*STATISTICA, anno LXII, n. 3, 2002*

# REGRESSIONE PER FUNZIONI DI RISCHIO CON COVARIATE TEMPO-DIPENDENTI: UNA PROPOSTA BASATA SULLA MASSIMA VEROSIMIGLIANZA LOCALE

Giuliano Galimberti

## 1. INTRODUZIONE

Il metodo della massima verosimiglianza costituisce un utile strumento per la soluzione di numerosi problemi di inferenza statistica. Uno dei principali difetti di questo metodo è tuttavia rappresentato dalla necessità di dover specificare un modello statistico sul quale basare la costruzione della funzione di verosimiglianza. Questa necessità risulta particolarmente penalizzante qualora l'obiettivo dell'analisi statistica sia di tipo esplorativo. Per ovviare a questo inconveniente sono state presentate in letteratura numerose proposte, soprattutto nell'ambito della stima di densità e dei modelli di regressione lineare. Tra queste, risultano essere particolarmente interessanti, sia dal punto di vista applicativo che da quello metodologico, quelle basate sui modelli polinomiali locali e, più in generale, sulla massima verosimiglianza locale (Fan *et al.*, 1998).

Lo scopo di questo lavoro è quello di estendere le soluzioni proposte dai metodi basati sulla verosimiglianza locale anche all'ambito dello studio delle relazioni esistenti tra la funzione di rischio condizionato associata ad una variabile di durata $T^*$ e una variabile tempo-dipendente $X(t)$ che assume valori nell'insieme $\aleph$.

L'idea di base consiste nell'approssimare localmente, in un intorno di ampiezza $h$, la funzione che esprime il legame tra la funzione di rischio di $T^*$ ed il valore $x_0 \subseteq \aleph$ mediante un polinomio di grado $p \geq 0$. In modo informale, ciò equivale a stimare il valore della funzione di rischio associato al particolare punto $x_0$ solo sulla base del tempo effettivamente trascorso dalle unità del campione nell'intorno di ampiezza $h$ di tale punto e del numero di eventi che in esso si sono verificati (una procedura analoga è stata proposta da Hjort (1992), nel contesto della stima non parametrica di funzioni di rischio).

Il ricorso alla verosimiglianza locale introduce nella procedura di stima un elemento di arbitrarietà legato alla necessità di specificare il valore di $h$ che determina l'ampiezza dell'intorno del punto $x_0$. La scelta dell'ampiezza di questo intorno risulta essere di cruciale importanza. In particolare, per aumentare le capacità esplorative di questo metodo, verrà preso in considerazione un criterio di scelta

automatico ed adattativo (Friedman, 1984), cioè in grado, sulla base delle informazioni disponibili, di far variare l'ampiezza dell'intorno a seconda dell'elemento di $\aleph$ preso in considerazione, in modo da poter eventualmente tener conto del diverso grado di complessità della funzione che esprime la relazione tra la funzione di rischio e la covariata tempo-dipendente in diversi sottoinsiemi di $\aleph$.

## 2. VEROSIMIGLIANZA LOCALE PER L'ANALISI DI DATI DI DURATA: MODELLI DI REGRESSIONE PER FUNZIONI DI RISCHIO CON COVARIATE TEMPO-DIPENDENTI

Sia $T^*$ una v. c. a valori in $\Re^+$ che descrive la durata (o tempo di attesa) tra due eventi. Tutte le informazioni relative a questa v. c. sono contenute nella sua funzione di probabilità cumulata (o funzione di ripartizione)

$$F(t) = \Pr\left(T^* \leq t\right). \tag{1}$$

Nell'ambito dell'analisi della durata si preferisce, per motivi interpretativi, fare ricorso alla funzione di rischio (condizionato):

$$\alpha(t) = \lim_{dt \to 0} \frac{1}{dt} \Pr\left(T^* \in [t, t+dt] \,|\, T^* \geq t\right). \tag{2}$$

Funzione di ripartizione e funzione di rischio costituiscono due rappresentazioni equivalenti della forma distributiva della v. c. $T^*$; si può infatti dimostrare che

$$\alpha(t) = \frac{\partial}{\partial t} \ln\left[1 - F(t)\right].$$

Una delle principali peculiarità dei dati di durata è la possibile presenza di osservazioni censurate. Estratto un campione casuale da $T^*$ di dimensione $n$, anziché osservare le realizzazioni delle $n$ v. c. $T_i^*$ ($i=1,\ldots,n$), in generale è possibile osservare le realizzazioni delle $n$ v. c. bidimensionali $(T_i, \delta_i)$, con $T_i=\min\{T_i^*, C_i\}$ e $\delta_i = 1_{\{T_i^* \leq C\}}\left(T_i^*\right)$. La v. c. $C$, che rappresenta il tempo di censura, è anch'essa a valori in $\Re^+$ ed è a sua volta caratterizzata da una particolare funzione di ripartizione (o equivalentemente da una particolare funzione di rischio).

Un'altra situazione tipica nel contesto dell'analisi di dati di durata è la presenza, per ciascun $i$, di informazioni circa una o più covariate tempo-dipendenti $X(t)$, cioè covariate che possono assumere valori differenti in istanti temporali differenti e che sono descritte da $n$ traiettorie definite su $[0, T_i] \subseteq \Re^+$ ($i=1,\ldots,n$) a valori in $\aleph$, l'insieme in cui le traiettorie $X_i(\bullet)$ assumono valori.

