STIMATORE DELLA FUNZIONE DI SOPRAVVIVENZA E SUE PROPRIETÀ

Per stimare la funzione di sopravvivenza $S(x)$, $x = a, a+1, \dots, \omega$, di un modello di sopravvivenza non parametrico, si esprime la funzione di sopravvivenza come prodotto di probabilità condizionate di sopravvivenza

$$S(x) = \frac{S(x)}{S(x-1)} \frac{S(x-1)}{S(x-2)} \cdots \frac{S(1)}{S(0)} = p_{x-1} p_{x-2} \cdots p_0 = \prod_{j<x} p_j$$

infatti

$$\frac{S(j)}{S(j-1)} = \frac{P(T_0 > j)}{P(T_0 > j-1)} = \frac{P(T_0 > j, T_0 > j-1)}{P(T_0 > j-1)} = P(T_0 > j \mid T_0 > j-1) = p_{j-1}$$

Siano

$\hat{p}_x = 1 - \hat{q}_x$ la stima di p_x

n'_x l'esposizione nella classe di età $]x, x+1]$

con $x = a, a+1, \dots, \omega - 1$

Si ottiene la seguente stima della funzione di sopravvivenza $S(x)$, $x = a, a+1, \dots, \omega$

$$\hat{S}(x) = \hat{p}_{x-1} \hat{p}_{x-2} \cdots \hat{p}_0$$

Sia

$\tilde{p}_x$ lo stimatore di p_x del quale $\hat{p}_x$ è la stima, $x = a, a+1, \dots, \omega-1$

Indichiamo con

$$\tilde{S}(x) = \prod_{j<x} \tilde{p}_j \,, \qquad x = a, a+1, \dots, \omega$$

lo stimatore del quale $\hat{S}(x)$ è la stima.

Per valutare speranza matematica e varianza dello stimatore $\tilde{S}(x)$ occorre formulare delle ipotesi sui n.a. $\tilde{p}_j$, $j = a, a+1, \dots, \omega-1$

Siano $\mathcal{I} = \{n'_a, n'_{a+1}, , \dots, n'_{\omega-1}\}$ le esposizioni nelle diverse classi di età.

Si formulano le seguenti <u>ipotesi</u> sui n.a. $\tilde{p}_x$

Condizionatamente a $\mathcal{I} = \{n'_a, n'_{a+1}, , \dots, n'_{\omega-1}\}$ *, i n.a.*

$\tilde{p}_x$, $x = a, a+1, \dots, \omega-1$ siano stocasticamente indipendenti

e siano

$$E(\tilde{p}_x | \mathcal{I}) = p_x \qquad Var(\tilde{p}_x | \mathcal{I}) = \frac{p_x\,(1-p_x)}{n'_x} \qquad x = a, a+1, \dots, \omega-1$$

Risulta allora che $\tilde{S}(x) = \prod_{j<x} \tilde{p}_j$ è uno stimatore non distorto, infatti

$$E\left(\tilde{S}(x) \middle| \mathcal{I}\right) = E\left(\prod_{j<x} \tilde{p}_j \middle| \mathcal{I} \right) = \prod_{j<x} p_j = S(x) \qquad x = a, a+1, \dots, \omega$$

La varianza dello stimatore $\tilde{S}(x)$ è

$$Var\left(\tilde{S}(x) \middle| \mathcal{I}\right) = [S(x)]^2 \left[\prod_{j<x} \left(1 + \frac{q_j}{p_j n'_j} \right) - 1 \right]$$

e può essere approssimata da

$$Var\left(\tilde{S}(x) \middle| \mathcal{I}\right) \cong [S(x)]^2 \sum_{j<x} \frac{q_j}{p_j n'_j}$$

dalla quale si ottiene la **formula di Greenwood**, che fornisce una stima della varianza dello stimatore $\tilde{S}(x)$

$$V\hat{a}r\left(\tilde{S}(x) \middle| \mathcal{I}\right) = \left[\hat{S}(x)\right]^2 \sum_{j<x} \frac{\hat{q}_j}{\hat{p}_j n'_j}$$

CONFRONTO DELLA TAVOLA DI SOPRAVVIVENZA STIMATA CON UNA TAVOLA STANDARD

Siano

$$\hat{q}_x, \qquad x = a, a+1, \dots, \omega - 1$$

le stime delle probabilità di morte q_x, $x = a, a+1, \dots, \omega - 1$, di un modello di sopravvivenza non parametrico.

