UNIVERSITÀ DEGLI STUDI DI PADOVA

Facoltà di Scienze Statistiche

Corso di Laurea in Statistica e tecnologie informatiche

Analisi statistiche in alcuni casi di studio per il Centro Oncoematologico Pediatrico

RELATORE
Prof. Laura Ventura

CORRELATORE
Dott. Gloria Tridello

TESI DI LAUREA DI
Michele Santacatterina
Matr. N. 557723

Anno Accademico 2008/2009

Dedicato a mio Padre.

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Introduzione

Questa tesi è il risultato di una esperienza di stage presso il Centro Oncoematologico Pediatrico di Padova. L'elaborato discute i risultati di analisi statistiche effettuate su due diversi insiemi di dati. Il primo dataset riguarda pazienti che hanno subito un trapianto di cellule staminali ematopoietiche (TCSE), mentre il secondo riguarda pazienti affetti da linfoma anaplastico a grandi cellule (ALCL). Entrambe le situazioni fanno riferimento a pazienti malati di tumore, ma le analisi svolte si possono estendere, con opportuni adattamenti, anche ad altri tipi di malattie.

Lo schema della tesi è il seguente. Nel Capitolo 1 viene discussa un'analisi di sopravvivenza, affrontando lo studio sia dal punto di vista dei trapianti, ovvero se l'esito del singolo trapianto risulta rilevante, sia da quello dei soggetti, nel quale si valuta la sopravvivenza del paziente nel complesso del periodo di osservazione. In entrambi gli studi ci si sofferma su di un fattore eziologico il quale, teoricamente, non rispetta l'assunto di proporzionalità. Nel Capitolo 2 viene considerato un modello *stress-strength* per valutare la quantità $R = P\{X < Y\}$, con X e Y variabili casuali esponenziali indipendenti. Questo modello serve per vedere se l'espressione della proteina Hsp70 legata alle cellule NPM-ALK+ in pazienti affetti da ALCL sono differenti o meno da quelle di un gruppo di controllo.

Capitolo 2

Analisi dei dati su TCSE

2.1 Considerazioni preliminari

In questo capitolo sono presentati i risultati di uno studio follow-up condotto dal Centro Oncoematologico Pediatrico di Padova, allo scopo di studiare i dati relativi ad un gruppo di soggetti sottoposti ad uno o più trapianti di cellule staminali ematopoietiche. Lo studio comprende 284 soggetti (314 trapianti; alcuni soggetti sono stati sottoposti a più trapianti) seguiti nel periodo tra il 1983 e il 2002. Come primo passo si presentano i momenti principali dello studio:

- Momento iniziale: Data TCSE[1];
- Evento d'interesse: Data morte soggetto / Data ultimo controllo per i vivi.

Poichè l'esperimento si basa sul tempo intercorrente tra il TCSE e la morte, i pazienti che non presentano l'evento morte entro la fine dello studio vengono trattati come dati censurati (vedi Klein e Moeschberger, 2003, p.63).

2.1.1 La malattia

Quando viene effettuato il TCSE?

Il Trapianto di Cellule Staminali Emopoietiche è diventato il trattamento di scelta per numerosi disordini ematologici neoplastici e non neoplastici, tumori solidi, errori congeniti del metabolismo ed immunodeficienze primitive. Il primo trapianto di cellule staminali emopoietiche è stato effettuato con successo ormai più di trenta anni fa in un bambino affetto da immunodeficienza severa combinata (SCID). Negli anni successivi il trapianto di cellule staminali è stato utilizzato nel trattamento di numerosissime altre patologie sia neoplastiche che non neoplastiche con risultati sempre migliori e offrendo, in molti casi, una straordinaria possibilità di guarigione in patologie altrimenti letali.

[1] Trapianto di Cellule Staminali Ematopoietiche

Cosa sono le cellule staminali emopoietiche?

Il midollo osseo è l'organo contenuto in tutte le ossa del corpo ed ha il compito di formare nuove cellule sanguigne (globuli rossi, globuli bianchi, piastrine) in sostituzione di quelle che muoiono naturalmente e terminano la loro funzione (emopoiesi). La cellula staminale emopoietica, all'interno del midollo osseo, è una cellula non ancora differenziata, pluripotente, capostipite di tutti gli elementi fondamentali del sangue: globuli rossi, globuli bianchi e piastrine. Si tratta di un tipo di cellula in grado di proliferare mantenendo intatta la potenzialità di replicarsi: è capace infatti di riprodurre se stessa e, contemporaneamente, produrre cellule figlie che, attraverso successivi processi di differenziazione e maturazione, daranno origine agli elementi maturi. Queste cellule staminali sono contenute in prevalenza nell'interno del midollo osseo, ma possono essere presenti anche nel sangue periferico, quando viene effettuata opportuna stimolazione farmacologica, e nel sangue del cordone ombelicale al momento della nascita. Da quanto detto, per quanto riguarda il trapianto, è più esatto il termine "trapianto di cellule staminali emopoietiche" che non quello di "trapianto di midollo osseo".

Come viene effettuato il trapianto?

Obiettivo del TCSE è fornire al ricevente una popolazione di cellule staminali sane che si differenzino in cellule ematiche per rimpiazzare gli elementi cellulari deficitari e/o patologici dell'ospite.

In quanto tempo si evidenziano gli effetti di un trapianto?

Le cellule staminali emopoietiche (CSE) presenti nel midollo osseo donato riescono a trovare da sole la strada per raggiungere la collocazione che compete loro per «iniziare a lavorare», in un periodo variabile tra le due e le tre settimane settimane in media. Dopo il trapianto, si incominciano a vedere i primi risultati, con la comparsa, nella circolazione sanguigna, di globuli bianchi con le caratteristiche nuove del donatore, e successivamente anche delle altre cellule del sangue (globuli rossi e piastrine). Ma il trapianto allogenico non si conclude con l'attecchimento delle CSE del donatore; è infatti necessario che le cellule del donatore si adattino nell'organismo del ricevente e convivano senza reagire dal punto di vista immunologico. Perché ciò avvenga sono necessari dei mesi e in alcuni rari casi degli anni.

Quali sono le possibili complicanze?

Le complicanze precoci comprendono: il rigetto delle cellule trapiantate da parte dell'ospite, la malattia del trapianto contro l'ospite (o reazione del trapianto verso l'ospite, GVHD - Graft versus Host Disease) le infezioni. Le complicanze tardive comprendono: la GVHD cronica, l'immunodeficienza prolungata, le recidive della

malattia di base. La GVHD è una patologia nella quale le cellule T immunologicamente competenti del donatore reagiscono contro gli antigeni di un ricevente immunologicamente depresso. Un problema fondamentale nei trapianti allogenici è costituito proprio dalla prevenzione e dal controllo della GVHD. Nonostante l'introduzione della ciclosporina nei primi anni '80 abbia enormemente ridotto sia l'incidenza sia la gravità della GVHD, essa continua a essere una delle principali cause di mortalità e di morbilità grave dopo trapianto; in alcuni casi la GVHD può insorgere più tardivamente con un decorso cronico. Dopo la somministrazione del regime di condizionamento, la conta dei globuli bianchi può impiegare da 2 a 3 settimane per tornare ai valori normali. Durante questo periodo, i pazienti sono molto suscettibili alle infezioni batteriche, fungine e virali. Anche dopo l'attecchimento del trapianto e quindi con globuli bianchi normali, i pazienti continuano a essere immunocompromessi con un rischio infettivo elevato a causa dei farmaci impiegati per trattare la GVHD. Tale rischio è maggiore per i pazienti che effettuano un trapianto da donatore non consanguineo.

2.1.2 Due tipi di studio

Una nota importante per quanto riguarda gli scopi di questo studio è quella di considerare che trapianti diversi sono stati eseguiti sullo stesso soggetto. Ha quindi una certa importanza valutare la sopravvivenza del paziente nel complesso del periodo di osservazione. Nello stesso tempo, però, è rilevante anche l'esito del singolo trapianto (es. mortalità a 100 gg dal trapianto). Interessano quindi anche i singoli trapianti, considerando il fatto che in alcuni casi sono riferiti allo stesso soggetto. Questa nota porta alla scelta di due strade per quanto riguarda lo studio: una riferita principalmente ai trapianti e l'altra ai soggetti. La modifica sostanziale sta nella creazione del database BMT, il quale nel primo tipo di studio il soggetto che ha subito, per esempio, due trapianti è ripetuto due volte, una per trapianto ricevuto. Quindi si ha una univocità per quanto riguarda il trapianto, ma non per il soggetto. Nell'altra si ha un'univocità per soggetto, il quale compare una unica volta. In uno si studia il tempo di sopravvivenza per trapianto nell'altro il tempo di sopravvivenza globale per soggetto considerando il numero di trapianti come una covariata (cfr. Figura 2.1).

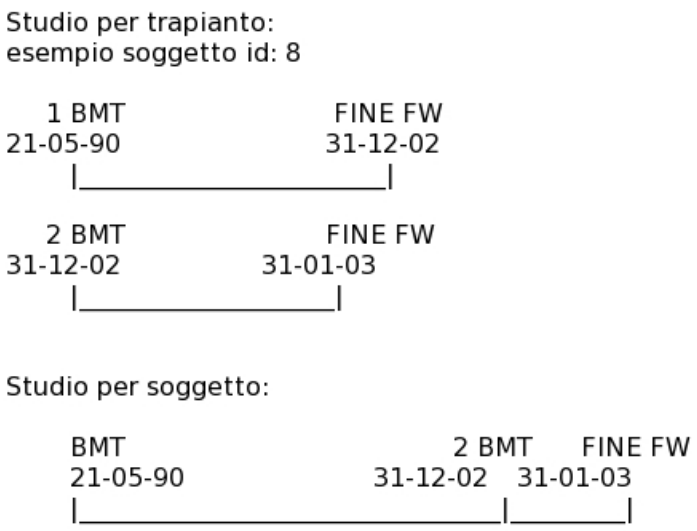


Figure 2.1: Descrizione grafica dei due diversi tipi di studio.

Nella Sezione 2.4 sarà trattato lo studio per trapianto, passando nella Sezione 2.5 a quello per soggetti.

2.1.3 Scopi

Gli scopi dello studio sono molteplici e riguardano:

1. studio dei dati e delle singole variabili contenute nel database;
2. considerazioni principali verso l'utilizzo di un modello di Cox rispetto ad altri tipi di modelli;
3. inferenza sui coefficienti del modello di Cox selezionato attraverso la procedura di selezione forward;
4. stima della funzione di sopravvivenza attraverso l'uso del metodo di Kaplan-Meier;
5. verifica dell'assunzione di rischio proporzionale per l'applicazione del modello di Cox;
6. stima e disegno della funzione di sopravvivenza attraverso il modello di Cox stratificato;
7. verifica bontà del modello attraverso residui di Cox-Snell e di devianza;
8. studio e applicazione del modello di Cox e relative conclusioni.

2.2 Dati e variabili

Nella Tabella 2.1 sono riportate le variabili che compongono il database BMT, una loro breve descrizione e la loro natura.

Variabile	Descrizione	Natura
ID_PZ	id paziente/soggetto sottoposto al TCSE	quantitativa
n_bmt	numero di TCSE effettuati dal paziente	quantitativa
VOD	infezione dopo TCSE 0=no infezione, 1=si infezione	dicotomica
durata_vod	durata VOD	quantitativa
SESSO	genere del paziente F=femmina, M=maschio	qualitativa
ETA	età del paziente alla morte	quantitativa
donor	tipo di donatore related=relazionato (es: fratello, sorella o parente) unrelated=non relazionato	dicotomica
FONTE	fonte TCSE ALLO=allogenico, APLO=aplo-identico, AUTO=autologo, CORD=da cordone ombelicale, MUD=Matched Unrelated Donor	qualitativa
morto	indicatore fallimento 0=censura, 1=paziente morto	dicotomica
PROF..VOD	tipo di profilassi VOD Eparina=profilassi Eparina, NO=profilassi NO, Pentossi=profilassi Pentossi, PGE=profilassi PGE	qualitativa
TEMPO	tempo tra momento iniziale ed evento d'interesse	quantitativa

Tabella 2.1: Tabella variabili database BMT

La variabile risposta è il tempo di sopravvivenza. La variabile *morto* dice se il dato è censurato o meno, mentre tutte le altre risultano essere variabili esplicative utili o meno allo studio. Dalla Tabella 2.1 si vede che nel database BMT si hanno alcune variabili qualitative altre quantitative.

2.2.1 La variabile VOD

La variabile VOD spiega la presenza o meno di un infezione dopo il TCSE. Particolare attenzione sarà dedicata a tale variabile, poichè potrebbe presentare una condizione di non verificabilità dell'assunto di proporzionalità, quindi un'assunzione di variabile tempo-dipendente. In questo caso, si vedrà il modello più adatto da utilizzare dopo aver verificato se questa variabile presenta una dipendenza dal tempo. Questa variabile è dicotomica. Il paziente avente TCSE in data gg-mm-aaaa ha, come valore iniziale di VOD(t), lo 0. Nel caso e solamente nel momento in cui c'è infezione allora si ha che VOD(t) passa ad 1, rimanendo costante fino alla fine del follow-up o alla morte del paziente. Quando non si ha mai infezione allora VOD(t) resta uguale a 0 (si veda Figura 2.2).

Figure 2.2: Rappresentazione variabile VOD(t)

2.3 Nozioni e terminologia statistica

Prima di entrare in merito allo studio di questi dati si presentano le principali nozioni relative all'analisi dei dati di sopravvivenza utilizzate per arrivare alle concluzioni presentate a fine capitolo (vedi Clark, et al., 2003).

2.3.1 Requisiti chiave

In questo paragrafo si presentano i principali requisiti chiave dell'analisi di sopravvivenza, i quali sono:

- la determinazione della variabile risposta;
- lo schema di censura utilizzato per i dati in studio;
- la completezza dei dati.

La variabile risposta risulta essere il tempo che passa finchè non accade un evento. Con evento si intende la morte, una ricaduta, un'infezione o altro. Solitamente nell'analisi di sopravvivenza ci si riferisce al "tempo di sopravvivenza", poichè è dato dal tempo che un individuo è sopravvissuto lungo un periodo di osservazione.

Per il tipo di dati trattati è opportuno specificare il fatto che esiste sempre la possibilità che una unità statistica venga "persa" prima dell'evento finale. Questo accade quando si ha informazione riguardo il tempo di sopravvivenza dell'individuo, ma non lo sa in modo esatto e preciso. Quindi la censura avviene proprio dal fatto che non si conosce il tempo di sopravvivenza esatto. Gli schemi di censura più ricorrenti, in pratica, possono essere raggruppati sostanzialmente in tre classi (vedi Klein e Moeschberger, 2003, p.64):

1. Censura 1° tipo: i soggetti sono osservati per un periodo di tempo fissato. Alla fine dello studio i soggetti che non presentano fallimento risultano censurati (es: esperimento di affidabilità);
2. Censura 2° tipo: simile al 1° tipo, ma il numero totale di fallimenti è stabilito a priori. La lunghezza dello studio quindi non risulta fissata;

3. Censura casuale: il totale del periodo di osservazione è fissato, ma i soggetti entrano in studio in tempi differenti. Alcuni individui falliscono, altri individui risultano persi dal follow-up, altri ancora non presentano fallimento alla fine dello studio (es. prova oncologica).

La completezza del dataset risulta associata al numero di dati censurati presenti nello studio. Qualsiasi stima effettuata su un campione con molti dati censurati non darà risultati attendibili. Maggiore è il numero di valori non censurati migliore saranno le stime dei coefficienti. La bontà delle stime è quindi relazionata al numero di eventi piuttosto che al numero di partecipanti.

2.3.2 Funzione di sopravvivenza

La funzione di sopravvivenza $S(t)$ fornisce la probabilità che una persona sopravviva più del tempo specificato in t; ossia fornisce la probabilità che la variabile casuale $T > t$. Questa funzione risulta essere di fondamentale importanza poichè ottenendo probabilità di sopravvivenza per differenti valori di t si possono avere informazioni riassuntive sui dati di sopravvivenza.

2.3.3 Funzione di rischio

La funzione di rischio $h(t)$ è data da

$$h(t) = lim_{\Delta t \to 0} \frac{P\left(t \leq T < t + \Delta t \mid T \geq t\right)}{\Delta t}, \tag{2.1}$$

con Δt che denota un intervallo di tempo piccolo.

In termini pratici la funzione di rischio $h(t)$ dà un *potenziale istantaneo* (vedi Kleinbaum e Klein, 2005, p.10) per unità di tempo di un determinato evento che avviene, dato che un individuo è sopravvissuto fino al tempo t. Nota che in contrasto alla funzione di sopravvivenza, la quale si focalizza sul *non* fallimento, la funzione di rischio si focalizza sul fallimento, che è l'evento accaduto.

Esiste una stretta relazione entro le due funzioni prima citate. Infatti se si conosce la forma di $S(t)$ allora si può derivare la corrispondente $h(t)$, e viceversa. La relazione tra $S(t)$ e $h(t)$ può essere espressa come

$$S(t) = exp\left(-\int_0^t h\left(u\right) du\right) \quad \Longrightarrow \quad h\left(t\right) = -\left[\frac{dS\left(t\right)/dt}{S\left(t\right)}\right]. \tag{2.2}$$

La prima formula descrive come la funzione di sopravvivenza può essere scritta in termini di un integrale che include la funzione di rischio.

