UNIVERSITA DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE STATISTICHE
CORSO DI LAUREA IN STATISTICA, ECONOMIA E FINANZA

TESI DI LAUREA

UN'APPLICAZIONE DELLA REGRESSIONE QUANTILE

Relatore: Ch.ma Prof.ssa Laura Ventura

Laureanda: Veronica Giro
Matricola: 616759

Anno accademico 2011-2012

Alla mia famiglia e
alle mie amiche

Indice

Introduzione

La sopravvivenza è ciò che ogni essere umano ha a cuore, più di qualsiasi altra cosa. E' un istinto, ossia una forma di difesa e autoconservazione, è un fenomeno insito negli stessi impulsi vitali dai quali trae la sua esistenza unicamente per assicurare agli stessi la continuazione ad essere. È proprio la continuazione ad essere a costituire motivo di discussione e ricerca tra antropologi, filosofi e scienziati. Tale questione, quindi, coinvolge numerosi campi di studio e ricerca, sia umanistici che scientifici. Le scienze umanistiche si concentrano sulla spiegazione e argomentazione di tale concetto, studiando i comportamenti umani in situazioni rischiose, ossia in cui l'istinto di sopravvivenza arriva a sovrastare ogni cosa. Dal punto di vista scientifico, la statistica è una scienza che procede di pari passo con la ricerca bio-medica. L'importanza di tale applicazione sta nel fatto che la statistica consente di trarre delle conclusioni e fare inferenza facendo riferimento ad un campione piuttosto che a tutte le unità costituenti una popolazione. È per questo motivo che tale scienza ha acquisito importanza sempre maggiore nel tempo, in campo medico come in molti altri campi di applicazione. L'analisi della sopravvivenza è parte della statistica inferenziale e la sua peculiarità consiste nel mettere in rapporto un certo esito o evento, nella maggior parte dei casi la morte, con il fattore tempo. Essa tenta di trovare, o almeno di ipotizzare, una risposta a questioni riguardanti: il tempo di sopravvivenza di uno o più campioni, le cause che hanno provocato l'evento di interesse, la stratificazione della popolazione di riferimento, e cosa più importante l'efficacia di cure sperimentali. Nuove cure e teorie mediche, infatti, devono sempre essere testate e studiate prima di poter essere applicate all'intera popolazione. È per tale motivo che i metodi di analisi della sopravvivenza sono in continua evoluzione; l'obbiettivo è quello di elaborare teorie e metodi statistici sempre più efficienti ed efficaci, in modo da ridurre al minimo il margine di errore.

L'elaborato presenta un'analisi generale dei metodi di analisi della sopravvivenza. Nel primo capitolo vengono descritti i principali metodi di analisi in materia. Nel secondo capitolo viene presentato un modello in particolare, ossia la regressione quantile, con anche un accenno a recenti sviluppi sull'argomento (Bottai e Zhang, 2010). Infine, nel terzo capitolo vengono applicate alcune delle teorie e tecniche di analisi, esposte nei primi due capitoli dell'elaborato, ad un dataset riguardante pazienti malati di leucemia.

Capitolo 1

Metodi di analisi della sopravvivenza

In diversi settori applicativi è di interesse analizzare dati che rappresentano, per ciascuna unità, il tempo trascorso, dall'inizio dell'esperimento o dell'osservazione, fino al verificarsi di un evento di interesse. L'analisi di dati di durata è frequente in ambito medico. Infatti, per una certa patologia si possono analizzare i tempi di sopravvivenza di un gruppo di pazienti dal momento della diagnosi, oppure si possono confrontare tali misurazioni con quelle di un altro gruppo di soggetti con la stessa patologia, ma con alcune caratteristiche diverse. Invece in ambito industriale è molto utile avere a disposizione informazioni sulla durata di corretto funzionamento di un dato prodotto o componente. L'utilizzo di dati di durata è frequente anche nelle scienze sociali, economiche, demografiche come ad esempio per quanto riguarda la durata di disoccupazione o matrimonio.

Lo scopo di questo capitolo è fornire una sintesi sui metodi di analisi della sopravvivenza. Verranno fornite nozioni basilari, come la durata e la censura, e verranno illustrati i metodi principali di trattamento e interpretazione dei dati. Alcuni utili riferimenti bibliografici sono Pace e Salvan (2001), Cox (1972), Klein e Moeschberger (2003), Lawless (1982) e Marubini e Valsecchi (1995).

1.1 Dati di durata

Sia T una variabile casuale univariata continua, con distribuzione tipicamente asimmetrica, che descrive una durata aleatoria, e sia $F_T(t) = \Pr\ (T \leq t)$ la funzione di ripartizione di T.
Il suo complemento a uno, ossia

$$S_T(t) = 1 - F_T(t)\ , \tag{1.1}$$

rappresenta la funzione di sopravvivenza. Infatti, $S_T(t)$ esprime la possibilità che la durata T sia maggiore di t, ossia la probabilità che la sopravvivenza dell'unità sperimentale sia superiore a t.
Sia, inoltre, $f_T(t)$ la funzione di densità di T, tale che $F_T(t) = \int_0^t f_T(u)du$.
Un'altra funzione importante per descrivere la distribuzione di T è la funzione di azzardo (o tasso di guasto), data da

$$\lambda_T(t) = \frac{f_T(t)}{1 - F_T(t)} = \frac{f_T(t)}{S_T(t)}, \qquad t \geq 0 \tag{1.2}$$

nei punti dove $\lambda_T(t)$ è continua. Per $\varepsilon \geq 0$ sufficientemente piccolo, $\varepsilon\lambda_T(t)$ esprime la probabilità $\Pr\ (t < T < t + \varepsilon | T > t)$, ossia la possibilità che l'evento di interesse si verifichi tra t e $t + \varepsilon$, dato che il soggetto è vivo al tempo t. Un andamento crescente di $\lambda_T(t)$ è tipico di unità soggette ad

invecchiamento: tanto maggiore è il tempo di sopravvivenza dell'unità, tanto più grande è la probabilità condizionata di guasto. Viceversa, se $\lambda_T(t)$ è descrescente, si è di fronte ad unità che hanno una maggiore probabilità condizionata di guastarsi all'inizio del periodo di osservazione; si dice in tal caso che vi è selezione iniziale. Un andamento costante di $\lambda_T(t)$ caratterizza, invece, unità non soggette né ad invecchiamento né a selezione iniziale. Naturalmente, la funzione di azzardo può anche avere andamento non monotono. Nella Figura 1.1 sono rappresentati alcuni tra gli andamenti descritti precedentemente che possono essere assunti dalla funzione di azzardo. In base ad essi, si comprende in che modo si verifica l'evento d'interesse.

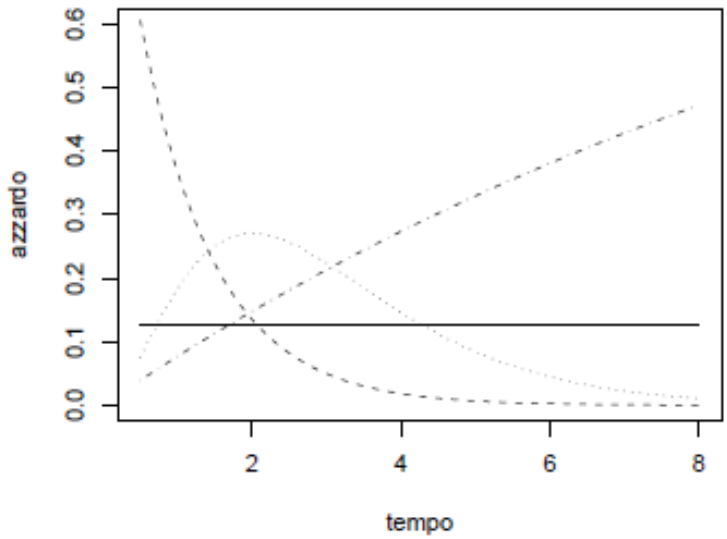

Figura 1.1: Possibili andamenti della funzione di azzardo rispetto al tempo.

Per $t \geq 0$, vale la relazione

$$f_T(t) = \lambda_T(t)\, exp\left\{-\int_0^t \lambda_T(u)du\right\} = \lambda_T(t)exp\{\Lambda_T(t)\}\ , \tag{1.3}$$

mentre $f_T(t) = 0$ per $t < 0$, dove $\Lambda_T(t) = \int_0^t f_T\,(u)du$ rappresenta la funzione di azzardo cumulato. Inoltre, per $t \geq 0$, si ha

$$F_T(t) = 1 - exp\left\{-\int_0^t \lambda_T(u)du\right\} = 1 - exp\{\Lambda_T(t)\} \tag{1.4}$$

e quindi la funzione di guasto determina univocamente la legge di probabilità.

1.2 Censura

Nell'ambito dell'analisi della sopravvivenza, il termine censura si riferisce a quelle unità statistiche su cui non si è verificato l'evento di interesse durante il periodo di osservazione. Si possono distinguere diversi tipi di censura:

- Censura di primo tipo. Al tempo $t = 0$, si mettono in osservazione n unità statistiche e si termina l'osservazione dopo un tempo prefissato $t_0 > 0$. Trascorso tale tempo, per $R \leq n$ unità si sarà verificato l'evento di interesse, mentre per le rimanenti $n - R$ unità, si saprà solamente che la durata è maggiore di t_0.
- Censura di secondo tipo. Al tempo t_0 si mettono in osservazione n unità e si termina la prova quando si è osservato l'evento di interesse per un numero r prefissato di unità.
- Censura variabile. È una generalizzazione della censura del primo tipo. In questo caso, il tempo prefissato di osservazione varia da unità a unità; i soggetti, infatti, entrano nella prova di durata a tempi diversi, ma lo studio deve essere completato entro un termine prefissato.
  Siano T_1,….,T_n le durate aleatorie delle n unità e siano $c_1, \ldots, c_n$ costanti positive assegnate, dove c_i rappresenta il tempo di osservazione per la i-esima unità. Si rappresentano i dati come realizzazioni delle variabili casuali (Y_i, δ_i), dove Y_i= min (T_i,c_i), e δ_i= $I_{(0,c_i)}(T_i)$, $i = 1, \ldots, n$, $I(\cdot)$ è una funzione indicatrice che indica se l'osservazione è censurata o meno. Da questo segue che δ_i=0 quando $Y_i = c_i$, ovvero quando la durata dell'evento non è interamente compresa entro il termine dello studio. Mentre δ_i=1 quando $Y_i = T_i$, ossia quando si osserva la durata del tempo di sopravvivenza, poiché esso si è concluso all'interno del periodo di tempo considerato.
- Censura casuale. In questo caso si prende in considerazione il fatto che, soprattutto negli studi clinici, il tempo di osservazione varia da soggetto a soggetto in modo non controllabile dallo sperimentatore. Un soggetto può uscire dallo studio per motivi personali, per trasferimento ad altra struttura, e per altri motivi non legati alla natura dello studio. E' perciò più realistico assumere che i tempi di censura siano a loro volta realizzazioni di variabili casuali.

1.3 Metodi di analisi della sopravvivenza

I metodi di analisi della sopravvivenza fanno riferimento a tre approcci:

- metodi non parametrici, per i quali la distribuzione di T non è specificata;
- metodi parametrici, in cui la distribuzione di T è completamente specificata con una particolare forma funzionale;
- metodi semi-parametrici, per i quali la distribuzione di T non è completamente specificata.

1.3.1 Metodi non parametrici

L'oggetto di primario interesse dell'inferenza è la funzione di sopravvivenza $S_T(t)$. In particolare, può essere importante confrontare stime della funzione di sopravvivenza calcolate per diversi strati di unità campionarie. Ciascuno strato corrisponde tipicamente a differenti condizioni sperimentali.

Gli indicatori di sintesi tradizionali non possono essere usati per la sopravvivenza poiché forniscono risultati distorti, in particolare sottostimati, per via dei dati censurati. Disporre di una stima non parametrica della funzione di ripartizione o della funzione di sopravvivenza è sempre utile, poiché costituisce uno strumento importante per il controllo empirico del modello statistico; inoltre, può anche rivestire un interesse autonomo, se le informazioni disponibili non permettono di assumere uno specifico modello parametrico.

Se si dispone di un campione $(t_1,\ldots..,t_n)$ completo, ossia senza censura, la stima non parametrica di $S_T(t)$ è

$$\hat{S}_n(t)= 1 - \hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I_{(t,+\infty)}(t_i), \qquad t > 0. \tag{1.5}$$

Poiché $n\hat{S}_n(t) \sim Bi\big(n, S^0(t)\big)$, con $S^0(t) = 1 - F^0(\cdot)$ che indica l'ignota funzione di sopravvivenza marginale di $T_i, i = 1,..,n$, una stima della varianza di $\hat{S}_n(t)$ è

$$\widehat{V}(\hat{S}_n(t))=\frac{\hat{S}_n(t)(1-\hat{S}_n(t))}{n} . \tag{1.6}$$

Si può ancora utilizzare $\hat{S}_n(t)$ come stima di $S^0(t)$, anche in presenza di censura del primo o del secondo tipo. Tuttavia per $t \geq t_0$, con censura del primo tipo, e per $t > t_r$, con censura del secondo tipo, con *r* che indica le unità per le quali si è verificato l'evento d'interesse entro il tempo stabilito, la stima è costante, al valore assunto in t_0 e t_r, rispettivamente.

Se invece le osservazioni di durata sono soggette a censura variabile o casuale, allora la stima della funzione di sopravvivenza va modificata nel modo seguente.

Si considerino il numero di unità a rischio al tempo t, ossia il numero di soggetti per cui $y_i = min\,(t_i,c_i)$ risulta maggiore di t, $i = 1,\ldots,n$, ovvero che potrebbero ancora presentare l'evento. Posto $t_0 = 0$, si considerino i *J* tempi all'evento ordinati in senso crescente, ossia $t_{(1)} < t_{(2)} < \cdots < t_{(J)}$, che rappresentano i tempi in cui si verificano uno o più eventi. All'origine t_0 sono a rischio n soggetti e il loro numero diminuisce nel tempo. Siano quindi $d_1, d_2, \ldots, d_J$ il numero di eventi al tempo $t_{(j)}$, e siano $n_1, n_2, \ldots n_J$ il numero di soggetti a rischio ai vari tempi. Per cui si può definire

$$\hat{S}_n^{KM}(t_{(j)})= \prod_{i=1}^{j} \frac{n_i-d_i}{n_i} , \tag{1.7}$$

con $j = 1, \dots, J$ e $J \leq n$. Si noti che il prodotto è esteso a tutti i tempi minori di t in cui si verificano guasti, ossia in cui si hanno osservazioni non censurate. Lo stimatore (1.7) è noto come stimatore di Kaplan-Meier della funzione di sopravvivenza. Si osservi che ogni singolo elemento della produttoria denota la stima relativa al generico tempo $t_{(j)}$ di sopravvivere, dato che si è sopravvissuti al tempo precedente. Quindi n_j indica i soggetti che potrebbero ancora presentare l'evento, che sono ancora vivi e sotto osservazione all'inizio del tempo $t_{(j)}$.

Uno stimatore della varianza dello stimatore di Kaplan-Meier è

$$\sqrt{\hat{V}(\hat{S}_n^{KM}(t_{(\mathrm{j})}))} = \hat{S}_n^{KM}(t_{(j)}) \sqrt{\sum_{i=1}^{J-1} \frac{d_i}{n_i(n_i - d_i)}} \tag{1.8}$$

Tale espressione è detta formula di Greenwood ed esprime la precisione della stima della probabilità di sopravvivenza. Essa è inversamente proporzionale, seppure in modo complesso, al numero di soggetti a rischio al tempo $t_{(j)}$.

Assumendo che la statistica $\hat{S}_n^{KM}(t_j)$ sia distribuita approssimativamente in modo gaussiano, gli intervalli di confidenza a livello (1-α)% per la stima di sopravvivenza al tempo t_j sono ottenuti, quindi, come:

$$\hat{S}_n^{KM}(t_j) \pm z_{1-\alpha/2} \sqrt{\hat{V}(\hat{S}_n^{KM}(\mathrm{t_j})} \ . \tag{1.9}$$

In generale, la stima di Kaplan-Meier della probabilità di sopravvivenza viene rappresentata con una curva a gradini che parte da 1, poiché al tempo 0 tutti i soggetti sono vivi, che decresce nel tempo e che cambia valore solamente in corrispondenza dei tempi in cui si osserva almeno un evento. Inoltre, l'altezza dei gradini dipende sia dal numero di eventi che dal numero di soggetti a rischio e di dati censurati. Le osservazioni censurate sono rappresentate per mezzo di una croce. Nella Figura 1.2 è rappresentata una curva di sopravvivenza e le relative bande di confidenza; tali elementi sono stati tracciati secondo le modalità descritte sopra.

