Introduzione ai Modelli di Durata: Stime Non-Parametriche (cenni)

a.a. 2009/2010 - Quarto Periodo

Prof. Filippo DOMMA

Corso di Laurea Specialistica/Magistrale in Economia Applicata

Facoltà di Economia – UniCal

Metodi Non Parametrici

Nel seguito saranno richiamate alcune tecniche di stima non parametrca di quantità descritte in precedenza, in particolare della funzione di sopravvivenza e della *hazard function*. L'utilità di queste stime si rilevano nella fase di scelta del modello parametrico in quanto forniscono sotto l'ipotesi di omogeneità delle osservazioni, ad esempio, l'andamento osservato della funzione di sopravvivenza e/o della *hazard function*.

- Tavola di sopravvivenza (*life table*);
- Kaplan-Meier (*Product-limit*).

Esempio.

n.	Data Trapianto (ingresso nello studio)	Data di cessata funzione del rene	Tempo di partecipaz. e stato alla data limite (31/12/81) (T – in mesi, δ)
1	30/01/1974	25/02/1974	(0,1)
2	13/04/1974	Vivo alla data limite	(93,0)
3	06/06/1974	08/07/1974	(1,1)
4	26/11/1974	23/12/1974	(0,1)
5	21/12/1974	Vivo alla data limite	(85,0)
6	16/02/1975	02/05/1975	(2,1)
7	16/02/1975	20/03/1978	(37,1)
8	28/03/1975	17/06/1975	(2,1)
9	18/09/1975	01/04/1976	(6,1)
10	15/01/1976	11/02/1976	(0,1)
11	18/01/1976	Morto il 15/07/1976 (con rene funzionante)	(5,0)
….	…….	…….	…….

Tavole di sopravvivenza

Per costruire la tavola di sopravvivenza suddividiamo l'asse temporale in intervalli,

$$I_j = [t_{j-1}, t_j)$$

con j=1,2,…,k+1, $t_0 = 0$, $t_k = L$ e $t_{k+1} = \infty$, dove L è il limite superiore dei tempi osservati. Indichiamo con t_{mj} e a_j , rispettivamente, il valore centrale e l'ampiezza dell'intervallo j-esimo.

In ogni intervallo così costruito si trovano sia osservazioni non-censurate (cioè che realizzano l'evento di interesse; ad esempio, l'individuo abbandona lo stato di disoccupazione) sia censurate (individui che non hanno ancora realizzato l'evento di interesse) che troncate (ad esempio, individui che escono dalla forza lavoro – non sono più alla ricerca di un posto di lavoro - o persi di vista).

In particolare, indichiamo con:

n'_j il numero di individui osservati nell'intervallo I_j
essi non hanno realizzato l'evento di interesse e sono ancora sotto osservazione
(sono gli individui disoccupati in cerca di occupazione – non sono persi di vista),

d_j il numero di individui che realizzano l'evento di interesse
(ad esempio, trovano un'occupazione cioè lasciano lo stato di disoccupato),

c_j il numero di osservazioni censurate
(cioè individui di cui non sappiamo quando lasceranno lo stato di disoccupazione),

l_j il numero di individui il cui tempo di partecipazione è troncato
perché persi di vista.

Chiaramente $n'_1 = n$ e

$$n'_j = n'_{j-1} - d_{j-1} - l_{j-1} - c_{j-1}$$

per j=2,3,….,k+1. Si definisce inoltre

$$n_j = n'_j - \frac{1}{2}(l_j + c_j)$$

il numero di individui "**esposti al rischio**" nell'intervallo Ij.
In assenza di censura e troncamento, evidentemente, si ha: $n_j = n'_j$

Indichiamo con q_j la probabilità condizionata che un individuo realizzi l'evento di interesse nell'**intervallo** I_j dato non lo ha realizzato in tutti i precedenti intervalli ed è sotto osservazione all'inizio dell'intervallo I_j.

