Dipartimento di Statistica
"Giuseppe Parenti"

Università degli Studi di Firenze

DIDATTICA 2001/01

# Metodi diagnostici basati sui residui nei modelli per dati di durata

Michela Baccini, Fabrizia Mealli

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze -

Firenze University Press

*Università degli Studi di Firenze*

PUBBLICAZIONI DIGITALI DEL DIPARTIMENTO DI STATISTICA «G. PARENTI»

SERIE DIDATTICA

# METODI DIAGNOSTICI BASATI SUI RESIDUI NEI MODELLI PER DATI DI DURATA

Michela Baccini, Fabrizia Mealli

**Firenze University Press**

**2001**

Metodi diagnostici basati sui residui nei
modelli per dati di durata / Michela Baccini,
Fabrizia Mealli. – Dati e programmi per
computer. – Firenze : Firenze University
Press, 2001.
(Pubblicazioni digitali del Dipartimento di Statistica "G. Parenti". Serie Didattica / Università degli
Studi di Firenze, Dipartimento di Statistica "G. Parenti")
Modalità di accesso: http://fulltext.casalini.it/is.asp?isbn=8884530253.htm
Tit. della schermata del titolo.

ISBN 88-8453-025-3

332 (ed. 20.)



Firenze University Press
Borgo Albizi, 28
50122 Firenze
Italy
http://www.unifi.it/e-press
email: e-press@unifi.it

# Indice

# Prefazione

L'analisi dei dati di durata ha ricevuto ampia attenzione nella letteratura statistica e i modelli per dati di durata sono ormai largamente impiegati nelle applicazioni di natura biometrica, economica, sociologica.

Così come per tutti gli altri modelli statistici, la verifica della loro adeguatezza rappresenta una fase essenziale del processo inferenziale. In presenza di errori di specificazione i risultati dello studio possono risultare compromessi in quanto aspetti essenziali del processo generatore dei dati, particolari strutture di dipendenza tra le osservazioni o relazioni tra variabili, possono non essere colte. Ciò si può ripercuotere sulle proprietà degli stimatori, inficiandone ad esempio la consistenza e l'efficienza.

Mentre nell'ambito della regressione lineare classica, e in parte per i modelli lineari generalizzati, esiste una vasta letteratura che tratta in modo sistematico dei metodi diagnostici (si veda ad esempio Cook e Weisberg, 1982; McCullagh e Nelder, 1989), per quanto riguarda i modelli per dati di durata, sono stati pochi i tentativi in letteratura di dare una forma organica all'argomento (Fleming e Harrington, 1991; Andersen *et al.*, 1993). Il presente lavoro, che trae spunto dalla tesi di laurea di Michela Baccini che ha ricevuto gli auspici di pubblicazione, vuole offrire un contributo in questa direzione. In particolare, ci occuperemo dei metodi basati sull'analisi dei residui, proposti nell'ambito del modello a rischio moltiplicativo, includendo i contributi più recenti sull'argomento.

Pur facendo spesso riferimento alla notazione classica (Kalbfleisch e Prentice, 1980; Cox e Oakes, 1984), i dati di durata saranno rappresentati come realizzazioni di processi di conteggio, secondo l'approccio che dalla seconda metà degli anni ottanta ha trovato uno sviluppo sempre crescente e che è trattato in modo organico nei lavori di Fleming e Harrington (1991) e Andersen *et al.* (1993). Questo consentirà di derivare in modo semplice i residui e i metodi diagnostici a essi associati, anche nel caso generale in cui l'evento oggetto di studio sia ricorrente.

Nel primo capitolo si mostrerà quindi come i dati di durata trovino una rappresentazione naturale nei processi di conteggio e, utilizzando risultati relativi alla teoria dei processi stocastici, in particolare alle martingale, saranno individuate in tali dati una componente sistematica su cui è possibile fare inferenza e una componente di innovazione non osservabile, la cui natura è simile a quella di un errore casuale. Sarà poi presentato un primo modello generale, il modello a intensità moltiplicativa, che usualmente costituisce la base per la definizione di modelli più particolari.

Saranno infine introdotti gli strumenti per l'inferenza non parametrica, semiparametrica e parametrica sui dati di durata. Dopo aver definito gli stimatori non parametrici di Nelson-Aalen e di Kaplan-Meier, che vengono comunemente utilizzati per stimare rispettivamente l'*hazard* integrato e la funzione di sopravvivenza relative a popolazioni omogenee, saranno presi in considerazione i modelli di regressione per dati di durata: in particolare il modello semiparametrico di Cox e altri modelli di regressione parametrici.

Nel secondo capitolo, dopo un breve cenno alla definizione di residuo nel modello di regressione lineare classico e a quella di residuo generalizzato, saranno definiti i diversi tipi di residuo proposti in letteratura per i modelli di durata (i residui martingala, *score*, di devianza, di Schoenfeld, *log-odds* e *probit*) e ne saranno analizzate le principali proprietà con particolare riferimento a quelle che, seppur debolmente, li accomunano ai residui del modello lineare classico.

Nel terzo capitolo si mostrerà come tali residui possano essere impiegati a fini diagnostici. Saranno presentate tecniche grafiche e numeriche che consentono sia di individuare eventuali osservazioni anomale o particolarmente influenti sulla stima del modello (Nardi e Schemper, 1999), sia di verificare le ipotesi su cui il modello stesso si basa. Ad esempio, si analizzerà come sia possibile utilizzare i residui per determinare la forma funzionale più adeguata attraverso la quale inserire le covariate nel modello di regressione (Therneau, Grambsch e Fleming, 1990; Lin, Wei e Ying, 1993) e saranno mostrati alcuni metodi atti a verificare la costanza dei coefficienti di regressione contro l'ipotesi che l'effetto di qualche covariata dipenda dal tempo, sottoponendo così a test l'ipotesi di *hazard* proporzionali (Therneau, Grambsch e Fleming, 1990; Grambsch e Therneau, 1994). Infine sarà presentato un test basato sui residui martingala il cui scopo è quello di individuare la presenza di eventuali fattori di eterogeneità non osservabile che influiscano in modo significativo sulla propensione dei soggetti a subire l'evento d'interesse (Lancaster, 1985). Si cercherà, ove possibile, di estendere i risultati, usualmente ricavati nel caso semiparametrico, anche ai modelli di tipo parametrico, e di evidenziare i legami tra i metodi sviluppati in ambito biometrico, con quelli proposti in ambito econometrico.

L'obiettivo che con il presente lavoro si è cercato di perseguire è stato quello di predisporre un testo di riferimento per quanti vogliano avvicinarsi alle problematiche dell'analisi diagnostica nell'ambito dei modelli per dati di durata e/o debbano applicare i metodi qui presentati in lavori empirici. In questo senso il volume è rivolto principalmente a ricercatori e studenti di dottorato, ma anche a studenti dei Corsi di Laurea in Scienze Statistiche che abbiano un'adeguata conoscenza di processi stocastici e inferenza statistica.

Per non appesantire eccessivamente l'esposizione e nel tentativo di conferirle un carattere di utilità e semplicità, abbiamo evitato di riportare la derivazione di diversi risultati teorici che avrebbero dato al lavoro un taglio non adatto alle esigenze di coloro ai quali esso è rivolto. Il lettore troverà comunque nel testo i riferimenti bibliografici necessari per eventuali approfondimenti teorici.

Gran parte dei metodi diagnostici presentati non sono ancora presenti nei principali testi di riferimento, nonostante diversi tipi di residui e altre quantità di interesse vengano calcolate automaticamente dai pacchetti statistici usualmente utilizzati per l'analisi dei dati di durata (SAS, S-Plus, STATA). Riteniamo pertanto di avere fatto cosa utile nel trattare l'argomento in forma organica, in un lavoro che speriamo sia di facile consultazione.

Desideriamo rivolgere un ringraziamento particolare ad Annibale Biggeri per averci seguito e fornito utili suggerimenti e ai membri della commissione di lettura per i preziosi commenti.

L'eventuale presenza di errori e carenze nel lavoro è unicamente di nostra responsabilità.

# Dati di durata: rappresentazione, modelli e metodi di stima

## 1.1 Introduzione

Col nome di dati di durata si indicano i dati relativi ai tempi in cui si verificano particolari eventi nelle storie di vita di un campione di individui.[1] Tali eventi possono essere interpretati come transizioni da uno stato $x$ a uno $y$, che in generale possono manifestarsi una o più volte per ogni soggetto osservato. Se l'evento si verifica al più una volta in ciascuna storia individuale (evento non ricorrente), come nel caso in cui $y$ è uno stato assorbente, si parla di dati di sopravvivenza e l'evento in questione viene chiamato genericamente "morte" del soggetto. In questo lavoro prendiamo in considerazione la situazione più semplice in cui la transizione oggetto d'interesse è unica. In generale è però possibile studiare transizioni di tipo diverso contemporaneamente. Per esempio si potrebbe essere interessati a studiare la mortalità per cause di un collettivo di individui. In questo caso eventi di morte sopravvenuti per cause diverse costituiscono transizioni di stato differenti. Molte delle tecniche statistiche presentate possono essere generalizzate a casi più complessi.

La storia di ogni individuo può essere rappresentata in modo schematico dalle durate che intercorrono dall'inizio della storia stessa fino a ogni eventuale transizione di stato. L'istante fissato come origine delle storie di vita dipende strettamente dal tipo di analisi da effettuare e, in generale, è diverso per ogni soggetto, dato che non necessariamente tutti gli individui entrano contemporaneamente nello studio. Ad esempio, in ambito medico spesso l'osservazione ha inizio al momento della diagnosi, che ovviamente non è lo stesso per tutti gli individui del campione. In seguito ogniqualvolta ci riferiremo a un generico tempo $t$, se non diversamente specificato, intenderemo indicare la durata $t$, che può realizzarsi in istanti di calendario diversi per ogni soggetto, ossia effettueremo un allineamento delle storie di vita alla durata 0.

[1]Col termine individuo o soggetto si intende in generale una delle unità che costituiscono la popolazione statistica studiata ed è usato indifferentemente per indicare, ad esempio, persone (come quando l'evento osservato è il manifestarsi di una malattia, oppure la transizione nello stato di disoccupazione), componenti elettroniche (in analisi di tipo industriale), o altro.

I dati di durata presentano sempre una certa approssimazione, poiché la loro registrazione è effettuata in tempo discreto a causa della limitatezza dello strumento di misura; dunque si possono osservare durate uguali per soggetti che in realtà hanno sperimentato l'evento in tempi diversi. Tuttavia, se, come noi supporremo, le approssimazioni sono contenute, si può assumere che la rilevazione avvenga in tempo continuo. A questo proposito è necessario osservare che non esiste un criterio unico per stabilire l'entità dell'approssimazione. Ad esempio, se in analisi di tipo clinico in genere durate espresse in termini di giorni si possono considerare prodotte da una registrazione continua, in analisi industriali spesso è necessaria almeno un'approssimazione di ore o addirittura minuti.

Nel presente capitolo, dopo una breve rassegna dei meccanismi che spesso impediscono l'osservazione completa di alcune storie di vita, vedremo in che modo i dati individuali possano essere rappresentati come traiettorie di due processi stocastici: un processo di conteggio sugli eventi e un processo che indica in quali istanti sia teoricamente possibile osservare una transizione di stato. L'applicazione del teorema di Doob-Meyer consentirà di scomporre il processo di conteggio in una componente di errore casuale e una componente sistematica, il processo di intensità cumulato.

Verrà poi introdotto il modello a intensità moltiplicativa, comunemente usato per l'intensità del processo di conteggio e ne sarà evidenziato il legame con l'*hazard function* definita in ambito classico. Analizzeremo, infine, i metodi statistici comunemente utilizzati per stimare l'*hazard* integrato nel caso in cui i dati a disposizione siano incompleti, iniziando dalla situazione in cui la popolazione oggetto d'indagine sia omogenea e non venga specificata a priori la forma funzionale dell'*hazard*, per poi passare all'analisi e alla stima del modello di regressione semiparametrico di Cox e di altri modelli di tipo parametrico.

## 1.2 Meccanismi di censura

Usualmente i dati di durata si presentano in forma incompleta, a causa di meccanismi di censura che impediscono di seguire interamente alcune storie.

Uno dei meccanismi che si incontra più frequentemente è quello di censura a destra che si verifica quando il soggetto può essere osservato solo su un intervallo di tempo $[0,U_i]$, dove $U_i$ è una variabile aleatoria non negativa. $U_i$ può essere definita in diversi modi e a ogni definizione di $U_i$ corrisponde un

diverso tipo di censura a destra; a titolo esemplificativo presentiamo alcuni dei casi più comuni:[2]

***Esempio 1.2.1. Censura a destra di I tipo semplice.*** Si verifica quando i soggetti entrano contemporaneamente nello studio e l'osservazione ha una durata deterministica uguale per tutti, ossia $U_i = u$ per ogni $i$. Questo meccanismo si incontra spesso in test di durata in ambito industriale. □

***Esempio 1.2.2. Censura a destra di I tipo progressiva.*** Questo tipo di censura si incontra ogni volta che l'istante in cui il singolo soggetto entra nello studio è casuale, mentre ne è fissata a priori la data di uscita; tale situazione è frequente in ambito clinico, quando per ciascun soggetto le durate sono rilevate a partire dal tempo di diagnosi e la durata dello studio è fissata. In pratica già al tempo 0 (nell'ottica di riallineamento dell'asse dei tempi) si conosce il tempo di censura deterministico $U_i = u_i$, dove $u_i$ corrisponde all'ampiezza del periodo di osservazione sull'individuo. □

***Esempio 1.2.3. Censura di II tipo semplice e progressiva in analisi di sopravvivenza.*** Entrambe sono comuni nei test di durata in ambito industriale, quando tutte le unità entrano nello studio contemporaneamente. Si parla di censura di II tipo semplice se, per tutti gli $i$, $U_i$ è il $k$-esimo più piccolo tempo di morte (questo equivale a fissare a priori il numero totale di osservazioni complete, $k$). Si parla invece di censura a destra di II tipo progressiva, quando a ogni tempo di morte, in ordine crescente, viene censurata una determinata frazione di individui a rischio. □

In questi esempi la variabile aleatoria $U_i$ è sempre espressa in termini dei precedenti tempi di manifestazione dell'evento studiato oppure è una costante. In generale, tuttavia, una storia potrebbe risultare censurata a destra a causa di una transizione diversa da quella d'interesse che fa uscire prematuramente il soggetto dallo studio. Ad esempio, in uno studio sulla fertilità alcune osservazioni possono essere censurate a causa dell'emigrazione o la morte del soggetto. Quando l'evento d'interesse e l'evento che causa la censura sono indipendenti, si parla di censura per rischi concorrenti.

La censura di I tipo (semplice e progressiva) e la censura per rischi concorrenti rappresentano casi particolari del cosiddetto modello generale di censura casuale che prevede l'indipendenza tra i tempi di censura $\mathbf{U}=(U_1, U_2, ...$

[2] Per un approfondimento si veda Kalbfleisch e Prentice (1980) e Andersen *et al.* (1993).

$,U_n$ ) e i tempi di manifestazione dell'evento oggetto di studio. In analisi di sopravvivenza la censura è casuale se il vettore **U** è indipendente dal vettore dei tempi di morte $\mathbf{T}=(T_1, T_2, \ldots, T_n)$.

Simmetrica alla censura a destra, è la cosiddetta censura a sinistra. Si parla di censura a sinistra quando il soggetto viene osservato a partire da una durata aleatoria $V_i$. Indicando con $\tau$ la durata massima dello studio, l'individuo $i$ è seguito solo nell'intervallo $(V_i, \tau]$.

Qualora un soggetto sia incluso nel campione solo a condizione che nella sua storia antecedente la durata $V_i$ si sia verificato un determinato evento, anziché di censura si parla propriamente di troncamento a sinistra. Ad esempio, supponiamo che, nell'ambito di uno studio sulla durata della disoccupazione, il campione sia estratto dalla popolazione dei disoccupati nell'istante di calendario $t$. Un soggetto può entrare nel campione solo a condizione che il suo tempo d'attesa per un lavoro non sia minore di $V$, durata dall'inizio della disoccupazione all'istante $t$. Il meccanismo di troncamento a sinistra è comune in diversi campi. Le stesse tavole di sopravvivenza classiche, utilizzate in demografia ed epidemiologia, prevedono che la storia dell'individuo sia seguita a partire da un'età di entrata.

La censura a destra e la censura a sinistra possono essere considerati casi particolari del meccanismo originato dalla loro combinazione, che prende il nome di censura a intervalli: l'$i$-esimo soggetto è osservabile solo su un intervallo o, in generale, su un'unione di intervalli di tipo $(V_i, U_i]$.

Per quanto concerne l'analisi di un evento ricorrente, è necessario fare un'ulteriore precisazione: la mancanza d'informazione causata dalla censura in $(U_i, V_i]$ può essere di due tipi: si può conoscere solo il numero di eventi verificatisi nell'intervallo, ma non gli istanti in cui si sono manifestati (nella pratica a volte si riesce a risalire al numero di transizioni effettuate nell'intervallo censurato attraverso interviste o questionari), oppure la lacuna può essere totale. Nel secondo caso, anziché di censura, è preferibile parlare di filtro.[3]

Se non diversamente specificato, i metodi di rappresentazione e analisi dei dati che illustreremo in questo lavoro sono da ritenersi validi qualunque sia il meccanismo di censura che agisce sui dati (censura in senso proprio, troncamento o filtro), a patto che soddisfi alcuni requisiti fondamentali. Uno di questi, che sarà definito in termini formali al paragrafo 1.7.1 e che è soddisfatto dalla censura casuale, è l'indipendenza. Dato che i meccanismi che producono dati incompleti costituiscono un elemento di disturbo nei confronti del manifestarsi dell'evento di interesse, è necessario ipotizzare, affinché le conclusioni che si traggono dall'analisi siano facilmente interpretabili, che in presenza di censura i risultati dell'inferenza siano ragionevolmente vicini a

[3] Andersen *et al.* (1993), cap. III.

quelli che si otterrebbero in assenza di censura. Questo si verifica se, in ogni istante, l'insieme degli individui osservabili in presenza di censura è "rappresentativo" dell'insieme corrispondente in assenza di censura. Un meccanismo che possiede questa proprietà si definisce meccanismo di censura indipendente. Supponiamo, ad esempio, che nello studio di sopravvivenza a una malattia la censura colpisca gli individui meno gravi, ossia quelli che presumibilmente sarebbero vissuti più a lungo. In questa situazione, poiché la censura non è indipendente, i soggetti non censurati non sono "rappresentativi" dell'intera popolazione e l'intensità della mortalità stimata dai dati incompleti è distorta rispetto a quella vera.

## 1.3 I dati di durata come processi di conteggio sugli eventi

Prima di dare la definizione di processo di conteggio è necessario introdurre il concetto di filtrazione e quello di processo adattato a una filtrazione.

***Definizione 1.3.1. (Filtrazione)*** Dato uno spazio di probabilità $(\Omega, F, P)$, si dice filtrazione una famiglia di sotto$\sigma$-algebre $\{F_t : t \geq 0\}$ della $\sigma$-algebra $F$ tali che : $F_t \subset F_s$ per ogni $s \geq t \geq 0$.

La filtrazione è lo strumento matematico adatto per descrivere l'accumularsi di informazioni col passare del tempo. Ad esempio, dato un qualsiasi processo stocastico $X = \{X(t): t \geq 0\}$, una filtrazione naturale è la cosiddetta storia del processo, ossia la famiglia crescente di $\sigma$-algebre il cui elemento al tempo $t$ è la $\sigma$-algebra generata dalle variabili del processo in $[0, t]$, $F_t = \sigma(X(s): 0 \leq s \leq t)$,. Nell'ambito dei dati di durata, un'opportuna filtrazione è quella il cui elemento in $t$ rappresenta le notizie raccolte attraverso l'osservazione dei soggetti in $[0, t]$, comprese quelle riguardanti la censura. Saremo in grado di definirla in termini formali una volta presentati i due processi fondamentali per la rappresentazione dei dati di durata.

Un processo stocastico si dice adattato a una filtrazione se, in ogni istante $t$, il fenomeno che descrive dipende al più dalle informazioni passate e presenti, rappresentate dalla $\sigma$-algebra $F_t$. Formalmente:

***Definizione 1.3.2. (Processo $F_t$-adattato)*** Un processo $X = \{X(t): t \geq 0\}$ si dice adattato alla filtrazione $\{F_t : t \geq 0\}$ se $X(t)$ è $F_t$-misurabile per ogni $t \geq 0$.

Ovviamente ogni processo è adattato alla sua storia e a qualsiasi filtrazione che la contenga.

***Definizione 1.3.3. (Processo di conteggio)*** Data una filtrazione $\{F_t : t \geq 0\}$, si dice processo di conteggio un processo stocastico $N = \{N(t): t \geq 0\}$ tale che:
1) $N$ è adattato alla filtrazione;
2) $N(0) = 0$ q.c.;
3) le traiettorie sono, con probabilità 1, continue a destra, a scalini con salti di ampiezza 1.

E' naturale che un processo con queste caratteristiche venga utilizzato per modellare situazioni in cui sia necessario contare le manifestazioni di un evento di interesse, memorizzando al tempo stesso gli istanti in cui esse si verificano.

Nell'ambito dei dati di durata, supponendo di essere in assenza di censura, la storia del soggetto $i$ può essere registrata come traiettoria di un processo di conteggio con punti di salto in corrispondenza degli istanti in cui si manifestano le transizione d'interesse, così che, per ogni $t > s \geq 0$, $N_i(t) - N_i(s)$ è il numero di transizioni che si osservano in $(s, t]$; in particolare $N_i(t)$ conta gli eventi osservati in $(0, t]$.

Se invece non è possibile registrare in modo esaustivo tutte le manifestazioni dell'evento, quello che si osserva è in realtà un processo di conteggio censurato, diverso da quello che si osserverebbe in assenza di censura. Quanto detto può essere chiarito con un semplice esempio.

***Esempio 1.3.1.*** In assenza di censura, sia $T_i$ la variabile aleatoria che indica il tempo di sopravvivenza a un evento non ricorrente per l'individuo $i$. Definiamo

$$N_i^*(t) = I(T_i \leq t)$$

dove con $I(A)$ si indica la funzione indicatrice dell'insieme A. E' immediato verificare che $\{N_i^*(t) : t \geq 0\}$ è un processo di conteggio. In presenza di censura a destra, sia $X_i = \min(U_i, T_i)$ il tempo in cui il soggetto esce dallo studio e $\delta_i = I(T_i \leq U_i)$ la variabile aleatoria che indica se il dato è completo ($\delta_i = 1$) o meno ($\delta_i = 0$). $X_i$ e $\delta_i$ sono le variabili osservabili per l'individuo $i$. Il processo di conteggio che si osserva non è $N_i^*$, ma $N_i$ tale che

$$N_i(t) = \delta_i I(X_i \leq t).$$

Si osservi che il processo osservabile $N_i$ non subisce nessun salto oltre $U_i$. □

## 1.4 Il processo indicatore dell'insieme di rischio

Un concetto importante in analisi di durata è quello di insieme di rischio al tempo $t$, definito come l'insieme $R_t$ costituito dai soggetti del campione per i quali potrebbe verificarsi una transizione osservabile in $t$. L'aggettivo "osservabile" è importante, perché esclude dall'insieme $R_t$ i soggetti che nell'istante $t$ risultano censurati, ma che in generale potrebbero subire l'evento d'interesse pur non essendo osservati. L'appartenenza o meno, istante per istante, dell'$i$-esimo individuo all'insieme a rischio può essere rappresentata per mezzo del processo stocastico osservabile $Y_i = \{Y_i(t) : t \geq 0\}$, dove

$$Y_i(t) = \begin{cases} 1 & i \in R_t \\ 0 & i \notin R_t \end{cases}.$$

Per capire meglio quali siano le informazioni contenute in $Y_i$, può essere utile definire il processo $C_i = \{C_i(t) : t \geq 0\}$ che assume valore 1 negli istanti in cui il soggetto è osservabile e 0 altrimenti e che quindi rappresenta lo stato del soggetto rispetto alla censura nel tempo. $Y_i$ è il prodotto di $C_i$ e del processo indicatore dell'insieme di rischio che si osserverebbe in assenza di censura, $Y_i^*$:

$$Y_i(t) = C_i(t) Y_i^*(t). \qquad (1.1)$$

Chiaramente in assenza di censura $C_i(t) = 1$ per ogni $t$ e di conseguenza $Y_i$ e $Y_i^*$ coincidono.

***Esempio 1.4.1.*** In analisi di sopravvivenza con censura a destra il soggetto è a rischio finché non muore o è censurato, dunque

$$Y_i(t) = I(X_i \geq t) = I(T_i \geq t)\, I(U_i \geq t).$$

In questo caso $C_i(t) = I(U_i \geq t)$ e $Y_i^*(t) = I(T_i \geq t)$. □

Come si deduce dall'esempio, in analisi di sopravvivenza con dati completi il processo indicatore dell'insieme di rischio non fornisce nessuna informazione aggiuntiva rispetto a $N_i$, poiché dipende solo dal tempo di morte ($C_i(t) = 1$ per ogni $t$). Se invece l'evento d'interesse è ricorrente, $Y_i^*(t)$ può assumere valore 1 anche dopo il verificarsi di una o più transizioni; in alcuni casi può anche

accadere che il soggetto entri ed esca dall'insieme a rischio più di una volta. Ad esempio, supponiamo di voler studiare gli episodi febbrili in un gruppo di individui: quando il soggetto contrae la febbre, esce dall'insieme di rischio, per poi rientrarvi dopo la guarigione, cioè quando è di nuovo nella situazione di subire l'evento. In presenza di dati incompleti, questo tipo d'informazione si fonde con quella riguardante la censura.

### 1.4.1 La prevedibilità del processo $Y_i$

Per motivi che chiariremo tra breve, è necessario ipotizzare che $Y_i$ sia un processo $F_t$–prevedibile, ossia che, data la storia dello studio in $[0, t)$, rappresentata dalla σ-algebra $F_{t-}$,[4] deve risultare automaticamente determinato il valore in $t$ di $Y_i$. La seguente definizione enuncia in modo formale questo concetto:[5]

***Definizione 1.4.1. (Processo prevedibile)*** Dato uno spazio di probabilità $(\Omega,F,P)$ e una filtrazione $\{F_t : t\geq 0\}$ su tale spazio, un processo si dice $F_t$-prevedibile se è misurabile rispetto alla più piccola σ-algebra su $\mathbf{R}^+ \times \Omega$ generata dai processi continui a sinistra adattati.

Il concetto di prevedibilità è strettamente legato a quello di filtrazione. Infatti un processo può essere prevedibile rispetto a una filtrazione e non prevedibile rispetto a un'altra. Tuttavia nella nostra trattazione spesso eviteremo di specificare la filtrazione cui si fa riferimento, sottintendendo che si tratta di quella generata dalla storia dello studio.

Se $X$ è adattato alla filtrazione, vale la seguente condizione sufficiente, ma non necessaria per la prevedibilità (Lemma 4.1.1 in Fleming e Harrington, 1991):

***Teorema 1.4.2. (Prevedibilità per un processo adattato)*** Data una filtrazione $\{F_t : t \geq 0\}$ su uno spazio di probabilità, se un processo è adattato alla filtrazione ed è continuo a sinistra,[6] allora è $F_t$-prevedibile.

---

[4]La notazione $F_{t-}$, indica il limite sinistro in $t$ della filtrazione, ossia la più piccola sottoσ-algebra di $F$ contenente $\cup_{h>0} F_{t-h}$.

[5]Fleming e Harrington (1991), proposizione 1.4.1.

[6]Si dice continuo a sinistra un processo stocastico con traiettorie continue a sinistra a meno di un insieme di probabilità nulla.

In analisi di sopravvivenza su dati completi, affinché $Y_i$ sia prevedibile, è sufficiente supporre che le sue traiettorie siano continue a sinistra, ossia considerare a rischio l'individuo anche nell'istante di morte.