2.3.4 Quale modello usare?

Una delle domande principali da porsi riguarda quale modello usare, o meglio quale modello risulterebbe migliore rispetto ad altri. A disposizione ci sono molti modelli: quelli non parametrici, quelli parametrici o semiparametrici. I modelli non-parametrici non richiedono alcuna ipotesi sulla struttura del modello da identificare. Potrebbero risultare utili in fase di analisi marginale, quando si vuole stabilire se esistono differenze tra le durate di vita dei pazienti[2]. I modelli di sopravvivenza parametrici, sono costruiti scegliendo una specifica distribuzione di probabilità per la funzione di sopravvivenza, e possono essere aggiustati per le eventuali covariate. Nei modelli semiparametrici non viene fatta nessuna ipotesi sulla funzione di sopravvivenza, e quindi sul rischio. Viene specificata solo la forma funzionale per valutare l'influenza delle covariate sui tempi di sopravvivenza e la forma dell'andamento del rischio viene lasciata non specificata.

Il modello più classico, e maggiormente usato, risulta essere il modello semiparametrico a rischi proporzionali di Cox. Si sceglie questo modello poichè:

- esiste la presenza di fattori che potrebbero influire sulla variabile risposta, e di questo bisogna tener conto;
- anche se la funzione $h_0(t)$ non è specificata si possono ottenere le stime della funzione di sopravvivenza, di rischio e di rischio cumulato, oltre quelle dei coefficienti di regressione e dei rapporti di rischio;
- l'utente non si deve preoccupare della forma specifica della distribuzione della funzione di rischio adottata;
- il modello a rischi proporzionali di Cox risulta molto popolare (vedi Kleinbaum e Klein, 2005, p.96) .

Quindi la scelta principale risulta essere quella di un modello semiparametrico a rischi proporzionali di Cox. Da notare però che per ora l'utilizzo di questo modello risulta essere solo un'ipotesi iniziale poichè non si esclude la possibilità di un'estensione o modifica di questo modello per la non verifica dell'assunto di proporzionalità da parte di alcune variabili.

2.3.5 Il modello a rischi proporzionali di Cox

In gran parte degli studi biomedici si ha la necessità di indagare e quantizzare l'effetto di alcuni fattori prognostici operanti sull'esperimento stesso. Noto il numero di variabili prognostiche, il tasso di rischio potrebbe essere espresso in funzione di tali variabili. Si passa allora all'adozione del modello parametrico, il quale presuppone che

[2]per esempio di sesso diverso

$h(t)$ e $S(t)$ possano essere espresse per mezzo di una funzione matematica nella variabile t ed eventualmente $\underline{z}$, dove $\underline{z}$ rappresenta un insieme di covariate $(z_1, z_2 \ldots, z_p)$. A differenza dei modelli parametrici, il modello di Cox divide la funzione di rischio in due quantità distinte:

- la prima, $h_0(t)$, chiamata *baseline function* che dipende solamente dal tempo;
- e la seconda che è l'espressione esponenziale della somma lineare dei $\beta_i z_i$, dove la somma è sulle p variabili esplicative, ossia

$$h(t, \underline{z}) = h_0(t) exp(\sum_i \beta_i z_i). \tag{2.3}$$

La parte esponenziale viene invocata per assicurare che il modello stimato possa sempre dare delle stime dei rischi non negative. Questo modello impone l'assunzione di rischio proporzionale, ovvero che il rapporto di rischio sia costante sul tempo, o equivalentemente, che il rischio per un individuo sia proporzionale al rischio di ogni altro individuo, dove la proporzionalità è indipendente dal tempo. Infatti se si rapportano due rischi si ha che

$$\widehat{HR} = \frac{h(t, \underline{z}^*)}{h(t, \underline{z})} = \frac{h_0(t) exp(\sum_i \beta_i z_i^*)}{h_0(t) exp(\sum_i \beta_i z_i)} = exp\left[\sum_i \beta_i (z_i^* - z_i)\right], \tag{2.4}$$

non dipende dal tempo (vedi Kleinbaum e Klein, 2005, p. 100). La quantità $\widehat{HR}$ viene chiamata *hazard ratio* (vedi Kleinbaum e Klein, 2005, p. 32).

2.3.6 Forma generale del modello a rischi proporzionali

Sia $h_0(t) = h(t, \underline{z})$ la funzione di rischio di base (vedi Kleinbaum e Klein, 2005, p. 94). Il modello prevede allora che

$$h(t, \underline{z}) = h_0(t)\, \mu\left(\underline{z}, \underline{\beta}\right), \tag{2.5}$$

dove il predittore $\mu(\cdot, \cdot)$ può assumere diverse forme. Generalmente si sceglie $\mu\left(\underline{z}, \underline{\beta}\right) = exp\left(\underline{z}'\underline{\beta}\right)$, come nella 2.3, ma può anche essere $\mu\left(\underline{z}, \underline{\beta}\right) = 1 + \underline{z}'\underline{\beta}$, oppure $\mu\left(\underline{z}, \underline{\beta}\right) = 1 + \left(\underline{z}'\underline{\beta}\right)^{-1}$. In ogni caso, l'ipotesi che sta sotto il modello è che *il rapporto tra i rischi relativi tra due soggetti distinti, caratterizzati da due diversi vettori di regressori, è costante nel tempo ovvero*

$$\frac{h_0(t, \underline{z}^*)}{h_0(t, \underline{z})} = \frac{\mu(\underline{z}^*, \beta)}{\mu(\underline{z}, \beta)}. \tag{2.6}$$

Assumere la (2.5) equivale a imporre per la funzione di sopravvivenza che

$$S\left(t, \underline{z}\right) = \left[S_0\left(t\right)\right]^{\mu\left(\underline{z}, \underline{\beta}\right)}. \tag{2.7}$$

Infatti considerando la funzione di rischio cumulato[3] $H\left(t, \underline{z}\right)$ (vedi Klein e Moeschberger, 2003, p.92), si ha che

$$S\left(t, \underline{z}\right) = exp\left\{-H\left(t, \underline{z}\right)\right\} = exp\left\{-\mu\left(\underline{z}, \underline{\beta}\right) H_o\left(t\right)\right\} = S\left(t, \underline{z}\right) = \left[S_0\left(t\right)\right]^{\mu\left(\underline{z}, \underline{\beta}\right)}. \tag{2.8}$$

Si parla anche di modello a rischi proporzionali (vedi Kleinbaum e Klein, 2005, p.107) poichè i rischi per differenti insiemi di covariate rimangono nella stessa proporzione per tutti gli istanti di tempo.

2.3.7 Selezione variabili

In molte applicazioni, quali in biostatistica, accade di dover trattare situazioni con un grande numero di variabili, delle quali, tuttavia, solo un piccolo numero è significativo per il problema. La valutazione della significatività di un coefficiente di regressione può avvenire attraverso il test di Wald (vedi Klein e Moeschberger, 2003, p.254). L'ipotesi che si vuole verificare con questo test risulta

$$\begin{cases} H_0: & \underline{\beta}_1 = \underline{\beta}_0 \\ H_1 & \underline{\beta}_1 \neq \underline{\beta}_0 \end{cases} \tag{2.9}$$

dove generalmente $\underline{\beta}_0 = 0$.

Nel test di Wald la stima di massima verosimiglianza del parametro di interesse $\underline{\beta}_1$ è confrontata con un valore proposto, $\underline{\beta}_0$. Il test per l'ipotesi (2.9) è

$$Q_W = \left(\hat{\underline{\beta}}_1 - \underline{\beta}_0\right)^t I\left(\hat{\beta}_1\right)\left(\hat{\underline{\beta}}_1 - \underline{\beta}_0\right) \tag{2.10}$$

dove $I\left(\hat{\beta}_1\right)$ è la matrice di informazione attesa (vedi Pace e Salvan, 2001, p. 139) e $\hat{\underline{\beta}}_1$ lo stimatore di massima verosimiglianza non vincolato per $\underline{\beta}_1$ (vedi Pace e Salvan, 2001, p. 132) . La (2.10) viene confrontata con una χ^2 con p gradi di libertà, dove p è il numero di componenti scalari.

Test alternativo risulta essere il test log-rapporto di verosimiglianza, il quale è dato da

[3] rapprenta il potenziale istantaneo di morte al tempo t

$$Q_{RV}\left(\underline{\beta}_0\right) = 2\left\{l\left(\underline{\hat{\beta}}_1\right) - l\left(\underline{\beta}_0\right)\right\}, \tag{2.11}$$

che sotto condizioni di regolarità, ha distribuzione asintotica χ^2 con p gradi di libertà (vedi Pace e Salvan, 2001, p. 140).

2.3.8 Stima curva di sopravvivenza attraverso il metodo di Kaplan-Meier

Come precedentemente accennato, la stima della curva di sopravvivenza si può semplicemente trovare attraverso l'uso di un modello non parametrico. Si suppone siano k gli istanti di morte distinti osservati. Si indica poi con

- τ_i l'istante di morte i-esimo;
- m_i il numero di morti osservate in τ_i;
- r_i il numero di "soggetti a rischio" a τ_i (cioè i soggetti nè morti nè persi appena prima di τ_i).

Si cerca allora uno stimatore per $S(t)$ che sia la funzione di sopravvivenza di una variabile casuale con supporto sui tempi osservati. Si sa che per una tale variabile casuale vale la relazione (vedi Kaplan e Meier, 1958)

$$S\left(t\right) = \prod_{\tau_i \leq t} \left[1 - h\left(\tau_i\right)\right], \tag{2.12}$$

dove $h\left(\tau_i\right)$ è la probabilità di morire in τ_i dato che si è a rischio a τ_i. Poichè $h\left(\tau_i\right)$ può essere stimata da $\frac{m_i}{r_i}$[4], lo stimatore per $S(t)$ cercato risulta essere

$$\hat{S}_{KM}\left(t\right) = \prod_{\tau_i \leq t} \left[1 - \frac{m_i}{r_i}\right]. \tag{2.13}$$

Tale stimatore è noto come *stimatore di Kaplan-Meier* o del *prodotto limite*. Esso rappresenta una generalizzazione al caso censurato della funzione di sopravvivenza empirica. Infatti, coincide con essa quando nel campione non sono presenti dati censurati. Il metodo del prodotto limite è un metodo puramente non parametrico che non implica la suddivisione dell'asse dei tempi in intervalli di ampiezza prefissata e quindi nemmeno il corrispondente raggruppamento di soggetti. Infatti si stima la probabilità condizionata di sopravvivenza in corrispondenza di ciascuno dei tempi in

[4] in base alla interpretazione sequenziale della vita di ogni individuo, nell'istante τ_i si osservano r_i prove di Bernoulli indipendenti con m_i morti

cui si verifica almeno un evento terminale. Può essere considerato un caso particolare della tavola di sopravvivenza nel quale ogni intervallo di tempo contiene una sola osservazione.

Risulta chiaro che, per costruzione, $\hat{S}_{KM}(t)$ è funzione a gradini con salti in corrispondenza degli istanti di morte osservati. Si noti però che lo stimatore di Kaplan-Meier può non tendere a 0 al tendere di t ad ∞. Ciò si verifica quando il più grande valore osservato $y_{(n)}$ è un dato censurato. In questo caso i dati non forniscono informazione completa per $t > y_{(n)}$. Pertanto se il più grande dato osservato è censurato, per $t > y_{(n)}$ lo stimatore di Kaplan-Meier non è definito. In sintesi

$$\hat{S}_{KM}(t) = \begin{cases} \prod_{\tau_i \leq t} \left[1 - \frac{m_i}{r_i}\right] & per\ t \leq y_{(n)} \\ 0 & per\ t > y_{(n)}\ se\ y_{(n)}\ non\ censurato \\ non\ definito & per\ t > y_{(n)}\ se\ y_{(n)}\ è\ censurato \end{cases} \tag{2.14}$$

L'ampiezza del salto, come già visto prima, corrisponde all'istante di morte τ_i ed è pari alla massa di probabilità che lo stimatore di Kaplan-Meier attribuisce a τ_i .

2.3.9 Verifica assunzione rischi proporzionali

Come detto nel paragrafo 2.3.4, il modello semiparametrico di Cox presuppone che i rischi siano proporzionali. Prima di utilizzare un modello semiparametrico di Cox, è opportuno verificare l'assunzione di proporzionalità (vedi Kleinbaum e Klein, 2005, p.131). Esistono tre approcci per verificare se una variabile risulta essere tempo-dipendente:

- analisi grafica;
- test di *Goodness Of Fit*;
- utilizzo di una variabile tempo-dipendente.

Grafico Questo approccio prevede di tracciare il logaritmo della funzione di rischio cumulata $H(t, \underline{z})$ per ogni strato della variabile in studio. Se le curve rappresentate risultano essere parallele, allora la variabile rispetta l'assunto di proporzionalità. Solo in caso di forte non parallelismo, invece, si può considerare la variabile dipendente dal tempo.

Test di *Goodness Of Fit* Per valutare l'assunzione di proporzionalità in letteratura sono stati proposti numerosi test. Un primo test è quello proposto da

Schoenfeld (1982), che è basato sui residui definiti da Schoenfeld. L'idea di base della statistica test è che se l'assunzione di proporzionalità è verificata per una particolare covariata, allora i residui di Schoenfeld per questa covariata non dovrebbero essere relazionati al tempo di sopravvivenza. Quindi se l'assunto è valido, allora i residui di Schoenfeld sono incorrelati con il tempo (vedi Kleinbaum e Klein, 2005, p.151).

Utilizzo di una variabile tempo-dipendente Quando una variabile tempo-dipendente è usata per verificare l'assunzione di proporzionalità per una determinata covariata, si ha un'estensione del modello di Cox. Il modello, dunque, contiene il prodotto tra la variabile inizialmente osservata e una funzione del tempo. L'idea base di questo metodo sta nel verificare la significatività del coefficiente del prodotto tra i due termini. L'ipotesi nulla è che il coefficiente del prodotto sia uguale a zero (vedi Kleinbaum e Klein, 2005, p.154).

2.3.10 Variabile tempo-dipendente: come trattarla?

Uno degli assunti di base del modello di Cox è che si hanno rischi proporzionali. Tuttavia, è possibile che diversi livelli di un dato fattore qualitativo producano consistenti scostamenti dalla situazione di proporzionalità. Scostamenti di questo tipo portano all'ipotesi di presenza nel modello di variabili dipendenti dal tempo. Per questo tipo di problema già da tempo in letteratura sono state esposte parecchie soluzioni (vedi Cox e Oakes, 1984, p.112), tra i quali quella di utilizzare un modello di Cox stratificato.

In questo caso il modello può essere aggiustato considerando la stratificazione dei dati in sottogruppi, ognuno dei quali è identificato da un livello del suddetto fattore. Si suppone che tali livelli siano B. Allora il modello di Cox stratificato assume

$$h_b\left(t \mid \underline{z}\right) = h_{0b}\left(t\right) \mu\left(\underline{z}, \underline{\beta}\right) \qquad con \qquad b = 1, .., B \tag{2.15}$$

Individui appartenenti a sottogruppi diversi possono avere, invece, rischi non proporzionali, potendo essere differenti e non proporzionali le funzioni di rischio di base $h_{01}\left(t\right), h_{02}\left(t\right), ..., h_{0B}\left(t\right)$ relative a ciascun strato.

La stratificazione, e quindi l'utilizzo di un modello di Cox stratificato, è una modifica del modello di Cox a rischi proporzionali che permette il controllo attraverso la "stratificazione" dei predittori che non soddisfano l'assunzione di proporzionalità. Si presume che i predittori che soddisfano l'assunzione di proporzionalità vengano inseriti nella parte parametrica del modello, mentre quelli che non la verificano ne siano esclusi (vedi Kleinbaum e Klein, 2005, p.173).

2.4 Studio per trapianti

In questo paragrafo si presenta lo studio riferito ai trapianti, ovvero se l'esito del singolo trapianto risulta rilevante.

Come detto nel paragrafo 2.3.1, i principali requisiti chiave per un'analisi di sopravvivenza sono:

1. la determinazione della variabile risposta;
2. lo schema di censura utilizzato per i dati in esame;
3. la completezza dei dati.

Per quanto riguarda il dataset TCSE:

1. la variabile risposta risulta il tempo di sopravvivenza di pazienti sottoposti a TCSE, quindi il tempo trascorso dal momento del trapianto al loro eventuale fallimento.
2. lo schema di censura utilizzato risulta essere quello di tipo casuale, che è quello che si adatta meglio al tipo di studio utilizzato per la raccolta dei dati.
3. su un totale di 314 trapianti, più della metà presentano uno stato di censura. Precisamente 124 pazienti hanno contribuito come dato "completo" mentre i restanti no.

2.4.1 Analisi esplorative

Obiettivi di una analisi esplorativa dei dati sono la ricerca di possibili fattori che influenzano il fenomeno in esame. In questo caso, l'analisi esplorativa è usata per vedere, attraverso rappresentazioni grafiche, se esistono possibili associazioni tra le covariate e la variabile di risposta.