N.B.: la definizione sopra riportata non deve essere confusa con la definizione di *hazard function* la quale fa riferimento alla probabilità condizionata della realizzazione dell'evento di interesse in un intervallo infinitesimale.

La stima della probabilità condizionata q_j è:

$$\hat{q}_j = \frac{d_j}{n_j} \qquad \text{per } j=2,3,\ldots,k+1.$$

Si dimostra che le stime della funzione di sopravvivenza e della funzione di densità al tempo tj e della *hazard function* al tempo t_{mj} sono, rispettivamente, date dalle seguenti:

$$\hat{S}(t_j) = \prod_{k=1}^{j}(1-\hat{q}_k) = \hat{S}(t_{j-1}) \times (1-\hat{q}_j)$$

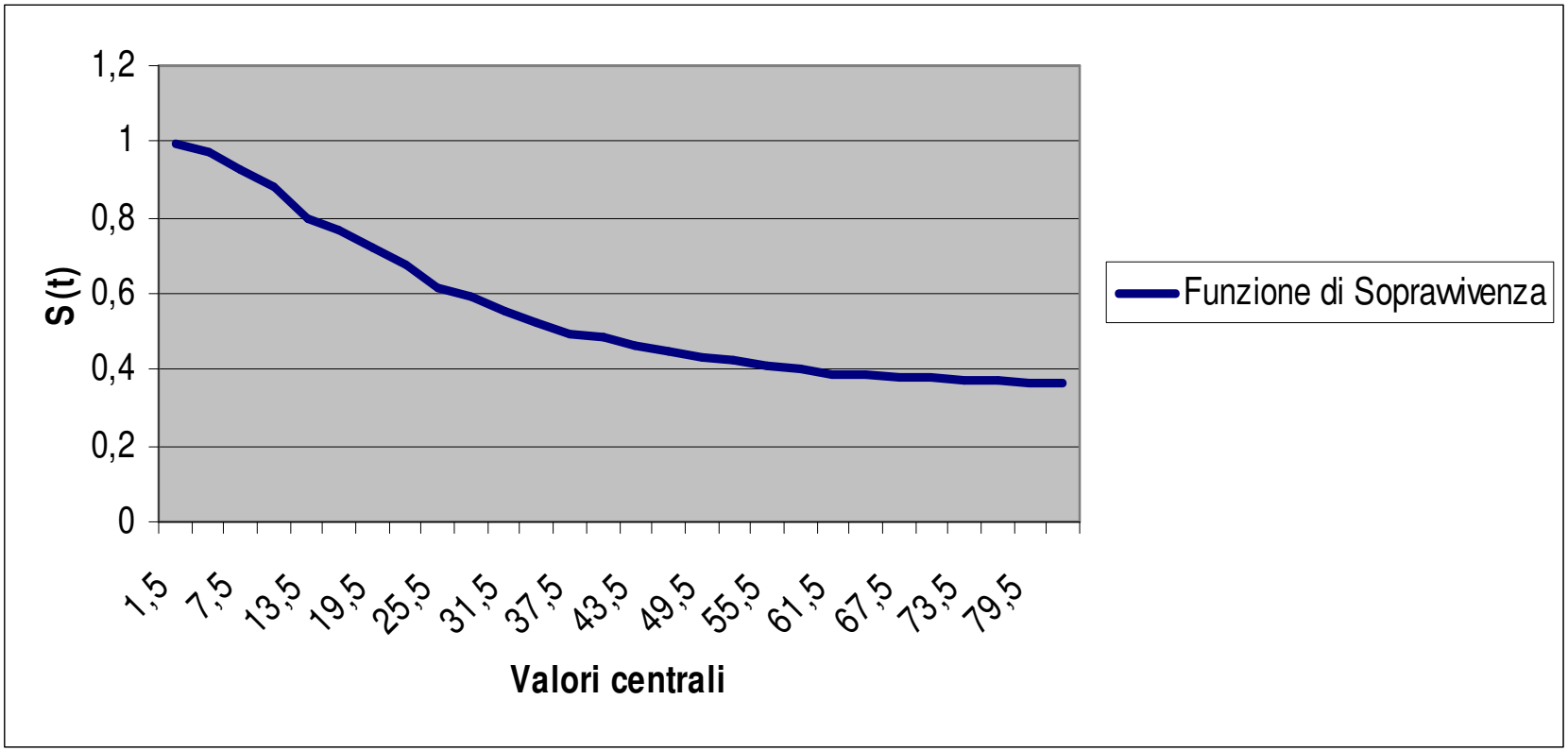
$$\hat{f}(t_{mj}) = \frac{\hat{F}(t_j) - \hat{F}(t_{j-1})}{a_j} = \frac{\hat{S}(t_{j-1}) - \hat{S}(t_j)}{a_j} = \frac{\hat{S}(t_{j-1})\hat{q}_j}{a_j}$$