Sulla base di queste informazioni può essere interessante analizzare l'eventuale relazione esistente tra le traiettorie descritte da $X(t)$ e la funzione di rischio associata a $T^*$. A tale scopo, è innanzitutto necessario esprimere in forma funzionale questa relazione. In particolare, nel presente lavoro si assume che esista una

funzione $g[\cdot]$ definita su $\aleph$ a valori in $\Re$ tale che, condizionatamente alla traiettoria descritta da $X(t)$ con $t \in \Re^+$, la v. c. $T^*$. abbia funzione di rischio pari a

$$\ln \alpha\left(t \mid X(t)\right) = g\left[X(t)\right]. \tag{3}$$

Tale formulazione implica che:

1. il valore della funzione di rischio al tempo $t$ dipende solo dal valore assunto dalla traiettoria della covariata in quel particolare istante;
2. le variazioni nel tempo della funzione di rischio sono imputabili esclusivamente a variazioni nel tempo della covariata tempo-dipendente (cioè a variazioni nelle traiettorie).

Per poter costruire la funzione di verosimiglianza (o equivalentemente di log-verosimiglianza) associata al campione osservato $(t_i,\ d_i,\ x_i(t)\ t \subseteq [0,\ T_i])\ \ i = 1,\ldots,n$, oltre alla densità di probabilità di $T^*$ (che, per quanto detto in precedenza, si ottiene una volta individuata la forma funzionale di $g[\cdot]$), è necessario individuare anche la legge distributiva della variabile $C$ e l'eventuale legame di dipendenza tra $T^*$ e $C$. Tuttavia, nell'ipotesi in cui il meccanismo di censura sia non informativo (ovvero $T^*$ e $C$ siano statisticamente indipendenti e la distribuzione di $C$ non contenga informazioni circa la distribuzione di $T^*$, si veda ad esempio Fahrmeir e Tutz, 1994), la funzione di log-verosimiglianza del campione osservato, a meno di una costante additiva, è pari a

$$\sum_{i=1}^{n}\left\{d_i g\left[x_i(t_i)\right] - \int_0^{t_i} \exp\left\{g\left[x_i(u)\right]\right\} du\right\}. \tag{4}$$

Nel caso in cui la forma funzionale di $g[\cdot]$ non sia nota, il metodo della massima verosimiglianza locale (Fan *et al.*, 1998) consente di ottenere una stima non parametrica di tale funzione. Fissato un punto $x_0 \in \aleph$ ed assumendo che $g[\cdot]$ abbia derivate continue fino all'ordine $p+1$ nel punto $x_0$, allora per i punti $x_i$ appartenenti ad un intorno $I(x_0, h)=[\,x_0 - h,\, x_0 + h] \subset \aleph$ è possibile approssimare $g[x_i]$, mediante espansione in serie di Taylor, con un polinomio di grado $p$:

$$g\left[x_i\right] \approx g\left[x_0\right] + g^{(1)}\left[x_0\right]\left(x_i - x_0\right) + \ldots + \frac{g^{(p)}\left[x_0\right]}{p!}\left(x_i - x_0\right)^p \equiv \mathbf{x}_i^{\mathrm{T}}\beta^{\mathbf{0}} \tag{5}$$

dove $\mathbf{x}_i^{\mathrm{T}} = \left(1, x_i - x_0, \ldots, \left(x_i - x_0\right)^p\right)$ e $\beta^{\mathbf{0}} = \left(\beta_0^0, \ldots, \beta_p^0\right)$, con $\beta_v^0 = \dfrac{g^{(v)}\left[x_0\right]}{v!}$, $v = 1,\ldots,p$. In questo modo, il contributo per ciascuna unità alla funzione di verosimiglianza locale rispetto a $x_0$ risulterà pari a

$$d_i\left[\mathbf{x}_i(t_i)^{\mathrm{T}}\beta^0 1_{\{I(x_0,h)\}}\left(x_i(t_i)\right)\right] - \int_0^{t_i} \exp\left\{\mathbf{x}_i(u)^{\mathrm{T}}\beta^0\right\} 1_{\{I(x_0,h)\}}\left(x_i(u)\right) du\,. \tag{6}$$

È interessante notare che, ponendo $p = 0$, ovvero approssimando localmente la funzione $g[\cdot]$ mediante una funzione costante, il contributo di ciascuna unità alla funzione di log-verosimiglianza locale si riduce a

$$d_i\left[\beta^0 1_{\{I(x_0,h)\}}(x_i(t_i))\right] - \exp\{\beta^0\} \times \int_0^{t_i} 1_{\{I(x_0,h)\}}(x_i(u))du\,, \tag{7}$$

il che equivale ad ipotizzare che, nell'intorno $I(x_0,h)$ la forma funzionale della legge di $T^*$ condizionata al valore assunto dalla covariata tempo-dipendente sia di tipo esponenziale di parametro $\exp\{\beta_0\}$. In questo caso la soluzione dell'equazione di verosimiglianza locale esiste in forma chiusa.

Quando $p > 0$, le soluzioni delle equazioni di verosimiglianza locale devono invece essere trovate per via numerica (ricorrendo ad esempio all'algoritmo di Newton-Raphson); ad ogni passo dell'algoritmo di massimizzazione, inoltre, è necessario ricorrere a tecniche di integrazione numerica (come la quadratura gaussiana).

Ricorrendo ad opportune tecniche di interpolazione, il metodo della massima verosimiglianza locale può essere adottato anche in situazione nelle quali le traiettorie della covariata siano note solo in corrispondenza di alcuni istanti temporali.

## 3. UNA PROPOSTA PER LA SCELTA DELL'AMPIEZZA DI BANDA

Il ricorso all'approssimazione locale di $g[\cdot]$ introduce nella procedura di stima un elemento di arbitrarietà legato alla necessità di specificare il valore di $h$ che determina l'ampiezza dell'intorno. Il parametro $h$ determina la complessità del modello: per $h = 0$ la stima prodotta risulta un'interpolazione dei dati osservati (e quindi individua il modello più complesso), mentre per $h = +\infty$ la funzione $g[\cdot]$ viene approssimata globalmente mediante un polinomio di grado $p$. Inoltre la scelta di $h$ è strettamente legata alla distorsione ed alla varianza del corrispondente stimatore: se da una lato gli stimatori associati a piccoli valori di questo parametro presentano una minore distorsione, dall'altro a valori elevati di $h$ sono associati stimatori con varianze meno elevate. In letteratura sono numerose le proposte di criteri per la scelta di questo parametro (si veda Fan e Gijbels, 1996, per una rassegna dei principali contributi).

