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

Statistica per l'Economia e l'Impresa

Modelli predittivi di sopravvivenza per l'apprendimento statistico nella Sclerosi Laterale Amiotrofica

Relatore: prof. Giuliana Cortese
Dipartimento di Scienze Statistiche

Laureanda: Erika Rettore

Matricola n. 1224560

Anno Accademico 2022/2023

Indice

Introduzione

La sclerosi laterale amiotrofica (SLA), è una malattia neurologica che colpisce le cellule nervose responsabili del controllo dei muscoli volontari. Nel corso della malattia, i motoneuroni degenerano progressivamente, causando una perdita graduale delle funzioni motorie e vitali.
Purtroppo, la SLA è caratterizzata da una grande varietà di sintomi e, soprattutto nelle fasi iniziali, è difficile diagnosticarla e fare una prognosi accurata.
Attualmente non esiste una terapia in grado di invertire o rallentare efficacemente la progressione della malattia.

Al fine di migliorare la situazione attuale diagnostica e prognostica, dovremmo progettare e sviluppare algoritmi di machine leaning in grado di predire la progressione della malattia in modo probabilistico, dipendente dal tempo. Nell'ambito di questa sfida, l'analisi si concentra sulla previsione di eventi avversi, tra cui la necessità di utilizzare la ventilazione non invasiva (NIV) o sottoporsi a una gastrotomia endoscopica percutanea (PEG), oltre alla morte (DEATH). L'obiettivo dell'analisi è quindi fornire una previsione degli eventi avversi, analizzando la sopravvivenza dei pazienti dalla diagnosi di SLA; Questo verrà svolto determinando se la variabile dipendente tempo di sopravvivenza è influenzata dalle variabili indipendenti presenti nel dataset come ad esempio genere, età, luogo di insorgenza, malattie già presenti ecc...
La tesi è organizzata nel modo seguente: Il primo capitolo descrive la malattia, le cause e i sintomi a essa associata; procedendo nel secondo capitolo troveremo una presentazione dei dati a disposizione per lo studio e l'analisi esplorativa dei questi ultimi; procediamo poi all'applicazione del modello nel terzo capitolo; il quarto capitolo è riservato all'applicazione di un modello di machine learning per fornire una previsione sulla malattia; alla fine dell'elaborato, vengono presentati i risultati ottenuti attraverso le analisi condotte e vengono tratte le conclusioni in base a tali risultati.

Capitolo 1

La Sclerosi Laterale Amiotrofica

La Sclerosi Laterale Amiotrofica (SLA) è una patologia neurodegenerativa progressiva che colpisce le cellule nervose responsabili del controllo dei muscoli volontari. E' conosciuta anche come malattia di Lou Gehrig, in onore del famoso giocatore di baseball statunitense che ne soffriva.
Si presenta come la degenerazione progressiva dei motoneuroni, che sono le cellule nervose che controllano i movimenti muscolari.
Questi ultimi si classificano in motoneuroni superiori, inferiori e misti.

- Motoneuroni superiori: sono situati nel sistema nervoso centrale (SNC), principalmente nella corteccia cerebrale (area motoria) e nel tronco cerebrale. Sono responsabili del controllo e della modulazione dell'attività dei motoneuroni inferiori. I segnali provenienti dai motoneuroni superiori scendono lungo le vie nervose chiamate tratti corticospinali e tratti corticonucleari per raggiungere i motoneuroni inferiori nella regione del midollo spinale o nel tronco cerebrale. I motoneuroni superiori svolgono un ruolo fondamentale nel coordinamento e nell'organizzazione dei movimenti volontari.

- Motoneuroni inferiori: si trovano nella porzione inferiore del sistema nervoso, inclusi il midollo spinale e il tronco cerebrale. Sono responsabili della trasmissione degli impulsi neurali ai muscoli scheletrici per innescare la contrazione muscolare e generare il movimento. I segnali provenienti dai motoneuroni superiori raggiungono i motoneuroni inferiori e da questi ultimi si propagano lungo gli assoni periferici per raggiungere e innervare i muscoli specifici. I motoneuroni inferiori sono responsabili dell'esecuzione e del controllo del movimento volontario.

- Motoneuroni misti: possono avere caratteristiche sia dei motoneuroni superiori che di quelli inferiori. Sono detti motoneuroni misti perché presentano connessioni e funzioni sia nel sistema nervoso centrale che nel sistema nervoso periferico.

La degenerazione dei motoneuroni porta ad una perdita graduale della capacità di controllare i muscoli volontari, incluso il movimento delle braccia, delle gambe, della bocca e della respirazione.

1.1 Classificazione

La SLA può essere classificata in base a diversi fattori, inclusi sintomi, segni clinici e i pattern di progressione della malattia.

1. **Classificazione clinica:** classificata in base ai soggetti colpiti dalla malattia. Le due forme principali sono:
    - SLA sporadica: è la forma più comune di SLA, che si verifica in modo casuale senza causa genetica nota. Comprende circa il 90-95% dei casi, manifestandosi principalmente tra i soggetti di età compresa tra i 40 e i 70 anni.
    - SLA familiare: rappresenta una piccola percentuale dei casi (5-10%), ed è caratterizzata da una possibile componente genetica. E' ereditata da uno o più geni mutati che possono essere trasmessi di generazione in generazione. Le persone affette da questo tipo di SLA hanno un rischio maggiore di sviluppare la malattia rispetto alla popolazione generale.
2. **Classificazione sintomatica:** classificazione in base all'insorgenza dei sintomi iniziali e alla distribuzione dei deficit neurologici. Alcune sottoclassificazioni includono:
    - Insorgenza bulbare: Nella forma bulbare della SLA, i sintomi iniziano nelle regioni del cervello coinvolte nel controllo dei muscoli della parola e della deglutizione. I motoneuroni superiori e inferiori coinvolti sono quelli localizzati nella regione del tronco cerebrale chiamata "bulbo". I sintomi possono includere difficoltà nella parola e nella deglutizione, cambiamenti nella voce, produzione eccessiva di saliva e affaticamento dei muscoli facciali.

- Insorgenza assiale: Nella forma assiale della SLA, i sintomi si manifestano principalmente nella regione del tronco cerebrale e del midollo spinale coinvolte nel controllo dei muscoli del tronco. I motoneuroni superiori e inferiori coinvolti sono quelli che controllano i muscoli della schiena, dell'addome e delle estremità superiori. I sintomi possono includere debolezza e atrofia dei muscoli del tronco, difficoltà nella respirazione e nella postura.
- Insorgenza agli arti : Nella forma degli arti (limbs) della SLA, i sintomi iniziano principalmente nei motoneuroni superiori e inferiori che controllano i muscoli degli arti, come braccia e gambe. I sintomi possono includere debolezza muscolare, crampi e difficoltà nel coordinamento e nel movimento degli arti.
- Insorgenza generalizzata: Nella forma generalizzata della SLA, i sintomi si manifestano in diverse regioni del sistema nervoso centrale, coinvolgendo sia i motoneuroni superiori che quelli inferiori in tutto il corpo. Questa forma presenta una progressione diffusa dei sintomi in diverse parti del corpo, coinvolgendo gli arti, il tronco e le funzioni vitali come la respirazione e la deglutizione.

Questi sono solo alcuni dei modi utilizzati per classificare la malattia e vengono utilizzate per descrivere e comprenderne meglio i diversi aspetti.

1.2 Sintomi e cure

I sintomi della SLA possono variare da persona a persona, ma ci sono alcune caratteristiche comuni associate alla malattia.

- Debolezza muscolare: La debolezza muscolare progressiva è uno dei sintomi principali della SLA, che può iniziare in una specifica regione del corpo (ad esempio mano, braccio o gamba) e poi diffondersi ad altre parti del corpo.
- Difficoltà nel parlare e nel deglutire: La SLA può causare disartria e disfagia a causa della progressiva debolezza dei muscoli coinvolti in queste funzioni
- Crampi muscolari e fascicolazioni

- Atrofia muscolare: La SLA porta alla progressiva atrofia (diminuzione del volume) dei muscoli a causa della degenerazione delle cellule nervose che li controllano.

- Difficoltà respiratorie: Con la progressione della malattia i muscoli respiratori si indeboliscono, portando a difficoltà respiratorie e alla necessità di assistenza ventilatoria.

E' una malattia incurabile e la sua progressione è generalmente rapida (due-quattro anni). La maggior parte delle persone colpite dalla SLA sviluppa una grave disabilità entro i primi anni dalla diagnosi. Sebbene attualmente non esista una cura, i trattamenti si concentrano principalmente sul miglioramento della qualità della vita del paziente e sulla gestione dei sintomi.
Alcune opzioni di trattamenti comuni includono:

- Gestione dei sintomi: Vengono utilizzati farmaci per affrontare sintomi come la spacisticità, i crampi muscolari, la depressione e la difficoltà di deglutizione.

- Terapia fisica e occupazionale: La terapia fisica può aiutare a mantenere la mobilità, la forza muscolare e la flessibilità. La terapia occupazionale può fornire supporto nell'affrontare le difficoltà quotidiane e trovare soluzioni per mantenere l'indipendenza.

- Dispositivi di assistenza: L'utilizzo di ausili e dispositivi di assistenza possono aiutare a compensare le difficoltà motorie e consentire al paziente di svolgere attività quotidiane.

- Terapia del linguaggio e della comunicazione: La terapia del linguaggio può aiutare a mantenere o migliorare le abilità di comunicazione, anche attraverso l'uso di dispositivi di comunicazione alternativi e aumentativi (AAC).

- Supporto psicologico: La SLA può avere un impatto emotivo significativo sul paziente e sulla famiglia. Il supporto psicologico, come la consulenza o il supporto di gruppo, può essere prezioso per affrontare le sfide emotive e psicologiche associate alla malattia.

Con la progressione della SLA, possono diventare necessarie terapie vitali per affrontare le sfide che la malattia presenta. Due di queste terapie sono la ventilazione non invasiva (NIV) e la gastrotomia endoscopica percutanea (PEG).
La **ventilazione non invasiva (NIV)** viene utilizzata quando i muscoli respiratori si indeboliscono, rendendo difficile la respirazione. Questa terapia prevede l'uso di una maschera che fornisce pressione positiva continua nelle vie aeree, migliorando

così la ventilazione polmonare e fornendo un adeguato supporto respiratorio. La NIV può contribuire a migliorare la qualità di vita, alleviare i sintomi respiratori e prolungare la sopravvivenza dei pazienti affetti da SLA.
La **gastrotomia endoscopica percutanea (PEG)** è una procedura chirurgica in cui viene inserito un tubo attraverso la parete addominale direttamente nello stomaco. Questo tubo consente l'alimentazione diretta nel tratto gastrointestinale, bypassando la necessità di deglutire il cibo. Con la progressione della SLA, la debolezza dei muscoli coinvolti nella deglutizione può rendere difficile l'assunzione adeguata di cibo e liquidi. La PEG fornisce un metodo sicuro ed efficace per garantire una nutrizione adeguata e l'idratazione del paziente.
La ventilazione non invasiva e la gastrotomia endoscopica percutanea sono terapie importanti per i pazienti con SLA in una fase più avanzata della malattia. Queste terapie mirano specificamente a gestire le difficoltà respiratorie e nutrizionali, contribuendo a mantenere la qualità di vita dei pazienti e a gestire le complicanze legate alla progressione della SLA.

La causa esatta della sclerosi laterale amiotrofica non è ancora completamente compresa. Tuttavia, sono stati identificati diversi fattori che potrebbero contribuire alla sua comparsa. Alcuni studi hanno evidenziato una componente genetica, con specifiche mutazioni genetiche associate a forme ereditarie della malattia. Inoltre, l'esposizione a determinate sostanze chimiche o metalli pesanti potrebbe aumentare il rischio di sviluppare la SLA. Si ritiene che un'anomala risposta del sistema immunitario e processi di neurodegenerazione siano coinvolti nella progressione della malattia. Durante la SLA, i neuroni motori, responsabili del controllo dei movimenti muscolari volontari, degenerano gradualmente, portando a debolezza muscolare e atrofia. La ricerca scientifica è in corso per comprendere meglio le cause esatte della malattia e per sviluppare nuovi trattamenti e strategie preventive.

1.3 Obiettivo dello studio

Nell'ambito di questa sfida, l'analisi si concentra sulla previsione di eventi avversi, tra cui la necessità di utilizzare la ventilazione non invasiva (NIV) o sottoporsi a una gastrotomia endoscopica percutanea (PEG), oltre alla morte (DEATH). Questi eventi sono di particolare importanza nella gestione della SLA, poiché possono influire sulla prognosi e sul benessere complessivo del paziente.
Classifichiamo quindi il rischio di compromissione in 3 attività:

- Compito 1: Ventilazione non invasiva (**NIV**) o DEATH (prendendo l'evento che si verifica per primo);
- Compito 2: Gastrotomia endoscopica percutanea (**PEG**) o DEATH (prendendo l'evento che si verifica per primo);
- Compito 3: **DEATH**

L'obiettivo dell'analisi è quindi fornire una previsione accurata degli eventi avversi per consentire ai medici di adottare le misure appropriate e garantire un supporto tempestivo e adeguato ai pazienti affetti da SLA.

Capitolo 2

I dati

Obiettivo principale dello studio di questi dati è analizzare la sopravvivenza dei pazienti dalla diagnosi di SLA, determinando se la variabile dipendente tempo di sopravvivenza è influenzata dalle variabili presenti nel dataset, quali il genere, l'età, il luogo di insorgenza ecc.. E' importante precisare che, come avviene molto frequentemente in ambito clinico, il dataset presenta dei dati censurati. Parliamo di censura quando i dati relativi ai tempi di sopravvivenza non forniscono un'osservazione completa dell'evento di interesse. Questo si verifica quando non sono presenti informazioni complete sulla durata del tempo di sopravvivenza per alcuni individui. La censura è di due tipi:

- Censura a destra: quando il tempo di osservazione termina prima che l'evento si verifichi per alcuni individui nello studio o nell'analisi. In altre parole, l'evento di interesse non si è ancora verificato per alcuni partecipanti alla fine del periodo di osservazione e questo può accadere per diversi motivi, come ad esempio la fine dello studio (quindi se il paziente non ha manifestato l'evento di interesse ma lo studio è finito), perdita di partecipanti (potrebbero ritirarsi dallo studio, essere persi nel follow-up o non essere disponibili per ulteriori osservazioni) ecc...

- Censura a sinistra: quando il tempo di inizio dell'osservazione di un individuo non è noto. Questo può accadere quando l'evento di interesse è già avvenuto prima dell'inizio dello studio, ma non si dispone di informazioni precise sul momento in cui è avvenuto.

Nel dataset sono presenti diverse variabili; andiamo quindi ad analizzare le più importanti.
La variabile *Occured* è una variabile dicotomica che indica con 1 se l'evento si è verificato o 0 se l'evento è stato censurato (a destra).

La variabile *Type* descrive invece il tipo di evento che si verifica (NIV, PEG, DEATH, NONE).
La variabile *Time*, descrive il tempo all'evento di interesse.
Le variabili *onsetDate* e *diagnosisDate* descrivono rispettivamente la data di insorgenza della malattia e la data di diagnosi.
La data di insorgenza indica il momento in cui i primi segni e sintomi della malattia compaiono e vengono percepiti dal paziente. È il punto iniziale in cui la malattia si manifesta e si presenta in modo riconoscibile.
La data di diagnosi invece, indica il momento in cui viene stabilita ufficialmente la presenza e la natura della malattia attraverso valutazioni cliniche, test diagnostici e consulenze mediche appropriate. La diagnosi può richiedere un certo periodo di tempo, durante il quale vengono eseguiti esami approfonditi, valutazioni specialistiche e analisi dei sintomi e dei risultati dei test. La data di diagnosi rappresenta quindi il momento in cui il paziente ottiene una conferma medica della sua condizione.
Nel nostro dataset le date di diagnosi e di insorgenza vengono rappresentate con segno negativo perché indicano il tempo trascorso tra un evento di riferimento (l'inizio dello studio) e l'evento di interesse (la diagnosi della malattia o l'insorgenza dei sintomi).
L'uso del segno negativo è una convenzione comune nell'analisi di sopravvivenza per distinguere le variabili temporali verso l'evento di interesse (diagnosi o insorgenza) rispetto al tempo trascorso dal punto di riferimento.
Nei casi in cui *diagnosisDate* è pari a 0, procediamo all'eliminazione della riga. Questo perché siamo nel caso di una censura a sinistra che si verifica quando il tempo di insorgenza dell'evento di interesse (nel nostro caso la diagnosi della malattia) è sconosciuto o non è stato osservato.
Le modalità per trattare una censura a sinistra sono diverse a seconda dei casi; nel nostro caso specifico procediamo all'eliminazione perché la censura è presente solo in una piccola percentuale dei campioni e non influisce significativamente sulla rappresentatività del dataset.
Procediamo con l'analisi delle variabili più importanti; creiamo una nuova variabile chiamata *insorgenza* che rappresenta il sito di insorgenza della malattia, raggruppando le quattro variabili che la descrivevano.

- **onset_bulbar** : una dilatazione o espansione arrotondata in un canale, un vaso o un organo

- **onset_axial** : La parte dello scheletro che comprende il cranio, la colonna vertebrale, lo sterno e le costole

- **onset_generalized** : Diffuso, largamente disperso, comune
- **onset_limbs** : Una regione del corpo che si riferisce a un'estremità superiore o inferiore

Allo stesso modo facciamo per le variabili che descrivono la presenza della malattia nel motoneurone.
Queste variabili sono tutte e tre variabili binarie, codificate quindi con 0 se l'evento non si verifica e 1 altrimenti.
Dato che descrivono lo stesso evento, abbiamo deciso di creare una nuova variabile qualitativa (*motoneurone*), che raggruppi queste tre variabili.

