

POLITECNICO DI TORINO

Corso di Laurea in Ingegneria Gestionale

Tesi di Laurea Magistrale

**Survival Analysis e Rischio di Insolvenza nel Settore
delle Apparecchiature Elettriche**



Relatore:

Franco Varetto

Candidato:

Maria Mangano

Luglio 2021

Sommario

Introduzione	6
Il Rischio di Credito	7
Analisi del settore	12
Codice Ateco	12
Database di riferimento.....	17
Censored Data	20
Censura di tipo I.....	20
Censura di tipo II.....	22
Censura di tipo III.....	22
Censura a destra, sinistra e ad intervallo	23
Survival Time	25
Survivorship Function	25
Probability Density Function	27
Hazard Function.....	28
Esempio: Stima delle funzioni di sopravvivenza.....	31
Relazione tra le funzioni di sopravvivenza	34
Metodi Non Parametrici per la Stima delle Funzioni di Sopravvivenza	37
Stima Product Limit della Funzione di Sopravvivenza	38
Esempio: Stima della funzione di sopravvivenza con il metodo Kaplan - Meier	39
Stima PL di Kaplan-Meier	41
Criticità del metodo di Kaplan-Meier	47
Analisi delle tavole di vita.....	48
Tabelle di vita della popolazione	49
Tavole di vita cliniche	53
Distribuzioni Parametriche di Sopravvivenza	57
Distribuzione Esponenziale.....	57
Distribuzione di Weibull	59
Distribuzione Log-normale	62
Distribuzione Log-logistica.....	63
Distribuzione Gamma	65
Distribuzione di Gompertz.....	67
Calcolo delle Stime per le Distribuzioni Parametriche di Sopravvivenza in assenza di Covariate	69
Procedure di stima (MLE) per i dati con osservazioni censurate a destra	69
Procedure di stima (MLE) per i dati con osservazioni censurate a destra, a sinistra e per intervallo	72
Metodi Grafici per l'Adattamento della Distribuzione di Sopravvivenza	83

Probability Plotting	84
Hazard Plotting	88
Metodo Dei Residui Di Cox-Snell	91
Metodi Parametrici Per L'adattamento Del Modello Di Regressione	94
L'identificazione Dei Fattori Prognostici	94
Struttura Generale Dei Modelli Di Regressione Parametrica	97
I residui di Cox- Snell in presenza delle Covariate	101
Cox Proportional Hazard Model	103
Funzione Di Verosimiglianza Parziale Per I Tempi Di Sopravvivenza	103
Procedure di Stima in Assenza di Tempi di Sopravvivenza Legati	107
Procedure di Stima in Presenza di Tempi di Sopravvivenza Legati	109
Scala Di Tempo Continua	109
Scala Di Tempo Discreta	110
Stima Delle Funzioni Di Sopravvivenza Con Le Covariate	111
Valutazione della bontà di adattamento dei residui	113
Modelli di Rischio Non Proporzionali	115
Modelli con Covariate che variano nel Tempo	115
Analisi dei risultati	117
Software e dati di partenza	117
Analisi Non Parametrica	121
Analisi Non Parametrica - Società	121
Analisi Non Parametrica - Fatturato	125
Analisi Non Parametrica - Area Geografica	127
Analisi Parametrica	130
Distribuzione Esponenziale	131
Distribuzione di Weibull	135
Distribuzione di Gompertz	139
Distribuzione Log-normale	141
Distribuzione Log-logistica	143
Distribuzione Gamma	145
Cox Proportional Hazard Model	147
Conclusioni	151
ALLEGATO A	153
ALLEGATO B	154
ALLEGATO C	155
ALLEGATO D	156

ALLEGATO E	157
ALLEGATO F	158
ALLEGATO G.....	160
ALLEGATO H.....	161
ALLEGATO I	163
Bibliografia.....	165

Introduzione

La Survival Analysis è un'applicazione statistica prevalentemente utilizzata in ambito medico per valutare la probabilità di sopravvivenza di un paziente affetto da una ben precisa malattia oppure sottoposto a ben precise cure mediche. Tuttavia, poiché si tratta di un modello molto flessibile, è possibile adattarlo a contesti differenti. Infatti, si vuole utilizzare la Survival Analysis per valutare la probabilità di sopravvivenza di un certo numero di imprese operanti nel settore delle apparecchiature elettriche, di conseguenza la probabilità che esse vadano in default. Per fare questo si utilizzano per prima cosa i metodi non parametrici, che registrano semplicemente se l'impresa nel corso del suo esercizio ha sperimentato il default o meno, poi i metodi parametrici, per i quali si prevede l'inserimento di un certo numero di covariate, o fattori prognostici, descrittive della situazione economico-patrimoniale dell'impresa e, infine, l'applicazione del modello regressivo dei rischi proporzionali proposto da Cox che prevede, oltre alle variabili indipendenti, alcune assunzioni forti come il fatto che le covariate siano indipendenti dal tempo. La probabilità di default è espressa in termini di anni di esercizio dell'impresa, valutati al momento della costituzione. La Survival Analysis è uno strumento molto potente che ci permette di trarre conclusioni affidabili anche quando ci troviamo di fronte ad osservazioni censurate o scorrette. Proprio per questo motivo può essere utilizzata come strumento per fare forecasting, ossia per capire cosa potrebbe succedere a delle imprese in un particolare contesto economico-patrimoniale all'interno di un ben preciso settore. Questo strumento può essere ulteriormente migliorato introducendovi delle variabili rappresentative della situazione economica di un certo periodo di tempo, ad esempio variabili macroeconomiche come il PIL, della situazione di un determinato Paese o settore oppure considerando l'evoluzione temporale dei fattori prognostici presi in esame.

Il Rischio di Credito

Il rischio di credito è indubbiamente uno dei rischi di mercato di più difficile quantificazione. Esso rappresenta l'eventualità che una delle parti di un contratto non rispetti gli obblighi finanziari assunti, cagionando una perdita per la controparte creditrice; questo può tradursi nel mancato pagamento degli interessi o anche nella mancata restituzione della quota capitale.

Il rischio di credito può essere definito come la possibilità che da una variazione inattesa del merito creditizio di un debitore derivi una variazione inattesa del valore del credito.

Questa tipologia di rischio è insita in ogni transazione finanziaria e rappresenta uno dei fattori rilevanti nella determinazione dei prezzi e dei rendimenti delle attività finanziarie.

Gli strumenti finanziari che possono essere soggetti al Rischio di Credito si possono distinguere in:

- Titoli di Debito
- Posizioni fuori bilancio

Nella prima categoria possiamo trovare i titoli di stato, titoli di debito emessi da enti pubblici, le corporate bonds che sono obbligazioni emesse da società private, altre forme di finanziamento concesse alle aziende, i mutui e in generale il credito al consumo.

Le posizioni fuori bilancio comprendono, invece, i titoli derivati trattati nei mercati regolamentari, i titoli derivati trattati nei mercati over the counter per cui esiste il rischio di controparte, i titoli derivati il cui sottostante comporta il rischio di credito e, infine, i derivati creditizi.

Ad esempio, una società che ha degli obiettivi di crescita ma non i sufficienti mezzi finanziari, può chiedere in prestito i fondi necessari ad una banca o può emettere dei titoli obbligazionari. Potrebbe succedere che, prima del pagamento delle cedole o del rimborso del debito, la controparte debitrice fallisca o incontri delle difficoltà finanziarie. Il debito sarà quindi rimborsato solo in parte o per niente. È proprio a causa di questo rischio di insolvenze che la controparte creditrice richiederà un tasso di interesse significativamente più alto rispetto a quello che caratterizza investimenti meno rischiosi. Quindi, a maggiore rischio corrisponde un maggiore tasso di interesse, come premio per l'investitore che ha deciso di sostenere l'impresa in difficoltà.

Secondo l'Accordo di Basilea, il rischio di credito può essere descritto da due variabili:

- Il **risk of recovery**, che si riferisce alla severità della perdita in caso di default ed è misurato dalla Loss Given Default (LGD).
- Il **risk of default**, che si riferisce al rischio di un cliente dell'istituzione finanziaria ed è misurato dalla Probabilità di Default (PD).

L'entità di questo rischio negli investimenti può essere valutata sfruttando due parametri:

- il **Rating**, ovvero il giudizio sul merito creditizio di un soggetto emittente titoli che incorporano il rischio di credito.
- il **CDS (Credit Default Swap)**, uno strumento finanziario attraverso il quale è possibile misurare il rischio di credito.

Il rating può far riferimento alla singola posizione creditoria ma anche al soggetto debitore. Può essere assegnato da alcune agenzie o anche dalla banca stessa, secondo diverse metodologie, ed in quest'ultimo caso si parla di rating interno. Per arrivare ad un giudizio si utilizzano sia informazioni qualitative che quantitative. Molto spesso non ci si affida ai soli calcoli matematici, ma, per l'assegnazione del rating, il giudizio e l'esperienza dell'analista creditizio giocano un ruolo fondamentale.

Sono diversi i fattori economici e non che vengono presi in considerazione in questa analisi, ad esempio:

- **Le prospettive di guadagni futuri e i futuri cash-flow:** maggiore è la capacità di generare redditi e flussi di cassa, migliore è la situazione dell'impresa rispetto al debito.
- **La crescita:** bisogna tenere a mente che se si prevede una crescita aziendale questa comporta elevati fabbisogni finanziari per investimenti e circolante.
- **La variabilità dei redditi e dei cash-flow:** maggiore è la volatilità dei redditi, maggiore è la probabilità di incontrare difficoltà impreviste.
- **Le riserve da utili:** più elevate sono le riserve da utili reinvestite, maggiore è stata la redditività passata e maggiore è la quota di attività finanziaria con risorse interne.
- **La struttura patrimoniale** (il livello di capitale proprio e quello di indebitamento): maggiori sono i debiti, maggiore sono le possibilità che l'impresa vada in difficoltà.
- **La struttura dell'indebitamento** (debito a breve, medio e lungo termine): più breve è la durata residua dei debiti, maggiore è la probabilità che l'impresa incorra in problemi di rimborso.
- **Il livello di liquidità:** maggiore è la liquidità più rapidamente sono disponibili i mezzi per rimborsare i debiti e pagare gli oneri finanziari.
- **Turnover dell'attivo:** più velocemente ruota l'attivo, più rapidamente l'impresa lo trasforma in liquidità.
- **La situazione politica e sociale di un Paese**
- **La situazione di mercato**
- **Il settore industriale** in cui la società opera o vi svolge le principali attività.
- **La qualità della classe dirigente** impatta molto anche sul potere contrattuale di una società.

- **La dimensione delle imprese:** le imprese più grandi hanno accesso ad un'ampia varietà di fonti finanziarie e quindi hanno più possibilità di diversificare.
- **Qualità utili e trasparenza nei bilanci:** imprese in difficoltà potrebbero manipolare i bilanci o cercare di nascondere delle informazioni.

Di seguito i livelli di giudizio delle principali agenzie di rating. Alla valutazione statica rappresentata dal rating, le agenzie aggiungono un'ulteriore valutazione ma di tipo dinamico e di monitoraggio sistematico del rischio di credito.

Tabella 1. Classi di Rating delle Principali Agenzie

Rating	Moody's	Fitch	S&P	Classificazione
AAA	Aaa	AAA	AAA, AAA-	Investment grade
AA	Aa, Aa1, Aa2, Aa3	AA, AA-, AA+	AA, AA-, AA+	Investment grade
A	A, A1, A2, A3	A, A-, A+	A, A-, A+	Investment grade
BBB	Baa, Baa1, Baa2, Baa3	BBB, BBB-, BBB+	BBB, BBB-, BBB+	Investment grade
BB	Ba, Ba1, Ba2, Ba3	BB, BB-, BB+	BB, BB-, BB+	Non investment grade
B	B, B1, B2, B3	B, B-, B+	B, B-, B+	Non investment grade
CCC	Caa, Caa1, Caa2, Caa3	CCC, CCC-, CCC+	CCC, CCC-, CCC+	Non investment grade
CC	Ca	CC, CC-, CC+	CC, CC-, CC+	Non investment grade
C	C	C, C-, C+	C	Non investment grade
DDD	DDD	DDD	DDD	Non investment grade
DD	DD	DD	DD	Non investment grade
D	D	D	D	Non investment grade

I CDS, **Credit Default Swap**, sono dei credit derivatives, ovvero dei contratti finanziari trattati over-the-counter. Consentono di negoziare i rischi associati al credito, permettendo di trasferirli fra soggetti economici indipendentemente dal sottostante. Il prezzo di un CDS è diretta conseguenza del rischio di insolvenza del debitore; è come se fosse un'assicurazione contro il fallimento di quest'ultimo. Ovviamente, tanto maggiore è il rischio, maggiore è il premio da corrispondere per assicurarsi dal fallimento. In pratica, i CDS permettono all'acquirente di trasferire, dietro pagamento di un premio, il rischio di credito riferito ad una attività sottostante. Il trasferimento è effettuato nei confronti del venditore che si impegna ad effettuare un pagamento all'acquirente nel caso in cui l'attività sottostante risulti inadempiente.

Proprio perché tutte le attività finanziarie sono esposte al rischio di credito, diventa di fondamentale importanza attuare delle **misure di sicurezza**, quali valutazioni di affidabilità del cliente, richieste di garanzie reali e/o personali, accantonamenti su base annua e fondi rischi sui credi. Ne consegue che il *rischio di credito* partecipa alla formazione del risultato economico gestionale tramite gli accantonamenti di idonee quote ai fondi.

Più in generale quindi, un creditore, per far fronte al rischio di credito, può richiedere di aggiustare il prezzo delle attività finanziarie in funzione del rischio, cioè andando a “caricare” la probabilità di default del debitore sul costo della transazione, oppure può richiedere dei collaterali a supporto della transazione stessa. In aggiunta è possibile ricorrere a forme di assicurazione contro il fallimento delle parti, come ad esempio ricorrendo ai CDS. Nei casi in cui il creditore ritenga che la copertura del rischio di credito non sia qualitativamente adeguata, o sia quantitativamente insufficiente, possono generarsi effetti negativi sulla concessione del credito stesso quali ad esempio la concessione di un ammontare inferiore rispetto a quello richiesto dal debitore.

Negli ultimi anni, l’attenzione della autorità di vigilanza, nazionali e internazionali, degli istituti creditizi e del comitato di Basilea, ha portato alla costruzione di modelli avanzati di valutazione dell’esposizione al *rischio di credito*, in base ai quali lo stesso viene quantificato non solo rispetto al singolo obbligato o operazione, ma anche rispetto al portafoglio crediti e obbligazionario nel suo complesso, derivando (con logiche [VAR](#) di tipo finanziario) misure attendibili e consistenti sul flusso di perdite medie attese e di perdite massime attese a un certo percentile di confidenza.

Analisi del settore

Codice Ateco

Le imprese di interesse per questa analisi fanno riferimento al codice **ATECO 27. Fabbricazioni di apparecchiature elettriche ed apparecchiature per uso domestico non elettriche**. Di seguito sono elencate tutte le sottocategorie considerate.

Figura 1.



Il comparto dell'elettronica e delle apparecchiature elettriche conta 22.035 aziende, delle quali 8.810 sono società di capitali e 3.162 risultano in crescita. Esse sono specializzate in una ben precisa sottocategoria, anche se non è raro trovare imprese che operano in più settori contemporaneamente. Le aziende che entrano a far parte di questo studio sono imprese italiane operanti in un intervallo di tempo che va dal 1995 al 2019.

Le principali aziende dello studio per dimensioni sono:

Tabella 2.

RAGIONE SOCIALE	FATTURATO 2019 (000 eur)	2019/2018 (%)	MOL 2019 (%)
SAMSUNG ELECTRONICS ITALIA S.P.A.	2.230.865	-10,7	2,96
ABB S.P.A.	1.987.711	0,0	10,63
STMICROELECTRONICS S.R.L.	1.829.599	4,0	10,75
SIEMENS S.P.A.	1.362.873	-1,0	5,71
HUAWEI TECHNOLOGIES ITALIA S.R.L.	1.309.651	-16,9	4,31
DE' LONGHI APPLIANCES S.R.L.	1.209.455	8,8	9,88
WHIRLPOOL EMEA S.P.A.	1.105.900	4,9	-20,83
CANDY HOOVER GROUP S.R.L.	990.146	87,9	0,11
ELECTROLUX ITALIA S.P.A.	969.540	4,0	9,44
BTICINO S.P.A.	872.020	2,1	26,32

La situazione del settore fino al 2019 può essere descritta dai grafici e dalle tabelle di seguito riportate.

Tabella 3. Numero di imprese, ricavi e addetti nel settore nel 2019

SRL	%	RICAVI	%2	ADDETTI	%3	FATT/ADD
3.297	4.2	14.754.611	4,4	57.984	4,4	254

Grafico 1. Clima di fiducia del settore e ciclo economico (2007-2019)

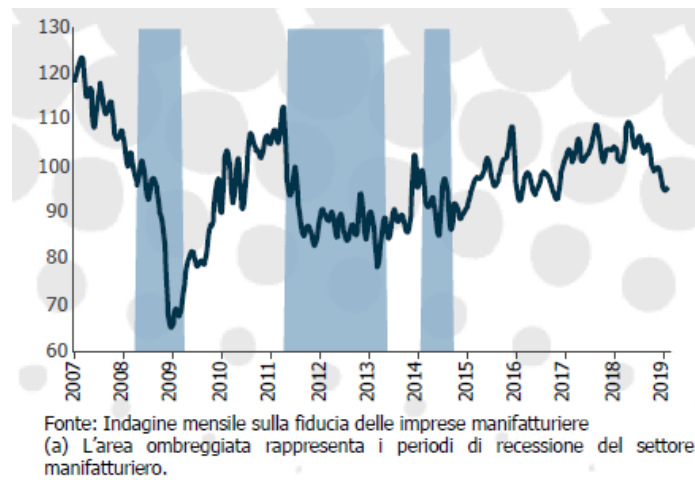


Grafico 2. Indicatore di tensione creditizia del settore e media manifatturiera- Anni 2013-2018

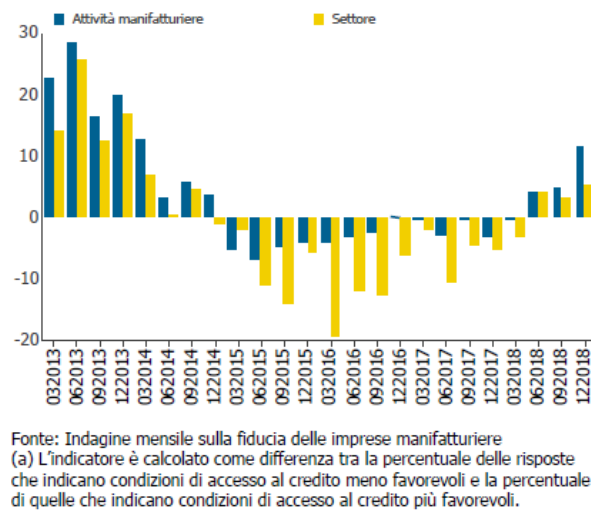


Grafico3. Esportazioni del settore e media manifattura dal 2006-2019 (variazioni rispetto allo stesso mese dell'anno precedente)

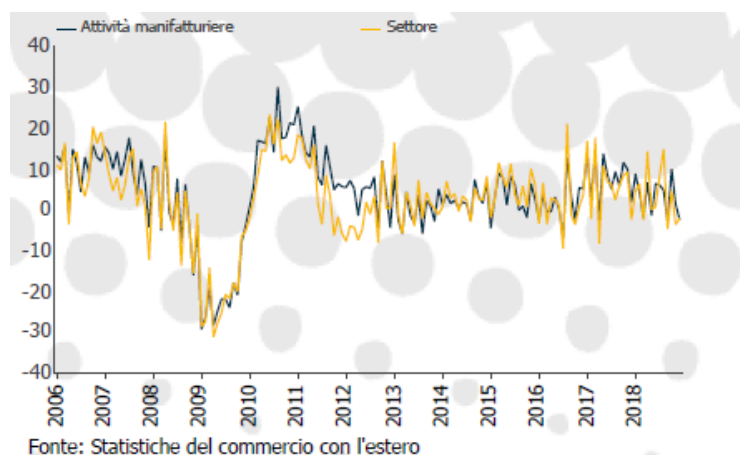


Grafico 4. Incidenza dell'export sul fatturato per le imprese esportatrici del settore e media manifattura 2014-2019 (In percentuale, dati destagionalizzati)

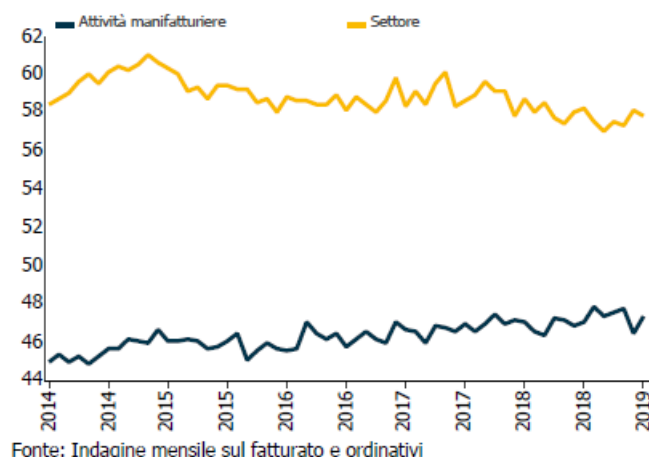
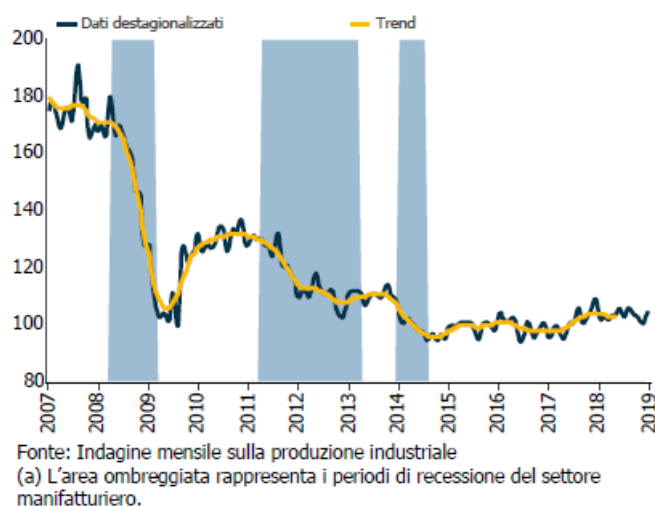
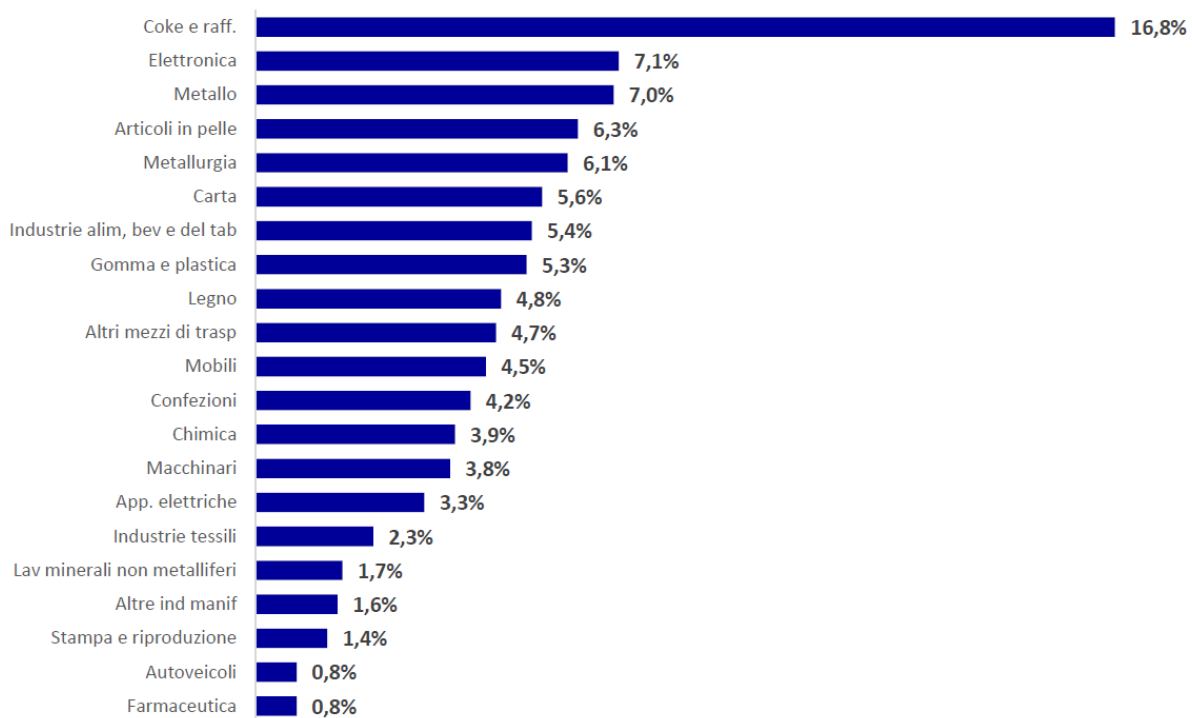


Grafico 5. Indice della produzione industriale del settore e ciclo economico dal 2007 al 2019 (Dati destagionalizzati)



Inoltre, è particolarmente interessante notare la variazione del fatturato nei vari comparti industriali e confrontarla con quella del settore dell'elettronica. Si nota un incremento molto significativo di fatturato rispetto al 2018 pari al 7,1%. Quello elettrico, pertanto, è un settore con buone prospettive e margine di crescita.

Grafico 6. Tasso di variazione annuale. Anno 2019



Database di riferimento

I dati utilizzati per la seguente elaborazione provengono dalla piattaforma AIDA. Il dataset è costituito da 8.671 imprese appartenenti al settore delle apparecchiature elettriche e apparecchiature non elettriche per uso domestico; per ciascuna di esse sono stati scaricati tutti gli anni di bilancio disponibili nella piattaforma. Per la maggior parte delle imprese sono disponibili 10 anni di bilancio, fino al 2019, e informazioni relative all'attuale stato giuridico e agli ultimi provvedimenti di cui l'azienda stessa è stata oggetto. Per altre imprese ci sono però a disposizione un minor numero di bilanci e informazioni mancanti circa i provvedimenti e lo stato giuridico. Tali imprese sono state ugualmente introdotte nella tesi perché, come vedremo più nel dettaglio in seguito, pur essendo caratterizzate da dati mancanti, possono essere portatrici di informazioni significative.

Per rendere fruibili queste informazioni, è stato utilizzato un file excel in cui sono state riportate tutte le informazioni utili per ogni azienda, quali:

- Ragione sociale
- Data di costituzione
- Anno di riferimento del bilancio
- Sede legale
- Forma Giuridica
- Descrizione attività svolta
- Stato giuridico
- Provvedimento con relativa data di inizio e/o di fine
- Codice Ateco
- Quotazione/non quotazione
- Modello di contabilità
- Le voci di Stato Patrimoniale

- Le voci di Conto Economico

Una volta riportate tutte queste informazioni sono stati calcolati gli indici di bilancio economici e patrimoniali. Di seguito sono riportati i principali indicatori utilizzati.

Tabella 4. Indicatori di Stato Patrimoniale e Conto Economico

Indicatori di Redditività	Indicatori di Produttività e Struttura operativa	Indicatori di Liquidità e Struttura finanziaria
EBITDA/Ricavi	Val Agg Oper/Ricavi	AC/PC
EBIT/ricavi	Consumi/Costi Operativi	AC-Mag/PC
Utile corrente/Ricavi	Servizi esterni/Costi operativi	Liq/PC
Risultato netto rettif/Ricavi	Costo Lavoro/Costi Operativi	Liq/AN
Risultato netto rettif/AN	Ammortam Materiali/Costi Operativi	Cap Circ/AN
EBITDA/AN	Val Agg/ITN	Patr netto/AN
ROA	Ricavi/AN	Riserve+Utile/AN
ROE	gg magazz	Patr Netto Tang/AN
ROE Ante Imposte		Patr Netto/Debiti Totali
		Patr Netto Tan/Debiti Tot+PN
		Patr Netto Tan/Debiti Tot-Liq+PN
		OF/RIC
		OFN/RIC
		OFN/EBITDA
		Deb Totali/VA
		Deb Totali/Ric
		PC/RIC
		Deb finanziari (stimati)/VA
		Deb Finanziari (stimati)/Ric
		Debiti Totali/EBITDA
		Debiti Finanziari/EBITDA
		Ebitda-Servizio Debito/AN
		Ln(AN)
		Ln(RIC)
		Autof Lordo/AN (=cash flow/attivo)
		Auto Lordo-Comp Straord/AN
		OFN/EBIT
		OFN/AN
		OFN/Autof lordo

Terminata questa fase di predisposizione dei dati, sono stati impostati i seguenti check:

- TOTALE ATTIVO = $\sum$ *Sottovoci dell'attivo*
- TOTALE PASSIVO = $\sum$ *Sottovoci del passivo*
- TOTALE ATTIVO = TOTALE PASSIVO
- CORRETTEZZA NEL CALCOLO DELL'UTILE DI ESERCIZIO

Sono emersi molti errori e la maggior parte delle squadrature è dovuta a dati mancanti, somme sbagliate, per cui i valori aggregati non corrispondono alla somma delle singole sottovoci, ed informazioni ridondanti. È stato necessario intervenire su ciascun errore messo in evidenza dai check.

Poiché si è intervenuto manualmente in molti bilanci, tra gli indicatori si è generato un problema di outlier. Per ripulire il database da questi errori, per ognuno di essi è stato calcolato il 5° ed il 95° percentile; i valori superiori al 95° percentile sono stati corretti inserendo al loro posto il valore del percentile. Medesimo ragionamento per i valori inferiori al 5° percentile.

Il passaggio conclusivo consiste nell'attribuire, per ogni azienda e per ogni anno di bilancio, un flag, sulla base dello stato giuridico e delle procedure cui la società è soggetta. L'attribuzione del flag segue le regole riportate in seguito.

Tabella 5.

FLAG	STATUS SOCIETA'	DESCRIZIONE
0	SANA SENZA PARTICOLARI SEGNALAZIONI	Nessun provvedimento in atto
1	ANOMALA	ES: Fallimento, Concordato preventivo, Bancarotta...
2	SANA IN CONDIZIONI PARTICOLARI	ES: Liquidazione volontaria, Scioglimento e liquidazione, Cessazione di ogni attività...
3	SANA IN CONDIZIONI PARTICOLARI	ES: Fusione mediante incorporazione in altra società, Scissione, Cessione d'Azienda...
4	SANA IN CONDIZIONI PARTICOLARI	ES: Cancellata dal registro di impresa, Trasferimento in altra provincia, Cessata...

Tuttavia, per distinguere le aziende che subiscono effettivamente l'evento che vogliamo studiare, ovvero le imprese che vanno in default, è necessario un ulteriore passaggio. Alle società sane, indipendentemente dallo status, viene attribuito il flag 0, alle società anomale, flag 1.

In questo modo tutte le imprese contraddistinte dal numero 1 sono quelle che subiscono l'evento di default e che quindi non entrano a far parte delle informazioni censurate.

Censored Data

Quando si effettua uno studio può accadere che alcuni soggetti non sperimentino l'evento oggetto di valutazione entro la fine dell'analisi. Se, ad esempio, si conduce un'indagine per un certo intervallo di tempo circa la mortalità di una determinata malattia, alcuni pazienti potrebbero essere ancora vivi allo scadere del tempo. Poiché terminato lo studio non si raccolgono più informazioni su queste persone, i dati raccolti sono incompleti ed i tempi esatti di sopravvivenza (di mortalità in questo caso) non sono noti, si parla di **osservazioni censurate**. Quando non sono presenti informazioni censurate, il dataset è **completo**.

Esistono 6 tipi di censura:

- Censura di tipo I
- Censura di tipo II
- Censura di tipo III
- Censura a destra
- Censura a sinistra
- Censura ad intervallo

Censura di tipo I

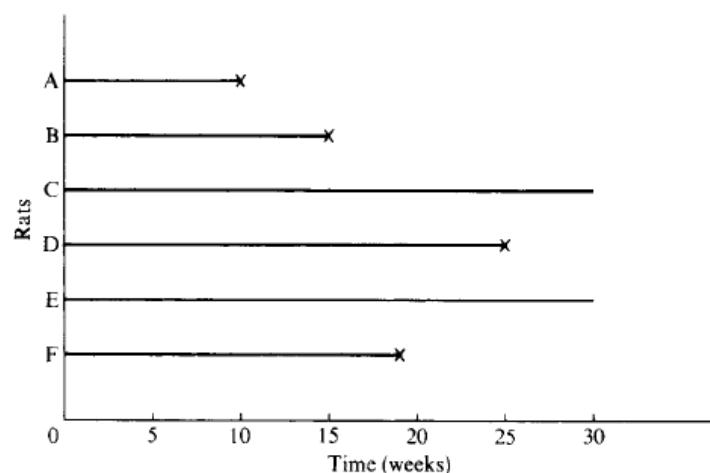
Supponiamo di eseguire un esperimento della durata di 30 settimane in cui un tumore viene iniettato in 6 topi e si osservano i tempi di manifestazione dei sintomi. Come è possibile notare dal grafico sottostante, tre dei sei topi sviluppano il tumore all'interno dell'intervallo di tempo previsto per lo studio (A, B, D), due topi non hanno ancora sviluppato la malattia entro la fine dello studio (C,E), ossia il tempo in cui sono ancora sani è superiore alle 30 settimane, ed infine il topo F muore accidentalmente senza

essersi ammalato dopo 19 settimane. I dati di sopravvivenza registrati sono 10, 15, 30+,25,30+,19+ settimane (il segno “+” identifica le osservazioni censurate).

- Per i topi A, B, D i tempi di sopravvivenza vanno dall’inizio dell’esperimento alla loro morte e pertanto si parla di **osservazioni esatte** o incensurate
- I tempi di sopravvivenza degli animali ancora vivi alla fine dell’esperimento non sono noti esattamente ma sono almeno pari alla durata dello studio. In questo caso ci troviamo di fronte ad **osservazioni censurate**.
- Le informazioni relative ad alcuni animali possono essere perse accidentalmente e anche in questo caso il loro tempo di sopravvivenza non è noto a priori e si tratta nuovamente di osservazioni censurate.

Quindi, nella censura di tipo I, in assenza di perdite accidentali, tutte le osservazioni censurate sono uguali alla durata del periodo di studio.

Grafico 7. Censura di tipo I

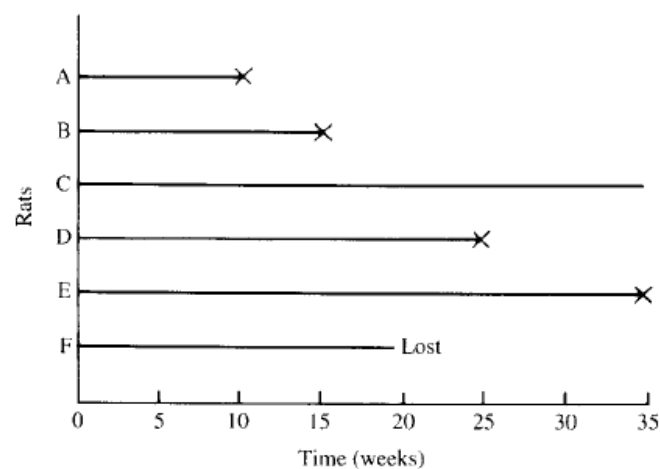


Censura di tipo II

Consideriamo uno studio simile al precedente, ma questa volta, invece di stabilirne la fine entro un intervallo di tempo determinato, si fissa la fine dello studio quando una quantità specifica di topi muore. Ad esempio, quando l'80% dei topi è deceduto ed il restante 20% viene sacrificato.

In questo caso, se non ci sono perdite accidentali, le osservazioni censurate sono equiparate alla più grande osservazione non censurata. Nell'esempio sottostante si termina lo studio dopo la morte di 4 topi su 6. I tempi di sopravvivenza in questo caso sono pari a: 10, 15, 35+; 25, 35, 19+ settimane.

Grafico 8. Censura di tipo II

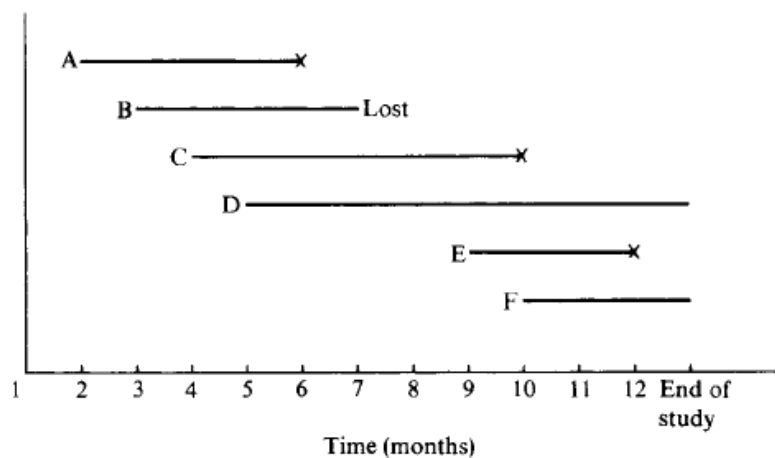


Censura di tipo III

Talvolta, il periodo dello studio è fisso ed i pazienti possono prendervene parte in momenti diversi durante quell'intervallo. Alcuni pazienti potrebbero morire entro la fine dello studio e i loro tempi di sopravvivenza sono noti, altri potrebbero abbandonare lo studio in qualsiasi momento ed altri ancora potrebbero essere ancora

vivi allo scadere del tempo di analisi. Per coloro che hanno lasciato lo studio, il tempo di sopravvivenza è pari almeno al tempo dall'ingresso nello studio fino all'ultimo momento in cui se ne hanno notizie; per coloro che sono ancora in vita, dall'ingresso alla fine dello studio. Questi ultimi due casi sono osservazioni censurate. Poiché i tempi di ingresso non sono simultanei, anche i tempi censurati sono diversi.

Grafico 9. Censura di tipo III



Le osservazioni censurate di tipo I e II sono anche chiamate **Singly censored data**, mentre quelle di tipo III **Progressively censored data**.

Censura a destra, sinistra e ad intervallo

Per quanto riguarda la censura a destra, e in questa casistica rientrano tutti e tre gli esempi descritti in precedenza, possiamo soltanto dire che il tempo di sopravvivenza supera un certo intervallo di tempo ma il valore esatto non è noto. Questo è il caso dei topi ancora in vita al termine dello studio: sappiamo che sopravvivono almeno fino alla fine dello studio, ma non possiamo dire altro.

La censura a sinistra si verifica invece quando si è a conoscenza del fatto che un certo evento di interesse si è verificato prima di un certo intervallo di tempo T , ma non è noto il momento esatto in cui si è verificato.

La censura ad intervallo si verifica quando è noto che l'evento di interesse si verifica in un intervallo compreso tra T_1 e T_2 , ma non si sa esattamente quando.

Nell'ambito di questo studio, per avere un database completo sarebbe stato necessario seguire un certo numero di imprese dal momento della loro costituzione fino al momento del default. Ovviamente un'analisi di questo tipo è impossibile ed è per questa ragione che il database usato in questa analisi è caratterizzato da osservazioni censurate. Infatti, ci sono imprese che entrano a far parte dell'analisi in momenti diversi dalla loro data di costituzione, imprese che al termine dell'analisi non hanno sperimentato il default e imprese le cui informazioni vengono a mancare durante lo studio. Nonostante il problema della censura però, come già accennato in precedenza, queste aziende sono state ugualmente riportate all'interno dello studio perché capaci di fornire indicazioni significative per le conclusioni che si vogliono trarre in questo elaborato.

Survival Time

I “Survival time data” rappresentano il tempo necessario perché si verifichi un determinato evento, come ad esempio il fallimento di un’azienda, la morte di un paziente, lo sviluppo di una certa malattia ecc....Questi tempi sono soggetti a variazioni casuali e, come tutte le variabili casuali, sono caratterizzati da una ben precisa distribuzione. La distribuzione dei tempi di sopravvivenza è rappresentata da tre funzioni:

- The Survivorship Function
- The Probability Density Function
- The Hazard Function

Queste funzioni sono matematicamente equivalenti, ovvero, data una di queste, le altre due possono essere derivate facilmente. Sono utilizzate per illustrare le diverse proprietà e caratteristiche dei dati ed una volta stimate è possibile trarre inferenze sul modello di sopravvivenza della popolazione.

Nel corso della seguente trattazione, si identifica con la lettera T il tempo di sopravvivenza la cui distribuzione è caratterizzata dalle 3 funzioni di seguito riportate.

Survivorship Function (or Survival Function)

Questa funzione, indicata da $S(t)$, rappresenta la probabilità che un individuo sopravviva oltre un certo tempo t :

$$\begin{aligned} S(t) &= P(\text{Probabilità che individuo sopravviva oltre un tempo } t) \\ &= P(T > t) \end{aligned} \quad (1.1)$$

Considerando la funzione di distribuzione cumulativa di T , $F(t)$:

$$S(t) = 1 - P(\text{Probabilità che un individuo muoia prima di } t) \quad (1.2)$$

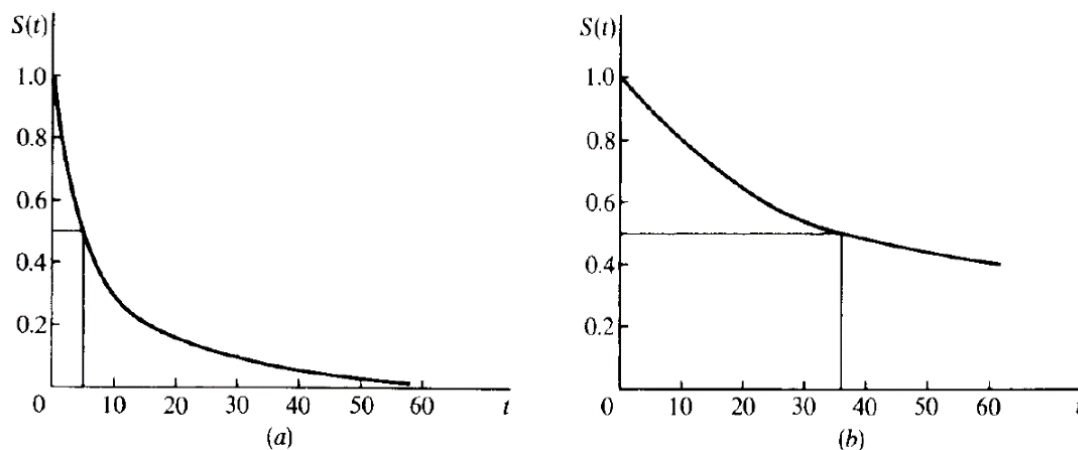
$$= 1 - F(t)$$

In questo caso, $S(t)$ è una funzione non crescente del tempo t con le seguenti proprietà:

$$S(t) = \begin{cases} 1, & t = 0 \\ 0, & t = \infty \end{cases} \quad (1.3)$$

La probabilità di sopravvivere almeno fino al tempo zero è pari ad uno e di sopravvivere all'infinito è pari a zero. La funzione $S(t)$ è conosciuta anche come **Cumulative Survival Rate** e la sua rappresentazione grafica come **Survival Curve**.

Grafico 10. Due esempi di survival curve



Il grafico (a) rappresenta dei bassi tassi di sopravvivenza o dei brevi Survival Time mentre il grafico (b) alti tassi di sopravvivenza o lunghi tempi di sopravvivenza.

Possiamo usare questa funzione per trovare il 50-esimo percentile, la mediana, o altri valori dei tempi di sopravvivenza ed anche per paragonare la distribuzione di diversi gruppi.

Se non ci fossero osservazioni censurate, la Survivorship Function potrebbe essere stimata come:

$$\widehat{S}(t) = \frac{\text{Numero di pazienti che sopravvivono oltre al tempo } t}{\text{Numero totale di pazienti}} \quad (1.4)$$

In presenza di osservazioni censurate non è sempre appropriato usare $S(t)$, in questo caso è opportuno usare dei metodi non parametrici.

