdot)$  è una specifica funzione del tempo, l'assunzione di proporzionalità degli azzardi viene testata saggiando l'ipotesi  $H_0 : \theta_k = 0$ , che equivale a verificare che i parametri del modello non siano tempo dipendenti (Moore 2016). Con lo stesso scopo, si riportano anche il test per la proporzionalità di Lin, Wei, Ying (Lin *et al.* 1993), che si basa sui residui cumulati di martingala.

### 5.2.1 Effetto delle variabili cliniche

#### Variabile AGE

La variabile AGE viene raggruppata per quartili nelle classi ordinate 1, 2, 3, 4 e rinominata con AGE.1. Nella Figura 5.4 si riportano le stime di Kaplan-Meier delle curve di sopravvivenza per ciascun gruppo. Il test dei ranghi logaritmici è significativo ( $p - value = 0.01$ ) e dunque la variabile AGE.1 ha un effetto marginale significativo sul tempo di sopravvivenza DFS. In particolare, dalle stime dei parametri del modello di Cox si nota che i pazienti appartenenti alla classe 2 hanno un rischio di recidiva di malattia più alto del 63.12% rispetto agli appartenenti alla classe 1 ( $HR = 1.63$ ; 95% CI: 1.04, 2.56), i pazienti appartenenti alla classe 3 hanno un rischio di recidiva di malattia più alto dell'89.98% rispetto agli appartenenti alla classe 1 ( $HR = 1.90$ ; 95% CI: 1.17, 3.07), i pazienti appartenenti alla classe 4 hanno un rischio di recidiva di malattia 2.17 volte superiore agli appartenenti alla classe 1 ( $HR = 2.17$ ; 95% CI: 1.30, 3.62). Il test di proporzionalità del modello di Cox basato sui residui di Schoenfeld non è significativo per il modello globale con  $g(t) = t$  ( $p - value > 0.1$ ), mentre è significativo al 10% con  $g(t) = \log(t)$  ( $p - value = 0.05$ ) e con i ranghi dei tempi all'evento ( $p - value = 0.06$ ). Tuttavia, il test sui residui cumulati di martingala per ciascuna variabile esplicativa non è significativo ( $p - value > 0.1$ ) e le curve rappresentate in Figura 5.4 non danno una particolare evidenza di violazione dell'assunzione di proporzionalità dei rischi. I test di Tarone-Ware, Peto-Peto e Peto-Peto modificato non presentano risultati significativi ( $p - value > 0.1$ ). Infine, il modello di Cox viene stimato con la variabi-



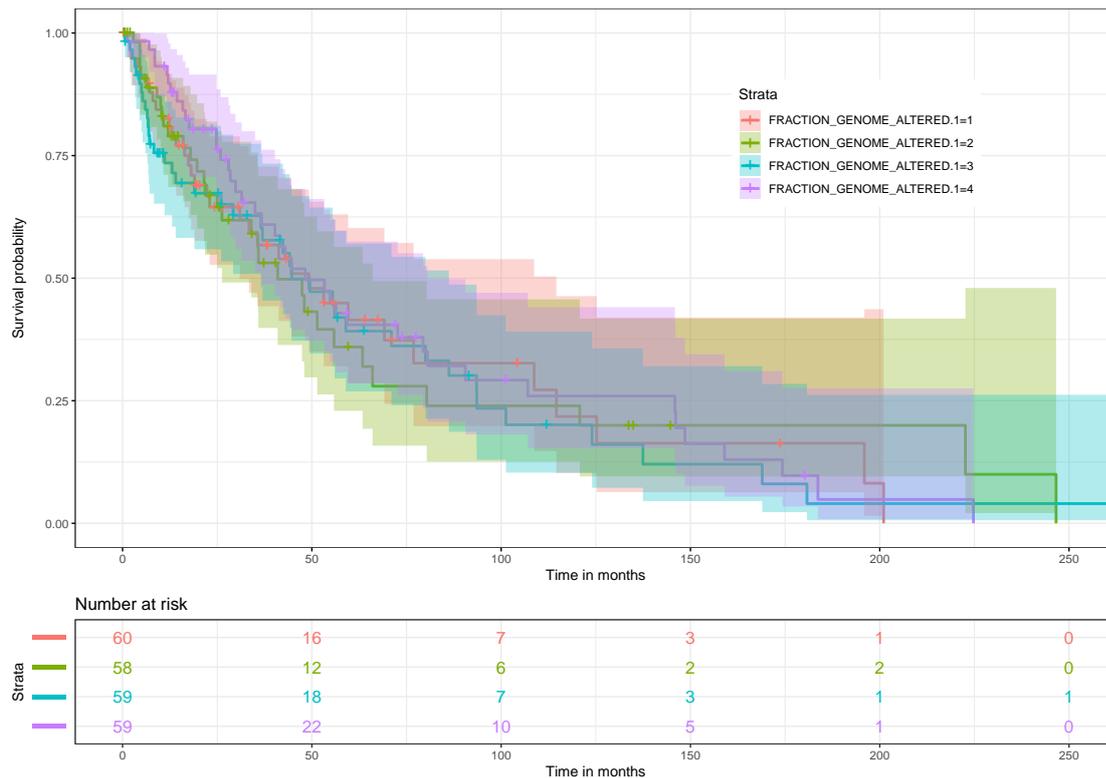
**Figura 5.4:** Stima delle curve di Kaplan-Meier per i gruppi della variabile AGE.1.

le mantenuta continua. Il test alla Wald per il coefficiente associato è significativo ( $p - value < 0.01$ ) e il rischio di recidiva di malattia è del 1.63% più alto per ciascun incremento unitario della variabile AGE ( $HR = 1.02$ ; 95% CI: 1.01, 1.03). Il test di proporzionalità basato sui residui di Schoenfeld è significativo al 10% per  $g(t) = t$  ( $p - value = 0.07$ ) e al 5% per  $g(t) = \log(t)$  ( $p - value < 0.01$ ) e per i ranghi dei tempi all'evento ( $p - value < 0.01$ ), mentre il test basato sui residui cumulati di martingala è significativo al 10% ( $p - value = 0.06$ ). Ciò comporta una possibile violazione dell'assunzione di proporzionalità dei rischi.

Si rileva un effetto significativo della variabile raggruppata e mantenuta continua. Tuttavia, l'assunzione di proporzionalità del sottostante modello di Cox non è sempre soddisfatta e i parametri del modello potrebbero dunque dipendere dal tempo.

Variabile `FRACTION_GENOME_ALTERED`

La variabile `FRACTION_GENOME_ALTERED` viene raggruppata per quartili nelle classi ordinate 1, 2, 3, 4 e rinominata con `FRACTION_GENOME_ALTERED.1`. Nella Figura 5.5 si riportano le stime di Kaplan-Meier delle curve di sopravvivenza per ciascun gruppo. Il test dei ranghi logaritmici non è significativo ( $p\text{-value} > 0.1$ ) e dunque



**Figura 5.5:** Stima delle curve di Kaplan-Meier per i gruppi della variabile `FRACTION_GENOME_ALTERED.1`.

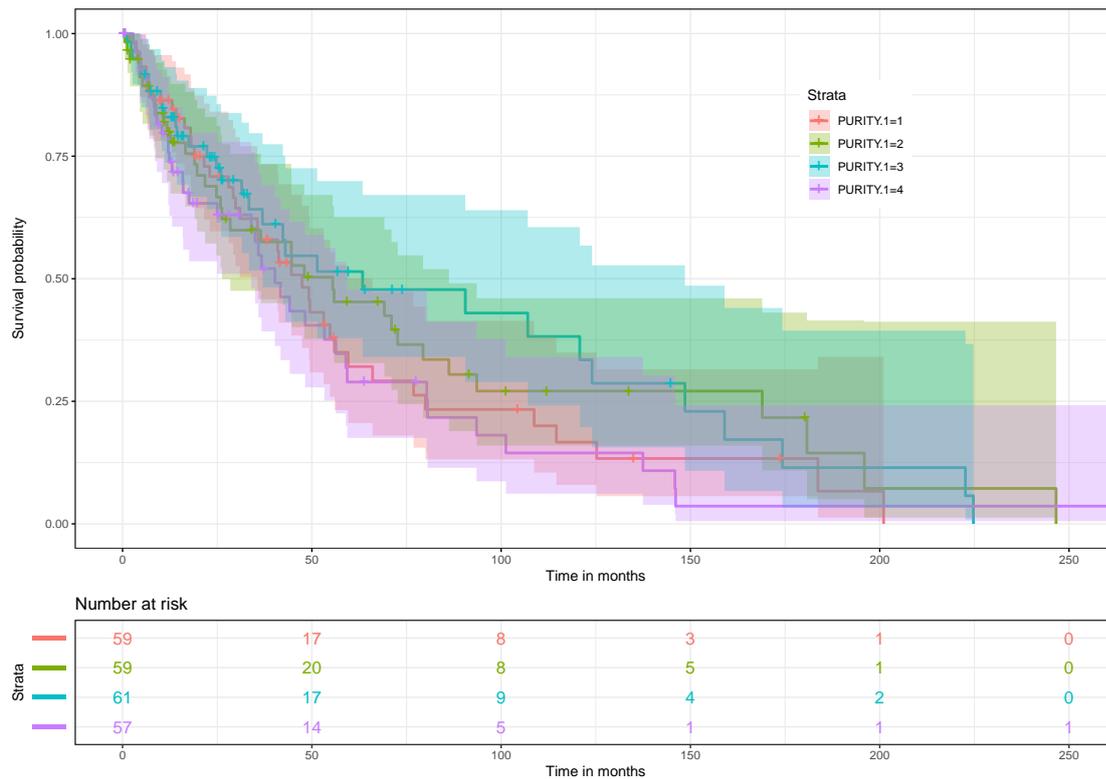
la variabile `FRACTION_GENOME_ALTERED.1` non ha un effetto marginale significativo sul tempo di sopravvivenza DFS. I test alla Wald per la stima dei coefficienti del modello di Cox non sono significativi ( $p\text{-value} > 0.1$ ). I test di proporzionalità del modello di Cox non danno alcuna evidenza di violazione dell'assunzione di proporzionalità dei rischi. Infatti, il test basato sui residui di Schoenfeld non è significativo ( $p\text{-value} > 0.1$ ) per il modello globale con  $g(t) = t$ , con  $g(t) = \log(t)$ , con i ranghi dei tempi all'evento e il test sui residui cumulati di

martingala per ciascuna variabile esplicativa non è significativo ( $p - value > 0.1$ ). I test di Tarone-Ware, Peto-Peto e Peto-Peto modificato non presentano risultati significativi ( $p - value > 0.1$ ). Infine, il modello di Cox viene stimato con la variabile mantenuta continua. Il test alla Wald per il coefficiente associato non è significativo ( $p - value > 0.1$ ). Il test di proporzionalità basato sui residui di Schoenfeld non è significativo ( $p - value > 0.1$ ) per  $g(t) = t$ , per  $g(t) = \log(t)$ , per i ranghi dei tempi all'evento, e il test basato sui residui cumulati di martingala non è significativo ( $p - value > 0.1$ ).

Non si rileva alcun effetto significativo della variabile raggruppata e mantenuta continua, infatti le curve di sopravvivenza riportate in Figura 5.5 presentano andamenti molto vicini e simili.

#### Variabile PURITY

La variabile PURITY viene raggruppata per quartili nelle classi ordinate 1, 2, 3, 4 e rinominata con PURITY.1. Nella Figura 5.6 si riportano le stime di Kaplan-Meier delle curve di sopravvivenza per ciascun gruppo. Il test dei ranghi logaritmici non è significativo ( $p - value > 0.1$ ) e dunque la variabile PURITY.1 non ha un effetto marginale significativo sul tempo di sopravvivenza DFS. I test alla Wald per la stima dei coefficienti del modello di Cox non sono significativi ( $p - value > 0.1$ ). I test di proporzionalità del modello di Cox non danno alcuna evidenza di violazione dell'assunzione di proporzionalità dei rischi. Infatti, il test basato sui residui di Schoenfeld non è significativo ( $p - value > 0.1$ ) per il modello globale con  $g(t) = t$ , con  $g(t) = \log(t)$ , con i ranghi dei tempi all'evento e il test sui residui cumulati di martingala per ciascuna variabile esplicativa non è significativo ( $p - value > 0.1$ ). Infine, il modello di Cox viene stimato con la variabile mantenuta continua. Il test alla Wald per il coefficiente associato non è significativo ( $p - value > 0.1$ ). Il test di proporzionalità basato sui residui di Schoenfeld non è significativo ( $p - value > 0.1$ ) per  $g(t) = t$ , per  $g(t) = \log(t)$ , per i ranghi dei tempi all'evento, e il test basato sui residui cumulati di martingala non è significativo ( $p - value > 0.1$ ).

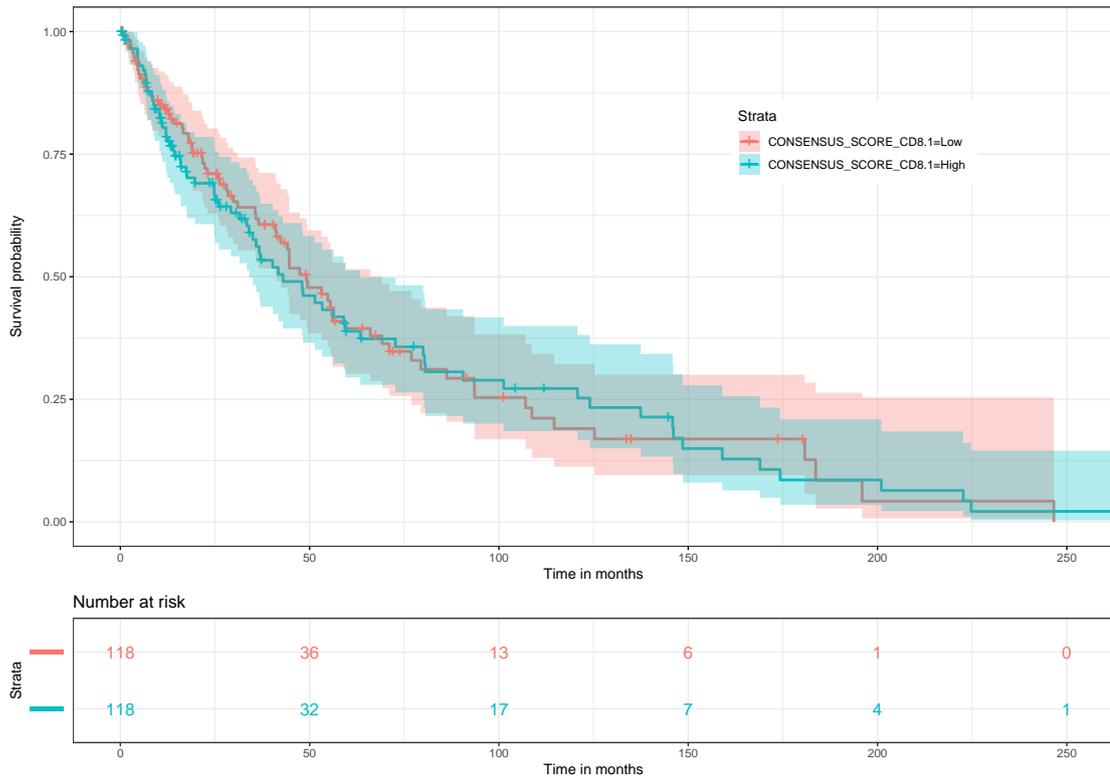


**Figura 5.6:** Stima delle curve di Kaplan-Meier per i gruppi della variabile Purity.1.

Non si rileva alcun effetto significativo della variabile raggruppata e mantenuta continua, infatti le curve di sopravvivenza riportate in Figura 5.6 si incrociano più volte e non si riscontrano particolari differenze.

#### Variabile CONSENSUS\_SCORE\_CD8

La variabile CONSENSUS\_SCORE\_CD8 viene raggruppata con la mediana nelle classi Low, High e rinominata con CONSENSUS\_SCORE\_CD8.1. Nella Figura 5.7 si riportano le stime di Kaplan-Meier delle curve di sopravvivenza per ciascun gruppo. Il test dei ranghi logaritmici non è significativo ( $p - value > 0.1$ ) e dunque la variabile CONSENSUS\_SCORE\_CD8.1 non ha un effetto marginale significativo sul tempo di sopravvivenza DFS. Il test alla Wald per la stima del coefficiente del modello di Cox non è significativo ( $p - value > 0.1$ ). I test di proporzionalità del modello di Cox non danno alcuna evidenza di violazione dell'assunzione di proporziona-



**Figura 5.7:** Stima delle curve di Kaplan-Meier per i gruppi della variabile CONSENSUS\_SCORE\_CD8.1.

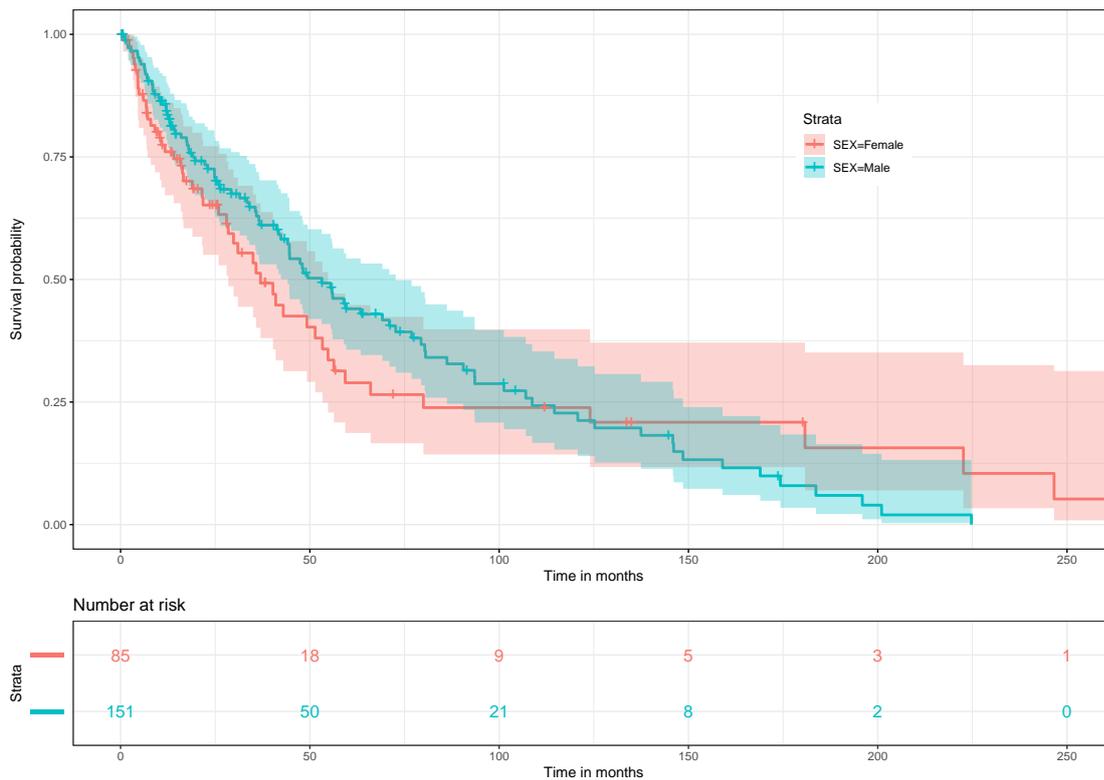
lità dei rischi. Infatti, il test basato sui residui di Schoenfeld non è significativo ( $p - value > 0.1$ ) per  $g(t) = t$ , per  $g(t) = \log(t)$ , per i ranghi dei tempi all'evento e il test sui residui cumulati di martingala per la variabile esplicativa non è significativo ( $p - value > 0.1$ ). I test di Tarone-Ware, Peto-Peto e Peto-Peto modificato non presentano risultati significativi ( $p - value > 0.1$ ). Infine, il modello di Cox viene stimato con la variabile mantenuta continua. Il test alla Wald per il coefficiente associato non è significativo ( $p - value > 0.1$ ). Il test di proporzionalità basato sui residui di Schoenfeld non è significativo ( $p - value > 0.1$ ) per  $g(t) = t$ , per  $g(t) = \log(t)$ , per i ranghi dei tempi all'evento, e il test basato sui residui cumulati di martingala non è significativo ( $p - value > 0.1$ ).