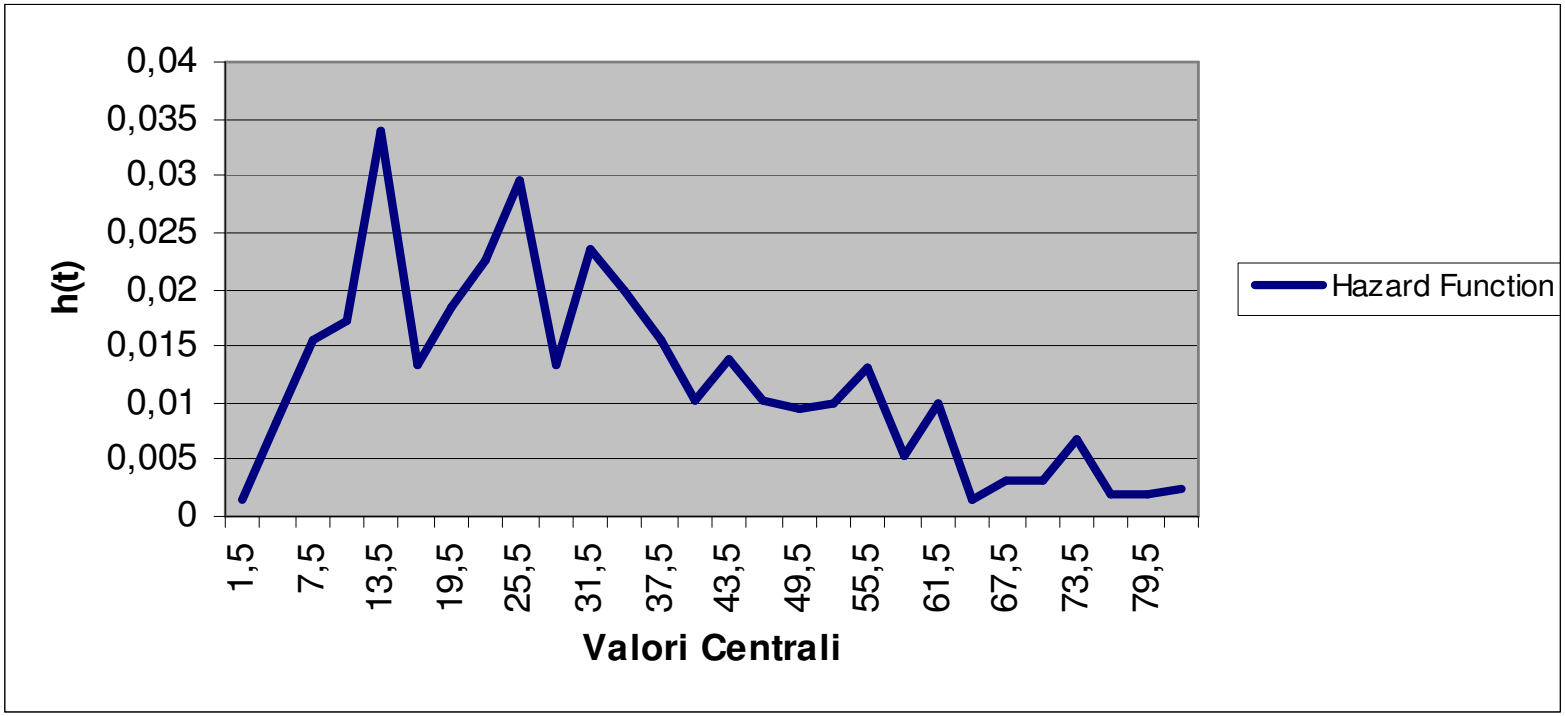
$$\hat{h}(t_{mj}) = \frac{\hat{f}(t_j)}{\tilde{S}(t_j)} = \frac{\frac{\hat{S}(t_{j-1})\hat{q}_j}{a_j}}{\frac{\hat{S}(t_{j-1}) + \hat{S}(t_j)}{2}} = \frac{2\hat{q}_j}{a_j(1+\hat{p}_j)}$$