Benché una ampiezza di banda fissa sia più semplice da implementare e da interpretare, il suo uso può determinare alcuni problemi in regioni dello spazio di definizione di $X(t)$ dove i dati sono sparsi, situazione tipica per i dati di durata. Un rimedio a questo problema può essere rappresentato dall'impiego di un metodo automatico per la scelta di $h$ che consenta a tale parametro di variare localmente nell'insieme $\aleph$.

In particolare, in questa sezione si presenta una proposta per la scelta dell'ampiezza di banda basata sul criterio nearest-neighbour (Tibshirani and Hastie, 1987): fissato il punto $x_0 \in \aleph$, si sceglie quel valore di $h$, dipendente da

$x_0$, tale che il numero di eventi osservati nell'intorno $I(x_0,h(x_0))$ sia uguale ad un intero positivo $J \in \left\{1,\ldots,\sum_{i=1}^{n} d_i\right\}$, ovvero ad un numero prefissato di eventi

$$\sum_{i=1}^{n} d_i 1_{\{I(x_0,h(x_0))\}}\left(x_i(t_i)\right) = J\ ,\ \forall x_0 \in \aleph\ . \tag{8}$$

In questo modo l'ampiezza di banda sarà maggiore laddove gli eventi sono più sparsi, e viceversa minore laddove gli eventi invece sono più densi. Chiaramente, procedendo in questo modo, l'attenzione viene spostata dalla scelta dell'ampiezza di banda alla scelta del valore del parametro $J$ (detto in letteratura "span"). Nel 1984 Friedman, nell'ambito della regressione non parametrica univariata con variabile dipendente continua, ha proposto una procedura (chiamata "supersmoother") che consente la scelta dello span ottimale in funzione del valore assunto dalla covariata usata come predittore. Tale procedura è basata sul concetto di cross-validation locale, ovvero scegliendo uno span in corrispondenza del valore del covariata $x_0$ sulla base dei residui leave-one-out relativi ad unità i cui corrispondenti valori della covariata appartengono all'intorno di $x_0$ (detti anche residui locali).

Per poter estendere la proposta originale di Friedman anche nell'ambito della regressione non parametrica per funzioni di rischio in presenza di una covariata tempo-dipendente è innanzitutto necessario definire opportunamente le quantità da impiegare come residui locali. Per questo scopo si può far riferimento ai cosiddetti residui martingala (Thernau *et al.*, 1990).

Posto $N_i(t) = 1_{\{T_i^* \le t\}}\left(T_i^*\right)$ uguale al processo di conteggio associato alla v. c. $T_i^*$ ed $Y_i(t)$ uguale al processo stocastico a valori in {0,1} che indica se l'unità $i$ è a rischio di sperimentare l'evento oggetto di interesse, la quantità

$$M_i(t) = N_i(t) - \int_0^t Y_i(u)\alpha\left[u \mid X_i(u)\right]du \tag{9}$$

risulta essere, sotto generali condizioni di misurabilità ed integrabilità (si veda ad esempio Andersen *et al.*, 1993) una martingala di quadrato integrabile.

Questa quantità può essere interpretata come la differenza tra il numero di eventi osservati nell'intervallo $[0,t]$ e il numero di eventi attesi per quello stesso intervallo. La martingala così definita, inoltre, gode di alcune proprietà molto simili a quelle dei termini di errore usati nell'ambito dei modelli lineari (Thernau e Grambsch, 2000).

Di particolare interesse ai fini della costruzione della procedura per la scelta dell'ampiezza di banda, oggetto di questa sezione, sono le proprietà relative agli incrementi $dM_i$ associati a tale martingala: essi risultano avere media nulla ed essere incorrelati.

Risulta inoltre che $\mathrm{Cov}[M_i(t) - M_i(t\text{-}s), M_i(t+u) - M_i(t)]=0$ per $t,u,s > 0$. In generale, quindi, individuata una partizione $I_1 = \left[t_1 = 0, t_2\right)$, …, $I_j = [t_j, t_{j+1})$,

…, $I_m = [t_{m-}, t_{m+1} = t]$, dell'intervallo temporale, la martingala $M_i$ può essere scomposta nella somma di quantità tra loro incorrelate:

$$\begin{aligned} M_i(t) &= \sum_{j=1}^{m} \left\{ \int_{t_{j-1}}^{t_j} dN_i(u) - \int_{t_{j-1}}^{t_j} Y_i(u)\alpha\left[u \mid X_i(u)\right]du \right\} = \\ &= \sum_{j=1}^{m} \left\{ \int_0^{\infty} 1_{\{I_j\}}(u)dN_i(u) - \int_0^{\infty} 1_{\{I_j\}}(u)Y_i(u)\alpha\left[u \mid X_i(u)\right]du \right\}. \end{aligned} \tag{10}$$

A fini operativi, i residui martingala possono allora essere definiti come

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(u)\hat{\alpha}\left[u \mid x_i(u)\right]du\,, \tag{11}$$

ovvero come la differenza tra il numero di eventi osservati nell'intervallo [0, $t$] ed la stima del numero di eventi attesi per quello stesso intervallo. Data una partizione dell'asse temporale, tale residuo martingala può essere scritto come

$$\hat{M}_i(t) = \sum_{j=1}^{m} \left\{ \int_0^{\infty} 1_{\{I_j\}}(u)dN_i(u) - \int_0^{\infty} 1_{\{I_j\}}(u)Y_i(u)\hat{\alpha}\left[u \mid x_i(u)\right]du \right\}. \tag{12}$$