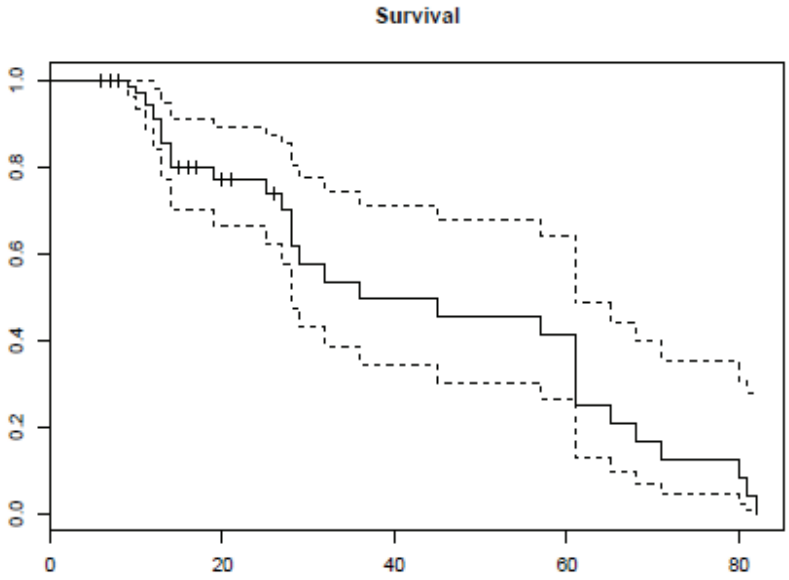

Figura 1.2 : Rappresentazione grafica della stima di Kaplan-Meier.

Uno degli assunti nella costruzione della curva di Kaplan-Meier è che il campione sia omogeneo, ma può capitare che i dati di uno studio siano eterogenei a causa del disegno sperimentale. Ad esempio, nel caso di uno studio in cui sia necessario confrontare due diversi tipi di trattamento, oppure a causa di caratteristiche intrinseche alle unità statistiche. In questi casi, risulta di interesse confrontare non parametricamente due o più curve di sopravvivenza. Per fare ciò, un possibile approccio è costituito dai test basati sui ranghi.

Il log-rank test è un test non parametrico basato sui ranghi ed è il più utilizzato per saggiare l'ipotesi nulla di uguaglianza delle funzioni azzardo in due o più gruppi, ovvero H_0: $\lambda_T^A(t) = \lambda_T^B(t)$, $\forall t$. Tale statistica test considera, per ogni tempo all'evento t_j , $j = 1 \dots J$, la distanza tra il numero di eventi osservati e quelli attesi se fosse vera H_0.

1.3.2 Metodi parametrici

Si assuma che T_1,…..,T_n siano variabili casuali, continue, indipendenti ed identicamente distribuite con funzione di sopravvivenza marginale $S_T(t;\theta)$, associata alla densità f_T $(t;\theta)$, con $\theta \in \Theta \subseteq \mathbb{R}^p$. Usualmente, $f_T(t;\theta)$ è una densità asimmetrica positiva, e la distribuzione esponenziale e la distribuzione di Weibull giocano un ruolo centrale nell'ambito dell'analisi della sopravvivenza.

Per l'inferenza su θ, e quindi anche su $S_T(t;\theta)$, anche in presenza di dati censurati, si utilizzano i metodi basati sulla funzione di verosimiglianza per θ. I dati sono rappresentabili come realizzazioni di variabili casuali (Y_i, δ_i), con Y_i = min (T_i, c_i), $i = 1, \dots, n$. Il contributo alla verosimiglianza per θ è dunque $f_T(t;\theta)$ quando δ_i=1, ossia quando si osserva l'evento, mentre è S_T $(t;\theta)$ quando δ_i=0. Allora le coppie (Y_i, δ_i) sono indipendenti con funzione di densità marginale $p(y_i, \delta_i; \theta) = f_T(y_i;\theta)^{\delta_i}(1 - F_T(y_i;\theta))^{(1-\delta_i)}$, per $0 \leq y_i \leq c_i$, $i = 1, \dots, n$.

La funzione di verosimiglianza per θ è quindi

$$L(\theta)=\prod_{i=1}^{n} f_T(y_i;\theta)^{\delta_i}(1-F_T(y_i;\theta))^{(1-\delta_i)} = \prod_{i=1}^{n} f_T(y_i;\theta)^{\delta_i} S_T(y_i;\theta)^{(1-\delta_i)} \quad . \quad (1.10)$$

Essa è ottenuta come prodotto delle densità per le osservazioni non censurate moltiplicato per il prodotto delle funzioni di sopravvivenza per le osservazioni censurate.

La funzione di log-verosimiglianza per θ è

$$l(\theta)=\sum_{i=1}^{n} \delta_i \log f_T(y_i;\theta) + \sum_{i=1}^{n}(1-\delta_i)\log S_T(y_{i;}\theta) = \sum_{i=1}^{n} \delta_i \lambda_T(t_i;\theta) + \sum_{i=1}^{n} \Lambda_T(t_i;\theta) . \quad (1.11)$$

La stima di massima verosimiglianza $\hat{\theta}$ si ottiene come massimo della funzione di log-verosimiglianza, tenendo eventualmente conto della censura. A partire da essa si arriva a calcolare la stima della funzione di sopravvivenza, $\hat{S}_T(t;\hat{\theta})$, e una stima del suo errore standard.

Come detto in precedenza, la distribuzione esponenziale e la distribuzione di Weibull sono ampiamente utilizzate per l'analisi dei dati di sopravvivenza.
Per quanto riguarda la distribuzione esponenziale, essa ha funzione di probabilità $f_T(t;\lambda) = \lambda e^{-\lambda t}$, con $\lambda > 0$, e la funzione azzardo è costante, ossia $\lambda_T(t;\lambda) = \lambda$. La funzione di azzardo cumulato è $\Lambda_T(t;\lambda) = \lambda t$ e la sopravvivenza è $S_T(t;\lambda) = e^{-\lambda y}$. La stima di massima verosimiglianza di λ è $\hat{\lambda} = d/(n\bar{t})$, dove $d = \sum_{i=1}^{n} \delta_i$ è il numero totale di eventi e $\bar{t} = \sum_{i=1}^{n} t_i/n$ è la media dei tempi dell'evento.
Per quanto riguarda, invece, la distribuzione di Weibull, essa ha come funzione di probabilità
$f_T(t;\lambda,k) = k\lambda^k t^{k-1} e^{-(\lambda t)^k}$, con $\lambda > 0$ e $k > 0$, e funzione di azzardo
$\lambda_T(t;\lambda,k) = k\lambda(\lambda t)^{k-1}$. Il parametro di forma k governa la forma della funzione di azzardo, che può essere monotona, crescente o decrescente a seconda del valore assunto da k. L'azzardo cumulato è $\Lambda_T(t;\lambda,k) = (\lambda t)^k$ e la sopravvivenza è $S_T(t;\lambda,) = e^{-(\lambda t)^k}$.

Nei modelli di regressione per dati di sopravvivenza, l'interesse si sposta allo studio della relazione fra il tempo all'evento e un insieme di variabili esplicative $(x_1, \dots \dots, x_p)$, allo scopo di individuare fattori prognostici o studiare eventuali interazioni tra covariate. Supponiamo di disporre di un campione di dati di sopravvivenza con censura, della forma (y_i, δ_i), con $\delta_i = 1$ se l'evento è osservato e $\delta_i = 0$ se il tempo è censurato, con $i = 1, \dots, n$. Spesso interessa valutare se e come la distribuzione di sopravvivenza è influenzata da variabili concomitanti $(x_{i1},\dots..,x_{ip})$, dove x_{ir} è il valore assunto dalla r-esima variabile concomitante per l'i-esima unità statistica. La modellazione più semplice si ha specificando un modello parametrico per la distribuzione marginale delle durate T_i, con parametro θ_i espresso in funzione delle variabili esplicative.

In particolare, se T_i ha distribuzione esponenziale con tasso di guasto λ_i, $i = 1, \dots, n$, un modello di regressione esponenziale esprime la dipendenza di λ_i da p valori esplicativi $(x_{i1},\dots..,x_{ip})$. Varie

specificazioni sono possibili per la relazione che collega λ_i ai valori esplicativi. Poiché λ_i può assumere solo valori positivi, può essere conveniente specificare un modello log-lineare della forma,

$$log\lambda_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \ , \tag{1.12}$$

dove $\beta_1, \ldots, \beta_p$ sono parametri reali, comuni a tutte le osservazioni. Se si desidera che il modello abbia intercetta, e quindi x_{i1} pari ad 1 per ogni i, si può scrivere $\lambda_T\ (t\ ; x_i) = \lambda_0 e^{\beta_2 x_{i2} + \cdots + \beta_p x_{ip}}$, con $\lambda_0 = e^{\beta_0}$. Per fare inferenza su $\beta = (\beta_0, \ldots, \beta_p)$ si usano gli usuali metodi di verosimiglianza. La funzione di log-verosimiglianza è

$$l(\beta) = \beta_1 \sum_{i=1}^{n} \delta_1 x_{i1} + \cdots + \beta_p \sum_{i=1}^{n} \delta_1 x_{ip} - \sum_{i=1}^{n} y_i exp\ (\beta_1 x_{i1} + \ldots + \beta_p x_{ip}), \tag{1.13}$$

la funzione di punteggio è

$$\frac{\partial l(\beta)}{\partial \beta_r} = \sum_{i=1}^{n} \delta_1 x_{ir} - \sum_{i=1}^{n} x_{ir} y_i\ exp\ (\beta_1 x_{i1} + \ldots + \beta_p x_{ip})\ , \tag{1.14}$$

con $r = 1, \ldots, p$. Quando esiste, la stima di massima verosimiglianza $\hat{\beta}$ è la soluzione del sistema di equazioni lineari $\partial l(\beta)/\partial \beta_r = 0$, per $r = 1, .., p$.

Inoltre, la matrice di informazione attesa $i(\beta)$ ha generico elemento

$$i_{rs}(\beta) = \sum_{i=1}^{n} x_{ir} x_{is}, \tag{1.15}$$

per $r, s = 1, \ldots, p$.

Vari altri modelli di regressione si possono definire utilizzando distribuzioni diverse dall'esponenziale per le distribuzioni marginali delle durate, ad esempio la distribuzione Gamma o la distribuzione di Weibull. Per specificare un modello di regressione è necessario formulare una relazione che esprima l'azzardo come una opportuna funzione delle variabili esplicative.

Il modello di regressione di Weibull prevede che

$$\lambda_T(t; x_i) = \lambda k (\lambda t)^{k-1} e^{\beta_1 x_{i1} + \ldots + \beta_p x_{ip}} = \lambda_0(t) e^{\beta_2 x_{i2} + \ldots + \beta_p x_{ip}}\ . \tag{1.16}$$

Il parametro k definisce la forma della funzione azzardo per ogni pattern di covariate. In generale, qualunque sia $\lambda_0(t)$, un modello di regressione di Weibull ha rapporto tra azzardi (HR) pari a

$$HR = \frac{\lambda_T(t; x_i^1)}{\lambda_T(t; x_i^2)} = e^{\beta_1 (x_{i1}^1 - x_{i1}^2) + \ldots + \beta_p (x_{ip}^1 - x_{ip}^2)}\ , \tag{1.17}$$

ossia costante nel tempo, con x^1 e x^2 vettori di covariate di due soggetti. Inoltre, l'effetto di ogni covariata sull'HR è moltiplicativo, ossia per una variazione unitaria nel valore della covariata x_r, HR è moltiplicativo di e^{β_r}.

1.3.3 Metodi semiparametrici

Nell'analisi dei dati di sopravvivenza si fa spesso ricorso a modelli di regressione semiparametrici. Rispetto alla formulazione interamente parametrica, si fornisce una specificazione meno dettagliata delle distribuzioni marginali dei tempi all'evento. Il modello più noto è senz'altro il modello di Cox con rischi proporzionali (Cox, 1972). Esso esprime l'azzardo in funzione del tempo e delle covariate, senza però formalizzare la dipendenza dal tempo. Tale modello assume che

$$\lambda_T(t; x_i) = \lambda_0(t) exp\ (\beta_1 x_{i1} + ... + \beta_p x_{ip})\ , \tag{1.18}$$

dove la funzione ignota $\lambda_0(t)$ non ha una forma specifica, dipende solo dal tempo, è la stessa per tutti i soggetti ed è detta azzardo di base. La dipendenza da t, ovvero dal tempo in cui è osservato l'evento, è inglobata interamente nella funzione $\lambda_0(t)$, che ha un'interpretazione diretta, molto simile a quella dell'intercetta in un modello lineare, ossia $\lambda_0(t) = \lambda_T(t; 0)$. Tale modello è definito come modello ad azzardi proporzionali (PH), con costanti di proporzionalità determinate dai termini esponenziali dipendenti dalle variabili esplicative, ma non da t. Infatti, il rapporto fra gli azzardi di due soggetti con vettori di covariate x^1 e x^2 è dato da

$$HR = \frac{\lambda_T(t; x_i^1)}{\lambda_T(t; x_i^2)} = e^{\beta_1(x_{i1}^1 - x_{i1}^2) + \cdots + \beta_p(x_{ip}^1 - x_{ip}^2)} \tag{1.19}$$

che non dipende dal tempo. Ciò sta a significare che le funzioni di azzardo cambiano nel tempo ma il loro rapporto non si modifica. Inoltre, le variabili esplicative hanno un effetto moltiplicativo sull'azzardo di base, mentre l'effetto è additivo sul logaritmo di quest'ultimo. La funzione di sopravvivenza si deriva dalla formulazione di base. L'azzardo cumulato, infatti, è pari a $\Lambda_T(t; x_i) = \Lambda_0(t)\exp\ (\beta_1 x_{i1} + ... + \beta_p x_{ip})$ e la funzione di sopravvivenza diventa $S_T(t; x_i) = \exp\big(-\Lambda_T(t; x_i)\big) = S_0(t)^{\exp\ (\beta_1 x_{i1} + \cdots + \beta_p x_{ip})}$.

Per la r-esima covariata, $\exp\ (\beta_r)$ esprime di quanto varia moltiplicativamente il tasso di evento per ogni variazione unitaria di x_r , a parità di tutte le altre covariate. In questo contesto, la funzione di verosimiglianza completa viene a dipendere da β, ma anche da $\lambda_0(t)$, che non è specificata e che costituisce quindi un parametro di disturbo infinito dimensionale.

Cox propose di fattorizzare la verosimiglianza completa come $L(\lambda_0, \beta) = L(\lambda_0)L(\beta|\lambda_0)$, e di considerare per l'inferenza su β solo la seconda componente, che dipende solamente da β e non contiene $\lambda_0(t)$, ossia di considerare solo la verosimiglianza parziale $L_P(\beta) = L(\beta|\lambda_0)$.

Siano $t_{(j)}$ i tempi all'evento ordinati e siano $R(t_{(j)})$ i set a rischio di evento in $t_{(j)}$, con $j = 1, \dots, J$. La verosimiglianza parziale è definita come:

$$L_P(\beta) = \prod_{j=1}^{J} \frac{\lambda_T(t_j; x_j)dt}{\sum_{i \in R(t_j)} \lambda_T(t_j; x_j)dt} = \prod_{j=1}^{J} \frac{exp\,(\beta_1 x_{j1} + \cdots + \beta_p x_{jp})}{\sum_{i \in R(t_j)} exp\,(\beta_1 x_{i1} + \cdots + \beta_p x_{ip})} \quad , \qquad (1.20)$$

dove l'argomento della produttoria indica la probabilità che un soggetto con covariate $x_{(j)}$ abbia un evento al tempo $t_{(j)}$ dato che ci sia un evento in $t_{(j)}$. La verosimiglianza parziale può quindi essere utilizzata per l'inferenza su β come una funzione di verosimiglianza genuina. Infatti, la verosimiglianza parziale si comporta come una verosimiglianza propria sotto molti punti di vista; ad esempio, il valore atteso della funzione di punteggio parziale è uguale a zero. Inoltre, vale l'approssimazione normale per la distribuzione dello stimatore di massima verosimiglianza parziale, la cui varianza asintotica è consistentemente stimata dall'inversa di $-\partial^2 log L_P(\beta)/(\partial\beta\partial\beta^T)$. Anche la statistica log-rapporto di verosimiglianza parziale ha l'usuale distribuzione asintotica nulla.