Come accennato inizialmente, i dataset di cui disponiamo riguardano tre attività: NIV, PEG, DEATH, che sono gli eventi avversi da prevedere nella sfida.
Andremo a svolgere le analisi in maniera separata per ogni attività e successivamente procederemo a confrontarle.
Tuttavia i rischi dei diversi tipi di evento non sono indipendenti, poiché sarebbe da analizzare gli eventi come eventi multipli e quindi considerando i rischi competitivi. Questo però richiede numerose analisi, ed essendo interessati principalmente alle prestazioni di predizione e non la relazione tra queste attività, nella nostra analisi verranno considerati come eventi separati.

2.1 Analisi esplorativa

Il dataset A contiene 1116 osservazioni, ciascuna rappresentante un paziente identificato con un codice esadecimale univoco, e 30 variabili che descrivono aspetti diversi legati alla malattia.
L'evento avverso che viene affrontato in questo set di dati è NIV o DEATH, a seconda di quale dei due eventi si verifica per primo.
Per quanto riguarda il dataset B invece abbiamo 1211 pazienti, e l'evento avverso da prevedere è PEG o DEATH (prendendo sempre l'evento che si verifica per primo).
Infine il dataset C contiene 1239, studiati grazie alle 30 variabili disponibili, e l'evento da prevedere nella sfida è DEATH.

L'obiettivo di questo capitolo è presentare una prima analisi generale delle variabili raccolte (qualitative e quantitative) per i pazienti affetti da sclerosi laterale amiotrofica (SLA). Nelle tabelle sottostanti, sono riportati i riassunti delle variabili, ottenuti con R, per ogni set di dati.

VARIABILI QUALITATIVE			
VAR	Categoria	N° oss	%
Type	DEATH	524	46.95%
	NIV	482	43.19%
	NONE	110	9.86%
motoneurone	superiore	193	17.30%
	inferiore	322	28.85%
	misto	601	53.85%
sex	M	564	50.54%
	F	552	49.46%
insorgenza	bulbar	361	32.35%
	limbs	750	67.20%
	axial	2	0.18%
	generalized	3	0.27%

VARIABILI QUANTITATIVE						
VAR	MIN	1° Quart.	Mediana	Media	3° Quart.	MAX
Time	1.60	11.93	19.27	27.87	33.72	190.17
onsetDate	-165.33	-16.28	-10.18	-14.11	-6.17	13.83
diagnosisDate	-162.27	-1.00	-0.63	-2.197	-0.33	17.90
height	1.36	1.58	1.65	1.64	1.70	1.98
weight_before_onset	39.00	60.00	70.00	70.25	79.00	133.00
weight	37.00	58.00	66.00	67.12	76.00	133.00
age_onset	20.33	57.30	65.24	63.93	71.92	88.81

Tabella 2.1: Riassunto delle variabili qualitative e quantitative per il dataset A

Il campione per il dataset A è composto da 564 maschi (50.54%) e 552 femmine (49.46%). L'età considerata è compresa tra 20.33 e 88.81 anni, con una media di 63.93 anni. Dalla tabella 2.1 notiamo che in media l'insorgenza della malattia si verifica circa dopo 14 mesi l'inizio dello studio, e viene diagnosticata solamente dopo 2 mesi.
Questo vuol dire che la malattia viene diagnosticata ma i sintomi in media si verificano in un momento successivo.

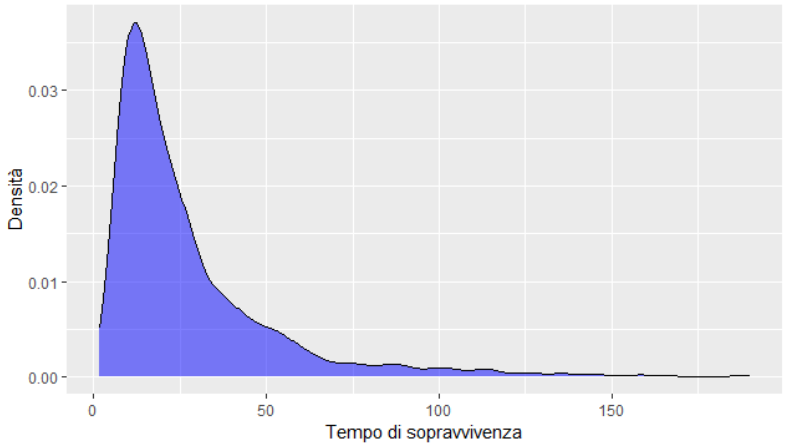

Figura 2.1: densità del tempo di sopravvivenza

Come notiamo dal grafico della densità dei tempi di sopravvivenza rappresentato in figura 2.1 è molto asimmetrica. La distribuzione è molto ripida ed è spostata verso sinistra, il che indica una maggior concentrazione di casi con tempi di sopravvivenza più brevi. Tuttavia, è interessante notare che la presenza di una coda molto lunga a destra indica anche la presenza di alcuni casi con tempi di sopravvivenza molto elevati. Questo suggerisce una variazione significativa nella durata della sopravvivenza dei soggetti nel nostro studio.

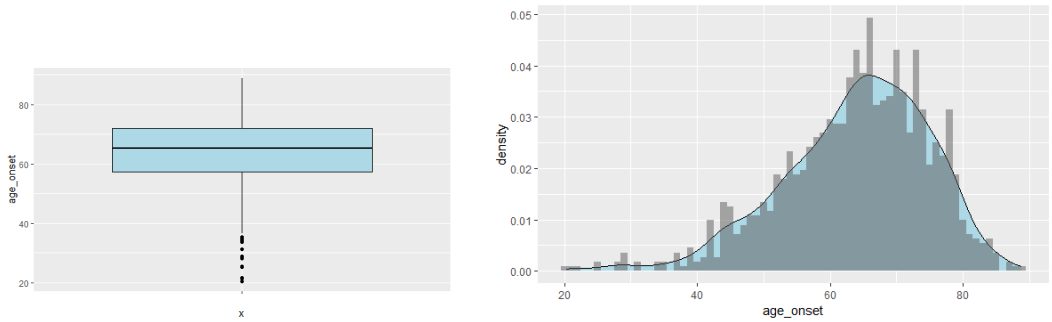

Figura 2.2: Istogramma e boxplot della variabile *age_ onset*

Dal boxplot dell'età, si può vedere che la fascia d'età considerata critica è tra i 59 e i 71 anni (1° e 3° quartile). Possiamo notare che tutti i valori inferiori a 35 vengono considerati anomali; questo perché non è comune osservare una complicanza della malattia in quella fascia d'età. Dall'istogramma invece, si nota un'asimmetria nella distribuzione di questa variabile. Il test di Shapiro_Wilk porta al rifiuto dell'ipotesi di normalità (W = 0.9738, p-value = 1.299e-13).

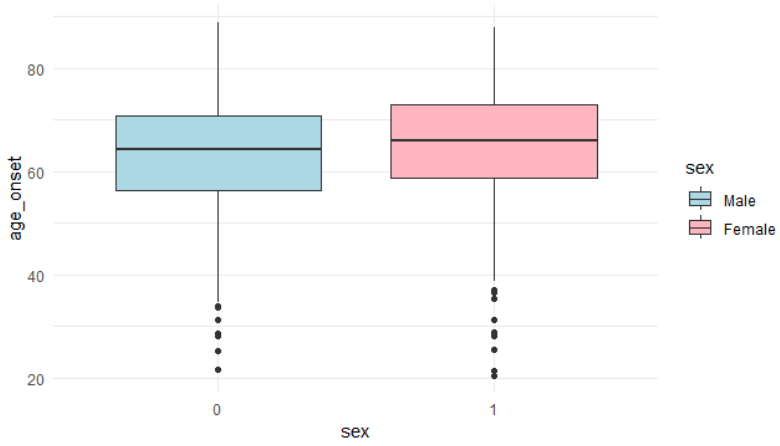

Figura 2.3: Boxplot dell'età distinta per sesso

Si procede quindi esaminando la variabile *age_ onset* per genere, per verificare un'eventuale differenza tra maschi e femmine (Figura 2.3). Come possiamo notare, non ci sono grandi differenze. Applichiamo ora il test di Wilcoxon basato sui ranghi per andare a verificare che non ci sia differenza, che è quello che ci aspettiamo. Questo test confronta le distribuzioni dei due gruppi utilizzando i ranghi dei dati anziché i valori reali. Questo significa che il test tiene conto dell'ordine dei valori piuttosto che dei valori numerici effettivi. Infatti anche se i boxplot sono molto simili, i risultati del test indicano che le distribuzioni dei due gruppi sono statisticamente diverse (W = 139422, p-value = 0.002553). Il valore p del test , che è inferiore al livello di significatività comune di 0.05, ci porta a rigettare l'ipotesi nulla che le

due distribuzioni siano simili e concludere che esiste un'associazione significativa tra il genere e l'età di insorgenza. Diamo ora uno sguardo alle variabili *insorgenza* e *motoneurone*, che sono state realizzate raggruppando le variabili che descrivevano queste caratteristiche.

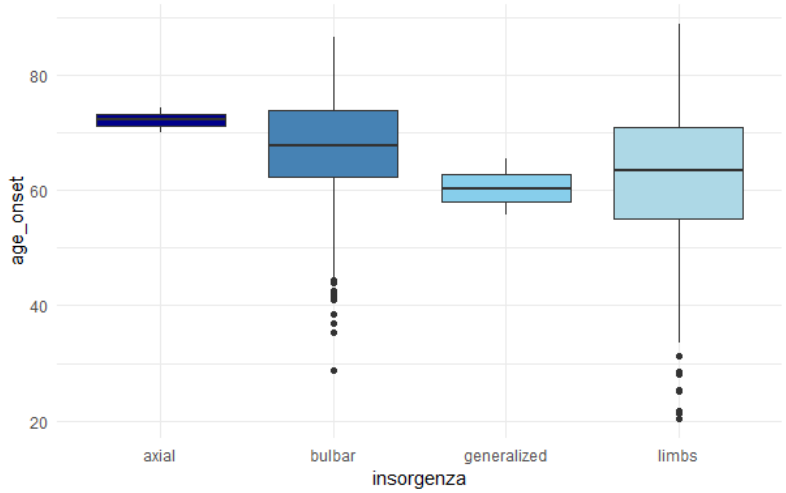

Figura 2.4: Boxplot del luogo di insorgenza della malattia

Come possiamo notare dalla Figura 2.4, per quanto riguarda la variabile *insorgenza* le modalità presenti nel dataset non si manifestano in modo proporzionale. La modalità axial (0.17%) e generalized (0.26%) sono poco presenti, a differenza invece delle modalità bulbar (32.70%) e limbs (66.87%) che invece descrivono quasi totalmente i luoghi di insorgenza della malattia. Può essere interessante vedere come queste modalità si manifestano in base al genere (Figura 2.5).

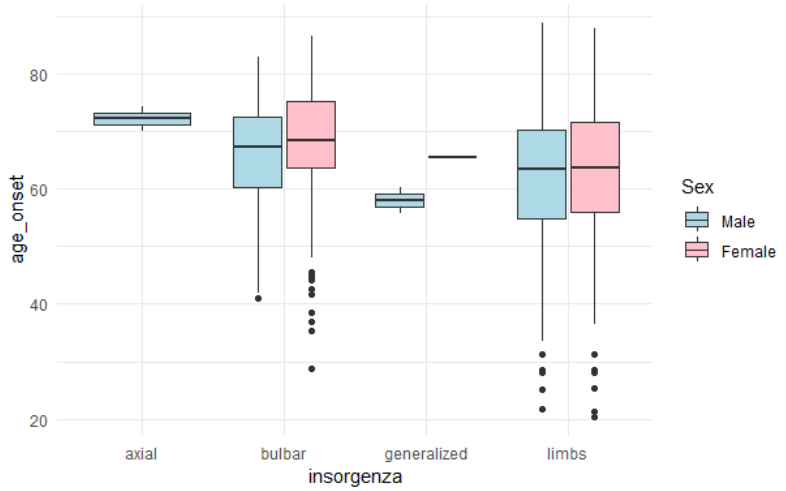

Figura 2.5: Boxplot del luogo di insorgenza della malattia diviso per sesso

Notiamo che axial e generalized, che sono le modalità più esigue, si manifestano solo nel genere maschile, mentre per le altre due modalità rimane abbastanza proporzionale. Confermiamo questa proporzionalità di limbs e bulbar con il test di Wilcoxon per vedere se la distribuzione dei dati è la stessa per i due generi. Per quanto riguarda bulbar il test (W = 14710, p-value = 0.03004) mi porta a rifiutare l'ipotesi di uguaglianza della distribuzione con una soglia di 0.05. Diverso è per limbs in quanto il valore p di 0.2478 ci porta all'accettazione dell'ipotesi nulla.
Andiamo ora a vedere come si distribuisce la malattia nei motoneuroni.

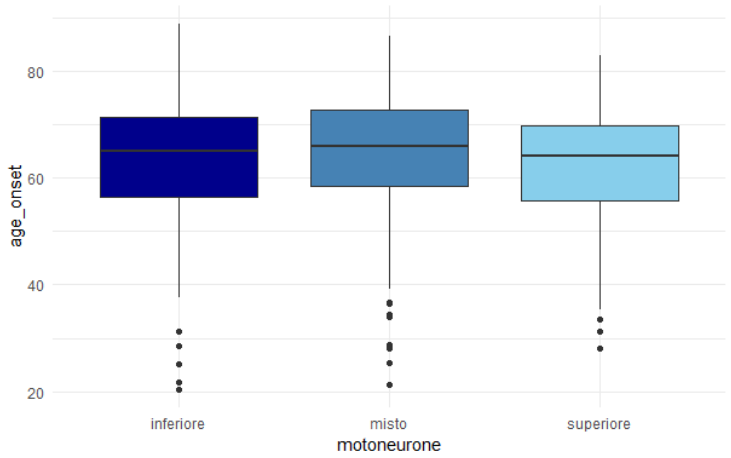

Figura 2.6: Boxplot del motoneurone coinvolto

La figura 2.6 non mette in evidenza le differenze della presenza dei tre motoneuroni, come invece possiamo notare dalla tabella 2.1, la quale ci permette di dire che la malattia si verifica in maniera più frequente nella modalità *misto* (53.85%).
Diverso invece è se guardiamo il grafico 2.7, che invece ci mette in risalto questa differenza.

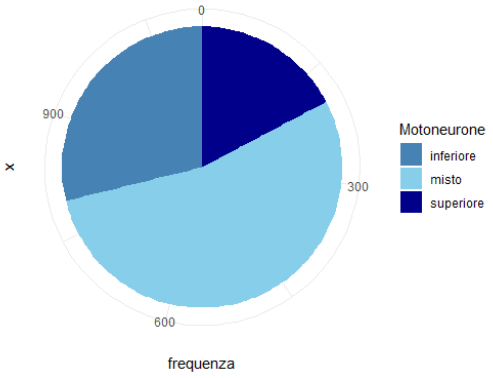

Figura 2.7: Grafico a torta delle modalità della variabile motoneurone

Anche in questo caso andiamo a vedere come queste modalità si manifestano in base al genere (figura 2.8).

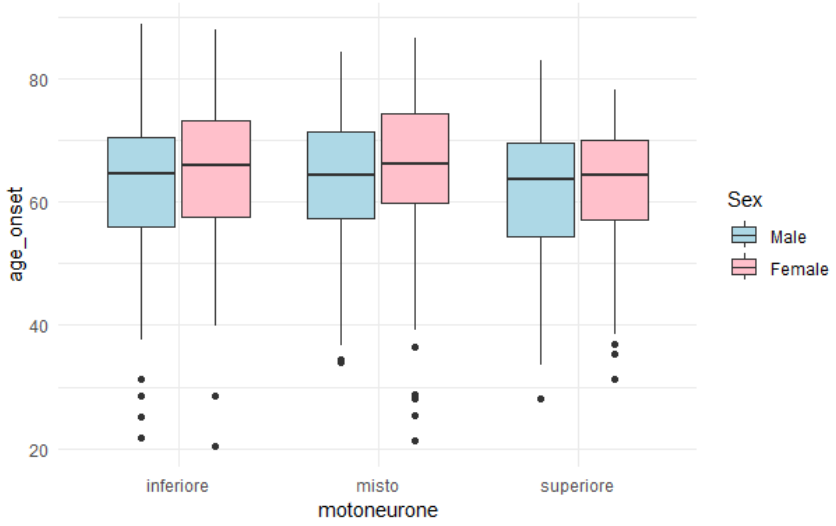

Figura 2.8: Boxplot del motoneurone coinvolto diviso per sesso

Notiamo che si distribuiscono in maniera abbastanza proporzionale; infatti non notiamo particolari differenze tra maschi e femmine, ad eccezione della media che per il genere femminile è un po più alta.
Confermiamo la proporzionalità con il test di Wilcoxon.
Per quanto riguarda la modalità *superiore* e *inferiore*, il test ci porta ad accettare l'ipotesi di proporzionalità, con un p-value rispettivamente di 0.4687 e 0.1345, entrambi superiori alla soglia presa in considerazione di 0.05.
Diverso è il caso per la modalità *misto*, in quanto il test ci porta a rifiutare la proporzionalità (W = 39537, p-value = 0.01125).

Per quanto riguarda il dataset B e C, la struttura dei dati è la stessa rispetto al dataset A, pertanto i grafici e i test risultano molto simili.
Procediamo quindi riportando le tabelle con i resoconti delle variabili qualitative e quantitative.