Per poter sfruttare la definizione di residuo martingala per la costruzione di un criterio automatico per la scelta dell'ampiezza di banda, in analogia con la proposta di Friedman del 1984, si può osservare che, data la natura tempo-dipendente della covariata $X_i(t)$, ciascuna partizione del suo spazio di definizione induce anche una partizione dell'asse temporale (quest'ultima non necessariamente uguale per tutte le unità del campione). In particolare, data una partizione $I_1 = [x_1 = x_{\min}, x_2)$, $I_2 = [x_2, x_3)$, …, $I_j = [x_{j-}, x_{j+1})$, …, $I_m = [x_{m-}, x_{m+1} = x_{\max}]$, il residuo martingala associato al punto $x_0$ ed al valore dello span $J$ può essere calcolato come

$$\hat{r}_{(0)}(J) = \sum_{i=1}^{n} \left| \int_0^{\infty} 1_{\{I_0\}}\left[x_i(u)\right]dN_i(u) - \int_0^{\infty} 1_{\{I_0\}}\left[x_i(u)\right]Y_i(u)\hat{\alpha}_J^{-i}\left[u \mid x_i(u)\right]du \right|, \tag{13}$$

dove $I_0$ è l'intervallo della partizione al quale $x_0$ appartiene, mentre $\hat{\alpha}_J^{-i}\left[t \mid x_i(t)\right]$ rappresenta la stima della funzione di rischio ottenuta escludendo l'unità $i$ dal campione e fissando lo span pari a $J$.

La procedura congiunta per la stima mediante verosimiglianza locale della funzione $g[\bullet]$ e per la scelta dell'ampiezza di banda locale può essere riassunta nei seguenti punti:

a) si fissa il grado del polinomio locale (ad esempio $p = 1$);

b) si fissano tre valori per lo span $J_1 = 0.05\sum_{i=1}^{n} d_i$, $J_2 = 0.2\sum_{i=1}^{n} d_i$, $J_3 = 0.5\sum_{i=1}^{n} d_i$ (la scelta di tali valori è giustificata, nel lavoro originale di Friedman, dal tentativo di riprodurre le tre parti principali dello spettro di $g[\bullet]$);
c) sulla base dei valori della covariata in corrispondenza dei quali si verifica almeno un evento, si individua una partizione dell'insieme $\aleph$ in $m$ intervalli $I_1 = [x_0, x_1)$, …, $I_j = [x_{j-1}, x_j)$, …, $I_m = [x_{m-1}, x_m)$. Data la natura tempo-dipendente della covariata, questa partizione individuerà a sua volta una partizione dell'asse temporale (non necessariamente uguale per ciascuna unità del campione);
d) si esegue una stima preliminare $g(x_j)$ per ciascuno dei tre valori di span fissati al punto a);
e) per ogni $x_j$ e per ogni $J_l$, calcolo dei residui locali cross-validati $\hat{r}_{(j)}(J_l)$, come in formula (13);
f) per ciascun intervallo, si individua il valore $\hat{J}(x_j)$ tra i tre proposti che corrisponde al più piccolo residuo locale cross-validato lisciato (con span pari a $J_2$) $\hat{r}_{(j)}(J_l)$, $l$=1, 2, 3;
g) si calcolano i valori $\tilde{J}(x_j)$ mediante lisciatura della sequenza dei valori $\hat{J}(x_j)$ rispetto agli $x_j$;
h) si individua la stima finale della funzione $g(x_j)$ mediante interpolazione tra le due stime iniziali $\hat{g}_{J_l}(x_j)$ associate ai due valori $J_l$ più vicini a $\tilde{J}(x_j)$.

## 4. UNO STUDIO DI SIMULAZIONE

Le performance del metodo proposto nelle precedenti sezioni sono state studiate mediante uno studio di simulazione costituito da tre diversi esperimenti. Le istruzioni per la generazione dei dati e la loro analisi sono state implementate in GAUSS. Per semplicità, si è scelta una covariata tempo-dipendente con traiettorie monotone non decrescenti. In particolare, le traiettorie di questa covariata risultano essere pari a $X_i(t) = 2\Phi\left(\frac{t - A_i}{B_i}\right)$, $t > 0$, dove $\Phi(\bullet)$ è la funzione di ripartizione di una v. c. normale standard; la traiettoria di ciascuna unità campionaria si caratterizza per un diverso valore della media $A$ e dello scarto quadratico medio $B$, generati da due v. c. uniformi indipendenti: $A \sim U(0,2)$ e $B \sim U(0.5,1)$. I tempi di censura $C$ sono stati estratti da una distribuzione esponenziale di parametro 5, indipendente da $T$.

Per ciascuno dei tre esperimenti, è stato preso in considerazione un diverso modello per la funzione $g[\bullet]$ e sono stati generati 100 campioni, ciascuno composto da 500 unità. La tabella 1 contiene i modelli usati nelle simulazioni.

TAVOLA 1
*Modelli considerati nelle simulazioni*

| Esperimento | g[X(t)] |
|---|---|
| I | 0.8[X(t) – 0.5] |
| II | sen[πX(t)] |
| III | 0.5{sen[2X(t) – 1] + exp[– 16(X(t) – 0.5)2]} |

I risultati di ciascuno dei tre esperimenti sono riassunti nelle seguenti figure. Le figure dalla 1 alla 4 si riferiscono all'esperimento I, per il quale la funzione $g[\bullet]$ è lineare e quindi rappresenta il modello più semplice tra i tre esaminati nello studio di simulazione.

Nella figura 1 viene riportata, per ogni valore della covariata, la media aritmetica delle stime di $g[\bullet]$ ottenute nei 100 campioni. Come è possibile vedere, il supersmoother può essere considerato uno stimatore non distorto per il modello I.

Nelle figure 2, 3 e 4 i risultati ottenuti con il supersmoother vengono confrontati con quelli ottenuti con riferimento alle tre stime iniziali, associate ai tre valori fissi dello span $J_1 = 0.05\sum_{i=1}^{n} d_i$, $J_2 = 0.2\sum_{i=1}^{n} d_i$, $J_3 = 0.5\sum_{i=1}^{n} d_i$.