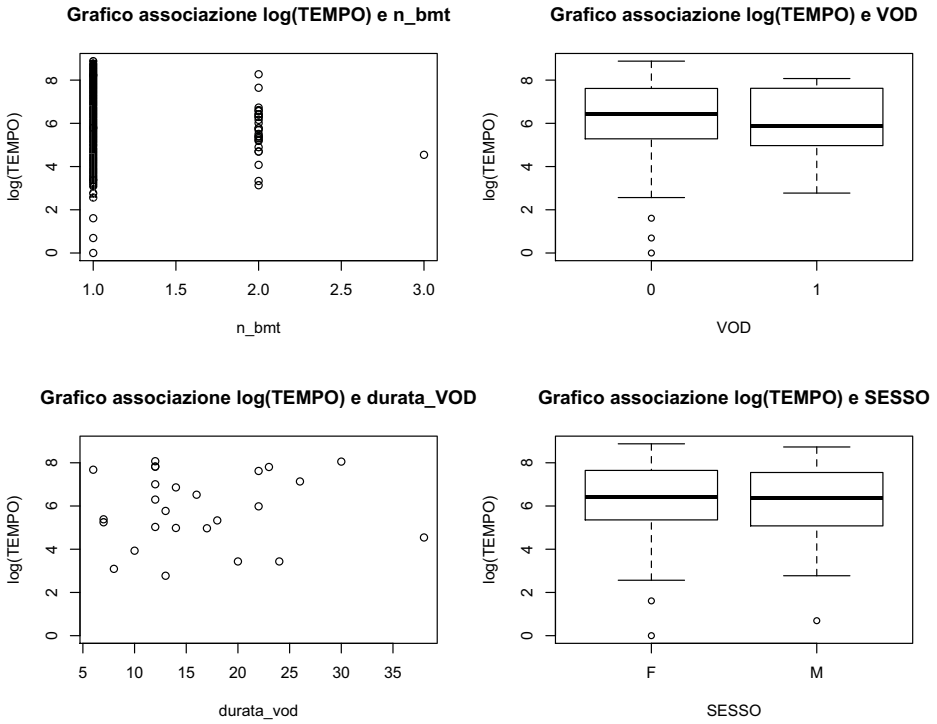


Figura 2.3: Analisi esplorativa studio per trapianti (a)

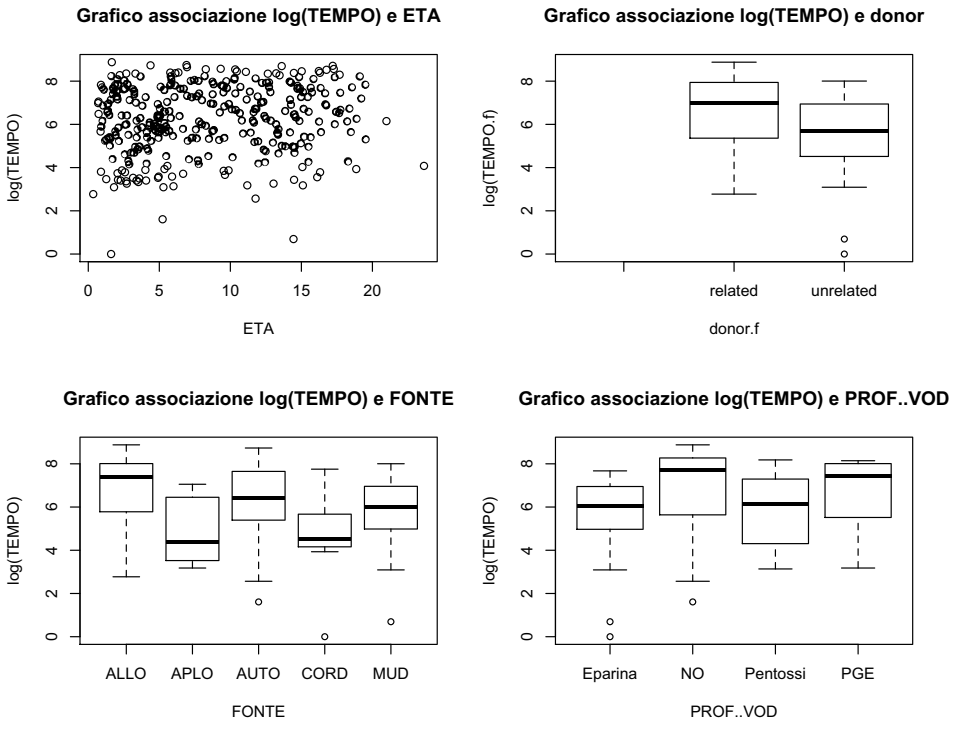


Figura 2.4: Analisi esplorativa studio per trapianti (b)

Si può notare che, anche se solo un paziente presenta tre trapianti, il tempo di sopravvivenza appare inversamente proporzionale al numero di trapianti. I pazienti affetti da infezione (VOD=1) sembrano avere sopravvivenza minore rispetto a quel-

li non affetti. Per quanto riguarda la durata dell'infezione, non sembra esserci un particolare andamento. Stesso discorso per quanto riguarda l'età. I tempi di sopravvivenza presentati dal gruppo maschi risultano più o meno equivalenti a quelli del gruppo femmine. Il grafico dei donatori relazionati sembra presentare una sopravvivenza maggiore di quelli non relazionati. La fonte del TCSE presenta differenze tra i vari gruppi mostrando vita maggiore per il gruppo appartenente al TCSE allogenico e peggiore per quello del TCSE aplo-identico. Anche il grafico della profilassi VOD mostra differenze tra i gruppi dei pazienti. Il gruppo PROF..VOD=NO presenta tempi di sopravvivenza maggiori degli altri, mentre i gruppi profilassi eparina e pentossi, presentano sopravvivenza minore. Si è potuto vedere come alcuni fattori sembrano influenzare più di altri i tempi di sopravvivenza. Conclusioni più precise si potranno ottenere dopo l'utilizzo di un modello semiparametrico di Cox.

2.4.2 Stima curva di sopravvivenza con il metodo di Kaplan-Meier

Prima di utilizzare il modello semiparametrico di Cox, è possibile valutare attraverso l'utilizzo del metodo di Kaplan-Meier, la differenza tra le curve di sopravvivenza stratificando per VOD. Questo permette di valutare possibili differenze grafiche tra i vari strati di una variabile. Quello che ci si aspetta è osservare un tempo di sopravvivenza minore da parte di chi ha presentato l'infezione. E' importante notare che il numero di soggetti che hanno presentato infezione risulta essere molto minore di quello che non l'hanno presentata. La Figura 2.5 rappresenta le due curve stimate.

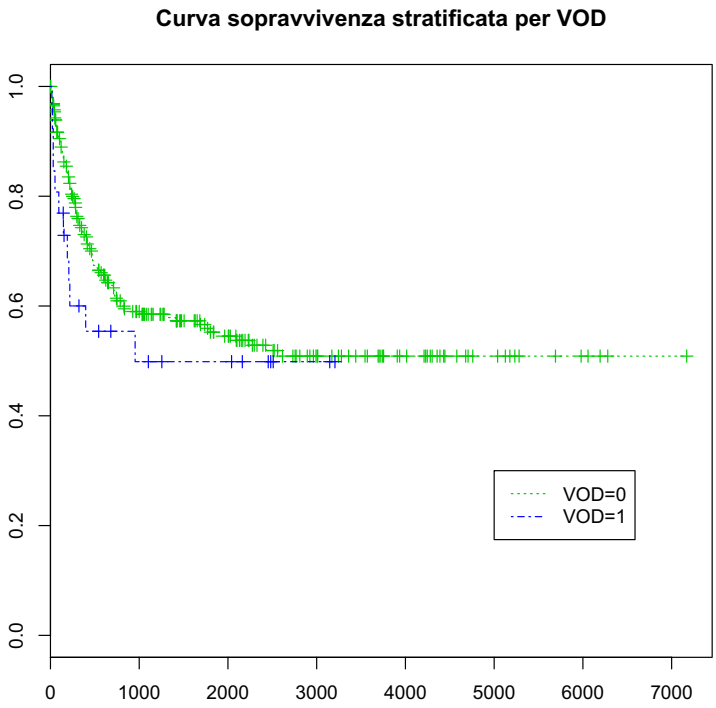


Figura 2.5: Curva di sopravvivenza stratificata per VOD

La curva di sopravvivenza, e quindi i tempi di sopravvivenza, degli individui con presenza di infezione risulta essere minore di quella dei soggetti senza. Le due curve non sembrano presentare punti incidenti, quindi sembra che la differenza tra i due livelli sia sempre presente. Il grafico lo si può interpretare nel seguente modo: per quanto riguarda i pazienti non affetti da infezione la probabilità stimata di sopravvivenza risulta essere, al tempo 1000 giorni, pari a 0.6, a differenza dei pazienti infetti per i quali risulta 0.5. Si nota inoltre che il valore minimo di probabilità attribuito ai due gruppi risulta circa 0.5, questa prova pratica afferma quanto detto nel paragrafo 2.3.8 che la curva di sopravvivenza non tende a 0 al tendere di t ad ∞. Inoltre si può vedere dal grafico la notevole quantità di valori censurati[5].

2.4.3 Selezione variabili

La selezione delle variabili è un'operazione importante per capire quali variabili risultano significative per il problema e quali no. Si ipotizza, comunque, che le scelte effettuate per l'entrata o l'uscita delle variabili che fanno parte dello studio siano basate su precise scelte suggerite dall'esperienza del ricercatore, e quindi non solo da una scelta di significatività statistica. Per quanto riguarda la loro scelta, viene utilizzata una procedura di selezione "in avanti" (forward) poichè data la notevole quantità di variabili qualitative, partire da un modello completo potrebbe risultare

[5]rappresentati nella curva da una croce

complesso da un punto di vista computazionale. La verifica della significatività di ogni variabile avviene attraverso il test di Wald (cfr. paragrafo 2.3.7). Il modello a cui si perviene contiene i seguenti regressori:

- n_bmt;
- FONTE;
- PROF..VOD;
- VOD.

La principale variabile in esame, ossia VOD (cfr. paragrafo 2.2.1) risulta essere non significativa (*p-value*=0.23). La variabile viene, comunque, inserita nel modello solo per un fattore di studio[6], consapevoli del fatto che non risulta essere significativa. Dopo essere pervenuti al modello con i regressori mostrati, si presentano i valori:

- dei coefficienti delle variabili;
- degli HR (cfr. paragrafo 2.3.5);
- dell'errore standard;
- della statistica test di Wald;
- del *p-value* della statistica test di Wald;

ottenuti dall'ambiente statistico di lavoro.

```
Call: coxph(formula = Surv(TEMPO, morto) ~
n_bmt + FONTE + PROF..VOD +      VOD)
  n= 312

                    coef      exp(coef)se(coef)   z Pr(>|z|)
n_bmt              0.9459    2.5752   0.2730 3.464 0.000531 ***
FONTEAPLO          2.4067   11.0970   0.4347 5.537 3.08e-08 ***
FONTEAUTO          0.5992    1.8208   0.2583 2.320 0.020325 *
FONTECORD          1.5996    4.9513   0.5160 3.100 0.001936 **
FONTEMUD           0.9904    2.6923   0.3731 2.654 0.007948 **
PROF..VODNO        0.6119    1.8440   0.2526 2.422 0.015429 *
PROF..VODPentossi  0.9222    2.5149   0.6170 1.495 0.134991
PROF..VODPGE       0.8416    2.3200   0.3055 2.755 0.005873 **
VOD1               0.3686    1.4457   0.3083 1.196 0.231814
```

[6]per cosi affrontare la problematica di una variabile dipendente dal tempo

Tutti i coefficienti tranne un livello della variabile PROF..VOD e VOD (come spiegato prima) risultano significativi contro l'ipotesi nulla. Come visto nella 2.4 del paragrafo 2.3.5, l'*hazard ratio* definisce il rischio di ogni individuo diviso per il rischio di un individuo diverso.

Si può notare come il livello del trapianto allogenico sia quello con maggiore *hazard ratio*, allora:

- il soggetto che presenta TCSE aplo-identico, ha 11 volte e mezza in più di probabilità di presentare l'evento morte di un paziente che non lo presenta;
- il soggetto che presenta TCSE da cordone ombelicale, ha 5 volte in più di probabilità di presentare l'evento morte di un paziente che non lo presenta;
- il soggetto che presenta due TCSE ha 5 volte in più, e chi ha tre TCSE 8 volte, di probabilità di presentare l'evento morte di un paziente che non lo presenta;

e cosi via.

2.4.4 Verifica assunzione rischi proporzionali

Si applicano i tre metodi prima descritti per vedere se le variabili presenti nel modello di Cox verificano l'assunto di proporzionalità. Il primo metodo utilizzato è quello grafico e le variabili verificate sono il numero di TCSE, la fonte del trapianto, la profilassi dell'infezione VOD e l'infezione VOD.

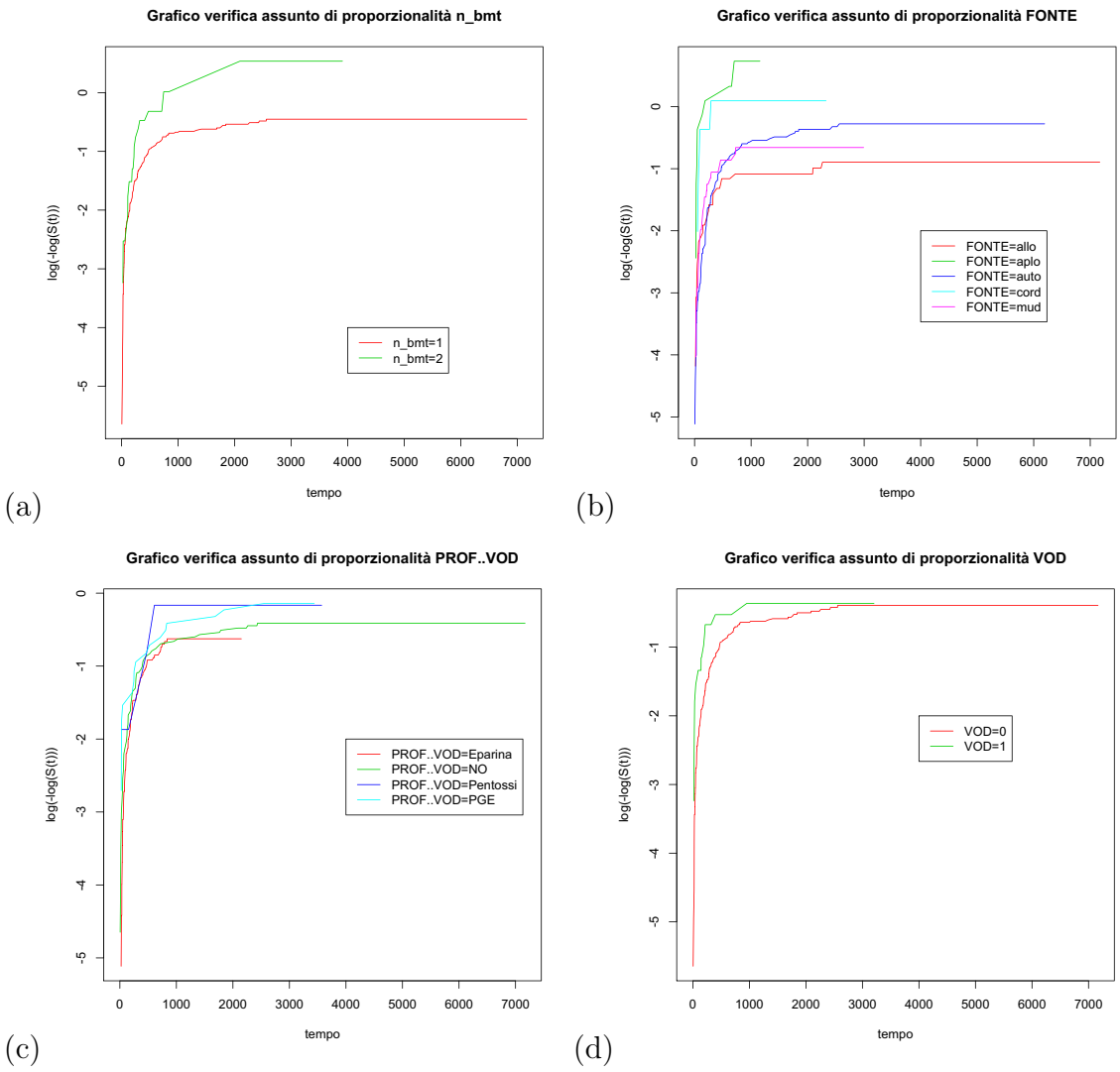


Figura 2.6: Verifica grafica assunto di proporzionalità per n_bmt, FONTE, PROF..VOD e VOD

L'assunto di proporzionalità è verificato quando si è in presenza di parallelismo tra le curve rappresentate. In questo caso, per quanto riguarda la variabile n_bmt l'assunto pare verificato anche se il gruppo con 3 trapianti non è considerato poichè composto da un solo paziente. Inoltre, la quantità del gruppo n_bmt=2 non assicura una stima della curva di sopravvivenza buona, dato il numero limitato di valori. Per quanto riguarda l'interpretazione degli altri grafici l'assunzione non sembra essere valida. In effetti si possono notare alcuni punti incidenti tra le varie curve. Si vuole far notare come però alcuni di questi gruppi presentano un numero limitato di soggetti, riducendo di conseguenza la bontà della stima della curva di sopravvivenza. Forse, quindi, una rappresentazione non corretta delle curve dovute alla scarsità di elementi del gruppo, potrebbe portare alla decisione di non validità delle variabili fonte del TCSE e profilassi di VOD. Per quanto riguarda la variabile VOD un perfetto parallelismo non esiste però non sembrano esserci punti di incidenza. Questo potrebbe far pensare, dunque, ad un'eventuale verificabilità dell'assunto della variabile VOD. Già da questi grafici si può intuire la difficoltà di interpretazione del

metodo grafico, il quale risulta essere tanto intuitivo quanto fuorviante. Per ovviare a questa problematica di decisione si passa ad un test di *Goodness Of Fit*.