VARIABILI QUALITATIVE			
VAR	Categoria	N° oss	%
Type	DEATH	650	53.67%
	PEG	397	32.78%
	NONE	164	13.55%
motoneurone	superiore	196	16.18%
	inferiore	384	31.71%
	misto	631	52.11%
sex	M	639	52.77%
	F	572	47.23%
insorgenza	bulbar	366	30.22%
	limbs	829	68.46%
	axial	12	0.99%
	generalized	4	0.33%

VARIABILI QUANTITATIVE						
VAR	MIN	1° Quart.	Mediana	Media	3° Quart.	MAX
Time	1.13	12.78	20.83	30.03	37.90	190.17
onsetDate	-165.33	-16.40	-10.20	-14.06	-6.17	13.83
diagnosisDate	-162.27	-1.03	-0.67	-2.21	-0.33	17.90
height	1.36	1.58	1.65	1.65	1.71	1.98
weight_before_onset	39.00	61.00	70.00	70.80	80.00	133.00
weight	37.00	58.00	67.00	67.53	76.00	133.00
age_onset	20.33	57.40	65.28	63.92	71.77	88.81

Tabella 2.2: Riassunto delle variabili qualitative e quantitative per il dataset B

VARIABILI QUALITATIVE			
VAR	Categoria	N° oss	%
Type	DEATH	1062	85.71%
	NONE	177	14.29%
motoneurone	superiore	205	16.55%
	inferiore	379	30.59%
	misto	655	52.86%
sex	M	649	52.38%
	F	590	47.62%
insorgenza	bulbar	405	32.69%
	limbs	818	66.02%
	axial	12	0.97%
	generalized	4	0.32%

VARIABILI QUANTITATIVE						
VAR	MIN	1° Quart.	Mediana	Media	3° Quart.	MAX
Time	1.13	14.12	23.30	31.61	39.90	190.17
onsetDate	-165.33	-16.12	-10.20	-13.93	-6.13	13.83
diagnosisDate	-162.27	-1.03	-0.67	-2.21	-0.37	17.90
height	1.36	1.58	1.65	1.65	1.71	1.98
weight_before_onset	39.00	61.00	70.00	70.64	79.00	145.00
weight	37.00	58.00	66.00	67.19	75.00	133.00
age_onset	20.33	57.89	65.60	64.24	72.17	88.81

Tabella 2.3: Riassunto delle variabili qualitative e quantitative per il dataset C

Come nel dataset A, i maschi sono leggermente di più rispetto alle femmine, e la presenza della malattia e il luogo di insorgenza, si distribuisce allo stesso modo in tutti e tre i set di dati.
Ugualmente succede per le variabili quantitative.

Per questo motivo, non sono stati riportati i grafici e i test svolti per il dataset A, poichè si ricorrebbe ad una ripetizione.

Capitolo 3

Analisi di sopravvivenza

L'analisi di sopravvivenza è un approccio statistico utilizzato per studiare il tempo tra un evento di partenza e un evento di interesse.
Questo tipo di analisi è comunemente utilizzato in campo medico per esaminare la sopravvivenza dei pazienti dopo una diagnosi o un trattamento.

3.1 Funzioni principali

La funzione di sopravvivenza, indicata come $S(t)$, ci fornisce la probabilità che un individuo sopravviva oltre un certo periodo di tempo t e viene indicata con

$$S(t) = 1 - F(t) = P(T > t) = \int_t^\infty f(s)ds, \tag{3.1}$$

con $F_T(t)$ funzione di ripartizione di T.
La funzione di ripartizione ci fornisce la probabilità che un individuo sperimenti l'evento di interesse entro un certo periodo di tempo t. Quest'ultima è definita come

$$F(t) = Pr(T \leq t) = \int_0^t f(s)ds, \tag{3.2}$$

È importante sottolineare che la funzione di sopravvivenza e la funzione di ripartizione sono strettamente legate. In particolare, la probabilità di sopravvivenza S(t) può essere calcolata sottraendo a 1 la probabilità di sperimentare l'evento di interesse F(t).

3.1.1 Funzione di rischio

I modelli per l'analisi di sopravvivenza sono utilizzati per valutare la relazione tra variabili predittive e il tempo di sopravvivenza. Questi modelli si basano sulla funzione di azzardo, indicata come $\lambda(t)$ (o *hazard function*), che svolge un ruolo fondamentale nell'analisi dei dati di durata, definita come

$$\lambda(t) = \lim_{\Delta_t \to 0^+} \frac{Pr\{t \leq T < t + \Delta_t | T \geq t\}}{\Delta_t} = \frac{f(t)}{S(t)} \tag{3.3}$$

con $f(t)$ densità di probabilità della variabile casuale T che descrive i tempi di sopravvivenza dei pazienti definita come

$$f(t) = \frac{dF(t)}{dt} = \frac{-dS(t)}{dt} \tag{3.4}$$

La funzione di azzardo fornisce informazioni sul rischio istantaneo che un evento si verifichi nell'intervallo di tempo compreso tra $[t \; e \; t+\Delta t]$, condizionatamente al fatto che il soggetto è ancora vivo al tempo t.

Oltre alla funzione di azzardo, un altro concetto importante nell'analisi di sopravvivenza è la funzione cumulativa di rischio, indicata come Λ.
La funzione cumulativa di rischio accumula il rischio istantaneo lungo tutto l'intervallo di tempo considerato, fornendo una stima della probabilità complessiva di sperimentare l'evento fino a un dato momento. Viene definita come

$$\Lambda(t) = \int_0^t \lambda(t) dt \tag{3.5}$$

Possiamo osservare il legame con la funzione di sopravvivenza e definire la funzione di rischio cumulativo come

$$\Lambda(t) = -logS(t) \tag{3.6}$$

3.1.2 Stima delle funzioni di base

Sia dato un campione di n soggetti. Si considera il caso in cui ciascun episodio termina il verificarsi dell'evento di interesse o con una censura a destra, dovuta al periodo di osservazione considerato (alcuni individui non sperimentano l'evento prima della fine dello studio) o altre cause (per esempio l'uscita dal follow-up o il decesso del soggetto). Si assume che il meccanismo di censura sia indipendente dal processo di formazione dei tempi di sopravvivenza (censura non informativa).

L'obiettivo dell'analisi di sopravvivenza può essere:

- stimare e interpretare funzioni di sopravvivenza e l'azzardo a partire dai dati di sopravvivenza;
- confrontare funzioni di sopravvivenza e l'azzardo in gruppi di soggetti con caratteristiche diverse;
- analizzare i fattori che influenzano la durata e di stimare la sopravvivenza di individui o gruppi di soggetti con particolari caratteristiche.

Un problema molto frequente nell'analisi di sopravvivenza è la presenza di osservazioni mancanti o incomplete, conosciute come osservazioni censurate.
Abbiamo tre classi principali di censura:

1. Censura di 1° tipo: i soggetti sono osservati per un periodo di tempo fissato.
2. Censura di 2° tipo: la lunghezza dello studio non è fissato a priori in quanto è fissato il numero totale di fallimenti.
3. Censura casuale: è fissato il periodo di osservazione ma i soggetti entrano in tempi differenti.

La bontà delle stime è relazionata al numero di eventi e non al numero di osservazioni: maggiore è il numero di valori non censurati migliori saranno le stime dei coefficienti.

3.2 Modelli parametrici, non parametri e semi parametrici

È naturale scegliere una distribuzione statistica che non ha supporto negativo, in quanto i tempi di sopravvivenza sono positivi.

3.2.1 Modelli parametrici

I modelli parametrici richiedono che la distribuzione del tempo di sopravvivenza sia nota e la funzione d'azzardo sia completamente specificata, ad esclusione dei valori di alcuni parametri. Abbiamo diverse distribuzioni che possono essere usate per modellarla. Nel nostro caso sono state applicate le distribuzioni Esponenziale(con funzione di rischio costante nel tempo), Weibull e Lognormale (con funzioni di rischio monotone nel tempo). Prima di procedere all'analisi vediamo le funzioni principali per ogni distribuzione trattata.

Funzioni principali			
Distribuzione	$f(t)$	$S(t)$	$h(t)$
Esponenziale	$\frac{1}{\lambda}e^{-\left(\frac{t}{\lambda}\right)}$	$e^{-\lambda t}$	λ
Weibull	$\frac{\alpha}{\lambda^{\alpha}}t^{\alpha-1}e^{-\left(\frac{t}{\lambda}\right)^{\alpha}}$	$e^{-\lambda t^{\alpha}}$	$\alpha\lambda t^{\alpha-1}$
Lognormale	$\frac{1}{t\cdot\sigma\sqrt{2\pi}}\cdot e^{-\frac{(\ln(t)-\mu)^2}{2\sigma^2}}$	$1-\Phi\left(\frac{\ln(t)-\mu}{\sigma}\right)$	$\frac{1}{t\cdot\sigma\sqrt{2\pi}}\cdot e^{-\frac{(\ln(t)-\mu)^2}{2\sigma^2}}\cdot\frac{1}{1-\Phi\left(\frac{\ln(t)-\mu}{\sigma}\right)}$

Tabella 3.1: Funzioni principali delle distribuzioni esponenziale e Weibull.

Vediamo ora come le tre distribuzioni si adattano ai dati dei dataset A, B e C.

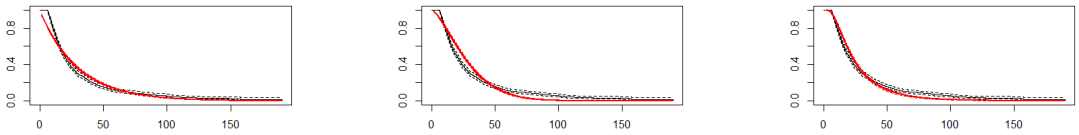

Figura 3.1: Modello Esponenziale, Weibull e Log-normale per il dataset A

I grafici ci mostrano che sia la distribuzione esponenziale che quella log-normale si adattano bene ai dati.
Andiamo quindi a confrontare l'indice di Akaike (AIC) per capire quale distribuzione si adatta meglio; sceglieremo quella con AIC inferiore.

Confronto tra modelli		
Distribuzione	AIC	BIC
Esponenziale	8640.533	8760.953
Weibull	8425.908	8551.346
Log-normale	8239.1	8364.537

Tabella 3.2: Confronto tra modelli parametrici per il dataset A

Come ci evidenzia la tabella 3.2 la distribuzione log-normale è quella che ci fornisce un AIC e un BIC inferiore, portandoci a preferire questo modello rispetto alla distribuzione esponenziale e di weibull.

Vediamo anche per gli altri dataset il modello parametrico che si adatta meglio ai dati.

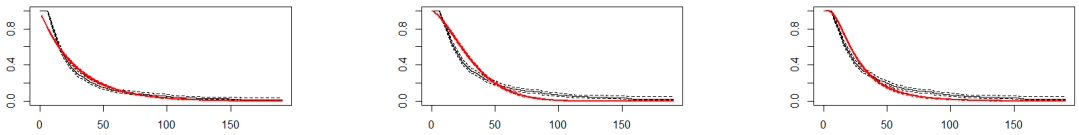

Figura 3.2: Modello Esponenziale, Weibull e Log-normale per il dataset B

Anche per il dataset B sembra che il modello log-normale è quello che si adatta meglio ai dati;
Andiamo a confrontare i risultati ottenuti con il test AIC e BIC.

Confronto tra modelli		
Distribuzione	AIC	BIC
Esponenziale	9066.614	9188.995
Weibull	8799.318	8926.798
Log-normale	8639.985	8767.465

Tabella 3.3: Confronto tra modelli parametrici per il dataset B

Anche in questo caso i test ci portano a preferire il modello log-normale.

Vediamo ora se anche per il dataset C la scelta del modello è la stessa.
Ci aspettiamo che sia uguale ai due casi precedenti poiché la struttura dei dati è la stessa, come abbiamo già notato in precedenza nell'analisi esplorativa.

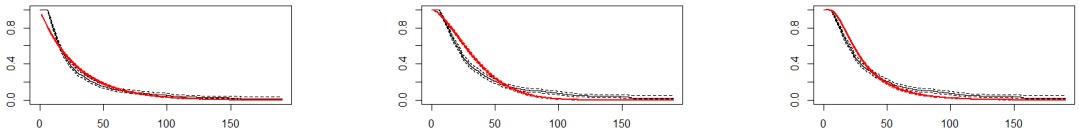

Figura 3.3: Modello Esponenziale, Weibull e Log-normale per il dataset C

Come per i due casi precedenti, graficamente il modello log-normale è quello segue più fedelmente la struttura dei dati.

Confronto tra modelli		
Distribuzione	AIC	BIC
Esponenziale	9305.395	9428.324
Weibull	8981.337	9109.389
Log-normale	8840.685	8968.737

Tabella 3.4: Confronto tra modelli parametrici per il dataset C

Come ci aspettavamo andiamo a preferire il modello log-normale anche per il dataset C che presenta risultati dei test AIC e BIC inferiori rispetto al modello esponenziale e di weibull.

È importante notare che l'uso di un modello parametrico richiede l'assunzione di una distribuzione specifica per i tempi di sopravvivenza. Se l'assunzione della distribuzione non è adeguata per i dati in esame, il modello parametrico può produrre risultati inaccurati. Pertanto, considerando che lo scopo principale della nostra analisi è fare previsioni sul verificarsi di un evento avverso, allora il modello più appropriato potrebbe essere il modello a rischi proporzionali (proportional hazards model) nell'analisi di sopravvivenza.
Il modello a rischi proporzionali assume che le variabili indipendenti abbiano un effetto proporzionale sul rischio di evento nel tempo. Ciò significa che il rapporto tra i rischi istantanei di due individui con valori diversi di una variabile indipendente rimane costante nel tempo. Il modello a rischi proporzionali è spesso basato sul modello di regressione di Cox, che non richiede specificazioni sulla distribuzione del tempo di sopravvivenza in quanto è un modello semi-parametrico Prima di procedere con l'analisi di questo modello andiamo a vedere come funzionano i modelli non parametrici.

3.2.2 Modelli non parametrici

I modelli non parametrici sono molto utilizzati in quanto non è necessario fare assunzioni sulla forma della funzione di rischio.
Lo strumento statistico più utilizzato per questo tipo di analisi è lo stimatore Kaplan-Meier (K-M).
L'analisi di Kaplan-Meier ci consente di stimare la sopravvivenza cumulativa nel tempo e di confrontare le curve di sopravvivenza tra i diversi gruppi o categorie. Questo tipo di analisi è particolarmente utile per fornire una descrizione visiva delle differenze di sopravvivenza tra i gruppi e per identificare eventuali differenze significative utilizzando il test di log-rank o altri test appropriati.
Come accennato nel paragrafo precedente, più essere utile utilizzare strumenti in

grado di gestire i dati censurati, e lo stimatore K-M è uno di questi.
Il concetto di base di questo stimatore è la probabilità condizionata, cioè la probabilità di sopravvivere fino ad uno specificato momento condizionata alla probabilità di essere vivo nei precedenti periodi temporali.
Per spiegare correttamente i dati andiamo a verificare che le assunzioni di K-M siano verificate:

- La prima assunzione prevede che i dati seguano una censura indipendente dal gruppo di classificazione;
- La seconda assunzione prevede un numero limitato di dati censurati in quanto questi ultimi influiscono sulla stima della curva K-M dato che diminuisce il numero di pazienti a rischio, rendendo la stima di sopravvivenza meno precisa di quanto non si avrebbe in presenza di un numero ridotto di dati censurati;
- La terza assunzione invece prevede un campione di dimensione sufficientemente grande per avere una maggiore precisione.

Ordinando i tempi relativi ad eventi accaduti tra gli n soggetti, avremo un ordinamento del tipo: $t_1 < t_2 < ... < t_j$, con $J \leq n$.
Sia d_j il numero di decessi che avvengono al tempo t_j e n_j il numero di soggetti a rischio al tempo $t_j (j = 1...J)$, la probabilità p_j di sopravvivere oltre il tempo t_j, condizionatamente all'essere sopravvissuti fino all'istante precedente a t_j, è stimata da

$$\hat{p}_j = \frac{n_j - d_j}{n_j} = 1 - \hat{q}_j, \tag{3.7}$$

dove $\hat{q}_j = \frac{d_j}{n_j}$ è la stima della probabilità condizionata di subire l'evento al tempo t_j, con $j = 1...J$.
La funzione di sopravvivenza che viene quindi stimata è un prodotto di probabilità di sopravvivenza, data da

$$\hat{S}_{KM}(t) = \prod_{t_j \leq t} \frac{n_j - d_j}{n_j}, \tag{3.8}$$

che cambia valore solo quando si verifica un evento.

Analisi dataset A

Partiamo da un'analisi generale senza inclusione di covariate per fornire una visione generale della distribuzione dei tempi di sopravvivenza nel campione e identificare eventuali differenze significative nella sopravvivenza tra i gruppi di interesse.

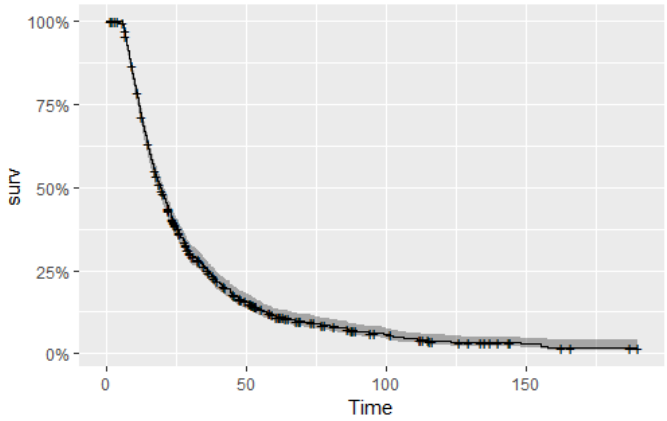

Figura 3.4: Curva di sopravvivenza Kaplan-Meier per un modello di sopravvivenza di base (senza covariate)

Sull'asse delle x abbiamo il tempo trascorso dal momento iniziale dell'osservazione; sull'asse delle y abbiamo la probabilità stimata di sopravvivenza, che rappresenta la proporzione di individui che non hanno ancora sperimentato l'evento di interesse nel corso del tempo.
La linea del grafico rappresenta la curva di sopravvivenza stimata. Inizialmente la curva parte da 1, che rappresenta una probabilità del 100% di sopravvivenza all'inizio del tempo di osservazione. Man mano che trascorre il tempo, la curva scende, indicando una diminuzione della probabilità di sopravvivenza.
Andiamo ora a studiare alcune delle variabili che riteniamo più interessanti per l'analisi e vediamo come queste influenzano il tempo di sopravvivenza.