In particolare, nella figura 2 il confronto viene effettuato, per ciascun valore della covariata, sulla base della distorsione degli stimatori (valutata sulla base delle stime ottenute per i 100 campioni). Come mostrato in questa figura, per quanto riguarda il modello I, anche i tre stimatori iniziali risultano essere pressoché non distorti.

Nella figura 3 il confronto viene eseguito prendendo in esame le varianze (considerate come funzione dei valori della covariata). In generale si può notare come queste tendano ad essere più elevate in corrispondenza degli estremi del dominio della covariata: è opportuno notare che questa è una caratteristica tipica dei metodi di regressione non parametrica (si veda ad esempio Fan e Gjibels, 1996). Lo stimatore ottenuto fissando il valore dello span pari a 0.5 è caratterizzato da una funzione di varianza uniformemente minore rispetto alle funzioni di varianza degli altri tre stimatori. Questo risultato è legato ai valori dell'ampiezza degli intorni locali associati al valore dello span: come già osservato in precedenza, localmente la varianza dello stimatore decresce al crescere del valore di $h$ (si noti inoltre che lo stimatore ottenuto fissando lo span pari a 0.2 è caratterizzato da una funzione di varianza che è uniformemente minore della funzione di varianza dello stimatore con valore di span più piccolo). Per quanto riguarda il modello I, si può notare che il supersmoother mostra una funzione di varianza molto simile a quella dello stimatore iniziale con span uguale a 0.2.

Le distorsioni e le varianze degli stimatori possono essere combinate nell'errore quadratico medio, sommando, in corrispondenza di ogni valore della covariata, la varianza e la distorsione al quadrato. Poiché le distorsioni (al quadrato) assumono valori molto vicini allo zero, le funzioni di errore quadratico medio mostrano un andamento del tutto simile a quello delle funzioni di varianza e quindi, anche con riferimento all'errore quadratico medio, lo stimatore migliore tra i quattro presi in esame risulta essere quello associato al valore fisso di span pari a 0.5. Le ragioni di

ciò sono strettamente legate alla particolare natura del modello I: in questo caso, la forma funzionale di $g[\cdot]$ coincide con la forma funzionale usata nella procedura di approssimazione locale e quindi quest'ultima darà risultati tanto migliori quanto più elevato sarà il valore dello span (o, equivalentemente, di $h$); come già accennato, nel caso limite (span uguale a 1) la procedura di approssimazione locale riprodurrà esattamente un modello parametrico globale.

I risultati relativi al modello II sono presentati nelle figure dalla 5 alla 8. Come per il precedente modello, la figura 5 riporta le medie aritmetiche delle stime ottenute con i 100 campioni. Si può notare come in questo caso il supersmoother, benché in grado di cogliere l'andamento di fondo della funzione $g[\cdot]$, risulti essere distorto. Confrontando le distorsioni dei tre stimatori ottenuti fissando i valori dello span e del supersmoother (figura 6) è possibile osservare che, per questo secondo modello, solo lo stimatore con span pari a 0.05 presenta una distorsione che risulta uniformemente vicina a zero al variare del valore di $X$, mentre gli altri tre stimatori presentano distorsioni non trascurabili. È interessante notare che le distorsioni assumono i valori più elevati in corrispondenza delle regioni del dominio di $X$ nelle quali la funzione $g[\cdot]$ presenta la maggior curvatura (cioè i valori più elevati della derivata seconda). Questo risultato non è del tutto sorprendente e dipende dal fatto che localmente la funzione $g[\cdot]$ viene approssimata mediante un polinomio di grado 1. In generale (Fan *et al.*, 1998), la distorsione degli stimatori di massima verosimiglianza locale è strettamente legata all'errore di approssimazione associato all'espansione in serie di Taylor:

$$R(x_i) = g(x_i) - \sum_{j=1}^{p} \frac{g^{(j)}[x_0]}{j!}(x_i - x_0)^j , \tag{14}$$

quantità che, nel caso in cui $g[\cdot]$ abbia derivate fino all'ordine $p + a + 1$ nel punto $x_0$ per $a > 0$, può essere approssimata da

$$R(x_i) \approx \sum_{j=p+1}^{p+a} \frac{g^{(j)}[x_0]}{j!}(x_i - x_0)^j . \tag{15}$$

Con riferimento alla variabilità degli stimatori considerati, dalla figura 7 emerge come il comportamento degli stimatori associati ai tre valori iniziali dello span sia simile a quello già visto per il modello I e, anche in questo caso, la funzione di varianza del supersmoother grosso modo non si discosta da quella associata allo stimatore con span uguale a 0.2.

Data la non trascurabilità delle distorsioni, risulta opportuno confrontare i quattro stimatori ricorrendo alle funzioni di errore quadratico medio. Da quanto emerge dalla figura 8, nessuno degli stimatori considerati presenta un comportamento uniformemente migliore rispetto agli altri. Si può però notare come, globalmente, il supersmoother sia caratterizzato da un miglior trade-off tra distorsione e varianza.

Le figure dalla 9 alla 12, infine, presentano i risultati relativi al modello III. Come si può notare, questi risultati sono qualitativamente molto simili a quelli ottenu-

ti per il modello II: il supersmoother è in grado di cogliere l'andamento di fondo della funzione $g[\bullet]$ (figura 9), ma solo lo stimatore associato al valore di span 0.05 mostra una distorsione trascurabile (figura 10). Le distorsioni degli altri stimatori sono più elevate in corrispondenza dei valori della covariata in cui $g[\bullet]$ presenta un valore della derivata seconda elevato. Anche in questo caso, sulla base delle funzioni di errore quadratico medio (figura 12) non è possibile individuare uno stimatore che presenti uniformemente un comportamento migliore degli altri; contrariamente a quanto visto per il modello II, tuttavia, lo stimatore che globalmente sembra mostrare il miglior trade-off tra distorsione e varianza quello associato al valore dello span pari a 0.2, anche se il supersmoother non sembra discostarsi di molto.