Maggiore attenzione merita il caso in cui le storie di vita sono incomplete. In questa situazione, affinché il processo indicatore dell'insieme di rischio conservi la prevedibilità, è necessario che anche il meccanismo di censura, rappresentato da $C_i$, sia prevedibile: date le informazioni raccolte in [0, $t$), si deve poter dedurre se in $t$ il soggetto è osservabile o meno. Se l'origine delle durate non è la stessa per tutti gli individui, quindi $t$ non corrisponde a un unico tempo di calendario, questo requisito non è ovvio. Per esempio, supponiamo di osservare solo tre storie di vita e che l'evento d'interesse sia la morte dell'individuo. L'origine dei tempi per l'individuo 1 corrisponde all'istante di calendario 0, per l'individuo 2 all'istante 3, per l'individuo 3 all'istante 4. Supponiamo che i tempi di sopravvivenza siano rispettivamente 5, 3 e 3. Se si stabilisce di interrompere l'osservazione una volta che si siano verificate le prime due morti, allora la storia del terzo soggetto risulta censurata con tempo di censura che, in un'ottica di allineamento dei tempi alla durata 0, dipende dal futuro.

Tutti i tipi di censura a destra presentati negli esempi al paragrafo 1.2 sono adattati alla filtrazione. In generale è comunque necessario verificare di volta in volta che il meccanismo possieda questa proprietà.

Una volta illustrate le caratteristiche dei due processi osservabili $N_i$ e $Y_i$, é possibile definire in modo formale la filtrazione che descrive l'accumularsi di notizie col passare del tempo. Essa deve contenere la storia di entrambi i processi. Inoltre, dato che il teorema di scomposizione, che enunceremo nel prossimo paragrafo, richiede come ipotesi che $N_i$ sia definito su una filtrazione che soddisfa i requisiti usuali di continuità a destra e completezza, è conveniente che essa sia costruita in modo tale da risultare continua a destra.[7] Indicando con $Y_i(u+)$ il limite destro di $Y(u)$ definiamo $F_t$ come la $\sigma$-algebra generata da $N_i(u)$ e $Y_i(u+)$, con $i = 1, 2,..., n$ e $0 \le u \le t$. L'utilizzazione del processo traslato $Y_i(u+)$ garantisce la continuità a destra della filtrazione.

[7]Una filtrazione $\{F_t : t \ge 0\}$ si dice continua a destra se è tale che $F_t = F_{t+}$, dove $F_{t+}$, è la $\sigma$-algebra limite destro della filtrazione in $t$ ($F_{t+} = \cap_{h>0} F_{t+h}$). Si dice completa se $A \in F_0 \ \forall A \in F$ con $P(A) = 0$.

## 1.5 Scomposizione del processo di conteggio

In questo paragrafo viene descritta un'importante proprietà dei processi di conteggio, in base alla quale ciascun $N_i$ può essere scomposto in modo univoco nella somma di due processi stocastici: una martingala locale, interpretabile come componente casuale, e un processo crescente prevedibile, interpretabile come componente sistematica di $N_i$. Dopo una breve definizione di processo martingala, suddetta scomposizione sarà analizzata nell'ambito di studi di sopravvivenza e successivamente, una volta introdotto il concetto di martingala locale, sarà estesa al caso più generale in cui l'evento studiato sia ricorrente.

### 1.5.1 Scomposizione di Doob-Meyer in analisi di sopravvivenza

***Definizione 1.5.1. (Martingala)*** Dato uno spazio di probabilità $(\Omega, F, P)$ e una filtrazione $\{F_t : t \geq 0\}$, un processo stocastico $\{X(t): t \geq 0\}$ si dice martingala rispetto a $\{F_t\}$ se:
1) $X(t)$ è $F_t$ -misurabile per ogni $t \geq 0$;
2) $E|X(t)| < \infty$ per ogni $t \geq 0$;
3) $E(X(t)|F_s) = X(s)$ q.c. per ogni $t \geq s \geq 0$.

Se al punto 3 della precedente definizione anziché l'uguaglianza vale $E(X(t)|F_s) \geq X(s)$, il processo prende il nome di sottomartingala; se vale $E(X(t)|F_s) \leq X(s)$, quello di supermartingala.

E' semplice dimostrare che tutte le variabili aleatorie di una martingala hanno stessa media e che il valore atteso dell'incremento $X(t) - X(t-\Delta t)$ condizionato alla storia passata è identicamente nullo a meno di un insieme di probabilità 0

$$E(X(t) - X(t-\Delta t) \mid F_{t-\Delta t}) = 0 \quad \text{q.c..}$$ [8]

Il teorema di scomposizione che utilizzeremo inizialmente è il seguente:[9]

---

[8] L'uguaglianza dei valori attesi delle variabili del processo martingala deriva dalla proprietà del valore atteso condizionale per la quale $EE(X|F)=EX$. Infatti

$$EX(t)=EE(X(t+h)|F_t)=EX(t+h) \qquad \text{per ogni } t \geq 0, \text{ per ogni } h > 0.$$

Per quanto riguarda il valore atteso condizionato al passato dell'incremento del processo martingala, dato che $E(X(t)|F_{t-\Delta t})= X(t-\Delta t)$ q.c. e che $E(X(t-\Delta t)|F_{t-\Delta t})=X(t-\Delta t)$ in quanto $X$ è adattato alla filtrazione, allora

$$E(X(t)-X(t-\Delta t)|F_{t-\Delta t})=0 \text{ q.c.}$$

[9] Fleming e Harrington (1991), teorema 1.4.1.

***Teorema 1.5.1. (Scomposizione di Doob-Meyer)*** Sia $X$ una sottomartingala non negativa continua a destra rispetto a una filtrazione $\{F_t : t\geq 0\}$ che soddisfa le usuali condizioni di completezza e continuità a destra. Allora esiste un unico processo crescente continuo a destra $F_t$-prevedibile $A$ , detto compensatore di $X$ , tale che:

1) $A(0)=0$ q.c.;
2) $EA(t)<\infty$
3) $\{M(t)=X(t)-A(t): t \geq 0\}$ è una $F_t$ -martingala continua a destra.

Un processo di conteggio tale che $EN(t)< \infty$ per ogni $t \geq 0$ è una sottomartingala non negativa continua a destra e come tale soddisfa le ipotesi del teorema 1.5.1.[10] Dunque il processo $N_i$ in analisi di sopravvivenza (in tale situazione $EN_i(t) < \infty$ per ogni $t$, visto che $N_i(t)$ assume sempre valore 0 o 1) ammette un'unica scomposizione come somma di una martingala $M_i$ e di un compensatore $A_i$, noto come processo di intensità cumulato o integrato.

$M_i$ può essere considerata la componente innovativa all'interno di $N_i$; infatti, essendo una martingala continua a destra, in generale non è prevedibile.[11] Come vedremo meglio in seguito $M_i$ possiede le caratteristiche di un processo di disturbo casuale, prima tra tutte quella di avere valore atteso nullo ($N_i(0) = 0$ e $A_i(0) = 0$ q.c., quindi $EM_i(t) = 0$ per ogni $t$).

Il processo d'intensità integrato costituisce invece la componente prevedibile sistematica di $N_i$. Esso specifica il modo in cui la media dell'incremento $dN_i(t)$ dipende dalle informazioni in $[0,t)$. In modo informale si può infatti scrivere:

$$E(dN_i(t)\,|\,F_{t-})=E(dA_i(t)\,|\,F_{t-})+E(dM_i(t)\,|\,F_{t-})=dA_i(t).$$

Dunque $A_i(t)$ può essere visto come la somma dei valori attesi condizionati al passato degli incrementi infinitesimi del processo di conteggio da 0 fino all'istante $t$.

---

[10] In generale un qualsiasi processo $X$ non decrescente adattato tale che $EX(t) < \infty$ per ogni $t$ è una sottomartingala. Infatti, se $X$ è non decrescente, per ogni $h>0$ $X(t+h) \geq X(t)$ q.c. e questo implica che

$$E(X(t+h)\,|\,F_t) \geq E(X(t)\,|\,F_t) = X(t) \text{ q.c..}$$

[11] Una martingala $M$ continua a destra è prevedibile solo se è continua (si veda par 1.4 in Fleming e Harrington, 1991).

### 1.5.2 Martingale locali e teorema di scomposizione generale

Si dice che una funzione $f$ da $[0,\infty)$ in **R** possiede localmente una proprietà se gode di tale proprietà su tutti gli intervalli $[0,s]$ per ogni $0 \leq s < \infty$. Anche per i processi stocastici esiste una definizione analoga, con la differenza che, poiché un processo associa a ogni $\omega \in \Omega$ una diversa funzione a valori reali, non è detto che per ogni traiettoria gli intervalli sui quali vale la proprietà siano gli stessi. La difficoltà è superata introducendo, come estremo destro dei suddetti intervalli, una sequenza crescente di tempi d'arresto (*stopping times*) $\tau_1,\tau_2,...,\tau_n...$, tali che $\lim_{n\to\infty} \tau_n = \infty$ q.c.. Si dice che il processo $X$ gode localmente di una proprietà se il processo arrestato (*stopped process*) $X_n = \{X(\min(t,\tau_n)) : 0 \leq t < \infty\}$ possiede tale proprietà per ogni $n$.[12] Dunque un processo $M$ è una martingala (sottomartingala) locale se esiste una sequenza crescente di tempi d'arresto tale che $\lim_{n\to\infty} \tau_n = \infty$ q.c. e tale che il processo $\{M(\min(t\,,\tau_n)) : 0 \leq t \leq \infty\}$ sia una martingala (sottomartingala) per ogni $n$.

Per le sottomartingale locali vale il seguente teorema:[13]

***Teorema 1.5.2. (Scomposizione di Doob-Meyer estesa)*** Data una filtrazione che soddisfa le condizioni usuali di continuità a destra e completezza, sia $X$ una sottomartingala locale non negativa continua a destra. Allora esiste un unico processo A crescente, continuo a destra, $F_t$-prevedibile, tale che:
1) $A(0)=0$ q.c.;
2)$P(A(t)<\infty)=1$ per ogni $t>0$
2) $\{M(t)=X(t)-A(t): t \geq 0\}$ è una $F_t$-martingala locale continua a destra.

Il teorema 1.5.1 è applicabile solo quando il processo di conteggio è una sottomartingala, ossia quando la media di $N_i(t)$ esiste per ogni $t$. Questa versione più generale rimuove la condizione sul valore atteso, estendendo il campo di applicazione della scomposizione di Doob-Meyer a processi di conteggio arbitrari (si può infatti dimostrare che un qualsiasi processo di conteggio è una sottomartingala locale, basta porre $\tau_n = \min(n, \sup\{t : N(t) < n\})$.[14] In base al teorema 1.5.2 è possibile affermare che, qualsiasi sia la natura dell'evento, non ricorrente o ricorrente, $N_i$ può essere rappresentato in modo

[12]Implicita nella nozione di tempi di arresto è la proprietà: $\{\tau_n \leq t\} \in F_t$. Questo garantisce che i processi $X_n$ siano $F_t$-misurabili. Per approfondimenti si veda Fleming e Harrington (1991), par. 2.2.
[13]Fleming e Harrington (1991), teorema 2.2.3.
[14]Fleming e Harrington (1991), pag.61.

unico come somma di un processo crescente prevedibile $A_i$ e una martingala locale quadrato integrabile $M_i$.[15]

Si noti che, sebbene la funzione valore atteso per una generica martingala locale non sia costante, nel caso in cui essa derivi dalla scomposizione di un processo di conteggio è semplice verificare che $EM_i(t) = 0$ per ogni $t$.[16]

## 1.6 Il processo intensità

In generale il compensatore $A_i$ è un processo continuo a destra, ma spesso è possibile verificarne la continuità assoluta in t. In questo caso si definisce processo intensità per $N_i$ rispetto alla filtrazione $\{F_t : t \geq 0\}$, il processo $l_i$ non negativo prevedibile tale che

$$A_i(t) = \int_0^t l_i(u)du \ .$$

Per capire meglio quale sia il significato euristico del processo intensità, può essere utile enunciare il seguente risultato, dimostrabile sotto alcune condizioni di regolarità:[17]

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathrm{P}\left(N_i(t+\Delta t) - N_i(t) = 1 \middle| F_t\right) = l_i(t+),$$

dove $l_i(t+) = \lim_{h \to 0^+} l_i(t+h)$. Il processo intensità rappresenta la probabilità condizionata al passato di osservare l'evento in un intervallo infinitesimo ($t$ , $t+\Delta t$); in quanto rapporto tra una probabilità e una durata, è interpretabile come un tasso istantaneo.

In generale in ogni istante il processo intensità dipende da tutte le informazioni raccolte in precedenza. Tuttavia se le storie di vita sono

---

[15]Un processo stocastico $X$ si dice quadrato integrabile se

$$\sup_{0 \leq t < \infty} EM^2(t) < \infty .$$

Per la dimostrazione della quadrato integrabilità della martingala locale ottenuta dalla scomposizione di Doob-Meier di un processo di conteggio, si veda Fleming e Harrington (1991), teorema 2.3.2.

[16]Fleming e Harrington (1991), par.1.4.

[17]Fleming e Harrington (1991), teorema 4.2.1.

indipendenti, come nel caso in cui i soggetti sono un campione indipendente e la censura è casuale, ogni elemento della filtrazione $F_t$ è il prodotto di $n$ σ-algebre indipendenti, ciascuna delle quali rappresenta la storia individuale di un unico soggetto fino all'istante $t$ e di conseguenza condizionarsi a $F_t$ equivale a condizionarsi alla storia passata del singolo individuo. In questo caso $l_i$ dipende solo dal passato del soggetto $i$ e l'interpretazione del processo intensità risulta notevolmente semplificata.

Il processo $l_i$ può descrivere situazioni molto diverse, tutto sta nel definirne il tipo di dipendenza dal passato. Per fare un esempio semplice, l'intensità con cui si manifesta un evento di tipo ricorrente all'istante $t$ può essere indipendente dal numero di transizioni già osservate nella storia del soggetto, oppure può essere una funzione crescente o decrescente di tale quantità. Attraverso particolari specificazioni del processo d'intensità si possono rappresentare entrambe queste situazioni.

### 1.6.1 Il processo intensità in analisi di sopravvivenza

Nell'analisi classica dei dati di sopravvivenza,[18] indicato con $T_i$ il tempo di permanenza nello stato iniziale per il soggetto $i$, si definisce *hazard function* la funzione del tempo non negativa

$$\lambda_i(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathrm{P}\left(t < T_i \le t + \Delta t \middle| T_i > t\right) = \lim_{\Delta t \to 0} \frac{P\left(t < T_i \le t + \Delta t\right)}{\Delta t} \frac{1}{\mathrm{P}\left(T_i > t\right)} \quad (1.2)$$

.

Essa esprime il valore istantaneo del rapporto tra la probabilità condizionata che l'individuo muoia nell'intervallo $(t, t + \Delta t]$ e l'ampiezza dell'intervallo stesso.[19] Per esempio, se $\lambda_i$ è crescente (decrescente) significa che il rischio di morire aumenta (diminuisce) col passare del tempo. Se $T_i$ è una variabile continua con funzione di densità $f_i(t)$, è semplice verificare che

$$\lambda_i(t) = \frac{f_i(t)}{S_i(t)} ,$$

[18]Tale approccio è alla base dei lavori di Kalbfleisch e Prentice (1980) e di Cox e Oakes (1984).

[21]In letteratura il termine *hazard function* viene a volte tradotto come funzione di rischio o, più propriamente, con riferimento alla natura stessa della quantità $\lambda_i$, come tasso istantaneo di mortalità.

dove $S_i(t) = 1 - \int_0^t f_i(u)du$ è chiamata funzione di sopravvivenza, poiché rappresenta la probabilità di sopravvivere oltre il tempo *t*.

Basta specificare una sola delle tre funzioni di densità, di *hazard* o di sopravvivenza, perché le restanti due possano essere determinate in modo univoco. In particolare vale la seguente relazione tra *hazard* e sopravvivenza:

$$\lambda_i(t) = \frac{f_i(t)}{S_i(t)} = -\frac{\partial S_i(t)}{\partial t}\frac{1}{S_i(t)} = -\frac{\partial \log S_i(t)}{\partial t} ,$$

da cui

$$S_i(t) = e^{-\Lambda_i(t)} ,$$

dove $\Lambda_i(t) = \int_0^t \lambda_i(u)du$ è il cosiddetto *hazard* cumulato o integrato.

Le funzioni che si utilizzano in ambito classico per caratterizzare il manifestarsi di un evento non ricorrente hanno uno stretto legame con il processo intensità definito nell'ambito dei processi di conteggio. Supponiamo di essere in presenza di un meccanismo di censura a destra e sia $U_i$ il tempo di censura. Si definisce *crude hazard* la funzione $\lambda_i^*(t)$:

$$\lambda_i^*(t) = \lim_{h \to 0} \frac{1}{h} \mathrm{P}(t \le T_i < t + h \,|\, T_i \ge t, U_i \ge t) .$$

Mentre l'*hazard* definito nella (1.2) rappresenta il tasso di mortalità "al netto" della censura, $\lambda_i^*(t)$ ne è influenzato. Se tasso lordo e tasso netto coincidono, è possibile dimostrare che il processo differenza

$$M_i(t) = N_i(t) - \int_0^t Y_i(u)d\Lambda_i(u),$$

dove $\Lambda_i$ è proprio l'*hazard* integrato per il soggetto *i*, è una martingala;[20] dunque per l'unicità della scomposizione di Doob-Meyer, il processo di intensità cumulato nel caso di dati di sopravvivenza è

$$A_i(t) = \int_0^t Y_i(u) d\Lambda_i(u) .$$

In particolare se $\Lambda_i(t)$ è assolutamente continua

$$l_i(t) = Y_i(t)\, \lambda_i(t). \qquad (1.3)$$

La traiettoria del processo intensità coincide con la funzione di rischio negli istanti in cui il soggetto può essere osservato e assume valore 0 altrimenti. Rispetto a $\lambda_i$, il processo intensità $l_i$ contiene anche le informazioni sulla censura, rappresentate dal processo indicatore dell'insieme a rischio $Y_i$. Il modello per l'intensità che sarà definito nel prossimo paragrafo generalizza la (1.3) al caso in cui l'evento è ricorrente.

## 1.7 Il modello a intensità moltiplicativa

Si definisce modello a intensità moltiplicativa il modello generale per la parte sistematica del processo di conteggio $N_i$ in cui $l_i$ è il prodotto di una funzione del tempo non negativa $\lambda_i$ e di un processo osservabile prevedibile $Y_i$ :

$$l_i(t) = Y_i(t)\, \lambda_i(t). \qquad (1.4)$$

La funzione $\lambda_i$ e il processo $Y_i$ rappresentano rispettivamente la componente deterministica e la componente stocastica del processo intensità. In genere nelle applicazioni $Y_i$ è il processo indicatore dell'insieme a rischio, che non a caso è stato definito in modo tale da essere $F_t$-prevedibile. Per analogia con la rappresentazione del processo intensità in analisi di sopravvivenza, chiameremo $\lambda_i$ *hazard function* o tasso istantaneo di transizione.

Il modello (1.4) è un modello generico che non contiene informazioni specifiche circa la dipendenza del processo intensità dal passato o da eventuali

[20] Fleming e Harrington (1991), teorema 1.3.1.

variabili esplicative. Questo tipo di informazioni è specificato attraverso ipotesi sulla forma funzionale di $\lambda_i$.

I modelli per dati di durata che saranno analizzati nei prossimi paragrafi sono tutti modelli a intensità moltiplicativa, caratterizzati da differenti specificazioni dell'*hazard function*.

### 1.7.1 Il concetto di censura indipendente nel modello a intensità moltiplicativa

Nel paragrafo 1.2 è stato definito indipendente un meccanismo di censura in presenza del quale i risultati inferenziali non sono diversi da quelli che si otterrebbero se i dati fossero completi. Nell'ambito del modello a intensità moltiplicativa, dove di fatto l'inferenza viene effettuata sull'*hazard function*, la censura è indipendente se $\lambda_i$ è la stessa, sia in presenza di dati completi che di dati incompleti, quindi se i compensatori del processo di conteggio completo e di quello censurato differiscono solo per il processo indicatore dell'insieme di rischio.

Nel caso particolare di dati di sopravvivenza, la censura è indipendente se rischio netto e rischio lordo coincidono, cioè se la probabilità che il soggetto subisca l'evento in $t$ è la stessa, sia esso osservabile o meno. E' interessante notare che l'indipendenza di $T_i$ e $U_i$ è condizione sufficiente, ma non necessaria affinché $\lambda_i$ e $\lambda_i^*$ siano uguali; ad esempio è possibile dimostrare che la censura di II tipo semplice (esempio 1.2.3) è indipendente, nonostante preveda la dipendenza tra tempo di morte e tempo di censura.[21] Questa considerazione può essere generalizzata: i meccanismi di censura casuale costituiscono un sottoinsieme dei meccanismi di censura indipendente.

## 1.8 Specificazione e stima dei modelli per dati di durata: il modello non parametrico

Nel precedenti paragrafi abbiamo definito il modello a intensità moltiplicativa il quale prevede che $l_i$ sia il prodotto di un fattore stocastico osservabile $Y_i$, indicatore dell'insieme di rischio, e di un fattore deterministico non osservabile $\lambda_i$ ($t$) che, in analisi di sopravvivenza, rappresenta l'*hazard function* della

[21] Fleming e Harrington (1991), esercizio 1.8.

variabile aleatoria tempo di morte. Analizzeremo adesso i metodi statistici comunemente utilizzati per stimare l'*hazard* integrato $\Lambda_i(t)$ nel caso in cui i dati siano incompleti. Inizieremo dalla situazione in cui la popolazione oggetto d'indagine è omogenea ($\Lambda_i(t) = \Lambda(t)$ per ogni $i$) e non è specificata a priori la forma funzionale dell'*hazard*, definendo uno stimatore empirico (o non parametrico) per $\Lambda(t)$, lo stimatore di Nelson-Aalen, che è strettamente legato allo stimatore non parametrico di Kaplan-Meier per $S(t)$, utilizzato nelle analisi di sopravvivenza.[22] Successivamente definiremo il modello di regressione semiparametrico di Cox, in cui $\Lambda_i(t)$ è il prodotto di un fattore completamente incognito comune a tutti gli individui, che è necessario stimare con un metodo non parametrico, e di un fattore dipendente dalle covariate, noto a meno di un vettore di coefficienti di regressione. Verrà infine dedicato un paragrafo ai modelli di tipo parametrico.

Si supponga che la popolazione studiata sia omogenea, ossia che tutti gli individui abbiano la stessa predisposizione a subire l'evento d'interesse nel corso della loro storia di vita. Questo comporta che $\Lambda_i(t) = \Lambda(t)$ per ogni $i$, quindi vale la relazione:

$$\int dN_i = \int Y_i d\Lambda + \int dM_i \,.$$

Qualora non venga formulata nessuna ipotesi a priori sulla forma funzionale di $\Lambda(t)$, è necessario effettuarne la stima in modo non parametrico, a partire dalle sole osservazioni campionarie.

Uno stimatore non parametrico per $\Lambda(t)$ è lo stimatore di Nelson-Aalen (Aalen, 1978):

$$\hat{\Lambda}(t) = \int_o^t \frac{d\overline{N}(u)}{\overline{Y}(u)} \,. \qquad (1.5)$$

dove

$$\overline{N}(t) = \sum_{i=1}^{n} N_i(t) \,.$$

$$\overline{Y}(t) = \sum_{i=1}^{n} Y_i(t).$$

[22] Antoniadis *et al.* (1999) hanno proposto un metodo non parametrico alternativo basato sulle *wavelet*. Anche nel caso di dati di durata a stati ed episodi multipli è possibile specificare modelli non parametrici; si veda , ad esempio, Fahmeir e Klinger (1998).

Il processo stocastico $\overline{N}$ conta il numero totale di transizioni di stato avvenute nel campione negli intervalli di tipo [0, $t$). Se, come abbiamo ipotizzato, l'analisi è in tempo continuo è ragionevole ipotizzare che la probabilità che due o più individui distinti subiscano l'evento contemporaneamente sia nulla. Sotto questa assunzione, $\overline{N}$ è un processo di conteggio, in quanto con probabilità 1 ogni sua traiettoria è caratterizzata da salti di ampiezza unitaria ($P(d\overline{N}(t) > 1) = 0$ per ogni $t$). In un'analisi di sopravvivenza, $\overline{N}(t)$ rappresenta il numero complessivo di morti registrate fino a $t$. $\overline{Y}$ indica invece il numero di soggetti a rischio in ogni istante di tempo; in quanto somma di processi prevedibili, è anch'esso un processo prevedibile.

Affinché l'integrando nella (1.5) sia ben definito è sufficiente assumere per convenzione che 0/0=0; infatti, quando il denominatore $\overline{Y}(t)$ è nullo, necessariamente in $t$ non si registra alcun evento. Ogni traiettoria del processo $\hat{\Lambda}$ è una funzione a scalini crescente, continua a destra, con discontinuità in corrispondenza degli istanti nei quali si verifica almeno una transizione di stato.

Nel caso particolare di analisi di sopravvivenza, anziché il rischio integrato può spesso essere utile stimare la funzione di sopravvivenza. In virtù della relazione biunivoca che lega la funzione di sopravvivenza all'*hazard* integrato, è possibile ricavare euristicamente uno stimatore empirico per $S(t)$, a partire da quello di Nelson-Aalen per $\Lambda(t)$. Infatti, dalla definizione di funzione di sopravvivenza, mediante uno sviluppo in serie di Taylor, si giunge alla seguente approssimazione:

$$S(t) = \exp\left(\int_0^t d\Lambda(s)\right) = \prod_{s\le t} \exp(d\Lambda(s)) \approx \prod_{s\le t} (1 - d\Lambda(s)),$$

e, sostituendo $\hat{\Lambda}(t)$ a $\Lambda(t)$, si ottiene uno stimatore empirico per la funzione di sopravvivenza:

$$\hat{S}(t) = \prod_{s\le t} \left(1 - \Delta\hat{\Lambda}(s)\right) = \prod_{s\le t} \left(1 - \frac{\Delta\overline{N}(s)}{\overline{Y}(s)}\right). \qquad (1.6)$$

$\hat{S}(t)$ è chiamato stimatore di Kaplan-Meier (Kaplan e Meier, 1958).[23]

[23]Storicamente lo stimatore di Kaplan-Meier è precedente a quello di Nelson-Aalen. In origine esso è stato introdotto come lo stimatore che massimizza la verosimiglianza sull'insieme di tutte le funzioni di sopravvivenza ammissibili (Kaplan e Meier, 1958).

Ciascun termine della produttoria (1.6) stima la probabilità di sopravvivere oltre *t*, dato che si è sopravvissuti fino a tale istante; dunque $\hat{S}(t)$ può essere interpretato come prodotto di probabilità condizionate calcolate in corrispondenza di ogni singola durata. Tali stime sono uguali ad 1 nei tempi in cui non si verifica nessun evento, che di conseguenza non contribuiscono al calcolo dello stimatore. Quindi, così come $\hat{\Lambda}(t)$, $\hat{S}(t)$ è una funzione a scalini, con salti nei punti in cui si registra la morte di qualche individuo. Per definizione la funzione di sopravvivenza è monotona decrescente, assume valore 1 in 0 e tende a 0 al crescere di *t*; anche $\hat{S}(t)$ è decrescente (i termini della produttoria sono tutti minori o uguali a 1) e tale che $\hat{S}(0)$=1. Tuttavia, al tendere di *t* all'infinito, $\hat{S}(t)$ può non annullarsi: questo si verifica quando la durata massima osservata corrisponde a un tempo di censura e, di conseguenza, l'ultimo fattore della produttoria è diverso da 0.