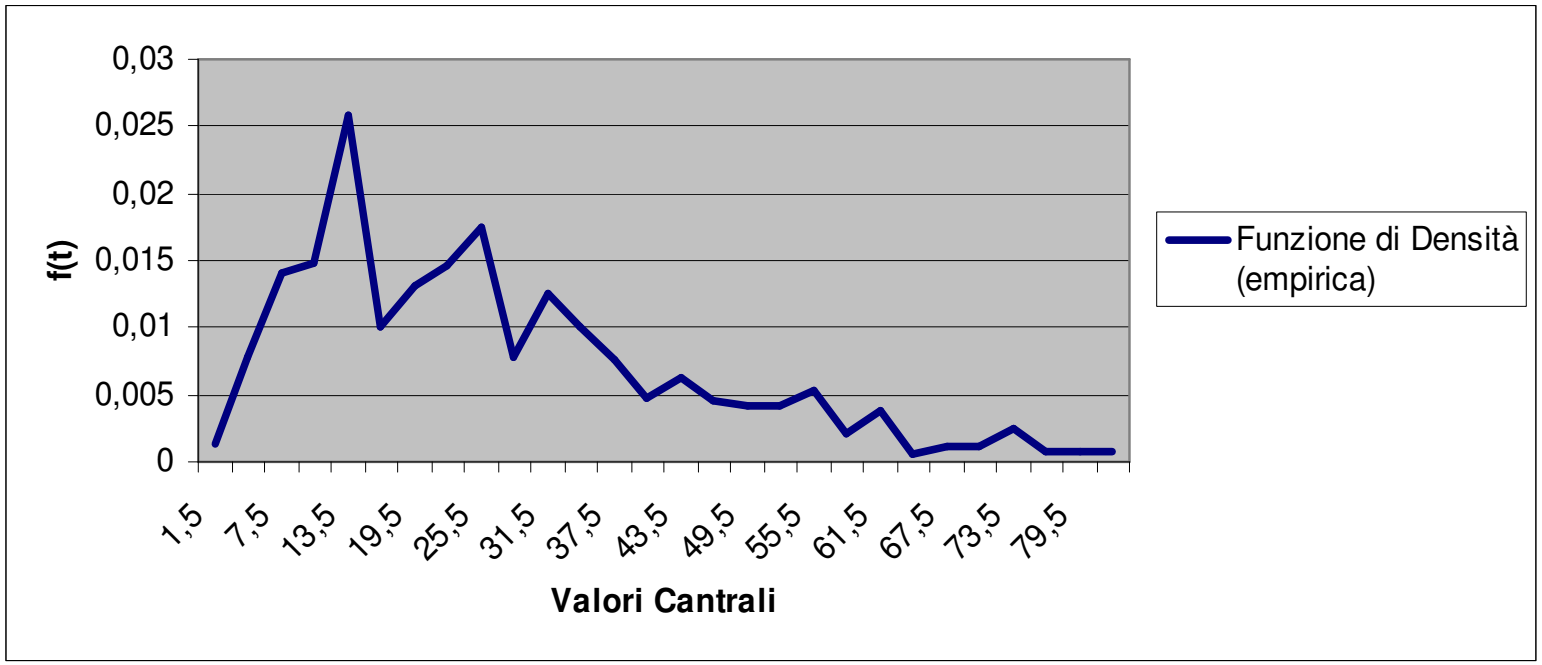
dove $\hat{p}_j = 1 - \hat{q}_j$

Esempio 1.2. Mercato del Lavoro (primo inserimento)

Intervalli di tempo	N. di individui ancora Occupati in Ij (nj')	Esposti a Rischio nj	N. Individui Censurati in Ij cj	N. di individui che escono dallo stato di Occupato dj	ampiezza intervallo	Prob. Condizionata realizzare l'evento qj	Prob. Condizionata non realizzare l'evento pj	Prob. Cumulativa di Sopravv in Ij Pj=pjPj-1	Hazard function nel punto centrale	Valori centrali	Funzione di Densità (Empirica)
[0,3)	703	703	0	3	3	0,00426743	0,995732575	0,995732575	0,001425517	1,5	0,001416
[3,6)	700	700	0	17	3	0,02428571	0,975714286	0,971550498	0,008194746	4,5	0,007865
[6,9)	683	683	0	31	3	0,04538799	0,954612006	0,92745377	0,015480649	7,5	0,014032
[9,12)	652	651,5	1	33	3	0,05065234	0,949347659	0,880476065	0,017322835	10,5	0,014866
[12,15)	618	618	0	60	3	0,09708738	0,902912621	0,794992952	0,034013605	13,5	0,025728
[15,18)	558	558	0	22	3	0,03942652	0,960573477	0,763649144	0,013406459	16,5	0,010036
[18,21)	536	536	0	29	3	0,05410448	0,945895522	0,722332306	0,018536274	19,5	0,013027
[21,24)	507	506	2	33	3	0,06521739	0,934782609	0,675223677	0,02247191	22,5	0,014679
[24,27)	472	471,5	1	40	3	0,08483563	0,915164369	0,617940651	0,029531192	25,5	0,017474
[27,30)	431	430,5	1	17	3	0,03948897	0,960511034	0,593538813	0,01342812	28,5	0,007813
[30,33)	413	410,5	5	28	3	0,0682095	0,931790499	0,553053827	0,023539302	31,5	0,012575