Ci si pone il problema se il fenomeno della mortalità osservata nella collettività possa essere descritto da una tavola di mortalità proveniente da altre esperienze statistiche. Tale tavola viene allora detta **tavola "standard"** e la indichiamo con

$$q'_x, \qquad x = a, a+1, \dots, \omega - 1$$

Per verificare se la tavola standard accosta bene le osservazioni si sottopone a verifica d'ipotesi la seguente ipotesi nulla:

H_0 : $\qquad q_x = q'_x \qquad x = a, a+1, \dots, \omega - 1$

Con riferimento alla classe di età $]x, x+1]$ siano

D_x n.a. di decessi

d_x numero di decessi osservati

E_x il numero di esposti al rischio

Per costruire la funzione test formuliamo le seguenti ipotesi

$$E(D_x) = E_x\ q_x \qquad Var(D_x) = E_x\ q_x\ (1-q_x)$$

Sotto l'ipotesi nulla i n.a.

$$Z_x = \frac{D_x - E_x\ q'_x}{\sqrt{E_x\ q'_x (1-q'_x)}} \qquad x = a,\ a+1,\ \dots,\ \omega-1$$

hanno distribuzione approssimata $N(0,1)$. In ipotesi di indipendenza stocastica dei n.a. Z_x , $x = a,\ a+1,\ \dots,\ \omega-1$, sia ha che il n.a.

$$\sum_{x=a}^{\omega-1} (Z_x)^2$$

ha distribuzione approssimata chi-quadrato con n gradi di libertà, essendo n il numero di classi di età.

Fissato un livello di significatività α si determina $\chi^2_{n,1-\alpha}$ tale che $P\left(\chi^2_n > \chi^2_{n,1-\alpha}\right) = 1-\alpha$ si rifiuta l'ipotesi nulla se

$$\sum_{x=a}^{\omega-1} (z_x)^2 > \chi^2_{n,1-\alpha}$$

essendo $\sum_{x=a}^{\omega-1} (z_x)^2$ la determinazione osservata del n.a. $\sum_{x=a}^{\omega-1} (Z_x)^2$

Tale test potrebbe non rilevare un buon accostamento della tavola standard a i dati osservati, e quindi non fare rifiutare l'ipotesi nulla, nelle seguenti situazioni:

– esistenza di scostamenti eccessivamente elevati per alcune età, controbilanciati da scostamenti molto ridotti per altre età;
– numero eccessivo di scostamenti tutti dello stesso segno (conseguenza di una mortalità rilevata "uniformemente" maggiore o minore di quella attesa in base alla tavola standard);
– gruppi eccessivamente numerosi di età consecutive con scostamenti tutti dello stesso segno.

Per evidenziare tali problematiche si utilizzano altri test, quale per esempio il test delle deviazioni cumulate.

Sotto l'ipotesi nulla

H_0 : $\quad q_x = q'_x \qquad x = a,\, a+1, \dots, \omega - 1$

si ha

$$E\left(D_x - E_x\ q'_x\right) = 0 \qquad \text{per ogni } x = a,\, a+1, \dots, \omega - 1$$

Consideriamo le deviazioni cumulate nell'intervallo di età da x_1 a x_2

$$\sum_{x=x_1}^{x_2}\left(D_x - E_x\ q'_x\right)$$

In ipotesi di indipendenza stocastica dei n.a. D_x si ha

$$E\left[\sum_{x=x_1}^{x_2}\left(D_x - E_x\ q'_x\right)\right] = 0 \qquad Var\left[\sum_{x=x_1}^{x_2}\left(D_x - E_x\ q'_x\right)\right] = \sum_{x=x_1}^{x_2} Var\left(D_x\right) = \sum_{x=x_1}^{x_2} E_x\ q'_x (1 - q'_x)$$

La distribuzione della deviazione cumulata standardizzata

$$\frac{\sum_{x=x_1}^{x_2}\left(D_x - E_x\, q'_x\right)}{\sqrt{\sum_{x=x_1}^{x_2} E_x\, q'_x (1-q'_x)}}$$

può essere approssimata mediante una $N(0,1)$ e quindi, fissato un livello di significatività α si determina il quantile $z_{1-\alpha/2}$ della distribuzione normale standard e si rifiuta l'ipotesi nulla se

$$\left|\frac{\sum_{x=x_1}^{x_2}\left(d_x - E_x\, q'_x\right)}{\sqrt{\sum_{x=x_1}^{x_2} E_x\, q'_x (1-q'_x)}}\right| > z_{1-\alpha/2}$$

Tale analisi va ripetuta su diversi intervalli di età che evidenziano criticità nell'accostamento della tavola standard ai dati osservati.