Lo stimatore di Nelson-Aalen e lo stimatore di Kaplan-Meier non sono stimatori corretti (rispettivamente per il rischio integrato e la funzione di sopravvivenza), ma la loro distorsione è trascurabile se la probabilità che l'insieme di rischio sia vuoto è sufficientemente piccola per ogni $s \leq t$. E' inoltre possibile dimostrare che sotto opportune condizioni relativamente deboli gli stimatori di Nelson-Aalen e di Kaplan-Meier sono uniformemente consistenti e asintoticamente normali.[24]

## 1.9 Il modello semiparametrico di Cox

Supponiamo che la popolazione oggetto d'indagine sia eterogenea e supponiamo di disporre di una stratificazione del campione rispetto a *k* variabili che teoricamente potrebbe influire sulla predisposizione dei soggetti a subire l'evento di interesse. Per verificare se esista una relazione tra le variabili e il rischio, potremmo stimare in ciascuno strato (sottogruppo omogeneo) l'*hazard function* o la funzione di sopravvivenza (se l'evento non è ricorrente), utilizzando lo stimatore di Nelson-Aalen o di Kaplan-Meier, quindi confrontare le curve così ottenute.[25] Purtroppo però, se la stratificazione è fine, oltre a dover stimare e confrontare molte curve, la numerosità del campione potrebbe non essere sufficiente a garantire un adeguato numero di soggetti per ciascuno

[24]Si veda Fleming e Harrington (1991).
[25]Il test di solito utilizzato per confrontare due curve di sopravvivenza è il cosiddetto *logrank* (o Mantel-Cox) test (Marubini e Valsecchi, 1995).

strato. In generale, per studiare gli effetti simultanei di $k$ possibili variabili esplicative su una variabile dipendente, è opportuno utilizzare un metodo di regressione multipla. Nell'ambito dell'analisi delle storie di vita, i modelli di regressione permettono di analizzare il legame tra il processo d'intensità e le covariate.

Supponiamo che, per ogni soggetto del campione, oltre al processo di conteggio e al processo indicatore dell'insieme di rischio, si osservi un vettore di covariate $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, ..., Z_{ik})$. Le covariate possono anche dipendere dal tempo, quindi in generale, $\mathbf{Z}_i$ è un processo stocastico $k$-variato.[26] La prima ipotesi che è necessario fare su tale processo è la prevedibilità: date le informazioni in $[0, t)$ il valore delle variabili esplicative al tempo $t$ deve essere completamente determinato. Supporremo inoltre che $\mathbf{Z}_i$ sia localmente limitato, ovvero limitato nel caso di covariate indipendenti dal tempo.

Assumendo che sia valido il modello a intensità moltiplicativa (1.4), supponiamo che l'*hazard function* dipenda dal valore delle variabili esplicative:

$$\lambda_i(t) = \lambda(t; \mathbf{Z}_i(t)).$$

Questo significa che il processo

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\lambda(s, \mathbf{Z}_i(s))ds$$

è una martingala locale rispetto alla filtrazione $F^Z{}_t$, che, oltre alle informazioni fornite dal processo di conteggio e da $Y_i$, contiene anche quelle derivanti dall'osservazione delle $k$ covariate. Formalmente tale filtrazione può essere definita nel modo seguente:

$$F^Z{}_t = \sigma\left(N_i(s), Y_i(s+), \mathbf{Z}_i(s+) : 0 \le s \le t, i = 1,2,...,n\right).^{27}$$

---

[26]Usualmente le covariate tempo dipendenti vengono distinte in covariate esterne (esogene) e interne (endogene). Le covariate esterne dipendono da fattori esterni alla storia di vita del soggetto sul quale vengono osservate (ad esempio il livello d'inquinamento atmosferico). Le covariate interne dipendono invece da una qualche variabile che influisce sulla predisposizione del soggetto a subire l'evento d'interesse (condizione di salute, pressione sanguigna,...). Si veda Andersen *et al.* (1993) e Lancaster (1990). In presenza di covariate interne l'interpretazione dei risultati inferenziali richiede molta attenzione, dal momento che l'effetto di tali covariate può mascherare quello di altre variabili esplicative incluse nel modello. Per un esempio si veda Marubini e Valsecchi (1995), paragrafo 6.8.2.

[27]L'ampliamento della filtrazione implica l'ampliamento della classe degli schemi di censura accettabili. Ad esempio, supponiamo che il modello includa come covariate il sesso e l'età. Un meccanismo che ogni anno censuri la donna più anziana è $F^c{}_t$ -prevedibile e indipendente.

In generale, in accordo con l'interpretazione del processo intensità fornita al paragrafo 1.6,

$$E\left(dN(t)\middle|F^{Z}{}_{t}\right)=Y_i(t)\lambda(t,\mathbf{Z}_i(t))dt\ .$$

Nel caso particolare di un'analisi di sopravvivenza su dati censurati a destra, con variabili esplicative costanti, se i tempi di morte e i tempi di censura sono indipendenti condizionatamente alle covariate, $\lambda_i(t,\mathbf{Z}_i)$ rappresenta l'*hazard function* condizionata a $\mathbf{Z}_i$ definita in ambito classico:

$$\lambda(t)=\lim_{\Delta t\to 0}\frac{\mathrm{P}\left(t<T_i\le t+\Delta t\middle|T_i>t,\mathbf{Z}_i\right)}{\Delta t}.$$ [28]

Il modello per l'*hazard function*, che sarà descritto in questo paragrafo e a cui faremo riferimento nel resto del lavoro, appartiene alla classe dei cosiddetti modelli di regressione a rischio moltiplicativo, in cui

$$\lambda_i(t)=\lambda_\circ(t)r(\boldsymbol{\beta}'\mathbf{Z}_i(t)), \tag{1.7}$$

dove $\lambda_\circ(t)$ è una funzione non negativa chiamata *baseline hazard*, $\beta$ un vettore di parametri incognito e *r*(.) una funzione nota non negativa, chiamata funzione di rischio relativo o *link*, che esprime il modo in cui il rapporto tra $\lambda_i(t)$ e il *baseline hazard* dipende dalla funzione lineare $\boldsymbol{\beta}'\mathbf{Z}_i(t)$:

$$\frac{\lambda_i(t)}{\lambda_\circ(t)}=r(\boldsymbol{\beta}'\mathbf{Z}_i(t)). \tag{1.8}$$

Dato che, nel caso di covariate indipendenti dal tempo, la (1.8) esprime la proporzionalità delle *hazard function* individuali, il modello di regressione a rischio moltiplicativo è chiamato anche modello ad *hazard* proporzionali.

La specificazione del modello (1.7) più frequentemente utilizzata è quella con funzione di rischio relativo esponenziale:

$$\lambda_i(t)=\lambda_\circ(t)\exp(\boldsymbol{\beta}'\mathbf{Z}_i(t)). \tag{1.9}$$

[28]Questo risultato costituisce una versione condizionata alle covariate della relazione (1.2). Si veda Fleming e Harrington (1991).

In questo caso $\lambda_{\circ}(t)$ rappresenta l'*hazard function* per un ipotetico individuo sul quale si osservano covariate identicamente uguali a 0; nel caso invece in cui le covariate siano standardizzate rispetto alla media, $\lambda_{\circ}(t)$ rappresenta l'*hazard* per un ipotetico soggetto medio.

Se $\lambda_{\circ}(t)$ non viene specificato, il modello (1.9) è semiparametrico, nel senso che richiede sia la stima di un parametro *k*-dimensionale, $\beta$, sia quella di un parametro a dimensione infinita, $\lambda_{\circ}(t)$. Nella sua versione semiparametrica, esso costituisce una generalizzazione del noto modello proposto da Cox (1972) in ambito classico per dati di sopravvivenza censurati a destra; da qui il nome di modello di regressione di Cox.

### 1.9.1 La stima del modello di Cox: la funzione di verosimiglianza parziale

Il metodo di massima verosimiglianza parziale (Cox, 1975) costituisce un procedimento generale che potenzialmente può essere usato in molti problemi inferenziali. Tuttavia la sua applicazione più frequente è nell'ambito del modello di regressione di Cox che, a causa della sua natura semiparametrica, non può essere stimato servendosi delle usuali tecniche d'inferenza basate sulla verosimiglianza standard. In particolare, l'utilizzo della funzione di verosimiglianza parziale consente di effettuare la stima dei coefficienti di regressione indipendentemente dal *baseline hazard*, che può essere visto come un termine di disturbo a dimensione infinita.

Consideriamo dei dati di sopravvivenza censurati a destra, per i quali sia ipotizzato il modello di regressione (1.9), con covariate indipendenti dal tempo e forma funzionale del *baseline hazard* non specificata. Si supponga inoltre che i dati non presentino parità, ossia che, in corrispondenza di una stessa durata *t*, nel campione sia osservata al più una sola una transizione di stato. Siano $t_1 < t_2 < \ldots < t_L$ le durate in corrispondenza delle quali si sono verificati i decessi e $t^*_{i1} < t^*_{i2} < \ldots < t^*_{ic_i}$ gli istanti in cui sono stati registrati gli episodi di censura tra $t_i$ e $t_{i+1}$. Indicando con *i* l'etichetta del soggetto morto in $t_i$ e con $(i, j)$ quella del soggetto con tempo di censura $t^*_{ij}$, la verosimiglianza parziale per i dati di sopravvivenza è data dalla seguente espressione:

$$L(\boldsymbol{\beta}) = \prod_{i=i}^{L} \frac{\exp(\boldsymbol{\beta}' \mathbf{Z}_i)}{\sum_{j \in R_i} \exp(\boldsymbol{\beta}' \mathbf{Z}_j)}. \tag{1.10}$$

La (1.10) non dipende dal *baseline hazard* e la sola quantità incognita in essa contenuta è il vettore dei coefficienti $\beta$. E' inoltre interessante osservare che i tempi di morte entrano nella verosimiglianza parziale unicamente attraverso il loro rango e quelli di censura attraverso la definizione dell'insieme a rischio.

Se i dati presentano parità (*ties*), come frequentemente accade a causa dell'errore di misura che usualmente accompagna la rilevazione delle durate, la (1.10) deve essere modificata. Se il numero delle parità non è eccessivamente alto, la verosimiglianza parziale è ben approssimata da

$$L(\boldsymbol{\beta}) = \prod_{i=i}^{L} \frac{\exp(\boldsymbol{\beta}' \mathbf{S}_i)}{\left[\sum_{j \in R_i} \exp(\boldsymbol{\beta}' \mathbf{Z}_i)\right]^{m_i}}, \tag{1.11}$$

dove $m_i$ indica il numero degli individui per i quali si registra la morte in $t_i$ e $\mathbf{S}_i$ la somma delle loro covariate. Si osservi che la verosimiglianza parziale approssimata (1.11), non tiene conto che, nonostante alcuni eventi vengano registrati nello stesso istante, in realtà possono essersi manifestati con un certo ordine, provocando una variazione nell'insieme a rischio. Come auspicabile, in assenza di parità la verosimiglianza parziale (1.10) e la sua approssimazione (1.11) coincidono.

Esprimendo i dati secondo la notazione dei processi di conteggio, l'estensione della verosimiglianza parziale (1.11) a eventi ricorrenti, meccanismi di censura più complessi e covariate tempo-dipendenti segue in modo naturale:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{s \geq 0} \left( \frac{Y_i(s) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(s))}{\sum_{j=1}^{n} Y_j(s) \exp(\boldsymbol{\beta}' \mathbf{Z}_j(s))} \right)^{\Delta N_i(s)}. \tag{1.12}$$

Sebbene la (1.12) descriva solo in parte le informazioni contenute nei dati, è utilizzata a fini inferenziali esattamente come una verosimiglianza standard.

Per semplificare la notazione, introduciamo la quantità

$$S^{(0)}(\boldsymbol{\beta},t)=\sum_{i=1}^{n}Y_i(t)\exp(\boldsymbol{\beta}'\mathbf{Z}_i(t)),$$

il vettore a *k* dimensioni

$$\mathbf{S}^{(1)}(\boldsymbol{\beta},t)=\sum_{i=1}^{n}\mathbf{Z}_i(t)Y_i(t)\exp(\boldsymbol{\beta}'\mathbf{Z}_i(t)),$$

e la matrice $k\times k$

$$\mathbf{S}^{(2)}(\boldsymbol{\beta},t)=\sum_{i=1}^{n}(\mathbf{Z}_i(t))^{\otimes 2}Y_i(t)\exp(\boldsymbol{\beta}'\mathbf{Z}_i(t)),$$

dove, per un vettore **a**, $\mathbf{a}^{\otimes 2}=\mathbf{aa}'$. Notiamo che $\mathbf{S}^{(1)}(\boldsymbol{\beta},t)$ e $\mathbf{S}^{(2)}(\boldsymbol{\beta},t)$ rappresentano rispettivamente la derivata prima e la derivata seconda rispetto a β di $S^{(0)}(\boldsymbol{\beta},t)$.

Si definiscano le quantità

$$\overline{\mathbf{Z}}(\boldsymbol{\beta},t)=\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta},t)}{S^{(0)}(\boldsymbol{\beta},t)}$$

e

$$\mathbf{V}(\boldsymbol{\beta},t)=\frac{\mathbf{S}^{(2)}(\boldsymbol{\beta},t)}{S^{(0)}(\boldsymbol{\beta},t)}-(\overline{\mathbf{Z}}(\boldsymbol{\beta},t))^{\otimes 2}=\frac{1}{n}\sum_{i=1}^{n}(\mathbf{Z}_i(t)-\overline{\mathbf{Z}}(\boldsymbol{\beta},t))^{\otimes 2}\frac{Y_i(t)\exp(\boldsymbol{\beta}'\mathbf{Z}_i(t))}{S^{(0)}(\boldsymbol{\beta},t)}$$

,

che possono essere interpretate rispettivamente come il vettore delle medie e la matrice di covarianza empirica delle covariate, calcolati sull'insieme a rischio all'istante *t*, attribuendo a ogni soggetto un peso proporzionale al suo rischio relativo in tale istante.

La funzione di log-verosimiglianza parziale ha la forma

$$\ln L(\boldsymbol{\beta})=\sum_{i=1}^{n}\int_0^{\infty}\left(\ln Y_i(s)+\boldsymbol{\beta}'\mathbf{Z}_i-\ln S^{(0)}(\boldsymbol{\beta},s)\right)dN_i(s).$$

Assumendo di poter scambiare l'ordine di integrale e derivata, il vettore *score* è quindi dato da

$$\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \int_0^{\infty} \{\mathbf{Z}_i(s) - \overline{\mathbf{Z}}(\boldsymbol{\beta}, s)\} dN_i(s).$$

e la derivata rispetto a β di $\mathbf{U}(\boldsymbol{\beta}, t)$, ossia la matrice hessiana per i coefficienti, è la seguente:

$$\mathbf{H}(\boldsymbol{\beta}, t) = -\sum_{i=1}^{n} \int_0^{t} \mathbf{V}(\boldsymbol{\beta}, s) dN_i(s).$$

Lo *score* parziale gode di proprietà simili allo *score* definito nell'ambito della teoria della verosimiglianza standard. In particolare è possibile dimostrare che i contributi individuali allo *score* hanno media 0 e sono tra loro incorrelati e che la covarianza dello *score* è uguale alla matrice d'informazione attesa di Fisher, definita come il valore atteso dell'Hessiano cambiato di segno. [29]

La stima dei coefficienti del modello di Cox viene effettuata risolvendo il sistema di *k* equazioni in *k* incognite:

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}.$$

Il vettore soluzione $\hat{\boldsymbol{\beta}}$ è detto stima di massima verosimiglianza parziale. Per semplicità supporremo sempre che tale stima esista e sia unica.

Il fatto di non utilizzare tutte le informazioni contenute nei dati causa una certa perdita di efficienza dello stimatore di massima verosimiglianza parziale, tuttavia è possibile dimostrare che sotto opportune condizioni esso gode della consistenza e della normalità asintotica, come gli stimatori di massima verosimiglianza standard.[30] Per quanto riguarda infine la verifica d'ipotesi, se la numerosità campionaria è sufficientemente elevata, possono essere utilizzati anche in ambito di verosimiglianza parziale i test classici: lo *score* test, il test del rapporto di verosimiglianza e il test di Wald.[31]

[29]Fleming e Harrington (1991).

[30]Cox (1975) fornisce una dimostrazione delle proprietà asintotiche secondo la notazione classica. Andersen e Gill (1982) ottengono gli stessi risultati sfruttando le proprietà dei processi di conteggio e martingala.

[31]Si veda Andersen *et al.* (1993), pagg. 486-7, e Fleming e Harrington (1991), teorema 8.3.4.

Una volta ottenute le stime dei parametri è necessario stimare in modo non parametrico il *baseline hazard* cumulato. Brevemente, lo stimatore di $\Lambda_{\circ}(t)$ comunemente utilizzato è il cosiddetto stimatore di Breslow (Breslow, 1974):

$$\hat{\Lambda}_{\circ}(t) = \int_0^t \left[ \sum_{i=1}^{n} Y_i(s) \exp\left(\hat{\boldsymbol{\beta}}' \mathbf{Z}_i(s)\right) \right]^{-1} d\overline{N}(s).$$

Notiamo che $\hat{\Lambda}_{\circ}(t)$ equivale a uno stimatore di Nelson-Aalen applicato a un campione omogeneo il cui insieme di rischio in *t* abbia numerosità pari a $\sum_{i=1}^{n} Y_i(t) \exp\left(\hat{\boldsymbol{\beta}}' \mathbf{Z}_i(t)\right)$. Come è ovvio, se le covariate sono identicamente uguali a 0 per ogni soggetto, il modello di regressione di Cox si riduce al modello non parametrico per l'*hazard function* all'interno di una popolazione omogenea e lo stimatore di Breslow allo stimatore di Nelson-Aalen per il rischio cumulato.

### 1.9.2 Il processo *score*

Il vettore *score* rappresenta il valore per $t = \infty$ del processo stocastico $\{\mathrm{U}(\boldsymbol{\beta}, t) : t \geq 0\}$, tale che

$$\mathbf{U}(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} \int_0^t \left(\mathbf{Z}_i(s) - \overline{\mathbf{Z}}(\boldsymbol{\beta}, s)\right) dN_i(s).$$

Questo processo prende il nome di processo *score*. Tenendo conto che la somma degli scarti dalla media è nulla, ove sia ponderata con gli stessi pesi utilizzati per calcolare il valor medio, è semplice verificare che

$$\mathbf{U}(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} \int_0^t \left(\mathbf{Z}_i(s) - \overline{\mathbf{Z}}(\boldsymbol{\beta}, s)\right) dM_i(s),$$

dove $M_i(t) = N_i(t) - \int_0^t Y_i(s) \exp\left(\boldsymbol{\beta}' \mathbf{Z}_i(s)\right) \lambda_o(s) ds$. Dunque il processo *score* è somma di *n* processi individuali $\mathbf{U}_i$ tali che

$$\mathbf{U}_i(\boldsymbol{\beta},t)=\int_0^t\left(\mathbf{Z}_i(s)-\overline{\mathbf{Z}}(\boldsymbol{\beta},s)\right)dM_i(s). \tag{1.13}$$

Nonostante l'uguaglianza (1.13) valga qualsiasi sia il valore di $\boldsymbol{\beta}$ e $\Lambda_\circ(t)$, la differenza $M_i$ è una martingala locale solo se calcolata sui veri coefficienti di regressione e la vera funzione di rischio. Supponendo che quest'ultima sia nota e indicando con $\boldsymbol{\beta}^*$ il vero valore di $\boldsymbol{\beta}$, per il teorema 2.4.3 in Fleming e Harrington (1991), i processi $\mathbf{U}_i(\boldsymbol{\beta}^*, t)$ sono delle martingale locali quadrato integrabili.

L'interpretazione martingala del processo *score* consente di individuarne in modo relativamente semplice le proprietà asintotiche. In particolare, dal teorema del limite centrale per martingale,[32] segue che, sotto opportune condizioni di regolarità, il processo *score* normalizzato, valutato in corrispondenza del vero valore dei parametri, converge in distribuzione a un processo gaussiano *k*-variato **W** a media nulla e incrementi indipendenti:

$$\frac{1}{\sqrt{n}}\mathbf{U}\left(\boldsymbol{\beta}^*,t\right)\xrightarrow{\ D\ }\mathbf{W}(t).$$

Inoltre, indicando con $\sum\left(\boldsymbol{\beta}^*,t\right)$ la matrice $k\times k$ il cui elemento $(i,j)$ descrive, al variare di $t$, la funzione di covarianza tra l'$i$-esima e la $j$-esima componente del processo **W**, si può dimostrare che, se $\hat{\boldsymbol{\beta}}$ è uno stimatore consistente di $\beta^*$, la matrice d'informazione osservata diviso $n$, $\frac{1}{n}\sum_{i=1}^{n}\int_0^t\mathbf{V}\left(\hat{\boldsymbol{\beta}},s\right)dN_i(s)$, rappresenta uno stimatore uniformemente consistente per $\sum\left(\boldsymbol{\beta}^*,t\right)$.[33]

Come diretta conseguenza della convergenza del processo *score* al processo **W**, il vettore *score* calcolato in $\boldsymbol{\beta}^*$ ha distribuzione asintotica normale con media nulla e matrice di covarianza $\sum\left(\boldsymbol{\beta}^*,\infty\right)$, che può essere stimata da $-n^{-1}\mathbf{H}\left(\hat{\boldsymbol{\beta}},\infty\right)$.

[32]Per l'enunciato e la dimostrazione di questo teorema si rimanda a Fleming e Harrington (1991) e ad Anderson e Gill (1982).

[33]Per l'enunciato formale e la dimostrazione completa di questi risultati rimandiamo a Fleming e Harrington (1991).

## 1.10 La specificazione e la stima di modelli parametrici

Consideriamo un generico modello parametrico in cui, per ogni *i*, l'*hazard function* sia definita a meno di un vettore di parametri θ, comune a tutti gli individui (se si suppone che la popolazione sia omogenea, $\lambda_i(t;\boldsymbol{\theta}) = \lambda(t;\boldsymbol{\theta})$ per ogni *i*). La funzione di verosimiglianza standard per le osservazioni può essere scritta come un prodotto di contributi infinitesimi,

$$\prod_{0\le s\le\infty} P\left(dN_1(s), dN_2(s), ..., dN_n(s) \middle| F_{s-}\right) \times P\left(\text{censura in ds} \middle| F_{s-}, dN_1(s), dN_2(s), ..., dN_n(s)\right). \tag{1.14}$$

Supponiamo che la censura sia non informativa, ossia che il secondo termine nella produttoria non dipenda dai parametri d'interesse e che quindi possa essere trascurato senza perdita d'informazioni. Dato che, in ogni intervallo di tempo infinitesimo, o non si verifica nessuna transizione di stato, o se ne verifica una per uno degli *n* individui osservati, il primo termine della produttoria (1.14) rappresenta la verosimiglianza di una variabile aleatoria (*n*+1)-variata con densità multinomiale di parametri ($k = 1, p_1, p_2, ..., p_n$) dove

$$p_i = P\left(dN_i(s) = 1 \middle| \mathrm{F_{s-}}\right) = l_i(s;\boldsymbol{\theta})ds = Y_i(s)\lambda_i(s;\boldsymbol{\theta})ds .$$

Quindi,

$$L(\boldsymbol{\theta}) = \prod_{0\le s\le\infty} \left\{ \prod_{i=1}^{n} \left[Y_i(s)\lambda_i(s;\boldsymbol{\theta})ds\right]^{dN_i(s)} \left[1 - \sum_{i=1}^{n} Y_i(s)\lambda_i(s;\boldsymbol{\theta})ds\right]^{1-d\bar{N}(s)} \right\}.$$

Infine, tenendo conto che $\exp(-x) \approx 1 - x$ e che il prodotto di esponenziali è l'esponenziale di una somma, dopo semplici passaggi algebrici, possiamo scrivere la funzione di verosimiglianza come prodotto di contributi individuali,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[ \prod_{0\le s\le\infty} \left[Y_i(s)\lambda_i(s;\boldsymbol{\theta})\right]^{dN_i(s)} \exp \int_0^{\infty} Y_i(s)\lambda_i(s;\boldsymbol{\theta})ds \right]. \tag{1.15}$$

E' interessante notare che, nel caso di un'analisi di sopravvivenza, il contributo alla verosimiglianza dell'*i*-esimo soggetto, qualora se ne osservi la morte all'istante *t*, è

$$L_i(\boldsymbol{\theta}) = \lambda_i(t;\boldsymbol{\theta})\exp\left[-\int_0^t \lambda_i(s;\boldsymbol{\theta})ds\right] = \lambda_i(t;\boldsymbol{\theta})S_i(t;\boldsymbol{\theta}) = f_i(t;\boldsymbol{\theta}),$$

ossia la funzione di densità del tempo di morte calcolata in *t*. Invece, se la storia di vita del soggetto è censurata all'istante $t^*$, $L_i(\boldsymbol{\theta})$ si riduce alla sola funzione di sopravvivenza

$$L_i(\boldsymbol{\theta}) = \exp\left[-\int_0^{t^*} \lambda_i(s;\boldsymbol{\theta})ds\right] = S_i(t^*;\boldsymbol{\theta});$$

questo è d'altra parte ovvio, dal momento che una storia non completa fornisce, come unica informazione riguardo all'evento d'interesse, che la sopravvivenza del soggetto è maggiore del tempo di censura.

Dalla (1.15) si ottiene la funzione di log-verosimiglianza

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log L_i(\boldsymbol{\theta}) = \sum_{i=1}^{n}\left[\int_0^{\infty} \log\left(Y_i(s)\lambda_i(s;\boldsymbol{\theta})\right)dN_i(s) - \int_0^{\infty} Y_i(s)\lambda_i(s;\boldsymbol{\theta})ds\right]$$

,

la cui derivata rispetto ai parametri rappresenta il vettore *score* $\mathbf{U}(\boldsymbol{\theta})$

$$\mathbf{U}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}) =$$

$$= \sum_{i=1}^{n}\left[\int_0^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}} \log\left(Y_i(s)\lambda_i(s;\boldsymbol{\theta})\right)dN_i(s) - \int_0^{\infty} Y_i(s)\frac{\partial}{\partial \boldsymbol{\theta}} \log \lambda_i(s;\boldsymbol{\theta})ds\right] =$$

$$= \sum_{i=1}^{n}\int_0^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}} \log \lambda_i(s;\boldsymbol{\theta})\left[dN_i(s) - Y_i(s)\lambda_i(s;\boldsymbol{\theta})ds\right] = \sum_{i=1}^{n}\int_0^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}} \log \lambda_i(s;\boldsymbol{\theta})dM_i(s).$$

Valutando il vettore *score* all'istante *t* anzichè in ∞, è possibile definire un processo *score* $\{\mathbf{U}(\boldsymbol{\theta}, t): t \geq 0\}$. Anche nel caso parametrico, ogni componente del processo *score* valutata in corrispondenza del vero valore dei parametri è una martingala locale. Inoltre, supponendo valide alcune condizioni di regolarità, è possibile dimostrare la normalità asintotica dello *score* e, di conseguenza, la consistenza e normalità asintotica dello stimatore di massima verosimiglianza $\hat{\boldsymbol{\theta}}$.[34]

### 1.10.1 Il modello di regressione a rischio moltiplicativo in ambito parametrico

I risultati enunciati sopra sono validi qualunque sia la specificazione parametrica di $\lambda_i$. In particolare supponiamo che valga il modello di regressione (1.9) con *baseline hazard* noto a meno di un vettore di parametri θ,

$$\lambda_i(t;\boldsymbol{\beta},\boldsymbol{\theta}) = \exp(\boldsymbol{\beta}' Z_i(t))\lambda_0(t;\boldsymbol{\theta}).$$

In questo caso non è necessario ricorrere alla definizione di una funzione di verosimiglianza parziale, poiché i regressori possono essere stimati congiuntamente a $\boldsymbol{\theta}$, massimizzando la funzione di verosimiglianza standard (1.15), ossia risolvendo il sistema di equazioni

$$\mathbf{U}(\boldsymbol{\beta},\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{U}^{(\boldsymbol{\beta})}(\boldsymbol{\beta},\boldsymbol{\theta}) \\ \mathbf{U}^{(\boldsymbol{\theta})}(\boldsymbol{\beta},\boldsymbol{\theta}) \end{bmatrix},$$

dove

$$\mathbf{U}^{(\boldsymbol{\beta})}(\boldsymbol{\beta},\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta},\boldsymbol{\theta}) = \sum_{i=1}^{\infty} \int_0^{\infty} \mathbf{Z}_i(s) dM_i(s)$$

e

[34]Per la dimostrazione delle proprietà asintotiche dello stimatore di massima verosimiglianza si veda Andersen *et al.* (1993), teorema VI.1.1. e teorema VI.1.2.

$$\mathbf{U}^{(\boldsymbol{\theta})}(\boldsymbol{\beta},\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\beta},\boldsymbol{\theta}) = \sum_{i=1}^{\infty} \int_0^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}} \log \lambda_0(s;\boldsymbol{\theta}) dM_i(s).$$[35]

***Esempio 1.10.1.*** Un'usuale specificazione parametrica è il modello di regressione Weibull, in cui il baseline hazard ha la seguente espressione:

$$\lambda_0(t;\alpha) = \alpha t^{\alpha-1}.$$

Qualsiasi sia il valore del parametro $\alpha$, $\lambda_0$ è monotono; in particolare, è crescente se $\alpha > 1$ e decrescente se $\alpha < 1$. Se $\alpha = 1$, il modello Weibull si riduce al modello di regressione Poisson (o esponenziale, limitatamente al caso di eventi non ricorrenti), caratterizzato da una funzione di rischio di base costante. □

[35]Un metodo di stima alternativo basato sulle equazioni di stima (Nadeau e Lawless, 1998) può essere applicato qualora vengano fatte ipotesi soltanto su alcuni momenti della distribuzione.