Probability Density Function (or Density Function)

Come per ogni variabile casuale continua, il tempo di sopravvivenza T ha una funzione di densità di probabilità definita come la probabilità di fallimento in un piccolo intervallo temporale per unità di tempo. In altri termini, può essere definita come il limite della probabilità che un individuo fallisca nel piccolo intervallo da t a $t+\Delta t$ per unità di tempo Δt :

$$f(t) = \frac{\lim_{\Delta t \rightarrow 0} P[\text{Probabilità che un individuo muoia nell'intervallo}(t, t + \Delta t)]}{\Delta t} \quad (1.5)$$

La funzione di densità, nota anche come **tasso di guasto incondizionato**, ha le seguenti proprietà:

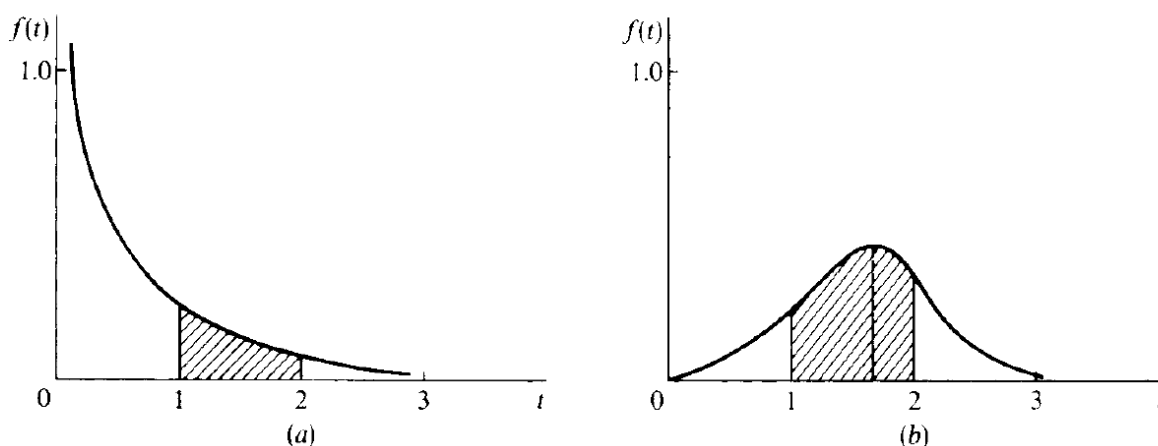
- $f(t)$ è una funzione non negativa

$$\begin{cases} f(t) = 0, & t < 0 \\ f(t) \geq 0, & t \geq 0 \end{cases}$$

- L'area tra la curva di densità e l'asse del tempo t è pari ad 1.

Il grafico di $f(t)$ prende il nome di **density curve**, di seguito ne sono riportati degli esempi.

Grafico 11. Due esempi di curve di densità.



La curva di densità nella figura (a) fornisce un modello di tasso di fallimento elevato all'inizio dello studio e tasso di fallimento decrescente con l'aumentare del tempo. Nella figura (b), il picco di alta frequenza di guasto si verifica a circa 1,7 unità di tempo. La proporzione di individui che falliscono tra la prima e la seconda unità di tempo è uguale all'area ombreggiata tra la curva di densità e l'asse.

In assenza di osservazioni censurate, $f(t)$ può essere stimata come:

$$\widehat{f(t)} = \frac{\text{Numero di pazienti che muoiono nell'intervallo che inizia al tempo } t}{(\text{numero totale di pazienti}) \times (\text{ampiezza dell'intervallo})} \quad (1.6)$$

Dalla funzione di densità è possibile calcolare la proporzione di individui che muoiono in ogni intervallo e il picco di alta frequenza di fallimento.

Hazard Function

La funzione di rischio $h(t)$ del tempo di sopravvivenza T ci permette di calcolare il **tasso di fallimento condizionato**. Possiamo definirla come la probabilità di fallimento in un piccolo intervallo di tempo, assumendo che l'individuo sia sopravvissuto all'inizio dell'intervallo stesso. In altri termini, può essere descritta come il limite della

probabilità che un individuo muoia in un intervallo molto piccolo, $t+\Delta t$, dato che l'individuo è sopravvissuto al tempo t :

$$h(t) = \frac{\lim_{\Delta t \rightarrow 0} P \left[\frac{\text{Probabilità che un individuo muoia nell'intervallo } t + \Delta t}{\text{Condizionato al fatto di essere sopravvissuto al tempo } t} \right]}{\Delta t} \quad (1.7)$$

La funzione di rischio può essere definita in termini di funzione di distribuzione cumulativa $F(t)$ e di funzione di densità di probabilità $f(t)$:

$$h(t) = \frac{f(t)}{1-F(t)} \quad (1.8)$$

$H(t)$ è anche chiamata **Instantaneous failure rate, force of mortality, conditional mortality rate o age-specific failure rate**.

Se la t nella formula (1.7) rappresenta l'età, la funzione di rischio è una misura della propensione al fallimento in funzione dell'età della persona; la quantità $\Delta t h(t)$ è la proporzione attesa di individui di età t che falliranno nel breve intervallo di tempo $t+\Delta t$.

La funzione di rischio restituisce il rischio di fallimento per unità di tempo durante il processo di invecchiamento. Quando non sono presenti dati censurati, la funzione può essere stimata come:

$$\begin{aligned} \widehat{h(t)} &= \frac{\text{Numero di pazienti morti nell'intervallo iniziato in } t}{(\text{numero di pazienti superstiti al tempo } t) \times (\text{ampiezza dell'intervallo})} \\ &= \frac{\text{Numero di pazienti che muoiono per unità di tempo nell'intervallo}}{\text{Numero di pazienti che sopravvivono a } t} \end{aligned} \quad (1.9)$$

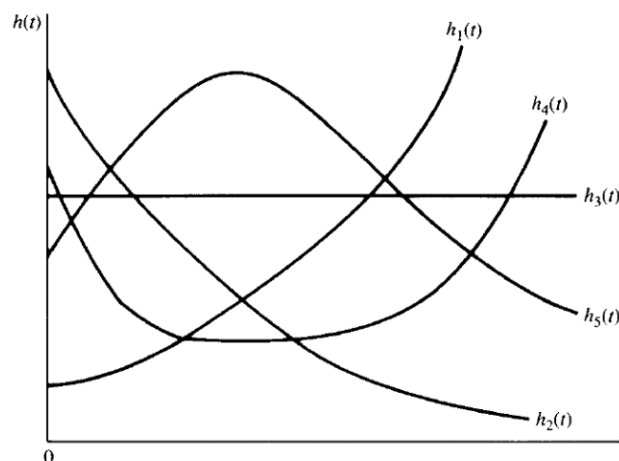
Spesso si preferisce effettuare un calcolo più prudente, che restituisce un tasso di fallimento più elevato, considerando il numero medio di sopravvissuti nel punto medio dell'intervallo, utilizzando la formula:

$$\widehat{h(t)} = \frac{\text{Numero di pazienti che muoiono per unità di tempo nell'intervallo}}{(\text{numero di pazienti che sopravvivono a } t) - \frac{\text{numero di morti nell'intervallo}}{2}} \quad (1.10)$$

La funzione di rischio può crescere, decrescere, rimanere costante o indicare un processo più complicato. Il grafico sottostante rappresenta diversi tipi di funzione di rischio. Ad esempio, pazienti affetti da leucemia acuta che non rispondono al trattamento, hanno un tasso di rischio crescente $h_1(t)$; $h_2(t)$ rappresenta una funzione di rischio decrescente, come nel caso di un soldato ferito da un proiettile che subisce un intervento chirurgico ed inizia un percorso di guarigione. $h_3(t)$, invece, è una funzione costante, può rappresentare il rischio di persone sane di età compresa tra i 18 e i 40 anni, i cui principali rischi di morte sono gli incidenti.

La curva $h_4(t)$, chiamata anche **bathtub curve**, descrive il processo della vita umana. In un periodo di tempo iniziale, il rischio è elevato per via della mortalità infantile, in seguito si mantiene approssimativamente costante per un certo periodo di tempo e alla fine vi è un incremento per via delle malattie o problemi di salute cui una persona va in contro. L'ultima curva, $h_5(t)$, rappresenta il caso di persone con una malattia, che durante la fase di diagnostica vedono il rischio crescere ma che rispondono positivamente alle cure e pertanto il rischio va via via scemando.

Grafico 12. Esempio di funzione di rischio



La funzione di rischio cumulativa è definita come:

$$H(t) = \int_0^t h(x)dx \quad (1.11)$$

Come vedremo in seguito, è verificata la relazione:

$$H(t) = -\log S(t) \quad (1.12)$$

Pertanto, è verificato che:

$$\begin{cases} t = 0, S(t) = 1, H(t) = 0; \\ t = \infty, S(t) = 0, H(t) = \infty; \end{cases}$$

Infatti, la funzione di rischio cumulativa può assumere qualsiasi valore compreso tra zero ed infinito.

Esempio: Stima delle funzioni di sopravvivenza

I dati presenti nelle prime tre colonne della tabella sottostante rappresentano i dati di sopravvivenza di 40 pazienti con il melanoma. I tempi di sopravvivenza sono raggruppati in intervalli di 5 mesi. Le ultime colonne riportano il calcolo della funzione di sopravvivenza, della funzione di densità e della funzione di rischio.

Tabella 6 Esempio di calcolo delle funzioni di sopravvivenza

Survival Time t (months)	Number of Patients Surviving at Beginning of Interval	Number of Patients Dying in Interval	$\hat{S}(t)$	$\hat{f}(t)$	$\hat{h}(t)$
0–5	40	5	1.000	0.025	0.027
5–10	35	7	0.875	0.035	0.044
10–15	28	6	0.700	0.030	0.048
15–20	22	4	0.550	0.020	0.040
20–25	18	5	0.450	0.025	0.065
25–30	13	4	0.325	0.020	0.072
30–35	9	4	0.225	0.020	0.114
35–40	5	0	0.125	0.000	0.000
40–45	5	2	0.125	0.010	0.100
45–50	3	1	0.075	0.005	0.080
≥ 50	2	2	0.050	—	—

Di seguito, sono riportati i rispettivi grafici.

Grafico 13. Stima di $\hat{S}(t)$

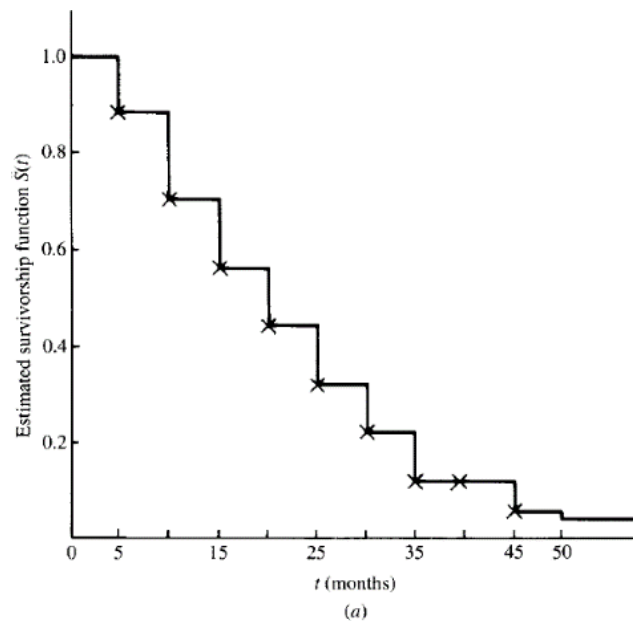


Grafico 14. Stima di $\hat{f}(t)$

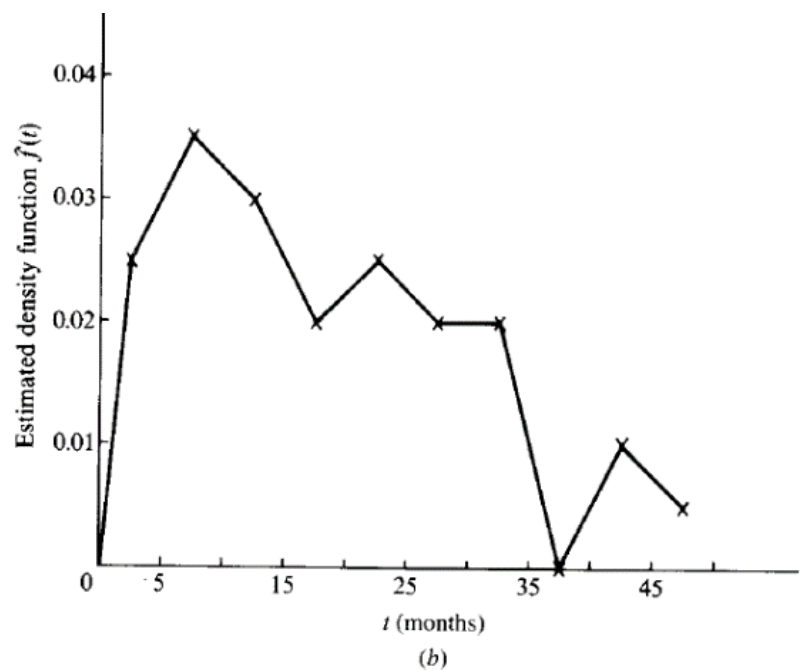
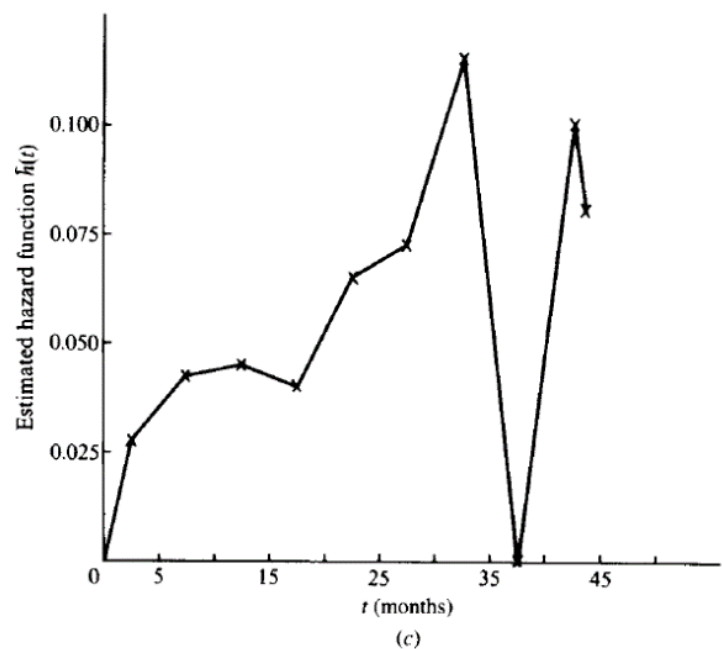


Grafico 15. Stima di $\hat{h}(t)$



La funzione di sopravvivenza stimata, $\widehat{S}(t)$, è calcolata all'inizio o alla fine di ogni intervallo; ad esempio, all'inizio del primo intervallo, quando tutti i pazienti sono ancora in vita, otteniamo $\widehat{S}(0) = 1$. All'inizio del secondo intervallo, invece, solo 35 dei 40 pazienti totali sono ancora in vita e pertanto, $\widehat{S}(5) = \frac{35}{40} = 0.875$. La funzione di densità stimata, $\widehat{f}(t)$, del primo intervallo (0-5), ad esempio, è pari a $\widehat{f}(t) = \frac{5}{40 \times 5} = 0.025$. Tale funzione è rappresentata graficamente nel punto medio di ciascun intervallo. La funzione di rischio stimata, $\widehat{h}(t)$, calcolata, sempre a titolo di esempio, per il primo intervallo è pari a $\frac{5}{\left[5\left(40 - \frac{5}{2}\right)\right]} = 0.027$. Come per la funzione di densità, la funzione di rischio è calcolata nel punto medio di ciascun intervallo.

Come è possibile notare dalla tabella o dal grafico della funzione, il tempo di sopravvivenza mediano dei pazienti affetti da melanoma è pari a circa 17,5 mesi ed il picco di alta frequenza di morte si verifica tra i 5 e i 10 mesi. Inoltre, la funzione di rischio mostra una tendenza crescente e raggiunge il suo picco a circa 32,5 mesi per poi oscillare.

Relazione tra le funzioni di sopravvivenza

Le tre funzioni precedentemente descritte sono matematicamente equivalenti, ovvero stimata la prima le altre due possono essere facilmente derivate. Di seguito sono riportate le principali relazioni:

- 1) Da (1.2) e (1.8) otteniamo:

$$h(t) = \frac{f(t)}{S(t)} \quad (1.13)$$

Questa relazione può essere anche derivata dalla (1.7) usando la nozione di base di probabilità condizionale.

- 2) Poiché la funzione di densità di probabilità è la derivata della funzione di distribuzione cumulativa:

$$f(t) = \frac{d}{dt} [1 - S(t)] = -S'(t) \quad (1.14)$$

- 3) Sostituendo la (1.14) nella (1.13)

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} \log S(t) \quad (1.15)$$

- 4) Integrando (1.15) da 0 a t considerando $S(0) = 1$, otteniamo:

$$-\int_0^t h(x) dx = \log S(t)$$

O

$$H(t) = -\log S(t)$$

O

$$S(t) = \exp[-H(t)] = \exp\left[-\int_0^t h(x) dx\right] \quad (1.16)$$

- 5) A partire dalla (1.13) e dalla (1.16) otteniamo:

$$f(t) = h(t) \exp[-H(t)] \quad (1.17)$$

Quindi, se $f(t)$ è noto, la funzione di sopravvivenza può essere determinata dalla relazione di base tra $f(t)$, $F(t)$, e (1.2). La funzione di rischio può essere calcolata a partire dalla (1.13). Se $S(t)$ è conosciuta, $f(t)$ e $h(t)$ possono essere determinate da (1.14) e (1.13) rispettivamente, o $h(t)$ può essere derivata prima da (1.15) e poi $f(t)$ da (1.13). Se la funzione di rischio è data, $S(t)$ e $f(t)$ è ottenuta da (1.16) e (1.17).

Metodi Non Parametrici per la Stima delle Funzioni di Sopravvivenza

Il metodo per la stima delle funzioni di sopravvivenza descritto nell'esempio precedente non è però esaustivo, non tiene infatti conto della presenza di dati censurati. Infatti, alla fine dello studio alcuni pazienti potrebbero essere ancora in vita e i loro tempi di sopravvivenza ignoti.

Per affrontare questo genere di problema, è possibile servirsi dei metodi non parametrici. Questi metodi, anche detti senza distribuzione, sono semplici da capire e applicare, tuttavia, risultano essere meno efficienti di quelli parametrici nel caso in cui i tempi di sopravvivenza seguono una ben precisa distribuzione teorica. Invece, quando non sono note distribuzioni teoriche adeguate è consigliabile servirsi dei metodi non parametrici. Per questo motivo, nel corso della seguente trattazione, prima di tentare di adattare ai dati di sopravvivenza una distribuzione teorica, si utilizzeranno i metodi senza distribuzione per analizzare il dataset. Le stime ottenute con metodi e grafici non parametrici possono essere utili nella scelta di una distribuzione.

Vedremo, per prima cosa, il metodo del limite di prodotto (PL) per stimare la funzione di sopravvivenza sviluppata da Kaplan e Meier nel 1958; questo metodo è applicabile sia per campioni piccoli, moderati e grandi ma, nell'eventualità che i dati siano già raggruppati in intervalli o la dimensione del campione è molto grande o siamo interessati ad una popolazione molto ampia, potrebbe essere più pratico e utile eseguire un'analisi delle tavole di vita. Infatti, il metodo immediatamente successivo che viene introdotto è quello delle tabelle di popolazione e di vita clinica.

Le stime PL e le stime delle tavole di vita sono essenzialmente le stesse, con la sola differenza che Kaplan e Meier si basano sui tempi di sopravvivenza individuali, mentre le tavole di vita sono costruite considerando degli intervalli di tempo.

La stima PL può essere considerata come un caso speciale della stima delle tavole di vita in cui ogni intervallo contiene una sola osservazione.

Stima Product Limit della Funzione di Sopravvivenza

Se consideriamo il caso semplice in cui ciascun paziente viene osservato fino alla sua morte, i tempi di sopravvivenza sono noti ed esatti. Siano $t_1, t_2, t_3, \dots, t_n$ i tempi di sopravvivenza degli n individui dello studio. Concettualmente, consideriamo questo gruppo di pazienti come un campione casuale di una popolazione più ampia di pazienti simili. Riorganizzando questi tempi in modo tale che

$$t_1 \leq t_2 \leq t_3 \leq \dots \leq t_n$$

la funzione di sopravvivenza calcolata in $t_{(i)}$ può essere stimata utilizzando la (1.2) e la (1.4):

$$\widehat{S(t_{(i)})} = \frac{n-i}{n} = 1 - \frac{i}{n} \quad (1.18)$$

Dove $n - i$ è il numero di persone nel campione che sopravvivono più a lungo di $t_{(i)}$. Nell'eventualità in cui due o più osservazioni risultino essere uguali (**osservazioni legate**), viene utilizzato il valore di (i) più grande. Ad esempio, se

$$t_2 = t_3 = t_4$$

Allora,

$$\widehat{S(t_{(2)})} = \widehat{S(t_{(3)})} = \widehat{S(t_{(4)})} = \frac{n-4}{n}$$

ciò fornisce una stima conservativa per le osservazioni legate.

Poiché ogni persona è viva all'inizio dello studio e nessuno sopravvive più di $t_{(n)}$, vale:

$$- \widehat{S}(t_{(0)}) = 1$$

$$- \widehat{S}(t_{(n)}) = 0$$

$\widehat{S}(t)$ viene calcolato per ciascun tempo di sopravvivenza; non è necessario preoccuparsi degli intervalli tra i distinti tempi di sopravvivenza in cui nessuno muore in quanto la funzione di sopravvivenza rimane costante. La funzione $\widehat{S}(t)$ graficamente è rappresentata da un gradino che parte da 1.0 e decresce in incrementi di $\frac{1}{n}$ (se non ci sono osservazioni legate) fino a 0. A partire dal grafico di $\widehat{S}(t)$ e del tempo è possibile valutare i percentili del tempo di sopravvivenza.

Esempio: Stima della funzione di sopravvivenza con il metodo Kaplan - Meier

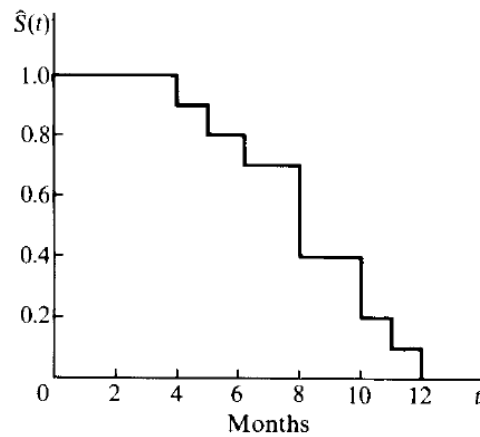
Consideriamo uno studio clinico in cui vengono seguiti fino alla morte 10 pazienti affetti da una malattia terminale e per ognuno di essi il tempo t di sopravvivenza è calcolato in mesi.

Tabella 7. Calcolo di $\widehat{S}(t)$ per i 10 pazienti in esame

t	i	$\widehat{S}(t)$
4	1	$\frac{9}{10} = 0.9$
5	2	$\frac{8}{10} = 0.8$
6	3	$\frac{7}{10} = 0.7$
8	4	$\frac{4}{10} = 0.4$
8	5	$\frac{4}{10} = 0.4$
8	6	$\frac{4}{10} = 0.4$
10	7	$\frac{2}{10} = 0.2$
10	8	$\frac{2}{10} = 0.2$
11	9	$\frac{1}{10} = 0.1$
12	10	$\frac{0}{10} = 0.0$

La funzione $\widehat{S}(t)$ è calcolata seguendo la (1.18) ed è rappresentata da una funzione a gradino, di seguito riportata.

Grafico 16. Stima della funzione $\widehat{S}(t)$



Il tempo di sopravvivenza mediano stimato è di 8 mesi circa, ma una stima più accurata può essere ottenuta usando un'interpolazione lineare:

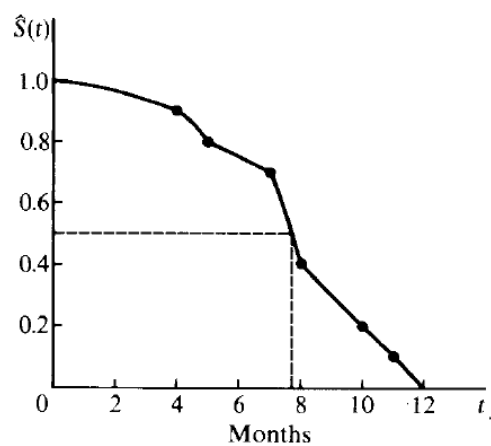
Figura 2. Calcolo interpolazione lineare

t	$\widehat{S}(t)$
6	0.7
m	0.5
8	0.4

$$\frac{8 - 6}{0.4 - 0.7} = \frac{8 - m}{0.4 - 0.5}$$

$$m = 8 - \frac{2(0.1)}{0.3} = 7.3(\text{months})$$

Grafico 17. Stima della funzione $\widehat{S}(t)$



Una funzione a gradini è migliore per mostrare quando la funzione di sopravvivenza rimane invariata nel tempo, mentre invece una funzione “liscia” consente di stimare con maggiore precisione la mediana.

Il metodo appena visto può essere utilizzato solo se tutti i pazienti sono seguiti fino alla morte, ossia non ci sono osservazioni censurate.

Se alcuni dei pazienti sono ancora in vita alla fine dello studio, è richiesto un diverso metodo di stima per $\widehat{S}(t)$, come ad esempio la stima PL fornita da Kaplan e Meier.

Stima PL di Kaplan-Meier

Ipotizzando che 10 persone prendano parte ad uno studio clinico a gennaio del 2000 e che durante l’anno 6 di essi perdano la vita a causa di una patologia. A dicembre altre 20 persone di uniscono allo studio. Nel 2001, tre pazienti che sono entrati all’inizio del 2000 e 15 pazienti che sono entrati in seguito muoiono, lasciando rispettivamente 1 e 5 sopravvissuti. Lo studio termina alla fine del 2001 e la stima di Kaplan Meier ci permette di calcolare la percentuale di pazienti che sopravvive per due anni o più.

La particolarità di uno studio simile consiste nel fatto che il primo gruppo viene seguito per due anni, mentre il secondo per uno soltanto.

Una prima stima, più semplicistica, consiste nel considerare un campione ridotto che ignora i 20 pazienti che sono entrati nello studio l’anno seguente, quindi la probabilità di sopravvivenza per due anni è data da:

$$\widehat{S}(2) = \frac{1}{10} = 0.1$$

Tuttavia, Kaplan e Meier ritengono che anche il secondo campione, pur essendo seguito per un anno soltanto, possa contribuire alla stima della funzione di sopravvivenza. Nello specifico, i pazienti che sono sopravvissuti a due anni possono essere considerati come sopravvissuti il primo anno e poi sopravvissuti per un altro

anno; pertanto, la probabilità di sopravvivere per due anni o più è uguale alla probabilità di sopravvivere il primo anno e poi sopravvivere un anno in più:

$$\widehat{S(2)} = P(\text{prob. di sopravvivere due anni dato che il paziente è sopravvissuto il primo anno}) \times P(\text{prob. di sopravvivere il primo anno}) \quad (1.18)$$

La stima di Kaplan-Meier è data da:

$$\widehat{S(2)} = \left(\text{Proporzione di pazienti che sopravvivono due anni dato che sono sopravvissuti per un anno} \right) \times (\text{proporzione di pazienti che sopravvivono un anno}) \quad (1.19)$$

Nel caso specifico, uno dei quattro pazienti sopravvissuti al primo anno è ancora vivo dopo due anni, quindi la prima proporzione della (1.19) è pari a $\frac{1}{4}$. Quattro su dieci pazienti entrati al primo anno dell'analisi e cinque su venti entrati alla fine del 2000 sono sopravvissuti un anno, quindi, la seconda proporzione è pari a $\frac{4+5}{10+20}$.

La stima PL è pari a:

$$\widehat{S(2)} = \frac{1}{4} \times \frac{4+5}{10+20} = 0.25 \times 0.3 = 0.075$$

Questa semplice regola può essere generalizzata come segue:

La probabilità di sopravvivere k (≥ 2) o più anni dall'inizio dello studio è il prodotto di k tassi di sopravvivenza osservati:

$$\widehat{S(k)} = p_1 \times p_2 \times p_3 \times \dots \times p_k \quad (1.20)$$

dove p_1 denota la proporzione di pazienti che sopravvivono almeno un anno, p_2 la proporzione di pazienti che sopravvive il secondo anno dopo che sono sopravvissuti un anno, p_3 la proporzione di pazienti che sopravvive il terzo anno dopo essere

sopravvissuti per due anni e p_k la proporzione di pazienti che sopravvive il k -esimo anno dopo essere sopravvissuti per $k - 1$ anni.

In pratica, la stima PL della probabilità di sopravvivenza per un determinato numero di anni dall'inizio dello studio è il prodotto della stessa valutata l'anno precedente a quello di interesse e il tasso di sopravvivenza osservato per il particolare anno, ovvero:

$$\widehat{S}(t) = \widehat{S}(t - 1) \times p_t \quad (1.21)$$

Le stime PL sono stime di **massima verosimiglianza** che nella pratica possono essere calcolate costruendo una tabella con 5 colonne seguendo lo schema di seguito riportato:

- 1) La **colonna 1** contiene tutti i tempi di sopravvivenza, sia censurati che no, in ordine dal più piccolo al più grande. Si inserisce un segno “+” in prossimità di osservazioni censurate. Se un’osservazione censurata ha lo stesso valore di una osservazione non censurata, quest’ultima viene riportata per prima.
- 2) La **colonna 2**, etichettata “ i ”, è formata dal rango corrispondente di ciascuna osservazione nella colonna 1.
- 3) La **colonna 3**, identificata da “ r ”, riguarda solo le osservazioni non censurate, tale per cui $r = i$.
- 4) La **colonna 4** riporta il calcolo $\frac{(n-r)}{(n-r+1)}$, o p_i , per ogni osservazione non censurata t_i in modo da fornire la proporzione di pazienti che sopravvive almeno fino a quel tempo.

5) Nella **colonna 5**, $\widehat{S}(t)$ è il prodotto di tutti i valori precedentemente calcolati fino e compreso il tempo di interesse. Se alcune osservazioni non censurate sono legate tra di loro, ossia assumono il medesimo valore, si dovrebbe usare il valore della funzione di sopravvivenza minore.

Riassumendo, sia n il numero totale di pazienti i cui tempi di sopravvivenza, sia che si tratta di dati censurati o meno, sono organizzati in ordine crescente, la funzione di sopravvivenza è data da:

$$\widehat{S}(t) = \prod_{t_{(r)} \leq t} \frac{n-r}{n-r+1} \quad (1.22)$$

Dove r è un numero intero positivo per cui $t_{(r)} \leq t$ e $t_{(r)}$ non è un'informazione censurata. Si tratta di numeri interi consecutivi solamente se non ci sono osservazioni censurate.

Il tempo di sopravvivenza mediano stimato è il 50-esimo percentile, ossia il valore di t a $\widehat{S}(t) = 0.50$.

Di seguito è riportato un esempio numerico del calcolo appena descritto.

Consideriamo un campione di 10 pazienti i cui tempi di sopravvivenza sono descritti nella tabella 8.

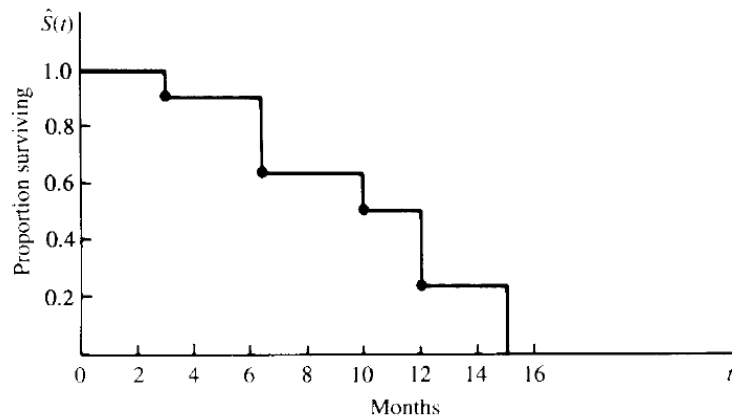
Tabella 8. Calcolo della stima PL della funzione di sopravvivenza.

Table 4.2 Calculation of the PL Estimate of $\hat{S}(t)$ for Data in Example 4.2

Remission Time t	Rank i	r	$(n - r)/(n - r + 1)$	$\hat{S}(t)$
3.0	1	1	$\frac{9}{10}$	$\frac{9}{10} = 0.900$
4.0+	2	—	—	—
5.7+	3	—	—	—
6.5	4	4	$\frac{6}{7}$	$\frac{9}{10} \times \frac{6}{7} = 0.771^a$
6.5	5	5	$\frac{5}{6}$	$\frac{9}{10} \times \frac{6}{7} \times \frac{5}{6} = 0.643^a$
8.4+	6	—	—	—
10.0	7	7	$\frac{3}{4}$	$\frac{9}{10} \times \frac{6}{7} \times \frac{5}{6} \times \frac{3}{4} = 0.482$
10.0+	8	—	—	—
12.0	9	9	$\frac{1}{2}$	$\frac{9}{10} \times \frac{6}{7} \times \frac{5}{6} \times \frac{3}{4} \times \frac{1}{2} = 0.241$
15.0	10	10	0	0

La funzione di sopravvivenza graficamente è descritta dalla seguente funzione a gradini:

Grafico 18. Funzione di sopravvivenza relativa all'esempio precedente.



Il tempo di remissione mediano stimato è di circa 9.8 mesi. Seguendo la (1.21) possiamo riscrivere:

$$\widehat{S}(t_{(i)}) = \widehat{S}(t_{(i-1)}) \times \frac{n-i}{n-i+1} \quad (1.23)$$

Tenendo conto che il calcolo è valido nel caso di osservazioni non censurate. Ad esempio,

$$\widehat{S}(12) = \widehat{S}(10) \times \frac{10-9}{10-9+1} = 0.241$$

In assenza di osservazioni censurate la (1.22) è equivalente alla (1.18).

La varianza della stima PL di $\widehat{S}(t)$ è approssimata da:

$$Var[\widehat{S}(t)] \simeq [\widehat{S}(t)]^2 \sum_r \frac{1}{(n-r)(n-r+1)} \quad (1.24)$$

dove r include quegli interi positivi per cui $t_{(r)} \leq t$.

Riprendendo il calcolo precedente:

$$Var[\widehat{S}(10)] \simeq [0.482]^2 \left(\frac{1}{9 \times 10} + \frac{1}{6 \times 7} + \frac{1}{5 \times 6} + \frac{1}{3 \times 4} \right) = 0.0352$$

e lo standard error stimato è pari a 0.1876. La varianza è molto utile per il calcolo di intervalli di confidenza per la funzione di sopravvivenza.

Infine, è possibile dimostrare che il **tempo di sopravvivenza medio** è pari all'area sottostante la funzione di sopravvivenza stimata:

$$\hat{\mu} = \int_0^{\infty} \widehat{S}(t) dt \quad (1.25)$$

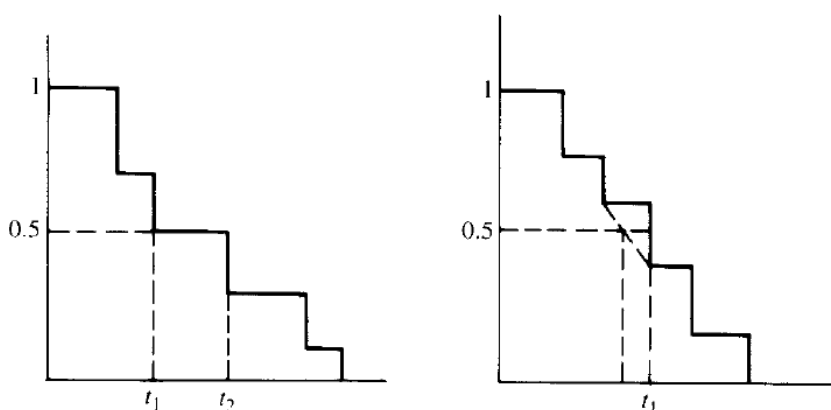
Quindi, se i tempi di sopravvivenza sono ordinati in modo crescente e se non sono presenti osservazioni censurate, il tempo medio è pari alla somma delle aree dei rettangoli sottostanti alla curva di sopravvivenza. Nel caso in cui, invece, l'osservazione più grande a disposizione è censurata e la si usa nel calcolo, la stima di μ rischia di essere più bassa.

Criticità del metodo di Kaplan-Meier

Il metodo di Kaplan-Meier fornisce stime molto utili della probabilità di sopravvivenza e delle rappresentazioni grafiche di facile interpretazione. Breslow e Crowley (1974) hanno dimostrato che in determinate condizioni, la stima è **consistente** e **asintoticamente normale**. Tuttavia, è possibile individuare delle caratteristiche critiche:

- 1) Le stime di Kaplan – Meier sono limitate all'intervallo di tempo in cui cadono le osservazioni. Se l'osservazione più grande non è censurata, la stima PL in quel preciso istante è pari a zero, la stima è corretta in quanto nessun soggetto in esame vive più a lungo. Ma se l'osservazione più grande che si ha a disposizione è censurata, la stima non potrà mai essere pari a zero ed è indefinita per un tempo che va oltre l'ultimo dato a disposizione.
- 2) La statistica riassuntiva maggiormente utilizzata è il tempo di sopravvivenza mediano. In maniera molto veloce, questo tempo può essere dedotto dalle curve di sopravvivenza stimate come il tempo t tale per cui $\widehat{S}(t) = 0.5$. L'aspetto critico di questa deduzione è che la soluzione potrebbe non essere univoca. Considerando l'esempio sottostante:

Figura 3. Stima della mediana con il metodo PL



In corrispondenza di $\widehat{S}(t) = 0.5$ la curva di sopravvivenza è orizzontale; qualsiasi valore compreso tra t_1 e t_2 è una ragionevole stima della mediana. La soluzione

più pratica consiste nel considerare come valore della mediana il punto medio dell'intervallo come mostrato in figura.

- 3) Se meno del 50% delle osservazioni non è censurata e l'osservazione più grande è censurata, il tempo di sopravvivenza mediano non può essere stimato.
- 4) Il metodo PL presuppone che i tempi censurati siano indipendenti dai tempi di sopravvivenza, ossia che il motivo per cui un'osservazione è censurata non è correlata al verificarsi dell'evento oggetto dello studio. Questa ipotesi può essere vera, ad esempio, se il paziente è ancora in vita al termine del periodo di studio. Risulta invece falsa in casi in cui il paziente è costretto ad abbandonare lo studio prima della sua morte o se è morto per cause diverse da quelle prese in esame. Quando è presente quest'ultimo tipo di censura, il metodo PL non è idoneo all'uso che se ne vuole fare.

Analisi delle tavole di vita

Il metodo delle tavole di vita è una delle più antiche tecniche per misurare la mortalità e descrivere l'esperienza di sopravvivenza di una popolazione. Quando queste tavole descrivono l'esperienza della mortalità di una specifica popolazione per uno specifico intervallo di tempo prendono il nome di **tabelle di vita della popolazione**. Quando invece questo metodo è applicato a pazienti con una determinata malattia, seguiti per un ben preciso intervallo di tempo, si parla di **tabelle di vita cliniche**. Questa differenziazione è necessaria perché, anche se da un punto di vista matematico sono equivalenti, le fonti dei dati richiesti sono diverse.

Tabelle di vita della popolazione

Esistono due tipi di tavole di vita della popolazione:

- Tavole di vita di coorte
- Tavole di vita correnti

Le **tavole di vita di coorte** descrivono l'esperienza di sopravvivenza o mortalità dalla nascita alla morte di una specifica coorte di persone nata all'incirca nello stesso periodo (ad esempio le persone nate nel 1950). Questo campione di persone viene seguito dalla nascita fino a quando non muoiono tutti coloro che sono sotto osservazione. La proporzione di morti (sopravvissuti) viene utilizzata per costruire le tabelle di vita per gli anni successivi. Tuttavia, questo tipo di tabella, che è particolarmente utile nella proiezione della popolazione futura e negli studi prospettici, non è utilizzata poiché richiede un lunghissimo periodo di follow-up.

La **tavola di vita corrente** è costruita applicando i tassi di mortalità specifici per età di una popolazione in un dato periodo di tempo ad un campione ipotetico di 100.000 o 1.000.000 di persone. Il punto di partenza è la nascita in un ben preciso anno. Sono necessarie principalmente due fonti di dati:

- 1) Dati del censimento sul numero di persone in vita a ciascuna età per un ben specifico anno (dati presi a metà anno)
- 2) Statistiche vitali sul numero di morti nell'anno stabilito per ciascuna età.

Ad esempio, consideriamo una tabella di vita che comprende una coorte ipotetica di 100.000 americani soggetti ai tassi di mortalità specifici per età sulla base dei dati osservati per gli Stati Uniti nel censimento del 1990. La tavola corrente di vita, basata sull'esperienza di vita di una popolazione effettiva in un breve periodo di tempo, fornisce un buon riepilogo della mortalità attuale.

Una delle statistiche più spesso riportate è l'**aspettativa di vita**. Inoltre, spesso è utile costruire delle tavole di vita considerando intervalli di durata superiore ad un anno (ad esempio considerando intervalli di 5 anni), ottenendo così delle **tavole di vita abbreviate**.

Le tavole correnti di vita, solitamente, sono costituite dalle seguenti informazioni:

- 1) **Intervallo di età. $[x ; x + t]$** Questo è l'intervallo tra due date esatte x e $x + t$, dove t rappresenta la lunghezza dell'intervallo di tempo preso in esame. Ad esempio, se considerassimo l'intervallo **$[20 ; 21[$** consideriamo il periodo di tempo che va dal ventesimo compleanno fino al ventunesimo compleanno escluso.
- 2) **Proporzione di persone vive all'inizio dell'intervallo d'età ma che muoiono durante l'intervallo (tq_x)**. Le informazioni possono essere reperite a partire dai dati del censimento decennale della popolazione e arricchite dal numero dei decessi che si verificano nell'intervallo di tempo specificato. Ad esempio, (tq_x) per l'intervallo d'età 20-21 è la proporzione di persone che sono morte durante e dopo il loro ventesimo compleanno, ma prima del ventunesimo. È una stima della probabilità condizionata di morire nell'intervallo, dato che la persona in esame è ancora viva all'inizio dell'intervallo stesso. A titolo di esempio, i tassi di mortalità riportati nella tabella 9, sono calcolati a partire dai dati del censimento della popolazione del 1990 e dei decessi avvenuti negli Stati Uniti nel triennio 1989-1991.

Tabella 9. Tabella di vita per la popolazione totale, Stati Uniti, 1989-1991

Age Interval	Proportion Dying	Of 100,000 Born Alive		Stationary Population		Average Remaining Lifetime
	Proportion of Persons Alive at Beginning of Age Interval	Number Living at Beginning of Age Interval	Number Dying During Age Interval	In the Age Interval	In This and All Subsequent Age Intervals	Average Number of Years of Life Remaining at Beginning of Age Interval
Period of Life Between Two Ages	Dying During Interval					
x to $x+t$	${}_tq_x$	l_x	${}_td_x$	${}_tL_x$	T_x	e_0
Days						
0-1	0.00351	100,000	351	274	7,536,614	75.37
1-7	0.00135	99,649	134	1,637	7,536,340	75.63
7-28	0.00104	99,515	104	5,722	7,534,703	75.71
28-365	0.00349	99,411	347	91,625	7,528,981	75.74
Years						
0-1	0.00936	100,000	936	99,258	7,536,614	75.37
1-2	0.00073	99,064	72	99,028	7,437,356	75.08
2-3	0.00048	98,992	48	98,968	7,338,328	74.13
3-4	0.00037	98,944	37	98,926	7,239,360	73.17
4-5	0.00030	98,907	30	98,892	7,140,434	72.19
5-6	0.00027	98,877	27	98,863	7,041,542	71.22
6-7	0.00025	98,850	24	98,839	6,942,679	70.23
7-8	0.00023	98,826	23	98,814	6,843,840	69.25
8-9	0.00020	98,803	20	98,794	6,745,026	68.27
9-10	0.00018	98,783	17	98,774	6,646,232	67.28

3) Numero di persone vive all'inizio dell'intervallo di età (l_x). Il valore iniziale di l_x , la dimensione della popolazione ipotetica, di solito è circa 100.000 o 1.000.000 di persone. I valori successivi vengono calcolati, per ogni intervallo, seguendo la formula:

$$l_x = l_{x-1}(1 - tq_{x-t}) \quad (1.26)$$

dove $1 - tq_{x-t}$ è la proporzione di persone che sono sopravvissute al precedente intervallo d'età.