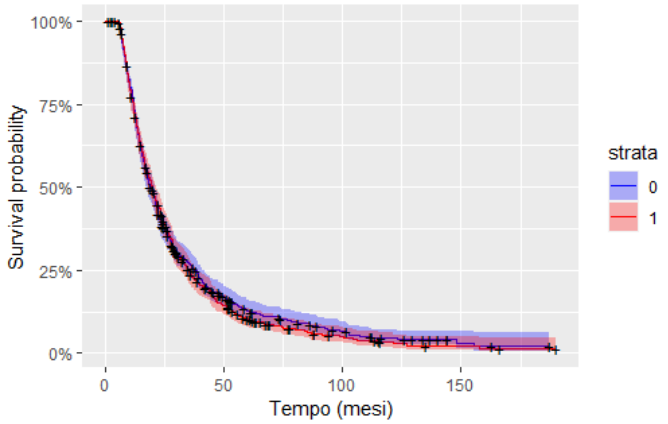

Figura 3.5: Curva di sopravvivenza Kaplan-Meier per un modello di sopravvivenza diviso per genere

Dal grafico possiamo notare che non c'è differenza significativa nella sopravvivenza tra i due gruppi (maschi indicati con 1 e femmine indicate con 0). Questo vuol dire

che la probabilità di sopravvivenza nel tempo è simile per entrambi i gruppi e non vi è un effetto significativo del genere sulla sopravvivenza. Andiamo a confermare questo risultato eseguendo dei test di significatività statistica:

- log-rank test: è il test più comunemente utilizzato; si basa sul confronto delle curve di sopravvivenza stimata per i gruppi di interesse calcolando una statistica del rapporto delle differenze osservate e attese tra i gruppi di sopravvivenza nel tempo.
- Peto-Peto test: è una variante del test precedente, che assegna pesi ai tempi di evento in base alla variazione del numero di individui a rischio nel tempo.
- Gehan-Breslow-Wilcoxon test: questo test attribuisce maggior peso ai tempi di evento che si verificano nei primi periodi di osservazione, rendendolo più sensibile a differenze iniziale nella sopravvivenza tra i gruppi.

Nel nostro caso abbiamo che tutti e tre i test (log-rang: Chisq= 0.4 on 1 degrees of freedom, p= 0.6; Peto-Peto: Chisq= 0 on 1 degrees of freedom, p= 1; Gehan-Breslow-Wilcoxon: Chisq= 0.1 on 1 degrees of freedom, p= 0.8)indicano che non vi è evidenza di una differenza significativa nella sopravvivenza tra i gruppi di genere nel dataset.

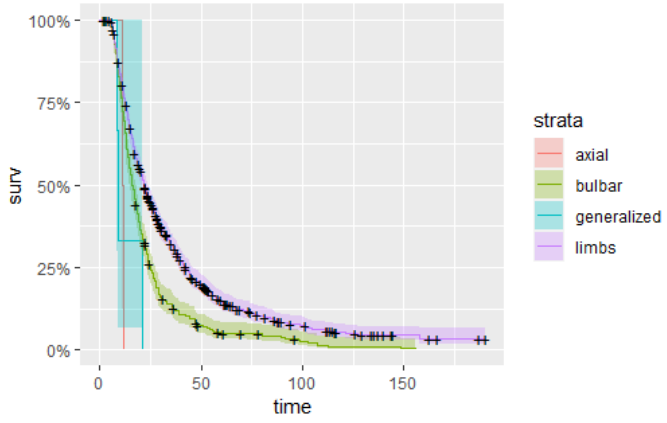

Figura 3.6: Curva di sopravvivenza Kaplan-Meier per un modello di sopravvivenza diviso per luogo insorgenza

Dal grafico, è possibile osservare che le modalità *axial* e *generalized* mostrano una diminuzione molto rapida nella sopravvivenza rispetto alle modalità *bulbar* e *limbs*. Questo è evidenziato dal fatto che le curve di sopravvivenza per *axial* e *generalized* si abbassano rapidamente già all'inizio dell'osservazione (circa 25 mesi) e continuano a scendere. D'altra parte, le curve di sopravvivenza per *bulbar* e *limbs* scendono in

modo più graduale nel corso del tempo.
Questi risultati sono supportati anche dai test di sopravvivenza, confermando la differenza tra i gruppi
Il log-rank test mostra una statistica del rapporto delle differenze osservate e attese per le diverse modalità di insorgenza, insieme a statistiche associate al chi-quadro. Il valore di p del test è estremamente piccolo (p= 3e-12), indicando che c'è una differenza significativa nella sopravvivenza tra le modalità di insorgenza.
Possiamo confermare quindi che ci sono differenze statisticamente significative nella sopravvivenza tra le diverse modalità di insorgenza della malattia nel dataset.

Per quanto riguarda la variabile *motoneurone*, dal grafico 3.7 possiamo notare una differenza tra le curve in base al motoneurone coinvolto.

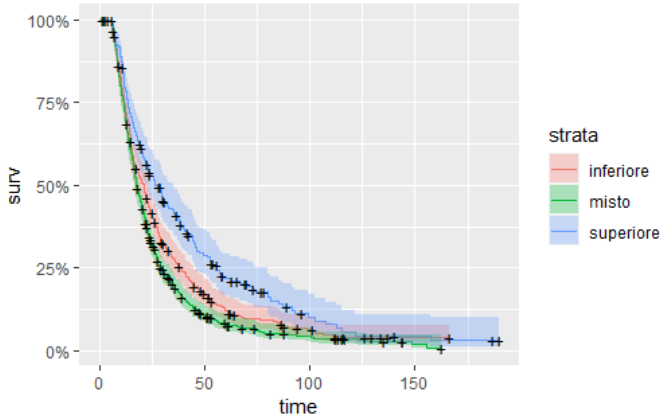

Figura 3.7: Curva di sopravvivenza Kaplan-Meier per un modello di sopravvivenza diviso per presenza della malattia nel motoneurone

Inoltre effettuando il test del log-rank, otteniamo un p-value di 3e-06 , confermando una differenza significativa tra i gruppi.
Vediamo ora solo alcune delle variabili presenti nel dataset, che sono quelle che dall'analisi hanno riportato una differenza significativa tra i gruppi.

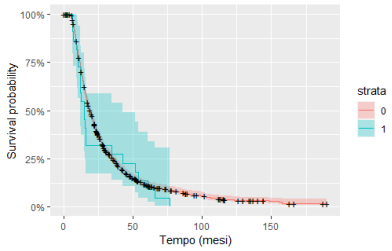

Figura 3.8: curva KM per *major_trauma_before_onset*

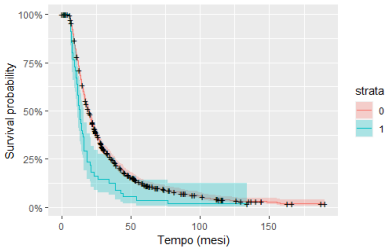

Figura 3.9: curva KM per *retired_at_diagnosis*

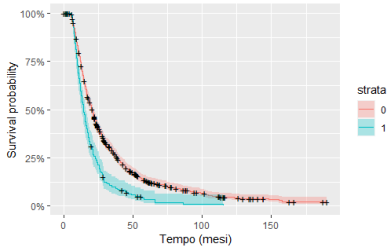

Figura 3.10: curva KM per *moreThan10PercentWeightloss*

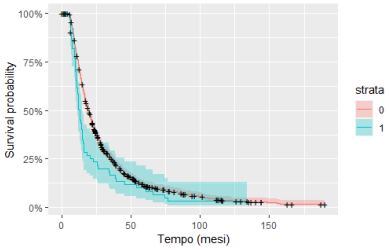

Figura 3.11: curva KM per *surgical_interventions_before_onset*

Di queste variabili possiamo notare che dove si verifica la variabile in questione non sono presenti censure. Questo vuol dire che l'evento (NIV o DEATH) si verifica sempre se si è verificata anche quella variabile.

La presenza di altre malattie (ipertensione, diabete, dislipidemia, disturbo della tiroide, malattia autoimmune, ictus), non comporta una compromissione nella malattia. Infatti dalle analisi svolte non notiamo differenze significative tra i gruppi.

Analisi dataset B

Anche in questo caso partiamo da un'analisi generale senza inclusione di covariate per fornire una visione generale della distribuzione dei tempi di sopravvivenza nel campione e identificare eventuali differenze significative nella sopravvivenza tra i gruppi di interesse.

Sull'asse delle x abbiamo il tempo trascorso dal momento iniziale dell'osservazione; sull'asse delle y abbiamo la probabilità stimata di sopravvivenza, che rappresenta la proporzione di individui che non hanno ancora sperimentato l'evento di interesse nel corso del tempo.

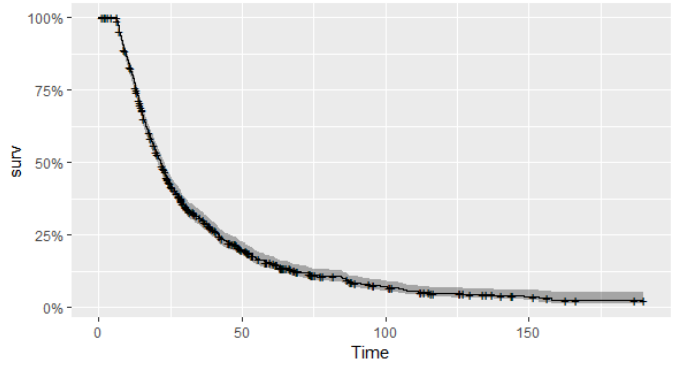

Figura 3.12: Curva di sopravvivenza Kaplan-Meier per un modello di sopravvivenza di base (senza covariate)

Come succede per il dataset A, la linea del grafico, che rappresenta la curva di sopravvivenza stimata, parte da 1, che rappresenta una probabilità del 100% di sopravvivenza all'inizio del tempo di osservazione. Man mano che trascorre il tempo, la curva scende, indicando una diminuzione della probabilità di sopravvivenza.
Andiamo a vedere come questa curva è influenza da altre variabili e se c'è differenza tra i gruppi presi in analisi.

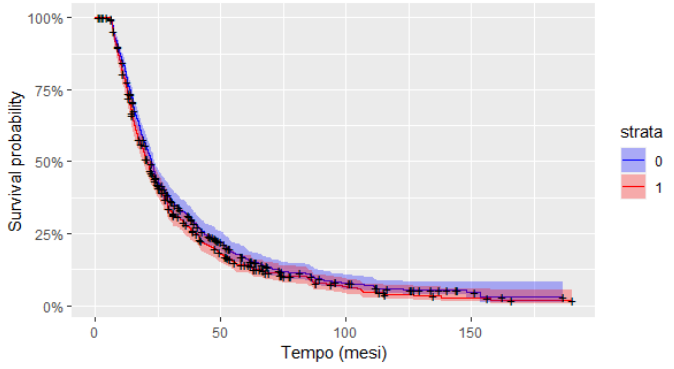

Figura 3.13: Curva di sopravvivenza Kaplan-Meier per un modello di sopravvivenza diviso per genere

Dal grafico 3.13 possiamo notare che non c'è differenza significativa nella sopravvivenza tra i due gruppi (maschi indicati con 1 e femmine indicate con 0). Questo vuol dire che la probabilità di sopravvivenza nel tempo è simile per entrambi i gruppi e non vi è un effetto significativo del genere sulla sopravvivenza. Confermiamo questa assunzione a livello grafico con il test del log-rank; il test ci fornisce un p-value di 0.08 che, essendo superiore ad una soglia fissata di 0.05, ci porta a confermare che non c'è differenza significativa tra il genere maschio e femmina.
Procediamo con l'analisi di altre variabili.

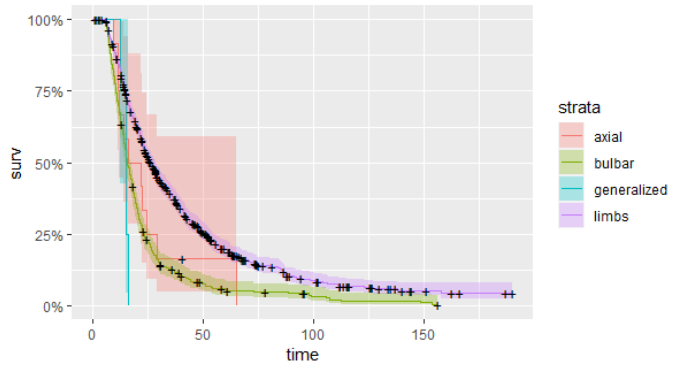

Figura 3.14: Curva di sopravvivenza Kaplan-Meier per un modello di sopravvivenza diviso per luogo insorgenza

La figura 3.14, ci fornisce informazioni sulla modalità di insorgenza della malattia, portandoci a dire che c'è una differenza tra i gruppi.
Infatti osserviamo che le modalità *axial* e *generalized* mostrano una diminuzione molto rapida nella sopravvivenza rispetto alle modalità *bulbar* e *limbs*.
Questi risultati sono supportati anche dai test di sopravvivenza, confermando la differenza tra i gruppi.

Per quanto riguarda la variabile *motoneurone*, dal grafico 3.15 possiamo notare una differenza tra le curve in base al motoneurone coinvolto.

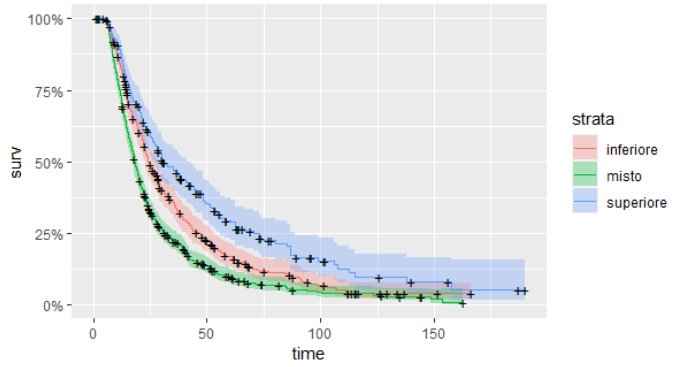

Figura 3.15: Curva di sopravvivenza Kaplan-Meier per un modello di sopravvivenza diviso per presenza della malattia nel motoneurone

La differenza, visibile graficamente, è confermata nei test di sopravvivenza; in particolare il test del log-rank ci fornisce un p-value pari a 2e-13, confermandoci la differenza.
Vediamo ora solo alcune delle variabili presenti nel dataset, che sono quelle che dall'analisi hanno riportato una differenza significativa tra i gruppi.

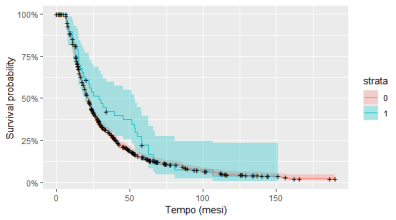

Figura 3.16: curva KM per *major_trauma_before_onset*

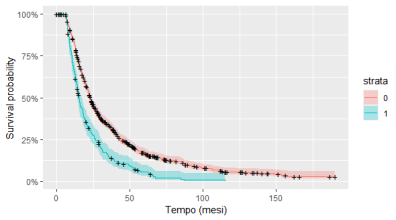

Figura 3.17: curva KM per *moreThan10PercentWeightloss*

La presenza di altre malattie (ipertensione, diabete, dislipidemia, disturbo della tiroide, malattia autoimmune, ictus), non comporta una compromissione nella malattia. Infatti dalle analisi svolte non notiamo differenze significative tra i gruppi.

Analisi dataset C

Come visto per i dataset A e B, procediamo all'analisi non parametrica di Kaplan-Meier anche per il dataset C. Anche in questo caso partiamo da un'analisi generale senza inclusione di covariate per fornire una visione generale della distribuzione dei tempi di sopravvivenza nel campione.

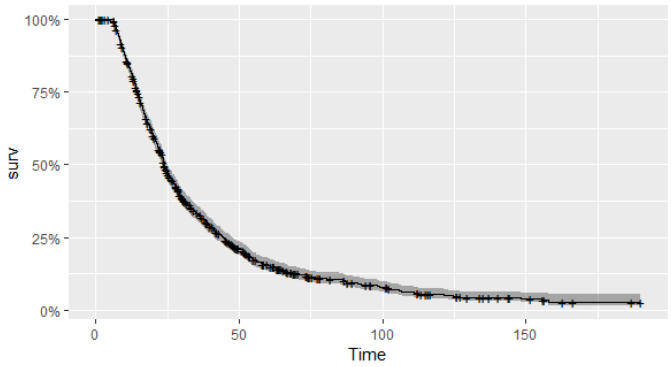

Figura 3.18: Curva di sopravvivenza Kaplan-Meier per un modello di sopravvivenza di base (senza covariate)

La curva, come per i dataset precedenti, diminuisce con l'aumentare del tempo, in quanto si riduce la probabilità di sopravvivenza alla malattia.
Non sono presenti differenze tra maschi e femmine, pertanto non viene riportata la curva di Kaplan-Meier.

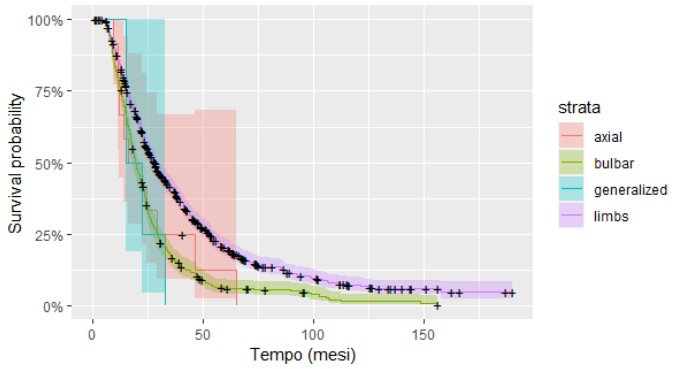

Figura 3.19: Curva di sopravvivenza Kaplan-Meier per un modello di sopravvivenza diviso per luogo insorgenza

Il grafico 3.19, ci fornisce informazioni sulla modalità di insorgenza della malattia. Graficamente possiamo notare una differenza tra i gruppi, che viene poi confermata dal test del log-rank.