# I residui nei modelli per dati di durata: definizioni e proprietà

## 2.1 Introduzione

Affinchè un'analisi statistica sia valida è necessario che il modello utilizzato sia appropriato, perché in caso di errata specificazione i risultati dello studio potrebbero essere compromessi. I residui sono stati introdotti per la prima volta nell'ambito della regressione lineare classica, come strumento per valutare la discrepanza tra le osservazioni di una variabile d'interesse e il loro valore stimato. Essi contengono molte informazioni riguardo l'adeguatezza delle ipotesi su cui il modello si basa e attraverso la loro analisi è possibile individuare sia problemi riguardanti l'adattamento globale del modello ai dati, sia scostamenti isolati di singole osservazioni dal loro valore previsto (par. 2.2). In seguito la definizione di residuo è stata estesa a classi più ampie di modelli (par. 2.3).

Per quanto concerne i modelli di regressione per dati di durata, la definizione più nota di residuo è quella di residuo martingala (par. 2.4).[36] Se da una parte i residui martingala costituiscono un'estensione naturale del concetto classico di residuo, in quanto differenza tra il numero osservato e il numero stimato di eventi per ciascun soggetto del campione (eventualmente anche in ogni istante di tempo), dall'altra hanno un'interessante interpretazione come indicatori di eterogeneità non osservabile.

Gli altri tipi di residuo proposti in letteratura derivano nella maggior parte dei casi da particolari trasformazioni dei residui martingala (par. 2.5). E' questo il caso dei residui *score* (parr. 2.6), che costituiscono il contributo individuale alla statistica *score* stimata, e dei residui di devianza (par. 2.8), che derivano da una definizione di residuo utilizzata per la diagnostica dei modelli lineari generalizzati (par. 2.3.2).

Nell'analisi di regressione di Cox, una definizione di natura completamente diversa dalle precedenti è quella di residuo parziale di Schoenfeld (par. 2.7). I residui parziali costituiscono gli incrementi del processo *score* stimato, quindi

[36] Recentemente la teoria delle martingale ha trovato ampia applicazione nell'ambito della verifica di ipotesi e di specificazione del modello, oltre a quelle presentate in questo lavoro. Per alcuni esempi si vedano ad esempio Lam ( 1998), Prentice (1999) e Glidden (1999).

sono relativi a istanti di tempo piuttosto che a individui; inoltre hanno la proprietà, coerente con le procedure d'inferenza basate sulla funzione di verosimiglianza parziale, di dipendere dagli istanti di transizione o censura solo in funzione del loro rango.

Limitatamente all'analisi di sopravvivenza, viene a volte utilizzato come residuo la stima del rischio cumulato valutata in corrispondenza del tempo di morte o di censura. Tuttavia i metodi diagnostici basati su questa quantità presentano dei problemi, soprattutto se il modello è semiparametrico (par. 2.9). Buone proprietà sembrano invece avere in questo ambito i residui *log-odds* e *probit* recentemente proposti per l'individuazione di osservazioni anomale in analisi di sopravvivenza (par. 2.10).

Nel presentare i vari tipi di residuo cercheremo di mettere in evidenza le proprietà che essi hanno in comune con i residui del modello di regressione classico. Vedremo inoltre come alcune definizioni nate nell'ambito del modello semiparametrico di Cox siano applicabili anche ai modelli parametrici.

## 2.2 L'analisi dei residui nell'ambito del modello lineare classico

Per ogni soggetto di un campione casuale di numerosità $n$ si osservi una variabile aleatoria $Y$ (variabile dipendente) e $k$-1 variabili $X_1, X_2, \ldots, X_{k-1}$ che per semplicità supporremo fissate a priori (variabili esplicative). La prima assunzione alla base del modello di regressione lineare normale è che il valore atteso della variabile dipendente, $\mu$, sia una funzione lineare delle variabili esplicative, ossia che, per ogni $i$=1,2,..,$n$

$$\mu_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{k-1i} .$$

La precedente relazione può essere scritta in modo compatto utilizzando la notazione matriciale:

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} ,$$

dove $\mu$ è il vettore $n \times 1$ dei valori attesi, $\mathbf{X}$ la matrice $n \times k$ delle variabili esplicative, la cui prima colonna è costituita dal vettore unitario, e $\boldsymbol{\beta}$ il vettore a $k$ dimensioni dei coefficienti di regressione. Il modello prevede inoltre che i cosiddetti errori o disturbi casuali, dati dalla differenza tra ciascuna variabile dipendente e il corrispondente valore atteso, siano indipendenti, omoschedastici, e distribuiti normalmente, ossia che, indicando con $\mathbf{y}$ il vettore i cui elementi

sono i valori osservati della variabile dipendente, $\mathbf{e} = \mathbf{y} - \mathbf{X}\beta$ sia una normale *n*-variata a media $\mathbf{0}$ e matrice di covarianza scalare $\sigma^2\mathbf{I}$; questa assunzione implica che $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$.

Sia $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ la stima dei minimi quadrati dei coefficienti di regressione, si definiscono residui gli errori stimati

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\boldsymbol{\mu}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} = (\mathbf{I} - \mathbf{H})\mathbf{y}, \qquad (2.1)$$

dove $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ è la matrice idempotente simmetrica (*hat matrix* o matrice di proiezione) che trasforma i valori osservati $\mathbf{y}$ nei valori stimati $\hat{\boldsymbol{\mu}}$. E' semplice verificare che i residui sono ortogonali ai regressori ( $\mathbf{X}'\hat{\mathbf{e}} = \mathbf{0}$ ) e che, se il modello ammette l'intercetta, la loro somma è nulla. Sotto l'ipotesi di corretta specificazione del modello, i residui, così come gli errori teorici, hanno distribuzione Normale a media nulla: $\hat{\mathbf{e}} \sim N(\mathbf{0}, (\mathbf{I}-\mathbf{H})\sigma^2)$, ma, mentre gli errori teorici sono per ipotesi indipendenti e omoschedastici, i residui sono eteroschedastici e presentano un certo grado di correlazione, poichè la *hat matrix* non è in generale scalare. Tuttavia, se il modello è valido, all'aumentare di *n*, gli elementi diagonali della matrice (**I-H**) tendono a 1 e gli elementi fuori dalla diagonale a 0; quindi, per campioni sufficientemente grandi, i residui si comportano approssimativamente come un campione casuale estratto da una popolazione normale a media 0 e varianza $\sigma^2$.

### 2.2.1 *Leverage*, influenza e *outlier* nel modello lineare classico

Indicando con $\hat{y}_i$ il valore di $y_i$ stimato dal modello, l'elemento $(i,j)$ della matrice di proiezione **H** rappresenta il contributo della *j*-esima osservazione alla stima $\hat{y}_i$. Inoltre, dato che **H** è simmetrica idempotente, l'elemento $h_{ii}$ sulla sua diagonale è tale che

$$h_{ii} = \sum_{j=1}^{n} h_{ij}^2,$$

e quindi può essere considerato una misura del peso (*leverage*) esercitato dall'individuo *i* nel calcolo dell'intero vettore $\hat{\mathbf{y}}$. In particolare, nell'ambito della regressione lineare semplice, è immediato verificare che

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}, \qquad (2.2)$$

dove $x_i$ è il valore del regressore per l'*i*-esimo individuo e $\bar{x} = \sum x_i / n$. La (2.2) mostra chiaramente che il peso è una funzione crescente dello scostamento della variabile esplicativa dalla sua media empirica calcolata sui soggetti del campione. [37]

Mentre la definizione di peso dipende dalle sole variabili indipendenti, il concetto di *outlier* è strettamente legato a quello di residuo. Si definisce *outlier* un soggetto per il quale si registri una notevole discrepanza tra il valore osservato e il valore stimato della variabile dipendente. Gli *outlier* sono facilmente individuabili attraverso grafici dei residui OLS. La loro presenza può essere sintomo di errata specificazione del modello o presenza di errori nella registrazione dei dati. Altre volte invece gli *outlier* sono semplicemente delle osservazioni anomale degne di attenzione.

E' sempre opportuno controllare se la rimozione di un *outlier* provoca una variazione rilevante delle stime del modello. A questo proposito è utile definire delle opportune misure dell'influenza che ciascun punto esercita sulla stima dei parametri.

Indicando con $\hat{\boldsymbol{\beta}}_{(-i)}$ la stima dei parametri effettuata sulla base di tutte le osservazioni tranne l'*i*-esima, si può valutare l'influenza dell'individuo *i* calcolando la differenza $\Delta\boldsymbol{\beta}_{(-i)} = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}$. Poichè questo approccio ha lo svantaggio di produrre ben *k* valori per ogni soggetto a volte è preferibile utilizzare delle misure che sintetizzino $\Delta\boldsymbol{\beta}_{(-i)}$ in un unico indice. Una tra le più comuni è la distanza di Cook:

$$D_i = \frac{\hat{e}_i^{\,2}}{(k+1)\hat{\sigma}^2} \times \frac{h_{ii}}{(1-h_{ii})},$$

dove $\hat{e}_i$ e $h_{ii}$ (*i*-esimo elemento sulla diagonale della matrice **H**) sono rispettivamente il residuo e il peso relativi all'individuo *i*, mentre $\hat{\sigma}^2 = \dfrac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n-2}$ è

[37] Fox (1984), par. 3.2.1.

la stima della varianza dei termini d'errore.[38] Dato che $D_i$ è una funzione crescente del *leverage* e del residuo del soggetto *i*, se un individuo presenta valori anomali delle variabili esplicative ed è un *outlier* rispetto alla variabile dipendente, è un individuo influente.

D'altra parte, un punto con un elevato *leverage* che giaccia sulla retta di regressione non è influente, così come un *outlier* le cui covariate assumano valori vicini alla media del campione. Così come gli *outlier*, anche i punti influenti dovrebbero essere analizzati con attenzione, quali indicatori di eventuali errori o anomalie nei dati spesso relative all'origine dell'osservazione.

### 2.2.2 La verifica della corretta specificazione del modello

I residui di regressione, in quanto stime dei disturbi casuali, contengono molte informazioni che possono essere utilizzate per verificare la validità di gran parte delle assunzioni fatte sulla distribuzione degli errori stessi e quindi sul modello.

Per quanto concerne la verifica dell'assunzione di normalità degli errori, i metodi diagnostici più semplici sono basati sulla distribuzione dei residui standardizzati $\tilde{\mathbf{e}} = \hat{\mathbf{e}}/\hat{\sigma}$ , che sotto l'ipotesi di normalità, per *n* sufficientemente grande approssimano un campione casuale da una distribuzione normale con media nulla e varianza unitaria. Ad esempio si possono confrontare graficamente i quantili empirici, calcolati sui residui standardizzati, con i quantili della normale standard (*normal probability plot*), oppure si possono effettuare test sulla curtosi e l'asimmetria della distribuzione dei residui.[39] Altre tecniche comunemente utilizzate si basano sui cosiddetti residui studentizzati o su opportune trasformazioni dei residui OLS, mirate a ottenere delle quantità che costituiscano esattamente un campione casuale normale.[40]

Il grafico dei residui rispetto a ciascun regressore costituisce uno strumento sensibile per individuare l'eventuale violazione dell'ipotesi di linearità del modello: se l'assunzione di linearità è valida, i residui non presentano nessun particolare andamento rispetto alle variabili esplicative. L'eventuale non linearità messa in luce dal grafico può essere spesso aggiustata introducendo come regressore una trasformazione non lineare della variabile esplicativa.[41]

I grafici dei residui rispetto ai regressori consentono anche di individuare l'eventuale eteroschedasticità della componente non osservabile. Tuttavia a tal

[38]La distanza di Cook relativa al soggetto *i* può essere pensata come il valore della statistica F utilizzata per sottoporre a test l'ipotesi nulla $H_o : \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{(-i)}$ . Si veda Fox (1984), pag.168.

[39]Johnston (1991).

[40]Per una rassegna sui possibili test di normalità si veda Fox (1984), par. 3.2.4.

[41]Fox (1984), par. 3.2.5.

scopo è più spesso utilizzato il grafico dei residui rispetto ai valori stimati della variabile dipendente: in presenza di eteroschedasticità si osserva una tendenza dei residui a essere in valore assoluto più piccoli (o più grandi) all'aumentare di $Y$.[42]

Un terzo strumento grafico per la diagnostica del modello è il cosiddetto *added-variable plot*. Si supponga di voler valutare se l'effetto di una variabile $Z$ non inclusa nel modello sia invece significativo. Il grafico dei residui rispetto a $Z$, non è in questo caso adeguato, a meno che $Z$ non sia indipendente dalle variabili incluse nel modello, nel qual caso l'assenza di *trend* indicherebbe che $Z$ è giustamente omessa. E' quindi necessario depurare $Z$ dalla dipendenza con $X_1, X_2, \ldots, X_{k-1}$ stimando il modello di regressione

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{e}^*.$$

Il grafico dei residui $\hat{\mathbf{e}}$ rispetto ai residui $\hat{\mathbf{e}}^*$ fornisce informazioni riguardo alla opportunità di inserire $Z$ nel modello. La presenza di un andamento sistematico significa che la variabile $Z$ "spiega" i residui $\hat{\mathbf{e}}$, quindi deve essere inclusa tra i regressori. Inoltre l'andamento dei residui $\hat{\mathbf{e}}$ rispetto ai residui $\hat{\mathbf{e}}^*$ indica in quale forma funzionale $Z$ debba essere inclusa nel modello.[43]

## 2.3 Alcune definizioni di residuo generalizzato

Fuori dall'ambito della regressione lineare standard non esistono definizioni di residuo naturali come quella fornita dalla (2.1). Per questo motivo diversi autori hanno proposto delle definizioni generali applicabili a un ampia classe di modelli statistici (Cox e Snell, 1968; McCullagh e Nelder, 1984; Gourieroux *et al*., 1987).

### 2.3.1 I residui di Cox-Snell

La prima definizione di residuo generalizzato si trova in Cox e Snell (1968). Si consideri un generico modello in cui le variabili osservate siano espresse in

[42]Fox (1984), par. 3.2.6.
[43]McCullagh e Nelder (1991), pag. 399.

termini di un vettore di parametri incogniti $\boldsymbol{\beta}$ e di un vettore **e** di termini di errore indipendenti e identicamente distribuiti. In particolare si assuma che ogni osservazione $Y_i$ dipenda su un solo disturbo casuale, così che si possa scrivere:

$$y_i = g_i(\boldsymbol{\beta}, e_i) \qquad i = 1,2,...,n.^{44}$$

(2.3)

Indicando con $\hat{\boldsymbol{\beta}}$ una stima di $\boldsymbol{\beta}$, si supponga che ogni equazione in $e_i$

$$y_i = g_i(\hat{\boldsymbol{\beta}}, e_i)$$

ammetta un'unica soluzione $r_i$. Cox e Snell (1968) definiscono $r_i$ residuo generalizzato per il modello (2.3). Anche se non è possibile ricavare in generale la distribuzione congiunta degli $r_i$, i due autori dimostrano le seguenti relazioni tra i momenti dei residui e degli errori teorici, nel caso in cui $\hat{\boldsymbol{\beta}}$ sia la stima di massima verosimiglianza per i parametri:

$$E(r_i) = E(e_i) + a_i$$

$$Var(r_i) = Var(e_i) + c_{ii}$$

$$Cov(r_i, r_j) = c_{ij},$$

dove $a_i$, $c_{ii}$ e $c_{ij}$ sono termini di ordine $n^{-1}$. Quindi, anche i residui di Cox-Snell soddisfano proprietà analoghe a quelle che abbiamo visto essere valide per i residui del modello lineare classico. All'aumentare della numerosità campionaria, il comportamento degli $r_i$ diventa sempre più simile a quello dei disturbi teorici relativamente ai momenti di primo e secondo ordine.

Osserviamo infine che i residui del modello di regressione classico rappresentano un caso particolare dei residui di Cox-Snell.

[44]Questo modello, seppur molto generale, non include come casi particolari modelli in cui ciascuna variabile dipendente dipende da più di un termine di errore.

### 2.3.2 I residui di devianza nell'ambito dei modelli lineari generalizzati

McCullagh e Nelder (1989) hanno proposto alcune definizioni di residuo nell'ambito dei modelli lineari generalizzati. Tra queste compare quella di residuo di devianza.

Un modello statistico si dice saturo se prevede un parametro per ogni osservazione del campione e, di conseguenza, si adatta perfettamente ai dati. Per fare un esempio semplice, in un modello lineare classico saturo ($k=n$), la matrice **X** è invertibile, quindi la matrice di proiezione **H** è la matrice identità. Questo comporta che i valori osservati **y** e i valori stimati $\hat{\boldsymbol{\mu}}$ coincidano, ossia che la funzione di regressione passi per tutte le osservazioni. Il modello saturo può essere preso come riferimento per valutare la discrepanza tra la stima del modello scelto per l'inferenza e le osservazioni.

Dato un modello lineare generalizzato, si supponga di effettuare una riparametrizzazione, così che la verosimiglianza sia espressa in funzione del vettore $\boldsymbol{\mu}$ dei valore attesi delle variabili dipendenti. Osserviamo che, sotto questa parametrizzazione, la massima verosimiglianza per il modello saturo può essere scritta come $L(\mathbf{y})$. Viene chiamata devianza la quantità:

$$D = 2\phi\left(\log L(\mathbf{y}) - \log L(\hat{\boldsymbol{\mu}})\right),$$

dove $\hat{\boldsymbol{\mu}}$ è la stima di massima verosimiglianza per $\boldsymbol{\mu}$ e $\phi$ è un parametro di dispersione.[45] Dato che $L(\mathbf{y})$ rappresenta la massima verosimiglianza possibile per le osservazioni, $D$ è una quantità sempre positiva che misura l'adattamento del modello ai dati.[46]

Indicando con $d_i$ il contributo del soggetto $i$ alla devianza ($D = \sum_i d_i$), McCullagh e Nelder definiscono residuo di devianza la radice quadrata con segno di $d_i$

$$r_i^D = \frac{|y_i - \hat{\mu}_i|}{y_i - \hat{\mu}_i}\sqrt{d_i} \ ,$$

[45]Nell'ambito dei modelli lineari generalizzati, la varianza delle variabili dipendenti è usualmente espressa come il prodotto di due termini, uno dipendente dalla media, l'altro da un parametro di dispersione $\phi$ costante sulle osservazioni:

$$\text{var}(y_i) = V(\mu_\iota) \times a_i(\phi)$$

McCullagh e Nelder (1989).

[46]McCullagh e Nelder (1989), par. 2.3.

dove, in accordo con la notazione usuale, $y_i$ e $\hat{\mu}_i$ sono rispettivamente il valore osservato e il valore stimato della variabile dipendente.

E' possibile dimostrare che la distribuzione asintotica dei residui di devianza approssima quella di un campione casuale normale.

***Esempio 2.3.1.*** Nell'ambito del modello lineare normale, tenuto conto che $\phi$ corrisponde alla varianza (comune) delle variabili esplicative, è semplice verificare che

$$D = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 .$$

Quindi i residui di devianza coincidono con i residui di regressione classici. □

***Esempio 2.3.2.*** Se **y** è un vettore di osservazioni indipendenti estratte da *n* distribuzioni Poisson ($\phi$ =1), la funzione di devianza è la seguente:

$$D = 2\sum_{i=1}^{n} \left( y_i \ln \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i) \right)$$

e

$$d_i = y_i \ln \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i) . \square$$

Nel paragrafo 2.12 la definizione di residuo di devianza sarà estesa ai modelli di regressione per dati di durata.

## 2.4 La definizione di residuo martingala

Come abbiamo visto nel primo capitolo, ogni processo di conteggio $N_i$ può essere scomposto in modo unico nella somma di una martingala locale quadrato integrabile $M_i$ e di un processo prevedibile $A_i$, il processo d'intensità

integrato. In particolare, nell'ambito del modello a intensità moltiplicativa, si assume che il processo di conteggio sia del tipo:

$$N_i = \int Y_i d\Lambda_i + M_i \tag{2.4}$$

dove $Y_i$ è il processo che indica l'appartenenza del soggetto all'insieme di rischio e $\Lambda_i$ è la funzione di rischio integrato. Essendo un processo non prevedibile, la martingala $M_i$ può essere interpretata come una componente di errore casuale. Per ogni $t$, $M_i(t)$ rappresenta la differenza tra il numero di eventi osservati in [0, $t$] per l'$i$-esimo individuo e il corrispondente numero teorico di eventi sotto le ipotesi del modello. E' possibile dimostrare che analogamente agli errori del modello di regressione semplice gli $M_i(t)$ hanno media nulla e sono tra loro incorrelati:

$$EM_i(t) = 0 \quad i = 1,2,...,n$$

$$E\left[M_i(t)M_j(t)\right] = 0 \quad i \neq j.^{47}$$

Si indichi con $\hat{\Lambda}_i$ una generica stima del rischio integrato per il soggetto $i$. Si definisce *residuo martingala* al tempo $t$ la quantità:

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s)d\hat{\Lambda}_i(s), \tag{2.5}$$

che rappresenta la differenza tra il numero osservato e il numero stimato di eventi che si sono verificati nella storia di vita del soggetto $i$ fino all'istante $t$.

E' possibile calcolare $n$ residui martingala, uno per ogni individuo del campione, a ogni diverso istante di tempo; tuttavia usualmente nelle applicazioni i residui martingala vengono valutati alla fine del periodo di osservazione:

$$\hat{M}_i(\infty) = N_i(\infty) - \hat{\Lambda}(\infty).$$

[47]Si veda il lemma 2.3.2 in Fleming e Harrington (1991).

In seguito, per semplificare la notazione porremo $\hat{M}_i(\infty) = \tilde{M}_i$.

***Esempio 2.4.1.*** In analisi di sopravvivenza su dati censurati a destra, dati $X_i = \min(T_i, U_i) = t_i$ e l'indicatore di censura $\delta_i$, i residui martingala hanno la forma seguente:

$$\tilde{M}_i = \delta_i - \int_0^{t_i} \hat{\lambda}_i(s)ds = \delta_i - \hat{\Lambda}_i(t_i).$$

In questo caso i residui martingala assumono valori nell'intervallo [1,-∞). □

### 2.4.1 I residui martingala nel modello di Cox

Il modello di regressione semiparametrico di Cox prevede che

$$\Lambda_i(t) = \exp(\boldsymbol{\beta}'\mathbf{Z}_i(t))\Lambda_\circ(t), \tag{2.6}$$

dove la forma funzionale del *baseline hazard* integrato $\Lambda_\circ$ non è specificata. Si indichi con $\hat{\beta}$ la stima di massima verosimiglianza parziale per i coefficienti di regressione e si supponga di utilizzare per la stima di $\Lambda_\circ$ lo stimatore di Breslow

$$\hat{\Lambda}(t) = \int_0^t \left[\sum_{i=1}^n Y_i(s)\exp\left(\hat{\boldsymbol{\beta}}'\mathbf{Z}_i(s)\right)\right]^{-1} d\overline{N}(s).$$

Il residuo martingala per il soggetto *i* valutato all'istante *t* è

$$\hat{M}_i(t) = N_i(t) - \int_0^t \hat{p}_i(s)d\overline{N}(s), \tag{2.7}$$

dove

$$\hat{p}_i(s) = \frac{Y_i(s)\exp\left(\hat{\boldsymbol{\beta}}'\mathbf{Z}_i(s)\right)}{\sum_{j=1}^{n} Y_j(s)\exp\left(\hat{\boldsymbol{\beta}}'\mathbf{Z}_j(s)\right)}. \qquad (2.8)$$

Si noti che $\hat{p}_i(s)$ è il contributo dell'*i*-esimo soggetto nell'istante *s* alla verosimiglianza parziale valutata in $\hat{\beta}$ e rappresenta una stima della probabilità condizionata all'insieme di rischio che l'individuo *i* subisca l'evento nell'istante *s*. Naturalmente $\sum_{i=1}^{n} \hat{p}_i(s) = 1$ in ogni istante di tempo *s*.

Nell'ambito del modello di regressione lineare con intercetta è possibile dimostrare che la somma dei residui è nulla. Una proprietà analoga è soddisfatta anche dai residui martingala (2.7), infatti, per ogni *t*,

$$\sum_{i=1}^{n} \hat{M}_i(t) = \sum_{i=1}^{n} N_i(t) - \int_0^t \sum_{i=1}^{n} \hat{p}_i(s)\, d\overline{N}(s) = \overline{N}(t) - \int_0^t d\overline{N}(s) = \overline{N}(t) - \overline{N}(t) = 0$$

.

Dunque, per ogni *t*, il numero di eventi che si osservano nel campione in [0,*t*] ($\overline{N}(t)$) coincide col numero previsto sotto il modello ($\sum \int_0^t Y_i(s)\, d\hat{\Lambda}_i(s)$). Si osservi che la proprietà $\sum \hat{M}_i(t) = 0$ deriva strettamente dall'utilizzazione dello stimatore di Breslow per $\Lambda_\circ$, e in generale non è da ritenersi valida qualora la stima del rischio di base venga effettuata con uno stimatore diverso.

Per campioni finiti, il comportamento dei residui martingala è diverso da quello degli errori teorici $M_i(t)$. La conseguenza più evidente della sostituzione dei veri parametri con delle stime, è rappresentata dalla correlazione degli $\hat{M}_i(t)$. Tuttavia dalla consistenza dello stimatore di massima verosimiglianza parziale $\hat{\beta}$ e dello stimatore di Breslow $\hat{\Lambda}_\circ(t)$, segue che, per $n \to \infty$,

$$E\hat{M}_i(t) \to 0 \qquad i=1,2,...,n$$

e

$$E\left[\hat{M}_i(t)\hat{M}_j(t)\right] \to 0 \quad i \neq j \,.$$

Quindi, se la numerosità campionaria è sufficientemente alta, i residui martingala possono essere considerati incorrelati, a media 0, esattamente come i residui teorici $M_i(t)$.