Per ottenere una stima della funzione di sopravvivenza, è necessaria una stima non parametrica di $S_0(t)$, oltre alla stima di β ottenuta a partire dalla verosimiglianza parziale. Si usa quindi

$$\hat{S}_0(t) = \prod_{t_j \le t} \left(1 - \frac{d_j}{\sum_{i \in R(t_j)} exp\,(\hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip})}\right) , \qquad (1.21)$$

dove d_j indica il numero di eventi in t_j. Si noti che per $\beta = 0$ si ottiene lo stimatore di Kaplan-Meier.

Una volta stimato il modello di Cox è necessario verificare l'adeguatezza di alcuni aspetti, tra cui la validità dell'assunto del PH, attraverso metodi grafici e con l'analisi dei residui; la forma funzionale secondo cui le covariate continue influenzano la variabile risposta, con l'analisi dei residui e la presenza di singole osservazioni influenti o di outliers.

1.4 Conclusioni

Il capitolo ha sinteticamente richiamato le nozioni basilari per procedere nell'analisi dei dati di durata. Sono stati presentati e descritti i tre metodi di studio più utilizzati per affrontare problemi relativi alla sopravvivenza. Il passo successivo sarà costituito dall'analisi di un ulteriore approccio: la regressione quantile. Tale argomento, che costituisce il tema principale della tesi, verrà illustrato nel prossimo capitolo.

Capitolo 2

La regressione Quantile

La regressione quantile è stata introdotta per la prima volta nel 1978 da Koenker e Basset, i quali affermarono che *"dai metodi classici di regressione, l'unica informazione che si ottiene sulla relazione tra Y e il vettore delle covariate X, è il modo in cui la media di Y varia al variare di X".* Il grande vantaggio della regressione quantile è la possibilità di stimare l'intera distribuzione dei quantili condizionati della variabile risposta, così da poter studiare l'influenza delle variabili esplicative sulla forma della distribuzione di Y. Alcuni utili riferimenti bibliografici sono Koenker (2005), Portnoy (2003), Koenker e Hallock (2001).

2.1 Quantile e Funzione quantile

Il quantile-τ di Y è un numero reale y_τ tale che Pr $(Y \leq y_\tau) \geq \tau$ e che Pr$(Y \geq y_\tau) \geq 1 - \tau$. Vi è un unico quantile-τ solo se l'equazione $F_Y(y) = \tau$ ha al più una soluzione. Spesso la distribuzione di una variabile casuale univariata viene tabulata sintetizzando alcuni quantili tipici, ad esempio sulla coda sinistra: il primo percentile $y_{0.01}$, il quinto percentile $y_{0.05}$, il primo decile $y_{0,10}$; al centro della distribuzione: il primo quartile $y_{0.25}$, la mediana $y_{0.50}$, il terzo quartile $y_{0.75}$; infine sulla coda destra: il nono decile $y_{0.90}$, il novantacinquesimo percentile $y_{0.95,}$ e il novantanovesimo percentile $y_{0.99}$.

Si dice funzione quantile di Y una applicazione $\tau \rightarrow F_Y^{-1}(\tau)$ che fa corrispondere a $0 \leq \tau \leq 1$ un opportuno quantile-τ della variabile casuale Y. Si definisce, quindi, la funzione quantile come:

$$Q(\tau) = F_Y^{-1}(\tau) = \inf\{y \in \mathbb{R} : F_Y(y) \geq \tau\}, \quad \text{per } 0 \leq \tau \qquad (2.1)$$

dove $F(\cdot)$ è la funzione di ripartizione della variabile casuale Y.
In altre parole, la funzione quantile è il meccanismo che lega le modalità osservate in un campione, ovvero i valori possibili di una variabile casuale (o popolazione) alle frequenze o probabilità con cui sono osservate o potrebbero essere osservate. Esprime, quindi, per ogni τ, il valore della variabile casuale cui è associata la probabilità p che si realizzi una osservazione ad esso inferiore o uguale e la probabilità $(1 - p)$ che si realizzi un valore ad essa superiore. La funzione quantile è definita come l'inversa della funzione di ripartizione, infatti $Q(\tau) = F^{-1}(\tau)$ e $F(x) = Q^{-1}(x)$, quando, ovviamente, le inverse esistono.

2.2 La regressione quantile

I quantili possono essere calcolati come soluzione di un semplice problema di ottimizzazione. Per ogni $0 \leq \tau \leq 1$, si definisce la "funzione di controllo" lineare a tratti come
$\rho_p(u) = u(p - I(u < 0))$, illustrata nella Figura 2.1.

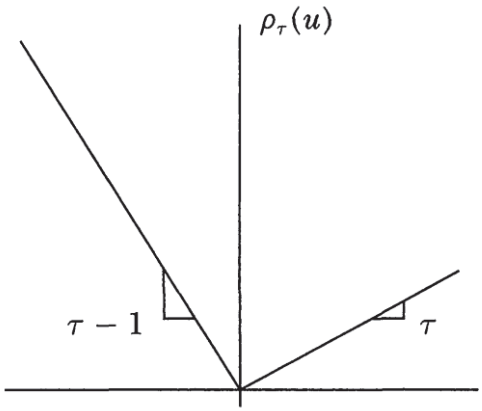

Figura 2.1: Funzione di controllo

Più precisamente, in analogia a quanto succede per la media campionaria, che può essere definita come la soluzione del problema di minimizzazione della somma degli scarti al quadrato, in questo caso possiamo definire ogni singolo quantile campionario $\xi(\tau)$, che è l'analogo di $Q(\tau)$, come soluzione del seguente problema di minimo

$$min_{\xi \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \xi). \tag{2.2}$$

È certamente di uso più comune definire i quantili campionari come una sorta di riordinamento del campione originale, $y_{(1)} \leq y_{(2)} \leq, \ldots\ldots, \leq y_{(n)}$; ma la loro formulazione come problema di minimizzazione ha il vantaggio di fornire una naturale generalizzazione dei quantili verso il contesto di regressione.

Come l'idea di stimare la media, vista come minimo di $\hat{\mu} = argmin_{\mu \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \mu)^2$, può essere estesa alla stima della funzione della media lineare condizionata $E(Y|X = x) = x'\beta$ risolvendo $\hat{\beta} = argmin_{\beta \in \mathbb{R}^p} \sum (y_i - x'_i\beta)^2$, la funzione quantile lineare condizionata, $Q_Y(\tau|X = x) = x'_i\beta(\tau)$, può essere stimata risolvendo

$$\hat{\beta}(\tau) = argmin_{\beta \in \mathbb{R}^p} \sum \rho_\tau(y_i - x'\beta)^2. \tag{2.3}$$

2.2.1 Interpretazione della regressione quantile

La stima dei minimi quadrati dei modelli di regressione media risponde alla domanda, " In che modo la media condizionata di Y dipende dalle variabili esplicative X?". La regressione quantile, invece, dà una risposta a questa domanda per ogni quantile della distribuzione condizionata fornendo una più completa descrizione di come la distribuzione condizionata di Y dato $X = x$ dipende da x. Invece di assumere che le variabili esplicative cambino solamente la posizione o la scala della distribuzione condizionata, i metodi di regressione quantile permettono di analizzare i potenziali effetti circa la forma della distribuzione.

La formulazione più semplice circa la regressione quantile è il modello a due campioni trattamento-controllo. Al posto del classico modello di sperimentazione, nel quale il trattamento comporta un semplice spostamento della distribuzione della variabile risposta, Lehmann (1974) propose il seguente modello generale di risposta al trattamento: " *Si supponga che il trattamento aggiunga la quantità* $\Delta(x)$ *quando la risposta dei soggetti non trattati è* x. *Dunque, la distribuzione* G *delle risposte al trattamento è data da* $X + \Delta(X)$, *dove* X *è distribuita in accordo a* F*."* Tra i casi particolari, troviamo il modello di traslazione, $\Delta(X) = \Delta_0$, e il modello di cambiamento di scala, $\Delta(X) = \Delta_0 X$,ma naturalmente il caso generale si trova all'interno del paradigma della regressione quantile.

Doksum (1974) mostrò che se $\Delta(x)$ è definito come la "distanza orizzontale" tra F e G, allora

$$F(x) = G(x + \Delta(x)) \tag{2.4}$$

e $\Delta(x)$ è definito unicamente e può essere espresso come

$$\Delta(x) = G^{-1}\big(F(x)\big) - x. \tag{2.5}$$

Cambiando la variabile come $\tau = F(x)$, si può definire l'effetto di trattamento quantile come

$$\delta(\tau) = \Delta\big(F^{-1}(\tau)\big) = G^{-1}(\tau) - F^{-1}(\tau). \tag{2.6}$$

Nell'impostazione a due campioni questa quantità è stimata come

$$\hat{\delta}(\tau) = \hat{G}_n^{-1}(\tau) - \hat{F}_m^{-1}(\tau), \tag{2.7}$$

dove G_ne F_m indicano rispettivamente le funzioni di distribuzione empirica delle osservazioni appartenenti ai gruppi trattamento e controllo.

Nel grafico in Figura 2.2 sono tracciate le funzioni di distribuzione (marginale) dei gruppi trattamento e controllo, la differenza tra le due curve rappresenta l'effetto del trattamento, stimato dalla 2.7.

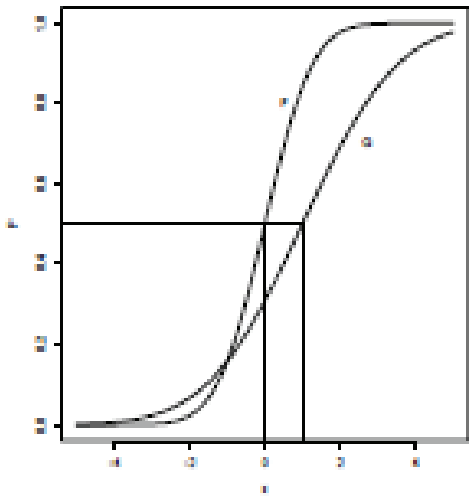

Figura 2.2 : Effetto di trattamento quantile: distanza orizzontale tra le funzioni di distribuzione (marginale) dei gruppi trattamento (G) e controllo (F).

Si formula, quindi, il modello di regressione quantile per il problema come

$$Q_{Y_i}(\tau|D_i) = \alpha(\tau) + \delta(\tau)D_i, \tag{2.8}$$

dove D_i denota l'indicatore del trattamento, ossia $D_i = 1$ indica il gruppo di trattamento e $D_i = 0$ il controllo. Inoltre, l'effetto quantile del trattamento può essere stimato risolvendo l'equazione

$$\left(\hat{\alpha}(\tau), \hat{\delta}(\tau)\right)' = argmin \sum_{i=1}^{n} \rho_\tau (y_i - \alpha - \delta D_i). \tag{2.9}$$

La soluzione dell'equazione $\left(\hat{\alpha}(\tau), \hat{\delta}(\tau)\right)'$ fornisce come risultati: $\hat{\alpha}(\tau) = \hat{F}_n^{-1}(\tau)$, che corrisponde al campione di controllo, e $\hat{\delta}(\tau) = \hat{G}_n^{-1}(\tau) - \hat{F}_n^{-1}(\tau)$.

Doksum (1974) suggerì di interpretare i soggetti del gruppo di controllo in termini di una caratteristica latente. Ad esempio, nelle applicazioni della *survival analysis*, i soggetti del gruppo di controllo possono essere etichettati come deboli se sono inclini a morire in tempi brevi, oppure robusti se sono inclini a morire in tempi più lunghi. Questa caratteristica latente è, quindi, implicitamente indicizzata da τ, il quantile della distribuzione di sopravvivenza in cui si trova il soggetto se non ha subito alcun trattamento, $(Y_i|D_i = 0) = \alpha(\tau)$. Per quanto riguarda il gruppo trattamento si assume di trasformare la risposta al trattamento di controllo, $\alpha(\tau)$, in $\alpha(\tau) + \delta(\tau)$. Se la caratteristica latente, ossia la propensione alla longevità, è stata osservata ex ante, si può interpretare l'effetto del trattamento $\delta(\tau)$ come un'interazione esplicita con questa variabile osservata. Mentre, in assenza di tale variabile osservata, l'effetto quantile del trattamento può essere considerato come una misura naturale della risposta al trattamento.

L'effetto del trattamento quantile trova un'efficace rappresentazione nel diagramma di dispersione a due campioni, il quale ha una lunga storia come dispositivo diagnostico grafico. La funzione $\Delta(x) = G^{-1}\big(F(x)\big) - x$ rappresenta esattamente ciò che viene disegnato in un diagramma di dispersione tradizionale a due campioni. Se F e G sono uguali allora la funzione $G^{-1}\big(F(x)\big)$ sarà disposta lungo la bisettrice del quadrante; se le due funzioni differiscono solamente in termini di locazione di scala, allora $G^{-1}\big(F(x)\big)$ si disporrà lungo un'altra linea con intercetta e pendenza determinate rispettivamente dalla differenza di posizione e scala. La regressione quantile può essere vista come un mezzo per estendere il diagramma di dispersione a due campioni tradizionale.

Quando la variabile trattamento presenta più di due valori possibili, l'effetto del trattamento quantile di Lehmann-Doksum necessita solamente di una piccola reinterpretazione. Se la variabile trattamento è continua come, ad esempio, negli studi di dose-risposta, allora è naturale considerare come assunzione il fatto che il suo effetto sia lineare, e scrivere

$$Q_{Y_i}(\tau|D_i) = \alpha(\tau) + \beta(\tau)x_i. \tag{2.10}$$

Assumiamo in tal modo che l'effetto del trattamento $\beta(\tau)$ di cambiare x da x_0 a $x_0 + 1$ abbia lo stesso effetto di un'alterazione da x_1 a $x_1 + 1$. Si noti che tale nozione riguardante l'effetto del trattamento quantile misura, per ogni τ, il cambiamento nella risposta richiesto per stare nel quantile τ della funzione quantile condizionata.

2.2.2 Caratteristiche e vantaggi della regressione quantile

Un'importante proprietà del modello di regressione quantile è il fatto che, per ogni funzione monotona $h(\cdot)$, vale

$$Q_{h(\tau)}(\tau|x) = h\big(Q_\tau(\tau|x)\big). \tag{2.11}$$

La 2.11 implica, quindi, che i quantili condizionati della risposta trasformata siano equivalenti ai quantili trasformati della variabile risposta.

Ciò segue immediatamente osservando che

$$Pr(T < t|x) = Pr(h(T) < h(t)|x).$$

Tale invarianza di fronte a trasformazioni monotone della funzione quantile condizionata è una caratteristica fondamentale. Tale proprietà, infatti, risulta essere più forte nella regressione quantile che negli altri tipi di regressione.

Un'applicazione particolarmente importante dell'invarianza, che si è dimostrata influente nei modelli di regressione quantile utilizzati in econometria, comporta la censura della variabile risposta osservata. Il più semplice modello di censura può essere formulato come segue.
Sia y_i^* una variabile risposta latente (inosservabile). Si assuma, inoltre, che sia stata generata dal modello lineare

$$y_i^* = x_i'\beta + u_i, \qquad i = 1, .., n \tag{2.12}$$

con $\{u_i\}$ v.c. iid proveniente da una funzione di distribuzione F. A causa della censura, le y_i^* non sono osservate direttamente, bensì si osserva

$$y_i = max\{0, y_i^*\}. \tag{2.13}$$

Powell (1986) notò che l'invarianza dei quantili di fronte a trasformazioni monotone implica che, in questo tipo di modello, la funzione quantile condizionata della risposta dipenda solamente dal punto di censura, ma sia indipendente da F. Formalmente, la funzione quantile τ della risposta osservata, y_i, in questo modello può essere espressa come

$$Q_i(\tau|x_i) = max\{0, x_i'\beta + F_u^{-1}(\tau)\}. \tag{2.14}$$

I parametri delle funzioni quantile condizionate possono essere stimati risolvendo:

$$min_b \sum_{i=1}^{n} \rho_\tau(y_i - max\{0, x_i'b\}), \tag{2.15}$$

dove è stato assunto che i vettori x_i, contengano un'intercetta che assorba l'effetto additivo di $F_u^{-1}(\tau)$. Questo modello è alquanto più esigente, in termini di calcolo, rispetto alla regressione quantile lineare classica poichè non è lineare nei parametri.