Utilizzando il modello trovato attraverso la selezione effettuata nella Sezione 2.4.3, e una specifica funzione di R per calcolare il test di *Goodness Of Fit*, si verifica se la variabile può essere considerata o meno una variabile tempo-dipendente , e quindi, verificare l'assunto di proporzionalità. Il risultato del test restituisce:

```
                       rho  chisq       p
n_bmt               0.0731  0.712 0.39887
FONTEAPLO          -0.0635  0.553 0.45709
FONTEAUTO           0.1832  4.258 0.03906
FONTECORD          -0.0778  0.705 0.40098
FONTEMUD           -0.0476  0.294 0.58742
PROF..VODNO        -0.1770  4.122 0.04233
PROF..VODPentossi  -0.0530  0.363 0.54709
PROF..VODPGE       -0.0761  0.767 0.38126
VOD1               -0.2412  7.145 0.00752
```

L'unico parametro che si presenta significativo contro l'ipotesi nulla è VOD.

Ipotizzando che le altre variabili (n_bmt, FONTE e PROF..VOD) verifichino l'assunto di proporzionalità, e supponendo che solo VOD non lo faccia, la costruzione del modello, per verificare se VOD è una variabile tempo-dipendente, risulta la seguente:

$$h\left(t, \underline{z}(t)\right) = h_0(t) exp\left[\sum_{i=1}^{p} \beta_i z_i + \delta VOD \cdot g\left(t\right)\right] \qquad (2.16)$$

con

- β_i coefficiente del regressore i-esimo;
- z_i regressore i-esimo;
- p numero di regressori presenti nel modello;
- δ coefficiente di VOD;
- $g\left(t\right)$ funzione che dipende dal tempo.

Considerando $g\left(t\right) = t$ si presenta di seguito l'adattamento del modello (2.16).

```
Call: coxph(formula = Surv(TEMPO, morto == 1) ~ n_bmt + FONTE
+ PROF..VOD + VOD + VOD_T)
```

```
 n= 312
                coef      exp(coef) se(coef)      z Pr(>|z|)
n_bmt         0.962325  2.617776  0.276778  3.477 0.000507 ***
FONTEAPLO     2.497109 12.147330  0.441624  5.654 1.56e-08 ***
FONTEAUTO     0.632081  1.881522  0.261301  2.419 0.015564 *
FONTECORD     1.663183  5.276079  0.515727  3.225 0.001260 **
FONTEMUD      1.025088  2.787341  0.379375  2.702 0.006892 **
PROF..VODNO   0.636954  1.890712  0.252724  2.520 0.011724 *
PROF..VODPen 0.692474  1.998654  0.623244  1.111 0.266534
PROF..VODPGE 0.817370  2.264537  0.306411  2.668 0.007640 **
VOD           3.113785 22.506062  0.528732  5.889 3.88e-09 ***
VOD_T        -0.005715  0.994301  0.001982 -2.883 0.003941 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Considerando invece $g(t) = log(t)$ risulta:

```
Call: coxph(formula = Surv(TEMPO, morto == 1) ~ n_bmt + FONTE
+ PROF..VOD + VOD + VOD_LOGT)
 n= 312
                coef       exp(coef)  se(coef)    z Pr(>|z|)
n_bmt         1.0101     2.7460     0.2747  3.677 0.000236 ***
FONTEAPLO     2.4576    11.6764     0.4403  5.581 2.39e-08 ***
FONTEAUTO     0.5849     1.7948     0.2597  2.252 0.024296 *
FONTECORD     1.6095     5.0002     0.5149  3.125 0.001775 **
FONTEMUD      0.9661     2.6276     0.3778  2.557 0.010561 *
PROF..VODNO   0.6047     1.8307     0.2560  2.362 0.018169 *
PROF..VODPen 0.7781     2.1773     0.6205  1.254 0.209848
PROF..VODPGE 0.7251     2.0650     0.3154  2.299 0.021508 *
VOD           9.3005 10943.2604     1.3360  6.962 3.37e-12 ***
VOD_LOGT     -1.5132     0.2202     0.2651 -5.707 1.15e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In entrambi i casi la variabile moltiplicata per il tempo, quindi la variabile tempo-dipendente, risulta essere significativa contro l'ipotesi nulla. In conclusione si può dire che si considera la variabile VOD una variabile tempo-dipendente. Quindi per quanto visto VOD risulta l'unica variabile tempo-dipendente presente nel modello.

2.4.5 Stima curva di sopravvivenza modello di Cox stratificato

Come visto nel precedente paragrafo, si ha che la variabile VOD non verifica l'assunto di proporzionalità. Ciò vuol dire che se si applica un modello semiparametrico di Cox si potrebbe arrivare a conclusioni errate. La cosa più semplice da fare quando si è in una situazione simile, rimane quella della stratificazione, operazione che in questo caso risulterebbe semplice dato il numero ridotto di livelli di VOD. La stratificazione è una modifica del modello di Cox a rischi proporzionali che permette il controllo dei predittori che non soddisfano l'assunzione di proporzionalità, attraverso la "stratificazione" (cfr. paragrafo 2.3.10). Il modello che si utilizza é

$$h_g(t, \underline{z}) = h_{0g}(t) exp(\sum_i \beta_i z_i), \tag{2.17}$$

dove

- $g = 1, 2$ denota lo strato #;

stratificando per VOD.

Si stima la curva di sopravvivenza con il modello stratificato di Cox dato nella (2.17).

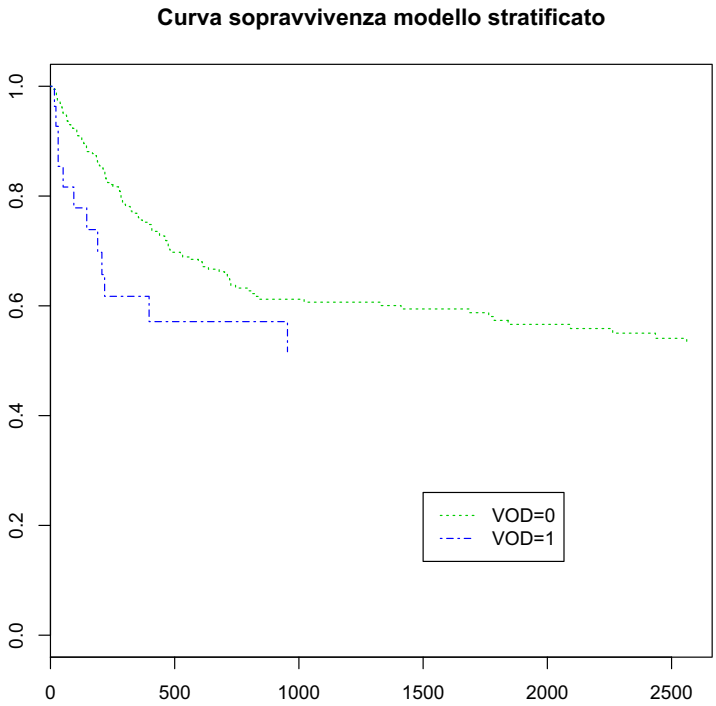


Figure 2.7: Stima curva di sopravvivenza.

La Figura 2.7 mostra che le due curve stratificate non sembrano essere molto differenti da quelle viste nel paragrafo 2.4.2 (Figura 2.5) utilizzando il metodo non parametrico di Kaplan-Meier. Anche qui la curva, e quindi i tempi di sopravvivenza, per i soggetti con presenza di infezione risultano minori di quelli senza. La differenza in questo caso risulta essere più enfatizzata rispetto alla stima effettuata attraverso il metodo di Kaplan-Meier.

2.5 Studio per soggetto

Come già spiegato nel paragrafo 2.1.2, la differenza sostanziale tra i due tipi di studio sta nella creazione del database. In questo studio si ha un'univocità per soggetto, il quale compare una unica volta. Si studia, quindi, il tempo di sopravvivenza globale per soggetto considerando il numero di trapianti come una covariata.

2.5.1 Analisi esplorative

Anche per questo studio si parte con un'analisi esplorativa. Questa è usata per valutare, attraverso rappresentazioni grafiche, se esistono possibili associazioni tra le covariate e la variabile risposta tempo di sopravvivenza.

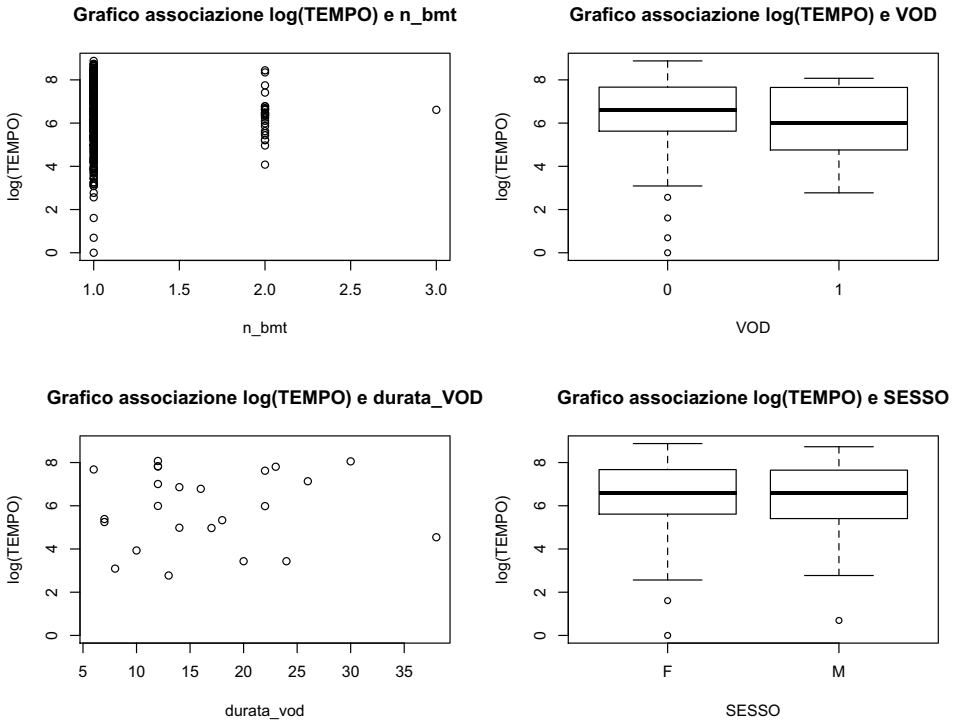


Figura 2.8: Analisi esplorative studio per soggetto (a)

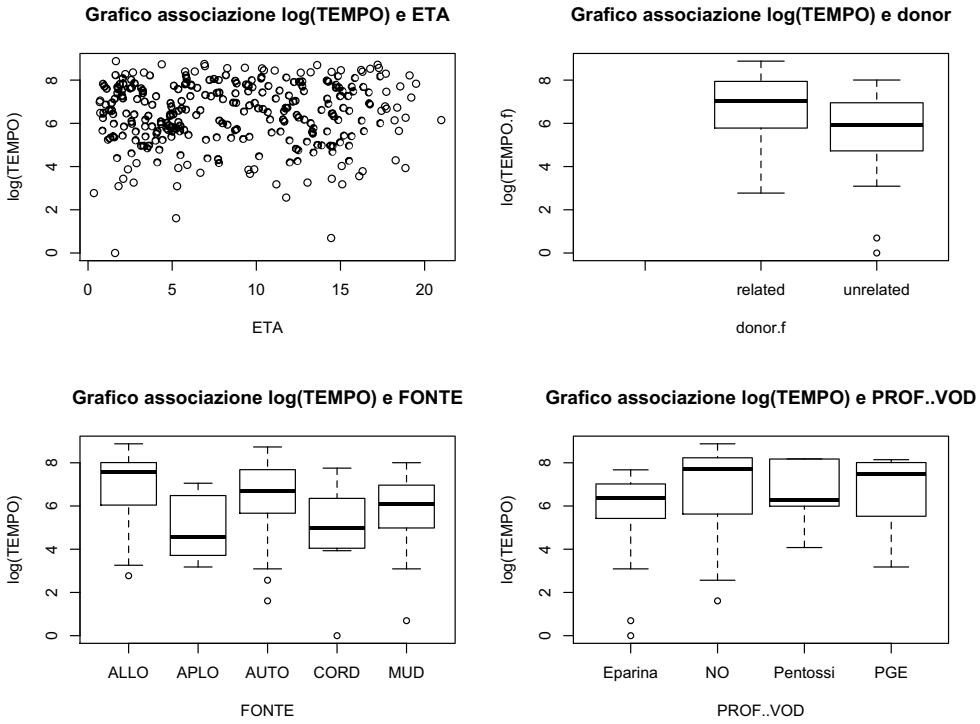


Figura 2.9: Analisi esplorative studio per soggetto (b)

Valgono le stesse considerazioni analoghe a quelle del paragrafo 2.4.1. Sebbene il gruppo con tre trapianti presenti un solo soggetto, si può notare come la sopravvivenza sia inversamente proporzionale al numero di trapianti subiti. Sembra non esserci una differenza sostanziale tra il gruppo che ha presentato infezione da quello che non l'ha presentata. Andamenti specifici della covariata durata dell'infezione VOD non ci sono. Stessa cosa per l'età. I tempi di sopravvivenza del gruppo femmine risulta simile a quello del gruppo maschi. Sembra differire, invece, il gruppo dei donatori relazionati al paziente da quello dei non relazionati. Inoltre sembrano differire tra loro i gruppi delle variabili fonte del TCSE e profilassi di VOD. I pazienti con TCSE allogenico hanno tempi di sopravvivenza più alti di tutti gli altri, mentre quelli con TCSE aplo-identico, più bassi. I pazienti con PROF..VOD=NO hanno tempi di sopravvivenza maggiori di tutti, mentre i pazienti con Profilassi eparina e pentossi minori.

2.5.2 Stima curva di sopravvivenza attraverso il metodo Kaplan-Meier

Per iniziare a farsi un'idea di quella che può essere la differenza tra i due livelli della variabile VOD, si stima la funzione di sopravvivenza separatamente per i due strati attraverso il metodo di Kaplan-Meier.

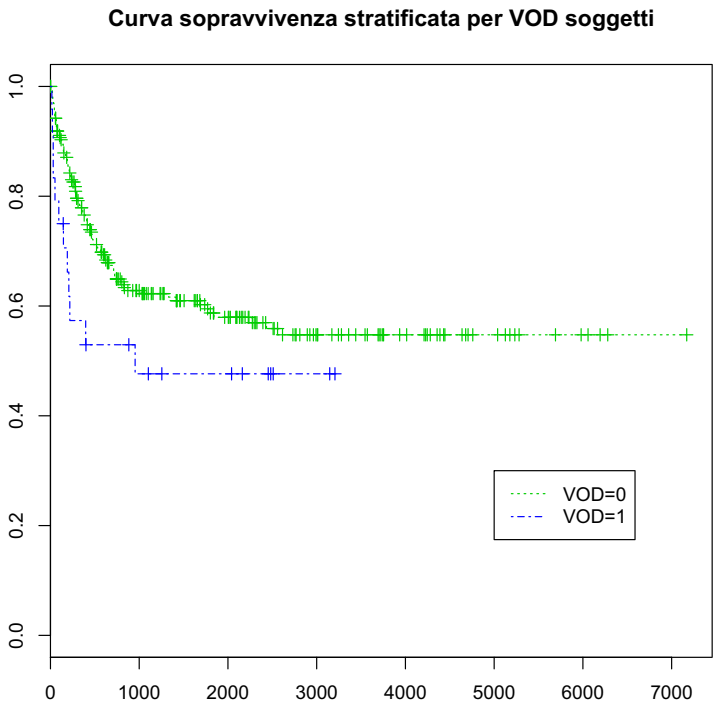


Figura 2.10: Stima curva sopravvivenza metodo Kaplan-Meier soggetti

Dalla Figura 2.10 si nota che la differenza tra i due gruppi appare netta. I pazienti che hanno presentato infezione mostrano un tempo di sopravvivenza minore di quelli che non l'hanno presentata. Si nota inoltre che le due curve sono ben separate, e la curva di sopravvivenza rispetto ai pazienti senza infezione presenta probabilità di sopravvivenza maggiore di quanto visto nella Figura 2.5.

2.5.3 Selezione variabili

Procedendo con una selezione *forward* si vogliono selezionare le variabili di maggiore interesse. Il modello che si ottiene non presenta grandi differenze da quello dello studio precedente. Infatti, il modello a cui si perviene contiene i seguenti regressori:

- ETA;
- FONTE;
- PROF..VOD;
- VOD.

Differentemente dallo studio precedente, la covariata che rappresenta il numero di trapianti effettuati dal singolo paziente non è contenuta nel modello. Questo vuol dire che il numero di trapianti sembra non incidere sul tempo di sopravvivenza

del paziente. E' opportuno tuttavia notare la notevole differenza tra il numero di pazienti con un solo trapianto effettuato e quelli con due o più trapianti.