## 5. CONCLUSIONI

Il metodo della massima verosimiglianza locale è stato ampiamente sfruttato per risolvere diversi problemi di stima non parametrica. In questo articolo è stata proposta una sua estensione anche al caso della regressione univariata per funzioni di rischio in presenza di una covariata tempo-dipendente e le capacità di questa estensione sono state analizzate mediante uno studio di simulazione. Sebbene i risultati di queste simulazioni sembrino essere incoraggianti, diversi aspetti della proposta richiedono ulteriori approfondimenti.

Un primo problema che richiede ulteriori studi è rappresentato dalla stima della distorsione e della varianza. La stima di queste quantità non solo può essere sfruttata per la costruzione di intervalli di confidenza o di bande di confidenza per la funzione obbiettivo, ma potrebbero anche essere sfruttate per derivare procedure di selezione dell'ampiezza di banda alternative a quella considerata in questo articolo.

Un altro aspetto da approfondire può essere rappresentato dall'introduzione di funzioni kernel per pesare in modo differente le osservazioni a seconda della loro distanza dal punto $x_0$ in corrispondenza del quale si vuole ottenere una stima della funzione obbiettivo. Con riferimento a questo contesto, è necessario però dedicare particolare attenzione alla definizione di distanza tra le osservazioni e il punto $x_0$: bisogna tener conto del fatto che le osservazioni non sono caratterizzate da punti nello spazio della covariata, ma da funzioni a valori in tale spazio.

Benché il metodo proposto si collochi nell'ambito della regressione univariata, infine, si può osservare che esso può essere impiegato anche in contesti di regressione multivariata, nell'ambito di algoritmi di backfitting per la stima dei coefficienti di modelli additivi (Hastie e Tibshirani, 1990). Il caso in cui non sia possibile ipotizzare l'additività degli effetti delle covariate rappresenta invece un ulteriore area di ricerca futura.

*Dipartimento di Scienze statistiche "Paolo Fortunati"*
*Università di Bologna*
GIULIANO GALIMBERTI

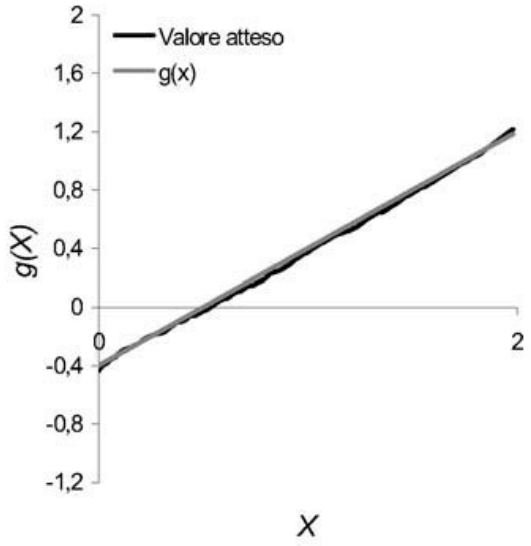


*Figura 1* – Modello I: valore atteso del supersmoother.

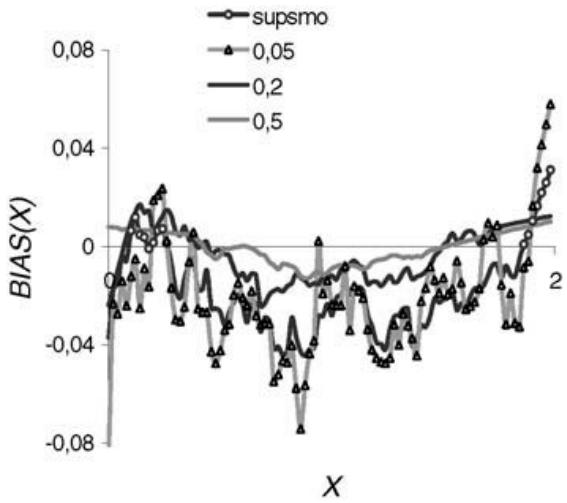


*Figura 2* – Modello I: distorsione degli stimatori.

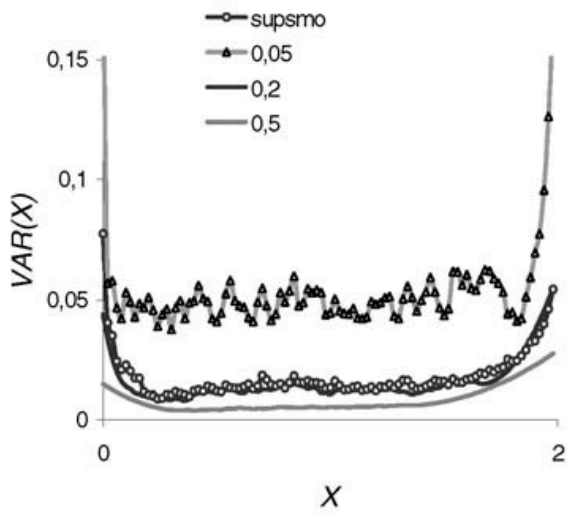


*Figura 3* – Modello I: varianza degli stimatori.

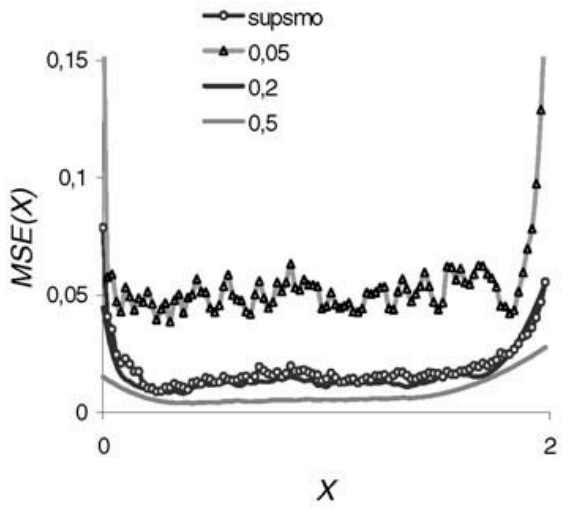


*Figura 4* – Modello I: errore quadratico medio degli stimatori.