La robustezza nelle ipotesi distributive è una caratteristica rilevante in statistica. Infatti, è importante sottolineare che la regressione quantile riceve in eredità alcune proprietà di robustezza dai quantili di un campione ordinario. Le stime e l'inferenza, ad esse associata, presentano una distribuzione libera dall'influenza degli outliers, contrariamente a quanto riguarda il modello di regressione classico. La stima del quantile, infatti, è influenzata solamente dal comportamento locale della distribuzione condizionata della risposta vicino al quantile specificato,. Le stime della regressione quantile sono intrinsecamente robuste rispetto a contaminazioni della variabile risposta, tuttavia potrebbero essere piuttosto sensibili a contaminazioni delle variabili esplicative, $\{x_i\}$.

Inoltre, quando i termini di errore del modello di regressione quantile non sono normalmente distribuiti, gli stimatori forniti da tale regressione possono risultare più efficienti degli stimatori dei minimi quadrati.

Per quanto riguarda gli aspetti computazionali di tale modello, il problema di minimo da cui si ottengono le stime dei parametri può essere risolto impiegando metodi di programmazione lineare. Un'importante caratteristica di tale formulazione è la possibilità di calcolare efficientemente l'intero range di soluzioni per $\tau \in (0,1)$. Per ogni soluzione $\hat{\beta}(\tau_0)$ c'è un intervallo di τ su cui tale soluzione resta ottimale. È semplice calcolare il punto finale di tale intervallo, in modo da trovare iterativamente $\hat{\beta}(\tau)$ per l'intero campione, calcolando un pivot alla fine di ognuno di questi intervalli.

Infine,il vantaggio principale di tale modello è il fatto che, osservando diverse stime per i diversi quantili considerati, si può comprendere come varia l'influenza delle covariate sulla dipendente, nei vari punti della distribuzione quantile condizionata.

2.3 La regressione quantile in R

L'analisi computazionale dei dataset in analisi e l'applicazione della regressione quantile è solitamente condotta mediante R, un programma ampiamente utilizzato per ogni tipo di analisi in qualsiasi ambito d'interesse. Alcuni riferimenti bibliografici sono Koenker (2005), Dalgaard (2004), Venables e Ripley (2002). Inoltre, informazioni più specifiche si trovano nella funzione `help()` del programma.

I metodi di calcolo computazionale si sono ampiamente evoluti nel corso degli anni fino a diventare parte integrante della ricerca statistica. Beran (2003), infatti, definì la statistica come *"lo studio di algoritmi per l'analisi dei dati"*. Per quanto riguarda la regressione quantile, lo sviluppo di teorie e algoritmi di programmazione lineare ha reso i metodi di regressione quantile competitivi rispetto al metodo dei minimi quadrati ordinari, in termini di sforzo computazionale. Tali sviluppi dell'ambito computazionale hanno rivestito un ruolo di fondamentale importanza nell'incoraggiare un approccio più approfondito dei vantaggi statistici di tale metodo.

Per lavorare in R con la regressione quantile è necessario, per prima cosa, installare il pacchetto `quantreg` mediante il comando:

```
>install.packages("quantreg")
```

Una volta installato tale pacchetto, è necessario renderlo accessibile alla sessione corrente del programma per mezzo del comando

```
>library(quantreg)
```

La stima di un modello di regressione quantile è fornita dal comando `rq(.)`. Segue una breve descrizione di tale comando, il cui uso è

```
>fit<-rq(formula, tau=.5, data, subset, weights,
na.action,method="br", model = TRUE, contrasts, ...)
```

Il primo argomento `formula` indica il modello che si desidera specificare, ponendo la variabile risposta a sinistra dell'operatore ~ e le variabili esplicative a destra. L'opzione `tau` si riferisce al quantile che si desidera stimare, generalmente un valore strettamente compreso tra 0 e 1. L'oggetto `data` indica il dataset di riferimento, in cui interpretare le variabili specificate nella precedente formula; `subset`, indica un vettore opzionale che specifica un sottoinsieme di osservazioni da usare nel processo di stima; `weights` si riferisce al peso che si desidera assegnare alle osservazioni, tale valore deve essere positivo ed è utilizzato nella minimizzazione della somma dei pesi moltiplicato nei residui in valore assoluto. `na.action` è una funzione che serve a filtrare i dati mancanti, in alternativa è possibile usare `na.omit` che elimina le osservazioni contenenti uno o più valori mancanti. L'opzione `method` indica l'algoritmo di calcolo impiegato per calcolare la stima, il metodo impiegato di default è `"br"`. Tale metodo risulta efficiente per problemi fino ad alcune migliaia di osservazioni e può essere utilizzato per calcolare l'intero processo di regressione quantile. E' presente, inoltre, uno schema di calcolo per ottenere gli intervalli di confidenza dei parametri stimati, basato sull'inversione del test a ranghi di Koenker (1994). Per problemi con più di migliaia di osservazioni è utile impiegare l'opzione `method="fn"`, che usa l'algoritmo di Frisch-Newton descritto da Portnoy e Koenker. Forme particolari dell'algoritmo di Frisch-Newton sono disponibili per risolvere problemi che includono vincoli di disuguaglianza lineari e per problemi con matrici sparse di progettazione. Per questo tipo di problemi è possibile utilizzare l'opzione `method="pfn"`.

In definitiva, tale funzione calcola una stima della τ-esima funzione quantile condizionata della variabile risposta, date le variabili esplicative, come specificato dall'argomento della formula. Come per `lm()`, la funzione presuppone una specificazione lineare per il modello di regressione quantile, ovvero una formula che definisce un modello lineare nei parametri. La funzione di calcolo utilizzata opera minimizzando una somma ponderata dei residui in valore assoluto, il che può essere formulato come un problema di programmazione lineare. Come affermato sopra, ci sono diverse opzioni di calcolo per ottenere le stime d'interesse. È, tuttavia, altamente consigliato utilizzare il metodo di default nei casi in cui n <5000 e $\tau < 20$. Per quanto riguarda gli intervalli di confidenza e altre statistiche associate sono presenti innumerevoli metodi di stima e modalità di calcolo.

Se si vuole ottenere un riassunto del modello ipotizzato, composto dalle stime dei coefficienti e da alcune informazioni di base, il comando da utilizzare è composto solamente dal nome del modello, in questo caso

```
>fit
```

Se, invece, si desidera ottenere qualche informazione in più, il comando da utilizzare è

```
>summary(object, se = NULL, covariance=FALSE, hs = TRUE,
...)
```

Il termine `object` si riferisce ad un elemento di classe `rq(.)`, specificato in precedenza; `se`, invece, specifica il metodo da utilizzare per calcolare gli standard error. Attualmente il programma dispone di cinque metodi disponibili che impiegano diversi algoritmi di calcolo per ottenere gli standard error. In particolare, se il comando è composto da `se=NULL` e `covariance=NULL` viene utilizzato il metodo per ranghi. L'opzione `covariance` specifica se la matrice di varianza covarianza debba essere restituita o meno, mentre `hs` si riferisce ad alcuni dettagli riguardo la stima di sparsità.

Uno strumento grafico di notevole importanza per questo tipo di analisi è il diagramma di dispersione del modello ipotizzato. Tale strumento consente di osservare come sono distribuiti i quantili del modello ipotizzato.

Si riporta in seguito un esempio proveniente da uno studio di Engel (1857) relativo al rapporto tra la spesa alimentare di un campione composto da 235 famiglie, appartenenti alla classe operaia Belga del XIX secolo, e il rispettivo reddito. I comandi utilizzati per produrre un risultato grafico dell'applicazione della regressione quantile (si veda la Figura 2.3) ai dati sono i seguenti:

```
>plot(income,foodexp,cex=.25,type="n",xlab="Household       Income",
ylab="Food Expediture")
> points(income, foodexp, cex=.5,col="blue")
> abline(rq(foodexp~income,tau=.5),col="blue")
> abline(lm(foodexp~income),lty=2,col="red")
> taus<-c(.05,.1,.25,.75,.90,.95)
> for(i in
1:length(taus)){abline(rq(foodexp~income,tau=taus[i]),col="green")}
```

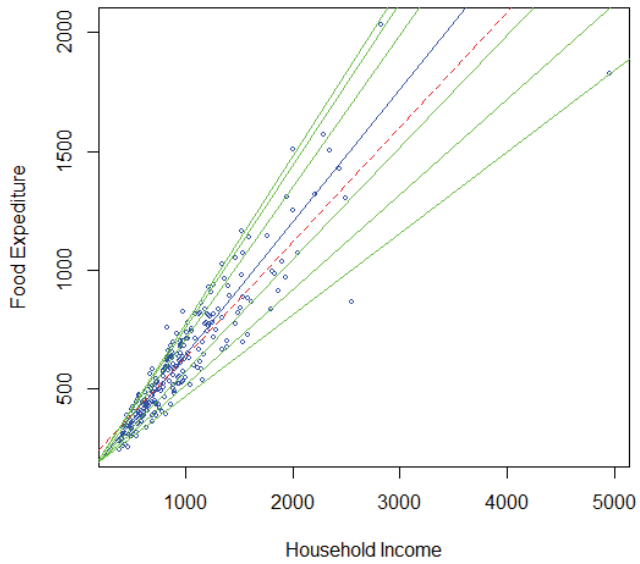

Figura 2.3: Diagramma di dispersione. I punti rappresentano le combinazioni tra variabile dipendente ed esplicativa, la linea rossa è ottenuta per mezzo della regressione lineare classica, mentre le linee verdi rappresentano le stime provenienti dalle regressione quantile, considerando vari livelli di τ {.05, .1, .25, .75, .90, .95}. In particolare la regressione mediana è rappresentata dalla linea blu.

Dal grafico è evidente come i dati si distribuiscano in gran parte nella parte in basso a sinistra del grafico, ciò sta a significare che a spese alimentari basse corrispondono redditi familiari altrettanto bassi. È presente, quindi, correlazione positiva tra variabile esplicativa e dipendente. Le rette stimate attraverso la regressione quantile comprendono quasi tutti i punti del grafico, ad eccezione di pochi. Tali rette sono disposte a raggiera in modo ordinato nel grafico. Esse risultano più vicine nella seconda metà della distribuzione, per il semplice fatto che in tale parte i punti risultano più concentrati. La retta di regressione lineare classica si trova al centro della distribuzione e si discosta di poco dalla mediana.

Se è d'interesse ottenere tutte le soluzioni distinte della regressione quantile per un particolare modello di regressione, è necessario specificare un livello di τ fuori dal range [0,1], ad esempio,

```
> z<-rq(foodexp~income,tau=-1)
```

Questa forma della funzione fornisce i passi di programmazione parametrica richiesti per trovare l'intero campione del processo di regressione quantile.

Un'altra funzione utile per analizzare il dataset è `akj`. Essa consente di associare le osservazioni ai valori prodotti dal processo di regressione quantile, fornendo una distribuzione di densità di tali valori, utilizzando una stima di densità univariata.

Il comando è composto dalle seguenti opzioni

```
>akj(x, z =, p =, h = -1, alpha = 0.5, kappa = 0.9, iker1
= 0)
```

`x` indica i punti da utilizzare per i centri di Kernel assunti da ordinare, `z`, invece, i punti in cui viene calcolata la densità, `p` è il vettore di probabilità, di default è $1/n$ per ogni x; `h` e `kappa` si riferiscono ad opzioni grafiche. `alpha` è un parametro riferito alla sensibilità; infine, `iker1` si riferisce all'indicatore di Kernel da impiegare.
Nella Figura 2.4 si riporta solamente il grafico ottenuto come output da tale funzione, tralasciando la specificazione dei comandi.

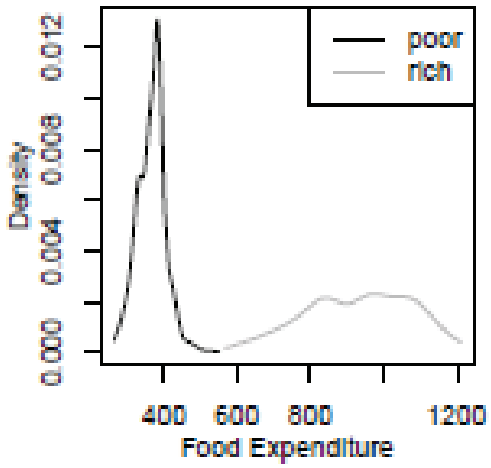

Figura 2.4: Funzione di densità stimata utilizzando due quantili della distribuzione, ottenuta per mezzo del comando `akj`. Tale funzione è stata applicata ai dati di Engel.

Per confrontare due modelli annidati a livello del medesimo quantile oppure per confrontare i coefficienti dello stesso modello stimati in quantili diversi è necessario impiegare la funzione `anova`. Essa si basa sul test di Wald e la matrice di covarianza asintotica è stimata utilizzando l'approccio di Hendricks e Koenker. Tale test è simile a quello impiegato per la regressione classica. L'ipotesi nulla è data dall'uguaglianza dei parametri dei modelli o dei quantili da confrontare, e fornisce, infatti, come output un test F e il corrispondente p-value in base al quale si accetta o meno tale ipotesi.

Le regressioni quantile non lineari possono essere stimate mediante l'opzione "`nlrq`"

```
>nlrq(formula, data=parent.frame(), start, tau=0.5,
      control, trace=FALSE,method="L-BFGS-B")
```

Tale funzione è composta da opzioni che consentono di stimare i parametri e la funzione obiettivo, ottenuta dalla soluzione migliore. Il comando è composto principalmente dai seguenti elementi: la `formula`, ovvero il modello utilizzato, la stima dei valori previsti dal modello, `tau`, ovvero il quantile utilizzato per produrre la stima ottenuta. Inoltre, tale metodo di sintesi utilizza un approccio bootstrap basato sulla linearizzazione finale del modello valutato dai parametri stimati.

Per quanto riguarda la regressione quantile non parametrica, il programma propone vari modi per affrontare questo problema. L'approccio più semplice impiega la funzione "`lprq`", cioè una funzione che mostra come funziona una regressione quantile locale polinomiale lisciata. Essa é ottenuta calcolando una stima della regressione quantile per un numero fissato di intervalli equispaziati sul supporto dei valori osservati sulle variabili esplicative. Si tratta di stimare le derivate di tale funzione e di fissare la larghezza delle bande.

Un altro modo per trattare regressioni quantile non parametriche è utilizzare regressioni spline. Si tratta di funzioni costituite da una serie di polinomi raccordati tra loro, il cui scopo è interpolare in un intervallo un insieme di punti, detti nodi, in modo da risultare continue, almeno fino ad un certo ordine di derivate, in ogni punto dell'intervallo. La funzione di riferimento è "`bs`" e si trova nel pacchetto `splines`. Questa procedura stima una funzione cubica polinomiale a tratti con 15 nodi (punti di interruzione nella terza derivata) fissata nei quantili. Un vantaggio di quest'approccio è dato dal fatto che è molto semplice aggiungere delle componenti al modello lineare.

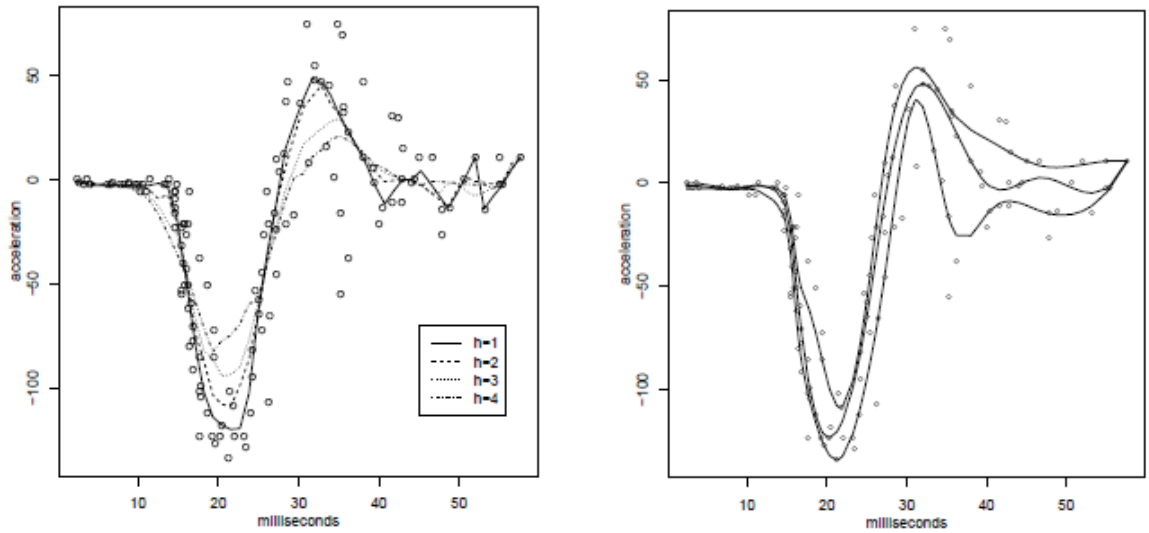

Figura 2.5: Mostrano rispettivamente l'applicazione grafica ottenuta in seguito ai comandi "`lprq`" (a sinistra) e "`bs`"(a destra).