Si presentano nel seguito i valori di:

- stime dei coefficienti delle variabili;
- HR (cfr. paragrafo 2.3.5);
- errori standard;
- statistica test di Wald;
- *p-values* della statistica test di Wald

ottenuti dall'ambiente statistico di lavoro.

```
Call: coxph(formula = Surv(TEMPO, morto == 1) ~ ETA + FONTE +
PROF..VOD + VOD)
 n= 285
               coef      exp(coef)se(coef)    z  Pr(>|z|)
ETA            0.04263   1.04355  0.01924 2.216  0.02671 *
FONTEAPLO      2.52876  12.53796  0.47835 5.286 1.25e-07 ***
FONTEAUTO      0.51599   1.67530  0.27670 1.865  0.06221 .
FONTECORD      1.68271   5.38010  0.56776 2.964  0.00304 **
FONTEMUD       0.96552   2.62614  0.39551 2.441  0.01464 *
PROF..VODNO    0.65576   1.92661  0.27084 2.421  0.01547 *
PROF..VODPen   0.38810   1.47418  0.74839 0.519  0.60405
PROF..VODPGE   0.82422   2.28009  0.32098 2.568  0.01023 *
VOD1           0.76185   2.14223  0.33003 2.308  0.02097 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tutti i coefficienti risultano significativi al livello 5%, tranne un livello (Pentossi) della variabile PROF..VOD. Di nuovo si può notare come il livello di FONTE = APLO sia quello con maggiore tasso di rischio. Si ha che:

- il soggetto che presenta TCSE aplo-identico, ha 13 volte in più di probabilità di presentare l'evento morte di un paziente che non lo presenta;
- il soggetto che presenta TCSE da cordone ombelicale, ha 5 volte e mezza in più di probabilità di presentare l'evento morte di un paziente che non lo presenta;
- il soggetto che presenta ETA = 2 ha 2 volte in più, e chi ha ETA= 13 ha 13 volte e mezza, di probabilità di presentare l'evento morte di un paziente che non lo presenta.

I risultati ottenuti coincidono con quelli dello studio precedente (cfr. paragrafo 2.4.3).

2.5.4 Verifica assunzione rischi proporzionali

Come nel paragrafo 2.4.4, la verifica viene effettuata con tre metodi e particolare attenzione sarà data alla variabile VOD. Una problematica legata al metodo grafico è come categorizzare una variabile continua come l'età. Per ovviare a questo problema si decide di studiare questa covariata solo con il test di *Goodness Of Fit* e non con il metodo grafico.

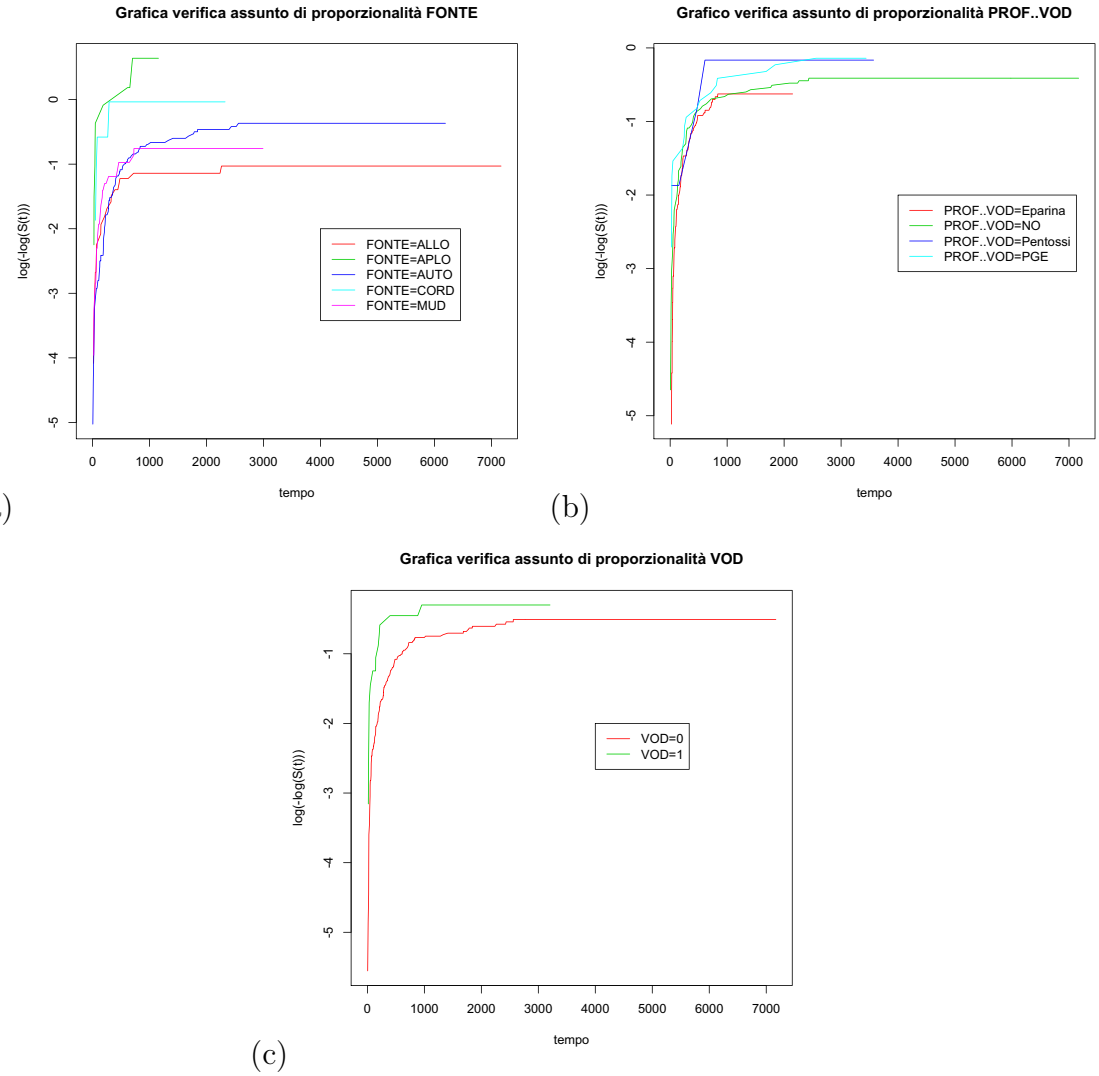


Figure 2.11: Verifica grafica assunto di proporzionalità per VOD soggetti.

Come nel paragrafo 2.4.4, il metodo grafico risulta essere di difficile interpretazione. Si possono notare alcuni punti incidenti tra le varie curve. Questi andamenti potrebbero, comunque, essere frutto di stime errate delle curve di sopravvivenza date

dalla scarsità di valori nei vari gruppi. Valutare, quindi, l'assunto di proporzionalità utilizzando il solo metodo grafico potrebbe portare a decisioni errate. Utilizzando il test di *Goodness Of Fit* si ottiene

```
                      rho  chisq       p
ETA                0.0975  1.021 0.31228
FONTEAPLO         -0.0376  0.174 0.67673
FONTEAUTO          0.2199  5.572 0.01825
FONTECORD         -0.0618  0.398 0.52829
FONTEMUD          -0.0672  0.528 0.46749
PROF..VODNO       -0.2295  6.550 0.01049
PROF..VODPentossi  0.0378  0.164 0.68595
PROF..VODPGE      -0.0865  0.882 0.34766
VOD1              -0.2441  6.087 0.01362
```

L'unico parametro che si presenta significativo contro l'ipotesi nulla è VOD. Ipotizzato che le altre variabili (n_bmt, FONTE e PROF..VOD) verifichino l'assunto di proporzionalità, e supponendo che solo VOD non lo faccia (come nel paragrafo 2.4.4), la costruzione del modello, per verificare se VOD è una variabile tempo-dipendente, risulta la seguente

$$h\left(t, \underline{z}(t)\right) = h_0(t) exp\left[\sum_{i=1}^{p} \beta_i z_i + \delta VOD \cdot g\left(t\right)\right] \tag{2.18}$$

con

- β_i coefficiente del regressore i-esimo;
- z_i regressore i-esimo;
- p numero di regressori presenti nel modello;
- δ coefficiente di VOD;
- $g\left(t\right)$ funzione che dipende dal tempo.

Considerando $g\left(t\right) = t$, di seguito sono presentati i risultati dell'adattamento del modello (2.18).

```
Call: coxph(formula = Surv(TEMPO, morto == 1) ~ ETA + FONTE
+ PROF..VOD + VOD + VOD_T)
 n= 285
                coef  exp(coef)  se(coef)     z  Pr(>|z|)
```