Per quanto riguarda la variabile *motoneurone*, dal grafico 3.20 possiamo notare una differenza tra le curve in base al motoneurone coinvolto che viene poi confermata anche dal test di sopravvivenza (p-value = 6e-14)

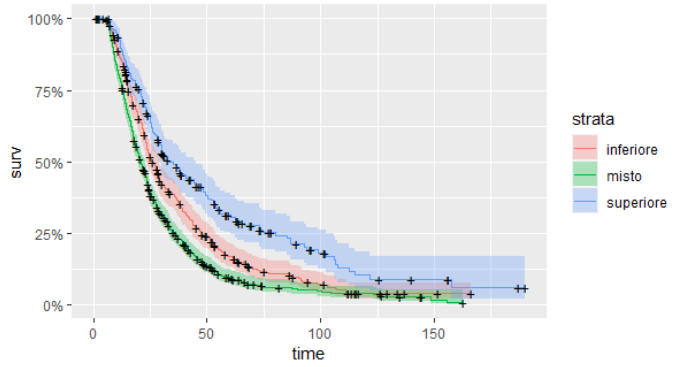

Figura 3.20: Curva di sopravvivenza Kaplan-Meier per un modello di sopravvivenza diviso per presenza della malattia nel motoneurone

Un'altra variabile in cui notiamo una differenza tra i gruppi è la variabile *moreThan10PercentWeightlos*, rappresentata nel grafico 3.21.

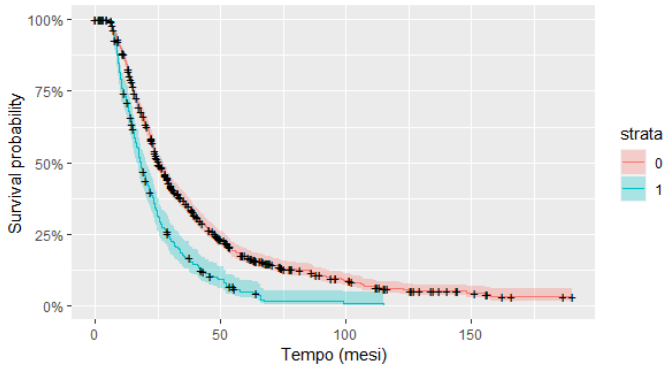

Figura 3.21: curva KM per la variabile *moreThan10PercentWeightloss*

Le altre variabili non presentano differenze significative tra i gruppi e non comportano una compromissione nella malattia. Infatti dalle analisi svolte non notiamo differenze significative tra i gruppi.

3.2.3 Modello semi-parametrico

Come accennato in precedenza, il modello che, a livello teorico, si adatta meglio ai dati, potrebbe essere un modello semi-parametrico.
E' un tipo di modello statistico che combina elementi di modelli parametrici e non parametrici.Questo approccio consente di ottenere un buon compromesso tra flessibilità e interpretabilità del modello. Alcune caratteristiche e vantaggi dei modelli semi-parametrici:

- **Flessibilità**: i modelli semi-parametrici sono più flessibili rispetto ai modelli parametrici tradizionali perché consentono di modellare le relazioni tra variabili in modo non lineare o complesso;
- **Riduzione del rischio di specificare erroneamente la forma funzionale**: nei modelli parametrici, spesso è necessario specificare una forma funzionale a priori per le variabili. Nei modelli semi-parametrici, l'uso di componenti non parametriche consente di evitare di specificare erroneamente la forma funzionale delle variabili di interesse;
- **Gestione dei dati con distribuzione complessa**: i modelli semi-parametrici sono adatti per affrontare dati con distribuzioni complesse o non normali, in quanto non richiedono l'assunzione di una distribuzione specifica;
- **Riduzione del rischio di overfitting**: i modelli semi-parametrici, grazie all'uso di componenti parametriche, possono ridurre il rischio di overfitting

rispetto ai modelli non parametrici pur mantenendo una certa flessibilità nel modellare le relazioni tra le variabili;

- **Interpretazione dei risultati**: a differenza dei modelli completamente non parametrici, i modelli semi-parametrici consentono di ottenere stime dei parametri che possono essere interpretate in modo tradizionale. Ciò facilita l'interpretazione dei risultati del modello e la comunicazione delle conclusioni;

- **Riduzione della dimensionalità**: nei modelli semi-parametrici è possibile utilizzare le componenti parametriche per ridurre la dimensionalità dei dati, consentendo di modellare relazioni complesse tra variabili con un numero limitato di parametri.

Tuttavia, è importante notare che l'uso di modelli semi-parametrici richiede un'attenta analisi dei dati e delle specifiche del modello per garantire che il modello sia appropriato per i dati in questione. Inoltre, la scelta tra un modello parametrico, semi-parametrico o non parametrico dipende dalle caratteristiche dei dati, dagli obiettivi dell'analisi e dalle ipotesi sottostanti.

Il più noto è il modello a rischi proporzionali di Cox, che esprime l'azzardo in funzione del tempo e delle covariate, senza però formalizzare la dipendenza dal tempo.
Tale modello assume

$$h(x|Z) = h_0(x)e^{\beta^T Z} \tag{3.9}$$

con $\beta^T = (\beta_1, ..., \beta_p)$ vettore di coefficienti di regressione.
Andando ad analizzare meglio questa formula possiamo notare che $h_0(x)$ è una funzione non parametrica e non specificata e $e^{\beta^T Z}$ è una funzione monotona, non negativa, non dipende dal tempo x (infatti è costante nel tempo).
Come per gli altri modelli, anche per il modello Cox devono essere verificate delle assunzioni; La prima assunzione riguarda la proporzionalità dei rischi:

$$\frac{h(x|Z = z_j)}{h(x|Z = z_h)} = \frac{h_0(x)e^{(\beta^T z_j)}}{h_0(x)e^{(\beta^T z_h)}} = e^{\beta^T (z_j - z_h)}; \tag{3.10}$$

questa implica che l'effetto delle covariate sul rischio dell'evento di interesse è costante nel tempo

Analisi

Nell'analisi di regressione di Cox, si considera un insieme di variabili predittive potenziali che potrebbero influenzare la sopravvivenza. Tuttavia, non tutte le variabili

potrebbero essere significative o contribuire in modo significativo alla modellizzazione della sopravvivenza.
Per selezionare le variabili predittive più significative da includere nel modello di Cox finale possono essere applicati tre metodi:

1. **forward selection**(Selezione in avanti): Inizia con un modello che include solo l'intercetta e successivamente aggiunge gradualmente le covariate una alla volta, selezionando quella che fornisce il miglior miglioramento al modello in termini di devianza o di un'altra misura di bontà di adattamento. Questo processo viene ripetuto fino a quando nessuna covariata aggiuntiva migliora ulteriormente il modello.

2. **backward selection**(Eliminazione all'indietro): Parte da un modello che include tutte le covariate disponibili e successivamente elimina gradualmente una covariata alla volta, basandosi su test di significatività o altre misure di importanza. Le covariate meno significative vengono eliminate una alla volta fino a quando tutte le covariate nel modello sono statisticamente significative o fino a quando il miglioramento nel modello non è più sostanziale.

3. **Stepwise selection**(Selezione passo-passo): Combina l'approccio forward e backward, eseguendo sia l'aggiunta che l'eliminazione delle covariate durante il processo di selezione. Inizia con un modello che include solo l'intercetta e successivamente aggiunge o elimina covariate in base a criteri predefiniti, come test di significatività o criteri di informazione come l'AIC (Akaike's Information Criterion) o il BIC (Bayesian Information Criterion). Questo processo viene iterato fino a raggiungere un modello finale che soddisfa i criteri di selezione stabiliti.

Dataset A

Partendo da un modello che include tutte le covariate otteniamo il seguente output:

	coef	exp(coef)	se(coef)	z	p	
sex1	-0.0112	0.9889	0.0941	-0.1190	0.9053	
height	0.3039	1.3552	0.4967	0.6119	0.5406	
weight_before_onset	0.0264	1.0267	0.0089	2.9764	0.0029	**
weight	-0.0252	0.9751	0.0091	-2.7820	0.0054	**
moreThan10PercentWeightloss1	0.1833	1.2012	0.1340	1.3683	0.1712	
major_trauma_before_onset1	-0.1724	0.8416	0.2844	-0.6064	0.5443	
surgical_interventions_before_onset1	0.3771	1.4580	0.2199	1.7146	0.0864	.
age_onset	0.0194	1.0196	0.0033	5.8619	0.0000	***
insorgenzabulbar	-1.0317	0.3564	0.7165	-1.4400	0.1499	
insorgenzageneralized	-0.4354	0.6470	0.9362	-0.4651	0.6419	
insorgenzalimbs	-1.4046	0.2455	0.7151	-1.9642	0.0495	*
motoneuronemisto	0.0506	1.0519	0.0848	0.5966	0.5508	
motoneuronesuperiore	-0.4726	0.6234	0.1062	-4.4494	0.0000	***
retired_at_diagnosis1	0.2726	1.3133	0.2092	1.3031	0.1925	
smoking1	-0.0395	0.9612	0.0702	-0.5635	0.5731	
hypertension1	-0.0369	0.9638	0.0723	-0.5105	0.6097	
diabetes1	-0.0756	0.9272	0.1127	-0.6707	0.5024	
dyslipidemia1	-0.0364	0.9643	0.1039	-0.3499	0.7264	
thyroid_disorder1	0.1055	1.1112	0.1000	1.0541	0.2918	
autoimmune_disease1	-0.0017	0.9983	0.2090	-0.0081	0.9936	
stroke1	-0.3882	0.6783	0.2036	-1.9067	0.0566	.
cardiac_disease1	0.0059	1.0059	0.1294	0.0457	0.9635	
primary_neoplasm1	0.0008	1.0008	0.1043	0.0073	0.9941	

—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

	exp(coef)	exp(-coef)	lower .95	upper .95
sex1	0.9889	1.0113	0.82235	1.1891
height	1.3552	0.7379	0.51195	3.5872
weight_before_onset	1.0267	0.9740	1.00905	1.0447
weight	0.9751	1.0255	0.95794	0.9926
moreThan10PercentWeightloss1	1.2012	0.8325	0.92379	1.5619
major_trauma_before_onset1	0.8416	1.1882	0.48198	1.4695
surgical_interventions_before_onset1	1.4580	0.6859	0.94746	2.2436
age_onset	1.0196	0.9808	1.01298	1.0262
insorgenzabulbar	0.3564	2.8060	0.08751	1.4514
insorgenzageneralized	0.6470	1.5455	0.10329	4.0530
insorgenzalimbs	0.2455	4.0738	0.06044	0.9970
motoneuronemisto	1.0519	0.9507	0.89087	1.2420
motoneuronesuperiore	0.6234	1.6041	0.50624	0.7677
retired_at_diagnosis1	1.3133	0.7614	0.87163	1.9789
smoking1	0.9612	1.0403	0.83776	1.1029
hypertension1	0.9638	1.0376	0.83648	1.1104
diabetes1	0.9272	1.0786	0.74334	1.1565
dyslipidemia1	0.9643	1.0370	0.78662	1.1821
thyroid_disorder1	1.1112	0.8999	0.91337	1.3519
autoimmune_disease1	0.9983	1.0017	0.66277	1.5038
stroke1	0.6783	1.4743	0.45514	1.0109
cardiac_disease1	1.0059	0.9941	0.78066	1.2962
primary_neoplasm1	1.0008	0.9992	0.81582	1.2276

Concordance= 0.622 (se = 0.01)
Likelihood ratio test= 179.7 on 23 df, p=<2e-16
Wald test = 186.8 on 23 df, p=<2e-16
Score (logrank) test = 191 on 23 df, p=<2e-16

Tabella 3.5: Modello di cox che include tutte le covariate

Il modello completo rappresentato nella tabella 3.5, ci riassume tutte le covariate presenti nel modello.
Nell'output, possiamo vedere i coefficienti stimati per ciascuna variabile indipendente nel modello, insieme ai relativi valori p e agli intervalli di confidenza.
I coefficienti rappresentano la relazione stimata tra ciascuna variabile e l'hazard ratio, che rappresenta il rapporto di rischio istantaneo per due gruppi di individui con una differenza unitaria nella variabile indipendente.

Possiamo valutare se le covariate risultano statisticamente significative (sulla base del valore p) e osservando se l'hazard ratio è superiore a 1 (indicando un aumento del rischio) o inferiore a 1 (indicando una riduzione del rischio).
I valori p rappresentano la probabilità di ottenere un risultato almeno o più estremo di quello osservato, assumendo che l'ipotesi nulla sia vera (l'ipotesi nulla afferma che non vi è alcuna relazione tra la variabile indipendente e l'outcome).
I valori sono rappresentati da simboli di asterischi (*) e punti (.) che indicano il livello di significatività statistica.
Le variabili che ci vengono evidenziate sono significative e hanno un effetto sull'outcome.
Infine, l'output fornisce anche informazioni sulla concordanza del modello (concordance), che indica quanto bene il modello si adatta ai dati, e i risultati dei test di significatività complessiva del modello (likelihood ratio test, Wald test e score test). In questo caso, i valori di p molto bassi indicano che il modello nel suo complesso è altamente significativo.

Proviamo a ridurre il modello tenendo solo le variabili che sono risultate significative nel test appena effettuato, ovvero tutte le variabili con un p-value inferiore a 0.05.
Abbiamo deciso di tenere la variabile *insorgenza*, seppur non significativa poiché nell'analisi non parametrica di Kaplan-Meier risultava interessante, mettendo in evidenza differenze tra i gruppi, creati in base a luogo in cui si manifestava la malattia.

	coef	exp(coef)	se(coef)	z	p
weight_before_onset	0.0372	1.0379	0.0057	6.5709	0.0000 ***
weight	-0.0360	0.9647	0.0057	-6.2624	0.0000 ***
age_onset	0.0174	1.0175	0.0030	5.8625	0.0000 ***
insorgenzabulbar	-0.9676	0.3800	0.7119	-1.3592	0.1741
insorgenzageneralized	-0.1453	0.8647	0.9150	-0.1588	0.8738
insorgenzalimbs	-1.3419	0.2614	0.7111	-1.8871	0.0591 .
motoneuronemisto	-0.0032	0.9968	0.0806	-0.0392	0.9687
motoneuronesuperiore	-0.4988	0.6072	0.1040	-4.7973	0.0000 ***

—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

	exp(coef)	exp(-coef)	lower .95	upper .95
weight_before_onset	1.0379	0.9635	1.02646	1.0495
weight	0.9647	1.0366	0.95388	0.9756
age_onset	1.0175	0.9828	1.01165	1.0235
insorgenzabulbar	0.3800	2.6316	0.09415	1.5337
insorgenzageneralized	0.8647	1.1564	0.14390	5.1965
insorgenzalimbs	0.2614	3.8261	0.06486	1.0532
motoneuronemisto	0.9968	1.0032	0.85120	1.1674
motoneuronesuperiore	0.6072	1.6468	0.49528	0.7445

Concordance= 0.613 (se = 0.01)
Likelihood ratio test= 163.1 on 8 df, p=<2e-16
Wald test = 166.5 on 8 df, p=<2e-16
Score (logrank) test = 170.4 on 8 df, p=<2e-16

Tabella 3.6: modello di cox ridotto

Per quanto riguarda il modello ridotto 3.6, notiamo che ci sono ancora delle variabili che non sono significative.

Per scegliere quale tra i due modelli è preferibile andiamo a misurare l'indice di Akaike e il log-rapporto di verosimiglianza.
Il test di Akaike (Akaike's Information Criterion, AIC) è un criterio di selezione del modello utilizzato per confrontare modelli statistici alternativi. È stato sviluppato da Hirotugu Akaike nel campo della teoria dell'informazione.
L'AIC è una misura della bontà di adattamento di un modello ai dati, che penalizza i modelli più complessi per evitare l'overfitting. L'obiettivo dell'AIC è trovare il modello che massimizza l'informazione ottenuta dai dati e allo stesso tempo minimizza la complessità del modello.
Si basa sul log-likelihood del modello e tiene conto del numero di parametri stimati nel modello. È definito come:

$$AIC = -2(logL_0 - logL) + 2p \tag{3.11}$$

dove p è il numero dei parametri e $-2(logL\_0 + logL)$ è il rapporto di log verosimigliaza con $logL_0$ valore del logaritmo della verosimiglianza del modello ridotto (modello con meno parametri) e $logL$ è il valore del logaritmo della verosimiglianza del modello completo (modello con più parametri)).
Quando si confrontano più modelli, il modello con l'AIC più basso viene considerato

come il modello migliore o più adatto ai dati. Il log-rapporto di verosimiglianza (lrt) è una misura che confronta la log-verosimiglianza di due modelli. L'idea è che se il log-rapporto di verosimiglianza è significativo e positivo, indica che il modello completo fornisce un miglior adattamento rispetto al modello ridotto.

Quindi, AIC e lrt forniscono informazioni diverse sui modelli. AIC tiene conto sia dell'adattamento sia della complessità del modello, mentre lrt confronta solo la log-verosimiglianza tra i due modelli.

Confronto tra modelli		
	AIC	Log-rapporto di verosimiglianza
fit_tot	12231.1	179.7039
fit_tot_ridotto	12217.71	163.0988

Tabella 3.7: Confronto tra i modelli di Cox dataset A

Come possiamo notare dalla tabella 3.7, il modello che preferiamo è il modello ridotto, che un un AIC inferiore; Tuttavia risulta difficile decidere quale modello tenere poiché differisco di poco.
Andiamo a calcolrae il test ANOVA e vediamo se ci fornisce maggiori informazioni.
Il test ANOVA (Analysis of Variance) è un test statistico utilizzato per confrontare le medie di due o più gruppi. L'obiettivo è quello di determinare se le differenze osservate tra le medie dei gruppi sono statisticamente significative o se possono essere attribuite alla variabilità casuale.
Si basa sull'ipotesi nulla (H0) che le medie di tutti i gruppi siano uguali. L'ipotesi alternativa (H1) sostiene che almeno una delle medie dei gruppi sia diversa dalle altre.
Il test ANOVA calcola la statistica F, che rappresenta il rapporto tra la varianza tra i gruppi (variabilità tra le medie dei gruppi) e la varianza all'interno dei gruppi (variabilità all'interno di ciascun gruppo). Se la variabilità tra i gruppi è significativamente maggiore rispetto alla variabilità all'interno dei gruppi, la statistica F sarà grande e il test ANOVA indicherà una significativa differenza tra le medie dei gruppi.
Nel nostro caso specifico, i risultati indicano che il test di rapporto di verosimiglianza è altamente significativo (p-value molto basso, $< 2.2e\text{-}16$), il che suggerisce che i due modelli non sono equivalenti. Questo significa che le variabili presenti nel modello completo forniscono un adattamento significativamente migliore rispetto al modello ridotto.