La Figura 2.5 è stata ottenuta applicando i due tipi di regressione quantile non parametrica descritti precedentemente ad un campione di dati composto da misurazioni effettuate sul tempo di acceleramento di un insieme di motociclette.

Un terzo possibile approccio è costituito dal metodo penalty. Si tratta di una particolare classe di algoritmi impiegata per risolvere problemi di ottimizzazione vincolata. Tale metodo consiste nel sostituire un problema di ottimizzazione vincolata con una serie di problemi vincolati le cui soluzioni convergono idealmente alla soluzione del problema iniziale vincolato. I problemi non vincolati sono costruiti aggiungendo un termine alla funzione obiettivo, un parametro di sanzione e una misura di violazione dei vincoli, che è pari a zero nella regione in cui i vincoli non sono violati e viceversa. In R la funzione di riferimento è "`rqss`". Essa offre una formula di interfaccia alla regressione quantile non parametrica

```
>rqss(formula, tau = 0.5, data = parent.frame(), weights,
na.action, method = "sfn", lambda = NULL, contrasts =
NULL, ztol = 1e-5, control, ...)
```

Un altro vantaggio di tale approccio è costituito dalla semplicità nell'imporre vincoli qualitativi addizionali alle funzioni di stima; le funzioni univariate possono essere vincolate ad essere monotone e/o convesse o concave.

2.4 La regressione di Laplace con dati censurati

In anni recenti è stata rivolta attenzione crescente alla stima dei quantili di una variabile risposta dato un insieme di esplicative. Koenker e Bassett (1978) svilupparono la regressione quantile, i cui vantaggi sono: non ci sono assunzioni sulla distribuzione del termine d'errore del modello di regressione; l'inferenza è invariante rispetto a trasformazioni monotone della variabile risposta; in presenza di valori anomali, può essere più efficiente dei minimi quadrati; e, più importante, permette una precisa inferenza dell'intera forma della distribuzione e non solo della media. Koenker (2005) fornì un'eccellente e completa presentazione dell'argomento.

In questo paragrafo, si considera l'inferenza sui quantili di una variabile risposta condizionata ad un insieme di esplicative, anche in presenza di osservazioni censurate. Il meccanismo di censura è assunto casuale e possibilmente dipendente dalle esplicative. Si considera un approccio di massima verosimiglianza basato su un modello di regressione, dove è assunto che i residui seguano una distribuzione asimmetrica di Laplace. In tale contesto, Robins e Ritov (1997) sostenevano che i coefficienti della regressione quantile non possono essere stimati senza ulteriori condizioni sia sul tempo che sulla distribuzione della censura. La assunzioni per la regressione di Laplace sono diverse da quelle di tutti i metodi disponibili presentati nel paragrafo precedente e possono fornire alla regressione di Laplace alcuni vantaggi in termini di errore quadratico medio e di tempo di calcolo. Il riferimento bibliografico su cui si basa questo paragrafo è costituito da un articolo pubblicato da Bottai e Zhang (2010).

2.4.1 La regressione di Laplace

Siano $T_i, i = 1, \dots, n$ le variabili risposta indipendenti e siano x_i i vettori k-dimensionali contenenti i valori delle variabili esplicative. Si assuma che T_i possa essere censurata. Invece di T_i, quindi si osserva $Y_i = \min(T_i, C_i)$, dove le C_i possono dipendere da x_i ma, condizionatamente a x_i, sono indipendenti da T_i. Si assuma che C_i non contenga alcuna informazione riguardo i parametri d'interesse. Sia $\delta_i = I(T_i, C_i)$, dove $I(A)$ è la funzione indicatrice dell'insieme A.
Si assuma che esista un vettore k-dimensionale di parametri $\beta(p)$ tale per cui

$$T_i = x_i'\beta(p) + \varepsilon_i\,, \tag{2.16}$$

dove $p \in (0,1)$ è una probabilità data e le ε_i sono indipendenti e identicamente distribuite con quantile-p nullo, ossia $P(\varepsilon_i \leq 0|x_i) = p$. Il modello (2.16) è equivalente ad assumere che $x_i'\beta(p)$ è il quantile-p della distribuzione condizionata di T_i dato x_i, ossia $P\big(T_i \leq x_i'\beta(p)|x_i\big) = p$.

Si ricorda che una proprietà desiderabile del quantile condizionato $x_i'\beta(p)$ è l'invarianza rispetto a trasformazioni monotone (non decrescenti) h della variabile T_i, in quanto $P\big(T_i \leq x_i'\beta(p)\big|x_i\big) = P[h(T_i) \leq h\{x_i'\beta(p)\}|x_i]$. Per alcuni insiemi di dati è possibile scegliere la trasformazione h per cui la relazione tra il quantile condizionato e le esplicative possa essere facilmente modellata. Nel seguito, la risposta T_i indicherà la risposta stessa o una qualsiasi trasformazione monotona di essa.

2.4.2 Stima

Si assuma che, condizionatamente a x_i, T_i segua una distribuzione asimmetrica di Laplace con funzione di densità di probabilità

$$f(t_i, x_i) = exp\left[\{I\big(t_i \leq x_i'\beta(p)\big) - p\}\frac{t_i - x_i'\beta(p)}{\sigma(p)}\right]\frac{p(1-p)}{\sigma(p)}\,, \tag{2.17}$$

e funzione di distribuzione cumulata

$$F(t_i, x_i) = exp\left[\{I\big(t_i \leq x_i'\beta(p)\big) - p\}\frac{t_i - x_i'\beta(p)}{\sigma(p)}\right]\{p - I(t_i > x_i'\beta(p))\} + I\big(t_i > x_i'\beta(p)\big), \tag{2.18}$$

con $\beta(p) \in \mathbb{R}^k$ e $\sigma(p) \in (0, +\infty)$. Si veda Bottai e Zhang (2010).

Se $T_i \sim f(t_i|x_i)$, vale $P\{T_i \leq x_i'\beta(p)|x_i\} = p$. Inoltre, la Laplace nella (2.17) è una famiglia di distribuzioni di posizione e scala. In altre parole, se T_i segue una distribuzione asimmetrica di Laplace con parametri $\beta(p)$ e $\sigma(p)$, allora la variabile standardizzata $\{T_i - x_i'\beta(p)\}/\sigma(p)$ segue una distribuzione asimmetrica di Laplace con parametri 0 e 1. Queste caratteristiche hanno reso celebre

la distribuzione di Laplace e quindi adatta ad essere applicata all'inferenza sui quantili. Si veda a questo proposito Kozubowski e Nadarajah (2008).

In presenza di osservazioni censurate, Y_i è osservata al posto di T_i, e la funzione di log-verosimiglianza è proporzionale a

$$l_n\{\beta(p), \sigma(p)|y_i, x_i, \delta_i\} = \sum_{i=1}^{n}[\delta_i log f(y_i|x_i) + (1-\delta_i) log\{1 - F(y_i|x_i)\}], \quad (2.19)$$

dove le funzioni $f(t|x_i)$ e $F(t|x_i)$ sono definite dalla (2.17) e dalla (2.18). Si può mostrare che (si veda Bottai e Zhang, 2010)

$$\begin{aligned} l_n\{\beta(p), \sigma(p)|y_i, x_i, \delta_i\} = \sum_{i=1}^{n} \delta_i \left\{(w_i - p)\frac{y_i - x_i'\beta(p)}{\sigma(p)} + log\frac{p(1-p)}{\sigma(p)}\right\} \\ +(1-\delta_i) w_i log\left[1 - p\, exp\left\{(1-p)\frac{y_i - x_i'\beta(p)}{\sigma(p)}\right\}\right] \\ +(1-\delta_i)(1-w_i)\left\{\log(1-p) - p\frac{y_i - x_i'\beta(p)}{\sigma(p)}\right\}, \end{aligned}$$

dove $w_i = I\{y_i \leq x_i'\beta(p)\}$. Le stime di massima verosimiglianza per i parametri $\beta(p)$ e $\sigma(p)$, ossia $\hat{\beta}(p)$ e $\hat{\sigma}(p)$ sono ottenute massimizzando la (2.19).

Le derivate parziali prime della funzione di log-verosimiglianza per i parametri $\beta(p)$ e $\sigma(p)$ sono date, rispettivamente, da

$$S_n\{\beta(p)\} = \frac{1}{\sigma(p)}\sum_{i=1}^{n} x_i \left\{p - w_i - w_i(1-\delta_i)\frac{p-1}{1-F(y_i|x_i)}\right\} \quad (2.20)$$

$$S_n\{\sigma(p)\} = \frac{1}{\sigma(p)}\sum_{i=1}^{n} \left[\frac{y_i - x_i'\beta(p)}{\sigma(p)}\left\{p - w_i - w_i(1-\delta_i)\frac{p-1}{1-F(y_i|x_i)}\right\} - \delta_i\right]. \quad (2.21)$$

Le derivate prime in (2.20) e (2.21) sono non continue rispetto ai parametri nei punti dove la regressione residua è nulla, ossia quando $y_i = x_i'\beta(p)$.

La distribuzione asimmetrica di Laplace può essere estesa al caso in cui il parametro di scala $\sigma(p)$ dipende dalle covariate attraverso una funzione nota $\sigma(\eta, x_i, p)$ parametrizzata dal vettore η. Notiamo come, quando il parametro di scala dipende dalle esplicative, il primo addendo nella (2.21) sia stato moltiplicato per la costante $\partial\sigma(\eta, x_i, p)/\partial\eta$. Questo si applica a tutte le derivate rispetto al parametro di scala.

2.4.3 Inferenza

In generale, le equazioni di verosimiglianza $S_n\{\beta(p)\}$ e $S_n\{\sigma(p)\}$ non ammettono una soluzione in forma chiusa. La stima dei parametri $\beta(p)$ e $\sigma(p)$ può essere risolta utilizzando una procedura iterativa che: (i) valuta la probabilità condizionata $F(y_i|x_i)$ nelle stime correnti dei parametri, (ii) la utilizza per aggiornare le stime dei parametri attraverso la regressione quantile pesata. Si veda Portnoy (2003) e Wang e Wang (2009). Tale procedura viene ripetuta fino alla convergenza.
In alternativa, Bottai e Zhang (2010) suggeriscono di massimizzare direttamente la funzione di log-verosimiglianza. Questa funzione è non differenziabile ed è concava. Tra gli altri algoritmi possibili, quello proposto da Nelder e Mead (1965) è un semplice metodo che non richiede l'uso di derivate e che generalmente opera meglio con le funzioni concave. Il metodo è disponibile nel software R. L'inferenza sui parametri può essere ottenuta stimando con il metodo bootstrap le stime puntuali per ogni p-quantile d'interesse.

2.4.4 Un caso particolare

In questo e nel prossimo paragrafo, vengono presentate alcune caratteristiche della regressione di Laplace, che possono aiutare per capire meglio questo metodo. In questo paragrafo, si considera il caso particolare in cui tutte le n osservazioni, sono non censurate, ossia $d_i = 1$. In questo caso , le derivate (2.20) e (2.21) si semplificano come segue

$$S_n\{\beta(p)\} = \frac{1}{\sigma(p)} \sum_{i=1}^{n} x_i(p - w_i), \tag{2.22}$$

$$S_n\{\sigma(p)\} = \frac{1}{\sigma(p)} \sum_{i=1}^{n} \left[\frac{y_i - x_i'\beta(p)}{\sigma(p)}(p - w_i) - 1\right]. \tag{2.23}$$

L'equazione di stima $S_n\{\beta(p)\} = 0$ ottenuta dalla (2.22) è uguale a quella della regressione quantile tradizionale. È non distorta per $\beta(p)$ e funzionalmente indipendente da $\sigma(p)$, il cui valore può essere posto pari ad una costante senza alcuna perdita di generalità per la stima di $\beta(p)$.

L'equazione di stima $S_n\{\sigma(p)\} = 0$ dalla (2.23) impone il vincolo per cui la somma pesata dei residui $\sigma(p)^{-1}\{y_i - x_i'\beta(p)\}(p - w_i)$ sia uguale a n, la dimensione del campione. La stima di massima verosimiglianza $\hat{\sigma}(p) = n^{-1} \sum_{i=1}^{n}\{y_i - x_i'\beta(p)\}(p - w_i)$ è uguale alla media campionaria della regressione residua pesata.

2.4.5 Il caso generale

Si considera ora il caso generale in cui alcune osservazioni possono essere censurate. Il contributo alla funzione di stima (2.20) delle osservazioni censurate ai valori maggiori o uguali al quantile-p è pari a quello delle osservazioni non censurate nella (2.22). Il contributo delle osservazioni censurate ai valori inferiori al quantile-p è pari a

$$S_n\{\beta(p)\} = \frac{1}{\sigma(p)} \sum_{i:\delta_i=0 \cap w_i=1} x_i \left\{p - 1 - \frac{p-1}{1-F(y_i|x_i)}\right\}$$

$$= \frac{1}{\sigma(p)} \sum_{i:\delta_i=0 \cap w_i=1} x_i \left\{(p-1)\frac{p-F(y_i|x_i)}{1-F(y_i|x_i)} + p\frac{1-p}{1-F(y_i|x_i)}\right\}. \quad (2.24)$$

La (2.24) indica che il contributo delle osservazioni censurate ai valori più piccoli del quantile-p è uguale a $(p-1)$ con probabilità $\{p - F(y_i|x_i)\}/\{1 - F(y_i|x_i)\}$ e uguale a p con probabilità complementare. Se $F(y_i|x_i)$ è la vera distribuzione di T_i, allora $\frac{\{p-F(y_i|x_i)\}}{\{1-F(y_i|x_i)\}} = P\{T_i \leq x_i'\beta(p) | T_i > C_i\}$. La regressione di Laplace ricava informazioni sulla funzione di ripartizione $F(y_i|x_i)$ dalla funzione di densità $f(y_i|x_i)$, alla quale è funzionalmente collegata. Si noti che $F(y_i|x_i) = f(y_i|x_i)\,\sigma(p)/(1-p)$ quando $y_i = x_i'\beta(p)$.

2.4.6 Algoritmo di calcolo in R

La regressione di Laplace può essere stimata con l'algoritmo di Nelder-Mead implementato nella funzione `optim` di R. Per il metodo di Portnoy e Peng e Huang, si può utilizzare la funzione `crq` disponibile nella libreria `quantreg` di R. Per il metodo di Wang e Wang, si può usare la funzione `LCRQ` fornita dagli autori.

2.4.7 Un esempio

Per illustrare la regressione di Laplace, in Bottai e Zhang (2010) viene presentato uno studio sul cancro ai polmoni. I dati sono stati studiati da Maksymiuk (1994) e sono stati precedentemente analizzati per mezzo della regressione mediana da Ying (1995) e Zhou (2006).

Al momento dello studio, la terapia standard per i pazienti malati di questo tipo di cancro era costituita da una combinazione di etoposide e cisplatino. Tuttavia, la sequenza e la somministrazione ottimali non erano ancora state stabilite. Lo scopo dello studio era quello di esaminare due regimi: cisplatino seguito da etoposide (gruppo A) e etoposide seguita da cisplatino (gruppo B). Un

campione di 121 pazienti con cancro al polmone era stato assegnato casualmente ad uno dei due gruppi, 62 pazienti al gruppo A e 59 pazienti al gruppo B.

È stato considerato inizialmente il modello proporzionale di azzardo. Attraverso l'analisi dei residui di Schoenfeld sul modello stimato, tuttavia, l'assunzione di azzardo proporzionale tra i due gruppi trattati risultava discutibile (p=0.0472). Quest'assunzione è ancora impropria quando il trattamento costituisce l'unica esplicativa del modello (p=0.0284). Quindi, è stato considerato un modello di sopravvivenza alternativo, il modello con rischi accelerati (AFT), tale per cui

$$logT = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \tag{2.25}$$

dove T_i denota il tempo all'evento, $x_{1i} = 0$ i pazienti del gruppo A, $x_{1i} = 1$ per i pazienti del gruppo B , e x_{2i} rappresenta l'età del paziente al momento dell'entrata nello studio.

I coefficienti di regressione del modello semiparametrico dell'AFT sono stati stimati con il metodo di stima per ranghi con il peso di Gehan e sono state trovate le stime: $\hat{\beta}_1 =$ -0.185, con standard error pari a 0.065 e $\hat{\beta}_2 =$ -0.006 con standard error pari a 0.004. L'effetto del trattamento è statisticamente significativo nel modello semiparamentrico AFT, ma non lo è nel modello proporzionale di azzardo.