I residui martingala (2.7) valutati in ∞ coincidono con i residui a informazione limitata definiti da Lancaster (1990) in ambito econometrico (*limited information residuals*). Si supponga di inserire nel modello di regressione di Cox *n* parametri $\alpha_i$ specifici per ogni soggetto. Il modello così ottenuto

$$\Lambda_i(t) = \exp(\alpha_i + \boldsymbol{\beta}'\mathbf{Z}_i(t))\Lambda_\circ(t) \qquad \text{i=1,2,...,}n \tag{2.9}$$

ha funzione di log-verosimiglianza parziale pari a

$$\log L(\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{i=1}^{n}\int_0^\infty \left( \log Y_i(s) + \alpha_i + \boldsymbol{\beta}'\mathbf{Z}_i(s) - \log\left(\sum_{i=1}^{n} Y_i(s)\exp(\alpha_i + \boldsymbol{\beta}'\mathbf{Z}_i(s))\right)\right) dN_i(s)$$

.

Lancaster definisce residuo a informazione limitata per il soggetto *i* la derivata parziale della log-verosimiglianza rispetto a $\alpha_i$ valutata in $\boldsymbol{\alpha} = \mathbf{0}$, che, come è semplice verificare corrisponde esattamente al residuo martingala $\tilde{M}_i$.

Questo risultato offre un'interessante interpretazione degli $\tilde{M}_i$. Infatti, se $\boldsymbol{\alpha} \neq \mathbf{0}$, il modello (2.9) descrive una situazione in cui le covariate non spiegano completamente l'eterogeneità del campione; dunque i residui martingala, in quanto misure della variazione della verosimiglianza attorno ad $\alpha = \mathbf{0}$, possono essere pensati come indicatori deterministici della presenza di eterogeneità non spiegata.

Quanto detto nel presente paragrafo è ovviamente valido anche nel caso particolare in cui la popolazione studiata sia omogenea (coefficienti di regressione nulli) e si utilizzi lo stimatore di Nelson-Aalen per la funzione di rischio integrato.

### 2.4.2 I residui martingala nei modelli parametrici

Supponiamo che nel modello (2.6) il *baseline hazard* sia specificato a meno di un vettore di parametri incogniti θ, ossia che il modello sia parametrico. I residui martingala sono dati dalla differenza

$$\hat{M}_i(t) = N_i(t) - \int_0^t \exp\left(\hat{\boldsymbol{\beta}}' \mathbf{Z}_i(t)\right) d\Lambda_\circ\left(t; \hat{\boldsymbol{\theta}}\right). \tag{2.10}$$

dove il vettore $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}\right)$ massimizza la funzione di verosimiglianza standard (1.15).

Dato che il *baseline hazard* viene stimato parametricamente anzichè con lo stimatore di Breslow, in generale la somma dei residui martingala (2.10) è diversa da 0; tuttavia, se il modello è del tipo:

$$\Lambda_i(t) = \exp\left(a + \boldsymbol{\beta}' \mathbf{Z}_i(t)\right) \Lambda_\circ(t),$$

si dimostra che $\sum \tilde{M}_i = 0$. Infatti, in tal caso, la componente del vettore *score* relativa all'intercetta *a* è data da

$$\frac{\partial \log L(a, \boldsymbol{\beta}, \boldsymbol{\theta})}{\partial a} = \bar{N}(\infty) - \sum_{i=1}^{n} \int_0^\infty Y_i(s) \exp\left(a + \boldsymbol{\beta}' \mathbf{Z}_i(s)\right) \lambda_\circ(s) ds = \sum_{i=1}^{n} M_i(\infty),$$

e, poichè, per definizione, tale funzione valutata in corrispondenza delle stime di massima verosimiglianza $\left(\hat{a}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}\right)$ è nulla, segue che

$$\sum_{i=1}^{n} \tilde{M}_i = 0\,. \tag{2.11}$$

Dunque, seppure in ogni singolo istante si registri una discrepanza tra il numero totale di eventi osservati, $\bar{N}(t)$, e la sua stima, $\sum \int_0^t Y_i(s) d\hat{\Lambda}_i(s)$, al termine dello studio ($t = \infty$) le due quantità coincidono. In generale, la presenza dell'intercetta è condizione sufficiente affinchè la somma dei residui martingala

sia nulla, esattamente come nell'ambito del modello di regressione lineare classico.[48]

Per quanto concerne il comportamento asintotico dei residui martingala parametrici, tenuto conto della consistenza degli stimatori di massima verosimiglianza dei parametri, valgono le stesse proprietà enunciate nell'ambito del modello di regressione di Cox.

Osserviamo infine che, secondo un ragionamento analogo a quello seguito nel caso semiparametrico, anche i residui martingala del modello parametrico possono essere interpretati come indicatori dell'eventuale presenza di eterogeneità non spiegata dalle covariate nel senso suggerito da Lancaster.[49]

## 2.5 Trasformazione integrale dei residui martingala

Per ogni soggetto *i*, sia $\{f_i(t): t \geq 0\}$ un processo stocastico prevedibile, localmente limitato, che abbia un significato nell'ambito del modello. Tale processo potrebbe in generale dipendere non solo dai dati relativi all'*i*-esimo soggetto, ma anche dai dati relativi agli altri individui del campione; potrebbe essere univariato o multivariato, osservabile o non osservabile. Una definizione generale di errore teorico motivata dalla (2.4) è data da

$$e_i(t) = \int_0^t f_i(s)dM_i(s) = \int_0^t f_i(s)dN_i(s) - \int_0^t f_i(s)Y_i(s)d\Lambda_i(s). \qquad (2.12)$$

L'errore teorico $e_i(t)$ rappresenta la discrepanza tra l'integrale su [0,*t*] di $f_i$ rispetto al processo di conteggio $N_i$ e il corrispondente integrale rispetto al valore atteso condizionato di $N_i$ sotto le ipotesi del modello.

Dalla teoria per i processi martingala segue che il processo $\{e_i(t): t \geq 0\}$ è in generale una martingala locale quadrato integrabile e gli $e_i(t)$ $i$=1,2,..,$n$

[48]Therneau, Grambsch e Fleming (1990) dimostrano che la (2.11) vale in generale per la classe dei modelli parametrici tali che, se $\hat{\Lambda}$ è una stima di massima verosimiglianza della funzione di rischio integrato, anche $k\hat{\Lambda}$ appartiene allo spazio delle soluzioni per ogni k>0. Osserviamo che condizione sufficiente, ma non necessaria, perchè un modello appartenga a questa classe è che il suo predittore lineare includa l'intercetta.

[49]Lancaster (1990), par. 11.2.

godono della proprietà di incorrelazione e media nulla come i residui martingala teorici $M_i(t)$:

$$E[e_i(t)] = 0 \quad i = 1,2,...,n$$

$$\text{cov}[e_i(t), e_j(t)] = 0 \qquad i \neq j.$$

Inoltre:

$$\text{var}[e_i(t)] = E\int_0^t [f_i(s)]^{\otimes 2} Y_i(s) d\Lambda_i(s).$$

Barlow e Prentice (1988) propongono una definizione generale di residuo, inserendo nella (2.12) una stima della funzione di rischio integrato e della funzione prevedibile $\underline{f}_1(t)$:

$$\hat{e}_i(t) = \int_0^\infty \hat{f}_i(s) dN_i(s) - \int_0^t \hat{f}_i(s) Y_i(s) d\hat{\Lambda}_i(s) = \int_0^t \hat{f}_i(s) d\hat{M}_i(s). \qquad (2.13)$$

Se $\hat{\Lambda}_i(t)$ e $\hat{f}_i(t)$ sono consistenti è logico aspettarsi che i residui (2.13) siano asintoticamente incorrelati a media 0 e che uno stimatore della loro varianza sia dato da

$$n^{-1} \sum_{i=1}^n \int_0^\infty \left[\hat{f}_i(s)\right]^{\otimes 2} Y_i(s) d\hat{\Lambda}_i(s).^{50}$$

***Esempio 2.5.1.*** Nell'ambito del modello di regressione di Cox

$$\hat{e}_i(t) = \int_0^t \hat{f}_i(s) dN_i(s) - \int_0^t \hat{f}_i(s) \hat{p}_i(s) d\overline{N}(s),$$

[50]Barlow e Prentice (1988).

dove $\hat{p}_j(s)$ è definito nella (2.8) e, in particolare, per dati di sopravvivenza censurati a destra, il residuo (2.13) valutato alla fine dello studio si riduce a

$$\hat{e}_i(\infty) = \hat{f}_i(t_i)\delta_i - \sum_{j=1}^{n} \delta_j \hat{f}_i(t_j)\hat{p}_i(t_j). \quad \square \tag{2.14}$$

Gli esempi 2.5.2 e 2.5.3 illustrano due particolari specificazioni del residuo (2.13).

***Esempio 2.5.2.*** Se, per ogni $t$, $f_1(t)= 1$, allora il residuo $\hat{e}_i(t)$ coincide col residuo martingala $\hat{M}_i(t)$. $\square$

***Esempio 2.5.3.*** In un'analisi di sopravvivenza, dove si osserva al più un evento per ogni soggetto, anzichè considerare i semplici residui martingala, può essere sensato confrontare il tempo di morte osservato col tempo di morte previsto sotto il modello. Definendo a questo scopo $f_1(t)=t$, il residuo valutato in $t=\infty$ è dato da

$$\hat{e}_i(\infty) = t_i\delta_i - \int_0^{t_i} s\, d\hat{\Lambda}_i(s).$$

Nell'ambito del modello di regressione di Cox, l'espressione di $\hat{e}_i(\infty)$ esprime in modo particolarmente chiaro il significato del residuo stesso:

$$\hat{e}_i(\infty) = \delta_i t_i - \sum_{j=1}^{n} t_j \delta_j \hat{p}_i(t_j); \tag{2.15}$$

( $\hat{p}_i(t_j)$ rappresenta la probabilità condizionata che il soggetto $i$ subisca l'evento in $t_j$). Osserviamo che, se il soggetto è censurato, il residuo (2.15) a esso relativo non può essere positivo, e, paradossalmente, è minore dei residui relativi a soggetti non censurati con stessa storia covariata e tempo di morte minore. $\square$

Anche i residui *score*, che saranno definiti nei prossimi due paragrafi rappresentano un caso particolare dei residui (2.13).

## 2.6 I residui *score*

Qualora la nostra attenzione sia rivolta non tanto allo scostamento delle osservazioni dal modello stimato, quanto sul contributo dei singoli soggetti alla stima del modello, è naturale utilizzare come residui in senso generalizzato i contributi individuali alla statistica *score* $\mathbf{U}\left(\hat{\boldsymbol{\beta}}\right)$. Poiché lo *score* ha una espressione diversa a seconda che si considerino modelli parametrici o semiparametrici, i residui *score* sono definiti in modo differente nei due casi.

### 2.6.1 I residui *score* nel modello di Cox

Nell'ambito del modello di regressione di Cox, sia $f_i(t) = \mathbf{Z}_i(t) - \overline{\mathbf{Z}}(t;\boldsymbol{\beta})$, dove $\overline{\mathbf{Z}}(t;\boldsymbol{\beta})$ è la media pesata del vettore delle covariate definita al paragrafo 1.9.1. Il residuo di tipo (2.13) basato sul processo $f_i$ rappresenta proprio il valore stimato del processo *score* individuale (1.13), valutato al tempo *t*:

$$\mathbf{U}_i\left(\hat{\boldsymbol{\beta}},t\right) = \int_0^t \left[\mathbf{Z}_i(s) - \overline{\mathbf{Z}}\left(\hat{\boldsymbol{\beta}},s\right)\right] d\hat{M}_i(s).$$

Sebbene sia possibile calcolare per tutti i soggetti del campione un residuo a *k* dimensioni $\mathbf{U}_i\left(\hat{\boldsymbol{\beta}},t\right)$ (una per ogni covariata) in ogni istante di tempo, spesso si fa riferimento al solo residuo valutato in $t=\infty$, che per semplicità indicheremo con $\mathbf{U}_i\left(\hat{\boldsymbol{\beta}}\right)$, chiamato usualmente residuo *score*.[51]

[51] Si vedano Barlow e Prentice (1988), Therneau, Grambsch e Fleming (1990), Andersen *et al.* (1994), Harrel (1997).

***Esempio 2.6.1.*** In un'analisi di sopravvivenza su dati censurati a destra, tenendo conto della (2.14) il residuo *score* può essere scritto come:

$$\mathbf{U}_i\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{U}_i\left(\hat{\boldsymbol{\beta}}, \infty\right) = \left[\mathbf{Z}_i\left(t_i\right) - \overline{\mathbf{Z}}\left(\hat{\boldsymbol{\beta}}, t_i\right)\right]\delta_i - \sum_{j=1}^{n} \delta_j \left[\mathbf{Z}_i\left(t_j\right) - \overline{\mathbf{Z}}\left(\hat{\boldsymbol{\beta}}, t_j\right)\right]\hat{p}_i\left(t_j\right). \square$$

Siccome la *p*-esima componente del residuo *score* si riferisce in particolare alla *p*-esima covariata inclusa nel modello (*p*=1,2,...,*k*), i residui *score* costituiscono un potente strumento per effettuare diagnosi specifiche per ciascun regressore.

Tutte le principali proprietà che caratterizzano i residui martingala sono soddisfatte anche dai residui *score*. Prima di tutto, per definizione di stima di massima verosimiglianza parziale,

$$\sum_{i=1}^{n} U_i\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{0}\,.^{52}$$

Inoltre, come caso particolare dei residui (2.13), i vettori dei residui *score* sono asintoticamente incorrelati, hanno media nulla e la loro varianza può essere stimata dalla quantità:

$$n^{-1}\sum_{i=1}^{n}\int_{0}^{\infty}\left(\mathbf{Z}_i(s) - \overline{\mathbf{Z}}\left(\hat{\boldsymbol{\beta}}, s\right)\right)^{\otimes 2} Y_i(s)\exp\left(\hat{\boldsymbol{\beta}}'\mathbf{Z}_i(s)\right)d\hat{\Lambda}_{\circ}(s).^{53}$$

### 2.6.2 I residui *score* nei modelli parametrici

Nell'ambito del generico modello di regressione parametrica

$$\Lambda_i(t) = \exp\left(\boldsymbol{\beta}'\mathbf{Z}_i(t)\right)\Lambda_{\circ}\left(t;\boldsymbol{\theta}\right),$$

si definisce residuo *score* il vettore

[52] Ovviamente tale somma è nulla anche al tempo 0.
[53] Barlow e Prentice (1988).

$$\mathbf{U}_i\left(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}}\right)=\mathbf{U}_i\left(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}},\infty\right)=\int_0^{\infty}\mathbf{Z}_i(s)d\hat{M}_i(s), \tag{2.16}$$

dove $\left(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}}\right)$ è la stima di massima verosimiglianza dei parametri e $\hat{M}_i(t)$ il residuo martingala. $U_i\left(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}}\right)$ rappresenta infatti il valore stimato del processo *score* individuale valutato alla fine del periodo di studio. E' immediato constatare che i residui *score* (2.16) soddisfano le stesse proprietà dei residui *score* definiti nel caso semiparametrico.

## 2.7 I residui parziali di Schoenfeld

Nell'ambito del modello di regressione di Cox, la statistica *score*, oltre che come somma dei contributi individuali $U_i\left(\hat{\boldsymbol{\beta}}\right)$, può essere scritta nel modo seguente (par. 1.9.2):

$$\mathbf{U}(\boldsymbol{\beta})=\sum_{i=1}^{n}\int_0^{\infty}\left(\mathbf{Z}_i(s)-\overline{\mathbf{Z}}(\boldsymbol{\beta},s)\right)dN_i(s).$$

In particolare, in un'analisi di sopravvivenza su dati censurati a destra

$$\mathbf{U}(\boldsymbol{\beta})=\sum_{i=1}^{n}\delta_i\left(\mathbf{Z}_i(t_i)-\overline{\mathbf{Z}}(\boldsymbol{\beta},t_i)\right),$$

quindi il contributo allo *score* di un individuo il cui tempo di morte sia $t_i$ è dato da

$$\mathbf{r}_i=\mathbf{Z}_i(t_i)-\overline{\mathbf{Z}}(\boldsymbol{\beta},t_i).$$

Schoenfeld (1982) definisce residuo parziale il valore stimato di suddetta quantità

$$\hat{\mathbf{r}}_i = \mathbf{Z}_i(t_i) - \overline{\mathbf{Z}}(\hat{\boldsymbol{\beta}}, t_i).^{54} \qquad (2.17)$$

Per definizione di stima di massima verosimiglianza parziale, la somma dei residui $\hat{\mathbf{r}}_i$ è nulla. Inoltre, mentre gli $\mathbf{r}_i$ hanno media nulla e sono incorrelati, i residui parziali (2.17) soddisfano solo asintoticamente tali proprietà.[55]

E' interessante osservare che i residui di Schoenfeld sono relativi ai tempi di morte più che agli individui, tanto che non sono definiti per i soggetti la cui storia di vita non sia completa.[56] Ogni $\hat{\mathbf{r}}_i$ è un vettore a $k$ dimensioni (una per ogni variabile esplicativa), che esprime la discrepanza tra il valore delle covariate osservato per l'individuo $i$ nel suo istante di morte e quello di un teorico soggetto medio che sia a rischio nello stesso istante. Se le covariate sono costanti al variare di $t$, i residui di Schoenfeld, analogamente alla funzione di verosimiglianza parziale, non dipendono dal tempo, ma solo dal rango delle osservazioni. Questa proprietà non è in generale soddisfatta dai residui martingala, a meno che non venga utilizzato lo stimatore di Breslow per il *baseline hazard*.

Dal momento che i residui parziali di Schoenfeld rappresentano gli incrementi del processo *score* stimato, la loro definizione può essere estesa in modo semplice anche al caso in cui l'evento studiato sia ricorrente.

Notiamo infine che, se la funzione di rischio è specificata parametricamente, i residui (2.17), seppur calcolabili, non hanno l'interpretazione e le proprietà valide in ambito semiparametrico.

## 2.8 I residui di devianza nei modelli per dati di durata

Therneau, Grambsch e Fleming (1990) estendono la definizione di devianza relativa ai modelli lineari generalizzati (par. 2.3.2) ai modelli di regressione a rischio moltiplicativo per dati di durata. Supponiamo che il *baseline* sia noto e che i regressori non dipendano dal tempo. La funzione di verosimiglianza ha la forma seguente:

[54]Schoenfeld (1982) ha definito i residui parziali limitatamente al caso in cui le covariate siano costanti nel tempo. Successivamente Barlow e Prentice (1988) hanno esteso tale definizione anche al caso di covariate tempo-dipendenti.
[55]Per la dimostrazione di queste proprietà si vedano Cox (1975), e Schoenfeld (1982).
[56]Barlow e Prentice (1988).

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[\int_0^{\infty}[\log Y_i(s) + \boldsymbol{\beta}'\mathbf{Z}_i + \log\lambda_\circ(s)]dN_i(s) - \int_0^{\infty} Y_i(s)\exp(\boldsymbol{\beta}'\mathbf{Z}_i)\lambda_\circ(s)ds\right]$$

.

Il corrispondente modello saturo può essere definito semplicemente come il modello che prevede una funzione di rischio diversa per ogni individuo del campione:

$$\Lambda_i(t) = \exp(c_i)\Lambda_\circ(t).$$

Sotto tale modello la funzione di log-verosimiglianza è data da

$$\log L(c_1, c_2, ..., c_n) = \sum_{i=1}^{n}\left[\int_0^{\infty}[\log Y_i(s) + c_i + \log\lambda_\circ(s)]dN_i(s) - \int_0^{\infty} Y_i(s)\exp(c_i)\lambda_\circ(s)ds\right]$$

.

E' semplice verificare che la stima di massima verosimiglianza per i parametri $(\hat{c}_1, \hat{c}_2, ..., \hat{c}_n)$ è tale che

$$N_i(\infty) = \int_0^{\infty} Y_i(s)\exp(\hat{c}_i)d\Lambda_\circ(s) \qquad i=1,2,...n. \tag{2.18}$$

Se il modello è saturo, il numero di eventi osservato e il numero di eventi stimato lungo il corso di ogni singola storia individuale coincidono.[57]

In questo contesto, la funzione di devianza può essere scritta come:

---

[57]Con ridondanza di notazione, Therneau, Grambsch e Fleming (1990) definiscono saturo il modello che prevede un vettore di parametri di regressione $\mathbf{h}_i$ distinto per ogni individuo del campione. Questa notazione ha il solo vantaggio di consentire la definizione di modello saturo anche se le covariate dipendono dal tempo. Tuttavia, in tal caso, è semplice verificare che la stima di massima verosimiglianza per i parametri è tale che:

$$\int_0^{\infty}\mathbf{Z}_i(s)dN_i(s) = \int_0^{\infty}\mathbf{Z}_i(s)Y_i(s)\exp(\hat{\mathbf{h}}_i\mathbf{Z}_i(s)')d\Lambda_\circ(s).$$

Questa condizione coincide con la (2.18) solo se le covariate sono indipendenti da *t* ed è in generale complessa da interpretare, tranne nel caso in cui le covariate tempo-dipendenti siano dicotomiche. Ad esempio, in uno studio clinico, in cui la variabile esplicativa sia unica e assuma valore 1 negli istanti in cui il soggetto è esposto a un trattamento e 0 altrimenti, essa stabilisce che il numero osservato e il numero stimato di eventi che si verificano mentre il soggetto è sottoposto a trattamento (non è sottoposto a trattamento) coincidono.

$$D = 2\sum_{i=1}^{n}\left[\int_0^{\infty}\left[\hat{c}_i - \hat{\boldsymbol{\beta}}'\mathbf{Z}_i\right]dN_i(s) - \int_0^{\infty} Y_i(s)\left(\exp\left(\hat{\boldsymbol{\beta}}'\mathbf{Z}_i\right) - \exp(\hat{c}_i)\right)d\Lambda_{\circ}(s)\right] =$$

$$= 2\sum_{i=1}^{n}\left[\int_0^{\infty}(\hat{c}_i - \boldsymbol{\beta}'\mathbf{Z}_i)dN_i(s) - \tilde{M}_i\right] = 2\sum_{i=1}^{n}\left[\log\frac{\exp(\hat{c}_i)}{\exp(\boldsymbol{\beta}'\mathbf{Z}_i)}N_i(\infty) - \tilde{M}_i\right] =$$

$$= 2\left[\sum_{i=1}^{n}\left(\log\frac{N_i(\infty)}{N_i(\infty) - \tilde{M}_i}\right)N_i(\infty) - \tilde{M}_i\right]. \qquad (2.19)$$

Per analogia con i modelli lineari generalizzati, si definisce residuo di devianza per l'*i*-esimo soggetto la radice quadrata con segno dell'*i*-esimo contributo alla devianza (2.19). E' interessante osservare che i residui di devianza nei modelli per dati di durata hanno espressione analoga ai residui di devianza per il modello Poisson (esempio 2.3.2).

***Esempio 2.8.1.*** Per dati di sopravvivenza censurati a destra, dopo semplici passaggi algebrici, si ottiene

$$D = -2\sum_{i=1}^{n}\left[\tilde{M}_i + \delta_i \log\left(\delta_i - \tilde{M}_i\right)\right],$$

quindi il residuo di devianza è rappresentato dalla quantità $d_i$

$$d_i = \frac{\left|\tilde{M}_i\right|}{\tilde{M}_i}\sqrt{\tilde{M}_i + \delta_i \log\left(\delta_i - \tilde{M}_i\right)}. \square$$

Mentre i residui martingala presentano un'accentuata asimmetria attorno allo 0 in quanto il loro campo di variazione si estende da 1 a -∞, nei residui di devianza, che possono essere considerati una trasformazione dei residui martingala, tale asimmetria risulta mitigata, poichè i residui vicini ad 1 vengono amplificati dal logaritmo, mentre quelli che assumono valori negativi molto grandi sono mitigati dalla radice quadrata.

## 2.9 La stima dell'*hazard* integrato come residuo

Spesso in analisi di sopravvivenza su dati censurati a destra vengono utilizzate come residui le quantità

$$\hat{\Lambda}_i(t_i) = \exp(\boldsymbol{\beta}'\mathbf{Z}_i)\hat{\Lambda}_\circ(t_i) \qquad i = 1,2,...,n \tag{2.20}$$

(si veda per esempio Crowley e Hu, 1977). La ragione principale di questa scelta è che, sotto l'ipotesi di validità del modello, la distribuzione delle variabili aleatorie $\Lambda_i(T_i)$ è nota e non dipende dalla particolare distribuzione dei tempi di censura. Infatti, poiché

$$S_i(t_i) = e^{-\Lambda_i(t_i)}$$

e, per la monotonia crescente della funzione di rischio cumulata,

$$S_i(t_i) = P(T_i \geq t_i) = P(\Lambda_i(T_i) \geq \Lambda_i(t_i)),$$

segue che la funzione di ripartizione per $\Lambda_i(T_i)$ valutata in $\Lambda_i(t_i)$ è

$$F_i(\Lambda_i(t_i)) = 1 - e^{-\Lambda_i(t_i)},$$

ossia $\Lambda_i(T_i) \sim \exp(1)$. Quindi se il modello è valido le quantità $\Lambda_i(t_i)$ costituiscono un campione casuale censurato da una popolazione esponenziale a media unitaria.[58]

La distribuzione dei residui (2.20) è ovviamente diversa da quella dei $\Lambda_i(T_i)$; per esempio, una prima differenza evidente è dovuta alla correlazione tra residui indotta dalla procedura di stima. Tuttavia si può assumere che, se *n* è sufficientemente grande, il comportamento dei residui stimati sia simile a quello dei residui teorici, quindi che, sotto l'ipotesi di validità del modello, il vettore dei $\hat{\Lambda}_i(t_i)$ approssimi un campione casuale estratto da una esponenziale unitaria.

Un metodo grafico frequentemente utilizzato per verificare la bontà di adattamento del modello è il cosiddetto *hazard plot*, ossia il grafico di

[58]Il campione è casuale semplice poichè i tempi di morte $T_i$ *i*=1,2,...,*n* sono per ipotesi indipendenti.

una stima non parametrica (usualmente quella di Nelson -Aalen) della funzione di rischio cumulato per i residui $\hat{\Lambda}_i(t_i)$, rispetto ai $\hat{\Lambda}_i(t_i)$ stessi. Poiché, come è semplice verificare, la funzione di rischio cumulato per una variabile aleatoria esponenziale di parametro $\lambda$ è $\Lambda(t) = \lambda\, t$, se i punti di salto della funzione giacciono approssimativamente sulla bisettrice del primo quadrante, si può concludere che i residui sono stati generati da una variabile aleatoria esponenziale di media unitaria e quindi il modello è valido.

Nel caso parametrico, se i tempi di morte sono di tipo esponenziale, il comportamento dei $\hat{\Lambda}_i(t_i)$ è molto simile a quello dei residui teorici anche per piccoli campioni, quindi i metodi diagnostici tipo l'*hazard plot* possono essere considerati validi. Per modelli anche poco più complessi, come ad esempio il modello Weibull, l'approssimazione è già meno buona. Ma i problemi più gravi sorgono nell'ambito del modello di Cox. Infatti, supponendo di utilizzare lo stimatore di Breslow per il *baseline hazard*, i residui (2.20) assumono la forma particolare:

$$g_i\left(\hat{\boldsymbol{\beta}}\right) = \exp\left(\hat{\boldsymbol{\beta}}'\mathbf{Z}_i\right)\sum_{j=1}^{n}\left[\delta_j / \sum_{j\in R_i}\exp\left(\hat{\boldsymbol{\beta}}'\mathbf{Z}_j\right)\right].$$

e non sono distribuiti esponenzialmente, neanche quando nella loro espressione si sostituiscono le stime di massima verosimiglianza parziale con i veri parametri del modello. In particolare, in assenza di censura è possibile dimostrare che il residuo teorico $g_i(\boldsymbol{\beta})$ corrisponde al valore atteso della variabile $\Lambda_i(T_i)$ (~Exp(1)) condizionato al rango delle osservazioni (Lagakos, 1980). Dunque in questo contesto non è appropriato utilizzare l'*hazard plot*, nè nessun'altra tecnica diagnostica basata sulla presunta distribuzione esponenziale dei residui teorici (2.20). Interessanti studi di simulazione sono stati effettuati a questo proposito da Baltzar-Aban e Pena (1995).