[33,36)	380	378	4	22	3	0,05820106	0,941798942	0,520865509	0,019981835	34,5	0,010105
[36,39)	354	352	4	16	3	0,04545455	0,954545455	0,497189804	0,015503876	37,5	0,007533
[39,42)	334	333,5	1	10	3	0,02998501	0,970014993	0,482281564	0,010147133	40,5	0,00482
[42,45)	323	320,5	5	13	3	0,04056162	0,959438378	0,462719441	0,013800425	43,5	0,006256
[45,48)	305	300	10	9	3	0,03	0,97	0,448837858	0,010152284	46,5	0,004488
[48,51)	286	283	6	8	3	0,02826855	0,971731449	0,436149862	0,009557945	49,5	0,00411
[51,54)	272	270,5	3	8	3	0,02957486	0,970425139	0,42325079	0,010006254	52,5	0,004173
[54,57)	261	259,5	3	10	3	0,03853565	0,961464355	0,406940548	0,013097577	55,5	0,005227
[57,60)	248	247,5	1	4	3	0,01616162	0,983838384	0,400363731	0,005431093	58,5	0,002157
[60,63)	243	236,5	13	7	3	0,02959831	0,970401691	0,388513642	0,010014306	61,5	0,003833
[63,66)	223	222	2	1	3	0,0045045	0,995495495	0,38676358	0,001504891	64,5	0,000581
[66,69)	220	218	4	2	3	0,00917431	0,990825688	0,383215291	0,003072197	67,5	0,001172
[69,72)	214	208,5	11	2	3	0,00959233	0,990407674	0,379539365	0,003212851	70,5	0,001214
[72,75)	201	196	10	4	3	0,02040816	0,979591837	0,371793663	0,006872852	73,5	0,002529
[75,78)	187	182	10	1	3	0,00549451	0,994505495	0,369750841	0,001836547	76,5	0,000677
[78,81)	176	168,5	15	1	3	0,00593472	0,994065282	0,367556474	0,001984127	79,5	0,000727
[81,84)	160	145	30	1	3	0,00689655	0,993103448	0,365021602	0,002306805	82,5	0,000839
>84	129		…	…							