Ecco riportato l'output del test:

	loglik	Chisq	Df	Pr(>|Chi|)
Modello completo	-6092.6			
Modello ridotto	-6100.9	16.605	15	0.343

Il test ANOVA presenta un p-value di 0.343 (non significativo rispetto ad una soglia fissata di 0.05). Pertanto non possiamo rifiutare l'ipotesi nulla che non ci siano differenze significative tra i modelli e questo mi suggerisce che hanno una simile capacità di adattarsi ai dati.

Questo viene confermato anche dai test AIC e lrt 3.7 che anch'essi mi portavano, sebbene di poco, a preferire il modello ridotto.

Adattabilità del modello

Una volta scelto il modello, possiamo andare a verificare quanto bene si adatta ai dati, utilizzando diversi test. Uno dei test maggiormente utilizzati è l'analisi dei residui, attraverso il test di Shoenfeld e di martingale. Il test di Schoenfeld valuta se l'andamento dei residui parziali del modello di Cox è indipendente dal tempo, cioè se non vi è correlazione tra i residui e il tempo trascorso. Se il test mostra una significativa correlazione tra i residui e il tempo, ciò suggerisce una violazione dell'assunzione di proporzionalità delle hazard. In pratica viene calcolato il coefficiente di correlazione tra i residui parziali e le stime dei coefficienti del modello di Cox al variare del tempo. Un valore p basso nel test di Schoenfeld suggerisce una violazione dell'assunzione di proporzionalità delle hazard. Se il test di Schoenfeld indica una violazione dell'assunzione di proporzionalità delle hazard, è possibile adottare diverse strategie per affrontare questo problema, come l'introduzione di termini di interazione tra la variabile indipendente e il tempo o l'utilizzo di modelli flessibili come i modelli di Cox con effetti non proporzionali delle hazard. Il test (applicato al modello completo) ci fornisce i seguenti risultati:

Variabile	Chi-quadro	Df	p-value
weight_before_onset	0.09840	1	0.754
weight	0.00456	1	0.946
age_onset	1.07471	1	0.300
insorgenza	7.78376	3	0.051
motoneurone	0.64462	2	0.724
GLOBAL	9.87161	8	0.274

Tabella 3.8: riassunto del calcolo dei residui mediante il test di shoenfeld per il dataset A

Come possiamo notare dalla tabella 3.8, per tutte le variabili è soddisfatta l'ipotesi di proporzionalità degli hazard.
Infatti osserviamo un p-value superiore a 0.05 (soglia fissata), suggerendoci che l'ipotesi di proporzionalità è soddisfatta. Anche a livello globale (p-value = 0.274), l'ipotesi è rispettata.

Andiamo ora a verificare questa ipotesi anche graficamente:

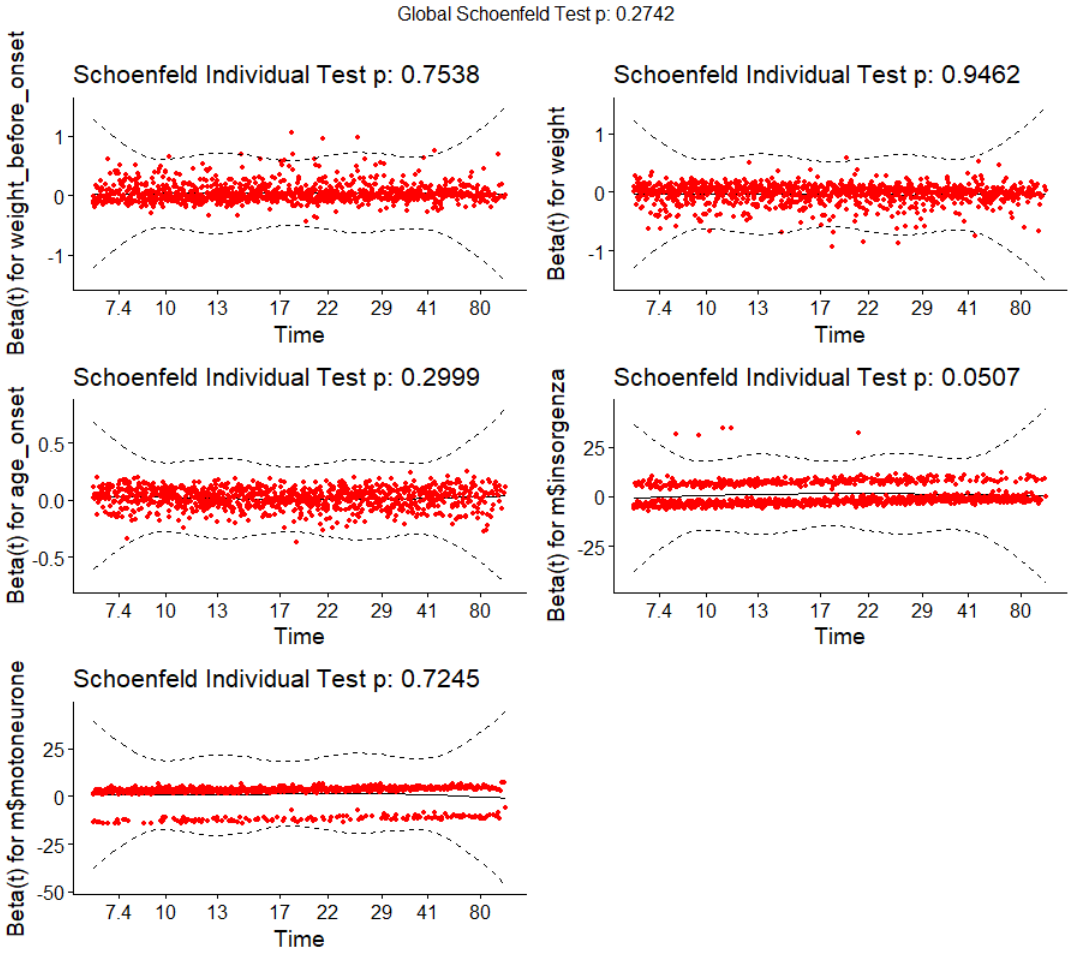

Figura 3.22: Residui di Shoenfeld per il dataset A

I punti dei residui di Schoenfeld seguono approssimativamente una linea orizzontale a zero e sono contenuti negli intervalli di confidenza; ciò suggerisce che non vi è evidenza di una violazione significativa dell'assunzione di proporzionalità degli effetti nel modello di Cox.
Confermiamo quindi quanto visto prima nella tabella 3.8 e affermiamo che la proporzionalità degli hazard è rispettata per tutte le variabili.

Dataset B

Per quanto riguarda il dataset B, applicando il modello di Cox otteniamo il seguente output:

	coef	exp(coef)	se(coef)	z	p	
sex1	0.1079	1.1140	0.0909	1.1871	0.2352	
height	0.8280	2.2888	0.4958	1.6702	0.0949	.
weight_before_onset	0.0300	1.0305	0.0086	3.4992	0.0005	***
weight	-0.0334	0.9672	0.0089	-3.7455	0.0002	***
moreThan10PercentWeightloss1	0.1741	1.1902	0.1298	1.3413	0.1798	
major_trauma_before_onset1	-0.3821	0.6824	0.2287	-1.6709	0.0947	.
surgical_interventions_before_onset1	0.3473	1.4153	0.1728	2.0094	0.0445	*
age_onset	0.0253	1.0256	0.0033	7.6197	0.0000	***
insorgenzabulbar	0.4444	1.5595	0.3147	1.4118	0.1580	
insorgenzageneralized	1.0835	2.9550	0.6025	1.7983	0.0721	.
insorgenzalimbs	-0.2009	0.8180	0.3117	-0.6447	0.5191	
motoneuronemisto	0.1020	1.1074	0.0827	1.2341	0.2172	
motoneuronesuperiore	-0.5637	0.5691	0.1072	-5.2605	0.0000	***
retired_at_diagnosis1	-0.1620	0.8505	0.1666	-0.9723	0.3309	
smoking1	0.0354	1.0361	0.0680	0.5207	0.6026	
hypertension1	-0.0508	0.9505	0.0700	-0.7257	0.4680	
diabetes1	-0.0853	0.9182	0.1125	-0.7586	0.4481	
dyslipidemia1	-0.0768	0.9261	0.1008	-0.7619	0.4461	
thyroid_disorder1	-0.0223	0.9779	0.1009	-0.2212	0.8249	
autoimmune_disease1	0.0497	1.0510	0.1914	0.2599	0.7950	
stroke1	-0.5797	0.5601	0.2156	-2.6883	0.0072	**
cardiac_disease1	-0.0669	0.9353	0.1265	-0.5289	0.5968	
primary_neoplasm1	0.0007	1.0007	0.1011	0.0069	0.9945	

—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

	exp(coef)	exp(-coef)	lower .95	upper .95
sex1	1.1140	0.8977	0.9322	1.3312
height	2.2888	0.4369	0.8662	6.0478
weight_before_onset	1.0305	0.9704	1.0133	1.0480
weight	0.9672	1.0339	0.9505	0.9842
moreThan10PercentWeightloss1	1.1902	0.8402	0.9228	1.5350
major_trauma_before_onset1	0.6824	1.4654	0.4359	1.0683
surgical_interventions_before_onset1	1.4153	0.7066	1.0086	1.9859
age_onset	1.0256	0.9750	1.0190	1.0323
insorgenzabulbar	1.5595	0.6412	0.8415	2.8899
insorgenzageneralized	2.9550	0.3384	0.9072	9.6253
insorgenzalimbs	0.8180	1.2226	0.4441	1.5067
motoneuronemisto	1.1074	0.9030	0.9417	1.3023
motoneuronesuperiore	0.5691	1.7572	0.4613	0.7021
retired_at_diagnosis1	0.8505	1.1758	0.6136	1.1788
smoking1	1.0361	0.9652	0.9067	1.1839
hypertension1	0.9505	1.0521	0.8286	1.0902
diabetes1	0.9182	1.0891	0.7365	1.1447
dyslipidemia1	0.9261	1.0798	0.7600	1.1284
thyroid_disorder1	0.9779	1.0226	0.8024	1.1918
autoimmune_disease1	1.0510	0.9515	0.7223	1.5294
stroke1	0.5601	1.7855	0.3670	0.8547
cardiac_disease1	0.9353	1.0692	0.7299	1.1984
primary_neoplasm1	1.0007	0.9993	0.8208	1.2200

Concordance= 0.657 (se = 0.009)
Likelihood ratio test= 293.6 on 23 df, p=<2e-16
Wald test = 297.5 on 23 df, p=<2e-16
Score (logrank) test = 307.2 on 23 df, p=<2e-16

Tabella 3.9: modello di cox che include tutte le covariate

Il modello completo rappresentato in tabella 3.9, ci riassume tutte le covariate presenti nel modello. Nell'output, possiamo vedere i coefficienti stimati per ciascuna variabile indipendente nel modello, insieme ai relativi valori p e agli intervalli di

confidenza. Le variabili che risultano significative sono quelle che verranno tenute nel modello di Cox ridotto; in aggiunta a queste, anche se non risulta significativa, teniamo la variabile *insorgenza*, poiché nell'analisi non parametrica di Kaplan-Meier abbiamo notato una differenza tra i gruppi che risulta interessante ai fini dell'analisi che stiamo svolgendo.
Il modello ridotto pertanto risulta essere:

	coef	exp(coef)	se(coef)	z	p	
weight_before_onset	0.0370	1.0377	0.0054	6.8735	0.0000	***
weight	-0.0403	0.9605	0.0055	-7.3032	0.0000	***
surgical_interventions_before_onset1	0.1235	1.1315	0.1344	0.9187	0.3583	
age_onset	0.0219	1.0221	0.0030	7.2826	0.0000	***
insorgenzabulbar	0.3578	1.4301	0.3087	1.1588	0.2465	
insorgenzageneralized	0.9473	2.5787	0.5873	1.6129	0.1068	
insorgenzalimbs	-0.2744	0.7600	0.3060	-0.8966	0.3699	
motoneuronemisto	0.1355	1.1451	0.0810	1.6735	0.0942	.
motoneuronesuperiore	-0.5481	0.5780	0.1060	-5.1717	0.0000	***
stroke1	-0.5897	0.5545	0.2136	-2.7606	0.0058	**

—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

	exp(coef)	exp(-coef)	lower .95	upper .95
weight_before_onset	1.0377	0.9637	1.0268	1.0487
weight	0.9605	1.0412	0.9501	0.9709
surgical_interventions_before_onset1	1.1315	0.8838	0.8694	1.4726
age_onset	1.0221	0.9784	1.0161	1.0281
insorgenzabulbar	1.4301	0.6992	0.7809	2.6193
insorgenzageneralized	2.5787	0.3878	0.8156	8.1535
insorgenzalimbs	0.7600	1.3157	0.4172	1.3846
motoneuronemisto	1.1451	0.8733	0.9771	1.3421
motoneuronesuperiore	0.5780	1.7300	0.4696	0.7115
stroke1	0.5545	1.8034	0.3648	0.8428

Concordance= 0.655 (se = 0.009)
Likelihood ratio test= 280.1 on 10 df, p=<2e-16
Wald test = 285.7 on 10 df, p=<2e-16
Score (logrank) test = 293.1 on 10 df, p=<2e-16

Tabella 3.10: modello di cox ridotto

Per confrontare i due modelli applichiamo un test ANOVA. Per questo dataset, il test non è statisticamente significativo (p-value = 0.4131), come possiamo notare dalla tabella sottostante, quindi non c'è una differenza significativa nell'adattamento tra i due modelli.

	loglik	Chisq	Df	Pr(>|Chi|)
Modello completo	-6395.5			
Modello ridotto	-6402.2	13.457	13	0.4131

Come visto per il dataset A, siamo portati a preferire il modello ridotto. Questo viene confermato anche dal test AIC che risulta 12836.92 per il modello completo e 12824.38 per il modello ridotto. Possiamo andare a verificare quanto bene si adatta ai dati, utilizzando il test di Shoenfeld.

Il test ci fornisce i seguenti risultati:

Variabile	Chi-quadro	Df	p-value
weight_before_onset	1.102	1	0.29
weight	1.984	1	0.16
surgical_interventions_before_onset	0.149	1	0.70
age_onset	2.360	1	0.12
insorgenza	8.975	3	0.03
motoneurone	2.085	2	0.35
stroke	1.446	1	0.23
GLOBAL	15.581	10	0.11

Tabella 3.11: riassunto del calcolo dei residui mediante il test di shoenfeld per il dataset B

Possiamo notare che solo la variabile *insorgenza* non rispetta la proporzionalità dei residui, poiché presenta un p-value di 0.03 (inferiore alla soglia fissata di 0.05). Tuttavia notiamo che, anche in questo caso, globalmente è rispettata.

Dataset C

Infine, per quanto riguarda il dataset C, applicando il modello di Cox otteniamo il seguente output:

	coef	exp(coef)	se(coef)	z	p	
sex1	-0.0579	0.9437	0.0905	-0.6398	0.5223	
height	0.8265	2.2853	0.4914	1.6821	0.0926	.
weight_before_onset	0.0362	1.0369	0.0087	4.1759	0.0000	***
weight	-0.0408	0.9600	0.0089	-4.5649	0.0000	***
moreThan10PercentWeightloss1	0.0754	1.0783	0.1277	0.5904	0.5549	
major_trauma_before_onset1	-0.3546	0.7014	0.2319	-1.5294	0.1262	
surgical_interventions_before_onset1	0.2480	1.2815	0.1667	1.4877	0.1368	
age_onset	0.0275	1.0279	0.0034	8.1600	0.0000	***
insorgenzabulbar	0.2308	1.2597	0.3141	0.7349	0.4624	
insorgenzageneralized	0.4974	1.6445	0.6008	0.8280	0.4077	
insorgenzalimbs	-0.2042	0.8153	0.3114	-0.6558	0.5119	
motoneuronemisto	0.1431	1.1538	0.0833	1.7178	0.0858	.
motoneuronesuperiore	-0.5696	0.5658	0.1072	-5.3111	0.0000	***
retired_at_diagnosis1	-0.1571	0.8546	0.1654	-0.9500	0.3421	
smoking1	-0.0053	0.9947	0.0679	-0.0783	0.9376	
hypertension1	-0.0091	0.9909	0.0696	-0.1310	0.8958	
diabetes1	-0.0873	0.9164	0.1104	-0.7909	0.4290	
dyslipidemia1	-0.0258	0.9746	0.0990	-0.2601	0.7948	
thyroid_disorder1	0.0639	1.0660	0.1011	0.6321	0.5273	
autoimmune_disease1	0.1320	1.1411	0.1909	0.6911	0.4895	
stroke1	-0.4655	0.6278	0.2073	-2.2454	0.0247	*
cardiac_disease1	-0.0488	0.9524	0.1236	-0.3946	0.6931	
primary_neoplasm1	0.0129	1.0130	0.1000	0.1290	0.8974	