4) Numero di morti durante l'intervallo d'età (td_x) calcolati come:

$$(td_x) = l_x(tq_x) = l_x - l_{x+1} \quad (1.27)$$

5) Popolazione stazionaria (tL_x) e T_x . In questo caso, tL_x è il numero totale di anni vissuti nell'i-esimo intervallo d'età o il numero di anni persona che l_x persone,

di età x , vivono nell'intervallo. Per coloro che sopravvivono all'intervallo in considerazione, il loro contributo a tL_x è pari alla lunghezza dell'intervallo stesso t ; per coloro che muoiono durante l'intervallo il tempo di sopravvivenza deve essere stimato se non è noto il momento esatto in cui una persona muore. L'ipotesi convenzionale è che vivano la metà dell'intervallo e che contribuiscano per $t/2$ al calcolo di tL_x come di seguito riportato:

$$tL_x = t(l_{x+1} + \frac{1}{2}td_x) \quad (1.28)$$

Il simbolo T_x rappresenta il numero totale di anni-persona vissuti oltre l'età t da persone in vita a quell'età, cioè:

$$T_x = \sum_{j \geq x} tL_j \quad (1.29)$$

e

$$T_x = tL_x + T_{x+t} \quad (1.30)$$

6) **Aspettativa di vita o numero medio di anni di vita rimanenti all'inizio dell'intervallo d'età ($\dot{e}_i$).** La vita media rimanente è calcolata come:

$$\dot{e}_x = \frac{T_x}{l_x} \quad (1.31)$$

L'aspettativa di vita di una persona di età x è pari a $x + \dot{e}_x$.

$\dot{e}_x$, a $x=0$, rappresenta l'aspettativa di vita alla nascita. È un indicatore generale della capacità di prolungare la vita, viene usata per identificare le tendenze e confrontare la longevità della popolazione rispetto agli anni passati.

Le tabelle di vita della popolazione possono essere costruite per vari sottogruppi, ad esempio differenziando per sesso, razza, causa di morte ecc....

Tavole di vita cliniche

Il metodo delle tabelle di vita richiede un numero elevato di osservazioni, in modo che i tempi di sopravvivenza possano essere raggruppati per intervallo.

Come per la stima PL, il metodo delle tavole di vita incorpora tutte le informazioni sulla sopravvivenza accumulate fino al termine dello studio, ad esempio informazioni sui pazienti che hanno preso parte allo studio anche in anni successivi all'inizio dello studio stesso. Questa tecnica utilizza anche i dati incompleti come le perdite al follow-up e persone che si sono ritirate.

Le tabelle sono costituite come riportato nella Tabella 10.

Tabella 10. Formato di una tavola di vita clinica

Interval	Midpoint	Width	Number Lost to Follow-up	Number Withdrawn Alive	Number Dying	Number Entering Interval	Number Exposed to Risk	Conditional Proportion Dying	Conditional Proportion Surviving	Cumulative Proportion Surviving	Probability Density $f(t_{mi})$	Hazard $\hat{h}(t_{mi})$
$t - t_2$	t_{m1}	b_1	l_1	w_1	d_1	n'_1	n_1	$\hat{q}_1$	$\hat{p}_1$	$\hat{S}(t_1) = 1.00$	$\hat{f}(t_{m1})$	$\hat{h}(t_{m1})$
$t_2 - t_3$	t_{m2}	b_2	l_2	w_2	d_2	n'_2	n_2	$\hat{q}_2$	$\hat{p}_2$	$\hat{S}(t_2)$	$\hat{f}(t_{m2})$	$\hat{h}(t_{m2})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$t_i = t_{i+1}$	t_{mi}	b_i	l_i	w_i	d_i	n'_i	n_i	$\hat{q}_i$	$\hat{p}_i$	$\hat{S}(t_i)$	$\hat{f}(t_{mi})$	$\hat{h}(t_{mi})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$t_{s-1} - t_s$	$t_{m,s-1}$	b_{s-1}	l_{s-1}	w_{s-1}	d_{s-1}	n'_{s-1}	n_{s-1}	$\hat{q}_{s-1}$	$\hat{p}_{s-1}$	$\hat{S}(t_{s-1})$	$\hat{f}(t_{m,s-1})$	$\hat{h}(t_{m,s-1})$
$t_s - \infty$	—	—	l_s	w_s	d_s	n'_s	n_s	1	0	$\hat{S}(t_s)$	—	—

- 1) Intervallo $[t_i ; t_{i+1}[$.** È l'intervallo in cui sono distribuiti i tempi di sopravvivenza, di ritiro e di perdita. L'intervallo va da t_i fino a t_{i+1} escluso. Si assume che questi intervalli siano fissi, l'ultimo intervallo ha una lunghezza infinita.
- 2) Punto Medio (t_{mi}).** Il punto medio di ogni intervallo viene calcolato poiché le funzioni di densità e di rischio sono tracciate in quel punto, tranne che per l'ultimo intervallo.
- 3) Ampiezza b_i .** L'ampiezza di ogni intervallo, $t_{i+1} - t_i$, con $i = 1, \dots, s - 1$, è necessaria per il calcolo delle funzioni di rischio e di densità. L'ampiezza dell'ultimo intervallo in linea teorica è infinita, pertanto non è possibile ottenere una stima del rischio o della funzione di densità per questo intervallo.

- 4) **Numero di persone perse al follow-up (l_i).** Questo è il numero di persone che vengono perse dall'osservazione e il cui stato di sopravvivenza è quindi sconosciuto nell' i -esimo intervallo ($i = 1, \dots, s$).
- 5) **Numero dei ritirati vivi (w_i).** Le persone che si ritirano dallo studio mentre sono ancora in vita nell' i -esimo intervallo, sono quelle per le quali è noto che siano vive alla fine dello studio. Il tempo di sopravvivenza registrato per queste persone è il periodo di tempo che va dall'ingresso alla data di chiusura dello studio.
- 6) **Numero dei morti nell' i -esimo intervallo (d_i).** Il tempo di sopravvivenza di queste persone è pari al tempo che va dall'ingresso nello studio fino alla loro morte.
- 7) **Numero di persone che entrano nell' i -esimo intervallo (n'_i).** Il numero di persone che entrano nel primo intervallo è pari alla dimensione totale del campione. Le altre entrate sono determinate da:

$$n'_i = (n'_{i-1}) - l_{i-1} - w_{i-1} - d_{i-1} \quad (1.32)$$

- 8) **Numero di persone esposte al rischio (n_i).** Il numero di persone che sono esposte al rischio nell' i -esimo intervallo definito come:

$$n_i = n'_i - \frac{1}{2} (l_i + w_i) \quad (1.33)$$

Si assume che i tempi di perdita o di ritiro siano distribuiti approssimativamente in modo uniforme nell'intervallo, pertanto, le persone perse o ritirate nell'intervallo sono esposte al rischio di morte per metà dell'intervallo. In assenza di perdite e/o di ritiri vale $n_i = n'_i$.

- 9) **Proporzione condizionale delle persone che muoiono ($\hat{q}_i$).** Si tratta di una stima della probabilità condizionale di morire nell' i -esimo intervallo data l'esposizione al rischio nello stesso intervallo, ed è calcolata come:

$$q_i = \frac{d_i}{n_i} \quad (1.34)$$

Tale per cui:

$$i = 1, \dots, s - 1 \text{ e } \widehat{q}_s = 1$$

10) Proporzione condizionale dei sopravvissuti ($\widehat{p}_i$). È una stima della probabilità condizionale di sopravvivere nell' i -esimo intervallo:

$$\widehat{p}_i = 1 - \widehat{q}_i \quad (1.35)$$

11) Proporzione cumulativa dei sopravvissuti $[\widehat{S}(t_i)]$. Consiste nella stima della funzione di sopravvivenza al tempo t_i , spesso indicato come **tasso di sopravvivenza cumulativo**, e si verifica che:

$$\begin{aligned} \text{Per } i = 1, \quad \widehat{S}(t_1) &= 1 \\ \text{Per } i = 2, \dots, s, \quad \widehat{S}(t_i) &= \widehat{p}_{i-1} \widehat{S}(t_{i-1}) \end{aligned}$$

12) Funzione di densità di probabilità stimata $[\widehat{f}(t_m)]$. Definita come la probabilità di morire nell' i -esimo intervallo per unità di tempo; la sua stima nel punto medio dell'intervallo è pari a:

$$\widehat{f}(t_m) = \frac{\widehat{S}(t_i) - \widehat{S}(t_{i-1})}{b_i} = \frac{\widehat{S}(t_i) \widehat{q}_i}{b_i} \quad \text{per } i = 1, \dots, s - 1 \quad (1.36)$$

13) Funzione di rischio $[\widehat{h}(t_{mi})]$. La funzione di rischio, stimata nel punto medio dell' i -esimo intervallo, è pari a:

$$\widehat{h}(t_{mi}) = \frac{d_i}{b_i(n_i - \frac{1}{2}d_i)} = \frac{2q_i}{b_i(1 + p_i)} \quad \text{per } i = 1, \dots, s - 1 \quad (1.37)$$

ossia, il numero medio di morti per unità di tempo nell'intervallo diviso il numero medio di sopravvissuti a metà dell'intervallo.

Un'altra possibile stima della funzione di rischio può essere ottenuta assumendo che il rischio sia costante all'interno di un intervallo, ma che vari tra gli intervalli, ed è calcolata come:

$$\hat{h}(t_{mi}) = -\frac{\log(p_i)}{b_i} \quad (1.38)$$

Le varianze, approssimate su ampio campione, delle funzioni di sopravvivenza stimate nell' i -esimo intervallo, sono date da:

$$Var [\hat{S}(t_i)] \simeq [\hat{S}(t_i)]^2 \sum_{j=1}^{i-1} \frac{\hat{q}_j}{\hat{p}_j n_j} \quad (1.39)$$

$$Var [\hat{f}(t_{mi})] \simeq \frac{[\hat{S}(t_i)q_i]^2}{b_i} \left(\sum_{j=1}^{i-1} \frac{\hat{q}_j}{\hat{p}_j n_j} + \frac{\hat{p}_j}{\hat{q}_j n_j} \right) \quad (1.40)$$

e

$$Var [\hat{h}(t_{mi})] \simeq [\hat{h}(t_{mi})]^2 \frac{1}{\hat{q}_i n_i} \left\{ 1 - \left[\frac{1}{2} \hat{h}(t_{mi}) b_i \right]^2 \right\} \quad (1.41)$$

Il calcolo della varianza può servire per ottenere degli intervalli di confidenza approssimati per le tre funzioni di sopravvivenza.

Infine, il grafico di $\hat{S}(t_i)$ può essere usato per trovare una stima della mediana. Un altro modo può essere quello di considerare l'intervallo (t_j, t_{j+1}) in modo tale $\hat{S}(t_j) \geq 0.5$ e $\hat{S}(t_{j+1}) < 0.5$, il tempo di sopravvivenza mediano t_m può essere stimato attraverso l'interpolazione lineare:

$$\hat{t}_m = t_j + \frac{[\hat{S}(t_j - 0.5)]b_j}{[\hat{S}(t_j) - \hat{S}(t_{j+1})]} = t_j + \frac{[\hat{S}(t_j - 0.5)]}{[\hat{f}(t_{mj})]} \quad (1.42)$$

Distribuzioni Parametriche di Sopravvivenza

Le cause fisiche che possono portare alla morte (o al fallimento) di una persona in un determinato momento possono essere diverse. È praticamente impossibile isolare le cause fisiche e spiegarle matematicamente, quindi si sceglie una distribuzione teorica per approssimare i dati di sopravvivenza. Tuttavia, si tratta di un'operazione complicata; sono di seguito descritte le principali distribuzioni per descrivere il tempo di sopravvivenza.

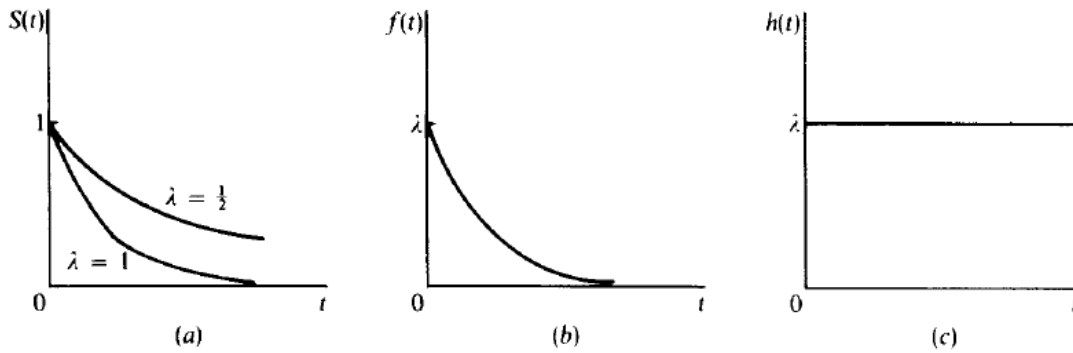
Distribuzione Esponenziale

Considerata la più semplice e la più importante distribuzione negli studi di sopravvivenza, l'esponenziale ha un'ampia applicazione negli studi circa la vita dei dispositivi elettrici. La distribuzione esponenziale assume in questo ambito lo stesso ruolo che la distribuzione normale ha nelle altre aree della statistica. È spesso definita come un modello di guasto puramente casuale ed è particolarmente nota per la sua proprietà di "assenza di memoria"; questo si traduce nel fatto che l'età del soggetto in esame non influenza la futura sopravvivenza. Sebbene molti dati di sopravvivenza non siano adeguatamente descritti dalla distribuzione esponenziale, il suo utilizzo rende più semplice la comprensione del contesto che si sta analizzando.

La distribuzione esponenziale è caratterizzata da un tasso di rischio costante λ , che è anche l'unico parametro che la caratterizza. Un alto valore di λ indica un rischio elevato cui è associata una sopravvivenza breve, un valore basso, invece, indica un basso rischio e di conseguenza una lunga sopravvivenza.

La figura 4, di seguito riportata, mostra la funzione di sopravvivenza, la funzione di densità e la funzione di rischio della distribuzione esponenziale con parametro λ .

Figura 4. Distribuzione esponenziale (a) funzione di sopravvivenza; (b) funzione di densità di probabilità; (c) funzione di rischio



Quando $\lambda=1$, la distribuzione viene chiamata anche **distribuzione esponenziale unitaria**.

Quindi, quando il tempo di sopravvivenza T segue una distribuzione esponenziale con parametro λ , la funzione di densità di probabilità è definita come:

$$f(x) = \begin{cases} \lambda e^{-\lambda t}, & t \geq 0, \quad \lambda > 0 \\ 0, & t < 0 \end{cases} \quad (1.43)$$

La funzione di distribuzione cumulativa è:

$$F(t) = 1 - e^{-\lambda t} \quad t \geq 0 \quad (1.44)$$

La funzione di sopravvivenza:

$$S(t) = e^{-\lambda t} \quad t \geq 0 \quad (1.45)$$

Pertanto, la funzione di rischio è data da:

$$h(t) = \lambda \quad t \geq 0 \quad (1.46)$$

ovvero una costante, indipendente dal tempo. Proprio per il tasso di rischio costante non ci sono fenomeni di invecchiamento o logoramento e il fallimento o la morte sono un evento casuale indipendente dal tempo.

Quando consideriamo i logaritmi naturali della funzione di sopravvivenza, $\log S(t) = -\lambda t$, otteniamo una funzione lineare di t . Pertanto, è facile determinare se i dati

possono essere descritti da una distribuzione esponenziale tracciando $\log \hat{S}(t)$ contro t . Un andamento lineare indica che i dati seguono una distribuzione esponenziale e la pendenza della retta è una stima del tasso di rischio.

La media e la varianza della distribuzione esponenziale con parametro λ sono, rispettivamente, $1/\lambda$ e $1/\lambda^2$. La mediana è data da $\frac{1}{\lambda} \log 2$; il coefficiente di variazione è pari ad uno.

Distribuzione di Weibull

La distribuzione di Weibull è una generalizzazione della distribuzione esponenziale, ma, a differenza di quest'ultima, non assume un tasso di rischio costante e quindi ha un campo di applicazione maggiore. Viene spesso utilizzata negli studi di affidabilità e di mortalità delle malattie umane.

Weibull è caratterizzata da due parametri, γ e λ . Il valore di γ determina la forma della curva della distribuzione mentre il valore di λ ne determina la scala. Di conseguenza, sono chiamati, rispettivamente, i parametri di forma e di scala.

La relazione tra γ e il tempo di sopravvivenza è illustrata nella figura 5, che mostra la funzione di rischio della distribuzione di Weibull, con $\lambda = 1$, al variare di γ .

Figura 5. Funzione di rischio della distribuzione di Weibull con $\lambda=1$

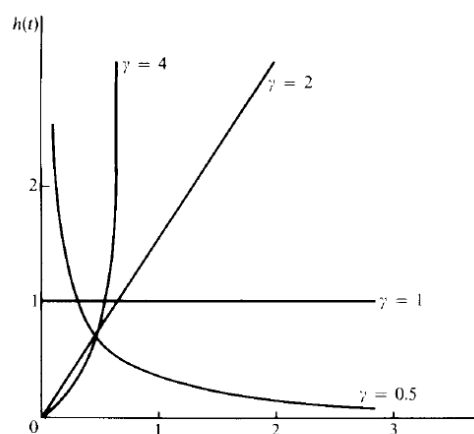


Figure 6.4 Hazard functions of Weibull distribution with $\lambda = 1$.

Quando $\Upsilon = 1$, la funzione di rischio rimane costante al crescere del tempo; questo è esattamente il caso esponenziale; il tasso di rischio aumenta quando $\Upsilon > 1$ e decresce quando $\Upsilon < 1$ al crescere del tempo. Pertanto, la distribuzione di Weibull può essere usata per modellare la distribuzione di sopravvivenza di una popolazione con rischio crescente, decrescente o costante.

La funzione di densità di probabilità e la funzione di distribuzione cumulativa sono rispettivamente:

$$f(t) = \lambda \Upsilon (\lambda t)^{\Upsilon-1} e^{-(\lambda t)^\Upsilon} \text{ con } t \geq 0, \Upsilon, \lambda > 0 \quad (1.47)$$

e

$$F(T) = 1 - e^{-(\lambda t)^\Upsilon} \quad (1.48)$$

La funzione di sopravvivenza e la funzione di rischio sono date da:

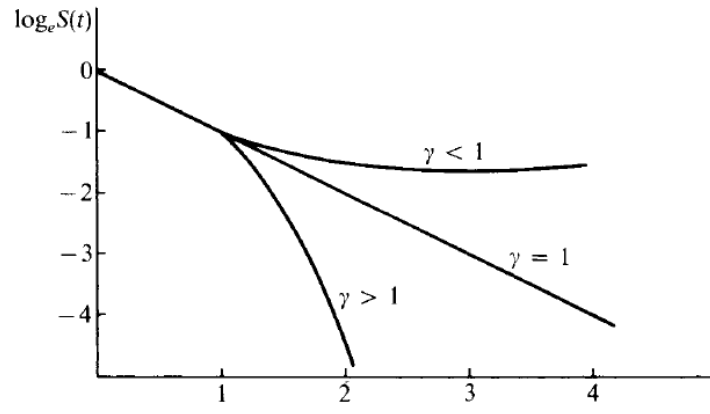
$$S(t) = e^{-(\lambda t)^\Upsilon} \quad (1.49)$$

$$h(t) = \lambda \Upsilon (\lambda t)^{\Upsilon-1} \quad (1.50)$$

Per la funzione di sopravvivenza, è più semplice plottare il logaritmo di $S(t)$, come mostrato in figura.

$$\log(S(t)) = -(\lambda t)^\Upsilon$$

Figura 6. Curve di $\log S(t)$ della distribuzione di Weibull con $\lambda=1$



Quando $\gamma=1$ otteniamo una linea dritta con pendenza negativa, quando $\gamma<1$, invecchiamento negativo, $\log(S(t))$ decresce molto lentamente da zero per poi avvicinarsi ad un valore costante. Infine, quando $\gamma>1$, invecchiamento positivo, $\log S(t)$ diminuisce bruscamente da 0 all'aumentare di t .

La media della distribuzione di Weibull è data da:

$$\mu = \frac{\Gamma(1+\frac{1}{\gamma})}{\lambda} \quad (1.51)$$

e la varianza:

$$\sigma^2 = \frac{1}{\lambda^2} \left[\Gamma\left(1 + \frac{2}{\gamma}\right) - \Gamma^2\left(1 + \frac{1}{\gamma}\right) \right] \quad (1.52)$$

dove $\Gamma(\gamma)$ rappresenta una funzione gamma, definita come:

$$\Gamma(\gamma) = \int_0^{\infty} x^{\gamma-1} e^{-x} dx = (\gamma - 1)! \quad (1.53)$$

Dove Υ è un intero positivo. Infine, il coefficiente di variazione è pari a:

$$CV = \left[\frac{\Gamma(1 + \frac{2}{\Upsilon})}{\Gamma^2(1 + \frac{1}{\Upsilon})} - 1 \right]^{1/2} \quad (1.54)$$

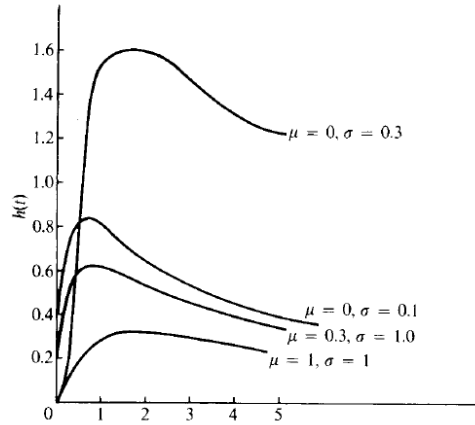
Distribuzione Log-normale

In maniera molto semplice la distribuzione log-normale può essere definita come la distribuzione di una variabile il cui logaritmo segue la distribuzione normale.

Considerando il tempo di sopravvivenza T ed il suo logaritmo normalmente distribuito con media μ e varianza σ^2 . Possiamo quindi dire che T è distribuito secondo una log-normale e si scrive $\Delta(\mu; \sigma^2)$. Bisogna precisare che μ e σ^2 non sono la media e la varianza della distribuzione lognormale.

La figura seguente mostra la funzione di rischio della distribuzione lognormale con differenti valori dei parametri. La funzione aumenta inizialmente fino a raggiungere un punto di massimo e poi diminuisce (quasi non appena viene superata la mediana) fino a zero quando il tempo tende all'infinito. Pertanto, questa distribuzione è adatta a modelli di sopravvivenza con un tasso di rischio inizialmente crescente e poi decrescente.

Figura 7. Funzione di rischio della distribuzione lognormale con differenti parametri



La funzione di densità di probabilità e la funzione di sopravvivenza sono stimate come:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2} (\log t - \mu)^2\right] \quad t > 0, \sigma > 2 \quad (1.55)$$

$$S(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_t^\infty \frac{1}{x} \exp\left[-\frac{1}{2\sigma^2} (\log x - \mu)^2\right] dx \quad (1.56)$$

La distribuzione log-normale è univocamente specificata attraverso i due parametri di media e di varianza che sono entrambi parametri di scala a differenza della distribuzione normale; ovviamente il tempo T non può assumere il valore zero poiché altrimenti il suo logaritmo non sarebbe definito.

Distribuzione Log-logistica

Il tempo di sopravvivenza T ha una distribuzione log-logistica se il $\log(T)$ ha una distribuzione logistica. Le funzioni di densità, sopravvivenza, rischio e di rischio cumulativo sono, rispettivamente,

$$f(t) = \frac{\alpha Y t^{Y-1}}{(1 + \alpha t^Y)^2} \quad (1.57)$$

$$S(t) = \frac{1}{(1 + \alpha t^\gamma)} \quad (1.58)$$

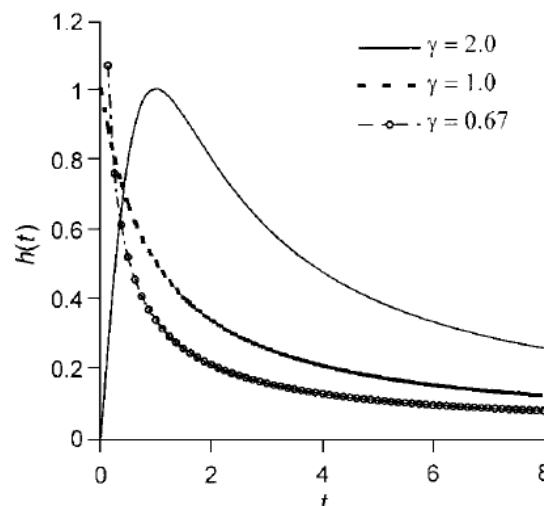
$$h(t) = \frac{\alpha \gamma t^{\gamma-1}}{(1 + \alpha t^\gamma)} \quad (1.59)$$

$$H(t) = \log(1 + \alpha t^\gamma) \quad (1.60)$$

$$t \geq 0; \alpha > 0; \gamma > 0$$

La distribuzione log-logistica è caratterizzata da due parametri α e γ . La mediana della distribuzione log-logistica è $\alpha^{-\frac{1}{\gamma}}$. A titolo di esempio, consideriamo la funzione di rischio con $\alpha=1$ e con differenti valori di γ (figura 8). Con $\gamma>1$ il rischio è pari a zero al tempo zero, aumenta fino a raggiungere il picco e poi diminuisce, similmente al rischio log-normale. Quando $\gamma=1$, la funzione al tempo zero è pari a $\alpha^{\frac{1}{\gamma}}$ e poi decresce monotonamente; quando $\gamma<1$, al tempo zero il rischio “parte da infinito” e poi decresce, in maniera simile alla Weibull, e la funzione tende a zero quando il tempo tende all’infinito.

Figura 8. Distribuzione log-logistica, funzione di rischio



Pertanto, la distribuzione log-logistica può essere utilizzata per descrivere un rischio prima crescente e poi decrescente o un rischio monotono decrescente.

Distribuzione Gamma

La distribuzione gamma, che comprende la distribuzione esponenziale e la chi-quadro, è stata spesso utilizzata per descrivere la vita dei bicchieri di vetro che circolano in una caffetteria e come modello statistico per la durata dei materiali. Da allora, questa distribuzione è stata usata frequentemente come modello per problemi di affidabilità industriale e per analisi circa la sopravvivenza umana.

Supponiamo che il fallimento o la morte avvengano in n fasi separate, al termine della prima fase, dopo il tempo T_1 , si verificano i primi “sotto-fallimenti”, dopodiché, inizia il secondo stadio e dopo il tempo T_2 si verificano altri fallimenti e così via. Il fallimento totale o la morte si verifica alla fine dell' n -esimo stadio, quando si verifica l' n -esimo sotto-fallimento. Il tempo di sopravvivenza totale, T , è quindi pari a $T_1 + T_2 + \dots + T_n$.

I tempi $T_1 + T_2 + \dots + T_n$ si assume siano distribuiti in maniera indipendente secondo una distribuzione esponenziale con funzione di densità di probabilità $\lambda \exp(-\lambda t_i)$ con $i = 1, \dots, n$. Cioè i sotto-fallimenti si verificano in maniera indipendente tra di loro ad un tasso costante λ . La distribuzione di T è chiamata **distribuzione di Erlangian**.

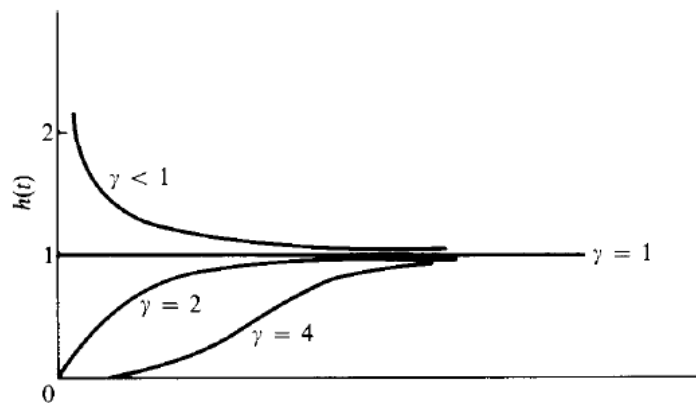
Una naturale generalizzazione di questa distribuzione consiste nella sostituzione del parametro n , formato dagli interi $1, 2, \dots$, con un parametro Υ che assume un qualsiasi valore reale positivo. Otteniamo così la distribuzione gamma.

La distribuzione gamma è caratterizzata da due parametri Υ e λ .

Quando $0 < \Upsilon < 1$, c'è un invecchiamento negativo e il tasso di rischio diminuisce monotonamente dall'infinito a λ quando il tempo va da 0 ad infinito. Quando $\Upsilon > 1$ c'è un invecchiamento positivo e il tasso di rischio aumenta monotonamente da 0 a λ mentre il tempo aumenta da 0 a infinito. Quando $\Upsilon = 1$, il tasso di rischio è uguale a λ ,

una costante, come nel caso esponenziale. Pertanto, la distribuzione gamma descrive un diverso tipo di modello di sopravvivenza in cui il tasso di rischio è in diminuzione o crescente ad un valore costante quando il tempo si avvicina all'infinito.

Figura 9. Distribuzione Gamma. Funzione di rischio con $\lambda=1$



La funzione di densità di probabilità della funzione gamma è:

$$f(t) = \frac{\lambda}{\Gamma(\gamma)} (\lambda t)^{\gamma-1} e^{-\lambda t} \quad \text{per } t > 0 \text{ e } \gamma, \lambda > 0 \quad (1.61)$$

I parametri γ e λ sono rispettivamente dei parametri di forma e di scala. La funzione di distribuzione cumulativa è data invece da:

$$F(t) = \int_0^t \frac{\lambda}{\Gamma(\gamma)} (\lambda x)^{\gamma-1} e^{-\lambda x} dx \quad (1.62)$$

Le funzioni di sopravvivenza e di rischio sono invece:

$$S(t) = \int_t^\infty \frac{\lambda}{\Gamma(\gamma)} (\lambda x)^{\gamma-1} e^{-\lambda x} dx \quad (1.63)$$

$$h(t) = \frac{\lambda(\lambda t)^{n-1}}{(n-1)! \sum_{k=0}^{n-1} \left(\frac{1}{k!}\right)(\lambda t)^k} \quad (1.64)$$

Quando $\Upsilon=1$ la distribuzione è esponenziale. Quando $\lambda = 1/2$ e $\Upsilon = (1/2)v$, con v numero intero, la distribuzione è una chi-quadro con v gradi di libertà. La media e la varianza della distribuzione gamma standard sono Υ/λ e Υ/λ^2 mentre il coefficiente di variazione è pari a $1/\sqrt{\Upsilon}$.

La distribuzione esponenziale, di Weibull, la log-normale e gamma sono casi speciali di una distribuzione gamma generalizzata con tre parametri λ, Υ e α , la cui funzione di densità è definita come:

$$f(t) = \frac{\alpha \lambda^{\alpha \Upsilon}}{\Gamma(\Upsilon)} t^{\alpha \Upsilon - 1} e^{-(\lambda t)^{\alpha}} \quad (1.65)$$

Per $t > 0$ e $\alpha, \Upsilon, \lambda > 0$

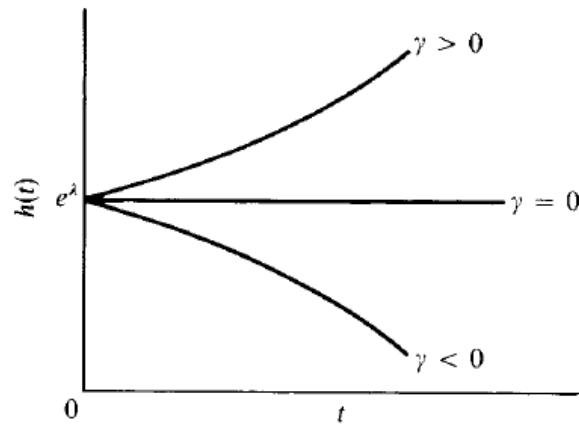
- Se $\alpha = \Upsilon = 1$, si ottiene una distribuzione esponenziale
- Se $\Upsilon = 1$, distribuzione Weibull
- Se $\Upsilon \rightarrow \infty$, distribuzione log-normale
- Se $\alpha = 1$, distribuzione gamma.

Distribuzione di Gompertz

La distribuzione di Gompertz è un'estensione della distribuzione esponenziale; descrive un modello di sopravvivenza con un tasso di rischio iniziale costante, che poi varia come funzione esponenziale del tempo. La distribuzione di Gompertz è caratterizzata da due parametri λ e Υ . La funzione di rischio (figura 10) è calcolata come:

$$h(t) = \exp(\lambda + \Upsilon t) \quad (1.66)$$

Figura 10. Funzione di rischio di Gompertz.



Quando $\gamma > 0$ c'è un invecchiamento positivo a partire da e^λ ; quando $\gamma < 0$ c'è invece un invecchiamento negativo. Infine, quando $\gamma = 0$, la funzione di rischio si riduce ad una costante e^λ . La funzione di sopravvivenza della distribuzione di Gompertz:

$$S(t) = \exp \left[- \frac{e^\lambda}{\gamma} (e^{\gamma t} - 1) \right] \quad (1.67)$$

Infine, la funzione di densità di probabilità:

$$f(t) = \exp \left[(\lambda + \gamma t) - \frac{1}{\gamma} (e^{\lambda + \gamma t} - e^\lambda) \right] \quad (1.68)$$

Calcolo delle Stime per le Distribuzioni Parametriche di Sopravvivenza in assenza di Covariate

In questa sezione si discutono le procedure analitiche per la stima delle distribuzioni di sopravvivenza precedentemente descritte. In particolar modo le **stime di massima verosimiglianza** (MLE) sono utilizzate per il calcolo dei parametri di queste distribuzioni. Il simbolo, che nell'algebra lineare prende il nome di vettore, $\mathbf{b} = (b_1, b_2, \dots, b_p)$ denota un insieme di parametri caratterizzanti una ben precisa distribuzione; ad esempio, parlando della distribuzione di Weibull si ha che $\mathbf{b} = (\lambda, \gamma)$. I risultati generali dell'inferenza di verosimiglianza asintotica più ampiamente utilizzati sono di seguito descritti.

Procedure di stima (MLE) per i dati con osservazioni censurate a destra

Si suppone di effettuare uno studio in cui un certo numero di persone sono seguite fino al momento della loro morte o vengano censurate. Siano $t_1, t_2, \dots, t_r, t_{r+1}^+, \dots, t_n^+$ i tempi osservati relativi alle n persone partecipanti all'analisi con r tempi noti e $(n - r)$ i tempi censurati a destra. Si assume che i tempi di sopravvivenza seguano una distribuzione con funzione di densità $f(t, \mathbf{b})$ e funzione di sopravvivenza $S(t, \mathbf{b})$, dove $\mathbf{b}$ denota p parametri sconosciuti della distribuzione.

Se il tempo di sopravvivenza è discreto, $f(t, \mathbf{b})$ rappresenta la probabilità di osservare t e $S(t, \mathbf{b})$ la probabilità che la sopravvivenza o il tempo in cui si verifica un certo evento sia maggiore di t . In altre parole, $f(t, \mathbf{b})$ e $S(t, \mathbf{b})$ rappresentano le informazioni che possono essere ottenute rispettivamente da un tempo di sopravvivenza osservato non censurato e da un tempo di sopravvivenza osservato censurato a destra.

Pertanto, il prodotto $\prod_{i=1}^n f(t_i, \mathbf{b})$ rappresenta la probabilità congiunta di osservare i tempi di sopravvivenza non censurati e $\prod_{i=r+1}^n S(t_i^+, \mathbf{b})$ la probabilità congiunta dei

tempi di sopravvivenza censurati a destra. Il prodotto di queste due probabilità, denotato con $L(\mathbf{b})$,

$$L(\mathbf{b}) = \prod_{i=1}^r f(t_i, \mathbf{b}) \prod_{i=r+1}^n S(t_i^+, \mathbf{b}) \quad (1.69)$$

rappresenta la probabilità congiunta di osservare $t_1, t_2, \dots, t_r, t_{r+1}^+, \dots, t_n^+$. Un'interpretazione analoga si applica nel caso di tempo continuo. $L(\mathbf{b})$ è detta **funzione di verosimiglianza di $\mathbf{b}$** , che può essere interpretata come una misura della probabilità di osservare uno specifico insieme di tempi di sopravvivenza, dato uno specifico insieme di parametri.

Il metodo MLE consiste nel trovare uno stimatore di $\mathbf{b}$ che massimizzi $L(\mathbf{b})$, o in altri termini, che è più probabile abbia prodotto i dati osservati $t_1, t_2, \dots, t_r, t_{r+1}^+, \dots, t_n^+$.

A partire dal logaritmo di $L(\mathbf{b})$:

$$l(\mathbf{b}) = \log L(\mathbf{b}) = \sum_{i=1}^r \log[f(t_i; \mathbf{b})] + \sum_{i=r+1}^n \log[S(t_i^+; \mathbf{b})] \quad (1.70)$$

Lo stimatore MLE $\hat{\mathbf{b}}$ di $\mathbf{b}$ è l'insieme $\hat{b}_1, \dots, \hat{b}_p$ che massimizza $l(\mathbf{b})$:

$$l(\hat{\mathbf{b}}) = \max_{all \mathbf{b}} (l(\mathbf{b}))$$

$\hat{\mathbf{b}}$ è la soluzione delle seguenti equazioni simultanee:

$$\frac{\partial l(\mathbf{b})}{\partial b_j} = 0, \quad j = 1, 2, \dots, p$$

Spesso non ci sono soluzioni chiuse per MLE $\hat{\mathbf{b}}$, ed è quindi necessario utilizzare un metodo numerico. Un metodo numerico comunemente usato è la procedura iterativa Newton-Raphson che può essere descritta come segue:

1. Porre i valori iniziali di $b_1, b_2, \dots, b_p$ pari a zero

$$b^{(0)} = 0$$

2. I cambiamenti per b ad ogni passo successivo, indicati con $\Delta^{(j)}$, si ottengono considerando la derivata seconda della funzione di verosimiglianza:

$$\Delta^{(j)} = \left[-\frac{\partial^2 l(b^{j-1})}{\partial b \partial b'} \right]^{-1} \frac{\partial l(b^{j-1})}{\partial b}$$

3. Usando $\Delta^{(j)}$, il valore di $b^{(j)}$ al j -esimo step è:

$$b^{(j)} = b^{(j-1)} + \Delta^{(j)} \quad j = 1, 2, \dots$$

L'iterazione termina all' m -esimo step se $\|\Delta^{(m)}\| < \delta$, dove δ è un valore dato di precisione, solitamente un valore molto piccolo, 10^{-4} o 10^{-5} . Allora lo stimatore $\hat{b}$ è dato da:

$$\hat{b} = b^{(m-1)} \quad (1.71)$$

La varianza dello stimatore è pari a:

$$\widehat{V(b)} = Cov(b) = \left[-\frac{\partial^2 l(b)}{\partial b \partial b'} \right]^{-1} \quad (1.72)$$

L'intervallo di confidenza stimato $100(1 - \alpha)\%$ per ogni parametro è dato da:

$$(\hat{b}_i - Z_{\frac{\alpha}{2}} \sqrt{v_{ii}}; \hat{b}_i + Z_{\frac{\alpha}{2}} \sqrt{v_{ii}}) \quad (1.73)$$

Dove v_{ii} è l' i -esimo elemento della diagonale della matrice $\widehat{V(b)}$ e $Z_{\frac{\alpha}{2}}$ è il $100 \left(1 - \frac{\alpha}{2}\right)$ percentile della distribuzione normale standard $\left[P \left(Z > Z_{\frac{\alpha}{2}} \right) = \frac{\alpha}{2} \right]$.

Procedure di stima (MLE) per i dati con osservazioni censurate a destra, a sinistra e per intervallo

Se i tempi $t_1, t_2, \dots, t_n$ osservati per le n persone sono formate da osservazioni non censurate, censurate a sinistra, a destra e per intervalli, le procedure di stima sono molto simili. Assumiamo che i tempi di sopravvivenza seguano una distribuzione con funzione di densità $f(t, \mathbf{b})$ e con funzione di sopravvivenza $S(t, \mathbf{b})$, con $\mathbf{b}$ che indica tutti i parametri sconosciuti della distribuzione. Allora la funzione di log-verosimiglianza è:

$$\begin{aligned} l(\mathbf{b}) = \log L(\mathbf{b}) = & \sum \log[f(t_i; \mathbf{b})] \\ & + \sum \log[S(t_i; \mathbf{b})] \\ & + \sum \log[1 - S(t_i; \mathbf{b})] + \sum \log[S(v_i; \mathbf{b}) - S(t_i; \mathbf{b})] \end{aligned} \quad (1.74)$$

dove la prima somma riguarda le osservazioni non censurate, la seconda le osservazioni censurate a destra, la terza le osservazioni censurate a sinistra e l'ultima somma le osservazioni censurate per intervallo, con v_i come estremità inferiore di un intervallo di censura. Gli altri passaggi per ottenere gli stimatori MLE sono simili ai passaggi descritti nella Sezione precedendo, sostituendo la funzione di verosimiglianza (1.70) con la funzione di verosimiglianza in (1.74).

Di seguito sono riportati i risultati ottenuti per le principali distribuzioni:

Distribuzione esponenziale

- In assenza di osservazioni censurate, lo stimatore λ MLE è pari:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^n t_i}$$

per il quale è possibile calcolare un intervallo di confidenza esatto $100(1-\alpha)\%$:

$$\frac{\hat{\lambda} \chi_{2n, 1-\alpha/2}^2}{2n} < \lambda < \frac{\hat{\lambda} \chi_{2n, \alpha/2}^2}{2n}$$

dove $\chi_{2n, \alpha/2}^2$ indica una distribuzione chi-quadro con 2n gradi di libertà, tale per cui $P(\chi_{2n}^2 > \chi_{2n, \alpha}^2) = \alpha$.

Nel caso di campioni molto numerosi ($n \gg 25$), $\hat{\lambda}$ è approssimativamente distribuita secondo una normale come media λ e con varianza $\frac{\lambda^2}{n}$, in questo caso l'intervallo di confidenza è dato da:

$$\hat{\lambda} - \frac{\hat{\lambda} Z_{\frac{\alpha}{2}}}{\sqrt{n}} < \lambda < \hat{\lambda} + \frac{\hat{\lambda} Z_{\frac{\alpha}{2}}}{\sqrt{n}}$$

dove $Z_{\alpha/2}$ rappresenta il 100 $\alpha/2$ punto percentuale, $P(Z > Z_{\alpha/2}) = \alpha/2$, della distribuzione normale standard.