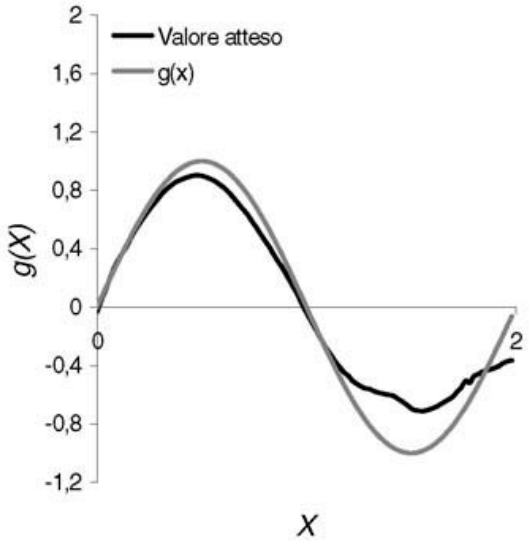


*Figura 5* – Modello II: valore atteso del supersmoother.

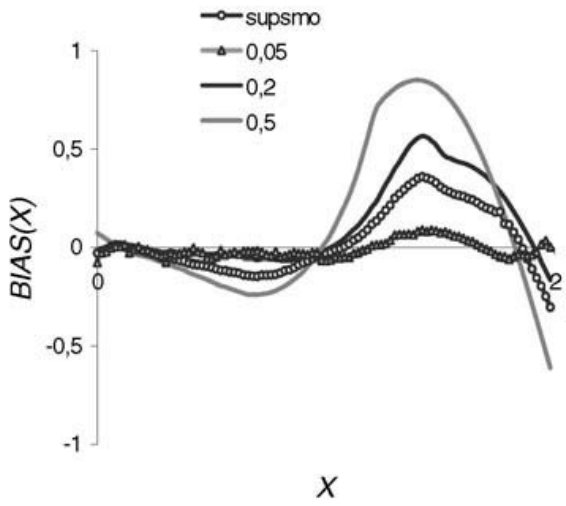


*Figura 6* – Modello II: distorsione degli stimatori.

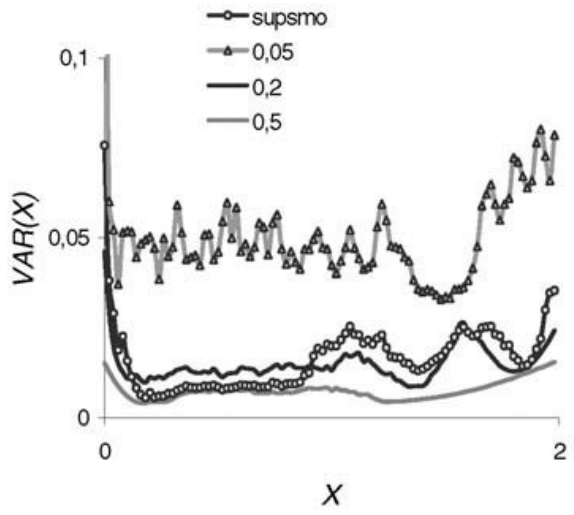


*Figura 7* – Modello II: varianza degli stimatori.

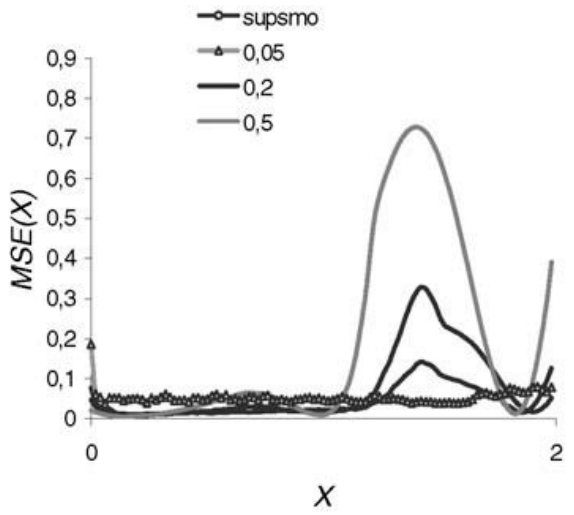


*Figura 8* – Modello II: errore quadratico medio degli stimatori.

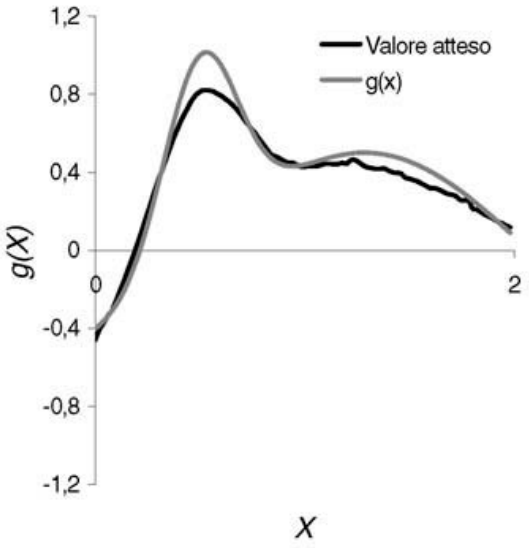


*Figura 9* – Modello III: valore atteso del supersmoother.