Non si rileva alcun effetto significativo della variabile raggruppata e mantenu-

ta continua, infatti le curve di sopravvivenza riportate in Figura 5.7 presentano andamenti molto vicini e simili.

### Variabile SEX

Nella Figura 5.8 si riportano le stime di Kaplan-Meier delle curve di sopravvivenza per ciascun gruppo definito dalla variabile **SEX**. Il test dei ranghi logaritmici



**Figura 5.8:** Stima delle curve di Kaplan-Meier per i gruppi della variabile **SEX**.

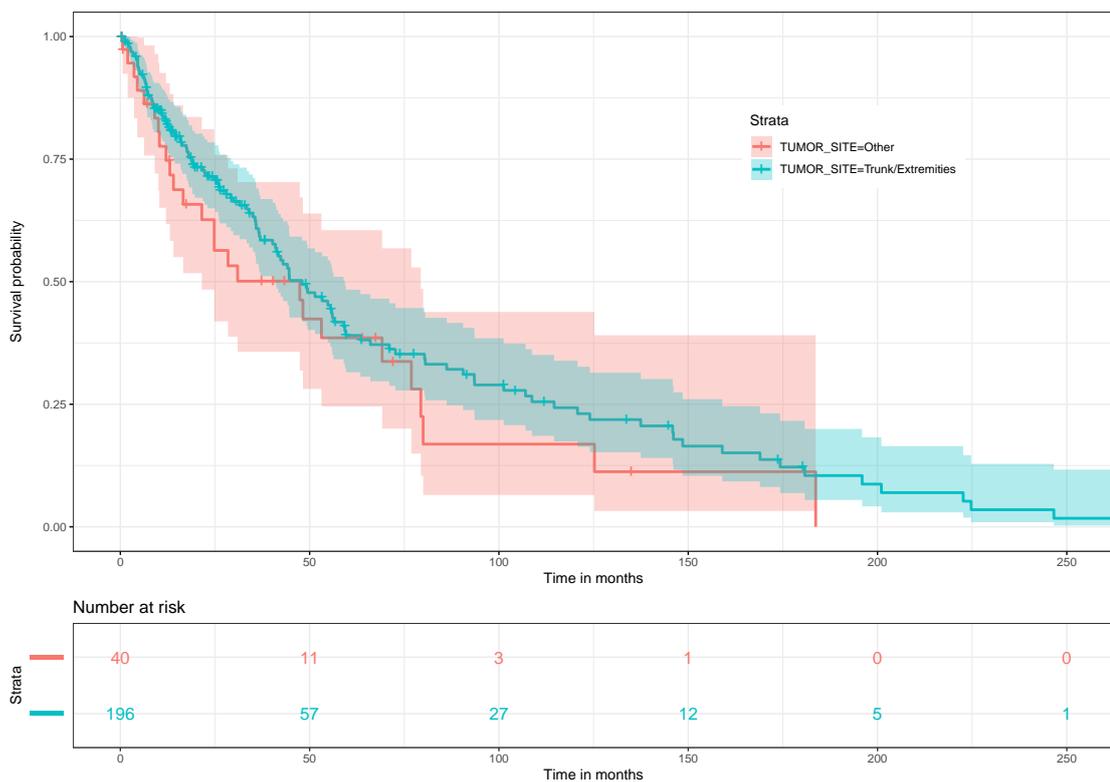
non è significativo ( $p - value > 0.1$ ) e dunque la variabile **SEX** non ha un effetto marginale significativo sul tempo di sopravvivenza DFS. Il test alla Wald per la stima del coefficiente del modello di Cox non è significativo ( $p - value > 0.1$ ). Il test di proporzionalità del modello di Cox basato sui residui di Schoenfeld è significativo per  $g(t) = t$  ( $p - value < 0.01$ ), per  $g(t) = \log(t)$  ( $p - value = 0.02$ ) e per i ranghi dei tempi all'evento ( $p - value = 0.01$ ). Il test sui residui cumulati di

martingala per la variabile esplicativa è significativo ( $p - value = 0.04$ ). Tali risultati e l'incrocio delle curve rappresentate in Figura 5.8 conducono a una possibile violazione dell'assunzione di proporzionalità dei rischi. I test di Tarone-Ware, Peto-Peto e Peto-Peto modificato non presentano risultati significativi ( $p - value > 0.1$ ).

Non si rileva alcun effetto significativo della variabile. Tuttavia, l'assunzione di proporzionalità del sottostante modello di Cox non è soddisfatta e i parametri del modello potrebbero dunque dipendere dal tempo.

### Variabile TUMOR\_SITE

Nella Figura 5.9 si riportano le stime di Kaplan-Meier delle curve di sopravvivenza per ciascun gruppo definito dalla variabile TUMOR\_SITE. Il test dei ranghi



**Figura 5.9:** Stima delle curve di Kaplan-Meier per i gruppi della variabile TUMOR\_SITE.

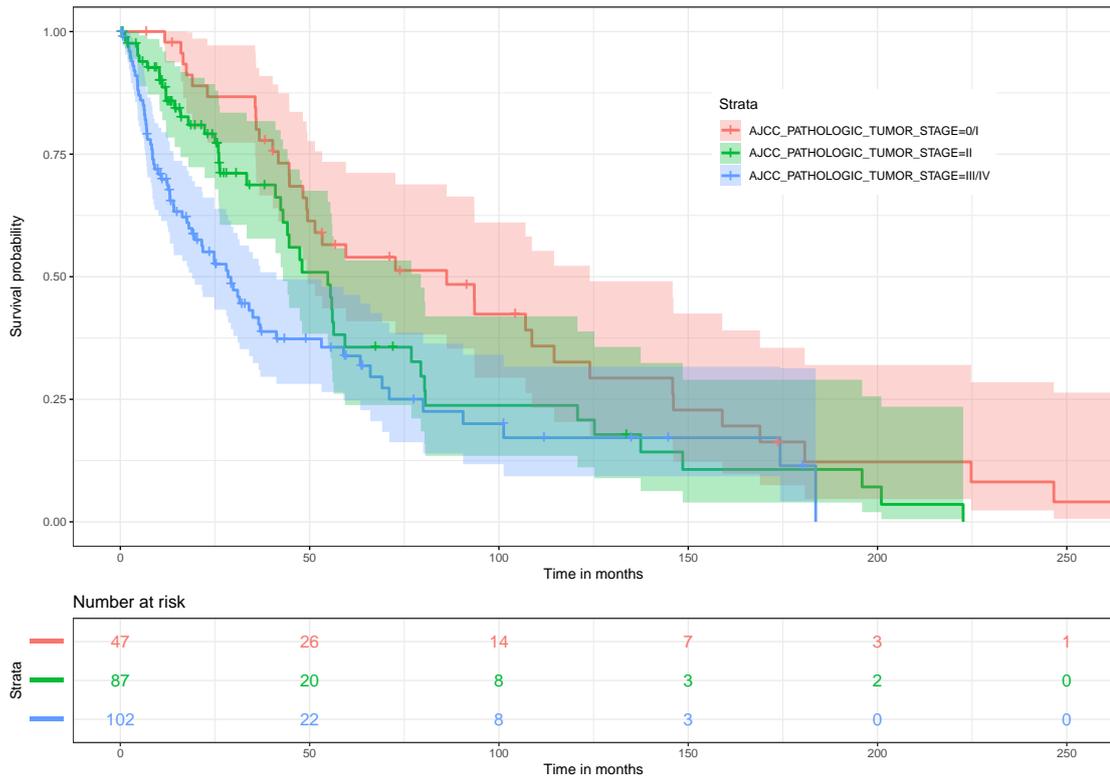
logaritmici non è significativo ( $p - value > 0.1$ ) e la variabile TUMOR\_SITE, dun-

que, non ha un effetto marginale significativo sul tempo di sopravvivenza DFS. Il test alla Wald per la stima del coefficiente del modello di Cox non è significativo ( $p - value > 0.1$ ). I test di proporzionalità del modello di Cox non danno alcuna evidenza di violazione dell'assunzione di proporzionalità dei rischi. Infatti, il test basato sui residui di Schoenfeld non è significativo ( $p - value > 0.1$ ) per  $g(t) = t$ , per  $g(t) = \log(t)$ , per i ranghi dei tempi all'evento e il test sui residui cumulati di martingala per la variabile esplicativa non è significativo ( $p - value > 0.1$ ). I test di Tarone-Ware, Peto-Peto e Peto-Peto modificato non presentano risultati significativi ( $p - value > 0.1$ ).

Non si rileva alcun effetto significativo della variabile, infatti le curve di sopravvivenza riportate in Figura 5.9 presentano andamenti molto vicini e simili.

#### Variabile AJCC\_PATHOLOGIC\_TUMOR\_STAGE

Nella Figura 5.10 si riportano le stime di Kaplan-Meier delle curve di sopravvivenza per ciascun gruppo definito dalla variabile AJCC\_PATHOLOGIC\_TUMOR\_STAGE. Il test dei ranghi logaritmici è significativo ( $p - value < 0.01$ ) e dunque la variabile AJCC\_PATHOLOGIC\_TUMOR\_STAGE ha un effetto marginale significativo sul tempo di sopravvivenza DFS. Il test alla Wald per la stima dei coefficienti del modello di Cox è significativo per il parametro associato alla variabile identificativa del gruppo III/IV ( $p - value < 0.01$ ), mentre non è significativo per il parametro associato alla variabile identificativa del gruppo II ( $p - value > 0.1$ ). In particolare, i pazienti appartenenti alla classe III/IV hanno un rischio di recidiva di malattia 2.22 volte superiore agli appartenenti alla classe 0/I ( $HR = 2.22$ ; 95% CI: 1.44, 3.40). Il test di proporzionalità del modello di Cox basato sui residui di Schoenfeld non è significativo per il modello globale con  $g(t) = t$  ( $p - value > 0.1$ ), mentre è significativo con  $g(t) = \log(t)$  ( $p - value < 0.01$ ) e con i ranghi dei tempi all'evento ( $p - value < 0.01$ ). Il test sui residui cumulati di martingala è significativo al 10% per la variabile identificativa del gruppo II ( $p - value = 0.06$ ) e al 5% per la variabile identificativa del gruppo III/IV ( $p - value < 0.01$ ). Tali risultati e l'andamento delle curve rappresentate in Figura 5.10 conducono a deviazione dell'assunzione di



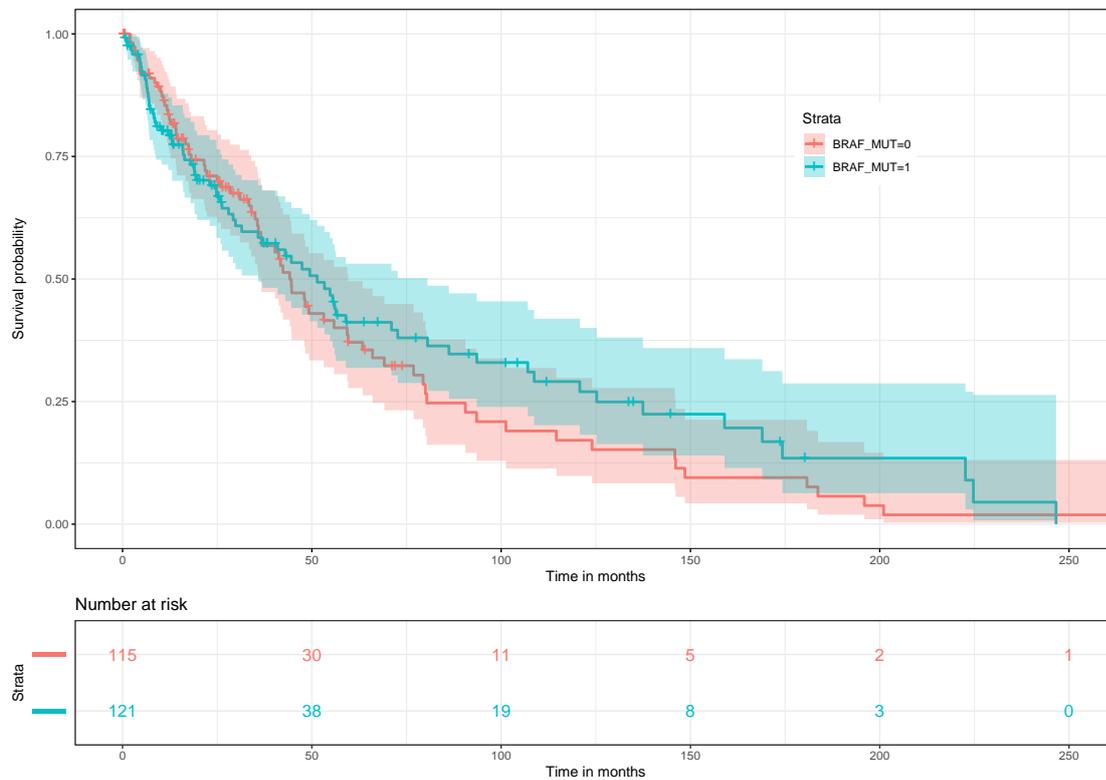
**Figura 5.10:** Stima delle curve di Kaplan-Meier per i gruppi della variabile `AJCC_PATHOLOGIC_TUMOR_STAGE`.

proporzionalità dei rischi. I test di Tarone-Ware, Peto-Peto e Peto-Peto modificato presentano risultati significativi ( $p\text{-value} < 0.1$ ) e si riscontrano quindi differenze significative tra le curve anche pesando maggiormente la parte iniziale del periodo di osservazione.

Si rileva un effetto significativo della variabile. Tuttavia l'assunzione di proporzionalità del sottostante modello di Cox non è sempre soddisfatta e i parametri del modello potrebbero dunque dipendere dal tempo. I test di Tarone-Ware, Peto-Peto e Peto-Peto risultano comunque significativi.

## Variabile BRAF\_MUT

Nella Figura 5.11 si riportano le stime di Kaplan-Meier delle curve di sopravvivenza per ciascun gruppo definito dalla variabile **BRAF\_MUT**. Il test dei ranghi



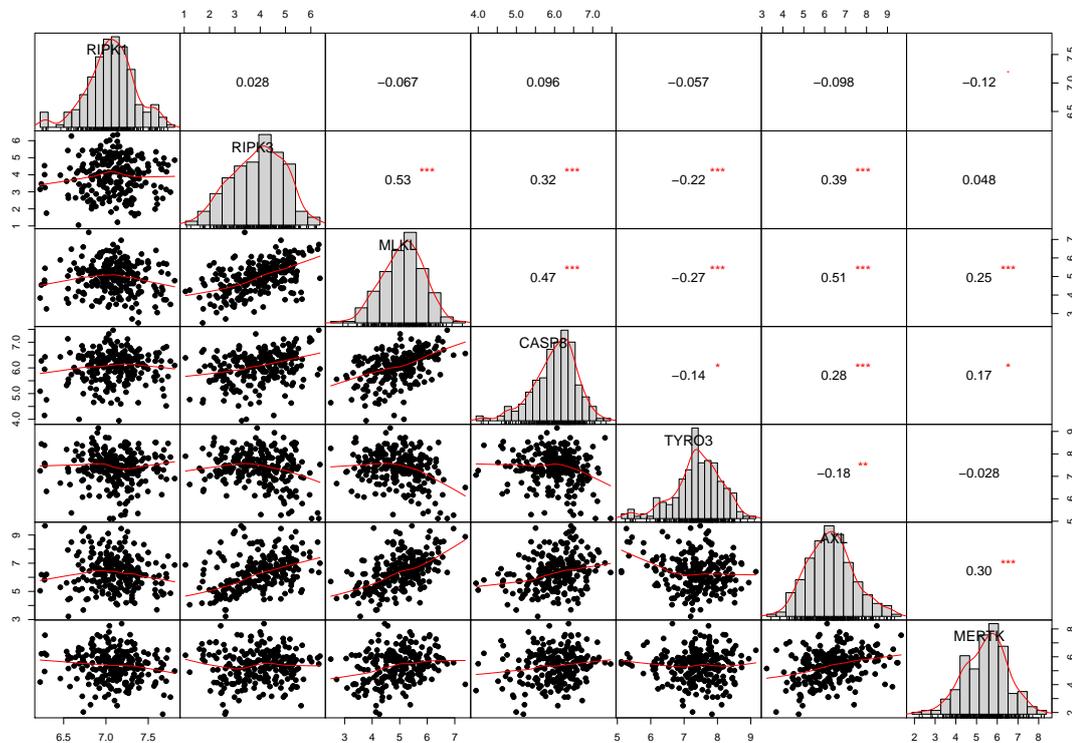
**Figura 5.11:** Stima delle curve di Kaplan-Meier per i gruppi della variabile **BRAF\_MUT**.

logaritmici non è significativo ( $p - value > 0.1$ ) e dunque la variabile **BRAF\_MUT** non ha un effetto marginale significativo sul tempo di sopravvivenza DFS. Il test alla Wald per la stima del coefficiente del modello di Cox non è significativo ( $p - value > 0.1$ ). I test di proporzionalità del modello di Cox non danno alcuna evidenza di violazione dell'assunzione di proporzionalità dei rischi. Infatti, il test basato sui residui di Schoenfeld non è significativo ( $p - value > 0.1$ ) per  $g(t) = t$ , per  $g(t) = \log(t)$ , per i ranghi dei tempi all'evento e il test sui residui cumulati di martingala per la variabile esplicativa non è significativo ( $p - value > 0.1$ ).