- In presenza di osservazioni censurate singolarmente

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^r t_i + \sum_{i=r+1}^n t_{(i)}^+}$$

La media e la varianza di $\hat{\lambda}$ sono pari a $\frac{r\lambda}{(r-1)}$ e $\frac{\lambda^2}{(r-1)}$. Il rispettivo intervallo di confidenza:

$$\frac{\hat{\lambda} \chi_{2r, 1-\alpha/2}^2}{2r} < \lambda < \frac{\hat{\lambda} \chi_{2r, \alpha/2}^2}{2r}$$

Nel caso di campioni molto numerosi ($n \gg 25$), $\hat{\lambda}$ è approssimativamente distribuita secondo una normale come media λ e con varianza $\frac{\lambda^2}{(r-1)}$, in questo caso l'intervallo di confidenza è dato da:

$$\hat{\lambda} - \frac{\hat{\lambda} Z_{\frac{\alpha}{2}}}{\sqrt{r-1}} < \lambda < \hat{\lambda} + \frac{\hat{\lambda} Z_{\frac{\alpha}{2}}}{\sqrt{r-1}}$$

- In presenza di osservazioni censurate progressivamente:

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^r t_i + \sum_{i=r+1}^n t_{(i)}^+}$$

Nel caso di campioni molto numerosi ($n \gg 25$), $\hat{\lambda}$ è approssimativamente distribuita secondo una normale come media λ e con varianza

$$Var(\hat{\lambda}) = \frac{\lambda^2}{\sum_{i=1}^n (1 - e^{-\lambda T_i})}$$

dove T_i è il tempo in cui la i -esima persona rimane sotto osservazione. In poche parole, si tiene conto del tempo che passa da quando la i -esima persona entra nello studio fino alla fine dello studio stesso. Il rispettivo intervallo di confidenza è pari a:

$$\hat{\lambda} - Z_{\frac{\alpha}{2}} \sqrt{Var(\hat{\lambda})} < \lambda < \hat{\lambda} + Z_{\frac{\alpha}{2}} \sqrt{Var(\hat{\lambda})}$$

Distribuzione di Weibull

- In assenza di osservazioni censurate gli stimatori MLE di λ e di Y possono essere ottenuti risolvendo contemporaneamente le seguenti equazioni, usando la procedura di Newton-Raphson:

$$n - \hat{\lambda}^{\hat{Y}} \sum_{i=1}^n t_i^{\hat{Y}} = 0$$

$$\frac{n}{\hat{Y}} + n \log(\hat{\lambda}) + \sum_{i=1}^n \log(t_i) - \hat{\lambda}^{\hat{Y}} \sum_{i=1}^n t_i^{\hat{Y}} (\log \hat{\lambda} + \log t_i) = 0$$

- In presenza di dati censurati singolarmente, le equazioni da risolvere sono date da:

$$r - \hat{\lambda}^{\hat{Y}} \left[\sum_{i=1}^r t_i^{\hat{Y}} + (n-r)t_{(r)}^{\hat{Y}} \right] = 0$$

$$\frac{n}{\hat{Y}} + r \log(\hat{\lambda}) + \sum_{i=1}^r \log(t_i) + \hat{\lambda}^{\hat{Y}} \left[\sum_{i=1}^n t_i^{\hat{Y}} (\log \hat{\lambda} + \log t_i) + (n-r)t_r^{\hat{Y}} (\log \hat{\lambda} + \log t_{(r)}) \right] = 0$$

- Infine, se i dati sono progressivamente censurati:

$$r - \hat{\lambda}^{\hat{Y}} \left[\sum_{i=1}^r t_i^{\hat{Y}} + \sum_{i=r+1}^n t_i^{+\hat{Y}} \right] = 0$$

$$\frac{n}{\hat{Y}} + r \log(\hat{\lambda}) + \sum_{i=1}^r \log(t_i) - \hat{\lambda}^{\hat{Y}} \left[\sum_{i=1}^r t_i^{\hat{Y}} (\log \hat{\lambda} + \log t_i) \right] - \hat{\lambda}^{\hat{Y}} \left[\sum_{i=r+1}^n t_i^{+\hat{Y}} (\log \hat{\lambda} + \log t_i^+) \right] = 0$$

Distribuzione Log-normale

Se il tempo di sopravvivenza T segue una distribuzione log-normale, la media e la varianza sono rispettivamente $\exp(\mu + 1/2\sigma^2)$ e $[\exp(\sigma^2) - 1]\exp(2\mu + \sigma^2)$. Sfruttando il fatto che $Y = \log T$ segue una distribuzione normale con media μ e varianza σ^2 .

- In assenza di osservazioni censurate, lo stimatore MLE di μ è la media semplice di Y :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \log(t_i)$$

Lo stimatore MLE di σ^2 è:

$$\widehat{\sigma^2} = \frac{1}{n} \left[\sum_{i=1}^n \log(t_i)^2 - \frac{(\sum_{i=1}^n \log t_i)^2}{n} \right]$$

La stima della media non è affetta da bias, mentre quella della varianza lo è. La migliore stima della varianza è quindi:

$$s^2 = \widehat{\sigma^2} \left[\frac{n}{(n-1)} \right]$$

Se n è moderatamente grande, la differenza tra s^2 e $\widehat{\sigma^2}$ è trascurabile. La media e la varianza MLE di T sono pari a $\exp\left(\widehat{\mu} + \frac{\widehat{\sigma^2}}{2}\right)$ e $[\exp(\widehat{\sigma^2} - 1)]\exp(2\widehat{\mu} + \widehat{\sigma^2})$. La media $\widehat{\mu}$ è normalmente distribuita con media μ e varianza $\frac{\sigma^2}{n}$; pertanto, se la varianza è nota, l'intervallo di confidenza per μ è $\widehat{\mu} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$, se invece non è nota possiamo usare la t-Student con $n-1$ gradi di libertà ($\widehat{\mu} \pm t_{\alpha/2, (n-1)} \frac{s}{\sqrt{n-1}}$). Infine, l'intervallo di confidenza per σ^2 :

$$\frac{n\widehat{\sigma^2}}{\chi^2_{(n-1), \alpha/2}} < \sigma^2 < \frac{n\widehat{\sigma^2}}{\chi^2_{(n-1), 1-\alpha/2}}$$

- In presenza di osservazioni singolarmente censurate, sfruttando nuovamente la distribuzione normale di Y , la stima di μ e σ^2 può essere ottenuta trasformando i dati $y_i = \log t_i$. La migliore stima lineare è quella proposta da Sarhan e Greenberg, combinazioni lineari dei logaritmi degli r tempi di sopravvivenza esatti:

$$\widehat{\sigma} = \sum_{i=1}^r b_i \log(t_i)$$

$$\hat{\mu} = \sum_{i=1}^r a_i \log(t_i)$$

dove i coefficienti a_i e b_i sono calcolati e tabulati da Sarhan e Greenberg per $n \leq 20$ e parzialmente riprodotto nell'Allegato A. La varianza e la covarianza di $\hat{\mu}$ e $\hat{\sigma}$ sono tabulate nell'Allegato B.

Nel caso ci trovassimo davanti un set di dati molto più ampio, $n > 20$, è possibile usare lo stimatore MLE di Cohen per la distribuzione normale

$$\bar{y} = \frac{1}{r} \sum_{i=1}^r \log(t_i)$$

e

$$s^2 = \frac{1}{r} \left[\sum_{i=1}^n \log(t_i)^2 - \frac{(\sum_{i=1}^n \log t_i)^2}{r} \right]$$

Gli stimatori MLE di μ e σ^2 sono:

$$\hat{\mu} = \bar{y} - \hat{\lambda}(\bar{y} - \log t_{(r)})$$

e

$$\hat{\sigma}^2 = s^2 + \hat{\lambda}(\bar{y} - \log t_{(r)})^2$$

Il valore di $\hat{\lambda}$ è stato tabulato da Cohen come una funzione di **a** e di **b**. La proporzione delle osservazioni censurate, **b**, è calcolata come:

$$b = \frac{n-r}{n}$$

e

$$a = \frac{1 - Y(Y - c)}{(Y - c)^2}$$

dove:

$$Y = \left[\frac{b}{1-b} \right] f(c)/F(c),$$

$f(c)$ e $F(c)$ rappresentano la funzione di densità e di distribuzione della normale standard valutata in $c = (\log t_r - \mu)/\sigma$. La tabella 11 mostra il valore di $\hat{\lambda}$ in corrispondenza di una ben precisa coppia di valori (a,b).

Tabella 11.

a	b																			
	.01	.02	.03	.04	.05	.06	.07	.08	.09	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60
.00	.010100	.020400	.030902	.041583	.052507	.063627	.074953	.086488	.09824	.11020	.17342	.24268	.31862	.4021	.4941	.5961	.7096	.8368	.9808	1.145
.05	.010551	.021294	.032225	.043350	.054670	.066189	.077909	.089834	.10197	.11431	.17935	.25033	.32793	.4130	.5066	.6101	.7252	.8540	.9994	1.166
.10	.010950	.022082	.033398	.044902	.056596	.068483	.080568	.092852	.10534	.11804	.18479	.25741	.33662	.4233	.5184	.6234	.7400	.8703	1.017	1.185
.15	.011310	.022798	.034466	.046318	.058356	.070586	.083009	.095629	.10845	.12148	.18985	.26405	.34480	.4330	.5296	.6361	.7542	.8860	1.035	1.204
.20	.011642	.023459	.035453	.047629	.059990	.072539	.085280	.098216	.11135	.12469	.19460	.27031	.35255	.4422	.5403	.6483	.7678	.9012	1.051	1.222
.25	.011952	.024076	.036377	.048858	.061522	.074372	.087413	.10065	.11408	.12772	.19910	.27626	.35993	.4510	.5506	.6600	.7810	.9158	1.067	1.240
.30	.012243	.024658	.037249	.050018	.062969	.076106	.089433	.10295	.11667	.13059	.20338	.28193	.36700	.4595	.5604	.6713	.7937	.9300	1.083	1.257
.35	.012520	.025211	.038077	.051120	.064345	.077756	.091355	.10515	.11914	.13333	.20747	.28737	.37379	.4676	.5699	.6821	.8060	.9437	1.098	1.274
.40	.012784	.025738	.038866	.052173	.065660	.079332	.093193	.10725	.12150	.13595	.21139	.29260	.38033	.4755	.5791	.6927	.8179	.9570	1.113	1.290
.45	.013036	.026243	.039624	.053182	.066921	.080845	.094958	.10926	.12377	.13847	.21517	.29765	.38665	.4831	.5880	.7029	.8295	.9700	1.127	1.306
.50	.013279	.026728	.040352	.054153	.068135	.082301	.096657	.11121	.12595	.14090	.21882	.30253	.39276	.4904	.5967	.7129	.8408	.9826	1.141	1.321
.55	.013513	.027196	.041054	.055089	.069306	.083708	.098298	.11308	.12806	.14325	.22235	.30725	.39870	.4976	.6051	.7225	.8517	.9950	1.155	1.337
.60	.013739	.027649	.041733	.055995	.070439	.085068	.099887	.11490	.13011	.14552	.22578	.31184	.40447	.5045	.6133	.7320	.8625	1.007	1.169	1.351
.65	.013958	.028087	.042391	.056874	.071538	.086388	.10143	.11666	.13209	.14773	.22910	.31630	.41008	.5114	.6213	.7412	.8729	1.019	1.182	1.366
.70	.014171	.028513	.043030	.057726	.072605	.087670	.10292	.11837	.13402	.14987	.23234	.32065	.41555	.5180	.6291	.7502	.8832	1.030	1.195	1.380
.75	.014378	.028927	.043652	.058556	.073643	.088917	.10438	.12004	.13590	.15196	.23550	.32489	.42090	.5245	.6367	.7590	.8932	1.042	1.207	1.394
.80	.014579	.029330	.044258	.059364	.074655	.090133	.10580	.12167	.13773	.15400	.23858	.32903	.42612	.5308	.6441	.7676	.9031	1.053	1.220	1.408
.85	.014755	.029723	.044848	.060153	.075642	.091319	.10719	.12325	.13952	.15599	.24158	.33307	.43122	.5370	.6515	.7761	.9127	1.064	1.232	1.422
.90	.014967	.030107	.045425	.060923	.076606	.092477	.10854	.12480	.14126	.15793	.24452	.33703	.43622	.5430	.6586	.7844	.9222	1.074	1.244	1.435
.95	.015154	.030483	.045989	.061676	.077549	.093611	.10987	.12632	.14297	.15983	.24740	.34091	.44112	.5490	.6656	.7925	.9314	1.085	1.255	1.448
1.00	.015338	.030850	.046540	.062413	.078471	.094720	.11116	.12780	.14465	.16170	.25022	.34471	.44592	.5548	.6724	.8005	.9406	1.095	1.267	1.461

Source: Cohen (1961).

* For all values $0 \leq a \leq 1$, $\lambda = 0$.

Le varianze e la covarianza sono le seguenti:

$$Var(\hat{\mu}) \simeq \frac{\sigma^2}{n} m_1$$

$$Var(\hat{\sigma}) \simeq \frac{\sigma^2}{n} m_2$$

$$Cov(\hat{\mu}; \hat{\sigma}) \simeq \frac{\sigma^2}{n} m_3$$

m_1, m_2 e m_3 sono coefficienti calcolati di Cohen e riportati nell'Allegato C, tenendo conto che $y = -\hat{c}$.

- Nel caso in cui abbiamo delle osservazioni progressivamente censurate, gli stimatori MLE possono essere ottenuti risolvendo contemporaneamente le seguenti due equazioni, applicando il metodo di Newton-Raphson:

$$\sum_{i=1}^r \frac{\log t_i - \mu}{\sigma^2} + \sum_{i=r+1}^n \frac{\int_{t_i^+}^{\infty} \frac{\log x - \mu}{x \sigma^2 \sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\log x - \mu)^2\right] dx}{\int_{t_i^+}^{\infty} \frac{1}{x \sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\log x - \mu)^2\right] dx} = 0$$

$$-\frac{n}{2\sigma^2} \sum_{i=1}^r \frac{(\log t_i - \mu)^2}{2\sigma^4} + \sum_{i=r+1}^n \frac{\int_{t_i^+}^{\infty} \frac{(\log x - \mu)^2}{x 2\sigma^4 \sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\log x - \mu)^2\right] dx}{\int_{t_i^+}^{\infty} \frac{1}{x \sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\log x - \mu)^2\right] dx} = 0$$

Distribuzione Log-Logistica

- Sia che si tratti di osservazioni censurate o meno, gli stimatori MLE sono ottenuti risolvendo le due equazioni seguenti, usando la procedura di Newton-Raphson:

$$r - \alpha \left[2 \sum_{i=1}^r \frac{t_i^Y}{1 + \alpha t_i^Y} + \sum_{i=r+1}^n \frac{t_i^{+Y}}{1 + \alpha t_i^Y} \right] = 0$$

e

$$\frac{r}{\bar{Y}} + \sum_{i=1}^r \log t_{(i)} - \alpha \left[2 \sum_{i=1}^r \frac{t_i^Y \log t_{(i)}}{1 + \alpha t_i^Y} + \sum_{i=r+1}^n \frac{t_i^{+Y} \log t_i^+}{1 + \alpha t_i^Y} \right] = 0$$

In mancanza di osservazioni censurate, è possibile utilizzare le medesime equazioni descritte, sostituendo ad r, n .

Distribuzione di Gompertz

- Sia in presenza che in assenza di osservazioni censurate, gli stimatori MLE possono essere calcolati risolvendo le seguenti equazioni,

$$r + \frac{e^\lambda}{\gamma} \left\{ \sum_{i=1}^r [1 - \exp(\gamma t_i)] + \sum_{i=r+1}^n [1 + \exp(\gamma t_i^+)] \right\} = 0$$

$$\sum_{i=1}^r t_i + \frac{e^\lambda}{\gamma^2} \left\{ \sum_{i=1}^r [1 - (\gamma t_i - 1) \exp(\gamma t_i)] + \sum_{i=r+1}^n [1 + (\gamma t_i^+ - 1) \exp(\gamma t_i^+)] \right\} = 0$$

usando la procedura iterativa di Newton-Raphson. Se non ci sono osservazioni censurate, è sufficiente sostituire ad r la dimensione totale del campione n .

Distribuzione Gamma

- In assenza di osservazioni censurate, lo stimatore MLE può essere calcolato risolvendo simultaneamente le seguenti due equazioni usando la procedura di Newton-Raphson:

$$\frac{n\hat{\gamma}}{\hat{\lambda}} - \sum_{i=1}^n t_i = 0$$

$$n \log \hat{\lambda} - \frac{n\Gamma'(\hat{\gamma})}{\Gamma(\hat{\gamma})} + \sum_{i=1}^n \log(t_i) = 0$$

dove $\Gamma'(\hat{\gamma})$ è la derivata di $\Gamma(\hat{\gamma})$,

$$\Gamma'(\hat{\gamma}) = \int_0^\infty x^{\gamma-1} \log(x) e^{-x} dx$$

Dalla prima equazione si ottiene

$$\hat{\lambda} = \frac{n\hat{Y}}{\sum_{i=1}^n t_i}$$

e sostituendolo nella seconda equazione da risolvere in funzione di $\hat{Y}$:

$$\frac{\Gamma'(\hat{Y})}{\Gamma(\hat{Y})} - \log \hat{Y} - \log \frac{(\prod_{i=1}^n t_i)^{\frac{1}{n}}}{\sum_{i=1}^n \frac{t_i}{n}} = 0$$

Le tavole di Greenwood e Durand ci permettono di trovare $\hat{Y}$ in funzione di R , dove quest'ultimo rappresenta il rapporto della media geometrica e di quella aritmetica delle n osservazioni:

$$R = \frac{(\prod_{i=1}^n t_i)^{\frac{1}{n}}}{\sum_{i=1}^n \frac{t_i}{n}}$$

È dimostrato che la relazione tra $\hat{Y}$ e $\frac{1}{(1-R)}$ è di tipo lineare ed i rispettivi valori sono tabulati nell'**allegato D**. Per i valori che non sono riportati nella tabella, è possibile usare l'interpolazione lineare. Determinato $\hat{Y}$, $\hat{\lambda}$ può essere calcolato di conseguenza.

- In presenza di osservazioni censurate singolarmente, posto che $\eta = \lambda t_{(r)}$, $P = [\prod_{i=1}^r t_{(i)}]^{1/r} / t_{(r)}$ e $S = \sum_{i=1}^r \frac{t_{(i)}}{rt_{(r)}}$. Gli stimatori MLE di η e di Y , possono essere ottenute risolvendo:

$$\log P = \frac{n\Gamma'(\hat{Y})}{r\Gamma(\hat{Y})} - \frac{n}{r} \log \eta - \left(\frac{n}{r} - 1\right) \frac{J'(Y, \eta)}{J(Y, \eta)}$$

e

$$S = \frac{Y}{\eta} - \frac{1}{\eta} \left(\frac{n}{r} - 1\right) \frac{e^{-\eta}}{J(Y, \eta)}$$

dove

$$J(Y, \eta) = \int_1^{\infty} t^{Y-1} e^{-\eta t} dt$$

e

$$J'(Y, \eta) = \frac{\partial}{\partial Y} J(Y, \eta) = \int_1^{\infty} t^{Y-1} \log t e^{-\eta t} dt$$

In allegato E vi è una griglia di valori di P, S e $\frac{n}{r}$ per poter calcolare $\hat{Y}$ e $\hat{\mu} = \frac{\hat{Y}}{\hat{\eta}}$. Per calcolare $\hat{Y}$ e $\hat{\lambda}$ bisogna prima calcolare P e S. Per specifici valori di P, S e $\frac{n}{r}$, $\hat{Y}$ e $\hat{\mu}$ possono essere calcolati dall'allegato E. Infine, $\hat{\lambda}$ può essere calcolato tenendo conto che $\lambda = \frac{\hat{Y}}{[\hat{\mu}t(r)]}$.

- Quando i dati sono progressivamente censurati, i parametri della distribuzione gamma possono essere calcolati servendosi della procedura di Newton-Raphson, risolvendo le due equazioni seguenti:

$$\frac{nY}{\lambda} - \sum_{i=1}^r t_i - \sum_{i=r+1}^n \frac{\int_{t_i^+}^{\infty} x^Y e^{-\lambda x} dx}{\int_{t_i^+}^{\infty} x^{Y-1} e^{-\lambda x} dx} = 0$$

$$n \log \lambda - \frac{n\Gamma'(\hat{Y})}{\Gamma(\hat{Y})} + \sum_{i=1}^r \log t_i + \sum_{i=r+1}^n \frac{\int_{t_i^+}^{\infty} x^{Y-1} e^{-\lambda x} \log(x) dx}{\int_{t_i^+}^{\infty} x^{Y-1} e^{-\lambda x} dx} = 0$$

Metodi Grafici per l'Adattamento della Distribuzione di Sopravvivenza

In questa sezione introduciamo tre metodi grafici per l'adattamento della distribuzione di sopravvivenza. I metodi grafici sono stati a lungo utilizzati per la visualizzazione e l'interpretazione dei dati e spesso al posto o in combinazione con l'analisi numerica.

L'idea di base è vedere se il tempo di sopravvivenza, o una sua funzione, ha una relazione lineare con la funzione di distribuzione e la funzione di rischio cumulativo di una data distribuzione parametrica.

Se esiste una tale relazione lineare, è dimostrata graficamente da una linea retta. Quindi, se si sceglie la distribuzione appropriata e si crea un grafico di probabilità o rischio, il risultato sarà una linea retta che si adatta ai dati. I parametri della distribuzione scelta possono essere stimati direttamente dai grafici di probabilità o di rischio. Tuttavia, le informazioni spesso non sono sufficienti per scegliere una distribuzione adeguata e il grafico potrebbe non essere una linea retta. In questo caso, non è necessario stimare i parametri e può essere selezionata una distribuzione alternativa. Invece, se si utilizza il metodo dei Residui di Cox-Snell, è necessario ottenere prima le stime dei parametri.

Un grafico non lineare può fornire diverse informazioni sui dati: in primo luogo, potrebbe essere stata utilizzata la distribuzione teorica sbagliata; in secondo luogo, il campione potrebbe provenire da popolazioni differenti ed è pertanto necessario separare i dati, di conseguenza, e creare un grafico separato per ciascuna popolazione.

La presenza di qualche borderline potrebbe essere il risultato di errori nella raccolta e registrazione dei dati o l'appartenenza ad altre popolazioni.

I vantaggi dei metodi grafici possono essere riassunti come segue:

- Sono veloci e semplici da usare, in contrasto con i metodi numerici. L'accuratezza dei metodi numerici di solito non è abbastanza grande in pratica da giustificare lo sforzo richiesto.
- I grafici di probabilità e di rischio forniscono una stima approssimativa dei parametri della distribuzione attraverso mezzi grafici.
- Consentono di valutare se una particolare distribuzione teorica fornisce un adattamento adeguato ai dati.
- Un grafico fornisce una rappresentazione visiva dei dati facile da comprendere. Questo è utile per presentare i dati a terze persone, poiché un grafico consente di valutare le conclusioni tratte dai dati da mezzi grafici o numerici.

Probability Plotting

Consideriamo il seguente esempio. Dato il numero di globuli bianchi (WBC) di 23 pazienti con leucemia pediatrica riportati nella tabella 12, che vanno da 8.000 a 120.000, una distribuzione cumulativa campionaria è costruita ordinando i dati dal più piccolo al più grande. È quindi possibile creare una curva di distribuzione cumulativa del campione tracciando ciascun valore WBC rispetto alla percentuale del campione uguale o inferiore a tale valore. Cioè, il valore i -esimo dei dati ordinati in un campione di n valori viene tracciato contro la percentuale $100i/n$.

Tabella 12. Dati di WBC ordinati e Distribuzione Cumulativa Campionaria

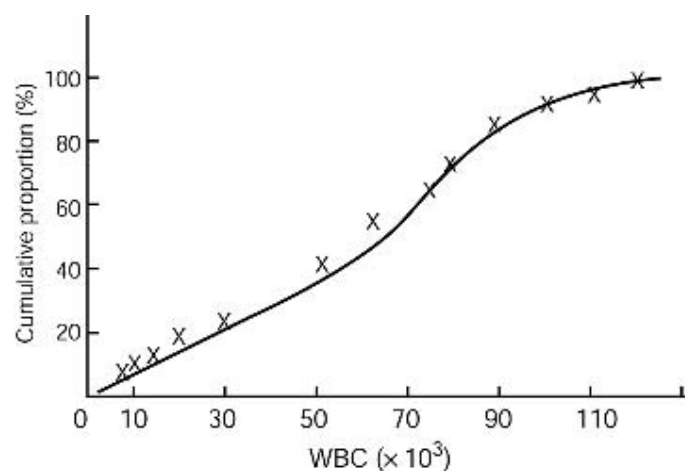
WBC (10 ³)	Order, <i>i</i>	Sample Distribution		$\Phi^{-1}(F)^a$
		$i/23$	$(i - 0.5)/23$	
8	1			
8	2	0.087	0.065	-1.512
10	3	0.130	0.109	-1.233
15	4	0.174	0.152	-1.027
20	5	0.217	0.196	-0.857
30	6	0.261	0.239	-0.709
50	7			
50	8			
50	9			
50	10			
50	11	0.478	0.457	-0.109
60	12			
60	13	0.565	0.543	0.109
75	14			
75	15	0.652	0.630	0.333
80	16			
80	17	0.739	0.717	0.575
90	18			
90	19			
90	20	0.870	0.848	1.027
100	21	0.913	0.891	1.233
110	22	0.957	0.935	1.512
120	23	1.000	0.978	2.019

^a $\Phi^{-1}(\cdot)$ denotes the inverse of the standard normal distribution function.

Si noti che per le osservazioni legate, si traccia la distribuzione campionaria solo per quella con il valore i più grande; ciò fornisce una stima conservativa della funzione di sopravvivenza. Ad esempio, il terzo valore di WBC, 10, viene tracciato contro una percentuale di $100 * 3/23 = 13\%$.

Un grafico della funzione di distribuzione cumulativa per le popolazioni più grandi contiene molti valori ravvicinati e può essere ben approssimato da una curva tracciata attraverso i punti. Al contrario, una funzione di distribuzione cumulativa campionaria ha un numero relativamente piccolo di punti e pertanto si cerca di ottenere ad occhio il migliore adattamento possibile, come si vede nella figura 11.

Figura 11. Curva di distribuzione cumulativa del campione



È una stima della funzione di distribuzione cumulativa della popolazione e viene utilizzata per ottenere ulteriori stime ed informazioni sulla popolazione. Graficamente è possibile effettuare una stima della mediana, che nell'esempio è circa 65.000 ma anche, ad esempio, una stima per il 25° percentile (40.000); per ottenere queste stime è sufficiente individuare sull'asse delle ordinate il valore che si intende valutare e muoversi orizzontalmente fino a toccare la curva, poi, a partire dal punto di intersezione, ci si muove in maniera verticale fino a leggere sull'asse delle x il risultato.

In modo analogo si può ottenere una stima della proporzione della popolazione che ha un WBC al di sotto di un valore specifico. Ad esempio, per trovare la proporzione della popolazione con un WBC di 10.000 o meno, si inserisce sull'asse orizzontale il valore dato, 10, ci si muove verticalmente fino alla linea adattata ai dati, quindi

orizzontalmente alla scala di probabilità, dove si legge la stima della proporzione della popolazione, 8%.

Come accennato in precedenza, una curva liscia può essere adattata ad occhio ad una funzione di distribuzione cumulativa campionaria per ottenere una stima della funzione di distribuzione della popolazione. Inoltre, è possibile adattare i dati con una funzione di distribuzione cumulativa teorica utilizzando un grafico di probabilità e poi utilizzare questo grafico per stimare i parametri nella funzione di distribuzione cumulativa teorica. La distribuzione può essere normale, log-normale, esponenziale, Weibull, gamma o log-logistica. Per creare un diagramma delle probabilità, si usa generalmente $(i - 0.5)/n$ oppure $i/(n + 1)$ per stimare la funzione di distribuzione cumulativa campionaria al valore i -esimo ordinato delle n osservazioni nel campione.

Il diagramma delle probabilità è costruito in modo tale che se la distribuzione teorica è adeguata per i dati, il grafico di una funzione di t (usata come asse y) rispetto a una funzione della funzione di distribuzione cumulativa campionaria (usata come asse x) sarà vicino a una linea retta. I parametri della distribuzione teorica possono quindi essere stimati da una linea fitta.

L'intera procedura può essere sintetizzata in questo modo:

STEP 1. Selezionare una distribuzione teorica per il tempo di sopravvivenza t .

STEP 2. La funzione di distribuzione cumulativa campionaria viene stimata utilizzando $\frac{i-0,5}{n}$ oppure $\frac{i}{n+1}$, $i = 1, 2, \dots, n$, per l' i -esimo valore t ordinato. Per le osservazioni che hanno lo stesso valore, la funzione di distribuzione cumulativa campionaria viene tracciata solo rispetto al t con il valore i più grande.

STEP 3. Tracciare t o una sua funzione rispetto alla distribuzione cumulativa campionaria stimata o una sua funzione.

STEP 4. Adattare una linea retta attraverso i punti ad occhio. La posizione della linea retta dovrebbe essere scelta per fornire un adattamento alla massa dei dati e può ignorare valori anomali o punti dati di dubbia validità.

Figura 12. Grafico di probabilità normale del numero di WBC

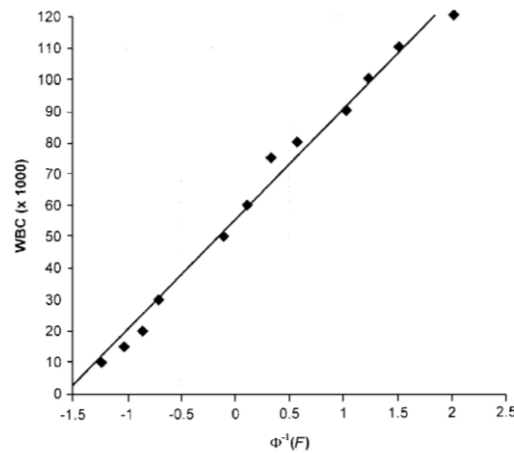


Figure 8.4 Normal probability plot of the WBC data in Example 8.1.

La figura 12 fornisce un normale diagramma di probabilità del WBC rispetto a $\varphi^{-1}(F)$, dove $\varphi^{-1}(\cdot)$ è l'inverso della funzione di distribuzione normale standard, i cui valori sono riportati nella tabella 12. La trama è ragionevolmente lineare, la distribuzione teorica scelta è quella corretta.

Hazard Plotting

Il grafico della funzione di rischio è analogo al grafico della probabilità, la differenza principale è che il tempo di sopravvivenza (o una sua funzione) è tracciato contro la funzione di rischio cumulativo (o una sua funzione) piuttosto che la funzione di distribuzione. Il grafico dei rischi è progettato per gestire i dati censurati.

Analogamente al caso precedente, le stime dei parametri della distribuzione possono essere determinate dal grafico dei rischi con uno sforzo computazionale minimo.

Per determinare se un insieme di tempi di sopravvivenza con osservazione censurate proviene da una data distribuzione teorica, costruiamo un diagramma di rischio tracciando il tempo di sopravvivenza (o una sua funzione) rispetto a una stima del rischio cumulativo (o una sua funzione). La funzione di rischio cumulativo può essere stimata seguendo i passaggi seguenti.

STEP 1. Ordinare le n osservazioni nel campione dalle più piccole alle più grandi, indipendentemente dal fatto che siano censurate. Se alcune osservazioni non censurate e censurate hanno lo stesso valore, devono essere elencate in ordine casuale. Nell'elenco dei valori ordinati, i dati censurati sono contrassegnati ciascuno da un segno più.

STEP 2. Numerare le osservazioni ordinate in ordine inverso, con n assegnato al valore di dati più piccolo, $n - 1$ al secondo più piccolo e così via. I numeri così ottenuti sono chiamati valori K o numeri di ordine inverso. Per l'osservazione non censurata, K è il numero di soggetti ancora a rischio in quel momento.

STEP 3. Ottenere il valore di rischio corrispondente per ogni osservazione non censurata. Le osservazioni censurate non hanno un valore di pericolo. Il valore di rischio per un'osservazione non censurata è $1/K$. Questa è la frazione degli individui K che sono sopravvissuti per quel periodo di tempo e poi hanno fallito. È una probabilità di fallimento condizionale per un'osservazione non censurata.

STEP 4. Per ogni osservazione non censurata, calcolare il valore di rischio cumulativo. Si tratta della somma dei valori di rischio dell'osservazione non censurata e di tutte le precedenti osservazioni non censurate. Per le osservazioni non censurate legate, il rischio cumulativo viene valutato solo per il più piccolo K tra le osservazioni non censurate.

Un esempio numerico è riportato nella tabella 13.

Tabella 13. Stima della funzione di rischio cumulativa

t	Reversed Order, K	Hazard, $1/K$	Cumulative Hazard, $\hat{H}(t)$
6	21	0.048	
6+	20		
6	19	0.053	
6	18	0.056	0.156
7	17	0.059	0.215
9+	16		
10	15	0.067	0.281
10+	14		
11+	13		
13	12	0.083	0.365
16	11	0.091	0.456
17+	10		
19+	9		
20+	8		
22	7	0.143	0.598
23	6	0.167	0.765
25+	5		
32+	4		
32+	3		
34+	2		
35+	1		

Metodo Dei Residui Di Cox-Snell

Il metodo dei Residui di Cox—Snell può essere applicato a qualsiasi modello parametrico. Il valore del residuo r_i per l' i -esimo individuo con tempo di sopravvivenza osservato t_i , non censurato o censurato, è definito come:

$$r_i = -\log \hat{S}(t_i) \quad i = 1, 2, \dots, n \quad (1.75)$$

dove $\hat{S}(t)$ è la funzione di sopravvivenza stimata basata sulla stima MLE dei parametri. Se la t_i osservata è censurata, anche la r_i corrispondente sarà censurata.

Poiché la funzione di rischio cumulativo è pari a $H(t) = -\log S(t)$, il residuo di Cox—Snell r_i è un valore di rischio cumulato stimato a t_i . L'importante proprietà del residuo di Cox-Snell è che se il modello selezionato si adatta ai dati, r_i segue la distribuzione esponenziale unitaria con funzione di densità $f_r(r) = e^{-r}$.

Sia $S_R(r)$ la funzione di sopravvivenza dei residui r_i ,

$$S_R = \int_r^\infty f_R(x) dx = \int_r^\infty e^{-x} dx = e^{-r}$$

e

$$-\log S_R(r) = -\log(e^{-r}) = r \quad (1.76)$$

Sia $\hat{S}_R(r)$ la stima di Kaplan-Meier di $S_R(r)$. Il grafico di r_i rispetto a $-\log \hat{S}_R(r_i)$ dovrebbe essere una linea retta con pendenza unitaria e intercetta zero se la distribuzione di sopravvivenza adattata è appropriata, indipendentemente dalla forma della distribuzione.

La procedura per l'utilizzo dei residui di Cox-Snell può essere riassunta come segue.

STEP 1. Trovare gli stimatori MLE dei parametri della distribuzione teorica selezionata.

STEP 2. Calcolare i residui di Cox-Snell $r_i = -\log \hat{S}(t_i)$, $i = 1, 2, \dots, n$, dove $\hat{S}(t_i)$ è la funzione di sopravvivenza stimata con i parametri MLE.

STEP 3. Applicare il metodo Kaplan-Meier per stimare la funzione di sopravvivenza $S_R(r)$ dei residui di Cox-Snell r_i , ottenuti nello step 2, quindi utilizzando la stima $\hat{S}_R(r)$, calcolare il $-\log \hat{S}_R(r_i)$.

STEP 4. Tracciare il grafico di r_i rispetto a $-\log \hat{S}_R(r_i)$, $i = 1, 2, \dots, n$. Se il grafico è una linea retta con pendenza unitaria e intercetta zero, la distribuzione scelta è appropriata.

Se un tempo di sopravvivenza individuale è censurato a destra, t_i^+ , e il modello adattato è corretto, il corrispondente log residuo di Cox—Snell $H(t_i^+) = -\log S(t_i^+)$ è minore del residuo valutato in un'osservazione non censurata con lo stesso valore t_i , poiché $H(t)$ è una funzione monotona crescente di t . Per tener conto di ciò, sono stati proposti due residui di Cox-Snell modificati per le osservazioni censurate: uno si basa sulla media e l'altro si basa sulla mediana ($= \log 2 = 0.693$) della distribuzione esponenziale unitaria assumendo che anche la differenza tra $H(t_i)$ e $H(t_i^+)$ segua la distribuzione esponenziale unitaria.

Per un'osservazione censurata t_i^+ , il residuo modificato r_i^+ , è definito come:

$$r_i^+ = r_i + 1$$

oppure

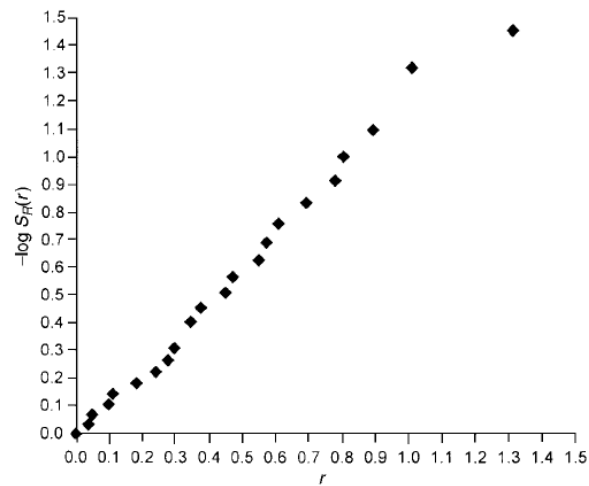
$$r_i^+ = r_i + 0.693$$

per cui vale $r_i = -\log \hat{S}(t_i)$.

La figura 13 mostra un esempio di applicazione di questo metodo; il grafico ottenuto è molto vicino ad una retta con pendenza unitaria e con intercetta pari a zero. Pertanto,

la distribuzione teorica scelta è appropriata per il campione che si sta analizzando e il risultato ottenuto è consistente.

Figura 13. Esempio di applicazione del metodo dei Residui di Cox-Snell



Un punto debole del metodo residuo di Cox-Snell è che il grafico non indica, però, il tipo di deviazione che i dati hanno dal modello selezionato se la configurazione non è lineare.

Metodi Parametrici Per L'adattamento Del Modello Di Regressione

L'identificazione Dei Fattori Prognostici

La previsione del futuro di un singolo paziente affetto da una ben precisa malattia gioca un ruolo importante nella pratica medica. Prima che un medico possa fare una prognosi e decidere sul trattamento, sono spesso necessari un'anamnesi, nonché dati patologici, clinici e di laboratorio. Pertanto, molte cartelle cliniche contengono un gran numero di caratteristiche del paziente, chiamate variabili indipendenti o **covariate**, ed è spesso difficile distinguere quali sono più strettamente correlate alla prognosi. Il medico può decidere quali caratteristiche sono irrilevanti, ma di solito è necessaria un'analisi statistica per preparare un riassunto compatto dei dati che possono rivelare la loro relazione. Un modo per raggiungere questo scopo è cercare un modello teorico che si adatti ai dati osservati e identifichi i fattori più importanti. Questi modelli, solitamente modelli di regressione, estendono i metodi discussi nelle sezioni precedenti per includere le covariate. Ci concentriamo sui modelli di regressione parametrica, cioè, assumiamo che il tempo di sopravvivenza segua una distribuzione teorica. Se si può assumere un modello appropriato, si può stimare la probabilità di sopravvivere in un dato momento quando le covariate sono incorporate.

La variabile dipendente, chiamata anche **variabile di risposta** o risultato della previsione, può essere dicotomica, policotomica o continua.

Esempi di variabili dipendenti **dicotomiche** sono risposta o non risposta, vita o morte e presenza o assenza di una data malattia. Le variabili dipendenti **policotomiche** includono diversi gradi di sintomi (ad esempio, nessuna evidenza di malattia, sintomo minore, sintomo principale) e punteggi di reazioni psichiatriche (ad esempio, sentirsi bene, tollerabile, depresso o molto depresso). Le variabili dipendenti **continue** possono

essere la durata della sopravvivenza dall'inizio del trattamento o la durata della remissione, entrambe misurate su una scala numerica di un intervallo continuo di valori. In questa analisi ci focalizziamo su variabili dipendenti continue, come il tempo di sopravvivenza e la durata della remissione.

La variabile **prognostica**, variabile indipendente, può essere numerica o non numerica. Nel primo caso può essere discreta o continua, nel secondo caso può essere non ordinate (ad es. razza o diagnosi) o ordinate (ad es., la gravità della malattia). Possono anche essere dicotomiche (ad esempio, un fegato è o non è ingrossato). Di solito, la raccolta di variabili prognostiche ne include alcune di ogni tipo.

Prima di eseguire un calcolo statistico, i dati devono essere esaminati accuratamente. I coefficienti di correlazione tra le variabili possono essere calcolati per rilevare variabili significativamente correlate. Nell'eliminare qualsiasi variabile altamente correlata, devono essere incorporate le informazioni provenienti da altri studi. Se altri studi mostrano che una data variabile ha un valore prognostico, dovrebbe essere conservata.

È possibile inserire nello studio anche variabili di tipo qualitativo, ad esempio se è necessario specificare il sesso di un paziente è possibile creare delle variabili fittizie ($x=1$ se il paziente è maschio, $x=0$ se femmina).

Nella pratica, c'è un numero molto elevato di possibili fattori prognostici associati agli esiti. Un modo per ridurre il numero di fattori consiste nell'esaminare la relazione tra ogni singolo fattore e la variabile dipendente (ad esempio, il tempo di sopravvivenza). Dall'analisi univariata, i fattori che hanno poco o nessun effetto sulla variabile dipendente possono essere esclusi dall'analisi multivariata.

Un problema frequente nell'analisi di regressione è la mancanza di dati. Questi possono essere distinti in:

- **Variabili dipendenti rispetto a variabili indipendenti:** se il valore della variabile dipendente (es. il tempo di sopravvivenza) è sconosciuto, non resta che eliminare quell'individuo dall'analisi e ridurre la dimensione del campione.
- **Molti dati mancanti rispetto a pochi:** il problema dei dati mancanti è di diversa entità a seconda di quanto grande sia la proporzione di dati mancanti, sia per la variabile dipendente che per le variabili indipendenti. Questo problema è ovviamente meno critico se manca l'1% dei dati per una variabile indipendente rispetto a quando manca il 40% dei dati per più variabili indipendenti. Quando una parte sostanziale di soggetti ha dati mancanti per una variabile, possiamo semplicemente scegliere di eliminarli ed eseguire l'analisi sul resto del campione.
- **Perdita di dati casuale o non casuale:** se i dati mancanti si verificano in un'ampia percentuale di persone e la dimensione del campione non è particolarmente grande, si può sollevare una questione di casualità. Se le persone con dati mancanti non mostrano differenze significative nella variabile dipendente, il problema non è serio. Se i dati non mancano in modo casuale, i risultati ottenuti possono essere fuorvianti. Pertanto, eliminare i casi non è sempre una soluzione adeguata al problema dei dati mancanti. Se la variabile indipendente viene misurata su una scala nominale o categoriale, un metodo alternativo consiste nel trattare gli individui di un gruppo con informazioni mancanti come un altro gruppo. Per le variabili misurate quantitativamente (ad es. età), la media dei valori disponibili può essere utilizzata per un valore mancante. Questo principio può essere applicato anche ai dati nominali. Ciò non significa che la media sia una buona stima del valore mancante, ma fornisce comodità per l'analisi.

Struttura Generale Dei Modelli Di Regressione Parametrica

Quando si considerano le covariate, si assume che il tempo di sopravvivenza, o una sua funzione, abbia una relazione esplicita con le covariate stesse. Inoltre, quando si considera un modello parametrico, si assume che il tempo di sopravvivenza (o una sua funzione) segua una data distribuzione teorica (o modello) e abbia una relazione esplicita con le variabili indipendenti.

A titolo di esempio, consideriamo la distribuzione di Weibull.

Siano $\mathbf{x} = (x_1, \dots, x_p)$ le p covariate prese in esame. Se il parametro nella distribuzione di Weibull λ è correlato a $\mathbf{x}$ come segue:

$$\lambda = e^{-(a_0 + \sum_{i=1}^p a_i x_i)} = \exp[-(a_0 + \mathbf{a}'\mathbf{x})]$$

dove il vettore $\mathbf{a} = (a_1, \dots, a_p)$ rappresenta l'insieme dei coefficienti di $\mathbf{x}$. La funzione di rischio della distribuzione di Weibull (1.50) può essere estesa per comprendere anche le covariate come:

$$h(t, \mathbf{x}) = \lambda^Y \Upsilon t^{Y-1} = \Upsilon t^{Y-1} e^{-(a_0 + \sum_{i=1}^p a_i x_i)Y} = \Upsilon t^{Y-1} \exp[-(a_0 + \mathbf{a}'\mathbf{x})Y] \quad (1.77)$$

La funzione di sopravvivenza (1.49) diventa:

$$S(t, \mathbf{x}) = (e^{-t^Y})^{\exp(-Y(a_0 + \mathbf{a}'\mathbf{x}))} \quad (1.78)$$

Introduciamo, a questo punto, un altro tipo di modello chiamato **modello del tempo di guasto accelerato (AFT)**.