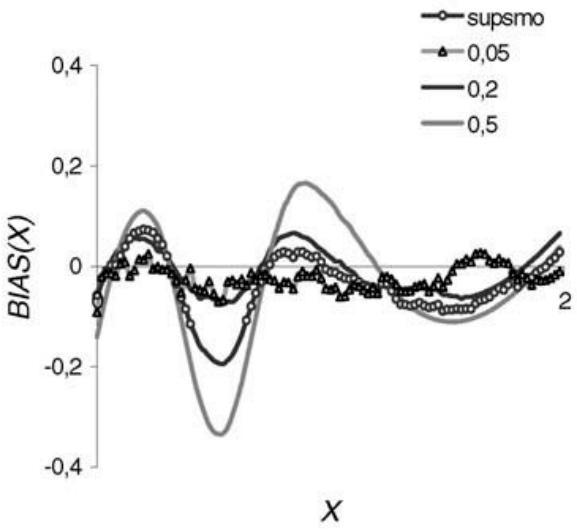


*Figura 10* – Modello III: distorsione degli stimatori.

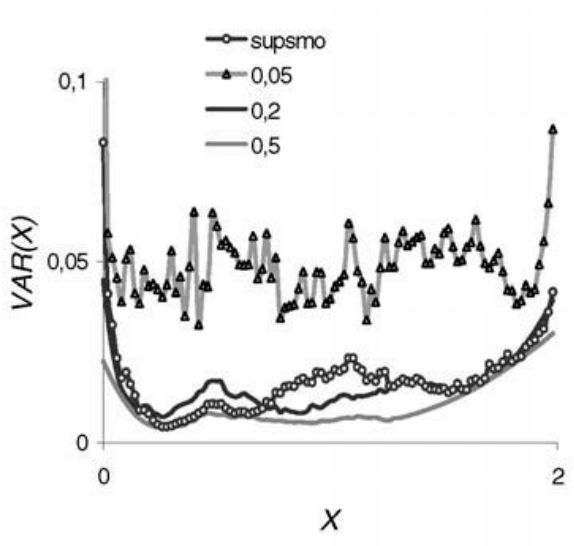


*Figura 11* – Modello III: varianza degli stimatori.

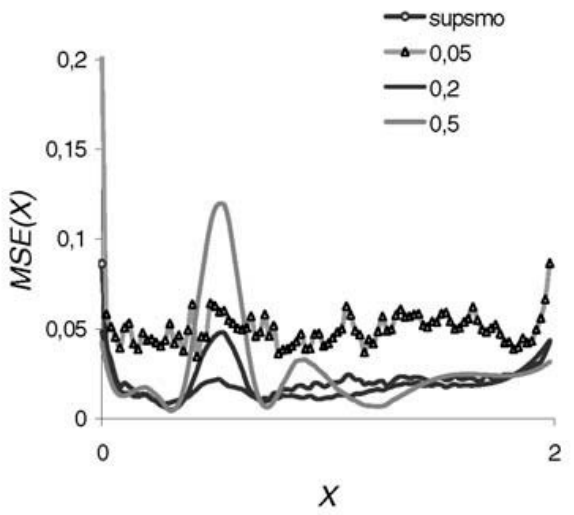


*Figura 12* – Modello III: errore quadratico medio degli stimatori.

## RIFERIMENTI BIBLIOGRAFICI

P. K. ANDERSEN, O. BORGAN, R. D. GILL, N. KEIDING, (1993), *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.

L. FAHRMEIR, G. TUTZ, (1994), *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer-Verlag, New York.

J. FAN, M. FARMEN, I. GIJBELS, (1998), *Local maximum likelihood estimation and inference*, "Journal of the Royal Statistical Society", B 60, pp. 591-608.

J. FAN, I. GIJBELS, (1995), *Data-driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation*, "Journal of the Royal Statistical Society", B 57, pp. 371-394.

J. FAN, I. GIJBELS, (1996), *Local Polinomial Modelling and Its Applications*, Chapman and Hall, London.

J. H. FRIEDMAN, (1984), *A variable span smoother*, Dept. of Statistics Technical Report LCS 05, Stanford University.

T. HASTIE, R. TIBSHIRANI, (1990), *Generalized Additive Models*, Chapman and Hall, London.

N. L. HJORT, (1992), *On Inference in Parametric Survival Data Models*, "International Statistical Review", 60, pp. 355-387.

T. M. THERNEAU, P. M. GRAMBSCH, (2000), *Modeling Survival Data*, Springer-Verlag, New York.

T. M. THERNEAU, P. M. GRAMBSCH, T. R. FLEMING, (1990), *Martingale-based residuals for survival models*, "Biometrika", 77, pp. 147-160.

R. TIBSHIRANI, T. HASTIE, (1987), *Local likelihood estimation*, "Journal of the American Statistical Association", 82, pp. 559-567.

## RIASSUNTO

*Regressione per funzioni di rischio con covariate tempo-dipendenti: una proposta basata sulla massima verosimiglianza locale*

L'obbiettivo di questo lavoro è quello di proporre un metodo flessibile per esplorare la possibile relazione tra la funzione di rischio associata ad una variabile durata $T^*$ ed una covariata tempo-dipendente. Questo metodo è basato sulla verosimiglianza locale, che non richiede una esplicita specificazione della forma funzionale di tale relazione. Per selezionare automaticamente l'ampiezza di banda, il metodo dello smoother a span variabile (Friedman, 1984), detto anche supersmoother, viene adattato al contesto dell'analisi di dati di durata.

## SUMMARY

*Hazard regression with time-dependent covariates: a proposal based on maximum local likelihood*

The aim of this paper is to propose a flexible method to explore the possible relationship between the hazard rate function associated to a duration time $T^*$ and a time-dependent covariate $X(t)$. This method is based on a local likelihood approach that does not require an explicit specification of the functional form of this relationship. In order to automatically select the bandwidth, the variable span smoother (Friedman, 1984), also called supersmoother, is adapted to the context of duration data analysis.