```
ETA          0.045011  1.046040  0.019781  2.275  0.02288 *
FONTEAPLO    2.566703 13.022814  0.480620  5.340 9.27e-08 ***
FONTEAUTO    0.424406  1.528682  0.279970  1.516  0.12955
FONTECORD    1.639325  5.151692  0.564975  2.902  0.00371 **
FONTEMUD     0.911522  2.488106  0.399428  2.282  0.02249 *
PROF..VODNO  0.716307  2.046861  0.279754  2.560  0.01045 *
PROF..VODPen 0.234049  1.263706  0.754160  0.310  0.75630
PROF..VODPGE 0.812093  2.252618  0.325065  2.498  0.01248 *
VOD1         3.426666 30.773860  0.481442  7.117 1.10e-12 ***
VOD_T       -0.004264  0.995745  0.001440 -2.961  0.00307 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Considerando $g(t) = log(t)$, invece si ottiene:

```
Call: coxph(formula = Surv(TEMPO, morto == 1) ~ ETA + FONTE
+ PROF..VOD +  VOD + VOD_LOGT)
 n= 285
                coef  exp(coef)   se(coef)      z Pr(>|z|)
ETA          0.04166    1.04253    0.01940  2.148  0.03175 *
FONTEAPLO    2.52974   12.55024    0.48068  5.263 1.42e-07 ***
FONTEAUTO    0.43352    1.54267    0.27879  1.555  0.11994
FONTECORD    1.62577    5.08235    0.56473  2.879  0.00399 **
FONTEMUD     0.87697    2.40361    0.39913  2.197  0.02800 *
PROF..VODNO  0.64743    1.91063    0.27874  2.323  0.02020 *
PROF..VODPen 0.34586    1.41321    0.75186  0.460  0.64551
PROF..VODPGE 0.71870    2.05176    0.33423  2.150  0.03153 *
VOD1         9.05183 8534.16841    1.27251  7.113 1.13e-12 ***
VOD_LOGT    -1.37273    0.25341    0.24708 -5.556 2.76e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Anche tale procedura conferma che la variabile VOD risulta essere una variabile che dipende dal tempo.

2.5.5 Stima curva di sopravvivenza modello di Cox stratificato

Utilizzando un modello stratificando rispetto alla variabile VOD, si ha il modello

$$h_g(t, \underline{z}) = h_{0g}(t)exp(\sum_i \beta_i z_i) \tag{2.19}$$

Tale modello, con la variabile età al posto del numero di TCSE, risulta identico al modello nella (2.17) del precedente studio. E' interessante stimare la curva di sopravvivenza nei due strati di VOD utilizzando questo tipo di modello.

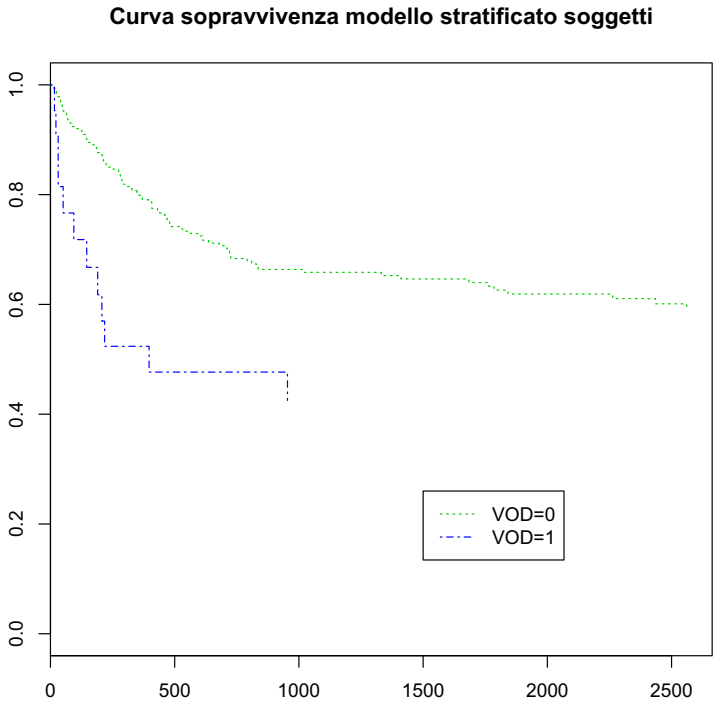


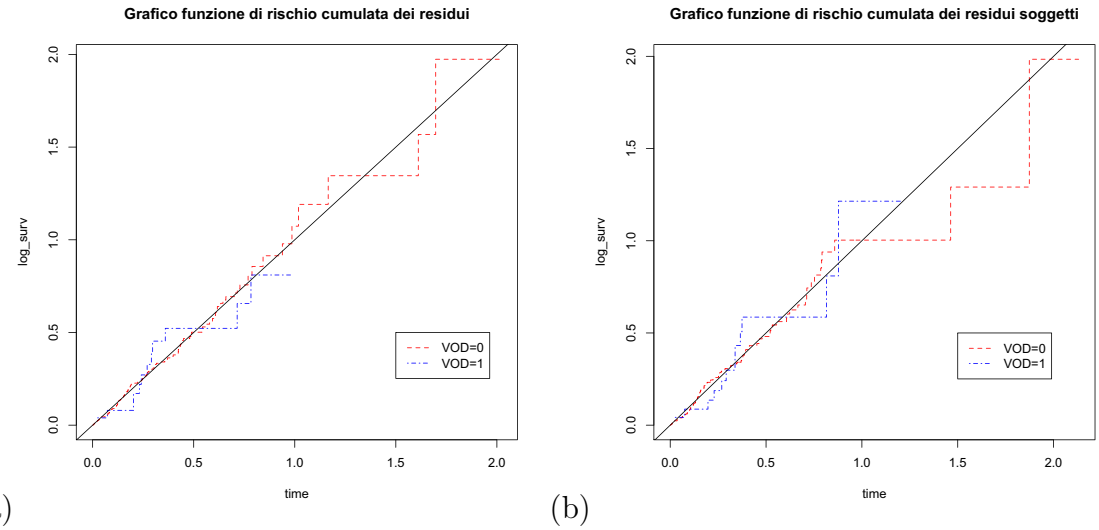
Figura 2.12: Stima curva di sopravvivenza soggetti

Dalla Figura 2.12 si nota che la curva, e i tempi di sopravvivenza, per i pazienti che presentano infezione sono minori di quelli che non la presentano. La differenza risulta essere ancora più evidente attraverso questo tipo di studio a differenza di quello precedente, soprattutto verso gli ultimi dati del gruppo VOD=1. Si nota, inoltre, che la curva di sopravvivenza per VOD=0 presenta probabilità di sopravvivenza maggiore rispetto a tutte le precedenti curve stimate (Figure 2.5, 2.7, 2.10).

2.6 Bontà del modello

I residui di Cox-Snell possono essere usati per verificare l'adattamento di un modello a rischi proporzionali di Cox (vedi Klein e Moeschberger, 2003, p.354). Per verificare se il modello si adatta bene, si rappresenta graficamente la stima della curva di rischio cumulata dei residui di Cox-Snell rispetto ai residui stessi. Se la curva è distribuita vicino alla bisettrice, il modello si adatta bene. I grafici saranno tanti

quanti gli strati della variabile VOD, quindi due. Si possono anche rappresentare i residui di devianza (vedi Therneau e Grambsch, 2000, p.83) per valutare la presenza di eventuali *outliers*. Tutti i residui vengono calcolati con il modello (2.17) per i trapianti e con il modello (2.19) per i soggetti.

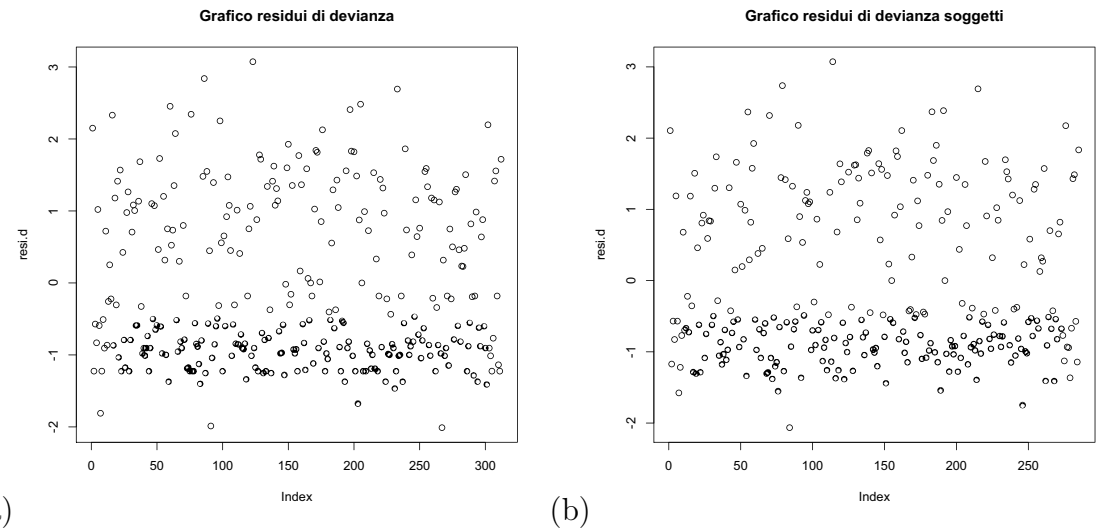


(a) (b)

Figura 2.13: Grafici funzione di rischio cumulato stimata dei residui di Cox-Snell: (a) trapianti (b) soggetti.

Si può notare che la parte finale della curva appena rappresentata, per quanto riguarda i pazienti senza infezione, risulta non adattarsi perfettamente. Ancora peggio la curva per i pazienti che hanno presentato infezione. Questi risultati sono simili per entrambi i tipi di studio.

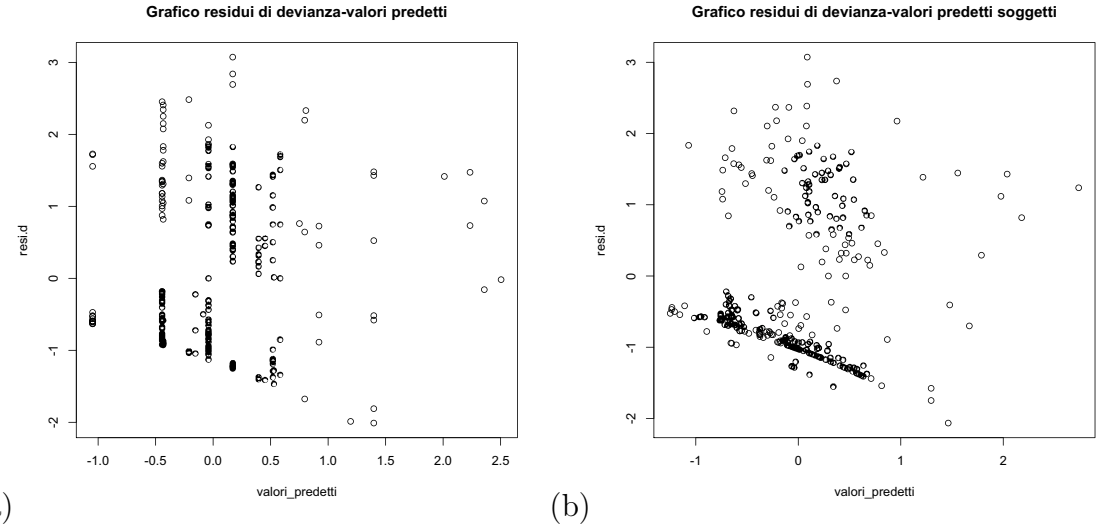
I residui di devianza servono per identificare individui la cui risposta è scarsamente predetta. In caso di bassa censura presentano una distribuzione normale. Se la censura risulta più ”pesante” la distribuzione è sempre approssimatamente simmetrica, ma l'approssimazione normale risulta meno accurata. Inoltre un residuo positivo suggerisce che l'evento accade prima di quello atteso. Un residuo negativo suggerisce che l'evento accade dopo che quello atteso (vedi Singer e Willett, 2003, p. 575). Nella Figura 2.14 vengono presentati i residui di devianza dei modelli (2.17) e (2.19).



(a) (b)

Figura 2.14: Grafico residui di devianza:(a) trapianti e (b) soggetti.

Se il modello si adatta ai dati in modo adeguato, i residui di devianza si distribuiscono in modo simmetrico intorno allo 0. Le Figure 2.14(a) e (b) mostrano una non perfetta simmetria. Non sembrano esserci valori anomali. I grafici risultano a vista d'occhio simili. Si vede ora il grafico sempre tra i residui di devianza e i valori predetti.



(a) (b)

Figura 2.15: Grafico residui di devianza vs valori predetti:
(a) trapianti e (b) soggetti.

Come visto in Figura 2.14 anche attraverso questi grafici si può notare che non sembrano esserci valori estremi. Si può notare che non c'è un'eccessiva variabilità nei dati, infatti sono tutti principalmente compresi tra | 3 |. I valori più estremi sono intorno al valore predetto=0.2 e 1.4 per entrambi i grafici mostrati. La Figura 2.15(b) presenta parecchi residui negativi a differenza della Figura 2.15(a).

2.7 Conclusioni

I principali quesiti posti in questa analisi erano lo studio di una variabile tempo-dipendente e quindi la difficoltà di utilizzo di un modello a rischi proporzionali come quello di Cox, e le possibili differenze di studio tra soggetti e trapianti. Esistono molte tecniche per estendere il modello di Cox per poter così utilizzare variabili dipendenti dal tempo. Si è scelto di utilizzare un modello di Cox stratificato poichè il numero di livelli della variabile che non verificava l'assunto di proporzionalità risultava piccolo.

Principali differenze tra i due tipi di studio non ce ne sono. Si può notare che, studiando il soggetto, la covariata che rappresenta il numero di trapianti risulta non significativa. Questo sembrerebbe mostrare che, guardando il soggetto, il numero di trapianti non sembra influenzare il tempo di sopravvivenza, forse dovuto anche al fatto che esiste una bassa numerosità di pazienti, che hanno subito più di un trapianto. In entrambi i tipi di studio, comunque, si è notato che i fattori principali, quelli più significativi e quindi quelli che dovrebbero influire maggiormente sul tempo di sopravvivenza risultano essere:

- la fonte del trapianto;
- la profilassi di VOD;
- il numero di TCSE;
- età dei soggetti.

In entrambi gli studi si può notare comunque che i pazienti TCSE aplo-identico presentano probabilità più alta rispetto agli altri di presentare fallimento. Il fattore TCSE aplo-identico quindi risulta uno dei maggiori fattori portatori di morte.

Capitolo 3

Analisi dati su linfoma ALCL

3.1 Introduzione

In questo capitolo vengono discussi dei dati sullo studio del Linfoma Anaplastico a Grandi Cellule (ALCL), eseguito dal Centro Oncoematologico Pediatrico di Padova. A differenza dello studio precedente, si ha in questo caso uno studio di tipo retrospettivo, in cui un gruppo di malati (casi) e un gruppo di sani (controlli) vengono confrontati sulla base del grado della loro esposizione ad un fattore di rischio. Problema principale risulta essere il numero di dati a disposizione, ovvero 14 pazienti in totale, di cui 4 per il gruppo dei controlli e 10 per quello dei casi.

3.1.1 La malattia

I geni sono le unità ereditarie fondamentali degli organismi viventi, dirigono lo sviluppo fisico e comportamentale di un essere vivente. I geni strutturali vengono trascritti e in genere determinano la sequenza amminoacidica delle proteine, molecole che svolgono una grande varietà di compiti. Per esempio, le proteine trasmettono messaggi tra le cellule, attivano e disattivano geni, sono fondamentali nella contrazione muscolare, formano strutture come capelli e peli. Alcuni geni non codificano proteine ma RNA, molecole deputate a svolgere funzioni precise e importantissime all'interno della cellula. Proprio da questo RNA parte lo studio effettuato dal Centro Oncologico di Padova riguardo un determinato linfoma: l'ALCL. In questo studio, il Centro voleva valutare la funzione della proteina Hsp70 nei controlli e nei casi, poiché la presenza (quando non necessaria) di questa proteina sembra limitare l'efficacia della terapia chemioterapica applicata al soggetto malato. Lo stesso paziente malato sembra avere di conseguenza un livello maggiore della proteina Hsp70 del paziente facente parte del gruppo di controllo, quindi sano.

3.2 I dati

In questo studio si hanno due gruppi, uno di controlli e uno di casi. Importante ricordare che i linfonodi sono presenti in qualsiasi individuo vivente, quindi i due gruppi si distinguono: uno per linfonodi sani, l'altro per quelli malati o tumorali. I valori presenti nella Tabella 3.1, descrivono il livello della proteina Hsp70 che, come detto, se presente quando non necessaria potrebbe causare una resistenza al trattamento farmacologico.

Controlli	Casi
0.23	2.8
0.44	1.4
0.19	0.13
0.08	0.2
	0.8
	0.56
	0.44
	5.2
	1.7
	1.14

Tabella 3.1: Osservazioni Casi-Controlli ALCL

Lo studio effettuato presenta una ridotta numerosità campionaria, soprattutto per quanto riguarda il numero di valori del gruppo di controllo. Lavorare con un numero così limitato di dati nasce dal fatto che, per svariati motivi, non se ne possono reperire altri. In campi quali la biologia, le scienze mediche, la psicologia, spesso ci si ritrova nella situazione in cui il numero di osservazioni disponibili è molto basso.

Il problema della limitata numerosità campionaria recentemente sta ricevendo diverse attenzioni in letteratura (vedi ad esempio Brazzale et al., 2007), soprattutto per quanto riguarda l'ambito medico. In effetti, in alcuni casi e in alcuni specifici settori, i dati a disposizione dai centri di ricerca in campo medico risultano essere molto scarsi. Tuttavia, esistono procedure di inferenza estremamente accurate anche per numerosità campionarie molto basse.

3.2.1 Scopi

I principali scopi di questo studio risultano essere:

- confronto fra il gruppo controlli e il gruppo casi, rispetto al livello della proteina presente nel soggetto;
- studio nei dati a disposizione della quantità $R = P(X < Y)$, dove
- X: livello valore della proteina Hsp70 nel gruppo casi;

- Y: livello valore della proteina Hsp70 nel gruppo controlli.

Pertanto, R esprime la probabilità che il livello della proteina Hsp70 nel gruppo dei casi sia minore del livello della proteina nel gruppo controlli.

Come sarà spiegato più avanti, per quanto riguarda il secondo scopo, si presenterà prima l'assunzione di distribuzione normale, poi quella esponenziale. I risultati verranno confrontati e si sceglierà quello più attendibile.

3.3 Nozione e terminologia

In questo paragrafo viene fornita una breve descrizione delle tecniche statistiche utilizzate per l'analisi dei dati ALCL.

3.3.1 Test di Kolmogorov-Smirnov

Molti test utilizzati in statistica sono test parametrici. Questi test sono basati su assunzioni importanti, quali un'adeguata dimensione campionaria e la distribuzione normale della variabile di interesse. Come detto nel paragrafo 3.2, in ambito medico la dimensione dei campioni studiati può essere piccola. Data questa limitazione, è inevitabile che molti studi di ricerca medica non verifichino le assunzioni necessarie per applicare test parametrici. Una soluzione a queste problematiche è l'uso di tecniche non-parametriche (vedi Pett, 1997, p. 8).

Il test di Kolmogorov-Smirnov (vedi Wayne, 1996, p. 550) è un test non parametrico che verifica la forma delle distribuzioni campionarie. È applicabile a dati per lo meno ordinali perchè richiede la costruzione di una distribuzione di frequenza cumulata (vedi Sheskin, 2004, p. 203). Ci sono, principalmente, due tipi di test: il primo chiamato *one-sample test* (vedi Gibbons e Chakraborti, 2003, p. 111) e il secondo *two-sample test* (vedi Gibbons e Chakraborti, 2003, p. 239). Entrambi sono usati per determinare se due insiemi di dati provengono dalla stessa distribuzione. La prima versione è comunemente usata per comparare dati sperimentali con distribuzioni attese. Le distribuzioni attese potrebbero derivare dai dati o essere completamente indipendenti da loro. Questo test può determinare se una popolazione differisce da una distribuzione, per esempio, normale.

Sia X una variabile casuale continua con funzione di ripartizione $F(x)$. Il test di Kolmogorov-Smirnov *one-sample* verifica che la variabile casuale X abbia funzione di ripartizione uguale ad una data funzione di ripartizione $F_0(x)$, ovvero

$$\begin{cases} H_0: & F(x) = F_0(x) \\ H_1: & F(x) \neq F_0(x) \end{cases} \tag{3.1}$$

Sia $x = (x_1, \dots x_n)$ un campione casuale di ampiezza n tratto dalla variabile casuale X. Poiché tale problema riguarda la funzione di ripartizione della variabile casuale X, è intuitivo basare la statistica test sulla funzione di ripartizione empirica. Dette quindi $x_{(1)}, \dots . x_{(n)}$ le n osservazioni ordinate, la funzione di ripartizione empirica è definita come

$$\hat{F}_n(x) = \begin{cases} 0 & x \leq x_{(1)} \\ \frac{k}{n} & x_{(k)} \leq x < x_{(k+1)} \\ 1 & x \geq x_{(n)} \end{cases} \tag{3.2}$$

La (3.2) è uno stimatore non distorto e consistente di $F(x)$.

La statistica test di Kolmogorov-Smirnov è data da

$$D_n = \sup_{-\infty<x<+\infty} \left| \hat{F}_n(x) - F_0(x) \right| \tag{3.3}$$

e risulta essere definita come la massima differenza (in valore assoluto) tra la funzione di ripartizione empirica $\hat{F}_n(x)$ e la funzione di ripartizione teorica $F_0(x)$. Le assunzioni sottostanti il test di Kolmogorov-Smornov sono:

- il campione è casuale semplice;
- la distribuzione ipotizzata $F_0(x)$ è continua.

L'idea del test di Kolmogorov-Smirnov è piuttosto semplice e intuitiva. Poiché $\hat{F}_n(x)$ stima la "vera" funzione di ripartizione $F(x)$, è logico basarsi su una qualche "distanza" tra $\hat{F}_n(x)$ e $F_0(x)$. Se $\hat{F}_n(x)$ e $F_0(x)$ sono "vicine", si accetta l'ipotesi nulla, mentre la si rifiuta se $\hat{F}_n(x)$ e $F_0(x)$ sono "lontane". Per valori "grandi" di D_n si rifiuta l'ipotesi nulla, mentre la si accetta per valori "piccoli" di D_n. Se c'è vicinanza tra le distribuzioni cumulative teorica ed empirica, viene avallata l'ipotesi che il campione provenga da una popolazione con una funzione di distribuzione cumulata specifica. Se il valore calcolato di D_n supera il valore riportato nella tabella dei quantili per la statistica test di Kolmogorov per una dimensione campionaria di n (vedi Miller, 1956), l'ipotesi nulla viene rifiutata ad un livello di significatività α.