Analogamente ai metodi di regressione convenzionali, anche il tempo di sopravvivenza può essere analizzato utilizzando il modello del tempo di guasto accelerato (AFT). Il

modello AFT per il tempo di sopravvivenza presuppone che la relazione tra il logaritmo del tempo di sopravvivenza T e le covariate sia lineare e che possa essere scritta come:

$$\log T = a_0 + \sum_{j=1}^p a_j x_j + \sigma \varepsilon \quad (1.79)$$

dove $x_j, j = 1, \dots, p$, rappresenta le covariate, $a_j, j = 1, \dots, p$, i coefficienti, $\sigma, (\sigma > 0)$, è un parametro di scala sconosciuto, e, infine, ε , il termine di errore, è una variabile casuale con funzione di densità $g(\varepsilon, \mathbf{d})$ e funzione di sopravvivenza $G(\varepsilon, \mathbf{d})$ note ma con parametro $\mathbf{d}$ sconosciuto. Questo vuol dire che la sopravvivenza è dipendente sia dalle covariate che dalla funzione sottostante g .

Consideriamo il caso semplice in cui c'è solo una covariata x con valore 0 e 1. Allora la (1.79) diventa:

$$\log T = a_0 + a_1 x + \sigma \varepsilon$$

Siano T_0 e T_1 i tempi di sopravvivenza per due individui con $x = 0$ e $x = 1$, rispettivamente, tale per cui $T_0 = \exp(a_0 + \sigma \varepsilon)$ e $T_1 = \exp(a_0 + a_1 + \sigma \varepsilon) = T_0 \exp(a_1)$.

Pertanto, $T_1 > T_0$ se $a_1 > 0$ e $T_1 < T_0$ se $a_1 < 0$.

Ciò significa che la covariata x "accelera" o "decelera" il tempo di sopravvivenza o il tempo al guasto, da cui il nome del modello.

Di seguito discutiamo la forma generale della funzione di verosimiglianza dei modelli AFT, le procedure di stima dei parametri di regressione $(a_0, \mathbf{a}, \sigma, \mathbf{d})$ della (1.79) e i test di significatività delle covariate sulla sopravvivenza tempo.

Siano $t_1, t_2, \dots, t_n$ i tempi di sopravvivenza osservati da n individui, inclusivi di osservazioni esatte, censurate a sinistra, a destra e per intervalli. Supponiamo che il logaritmo del tempo di sopravvivenza possa essere modellato dalla (1.79) e che siano $\mathbf{a}' = (a_1, \dots, a_p)$, e $\mathbf{b}' = (\mathbf{a}', a_0, \sigma, \mathbf{d}')$.

La funzione di log-verosimiglianza in termini di funzione di densità $g(\varepsilon)$ e funzione di sopravvivenza $G(\varepsilon)$ di ε è:

$$l(\mathbf{b}) = \log L(\mathbf{b}) = \sum \log[g(\varepsilon_i)] + \sum \log[G(\varepsilon_i)] + \sum \log[1 - G(\varepsilon_i)] + \sum \log[G(\eta_i) - G(\varepsilon_i)] \quad (1.80)$$

dove

$$\varepsilon_i = \frac{\log t_i - a_0 - \sum_{j=1}^p a_j x_{ji}}{\sigma}$$

$$\eta_i = \frac{\log v_i - a_0 - \sum_{j=1}^p a_j x_{ji}}{\sigma}$$

Il primo termine nella funzione di verosimiglianza riguarda le osservazioni non censurate, il secondo termine osservazioni censurate a destra, il terzo termine osservazioni censurate a sinistra e l'ultimo termine su osservazioni censurate a intervalli.

Alternativamente,

$$\mu_i = a_0 + \sum_{j=1}^p a_j x_j, \quad i = 1, \dots, n \quad (1.81)$$

Quindi la (1.79) diventa:

$$\log T = \mu + \sigma \varepsilon \quad (1.82)$$

La rispettiva funzione di log-verosimiglianza in termini di funzione di densità $f(t, \mathbf{b})$ e funzione di sopravvivenza $S(t, \mathbf{b})$ di T è:

$$\begin{aligned}
 l(\mathbf{b}) &= \log L(\mathbf{b}) \\
 &= \sum \log[f(t_i, \mathbf{b})] \\
 &+ \sum \log[S(t_i, \mathbf{b})] \\
 &+ \sum \log[1 - S(t_i, \mathbf{b})] + \sum \log[S(v_i, \mathbf{b}) - S(t_i, \mathbf{b})]
 \end{aligned} \tag{1.83}$$

tenendo conto che:

$$f(t, \mathbf{b}) = \frac{g\left(\frac{\log t - \mu}{\sigma}\right)}{\sigma t}$$

e che $S(t, \mathbf{b})$ è la corrispondente funzione di sopravvivenza. Il vettore $\mathbf{b}$ contiene i coefficienti di regressione e altri parametri della distribuzione.

Entrambi i metodi usati per calcolare la funzione di verosimiglianza possono essere usati per calcolare gli stimatori MLE dei parametri del modello. Per una data funzione $l(\mathbf{b})$, lo stimatore MLE $\hat{\mathbf{b}}$ è una soluzione della seguente equazione:

$$\frac{\partial(l(\mathbf{b}))}{\partial b_i} = 0, \text{ per ogni } i$$

Di solito, non c'è una soluzione chiusa e bisogna applicare la procedura iterativa Newton-Raphson per ottenere $\hat{\mathbf{b}}$. Sostituendo i parametri $\mathbf{b}$ con il suo rispettivo stimatore in $S(t_i, \mathbf{b})$, abbiamo una funzione di sopravvivenza stimata $S(t_i, \hat{\mathbf{b}})$, che tiene conto delle covariate.

A partire da questo modello, è possibile eseguire tutti i test di ipotesi che si ritiene necessario e costruire gli intervalli di confidenza. Quella appena descritta è la regola generale che può essere adattata alle diverse distribuzioni teoriche, secondo le regole definite dalla tabella 14.

Tabella 14. Regressione Parametrica e Relative Metriche Applicabili alle varie Distribuzioni

Distribution	Metric	Survivor function	Parameterization	Ancillary parameters
Exponential	PH	$\exp(-\lambda_j t_j)$	$\lambda_j = \exp(\mathbf{x}_j \boldsymbol{\beta})$	
Exponential	AFT	$\exp(-\lambda_j t_j)$	$\lambda_j = \exp(-\mathbf{x}_j \boldsymbol{\beta})$	
Weibull	PH	$\exp(-\lambda_j t_j^p)$	$\lambda_j = \exp(\mathbf{x}_j \boldsymbol{\beta})$	p
Weibull	AFT	$\exp(-\lambda_j t_j^p)$	$\lambda_j = \exp(-p \mathbf{x}_j \boldsymbol{\beta})$	p
Gompertz	PH	$\exp\{-\lambda_j \gamma^{-1}(e^{\gamma t_j} - 1)\}$	$\lambda_j = \exp(\mathbf{x}_j \boldsymbol{\beta})$	γ
Lognormal	AFT	$1 - \Phi\left\{\frac{\log(t_j) - \mu_j}{\sigma}\right\}$	$\mu_j = \mathbf{x}_j \boldsymbol{\beta}$	σ
Loglogistic	AFT	$\{1 + (\lambda_j t_j)^{1/\gamma}\}^{-1}$	$\lambda_j = \exp(-\mathbf{x}_j \boldsymbol{\beta})$	γ
Generalized gamma				
if $\kappa > 0$	AFT	$1 - I(\gamma, u)$	$\mu_j = \mathbf{x}_j \boldsymbol{\beta}$	σ, κ
if $\kappa = 0$	AFT	$1 - \Phi(z)$	$\mu_j = \mathbf{x}_j \boldsymbol{\beta}$	σ, κ
if $\kappa < 0$	AFT	$I(\gamma, u)$	$\mu_j = \mathbf{x}_j \boldsymbol{\beta}$	σ, κ

I residui di Cox- Snell in presenza delle Covariate

I modelli AFT assumono una relazione lineare tra $\log T$ e le covariate:

$$\log T_i = a_0 + \sum_{k=1}^p a_k x_{ki} + \sigma \varepsilon_i = \mu_i + \sigma \varepsilon_i \quad (1.84)$$

dove ε_i ha funzione di sopravvivenza $G(\varepsilon)$. Una volta selezionati un modello parametrico specificato e un sottoinsieme di covariate, per valutare la bontà di adattamento di questo modello, un approccio consiste nel calcolare i residui di regressione:

$$\hat{\varepsilon}_i = \frac{\log t_i - \hat{\mu}_i}{\hat{\sigma}} \quad i = 1, 2, \dots, n \quad (1.85)$$

considerando che $\hat{\mu}_i = \hat{a}_0 + \sum_{k=1}^p \hat{a}_k x_{ki}$ e che $\hat{a}_i$ e $\hat{\sigma}$ sono gli stimatori MLE di a_i e σ rispettivamente e t_i sono i tempi di sopravvivenza osservati. $\hat{\varepsilon}_i$ è censurato se il t_i corrispondente è censurato.

Se il modello adattato è corretto, la funzione di sopravvivenza corrispondente $G(\varepsilon)$ è la funzione di sopravvivenza del modello adattato. Ad esempio, se T segue effettivamente il modello di regressione log-logistica con un sottoinsieme selezionato di covariate, i corrispondenti $\hat{\varepsilon}_i$ dovrebbero seguire la distribuzione log-logistica.

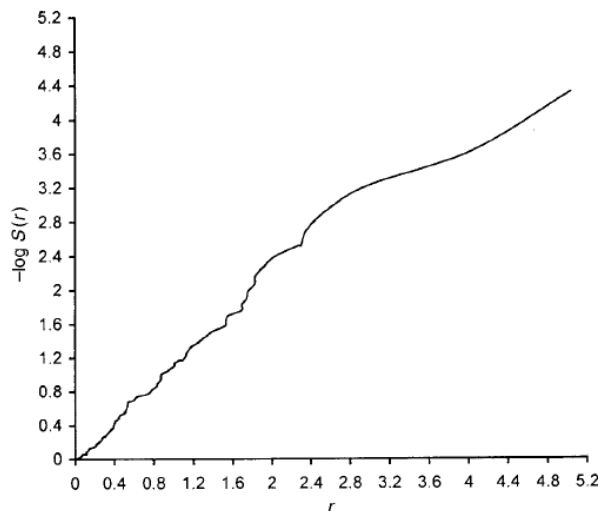
Inoltre, se il modello adattato è corretto, i residui di Cox—Snell definiti in (1.75) sono definiti come:

$$r_i = -\log G(\hat{\varepsilon}_i; \hat{\mathbf{d}}) = -\log G\left(\frac{\log t_i - \hat{\mu}_i}{\hat{\sigma}}, \hat{\mathbf{d}}\right); i = 1, 2, \dots, n \quad (1.86)$$

dove $\hat{\mathbf{d}}$ è lo stimatore MLE dei parametri della distribuzione.

Sia $\hat{\mathbf{S}}(r)$ la funzione di sopravvivenza stimata di r_i . Il grafico di r_i rispetto $-\log \hat{\mathbf{S}}(r_i)$, $i = 1, 2, \dots, n$, dovrebbe essere una linea retta con pendenza unitaria e intercetta zero se il modello adattato per il tempo di sopravvivenza T è corretto. Questo metodo grafico può essere utilizzato per valutare la bontà di adattamento del modello di regressione parametrica.

Figura 14. Grafico dei residui di Cox-Snell per l'adattamento del modello di Weibull



Cox Proportional Hazard Model

I metodi di sopravvivenza parametrica sono molto potenti per l'adattamento del modello e per l'identificazione di fattori prognostici significativi se è nota la distribuzione di sopravvivenza sottostante. Tuttavia, nella pratica, la forma esatta della distribuzione di sopravvivenza sottostante è solitamente sconosciuta e potremmo non essere in grado di trovare un modello appropriato. Pertanto, l'uso di metodi parametrici nell'identificazione di fattori prognostici significativi è alquanto limitato. In questo capitolo discutiamo di un modello più comunemente usato: il modello dei rischi proporzionali di Cox (1972) e la relativa inferenza statistica. Questo modello non richiede la conoscenza della distribuzione sottostante e la funzione di rischio può assumere qualsiasi forma, inclusa quella di una funzione a gradino, ma si presume che le funzioni di rischio di individui diversi siano proporzionali e indipendenti dal tempo. La normale funzione di verosimiglianza viene sostituita dalla funzione di verosimiglianza parziale anche se l'inferenza statistica basata su quest'ultima è molto simile a quella basata sulla funzione di verosimiglianza.

Funzione Di Verosimiglianza Parziale Per I Tempi Di Sopravvivenza

Il modello dei rischi proporzionali di Cox possiede una proprietà tale per cui individui diversi hanno funzioni di rischio proporzionali, cioè $[h(t|x_1)/h(t|x_2)]$, ossia, il rapporto delle funzioni di rischio per due individui con covariate $x_1 = (x_{11}, x_{21}, \dots, x_{p1})'$ e $x_2 = (x_{12}, x_{22}, \dots, x_{p2})'$ è una costante (non varia con il tempo t). Ciò significa che il rapporto tra il rischio di morte di due individui è lo stesso, non importa quanto tempo sopravvivano. I modelli di regressione esponenziale e di Weibull possiedono questa stessa proprietà.

Ciò implica che la funzione di rischio, dato un insieme di covariate $\mathbf{x} = (x_1, x_2, \dots, x_p)'$, può essere riscritta come una funzione di una funzione di rischio sottostante e una funzione, diciamo $g(x_1, x_2, \dots, x_p)$, di solo le covariate, cioè:

$$h(t|x_1, x_2, \dots, x_p) = h_0(t)g(x_1, x_2, \dots, x_p)$$

o in maniera equivalente:

$$h(t|\mathbf{x}) = h_0(t)g(\mathbf{x}) \quad (1.87)$$

La funzione di rischio sottostante, $h_0(t)$, rappresenta come il rischio cambia nel tempo e $g(\mathbf{x})$ rappresenta l'effetto delle covariate. $h_0(t)$ può essere interpretato come la funzione di rischio quando tutte le covariate vengono ignorate o quando $g(\mathbf{x}) = 1$, ed è anche chiamata **funzione di rischio di base**. Il rapporto di rischio di due individui con diverse covariate $\mathbf{x}_1$ e $\mathbf{x}_2$ è una costante, indipendente dal tempo:

$$\frac{h(t|\mathbf{x}_1)}{h(t|\mathbf{x}_2)} = \frac{h_0(t)g(\mathbf{x}_1)}{h_0(t)g(\mathbf{x}_2)} = \frac{g(\mathbf{x}_1)}{g(\mathbf{x}_2)} \quad (1.88)$$

Il modello di Cox assume che $g(\mathbf{x})$ è una funzione esponenziale, così caratterizzata:

$$g(\mathbf{x}) = \exp\left(\sum_{j=1}^p b_j x_j\right) = \exp(\mathbf{b}'\mathbf{x})$$

e con funzione di rischio:

$$h(t|\mathbf{x}) = h_0(t) \exp(\sum_{j=1}^p b_j x_j) = h_0(t) \exp(\mathbf{b}'\mathbf{x}) \quad (1.89)$$

e $\mathbf{b} = (b_1, \dots, b_p)$ denota i coefficienti delle covariate. Questi coefficienti possono essere stimati dai dati osservati e indicano l'entità degli effetti delle loro corrispondenti covariate.

Ad esempio, se esiste solo una covariata “trattamento” che è pari a $x = 0$ se una persona riceve il placebo, $x = 1$ se una persona riceve un farmaco sperimentale. Il rapporto di rischio del paziente che riceve il farmaco sperimentale e quello che riceve il placebo, tenendo conto di quanto appena detto, è pari a:

$$\frac{h(t|x_1 = 1)}{h(t|x_2 = 0)} = \exp(b_1)$$

Quindi, i due trattamenti hanno uguale effetto se $b_1 = 0$ e la cura sperimentale assume un più basso o un più alto rischio per la sopravvivenza se $b_1 < 0$ o $b_1 > 0$.

È possibile dimostrare che la (1.89) è equivalente a:

$$S(t|x) = [S_0(t)]^{\exp(\sum_{j=1}^p b_j x_j)} = [S_0(t)]^{\exp(b'x)} \quad (1.90)$$

Quindi le covariate possono essere incluse nella funzione di sopravvivenza. L'uso della (1.89) può essere semplificato come segue:

1. **Problema a due campioni.** Supponiamo che $p = 1$, ovvero c'è solo una covariata x_1 , che è un indicatore:

$$x_{1i} = \begin{cases} 0 & \text{se l'i-esimo individuo appartiene al gruppo 0} \\ 1 & \text{se l'i-esimo individuo appartiene al gruppo 1} \end{cases}$$

Secondo la (1.89) le funzioni di rischio per il gruppo 1 e per il gruppo 2 sono rispettivamente $h_0(t)$ e $h_1(t) = h_0(t) \exp(b_1)$. La funzione di rischio del gruppo 1 è uguale alla funzione di rischio del gruppo 0 moltiplicata per una costante $\exp(b_1)$, o, in altri termini, le due funzioni di rischio sono proporzionali. In termini di funzioni di sopravvivenza,

$$S(t) = [S_0(t)]^c$$

dove $c = \exp(b_1)$ rappresenta la costante.

2. **Problema a due campioni con le covariate.** Le covariate nella (1.89) possono essere variabili indicatrici come x_1 nel problema dei due gruppi di cui sopra. Avere una o più covariate che rappresentano fattori prognostici consente di esaminare la relazione tra due gruppi, aggiustando per la presenza di questi fattori.
3. **Problemi di Regressione.** Dividendo entrambi i membri della (1.89) per $h_0(t)$ e calcolandone il logaritmo, otteniamo:

$$\log \frac{h_1(t)}{h_0(t)} = b_1 x_{1i} + b_2 x_{2i} + \dots + b_p x_{pi} = \sum_{j=1}^p b_j x_{ji} = b' x_i \quad (1.91)$$

Dove le x sono le covariate per i i-esimi individui. La (1.91) è formata dalla funzione di rischio e dalla funzione lineare delle covariate e dai rispettivi coefficienti.

$h_0(t)$ è la funzione di rischio quanto tutte le covariate sono ignorate. Se le covariate sono standardizzate rispetto alla loro media e il modello usato è:

$$\log \frac{h_1(t)}{h_0(t)} = b_1 (x_{1i} - \bar{x}_1) + b_2 (x_{2i} - \bar{x}_2) + \dots + b_p (x_{pi} - \bar{x}_p) = b' (x_i - \bar{x}) \quad (1.92)$$

Dove $\bar{x}' = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)$ e $\bar{x}_j$ è la media delle j-esime covariate per tutti i pazienti; la (1.92) rappresenta il logaritmo del rapporto del rischio di fallimento tra un paziente con un dato insieme di valori $x'_i = (x_{1i}, x_{2i}, \dots, x_{pi})$ ed un paziente che ha un valore medio per ogni covariata.

L'interesse principale qui è identificare importanti fattori prognostici, in altre parole, si vuole identificare dall'insieme delle covariate p un sottoinsieme di variabili che influenzano in modo più significativo il rischio e, di conseguenza, la durata della sopravvivenza del paziente. Ci occupiamo dei coefficienti di regressione: se b_i è zero, la covariata corrispondente non è correlata alla sopravvivenza, ma se b_i non è zero,

questo rappresenta la grandezza dell'effetto di x_i sul rischio quando le altre covariate sono considerate simultaneamente.

Per stimare i coefficienti, $b_1, b_2, \dots, b_p$, Cox propone una parziale funzione di verosimiglianza basata su una probabilità condizionata di fallimento, assumendo che non vi siano valori legati nei tempi di sopravvivenza. Tuttavia, in pratica, i tempi di sopravvivenza legati sono comunemente osservati e la funzione di probabilità parziale di Cox è stata modificata per gestire tali legami.

Procedure di Stima in Assenza di Tempi di Sopravvivenza Legati

Supponiamo che k tempi di sopravvivenza di n individui sono esatti e distinti e $n - k$ sono censurati a destra. Siano $t_1 \leq t_2 \leq t_3 \leq \dots \leq t_k$ i tempi di fallimento esatti k in ordine crescente con le rispettive covariate $x_1, x_2, \dots, x_k$. Sia $R(t_i)$ i valori del rischio al tempo t_i . $R(t_i)$ è costituito da tutte le persone i cui tempi di sopravvivenza sono almeno pari a t_i . Per il particolare caso in cui il fallimento è pari al tempo t_i , condizionalmente a $R(t_i)$, la probabilità che il guasto sia sull'individuo osservato è:

$$\frac{\exp(\sum_{j=1}^p b_j x_{j(i)})}{\sum_{l \in R(t_i)} \exp(\sum_{j=1}^p b_j x_{jl})} = \frac{\exp(\mathbf{b}' \mathbf{x}_{(i)})}{\sum_{l \in R(t_i)} \exp(\mathbf{b}' \mathbf{x}_{(l)})}$$

Ogni guasto contribuisce con un fattore e quindi la funzione di verosimiglianza parziale è data da:

$$L(\mathbf{b}) = \prod_{i=1}^k \frac{\exp(\sum_{j=1}^p b_j x_{j(i)})}{\sum_{l \in R(t_i)} \exp(\sum_{j=1}^p b_j x_{jl})} \quad (1.93)$$

La log funzione di verosimiglianza parziale è:

$$\begin{aligned}
 l(\mathbf{b}) &= \log L(\mathbf{b}) \\
 &= \sum_{i=1}^k \sum_{j=1}^p b_j x_{ji} \\
 &\quad - \sum_{i=1}^k \log \left[\sum_{l \in R(t_i)} \exp \left(\sum_{j=1}^p b_j x_{jl} \right) \right] \\
 &= \sum_{i=1}^k \left\{ \mathbf{b}' \mathbf{x}_{(i)} - \log \left[\sum_{l \in R(t_i)} \exp(\mathbf{b}' \mathbf{x}_l) \right] \right\}
 \end{aligned} \tag{1.94}$$

Lo stimatore di massima verosimiglianza parziale (MPLE) $\hat{\mathbf{b}}$ di $\mathbf{b}$ può essere ottenuto risolvendo le seguenti equazioni simultanee, applicando la procedura iterativa di Newton-Raphson.:

$$\frac{\partial(l(\mathbf{b}))}{\partial \mathbf{b}} = 0$$

La matrice di covarianza dello stimatore MLE $\hat{\mathbf{b}}$ è così definita:

$$\hat{V}(\hat{\mathbf{b}}) = \widehat{Cov}(\hat{\mathbf{b}}) = \left[-\frac{\partial^2 l(\hat{\mathbf{b}})}{\partial \mathbf{b} \partial \mathbf{b}'} \right]^{-1}$$

e $-\frac{\partial^2 l(\hat{\mathbf{b}})}{\partial \mathbf{b} \partial \mathbf{b}'}$ prende il nome di **matrice delle informazioni osservate**; si definisce la generica coppia di valori $(i; j)$ della matrice come v_{ij} , allora l'intervallo di confidenza 100(1- α)% di $\mathbf{b}_i$,

$$\left(\hat{b}_i - Z_{\frac{\alpha}{2}} \sqrt{v_{ii}}, \hat{b}_i + Z_{\frac{\alpha}{2}} \sqrt{v_{ii}} \right) \tag{1.95}$$

Procedure di Stima in Presenza di Tempi di Sopravvivenza Legati

Supponiamo che tra gli n tempi di sopravvivenza osservati ci siano k tempi non censurati e che $m_{(i)}$ denoti il numero di persone che falliscono al tempo t_i ; $m_{(i)} > 1$ se ci sono più di una osservazione al tempo t_i . Sia $R(t_i)$ il set di dati delle persone esposte al rischio al tempo t_i e r_i il numero di quelle persone. Per ogni $R(t_i)$ possiamo selezionare in maniera casuale $m_{(i)}$ soggetti, si associa a ciascuno di essi un determinato $u_{(j)}$. Ci sono $r_i C_{m_{(i)}} = \frac{r_i!}{[m_{(i)}!(r_i - m_{(i)})!]}$ possibili $u_{(j)}$. U_i indica il set di dati comprensivo di tutti gli $u_{(j)}$.

Per tenere conto delle osservazioni legate, si indicano con $x_k = (x_{1k}, x_{2k}, \dots, x_{pk})'$ le covariate dei k individui e $z_{u_{(j)}} = \sum_{k \in u_{(j)}} x_k = (z_{1u_j}, z_{2u_j}, \dots, z_{pu_j})'$, ossia z_{lu_j} è la somma delle l covariate delle $m_{(i)}$ persone che sono in $u_{(j)}$. Infine, indichiamo con $u_{(j)}^*$ il set delle $m_{(i)}$ persone che falliscono al tempo t_i e $z_{u_{(j)}^*}$ la somma delle covarianze delle $m_{(i)}$ persone che sono in $u_{(j)}^*$.

Scala Di Tempo Continua

Nel caso di una scala temporale continua, per le $m_{(i)}$ persone che falliscono al tempo t_i , è ragionevole dire che i tempi di sopravvivenza delle $m_{(i)}$ persone non sono identici poiché è molto probabile che i legami siano il risultato di misurazioni imprecise. Se si potessero fare le misurazioni precise, si potrebbero ordinare questi $m_{(i)}$ tempi di sopravvivenza e si potrebbe usare la funzione di verosimiglianza. In assenza di conoscenza del vero ordine (il caso reale), dobbiamo considerare tutti i possibili ordini di questi tempi di sopravvivenza legati osservati. Per ogni t_i , il tempo di sopravvivenza legato $m_{(i)}$ osservato può essere ordinato in $m_{(i)}!$ diversi modi possibili.

Per ognuno di questi possibili ordini avremo un prodotto come in (1.93) per i corrispondenti $m_{(i)}$ tempi di sopravvivenza. Pertanto, quando il tempo di sopravvivenza viene misurato su una scala temporale continua, la costruzione e il calcolo dell'esatta funzione di verosimiglianza parziale è un compito molto lungo se $m_{(i)}$ è molto grande.

Per approssimare l'esatta funzione di verosimiglianza parziale, possiamo utilizzare la seguente funzione di verosimiglianza, fornita da Breslow (1974), quando $m_{(i)}$ è piccolo rispetto a $r_{(i)}$:

$$L_B(b) = \prod_{i=1}^k \frac{\exp(z_{u^*(j)}^{*'} b)}{[\sum_{l \in R(t_i)} \exp(x_l' b)]^{m_i}} \quad (1.96)$$

Scala Di Tempo Discreta

Se i tempi di sopravvivenza sono osservati per un tempo discreto, le osservazioni legate sono veramente tali: gli eventi sono effettivamente avvenuti nel medesimo momento. In questo ambito, Cox propone il seguente modello logistico:

$$\frac{h_i(t) dt}{1 - h_i(t) dt} = \frac{h_0(t) dt}{1 - h_0(t) dt} \exp\left(\sum_{j=1}^p b_j x_{ji}\right) = \frac{h_0(t) dt}{1 - h_0(t) dt} \exp\left(\sum_{j=1}^p b' x_j\right)$$

Questo modello riduce alla (1.89) in una scala di tempo continua. Usando il modello e sostituendo l'i-esimo termine nella (1.93) con il seguente termine con le osservazioni legate a t_i :

$$\frac{\exp(z_{u^*(i)}^{*'} b)}{\sum_{u \in U(i)} \exp(z_{u(j)}' b)}$$

La funzione di verosimiglianza parziale in presenza di osservazioni legate per una scala temporale discreta è:

$$L_d(d) = \prod_{i=1}^k \frac{\exp(z_{u^*(j)}^{*'} b)}{\sum_{u(j) \in U_i} \exp(z_{u(j)}' b)} \quad (1.97)$$

Il termine i -esimo in questa espressione rappresenta la probabilità condizionata di osservare gli $m_{(i)}$ guasti dato che ci sono $m_{(i)}$ guasti al tempo $t_{(i)}$ e l'insieme di rischio $R(t_i)$ in $t_{(i)}$. La (1.97) può essere considerata come un'approssimazione della funzione di verosimiglianza parziale per i tempi di sopravvivenza continui con legami assumendo che i legami siano veri come se fossero osservati su una scala temporale discreta.

Le funzioni di verosimiglianza parziale, di cui sopra, sono approssimazioni ragionevolmente buone dell'esatta funzione di verosimiglianza parziale per il tempo di sopravvivenza continuo con legami.

Stima Delle Funzioni Di Sopravvivenza Con Le Covariate

Quando vengono utilizzati i modelli di regressione parametrica, possiamo stimare la funzione di sopravvivenza semplicemente sostituendo i parametri e i coefficienti nella funzione di sopravvivenza con le loro stime. Questo non è possibile se si utilizza il modello dei rischi proporzionali di Cox poiché non conosciamo la forma esatta della funzione di rischio di base o della funzione di sopravvivenza. Tuttavia, esiste uno stimatore della funzione di sopravvivenza, proposto da Breslow nel 1974.

Come indicato in precedenza, sotto il modello di Cox, la funzione di sopravvivenza inclusiva delle covariate, è pari a:

$$S(t, \mathbf{x}) = [S_0(t)]^{\exp(\sum_{j=1}^p b_j x_j)} \quad (1.98)$$

Una volta stimati i coefficienti di regressione, è sufficiente stimare la funzione di sopravvivenza sottostante, $S_0(t)$. Dalla funzione di sopravvivenza stimata, possiamo facilmente stimare la probabilità di sopravvivere più a lungo di un dato tempo per un paziente con un dato insieme di covariate.

Assumendo che la funzione di rischio di base sia costante tra ciascuna coppia di successivi tempi di guasto osservati, Breslow ha proposto il seguente stimatore della funzione di rischio cumulativo di base:

$$\widehat{H}_0(t) = \sum_{t_i \leq t} \frac{m_i}{\sum_{l \in R(t_i)} \exp(x_l' \widehat{b})} \quad (1.99)$$

La funzione di sopravvivenza di base può essere stimata come segue.

$$\widehat{S}_0(t) = \exp[-\widehat{H}_0(t)] = \prod_{t_i \leq t} \left\{ \exp \left[\frac{m_i}{\sum_{l \in R(t_i)} \exp(x_l' \widehat{b})} \right] \right\} \quad (2.01)$$

e la funzione di sopravvivenza per una persona con un set di covariate $\mathbf{x} = (x_1, \dots, x_p)$ è:

$$\widehat{S}(t, \mathbf{x}) = [S_0(t)]^{\exp(\sum_{j=1}^p b_j x_j)} = [S_0(t)]^{\exp(\widehat{b}' \mathbf{x})} \quad (2.02)$$

$\hat{S}(t, \mathbf{x})$ ha una distribuzione normale asintotica con media $S(t, \mathbf{x}) = \exp [-H(t, \mathbf{x})]$ e varianza $\widehat{Var}(\hat{S}(t, \mathbf{x})) \sim [\hat{S}(t, \mathbf{x})]^2 \widehat{Var}(\hat{H}(t, \mathbf{x}))$. L'intervallo di confidenza per la funzione di sopravvivenza è pari a:

$$\left(\hat{S}(t, \mathbf{x}) - Z_{\alpha/2} \sqrt{\widehat{Var}(\hat{S}(t, \mathbf{x}))}, \hat{S}(t, \mathbf{x}) + Z_{\alpha/2} \sqrt{\widehat{Var}(\hat{S}(t, \mathbf{x}))} \right)$$

Valutazione della bontà di adattamento dei residui

Sono disponibili molti metodi grafici per valutare la bontà di adattamento di un modello di rischi proporzionali. Questi metodi sono basati sui residui e sono spesso usati come strumenti diagnostici. Il grafico fornisce alcune informazioni sull'adeguatezza del modello e su valori anomali. Analogamente ad altri metodi grafici, l'interpretazione dei grafici residui può essere soggettiva.

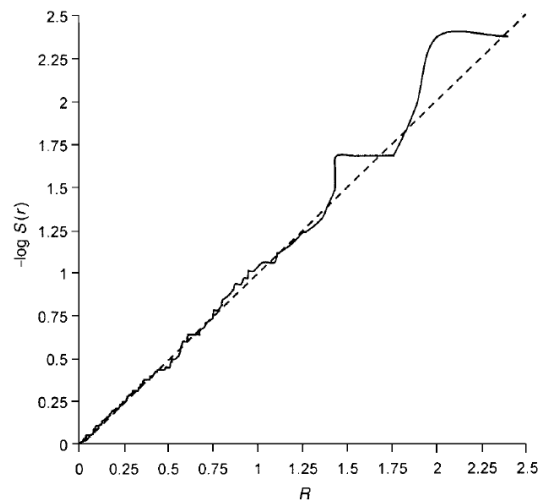
Nei metodi di regressione multipla, i residui sono indicati come la differenza tra i valori osservati e quelli previsti (basati sul modello di regressione) della variabile dipendente.

Tuttavia, quando sono presenti osservazioni censurate e viene utilizzata solo una funzione di verosimiglianza parziale nel modello dei rischi proporzionali, il consueto concetto di residui non è applicabile.

Il metodo Cox-Snell, precedentemente discusso, può essere facilmente esteso al modello dei rischi proporzionali. Il residuo esteso di Cox—Snell, R_i , per l'i-esimo individuo con tempo di sopravvivenza osservato t_i e covariate ai valori $\mathbf{x}_i$ è definito come $R_i = -\log S(t_i; \mathbf{x}_i)$, che è il rischio accumulato stimato basato sul modello dei rischi proporzionali. Se la t_i osservata è censurata, sarà censurata anche la corrispondente R_i . Se il modello dei rischi proporzionali è appropriato, il grafico di R_i e la sua stima di Kaplan-Meier della funzione di sopravvivenza ($\hat{S}(R)$) apparirà come

una linea retta a 45°(come mostrato nel grafico 19). Il metodo residuo di Cox-Snell è utile per valutare la bontà dell'adattamento di un modello parametrico, ma non è così desiderabile per un modello a rischi proporzionali in cui viene utilizzata una funzione di verosimiglianza parziale e la funzione di sopravvivenza è stimata con metodi non parametrici.

Grafico 19. Grafico Dei Residui Di Cox Snell E Del Modello Dei Rischi Proporzionali Di Cox



Modelli di Rischio Non Proporzionali

Il modello dei rischi proporzionali per l'identificazione di importanti fattori prognostici presume che le covariate siano indipendenti dal tempo. Assumiamo anche che ci sia solo una causa di fallimento; cioè, l'evento o il guasto può verificarsi solo una volta per ogni persona e non c'è correlazione tra i tempi di guasto di persone diverse. Tuttavia, in pratica, le covariate possono essere osservate più di una volta durante lo studio e i loro valori cambiano nel tempo, il fallimento può essere dovuto a più di un evento o causa, lo stesso evento o fallimento può ripresentarsi durante uno studio di follow-up e l'evento o fallimento osservato può provenire da persone collegate in una famiglia o dalla stessa persona in momenti diversi.

Modelli con Covariate che variano nel Tempo

Nel modello dei rischi proporzionali di Cox, si assume che il rapporto delle funzioni di rischio per due persone qualsiasi sia indipendente dal tempo t , oppure le covariate non sono dipendenti dal tempo. Tuttavia, è comune nella pratica che uno studio includa sia covariate dipendenti dal tempo che indipendenti dal tempo. Ad esempio, in uno studio longitudinale sulle malattie cardiache, alcune variabili demografiche, come il sesso e la razza, non cambiano nel tempo e di solito vengono raccolte solo una volta all'esame di riferimento. Altre variabili possono variare con il tempo e sono spesso raccolti in esami successivi. La funzione di verosimiglianza parziale che riguarda le covariate dipendenti dal tempo ha la stessa forma di quella in (1.93) eccetto che le covariate sono ora una funzione del tempo. Cioè, la funzione di verosimiglianza parziale con covariate dipendenti dal tempo è:

$$L(\mathbf{b}) = \prod_{i=1}^k \frac{\exp(\sum_{j=1}^p b_j x_{j(i)}(t_i))}{\sum_{l \in R(t_i)} \exp(\sum_{j=1}^p b_j x_{jl}(t_i))} \quad (2.03)$$

dove k è il numero di distinti tempi di fallimento, $R(t_i)$ è il risk set che contiene tutte le persone a rischio al tempo t_i , $x_l(t_i) = (x_{1l}(t_i), x_{2l}(t_i), \dots, x_{pl}(t_i))$ indicano le covariate osservate dalla persona l all'istante dell'evento ordinato non censurato t_i , e $b' = (b_1, b_2, \dots, b_p)'$ indica i coefficienti incogniti. Per le covariate che non variano nel tempo, i loro valori sono costanti nel tempo. Di solito, abbiamo una combinazione di covariate non dipendenti dal tempo e dipendenti dal tempo nella funzione di verosimiglianza. La procedura di stima dei coefficienti, b_j , è simile a quella discussa precedentemente.

Esistono due tipi di covariate dipendenti dal tempo: (1) covariate osservate ripetutamente in diversi momenti di follow-up prima del verificarsi dell'evento o della fine di uno studio o del tempo censurato; e (2) covariate che cambiano nel tempo secondo una funzione matematica nota e covariate che hanno valori diversi a causa della terapia, dell'età o dei cambiamenti nelle condizioni mediche.

Analisi dei risultati

Software e dati di partenza

STATA è il software utilizzato per l'intera analisi, dall'acquisizione dei dati, alla generazione e esportazione dei risultati, grafici e tabelle, attraverso un'interfaccia grafica user-friendly. Si tratta di un moderno e ricco programma per un ampio panorama di analisi statistiche.

L'uso di questo software ha permesso di ottenere risultati circa le stime non parametriche di Kaplan-Meier e l'analisi delle tabelle di vita, di confrontare l'adattamento ai dati di diverse distribuzioni parametriche e di applicare modelli semi-parametrici al dataset disponibile. In quest'ultimo caso, come vedremo in seguito, Stata è uno strumento prezioso in modelli "statici" come quello di Cox. Quando si cerca di affinare il modello della Survival Analysis, considerando covariate time varying, il software risulta meno affidabile. La complessità delle funzioni stimate e l'elevata dimensione del campione non permettono a Stata di generare risultati soddisfacenti ai fini di questa analisi. D'altro canto, se avessimo usato informazioni relative ad un settore con un numero minore di imprese, i risultati sarebbero poco credibili.

Il database utilizzato per questa analisi comprende informazioni relative a circa 8657 imprese operanti nel settore delle apparecchiature elettriche, attive in tutte le zone d'Italia e che hanno depositato il bilancio annuale in un periodo che va dal 1995 al 2019. Sono riportate le seguenti informazioni:

OSSERVAZIONI	ETA'	EVENTO	CENSURA	COVARIATE
--------------	------	--------	---------	-----------

- La variabile **OSSERVAZIONI** effettua il conteggio delle informazioni riportate, ma, poiché per ogni impresa è considerato soltanto il bilancio in cui si sperimenta

l'evento di default o la censura, tale variabile coincide con il numero di imprese del dataset.

- La variabile **ETÀ** rappresenta gli anni di operatività dell'impresa al momento di default o di censura. È così caratterizzata:

Variabile	Media	Dev. Standard	Min	Max
ETA'	15.65	13.4379	0	110

- La variabile **EVENTO** è una variabile binaria che assume valore 1 se l'impresa sperimenta l'evento di interesse, in questo caso il default, 0 se invece non lo sperimenta.
- La variabile **CENSURA** è il complemento ad 1 della variabile evento. Se l'impresa va in default, la variabile evento assume valore 1 e di conseguenza questo tipo di informazione non sarà censurata, quindi la variabile Censura assume valore 0. Ma se invece l'impresa non sperimenta mai il default (evento=0), questa informazione sarà censurata, quindi la variabile CENSURA assume valore 1.
- Infine, sono riportate le **COVARIATE** introdotte nello studio, ossia gli indicatori di stato patrimoniale e di conto economico precedentemente descritti. Il valore di queste variabili è quello registrato nel momento del default o di censura.

Nel corso della seguente analisi, non sono state utilizzate tutte le variabili di stato patrimoniale e di conto economico in quanto, una parte di queste risultava essere poco correlata con la variabile EVENTO oggetto di interesse, altre invece presentavano un'elevata correlazione tra di loro e pertanto si generavano problemi di collinearità nell'applicazione dei modelli. È stato necessario fare una selezione tra le variabili.

Per prima cosa è stata calcolata la correlazione tra le singole covariate e la variabile evento e sono state escluse tutte quelle con bassa correlazione (<0.15).

È stata poi calcolata la correlazione tra le variabili restanti; tra quelle molto correlate tra di loro sono state eliminate quelle che presentavano una correlazione minore con la variabile evento.

Le più indicative variabili “scelte” per far parte dello studio sono le seguenti:

- Ebitda/AN
- Patrimonio Netto Tangibile /AN
- Oneri finanziari/ EBIT
- Passività Correnti/ Ricavi

I rispettivi valori di correlazione con la variabile evento sono pari a:

COVARIATA	EVENTO
Ebitda/AN	-0.2379
Patrimonio Netto Tangibile /AN	-0.1869
Oneri finanziari/ EBIT	0.2519
Passività Correnti/ Ricavi	0.1645

I valori di correlazione tra le singole variabili sono invece pari a:

	Ebitda/AN	Patrimonio Netto Tangibile /AN	Oneri finanziari/ EBIT	Passività Correnti/ Ricavi
Ebitda/AN	1			
Patrimonio Netto Tangibile /AN	0.3365	1		
Oneri finanziari/ EBIT	-0.7290	-0.2886	1	
Passività Correnti/ Ricavi	-0.3072	-0.0942	0.3572	1

Le informazioni di partenza di queste variabili, descritte da Stata, sono:

	Media	Dev. Standard	Min	Max
Ebitda/AN	2,537	14,022	-23,987	273912
Patrimonio Netto Tangibile /AN	17,345	38,07	-52,034	78,088
Oneri finanziari/ EBIT	173,65	187,88	-3,759	400
Passività Correnti/ Ricavi	38351,58	107152,4	16,222	363305

Stabilite le variabili di partenza, è possibile partire con l'analisi vera e propria. Prima di poter analizzare i risultati, però, è necessario definire qual è la variabile temporale dello studio e quale la variabile di fallimento. Le due variabili cui si fa riferimento sono rispettivamente la variabile ETA' e la variabile EVENTO.

```
. stset ETA, failure(evento)
```

```
      failure event:  evento != 0 & evento < .
obs. time interval:  (0, ETA]
exit on or before:  failure
```

```
-----
      8657  total obs.
```

```
      337  obs. end on or before enter()
-----
```

```
      8320  obs. remaining, representing
```

```
      1794  failures in single record/single failure data
```

```
135402  total analysis time at risk, at risk from t =          0
              earliest observed entry t =          0
              last observed exit t =          110
```

Prima di procedere con l'illustrazione dei risultati ottenuti, è importante fare le seguenti osservazioni:

- I valori numerici dell'età, compresi tra 0 e 110, possono sembrare molto elevati; la ragione è legata al fatto che sono calcolati in base all'anno di costituzione della società, che è come richiede l'analisi di sopravvivenza, e non in base all'età che le imprese hanno nel momento in cui entrano nello studio. Imprese con elevati anni di esercizio non si esclude abbiano subito precedentemente altre procedure negli anni passati.

Per un'analisi più completa sarebbe opportuno osservare società costituite da poco, entro il periodo di tempo usato per lo scarico dei bilanci, ma questi non erano disponibili nel database AIDA.

- Il campione delle imprese per questa analisi risente del Survival Bias. Infatti, i dati presenti nel database AIDA sono relativi principalmente alle società di capitali, mancano informazioni relative a società di persone e ditte individuali ad esempio, e, inoltre, molti bilanci riportano dei dati mancanti o inesatti. Di conseguenza questo aspetto ha un impatto sull'analisi che si sta svolgendo.

Analisi Non Parametrica

Quella Non Parametrica è l'analisi più semplice che ci possa essere, non prevede la presenza delle covariate all'interno del dataset. Le uniche informazioni richieste sono quelle che riguardano la variabile temporale (ETÀ), la variabile di fallimento che ci dice se una determinata azienda ha sperimentato o meno l'evento di default ed infine, la variabile che segnala la presenza di informazioni censurate.

La stima non parametrica è condotta considerando in primis l'insieme di tutte le società del database, poitenendo conto della differenza dimensionale delle imprese sulla base del fatturato e dell'area geografica di provenienza. Questo ci consente di verificare se l'appartenenza ad una precisa zona o la dimensione delle imprese hanno un impatto significativo sulla sopravvivenza delle imprese stesse.