Non si rileva alcun effetto significativo della variabile, infatti le curve di sopravvivenza riportate in Figura 5.11 presentano andamenti molto vicini e simili.

## 5.2.2 Effetto delle variabili di espressione genica

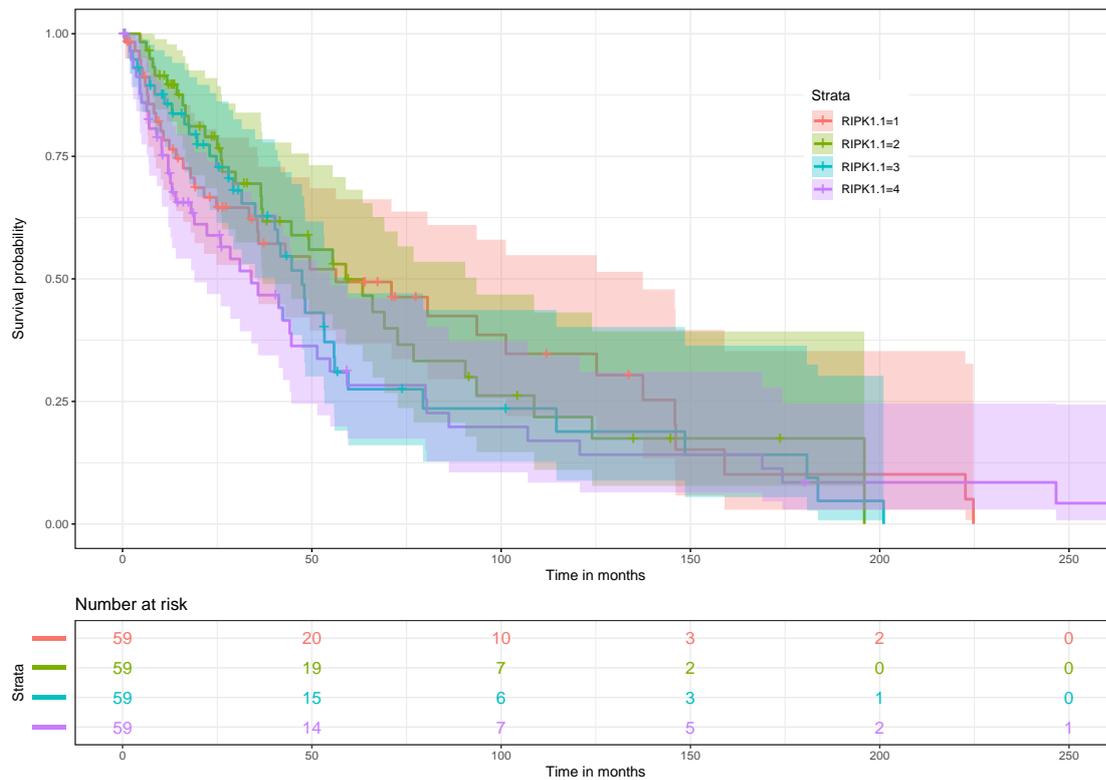
Per correggere la forte asimmetria a destra delle variabili di espressione genica, si decide di considerare la trasformata logaritmica dei valori di espressione. Nella Figura 5.12 si riportano le distribuzioni dei logaritmi di ciascun valore di espressione e gli indici di correlazione di Pearson, con il relativo test  $t$  di Student per valutare la significatività statistica. Dai grafici, la distribuzione delle variabili appare approssimativamente simmetrica e non si notano correlazioni elevate.



**Figura 5.12:** Trasformate logaritmiche della variabili di espressione genica: distribuzione di ciascuna variabile (diagonale), indici di correlazione (parte superiore), grafici di dispersione (parte inferiore).  $p$ -value < .001 ‘\*\*\*’,  $p$ -value < .01 ‘\*\*’,  $p$ -value < .05 ‘\*’.

## Variabile RIPK1

La variabile **RIPK1** viene raggruppata per quartili nelle classi ordinate 1, 2, 3, 4 e rinominata con **RIPK1.1**. Nella Figura 5.13 si riportano le stime di Kaplan-Meier delle curve di sopravvivenza per ciascun gruppo. Il test dei ranghi logaritmici non



**Figura 5.13:** Stima delle curve di Kaplan-Meier per i gruppi della variabile **RIPK1.1**.

è significativo ( $p - value > 0.1$ ) e dunque la variabile **RIPK1.1** non ha un effetto marginale significativo sul tempo di sopravvivenza DFS. I test alla Wald per la stima dei coefficienti del modello di Cox non sono significativi ( $p - value > 0.1$ ). I test di proporzionalità del modello di Cox non danno forti evidenze di violazione dell'assunzione di proporzionalità dei rischi. Infatti, seppur il test basato sui residui di Schoenfeld è significativo al 10% per il modello globale con  $g(t) = \log(t)$  ( $p - value = 0.06$ ) e con i ranghi dei tempi all'evento ( $p - value = 0.05$ ), non è significativo con  $g(t) = t$  ( $p - value > 0.1$ ) e il test sui residui cumulati di martingala per ciascuna variabile esplicativa non è significativo ( $p - value > 0.1$ ).

Infine, il modello di Cox viene stimato con la trasformata logaritmica della variabile continua. Il test alla Wald per il coefficiente associato non è significativo ( $p - value > 0.1$ ). Il test di proporzionalità basato sui residui di Schoenfeld non è significativo ( $p - value > 0.1$ ) per  $g(t) = t$ , per  $g(t) = \log(t)$ , per i ranghi dei tempi all'evento, e il test basato sui residui cumulati di martingala non è significativo ( $p - value > 0.1$ ).

Non si rileva alcun effetto significativo della variabile raggrupata e mantenuta continua, infatti le curve di sopravvivenza riportate in Figura 5.13 si incrociano più volte e non si riscontrano particolari differenze.

### Variabile RIPK3

La variabile RIPK3 viene raggrupata per quartili nelle classi ordinate 1, 2, 3, 4 e rinominata con RIPK3.1. Nella Figura 5.14 si riportano le stime di Kaplan-Meier delle curve di sopravvivenza per ciascun gruppo. Il test dei ranghi logaritmici è significativo ( $p - value = 0.03$ ) e dunque la variabile RIPK3.1 ha un effetto marginale significativo sul tempo di sopravvivenza DFS. Il test alla Wald per la stima dei coefficienti del modello di Cox è significativo per il parametro associato alla variabile identificativa del gruppo 2 ( $p - value < 0.01$ ), mentre non è significativo ( $p - value > 0.1$ ) per i parametri associati alle variabili identificative dei gruppi 3 e 4. In particolare, i pazienti appartenenti alla classe 2 hanno un rischio di recidiva di malattia più basso del 50.50% rispetto agli appartenenti alla classe 1 ( $HR = 0.49$ ; 95% CI: 0.30, 0.80). Il test di proporzionalità del modello di Cox basato sui residui di Schoenfeld non è significativo ( $p - value > 0.1$ ) per il modello globale con  $g(t) = t$ , con  $g(t) = \log(t)$ , con i ranghi dei tempi all'evento. Il test sui residui cumulati di martingala non è significativo ( $p - value > 0.1$ ) per la variabile identificativa del gruppo 2 e 3, ma è significativo per la variabile identificativa del gruppo 4 ( $p - value = 0.01$ ). Tali risultati e l'andamento delle curve rappresentate in Figura 5.14 conducono a una possibile deviazione dell'assunzione di proporzionalità dei rischi. I test di Tarone-Ware, Peto-Peto e Peto-Peto modificato non presentano risultati significativi ( $p - value > 0.1$ ). Infine, il modello di Cox viene

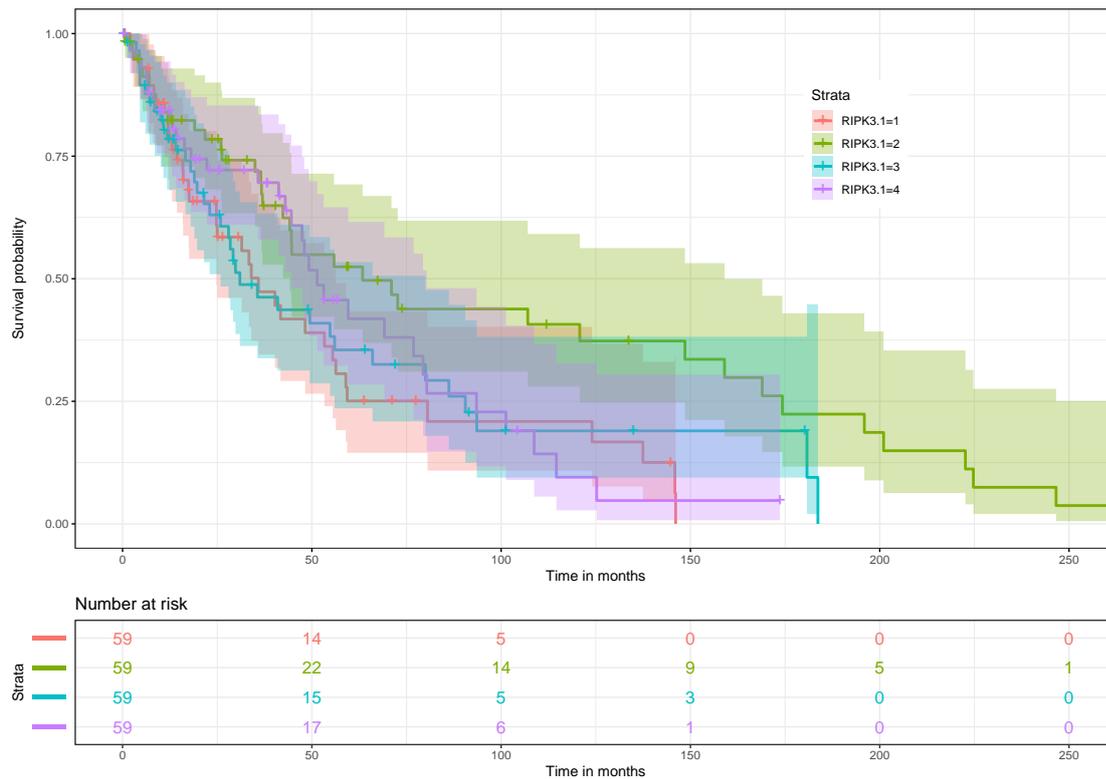


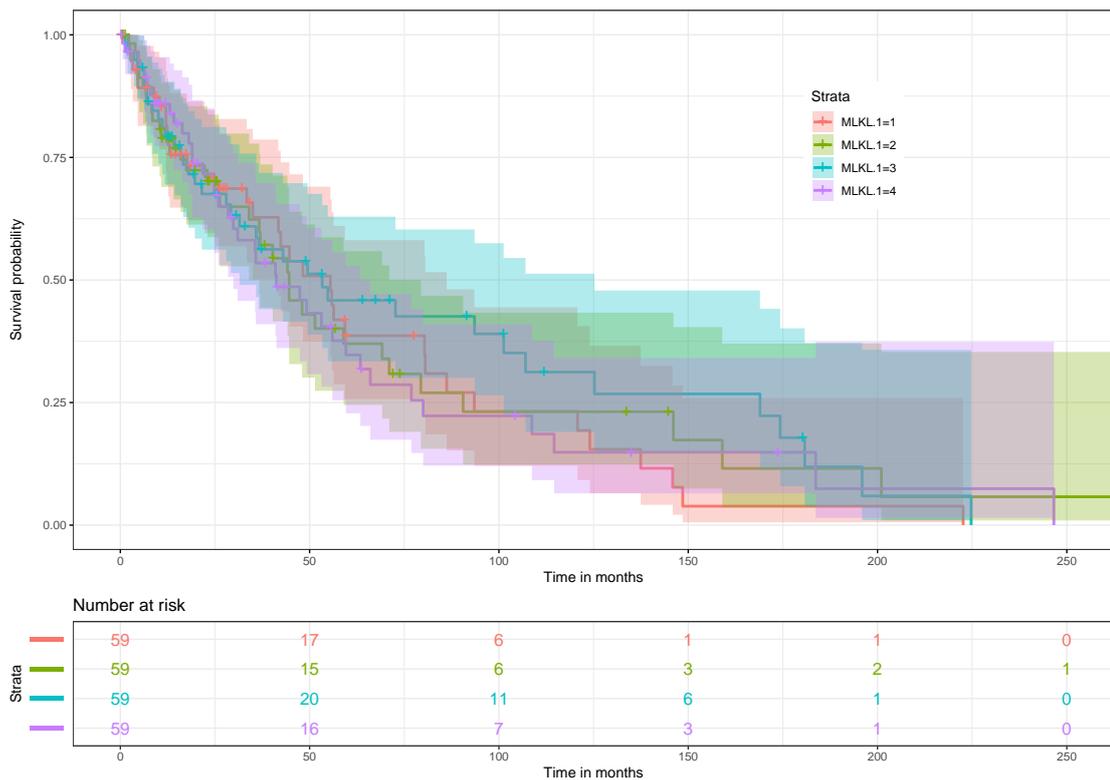
Figura 5.14: Stima delle curve di Kaplan-Meier per i gruppi della variabile RIPK3.1.

stimato con la trasformata logaritmica della variabile continua. Il test alla Wald per il coefficiente associato non è significativo ( $p - value > 0.1$ ). Il test di proporzionalità basato sui residui di Schoenfeld non è significativo ( $p - value > 0.1$ ) per  $g(t) = t$ , per  $g(t) = \log(t)$ , per i ranghi dei tempi all'evento, e il test basato sui residui cumulati di martingala non è significativo ( $p - value > 0.1$ ).

Si rileva un effetto significativo della variabile raggruppata e l'assunzione di proporzionalità del sottostante modello di Cox non è sempre soddisfatta, dunque i parametri del modello potrebbero dipendere dal tempo. L'effetto non è invece significativo per la variabile mantenuta continua.

## Variabile MLKL

La variabile MLKL viene raggruppata per quartili nelle classi ordinate 1, 2, 3, 4 e rinominata con MLKL.1. Nella Figura 5.15 si riportano le stime di Kaplan-Meier delle curve di sopravvivenza per ciascun gruppo. Il test dei ranghi logaritmici non



**Figura 5.15:** Stima delle curve di Kaplan-Meier per i gruppi della variabile MLKL.1.

è significativo ( $p - value > 0.1$ ) e dunque la variabile MLKL.1 non ha un effetto marginale significativo sul tempo di sopravvivenza DFS. I test alla Wald per la stima dei coefficienti del modello di Cox non sono significativi ( $p - value > 0.1$ ). I test di proporzionalità del modello di Cox non danno alcuna evidenza di violazione dell'assunzione di proporzionalità dei rischi. Infatti, il test basato sui residui di Schoenfeld non è significativo ( $p - value > 0.1$ ) per il modello globale con  $g(t) = t$ , con  $g(t) = \log(t)$ , con i ranghi dei tempi all'evento e il test sui residui cumulati di martingala per ciascuna variabile esplicativa non è significativo ( $p - value > 0.1$ ).

Infine, il modello di Cox viene stimato con la trasformata logaritmica della variabile continua. Il test alla Wald per il coefficiente associato non è significativo ( $p - value > 0.1$ ). Il test di proporzionalità basato sui residui di Schoenfeld non è significativo ( $p - value > 0.1$ ) per  $g(t) = t$ , per  $g(t) = \log(t)$ , per i ranghi dei tempi all'evento, e il test basato sui residui cumulati di martingala non è significativo ( $p - value > 0.1$ ).

Non si rileva alcun effetto significativo della variabile raggruppata e mantenuta continua, infatti le curve di sopravvivenza riportate in Figura 5.15 presentano andamenti molto vicini e simili.

### Variabile CASP8

La variabile CASP8 viene raggruppata per quartili nelle classi ordinate 1, 2, 3, 4 e rinominata con CASP8.1. Nella Figura 5.16 si riportano le stime di Kaplan-Meier delle curve di sopravvivenza per ciascun gruppo. Il test dei ranghi logaritmici non è significativo ( $p - value > 0.1$ ) e dunque la variabile CASP8.1 non ha un effetto marginale significativo sul tempo di sopravvivenza DFS. I test alla Wald per la stima dei coefficienti del modello di Cox non sono significativi ( $p - value > 0.1$ ). I test di proporzionalità del modello di Cox non danno alcuna evidenza di violazione dell'assunzione di proporzionalità dei rischi. Infatti, il test basato sui residui di Schoenfeld non è significativo ( $p - value > 0.1$ ) per il modello globale con  $g(t) = t$ , con  $g(t) = \log(t)$ , con i ranghi dei tempi all'evento e il test sui residui cumulati di martingala per ciascuna variabile esplicativa non è significativo ( $p - value > 0.1$ ). I test di Tarone-Ware, Peto-Peto e Peto-Peto modificato non presentano risultati significativi ( $p - value > 0.1$ ). Infine, il modello di Cox viene stimato con la trasformata logaritmica della variabile continua. Il test alla Wald per il coefficiente associato non è significativo ( $p - value > 0.1$ ). Il test di proporzionalità basato sui residui di Schoenfeld non è significativo ( $p - value > 0.1$ ) per  $g(t) = t$ , per  $g(t) = \log(t)$ , per i ranghi dei tempi all'evento, e il test basato sui residui cumulati di martingala non è significativo ( $p - value > 0.1$ ).