3.3.2 *Stress-Strength model*

In anni recenti, la letteratura statistica ha ampiamente discusso il problema dell'inferenza su $R = P(X < Y)$, quantità nota come *stress-strength model*. Questo tipo di modello può essere applicato in qualsiasi tipo di contesto, da quello meccanico, finanziario, sociale e per esempio medico.

Lo *stress-strength model* è nato inizialmente da un classico problema risolto tramite un semplice test non-parametrico (vedi Kotz, 2002, p. 10). Se inizialmente l'uso di R si è dimostrato utile in ambito ingegneristico, sotto il termine di "affidabilità" (vedi Kotz, 2002, p. 205), successivamente si è visto come l'applicazione di questo tipo di problema non era confinata solo nell'ambito ingegneristico. Lo sviluppo di queste nuove applicazioni statistiche ha innescato numerose applicazioni orientate anche in campo medico. Ad esempio, in uno studio clinico, X potrebbe rappresentare la risposta di un gruppo controllo e Y quella di un gruppo trattamento. Pertanto, R misura l'efficacia del trattamento. Esiste inoltre una relazione tra la curva ROC e lo *stress-strength model* (vedi Kotz, 2002, p. 223). Il modello *stress-strength* è, quindi, una relazione flessibile, universale e facilmente adattabile in vari campi di fenomeni sia naturali che umani.

Stima di massima verosimiglianza di R. La stima di massima verosimiglianza risulta senza dubbio la più popolare procedura di stima parametrica per $R = P(X < Y)$, (vedi Pace e Salvan, 2001, p.131). Una descrizione dettagliata del metodo della stima di massima verosimiglianza è presentato, per esempio, in Casella e Berger, (1990). Siano dati due campioni casuali semplici $(x_1, \ldots, x_n)$ e $(y_1, \ldots, y_m)$, tratti, rispettivamente, da due variabili casuali X e Y indipendenti con funzioni di densità $f_X(x \mid \theta_X)$ e $f_Y(y \mid \theta_Y)$, con $\theta_X \, \epsilon \, \Theta_X$ e $\theta_Y \, \epsilon \, \Theta_Y$. Lo scopo è stimare R sulla base delle osservazioni $(x_1, \ldots, x_n)$, e $(y_1, \ldots, y_m)$.

Sia $L(\theta) = L(\theta \mid x, y)$ la funzione di verosimiglianza per $\theta = (\theta_x, \theta_y)$.

La stima di massima verosimiglianza $\hat{\theta}$ del parametro θ è il valore del parametro per cui $L(\theta)$ raggiunge il suo massimo. Una proprietà importante della stima di massima verosimiglianza è l'invarianza rispetto a qualunque funzione $\varphi(\theta)$. Infatti, se $\hat{\theta}$ è la stima di massima verosimiglianza di θ, allora la stima di massima verosimiglianza di $\varphi(\theta)$ è $\varphi(\hat{\theta})$ (vedi Casella e Berger, 1990, p.294).

Se X e Y sono indipendenti con funzioni di densità di probabilità $f_X(x \mid \theta_X)$ e $f_Y(y \mid \theta_Y)$ e funzioni di ripartizione $F_X(x \mid \theta_X)$ e $F_Y(x \mid \theta_Y)$, rispettivamente, $R = R(\theta)$ può essere scritto come

$$R(\theta) = \int_{-\infty}^{\infty} F_X(z \mid \theta_X) \, f_Y(z \mid \theta_Y) \, dz = \int_{-\infty}^{\infty} (1 - F_Y(z \mid \theta_Y)) \, f_x(z \mid \theta_X) \, dz. \tag{3.4}$$

Per la proprietà di invarianza della stima di massima verosimiglianza si ha che la stima di massima verosimiglianza di R è $\hat{R} = R\left(\hat{\theta}\right)$. In molte applicazioni una stima puntuale non è sufficiente, ma è preferibile considerare un intervallo di confidenza per R. La procedura più comune è quella di considerare un intervallo di confidenza alla Wald di livello approssimato $(1 - \alpha)$, (vedi Ott, 2008, p.502)

L'inferenza su R è stata ampiamente discussa in letteratura sotto varie assunzioni distributive per X e Y. Sebbene le classiche procedure basate sulla verosimiglianza per fare inferenza su R sono semplici e generali, esse possono essere inaccurate quando la grandezza del campione risulta piccola, in particolare quando la dimensione dei parametri ignoti è elevata. Esistono procedure per ovviare a questo problema, basate su recenti tecniche di verosimiglianza e metodi asintotici (vedi Cortese Ventura, 2009) a cui si rimanda e che non saranno trattate in questa tesi. Di seguito si presentano le espressioni di R e delle stime di massima verosimiglianza, sotto assunzioni di distribuzione normali e/o esponenziali per X e Y.

Esempio 1: Assunzione normale Siano X e Y due variabili normali indipendenti con medie μ_1 e μ_2 e varianze σ_1^2 e σ_2^2, rispettivamente. Allora si ha

$$R(\theta) = P(X < Y) = \Phi\left(\frac{\mu_2 - \mu_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right), \quad con \quad \theta = \left(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right), \tag{3.5}$$

dove $\Phi(z)$ è la funzione di ripartizione della distribuzione normale. La stima di massima verosimiglianza di R è

$$\hat{R} = \Phi\left(\frac{\overline{y} - \overline{x}}{\sqrt{\frac{n_2-1}{n_2} \cdot s_Y^2 + \frac{n_1-1}{n_1} \cdot s_X^2}}\right) \tag{3.6}$$

con

$$\overline{y} = \sum_{i=1}^{n_2} \frac{y_i}{n_2}, \overline{x} = \sum_{i=1}^{n_1} \frac{x_i}{n_1} \tag{3.7}$$

$$s_Y^2 = \frac{\sum (y_i - \bar{y})^2}{n_2 - 1}, s_X^2 = \frac{\sum (x_i - \bar{x})^2}{n_1 - 1}. \tag{3.8}$$

Esempio 2: Assunzione esponenziale Siano X e Y due variabili casuali esponenziali con funzioni di densità di probabilità $f_X(x \mid \alpha) = \alpha exp(-\alpha X)$ e $f_Y(y \mid \beta) = \beta exp(-\beta Y)$, rispettivamente. Dalla 3.4 si ha che

$$R(\theta) = \int_0^\infty \left(1 - e^{-\alpha x}\right) \beta e^{-\beta x} dx = \frac{\alpha}{\alpha + \beta}. \tag{3.9}$$

Se $(x_1 \dots x_{n_1})$ e $(y_1 \dots y_{n_2})$ sono campioni indipendenti da $f_X(x \mid \alpha)$ e $f_Y(y \mid \beta)$, rispettivamente, si ha $f(x, y \mid \alpha, \beta) = \alpha^{n_1} \beta^{n_2} exp\{-\alpha n_1 \bar{x} - \beta n_2 \bar{y}\}$, con $\bar{x} = n_1^{-1} \sum_{i=1}^{n_1} x_i$ e $\overline{y} = n_2^{-1} \sum_{j=1}^{n_2} y_j$.

Le stime di massima verosimiglianza $\hat{\alpha}$ e $\hat{\beta}$ sono $\hat{\alpha} = \frac{1}{\bar{x}}$ e $\hat{\beta} = \frac{1}{\bar{y}}$. Pertanto

$$\hat{R} = \frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}} = \frac{\overline{y}}{\overline{y} + \overline{x}} = \frac{1}{1 + \frac{\bar{x}}{\bar{y}}}. \tag{3.10}$$

3.4 Il caso di studio ALK+

Si presenta in questo paragrafo l'analisi dei dati presentati nel paragrafo 3.2. Verrà anche presentata un'analisi preliminare dei dati, per dare un'idea sulle eventuali differenze tra i due gruppi. Sarà poi discusso il calcolo di $R(\theta) = P(X < Y)$ considerando

- X livello della proteina nei casi;
- Y livello della proteina nei controlli;
- $R = P(X < Y)$.

3.4.1 Analisi preliminare

In questo paragrafo viene presentata un'analisi grafica esplorativa. Inoltre, viene presentata un'applicazione del test statistico di Kolmogorov-Smirnov (cfr. Paragrafo 3.3.1) per verificare se, casi e controlli, differiscono da una distribuzione normale e/o da una distribuzione esponenziale. Si ricorda tuttavia che i dati a disposizione hanno bassa numerosità.

Di facile intuizione grafica risulta il box-plot, che permette di visualizzare eventuali differenze tra casi e controlli.

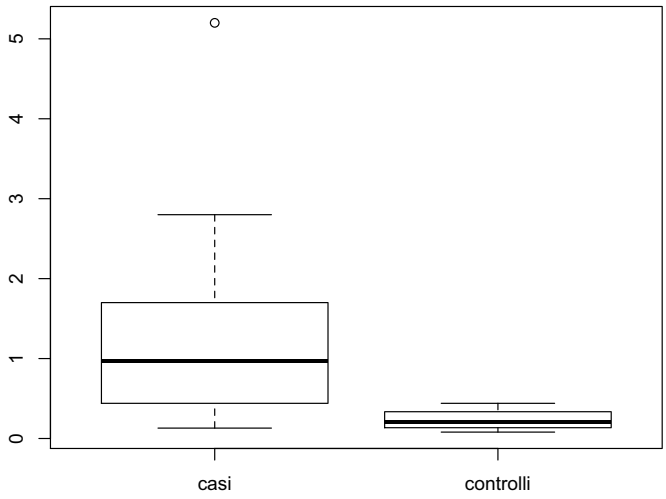


Figura 3.1: Boxplot casi controlli ALK+

Nella Figura 3.1 si nota un valore anomalo per i casi, una maggiore variabilità per i casi e una differenza tra il valore della mediana dei casi con quella dei controlli. Da qui già si potrebbe pensare che ci sia una differenza, significativa, tra casi e controlli: sembra, infatti, che il livello della proteina presente nei controlli sia nettamente minore di quello dei casi. Si ricorda ancora che i dati a disposizione hanno bassa numerosità.

Distribuzione normale Attraverso la rappresentazione dei boxplot si è potuto notare una simmetria da parte dei dati nei due gruppi. Questo potrebbe portare alla scelta di una distribuzione normale. Osservando gli istogrammi dei due gruppi, e applicando il metodo del nucleo per stimare la funzione di densità di probabilità di una variabile casuale, si vede se la scelta dell'assunzione di normalità potrebbe essere una buona scelta o meno (Figura 3.2 e 3.4). Inoltre applicando il grafico quantile-quantile, si confrontano i quantili della distribuzione empirica con quelli della distribuzione normale (Figura 3.3 e 3.5).

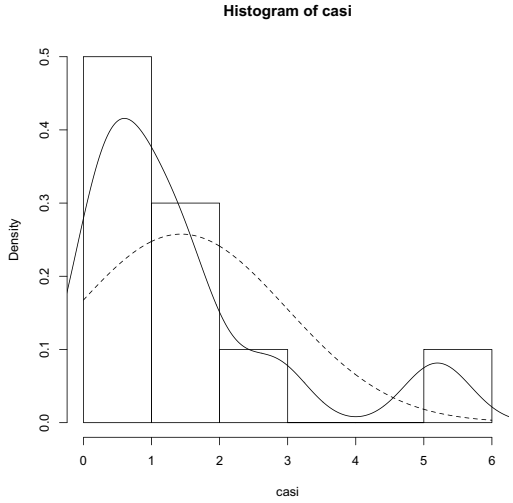


Figure 3.2: Istogramma gruppo casi.

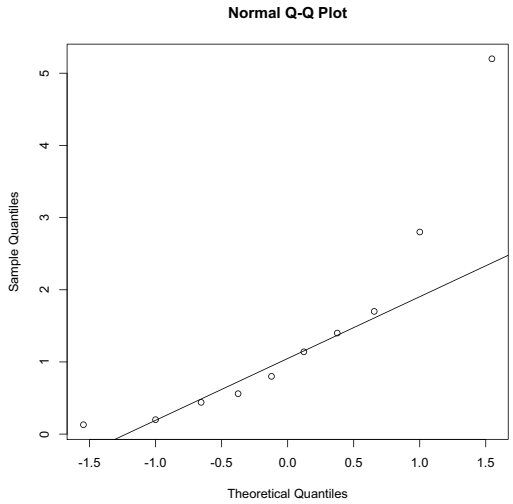


Figure 3.3: Grafico quantile contro quantile casi.

Tra i due gruppi, quello dei casi è quello con numerosità maggiore. Si nota che un andamento normale non è verificato. Anche utilizzando il grafico quantile contro quantile (Figura 3.3) i valori nelle code sembrano poco adattarsi ad una distribuzione normale. L'andamento sembra, piuttosto, seguire una distribuzione esponenziale.

Per quanto riguarda il gruppo controlli le cose sono addirittura più complicate, dato che i valori a disposizione sono solamente quattro (vedi Figura 3.4 e 3.5).

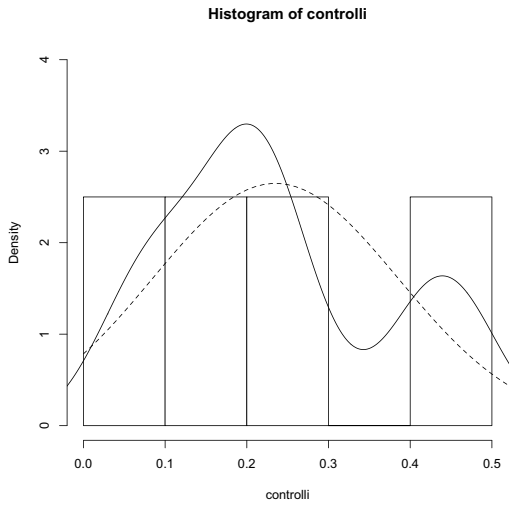


Figure 3.4: Istogramma gruppo controlli.

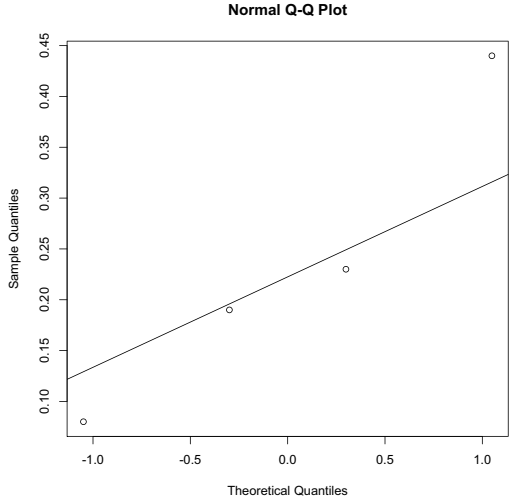


Figure 3.5: Grafico quantile contro quantile controlli.

Inoltre se si rappresentano casi e controlli nello stesso grafico si ha

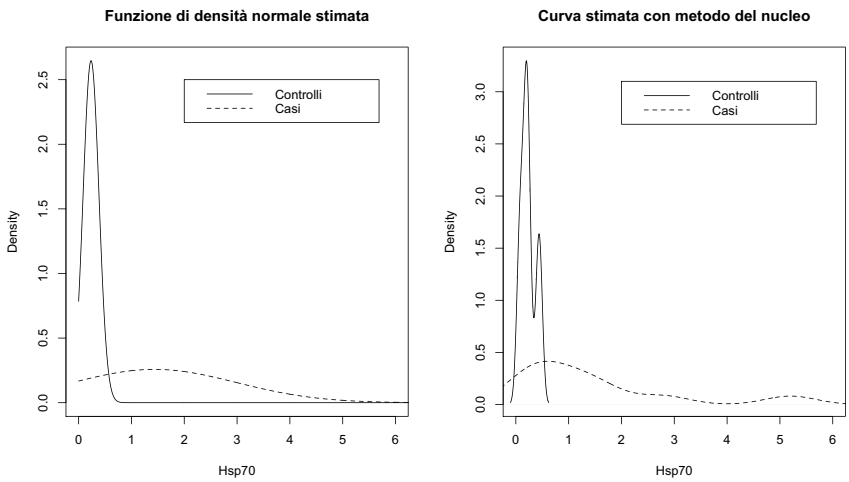


Figure 3.6: Grafico Casi e Controlli a confronto.

Dalla Figura 3.6 si può vedere come i casi presentano valore di Hsp70 maggiore rispetto ai controlli.

Come spiegato ad inizio capitolo, il test statistico di Kolmogorov-Smirnov (cfr. Paragrafo 3.3.1) viene applicato per verificare se, casi e controlli, differiscono da una distribuzione normale e/o da una distribuzione esponenziale.

Dalla 3.1, il test di ipotesi per la distribuzione normale diventa

$$\begin{cases} H_0: & F(X) = N(\mu, \sigma^2) \\ H_1: & F(X) \neq \overline{H_0}. \end{cases} \tag{3.11}$$

Il *p-value* del test per i casi risulta 0.002180, ovvero il risultato è significativo contro H_0.

Il *p-value* del test per i controlli risulta 0.1402, ovvero il risultato è non significativo contro H_0. Quindi, il test verifica che, i controlli, non differiscono significativamente da una distribuzione normale. Si ricorda, comunque, che il risultato del test, con una così bassa numerosità, può essere inaccurato.

Distribuzione esponenziale Oltre all'assunzione di distribuzione normale, si valuta anche l'adattamento ad una distribuzione esponenziale. In effetti, già da quanto visto nella Figura 3.2 si ha che forse una distribuzione esponenziale potrebbe adattarsi in modo migliore ai dati. Anche qui vengono rappresentati gli istogrammi per casi e controlli aventi curva stimata con il metodo del nucleo e funzione di densità esponenziale stimata (Figura 3.7 e 3.8). .

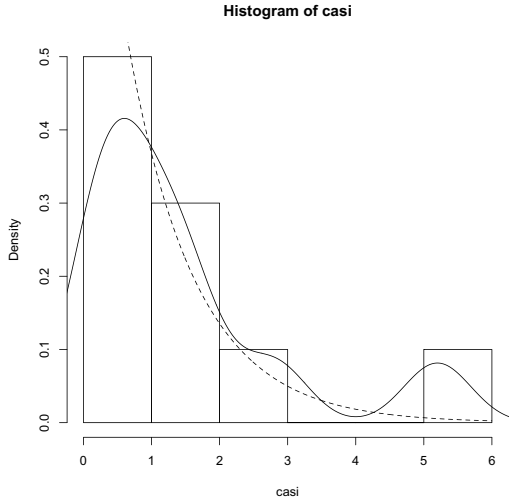


Figure 3.7: Istogramma gruppo casi

Tranne che per l'ultima parte, questa distribuzione sembra decisamente meglio adattarsi ai dati. Per quanto riguarda il gruppo controlli, la ridotta numerosità campionaria non permette di trarre conclusioni accurate.

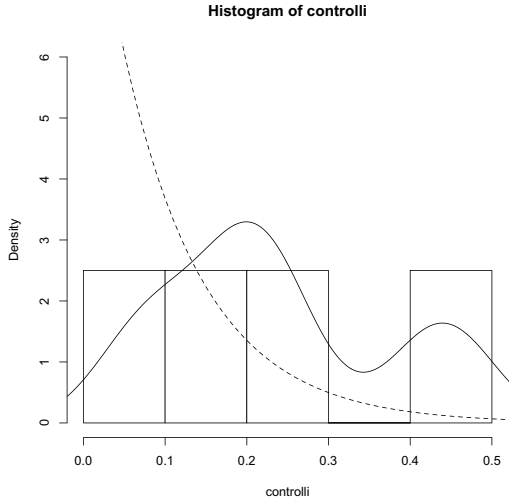


Figure 3.8: Istogramma gruppo controlli

Dalla (3.1), il test di ipotesi per la distribuzione esponenziale diventa

$$\begin{cases} H_0: & F(X) = Exp(\lambda) \\ H_1: & F(X) \neq \overline{H_0} \end{cases} \tag{3.12}$$

Il *p-value* del test per i casi risulta 0.8465, ovvero il risultato è non significativo contro H_0. Quindi, il test verifica che i casi si distribuiscono come una esponenziale.

Il *p-value* del test per i controlli risulta 0.03824, ovvero il risultato è significativo contro H_0 a livello $\alpha = 0.05$. Ancora una volta si ricorda che i risultati inferenziali dipendono dalla numerosità campionaria, e quindi, possono essere non accurati.

Quello che si è potuto vedere da un'analisi preliminare è che per i casi si potrebbe adattare una distribuzione esponenziale. Per quanto riguarda i controlli, la bassa numerosità campionaria, può non portare ad assunzioni accurate.

3.4.2 Valutazione di $R(\theta)$

Il calcolo di $R(\theta)$ è stato esemplificato sotto due assunzioni distributive: una normale e l'altra esponenziale. In letteratura sono, tuttavia, presenti altri risultati che discutono come stimare R con altre distribuzioni parametriche.

Nel seguito si considerano:

- la distribuzione normale poichè risulta essere una assunzione classica, anche se dai risultati del test (3.11), i casi sono significativi contro l'ipotesi nulla;
- la distribuzione esponenziale per il gruppo casi, dai risultati del test (3.12), sembra adattarsi in modo migliore.

Assumendo la distribuzione normale, si ha che dopo aver stimato la varianza e la media per il gruppo dei casi e per quello dei controlli, si trova $\hat{R} = 0.2$. La probabilità dell'evento $X < Y$ è, quindi, bassa.

Assumendo la distribuzione esponenziale si ha che per la (3.10) la stima di $R = P(X < Y)$ risulta pari a 0.14. La probabilità dell'evento $X < Y$, in accordo a quanto detto con la distribuzione normale, è bassa. L'intervallo di confidenza alla Wald è $(0.01, 0.28)$.

3.4.3 Conclusioni

I due gruppi sotto studio sembrano presentare, attraverso una semplice rappresentazione grafica, un andamento diverso da quello normale. Per i casi la distribuzione esponenziale sembrerebbe adattarsi meglio ai dati. Queste assunzioni vengono verificate attraverso il test non-parametrico di Kolmogorv-Smirnov. Il problema principale rimane comunque quello della bassa numerosità campionaria. Per espandere lo studio si è deciso di applicare un altro tipo di modello il quale potrebbe spiegare meglio un determinato comportamento dei due gruppi. Lo *stress-strength model* applicato ai gruppi casi e controlli viene utilizzato prima con un'assunzione di distribuzione normale, poi esponenziale. Questo modello applicato sotto assunzione normale fornisce un valore di R pari a 0.2. Per quanto riguarda il modello sotto assunzione esponenziale i risultati non cambiano di molto, e si ha un valore di R pari a 0.14. In entrambe le assunzioni si ha che la probabilità dell'evento $X < Y$ è bassa. Quindi, nella maggior parte delle volte, il livello della proteina nei casi risulta

maggiore del livello nei controlli. Questa conclusione va a dimostrare quanto detto nel paragrafo 3.1.1 dove il paziente malato sembra avere un livello maggiore della proteina Hsp70.

Appendice A

Appendice

Viene di seguito riportato e commentato il codice R utile per effettuare gli studi presentati in questo elaborato.

A.1 Dati TCSE

Come prima cosa si caricano la libreria `survival`, utilizzata per effettuare tutte le operazioni di analisi di sopravvivenza, e `gdata`, utile per caricare files in formato xls.