Analisi Non Parametrica - Società

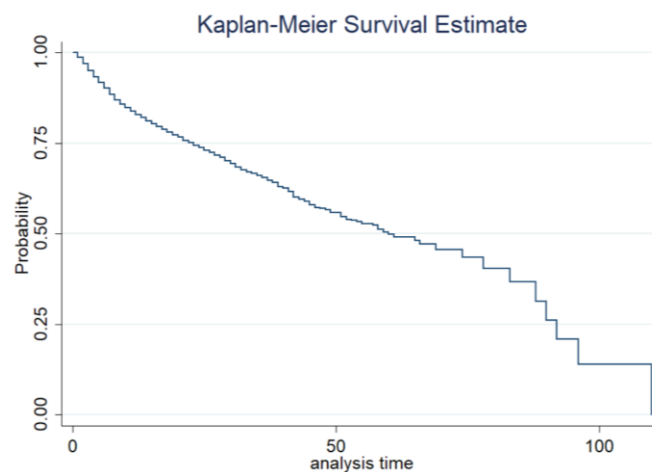
Il primo passo è quello di considerare nell'interezza il database costruito per questo studio. In pratica, non si mettono in evidenza le differenze tra le imprese ma si assume che tutte contribuiscano in egual modo all'analisi della sopravvivenza.

La stima di Kaplan-Meier, allegata in forma tabellare nell'**Allegato F**, riporta l'intervallo temporale (lo studio riguarda le imprese che hanno da 1 fino a 110 anni di attività), il totale delle imprese aggiornato per ogni singolo anno al quale vengono sottratte le

imprese che subiscono il fallimento e la censura, la funzione di sopravvivenza stimata per quell'anno, i rispettivi standard error e intervalli di confidenza.

Come mostrato nel grafico 20 sottostante, quella della sopravvivenza è una funzione a gradini, con un valore molto alto al primo anno, circa il 98.7% che è pari a 0 nell'ultimo anno; tuttavia, sappiamo che in presenza di osservazioni censurate una funzione di sopravvivenza pari a zero non ha molto significato.

Grafico 20. Stima di sopravvivenza di Kaplan-Meier

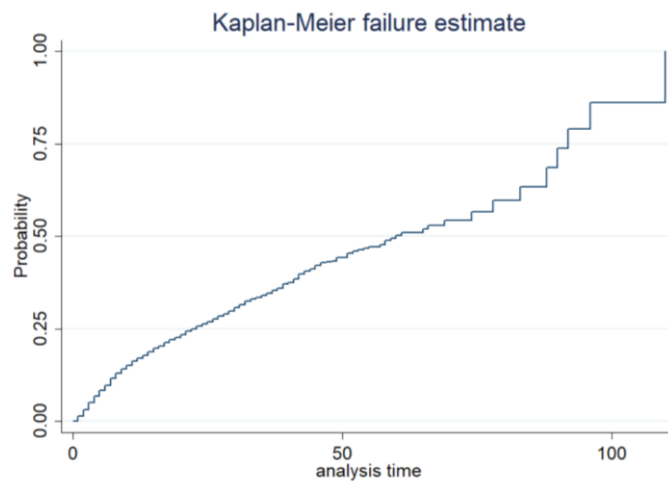


Nei punti in cui assume un andamento orizzontale, la funzione di sopravvivenza rimane costante, come nel caso degli anni che vanno dai 95 ai 110 per poi annullarsi del tutto.

Un modo del tutto analogo per descrivere quanto appena detto, è quello di considerare non la curva di sopravvivenza, bensì quella di fallimento (**grafico 21**). Quest'ultima ha un andamento monotono crescente che descrive la probabilità di fallimento che aumenta e che nei tratti orizzontali rimane costante.

La probabilità di fallimento cresce (la probabilità di sopravvivenza decresce) molto rapidamente fino a circa 70 anni di attività, per poi subire dei rallentamenti negli anni successivi.

Grafico 21. Stima di fallimento di Kaplan-Meier



L'allegato G riporta la costruzione delle Tabelle di Vita per queste aziende, dove la variabile temporale non è più considerata puntualmente per ogni singolo anno di età, ma per intervalli di tempo che per convenzione sono pari a 5 anni. Come si vede nei grafici 22 e 23, i risultati ottenuti sono analoghi a quanto detto finora, con l'unica differenza che i dati sono aggregati e non più puntuali.

Grafico 22. Proportion Surviving – Tavole di Vita

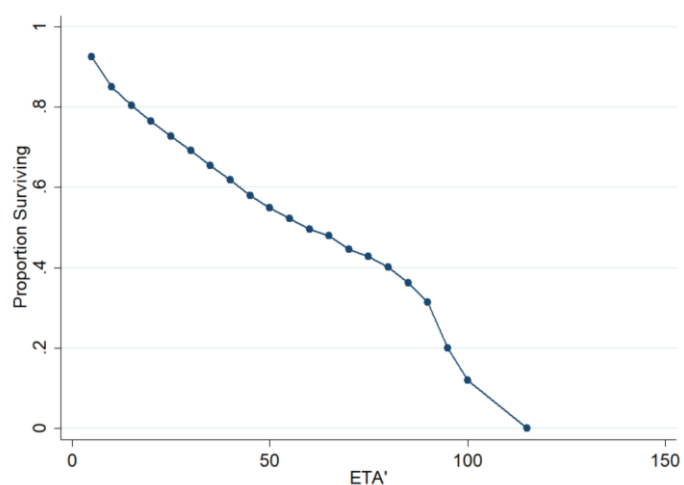
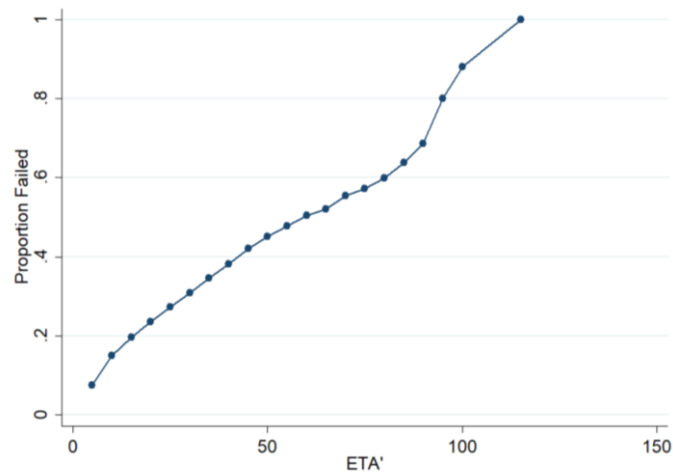
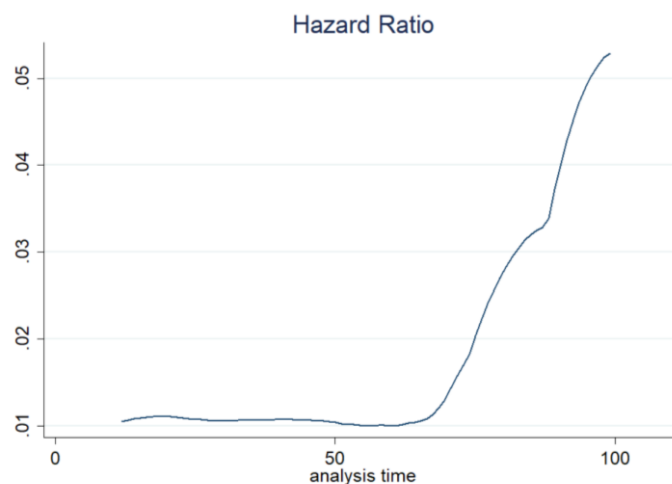


Grafico 22. Proportion Failed – Tavole di Vita



Infine, considerando la funzione di rischio, **grafico 24**, possiamo notare come il tasso di rischio oscilli intorno allo 0.01 fino ai 70 anni circa, per poi aumentare drasticamente dai 71 anni, subendo un rallentamento dove la funzione di sopravvivenza rimane costante.

Grafico 24. Hazard Ratio



Analisi Non Parametrica - Fatturato

In questa sezione, proviamo a differenziare le imprese in base alle loro dimensioni per vedere se ci sono delle differenze significative. Le aziende sono così suddivise:

- **GRANDI IMPRESE**, imprese con fatturato superiore ai 50 mln di euro;
- **MEDIE IMPRESE**, imprese con fatturato minore o pari a 50 mln di euro;
- **PICCOLE IMPRESE**, imprese con fatturato minore o pari a 10 mln di euro;
- **MICRO IMPRESE**, imprese con fatturato minore o pari a 2 mln di euro.

Le stime differenziate di Kaplan- Meier sono riportate nei **grafici 25 e 26**.

Grafico 25. Funzione di sopravvivenza stimata

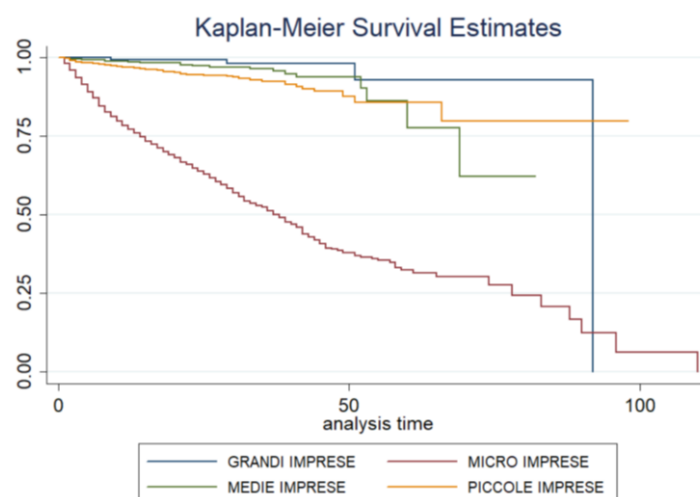
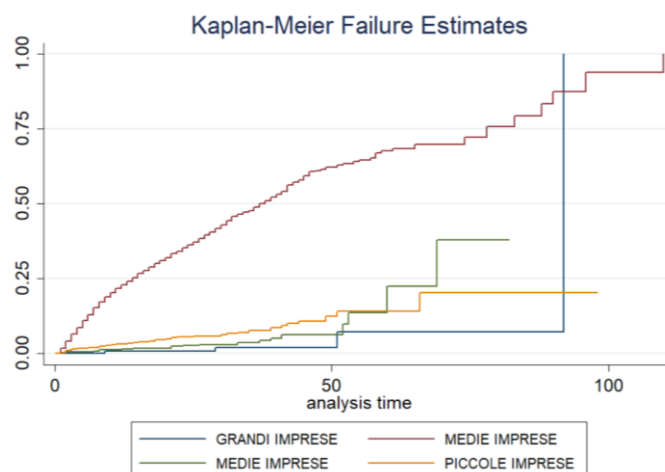


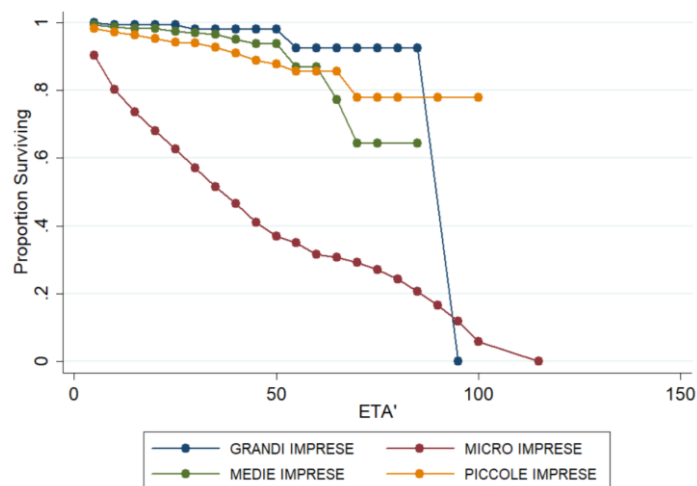
Grafico 26. Funzione di fallimento stimata



Le imprese che in questo settore sembrano essere maggiormente in difficoltà sono le microimprese. Queste, infatti, sono caratterizzate da una funzione di sopravvivenza che decresce rapidamente fino ai 70 anni di vita, per poi subire un rallentamento e annullarsi oltre i 100 anni di attività. Al contrario, le grandi imprese hanno una funzione di sopravvivenza quasi prossima al 100% per i primi 50 anni di attività, cosa che le rende particolarmente solide, e poi, nel restante mezzo secolo, la probabilità di sopravvivenza subisce una piccola riduzione. Essa, infatti, rimane al di sotto del 25%. Le PMI, invece, vedono una realtà molto simile a quella delle grandi imprese. Queste ultime hanno una funzione di sopravvivenza migliore delle microimprese ma sperimentano maggiori casi di fallimento rispetto alle grandi imprese.

L' **allegato H** mostra le tabelle di vita differenziate in base al fatturato e il **grafico 27** la proporzione di sopravvissuti per le quattro classi di aziende.

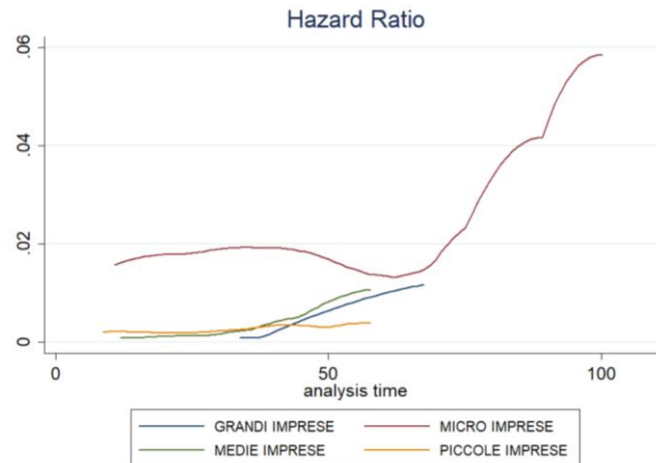
Grafico 27. Proporzione di sopravvissuti in base al fatturato



Il grafico 28 mostra la funzione di rischio stimata in base al fatturato delle imprese. Le grandi imprese, le cui età, tipicamente vanno dai 30 ai 70 anni, hanno un tasso di rischio crescente ma al di sotto dello 0.02. Le piccole e medie imprese hanno un tasso di rischio piatto fino ai 30 anni circa che poi cresce fino ai 70 pur rimanendo al di sotto

dello 0.02. Le micro imprese, che sono quelle con anni di attività maggiori, hanno un rischio maggiore che aumenta notevolmente dopo i 60 anni di attività.

Grafico 28. Stima della funzione di rischio



Pertanto, la categoria maggiormente esposta al rischio in questo settore è quella delle microimprese e quella meno esposta è quella delle grandi imprese.

Analisi Non Parametrica - Area Geografica

Classificando le imprese in base a dove è collocata la loro sede legale, avremo:

- Imprese del Nord-est
- Imprese del Nord-ovest
- Imprese del Centro
- Imprese del Sud & Isole

L'obiettivo è quello di capire se la provenienza geografica pregiudica la probabilità di sopravvivere di un'impresa.

Le stime di Kaplan-Meier sono descritte nei **grafici 29 e 30**.

Grafico 29. Stima di Sopravvivenza di Kaplan-Meier

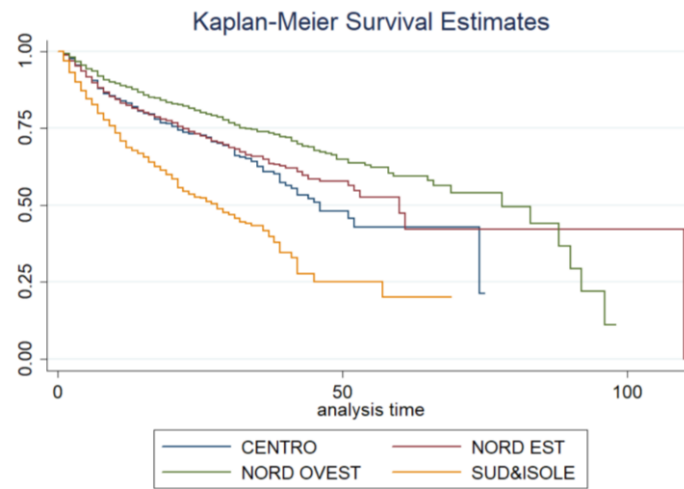
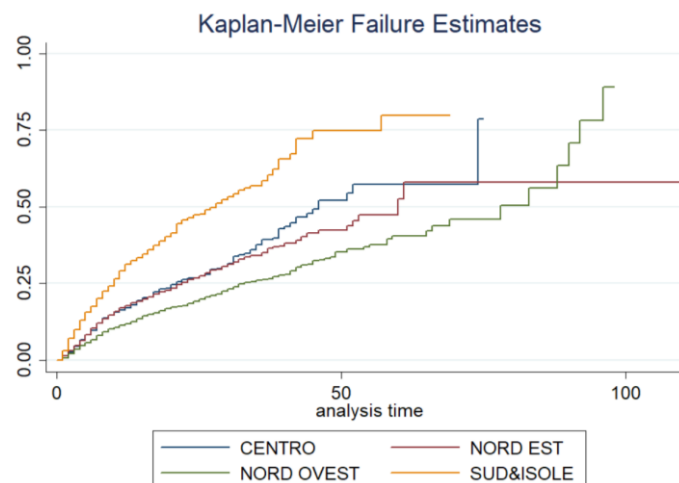


Grafico 30. Stima di Fallimento di Kaplan-Meier

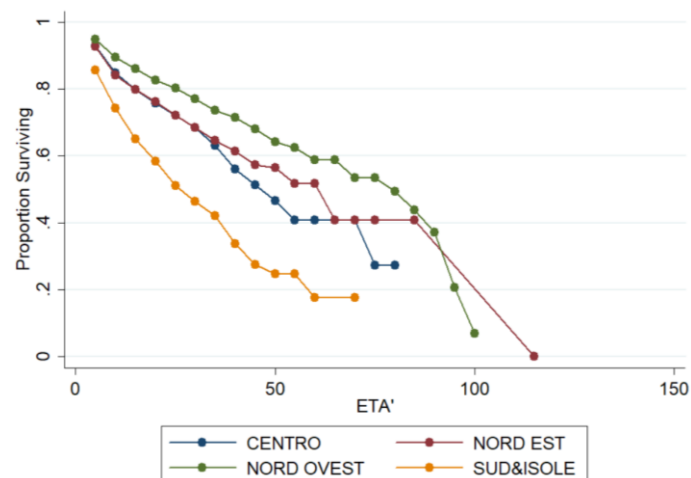


Le funzioni sono molto simili tra di loro, hanno tutte un andamento monotono crescente nel caso della sopravvivenza e monotono decrescente nel caso del fallimento. Tuttavia, a parità di età, le imprese del nord-ovest sembrano avere maggiori possibilità di sopravvivenza rispetto alle altre; le imprese del sud e delle isole sono quelle che invece hanno una maggior probabilità di fallire. Per queste ultime, la funzione di sopravvivenza è costante tra i 60 e i 70 anni. Non ci sono imprese, in questo campione che superano questi anni di attività. Le imprese del Nord Italia sono quelle

più longeve tra quelle in esame. Da i 60 anni in poi, le imprese del Nord-est hanno una funzione di sopravvivenza costante, di poco inferiore allo 0.50; al di sotto di quell'età sono molto simili alle imprese del Centro.

Le tabelle di vita per queste imprese sono presenti nell'**Allegato I** ed il **grafico 31** aggrega le informazioni appena descritte in intervalli con ampiezza di cinque anni. Considerando i primi anni di esercizio, le imprese del nord-ovest hanno una percentuale di sopravvivenza decisamente maggiore di quelle del Sud e delle Isole e poco più grande di quelle del centro e del nord est. Questo divario cresce con l'aumentare del tempo, fino all'età di circa 85 anni in cui troviamo soltanto imprese del Nord.

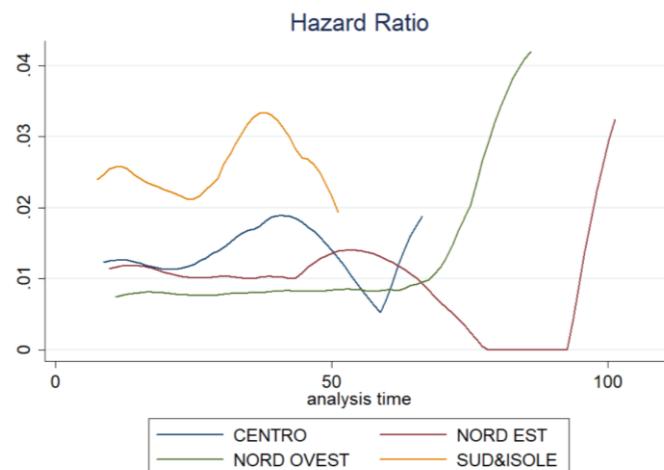
Grafico 31. Stima della proporzione di sopravvivenza



Guardando al grafico 32, di seguito riportato, le stime del tasso di rischio confermano quanto appena detto. Le imprese del Sud & Isole, quelle relativamente più giovani, sono soggette ad un tasso di rischio maggiore, che oscilla tra lo 0.2 e lo 0.4. Le imprese del nord-ovest, considerate le più "solide" hanno il tasso di rischio più basso (meno dello 0.1) di tutte le altre, fino all'età di 70 anni circa, per poi aumentare rapidamente. Un caso interessante è rappresentato dalle imprese del nord-est; dopo i 75 anni la

funzione di rischio rimane costante e prossima allo zero, per poi aumentare superati i 95 anni.

Grafico 32. Stima del tasso di rischio



Analisi Parametrica

L'analisi parametrica prevede l'introduzione delle covariate nello studio. Il database di partenza, su cui si basano questi modelli, contiene molte più informazioni rispetto a quelle richieste per l'analisi non parametrica. Oltre a considerare le variabili dummy di evento e censura e la variabile temporale dell'età, in corrispondenza di ogni azienda, sono riportati i valori relativi all'ultimo anno di esercizio a disposizione dei quattro indicatori precedentemente selezionati tra i più descrittivi dell'esperienza della sopravvivenza.

Secondo lo schema precedentemente descritto nella **Tabella 14**, sono di seguito riportate le applicazioni di questo modello per le varie distribuzioni, applicando, ove possibile, le differenti metriche ammissibili. L'obiettivo è quello di individuare quale, tra queste distribuzioni, meglio rappresenta l'esperienza di default (o di sopravvivenza) di queste aziende. Per ciascuno di questi casi, l'approssimazione di Breslow è usata per trattare le informazioni legate.

Distribuzione Esponenziale

La distribuzione esponenziale è caratterizzata da un solo parametro λ che rappresenta il tasso di rischio. Per questa distribuzione, sono calcolate entrambe le metriche disponibili. L'output generato da STATA può essere espresso sia in termini di hazard ratio che, come più frequentemente avviene, in termini di coefficienti.

PH

Come possiamo vedere dai risultati di regressione di seguito riportati, l'insieme delle variabili indipendenti prevede in maniera affidabile la variabile dipendente di sopravvivenza.

Gli hazard ratio, tutti statisticamente significativi, rappresentano l'incremento di rischio conseguente all'aumentare della misura considerata di una unità. In corrispondenza di ciascuna covariata e anche della costante, per ogni tasso di rischio sono riportati lo standard error ed il rispettivo intervallo di confidenza.

Exponential regression -- log relative-hazard form						
No. of subjects	=	8320	Number of obs	=	8320	
No. of failures	=	1794				
Time at risk	=	135402				
Log pseudolikelihood	=	-5050.9274	Wald chi2(4)	=	1493.70	
			Prob > chi2	=	0.0000	
_t	Haz. Ratio	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
EBITDAAN	.9792635	.0028078	-7.31	0.000	.9737758	.9847822
PatrnettotangAN	.9923287	.0007081	-10.79	0.000	.9909418	.9937175
OFNEBIT	1.002534	.0002004	12.66	0.000	1.002141	1.002926
PCric	1.000001	1.72e-07	4.91	0.000	1.000001	1.000001
_cons	.0076629	.0004244	-87.96	0.000	.0068747	.0085415

I coefficienti di regressione sono significativamente diversi da zero ed i loro segni sono coerenti ai risultati attesi. Il segno negativo del coefficiente di regressione, come per

gli indicatori EBITDA/AN e Patrimonionettotangibile/AN, rappresenta l'effetto negativo sul tempo di sopravvivenza che ha l'aumento di una unità della variabile indipendente. Il ragionamento opposto vale per i coefficienti con segno positivo.

Le aziende che hanno un maggiore valore delle variabili con coefficienti positivi, a parità di quelle con coefficienti negativi, hanno un maggiore effetto positivo sul tempo di sopravvivenza. Di conseguenza, a parità di variabili con segno positivo, le imprese con un più alto valore nelle variabili con coefficienti negativi hanno un peggiore effetto sul tempo di sopravvivenza.

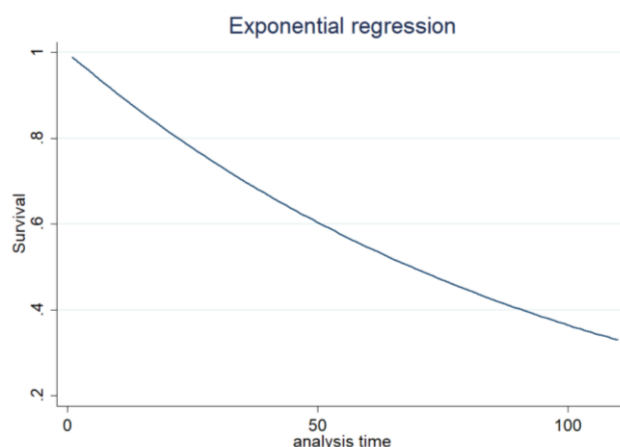
Exponential regression -- log relative-hazard form						
No. of subjects	=	8320	Number of obs	=	8320	
No. of failures	=	1794				
Time at risk	=	135402				
Log pseudolikelihood	=	-5050.9274	Wald chi2(4)	=	1493.70	
			Prob > chi2	=	0.0000	
_t	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
EBITDAAN	-.0209545	.0028672	-7.31	0.000	-.0265742	-.0153348
PatrnettotangAN	-.0077009	.0007136	-10.79	0.000	-.0090994	-.0063023
OFNEBIT	.0025304	.0001999	12.66	0.000	.0021385	.0029222
PCric	8.43e-07	1.72e-07	4.91	0.000	5.06e-07	1.18e-06
_cons	-4.871365	.0553836	-87.96	0.000	-4.979914	-4.762815

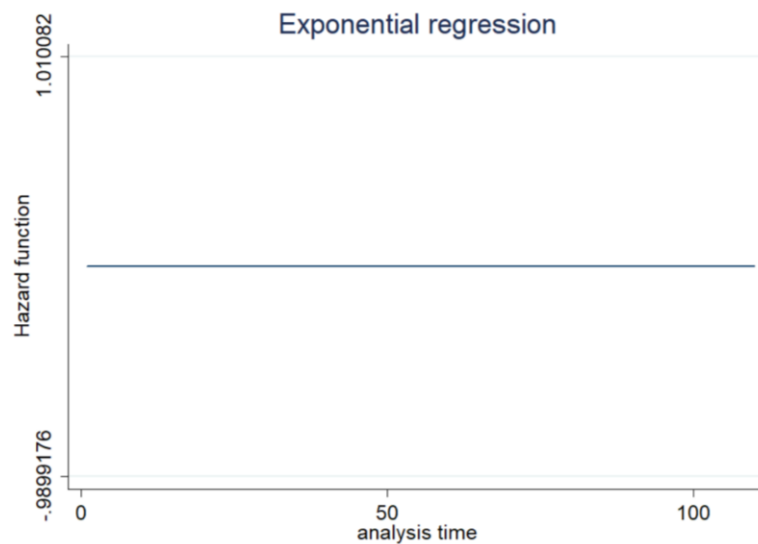
AFT

Utilizzando il tempo di fallimento accelerato, otteniamo risultati analoghi a quelli registrati con la metrica PH ma con tutti i segni invertiti. La particolarità di questa metrica sta nell'interpretazione dei coefficienti di regressione. Ad esempio, il segno più dell'indicatore EBITDA/AN indica che al crescere del suo valore accelera il tempo di sopravvivenza, ossia "il tempo al guasto" è più vicino. Se, invece, consideriamo Pc/Ric, all'aumentare del suo valore, il tempo di sopravvivenza decelera, ossia il "tempo al guasto" è più lontano.

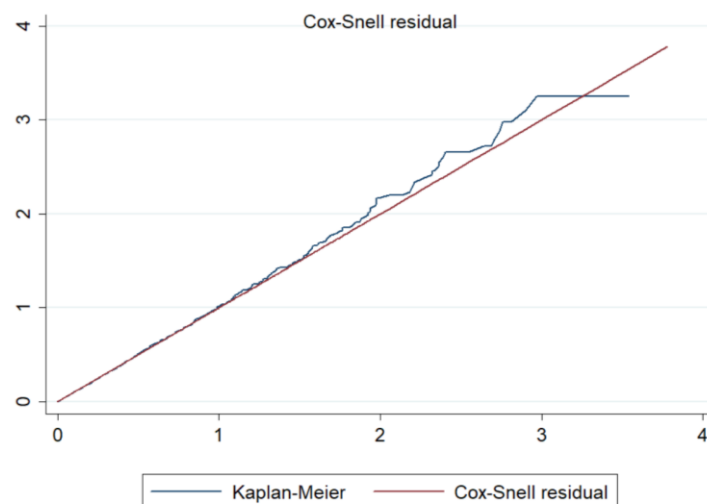
No. of subjects	=	8320	Number of obs	=	8320
No. of failures	=	1794			
Time at risk	=	135402			
			Wald chi2(4)	=	1493.70
Log pseudolikelihood	=	-5050.9274	Prob > chi2	=	0.0000

_t	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
EBITDAAN	.0209545	.0028672	7.31	0.000	.0153348	.0265742
PatrnettotangAN	.0077009	.0007136	10.79	0.000	.0063023	.0090994
OFNEBIT	-.0025304	.0001999	-12.66	0.000	-.0029222	-.0021385
PCric	-8.43e-07	1.72e-07	-4.91	0.000	-1.18e-06	-5.06e-07
_cons	4.871365	.0553836	87.96	0.000	4.762815	4.979914





Infine, il calcolo dei residui di Cox Snell e il loro confronto con la stima di Kaplan-Meier, ci consente di valutare la bontà di adattamento del modello al database. Nel caso della distribuzione esponenziale, il grafico seguente mette in evidenza l'adeguatezza del modello preso in esame, salvo qualche scostamento ritenuto accettabile.



Distribuzione di Weibull

La distribuzione di Weibull, considerata una generalizzazione di quella esponenziale, è caratterizzata dai due parametri, λ e γ . Non assumendo il tasso di rischio costante, questa distribuzione ha un campo di applicazione molto più ampio di quella esponenziale e dispone di entrambe le metriche. L'output generato è espresso sia in termini di hazard ratio che in termini di coefficienti.

PH

Come possiamo vedere dai risultati di regressione di seguito riportati, l'insieme delle variabili indipendenti mostra una relazione statisticamente significativa con la variabile dipendente.

Gli hazard ratio calcolati per la distribuzione di Weibull sono significativamente diversi da zero. Inoltre, nel fondo della tabella è riportata anche la stima del parametro ausiliario p , del suo reciproco e del logaritmo.

Weibull regression -- log relative-hazard form						
No. of subjects	=	8320	Number of obs	=	8320	
No. of failures	=	1794				
Time at risk	=	135402				
Log pseudolikelihood	=	-5048.2557	Wald chi2 (4)	=	1419.86	
			Prob > chi2	=	0.0000	
_t	Haz. Ratio	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
EBITDAAN	.9790764	.0028554	-7.25	0.000	.9734958	.9846889
PatrnettotangAN	.9921959	.000721	-10.78	0.000	.9907838	.99361
OFNEBIT	1.002532	.0002022	12.54	0.000	1.002136	1.002928
PCric	1.000001	1.75e-07	4.84	0.000	1.000001	1.000001
_cons	.0066959	.0005121	-65.45	0.000	.0057637	.0077788
/ln_p	.0433621	.0155538	2.79	0.005	.0128772	.073847
p	1.044316	.0162431			1.01296	1.076642
1/p	.9575646	.0148938			.9288138	.9872054

Come per la distribuzione esponenziale, i coefficienti di regressione sono significativamente diversi da zero ed i loro segni coerenti con i risultati attesi. Il segno positivo del coefficiente di regressione, come per OFN/EBIT e Pc/RIC, rappresenta l'effetto positivo sul tempo di sopravvivenza che ha l'aumento di una unità della variabile indipendente. Il ragionamento opposto vale per i coefficienti con segno negativo.

Weibull regression -- log relative-hazard form						
No. of subjects	=	8320	Number of obs	=	8320	
No. of failures	=	1794				
Time at risk	=	135402				
Log pseudolikelihood	=	-5048.2557	Wald chi2(4)	=	1419.86	
			Prob > chi2	=	0.0000	
_t	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
EBITDAAN	-.0211456	.0029165	-7.25	0.000	-.0268618	-.0154295
PatrnettotangAN	-.0078347	.0007267	-10.78	0.000	-.0092589	-.0064105
OFNEBIT	.0025286	.0002017	12.54	0.000	.0021334	.0029239
PCric	8.48e-07	1.75e-07	4.84	0.000	5.04e-07	1.19e-06
_cons	-5.006264	.0764858	-65.45	0.000	-5.156174	-4.856355
/ln_p	.0433621	.0155538	2.79	0.005	.0128772	.073847
p	1.044316	.0162431			1.01296	1.076642
1/p	.9575646	.0148938			.9288138	.9872054

AFT

Per quanto riguarda la metrica AFT (accelerated failure time), la sua interpretazione ci permette di valutare il tempo di sopravvivenza come accelerato o decelerato rispetto al segno dei coefficienti di regressione. Ovviamente, i segni sono invertiti rispetto a quelli ottenuti con la stima PH e i rispettivi valori sono statisticamente significativi.

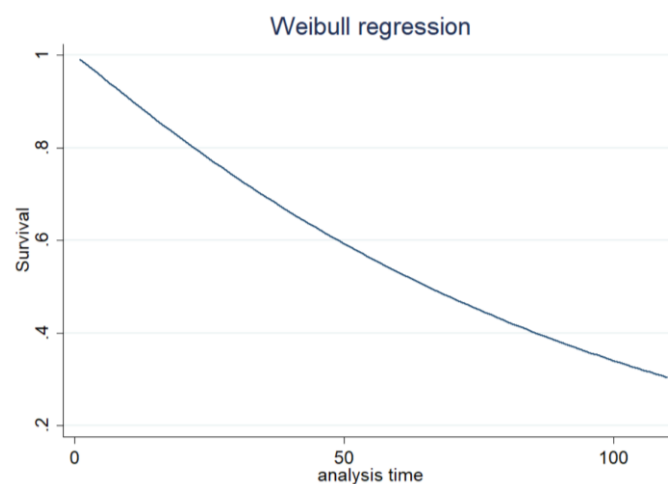
Weibull regression -- accelerated failure-time form

No. of subjects	=	8320	Number of obs	=	8320
No. of failures	=	1794			
Time at risk	=	135402			
Log pseudolikelihood	=	-5048.2557	Wald chi2(4)	=	1374.14
			Prob > chi2	=	0.0000

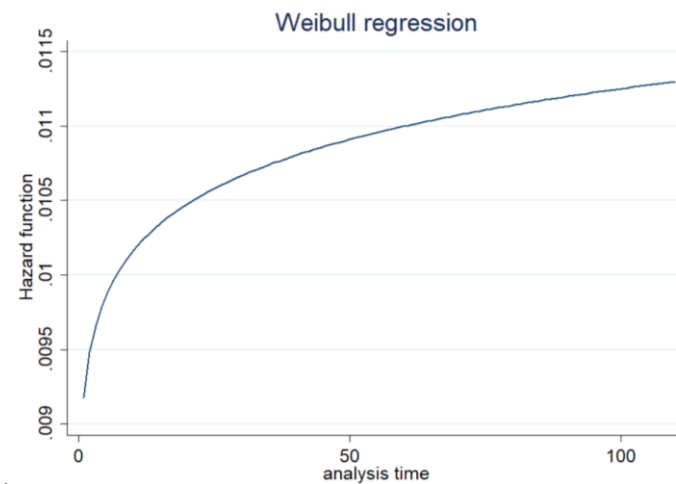
_t	Robust		z	P> z	[95% Conf. Interval]	
	Coef.	Std. Err.				
EBITDAAN	.0202483	.0027895	7.26	0.000	.014781	.0257157
PatrnettotangAN	.0075022	.0006903	10.87	0.000	.0061493	.0088552
OFNEBIT	-.0024213	.0001956	-12.38	0.000	-.0028046	-.002038
PCric	-8.12e-07	1.68e-07	-4.83	0.000	-1.14e-06	-4.82e-07
_cons	4.793822	.0585653	81.85	0.000	4.679036	4.908608
/ln_p	.0433621	.0155538	2.79	0.005	.0128772	.073847
p	1.044316	.0162431			1.01296	1.076642
1/p	.9575646	.0148938			.9288138	.9872054

Come è possibile notare, data l'elevata numerosità del campione, il valore degli standard error, e di conseguenza l'ampiezza degli intervalli di confidenza, sono molto piccoli, ne consegue che siamo in presenza di un modello robusto.

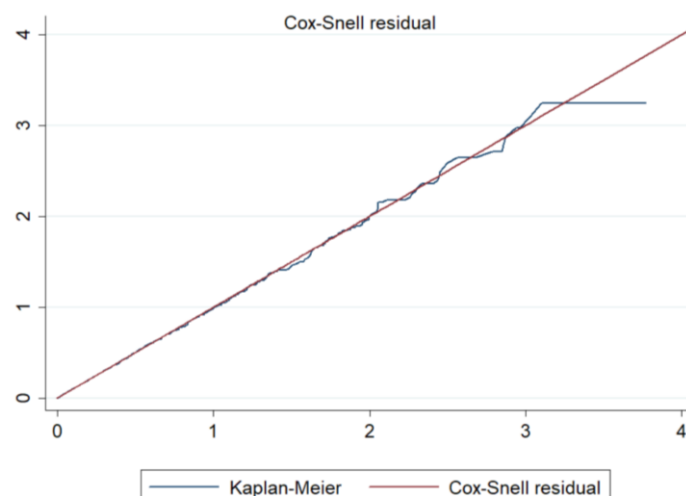
La funzione di sopravvivenza ha un andamento monotono decrescente. Le imprese con 50 anni di attività hanno una probabilità di sopravvivenza minore dello 0.6. La curva di sopravvivenza stimata con la distribuzione di Weibull è un pochino più "bassa" rispetto a quella stimata con la distribuzione esponenziale; pertanto i valori di sopravvivenza sono inferiori al caso esponenziale.



Come messo in evidenza dalla (1.50), ad eccezione del caso in cui $\Upsilon = 1$, la funzione di rischio varia con il tempo. Nel caso specifico, $\Upsilon > 1$, la funzione di rischio è crescente nel tempo. I tassi di rischio, pur assumendo dei valori molto piccoli (asintoticamente inferiori allo 0.0115), smentiscono l'ipotesi precedente di una funzione costante nel tempo.



Dal confronto tra la stima di Kaplan-Meier e i residui di Cox-Snell, notiamo come la distribuzione di Weibull si adatta meglio ai dati rispetto alla distribuzione esponenziale. La configurazione è molto vicina a una linea di 45°, indicando che la distribuzione di Weibull fornisce un adattamento ragionevole.



Distribuzione di Gompertz

La distribuzione di Gompertz, anch'essa considerata una generalizzazione di quella esponenziale, è caratterizzata da due parametri, λ e Υ . Essa assume un tasso di rischio costante inizialmente, ma che poi varia in maniera esponenziale secondo la (1.66).

Per questa distribuzione è disponibile la sola metrica PH e non è possibile il calcolo della stima AFT.

PH

Come possiamo vedere dai risultati di regressione di seguito riportati, l'insieme delle variabili indipendenti prevede in maniera affidabile la variabile dipendente.

Gompertz regression -- log relative-hazard form						
No. of subjects	=	8320	Number of obs	=	8320	
No. of failures	=	1794				
Time at risk	=	135402				
Log pseudolikelihood	=	-5050.2656	Wald chi2(4)	=	1453.39	
			Prob > chi2	=	0.0000	
_t	Haz. Ratio	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
EBITDAAN	.9793326	.002787	-7.34	0.000	.9738854	.9848103
PatrnettotangAN	.9924123	.0007096	-10.65	0.000	.9910224	.9938042
OFNEBIT	1.002538	.0001995	12.74	0.000	1.002147	1.002929
PCric	1.000001	1.70e-07	4.93	0.000	1.000001	1.000001
_cons	.0078856	.0004816	-79.30	0.000	.006996	.0088883
/gamma	-.0022784	.0019106	-1.19	0.233	-.0060231	.0014662

Gli hazard ratio calcolati per la distribuzione di Weibull sono significativamente diversi da zero. Inoltre, nel fondo della tabella è riportata anche la stima del parametro ausiliario Υ .

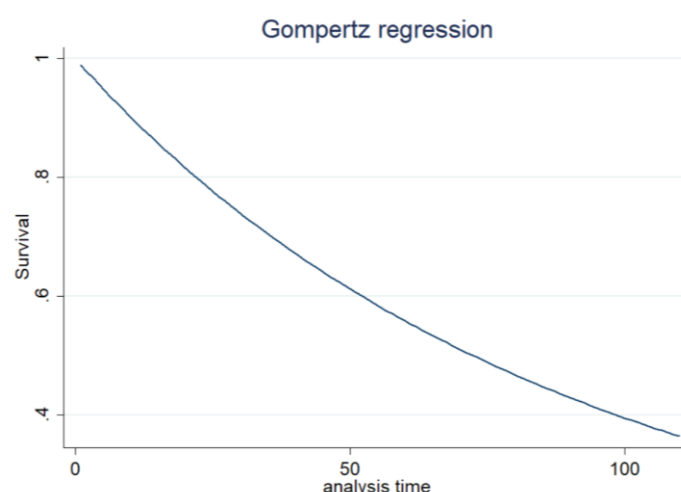
Come per la distribuzione esponenziale, i coefficienti di regressione sono significativamente diversi da zero ed i loro segni coerenti con i risultati attesi.

Gompertz regression -- log relative-hazard form

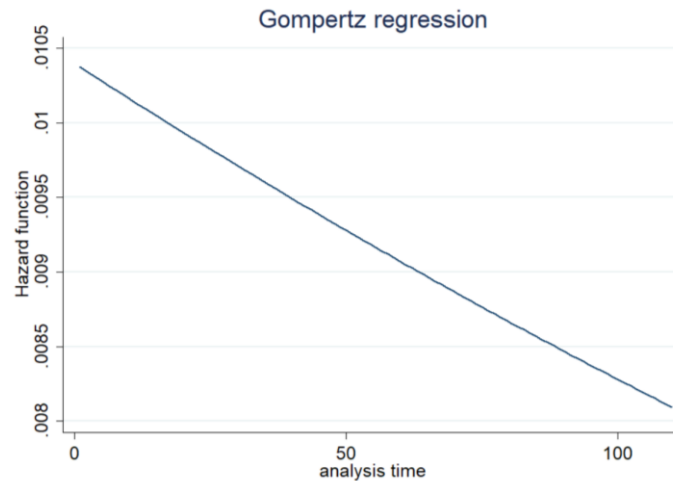
No. of subjects	=	8320	Number of obs	=	8320
No. of failures	=	1794			
Time at risk	=	135402			
Log pseudolikelihood	=	-5050.2656	Wald chi2(4)	=	1453.39
			Prob > chi2	=	0.0000

_t	Robust		z	P> z	[95% Conf. Interval]	
	Coef.	Std. Err.				
EBITDAAN	-.0208839	.0028458	-7.34	0.000	-.0264616	-.0153062
PatrnettotangAN	-.0076166	.0007151	-10.65	0.000	-.0090181	-.0062151
OFNEBIT	.0025346	.000199	12.74	0.000	.0021447	.0029246
PCric	8.40e-07	1.70e-07	4.93	0.000	5.06e-07	1.17e-06
_cons	-4.842718	.0610694	-79.30	0.000	-4.962412	-4.723024
/gamma	-.0022784	.0019106	-1.19	0.233	-.0060231	.0014662

La funzione di sopravvivenza ha un andamento monotono decrescente.