—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

	exp(coef)	exp(-coef)	lower .95	upper .95
sex1	0.9437	1.0596	0.7903	1.1269
height	2.2853	0.4376	0.8724	5.9869
weight_before_onset	1.0369	0.9644	1.0194	1.0547
weight	0.9600	1.0416	0.9434	0.9770
moreThan10PercentWeightloss1	1.0783	0.9274	0.8395	1.3850
major_trauma_before_onset1	0.7014	1.4256	0.4453	1.1050
surgical_interventions_before_onset1	1.2815	0.7803	0.9243	1.7767
age_onset	1.0279	0.9728	1.0211	1.0347
insorgenzabulbar	1.2597	0.7939	0.6806	2.3314
insorgenzageneralized	1.6445	0.6081	0.5066	5.3387
insorgenzalimbs	0.8153	1.2266	0.4429	1.5009
motoneuronemisto	1.1538	0.8667	0.9800	1.3584
motoneuronesuperiore	0.5658	1.7676	0.4585	0.6981
retired_at_diagnosis1	0.8546	1.1701	0.6180	1.1818
smoking1	0.9947	1.0053	0.8707	1.1363
hypertension1	0.9909	1.0092	0.8646	1.1357
diabetes1	0.9164	1.0913	0.7380	1.1378
dyslipidemia1	0.9746	1.0261	0.8027	1.1833
thyroid_disorder1	1.0660	0.9381	0.8743	1.2997
autoimmune_disease1	1.1411	0.8764	0.7849	1.6589
stroke1	0.6278	1.5929	0.4181	0.9425
cardiac_disease1	0.9524	1.0500	0.7475	1.2134
primary_neoplasm1	1.0130	0.9872	0.8326	1.2324

Concordance= 0.648 (se = 0.009)
Likelihood ratio test= 282.8 on 23 df, p=<2e-16
Wald test = 278 on 23 df, p=<2e-16
Score (logrank) test = 285.2 on 23 df, p=<2e-16

Tabella 3.12: Modello di cox che include tutte le covariate

Andiamo a tenere solo le variabili che risultano significative (più la variabile insorgenza, per i motivi già citati).
Il modello ridotto pertanto risulta essere:

	coef	exp(coef)	se(coef)	z	p
weight_before_onset	0.0404	1.0412	0.0053	7.6228	0.0000 ***
weight	-0.0420	0.9588	0.0055	-7.7059	0.0000 ***
age_onset	0.0249	1.0252	0.0030	8.1762	0.0000 ***
m3$insorgenzabulbar	0.1107	1.1171	0.3082	0.3593	0.7193
m3$insorgenzageneralized	0.3522	1.4222	0.5853	0.6018	0.5473
m3$insorgenzalimbs	-0.3124	0.7317	0.3055	-1.0227	0.3065
m3$motoneuronemisto	0.1407	1.1511	0.0768	1.8332	0.0668 .
m3$motoneuronesuperiore	-0.5676	0.5669	0.1042	-5.4467	0.0000 ***
stroke1	-0.4267	0.6527	0.2047	-2.0847	0.0371 *

—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

	exp(coef)	exp(-coef)	lower .95	upper .95
weight_before_onset	1.0412	0.9604	1.0305	1.0521
weight	0.9588	1.0429	0.9487	0.9692
age_onset	1.0252	0.9754	1.0191	1.0314
insorgenzabulbar	1.1171	0.8952	0.6106	2.0438
insorgenzageneralized	1.4222	0.7031	0.4516	4.4788
insorgenzalimbs	0.7317	1.3667	0.4021	1.3315
motoneuronemisto	1.1511	0.8687	0.9903	1.3380
motoneuronesuperiore	0.5669	1.7641	0.4621	0.6953
stroke1	0.6527	1.5322	0.4370	0.9748

Concordance= 0.644 (se = 0.009)
Likelihood ratio test= 269.3 on 9 df, p=<2e-16
Wald test = 267 on 9 df, p=<2e-16
Score (logrank) test = 272.5 on 9 df, p=<2e-16

Tabella 3.13: Modello di cox ridotto

Per confrontare i due modelli applichiamo un test ANOVA. Anche per questo dataset, il test non è statisticamente significativo (p-value = 0.4823) e ci porta a dire che non c'è una differenza significativa nell'adattamento tra i due modelli.

	loglik	Chisq	Df	Pr(>|Chi|)
Modello completo	-6514.8			
Modello ridotto	-6521.6	13.568	14	0.4823

Andiamo a vedere anche cosa ci dice il test AIC che risulta pari a 13075.56 per il modello completo e 13061.13 per il modello ridotto.
Questo ci porta a preferire, anche se leggermente, il modello ridotto.
Attraverso il test di Shoenfeld, andiamo a verificare se le variabili incluse nel modello rispettano la proporzionalità.

Il test svolto ci fornisce il seguente output:

Variabile	Chi-quadro	Df	p-value
weight_before_onset	0.263	1	0.608
weight	0.528	1	0.467
age_onset	3.421	1	0.064
insorgenza	3.438	3	0.329
motoneurone	1.512	2	0.470
stroke	4.435	1	0.035
GLOBAL	15.573	9	0.076

Tabella 3.14: riassunto del calcolo dei residui mediante il test di shoenfeld per il dataset C

Possiamo notare che la variabile *stroke* non rispetta la proporzionalità dei residui, poiché presenta un p-value di 0.035.
Tuttavia, la proporzionalità è rispettata globalmente, presentando un p-value di 0.076.

Capitolo 4

Metodo per la previsione

Nei capitoli precedenti abbiamo analizzato i dati di sopravvivenza con metodi standard, utilizzando modelli semi-parametrici (modello di Cox), modelli non parametrici (Kaplan-meier) e modelli parametrici.
I modelli parametrici (come Gamma, Weibull, ecc.) e i modelli semiparametrici come il modello di rischio proporzionale di Cox sono utili e ampiamente discussi nella letteratura. Tuttavia, questi modelli hanno il limite di richiedere una specificazione anticipata della relazione funzionale tra il tempo di sopravvivenza e le covariate.
Per superare questa limitazione, sono stati proposti metodi non parametrici più flessibili, come le foreste di sopravvivenza causali (random survival forests), che consentono ai dati di determinare automaticamente la struttura del modello. Questi metodi si basano sulla costruzione di alberi di decisione, adattati specificamente per i dati di sopravvivenza censurati. Questo modello è un'estensione del random forest al quale vengono applicati i dati di sopravvivenza.
Prima di vedere questa estensione cerchiamo di capire cos'è il random forest e come funziona.

4.1 Random Forest

Il random forest (RF) è un insieme (ensemble) di alberi di regressione costruiti utilizzando il metodo CART (Classification and Regression Trees) per generare alberi decisionali all'interno dell'ensemble.
Il metodo CART è una tecnica di apprendimento automatico che costruisce alberi di decisione binari in cui ogni nodo interno effettua una divisione sui dati utilizzando una variabile predittiva. Il processo viene ripetuto ricorsivamente fino a quando si raggiungono le foglie, che forniscono le previsioni. Nel contesto del Random Forest,

vengono creati numerosi alberi di decisione indipendenti, ognuno addestrato su un campione casuale dei dati di addestramento. Durante la costruzione di ogni albero, viene effettuata una selezione casuale di un sottoinsieme di variabili predittive per determinare la migliore suddivisione in ciascun nodo. Questo processo di campionamento casuale delle variabili e dei dati di addestramento contribuisce a ridurre l'overfitting e aumenta la diversità tra gli alberi all'interno dell'ensemble.

Il RF è un estensione del metodo di bagging, una tecnica di campionamento che coinvolge la costruzione di un insieme di modelli (solitamente alberi di regressione o di classificazione) su campioni selezionati casualmente con sostituzione dal set di addestramento. La selezione casuale con sostituzione consente di creare diversi campioni di dati, detti campioni bootstrap, che possono includere osservazioni duplicate o differire l'uno dall'altro.
Per ogni campione bootstrap viene costruito un modello predittivo utilizzando l'intero set di addestramento. Questi modelli individuali, noti come "base learner" sono costruiti in modo indipendente e spesso utilizzano la stessa configurazione di modello e gli stessi iperparametri.
Le previsioni dei modelli individuali vengono poi combinati attraverso una media delle previsioni (nel caso della regressione) o una votazione maggioritaria (nel caso della classificazione) per ottenere la previsione finale del modello di bagging.
Il bagging è particolarmente utile per ridurre la varianza delle previsioni dei modelli. Poiché i campioni bootstrap vengono selezionati casualmente con sostituzione, ogni modello di bagging viene addestrato su un insieme di dati leggermente diverso, portando a una diversificazione delle previsioni dei singoli modelli e alla riduzione della varianza complessiva.
Il random forest, invece di considerare tutte le variabili predittive disponibili per la scelta del miglior split [1] in ogni nodo dell'albero, seleziona solo un sottoinsieme casuale i variabili predittive.
In questo modo le variabili vengono randomizzate attraverso una selezione casuale di un sottoinsieme di variabili predittive da utilizzare.
Il numero di variabili selezionate può essere fissato in anticipo o può essere determinato tramite tecniche come la selezione casuale o la selezione basata sulla importanza delle variabili. Quindi, solo le variabili selezionate vengono considerate per la scelta del miglior split in ogni nodo dell'albero.
Questa randomizzazione delle variabili aiuta a ridurre la correlazione tra gli alberi; questo perché, selezionando solo alcune variabili si introduce diversità nel processo

[1] strategia utilizzata per suddividere i nodi degli alberi durante la fase di addestramento

di addestramento. Pertanto, si riduce la correlazione tra le previsioni degli alberi, poiché ogni albero contribuirà in modo diverso alle previsioni finali. Questa riduzione della correlazione è vantaggiosa perché gli alberi altamente correlati tendono a fornire informazioni ridondanti e non apportano un beneficio significativo al Random Forest. Riducendo la correlazione tra gli alberi, si aumenta la diversità e l'indipendenza delle previsioni, consentendo al Random Forest di raggiungere una migliore accuratezza e una maggiore capacità di generalizzazione sui dati di test[2] .

I vantaggi di questo modello sono numerosi, tra cui:

- capacità di gestire un gran numero di variabili predittive;
- robustezza all'overfitting; L'overfitting è il fenomeno che si verifica quando un modello di machine learning si adatta troppo bene ai dati di addestramento, al punto da perdere la sua capacità di generalizzazione su nuovi dati. Questo può portare a prestazioni scadenti del modello quando viene applicato a dati non visti durante l'addestramento, rendendo le sue previsioni inaccurate e poco affidabili;
- possibilità di ottenere stime dell'importanza delle variabili predittive;
- riduzione della varianza rispetto a un singolo albero decisionale.

L'obiettivo del Random Forest è quello di creare un insieme di alberi di decisione, ognuno dei quali viene addestrato su un campione di dati selezionato casualmente dal set di addestramento.

4.1.1 Algoritmo RF

L'algoritmo del Random Forest funziona nel seguente modo [3].
Prima di tutto viene creato un insieme di campioni bootstrap utilizzando il processo di bagging sopra descritto. Per ogni campione, viene costruito un albero decisionale completo utilizzando l'algoritmo CART.
Durante la costruzione di ciascun albero, viene effettuato uno split in ogni nodo, per separare i dati in base alle variabili predittive.
Lo split viene determinato utilizzando una misura di impurità al fine di massimizzare l'omogeneità delle sottopopolazioni create dallo split. Per i problemi di regressione, come nel nostro caso, viene utilizzato l'errore quadratico medio (MSE). Altri criteri

[2] James *et al.* (2013)
[3] Hastie *et al.* (2009)

di impurità possono includere l'entropia o la deviazione standard dei valori target. L'algoritmo per determinare lo split in un nodo può essere descritto come segue:

- Per ogni variabile predittiva disponibile nel sottoinsieme casuale selezionato per lo split, si considerano diversi punti di suddivisione possibili. Questi punti di suddivisione possono essere fissati in base ai valori unici della variabile o possono essere selezionati in modo più intelligente, ad esempio, considerando le medie dei valori di due punti adiacenti;
- Per ogni punto di suddivisione, i dati vengono separati in due gruppi, uno che soddisfa la condizione di suddivisione e l'altro che non la soddisfa;
- Viene calcolata la misura di impurità per ciascuno dei due gruppi ottenuti utilizzando la misura di impurità scelta (MSE);
- Viene calcolata una misura di impurità ponderata per la suddivisione, che tiene conto delle dimensioni dei due gruppi ottenuti. Questa misura è ottenuta combinando le misure di impurità dei due gruppi pesate in base alle loro dimensioni relative;
- Vengono valutati tutti i punti di suddivisione possibili e viene selezionato il punto di suddivisione che minimizza la misura di impurità ponderata. Questo punto di suddivisione viene quindi utilizzato per dividere i dati nel nodo corrente.

Questo processo viene ripetuto ricorsivamente per ciascun nodo dell'albero fino a quando viene soddisfatto un criterio di stop, ad esempio, raggiungendo una profondità massima o un numero minimo di campioni in un nodo. Successivamente viene creato un ensemble di alberi indipendenti, ognuno costruito su un campione bootstrap diverso. Infine, per effettuare una previsione sui nuovi dati, i campioni vengono passati attraverso ciascuno degli alberi del Random Forest, e viene calcolata la previsione di ogni albero. La previsione finale è ottenuta facendo una media tra le previsioni di tutti gli alberi.

4.2 Random Survival Forest

Introduciamo ora il modello Random Survival Forest (RSF)[1].
Come abbiamo già affrontato e discusso, la costruzione di ensamble a partire da

[1] Ishwaran *et al.* (2008)

modelli base, come gli alberi, può migliorare sostanzialmente le prestazioni di previsione.
Recentemente, Breiman (2001) ha dimostrato che l'apprendimento di un ensemble può essere ulteriormente migliorato inserendo casualità nel processo di apprendimento di base, approccio chiamato Random Forest (RF).
La metodologia delle RSF estende il modello del RF di Breiman.
Vengono introdotte nuove regole di suddivisione per la sopravvivenza nella creazione degli alberi. Queste regole sono basate sulla massimizzazione della differenza di sopravvivenza tra i nodi figli durante la creazione degli alberi.
Viene inoltre introdotto il principio di conservazione degli eventi per le random forest di sopravvivenza, utilizzato per definire la "mortalità complessiva dell'ensable". Questa misura rappresenta un'indicazione sintetica e interpretabile della mortalità prevista dall'ensemble delle RSF. In sostanza, tiene conto delle previsioni di sopravvivenza fatte da tutti gli alberi e le combina per ottenere una stima complessiva della mortalità.
L'obiettivo principale è comprendere come variabili specifiche influenzino il tempo di sopravvivenza. Pertanto, durante la suddivisione di un nodo in un albero di sopravvivenza, viene selezionata la variabile candidata che massimizza la differenza tra i nodi figli generati dalla suddivisione (split).
Un meccanismo di split molto efficace per la Random Survival Forest è il *log-rank splitting*.
Il log-rank splitting è una modalità di split utilizzata nel contesto del RSF che viene introdotto come una regola di suddivisione specifica per i nodi degli alberi di sopravvivenza.
Viene utilizzato per confrontare le distribuzioni di sopravvivenza tra due o più gruppi.
L'obiettivo del log-rank splitting è identificare le divisioni che massimizzano la significatività statistica tra i gruppi di sopravvivenza. Ciò significa che le suddivisioni selezionate saranno quelle che separano meglio i pazienti con diverse durate di sopravvivenza e che presentano differenze statisticamente significative tra di loro.
Questo ci consente di identificare le variabili che influenzano in modo significativo la sopravvivenza e di creare alberi di sopravvivenza più informativi. Questo approccio contribuisce a migliorare la precisione e l'interpretabilità delle previsioni dei modelli RSF per i dati di sopravvivenza censurati.
Il metodo appena descritto può avere una significativa perdita di potenza in alcune circostanze, soprattutto quando le funzioni di sopravvivenza e di rischio si incrociano nei due gruppi confrontati.
Sarebbe possibile utilizzare l'uso della differenza assoluta integrata tra le funzioni

di sopravvivenza dei due nodi figli come regola di divisione.

$$L_1 = (n_L n_R) \int_t |\hat{S_L}(t) - \hat{S_R}(t)| dt \tag{4.1}$$

dove $\hat{S_L}(t)$ e $\hat{S_R}(t)$ rappresentano le stime della funzione di sopravvivenza Kaplan-Meier relative al nodo sinistro e destro e n_L e n_R il numero di osservazioni del nodo sinistro e destro.

Chiamiamo questa regola $L_1 splittingrule$.
Questa è correlata alla statistica di prova proposta da Lin e Xu (2010) (Moradian *et al.* (2016)).

4.2.1 Algoritmo RSF

Prima di procedere alla spiegazione del funzionamento dell'algoritmo vediamo come devono essere strutturati i dati.
Si parte da un set di dati di sopravvivenza con informazioni sul tempo di sopravvivenza e lo stato di censura. I dati vengono poi suddivisi in un insieme di addestramento e un insieme di test.
Questa suddivisione è una pratica fondamentale nell'apprendimento automatico per valutare l'efficacia di un modello predittivo su dati non visti durante la fase di addestramento. L'obiettivo principale è quello di valutare quanto bene il modello generalizza su nuovi dati, ossia dati che non sono stati utilizzati per addestrare il modello.

Come funziona l'algoritmo?

Si definisce il numero desiderato di alberi, indicato come B. Per ogni albero si estrae un campione bootstrap con ricampionamento dei dati dall'insieme di addestramento. In media ogni campione bootstrap esclude il 37% dei dati, che vengono chiamati out of bag (OOB).
Quando si utilizza il bagging, viene creato un insieme di alberi decisionali utilizzando campioni di addestramento selezionati casualmente con sostituzione dal set di addestramento originale. Ciò significa che alcuni campioni possono essere selezionati più volte, mentre altri possono non essere selezionati affatto.
Il concetto di OOB sfrutta questa selezione casuale per valutare le prestazioni del modello senza la necessità di un set di dati di test separato. Durante il processo di addestramento, alcuni campioni non vengono mai selezionati per la creazione di un

determinato albero. Questi campioni non selezionati costituiscono l'OOB set.

Successivamente si costruisce un albero di sopravvivenza utilizzando il campione estratto e in ogni nodo si seleziona casualmente un sottoinsieme di p variabili candidate.
Si sceglie la variabile candidata che massimizza la differenza di sopravvivenza tra i nodi figli come regola di suddivisione e si continua a suddividere i nodi fino a quando si raggiunge una condizione di stop (ad esempio quando un nodo terminale contiene un numero minimo di eventi di morte unici).
Viene poi calcolata la funzione di sopravvivenza cumulativa (CHF), una misura utilizzata nell'analisi di sopravvivenza per stimare la probabilità cumulativa di raggiungere un evento avverso o di fallimento entro un dato istante temporale.
Successivamente si calcola la media delle CHF di tutti gli alberi per ottenere quella dell'ensemble.
Una volta addestrato il modello RSF, è possibile utilizzarlo per fare previsioni sui nuovi dati di sopravvivenza e valutare l'effetto delle variabili sul tempo di sopravvivenza.