PEREQUAZIONE MEDIANTE TAVOLE STANDARD

Se è stata rifiutata l'ipotesi nulla sulla bontà di accostamento della tavola standard ai dati, si può decidere di "adattare" la tavola standard ai dati osservati.

Si ipotizza quindi un legame funzionale tra le probabilità di morte della collettività in esame

$$q_x, \ x = a, a+1, \dots, \omega - 1,$$

e le probabilità di morte riportate nella tavola standard

$$q'_x, \ x = a, a+1, \dots, \omega - 1$$

Si assume che tale funzione

$$q_x = f\left(x, q'_x\right)$$

dipenda da alcuni parametri che devono essere stimati, per esempio con il metodo dei minimi quadrati pesati.

Siano

$$\hat{q}_x, \qquad x = a, a+1, \dots, \omega - 1$$

le stime delle probabilità di morte q_x, $x = a, a+1, \dots, \omega - 1$, di un modello di sopravvivenza non parametrico.

Dall'analisi grafica dei rapporti $\dfrac{\hat{q}_x}{q'_x}, \qquad x = a, a+1, \dots, \omega - 1$

si individua un possibile legame funzionale $q_x = f(x, q'_x)$.

Per esempio, nel caso di andamento approssimativamente lineare si può ipotizzare

$$q_x = q'_x (a + b\,x)$$

oppure

$$q_x = a\,q'_x + b$$

Più in generale si può ipotizzare anche un legame con due tavole standard $\{q'_x\}$ e $\{q''_x\}$

$$q_x = a'\,q'_x + a''\,q''_x$$

In alternativa, Lidstone ha proposto di considerare per l’analisi grafica la seguente trasformazione

$$\log\left(\frac{p'_x}{\hat{p}_x}\right), \qquad x = a, a+1, \dots, \omega-1$$

con

$$p'_x = 1 - q'_x \qquad \hat{p}_x = 1 - \hat{q}_x$$

in quanto i $\frac{p'_x}{\hat{p}_x}$ presentano un andamento più regolare rispetto ai $\frac{\hat{q}_x}{q'_x}$

Se evidenziano un andamento approssimativamente costante si può ipotizzare

$$\log\left(\frac{p'_x}{p_x}\right) = c \quad \Rightarrow \quad q_x = 1 - \frac{p'_x}{e^c}$$

Dopo avere individuato la funzione $f(x;\, a, b, \ldots)$ che esprime il legame tra q_x e q'_x, per la stima dei parametri della funzione si può utilizzare, per esempio, il metodo dei minimi quadrati.

$$\min_{a,b,\ldots} F(a,b,\ldots) \qquad \text{con } F(a,b,\ldots) = \sum_{x=a}^{\omega-1} w_x \left[\hat{q}_x - f(x;\, a,b,\ldots)\right]^2$$

essendo

$w_x = 1$ nel caso di minimi quadrati non pesati

$w_x = \dfrac{n'_x}{\hat{q}_x}$ nel caso di minimi quadrati pesati

Si noti che nell'ipotesi $Var(D_x) = n'_x\, q_x\, (1 - q_x)$ si ha

$$Var\left(\frac{D_x}{n'_x}\right) = \frac{q_x\,(1 - q_x)}{n'_x} \cong \frac{q_x}{n'_x}$$

quindi il peso $w_x = \dfrac{n'_x}{\hat{q}_x}$ è approssimativamente pari al reciproco della varianza dello stimatore di q_x

Se la funzione $f(x;a,b,\ldots)$ è lineare, le stime dei minimi quadrati dei parametri si ottengono agevolmente risolvendo un sistema lineare.

Sia u_x tale che il legame funzionale tra q_x e q'_x sia espresso mediante la funzione lineare

$$f(x;a,b) = a + b\,x$$

Per esempio $u_x = \dfrac{q_x}{q'_x}$ nel caso in cui sia $q_x = q'_x(a + b\,x)$

Si ha allora

$$\min_{a,b} F(a,b) \qquad \text{con } F(a,b) = \sum_{x=a}^{\omega-1} w_x [\hat{u}_x - (a + b\,x)]^2$$

essendo $\hat{u}_x = \dfrac{\hat{q}_x}{q'_x}$

Si ottiene:

$$\begin{cases} \dfrac{\partial F}{\partial a} = 0 \\ \dfrac{\partial F}{\partial b} = 0 \end{cases} \Leftrightarrow \begin{cases} \displaystyle\sum_{x=a}^{\omega-1} w_x [\hat{u}_x - (a + b\,x)] = 0 \\ \displaystyle\sum_{x=a}^{\omega-1} w_x [\hat{u}_x - (a + b\,x)]x = 0 \end{cases}$$