## 2.10 Residui *log-odds* e residui *probit*

Recentemente Nardi e Schemper (1999) hanno proposto due nuovi tipi di residuo da utilizzare per l'individuazione di osservazioni anomale nell'ambito dell'analisi di sopravvivenza (parametrica, semiparametrica e non parametrica),

i residui *log-odds* e i residui n*ormal deviate*. Diversamente dai residui martingala che misurano lo scostamento dal modello in termini di eventi osservati e attesi, questi residui sono misure di discrepanza basate sul tempo di sopravvivenza. L'idea di confrontare dei tempi, anziché il numero di eventi, non è nuova, infatti, come abbiamo visto nell'esempio 2.5.3, anche se non frequentemente utilizzata, esiste una definizione di residuo basata sulla differenza tra tempo di sopravvivenza osservato e tempo di sopravvivenza medio stimato sotto le ipotesi del modello. Quello che è nuovo in questa definizione di residuo è il tipo di confronto su cui essa si basa: anziché misurare la differenza tra tempo osservato e tempo medio, i residui *log-odds* e normal deviate misurano la discrepanza tra tempo osservato e tempo mediano predetto dal modello.

Supponiamo per il momento di essere in assenza di censura. Sia $t_i$ il tempo di morte del soggetto *i* e sia $\hat{S}_i(t)$ il valore stimato della funzione di sopravvivenza associata all'individuo *i*. Il tempo di sopravvivenza mediano stimato sotto le ipotesi del modello è definito come il tempo $\hat{t}_i^M$ tale che $\hat{S}_i(\hat{t}_i^M) = 0.5$. Quindi confrontare $t_i$ con $\hat{t}_i^M$ equivale a confrontare $\hat{S}_i(t_i)$ con il valore 0.5.

Nardi e Schemper propongono di utilizzare come residuo la trasformazione logit o la trasformazione *probit* della quantità $\hat{S}_i(t_i)$, ottenendo nel primo caso il cosiddetto residuo *log-odds*

$$\hat{l}_i = \log \frac{\hat{S}_i(t_i)}{1 - \hat{S}_i(t_i)} \tag{2.21}$$

e nel secondo il cosiddetto residuo normale deviato che potrebbe essere chiamato più brevemente residuo *probit*:

$$\hat{p}_i = \Phi^{-1}\left(\hat{S}_i(t_i)\right) \tag{2.22}$$

dove $\Phi$ è la funzione di ripartizione della normale standard. In quest'ottica di confronto con il tempo mediano, un'osservazione è considerata esattamente predetta dal modello se $\hat{S}_i(t_i) = 0.5$, quindi se l'adattamento del modello alla singola osservazione è perfetto sia il residuo *log-odds* che il residuo *probit* sono uguali a 0, mentre, all'aumentare della discrepanza tra tempo osservato e tempo mediano, il loro valore aumenta in valore assoluto.

Oltre a queste caratteristiche descrittive, i residui *probit* e *log-odds* possiedono anche interessanti proprietà riguardanti la loro distribuzione sotto l'ipotesi nulla di validità del modello. Consideriamo i residui teorici $l_i$ e $p_i$

$$l_i = \log \frac{S_i(t_i)}{1 - S_i(t_i)}$$

$$p_i = \Phi^{-1}(S_i(t_i)).$$

Supponendo nota la vera funzione di sopravvivenza, la distribuzione campionaria di $l_i$ e $p_i$ è perfettamente specificata. Infatti, come noto, la distribuzione della variabile aleatoria $S(T)$ è uniforme in [0,1], e di conseguenza, come è possibile verificare attraverso una semplice trasformazione di variabile, i residui teorici *probit* e *log-odds* costituiscono rispettivamente campioni casuali da una normale standard e da una logistica con media 0 e varianza $\pi_i^2 / 3$.

Se dai residui teorici si passa ai residui stimati, sotto l'ipotesi di corretta specificazione del modello, il precedente risultato è valido asintoticamente (Nardi e Schemper, 1999). Questa caratteristica dei residui *log-odds* e *probit* è estremamente importante, poiché l'esistenza di una distribuzione teorica nota a cui fare riferimento consente di affiancare all'analisi grafica dei residui misure statistiche più affidabili e meno soggettivi. Si ricordi inoltre che né per i residui martingala, né per i residui di devianza esistono risultati soddisfacenti da questo punto di vista.

Finora i residui *log-odds* e *probit* sono stati definiti solo nel caso semplice di dati completi. In effetti questi residui non possono essere calcolati per i soggetti la cui storia di vita risulti censurata, essendo in quei casi il tempo di morte incognito. Nardi e Schemper propongono di stimare i residui delle osservazioni incomplete sostituendo nelle espressioni (2.21) e (2.22) il tempo di morte incognito con il suo valore mediano atteso, condizionato al tempo di censura.

Si indichi con $t_i^C$ il tempo di censura e con $T_i$ la variabile aleatoria che indica tempo di morte del soggetto *i*. Come è semplice verificare

$$\Pr\left(T_i > t \middle| T_i > t_i^C\right) = \frac{\Pr\left(T_i > t, T_i > t_i^C\right)}{S_i\left(t_i^C\right)}$$

e, in particolare, se *t* è un istante successivo a $t_i^C$,

$$\Pr\left(T_i > t \middle| T_i > t_i^C\right) = \frac{S_i(t)}{S_i\left(t_i^C\right)} \tag{2.23}$$

Il tempo di morte mediano condizionato al tempo di censura corrisponde al valore $t_i^*$ tale che $\Pr\left(T_i > t_i^* \middle| T_i > t_i^C\right) = 0.5$. Quindi, tenuto conto della (2.23), il tempo mediano condizionato predetto dal modello può essere stimato come la durata $\hat{t}_i^*$ tale che $\hat{S}_i\left(\hat{t}_i^*\right) = 0.5 \times \hat{S}_i\left(t_i^C\right)$. Sostituendo questo valore della funzione di sopravvivenza nelle definizioni di residuo (2.21) e (2.22) si ottengono le seguenti stime dei residui per i soggetti censurati:

$$\hat{l}_i^C = \log \frac{\hat{S}_i\left(t_i^C\right)}{2 - \hat{S}_i\left(t_i^C\right)} \tag{2.24}$$

$$\hat{p}_i^C = \Phi^{-1}\left(\frac{\hat{S}_i\left(t_i^C\right)}{2}\right). \tag{2.25}$$

Poiché sono calcolati utilizzando la sopravvivenza mediana, $\hat{l}_i^C$ e $\hat{p}_i^C$ assumono valori meno estremi dei corrispondenti residui non osservabili e di conseguenza la loro distribuzione empirica è più concentrata di quella dei residui calcolati in assenza di censura.

E' interessante notare che i residui *log-odds* e normali valutati in base alla (2.24) e alla (2.25) relativi a soggetti censurati precocemente assumono valori molto prossimi a zero e quindi non possono costituire in nessun caso osservazioni anomale. Questo è tuttavia un problema solo apparente, perché le osservazioni di questo tipo, pur essendo teoricamente compatibili con osservazioni anomale di soggetti che "muoiono troppo presto" contengono una quantità di informazioni sull'evento d'interesse troppo esigua per essere ritenute interessanti.

I due autori propongono infine di affiancare alla stima dei residui relativi alle osservazioni incomplete, delle valutazione di probabilità. In particolare suggeriscono di stimare per ciascuna storia censurata con quale probabilità, se l'osservazione fosse completa, la funzione di sopravvivenza valutata al tempo di morte sarebbe minore di un certo valore prefissato $\alpha$ (per esempio 0.05 o 0.025) o, analogamente con quale probabilità il residuo *log-odds* o *probit* farebbe parte della percentuale $\alpha$ di residui che assumono i valori negativi più elevati. Tenendo conto della relazione (2.23) è semplice verificare che la probabilità

che, condizionatamente al tempo di censura, la funzione di sopravvivenza assuma valori minori di $\alpha$ è $\alpha / S_i\left(t_i^C\right)$, se $\alpha < S_i\left(t_i^C\right)$, ed è pari a 1, se $\alpha > S_i\left(t_i^C\right)$:

$$P_i = \min\left(1, \alpha / S_i\left(t_i^C\right)\right).$$

I soggetti usciti tardi dallo studio a causa della censura sono caratterizzati da elevati valori di $P_i$. Potenzialmente sono individui vissuti "troppo a lungo" e come tali sono oggetto d'interesse nell'ottica dell'individuazione di osservazioni anomale. Al contrario, come ovvio, la probabilità relativa ai soggetti usciti dallo studio precocemente senza aver subito l'evento ( $S_i\left(t_i^C\right) \approx 1$ ) è circa pari ad $\alpha$.

# Metodi diagnostici basati sui residui

## 3.1 Introduzione

Come accade in generale per qualsiasi modello statistico, la violazione delle assunzioni alla base del modello di regressione

$$\Lambda_i(t) = \Lambda_\circ(t)\exp(\beta' \mathbf{Z}_i(s)), \qquad (3.1)$$

può compromettere i risultati dell'inferenza. Ad esempio, è possibile dimostrare che, in presenza di eventuali errori di specificazione, lo stimatore di massima verosimiglianza (parziale o standard) perde le ottime proprietà di consistenza e di efficienza rispetto allo stimatore del modello corretto (Lin e Wei, 1989; Struthers e Kalbfleisch, 1986; Lagakos, 1988). E' quindi utile disporre di tecniche grafiche e numeriche per l'individuazione di eventuali errori nella formulazione delle ipotesi di base.

Nel presente capitolo descriveremo il ruolo dei residui (martingala, di devianza, di Schoenfeld, *score*, *log-odds* e *probit*) nella fase di diagnosi del modello, facendo riferimento sia all'ambito semiparametrico che parametrico. Anche in questo caso, come nel modello di regressione classico, l'analisi dei residui consente di individuare osservazioni anomale e punti influenti (par. 3.2 e 3.3), nonchè eventuali errori di specificazione. Per quanto concerne questo secondo punto, l'inadeguatezza del modello può riguardare diversi elementi:

a) l'errata forma funzionale con cui le covariate entrano nel predittore lineare (par. 3.5)

b) la scelta inappropriata del *link* esponenziale (par. 3.6)

c) la violazione dell'ipotesi di rischi proporzionali (PH), se le variabili esplicative sono costanti nel tempo, o, più in generale, la presenza di coefficienti di regressione che dipendono da *t* (par. 3.7).

In ambito parametrico un'ulteriore fonte d'errore può essere l'errata specificazione del *baseline hazard*, ma questo argomento non è stato preso in considerazione in modo approfondito nel presente lavoro.

Come vedremo, alcune delle tecniche descritte ricordano quelle usualmente utilizzate per la diagnostica del modello lineare classico (ad esempio l'*added variable plot* o il grafico dei residui OLS per la linearità), altre, come ad

esempio quelle utilizzate per la verifica dell'ipotesi PH, sono invece strettamente legate alla natura del modello di regressione a rischio moltiplicativo.

Il paragrafo 3.4 sarà interamente dedicato alla descrizione delle proprietà di particolari somme cumulate dei residui martingala (Lin, Wei e Ying, 1993) che, opportunamente specificate, possono essere utilizzate in svariati modi per la verifica del modello.

Infine definiremo un test di specificazione basato sui residui martingala (Lancaster, 1990) il cui scopo è quello di individuare l'eventuale presenza di eterogeneità non osservabile (o *frailty*) nei dati.

## 3.2 Individuazione degli *outlier*

Nell'ambito dei modelli per dati di durata l'individuazione di eventuali *outlier* potrebbe teoricamente essere basata sull'analisi dei residui martingala $\widetilde{M}_i$. Infatti ciascun residuo $\widetilde{M}_i = \int_0^\infty d\hat{M}_i(s)$, rappresenta una misura globale dello scostamento dell'*i*-esima osservazione dal modello lungo tutto il periodo d'osservazione; in particolare, se $\widetilde{M}_i$ assume un elevato valore positivo (negativo) il modello sottostima (sovrastima) eccessivamente il numero di eventi che si verificano nella storia di vita del soggetto. Purtroppo la forte asimmetria che caratterizza i residui martingala rende difficile l'individuazione grafica degli *outlier*. Ad esempio, nelle analisi di sopravvivenza, il grafico degli $\widetilde{M}_i$ presenta di solito una forte concentrazione di punti vicino ad 1, che maschera quasi sempre gli *outlier*. Inoltre, l'estensione del campo di variazione dei residui martingala fino a -∞, può indurre a considerare *outlier* negativi anche osservazioni che in realtà non lo sono.

Un modo semplice e frequentemente proposto in letteratura per risolvere il problema dell'asimmetria consiste nell'analizzare i residui di devianza (par. 2.8). Questi residui specialmente se la percentuale di dati censurati sul campione non è alta, sono circa simmetrici attorno a 0, e, di conseguenza, il loro grafico (usualmente effettuato rispetto al predittore lineare) è caratterizzato da una minore concentrazione dei punti con ordinata maggiore di 0, e una minore dispersione dei punti con ordinata minore di 0. Ciò consente di individuare meglio gli eventuali *outlier* positivi e riduce la possibilità di definire *outlier* negativi spuri, anche se,

come è stato dimostrato attraverso studi di simulazione da Therneau, Grambsch e Fleming (1990), mentre le osservazioni anomale negative (individui che "vivono troppo a lungo") appaiono sempre come punti isolati anche nel grafico dei residui martingala, talvolta l'individuazione delle osservazioni anomale positive (soggetti che "muoiono troppo presto") è difficoltosa anche se si utilizzano i residui di devianza.

I grafici dei residui *log-odds* e *probit* (par. 2.10) appaiono ancora meno asimmetrici dei grafici dei residui di devianza e di conseguenza consentono una migliore individuazione delle eventuali osservazioni anomale.[59] Inoltre, essendo nota la loro distribuzione asintotica in assenza di *outlier*, è anche possibile confrontare i quantili osservati con i quantili teorici o calcolare statistiche test per valutare la discrepanza tra la distribuzione osservata e quella teorica sotto l'ipotesi nulla.

## 3.3 Una misura d'influenza basata sui residui *score*

Nei modelli di regressione per dati di durata, così come nella regressione lineare, l'influenza dell'*i*-esima osservazione sulla stima del modello è espressa dal vettore

$$\Delta \boldsymbol{\beta}_i = \hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}},$$

che costituisce la variazione indotta nella stima dei coefficienti dalla rimozione dell'osservazione *i* dal campione.

Poichè il calcolo diretto di queste differenze (metodo *jacknife*), richiede *n* stime supplementari del modello, in letteratura sono stati proposti dei metodi per approssimare i $\Delta\boldsymbol{\beta}_i$, meno dispendiosi dal punto di vista computazionale (Storer e Crowley, 1985). Uno di questi è basato sui residui *score* definiti nel paragrafo 2.6.

Nell'ambito del modello di regressione a rischi proporzionali con *link* esponenziale, si supponga di assegnare a ciascuna osservazione un peso $w_i$. Se il *baseline hazard* non è specificato parametricamente la funzione di log-verosimiglianza parziale pesata è data da

[59]Per un'analisi comparata tra residui maringala, di devianza, *log-odds* e *probit* si veda Nardi e Schemper, 1999.

$$\ln L(\boldsymbol{\beta},\mathbf{w}) = \sum_{i=1}^{n}\int_0^{\infty} w_i\left[\ln Y_i(s) + \boldsymbol{\beta}'\mathbf{Z}_i(s) - \ln\left(\sum_{j=1}^{n} w_j Y_j(s)\exp\left(\boldsymbol{\beta}'\mathbf{Z}_j(s)\right)\right)\right]dN_i(s)$$

quindi il vettore *score* è

$$\mathbf{U}(\boldsymbol{\beta},\mathbf{w}) = \sum_{i=1}^{n}\int_0^{\infty} w_i\left[Z_i(s) - \frac{\sum_{j=1}^{n} w_j\mathbf{Z}_j(s)Y_j(s)\exp\left(\boldsymbol{\beta}'\mathbf{Z}_j(s)\right)}{\sum_{j=1}^{n} w_j Y_j(s)\exp\left(\boldsymbol{\beta}'\mathbf{Z}_j(s)\right)}\right]dN_i(s).$$[60]

Secondo questa notazione $\Delta\boldsymbol{\beta}_i$ è dato dalla differenza tra la stima dei coefficienti di regressione effettuata quando $w_i$=1 per ogni *i* e la stima ottenuta ponendo $w_j$=1 per ogni $j \neq i$ e $w_i$= 0. Dunque la derivata parziale di $\hat{\boldsymbol{\beta}}$ rispetto a $w_i$, valutata in corrispondenza dei pesi unitari costituisce una stima *jacknife* infinitesima dell'influenza del soggetto *i*. In particolare è semplice verificare che, se la stima del *baseline hazard* è effettuata utilizzando lo stimatore di Breslow, vale la seguente uguaglianza

$$\left[\frac{\partial\hat{\boldsymbol{\beta}}}{\partial w_i}\right]_{\mathbf{w}=\mathbf{1}} = \left[\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\mathbf{U}\left(\hat{\boldsymbol{\beta}},\mathbf{w}\right)}\times\frac{\partial\mathbf{U}\left(\hat{\boldsymbol{\beta}},\mathbf{w}\right)}{\partial w_i}\right]_{\mathbf{w}=\mathbf{1}} = \left[\mathbf{I}\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1}\times\mathbf{S}_i \qquad (3.2)$$

dove $\mathbf{I}\left(\hat{\boldsymbol{\beta}}\right) = -\mathbf{H}\left(\hat{\boldsymbol{\beta}}\right)$ e $\mathbf{S}_i$ è il residuo *score* per il soggetto *i*.[61]

[60] La log-verosimiglianza pesata può essere interpretata come una log-verosimiglianza valutata su un campione teorico in cui ciascun soggetto *i* compare $w_i$ volte, *i* = 1, 2,..., *n*.

[61] L'uguaglianza $\left[\frac{\partial\mathbf{U}\left(\hat{\boldsymbol{\beta}},\mathbf{w}\right)}{\partial w_i}\right]_{\mathbf{w}=\mathbf{1}} = \mathbf{S}_i$ è soddisfatta solo se si effettua la stima del *baseline hazard* integrato con lo stimatore di Breslow. Infatti

$$\frac{\partial\mathbf{U}\left(\hat{\boldsymbol{\beta}},\mathbf{w}\right)}{\partial w_i} = \int_0^{\infty}\left(\mathbf{Z}_i(s) - \mathbf{E}\left(\hat{\boldsymbol{\beta}},s\right)\right)dN_i(s) +$$

La diagnostica (3.2) coincide esattamente con la funzione d'influenza empirica calcolata per altra via da Reid e Crepeau (1985). Anch'essi utilizzano tale quantità per valutare in modo informale l'influenza dell'*i*-esima osservazione sulla stima dei coefficienti.

E' interessante notare che, in ambito semiparametrico, il ruolo giocato dai residui *score* nella valutazione dell'influenza trova una spiegazione euristica nell'espressione stessa degli $\mathbf{S}_i$. Per prima cosa si osservi che la *leverage* del soggetto *i* (inteso come scostamento delle covariate dalla loro media empirica) non è costante: anche nel caso in cui le covariate siano indipendenti dal tempo, il peso dell'individuo si modifica al variare dell'insieme di rischio. Dunque la leverage dell'osservazione *i* al tempo *t* può essere espressa dalla quantità $\mathbf{Z}_i(t) - \mathbf{E}(\boldsymbol{\beta}, t)$. Tenuto conto inoltre che la misura di scostamento dalla stima del modello nell'istante *t* relativa alla stessa osservazione è costituita dal residuo martingala $\hat{M}_i(t)$, il residuo *score* $\mathbf{S}_i = \int_0^\infty \left(\mathbf{Z}_i(s) - \mathbf{E}(\hat{\boldsymbol{\beta}}, s)\right) d\hat{M}_i(s)$ ha natura simile agli usuali indici d'influenza (ad es. la distanza di Cook) definiti per il modello di regressione lineare normale.

Dato che le componenti dei $\Delta\boldsymbol{\beta}_i$ in generale assumono valori su scale anche molto diverse, può essere utile effettuare una standardizzazione delle misure d'influenza. Si indichi con **D** la matrice diagonale il cui elemento (*j*, *j*) è costituito dall'errore standard della stima del *j*-esimo coefficiente di regressione. Moltiplicando la (3.2) per l'inversa della matrice **D**, si ottiene la seguente misura d'influenza standardizzata

---

$$-\sum_{l=1}^{n} w_l \int_0^\infty Y_i(s)\left(\mathbf{Z}_i(s) - \mathbf{E}(\hat{\boldsymbol{\beta}}, s)\right)\left[\frac{\exp(\hat{\boldsymbol{\beta}}'\mathbf{Z}_i(s))}{\sum_{j=1}^{n} Y_j(s) w_j \exp(\hat{\boldsymbol{\beta}}'\mathbf{Z}_j(s))}\right] dN_l(s) =$$

$$= \int_0^\infty \left(\mathbf{Z}_i(s) - \mathbf{E}(\hat{\boldsymbol{\beta}}, s)\right) dN_i(s) - \int_0^\infty \frac{\left(\mathbf{Z}_i(s) - \mathbf{E}(\hat{\boldsymbol{\beta}}, s)\right) Y_i(s) \exp(\hat{\boldsymbol{\beta}}'\mathbf{Z}_i(s))}{\sum_{j=1}^{n} w_j Y_j(s) \exp(\hat{\boldsymbol{\beta}}'\mathbf{Z}_j(s))} \left[\sum_{l=1}^{n} w_l dN_l(s)\right].$$

Quindi

$$\left[\frac{\partial \mathbf{U}(\hat{\boldsymbol{\beta}}, \mathbf{w})}{\partial w_i}\right]_{\mathbf{w}=\mathbf{1}} = \int_0^\infty \left(\mathbf{Z}_i(s) - \mathbf{E}(\hat{\boldsymbol{\beta}}, s)\right)\left[dN_i(s) - Y_i(s)\exp(\hat{\boldsymbol{\beta}}'\mathbf{Z}_i(s))\, d\hat{\Lambda}_\circ(s)\right],$$

dove $\hat{\Lambda}_\circ(t)$ è lo stimatore di Breslow.

$$\left[\mathbf{I}\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1} \times \mathbf{S}_i \times \mathbf{D}\left(\sigma_1, \sigma_2, ..., \sigma_k\right). \qquad (3.3)$$

Possono essere considerate influenti tutte le osservazioni per le quali ogni componente del vettore (3.3) sia maggiore di una valore soglia *u*, ossia quelle la cui eliminazione induca una variazione nei parametri maggiore di *u* deviazioni standard.

Nell'ambito dei modelli per dati di durata non sono state proposte misure che sintetizzino le informazioni contenute nei $\Delta\boldsymbol{\beta}_i$ in un unico valore per ogni soggetto. Specialmente se il numero (*k*) delle covariate incluse nel modello è alto, il fatto di avere, per ogni individuo, *k* misure distinte comporta dei problemi nell'individuazione dei punti influenti, poichè l'eliminazione di un soggetto dal campione può incidere in modo sensibile sulla stima di alcuni parametri, ma non su quella di altri. Una regola ragionevole consiste nel concentrare l'attenzione solo sulle variabili esplicative il cui effetto risulti maggiormente significativo.

Con semplici modifiche, è possibile estendere questi risultati anche in ambito parametrico. E' sufficiente tenere conto che la verosimiglianza pesata è data da

$$\ln L(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{w}) = \sum_{i=1}^{\infty} w_i \left[ \int_0^{\infty} \left[ \ln Y_i(s) + \boldsymbol{\beta}' \mathbf{Z}_i(s) + \ln \lambda_0(s) \right] dN_i(s) - \int_0^{\infty} Y_i(s) \exp\left(\boldsymbol{\beta}' \mathbf{Z}_i(s)\right) \lambda_0(s) ds \right]$$

e che

$$\frac{\partial \ln L(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{w})}{\partial \boldsymbol{\beta}} = \mathbf{U}^{(\boldsymbol{\beta})}(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}) = \sum_{i=1}^{n} w_i \int_0^{\infty} \mathbf{Z}_i(s) dM_i(s).$$

Quindi

$$\left[ \frac{\partial \hat{\boldsymbol{\beta}}}{\partial w_i} \right]_{\mathbf{w}=\mathbf{1}} = \left[ \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \mathbf{U}^{(\boldsymbol{\beta})}\left(\hat{\boldsymbol{\beta}}, \boldsymbol{\theta}, \mathbf{w}\right)} \times \frac{\partial \mathbf{U}^{(\boldsymbol{\beta})}\left(\hat{\boldsymbol{\beta}}, \boldsymbol{\theta}, \mathbf{w}\right)}{\partial w_i} \right]_{\mathbf{w}=\mathbf{1}} = \mathbf{I}_{(\boldsymbol{\beta})}^{-1}\left(\hat{\boldsymbol{\beta}}, \boldsymbol{\theta}\right) \times \mathbf{S}_i \ ,$$

dove $\mathbf{I}_{(\boldsymbol{\beta})}^{-1}\left(\hat{\boldsymbol{\beta}}, \boldsymbol{\theta}\right)$ è la porzione dell'inversa della matrice d'informazione osservata relativa ai coefficienti di regressione.

## 3.4 Proprietà delle somme cumulate di residui martingala

Mentre le informazioni riguardanti la distribuzione asintotica dei residui martingala si riducono essenzialmente a quelle relative ai momenti di primo e secondo ordine (par. 2.4) esistono risultati teorici sulla base dei quali è piuttosto semplice approssimare la distribuzione asintotica di processi stocastici ottenuti dalla somma cumulata dei residui martingala rispetto al tempo o al valore delle covariate.

Per semplicità si supponga che le covariate incluse nel modello di regressione siano indipendenti dal tempo e si consideri il generico processo multiparametrico $W_z = \{W_z(t,\mathbf{z}) : t \geq 0, \mathbf{z} \in \mathbf{R}^k\}$ tale che

$$W_z(t,\mathbf{z}) = \sum_{i=1}^{n} f(\mathbf{Z}_i) I(\mathbf{Z}_i \leq \mathbf{z}) \hat{M}_i(t), \tag{3.4}$$

dove *f*(.) è una funzione nota che liscia il vettore delle covariate, $I(\mathbf{Z}_i \leq \mathbf{z})$ è la funzione indicatrice che assume valore 1 se ogni elemento del vettore $\mathbf{Z}_i$ è minore o uguale al corrispondente elemento del vettore **z** e 0 altrimenti, e $\hat{M}_i(t)$ è il residuo martingala per il soggetto *i* valutato all'istante *t*. Lin, Wei e Ying (1993) dimostrano che, sotto l'ipotesi di corretta specificazione del modello, il processo normalizzato $n^{-1/2} W_z$, all'aumentare della numerosità campionaria, converge in distribuzione a un processo gaussiano a media nulla. In particolare, nell'ambito del modello di Cox, è possibile verificare che tale limite è lo stesso a cui tende la distribuzione condizionata alle osservazioni del processo gaussiano $n^{-1/2} \hat{W}_z$ tale che

$$\hat{W}_z(t,\mathbf{z}) = \sum_{h=1}^{n} \int_0^t \left[ f(\mathbf{Z}_h) I(\mathbf{Z}_h \leq \mathbf{z}) - g(\hat{\boldsymbol{\beta}}, s, \mathbf{z}) \right] G_h dN_h(s) +$$

$$- \sum_{i=1}^{n} \int_0^t Y_i(s) \exp(\hat{\boldsymbol{\beta}}' \mathbf{Z}_i) f(\mathbf{Z}_i) I(\mathbf{Z}_i \leq \mathbf{z}) \left[ \mathbf{Z}_i - E(\hat{\boldsymbol{\beta}}, s) \right] d\hat{\Lambda}_\circ(s) \times$$

$$\times \left( \mathbf{I}(\hat{\boldsymbol{\beta}}) \right)^{-1} \sum_{j=1}^{n} \int_0^\infty \left[ \mathbf{Z}_j - E(\hat{\boldsymbol{\beta}}, s) \right] G_j dN_j(s) \ , \tag{3.5}$$

dove

$$g(\boldsymbol{\beta},t,\mathbf{z}) = \frac{\sum_{i=1}^{n} Y_i(t) I(\mathbf{Z}_i \le \mathbf{z}) f(\mathbf{Z}_i) \exp(\boldsymbol{\beta}'\mathbf{Z}_i)}{\sum_{i=1}^{n} Y_i(t) \exp(\boldsymbol{\beta}'\mathbf{Z}_i)}$$

e $G_1, G_2, ..., G_n$ è un campione casuale estratto da una popolazione normale standard.[62] Sulla base di questo risultato, se il modello è adeguato e *n* è sufficientemente grande, la distribuzione asintotica del processo $W_z$ può essere approssimata attraverso delle simulazioni del processo gaussiano $\hat{W}_z$ , ottenute generando ripetutamente dei campioni casuali normali.