Si consideri ora la regressione di Laplace per il modello (2.25) per quattro quantili, ossia {0.25, 0.50, 0.75, e 0.90}. La Tabella 2.1 riassume le stime dei coefficienti e gli intervalli di confidenza al 95% per la regressione di Laplace e per il metodo di Wang e Wang. L'effetto del trattamento è significativo per i quantili 25% e 50%. Dopo l'aggiustamento per l'età al momento dell'entrata nello studio, si è stimato che il tempo mediano di sopravvivenza per il gruppo A è circa $10^{0,160} - 1 = 0.445$ volte più grande di quello del gruppo B.

Table 1 Estimates and 95% bootstrap confidence intervals (CI) for the regression coefficients associated with treatment regimen (β_1) and age at study entry (β_2) for four quantiles (0.25, 0.50, 0.75, 0.90) in the small cell lung cancer patients.

τ		Laplace regression		Wang and Wang	
		Estimate	95% CI	Estimate	95% CI
0.25	$\hat{\beta}_0$	2.874	(2.142, 3.605)	2.873	(2.512, 3.235)
	β_1	−0.202	(−0.377, −0.027)	−0.202	(−0.344, −0.061)
	β_2	−0.003	(−0.016, 0.009)	−0.003	(−0.010, 0.003)
0.50	β_0	3.040	(2.385, 3.695)	3.035	(2.448, 3.621)
	β_1	−0.160	(−0.299, −0.021)	−0.163	(−0.304, −0.022)
	β_2	−0.004	(−0.014, 0.006)	−0.004	(−0.013, 0.005)
0.75	β_0	3.996	(3.055, 4.937)	3.341	(2.575, 4.107)
	β_1	−0.170	(−0.393, 0.054)	−0.149	(−0.331, 0.032)
	β_2	−0.015	(−0.030, 0.001)	−0.005	(−0.018 0.007)
0.90	β_0	3.908	(1.402, 6.415)	3.647	(2.985, 4.309)
	β_1	−0.189	(−0.482, 0.105)	−0.136	(−0.303, 0.031)
	β_2	−0.010	(−0.047, 0.027)	−0.007	(−0.018, 0.003)

Tabella 2.1: Stime dei coefficienti e intervalli di confidenza al 95% per la regressione di Laplace e per il metodo di Wang e Wang (da Bottai e Zhang, 2010).

I quantili stimati $\hat{Q}_{T(p)} = 10^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2}$ sono rappresentati contro $1-p$, per p={0.01,…..,0.99}. La probabilità di sopravvivenza all'età di 65 anni è più alta nel gruppo A che nel gruppo B. Il tempo mediano di sopravvivenza è circa $10^{0,160} - 1 = 0.445$ volte più alto nel gruppo A che nel gruppo B, come riportato nella tabella 2.1.

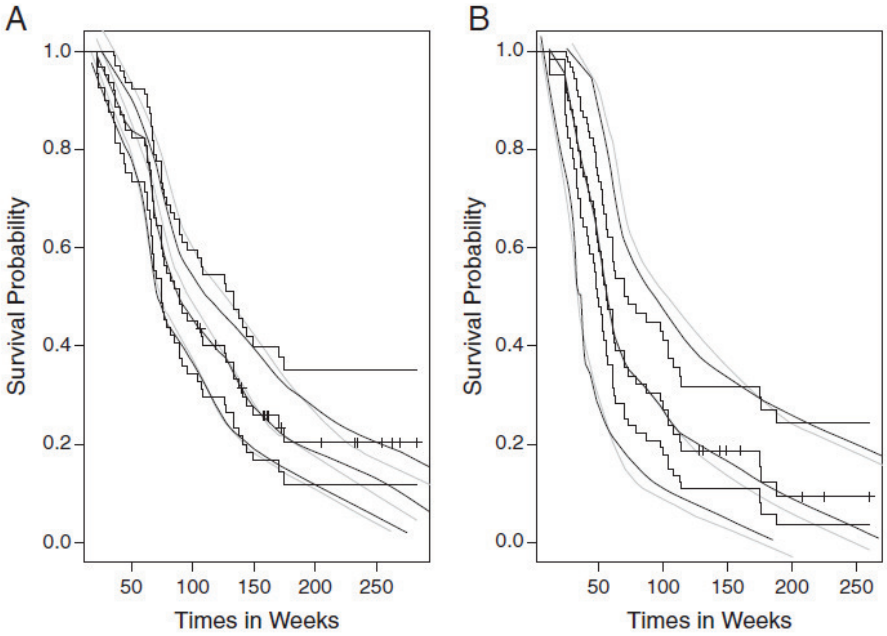

Figura 2.6: Curve di sopravvivenza levigate dei gruppi A e B rispettivamente. Le linee in nero rappresentano la regressione di Laplace, quelle in grigio il metodo di Wang e Wang, mentre le linee a gradini sono tracciate utilizzando la stima di Kaplan-Meier.

Dai grafici è evidente come le curve di sopravvivenza del gruppo A, stimate con entrambi i metodi utilizzati, si trovino in una posizione superiore rispetto a quelle del gruppo B. Ciò sta a significare che il trattamento somministrato al gruppo A è più efficace rispetto a quello somministrato al gruppo B e conferma i risultati trovati per mezzo della regressione. Le curve di sopravvivenza dei due gruppi, infatti, si allontanano una dall'altra dalla 30-esima settimana in poi, quindi la differenza tra i due tipi di trattamento si nota fin dai primi quantili osservati.

In sintesi, si conclude che l'uso di cisplatino seguito da etoposide è preferibile nel trattare il tumore al polmone, e la regressione di Laplace fornisce un'inferenza precisa circa le differenza tra i due regimi considerati.

2.5 Conclusioni

Il capitolo ha arricchito le nozioni già fornite nel Capitolo 1 sull'analisi dei dati di durata con la presentazione dell'argomento principale dell'elaborato, ossia la regressione quantile. Ad essa è stata affiancato un approfondimento legato sempre a tale materia, cioè la regressione di Laplace per dati censurati.

Nel prossimo capitolo viene analizzato un dataset proveniente da un campione di soggetti malati di leucemia e ad esso verranno applicate alcuni metodi di analisi della sopravvivenza presentati nel corso dell'elaborato.

Capitolo 3

Un caso di studio

In questo capitolo si discute un dataset su un gruppo di pazienti malati di leucemia. Lo scopo dello studio è capire se esiste una relazione tra il tempo di sopravvivenza e il numero di globuli bianchi presenti nel sangue dei pazienti appartenenti al campione in analisi. Per i rifermenti bibliografici si vedano Cox e Snell (1981) e Brazzale et al. (2007). Lo studio del campione sarà effettuato applicando due metodi presentati nel capitoli precedenti, ossia la regressione esponenziale e la regressione quantile.

3.1 Leucemia

Il termine leucemia indica un insieme di malattie maligne, ossia vari tipi di tumori caratterizzati dalla proliferazione neoplastica di una cellula staminale emopoietica. Col termine leucemia viene quindi comunemente indicato il tumore "del sangue". Le cellule che normalmente si trovano nel sangue, ossia (globuli banchi, globuli rossi e piastrine) prendono origine da cellule immature, dette cellule staminali o blasti, che si trovano nel modello osseo, ovvero in quella parte di tessuto spugnoso contenuto nelle ossa. Nelle persone affette da tale malattia vi è una proliferazione incontrollata di queste cellule, che interferisce con la crescita e lo sviluppo delle normali cellule del sangue.

Le leucemie vengono comunemente distinte in acute e croniche, sulla base della velocità di progressione della malattia. Nella leucemia acuta il numero di cellule tumorali aumenta più velocemente e la comparsa dei sintomi è precoce; nella leucemia cronica invece le cellule maligne tendono a proliferare più lentamente. Con il tempo, però, anche queste ultime diventano più aggressive e provocano un aumento delle cellule leucemiche all'interno del flusso sanguigno. Un'altra importante distinzione riguarda le cellule da cui prende origine il tumore. La cellula staminale, durante le varie fasi di maturazione, dà origine a cellule di tipo mieloide e cellule di tipo linfoide: da queste si differenzieranno successivamente i globuli rossi o eritrociti, le piastrine e i globuli bianchi (leucociti e linfociti). Pertanto avremo quattro tipi comuni di leucemia: la leucemia linfoblastica acuta (LLA), la leucemia linfocitica cronica (LLC), la leucemia mieloide acuta (LMA) e la leucemia mieloide cronica (LMC). Esistono poi altri tipi di leucemia più rari, come la leucemia a cellule capellute. La leucemia cronica può non dare sintomi nelle fasi iniziali perché le cellule leucemiche sono ancora in grado di svolgere il loro normale lavoro, non interferendo con le funzioni delle altre cellule. Invece, nella leucemia acuta i sintomi si presentano precocemente e possono peggiorare con rapidità. Le cellule leucemiche, al pari delle altre cellule presenti nel sangue, si spostano all'interno dell'organismo. Sulla base del loro numero e della loro localizzazione si avranno diverse manifestazioni quali, per esempio, la febbre, le sudorazioni notturne, la stanchezza e

l'affaticamento, il mal di testa, i dolori ossei e articolari, la perdita di peso, la suscettibilità alle infezioni, la facilità al sanguinamento oppure l'ingrossamento della milza e dei linfonodi, in modo particolare a livello del collo e delle ascelle. Talvolta la leucemia può coinvolgere anche lo stomaco, l'intestino, i reni e i polmoni. Tutti questi sintomi non sono sicuri segni di leucemia, perchè sono comuni a molte altre malattie; occorre quindi rivolgersi al medico per approfondire la natura di eventuali disturbi.

I tumori che colpiscono le cellule del sangue sono molto più frequenti nell'età infantile rispetto a quella adulta. Le leucemie acute, in particolare, rappresentano il 25 per cento di tutti i tumori dei bambini e si collocano quindi al primo posto. Tra le leucemie acute, la leucemia linfoblastica è il tipo più frequente nei bambini, ma può anche colpire gli adulti. Anche la leucemia mieloide acuta si può presentare sia in età infantile sia in età più avanzata. Le leucemie croniche sono invece più caratteristiche dell'età adulta. Solo il 2 per cento delle leucemie mieloidi croniche si manifesta sotto i 20 anni e la sua incidenza aumenta con l'età: infatti nella prima decade di vita dei bambini ha un'incidenza di circa 1 caso su 1.000.000, a 40 anni di 1 caso su 100.000 e all'età di 80 anni di 1 caso ogni 10.000. In Italia l'incidenza complessiva è di circa 15 nuovi casi per milione di persone all'anno. Gran parte delle leucemie che insorgono in età pediatrica dipendono da anomalie del DNA, sia a livello dei cromosomi, come accade nella leucemia mieloide cronica, sia a livello di singoli geni (un esempio è rappresentato dal gene p53). Inoltre alcune malattie, come la sindrome di Down, sono collegate a un rischio da 10 a 20 volte superiore di sviluppare una leucemia nei primi dieci anni di vita. Per quanto riguarda gli adulti, esiste un collegamento tra l'esposizione a dosi massicce di radiazioni e alcuni tipi di leucemia. Esiste inoltre un'associazione con l'esposizione a sostanze come il benzene e la formaldeide, utilizzate nell'industria chimica. Infine altri fattori di rischio che sono noti con certezza sono la chemioterapia e la radioterapia, effettuate in precedenza per curare altre forme tumorali.

Per quanto riguarda la diagnosi di tale malattia, la visita medica è molto importante per controllare se vi è un ingrossamento dei linfonodi, del fegato oppure della milza, e per scoprire eventuali emorragie. Gli esami del sangue, e in particolare l'emocromo, e gli indicatori del funzionamento di reni e fegato danno informazioni molto utili: nella leucemia aumenta il numero dei globuli bianchi e diminuiscono le piastrine e l'emoglobina che si trova all'interno dei globuli rossi. Per completare la diagnosi possono essere necessarie una biopsia ossea e una rachicentesi. A questi esami vanno associati infine una radiografia del torace e un'ecografia dell'addome. La gravità della leucemia dipende dallo stadio di malattia e quindi dall'estensione e dal coinvolgimento dei vari organi, nonché dalla risposta alla terapia medica. Vi sono leucemie che si presentano con un andamento meno aggressivo e altre, come quelle acute, che danno segno di sé più precocemente creando seri disturbi a chi ne è colpito.

La sopravvivenza a cinque anni nella leucemia linfatica supera il 63 per cento, e nella leucemia mieloide arriva al 26 per cento. In generale, la sopravvivenza a cinque anni per tutte le forme di leucemia si aggira intorno al 45 per cento nell'adulto, ma arriva ad oltre il 70 per cento nei bambini, e supera l'80 per cento nella leucemia mieloide infantile, la più comune. La terapia dipende dal tipo di

leucemia, dal suo stadio e dal fatto che la malattia sia in fase acuta o cronica. Importante è anche l'età al momento della diagnosi. Il trattamento delle leucemie si avvale dell'utilizzo di più terapie in combinazione o in sequenza, con lo scopo di ottenere una migliore qualità di vita e la guarigione. Vi sono poi le cosiddette terapie biologiche, che stimolano il sistema immunitario a riconoscere e a distruggere le cellule leucemiche. Negli ultimi anni si è sviluppato anche il trapianto di cellule staminali, che oggi è diventato lo standard terapeutico per le forme che non rispondono più alla chemioterapia. Questo trapianto permette al malato di ricevere dosi molto elevate di farmaci chemioterapici e di radiazioni, in grado di distruggere le cellule leucemiche che popolano il midollo osseo ma anche quelle sane. Le cellule staminali, ovvero progenitrici di tutte le altre, possono essere prelevate dal malato stesso e poi reinfuse dopo la chemio-radioterapia, oppure raccolte da un donatore compatibile (che può essere un fratello oppure uno sconosciuto).

3.2 Analisi del dataset

Il dataset in analisi è composto da due variabili, il tempo di sopravvivenza in settimane e una variabile relativa al numero di globuli bianchi nel sangue, osservate su diciassette pazienti. Per ulteriori approfondimenti circa la natura e composizione del dataset si vedano Cox e Snell (1981), oppure Brazzale et al. (2007).

Il dataset si presenta come segue

	time	x
1	65	3.36
2	156	2.88
3	100	3.63
4	134	3.41
5	16	3.78
6	108	4.02
7	121	4.00
8	4	4.23
9	39	3.73
10	143	3.85
11	56	3.97
12	26	4.51
13	22	4.54
14	1	5.00
15	1	5.00
16	5	4.72
17	65	5.00

La variabile `time` indica il tempo di sopravvivenza in settimane osservato sui pazienti appartenenti al campione, mentre la variabile `x` fornisce i valori provenienti da un indicatore del numero di globuli bianchi presenti nel sangue dei pazienti.

Nel seguito, per effettuare l'analisi è stata utilizzata una trasformazione della variabile `x`, ossia

```
> x1=x-mean(x)
```

Il dataset diventa quindi

```
> leucemia
   time          x1
1    65 -0.73588235
2   156 -1.21588235
3   100 -0.46588235
4   134 -0.68588235
5    16 -0.31588235
6   108 -0.07588235
7   121 -0.09588235
8     4  0.13411765
9    39 -0.36588235
10  143 -0.24588235
11   56 -0.12588235
12   26  0.41411765
13   22  0.44411765
14    1  0.90411765
15    1  0.90411765
16    5  0.62411765
17   65  0.90411765
```

Per prima cosa, è opportuno condurre un'analisi esplorativa dei dati. Media, mediana e quartili sono in grado di fornire indicazioni utili sulla forma della distribuzione. Forti scarti tra i valori di questi indici, infatti, possono indicare uno sbilanciamento eccessivo della distribuzione verso destra o verso sinistra.

Nel seguito si riportano le principali statistiche di sintesi riferite alle variabili presenti nel dataset.

```
> summary(time)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.00   16.00   56.00   62.47  108.00  156.00
```

```
> sd(time)
[1] 54.3531
```

```
> summary(x1)
 Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.21600 -0.36590 -0.09588  0.00000  0.44410  0.90410
```

```
> sd(x1)
[1] 0.6255404
```

Nella Figura 3.1 sono riportati i diagrammi a scatola con baffi relativi alle variabili `time` e `x1`. È d'interesse, per il tipo di analisi che si intende svolgere, avere una rappresentazione schematica dei quartili riportati nel summary. Tale strumento grafico consente, inoltre, di valutare le caratteristiche della distribuzione da cui provengono le osservazioni.