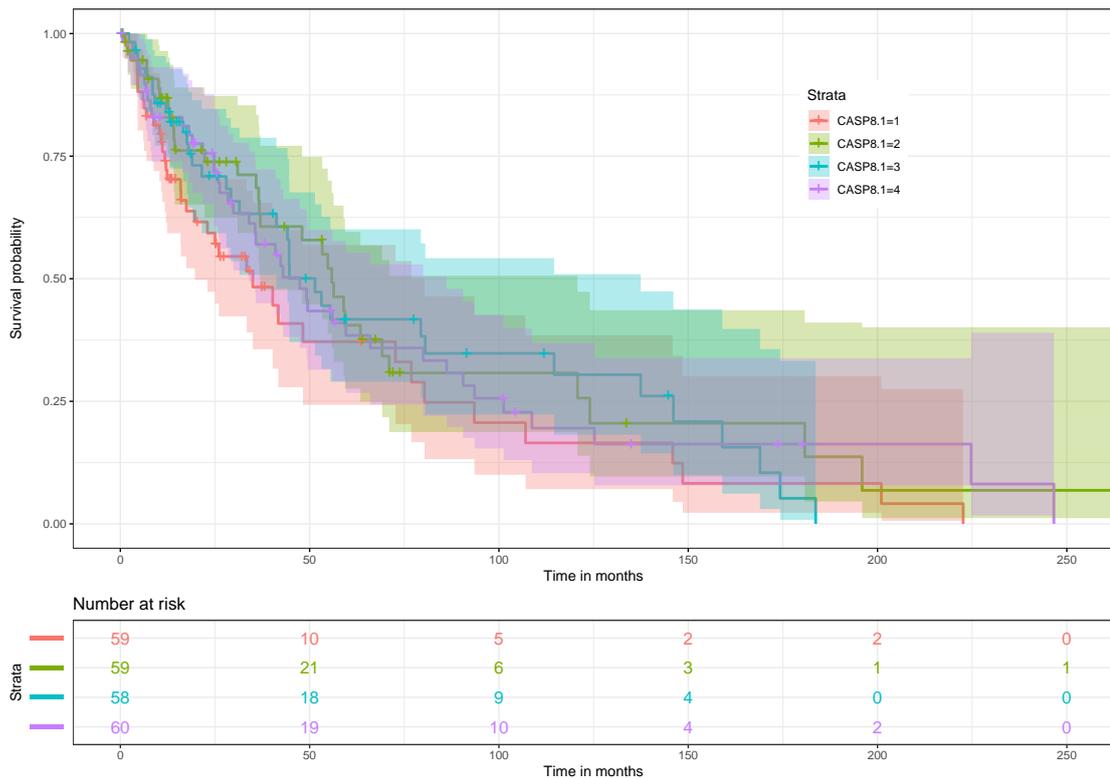


Figura 5.16: Stima delle curve di Kaplan-Meier per i gruppi della variabile CASP8.1.

Non si rileva alcun effetto significativo della variabile raggruppata e mantenuta continua, infatti le curve di sopravvivenza riportate in Figura 5.16 presentano andamenti molto vicini e simili.

### Variabile TYR03

La variabile TYR03 viene raggruppata per quartili nelle classi ordinate 1, 2, 3, 4 e rinominata con TYR03.1. Nella Figura 5.17 si riportano le stime di Kaplan-Meier delle curve di sopravvivenza per ciascun gruppo. Il test dei ranghi logaritmici non è significativo ( $p - value > 0.1$ ) e dunque la variabile TYR03.1 non ha un effetto marginale significativo sul tempo di sopravvivenza DFS. I test alla Wald per la stima dei coefficienti del modello di Cox non sono significativi ( $p - value > 0.1$ ). I test di proporzionalità del modello di Cox non danno alcuna evidenza di violazione dell'assunzione di proporzionalità dei rischi. Infatti, il test basato sui residui di

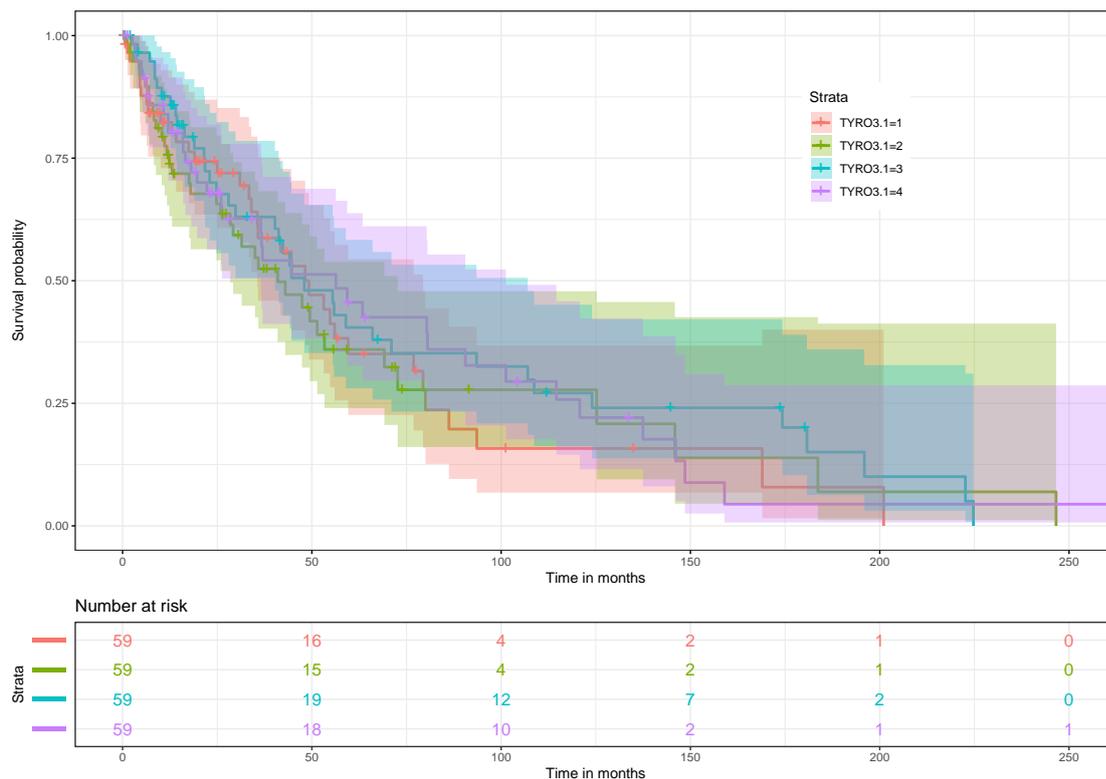


Figura 5.17: Stima delle curve di Kaplan-Meier per i gruppi della variabile TYR03.1.

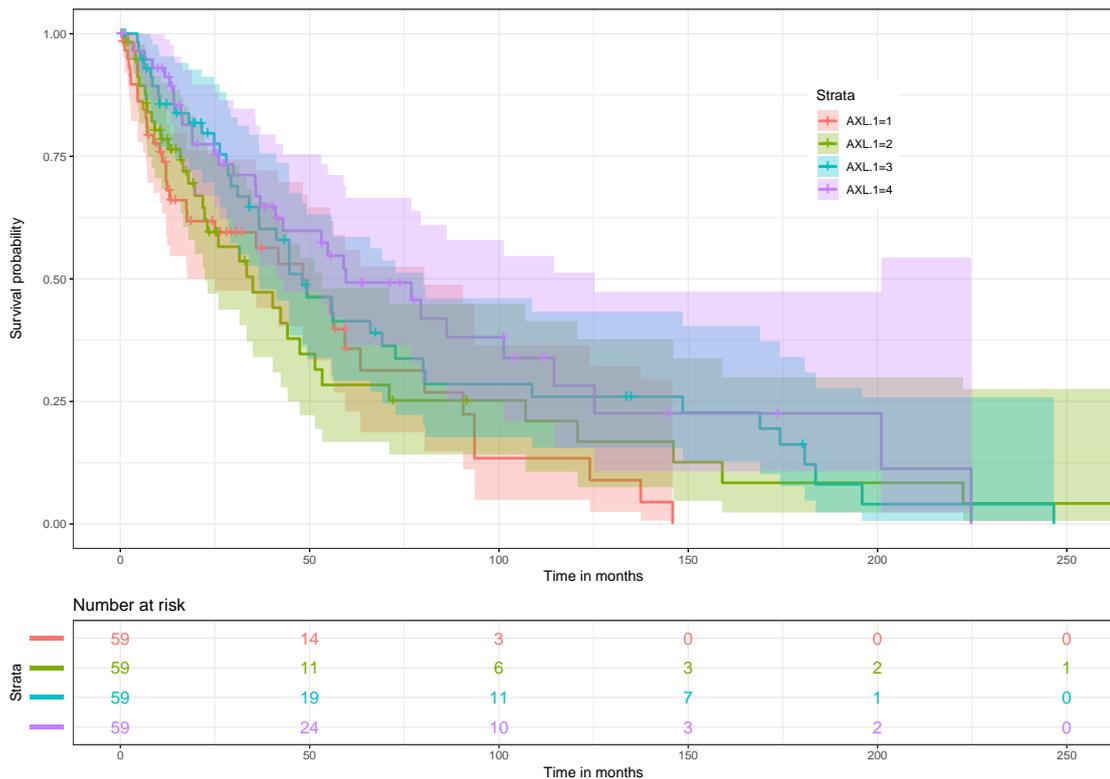
Schoenfeld non è significativo ( $p - value > 0.1$ ) per il modello globale con  $g(t) = t$ , con  $g(t) = \log(t)$ , con i ranghi dei tempi all'evento e il test sui residui cumulati di martingala per ciascuna variabile esplicativa non è significativo ( $p - value > 0.1$ ). I test di Tarone-Ware, Peto-Peto e Peto-Peto modificato non presentano risultati significativi ( $p - value > 0.1$ ). Infine, il modello di Cox viene stimato con la trasformata logaritmica della variabile continua. Il test alla Wald per il coefficiente associato non è significativo ( $p - value > 0.1$ ). Il test di proporzionalità basato sui residui di Schoenfeld non è significativo ( $p - value > 0.1$ ) per  $g(t) = t$ , per  $g(t) = \log(t)$ , per i ranghi dei tempi all'evento, e il test basato sui residui cumulati di martingala non è significativo ( $p - value > 0.1$ ).

Non si rileva alcun effetto significativo della variabile raggruppata e mantenuta continua, infatti le curve di sopravvivenza riportate in Figura 5.17 presentano

andamenti molto vicini e simili.

### Variabile AXL

La variabile AXL viene raggruppata per quartili nelle classi ordinate 1, 2, 3, 4 e rinominata con AXL.1. Nella Figura 5.18 si riportano le stime di Kaplan-Meier delle curve di sopravvivenza per ciascun gruppo. Il test dei ranghi logaritmici è



**Figura 5.18:** Stima delle curve di Kaplan-Meier per i gruppi della variabile AXL.1.

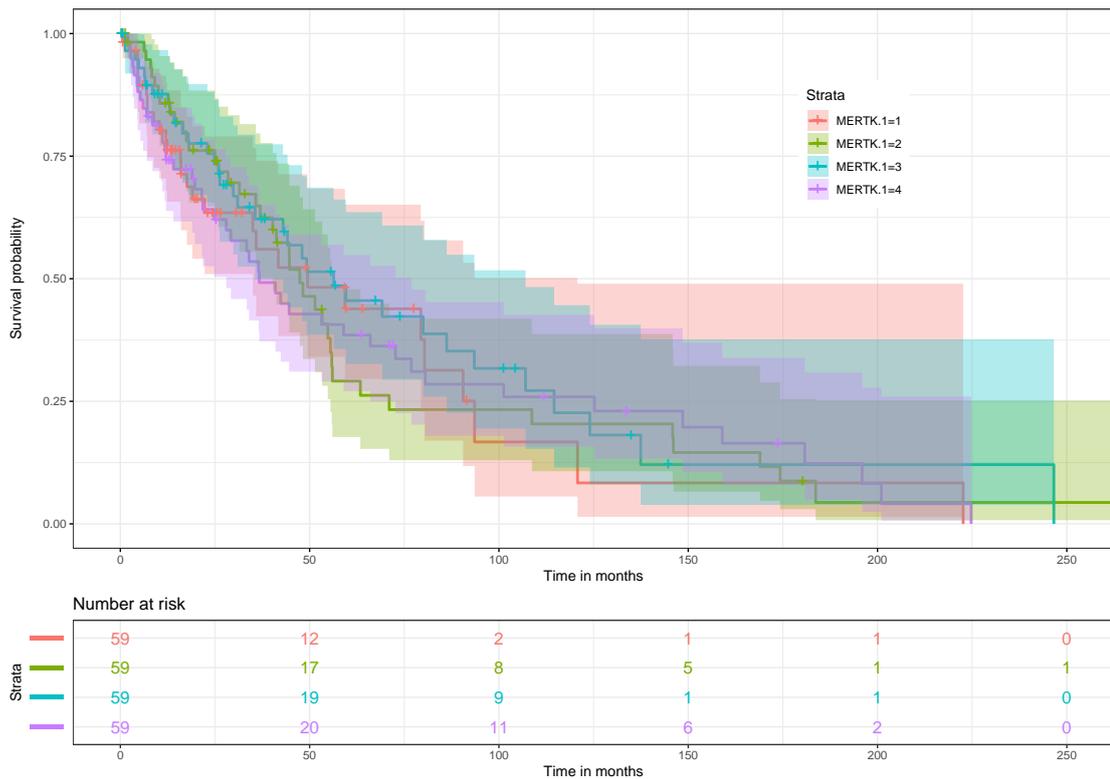
significativo al 10% ( $p - value = 0.06$ ) e dunque la variabile AXL.1 ha un effetto marginale significativo al 10% sul tempo di sopravvivenza DFS. Il test alla Wald per la stima dei coefficienti del modello di Cox è significativo al 5% per il parametro associato alla variabile identificativa del gruppo 4 ( $p - value = 0.01$ ), è significativo al 10% per il parametro associato alle variabile identificativa del gruppo 3 ( $p - value = 0.07$ ) e non è significativo per il parametro associato alle variabile identificativa del gruppo 2 ( $p - value > 0.1$ ). In particolare, i pazienti

appartenenti alla classe 4 hanno un rischio di recidiva di malattia più basso del 44.73% rispetto agli appartenenti alla classe 1 ( $HR = 0.55$ ; 95% CI: 0.34, 0.89). I test di proporzionalità del modello di Cox non danno alcuna evidenza di violazione dell'assunzione di proporzionalità dei rischi. Infatti, il test basato sui residui di Schoenfeld non è significativo ( $p - value > 0.1$ ) per il modello globale con  $g(t) = t$ , con  $g(t) = \log(t)$ , con i ranghi dei tempi all'evento e il test sui residui cumulati di martingala per ciascuna variabile esplicativa non è significativo ( $p - value > 0.1$ ). I test di Tarone-Ware, Peto-Peto e Peto-Peto modificato presentano risultati significativi ( $p - value < 0.05$ ) e si riscontrano quindi differenze significative tra le curve anche pesando maggiormente la parte iniziale del periodo di osservazione. Infine, il modello di Cox viene stimato con la trasformata logaritmica della variabile continua. Il test alla Wald per il coefficiente associato è significativo ( $p - value < 0.01$ ) e il rischio di recidiva di malattia è del 16.20% più basso per ciascun incremento unitario della variabile  $\log(\text{AXL})$  ( $HR = 0.84$ ; 95% CI: 0.73, 0.96). Il test di proporzionalità basato sui residui di Schoenfeld non è significativo ( $p - value > 0.1$ ) per  $g(t) = t$ , per  $g(t) = \log(t)$ , per i ranghi dei tempi all'evento, e il test basato sui residui cumulati di martingala non è significativo ( $p - value > 0.1$ ).

Si rileva un effetto significativo della variabile raggruppata e mantenuta continua, non si riscontrano violazioni dell'assunzione di proporzionalità del modello di Cox sottostante.

### Variabile MERTK

La variabile MERTK viene raggruppata per quartili nelle classi ordinate 1, 2, 3, 4 e rinominata con MERTK.1. Nella Figura 5.19 si riportano le stime di Kaplan-Meier delle curve di sopravvivenza per ciascun gruppo. Il test dei ranghi logaritmici non è significativo ( $p - value > 0.1$ ) e dunque la variabile MERTK.1 non ha un effetto marginale significativo sul tempo di sopravvivenza DFS. I test alla Wald per la stima dei coefficienti del modello di Cox non sono significativi ( $p - value > 0.1$ ). I test di proporzionalità del modello di Cox non danno alcuna evidenza di violazione dell'assunzione di proporzionalità dei rischi. Infatti, il test basato sui residui di



**Figura 5.19:** Stima delle curve di Kaplan-Meier per i gruppi della variabile MERTK.1.

Schoenfeld non è significativo ( $p - value > 0.1$ ) per il modello globale con  $g(t) = t$ , con  $g(t) = \log(t)$ , con i ranghi dei tempi all'evento e il test sui residui cumulati di martingala per ciascuna variabile esplicativa non è significativo ( $p - value > 0.1$ ). Infine, il modello di Cox viene stimato con la trasformata logaritmica della variabile continua. Il test alla Wald per il coefficiente associato non è significativo ( $p - value > 0.1$ ). Il test di proporzionalità basato sui residui di Schoenfeld non è significativo ( $p - value > 0.1$ ) per  $g(t) = t$ , per  $g(t) = \log(t)$ , per i ranghi dei tempi all'evento, e il test basato sui residui cumulati di martingala non è significativo ( $p - value > 0.1$ ).