1,2
1
0,8
0,6
0,4
0,2
0
S(t)
1,5
7,5
13,5
19,5
25,5
31,5
37,5
43,5
49,5
55,5
61,5
67,5
73,5
79,5
Valori centrali
Funzione di Sopravvivenza


0,04
0,035
0,03
0,025
0,02
0,015
0,01
0,005
0
h(t)
1,5
7,5
13,5
19,5
25,5
31,5
37,5
43,5
49,5
55,5
61,5
67,5
73,5
79,5
Valori Centrali
Hazard Function


0,03
0,025
0,02
0,015
0,01
0,005
0
f(t)
1,5
7,5
13,5
19,5
25,5
31,5
37,5
43,5
49,5
55,5
61,5
67,5
73,5
79,5
Valori Cantrali
Funzione di Densità (empirica)

Kaplan-Meier (Product-limit).

Il metodo del prodotto limite stima la curva di sopravvivenza in base al criterio della massima verosimiglianza.
A differenza della procedura utilizzata nella costruzione della tavola di sopravvivenza, il metodo di Kaplan-Meier **non** implica la suddivisione dell'asse temporale in intervalli di ampiezza prefissata e, quindi, nemmeno il corrispondente raggruppamento di soggetti.
Infatti, Kaplan e Meier stimano la probabilità condizionata di sopravvivenza in corrispondenza di ciascuno dei tempi in cui si verifica almeno un evento di interesse.
Tale metodo può essere visto come caso limite del metodo attuariale (tavole di sopravvivenza) in cui si costruisco intervalli infinitesimali.

Sia :

N — il numero di soggetti ammessi allo studio;

J — (con $J \leq N$) il numero di tempi distinti in cui si verificano gli eventi di interesse rilevati nel campione e ordinati in modo crescente: $t_{(1)} < t_{(2)} < \ldots < t_{(J)}$;

d_j — il numero di soggetti che realizzano l'evento al tempo $t_{(j)}$, $j=1,2,\ldots,J$. E' evidente che $d_j > 1$ solo nel caso in cui più soggetti realizzano l'evento di interesse in $t_{(j).}$

n_j — il numero di individui esposti a rischio al tempo $t_{(j)}$. Sono tutti i soggetti vivi e sotto osservazione appena prima di $t_{(j).}$

N.B.: i soggetti persi di vista o usciti vivi vengono inclusi negli esposti a rischio (si ipotizza che tali soggetti abbiano una esperienza di vita identica a quelli che realizzano l'evento di interesse).

Una stima della probabilità condizionata q_j di realizzare l'evento di interesse all'istante $t_{(j)}$ dato che il soggetto non ha realizzato l'evento fino all'istante immediatamente precedente, è data da:

$$\hat{q}_j = \frac{d_j}{n_j} \qquad j=1,2,\ldots,J$$

Si osservi che in tale contesto la stima della hazard function coincide con quella della probabilità condizionata sopra definita, cioè

$$\hat{h}(t_j) = \hat{q}_j = \frac{d_j}{n_j}$$