Secondo la (1.66), a seconda del valore di gamma, la funzione di rischio può crescere, decrescere o rimanere costante nel tempo. Nel caso specifico, il valore di gamma è negativo, motivo per cui la funzione di rischio decresce con il tempo. Questo è in contrasto con i risultati che ci aspettiamo di avere. Pertanto, la distribuzione di Gompertz non è idonea agli scopi di questa tesi.



Distribuzione Log-normale

Come si evince dall'output sottostante, le variabili indipendenti prevedono in modo affidabile la variabile dipendente. Per la distribuzione Log-Normale è possibile utilizzare solo la metrica AFT. I coefficienti, con i segni invertiti, sono statisticamente significativi. Un incremento nella variabile Ebitda/an, ad esempio, accelera il tempo di sopravvivenza e quindi riduce il tempo necessario affinché il fallimento si verifichi.

Lognormal regression -- accelerated failure-time form

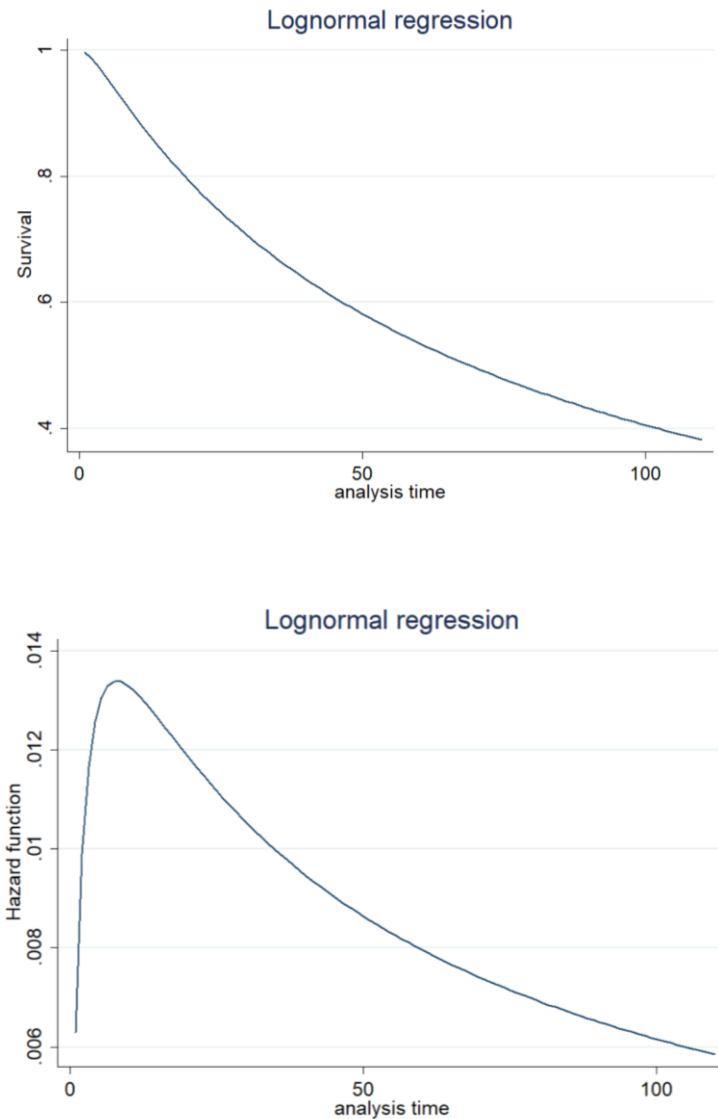
```

No. of subjects =      8320                Number of obs   =      8320
No. of failures =      1794
Time at risk    =     135402
Log likelihood   =    -5127.2269
LR chi2(4)      =     1291.10
Prob > chi2     =      0.0000

```

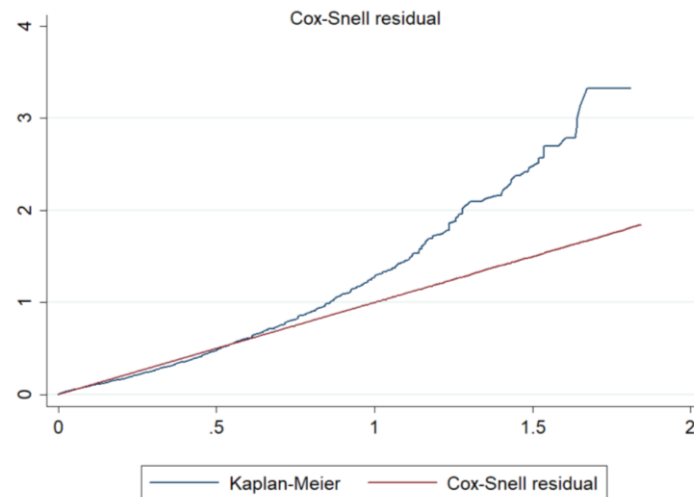
_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
EBITDAAN	.0169791	.0027649	6.14	0.000	.0115599	.0223982
PatrnettotangAN	.0077765	.0006547	11.88	0.000	.0064934	.0090596
OFNEBIT	-.0024087	.0002131	-11.30	0.000	-.0028263	-.001991
PCric	-1.00e-06	2.11e-07	-4.75	0.000	-1.41e-06	-5.88e-07
_cons	4.501246	.0625098	72.01	0.000	4.378729	4.623763
/ln_sig	.4403731	.0175535	25.09	0.000	.4059689	.4747772
sigma	1.553287	.0272656			1.500756	1.607656

Di seguito sono riportati il grafico della funzione di sopravvivenza e della funzione di rischio. La funzione di rischio aumenta fino a raggiungere un punto di massimo e poi diminuisce non appena superato il valore mediano. La funzione tende a zero quando il tempo tende all'infinito.



Il metodo dei residui, di seguito riportato, mette in evidenza lo scarso adattamento del modello ai dati del campione.

Questo modello non è particolarmente indicato per questo studio.



Distribuzione Log-logistica

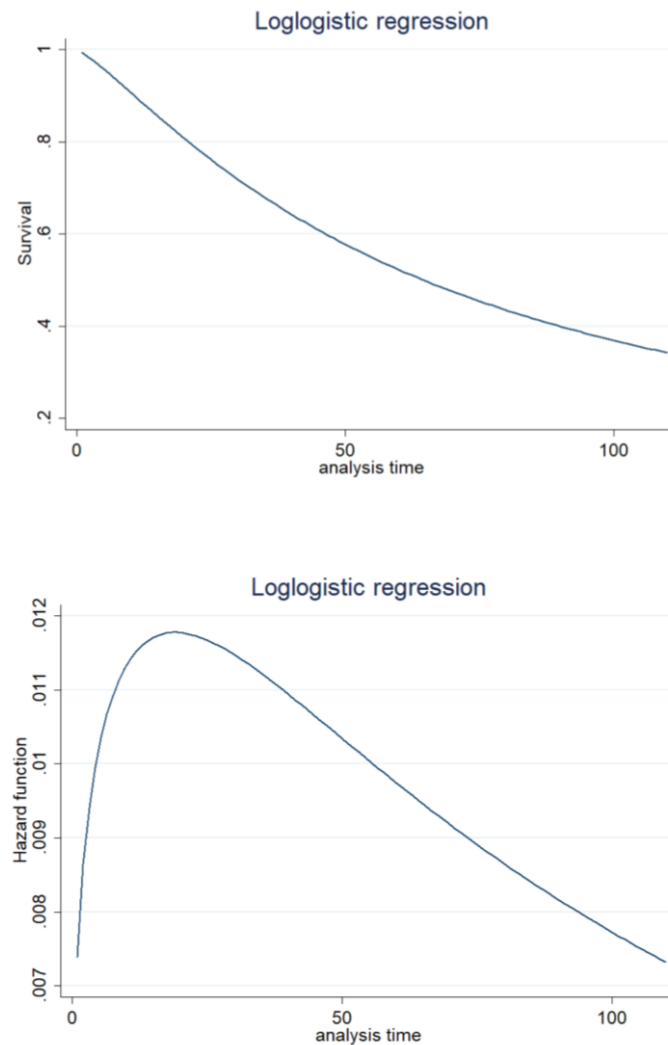
Nel caso della distribuzione log-logistica, è utilizzata la sola metrica AFT. Questa distribuzione, caratterizzata da due parametri α e γ , genera risultati significativamente diversi da zero.

Loglogistic regression -- accelerated failure-time form

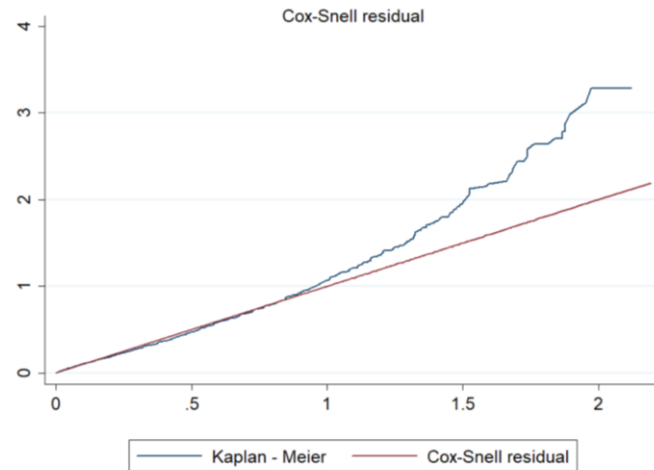
No. of subjects	=	8320	Number of obs	=	8320
No. of failures	=	1794			
Time at risk	=	135402			
Log pseudolikelihood	=	-5078.4136	Wald chi2(4)	=	1432.86
			Prob > chi2	=	0.0000

_t	Robust		z	P> z	[95% Conf. Interval]	
	Coef.	Std. Err.				
EBITDAAN	.0201213	.0030182	6.67	0.000	.0142057	.0260369
PatrnettotangAN	.0076838	.0006937	11.08	0.000	.0063242	.0090433
OFNEBIT	-.0023642	.0002026	-11.67	0.000	-.0027613	-.0019672
PCric	-9.41e-07	1.81e-07	-5.19	0.000	-1.30e-06	-5.86e-07
_cons	4.419782	.0545401	81.04	0.000	4.312886	4.526679
/ln_gam	-.202638	.016857	-12.02	0.000	-.2356772	-.1695988
gamma	.8165738	.013765			.7900357	.8440034

La funzione di sopravvivenza è coerente con la stima delle altre distribuzioni e , per quanto riguarda la funzione di rischio, un valore di gamma positivo indica una funzione prima crescente e poi decrescente rispetto al tempo, come si può vedere dal grafico sottostante.



Come per il caso precedente, questa distribuzione non è particolarmente indicata per i nostri scopi, il grafico dei residui sottostante evidenzia lo scarso adattamento del modello ai dati.



Distribuzione Gamma

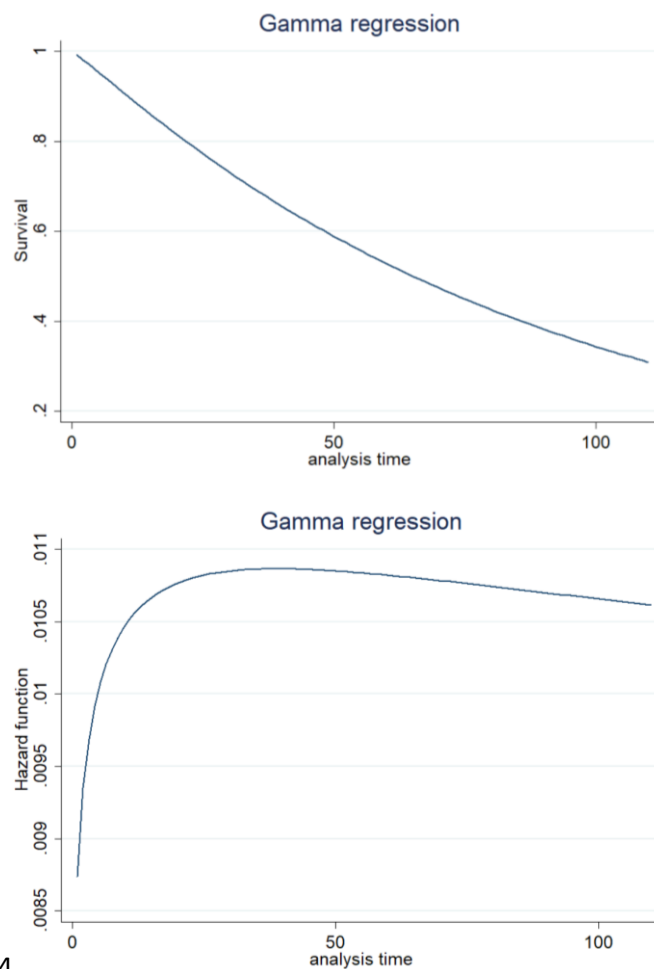
Infine, per la distribuzione gamma, caratterizzata da due parametri Υ e λ , è applicabile soltanto il metodo AFT. I risultati di regressione sono tutti statisticamente significativi.

Gamma regression -- accelerated failure-time form

No. of subjects	=	8320	Number of obs	=	8320
No. of failures	=	1794			
Time at risk	=	135402			
			Wald chi2(4)	=	1379.32
Log pseudolikelihood	=	-5047.0745	Prob > chi2	=	0.0000

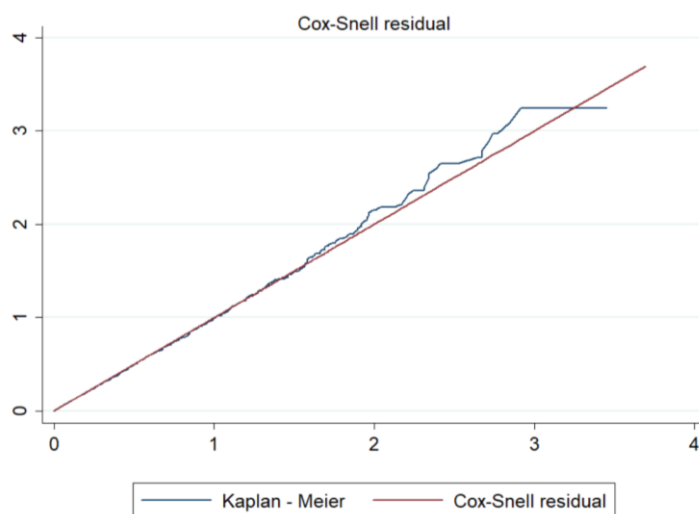
_t	Robust					[95% Conf. Interval]	
	Coef.	Std. Err.	z	P> z			
EBITDAAN	.0203342	.0028185	7.21	0.000	.01481	.0258584	
PatrnettotangAN	.0075878	.0006927	10.95	0.000	.0062302	.0089454	
OFNEBIT	-.0024041	.0001948	-12.34	0.000	-.0027858	-.0020224	
PCric	-8.39e-07	1.71e-07	-4.90	0.000	-1.18e-06	-5.04e-07	
_cons	4.756837	.058993	80.63	0.000	4.641213	4.872462	
/ln_sig	.0149801	.0425338	0.35	0.725	-.0683846	.0983448	
/kappa	.8878598	.0700812	12.67	0.000	.7505032	1.025216	
sigma	1.015093	.0431757			.9339013	1.103343	

La distribuzione gamma è caratterizzata da un valore Υ positivo; per questo motivo la funzione di rischio cresce per poi decrescere lentamente superata la mediana.



4

Dal grafico dei residui si evince che la distribuzione è adeguata al modello che stiamo applicando, tuttavia, secondo questo criterio la distribuzione di Weibull è quella che fornisce il miglior adattamento ai dati.



Cox Proportional Hazard Model

Considerato che per la maggior parte dei casi la forma esatta della distribuzione di sopravvivenza non è nota, potrebbe risultare molto difficile trovare il modello più appropriato. Il modello di Cox ci permette di superare questa limitazione assumendo che le funzioni di rischio delle diverse imprese siano proporzionali tra di loro e indipendenti dal tempo. La funzione di verosimiglianza parziale ci consente di ottenere degli stimatori migliori, come è possibile notare dal valore del logaritmo della verosimiglianza parziale rispetto a quello delle distribuzioni parametriche basate sulla funzione di verosimiglianza. Anche in questo caso, in presenza di osservazioni legate, si utilizza l'approssimazione proposta da Breslow.

Di seguito è riportato l'output generato da STATA, in funzione dei quattro fattori prognostici precedentemente scelti, espresso sia in termini di tasso di rischio che di coefficienti.

Cox regression -- Breslow method for ties

No. of subjects =	8320	Number of obs =	8320
No. of failures =	1794		
Time at risk =	135402		
Log likelihood =	-14141.382	LR chi2(4) =	1507.06
		Prob > chi2 =	0.0000

	_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
EBITDAAN		.9796436	.0025582	-7.88	0.000	.9746423 .9846705
Patrnettotangan		.9925989	.0006381	-11.56	0.000	.9913491 .9938502
OFNEBIT		1.002524	.0002064	12.25	0.000	1.00212 1.002929
PCric		1.000001	1.71e-07	4.89	0.000	1.000001 1.000001

Cox regression -- Breslow method for ties

No. of subjects =	8320	Number of obs =	8320
No. of failures =	1794		
Time at risk =	135402		
Log likelihood =	-14141.382	LR chi2(4) =	1507.06
		Prob > chi2 =	0.0000

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
EBITDAAN	-.0205665	.0026114	-7.88	0.000	-.0256847	-.0154483
PatrnettottangAN	-.0074287	.0006428	-11.56	0.000	-.0086886	-.0061688
OFNEBIT	.002521	.0002059	12.25	0.000	.0021175	.0029245
PCric	8.38e-07	1.71e-07	4.89	0.000	5.02e-07	1.17e-06

Dal test di significatività globale è possibile notare come il gruppo di variabili indipendenti preveda in modo affidabile la variabile dipendente.

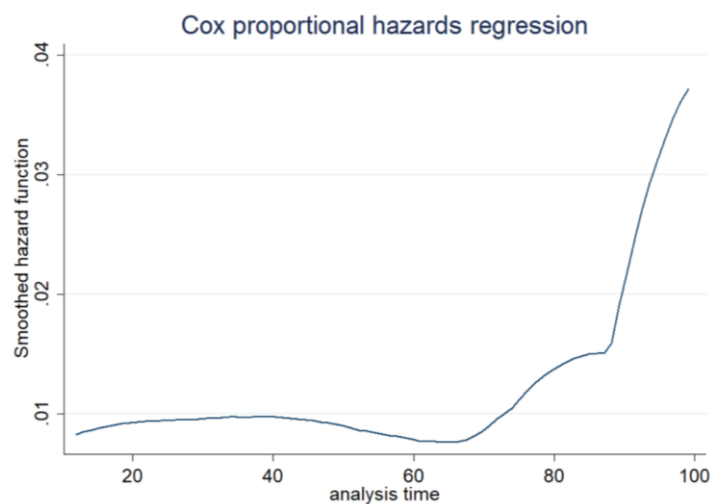
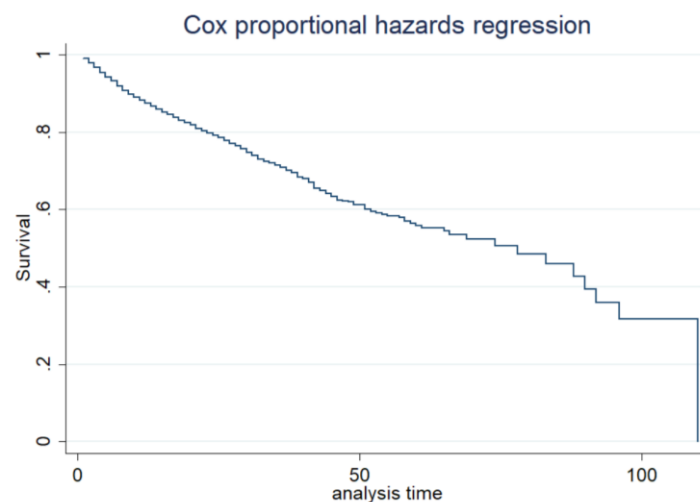
Come si può notare, il segno dei coefficienti è opposto rispetto a quello che si ottiene con la stima PH del metodo parametrico. Infatti, il segno positivo dei coefficienti di regressione indica che le aziende con valori molto alti di OFN/EBIT e PC/RIC sono soggette ad un rischio molto più alto di default. Le imprese con un segno negativo ed elevati valori di EBITDA/AN e Patrimoniottotang/an sono quelle che hanno una più bassa probabilità di fallimento.

I quattro indicatori sono significativamente correlati alla sopravvivenza ($p < 0.5$) ed i risultati sono consistenti con quelli ottenuti con la distribuzione di Weibull, essendo essa stessa un caso particolare dei modelli di rischio proporzionale.

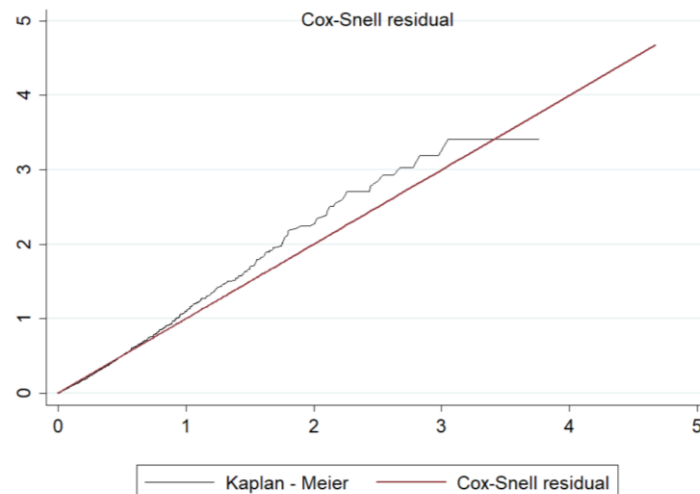
L'elevato numero di aziende determinare un basso valore nello standard error (che rappresenta la precisione con cui è misurato il coefficiente di regressione) di ciascuna variabile e di conseguenza un piccolo intervallo di confidenza, per questo motivo possiamo considerare le stime come affidabili e il modello robusto.

Di seguito sono riportati il grafico di sopravvivenza e di rischio ottenuti attraverso il modello dei rischi proporzionali. La funzione di sopravvivenza è coerente con quanto

precedentemente valutato con l'approccio non parametrico, mentre, invece, la funzione di rischio mostra un andamento "più dettagliato" del tasso di rischio. Non si tratta semplicemente di una funzione monotona crescente o decrescente, ma di una funzione che fino ai 40 anni di attività ha un andamento crescente, seppure si mantenga al di sotto dello 0.01; tra i 40 ed i 65 anni vi è una piccola riduzione nel tasso di rischio per poi diventare una funzione di rischio crescente dal 65 anni in poi. Dopo circa 90 anni di attività, il rischio tende a crescere più velocemente.



In ultimo, applicando il modello dei residui di Cox Snell al modello dei rischi proporzionali, notiamo che la configurazione si avvicina abbastanza ad una linea di 45°, suggerendo che il modello di Cox fornisce un adattamento ragionevole. Lo scostamento che si osserva rispetto alla linea retta, è in parte dovuto ai survival bias di cui il modello è affetto.



Conclusioni

Nel corso della seguente trattazione si è cercato di individuare quale modello fosse più idoneo per descrivere l'esperienza di sopravvivenza e/o di default di un certo numero di imprese operanti nel settore delle apparecchiature elettriche. Si è cercato di quantificare la probabilità di subire il rischio di credito da parte dell'impresa; tale strumento può essere particolarmente utile per prepararsi ed eventualmente intraprendere delle azioni correttive. Lo studio riguarda le società di capitali operanti in questo settore dal 1995 al 2019; molte di queste hanno subito diverse procedure nel corso di questi anni, una parte è andata in default ed una parte risulta essere ancora attiva. In quest'ultimo caso si parla di osservazioni censurate; pur trattandosi di fatto di dati mancanti, anche queste sono state incluse nello studio in quanto in grado di apportare informazioni significative allo studio.

Il tempo di sopravvivenza e le sue funzioni sono state valutate inizialmente attraverso il metodo non parametrico, ovvero la stima di Kaplan-Meier e le tavole di vita. Per ogni anno di bilancio e per ogni intervallo, rispettivamente, sono state valutate la funzione di sopravvivenza e quella di rischio. Il quadro che emerge è quello di una sopravvivenza che decresce nel tempo seguendo una funzione a gradini, rallentando la sua decrescita negli ultimi anni in esame e una funzione di rischio piatta per più della metà dello studio (prossima ad un valore molto piccolo pari allo 0.01), che poi cresce molto velocemente fino a valori pari allo 0.05. Questo ci suggerisce che siamo di fronte ad un settore non particolarmente rischioso e piuttosto stabile. L'analisi si è svolta poi considerando separatamente micro, piccole, medie e grandi imprese. Quello che emerge è che le imprese più grandi sono le più solide, quelle micro sono quelle maggiormente esposte al rischio di fallimento. Considerando l'area geografica di appartenenza, le imprese del Nord-Ovest sono quelle più stabili a discapito di quelle del Sud Italia e delle isole.

La fase successiva dello studio prevede la ricerca di una distribuzione teorica rappresentativa della sopravvivenza delle imprese, introducendo quattro indicatori economici-patrimoniali individuati come i più rappresentativi dello status della società. Per l'approccio parametrico sono state usate le più note distribuzioni: esponenziale, Weibull, log-logistica, log-normale, gamma e di Gompertz. Sebbene sia la distribuzione di Weibull che la distribuzione Gamma diano risultati soddisfacenti, la distribuzione di Weibull è quella che meglio descrive l'esperienza di sopravvivenza delle imprese in base anche ai risultati ottenuti con il metodo dei residui di Cox-Snell.

Infine, il metodo dei rischi proporzionali di Cox surclassa gli altri due. I risultati di verosimiglianza ottenuti sono migliori rispetto a quelli ottenuti con i metodi parametrici, seppur coerenti. Questo metodo è quello maggiormente indicato per un'analisi di questo tipo. Trattandosi di un modello regressivo, può essere facilmente utilizzato come strumento per fare previsioni sul futuro delle imprese considerando le informazioni presenti nei bilanci delle stesse.

A causa delle limitazioni imposte dal software, l'analisi termina qui. Tuttavia, un possibile miglioramento dello studio consiste nel considerare le time-varying covariates, ossia considerare lo sviluppo temporale dei fattori prognostici; in questo modo viene meno l'assunzione del modello proporzionale che vuole le variabili indipendenti dal tempo ed è più semplice fare previsioni sul futuro. Inoltre, potrebbe essere particolarmente utile inserire variabili che tengano conto della situazione economico-finanziaria di un determinato Paese e del livello di rischio di un ben preciso settore.

n = 18																			
n	r	t ₍₁₁₎	t ₍₁₂₎	t ₍₁₃₎	t ₍₁₄₎	t ₍₁₅₎	t ₍₁₆₎	t ₍₁₇₎	t ₍₁₈₎	t ₍₁₉₎	t ₍₁₁₀₎	t ₍₁₁₁₎	t ₍₁₁₂₎	t ₍₁₁₃₎	t ₍₁₁₄₎	t ₍₁₁₅₎	t ₍₁₁₆₎	t ₍₁₁₇₎	t ₍₁₁₈₎
0	μ	-.0556	-.0556	-.0556	-.0556	-.0556	-.0556	-.0556	-.0556	-.0556	-.0556	-.0556	-.0556	-.0556	-.0556	-.0556	-.0556	-.0556	-.0556
	σ	-.1235	-.0822	-.0645	-.0512	-.0401	-.0302	-.0211	-.0125	-.0041	-.0041	-.0125	-.0211	-.0302	-.0401	-.0512	-.0645	-.0822	-.1235
1	μ	-.0509	-.0526	-.0534	-.0540	-.0544	-.0548	-.0552	-.0556	-.0559	-.0563	-.0566	-.0570	-.0574	-.0577	-.0582	-.0586	-.0586	-.0586
	σ	-.1338	-.0887	-.0693	-.0548	-.0426	-.0318	-.0219	-.0124	-.0033	-.0058	-.0149	-.0243	-.0342	-.0450	-.0570	-.0712	-.0861	-.1338
2	μ	-.0449	-.0489	-.0507	-.0520	-.0531	-.0540	-.0549	-.0558	-.0566	-.0574	-.0582	-.0590	-.0598	-.0607	-.0616	-.0623	-.0623	-.0623
	σ	-.1449	-.0955	-.0743	-.0584	-.0451	-.0333	-.0224	-.0121	-.0021	-.0078	-.0178	-.0281	-.0389	-.0505	-.0634	-.0788	-.0955	-.1449
3	μ	-.0373	-.0443	-.0474	-.0496	-.0515	-.0532	-.0547	-.0562	-.0576	-.0589	-.0603	-.0617	-.0631	-.0646	-.0660	-.0674	-.0674	-.0674
	σ	-.1573	-.1030	-.0798	-.0623	-.0477	-.0342	-.0228	-.0115	-.0005	-.0103	-.0213	-.0325	-.0442	-.0568	-.0706	-.0855	-.1030	-.1573
4	μ	-.0279	-.0387	-.0433	-.0468	-.0497	-.0522	-.0546	-.0568	-.0589	-.0610	-.0631	-.0652	-.0674	-.0698	-.0724	-.0750	-.0750	-.0750
	σ	-.1713	-.1114	-.0858	-.0665	-.0504	-.0361	-.0230	-.0105	-.0015	-.0135	-.0254	-.0377	-.0505	-.0639	-.0788	-.0948	-.1114	-.1713
5	μ	-.0161	-.0317	-.0384	-.0434	-.0475	-.0512	-.0546	-.0578	-.0609	-.0639	-.0669	-.0698	-.0729	-.0759	-.0789	-.0819	-.0819	-.0819
	σ	-.1872	-.1209	-.0925	-.0711	-.0533	-.0373	-.0230	-.0102	-.0002	-.0173	-.0305	-.0440	-.0578	-.0729	-.0888	-.1058	-.1209	-.1872
6	μ	-.0013	-.0230	-.0323	-.0392	-.0450	-.0502	-.0549	-.0593	-.0636	-.0677	-.0718	-.0758	-.0798	-.0838	-.0878	-.0918	-.0918	-.0918
	σ	-.2058	-.1318	-.1001	-.0763	-.0564	-.0388	-.0226	-.0073	-.0005	-.0221	-.0367	-.0518	-.0674	-.0838	-.1009	-.1188	-.1318	-.2058
7	μ	-.0176	-.0121	-.0247	-.0342	-.0421	-.0491	-.0555	-.0615	-.0673	-.0729	-.0782	-.0835	-.0888	-.0941	-.0994	-.1047	-.1047	-.1047
	σ	-.2278	-.1445	-.1089	-.0821	-.0598	-.0400	-.0219	-.0047	-.0005	-.0282	-.0439	-.0609	-.0788	-.0974	-.1168	-.1368	-.1445	-.2278
8	μ	-.0419	-.0019	-.0153	-.0281	-.0387	-.0482	-.0568	-.0648	-.0726	-.0794	-.0862	-.0930	-.0998	-.1066	-.1134	-.1202	-.1202	-.1202
	σ	-.2543	-.1597	-.1191	-.0888	-.0635	-.0411	-.0205	-.0011	-.0005	-.0282	-.0439	-.0609	-.0788	-.0974	-.1168	-.1368	-.1597	-.2543
9	μ	-.0741	-.0200	-.0031	-.0204	-.0348	-.0474	-.0590	-.0698	-.0797	-.0888	-.0974	-.1066	-.1158	-.1250	-.1342	-.1434	-.1434	-.1434
	σ	-.2869	-.1781	-.1315	-.0966	-.0676	-.0418	-.0183	-.0039	-.0005	-.0282	-.0439	-.0609	-.0788	-.0974	-.1168	-.1368	-.1781	-.2869
10	μ	-.1176	-.0441	-.0128	-.0106	-.0100	-.0471	-.0627	-.0741	-.0824	-.0907	-.0990	-.1073	-.1156	-.1239	-.1322	-.1405	-.1405	-.1405
	σ	-.3283	-.2010	-.1466	-.1059	-.0720	-.0421	-.0147	-.0007	-.0005	-.0282	-.0439	-.0609	-.0788	-.0974	-.1168	-.1368	-.2010	-.3283
11	μ	-.1788	-.0735	-.0143	-.0032	-.0345	-.0479	-.0590	-.0682	-.0765	-.0848	-.0930	-.1013	-.1096	-.1179	-.1262	-.1345	-.1345	-.1345
	σ	-.3827	-.2307	-.1657	-.1173	-.0770	-.0414	-.0118	-.0005	-.0005	-.0282	-.0439	-.0609	-.0788	-.0974	-.1168	-.1368	-.2307	-.3827
12	μ	-.2689	-.1257	-.0648	-.0195	-.0180	-.1460	-.1609	-.1758	-.1907	-.2056	-.2205	-.2354	-.2503	-.2652	-.2801	-.2950	-.2950	-.2950
	σ	-.4576	-.2707	-.1911	-.1317	-.0824	-.0414	-.0118	-.0005	-.0005	-.0282	-.0439	-.0609	-.0788	-.0974	-.1168	-.1368	-.2707	-.4576
13	μ	-.4113	-.2001	-.1106	-.0442	-.1762	-.2158	-.2554	-.2950	-.3346	-.3742	-.4138	-.4534	-.4930	-.5326	-.5722	-.6118	-.6118	-.6118
	σ	-.5681	-.3285	-.2266	-.1569	-.1042	-.0642	-.0346	-.0147	-.0005	-.0282	-.0439	-.0609	-.0788	-.0974	-.1168	-.1368	-.3285	-.5681
14	μ	-.6615	-.3276	-.1867	-.1158	-.0758	-.0452	-.0246	-.0140	-.0034	-.0028	-.0022	-.0016	-.0010	-.0004	-.0004	-.0004	-.0004	-.0004
	σ	-.7436	-.4205	-.2815	-.14505	-.09505	-.06405	-.04305	-.02205	-.01105	-.0005	-.0005	-.0005	-.0005	-.0005	-.0005	-.0005	-.4205	-.7436
15	μ	-.11905	-.5891	2.7796															
	σ	-.11013	-.5948	1.6960															
16	μ	-.28756	3.8756																
	σ	-.21294	2.1294																

ALLEGATO B

Table B-9 Variances and Covariances of the Best Linear Estimates of the Mean ($\hat{\mu}$) and Standard Deviation ($\hat{\sigma}$) for Censored Samples Up to Size 20 from a Normal Population

		n-r												
n		0	1	2	3	4	5	6	7	8	9	10	11	12
2	V($\hat{\mu}$)	.5000												
	V($\hat{\sigma}$)	.5708												
	Cov($\hat{\mu}, \hat{\sigma}$)	0												
3	V($\hat{\mu}$)	.3333	.4487											
	V($\hat{\sigma}$)	.2755	.6378											
	Cov($\hat{\mu}, \hat{\sigma}$)	0	.2044											
4	V($\hat{\mu}$)	.2500	.2870	.5130										
	V($\hat{\sigma}$)	.1801	.3021	.6730										
	Cov($\hat{\mu}, \hat{\sigma}$)	0	.0672	.3567										
5	V($\hat{\mu}$)	.2000	.2177	.2839	.6112									
	V($\hat{\sigma}$)	.1333	.1948	.3181	.6957									
	Cov($\hat{\mu}, \hat{\sigma}$)	0	.0330	.1234	.4749									
6	V($\hat{\mu}$)	.1667	.1769	.2068	.2999	.7186								
	V($\hat{\sigma}$)	.1057	.1428	.2044	.3292	.7119								
	Cov($\hat{\mu}, \hat{\sigma}$)	0	.0195	.0624	.1702	.5705								
7	V($\hat{\mu}$)	.1429	.1494	.1660	.2071	.3248	.8264							
	V($\hat{\sigma}$)	.0875	.1123	.1493	.2114	.3375	.7243							
	Cov($\hat{\mu}, \hat{\sigma}$)	0	.0128	.0375	.0881	.2099	.6503							
8	V($\hat{\mu}$)	.1250	.1295	.1399	.1623	.2138	.3541	.9310						
	V($\hat{\sigma}$)	.0746	.0924	.1171	.1542	.2168	.3441	.7342						
	Cov($\hat{\mu}, \hat{\sigma}$)	0	.0090	.0250	.0538	.1106	.2442	.7186						
9	V($\hat{\mu}$)	.1111	.1144	.1214	.1352	.1629	.2241	.3854	1.0313					
	V($\hat{\sigma}$)	.0650	.0784	.0960	.1207	.1581	.2212	.3494	.7423					
	Cov($\hat{\mu}, \hat{\sigma}$)	0	.0067	.0178	.0362	.0684	.1305	.2743	.7781					
10	V($\hat{\mu}$)	.1000	.1025	.1075	.1167	.1336	.1664	.2366	.4174	1.1269				
	V($\hat{\sigma}$)	.0576	.0681	.0813	.0949	.1237	.1613	.2248	.3539	.7491				
	Cov($\hat{\mu}, \hat{\sigma}$)	0	.0051	.0132	.0260	.0465	.0816	.1483	.3011	.8306				
11	V($\hat{\mu}$)	.0909	.0929	.0966	.1031	.1143	.1342	.1718	.2504	.4493	1.2179			
	V($\hat{\sigma}$)	.0517	.0601	.0704	.0836	.1013	.1262	.1640	.2279	.3577	.7550			
	Cov($\hat{\mu}, \hat{\sigma}$)	0	.0041	.0102	.0195	.0336	.0559	.0936	.1644	.3251	.8777			
12	V($\hat{\mu}$)	.0833	.0849	.0878	.0926	.1004	.1136	.1363	.1784	.2650	.4809	1.3044		
	V($\hat{\sigma}$)	.0469	.0538	.0620	.0723	.0855	.1033	.1283	.1663	.2305	.3610	.7601		
	Cov($\hat{\mu}, \hat{\sigma}$)	0	.0033	.0081	.0152	.0254	.0406	.0645	.1045	.1791	.3469	.9202		

Table B-9 (continued)															
13	V($\hat{\mu}$)	.0769	.0782	.0805	.0841	.0899	.0991	.1141	.1395	.1858	.2799	.5118	1.3868		
	V($\hat{\sigma}$)	.0429	.0486	.0554	.0636	.0739	.0872	.1050	.1302	.1688	.2329	.3640	.7646		
	Cov($\hat{\mu}, \hat{\sigma}$)	0	.0027	.0066	.0121	.0198	.0309	.0472	.0725	.1146	.1925	.3668	.9589		
14	V($\hat{\mu}$)	.0714	.0725	.0743	.0772	.0816	.0883	.0988	.1155	.1435	.1938	.2950	.5419	1.4654	
	V($\hat{\sigma}$)	.0395	.0444	.0500	.0568	.0650	.0753	.0886	.1065	.1318	.1702	.2350	.3666	.7687	
	Cov($\hat{\mu}, \hat{\sigma}$)	0	.0023	.0055	.0099	.0159	.0242	.0360	.0533	.0799	.1239	.2048	.3851	.9945	
15	V($\hat{\mu}$)	.0667	.0676	.0691	.0714	.0748	.0799	.0875	.0992	.1176	.1480	.2022	.3101	.5713	1.5404
	V($\hat{\sigma}$)	.0366	.0408	.0456	.0512	.0580	.0662	.0765	.0899	.1107	.1332	.1718	.2368	.3689	.7723
	Cov($\hat{\mu}, \hat{\sigma}$)	0	.0019	.0046	.0082	.0130	.0195	.0284	.0409	.0590	.0868	.1325	.2163	.4020	1.0273

		n-r																
n		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
16	V($\hat{\mu}$)	.0625	.0633	.0645	.0664	.0692	.0731	.0789	.0874	.1002	.1202	.1529	.2107	.3250	.5998	1.6121		
	V($\hat{\sigma}$)	.0341	.0378	.0419	.0467	.0523	.0590	.0672	.0776	.0910	.1090	.1345	.1732	.2365	.3710	.7756		
	Cov($\hat{\mu}, \hat{\sigma}$)	0	.0017	.0039	.0069	.0109	.0160	.0229	.0323	.0454	.0644	.0933	.1406	.2269	.4177	1.0578		
17	V($\hat{\mu}$)	.0588	.0595	.0605	.0621	.0644	.0675	.0720	.0784	.0877	.1016	.1231	.1581	.2194	.3398	.6274	1.6808	
	V($\hat{\sigma}$)	.0320	.0352	.0387	.0428	.0476	.0532	.0599	.0682	.0786	.0920	.1101	.1357	.1745	.2400	.3729	.7786	
	Cov($\hat{\mu}, \hat{\sigma}$)	0	.0015	.0034	.0059	.0092	.0134	.0189	.0262	.0360	.0497	.0694	.0993	.1481	.2369	.4324	1.0862	
18	V($\hat{\mu}$)	.0556	.0561	.0570	.0584	.0602	.0628	.0664	.0713	.0784	.0885	.1034	.1264	.1635	.2282	.3544	.6543	1.7467
	V($\hat{\sigma}$)	.0301	.0329	.0360	.0395	.0436	.0483	.0540	.0607	.0690	.0794	.0929	.1111	.1368	.1757	.2414	.3746	.7814
	Cov($\hat{\mu}, \hat{\sigma}$)	0	.0013	.0030	.0051	.0079	.0114	.0158	.0216	.0293	.0395	.0538	.0742	.1050	.1552	.2463	.4462	1.1127

		n-r																		
n		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
19	V($\hat{\mu}$)	.0526	.0531	.0539	.0550	.0566	.0587	.0616	.0656	.0710	.0787	.0895	.1055	.1299	.1691	.2369	.3687	.6804	1.8101	
	V($\hat{\sigma}$)	.0284	.0309	.0336	.0367	.0403	.0443	.0491	.0547	.0614	.0698	.0802	.0938	.1120	.1377	.1768	.2426	.3763	.7839	
	Cov($\hat{\mu}, \hat{\sigma}$)	0	.0011	.0026	.0045	.0068	.0098	.0135	.0182	.0242	.0322	.0429	.0576	.0787	.1104	.1619	.2551	.4591	1.1377	
20	V($\hat{\mu}$)	.0500	.0504	.0511	.0521	.0534	.0552	.0576	.0603	.0651	.0710	.0792	.0909	.1078	.1335	.1747	.2456	.3828	.7056	1.8710
	V($\hat{\sigma}$)	.0268	.0291	.0316	.0343	.0374	.0409	.0450	.0497	.0553	.0621	.0705	.0810	.0945	.1128	.1386	.1778	.2438	.3778	.7863
	Cov($\hat{\mu}, \hat{\sigma}$)	0	.0010	.0023	.0039	.0060	.0085	.0116	.0155	.0204	.0267	.0350	.0460	.0612	.0829	.1155	.1682	.2634	.4713	1.1613

Source: Up to $n = 15$ of this table is reproduced from A. E. Sarhan and B. G. Greenberg, "Estimation of Location and Scale Parameters by Order Statistics from Singly and Censored Samples, Parts I and II," *Ann. Math. Statist.*, Vol. 27, pp. 427-451 (1956), and Vol. 29, pp. 79-105 (1958), with permission of the editor of the *Annals of Mathematical Statistics*. The rest of the table is produced from A. E. Sarhan and B. G. Greenberg, "Estimation of Location and Scale Parameters by Order Statistics from Singly and Doubly Censored Samples, Part III," Tech. Rep. 4-OOR, Project 1597, U.S. Army Research Office.