Non si rileva alcun effetto significativo della variabile raggruppata e mantenuta continua, infatti le curve di sopravvivenza riportate in Figura 5.19 si incrociano più volte e non si riscontrano particolari differenze.

### 5.2.3 Alcune considerazioni

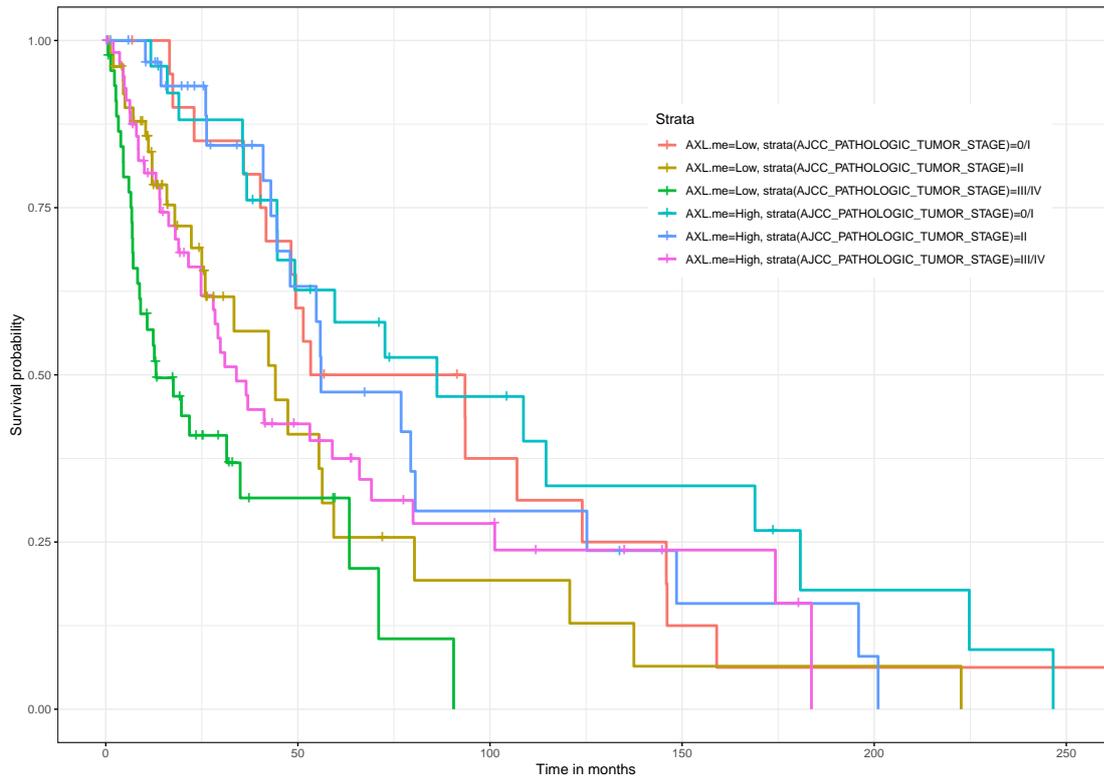
La stime delle curve di Kaplan-Meier riportate nei paragrafi 5.2.1, 5.2.2 sono meno precise e presentano intervalli di confidenza molto ampi nei periodi di osservazione finale. Tale risultato è dovuto al tasso di eventi osservati, che è alto per brevi periodi di osservazione e si abbassa drasticamente per lunghi periodi. Per tale motivo, si conduce un'analisi con censura amministrativa, in cui gli effetti delle variabili sul tempo DFS e i modelli di Cox stimati nei paragrafi precedenti vengono riproposti escludendo l'ultima parte del periodo di osservazione arrestando lo studio a 200 mesi. Come nei paragrafi citati, il test dei ranghi logaritmici è significativo solo per le variabili **AGE.1** ( $p - value = 0.07$ ), **AGE** ( $p - value = 0,02$ ), **AJCC\_PATHOLOGIC\_TUMOR\_STAGE** ( $p - value = 0.02$ ), **AXL.1** ( $p - value = 0.02$ ),  $\log(\mathbf{AXL})$  ( $p - value = 0.03$ ) e la quantificazione dell'effetto tramite la stima dei coefficienti del modello di Cox è concorde con le analisi precedenti. Arrestando anticipatamente lo studio, il test dei ranghi logaritmici non è più significativo per la variabile **RIPK3.1** ( $p - value > 0.1$ ), mentre lo è al 10% per la variabile **CASP8.1** ( $p - value = 0.09$ ). Per quest'ultima, il test alla Wald per la stima dei coefficienti del modello di Cox è significativo per il parametro associato alla variabile identificativa del gruppo 2 ( $p - value = 0.04$ ) e 3 ( $p - value = 0.03$ ), mentre non è significativo per il parametro associato alle variabile identificativa del gruppo 4 ( $p - value > 0.1$ ). In particolare, i pazienti appartenenti alla classe 2 hanno un rischio di recidiva di malattia più basso del 40.47% rispetto agli appartenenti alla classe 1 ( $HR = 0.59$ ; 95% CI: 0.37, 0.97), i pazienti appartenenti alla classe 3 hanno un rischio di recidiva di malattia più basso del 41.43% rispetto agli appartenenti alla classe 1 ( $HR = 0.59$ ; 95% CI: 0.36, 0.94).

Vista l'influenza non trascurabile della variabile **AJCC\_PATHOLOGIC\_TUMOR\_STAGE**, si decide di studiare l'effetto delle variabili di espressione genica sui tempi DFS per ogni sottogruppo della variabile **AJCC\_PATHOLOGIC\_TUMOR\_STAGE**. A tal fine, le variabili di espressione genica **RIPK1**, **RIPK3**, **MLKL**, **CASP8**, **TYR03**, **AXL**, **MERTK** vengono raggruppate con la mediana nelle classi **Low**, **High** e rinominate con **RIPK1.me**, **RIPK3.me**, **MLKL.me**, **CASP8.me**, **TYR03.me**, **AXL.me**, **MERTK.me**. Per ciascun gene, si confrontano le curve di sopravvivenza per i sottogruppi formati dalle due classi definite per la

variabile di espressione genica e le tre classi di `AJCC_PATHOLOGIC_TUMOR_STAGE`. I test dei ranghi logaritmici e i test alternativi di Tarone-Ware, Peto-Peto e Peto-Peto modificato sono sempre significativi ( $p - value < 0.01$ ). In aggiunta, per comparare le curve di sopravvivenza delle due classi `Low` e `High` delle variabili di espressione genica aggiustando per `AJCC_PATHOLOGIC_TUMOR_STAGE`, si propongono test dei ranghi logaritmici stratificati per la variabile `AJCC_PATHOLOGIC_TUMOR_STAGE` (Feng *et al.* 2010). Il test saggia l'ipotesi  $H_0 : S_{0k}(t) = S_{1k}(t) = \dots = S_{pk}(t)$  per  $k = 1, \dots, M$ , dove  $M$  sono le modalità della variabile di stratificazione (`AJCC_PATHOLOGIC_TUMOR_STAGE`) e  $p$  le modalità della variabile di cui si vuole testare l'effetto (variabile di espressione genica raggruppata). Tale aggiustamento comporta che l'effetto della variabile di espressione genica raggruppata sia simile tra gli strati della variabile `AJCC_PATHOLOGIC_TUMOR_STAGE`. Il test risulta significativo solamente per la variabile `AXL.me` ( $p - value < 0.01$ ), di cui si riportano le relative curve stimate in Figura 5.20. Per quantificare l'effetto, si stima il relativo modello di Cox stratificato avente come covariata la variabile `AXL.me`. Per ogni strato della variabile `AJCC_PATHOLOGIC_TUMOR_STAGE`, i pazienti che presentano valori di `AXL.me` nella fascia `High` hanno un rischio di recidiva di malattia più basso del 38.71% rispetto a quelli nella fascia `Low` ( $HR = 0.61$ ; 95% CI: 0.44, 0.86). Infine, per ogni variabile di espressione genica viene valutata la possibile presenza di interazione tra la variabile di espressione genica e `AJCC_PATHOLOGIC_TUMOR_STAGE`. A tal fine, per ogni gene vengono stimati dei modelli di Cox aventi come covariate l'espressione del gene raggruppata e `AJCC_PATHOLOGIC_TUMOR_STAGE` (effetti principali), e la variabile di interazione. Il test alla Wald per la stima dei coefficienti relativi all'interazione non è significativo per alcun modello ( $p - value > 0.05$ ).

In conclusione, dalle analisi svolte emerge la potenziale importanza delle variabili di espressione dei geni `Axl`, `Ripk3`, `Casp8` sul tasso di recidiva di malattia, sulle quali si concentrerà particolare attenzione nelle analisi successive.

Le tecniche trattate in questo paragrafo e nei paragrafi 5.2.1, 5.2.2 danno un'idea dell'effetto marginale delle variabili sul tasso di recidiva. Tuttavia, i modelli di Cox utilizzati non è detto siano adeguati per testare gli effetti considerati. Seppur i test di proporzionalità dei rischi dei modelli spesso non presentano significati-

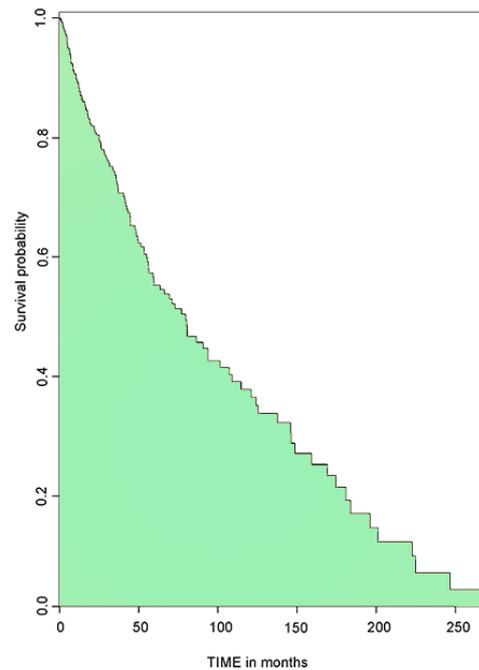


**Figura 5.20:** Stima delle curve di Kaplan-Meier per i sottogruppi formati dalle variabili `AXL.me` e `AJCC_PATHOLOGIC_TUMOR_STAGE`.

ve deviazioni dall'assunzione di proporzionalità, la non significatività dei test non implica che il modello a rischi proporzionali sia corretto. Le curve di Kaplan-Meier stimate, infatti, spesso si incrociano tra di loro conducendo a una possibile violazione di tale ipotesi. Inoltre, non è detto che la forma funzionale delle variabili utilizzate come continue sia corretta e altri approcci potrebbero condurre a soluzioni migliori.

### 5.3 Stima univariata del *Residual Mean Time*

Nei paragrafi 5.2.1 e 5.2.2 sono stati stimati dei modelli di Cox con una sola covariata, per analizzare l'influenza delle variabili sui tempi DFS. L'effetto è risultato significativo per le variabili `AGE.1`, `AGE`, `AJCC_PATHOLOGIC_TUMOR_STAGE`,



**Figura 5.21:** Stima della curva di sopravvivenza fino a  $\tau = 266.95$  mesi per la classe 1 di AGE.1, basata sul modello di Cox con covariata AGE.1. Area verde: stima del *restricted mean time* per la classe 1.

RIPK3.1, AXL.1 e  $\log(\text{AXL})$ . Attraverso i modelli di Cox stimati con tali variabili, è possibile fornire una stima indiretta del *residual mean time* seguendo i metodi proposti nel Capitolo 4. La funzione di sopravvivenza viene stimata tramite lo stimatore di Breslow e integrata numericamente su  $[0, \tau]$ . Ciò permette di calcolare l'area al di sotto della curva nell'intervallo  $[0, \tau]$  e di stimare il *restricted residual mean time* per  $t = 0$ , che corrisponde al *restricted mean time*. Per la scelta dell'istante  $\tau$ , si sceglie il tempo osservato più lungo dello studio, ossia 266.95 mesi (22.25 anni). Per quanto riguarda la variabile AGE.1, ad esempio, per la classe 1 si riporta la curva di sopravvivenza stimata fino all'istante  $\tau = 266.95$  mesi col modello di Cox (Figura 5.21). L'associata area colorata sotto la curva corrisponde alla stima del relativo *restricted mean time*. Il tempo medio alla recidiva di malattia ristretto a 266.95 mesi vale  $99.68 (\pm 12.26)$  mesi per i pazienti appartenenti

alla classe 1 (pazienti con un età minore di 48.00 anni), 66.71 ( $\pm 9.56$ ) mesi per i pazienti appartenenti alla classe 2 (pazienti con un età compresa tra 48.00 e 60.00 anni), 57.99 ( $\pm 8.00$ ) mesi per i pazienti appartenenti alla classe 3 (pazienti con un età compresa tra 60.00 e 72.00 anni), 51.12 ( $\pm 8.06$ ) mesi per i pazienti appartenenti alla classe 4 (pazienti con un età maggiore di 72.00 anni). Confrontando la classe 1 con la 4, la stima della differenza è di 48.56 (95% CI: 19.79, 77.32) mesi, indicando che i pazienti nella fascia più giovane hanno un tempo medio alla recidiva di malattia 48.56 mesi più lungo. Tali stime, tuttavia, si basano sul modello di Cox sottostante e la relativa assunzione di proporzionalità dei rischi. Inoltre, con questo approccio, l'effetto delle covariate sul *restricted residual mean time* non è direttamente interpretabile dai coefficienti di regressione del modello di Cox.

Per questo motivo, si decide di stimare il *restricted mean time* tramite modelli diretti additivi del tipo (4.14), sotto assunzione di censura indipendente dalle covariate. La stima dell'effetto di ciascuna variabile sul tempo medio alla recidiva di malattia viene analizzato stimando dei modelli diretti additivi per il *restricted mean time*, con  $t = 0$  e  $\tau = 266.95$  mesi, costruiti con una sola covariata. Risultano significativi i coefficienti del modello con covariata **AGE . 1**, di cui si riportano i risultati in Tabella 5.3, e i coefficienti del modello con covariata **AGE** (Tabella 5.4), **RIPK3 . 1** (Tabella 5.5), **AXL . 1** (Tabella 5.6), **AJCC\_PATHOLOGIC\_TUMOR\_STAGE** (Tabella 5.7). La stima dei parametri del modello diretto per la variabile **AJCC\_PATHOLOGIC\_TUMOR\_STAGE**, ad esempio, mostrano che il tempo medio alla recidiva di malattia ristretto a 266.95 mesi diminuisce all'aumentare dello stadio. In particolare, il tempo medio stimato ristretto vale 108.66 per i pazienti appartenenti allo stadio **0/I**, 76.15 (=  $108.66 - 32.51$ ) per i pazienti appartenenti allo stadio **II** e 38.98 (=  $108.66 - 69.68$ ) per i pazienti appartenenti allo stadio **III/IV**. Si noti che tali stime equivalgono a quelle che si sarebbero ottenute utilizzando le stime del *restricted mean time* ricavate tramite lo stimatore di Kaplan-Meier per ciascuna classe della variabile. Tale risultato è dovuto all'equivalenza tra lo stimatore IPCW e lo stimatore di Kaplan-Meier (paragrafo 4.2.3). Così come per le altre variabili, i risultati sono in accordo con gli effetti individuati con i modelli di Cox a una sola covariata stimati nei paragrafi 5.2.1 e 5.2.2.

	$\hat{\beta}$	<i>se</i>	<i>z</i>	<i>p - value</i>
Intercept	109.34	14.78	7.39	< 0.01
AGE.1=2	-40.71	19.50	-2.09	0.04
AGE.1=3	-59.54	16.75	-3.55	< 0.01
AGE.1=4	-75.95	15.43	-4.92	< 0.01

**Tabella 5.3:** Modello additivo diretto con covariata AGE.1: stima dei parametri ( $\hat{\beta}$ ) e *standard error* (*se*), test alla Wald (*z*, *p - value*).

	$\hat{\beta}$	<i>se</i>	<i>z</i>	<i>p - value</i>
Intercept	156.75	27.99	5.60	< 0.01
AGE	-1.51	0.41	-3.64	< 0.01

**Tabella 5.4:** Modello additivo diretto con covariata AGE: stima dei parametri ( $\hat{\beta}$ ) e *standard error* (*se*), test alla Wald (*z*, *p - value*).

	$\hat{\beta}$	<i>se</i>	<i>z</i>	<i>p - value</i>
Intercept	51.01	9.38	5.43	< 0.01
RIPK3.1=2	63.71	18.62	3.42	0.04
RIPK3.1=3	3.50	15.61	0.22	0.82
RIPK3.1=4	2.65	11.67	0.23	0.82

**Tabella 5.5:** Modello additivo diretto con covariata RIPK3.1: stima dei parametri ( $\hat{\beta}$ ) e *standard error* (*se*), test alla Wald (*z*, *p - value*).

	$\hat{\beta}$	<i>se</i>	<i>z</i>	<i>p - value</i>
Intercept	51.02	8.98	5.68	< 0.01
AXL.1=2	21.39	19.51	1.10	0.27
AXL.1=3	36.59	16.42	2.23	0.03
AXL.1=4	20.43	17.36	1.18	0.24

**Tabella 5.6:** Modello additivo diretto con covariata AXL.1: stima dei parametri ( $\hat{\beta}$ ) e *standard error* (*se*), test alla Wald (*z*, *p - value*).

	$\hat{\beta}$	<i>se</i>	<i>z</i>	<i>p - value</i>
Intercept	108.66	14.24	7.63	< 0.01
AJCC_PATHOLOGIC_TUMOR_STAGE=II	-32.51	18.93	-1.72	0.09
AJCC_PATHOLOGIC_TUMOR_STAGE=III/IV	-69.68	16.40	-4.25	< 0.01

**Tabella 5.7:** Modello additivo diretto con covariata AJCC\_PATHOLOGIC\_TUMOR\_STAGE: stima dei parametri ( $\hat{\beta}$ ) e *standard error* (*se*), test alla Wald (*z*, *p - value*).