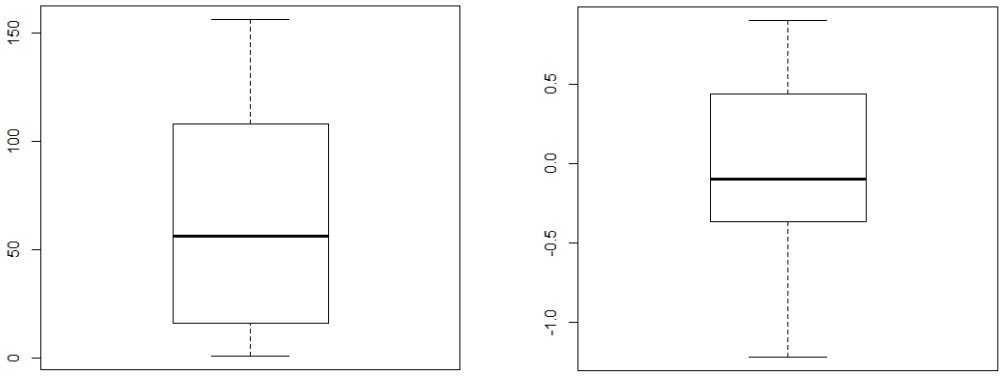

Figura 3.1: Boxplot di `time` (a sinistra) e `x1`(a destra).

Dalle statistiche di sintesi e dal boxplot, si può notare che per la variabile `time` la coda di sinistra della distribuzione è leggermente più corta rispetto a quella di destra. Tuttavia, media e mediana differiscono di poco, il che conferma l'assenza di asimmetria. Lo standard error è piuttosto elevato rispetto al campo di variazione della variabile, ciò significa che la distribuzione non è molto concentrata attorno al valore medio.

Per quanto riguarda la variabile `x1`, la distribuzione empirica non presenta asimmetrie. Media e mediana differiscono di poco e lo standard error risulta abbastanza elevato anche in questo caso rispetto al campo di variazione di `x1`.

Nella Figura 3.2 è riportato il diagramma di dispersione relativo al campione. Il grafico rappresenta le coppie di punti su un piano euclideo e consente di stabilire la presenza e il tipo di correlazione tra le variabili rappresentate.

La correlazione vale

```
> cor(x1,time)
[1] -0.6848642
```

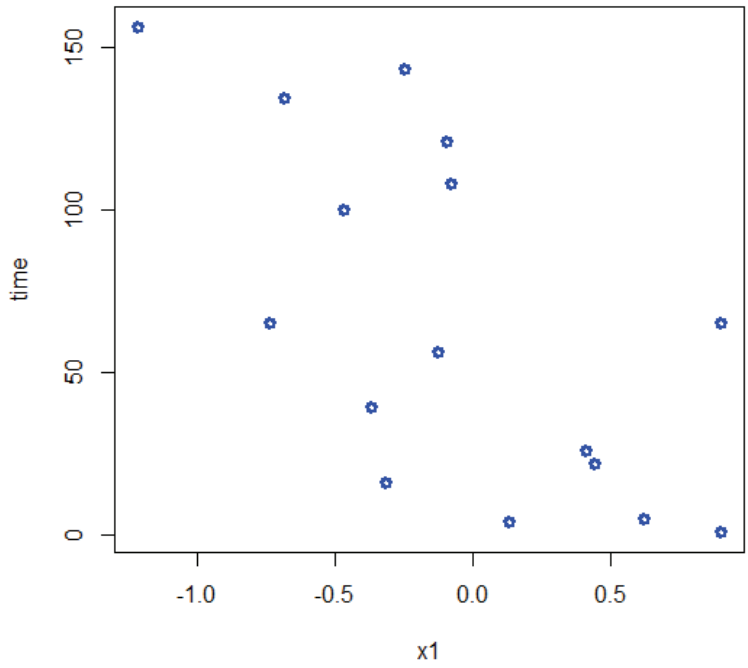

Figura 3.2: Diagramma di dispersione.

Dalla Figura 3.2 si nota come a tempi di sopravvivenza minori corrispondano valori della variabile `x1` maggiori. Si può ipotizzare, quindi, che i pazienti per i quali si è osservato il decesso in tempi brevi avessero un numero di globuli bianchi nel sangue più elevato rispetto ai pazienti per i quali il tempo di sopravvivenza è stato maggiore. Il valore assunto dal coefficiente di correlazione indica che c'è dipendenza tra variabile dipendente ed esplicativa. La negatività del coefficiente sta a significare che la variabile `time` varia negativamente all'aumentare di `x1`.

Lo scopo dell'analisi è capire se e come il livello dell'indicatore del numero di globuli banchi influisce sul tempo di sopravvivenza dei pazienti. Il dataset non contiene dati censurati: i valori della variabile `time` sono, quindi, relativi alle settimane di vita dei pazienti fino al momento del decesso.

Per svolgere l'analisi su questo campione si considera inizialmente un approccio parametrico. Nell'ambito dell'analisi della sopravvivenza, le distribuzioni maggiormente impiegate sono l'esponenziale e la Weibull (si veda il paragrafo 1.3.2). La distribuzione di Weibull è un'estensione della distribuzione esponenziale; infatti per k=1, con k parametro di forma, le due distribuzioni coincidono.

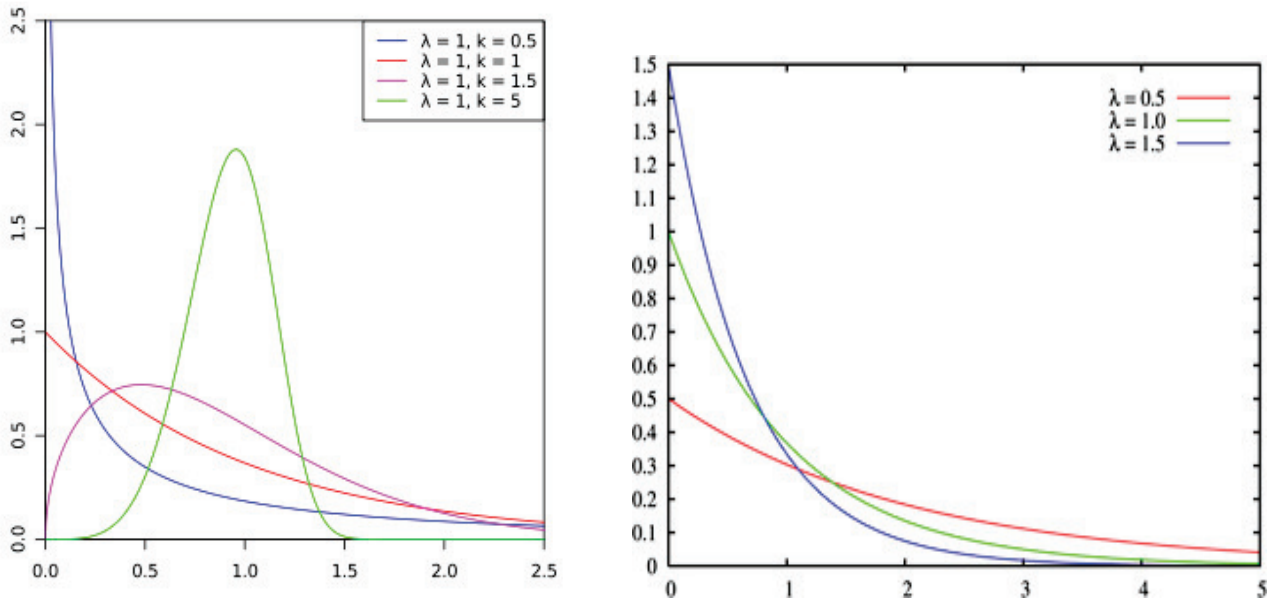

Figura 3.3 : Distribuzioni di Weibull (a sinistra) ed esponenziale (a destra), rispettivamente.

Per quanto riguarda la distribuzione esponenziale, essa ha funzione di probabilità f_T (t;λ) =$\lambda e^{-\lambda t}$, con $\lambda > 0$, e la funzione azzardo è costante, ossia $\lambda_T(t;\lambda) = \lambda$. Un modello di regressione esponenziale esprime la dipendenza di λ_i dai p valori esplicativi. Poiché λ_i può assumere solamente valori postivi, è conveniente specificare un modello della forma logλ_i= $\beta_1 x_{i1}$+….+$\beta_p x_{ip}$, dove β_1,….,β_p sono parametri reali, comuni a tutte le osservazioni. Il modello ipotizzato sul dataset è

$$\lambda_i = \lambda_0 e^{\beta_2 x_{i2}},\ i = 1, \dots, 17 \qquad (3.1)$$

dove x_{i2} indica la variabile esplicativa `x1`.

La stima del modello in R è la seguente

```
> fit1=survreg(time~x1,dist="exponential")

> summary(fit1)

Call:

survreg(formula = time ~ x1, dist = "exponential")

             Value Std. Error     z        p

(Intercept)  3.93      0.243 16.22 3.64e-59

x1          -1.11      0.414 -2.68 7.31e-03

Scale fixed at 1

Exponential distribution

Loglik(model)= -83.9   Loglik(intercept only)= -87.3

        Chisq= 6.83 on 1 degrees of freedom, p= 0.009

Number of Newton-Raphson Iterations: 5

n= 17
```

Entrambi i coefficienti stimati risultano significativi contro l'ipotesi nulla ($H_o: \beta_j = 0$). Inoltre, il valore del `Chisq= 6.83` (e il rispettivo `p-value= 0.009`) fornisce il valore del test del log-rapporto di verosimiglianza per il confronto tra il modello con la sola intercetta e il modello corrente. L'ipotesi nulla di uguaglianza dei due modelli, ossia quello con la sola intercetta e il modello completo, viene rifiutata per qualsiasi livello di α usuale.

I parametri stimati dal modello (`fit1`) risultano

$\hat{\lambda}_0 = \exp(-(3.93)) = 0.01964367$

$\hat{\beta}_2 = \exp\big(-(-1.11)\big) = 3.034358$

Il modello stimato risulta, quindi $log\hat{\lambda}_i = 0.01964367 + 3.034358x_i$.

Per quanto riguarda la distribuzione di Weibull, essa ha come funzione di probabilità $f_T(t;\lambda,k) = k\lambda^k t^{k-1} e^{-(\lambda t)^k}$, con $\lambda > 0$ e $p > 0$, e funzione di azzardo $\lambda_T(t;\lambda,p) = k\lambda(\lambda t)^{k-1}$.

Il modello di regressione di Weibull prevede che

$\lambda_i = \lambda k(\lambda t)^{k-1} e^{\beta_1 x_{i1} + \cdots + \beta_k x_{ip}} = \lambda_0(t) e^{\beta_2 x_{i2} + \cdots + \beta_p x_{ip}}$.

Il parametro k definisce la forma della funzione azzardo per ogni pattern di covariate. In generale, qualunque sia $\lambda_0(t)$, un modello di regressione di Weibull ha rapporto tra azzardi (HR) pari a:

$HR = \frac{\lambda_T(t;x_i^1)}{\lambda_T(t;x_i^2)} = e^{\beta_1(x_{i1}^1 - x_{i1}^2) + \cdots + \beta_p(x_{ip}^1 - x_{ip}^2)}$, ossia costante nel tempo, con x^1 e x^2 vettori di covariate di due soggetti.

Il modello ipotizzato per il dataset è

$$\lambda_i = \lambda_0(t) e^{\beta_2 x_{i2}}, \quad i = 1, \ldots, 17 \tag{3.2}$$

dove x_{i2} indica la variabile esplicativa `x1`.

La stima del modello in R è la seguente

```
> fit2=survreg(time~x1,dist="weibull")
> summary(fit2)
Call:
survreg(formula = time ~ x1, dist = "weibull")
              Value Std. Error      z        p
(Intercept)  3.9425      0.251 15.736 8.51e-56
x1          -1.0982      0.418 -2.630 8.53e-03
Log(scale)  -0.0216      0.202 -0.107 9.15e-01
Scale= 0.979
Weibull distribution
Loglik(model)= -83.9   Loglik(intercept only)= -87.1
        Chisq= 6.48 on 1 degrees of freedom, p= 0.011
Number of Newton-Raphson Iterations: 6
```

n= 17

Per quanto riguarda i coefficienti, l'intercetta e il valore relativo alla variabile esplicativa x1 risultano significativi contro l'ipotesi nulla ($H_o: \beta_j = 0$). Il terzo coefficiente, invece, è relativo al logaritmo del parametro di scala, proprio della distribuzione di Weibull, e non risulta significativo contro l'ipotesi nulla. Ciò sta a significare che se $p = log(k) = 0$, k è pari ad 1, quindi, la distribuzione di Weibull si riconduce al caso in cui coincide con la distribuzione esponenziale.

Sono riportate in seguito le stime dei coefficienti del modello di Weibull:

$\hat{\lambda} = \exp(-3.9425) = 0.01939965$

$\hat{\beta}_2 = -(-1.0982)/\ 0.979 = 1.121757$

$\hat{p} = -\left(-\frac{0.0216}{0.979}\right) = 0.02206333$

Il modello stimato risulta, quindi $\lambda_i = 0.0194 * 0.022(0.0194t)^{0.022-1} e^{1.22x_{i2}}$.

Il rapporto tra azzardi è costante nel tempo, poiché dipende solamente dall'effetto della variabile esplicativa che risulta essere moltiplicativo, ed è pari a

HR= exp(1.211757)= 3.359382

Il valore del test del log-rapporto di verosimiglianza, ossia Chisq= 6.48 (e il rispettivo p-value= 0.011), porta ad accettare l'ipotesi nulla di uguaglianza del modello completo al modello con la sola intercetta ad un livello di α pari all'1%. Tuttavia al 5% il test rifiuta l'ipotesi nulla.

Tra il modello esponenziale e il modello di Weibull è più conveniente considerare il modello esponenziale sostanzialmente perché, a parità di efficacia e applicabilità, è più semplice per quanto riguarda l'interpretazione dei dati.

Si riporta nella Figura 3.4 il grafico della distribuzione esponenziale adattata ai dati del campione in analisi.

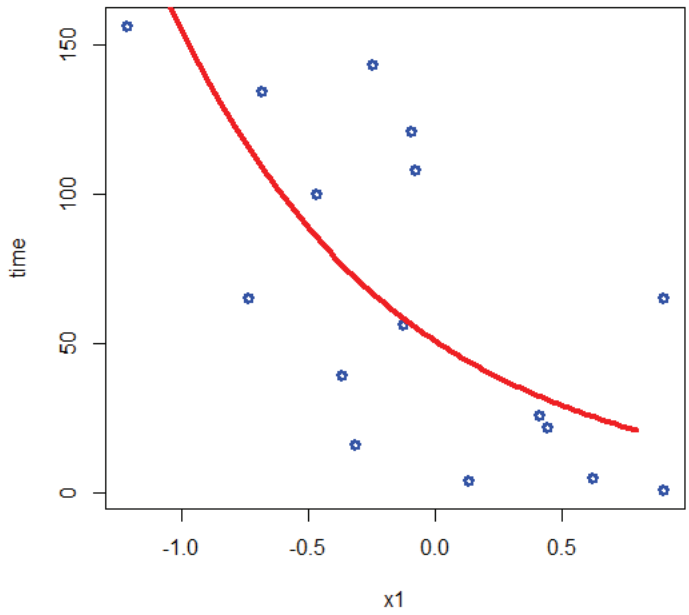

Figura 3.4: Grafico del modello esponenziale stimato.

Oltre ai metodi applicati fin'ora, è possibile impiegare anche la tecnica di analisi descritta nel Capitolo 2, ovvero la regressione quantile. Il grande vantaggio della regressione quantile è la possibilità di stimare l'intera distribuzione dei quantili condizionati della variabile risposta. In questo modo si può capire come la variabile esplicativa `x1` influisce sulla sopravvivenza dei pazienti e, cosa più importante, se a valori elevati di tale indicatore corrispondono tempi bassi di sopravvivenza.