Come vedremo nel prossimo paragrafo, il confronto delle traiettorie simulate con la traiettoria osservata costituisce un possibile strumento per verificare la corretta specificazione del modello. Dato che nei metodi diagnostici che noi descriveremo porremo sempre o $t = \infty$ o $\mathbf{z} = \infty$, il confronto tra le traiettorie potrà essere effettuato attraverso grafici a due dimensioni.

Le proprietà del processo $W_z$ possono essere estese, con modifiche minime, al processo $W_r$ ottenuto effettuando le somme cumulate dei residui martingala rispetto al predittore lineare $\boldsymbol{\beta}'\mathbf{Z}_i$, anzichè rispetto al vettore delle variabili esplicative:

$$W_r(t,r) = \sum_{i=1}^{n} f(\mathbf{Z}_i) I(\boldsymbol{\beta}'\mathbf{Z}_i \le r) \hat{M}_i(t). \qquad (3.6)$$

Il processo $\hat{W}_r$, da utilizzare per approssimare la distribuzione di $W_r$, si ottiene semplicemente sostituendo nella (3.5) le funzioni indicatrici $I(\mathbf{Z}_i \le \mathbf{z})$ con le funzioni indicatrici $I(\boldsymbol{\beta}'\mathbf{Z}_i \le r)$. [63]

I risultati sulle somme cumulate dei residui martingala possono essere estesi anche ai modelli parametrici. In tal caso però i processi gaussiani utilizzati per

[62]Per la dimostrazione di questo risultato si veda l'Appendice 1 in Lin, Wei e Ying (1993).

[63]Per i risultati sulla convergenza dei processi $W_r$ e $\hat{W}_r$ si veda l'Appendice 2 in Lin, Wei e Ying (1993).

le simulazioni avranno forma diversa da quelli ricavati in ambito semiparametrico.

## 3.5 La scelta della forma funzionale delle covariata

Sia $X$ una possibile variabile esplicativa indipendente dal tempo. In generale, il termine in esponente nel modello di regressione a rischio moltiplicativo (3.1), anzichè essere lineare rispetto a $X$ potrebbe esserlo rispetto a una trasformazione di $X$, ad esempio $X^2$, ln ($X$) o I($X > c$). Nell'ambito del modello lineare classico, il problema della scelta della forma funzionale per una covariata esclusa dal modello può essere risolto per via grafica: il grafico dei residui OLS rispetto alla covariata oggetto d'interesse, qualora essa sia approssimativamente incorrelata con le altre variabili esplicative, rivela la trasformazione più appropriata per $X$ (par. 2.2.2). Therneau, Grambsch e Fleming (1990) dimostrano che un metodo analogo può essere utilizzato anche nell'ambito dei modelli per dati di durata.

Si supponga che il "vero" *hazard* integrato per un generico soggetto con covariate $\mathbf{Z}(t)$ e $X$ sia dato da

$$\Lambda(t) = \Lambda_{\circ}(t)\exp(\boldsymbol{\beta}'\mathbf{Z}(t))\exp(f(X)) = \Lambda_{\circ}(t)\exp(\boldsymbol{\beta}'\mathbf{Z}(t))h(X),$$

ma che si stimi il modello

$$\Lambda(t) = \Lambda_{\circ}(t)\exp(\boldsymbol{\beta}'\mathbf{Z}(t))$$

che non include tra le variabili esplicative $X$. Sia $\bar{h}(t,\mathbf{Z})$ la quantità

$$\bar{h}(t,\mathbf{Z}) = \frac{E(h(X)Y(t)|\mathbf{Z})}{E(Y(t)|\mathbf{Z})},$$

dove $Y$ rappresenta il processo indicatore dell'insieme di rischio per un generico individuo. Si noti che, se $X$ e $Y$ sono congiuntamente indipendenti da $\mathbf{Z}$ (o $\mathbf{Z}$ è nullo) $\bar{h}(t,\mathbf{Z}) = \bar{h}(t)$ è la media di $h(X)$ sull'insieme dei soggetti a rischio nell'istante t. Attraverso passaggi algebrici piuttosto complessi e considerazioni di tipo probabilistico, è possibile dimostrare che, per ogni $t$,

$$1-\frac{E\left(\hat{M}(t)\middle|X\right)}{E\left(N(t)\middle|X\right)}\approx\frac{\bar{h}}{h(X)},$$

dove

$$\bar{h}=\frac{E\left(\int_0^t \bar{h}(s,Z)\exp(\boldsymbol{\beta}'Z)Y(s)h(X)d\Lambda_\circ(s)\middle|X\right)}{E\left(\int_0^t \exp(\boldsymbol{\beta}'Z)Y(s)h(X)d\Lambda_\circ(s)\middle|X\right)},$$

o analogamente, ponendo $\bar{f}=\ln\bar{h}$, che

$$-\ln\left[1-\frac{E\left(\hat{M}(t)\middle|X\right)}{E\left(N(t)\middle|X\right)}\right]\approx f(X)-\bar{f}\ .^{64} \tag{3.7}$$

Da uno sviluppo in serie di Taylor del primo termine della (3.7) ($\ln(1-x)\approx -x$), segue che

$$E\left(\hat{M}(t)\middle|X\right)\approx E\left(N(t)\middle|X\right)\left(f(X)-\bar{f}\right).^{65} \tag{3.8}$$

Se la variazione di $E\left(N(t)\middle|X\right)$ rispetto a $X$ è trascurabile, come in analisi di sopravvivenza con una bassa percentuale di dati censurati, la (3.8) può essere scritta come

$$E\left(\hat{M}(t)\middle|X\right)\approx c\left(f(X)-\bar{f}\right), \tag{3.9}$$

dove $c$ è dato dal rapporto tra il numero totale di eventi verificatisi nel campione in $[0,t]$ e la numerosità del campione stesso.

---

[64] Per la dimostrazione si rimanda all'Appendice 1 in Therneau, Grambsch e Fleming (1990).

[65] Therneau, Grambsch e Fleming (1990) trovano che, qualora $N$ sia indipendente da $Y$, come nel caso di dati Poisson, un' approssimazione alternativa è data da

$$E\left(\hat{M}(t)\middle|X\right)\approx E\left(N(t)\middle|X\right)\left(h(X)-\bar{h}\right).$$

Sebbene in generale $\bar{h}$ possa variare in modo anche sensibile al variare di $X$, se $X$ e $\mathbf{Z}$ sono indipendenti, la variazione di $\bar{h}$ è trascurabile rispetto alla variazione di $h(X)$. Questa proprietà è particolarmente evidente quando le covariate non variano nel tempo e la censura è casuale. In questo caso infatti, se $X$ e $\mathbf{Z}$ sono indipendenti, $\bar{h}$ è completamente indipendente da $X$:

$$\bar{h} = \frac{\int_0^t \bar{h}(s)P(Y(s)=1)d\Lambda_\circ(s)}{\int_0^t P(Y(s)=1)d\Lambda_\circ(s)}.$$

Lo stesso vale ovviamente per $\bar{f}$ e $f(X)$. Facendo riferimento alla relazione (3.9), questo implica che l'andamento del valore atteso condizionato $E(\hat{M}(t)|X)$ rispetto a $X$ approssima la trasformazione corretta della variabile $X$ stessa. Quindi, se si effettua la stima di $E(\hat{M}(t)|X)$ lisciando, attraverso una qualsiasi tecnica di *smoothing*, il grafico dei residui martingala rispetto a $X$, si ottiene una curva che approssima la forma funzionale adeguata per tale covariata.[66] Analogamente, lo *smoothed plot* dei residui martingala rispetto alla trasformazione corretta di $X$ risulterà approssimativamente lineare. In particolare, qualora $X$ non fornisca informazioni aggiuntive rispetto a $\mathbf{Z}$, si otterrà una curva circa costante.

Il grafico lisciato dei residui martingala, oltre a essere facilmente interpretabile (i residui rappresentano infatti la differenza tra il numero di eventi osservati e il numero di eventi previsti dal modello), consente anche di capire quali siano gli individui che influenzano maggiormente la scelta della trasformazione.

Se l'incorrelazione tra $X$ e $\mathbf{Z}$ non è trascurabile, può essere opportuno utilizzare i residui martingala del modello non parametrico (Nelson-Aalen), così da ottenere un grafico "al lordo" di $\mathbf{Z}$ sia in ascissa che in ordinata; oppure,

[66]Una tecnica di lisciamento (*smoothing*) è un metodo per stimare in modo non parametrico la relazione che lega una variabile risposta con una o più variabili esplicative. In questo senso può essere considerata una tecnica di regressione non parametrica (Hardle, 1989; Hastie e Tibshirani, 1990).

estendendo in modo informale l'idea alla base dell'*added variable plot*,. si potrebbe individuare la trasformazione corretta per $X$ attraverso il grafico dei residui martingala del modello (3.1), rispetto ai residui del modello di regressione

$$X_i = \alpha \mathbf{Z}_i + u_i \quad i = 1,2,..,n.$$

stimato col metodo dei minimi quadrati.

L'utilizzazione dei residui martingala per la determinazione della forma funzionale della covariata ha trovato applicazione soprattutto su dati di tipo medico. In questo ambito usualmente l'interesse è rivolto alla valutazione dell'effetto prognostico di una o più variabili sulla sopravvivenza ed è rilevante individuare l'andamento della relazione sottostante (si veda per esempio Estey, 1997).

Il metodo per l'identificazione della forma funzionale di una covariata proposto da Therneau, Grambsch e Fleming (1990) possiede l'ottima proprietà di essere estremamente semplice. Tuttavia, alcuni autori (ad es. Lin, Wei e Ying, 1993) sostengono che esso ha il difetto di dipendere dalla particolare procedura utilizzata per lisciare il grafico dei residui.

Una tecnica diagnostica alternativa meno soggettiva, è basata su una particolare specificazione del processo $W_z$ definito nella (3.4). Mentre il grafico dei residui martingala serve in fase di costruzione del modello per scegliere l'opportuna trasformazione di una variabile non ancora inclusa nel predittore lineare, questo secondo metodo ha piuttosto l'obiettivo di verificare se una covariata sia stata inclusa nel modello di regressione nella forma funzionale appropriata.

Si indichi con $X$ la variabile d'interesse (indipendente dal tempo) e si supponga di stimare il modello di regressione in cui $X$ entri nel predittore lineare tramite una trasformazione $g(X)$. Considerando le somme cumulate dei residui martingala $\tilde{M}_i$ rispetto al valore di $g(X)$, si ottiene il processo $W_x = \{W_x(x) : x \in \mathbf{R}\}$ tale che

$$W_x(x) = \sum_{i=1}^{n} I(g(X) \le x)\tilde{M}_i .$$

$W_x$ rappresenta un caso particolare del processo multiparametrico $W_z$, per $f(.) = 1$, $t = \infty$ e $\mathbf{z}$ pari a un vettore le cui componenti siano tutte uguali a $\infty$ tranne quella corrispondente alla covariata $g(X)$. In accordo con i risultati generali

descritti nel paragrafo 3.4, se la forma funzionale di X è corretta e il modello non presenta altri errori di specificazione, la traiettoria del processo $W_x$ oscillerà attorno a 0 e non apparirà inusuale se confrontata con le simulazioni del corrispondente processo gaussiano $\hat{W}_x$. Qualora invece la scelta della funzione $g(.)$ non sia appropriata, la curva osservata si discosterà eccessivamente dallo 0 e quindi dalle curve simulate.[67]

Una statistica test naturale per la verifica della correttezza della forma funzionale $g(.)$ è

$$S_x = \sup_x \left| W_x(x) \right| .$$

La probabilità (*p-value*) a essa associata, $P(S_x \geq s)$, può essere approssimata da $P(\hat{S}_x \geq s)$, dove $\hat{S}_x = \sup_x \left| \hat{W}_x(x) \right|$. A sua volta, una stima di $P(\hat{S}_x \geq s)$ può essere ottenuta generando un grande numero di campioni casuali normali. Lin, Wei e Ying giustificano in modo formale l'utilizzo della statistica test $S_x$ per verificare l'ipotesi nulla di corretta forma funzionale, dimostrando che, sotto l'ipotesi alternativa di errata specificazione della funzione $g(.)$, se le componenti di $\hat{\beta}$ relative alle altre covariate del modello tendono al vero valore dei coefficienti, il $\sup_x \left| n^{-1} W_x(x) \right|$ converge quasi certamente a un valore maggiore di 0.[68]

## 3.6 La verifica dell'ipotesi di *link* esponenziale

La specificazione più comune del modello a rischio moltiplicativo prevede che la funzione *link* sia esponenziale, ma in alcune situazioni potrebbe essere opportuno scegliere una funzione di rischio relativo diversa (ad esempio $r(x) = 1 + x$).

Alcuni autori propongono di diagnosticare l'eventuale violazione dell'ipotesi di *link* esponenziale utilizzando il grafico lisciato dei residui martingala rispetto al valore stimato del predittore lineare: se la curva così ottenuta oscilla attorno a 0 l'ipotesi è valida (Valsecchi, Silvestri e Sisieni, 1996). Questo metodo è equivalente a quello adottato nell'ambito del modello di regressione classico per

[67] Ovviamente non ha senso definire il processo $W_x$ se $g(.)$ è una trasformazione dicotomica.

[68] Si veda l'Appendice 3 in Lin, Wei e Ying (1993).

diagnosticare l'eventuale violazione dell'ipotesi di linearità (si veda il par. 2.2.2).

Tuttavia, a nostro parere, uno curva che si discosti eccessivamente da 0 o che presenti un andamento sistematico rispetto al predittore, indica un'inadeguatezza generale del modello di regressione che non necessariamente è da imputarsi all'errata scelta della funzione di rischio relativo. Ad esempio, potrebbe derivare da un errore nella specificazione della forma funzionale di qualche covariata, essendo la violazione dell'ipotesi di *link* esponenziale e l'inadeguatezza della forma funzionale delle covariate due tipi di violazione alle assunzioni del modello strettamente legati tra loro.

Un approccio più formale alla verifica dell'ipotesi di *link* esponenziale è basato sulla somme cumulate dei residui martingala.[69] Per sottoporre a test l'ipotesi nulla si può utilizzare una particolare specificazione del processo $W_r$ definito nella (3.6) ($t = \infty$, $f(.) = 1$):

$$W_l(r) = \sum_{i=1}^{n} I(\boldsymbol{\beta}'\mathbf{Z}_i \le r)\tilde{M}_i \, .$$

Come nel caso della forma funzionale, la distribuzione di $W_r$ può essere approssimata attraverso delle simulazioni del corrispondente processo gaussiano $\hat{W}_r$ e una statistica test ragionevole per la corretta specificazione della funzione *link* è costituita dal $\sup_r |W_l(r)|$.[70]

## 3.7 Metodi diagnostici per la verifica dell'ipotesi PH (*proportional hazards*)

Come già abbiamo osservato nel paragrafo 1.9, se le covariate non variano nel tempo, il modello (3.1) esprime la proporzionalità dell'*hazard* tra gli individui della popolazione, ossia che, per ogni *i*, il rapporto tra $\Lambda_i(t)$ e il *baseline hazard* integrato è costante nel tempo. Tuttavia, soprattutto se la durata dello studio osservazionale è lunga, questa assunzione può essere violata. Ad esempio, l'effetto di una covariata sulla propensione dei soggetti a subire un dato evento potrebbe aumentare o diminuire col passare del tempo. In altri casi

[69] Lin,Wei e Ying (1993).

[70] Per una giustificazione formale dell'utilizzo di tale test si veda l'Appendice 3 in Lin, Wei e Ying (1993).

il *baseline hazard* potrebbe non essere lo stesso per tutti gli individui della popolazione.

In letteratura sono stati proposti molti metodi per la verifica dell'ipotesi PH.[71] A titolo esemplificativo ne descriveremo alcuni.

Il più banale di questi consiste nello stratificare il campione e nell'effettuare la stima non parametrica (Nelson-Aalen) dell'*hazard* integrato in ciascuno strato della popolazione. Se l'assunzione di proporzionalità è valida le curve stimate risulteranno approssimativamente parallele una volta che se ne sia effettuata la trasformazione logaritmica. Questo metodo teoricamente semplice ha il grave difetto di essere inaffidabile nel caso in cui il numero di individui in ciascuno strato sia piccolo e di conseguenza la variabilità delle stime sia alta.[72]

Un secondo metodo per diagnosticare eventuali scostamenti dalla proporzionalità è basato sulla stima di modelli che contengano appropriate covariate dipendenti dal tempo. Si supponga di voler verificare l'ipotesi PH relativamente a una covariata $X$, che per semplicità assumeremo essere l'unica inclusa nel modello. A tal fine si può aggiungere nel predittore lineare la variabile esplicativa $X(t) = Xg(t)$, ad esempio $X(t) = X\ln(t)$ o $X(t) = Xt$, ottenendo il modello di regressione ampliato

$$\Lambda_i(t) = \exp(\boldsymbol{\beta}X + \gamma\ Xg(t))\Lambda_0(t).$$

Sottoporre a test l'ipotesi PH equivale in questo caso a sottoporre a test l'ipotesi nulla che il coefficiente $\gamma$ relativo a $X(t)$ sia uguale a 0, contro l'ipotesi alternativa che $\gamma$ sia diverso da 0. Purtroppo i risultati di questa tecnica diagnostica dipendono strettamente dalla scelta della funzione $g(t)$.

Alcuni autori propongono di suddividere l'asse temporale in $h$ intervalli e di stimare il modello di regressione separatamente in ognuno di essi, assumendo che l'ipotesi PH sia valida solo localmente. Si ottengono così $h$ stime dei parametri del modello ciascuna relativa a un periodo di tempo diverso. Se

[71] Per una breve rassegna di tali metodi si veda Harrel (1997).

[72] In ambito parametrico il confronto delle stime di Nelson-Aalen fornisce informazioni anche riguardo alla specificazione del *baseline hazard*. Ad esempio, per sottoporre a test l'ipotesi nulla che il modello sia Weibull, si possono confrontare i logaritmi delle stime non parametriche del rischio integrato, effettuate in ogni strato della popolazione, in un grafico con ascissa ln$t$. Se le curve così ottenute appaiono approssimativamente lineari e parallele, l'ipotesi nulla è valida. Curve lineari ma non parallele indicano invece la violazione della sola assunzione PH e, in particolare, che i parametri della Weibull assumono valori diversi da strato a strato. Infine, se si ottengono curve parallele, ma non lineari, l'ipotesi di proporzionalità può essere ritenuta valida, ma il modello Weibull non è adeguato (Harrel, 1997). Per un contributo recente al problema della specificazione del *baseline hazard* si veda per esempio Pena (1998).

l'ipotesi di proporzionalità è valida tali stime non presenteranno un andamento sistematico nel tempo e oscilleranno attorno alla stima dei coefficienti effettuata sull'intero asse temporale.

Nel caso in cui vi sia evidenza che l'assunzione di *hazard proporzionali* relativa a una covariata $X$ non è appropriata, il modo più semplice per tener conto di questa violazione consiste nella categorizzazione della variabile $X$ e nella specificazione di un modello in cui il *baseline hazard* è diverso in ogni strato:

$$\Lambda_i(t) = \Lambda_{\circ j}(t)\exp(\boldsymbol{\beta}'\mathbf{Z}_i) \quad j=1,2,...,h \quad i=1,2,...,n_j . \tag{3.10}$$

Un modello di questo tipo è detto modello stratificato.[73] A questo proposito esistono test per la proporzionalità basati sul confronto tra le stime dei coefficienti del modello non stratificato e del modello stratificato (3.10).

Recentemente sono stati proposti metodi più flessibili che utilizzano funzioni non parametriche, per esempio *spline* cubiche, per modellare l'effetto tempo-dipendente delle covariate (Quantin *et al.*, 1999). Sargent (1997) ha proposto invece un approccio al problema della violazione dell'ipotesi PH basato sulla specificazione bayesiana di un modello lineare dinamico.

Nei prossimi paragrafi vengono presentate alcune interessanti tecniche diagnostiche per la verifica dell'ipotesi PH basate sui residui e più in particolare sul processo *score* stimato.[74]

### 3.7.1 L'utilizzo dei residui di Schoenfeld

L'analisi dei residui riveste un ruolo molto importante anche nell'ambito della verifica dell'assunzione PH. In particolare, per quanto concerne il modello semiparametrico di Cox, l'attenzione è rivolta ai residui parziali di Schoenfeld

$$\hat{\mathbf{r}}_i = \mathbf{Z}_i(t_i) - E(\hat{\boldsymbol{\beta}}, t_i).$$

[73]Il procedimento di stima del modello stratificato prevede la massimizzazione di una funzione di verosimiglianze (o verosimiglianza parziale) opportunamente modificata per tenere conto della presenza di strati nella popolazione, e di due differenti *baseline hazard*. Si veda Fleming e Harrington (1991), pag.144.

[74]Per alcune interessanti applicazioni dei metodi esposti in questo lavoro si veda Valsecchi, Silvestri e Sisieni (1996) e Hilsenbeck *et al.* (1998). Per un confronto dei vari approcci per la verifica dell'ipotesi PH si veda Ng'andu (1997).

Il fatto che i residui di Schoenfeld siano definiti su istanti piuttosto che su individui li rende adatti a verificare un'ipotesi che coinvolge il concetto di tempo, come quella di proporzionalità.

Un semplice esempio mostra chiaramente quali siano le informazioni contenute nei residui parziali. Per semplicità assumiamo di effettuare la stima di un modello di regressione che includa nel predittore lineare una sola covariata dicotomica $X$ indipendente dal tempo che assuma valore 1 sui soggetti appartenenti a un insieme A e 0 altrimenti. Il rapporto tra l'*hazard function* dei soggetti tali che $X$=1 e quelli tali che $X$=0 è stimato dalla quantità

$$\frac{\hat{\lambda}\left(t|X=1\right)}{\hat{\lambda}\left(t|X=0\right)} = \exp\hat{\boldsymbol{\beta}}. \tag{3.11}$$

Supponiamo però che il "vero" rapporto di rischio abbia andamento monotono decrescente rispetto al tempo, quindi esista un istante $t^*$ tale che:

$$\text{se } t \in \left[0,t^*\right] \ \frac{\lambda\left(t|Z=1\right)}{\lambda\left(t|Z=0\right)} \geq \exp\hat{\boldsymbol{\beta}}$$

$$\text{se } t \in \left(t^*,\infty\right) \ \frac{\lambda\left(t|Z=1\right)}{\lambda\left(t|Z=0\right)} < \exp\hat{\boldsymbol{\beta}}$$

In $[0,t^*]$ il modello sottostima il rischio per i soggetti appartenenti all'insieme A, di conseguenza $E\left(\hat{\boldsymbol{\beta}},t\right) \leq E\left(\boldsymbol{\beta},t\right)$ e i residui parziali tenderanno a essere positivi. Viceversa, in $(t^*,\infty)$ i residui tenderanno a essere negativi. Dunque, in generale, in caso di violazione dell'ipotesi PH, il grafico dei residui di Schoenfeld rispetto al tempo, sarà caratterizzato da un eccesso di residui positivi o negativi in qualche regione dell'asse dei tempi, anche se complessivamente i due eccessi si compenseranno lungo l'intero periodo d'osservazione, essendo la somma dei residui di Schoenfeld nulla a prescindere dalla correttezza o meno del modello. [75]

[75]Per un'applicazione basata sul grafico dei residui parziali si veda Schoenfeld (1982).

### 3.7.2 La somma cumulata dei residui di Schoenfeld

Therneau, Grambsch e Fleming (1990) propongono di verificare l'assunzione PH utilizzando, anzichè i semplici residui parziali, la loro somma cumulata rispetto al tempo. Tale somma cumulata costituisce il processo *score* valutato in $\hat{\boldsymbol{\beta}}$, ossia la somma dei processi individuali $\mathbf{S}_i\left(\hat{\boldsymbol{\beta}},t\right)$ (par. 2.7).

E' interessante osservare che se l'unica covariata del modello è un fattore a due livelli (covariata dicotomica), la somma cumulata dei residui di Schoenfeld equivale alla somma cumulata dei residui martingala sullo strato della popolazione in cui la covariata assume valore 1. Infatti, supponendo per semplicità che Z sia l'unica variabile esplicativa inclusa nel modello, indicando con $\hat{\Lambda}_{\circ}$ lo stimatore di Breslow, si trova che:

$$\sum_{i=1}^{n}\int_0^t\left(Z_i - E\left(\hat{\beta},s\right)\right)dN_i(s) = \sum_{i=1}^{n}\int_0^t Z_i dN_i(t) - \int_0^t E\left(\hat{\beta},s\right)d\overline{N}(s) =$$

$$= \sum_{i=!}^{n}\int_0^t Z_i dN_i(s) - \sum_{i=1}^{n}\int_0^t Z_i \exp\left(\hat{\beta}Z_i\right)d\hat{\Lambda}_{\circ}(s) = \sum_{i=!}^{n}\int_0^t Z_i d\hat{M}_i(s) = \sum_{i:Z_i=1}\hat{M}_i(t)$$

(Valsecchi, Silvestri e Sisieni, 1996). Con analogo ragionamento si può verificare che la somma cumulata dei residui parziali corrisponde a una somma di residui martingala anche se le covariate dicotomiche sono più di una.

Si supponga che, come nell'esempio utilizzato nel paragrafo precedente, il modello stimato preveda la proporzionalità dei rischi, ma che il modello corretto sia caratterizzato da un andamento monotono decrescente del rapporto $\lambda\left(t|Z=1\right)/\lambda\left(t|Z=0\right)$. La traiettoria del processo $\mathbf{U}\left(\hat{\boldsymbol{\beta}}\right)$ tenderà ad allontanarsi dallo 0 fino a raggiungere elevati valori positivi, per poi tornare nuovamente a 0.[76] Affinchè queste considerazioni di tipo euristico possano essere espresse in modo formale, è necessario disporre di informazioni sulla distribuzione del processo *score* stimato, perlomeno di tipo asintotico.