## 5.4 Modelli di regressione

Sulla base delle informazioni ottenute nei paragrafi precedenti, si propongono dei modelli di regressione che permettano di valutare l'effetto congiunto di più variabili sui tempi di sopravvivenza liberi dalla malattia (DFS). Per le procedure di selezione delle covariate, le variabili quantitative vengono mantenute continue e non raggruppate, a eccezione della variabile `CONSENSUS_SCORE_CD8` che è stata raggruppata nei paragrafi precedenti (`CONSENSUS_SCORE_CD8.1`) per distinguere i campioni ad alto (=Low) e basso (=High) infiltrato immunitario. Le variabili relative alle espressioni geniche vengono utilizzate con la trasformata logaritmica. Per semplicità, nella variabile AJCC\_PATHOLOGIC\_TUMOR\_STAGE si decide di accorpate i pazienti negli stadi 0, I, II e considerare così le classi 0/I/II, III/IV.

### 5.4.1 Modello di Cox

Per prima cosa, viene specificato un modello semiparametrico di Cox a rischi proporzionali. Per quanto riguarda la scelta delle variabili si fa riferimento ai test di log-rapporto di verosimiglianza, implementando la procedura di selezione in avanti (*forward*) e all'indietro (*backward*), utilizzando un livello di significatività del 5%.

Entrambe le procedure conducono a un modello avente come covariate le variabili AJCC\_PATHOLOGIC\_TUMOR\_STAGE,  $\log(\text{AXL})$ , AGE. In accordo con i risultati dei paragrafi precedenti, dai test basati sui residui di Schoenfeld e i residui cumulati emerge che la variabili AGE e AJCC\_PATHOLOGIC\_TUMOR\_STAGE violano l'assunzione di proporzionalità dei rischi. Infatti, il test sui residui di Schoenfeld è significativo

per entrambe le covariate al 10% con  $g(t) = t$ , e al 5% con  $g(t) = \log(t)$  e con i ranghi dei tempi all'evento, mentre non è significativo per  $\log(\text{AXL})$ . Allo stesso modo, il test basato sui residui cumulati di martigala è significativo per le variabili `AJCC_PATHOLOGIC_TUMOR_STAGE`, `AGE` e non è significativo per  $\log(\text{AXL})$ .

Per tale motivo, si propone la procedura di selezione in avanti e all'indietro con il modello stratificato per la variabile `AJCC_PATHOLOGIC_TUMOR_STAGE` e la variabile `AGE` raggruppata in due classi con la mediana. Per i quattro sottogruppi formati dalla variabile `AJCC_PATHOLOGIC_TUMOR_STAGE` e dalla variabile `AGE` raggruppata, il modello prevede quattro funzioni di rischio di base diverse e coefficienti di regressione comuni. I risultati del modello ottenuto sono riportati in Tabella 5.8. Dalla

	$\hat{\beta}$	<i>se</i>	<i>z</i>	<i>p - value</i>
$\log(\text{AXL})$	-0.21	0.07	-2.92	< 0.01

**Tabella 5.8:** Modello di Cox: stima dei parametri ( $\hat{\beta}$ ) e *standard error* (*se*), test alla Wald (*z*, *p - value*).

stima del coefficiente risulta che il rischio di recidiva di malattia per ogni strato formato dalle variabili `AJCC_PATHOLOGIC_TUMOR_STAGE` e `AGE` raggruppata è del 18.78% più basso per ciascun incremento unitario della variabile  $\log(\text{AXL})$  ( $HR = 0.81$ ; 95% CI: 0.71, 0.93). Il modello di Cox stratificato permette di tener conto di effetti non proporzionali, tuttavia presenta alcuni limiti. Infatti, gli effetti delle variabili di stratificazione non sono stimabili e i risultati dipendono dal modo in cui le covariate di stratificazione sono state categorizzate.

#### 5.4.2 Modellazione del *Residual Mean Time*

Le stime basate sul modello di Cox del paragrafo precedente si affidano sulla assunzione di proporzionalità degli azzardi. Al contrario, l'approccio di regressione diretta per il *residual mean time* proposto nella tesi, richiede assunzioni meno stringenti. Tra queste, vi è la cruciale assunzione sui sottostanti modelli utilizzati per la distribuzione della censura.

Viene dunque stimato un modello diretto additivo per il *restricted residual mean time* riportato nella (4.14), fissando l'orizzonte  $\tau$  pari al tempo osservato più lungo dello studio (266.95 mesi). Il tempo di sopravvivenza  $T$  libero da malattia viene studiato condizionatamente a  $T \geq t$ , con  $t = 0$  mesi. La stima dei parametri di regressione è data dall'equazione (4.19) e la costruzione dei pesi è ottenuta inizialmente tramite stime non parametriche dei tempi alla censura, indipendentemente dalle covariate. Per la selezione delle variabili del modello, viene implementata la procedura di selezione all'indietro utilizzando il test alla Wald e un livello di significatività del 5%. Dalla Tabella 5.9 (Modello A) risulta che le variabili AGE,  $\log(\text{AXL})$  e AJCC\_PATHOLOGIC\_TUMOR\_STAGE hanno un impatto congiunto significativo sul *restricted residual mean time*. Si noti che nella costruzione del modello si è tenuto conto di eventuali effetti di interazione, che non sono risultati significativi.

In aggiunta, viene investigata la possibile dipendenza della censura dalle variabili. Come riportato nel paragrafo 2.3, gli individui che sperimentano l'evento d'interesse vengono considerati come osservazioni censurate e viceversa. La distribuzione della censura si assume segua un modello di Cox avente come covariate le variabili  $\log(\text{MERTK})$ , AGE, FRACTION\_GENOME\_ALTERED, SEX, scelte con le procedure di selezione in avanti e all'indietro. I test di proporzionalità non danno alcuna evidenza di violazione dell'assunzione di proporzionalità dei rischi: per ogni covariata il test basato sui residui di Schoenfeld non è significativo ( $p - \text{value} > 0.1$ ) con  $g(t) = t$ , con  $g(t) = \log(t)$ , con i ranghi dei tempi all'evento e il test sui residui cumulati di martingala non è significativo ( $p - \text{value} > 0.1$ ).

Con i risultati ottenuti, il Modello A viene stimato nuovamente con pesi dipendenti dalle covariate nell'equazione di stima (4.19), costruiti con il modello di Cox specificato per il tempo alla censura. Le stime sono riportate in Tabella 5.9 (Modello B). Tale modello ha il vantaggio di tener conto della possibile presenza di censura dipendente nello studio. Tuttavia, nel modello di Cox considerato per la costruzione dei pesi, seppur non vi sia evidenza di violazione della proporzionalità dei rischi, non è detto che la forma funzionale delle covariate utilizzate sia corretta e è opportuno notare che potrebbero esistere altri approcci migliori per modellare

il meccanismo di censura (paragrafo 2.3).

	$\hat{\beta}$	<i>se</i>	<i>z</i>	<i>p - value</i>
<b>Modello A</b>				
Intercept	102.80	34.34	2.99	< 0.01
AGE	-1.33	0.36	-3.74	< 0.01
log(AXL)	9.79	4.54	2.15	0.03
AJCC_PATHOLOGIC_TUMOR_STAGE=III/IV	-49.46	11.37	-4.35	< 0.01
<b>Modello B</b>				
Intercept	36.80	19.33	1.90	0.06
AGE	-0.58	0.19	-3.02	< 0.01
log(AXL)	8.19	3.05	2.68	< 0.01
AJCC_PATHOLOGIC_TUMOR_STAGE=III/IV	-34.20	6.41	-5.34	< 0.01

**Tabella 5.9:** Modelli additivi diretti ( $t = 0$ ) con pesi non dipendenti (Modello A) e dipendenti (Modello B) dalle covariate: stima dei parametri ( $\hat{\beta}$ ) e *standard error* (*se*), test alla Wald (*z*, *p - value*).

Confrontando il Modello A con il Modello B, si nota che la modellazione del tempo di censura non ha un ruolo particolarmente rilevante. Infatti, per entrambi i modelli, dalle stime dei parametri si conclude che il tempo medio alla recidiva di malattia ristretto a 266.95 mesi si riduce significativamente all'aumentare dell'età del paziente, cresce significativamente all'aumentare dell'espressione del gene Axl e si riduce significativamente all'aumentare dello stadio del tumore. Tuttavia, la quantificazione degli effetti è minore nel Modello B, dove la stima dei coefficienti del modello, interpretata come aumento o diminuzione attesa del *residual mean time* dovuta ai fattori di rischio, è di poco più bassa. Dalle stime dei coefficienti di regressione del Modello A, per esempio, risulta che il tempo medio alla recidiva ristretto diminuisce di 1.33 mesi per un incremento unitario della variabile AGE a parità delle alte covariate, aumenta di 9.79 mesi per un incremento unitario della variabile log(AXL) a parità delle alte covariate, diminuisce di 49.46 mesi per gli individui appartenenti alla fascia III/IV della variabile AJCC\_PATHOLOGIC\_TUMOR\_STAGE rispetto a quelli appartenenti alla fascia 0/I/II a parità delle altre covariate.

---

In conclusione, in accordo col modello di Cox stimato nel paragrafo precedente, emerge l'importanza dell'espressione del gene *Axl*. Tale espressione ha un effetto marginale significativo sul tasso di recidiva di malattia (paragrafo 5.2.2) e sul tempo medio alla recidiva ristretto (paragrafo 5.3). Con i modelli di regressione ottenuti nei paragrafi 5.4.1 e 5.4.2 l'effetto sul tasso di recidiva di malattia e sul tempo medio alla recidiva viene aggiustato con le variabili `AGE` e `AJCC_PATHOLOGIC_TUMOR_STAGE`.



# Capitolo 6

## Studi di simulazione

In questo capitolo sono riportati alcuni studi di simulazione al fine di valutare le proprietà inferenziali dei modelli di regressione diretti e indiretti per il *restricted residual mean time*. Le simulazioni sono implementate tramite il software R (ver. 3.6.1, [www.r-project.org](http://www.r-project.org)), utilizzando i pacchetti R `survival` e `timereg`. In Appendice B si riporta il codice sviluppato.

Gli studi di simulazione proposti di basano su 2000 repliche di dataset con numerosità campionaria  $n$ . Si assume che il tempo alla censura a destra  $C$  e il tempo di sopravvivenza  $T$  siano indipendenti condizionatamente alle covariate, e che  $C$  sia indipendente dalle covariate. Per la scelta delle covariate, si considerano due variabili esplicative dicotomiche  $X$  e  $Z$ , entrambe generate da una distribuzione di Bernoulli di parametro  $p = 0.5$ . Il tempo di sopravvivenza  $T$  è generato da una distribuzione esponenziale di parametro  $\lambda = \alpha e^{X\beta_1 + Z\beta_2}$ , mentre  $C$  segue una distribuzione continua Uniforme definita nell'intervallo  $[0, a]$ , dove il parametro  $a$  viene fissato in base alla percentuale di censura che si desidera.

Negli studi di simulazione presentati, il parametro d'interesse è il *restricted residual mean time*  $\mu$  condizionato alle covariate e si studia il caso in cui  $t = 0$ . Dunque, come specificato nel paragrafo 3.1,

$$\mu = E(\min(T, \tau) | X, Z) = \int_0^\tau S(t | X, Z) dt. \quad (6.1)$$

Poiché  $T$  è distribuito esponenzialmente, il relativo parametro  $\lambda$  esprime il tasso con il quale si verificano gli eventi, che per definizione è costante. Per come è stato definito, tale parametro dipende dalle variabili esplicative  $X$ ,  $Z$  e la funzione di sopravvivenza di  $T$  è allora definita come

$$S(t | X, Z) = \exp(-\lambda t) = \exp(-\alpha e^{X\beta_1 + Z\beta_2} t). \quad (6.2)$$

Dalle espressioni (6.1) e (6.2), consegue che

$$\begin{aligned} \mu &= E(\min(T, \tau) | X, Z) = \int_0^\tau \exp(-\alpha e^{X\beta_1 + Z\beta_2} t) dt \\ &= \frac{1 - \exp(-\tau \alpha e^{X\beta_1 + Z\beta_2})}{\alpha e^{X\beta_1 + Z\beta_2}}. \end{aligned} \quad (6.3)$$

Gli studi di simulazione vengono implementati per  $n = 200$  (Tabella 6.1) e  $n = 400$  (Tabella 6.2). Il tempo di restrizione è fissato a  $\tau = 10$ , mentre  $\alpha = 0.05$ ,  $\beta_1 = 1$  e  $\beta_2 = -1$ . Il parametro  $a$  della distribuzione di  $C$  viene scelto pari a 116 per avere una percentuale media di censura del 20% ( $C\% = 20$ ) e a 70 per avere una percentuale media del 30% ( $C\% = 30$ ). I veri valori dei *restricted residual mean time* sono calcolati per diversi strati delle covariate tramite l'espressione (6.3). In particolare, si definiscono gli strati 0 ( $X = 0$  e  $Z = 0$ ), 1 ( $X = 0$  e  $Z = 1$ ), 2 ( $X = 1$  e  $Z = 0$ ), 3 ( $X = 1$  e  $Z = 1$ ). Tali valori vengono confrontati con le stime IPCW di un modello di regressione diretta additivo per il *restricted residual mean time* (4.14) con variabili esplicative  $X$  e  $Z$ , dove i pesi IPCW non dipendono dalle covariate, e le stime dei *restricted residual mean time* ottenute indirettamente attraverso un modello di Cox a rischi proporzionali (4.8) con variabili esplicative  $X$  e  $Z$ .

Per i due modelli citati e per ciascuno dei quattro gruppi definiti (strato 0, 1, 2, 3), si riportano la media  $\hat{\mu}$  delle 2000 stime dei *restricted residual mean time*, lo scarto quadratico medio empirico  $sd(\hat{\mu})$  dei *restricted residual mean time* stimati, la media  $\bar{se}$  delle 2000 stime degli *standard error* espresse dalla formula (4.20), il rapporto tra  $sd(\hat{\mu})$  e  $\bar{se}$ , la probabilità di copertura empirica per gli intervalli di confidenza alla Wald al 95% (95% CP). Quest'ultima quantità specifica la percentuale di volte in cui il vero valore d'interesse cade all'interno dell'intervallo in ogni replicazione.

Dai risultati numerici riportati nelle Tabelle 6.1, 6.2 si può notare come entrambi i metodi di stima proposti diano risultati soddisfacenti e simili al variare della percentuale di censura e della numerosità campionaria. Le stime puntuali dei *restricted residual mean time* sono molto accurate per entrambi i modelli e tendono a coincidere con i veri valori. Le stime degli *standard error* proposte sono in accordo con lo *standard error* empirico, e il rapporto  $sd(\hat{\mu})/\bar{se}$  è sempre vicino a 1. Le probabilità di copertura sono ragionevoli e vicine al livello nominale del 95%. Per ogni strato, il rapporto  $sd(\hat{\mu})/\bar{se}$  è più vicino a 1 e la probabilità di copertura è più alta all'aumentare della numerosità campionaria, anche se le differenze riscontrate sono minime.

	Strato	$\mu$	$\hat{\mu}$	$sd(\hat{\mu})$	$\bar{se}$	$sd(\hat{\mu})/\bar{se}$	95% CP
$C\% = 20$							
Modello diretto	0	7.87	7.87	0.47	0.46	1.03	0.93
	1	9.13	9.13	0.32	0.31	1.03	0.92
	2	5.47	5.46	0.53	0.52	1.02	0.94
	3	7.87	7.86	0.47	0.46	1.02	0.94
Modello indiretto	0	7.87	7.87	0.34	0.33	1.04	0.94
	1	9.13	9.14	0.19	0.19	1.01	0.94
	2	5.47	5.47	0.47	0.45	1.05	0.93
	3	7.87	7.87	0.34	0.33	1.03	0.94
$C\% = 30$							
Modello diretto	0	7.87	7.87	0.46	0.46	1.00	0.94
	1	9.13	9.11	0.33	0.32	1.02	0.92
	2	5.47	5.43	0.54	0.52	1.02	0.94
	3	7.87	7.85	0.47	0.46	1.02	0.94
Modello indiretto	0	7.87	7.87	0.35	0.34	1.04	0.93
	1	9.13	9.13	0.21	0.20	1.03	0.93
	2	5.47	5.45	0.47	0.45	1.03	0.93
	3	7.87	7.86	0.36	0.35	1.03	0.94

**Tabella 6.1:** Risultati delle simulazioni con numerosità campionaria  $n = 200$ .

	Strato	$\mu$	$\hat{\mu}$	$sd(\hat{\mu})$	$\bar{s}e$	$sd(\hat{\mu})/\bar{s}e$	95% CP
$C\% = 20$							
Modello diretto	0	7.87	7.87	0.32	0.32	0.99	0.95
	1	9.13	9.13	0.23	0.23	1.00	0.94
	2	5.47	5.46	0.37	0.37	1.01	0.95
	3	7.87	7.88	0.33	0.32	1.02	0.94
Modello indiretto	0	7.87	7.87	0.23	0.23	1.00	0.95
	1	9.13	9.14	0.13	0.13	1.01	0.94
	2	5.47	5.47	0.33	0.32	1.02	0.94
	3	7.87	7.88	0.24	0.24	1.02	0.94
$C\% = 30$							
Modello diretto	0	7.87	7.87	0.33	0.33	1.00	0.94
	1	9.13	9.14	0.23	0.23	1.02	0.93
	2	5.47	5.45	0.37	0.37	0.99	0.95
	3	7.87	7.87	0.33	0.33	1.01	0.95
Modello indiretto	0	7.87	7.87	0.24	0.24	0.99	0.94
	1	9.13	9.13	0.14	0.14	0.99	0.94
	2	5.47	5.47	0.32	0.32	0.98	0.95
	3	7.87	7.87	0.25	0.25	1.01	0.94

**Tabella 6.2:** Risultati delle simulazioni con numerosità campionaria  $n = 400$ .



# Capitolo 7

## Conclusioni

Nell'analisi di uno studio longitudinale in cui la variabile risposta d'interesse è il tempo all'evento, è stato dimostrato come il *residual mean time* sia una quantità facilmente interpretabile e riassuntiva della funzione di sopravvivenza in presenza di censura a destra. Si è visto come, in ambito non parametrico, questa possa essere stimata in modo consistente dall'area al di sotto della curva di Kaplan-Meier. Tuttavia, poiché la coda di tale curva è spesso difficile da stimare correttamente a causa della censura, è preferibile studiare la quantità *restricted residual mean time*, considerando l'area al di sotto della curva fino a un certo istante  $\tau > 0$ . Essa è interpretata come l'aspettativa di vita residua fino a  $\tau$  per un soggetto sopravvissuto fino al tempo  $t$ . La differenza (o il rapporto) tra la stima dei *restricted residual mean time* tra gruppi di soggetti con caratteristiche diverse permette di confrontare curve di sopravvivenza differenti, e costituisce un'alternativa robusta all'*hazard ratio* nel caso in cui quest'ultimo non sia costante nel tempo. Il *restricted residual mean time* per ciascun gruppo può essere calcolato indirettamente attraverso i modelli di Cox, integrando numericamente le funzioni di sopravvivenza stimate a partire dal modello stesso. Si noti che tale procedura dipende dal modello di Cox sottostante e produce risultati distorti nel caso in cui la relativa assunzione di proporzionalità del modello sia violata. Tale approccio può essere esteso aggiungendo per ulteriori covariate come riportato da Karrison (1987, 1997) e Zucker (1998), in modo da tener conto di eventuali fattori di confondimento. In questi

approcci, il *restricted residual mean time* non è tuttavia modellato direttamente con le covariate, ma per mezzo della funzione azzardo, e dunque l'influenza delle covariate sulle quantità d'interesse non è descritta esplicitamente da dei coefficienti di regressione.