Nella Tabella 3.1 sono riportate le stime dei coefficienti i regressione ottenute dalle regressioni quantile per i quantili di livello: {0.05, 0.1, 0.25, 0.5, 0.75, 0.90, 0.95}.

modello	quantile		coefficienti
Fit 3	0.05	intercetta	7.576
		x1	-26.667
Fit 4	0.10	intercetta	12.116
		x1	-12.295
Fit 5	0.25	intercetta	36.283
		x1	-39.024
Fit 6	0.50	intercetta	66.334
		x1	-72.263
Fit 7	0.75	intercetta	104.235
		x1	-43.396
Fit 8	0.90	intercetta	115.631
		x1	-56.000
Fit 9	0.95	intercetta	115.631
		x1	-67.826

Nella Figura 3.5 sono riportate le rette di regressione stimate con la regressione quantile. Per chiarezza interpretativa queste sono state rappresentate con vari colori, attraverso i comandi

```
> plot(x1,time)
> abline(rq(time~x1,tau=.5),col="blue")
> abline(rq(time~x1,tau=.05),col="red")
> abline(rq(time~x1,tau=.1),col="green")
> abline(rq(time~x1,tau=.25),col="gray")
> abline(rq(time~x1,tau=.75),col="yellow")
> abline(rq(time~x1,tau=.90),col="violet")
> abline(rq(time~x1,tau=.95),col="brown")
```

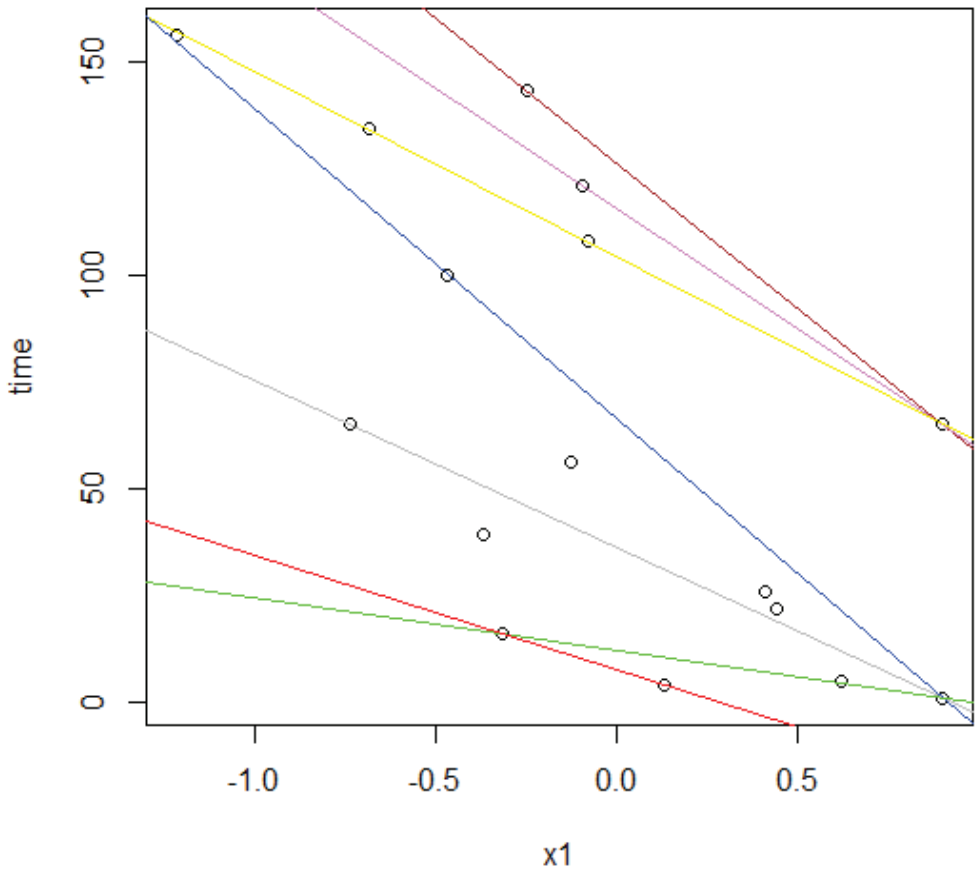

Figura 3.5: Rette stimate dalla regressione quantile.

L'output del comando `rq(.)` fornisce come risultato solamente una stima dei coefficienti del modello (ed eventualmente le relative bande di significatività). Si nota come il valore del coefficiente relativo all'intercetta aumenti man mano che si considerano quantili di livello sempre più alto e assume sempre valori positivi per tutti i quantili considerati. Questo perché la regressione attribuisce il tempo di sopravvivenza ai quantili in modo crescente. Non accade lo stesso per il coefficiente relativo alla variabile esplicativa `x1`. I valori assunti da esso, infatti, risultano essere piuttosto variabili. Tuttavia, sono accomunati dal fatto di essere tutti negativi. Ciò sta a significare che la variabile esplicativa ha un effetto negativo sul tempo di sopravvivenza dei pazienti. Tali coefficienti stimano l'intercetta e il coefficiente angolare delle rette rappresentate nella Figura 3.5. Valori variabili e non ordinati dei coefficienti angolari delle rette fanno si che esse non presentino un andamento a raggiera ordinato secondo i quantili. Come è visibile dal grafico, infatti, le rette stimate arrivano addirittura ad incrociarsi in alcuni punti, il che può essere attribuito alla scarsa numerosità campionaria. Il grafico, inoltre, conferma l'ipotesi formulata precedentemente: è evidente, infatti, come a valori più elevati della variabile esplicativa, corrispondano tempi di sopravvivenza generalmente più bassi. Ad esempio, in corrispondenza del quantile {0.05} sono situate le seguenti coppe (`time,x1`) di valori : (1, 0.90411765), (1, 0.90411765), (5, 0.62411765). Ciò è visibile nel grafico, la linea a cui si fa riferimento è quella di colore verde. I tempi di sopravvivenza per questi tre pazienti sono molto bassi e presentano valori della variabile esplicativa tra i più alti. Le altre coppie di punti che si trovano sotto la mediana, cioè nella prima metà della distribuzione presentano valori di `x1` che oscillano circa tra (0.5 e -0.5) e presentano valori relativi al tempo di sopravvivenza sotto le 65 settimane. Per quanto riguarda, invece, la seconda metà della distribuzione, i punti presentano tempi di sopravvivenza elevati, ossia superiori a 65, in relazione a valori della variabile esplicativa compresi tra (0 e -1.3). L'unica eccezione è costituita dal punto (65, 0.90411765); infatti, ad un tempo di sopravvivenza medio, rispetto ai valori presentati dal campione, corrisponde un valore della variabile esplicativa alto, pari circa a quello assunto dai pazienti appartenenti ai primi quantili della distribuzione, per i quali il decesso è avvenuto in poche settimane. Probabilmente, il punto in questione può essere considerato un outlier. Si può ipotizzare che tale paziente abbia contratto una forma particolare di leucemia, diversa da quella contratta dagli altri pazienti appartenenti al campione. Nella parte centrale della distribuzione si nota che a valori della variabile esplicativa compresi tra (0 e -0.5) corrispondono valori della variabile dipendente compresi tra (40 e 150). Un intervallo centrale così ampio suggerisce che per valori non estremi della variabile esplicativa possono presentarsi varie situazioni. Alcuni pazienti, infatti, sono sopravvissuti molto, altri invece sono deceduti in tempi brevi. Non avendo a disposizione un valore di riferimento con cui confrontare le osservazioni relative ad `x1`, non si è in grado di stabilire se tali valori siano nella norma. Probabilmente il problema è costituito anche in questo caso dalla bassa numerosità campionaria che non consente una stratificazione più precisa del campione. Un altro problema è il fatto che i valori più alti assunti dalla variabile `time` e corrispondenti a valori bassi di `x1`, come ad esempio le coppie (156, -1.21588235) e (134, -0.68588235) sono situate a livello del quantile {0.75}

invece che dei quantili più alti. Probabilmente, ciò è dovuto alla scarsità di osservazioni appartenenti al campione. Ciò costituisce un problema a livello di stima dei coefficienti delle rette. Si conclude, quindi, che le regressioni ottenute possono essere considerate soddisfacenti data la bassa numerosità del campione con cui si sta operando.

Per confrontare i due metodi di studio affrontati in questo capitolo, viene riportato in seguito un grafico in cui compaiono entrambi (si veda la Figura 3.6).

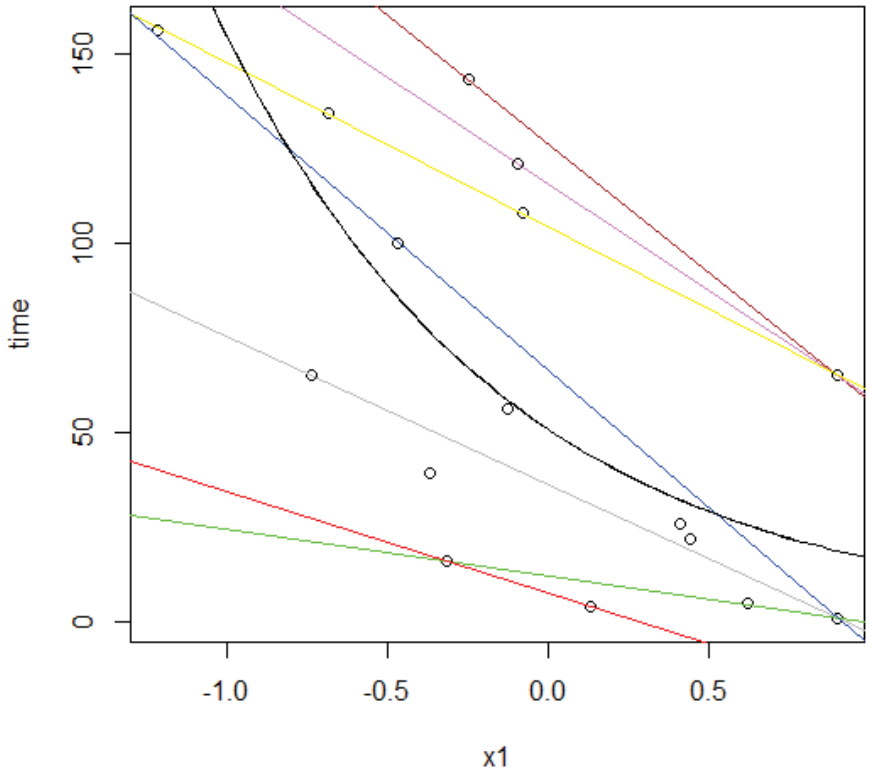

Figura 3.6: Grafico di confronto tra regressione quantile ed esponenziale.

La curva esponenziale è situata circa allo stesso livello della retta di regressione mediana. La regressione esponenziale, infatti, ha stimato la dipendenza del tasso di guasto dai valori assunti dalle variabili esplicative, dando come risultato due parametri per l'intera distribuzione. Tali parametri sono calcolati in media, è per questo motivo che la curva, pur mantenendo il naturale andamento della distribuzione a cui appartiene, è situata attorno alla mediana. La regressione quantile, invece, fornisce una stima dei parametri per tutti i quantili che si desidera stimare. Ha, quindi, il vantaggio di dare un'idea di come si comporta la variabile risposta rispetto ai possibili valori che le variabili esplicative possono assumere. Un altro vantaggio della regressione quantile è costituito dalla robustezza che esse presenta di fronte a valori anomali presenti all'interno della distribuzione. Ciò costituisce una caratteristica importante e un vantaggio rispetto alla regressione lineare classica. Quest'ultima, infatti, non è robusta in presenza di valori anomali, le stime della regressione, quindi,

potrebbero risultare distorte per colpa di questi valori. Dal grafico si nota come il valore (65, 0.90411765), anomalo rispetto agli altri valori della distribuzione, non influenzi tutte le rette stimate con la regressione quantile, dato che la stima dei quantili è influenzata solamente dal comportamento locale della distribuzione condizionata della risposta vicino al quantile specificato.

3.3 Conclusioni

Entrambe le regressioni condotte sul dataset permettono di giungere alle stesse conclusioni: il numero di globuli bianchi nel sangue dei pazienti malati di leucemia può essere considerato un indicatore significativo per quanto riguarda il tempo di sopravvivenza. In particolare, si è osservato che valori elevati di tale indicatore sono strettamente connessi a tempi di sopravvivenza bassi. Tale osservazione, derivante dal presente studio, è confermata dalla letteratura medica relativa alla sindrome leucemica.

Per quanto riguarda le due tecniche di analisi condotte, è stato verificato come entrambe siano ugualmente efficaci. L'unica differenza è costituita dal fatto che la regressione quantile, a differenza della regressione esponenziale, fornisce stime del modello di regressione per qualsiasi quantile richiesto. Ciò è molto utile poiché consente di osservare come varia la variabile dipendente in relazione alle variazioni dell'esplicativa. Nel caso considerato, quindi, il modello di regressione quantile è preferibile. A livello interpretativo, infatti, è di notevole importanza conoscere come si dispongono le rette di regressione stimate a livello dei diversi quantili considerati. Ciò fornisce una visione generale relativa alle caratteristiche e alla forma della funzione di distribuzione del dataset. Tuttavia, la scarsa numerosità campionaria ha generato qualche problema. Le rette stimate a partire dai pochi valori appartenenti al dataset non sono disposte in modo ordinato rispetto al livello dei quantili. In alcuni casi, infatti, le rette si incrociano. Ciò suggerisce che, se il campione fosse stato composto da un numero maggiore di osservazioni, si sarebbero ottenute stime e di conseguenza rette migliori e sarebbe stato possibile formulare un'interpretazione più soddisfacente dei risultati ottenuti.

Bibliografia

Beran, R. (2003). Impact of the Bootstrap on Statistical Algorithms and Theory. *Statistical Science*, 175-184.

Bottai, M. e Zhang, J. (2010). Laplace regression with censored data. *Biometrical Journal*, **52**, 487-503.

Brazzale, A.R., Davison, A.C. e Reid, N. (2007). *Applied Asymptotics: Case Studies in Small-Sample Statistics*. Cambridge University Press, Cambridge.

Broccoli, S., Cavrini, G. e Zoli, M. (2005). Il modello di regressione quantile nell'analisi delle determinanti della qualità della vita in una popolazione anziana. *STATISTICA,* **4**, 419-437.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, Series B (Methodological), **34**, 187-220.

Cox, D.R. e Snell, E.J. (1981). *Applied statistics: principles and examples*. Chapman & Hall, London.

Dalgaard, P. (2004). *Introductory statistics with R*, (3rd edition). Springer, New York.

Docksum, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two sample case. *Annals of Statistics*, **2**, 267-77.

Engel, E. (1857). Die Production und Consumtionsverhältnisse des Königreichs Sachsen. *Zeitschrift des Statistischen Bureaus des Königlich Sächsischen Ministeriums des Innern*, **8**, 1-54.

Klein, J. e Moeschberger, M.L. (2003). *Survival Analysis. Techniques for Censored and Truncated Data*. Springer, New York.

Koenker, R. (2005). *Quantile Regression in R: a vignette*. Cambridge University Press, New York.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, New York.

Koenker, R. e Basset, G. (1978). Quantile regression. *Econometria*, **46**, 33-50.

Koenker, R. e Hallock, K.F. (2001). Quantile Regression. *Journal of Economic Perspectives*, **15**, 143-156.

Koenker, R. e Portnoy, S. (1994). Quantile Smoothing Splines. *Biometrika*, **81**, 672-80.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. Wiley, New York.

Lehmann, E. (1974). *Nonparametrics: Statistical methods Based on Ranks.* Holden-Day, San Francisco.

Maksymiuk, A., Jett, J., Earle, J. D., Su, J., Diegert, F., Mailliard, J., Kardinal, C., Krook, J., Veeder, M. H. e Wiesenfeld, M. (1994). Sequencing and schedule effects of cisplatin plus etoposide in small-cell lung cancer: results of a north central cancer treatment group randomized clinical trial. *Journal of Clinical Oncology,* **12**, 70–76.

Marubini, E. e Valsecchi, M.G. (1995). *Analysing survival data from clinical trials and observational studies*. Wiley, Chichester.

Nelder, J.A. e Mead, R. (1965). A simplex method for function minimization. *Computer Journal,* **7**, 308–313.

Pace, L. e Salvan, A. (2001). *Introduzione alla statistica –II - Inferenza, verosimiglianza, modelli.* CEDAM, Padova.

Portnoy, S. (2003). Censored Regression Quantiles. *Journal of the American Statistical Association*, **98**, 1001-1012.

Powell, J.L. (1986). Censored regression quantiles. *Journal of Econometrics*, **32**, 143-55.

Robins, J. M. e Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine,* **16**, 285–319.

Venables, W.N. e Ripley, B.D. (2002). *Modern applied statistics with S* (4th edition). Springer, New York.

Wang, H. J. e Wang, L. (2009). Locally weighted censored quantile regression. *Journal of the American Statistical Association*, **1046**, 1117–1128.

Ying, Z., Jung, S. e Wei, L. (1995). Survival analysis with median regression models. *Journal of the American Statistical Association,* **90**, 178–184.

Zhou, L. (2006). A simple censored median regression estimator. *Statistica Sinica,* **16**, 1043–1058.