4.3 Applicazione del modello RSF ai dati

Prima di procedere all'applicazione del modello ai dati, è necessario suddividere il dataset in set di addestramento e di test. In questo modo il modello verrà addestrato nel set di training e successivamente verrà verificata la sua validità con set di test.
Bisogna garantire che la divisione del dataset mantenga proporzionalità tra eventi e censure. In questo modo si contribuisce a evitare una possibile distorsione delle stime di sopravvivenza e a ottenere una valutazione accurata delle prestazioni del modello.
Nel nostro caso, abbiamo prima creato due dataset diversi, uno contente solo gli eventi e uno solo le censure.
Successivamente abbiamo diviso entrambi i dataset, creando il dataset di addestramento (75%) e il dataset di test (25%) per entrambi i set.
Infine abbiamo combinato i due dataset di training (per eventi e censure) e creato il dataset che useremo come addestramento. Allo stesso modo abbiamo fatto per quello di test.

I dati sono così distribuiti:

Dataset di addestramento		
Dataset	N. pazienti	Outcome
A	839	NIV: 371 (44.22%) DEATH: 384 (45.77%) NONE: 84 (10.01%)
B	911	PEG: 308 (33.81 %) DEATH: 479 (52.58%) NONE: 124 (13.61%)
C	931	DEATH: 798 (85.71%) NONE: 133 (14.29%)
Dataset di test		
A	277	NIV: 111 (40.07%) DEATH: 140 (50.54%) NONE: 26 (9.39%)
B	300	PEG: 89 (29.67%) DEATH: 171 (57,00%) NONE: 40 (13.33%)
C	308	DEATH: 264 (85.71%) NONE: 44 (14.29%)

Tabella 4.1: Set di dati di addestramento e test per i dataset A, B e C

Possiamo ora procedere all'applicazione del Random Survival Forest ai dataset A, B e C.

4.3.1 Applicazione del modello per il Dataset A

Il modello Random Survival Forest produce il seguente output[4]:

Sample size	839
Number of events	755
Number of trees	1000
Forest terminal node size	15
Average no. of terminal nodes	44.437
No. of variables tried at each split	3
Total no. of variables	20
Resampling used to grow trees	swor
Resample size used to grow trees	530
Analysis	RSF
Family	surv
Splitting rule	logrank *random*
Number of random split points	10
(OOB) CRPS	0.08406213
(OOB) Requested performance error	0.40030369

Tabella 4.2: Random Survival Forest per il dataset A

Il modello RSF (Random Survival Forest) è stato addestrato utilizzando un insieme di 839 osservazioni. Tra queste, 755 sono eventi (NIV o DEATH). Sono stati costruiti 1000 alberi.
La dimensione dei nodi terminali dell'albero è stata impostata a 15, mentre in media ogni albero ha circa 44.4 nodi terminali.
Durante la costruzione degli alberi, sono stati considerati 3 predittori (variabili) per ogni suddivisione, tra un totale di 21 variabili disponibili.
La tecnica di campionamento utilizzata per la crescita degli alberi è stata "swor" (sampling without replacement), e la dimensione del campione utilizzata per la crescita degli alberi è stata di 530 osservazioni. Le resanti 309 osservazioni compongono l'Out-Of-Bag.
L'OOB è il termine usato per descrivere i dati che non vengono selezionati per l'addestramento di un particolare albero. Poiché questi dati non vengono utilizzati per costruire l'albero, potrebbero essere utilizzati per valutare la performance del modello.
In pratica, l'OOB fornisce una stima dell'errore di previsione del modello basata sui dati non utilizzati per addestrare ciascun albero.
Viene utilizzato come una misura di validazione interna del modello Random Forest, fornendo un'indicazione approssimativa dell'accuratezza delle previsioni sul set di dati di addestramento.

[4]Ishwaran e Kogalur (2023)

Il modello è stato addestrato utilizzando l'analisi RSF (Random Survival Forest) per la modellazione di dati di sopravvivenza. La regola di suddivisione utilizzata è stata il logrank test con selezione casuale dei punti di suddivisione.

Sample size of test (predict) data	277
Number of grow trees	1000
Average no. of grow terminal nodes	44.437
Total no. of grow variables	20
Resampling used to grow trees	swor
Resample size used to grow trees	530
Analysis	RSF
Family	surv
CRPS	0.06102574
Requested performance error	0.39665305

Tabella 4.3: Previsione sul RFS dataset A

I risultati della predizione mostrano che sono stati utilizzati 1000 alberi nella costruzione del modello RSF. La dimensione del campione di dati di test è di 277 osservazioni.

L'indice di errore di performance richiesto, che nel nostro caso è di 0.39665305, rappresenta il livello di errore massimo accettabile per la valutazione del modello. In altre parole, è l'obiettivo che il modello cerca di raggiungere in termini di accuratezza delle previsioni.
L'indice di errore di performance ottenuto invece, misurato come CRPS (Cumulative Random Probability Score), ha un valore di 0.06102574. Il CRPS è una misura di valutazione della calibrazione di un modello di previsione di sopravvivenza. Viene utilizzato principalmente nei modelli di sopravvivenza per valutare quanto bene il modello è in grado di stimare la probabilità di sopravvivenza cumulativa nel tempo. Viene calcolato confrontando le previsioni del modello con le osservazioni reali. Per ogni osservazione di sopravvivenza, il CRPS tiene conto della probabilità prevista dal modello e della probabilità effettiva di osservare un evento o una censura in quel punto temporale.
In pratica, il CRPS misura la discrepanza tra la previsione del modello e la realtà osservata, tenendo conto sia delle probabilità di sopravvivenza che delle probabilità di evento o censura. Un valore più basso del CRPS indica una migliore calibrazione del modello, ovvero una previsione più accurata delle probabilità di sopravvivenza cumulativa nel tempo.

Possiamo dire quindi che l'errore ottenuto è inferiore all'errore di performance che ci aspettavamo di ottenere.

Attraverso il Random Survival Forest è possibile andare a valutare anche l'importanza delle variabili, come possiamo osservare nel grafico 4.1

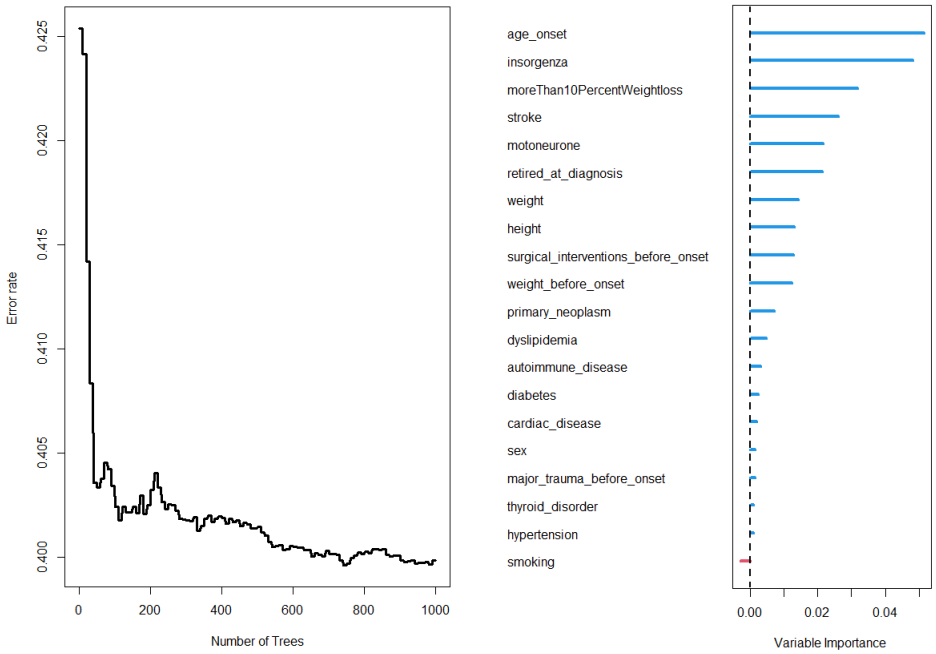

Figura 4.1: Importanza delle variabili per il dataset A

Notiamo che il numero di alberi che minimizza l'errore è circa 750 e che le variabili *age_ onset* e *insorgenza* sono quelle che risultano più importanti.

4.3.2 Applicazione del modello per il Dataset B e C

Per quanto riguarda il dataset B e C, sono riportati nella tabella 4.4 e 4.5 gli output dei risultati delle previsioni.

I risultati della previsione mostrano che sono stati utilizzati 1000 alberi nella costruzione del modello RSF.
La dimensione del campione di dati di test è di 300 e 308 osservazioni rispettivamente per il dataset B e C.

Sample size of test (predict) data	300
Number of grow trees	1000
Average no. of grow terminal nodes	47.863
Total no. of grow variables	20
Resampling used to grow trees	swor
Resample size used to grow trees	576
Analysis	RSF
Family	surv
CRPS	0.07277872
Requested performance error	0.38313186

Tabella 4.4: Previsioni per il dataset B

Sample size of test (predict) data	308
Number of grow trees	1000
Average no. of grow terminal nodes	48.693
Total no. of grow variables	20
Resampling used to grow trees	swor
Resample size used to grow trees	588
Analysis	RSF
Family	surv
CRPS	0.08354083
Requested performance error	0.36201794

Tabella 4.5: Previsioni per il dataset C

In entrambi i casi, come per il dataset A, l'indice di errore di performance che ci aspettavamo di ottenere (Requested performance error), è superiore a quello che abbiamo ottenuto (CRPS).

Andiamo ora a vedere l'importanza delle variabili e il numero di alberi necessari per minimizzare l'errore.

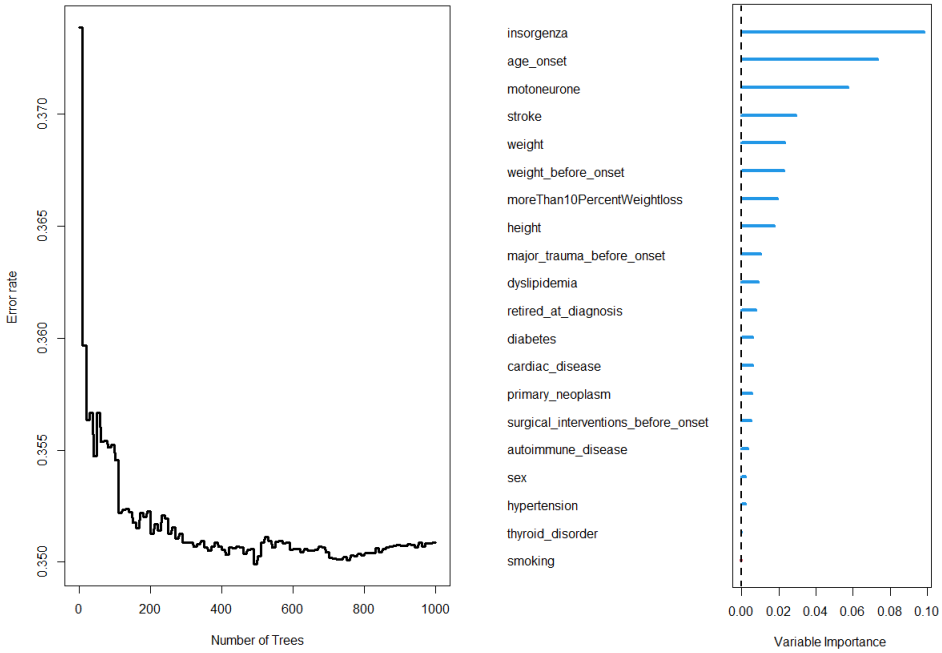

Figura 4.2: Importanza delle variabili per il dataset B

Per quanto riguarda il dataset B, possiamo notare dal grafico 4.2 che il numero ottimale di alberi è circa 500, e le variabili ritente più importanti sono *insorgenza*, *age_ onset* e *motoneurone*.

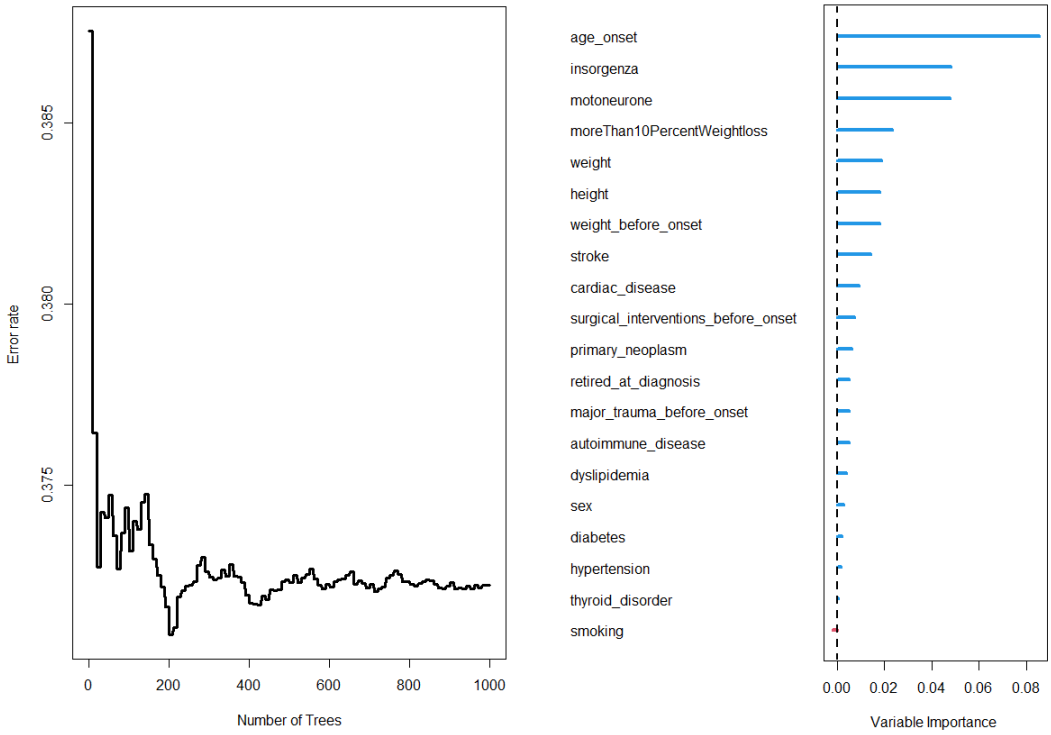

Figura 4.3: Importanza delle variabili per il dataset C

Per quanto riguarda il dataset C invece, il numero di alberi scende a 200 e le variabili ritenute più importanti sono anche in questo caso (seppur in ordine diverso) *age_ onset*, *insorgenza* e *motoneurone*.

Risultati e conclusioni

Nel presente lavoro, attraverso l'utilizzo di modelli parametrici, non parametrici e semiparametrici, idonei alla gestione dei dati di sopravvivenza, abbiamo cercato di capire quali sono le variabili che influenzavano maggiormente il nostro evento di interesse, ovvero la previsione di NIV (ventilazione non invasiva), PEG (gastrotomia endoscopica percutanea) o DEATH (morte).

Da una prima analisi esplorativa, abbiamo visto che, indipendentemente dall'evento sotto esame, il genere (maschio o femmina) non è una caratteristica che influenza il tempo trascorso al verificarsi dell'evento .
In generale, le variabili che abbiamo notato avere un impatto significativo riguardano l'età dei pazienti, il luogo di insorgenza e la modalità con cui si presenta la malattia.
Una volta analizzate queste variabili, e verificato il divario tra i gruppi attraverso l'analisi non parametrica di Kaplan-Meier e i relativi test di verifica, abbiamo proceduto applicando il modello semiparametrico di Cox e il modello non parametrico Random Survival Forest.

Il Random Survival Forest (RSF) è un potente algoritmo di apprendimento automatico utilizzato per analizzare dati di sopravvivenza. Tuttavia, a differenza di altri modelli di sopravvivenza, come il modello di Cox, il RSF può essere di difficile interpretazione.
Una delle ragioni principali è che il RSF combina l'output di un insieme di alberi decisionali, ognuno dei quali stima la sopravvivenza in modo indipendente. Pertanto, l'interpretazione diretta dei singoli alberi può essere complessa. Inoltre, la complessità del RSF aumenta con il numero di alberi utilizzati nel modello, rendendo ancora più difficile la comprensione del modello nel suo complesso.
Inoltre, il RSF non fornisce stime dirette dei coefficienti associati alle variabili predittive, come avviene nel modello di Cox.
D'altra parte, il modello di Cox fornisce stime dei coefficienti associati a ciascuna

variabile, consentendo di interpretare l'effetto di ogni variabile sulla sopravvivenza dei pazienti. Inoltre, il modello di Cox può tener conto delle interazioni tra variabili, offrendo una visione più completa dei fattori che influenzano la sopravvivenza.

In conclusione, sebbene il Random Survival Forest sia potente per l'analisi dei dati di sopravvivenza, la sua complessità e la mancanza di stime dirette dei coefficienti e delle interazioni possono renderlo di difficile interpretazione. In confronto, il modello di Cox offre una maggiore interpretabilità grazie alle stime dei coefficienti e alla considerazione delle interazioni tra variabili.
Per questo motivo, la scleta del modello per questa analisi, cade sul modello di Cox.

Bibliografia

Hastie T.; Tibshirani R.; Friedman J. (2009). *The Elements of Statistical Learning*. Springer, New York.

Ishwaran H.; Kogalur U. B. (2023). *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*.

Ishwaran H.; Kogalur U. B.; Blackstone E. H.; Lauer M. S. (2008). The annals of applied statistics. pp. 841–860.

James G.; Witten D.; Hastie T.; Tibshirani R. (2013). *An Introduction to Statistical Learning*. Springer, New York.

Moradian H.; Larocque D.; Bellavance F. (2016). L1 splitting rules in survival forests. *Springer Science+Business Media*.