In questo lavoro, si è presa in considerazione una classe di modelli addittivi diretti per il *restricted residual mean time*. Al contrario dei modelli di regressione indiretti citati, il predittore lineare è in relazione con il *restricted residual mean time* tramite una funzione legame, dunque i coefficienti di regressione sono direttamente interpretabili come aumento o diminuzione attesa della vita residua causata dai fattori di rischio. La scelta della funzione legame definisce una vasta classe di modelli. Tra queste è stata proposta la forma additiva e moltiplicativa, che permettono una semplice interpretazione dei coefficienti di regressione. Uno dei vantaggi principali di questa famiglia di modelli è la possibilità di tener conto del meccanismo di censura. La stima dei parametri, infatti, avviene tramite equazioni di stima pesate con tecnica *Inverse Probability of Censoring Weighting* (IPCW), in cui ai soggetti non censurati sono assegnati dei pesi dipendenti dalle covariate o meno a seconda che il meccanismo di censura sia dipendente o indipendente dalle covariate. Ciò nonostante, quando si applica la stima IPCW l'indipendenza tra i tempi alla censura e le covariate per la scelta dei pesi dev'essere investigata attentamente e la corretta modellazione della distribuzione della censura è cruciale (come discusso nel Capitolo 2). Infine, viene dimostrata l'equivalenza tra la stima IPCW del *restricted residual mean time* fornita dal modello con una singola covariata qualitativa e quella fornita dello stimatore di Kaplan-Meier.

Nel Capitolo 5 di applicazione ai dati, è stato valutato l'effetto marginale e congiunto delle variabili sui tempi di sopravvivenza liberi da malattia (DFS), con particolare attenzione all'effetto delle variabili di espressione dei geni considerate. Si è concluso che l'età del paziente, lo stadio del tumore, l'espressione del geni Ripk3 e Axl hanno un effetto marginale sul tasso di recidiva di malattia. Attraverso i modelli di Cox stimati per misurare tali effetti, sono state trovate le stime indirette dei *restricted mean time*. Tali stime tuttavia, dipendono dal modello di Cox sottostante e dalla relativa assunzione di proporzionalità degli azzardi del mo-

---

dello, che è risultata in alcuni casi violata. Per questo si è ricorso ai modelli diretti additivi del paragrafo 4.2, che misurano l'effetto marginale diretto delle variabili sul *restricted mean time*. I risultati si sono comunque dimostrati in linea con quelli precedenti. Per valutare l'effetto congiunto di più variabili sui tempi DFS, vengono stimati un modello di Cox finale e due modelli diretti additivi per il *restricted residual mean time*. Il modello di Cox esprime gli effetti delle covariate sul tasso di rischio, che è una quantità istantanea e può variare nel tempo. I modelli additivi considerati, al contrario, hanno il vantaggio di esprimere gli effetti delle covariate sul *restricted residual mean time*, che rappresenta una misura riassuntiva globale dell'intervallo di tempo considerato e non è vincolata dal tempo specifico. Dai modelli implementati emerge l'importanza dell'espressione del gene Axl. Per correggere la violazione della proporzionalità degli azzardi, il modello di Cox è stato stratificato per le variabili identificative dello stadio tumorale e della classe di età del paziente, concludendo che al diminuire dell'espressione del gene Axl il rischio di recidiva di malattia aumenta. Il modello additivo diretto per il *restricted residual mean time* è stato stimato con pesi dipendenti e indipendenti dalle covariate. Per entrambi i modelli, risulta che il tempo medio alla recidiva ristretto a 266.95 mesi cresce all'aumentare dell'espressione del gene Axl, a parità di età e stadio tumorale del paziente.

Per concludere, negli studi di simulazione del Capitolo 6 sono state valutate le proprietà inferenziali dei modelli di regressione diretti e indiretti per il *restricted residual mean time* presentati nel Capitolo 4, al variare della percentuale di censura e della numerosità campionaria. In ogni scenario considerato, per entrambi gli approcci si riscontrano risultati simili e soddisfacenti.



# Bibliografia

- Aalen, Odd O (1989). «A linear regression model for the analysis of life times». In: *Statistics in medicine* 8.8, pp. 907–925.
- Andersen, Per Kragh, Mette Gerster Hansen e John P Klein (2004). «Regression analysis of restricted mean survival time based on pseudo-observations». In: *Lifetime data analysis* 10.4, pp. 335–350.
- Angrist, Joshua D e Jörn-Steffen Pischke (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.
- Aran, Dvir, Marina Sirota e Atul J Butte (2015). «Systematic pan-cancer analysis of tumour purity». In: *Nature Communications* 6, p. 8971.
- Austin, Peter C (2011). «An introduction to propensity score methods for reducing the effects of confounding in observational studies». In: *Multivariate behavioral research* 46.3, pp. 399–424.
- Balch, Charles M et al. (2001). «Final version of the American Joint Committee on Cancer staging system for cutaneous melanoma». In: *Journal of Clinical Oncology* 19.16, pp. 3635–3648.
- Chen, Y Q et al. (2005). «Semiparametric estimation of proportional mean residual life model in presence of censoring». In: *Biometrics* 61.1, pp. 170–178.
- Chen, Ying Qing (2007). «Additive expectancy regression». In: *Journal of the American Statistical Association* 102.477, pp. 153–166.
- Chen, Ying Qing e Seu Cheng (2006). «Linear life expectancy regression with censored data». In: *Biometrika* 93.2, pp. 303–313.

- Cortese, Giuliana, Stine A Holmboe e Thomas H Scheike (2017). «Regression models for the restricted residual mean life for right-censored and left-truncated data». In: *Statistics in medicine* 36.11, pp. 1803–1822.
- Cortese, Giuliana, Thomas H Scheike e Torben Martinussen (2010). «Flexible survival regression modelling». In: *Statistical Methods in Medical Research* 19.1, pp. 5–28.
- Cox, David R (1972). «Regression models and life-tables». In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2, pp. 187–202.
- Davison, Anthony Christopher (2003). *Statistical models*. Vol. 11. Cambridge University Press.
- Efron, Bradley (1967). «The two sample problem with censored data». In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 4. University of California Press, Berkeley, CA, pp. 831–853.
- Feng, Changyong, Hongyue Wang e Xin M Tu (2010). «Power loss of stratified log-rank test in homogeneous samples». In: *International Journal of Quality, Statistics, and Reliability* 2010.
- Finotello, Francesca e Zlatko Trajanoski (2018). «Quantifying tumor-infiltrating immune cells from transcriptomics data». In: *Cancer Immunology, Immunotherapy* 67.7, pp. 1031–1040.
- Flaherty, Keith T et al. (2010). «Inhibition of mutated, activated BRAF in metastatic melanoma». In: *New England Journal of Medicine* 363.9, pp. 809–819.
- Friedman, Jerome, Trevor Hastie e Robert Tibshirani (2001). *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York.
- Gilbert, Peter B et al. (2002). «Simultaneous inferences on the contrast of two hazard functions with censored observations». In: *Biometrics* 58.4, pp. 773–780.
- Grambsch, Patricia M e Terry M Therneau (1994). «Proportional hazards tests and diagnostics based on weighted residuals». In: *Biometrika* 81.3, pp. 515–526.

- Irwin, J O (1949). «The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice». In: *Epidemiology & Infection* 47.2, pp. 188–189.
- Kalbfleisch, John D e Ross L Prentice (2002). *The statistical analysis of failure time data*. Vol. 360. John Wiley & Sons.
- Kaplan, Edward L e Paul Meier (1958). «Nonparametric estimation from incomplete observations». In: *Journal of the American statistical association* 53.282, pp. 457–481.
- Karadeniz, Pinar Gunel e Ilker Ercan (2017). «Examining tests for comparing survival curves with right censored data». In: *Stat Transit* 18.2, pp. 311–328.
- Karrison, Theodore (1987). «Restricted mean life with adjustment for covariates». In: *Journal of the American Statistical Association* 82.400, pp. 1169–1176.
- Karrison, Theodore G (1997). «Use of Irwin’s restricted mean as an index for comparing survival in different treatment groups—interpretation and power considerations». In: *Controlled clinical trials* 18.2, pp. 151–167.
- Klein, John P. e Melvin L. Moeschberger (2003). *Survival analysis: Techniques for censored and truncated data*, p. 536.
- Liang, Kung-Yee e Scott L Zeger (1986). «Longitudinal data analysis using generalized linear models». In: *Biometrika* 73.1, pp. 13–22.
- Lin, D Y e Zhiliang Ying (1994). «Semiparametric analysis of the additive risk model». In: *Biometrika* 81.1, pp. 61–71.
- Lin, Danyu Y, Lee-Jen Wei e Zhiliang Ying (1993). «Checking the Cox model with cumulative sums of martingale-based residuals». In: *Biometrika* 80.3, pp. 557–572.
- Maguluri, Gangaji e Cun-Hui Zhang (1994). «Estimation in the mean residual life regression model». In: *Journal of the Royal Statistical Society: Series B (Methodological)* 56.3, pp. 477–489.
- Mansournia, Mohammad Ali e Douglas G Altman (2016). «Inverse probability weighting». In: *BMJ* 352.
- Moore, Dirk F (2016). *Applied survival analysis using R*. Springer.

- Pace, Luigi e Alessandra Salvan (2001). *Introduzione alla statistica: Inferenza, verosimiglianza, modelli*. Cedam.
- Robins, James M e Andrea Rotnitzky (1992). «Recovery of information and adjustment for dependent censoring using surrogate markers». In: *AIDS epidemiology*. Springer, pp. 297–331.
- Rosenbaum, Paul R e Donald B Rubin (1983). «The central role of the propensity score in observational studies for causal effects». In: *Biometrika* 70.1, pp. 41–55.
- Royston, Patrick e Mahesh K B Parmar (2002). «Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects». In: *Statistics in medicine* 21.15, pp. 2175–2197.
- (2011). «The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt». In: *Statistics in medicine* 30.19, pp. 2409–2421.
- Satten, Glen A e Somnath Datta (2001). «The Kaplan–Meier estimator as an inverse-probability-of-censoring weighted average». In: *The American Statistician* 55.3, pp. 207–210.
- Seaman, Shaun R e Ian R White (2013). «Review of inverse probability weighting for dealing with missing data». In: *Statistical methods in medical research* 22.3, pp. 278–295.
- Sheldon, Emily H (2013). *Choosing the Cut Point for a Restricted Mean in Survival Analysis, a Data Driven Method*.
- Sun, Liuquan, Xinyuan Song e Zhigang Zhang (2011). «Mean residual life models with time-dependent coefficients under right censoring». In: *Biometrika* 99.1, pp. 185–197.
- Sun, Liuquan e Zhigang Zhang (2009). «A class of transformed mean residual life models with censored survival data». In: *Journal of the American Statistical Association* 104.486, pp. 803–815.

- Tian, Lu, Lihui Zhao e L J Wei (2013). «Predicting the restricted mean event time with the subject's baseline covariates in survival analysis». In: *Biostatistics* 15.2, pp. 222–233.
- Uno, Hajime et al. (2014). «Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis». In: *Journal of clinical Oncology* 32.22, p. 2380.
- Willems, Sanne J W e M Fiocco (2014). *Inverse Probability Censoring Weights for Routine Outcome Monitoring Data*.
- Zhao, Lihui et al. (2016). «On the restricted mean survival time curve in survival analysis». In: *Biometrics* 72.1, pp. 215–221.
- Zucker, David M (1998). «Restricted mean life with covariates: modification and extension of a useful survival analysis method». In: *Journal of the American Statistical Association* 93.442, pp. 702–709.



# Appendice A

## Codice R - Applicazione ai dati

In questa appendice è contenuto parte del codice R utilizzato per le analisi svolte nel Capitolo 5.

---

```
# Caricamento del dataset
-----

#Caricamento del dataset relativo alle variabili di espressione dei geni
dati0=read.delim("F:/Pregetto melanomi/Gene_Expression_Data_SKCM.txt")
#Caricamento dataset relativo alle variabili cliniche
dati1=read.delim("F:/Pregetto melanomi/Clinical_Data_SKCM.txt")

#Pulizia del dataset (gestione dei valori mancanti, selezione e ricodifica
  delle variabili utilizzate nelle analisi)
dati2=dati0[c(212,13,31,201,177,130,10),]
dati21=dati2[,-1]
dati21=t(dati21)
dati21=data.frame(dati21)
colnames(dati21)=(dati2[,1])
rownames(dati21)=rownames(dati1)
dati=cbind(dati1,dati21)
dati=dati[,-c(1,9,10,11,12,16,18,21:27)]
```

```
dati=dati[,-c(9)]
dati$DFS_STATUS[dati$DFS_STATUS==""]=NA
dati$DFS_STATUS=factor(dati$DFS_STATUS)
dati$TUMOR_SITE[dati$TUMOR_SITE==""]=NA
dati$TUMOR_SITE=factor(dati$TUMOR_SITE)
dati$AJCC_PATHOLOGIC_TUMOR_STAGE[dati$AJCC_PATHOLOGIC_TUMOR_STAGE==""]=NA
dati$AJCC_PATHOLOGIC_TUMOR_STAGE=factor(dati$AJCC_PATHOLOGIC_TUMOR_STAGE)
dati$BRAF_MUT=factor(dati$BRAF_MUT)
dati=na.omit(dati)
which(dati$OS_MONTHS==-0.07)
dati[186,]
dati=dati[-186,]
levels(dati$OS_STATUS)=c(1,0)
dati$OS_STATUS=as.numeric(as.character(dati$OS_STATUS))
levels(dati$DFS_STATUS)=c(0,1)
dati$DFS_STATUS=as.numeric(as.character(dati$DFS_STATUS))
levels(dati$TUMOR_SITE)=c("Other","Trunk/Extremities","Trunk/Extremities","
  Trunk/Extremities","Other","Other","Other","Other","Trunk/Extremities","
  Trunk/Extremities")
levels(dati$AJCC_PATHOLOGIC_TUMOR_STAGE)=c("II","0/I","0/I","0/I","0/I","II","
  II","II","II","III/IV","III/IV","III/IV","III/IV","III/IV")
CONSENSUS_SCORE_CD8.1=dati$CONSENSUS_SCORE_CD8
CONSENSUS_SCORE_CD8.1[dati$CONSENSUS_SCORE_CD8>median(dati$CONSENSUS_SCORE_CD8)
  ]="High"
CONSENSUS_SCORE_CD8.1[dati$CONSENSUS_SCORE_CD8<=median(dati$CONSENSUS_SCORE_CD8
  )]="Low"
CONSENSUS_SCORE_CD8.1=factor(CONSENSUS_SCORE_CD8.1)

#Statistiche descrittive sulle variabili
summary(dati)

#Caricamento librerie
library(survival)
```

```
library(timereg)
library(survminer)

# Analisi univariata di OS e DFS
-----

#Stimatore di Kaplan-Meier per OS
fit=survfit(Surv(DFS_MONTHS,DFS_STATUS)~1,data=dati)
ggsurvplot(fit,risk.table.col="strata",xlab="Time in months",risk.table=TRUE,
  ggtheme=theme_bw(),legend=c(0.75,0.75),risk.table.y.text.col=T,risk.table.y
  .text=FALSE)
quantile(fit)

#Stimatore di Kaplan-Meier per DFS
fit=survfit(Surv(DFS_MONTHS,DFS_STATUS)~1,data=dati)
ggsurvplot(fit,risk.table.col="strata",xlab="Time in months",risk.table=TRUE,
  ggtheme=theme_bw(),legend=c(0.75,0.75),risk.table.y.text.col=T,risk.table.y
  .text = FALSE)
quantile(fit)

#Confronto tra i tempi OS e DFS
table(dati$OS_STATUS,dati$DFS_STATUS)
summary(dati[dati$DFS_STATUS==1,"DFS.MONTHS"])
summary(dati[dati$DFS_STATUS==0,"DFS.MONTHS"])
summary(dati[dati$DFS_STATUS==1 & dati$OS_STATUS==1,"DFS.MONTHS"])
summary(dati[dati$DFS_STATUS==1 & dati$OS_STATUS==0,"DFS.MONTHS"])
boxplot(dati[dati$DFS_STATUS==1,"DFS.MONTHS"],dati[dati$DFS_STATUS==0,"DFS_
  MONTHS"],dati[dati$DFS_STATUS==1 & dati$OS_STATUS==1,"DFS.MONTHS"],dati[
  dati$DFS_STATUS==1 & dati$OS_STATUS==0,"DFS.MONTHS"],names=c("A", "B", "C", "
  D"))

#Effetto marginale delle variabili sui tempi DFS. Per semplicita' si riporta il
  codice per la sola variabile AGE. Il codice per le altre variabili ne e'
```