```
> library(survival)
> library(gdata)
```

Studio trapianto

Si crea il nuovo dataset attraverso il database in formato .xls.

```
> bmt = read.xls(file.choose())
> names(bmt)
> attach(bmt)
```

Si controlla se le variabili qualitative sono considerate tali da parte dell'ambiente R, ed eventualmente si ricodificano le variabili.

Il comando per la creazione del modello di Cox, dopo aver effettuato la selezione delle variabili, è il seguente:

```
> fit.cox = coxph(Surv(TEMPO,morto==1)~n_bmt+FONTE+PROF..VOD+VOD)
```

Per rappresentare la curva di sopravvivenza stimata attraverso il metodo Kaplan-Meier stratificata per VOD (mostrata in Figura 2.5), si esegue:

```
> s.hat = survfit(Surv(TEMPO,morto==1)~VOD)
> s.hat$strata
> plot(s.hat,col=3:4,lty=3:4,main="Curva sopravvivenza
  stratificata per VOD")
> legend(5000,0.3,c("VOD=0","VOD=1"),col=3:4,lty=3:4)
```

La procedura grafica di verifica assunto di proporzionalità (Figura 2.6(d)):

```
> strato = c(rep(0,263),rep(1,25))
> time = s.hat$time
> log_log_surv = log(-log(s.hat$surv))
> plot(time,log_log_surv,type="n")
> lines(s.hat$time[strato==0],log(-log(s.hat$surv[strato==0]))
  ,col=2,lty=2)
> lines(s.hat$time[strato==1],log(-log(s.hat$surv[strato==1]))
  ,col=3,lty=3)
```

il test di *Goodness Of Fit*:

```
> cox.zph(fit.cox)
```

ultimo test da effettuare risulta quello dell'utilizzo di una variabile tempo-dipendente. Quindi si creano due variabili dipendenti dal tempo

```
> VOD_T = VOD*TEMPO;
> VOD_LOGT = VOD*log(TEMPO);
```

e si creano i due modelli descritti 2.17(a) e (b)

```
> fit.cox.t = coxph(Surv(TEMPO,morto==1)~n_bmt+FONTE+PROF..VOD
  +VOD+VOD_T)
> fit.cox.logt = coxph(Surv(TEMPO,morto==1)~n_bmt+FONTE
  +PROF..VOD+VOD+VOD_LOGT)
```

e si verifica se VOD_T e VOD_LOGT risultano significative contro l'ipotesi nulla nel test di Wald.

Dal paragrafo 2.4.4 si ha che la variabile VOD non rispetta l'assunto di proporzionalità quindi si sceglie di stratificare per VOD, l'operazione di scelta delle variabili ridà le stesse variabili e il modello diventa:

```
> fit.cox.strata = coxph(Surv(TEMPO,morto==1)~strata(VOD)+n_bmt
  +FONTE+PROF..VOD)
```

Si passa ora a rappresentare la curva di sopravvivenza stimata del modello appena trovato (Figura 2.7):

```
> s.fit = survfit(fit.cox.strata)
> plot(s.fit,col=3:4,lty=3:4,main="Curva sopravvivenza modello
  stratificato")
```

Stima dei residui di Cox-Snell, di devianza per verificare la bontà del modello utilizzato: si inizia con il grafico della funzione di rischio cumulata dei residui (Figura 2.13(a))

```
> res.m = residuals(fit.cox.strata,type="mart")
> res.cs = morto-res.m
> s.res<-survfit(Surv(res.cs,morto)~VOD)
> s.res$strata
> strato = c(rep(0,196),rep(1,21))
> time = s.res$time
> log_surv = -log(s.res$surv)
> plot(time,log_surv,type="n",main="Grafico funzione di
  rischio cumulata dei residui")
> lines(s.res$time[strato==0],-log(s.res$surv[strato==0]),
  type="s",col=2,lty=2)
> lines(s.res$time[strato==1],-log(s.res$surv[strato==1]),
  type="s",col=4,lty=4)
> abline(0,1)
> legend(1.5,0.5,c("VOD=0","VOD=1"),col=c(2,4),lty=c(2,4))
```

e si passa alla rappresentazione del grafico dei residui di devianza (Figura 2.14(a) e 2.15(a))

```
> resi.d = residuals(fit.cox,type="dev",data=bmt)
> plot(resi.d,main="Grafico residui di devianza")
> plot(fit.cox$linear,resi.d,main="Grafico residui di devianza
  - Valori predetti")
```

Studio soggetto

Si crea il nuovo dataset attraverso il database in formato .xls.

```
> bmt = read.xls(file.choose())
> names(bmt)
> attach(bmt)
```

si controlla che tutte le variabili qualitative siano considerate fattori da parte dell'ambiente R.

Comando per la creazione del modello di Cox, dopo aver effettuato la selezione delle variabili.

```
> fit.cox = coxph(Surv(TEMPO,morto==1)~ETA+FONTE+PROF..VOD+VOD)
```

Rappresentazione della curva di sopravvivenza stimata attraverso il metodo Kaplan-Meier stratificata per VOD (mostrata in Figura 2.10):

```
> s.hat = survfit(Surv(TEMPO,morto==1)~VOD)
> s.hat$strata
> plot(s.hat,col=3:4,lty=3:4,main="Curva sopravvivenza
  stratificata per VOD soggetti")
> legend(5000,0.3,c("VOD=0","VOD=1"),col=3:4,lty=3:4)
```

Procedura grafica di verifica assunto di proporzionalità (Figura 2.11(c)):

```
> strato = c(rep(0,246),rep(1,23)) grafico
> time = s.hat$time > log_log_surv = log(-log(s.hat$surv))
> plot(time,log_log_surv,type="n")
> lines(s.hat$time[strato==0],log(-log(s.hat$surv[strato==0]))
  ,col=2,lty=2)
> lines(s.hat$time[strato==1],log(-log(s.hat$surv[strato==1]))
  ,col=3,lty=3)
> legend(5000,-3,c("VOD=0","VOD=1"),col=3:4,lty=2:3)
```

Test gof:

```
> cox.zph(fit.cox)
```

Ultimo test da effettuare risulta quello dell'utilizzo di una variabile tempo-dipendente quindi si creano due variabili dipendenti dal tempo

```
> VOD_T = VOD*TEMPO;
> VOD_LOGT = VOD*log(TEMPO);
```

Si creano i due modelli descritti 2.19(a) e (b)

```
> fit.cox.t = coxph(Surv(TEMPO,morto==1)~ETA+FONTE+PROF..VOD
  +VOD+VOD_T)
> fit.cox.logt = coxph(Surv(TEMPO,morto==1)~ETA+FONTE+PROF..VOD
  +VOD+VOD_LOGT)
```

e si verifica se VOD_T e VOD_LOGT risultano significative contro l'ipotesi nulla nel test di Wald:

Poichè la variabile VOD non rispetta l'assunto di proporzionalità, si sceglie di stratificare per VOD. L'operazione di scelta delle variabili fornisce le stesse variabili e il modello diventa:

```
> fit.cox.strata = coxph(Surv(TEMPO,morto==1)~strata(VOD)+ETA
  +FONTE+PROF..VOD)
```

VOD non rispetta l'assunto di proporzionalità. Rappresentazione della curva di sopravvivenza stimata del modello appena trovato (Figura 2.12):

```
> s.fit = survfit(fit.cox.strata)
> plot(s.fit,col=3:4,lty=3:4,main="Curva sopravvivenza modello
  stratificato soggetti")
```

Stima dei residui di Cox-Snell, di devianza per verificare la bontà del modello utilizzato. Grafico della funzione di rischio cumulata dei residui (Figura 2.13(b)):

```
> res.m = residuals(fit.cox.strata,type="mart")
> res.cs = morto-res.m
> s.res<-survfit(Surv(res.cs,morto)~VOD)
> s.res$strata
> strato = c(rep(0,260),rep(1,24))
> time = s.res$time
> log_surv = -log(s.res$surv)
> plot(time,log_surv,type="n",main="Grafico funzione
  di rischio cumulata dei residui soggetti")
> lines(s.res$time[strato==0],-log(s.res$surv[strato==0]),
  type="s",col=2,lty=2)
> lines(s.res$time[strato==1],-log(s.res$surv[strato==1]),
  type="s",col=4,lty=4)
> abline(0,1)
> legend(1.5,0.5,c("VOD=0","VOD=1"),col=c(2,4),lty=c(2,4))
```

Rappresentazione del grafico dei residui di devianza (Figura 2.14(b) e 2.15(b)):

```
> resi.d = residuals(fit.cox,type="dev",data=bmt)
> plot(resi.d,main="Grafico residui di devianza soggetti")
> plot(fit.cox$linear,resi.d,main="Grafico residui di devianza
  - Valori predetti soggetti")
```

A.2 Dati su linfoma ALCL

Si caricano i dati che fanno parte del gruppo casi e controlli

```
> controlli = c(0.23,0.44,0.19,0.08)
> casi = c(2.8,1.4,0.13,0.2,0.8,0.56,0.44,5.2,1.7,1.14)
```

Calcolo delle mediane nei due gruppi

```
> median(casi)
> median(controlli)
```

Analisi iniziale trovando il boxplot che raffigura una principale analisi per la differenza tra i due gruppi. Inoltre grafico quantile contro quantile per vedere un'eventuale assunzione di normalità

```
> boxplot(casi,controlli,names=c("casi","controlli"))
> qqnorm(casi)
> qqline(casi)
> qqnorm(controlli)
> qqline(controlli)
```

Test non parametrico di Kolmogorov-Smirnov

```
> ks.test(casi,pnorm)
> ks.test(controlli,pnorm)
> ks.test(casi,pexp)
> ks.test(controlli,pexp)
```

Assunzione Normale Vengono rappresentati gli istogrammi per casi e controlli aventi curva stimata con il metodo del nucleo e funzione di densità stimata di una normale:

```
> hist(controlli,probability=T)
> lines(density(controlli))
> lines(x,dnorm(x,mean(controlli),sqrt(var(controlli))),lty=2)
> hist(casi,probability=T)
> lines(density(casi))
> lines(x,dnorm(x,mean(casi),sqrt(var(casi))),lty=2)
```

Funzioni che servono per calcolare $\hat{R}$ sotto assunzione di distribuzione normale:

```
R_hat = function(medie, varianze,n){
temp_controlli = ((n[1]-1)/n[1])*varianze[1]
temp_casi = ((n[2]-1)/n[2])*varianze[2]
temp_R = (medie[2]-medie[1])/(sqrt(temp_controlli+temp_casi))
temp_R = pnorm(temp_R)
    temp_R
}
mean_casi = mean(casi)
mean_controlli = mean(controlli)
var_casi = var(casi)
var_controlli = var(controlli)
medie = c(mean_controlli,mean_casi)
varianze = c(var_controlli,var_casi)
n = c(length(controlli),length(casi))
R = R_hat(medie,varianze,n)
```

Assunzione Esponenziale Istogrammi per casi e controlli aventi curva stimata con il metodo del nucleo e funzione di densità stimata di una esponenziale:

```
> hist(controlli,probability=T)
> lines(density(controlli))
> lines(x,dexp(x,1),lty=2)
> hist(casi,probability=T)
> lines(density(casi))
> lines(x,dexp(x,10),lty=2)
```

Funzioni per calcolare $\hat{R}$ sotto assunzione di distribuzione esponenziale:

```
R_exp = function(casi,controlli){
    mean_casi = mean(casi)
    mean_controlli = mean(controlli)
temp = mean_casi/mean_controlli
    R_hat = 1/(1+temp)
    R_hat
}
R_exp(casi,controlli)
```

Ringraziamenti

L'autore vorrebbe ringraziare:

- La professoressa Laura Ventura per la costante disponibilità e cortesia avute nei miei confronti;
- Il Centro Oncoematologico Pediatrico di Padova, e soprattutto alla Dott. Tridello Gloria e Pillon Marta, per avermi dato la possibilità di fare un'esperienza di stage, molto interessante, costruttiva ed educativa, sia dal punto di vista professionale che umano.
- Il Dott. Raccanelli Giorgio che senza i suoi consigli non avrei mai intrapreso la carriera universitaria.
- Ai miei genitori, ai miei amici e alla mia fidanzata che mi sono sempre rimasti vicini.

Bibliography

[1] Clark, T.G. et al., (2003), "Survival Analysis Part I: Basic concepts and first analyses", *British Journal of Cancer*, 89, pp. 232–238.

[2] Collet, D. (2003) *Modelling survival data in medical research*, London, Chapman & Hall.

[3] Cortese, G. & Ventura, L. (2009) "Likelihood asymptotics for the stress-strength model P(X<Y)", Padova.

[4] Cox, D. R. & Oakes, D. (1984) *Analysis of Survival Data*, New York, Chapman & Hall.

[5] Gibbons, J.D. & Chakraborti, S. (2003) *Nonparametric statistical inference*, New York, Marcel Dekker.

[6] Hedeker, D.R. & Gibbons, R.D. (2006) *Longitudinal data analysis*, New York, J.Wiley & Sons.

[7] Kaplan, E.L. & Meier, P. (1958) "Nonparametric estimation from incomplete observations", *Journal of the American Statistical Association* 53: 457–481.

[8] Klein, J.P. & Moeschberger, M.L. (1997) *Survival analysis: techniques for censored and truncated data*, New York: Springer.

[9] Kleinbaum, D.G. & Klein, M. (2005) *Survival Analysis: A Self-Learning Text*, New York, Springer.

[10] Kolmogorov, A.N. (1933) *Sulla determinazione empirica di una legge di distribuzione*, Giornale dell'Istituto Italiano degli Attuari, 4, pp. 83-91.

[11] Kotz, S. et al. (2003) *The Stress–Strength Model and Its Generalizations: Theory and Applications*, Singapore, World Scientific.

[12] Marubini, E & Valsecchi, M.G. (2004), *Analysing Survival Data from Clinical Trials and Observational Studies,* New York, Statistics in Practice.

[13] Miller, L.H. (1956) *Table of percentage points of Kolmogorov Statistics*, Journal of the American Statistical Association, 51, pp. 111-121.

[14] Ott, L. & Longnecker, M. (2008) *An Introduction to Statistical Methods and Data Analysis*, Belmont, CA, Brooks/Cole Cengage Learning.

[15] Pett, M.A. (1997) *Nonparametric statistics for health care research: statistics for small sample and unusual distribution*, Thousand Oaks, CA, Sage Publications.

[16] Sheskin, D.J. (2004) *Handbook of parametric and nonparametric statistical procedures*, Boca Raton, FL, Chapman & Hall.

[17] Singer, J.D. & Willett, J.B. (2003) *Applied Longitudinal Data Analysis*, New York, Oxford University Press.

[18] Smirnov, N.V. (1939) *Estimate of deviation between empirical distribution function in two inpipendent samples*, Bulletin Moscow University 2, pp. 3-16.

[19] Therneau, T.M. & Grambsch, P.M. (2000) *Modeling Survival Data: Extending the Cox Model*, New York, Springer.

[20] Wayne W.D. (1996) *Biostatistica: concetti di base per l'analisi statistica delle scienze dell'area medico-sanitaria*, Napoli, EdiSES.