Dai risultati sul comportamento asintotico del processo *score* valutato in corrispondenza del vero valore dei parametri (par. 1.9.2), è possibile ricavare, sotto opportune ipotesi, la distribuzione asintotica del processo $\mathbf{U}(\hat{\boldsymbol{\beta}},t)$. A tal

[76]Ricordiamo che, per definizione, il processo *score* parte da 0 e torna a 0.

fine si indichi con **v** il limite in probabilità della matrice **V** di covarianza empirica pesata definita nel capitolo 1. Therneau, Grambsch e Fleming (1990), generalizzando un precedente risultato di Wei (1984), valido per modelli con un'unica covariata dicotomica, dimostrano che, sotto opportune condizioni di regolarità,[77] se, per qualche $j \in \{1,2,...k\}$, $v_{jk}$ è proporzionale a $v_{jj}$ per ogni $k \neq j$ e per ogni $t$, allora la componente del processo *score* stimato relativa alla $j$-esima covariata opportunamente standardizzata,

$$\left[\mathbf{I}^{-1}\left(\hat{\boldsymbol{\beta}},\infty\right)\right]_{jj}^{1/2} U_j\left(\hat{\boldsymbol{\beta}},t\right),$$

all'aumentare della numerosità campionaria converge in distribuzione a un Ponte Browniano standard $W^0$.[78]

Sulla base di questo risultato una semplice statistica test per la verifica dell'ipotesi PH è data da

$$\sup_t \left[\mathbf{I}^{-1}\left(\hat{\boldsymbol{\beta}},t\right)_{jj}\right]^{-1/2} \left|U_j\left(\hat{\boldsymbol{\beta}},t\right)\right|, \tag{3.12}$$

la cui distribuzione può essere approssimata da quella ben nota di $\sup_t \left|W^0(t)\right|$.[79] Come si può intuire dall'esempio del paragrafo precedente, questa statistica test è particolarmente sensibile nel caso in cui l'andamento del rapporto (3.11) sia strettamente monotono rispetto al tempo.[80]

E' tuttavia necessario osservare che l'ipotesi $v_{jk}/v_{jj}$ = *costante* è piuttosto forte e costituisce un limite per l'applicazione del test del sup. In particolare, poichè **V** costituisce la matrice di covarianza empirica pesata tra le variabili incluse nel modello, tale ipotesi è soddisfatta se, in ogni istante di tempo, la $j$-esima covariata è incorrelata con le altre, come nel caso in cui rappresenti un trattamento assegnato casualmente e non esistano interazioni forti tra il trattamento stesso e gli altri fattori esplicativi.

Qualora l'ipotesi di incorrelazione non sia valida, è comunque possibile approssimare la distribuzione del processo *score* stimato attraverso opportune simulazioni.[81] Infatti, se si utilizza lo stimatore di Breslow per il *baseline hazard* integrato e i regressori non dipendono dal

[77] Andersen e Gill (1982).
[78] Per la dimostrazione di questo teorema si veda Fleming e Harrington (1991) o Therneau, Grambsch e Fleming (1990).
[79] Si vedano le relative tavole di probabilità in Koziol e Byar (1975).
[80] Fleming e Harrington (1991), pagg. 175-177.
[81] Lin, Wei e Ying (1993).

tempo, $\mathbf{U}(\hat{\boldsymbol{\beta}},t)=\sum_{i=1}^{n}\mathbf{Z}_i\hat{M}_i(t)$ e, di conseguenza, $\{\mathbf{U}(\boldsymbol{\beta},t):t\geq 0\}$ rappresenta un caso particolare del processo $W_z$ definito nella (3.4), in cui si sia posto $f(x)=x$, $\mathbf{z}=\infty$. Quindi la verifica dell'assunzione PH può avvenire attraverso il confronto grafico della traiettoria osservata con le traiettorie simulate del corrispondente processo gaussiano $\hat{W}_z$. Poichè i processi $W_z$ e $\hat{W}_z$ hanno *k* componenti si può tracciare un grafico distinto per ognuna delle *k* covariate del modello.

Mediante simulazione, è inoltre possibile approssimare la distribuzione della statistica test (3.12).[82]

### 3.7.3 L'utilizzo dei residui di Schoenfeld pesati

Si indichi con $r_i$ il residuo parziale teorico relativo all'istante $t_i$ nel modello di regressione a rischio moltiplicativo (3.1). Supponiamo però che il "vero" modello non soddisfi l'assunzione PH e che sia caratterizzato da coefficienti di regressione dipendenti dal tempo:

$$\boldsymbol{\beta}(t)=\boldsymbol{\beta}+\mathbf{H}(t)\boldsymbol{\theta}\,,^{83} \qquad (3.13)$$

dove $\mathbf{H}(t)$ è una matrice diagonale $k\times k$ i cui generici elementi $(i,i)$ sono funzioni del tempo note che, per motivi di identificabilità assumeremo variare attorno a 0. Si indichi con $\overline{\mathbf{Z}}(\boldsymbol{\beta}(t),t)$ la media pesata

$$E(\boldsymbol{\beta}(t),t)=\frac{\sum_{i=1}^{n}Y_i(t)\mathbf{Z}_i\exp(\boldsymbol{\beta}(t)'\mathbf{Z}_i)}{\sum_{i=1}^{n}Y_i(t)\exp(\boldsymbol{\beta}(t)'\mathbf{Z}_i)}\,.$$

Il residuo parziale teorico relativo al modello (3.1) può essere scomposto nella somma di due termini:

[82] Lin, Wei e Ying (1993).

[83] Il vero modello può essere interpretato come un modello in cui le covariate dipendono dal tempo o analogamente come un modello in cui i coefficienti di regressione dipendono dal tempo.

$$\mathbf{r}_i = \left[\mathbf{Z}_i - E\left(\boldsymbol{\beta}(t_i), t_i\right)\right] + \left[E\left(\boldsymbol{\beta}(t_i), t_i\right) - E\left(\boldsymbol{\beta}, t_i\right)\right],$$

il primo dei quali, in quanto residuo parziale teorico relativo al modello correttamente specificato, ha media nulla.[84]

Come è semplice verificare, dall'espansione in serie di Taylor di $E\left(\boldsymbol{\beta}(t), t\right)$ attorno a $\boldsymbol{\beta}(t_i) = \boldsymbol{\beta}$ si ottiene la seguente approssimazione:

$$E\left(\boldsymbol{\beta}(t_i), t_i\right) \approx E\left(\boldsymbol{\beta}, t_i\right) + \mathbf{V}\left(\boldsymbol{\beta}, t_i\right)\mathbf{H}(t_i)\boldsymbol{\theta} \ ,$$

dove $\mathbf{V}\left(\boldsymbol{\beta}, t_i\right)$ è la matrice definita nel paragrafo 1.9.1 valutata in $t_i$. Quindi

$$E\left(\mathbf{r}_i\right) \approx \mathbf{V}\left(\boldsymbol{\beta}, t_i\right)\mathbf{H}(t_i)\boldsymbol{\theta}$$

e il residuo scalato $\mathbf{r}_i^* = \mathbf{V}^{-1}\left(\boldsymbol{\beta}, t_i\right)\mathbf{r}_i$ è tale che

$$E\left(\mathbf{r}_i^*\right) \approx \mathbf{H}(t_i)\boldsymbol{\theta}$$

Questa relazione, unitamente alla (3.13), suggerisce che lo *smoothed plot* rispetto al tempo della *j*-esima componente del vettore

$$\hat{\boldsymbol{\beta}} + \mathbf{V}^{-1}\left(\hat{\boldsymbol{\beta}}, t_i\right)\hat{\mathbf{r}}_i \qquad (3.14)$$

approssima la dipendenza del *j*-esimo parametro dal tempo.[85] Le componenti del vettore (3.14) sono chiamate residui di Schoenfeld pesati. In particolare l'assenza di andamento sistematico, indicherà la validità dell'ipotesi PH.[86]

Dal punto di vista computazionale, il calcolo di $\mathbf{V}^{-1}\left(\hat{\boldsymbol{\beta}}, t_i\right)$ in ciascun tempo di morte è piuttosto oneroso, tuttavia, nella maggior parte dei casi, la matrice $\mathbf{V}$ varia lentamente e può quindi essere sostituita da $\overline{\mathbf{V}} = \mathbf{I}\left(\hat{\boldsymbol{\beta}}\right)/k$. Utilizzando la matrice d'informazione osservata per la standardizzazione dei residui si evita anche il problema della singolarità di $\mathbf{V}\left(\hat{\boldsymbol{\beta}}, t_i\right)$, qualora, in corrispondenza dei

[84] Si veda il par. 2.7.
[85] Grambsch e Therneau (1994).
[88] Per un'analisi delle proprietà dei residui pesati, si vedano gli esempi in Grambsch e Therneau (1994).

tempi più lunghi, il numero dei soggetti a rischio diventi minore del numero delle covariate.[87]

### 3.7.4 La verifica dell'ipotesi PH nel caso parametrico

I risultati relativi alla distribuzione asintotica del processo *score* enunciati nel paragrafo (3.7.2) restano sostanzialmente validi anche se il modello di regressione è parametrico. Dunque, se le covariate sono incorrelate, è possibile utilizzare la statistica test (3.12), tenendo conto però che in tale contesto il processo *score* e la matrice d'informazione assumono una forma diversa da quella che li caratterizza in ambito semiparametrico. Sono applicabili anche in questo caso le tecniche diagnostiche basate sulla somma cumulata di residui martingala.

## 3.8 Un test per l'eterogeneità non osservabile

Una delle ipotesi alla base del modello di regressione (3.1) è che i dati siano omogenei condizionatamente al vettore dei regressori **Z**. Qualora il rischio sia significativamente influenzato anche da qualche fattore non osservabile, tale assunzione risulta violata e il modello mal specificato.[88] In ambito econometrico si parla in questo caso di presenza di eterogeneità non osservabile (si veda per esempio Lancaster, 1985), mentre in ambito biometrico si utilizza più frequentemente il termine *frailty*, facendo esplicito riferimento alla "fragilità" dell'individuo, ossia la sua propensione non osservabile a subire l'evento d'interesse, ad esempio una malattia.

Un modo per tenere conto dell'eventuale presenza di eterogeneità consiste nello stimare il modello

$$\Lambda_i(t) = \Lambda_\circ(t)\exp(\alpha_i + \boldsymbol{\beta}'\mathbf{Z}_i(t)) = \Lambda_\circ(t)\nu_i \exp(\boldsymbol{\beta}'\mathbf{Z}_i(t)), \qquad (3.15)$$

[87]Per ulteriori commenti a questa sostituzione e relativi studi di simulazione, si veda Grambsch e Therneau (1994).

[88]Sugli effetti di questo tipo di violazione si veda Lancaster (1985, 1990) e Henderson e Oman (1999).

dove $\nu_i$ è una variabile aleatoria non osservabile che rappresenta la componente di eterogeneità del soggetto *i*.[89] Usualmente si assume che la distribuzione dei $\nu_i$ sia tale che, per ogni *i*,

$$E(\nu_i) = 1$$

$$E(\nu_i - 1)^2 = \sigma^2$$

$$E(\nu_i \nu_j) = 0 \quad i \neq j \, .$$

Ad esempio le componenti di eterogeneità potrebbero essere un campione casuale estratto da una distribuzione Gamma con media unitaria.

Una particolare specificazione del modello (3.15) frequentemente utilizzata in econometria è rappresentata dal modello Burr,

$$\lambda_i(t;\boldsymbol{\beta},\alpha,\sigma^2) = \frac{\exp(\boldsymbol{\beta}'\mathbf{Z}_i(t))\alpha t^{\alpha-1}}{1+\sigma^2 \exp(\boldsymbol{\beta}'\mathbf{Z}_i(t))t^{\alpha}} \qquad \alpha \geq 0, \sigma^2 \geq 0 \, .$$

Il modello Burr può essere interpretato sia come un modello Weibull in cui sia stata introdotta una componente di eterogeneità Gamma, sia, come un modello ad *hazard* non proporzionali.[90] In questo caso si può verificare la presenza di eterogeneità non osservabile (o equivalentemente) la violazione dell'assunzione PH, sottoponendo a test l'ipotesi nulla $\sigma = 0$ (Lancaster, 1990).

Anche nell'ambito del modello di Cox usualmente la distribuzione della componente di *frailty* viene specificata in forma parametrica (si vedano per esempio Nielsen *et al*., 1992; Clayton, 1991), tuttavia di recente Horowitz (1999) ha proposto un metodo di stima per un modello in cui sia la funzione di rischio di base che la distribuzione della componente di eterogeneità siano in forma non parametrica.

Tornando al caso generale, il modello (3.15) può essere utilizzato per verificare la correttezza del modello senza *frailty*; tuttavia, poichè le procedura

[89]Si noti che il modello (3.15) coincide con quello di cui ci siamo serviti nel paragrafo 2.4 per definire i residui secondo l'approccio di Lancaster (1990), tranne per il fatto che, in quel caso, gli $\alpha_i$ erano visti come parametri e non come elementi stocastici.

[90]Nel caso in cui le covariate siano indipendenti dal tempo, il modello Burr può essere interpretato come un particolare modello parametrico a odds proporzionali (*Proportional Odds models*). Si veda Marubini e Valsecchi (1995), cap. 8.

di stima sono spesso piuttosto complesse, è preferibile individuare l'eventuale presenza di eterogeneità attraverso un opportuno test.

In questo paragrafo esamineremo il test *score* proposto da Lancaster (1985 e 1990) in ambito econometrico,[91] prendendo in considerazione, per semplicità, solo la situazione in cui i dati siano relativi a un'analisi di sopravvivenza con censura a destra e le covariate siano indipendenti dal tempo. In accordo con l'interpretazione dei residui martingala come indicatori dell'eterogeneità non osservabile (par. 2.4.1), vedremo che tale test dipende proprio dagli $\widetilde{M}_i$.

Consideriamo prima il caso in cui il *baseline hazard* nel modello (3.15) sia specificato parametricamente.[92] Si indichi con $L^*$ la verosimiglianza condizionata alle componenti di eterogeneità. Il contributo individuale del soggetto *i* a tale verosimiglianza è

$$L_i^* = \left[\lambda_\circ(t;\boldsymbol{\theta})\nu_i \exp(\boldsymbol{\beta}'\mathbf{Z}_i)\right]^{\delta_i} \exp\left(-\Lambda_\circ(s;\boldsymbol{\theta})\nu_i \exp(\boldsymbol{\beta}'\mathbf{Z}_i)\right). \qquad (3.16)$$

Effettuando l'espansione di Taylor fino al secondo ordine della (3.16) intorno al punto $\nu_i$ =1 e facendone il valore atteso rispetto a $\nu_i$ stesso, si trova

$$E(L_i^*) \approx \left[\lambda_\circ(t_i;\boldsymbol{\theta})\exp(\boldsymbol{\beta}'\mathbf{Z}_i)\right]^{\delta_i} \exp\left(-\Lambda_\circ(t_i;\boldsymbol{\theta})\exp(\boldsymbol{\beta}'\mathbf{Z}_i)\right)\times$$

$$\times\left[1-2\delta_i\Lambda_\circ(t_i;\boldsymbol{\theta})\exp(\boldsymbol{\beta}'\mathbf{Z}_i)\frac{\sigma^2}{2}+\left(\Lambda_\circ(t_i;\boldsymbol{\theta})\exp(\boldsymbol{\beta}'\mathbf{Z}_i)\right)^2\frac{\sigma^2}{2}\right].$$

$E(L_i^*)$ è un'approssimazione locale della verosimiglianza individuale non condizionata relativa al modello (3.15), in cui il fattore di fragilità compare solo attraverso la sua varianza $\sigma^2$.[93]

Per diagnosticare la violazione dell'assunzione di omogeneità, si può costruire il test *score* per l'ipotesi nulla $H_\circ : \sigma^2 = 0$ (la distribuzione che genera i $\nu_i$ è degenere). A tal scopo, è necessario calcolare la derivata parziale di $E(L_i^*) \approx L_i$ rispetto a $\sigma^2$, in $\sigma^2 = 0$:

[91] Un test analogo a quello qui presentato può essere ottenuto per la presenza di eterogeneità dovuta a covariate omesse; Orme (1998) analizza il comportamento del test attraverso simulazioni.

[92] Il procedimento che seguiremo per ricavare il test è stato applicato nell'ambito più complesso dei dati di durata a destinazioni multiple da Mealli (1994).

[93] A differenza dell'approssimazione qui introdotta, la verosimiglianza non condizionata dipende dalla particolare distribuzione del fattore di eterogeneità.

$$\left[\frac{\partial L_i}{\partial \sigma^2}\right]_{\sigma^2=0} = \frac{1}{2}\left[-2\delta_i \Lambda_0(t_i;\boldsymbol{\theta})\exp(\boldsymbol{\beta}'\mathbf{Z}_i) + (\Lambda_0(t_i;\boldsymbol{\theta})\exp(\boldsymbol{\beta}'\mathbf{Z}_i))^2\right] =$$

$$= \frac{1}{2}\left[(\delta_i - \Lambda_0(t_i;\boldsymbol{\theta})\exp(\boldsymbol{\beta}'\mathbf{Z}_i))^2 - \delta_i\right] = \frac{1}{2}\left[M_i(\infty)^2 - \delta_i\right].$$

Tale quantità, valutata in corrispondenza delle stime di massima verosimiglianza vincolata dei parametri $\beta$ e $\boldsymbol{\theta}$, costituisce la base per il calcolo della statistica test *score*:

$$\frac{1}{4}\sum_{i=1}^{n}\left(\tilde{M}_i^2 - \delta_i\right)\left[\boldsymbol{I}^{-1}\left(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}}\right)\right]_{\sigma^2}\sum_{i=1}^{n}\left(\tilde{M}_i^2 - \delta_i\right), \tag{3.17}$$

dove $\left[\mathbf{I}^{-1}\left(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}}\right)\right]_{\sigma^2}$ stima la porzione dell'inversa della matrice d'informazione osservata relativa a $\sigma^2$. In assenza di eterogeneità la (3.17) è distribuita come un $\chi^2_{(1)}$.

E' interessante osservare che, in assenza di censura,

$$\sum_{i=1}^{n}\left(\tilde{M}_i^2 - \delta_i\right) = n\left[n^{-1}\sum_{i=1}^{n}\tilde{M}_i^2 - 1\right],$$

quindi, attraverso il test *score* viene confrontata la varianza empirica dei residui martingala con la varianza dei residui teorici sotto l'ipotesi nulla, che è in questo caso unitaria, poichè

$$\operatorname{var} M_i = \operatorname{var}(1 - \Lambda_i(t_i)) = \operatorname{var}(\Lambda_i(t_i)) = 1.$$

Sulla base di questa considerazione si può derivare in modo semplice la statistica test *score* anche in ambito semiparametrico, in particolare nel caso in cui la stima del *baseline hazard* sia effettuata con lo stimatore di Breslow. Infatti in tale situazione il residuo martingala teorico può essere scritto come:

$$M_i = \delta_i - \exp(\boldsymbol{\beta}'\mathbf{Z}_i)\sum_{j:t_j \le t_{ii}}\frac{\delta_j}{\sum_{k\in R_j}\exp(\boldsymbol{\beta}'\mathbf{Z}_k)}$$

ed è possibile dimostrare che la varianza di tale residuo è uguale al valore atteso della quantità:

$$H_i = \delta_i - \exp 2\left(\boldsymbol{\beta}'\mathbf{Z}_i\right) \sum_{j:t_j \le t_i} \frac{\delta_j}{\left[\sum_{k \in R_j} \exp\left(\boldsymbol{\beta}'\mathbf{Z}_k\right)\right]^2}.^{94} \qquad (3.18)$$

Dal confronto tra la varianza empirica e la stima della varianza teorica, ottenuta inserendo nella (3.18) $\hat{\boldsymbol{\beta}}$ ed effettuando la media delle quantità così ottenute, si ricava la statistica

$$\sum_{i=!}^{n} \left(\tilde{M}_i^2 - \hat{H}_i\right). \qquad (3.19)$$

Lancaster dimostra che la (3.19) coincide con la statistica *score* valutata in $\sigma^2 = 0$. Di conseguenza, è possibile anche in questo caso applicare i risultati standard relativi al test *score*.

[94]Lancaster (1990), par. 11.3.

# Bibliografia


AALEN, O. (1978). Nonparametric estimation of partial transition probabilities in multiple decrement models. *The Annals of Statistics*, 6, 534–545.

AALEN, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6, 701-726.

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10, 1100-1120.

ANDERSEN, P. K., BORGAN, O., GILL, R. D. and KEIDING, N. (1993). *Statistical models based on counting processes*. Springer-Verlag, New York.

ANTONIADIS, A., GREGOIRE, G. and NASON, G. (1999). Density and hazard rate estimation for right-censored data by using wavelet methods. *J. R. Statist. Soc.* B, 61, 63-84.

BALTAZAR-ABAN, I. and PEÑA, E. A. (1995). Properties of hazard-based residuals and implications in model diagnostics. *J. Am. Statist. Assoc*., 90, 185–197.

BARLOW, W. E. and PRENTICE R.L. (1988). Residuals for relative risk regression. *Biometrika,* 75, 65–74.

BRESLOW, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 30, 89–99.

BRESLOW, N. and CROWLEY, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, 2, 437–453.

CLAYTON, D. G. (1991). A Monte Carlo method for Bayesian Inference in Frailty Models. *Biometrics*, 47, 467-485.

COOK, R. D. and WEISBERG, S. (1982). *Residuals and influence in regression*. Chapman & Hall, New York.

COX, D. R. (1972). Regression models and life-tables. *J. R. Statist. Soc.* B, 34, 187–220.

COX, D. R. (1975). Partial likelihood. *Biometrika*, 62, 269–276.

COX, D. R. and OAKES, D. (1984). *Analysis of survival data*. Chapman and Hall, London.

COX, D. R. and SNELL, E. J. (1968). A general definition of residuals. *J. R. Statist. Soc.* B, 30, 248-275

CROWLEY, J. And HU, M. (1977). Covariance analysis of heart transplant survival data. *J. Am. Statist. Assoc*., 72, 27–36.

ESTEY, E. (1997). Prognostic factor in clinical cancer trials. *Clin. Cancer Res.*, 3, 12, 2591-2593.

FAHRMEIR, L. and KLINGER, A. (1998). A non parametric multiplicative hazard model for event history analysis. *Biometrika*, 85, 581-592.

FLEMING, T. R. and HARRINGTON, D. P. (1991). *Counting processes and survival analysis*. Wiley, New York.

FOX, J. (1984). *Linear statistical models and related methods with applications to social research*. Wiley, New York.

GLIDDEN, D. V. (1999). Checking the adequacy of the Gamma frailty model for multivariate failure time, *Biometrika*, 86, 381-393.

GOURIEROUX, C., MONFORT, A., RENAULT, E. and TROGNON, A. (1987). Generalised residuals. *Journal of Econometrics*., 34, 5–32.

GRAMBSCH, P. M. and THERNEAU, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81, 515–526

HARREL, F. E. Jr. (1997). *Predicting outcomes: applied survival analysis and logistic regression.* Division of Biostatistics and Epidemiology, Department of Health Evaluation Sciences, School of Medicine, University of Virginia.

HARDLE, W. (1989). *Applied nonparametric regression*. Cambridge University Press, Cambridge.

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized additive models*. Chapman and Hall, London.

HENDERSON, R. and OMAN, P. (1999). Effect of frailty on marginal regression estimates in survival analysis, *J. R. Statist. Soc.* B, 61, 367-379.

HILSENBECK, S. G., RAVDIN, P. M., DE MOOR *et al* (1998). Time – dependence of hazard ratios for prognostic factors in prymary breast cancer. *Breast Cancer Res Treat.,* 52, 1-3, 227-237.

HOROWITZ, J. L. (1999). Semiparametric Estimation of a Proportional Hazard Model with Unobserved Heterogeneity. *Econometrica*, 67, 1001-1028.

JOHNSTON, J. (1991). *Econometric methods*. Terza edizione. McGraw-Hill, Auckland.

KALBFLEISCH, J. D. and PRENTICE, R. L. (1980). *The statistical analysis of failure time data*. Wiley, New York.

KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Statist. Assoc*., 53, 457-481

LANCASTER, T. (1985). Generalised residuals and heterogeneous duration models. *Journal of Econometrics*, 28, 155-169.

LANCASTER, T. (1990). *The econometric analysis of transition data.* Cambridge University Press, Cambridge.

LAKAGOS, S. W. (1988). The loss in efficiency from misspecifying covariates in proportional hazards regression models. *Biometrika*, 75, 156–160.

LAM, K. F. (1998), A class of tests for the equality of k cause-specific hazard rates in a competing risks model, *Biometrika*, 85, 179-188.

LIN, D. Y. And WEI, L. J. (1989). The robust inference for the Cox proportional hazards model. *J. Am. Statist. Assoc*., 84, 1074–1078.

LIN, D. Y., WEI, L. J. and YING, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*., 80, 557–572.

MARUBINI, E. e VALSECCHI, M. G. (1995*). Analyzing Survival Data from Clinical Trials and Observational Studies*. Wiley, West Sussex.

McCULLAGH, P. and NELDER, J. A. (1989). *Generalized linear models*. Seconda edizione. Chapman & Hall, London.

MEALLI, F. (1994). *La specificazione di modelli di durata a stati ed episodi multipli*. Quaderni del Dipartimento Statistico, Università di Firenze.

NADEAU, C. and LAWLESS, J. F. (1998). Inference for means and covariances of point processes through estimating functions. *Biometrika*, 85, 893-906.

NARDI e SCHEMPER (1999). New residuals for Cox regression and their application to outliers screening. *Biometrics*, **55**, 523-529.

NG'ANDU, N. H. (1997). An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. Statistics in Medicine, 16, 611-626.

NIELSEN, G. G., GILL, R. D., ANDERSEN, P. K. and SØRENSEN, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scand j. Statist.*, 19, 25–43.

ORME, C.D. (1998). On the insensitivity of the score test for heterogeneity to omitted covariates in multivariate failure time models. *Biometrika*, 85, 457-461.

PENA, E. A. (1998). Smooth Goodness of Fit Tests for the Baseline Hazard in Cox's Proportional Hazard Model, *J. Am. Statist. Assoc*., 93, 673-692.

PRENTICE, Y (1999). Semiparametric inference in the proportional odds regression model, *J. Am. Statist. Assoc*., 94, 125-136.

QUANTIN, C., ABRAHAMOWICZ, M., MOREAU. T. *et al.* (1999). Variation over time of the effects of prognostic factors in a population-based study of colon cancer: comparison of statistical models. *Am. J. Epidemiol.*, 150, 11, 118-200.

REID, N. and CRÉPEAU, H. (1985). Influence function for proportional hazards regression. *Biometrika*, 72, 1–9.

SARGENT, D. J. (1997). A felxible approach to time-varying coefficients in the Cox regression setting. *Lifetime Data Anal.*, 3, 1, 13-25.

SCHOENFELD, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69, 239–241.

STORER, B. E. and CROWLEY, J. (1985). A diagnostic for Cox regression and general conditional likelihoods. *J. Am. Statist. Assoc*., 80, 139–147.

STRUTHERS, C. A. and KALBFLEISCH, J. D. (1986). Misspecified proportional hazard models. *Biometrika*, 73, 363–369.

THERNEAU, T. M., GRAMBSCH, P. M. and FLEMING, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77, 147 – 160.

VALSECCHI, M. G., SILVESTRI, D. and SASIENI, P. (1996). Evaluation of long-term survival: use of diagnostics and robust estimators with Cox's proportional hazards model. *Statistics in Medicine*, 15, 2763–2780.

WEY, L. J. (1984). Testing goodness of fit for proportional hazards model with censored observations. *J. Am. Statist. Assoc*., 79, 649–652.

L’analisi dei dati di durata ha ricevuto ampia attenzione nella letteratura statistica e i modelli per dati di durata sono ormai largamente impiegati nelle applicazioni di natura biometrica, economica, sociologica. Mentre nell’ambito della regressione lineare classica esiste una vasta letteratura che tratta in modo sistematico dei metodi diagnostici, per quanto riguarda i modelli per dati di durata, sono stati pochi i tentativi in letteratura di dare una forma organica all’argomento. Il presente lavoro offre un contributo in questa direzione: vengono presentati i principali metodi basati sull’analisi dei residui proposti nell’ambito del modello a rischio moltiplicativo, inclusi i contributi più recenti sull’argomento. I dati di durata saranno rappresentati come realizzazioni di processi di conteggio, consentendo di derivare in modo semplice i residui e i metodi diagnostici a essi associati. Alcuni risultati, usualmente ricavati nel caso semiparametrico, saranno estesi anche ai modelli di tipo parametrico e saranno evidenziati i legami tra i metodi sviluppati in ambito biometrico con quelli proposti in ambito econometrico.