ALLEGATO C

y	m_1	m_2	m_3	Percentage Censored
-4.0	1.00000	0.500030	0.000006	0.00
-3.5	1.00001	0.500208	0.000052	0.02
-3.0	1.00010	0.501180	0.000335	0.13
-2.5	1.00056	0.505280	0.001712	0.62
-2.4	1.00078	0.506935	0.002312	0.82
-2.3	1.00107	0.509030	0.003099	1.07
-2.2	1.00147	0.511658	0.004121	1.39
-2.1	1.00200	0.514926	0.005438	1.79
-2.0	1.00270	0.518960	0.007123	2.28
-1.9	1.00363	0.523899	0.009266	2.87
-1.8	1.00485	0.529899	0.011971	3.59
-1.7	1.00645	0.537141	0.015368	4.46
-1.6	1.00852	0.545827	0.019610	5.48
-1.5	1.01120	0.556186	0.024884	6.68
-1.4	1.01467	0.568417	0.031410	8.08
-1.3	1.01914	0.582981	0.039460	9.68
-1.2	1.02488	0.600046	0.049355	11.51
-1.1	1.03224	0.620049	0.061491	13.57
-1.0	1.04168	0.643438	0.076345	15.87
-0.9	1.05376	0.670724	0.094501	18.41
-0.8	1.06923	0.702513	0.116674	21.19
-0.7	1.08904	0.739515	0.143744	24.20
-0.6	1.11442	0.782574	0.176698	27.43
-0.5	1.14696	0.832691	0.217183	30.85
-0.4	1.18876	0.891077	0.266577	34.46
-0.3	1.24252	0.959181	0.327080	38.21
-0.2	1.31180	1.03877	0.401326	42.07
-0.1	1.40127	1.13198	0.492641	46.02
0.0	1.51709	1.24145	0.605233	50.00
0.1	1.66743	1.37042	0.744459	53.98
0.2	1.86310	1.52288	0.917165	57.93
0.3	2.11857	1.70381	1.13214	61.79
0.4	2.45318	1.91942	1.40071	65.54
0.5	2.89293	2.17751	1.73757	69.15
0.6	3.47293	2.48793	2.16185	72.57
0.7	4.24075	2.86318	2.69858	75.80
0.8	5.2612	3.3192	3.3807	78.81
0.9	6.6229	3.8765	4.2517	81.59
1.0	8.4477	4.5614	5.3696	84.13
1.1	10.903	5.4082	6.8116	86.43
1.2	14.224	6.4616	8.6818	88.49
1.3	18.735	7.7804	11.121	90.32
1.4	24.892	9.4423	14.319	91.92

Table B-10 $1/(1-R)$ and $\hat{\gamma}$ for the Estimation of the Parameters of the Gamma Distribution When There Are No Censored Observations

$1/(1-R)$	$\hat{\gamma}$	$1/(1-R)$	$\hat{\gamma}$	$1/(1-R)$	$\hat{\gamma}$
1.000	0.00000	1.11	0.29854	2.4	1.06335
1.001	0.11539	1.12	0.30749	2.5	1.11596
1.002	0.12663	1.13	0.31616	2.6	1.16833
1.003	0.13434	1.14	0.32461	2.7	1.22050
1.004	0.14043	1.15	0.33285	2.8	1.27248
1.005	0.14556	1.16	0.34090	2.9	1.32430
1.006	0.15005	1.18	0.35654	3.0	1.37599
1.007	0.15408	1.20	0.37165	3.2	1.47899
1.008	0.15775	1.22	0.38631	3.4	1.58158
1.009	0.16115	1.24	0.40060	3.6	1.68385
1.010	0.16432	1.26	0.41457	3.8	1.78585
1.012	0.17011	1.28	0.42825	4.0	1.88763
1.014	0.17535	1.30	0.44170	4.2	1.98921
1.016	0.18017	1.32	0.45492	4.4	2.09063
1.018	0.18464	1.34	0.46795	4.6	2.19191
1.020	0.18884	1.36	0.48081	4.8	2.29308
1.022	0.19282	1.38	0.49351	5.0	2.39414
1.024	0.19659	1.40	0.50608	5.5	2.64643
1.026	0.20020	1.45	0.53694	6.0	2.89830
1.028	0.20366	1.50	0.56715	6.5	3.14984
1.030	0.20700	1.55	0.59682	7.0	3.40115
1.035	0.21486	1.60	0.62604	8.0	3.90325
1.040	0.22217	1.65	0.65488	9.0	4.40482
1.045	0.22904	1.70	0.68340	10.0	4.90608
1.050	0.23554	1.75	0.71163	12.0	5.90790
1.055	0.24175	1.80	0.73961	14.0	6.90914
1.060	0.24771	1.85	0.76737	16.0	7.91005
1.065	0.25345	1.90	0.79494	18.0	8.91073
1.070	0.25899	1.95	0.82233	20.0	9.91125
1.075	0.26437	2.00	0.84957	30.0	14.91301
1.080	0.26959	2.10	0.90364	40.0	19.91401
1.090	0.27966	2.20	0.95724	50.0	24.91457
1.100	0.28928	2.30	1.01046		

Source: "Estimation of Parameters of the Gamma Distribution Using Order Statistics," by M. B. Wilk, R. Gnanadesikan, and Marilyn J. Huyett, *Biometrika*, Vol. 49, pp. 525-545 (1962). Reproduced by permission of the editor of *Biometrika*.

Table B-11 $\hat{\gamma}(P, S)$ and $\hat{\mu}(P, S)$ for Various Values of n/r : $n/r = 1.0$

			0.96	0.92	0.88	0.84	0.80	0.76	0.72	0.68	0.64	0.60	0.56	P/S
0.08	γ	0.850	12.412	6.158	4.070	3.024	2.394	1.972	1.669	1.440	1.261	1.116	0.996	1.00
	μ	0.080	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	
0.12	γ	0.567	1.376	11.912	5.908	3.903	2.898	2.293	1.888	1.596	1.376	1.203	1.063	0.96
	μ	0.120	0.120	0.960	0.960	0.960	0.960	0.960	0.960	0.960	0.960	0.960	0.960	
0.16	γ	0.464	0.850	1.887	11.412	5.657	3.736	2.772	2.192	1.803	1.523	1.311	1.145	0.92
	μ	0.160	0.160	0.160	0.920	0.920	0.920	0.920	0.920	0.920	0.920	0.920	0.920	
0.20	γ	0.408	0.664	1.116	2.394	10.912	5.407	3.509	2.646	2.091	1.718	1.450	1.247	0.88
	μ	0.200	0.200	0.200	0.200	0.880	0.880	0.880	0.880	0.880	0.880	0.880	0.880	
0.24	γ	0.372	0.567	0.850	1.376	2.898	10.411	5.157	3.401	2.520	1.989	1.633	1.376	0.84
	μ	0.240	0.240	0.240	0.240	0.240	0.840	0.840	0.840	0.840	0.840	0.840	0.840	
0.28	γ	0.346	0.506	0.712	1.028	1.633	3.401	9.911	4.906	3.234	2.394	1.888	1.547	0.80
	μ	0.280	0.280	0.280	0.280	0.280	0.280	0.800	0.800	0.800	0.800	0.800	0.800	
0.32	γ	0.327	0.464	0.626	0.850	1.203	1.888	3.903	9.411	4.655	3.066	2.268	1.786	0.76
	μ	0.320	0.320	0.320	0.320	0.320	0.320	0.320	0.760	0.760	0.760	0.760	0.760	
0.36	γ	0.312	0.432	0.567	0.740	0.984	1.376	2.141	4.405	8.911	4.405	2.898	2.141	0.72
	μ	0.360	0.360	0.360	0.360	0.360	0.360	0.360	0.360	0.720	0.720	0.720	0.720	
0.40	γ	0.300	0.408	0.524	0.664	0.850	1.116	1.547	2.394	4.906	8.410	4.154	2.730	0.68
	μ	0.400	0.400	0.400	0.400	0.400	0.400	0.400	0.400	0.400	0.680	0.680	0.680	
0.44	γ	0.288	0.388	0.490	0.609	0.758	0.957	1.247	1.718	2.646	5.407	7.910	3.903	0.64
	μ	0.440	0.440	0.440	0.440	0.440	0.440	0.440	0.440	0.440	0.440	0.640	0.640	
0.48	γ	0.281	0.372	0.464	0.567	0.691	0.850	1.063	1.376	1.888	2.898	5.908	7.409	0.60
	μ	0.480	0.480	0.480	0.480	0.480	0.480	0.480	0.480	0.480	0.480	0.600	0.600	
0.52	γ	0.273	0.358	0.442	0.534	0.641	0.771	0.929	1.168	1.505	2.057	3.150	6.408	
	μ	0.520	0.520	0.520	0.520	0.520	0.520	0.520	0.520	0.520	0.520	0.520	0.520	
0.56	γ	0.266	0.346	0.423	0.506	0.600	0.712	0.850	1.028	1.272	1.633	2.226	3.401	6.909
	μ	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560	0.560
0.60	γ	0.261	0.336	0.408	0.483	0.567	0.664	0.781	0.927	1.116	1.376	1.760	2.394	3.652
	μ	0.600	0.600	0.600	0.600	0.600	0.600	0.600	0.600	0.600	0.600	0.600	0.600	0.600
0.64	γ	0.255	0.327	0.394	0.464	0.540	0.626	0.727	0.850	1.003	1.203	1.479	1.888	2.562
	μ	0.640	0.640	0.640	0.640	0.640	0.640	0.640	0.640	0.640	0.640	0.640	0.640	0.640
0.68	γ	0.251	0.319	0.382	0.447	0.516	0.594	0.683	0.789	0.917	1.078	1.290	1.582	2.015
	μ	0.680	0.680	0.680	0.680	0.680	0.680	0.680	0.680	0.680	0.680	0.680	0.680	0.680
0.72	γ	0.246	0.312	0.372	0.432	0.496	0.567	0.647	0.740	0.850	0.984	1.153	1.376	1.684
	μ	0.720	0.720	0.720	0.720	0.720	0.720	0.720	0.720	0.720	0.720	0.720	0.720	0.720
0.76	γ	0.242	0.305	0.362	0.419	0.479	0.544	0.616	0.699	0.795	0.910	1.050	1.228	1.462
	μ	0.760	0.760	0.760	0.760	0.760	0.760	0.760	0.760	0.760	0.760	0.760	0.760	0.760
0.80	γ	0.239	0.300	0.354	0.408	0.464	0.524	0.590	0.664	0.750	0.850	0.969	1.116	1.302
	μ	0.800	0.800	0.800	0.800	0.800	0.800	0.800	0.800	0.800	0.800	0.800	0.800	0.800
0.84	γ	0.236	0.294	0.348	0.397	0.450	0.506	0.567	0.635	0.712	0.800	0.904	1.028	1.181
	μ	0.840	0.840	0.840	0.840	0.840	0.840	0.840	0.840	0.840	0.840	0.840	0.840	0.840
0.88	γ	0.232	0.289	0.339	0.388	0.438	0.490	0.547	0.609	0.679	0.758	0.850	0.957	1.087
	μ	0.880	0.880	0.880	0.880	0.880	0.880	0.880	0.880	0.880	0.880	0.880	0.880	0.880
0.92	γ	0.230	0.285	0.333	0.379	0.427	0.476	0.529	0.587	0.651	0.722	0.804	0.899	1.010
	μ	0.920	0.920	0.920	0.920	0.920	0.920	0.920	0.920	0.920	0.920	0.920	0.920	0.920
0.96	γ	0.227	0.281	0.327	0.372	0.417	0.464	0.513	0.567	0.626	0.691	0.765	0.850	0.948
	μ	0.960	0.960	0.960	0.960	0.960	0.960	0.960	0.960	0.960	0.960	0.960	0.960	0.960
1.00	γ	0.225	0.277	0.322	0.365	0.408	0.452	0.499	0.549	0.604	0.664	0.732	0.808	0.895
	μ	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
S/P		0.04	0.08	0.12	0.16	0.20	0.24	0.28	0.32	0.36	0.40	0.44	0.48	0.52

For $P \leq 0.52$ read S from the left-hand margin, and for $P \geq 0.56$ read S from the right-hand margin. Note that the figures in region 2 are printed in bold roman type and those in region 3 in bold italic type; the remainder of the table (outside of regions 2 and 3) is region 1.

ALLEGATO F

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf. Int.]	
1	8320	116	325	0.9861	0.0013	0.9833	0.9884
2	7879	141	362	0.9684	0.0019	0.9644	0.9720
3	7376	140	309	0.9500	0.0024	0.9450	0.9546
4	6927	125	286	0.9329	0.0028	0.9271	0.9382
5	6516	114	281	0.9166	0.0032	0.9101	0.9226
6	6121	93	248	0.9026	0.0034	0.8957	0.9092
7	5780	119	229	0.8841	0.0038	0.8764	0.8912
8	5432	90	258	0.8694	0.0040	0.8613	0.8771
9	5084	65	229	0.8583	0.0042	0.8499	0.8663
10	4790	56	200	0.8483	0.0043	0.8395	0.8566
11	4534	56	192	0.8378	0.0045	0.8287	0.8464
12	4286	44	184	0.8292	0.0047	0.8198	0.8381
13	4058	40	150	0.8210	0.0048	0.8114	0.8302
14	3868	43	156	0.8119	0.0049	0.8020	0.8213
15	3669	41	148	0.8028	0.0051	0.7926	0.8125
16	3480	28	133	0.7963	0.0052	0.7860	0.8063
17	3319	36	136	0.7877	0.0053	0.7771	0.7979
18	3147	32	143	0.7797	0.0054	0.7688	0.7902
19	2972	24	174	0.7734	0.0056	0.7623	0.7841
20	2774	26	108	0.7662	0.0057	0.7548	0.7771
21	2640	33	125	0.7566	0.0058	0.7449	0.7678
22	2482	19	110	0.7508	0.0060	0.7389	0.7622
23	2353	23	96	0.7434	0.0061	0.7313	0.7552
24	2234	18	97	0.7375	0.0062	0.7251	0.7494
25	2119	19	109	0.7308	0.0063	0.7182	0.7430
26	1991	20	116	0.7235	0.0065	0.7106	0.7360
27	1855	19	115	0.7161	0.0066	0.7029	0.7288
28	1721	14	103	0.7103	0.0068	0.6968	0.7233
29	1604	19	98	0.7019	0.0069	0.6880	0.7152
30	1487	18	111	0.6934	0.0071	0.6791	0.7071
31	1358	18	105	0.6842	0.0074	0.6695	0.6984
32	1235	16	80	0.6753	0.0076	0.6602	0.6899
33	1139	9	85	0.6700	0.0077	0.6545	0.6849
34	1045	7	83	0.6655	0.0079	0.6498	0.6807
35	955	7	67	0.6606	0.0080	0.6446	0.6761
36	881	9	65	0.6539	0.0083	0.6374	0.6698
37	807	9	69	0.6466	0.0085	0.6296	0.6630
38	729	7	77	0.6404	0.0088	0.6229	0.6572
39	645	11	61	0.6294	0.0092	0.6111	0.6472
40	573	4	62	0.6250	0.0094	0.6063	0.6431
41	507	7	53	0.6164	0.0098	0.5969	0.6353
42	447	11	42	0.6012	0.0106	0.5802	0.6216
43	394	4	28	0.5951	0.0109	0.5734	0.6162
44	362	4	32	0.5886	0.0113	0.5661	0.6103
45	326	5	41	0.5795	0.0118	0.5560	0.6023

ALLEGATO F – CONTINUAZIONE

46	280	4	23	0.5713	0.0123	0.5467	0.5950
47	253	1	28	0.5690	0.0125	0.5441	0.5931
48	224	1	11	0.5665	0.0127	0.5412	0.5909
49	212	3	10	0.5584	0.0133	0.5319	0.5841
50	199	0	13	0.5584	0.0133	0.5319	0.5841
51	186	4	18	0.5464	0.0143	0.5179	0.5740
52	164	2	9	0.5398	0.0149	0.5101	0.5685
53	153	1	18	0.5362	0.0152	0.5059	0.5656
54	134	1	10	0.5322	0.0156	0.5011	0.5623
55	123	1	10	0.5279	0.0161	0.4959	0.5589
56	112	0	11	0.5279	0.0161	0.4959	0.5589
57	101	1	10	0.5227	0.0168	0.4893	0.5549
58	90	2	7	0.5111	0.0183	0.4746	0.5463
59	81	1	8	0.5048	0.0191	0.4667	0.5416
60	72	1	2	0.4977	0.0201	0.4577	0.5364
61	69	1	2	0.4905	0.0211	0.4486	0.5311
62	66	0	8	0.4905	0.0211	0.4486	0.5311
63	58	0	6	0.4905	0.0211	0.4486	0.5311
65	52	1	3	0.4811	0.0227	0.4360	0.5247
66	48	1	7	0.4711	0.0243	0.4227	0.5178
67	40	0	4	0.4711	0.0243	0.4227	0.5178
68	36	0	3	0.4711	0.0243	0.4227	0.5178
69	33	1	1	0.4568	0.0274	0.4023	0.5096
70	31	0	1	0.4568	0.0274	0.4023	0.5096
71	30	0	4	0.4568	0.0274	0.4023	0.5096
72	26	0	1	0.4568	0.0274	0.4023	0.5096
73	25	0	5	0.4568	0.0274	0.4023	0.5096
74	20	1	1	0.4340	0.0343	0.3661	0.4998
75	18	0	3	0.4340	0.0343	0.3661	0.4998
77	15	0	1	0.4340	0.0343	0.3661	0.4998
78	14	1	1	0.4030	0.0437	0.3172	0.4870
82	12	0	1	0.4030	0.0437	0.3172	0.4870
83	11	1	0	0.3663	0.0529	0.2642	0.4687
84	10	0	2	0.3663	0.0529	0.2642	0.4687
85	8	0	1	0.3663	0.0529	0.2642	0.4687
88	7	1	0	0.3140	0.0663	0.1909	0.4448
90	6	1	0	0.2617	0.0731	0.1331	0.4101
92	5	1	0	0.2093	0.0749	0.0864	0.3683
93	4	0	1	0.2093	0.0749	0.0864	0.3683
96	3	1	0	0.1396	0.0758	0.0340	0.3175
98	2	0	1	0.1396	0.0758	0.0340	0.3175
110	1	1	0	0.0000	.	.	.

Interval	Beg. Total	Deaths	Lost	Survival Function	Std. Error	[95% Conf. Int.]	Cum. Failure	Std. Error	[95% Conf. Int.]	Hazard Function	Std. Error	[95% Conf. Int.]
0 5	8657	591	1550	0.9250	0.0030	0.9190 0.9306	0.0750	0.0030	0.0694 0.0810	0.0156	0.0006	0.0143 0.0168
5 10	6516	481	1245	0.8495	0.0043	0.8409 0.8577	0.1505	0.0043	0.1423 0.1591	0.0170	0.0008	0.0155 0.0185
10 15	4790	239	882	0.8028	0.0050	0.7928 0.8124	0.1972	0.0050	0.1876 0.2072	0.0113	0.0007	0.0099 0.0127
15 20	3669	161	734	0.7637	0.0056	0.7525 0.7745	0.2363	0.0056	0.2255 0.2475	0.0100	0.0008	0.0085 0.0115
20 25	2774	119	536	0.7274	0.0063	0.7149 0.7395	0.2726	0.0063	0.2605 0.2851	0.0097	0.0009	0.0080 0.0115
25 30	2119	91	541	0.6916	0.0070	0.6777 0.7051	0.3084	0.0070	0.2949 0.3223	0.0101	0.0011	0.0080 0.0122
30 35	1487	68	464	0.6541	0.0080	0.6383 0.6695	0.3459	0.0080	0.3305 0.3617	0.0111	0.0014	0.0085 0.0138
35 40	955	43	339	0.6183	0.0092	0.6000 0.6361	0.3817	0.0092	0.3639 0.4000	0.0113	0.0017	0.0079 0.0146
40 45	573	30	217	0.5784	0.0111	0.5563 0.5999	0.4216	0.0111	0.4001 0.4437	0.0133	0.0024	0.0086 0.0181
45 50	326	14	113	0.5484	0.0131	0.5222 0.5737	0.4516	0.0131	0.4263 0.4778	0.0107	0.0028	0.0051 0.0163
50 55	199	8	68	0.5218	0.0155	0.4909 0.5516	0.4782	0.0155	0.4484 0.5091	0.0099	0.0035	0.0031 0.0168
55 60	123	5	46	0.4957	0.0186	0.4587 0.5315	0.5043	0.0186	0.4685 0.5413	0.0103	0.0046	0.0013 0.0192
60 65	72	2	18	0.4799	0.0211	0.4380 0.5206	0.5201	0.0211	0.4794 0.5620	0.0065	0.0046	0.0000 0.0154
65 70	52	3	18	0.4465	0.0271	0.3928 0.4986	0.5535	0.0271	0.5014 0.6072	0.0145	0.0083	0.0000 0.0308
70 75	31	1	12	0.4286	0.0313	0.3667 0.4890	0.5714	0.0313	0.5110 0.6333	0.0082	0.0082	0.0000 0.0242
75 80	18	1	5	0.4009	0.0397	0.3230 0.4775	0.5991	0.0397	0.5225 0.6770	0.0133	0.0133	0.0000 0.0395
80 85	12	1	3	0.3628	0.0511	0.2642 0.4619	0.6372	0.0511	0.5381 0.7358	0.0200	0.0200	0.0000 0.0592
85 90	8	1	1	0.3144	0.0631	0.1967 0.4389	0.6856	0.0631	0.5611 0.8033	0.0286	0.0285	0.0000 0.0844
90 95	6	2	1	0.2001	0.0760	0.0777 0.3631	0.7999	0.0760	0.6369 0.9223	0.0889	0.0613	0.0000 0.2090
95 100	3	1	1	0.1200	0.0769	0.0216 0.3097	0.8800	0.0769	0.6903 0.9784	0.1000	0.0968	0.0000 0.2898
110 115	1	1	0	0.0000	.	.	10.000	.	.	0.4000	.	.

Interval	Beg. Total	Deaths	Lost	Survival Function	Std. Error	[95% Conf. Int.]	Cum. Failure	Std. Error	[95% Conf. Int.]	Hazard Function	Std. Error	[95% Conf. Int.]
GRANDI IMPRESE												
0 5	151	0	4	10.000	0.0000	. . .	0.0000	0.0000	. . .	0.0000	. . .	. . .
5 10	147	1	9	0.9930	0.0070	0.9512 0.9990	0.0070	0.0070	0.0010 0.0488	0.0014	0.0014	0.0000 0.0042
10 15	137	0	16	0.9930	0.0070	0.9512 0.9990	0.0070	0.0070	0.0010 0.0488	0.0000	. . .	. . .
15 20	121	0	14	0.9930	0.0070	0.9512 0.9990	0.0070	0.0070	0.0010 0.0488	0.0000	. . .	. . .
20 25	107	0	16	0.9930	0.0070	0.9512 0.9990	0.0070	0.0070	0.0010 0.0488	0.0000	. . .	. . .
25 30	91	1	17	0.9809	0.0138	0.9224 0.9954	0.0191	0.0138	0.0046 0.0776	0.0024	0.0024	0.0000 0.0072
30 35	73	0	17	0.9809	0.0138	0.9224 0.9954	0.0191	0.0138	0.0046 0.0776	0.0000	. . .	. . .
35 40	56	0	19	0.9809	0.0138	0.9224 0.9954	0.0191	0.0138	0.0046 0.0776	0.0000	. . .	. . .
40 45	37	0	12	0.9809	0.0138	0.9224 0.9954	0.0191	0.0138	0.0046 0.0776	0.0000	. . .	. . .
45 50	25	0	4	0.9809	0.0138	0.9224 0.9954	0.0191	0.0138	0.0046 0.0776	0.0000	. . .	. . .
50 55	21	1	7	0.9249	0.0560	0.7000 0.9831	0.0751	0.0560	0.0169 0.3000	0.0118	0.0118	0.0000 0.0348
55 60	13	0	6	0.9249	0.0560	0.7000 0.9831	0.0751	0.0560	0.0169 0.3000	0.0000	. . .	. . .
60 65	7	0	1	0.9249	0.0560	0.7000 0.9831	0.0751	0.0560	0.0169 0.3000	0.0000	. . .	. . .
65 70	6	0	2	0.9249	0.0560	0.7000 0.9831	0.0751	0.0560	0.0169 0.3000	0.0000	. . .	. . .
70 75	4	0	1	0.9249	0.0560	0.7000 0.9831	0.0751	0.0560	0.0169 0.3000	0.0000	. . .	. . .
75 80	3	0	1	0.9249	0.0560	0.7000 0.9831	0.0751	0.0560	0.0169 0.3000	0.0000	. . .	. . .
80 85	2	0	1	0.9249	0.0560	0.7000 0.9831	0.0751	0.0560	0.0169 0.3000	0.0000	. . .	. . .
90 95	1	1	0	0.0000	. . .	. . .	10.000	. . .	. . .	0.4000	. . .	. . .
MICRO IMPRESE												
0 5	6577	564	1406	0.9040	0.0038	0.8962 0.9112	0.0960	0.0038	0.0888 0.1038	0.0202	0.0008	0.0185 0.0218
5 10	4607	464	1023	0.8016	0.0056	0.7903 0.8123	0.1984	0.0056	0.1877 0.2097	0.0240	0.0011	0.0218 0.0262
10 15	3120	228	668	0.7360	0.0066	0.7227 0.7487	0.2640	0.0066	0.2513 0.2773	0.0171	0.0011	0.0149 0.0193
15 20	2224	152	497	0.6793	0.0075	0.6643 0.6939	0.3207	0.0075	0.3061 0.3357	0.0160	0.0013	0.0135 0.0185
20 25	1575	109	329	0.6268	0.0085	0.6100 0.6432	0.3732	0.0085	0.3568 0.3900	0.0161	0.0015	0.0131 0.0191
25 30	1137	87	302	0.5715	0.0096	0.5525 0.5901	0.4285	0.0096	0.4099 0.4475	0.0185	0.0020	0.0146 0.0223
30 35	748	62	228	0.5156	0.0110	0.4939 0.5369	0.4844	0.0110	0.4631 0.5061	0.0206	0.0026	0.0155 0.0257
35 40	458	37	156	0.4654	0.0126	0.4405 0.4900	0.5346	0.0126	0.5100 0.5595	0.0205	0.0034	0.0139 0.0271
40 45	265	26	89	0.4106	0.0150	0.3810 0.4399	0.5894	0.0150	0.5601 0.6190	0.0251	0.0049	0.0154 0.0347
45 50	150	13	44	0.3689	0.0174	0.3348 0.4029	0.6311	0.0174	0.5971 0.6652	0.0214	0.0059	0.0098 0.0330
50 55	93	4	29	0.3501	0.0189	0.3132 0.3871	0.6499	0.0189	0.6129 0.6868	0.0105	0.0052	0.0002 0.0207
55 60	60	5	17	0.3161	0.0223	0.2728 0.3602	0.6839	0.0223	0.6398 0.7272	0.0204	0.0091	0.0025 0.0383
60 65	38	1	12	0.3062	0.0237	0.2604 0.3531	0.6938	0.0237	0.6469 0.7396	0.0063	0.0063	0.0000 0.0188
65 70	25	1	8	0.2916	0.0267	0.2404 0.3446	0.7084	0.0267	0.6554 0.7596	0.0098	0.0098	0.0000 0.0289
70 75	16	1	4	0.2708	0.0319	0.2103 0.3346	0.7292	0.0319	0.6654 0.7897	0.0148	0.0148	0.0000 0.0438
75 80	11	1	3	0.2423	0.0393	0.1697 0.3220	0.7577	0.0393	0.6780 0.8303	0.0222	0.0222	0.0000 0.0657
80 85	7	1	1	0.2050	0.0477	0.1208 0.3048	0.7950	0.0477	0.6952 0.8792	0.0333	0.0333	0.0000 0.0984
85 90	5	1	0	0.1640	0.0530	0.0769 0.2797	0.8360	0.0530	0.7203 0.9231	0.0444	0.0444	0.0000 0.1310
90 95	4	1	1	0.1171	0.0548	0.0374 0.2469	0.8829	0.0548	0.7531 0.9626	0.0667	0.0667	0.0000 0.1955
95 100	2	1	0	0.0586	0.0497	0.0061 0.2060	0.9414	0.0497	0.7940 0.9939	0.1333	0.1257	0.0000 0.3797
110 115	1	1	0	0.0000	. . .	. . .	10.000	. . .	. . .	0.4000	. . .	. . .

ALLEGATO H – CONTINUAZIONE

MEDIE IMPRESE																
0	5	521	3	30	0.9941	0.0034	0.9817	0.9981	0.0059	0.0034	0.0019	0.0183	0.0012	0.0007	0.0000	0.0025
5	10	488	3	41	0.9877	0.0050	0.9728	0.9945	0.0123	0.0050	0.0055	0.0272	0.0013	0.0007	0.0000	0.0027
10	15	444	2	41	0.9830	0.0060	0.9663	0.9915	0.0170	0.0060	0.0085	0.0337	0.0009	0.0007	0.0000	0.0023
15	20	401	0	56	0.9830	0.0060	0.9663	0.9915	0.0170	0.0060	0.0085	0.0337	0.0000	.	.	.
20	25	345	3	60	0.9737	0.0080	0.9524	0.9855	0.0263	0.0080	0.0145	0.0476	0.0019	0.0011	0.0000	0.0041
25	30	282	1	54	0.9698	0.0088	0.9467	0.9830	0.0302	0.0088	0.0170	0.0533	0.0008	0.0008	0.0000	0.0023
30	35	227	1	68	0.9648	0.0101	0.9384	0.9800	0.0352	0.0101	0.0200	0.0616	0.0010	0.0010	0.0000	0.0031
35	40	158	2	55	0.9500	0.0144	0.9126	0.9717	0.0500	0.0144	0.0283	0.0874	0.0031	0.0022	0.0000	0.0074
40	45	101	1	41	0.9382	0.0184	0.8900	0.9657	0.0618	0.0184	0.0343	0.1100	0.0025	0.0025	0.0000	0.0074
45	50	59	0	27	0.9382	0.0184	0.8900	0.9657	0.0618	0.0184	0.0343	0.1100	0.0000	.	.	.
50	55	32	2	10	0.8687	0.0503	0.7298	0.9391	0.1313	0.0503	0.0609	0.2702	0.0154	0.0109	0.0000	0.0367
55	60	20	0	10	0.8687	0.0503	0.7298	0.9391	0.1313	0.0503	0.0609	0.2702	0.0000	.	.	.
60	65	10	1	2	0.7722	0.1014	0.4968	0.9089	0.2278	0.1014	0.0911	0.5032	0.0235	0.0235	0.0000	0.0696
65	70	7	1	2	0.6435	0.1447	0.3018	0.8503	0.3565	0.1447	0.1497	0.6982	0.0364	0.0362	0.0000	0.1073
70	75	4	0	3	0.6435	0.1447	0.3018	0.8503	0.3565	0.1447	0.1497	0.6982	0.0000	.	.	.
80	85	1	0	1	0.6435	0.1447	0.3018	0.8503	0.3565	0.1447	0.1497	0.6982	0.0000	.	.	.
PICCOLE IMPRESE																
0	5	1408	24	110	0.9823	0.0036	0.9737	0.9881	0.0177	0.0036	0.0119	0.0263	0.0036	0.0007	0.0021	0.0050
5	10	1274	13	172	0.9715	0.0046	0.9609	0.9793	0.0285	0.0046	0.0207	0.0391	0.0022	0.0006	0.0010	0.0034
10	15	1089	9	157	0.9629	0.0054	0.9506	0.9721	0.0371	0.0054	0.0279	0.0494	0.0018	0.0006	0.0006	0.0030
15	20	923	9	167	0.9525	0.0064	0.9384	0.9635	0.0475	0.0064	0.0365	0.0616	0.0022	0.0007	0.0007	0.0036
20	25	747	7	131	0.9428	0.0073	0.9266	0.9554	0.0572	0.0073	0.0446	0.0734	0.0021	0.0008	0.0005	0.0036
25	30	609	2	168	0.9392	0.0077	0.9222	0.9526	0.0608	0.0077	0.0474	0.0778	0.0008	0.0005	0.0000	0.0018
30	35	439	5	151	0.9262	0.0095	0.9052	0.9428	0.0738	0.0095	0.0572	0.0948	0.0028	0.0012	0.0003	0.0052
35	40	283	4	109	0.9100	0.0123	0.8826	0.9313	0.0900	0.0123	0.0687	0.1174	0.0035	0.0018	0.0001	0.0070
40	45	170	3	75	0.8894	0.0168	0.8514	0.9182	0.1106	0.0168	0.0818	0.1486	0.0046	0.0026	0.0000	0.0098
45	50	92	1	38	0.8772	0.0205	0.8303	0.9119	0.1228	0.0205	0.0881	0.1697	0.0028	0.0028	0.0000	0.0082
50	55	53	1	22	0.8564	0.0288	0.7889	0.9036	0.1436	0.0288	0.0964	0.2111	0.0048	0.0048	0.0000	0.0143
55	60	30	0	13	0.8564	0.0288	0.7889	0.9036	0.1436	0.0288	0.0964	0.2111	0.0000	.	.	.
60	65	17	0	3	0.8564	0.0288	0.7889	0.9036	0.1436	0.0288	0.0964	0.2111	0.0000	.	.	.
65	70	14	1	6	0.7785	0.0787	0.5755	0.8927	0.2215	0.0787	0.1073	0.4245	0.0190	0.0190	0.0000	0.0563
70	75	7	0	4	0.7785	0.0787	0.5755	0.8927	0.2215	0.0787	0.1073	0.4245	0.0000	.	.	.
75	80	3	0	1	0.7785	0.0787	0.5755	0.8927	0.2215	0.0787	0.1073	0.4245	0.0000	.	.	.
85	90	2	0	1	0.7785	0.0787	0.5755	0.8927	0.2215	0.0787	0.1073	0.4245	0.0000	.	.	.
95	100	1	0	1	0.7785	0.0787	0.5755	0.8927	0.2215	0.0787	0.1073	0.4245	0.0000	.	.	.

Interval	Beg. Total	Deaths	Lost	Survival Function	Std. Error	[95% Conf. Int.]	Cum. Failure	Std. Error	[95% Conf. Int.]	Hazard Function	Std. Error	[95% Conf. Int.]
CENTRO												
0 5	1384	86	310	0.9300	0.0073	0.9143 0.9430	0.0700	0.0073	0.0570 0.0857	0.0145	0.0016	0.0114 0.0176
5 10	988	77	224	0.8483	0.0111	0.8251 0.8687	0.1517	0.0111	0.1313 0.1749	0.0184	0.0021	0.0143 0.0225
10 15	687	36	143	0.7987	0.0132	0.7714 0.8231	0.2013	0.0132	0.1769 0.2286	0.0121	0.0020	0.0081 0.0160
15 20	508	23	125	0.7574	0.0150	0.7264 0.7854	0.2426	0.0150	0.2146 0.2736	0.0106	0.0022	0.0063 0.0149
20 25	360	15	82	0.7218	0.0169	0.6871 0.7534	0.2782	0.0169	0.2466 0.3129	0.0096	0.0025	0.0048 0.0145
25 30	263	12	70	0.6838	0.0193	0.6444 0.7199	0.3162	0.0193	0.2801 0.3556	0.0108	0.0031	0.0047 0.0169
30 35	181	12	51	0.6311	0.0230	0.5841 0.6743	0.3689	0.0230	0.3257 0.4159	0.0161	0.0046	0.0070 0.0251
35 40	118	11	39	0.5606	0.0286	0.5026 0.6145	0.4394	0.0286	0.3855 0.4974	0.0237	0.0071	0.0097 0.0376
40 45	68	5	21	0.5118	0.0334	0.4445 0.5751	0.4882	0.0334	0.4249 0.5555	0.0182	0.0081	0.0023 0.0341
45 50	42	3	19	0.4646	0.0399	0.3848 0.5405	0.5354	0.0399	0.4595 0.6152	0.0194	0.0112	0.0000 0.0412
50 55	20	2	7	0.4083	0.0512	0.3076 0.5063	0.5917	0.0512	0.4937 0.6924	0.0258	0.0182	0.0000 0.0615
55 60	11	0	2	0.4083	0.0512	0.3076 0.5063	0.5917	0.0512	0.4937 0.6924	0.0000	.	.
60 65	9	0	3	0.4083	0.0512	0.3076 0.5063	0.5917	0.0512	0.4937 0.6924	0.0000	.	.
65 70	6	0	2	0.4083	0.0512	0.3076 0.5063	0.5917	0.0512	0.4937 0.6924	0.0000	.	.
70 75	4	1	2	0.2722	0.1163	0.0841 0.5046	0.7278	0.1163	0.4954 0.9159	0.0800	0.0784	0.0000 0.2336
75 80	1	0	1	0.2722	0.1163	0.0841 0.5046	0.7278	0.1163	0.4954 0.9159	0.0000	.	.
NORD EST												
0 5	2574	173	465	0.9261	0.0054	0.9148 0.9360	0.0739	0.0054	0.0640 0.0852	0.0153	0.0012	0.0131 0.0176
5 10	1936	159	363	0.8422	0.0080	0.8257 0.8572	0.1578	0.0080	0.1428 0.1743	0.0190	0.0015	0.0160 0.0219
10 15	1414	69	242	0.7972	0.0092	0.7784 0.8147	0.2028	0.0092	0.1853 0.2216	0.0110	0.0013	0.0084 0.0136
15 20	1103	43	237	0.7624	0.0103	0.7416 0.7818	0.2376	0.0103	0.2182 0.2584	0.0089	0.0014	0.0063 0.0116
20 25	823	40	175	0.7210	0.0116	0.6975 0.7430	0.2790	0.0116	0.2570 0.3025	0.0112	0.0018	0.0077 0.0146
25 30	608	27	170	0.6837	0.0130	0.6574 0.7085	0.3163	0.0130	0.2915 0.3426	0.0106	0.0020	0.0066 0.0146
30 35	411	19	146	0.6453	0.0150	0.6151 0.6738	0.3547	0.0150	0.3262 0.3849	0.0116	0.0027	0.0064 0.0168
35 40	246	10	89	0.6133	0.0173	0.5783 0.6462	0.3867	0.0173	0.3538 0.4217	0.0102	0.0032	0.0039 0.0165
40 45	147	8	54	0.5724	0.0214	0.5293 0.6130	0.4276	0.0214	0.3870 0.4707	0.0138	0.0049	0.0042 0.0233
45 50	85	1	37	0.5638	0.0227	0.5180 0.6069	0.4362	0.0227	0.3931 0.4820	0.0030	0.0030	0.0000 0.0090
50 55	47	3	23	0.5161	0.0335	0.4485 0.5795	0.4839	0.0335	0.4205 0.5515	0.0176	0.0102	0.0000 0.0376
55 60	21	0	11	0.5161	0.0335	0.4485 0.5795	0.4839	0.0335	0.4205 0.5515	0.0000	.	.
60 65	10	2	1	0.4075	0.0732	0.2647 0.5453	0.5925	0.0732	0.4547 0.7353	0.0471	0.0330	0.0000 0.1118
65 70	7	0	4	0.4075	0.0732	0.2647 0.5453	0.5925	0.0732	0.4547 0.7353	0.0000	.	.
70 75	3	0	1	0.4075	0.0732	0.2647 0.5453	0.5925	0.0732	0.4547 0.7353	0.0000	.	.
80 85	2	0	1	0.4075	0.0732	0.2647 0.5453	0.5925	0.0732	0.4547 0.7353	0.0000	.	.
110 115	1	1	0	0.0000	.	.	10.000	.	.	0.4000	.	.

ALLEGATO I – CONTINUAZIONE

NORD OVEST															
0	5	3385	162	512	0.9482	0.0040	0.9399	0.9554	0.0518	0.0040	0.0446	0.0601	0.0106	0.0008	0.0090
5	10	2711	143	445	0.8937	0.0058	0.8818	0.9045	0.1063	0.0058	0.0955	0.1182	0.0118	0.0010	0.0099
10	15	2123	72	363	0.8606	0.0068	0.8467	0.8733	0.1394	0.0068	0.1267	0.1533	0.0076	0.0009	0.0058
15	20	1688	61	279	0.8267	0.0078	0.8109	0.8413	0.1733	0.0078	0.1587	0.1891	0.0080	0.0010	0.0060
20	25	1348	37	235	0.8018	0.0085	0.7845	0.8180	0.1982	0.0085	0.1820	0.2155	0.0061	0.0010	0.0041
25	30	1076	38	252	0.7698	0.0097	0.7502	0.7880	0.2302	0.0097	0.2120	0.2498	0.0082	0.0013	0.0056
30	35	786	29	221	0.7367	0.0110	0.7144	0.7576	0.2633	0.0110	0.2424	0.2856	0.0088	0.0016	0.0056
35	40	536	13	193	0.7149	0.0122	0.6901	0.7381	0.2851	0.0122	0.2619	0.3099	0.0060	0.0017	0.0027
40	45	330	13	129	0.6799	0.0150	0.6495	0.7083	0.3201	0.0150	0.2917	0.3505	0.0100	0.0028	0.0046
45	50	188	9	55	0.6418	0.0188	0.6037	0.6773	0.3582	0.0188	0.3227	0.3963	0.0115	0.0038	0.0040
50	55	124	3	35	0.6237	0.0210	0.5811	0.6633	0.3763	0.0210	0.3367	0.4189	0.0057	0.0033	0.0000
55	60	86	4	30	0.5886	0.0261	0.5355	0.6377	0.4114	0.0261	0.3623	0.4645	0.0116	0.0058	0.0002
60	65	52	0	14	0.5886	0.0261	0.5355	0.6377	0.4114	0.0261	0.3623	0.4645	0.0000	.	.
65	70	38	3	11	0.5342	0.0381	0.4567	0.6056	0.4658	0.0381	0.3944	0.5433	0.0194	0.0112	0.0000
70	75	24	0	9	0.5342	0.0381	0.4567	0.6056	0.4658	0.0381	0.3944	0.5433	0.0000	.	.
75	80	15	1	4	0.4931	0.0529	0.3860	0.5915	0.5069	0.0529	0.4085	0.6140	0.0160	0.0160	0.0000
80	85	10	1	2	0.4384	0.0699	0.2999	0.5685	0.5616	0.0699	0.4315	0.7001	0.0235	0.0235	0.0000
85	90	7	1	1	0.3709	0.0857	0.2090	0.5335	0.6291	0.0857	0.4665	0.7910	0.0333	0.0332	0.0000
90	95	5	2	1	0.2061	0.0991	0.0568	0.4190	0.7939	0.0991	0.5810	0.9432	0.1143	0.0774	0.0000
95	100	2	1	1	0.0687	0.0859	0.0012	0.3422	0.9313	0.0859	0.6578	0.9988	0.2000	0.1732	0.0000
SUD&ISOLE															
0	5	1314	170	263	0.8562	0.0102	0.8349	0.8750	0.1438	0.0102	0.1250	0.1651	0.0310	0.0024	0.0263
5	10	881	102	213	0.7435	0.0137	0.7155	0.7691	0.2565	0.0137	0.2309	0.2845	0.0282	0.0028	0.0227
10	15	566	62	134	0.6511	0.0162	0.6182	0.6819	0.3489	0.0162	0.3181	0.3818	0.0265	0.0034	0.0199
15	20	370	34	93	0.5827	0.0183	0.5459	0.6175	0.4173	0.0183	0.3825	0.4541	0.0222	0.0038	0.0147
20	25	243	27	44	0.5115	0.0206	0.4705	0.5509	0.4885	0.0206	0.4491	0.5295	0.0260	0.0050	0.0162
25	30	172	14	49	0.4629	0.0223	0.4186	0.5060	0.5371	0.0223	0.4940	0.5814	0.0199	0.0053	0.0095
30	35	109	8	46	0.4199	0.0249	0.3708	0.4681	0.5801	0.0249	0.5319	0.6292	0.0195	0.0069	0.0060
35	40	55	9	18	0.3377	0.0317	0.2764	0.4000	0.6623	0.0317	0.6000	0.7236	0.0434	0.0144	0.0152
40	45	28	4	13	0.2749	0.0383	0.2028	0.3517	0.7251	0.0383	0.6483	0.7972	0.0410	0.0204	0.0010
45	50	11	1	2	0.2474	0.0432	0.1678	0.3352	0.7526	0.0432	0.6648	0.8322	0.0211	0.0210	0.0000
50	55	8	0	3	0.2474	0.0432	0.1678	0.3352	0.7526	0.0432	0.6648	0.8322	0.0000	.	.
55	60	5	1	3	0.1767	0.0673	0.0696	0.3240	0.8233	0.0673	0.6760	0.9304	0.0667	0.0657	0.0000
65	70	1	0	1	0.1767	0.0673	0.0696	0.3240	0.8233	0.0673	0.6760	0.9304	0.0000	.	.

Bibliografia

Elisa T. Lee, John Wenju Wang, **Statistical Methods for Survival Data Analysis**, United States of America, John Wiley & Sons, (2003), Third Edition, 523.

David G. Kleinbaum, Mitchel Klein, **Survival Analysis, A self-Learning Test**, USA, Springer, (2012), Third Edition, 700.

Stata Survival Analysis Reference Manual, Texas, Stata Press, (2019), Release 16, 488

Database AIDA

<https://login.ezproxy.biblio.polito.it/login?qurl=https://aida.bvdinfo.com%2fversion-2020331%2fhome.serv%3fproduct%3dAidaNeo>

Istituto Nazionale di Statistica

<https://www.istat.it/>

Confindustria

<https://www.confindustria.it/>