una semplice ripetizione

```
-----

#Variabile categorizzata
summary(dati$AGE)
dati$AGE.1[dati$AGE<=48.00]="1"
dati$AGE.1[dati$AGE>48.00 & dati$AGE<=60.00]="2"
dati$AGE.1[dati$AGE>60.00 & dati$AGE<=72.00]="3"
dati$AGE.1[dati$AGE>72.00]="4"
dati$AGE.1=factor(dati$AGE.1)

#Curve di Kaplan-Meier per ciascun gruppo
ggsurvplot(survfit(Surv(DFS_MONTHS,DFS_STATUS)~AGE.1,data=dati),risk.table.col=
  "strata",xlab="Time in months",risk.table=TRUE,ggtheme=theme_bw(),legend=c
  (0.75,0.75),risk.table.y.text.col=T,risk.table.y.text = FALSE,conf.int=T)

#Stima del modello di Cox con variabile esplicativa categorizzata
out.survival=coxph(Surv(DFS_MONTHS,DFS_STATUS)~AGE.1,data=dati)
#Test di proporzionalita'
ggcoxzph(cox.zph(out.survival, transform="identity"),ggtheme=theme_bw())
ggcoxzph(cox.zph(out.survival, transform="log"),ggtheme=theme_bw())
ggcoxzph(cox.zph(out.survival, transform="rank"),ggtheme=theme_bw())
summary(cox.aalen(Surv(DFS_MONTHS,DFS_STATUS)~prop(AGE.1),data=dati))

#Test di Tarone-Ware, Peto-Peto, Peto-Peto modificato
surv_pvalue(survfit(Surv(DFS_MONTHS,DFS_STATUS)~AGE.1,data=dati),method="sqrtN"
  )
surv_pvalue(survfit(Surv(DFS_MONTHS,DFS_STATUS)~AGE.1,data=dati),method="S1")
surv_pvalue(survfit(Surv(DFS_MONTHS,DFS_STATUS)~AGE.1,data=dati),method="S2")

#Stima del modello di Cox con variabile esplicativa continua
out.survival=coxph(Surv(DFS_MONTHS,DFS_STATUS)~AGE,data=dati)
#Test di proporzionalita'
```

```
ggcoxzph(cox.zph(out.survival, transform="identity"),ggtheme=theme_bw())
ggcoxzph(cox.zph(out.survival, transform="log"),ggtheme=theme_bw())
ggcoxzph(cox.zph(out.survival, transform="rank"),ggtheme=theme_bw())
summary(cox.aalen(Surv(DFS_MONTHS,DFS_STATUS)~prop(AGE),data=dati))

#Stima univariata del restricted mean time
-----

#Stima indiretta tramite modello di Cox
out=cox.aalen(Surv(DFS_MONTHS,DFS_STATUS)~prop(AGE.1),data=dati,max.timepoint.
  sim=NULL,resample.iid=1)
coxrm=restricted.residual.mean(out,tau=266.95,x=rbind(c(0,0,0),c(1,0,0)),c
  (0,1,0),c(0,0,1)),iid=1)
plot(coxrm)

#Modelli di regressione additivi diretti a una sola covariata (sotto assunzione
  di censura indipendente dalle covariate)
out=res.mean(Event(DFS_MONTHS,DFS_STATUS)~AGE.1,data=dati,cause=1,
  times=0,restricted= 266.95 ,n.sim=0,model="additive",estimator=1)
out=res.mean(Event(DFS_MONTHS,DFS_STATUS)~AGE,data=dati,cause=1,
  times=0,restricted= 266.95 ,n.sim=0,model="additive",estimator=1)
out=res.mean(Event(DFS_MONTHS,DFS_STATUS)~RIPK3.1,data=dati,cause=1,times=0,
  restricted= 266.95 ,n.sim=0,model="additive",estimator=1)
out=res.mean(Event(DFS_MONTHS,DFS_STATUS)~AXL.1,data=dati,cause=1,times=0,
  restricted= 266.95,n.sim=0,model="additive",estimator=1)
out=res.mean(Event(DFS_MONTHS,DFS_STATUS)~AJCC_PATHOLOGIC_TUMOR_STAGE,data=dati
  ,cause=1,times=0,restricted=266.95,n.sim=0,model="additive",estimator=1)

#Modelli di regressione multipla
-----

#Dicotomizzazione della variabile AJCC_PATHOLOGIC_TUMOR_STAGE
levels(dati$AJCC_PATHOLOGIC_TUMOR_STAGE)=c("0/I/II","0/I/II","III/IV")
```

```
#Modello di Cox
cx0=coxph(Surv(DFS_MONTHS,DFS_STATUS)~AJCC_PATHOLOGIC_TUMOR_STAGE+log(AXL)+AGE,
  data=dati)
ggcoxzph(cox.zph(cx0, transform="identity"),ggtheme=theme_bw())
ggcoxzph(cox.zph(cx0, transform="log"),ggtheme=theme_bw())
ggcoxzph(cox.zph(cx0, transform="rank"),ggtheme=theme_bw())
summary(cox.aalen(Surv(DFS_MONTHS,DFS_STATUS)~prop(AJCC_PATHOLOGIC_TUMOR_STAGE)
  +prop(log(AXL))+prop(AGE),data=dati))
#Modello di Cox stratificato
dati$STRAT[dati$AJCC_PATHOLOGIC_TUMOR_STAGE=="0/I/II" & (dati$AGE.1=="1" | dati
  $AGE.1=="2")]="A"
dati$STRAT[dati$AJCC_PATHOLOGIC_TUMOR_STAGE=="0/I/II" & (dati$AGE.1=="3" | dati
  $AGE.1=="4")]="B"
dati$STRAT[dati$AJCC_PATHOLOGIC_TUMOR_STAGE=="III/IV" & (dati$AGE.1=="1" | dati
  $AGE.1=="2")]="C"
dati$STRAT[dati$AJCC_PATHOLOGIC_TUMOR_STAGE=="III/IV" & (dati$AGE.1=="3" | dati
  $AGE.1=="4")]="D"
dati$STRAT=factor(dati$STRAT)
cx0=coxph(Surv(DFS_MONTHS,DFS_STATUS)~strata(STRAT)+log(AXL),data=dati)

#Modello A: modello additivo diretto per il restricted residual mean time (
  sotto assunzione di censura indipendente dalle covariate)
out=res.mean(Event(DFS_MONTHS,DFS_STATUS)~AGE+log(AXL)+AJCC_PATHOLOGIC_TUMOR_
  STAGE,data=dati,cause=1,times=0,restricted=266.95,n.sim=0,model="additive",
  estimator=1)

#Modello di Cox per la censura
cx0=coxph(Surv(DFS_MONTHS,(1-DFS_STATUS))~AGE+FRACTION_GENOME_ALTERED+log(MERTK
  )+SEX,data=dati)
ggcoxzph(cox.zph(cx0, transform="identity"),ggtheme=theme_bw())
ggcoxzph(cox.zph(cx0, transform="log"),ggtheme=theme_bw())
ggcoxzph(cox.zph(cx0, transform="rank"),ggtheme=theme_bw())
```

---

```
summary(cox.aalen(Surv(DFS_MONTHS,1-DFS_STATUS)~prop(log(MERTK))+prop(AGE)+prop
(FRACTION_GENOME_ALTERED)+prop(SEX),data=dati))
#Stima della funzione di sopravvivenza associata per la costruzione dei pesi
IPCW del Modello B
a=exp(-predict(cx0, type="expected"))

#Modello B: modello additivo diretto per il restricted residual mean time (
sotto assunzione di censura dipendente dalle covariate)
out=res.mean(Event(DFS_MONTHS,DFS_STATUS)~AGE+log(AXL)+AJCC_PATHOLOGIC_TUMOR_
STAGE,data=dati,cause=1,times=0,restricted=266.95,n.sim=0,model="additive",
estimator=1,cens.model="weights",cens.weights=1/a)
```

---



# Appendice B

## Codice R - Studi di simulazione

In questa appendice è contenuto il codice R utilizzato per implementare gli studi di simulazione del Capitolo 6.

---

```
#Creazione della funzione per condurre la simulazione
-----

simulazione=function(a,alfa,beta1,beta2,tau,k,n){
deltavec=rep(0,k)
mu=matrix(NA, nrow=k, ncol=4)
sd_mu=matrix(NA,nrow=k,ncol=4)
CI=matrix(NA,nrow=k,ncol=4)
mu.ind=matrix(NA,nrow=k,ncol=4)
sd_mu.ind=matrix(NA,nrow=k,ncol=4)
CI.ind=matrix(NA,nrow=k,ncol=4)
truemu.strato0=(1-exp(-tau*alfa))/(alfa)
truemu.strato1=(1-exp(-tau*alfa*exp(beta2)))/(alfa*exp(beta2))
truemu.strato2=(1-exp(-tau*alfa*exp(beta1)))/(alfa*exp(beta1))
truemu.strato3=(1-exp(-tau*alfa*exp(beta1+beta2)))/(alfa*exp(beta1+beta2))
for(i in 1:k){
if((i %% 100)==0){
print(paste("Completate ", i, " simulazioni su ", k))
```

```

}
X=rbinom(n=n,size=1,prob=0.5)
Z=rbinom(n=n,size=1,prob=0.5)
lambda=alfa*exp(X*beta1+Z*beta2)
TIME=rexp(n=n,rate=lambda)
C=runif(n=n,min=0,max=a)
Y=pmin(TIME,C)
delta=rep(0,n)
delta[Y==TIME]=1
dataset=data.frame(Y,X,Z,delta)
require(timereg)
X=factor(X)
Z=factor(Z)
XZ=c()
XZ[X=="0" & Z=="0"]="0"
XZ[X=="0" & Z=="1"]="1"
XZ[X=="1" & Z=="0"]="2"
XZ[X=="1" & Z=="1"]="3"
XZ=factor(XZ)
modello=res.mean(Event(Y,delta)~-1+XZ,data=dataset,cause=1,times=0,restricted=
  tau,n.sim=0,model="additive",estimator=1)
mu[i,1]=modello$cum[2]
sd_mu[i,1]=sqrt(modello$var.cum[2])
lowerCI0=mu[i,1]-qnorm(0.975)*sd_mu[i,1]
upperCI0=mu[i,1]+qnorm(0.975)*sd_mu[i,1]
CI[i,1]=ifelse(lowerCI0<truemu.strato0&truemu.strato0<upperCI0, 1, 0)
mu[i,2]=modello$cum[3]
sd_mu[i,2]=sqrt(modello$var.cum[3])
lowerCI1=mu[i,2]-qnorm(0.975)*sd_mu[i,2]
upperCI1=mu[i,2]+qnorm(0.975)*sd_mu[i,2]
CI[i,2]=ifelse(lowerCI1<truemu.strato1&truemu.strato1<upperCI1, 1, 0)
mu[i,3]=modello$cum[4]
sd_mu[i,3]=sqrt(modello$var.cum[4]) #stima sd strato 2 (x=1, z=0)

```

```

lowerCI2=mu[i,3]-qnorm(0.975)*sd_mu[i,3]
upperCI2=mu[i,3]+qnorm(0.975)*sd_mu[i,3]
CI[i,3]=ifelse(lowerCI2<truemu.strato2&truemu.strato2<upperCI2, 1, 0)
mu[i,4]=modello$cum[5]
sd_mu[i,4]=sqrt(modello$var.cum[5])
lowerCI3=mu[i,4]-qnorm(0.975)*sd_mu[i,4]
upperCI3=mu[i,4]+qnorm(0.975)*sd_mu[i,4]
CI[i,4]=ifelse(lowerCI3<truemu.strato3&truemu.strato3<upperCI3, 1, 0)
modello.ind=cox.aalen(Surv(Y,delta)~prop(XZ),max.timepoint.sim=NULL,resample.
  iid=1)
out.ind=restricted.residual.mean(modello.ind,tau=tau,x=rbind(c(0,0,0),c(1,0,0),
  c(0,1,0),c(0,0,1)),iid=1)
mu.ind[i,1]=out.ind$mean[1]
sd_mu.ind[i,1]=out.ind$se[1]
lowerCI0.ind=mu.ind[i,1]-qnorm(0.975)*sd_mu.ind[i,1]
upperCI0.ind=mu.ind[i,1]+qnorm(0.975)*sd_mu.ind[i,1]
CI.ind[i,1]=ifelse(lowerCI0.ind<truemu.strato0&truemu.strato0<upperCI0.ind, 1,
  0)
mu.ind[i,2]=out.ind$mean[2]
sd_mu.ind[i,2]=out.ind$se[2]
lowerCI1.ind=mu.ind[i,2]-qnorm(0.975)*sd_mu.ind[i,2]
upperCI1.ind=mu.ind[i,2]+qnorm(0.975)*sd_mu.ind[i,2]
CI.ind[i,2]=ifelse(lowerCI1.ind<truemu.strato1&truemu.strato1<upperCI1.ind, 1,
  0)
mu.ind[i,3]=out.ind$mean[3]
sd_mu.ind[i,3]=out.ind$se[3]
lowerCI2.ind=mu.ind[i,3]-qnorm(0.975)*sd_mu.ind[i,3]
upperCI2.ind=mu.ind[i,3]+qnorm(0.975)*sd_mu.ind[i,3]
CI.ind[i,3]=ifelse(lowerCI2.ind<truemu.strato2&truemu.strato2<upperCI2.ind, 1,
  0)
mu.ind[i,4]=out.ind$mean[4]
sd_mu.ind[i,4]=out.ind$se[4]
lowerCI3.ind=mu.ind[i,4]-qnorm(0.975)*sd_mu.ind[i,4]

```

```

upperCI3.ind=mu.ind[i,4]+qnorm(0.975)*sd_mu.ind[i,4]
CI.ind[i,4]=ifelse(lowerCI3.ind<truemu.strato3&truemu.strato3<upperCI3.ind, 1,
  0)
deltavec[i]=mean(delta)
}
print("STIME DIRETTE")
print(paste("Media delle k stime di mu – Strato 0:", mean(mu[,1])))
print(paste("Standard error empirico – Strato 0:", sd(mu[,1])))
print(paste("Media dei k standard error stimati – Strato 0:", mean(sd_mu[,1])))
print(paste("Rapporto – Strato 0:", sd(mu[,1])/mean(sd_mu[,1])))
print(paste("95% CP – Strato 0:", mean(CI[,1])))
print("_____")
print(paste("Media delle k stime di mu – Strato 1:", mean(mu[,2])))
print(paste("Standard error empirico – Strato 1:", sd(mu[,2])))
print(paste("Media dei k standard error stimati – Strato 1:", mean(sd_mu[,2])))
print(paste("Rapporto – Strato 1:", sd(mu[,2])/mean(sd_mu[,2])))
print(paste("95% CP – Strato 1:", mean(CI[,2])))
print("_____")
print(paste("Media delle k stime di mu – Strato 2:", mean(mu[,3])))
print(paste("Standard error empirico – Strato 2:", sd(mu[,3])))
print(paste("Media dei k standard error stimati – Strato 2:", mean(sd_mu[,3])))
print(paste("Rapporto – Strato 2:", sd(mu[,3])/mean(sd_mu[,3])))
print(paste("95% CP – Strato 2:", mean(CI[,3])))
print("_____")
print(paste("Media delle k stime di mu – Strato 3:", mean(mu[,4])))
print(paste("Standard error empirico – Strato 3:", sd(mu[,4])))
print(paste("Media dei k standard error stimati – Strato 3:", mean(sd_mu[,4])))
print(paste("Rapporto – Strato 3:", sd(mu[,4])/mean(sd_mu[,4])))
print(paste("95% CP – Strato 3:", mean(CI[,4])))
print("_____")
print("STIME INDIRETTE")
print(paste("Media delle k stime di mu – Strato 0:", mean(mu.ind[,1])))
print(paste("Standard error empirico – Strato 0:", sd(mu.ind[,1])))

```

```
print(paste("Media dei k standard error stimati – Strato 0:", mean(sd_mu.ind
  [,1])))
print(paste("Rapporto – Strato 0:", sd(mu.ind[,1])/mean(sd_mu.ind[,1])))
print(paste("95% CP – Strato 0:", mean(CI.ind[,1])))
print("-----")
print(paste("Media delle k stime di mu – Strato 1:", mean(mu.ind[,2])))
print(paste("Standard error empirico – Strato 1:", sd(mu.ind[,2])))
print(paste("Media dei k standard error stimati – Strato 1:", mean(sd_mu.ind
  [,2])))
print(paste("Rapporto – Strato 1:", sd(mu.ind[,2])/mean(sd_mu.ind[,2])))
print(paste("95% CP – Strato 1:", mean(CI.ind[,2])))
print("-----")
print(paste("Media delle k stime di mu – Strato 2:", mean(mu.ind[,3])))
print(paste("Standard error empirico – Strato 2:", sd(mu.ind[,3])))
print(paste("Media dei k standard error stimati – Strato 2:", mean(sd_mu.ind
  [,3])))
print(paste("Rapporto – Strato 2:", sd(mu.ind[,3])/mean(sd_mu.ind[,3])))
print(paste("95% CP – Strato 2:", mean(CI.ind[,3])))
print("-----")
print(paste("Media delle k stime di mu – Strato 3:", mean(mu.ind[,4])))
print(paste("Standard error empirico – Strato 3:", sd(mu.ind[,4])))
print(paste("Media dei k standard error stimati – Strato 3:", mean(sd_mu.ind
  [,4])))
print(paste("Rapporto – Strato 3:", sd(mu.ind[,4])/mean(sd_mu.ind[,4])))
print(paste("95% CP – Strato 3:", mean(CI.ind[,4])))
print("-----")
print(paste("Percentuale media di censurati:", 1-mean(deltavec)))
}

#Risultati
-----

set.seed(123)
```

```
#Risultati Tabella 6.1
```

```
simulazione(a = 116, alfa = 0.05, beta1 = 1, beta2 = -1, tau = 10, k = 2000, n = 200)
```

```
simulazione(a = 70, alfa = 0.05, beta1 = 1, beta2 = -1, tau = 10, k = 2000, n = 200)
```

```
#Risultati Tabella 6.2
```

```
simulazione(a = 116, alfa = 0.05, beta1 = 1, beta2 = -1, tau = 10, k = 2000, n = 400)
```

```
simulazione(a = 70, alfa = 0.05, beta1 = 1, beta2 = -1, tau = 10, k = 2000, n = 400)
```

---