

LA STIMA DI KAPLAN-MEIER PER LA FUNZIONE DI SOPRAVVIVENZA IN UN MODELLO DISCRETO

Rilevazione trasversale: per ogni individuo si osserva l'istante iniziale; alcuni individui danno luogo ad osservazioni censurate a destra. Quindi i dati sono incompleti.

Osservazioni: n individui osservati; m decessi osservati alle età $t_1 < t_2 < \dots < t_k$

d_1, d_2, \dots, d_k numeri di decessi osservati alle età $t_1 < t_2 < \dots < t_k$

$$d_1 + d_2 + \dots + d_k = m$$

$n - m$ numero di osservazioni censurate

Sia t_0 l'età di ingresso in osservazione corrispondente all'istante iniziale

Si ripartisce l'intervallo $[t_0, +\infty[$ negli intervalli

$$[t_j, t_{j+1}[\quad j = 0, 1, \dots, k \quad \text{con} \quad t_{k+1} = +\infty$$

Sia $c_j, j = 0, 1, \dots, k$, il numero di osservazioni censurate nell'intervallo $[t_j, t_{j+1}[$

$$c_0 + c_1 + \dots + c_k = n - m$$

La stima di Kaplan-Meier per la funzione di sopravvivenza in un modello discreto

Con riferimento all'intervallo $[t_j, t_{j+1}[$, $j = 0, 1, \dots, k$, indichiamo con

$$t_{j1} \leq t_{j2} \leq \dots \leq t_{jc_j}$$

le età di uscita delle c_j osservazioni censurate.

Sia

$$n_j, \quad j = 1, \dots, k,$$

il numero di individui in vita all'età t_j che danno luogo ai d_j decessi

Si ha

$$n_j = (d_j + d_{j+1} + \dots + d_k) + (c_j + c_{j+1} + \dots + c_k) \quad j = 1, \dots, k$$

Obiettivo: stimare la funzione di ripartizione $F(t)$
del n.a. T durata aleatoria di vita dall'età t_0

Disponendo delle osservazioni: decesso all'età t_j , $j = 1, \dots, k$

osservazione censurata all'età

$$t_{ji}, \quad j = 0, \dots, k, \quad i = 1, \dots, c_j$$

La stima di Kaplan-Meier per la funzione di sopravvivenza in un modello discreto

Nell'ipotesi che i numeri aleatori delle durate osservate siano stocasticamente indipendenti ed ugualmente distribuiti con funzione di ripartizione $F(t)$, la verosimiglianza delle osservazioni è

$$L = \prod_{j=1}^k [F(t_j) - F(t_j^-)]^{d_j} \prod_{j=0}^k \prod_{i=1}^{c_j} [1 - F(t_{ji})]$$

essendo

$$F(t_j^-) = \lim_{t \rightarrow t_j^-} F(t) \quad j = 1, \dots, k$$

Obiettivo: determinare la stima $\hat{F}(t)$ di massima verosimiglianza della funzione di ripartizione $F(t)$

Osservazione: $L \neq 0 \Leftrightarrow F(t_j) - F(t_j^-) > 0 \quad \forall j = 1, \dots, k$
inoltre $1 - F(t_{ji})$ è massimo se $F(t_{ji}) = F(t_j) \quad \forall j = 1, \dots, k, \quad \forall i = 1, \dots, c_j$

Quindi la verosimiglianza diventa

$$L = \prod_{j=1}^k [F(t_j) - F(t_j^-)]^{d_j} \cdot \prod_{j=0}^k [1 - F(t_j)]^{c_j}$$

Osservazioni:

- $F(t)$ è la funzione di ripartizione di un n.a. con distribuzione discreta e determinazioni t_1, t_2, \dots, t_k

- per stimare $F(t)$ si tratta quindi di stimare le probabilità

$$q_j = F(t_j) - F(t_j^-) \quad j = 1, \dots, k$$

- ovvero di stimare la funzione di rischio

$$\lambda_j = \lambda(t_j) = \frac{q_j}{S(t_{j-1})} = \frac{F(t_j) - F(t_j^-)}{1 - F(t_{j-1})} \quad j = 1, \dots, k$$

essendo

$$S(t_j) = 1 - F(t_j) = \prod_{t_i \leq t_j} (1 - \lambda(t_i)) = \begin{cases} 1 & \text{se } j = 0 \\ (1 - \lambda_1)(1 - \lambda_2) \dots (1 - \lambda_j) & \text{se } j = 1, \dots, k \end{cases}$$

Si ottiene allora la seguente espressione per la verosimiglianza

$$L = \prod_{j=1}^k [\lambda(t_j)]^{d_j} \cdot \prod_{j=1}^k [1 - F(t_{j-1})]^{d_j} \cdot \prod_{j=0}^k [1 - F(t_j)]^{c_j} = \prod_{j=1}^k [\lambda(t_j)]^{d_j} \cdot \prod_{j=1}^k [1 - \lambda(t_j)]^{n_j - d_j}$$

La stima di Kaplan-Meier per la funzione di sopravvivenza in un modello discreto

Data la verosimiglianza

$$L = \prod_{j=1}^k \lambda_j^{d_j} [1 - \lambda_j]^{n_j - d_j}$$

Osservazione

$\lambda_j^{d_j} [1 - \lambda_j]^{n_j - d_j}$ è proporzionale alla verosimiglianza di un n.a. con distribuzione Binomiale(n_j, λ_j)

Sia D_j il n.a. dei decessi all'età t_j a partire dagli n_j individui in vita all'età t_j prima che si verificano i d_j decessi

Si ipotizza per D_j la distribuzione Binomiale(n_j, λ_j)

La stima di massima verosimiglianza di λ_j è

$$\hat{\lambda}_j = \frac{d_j}{n_j}$$

La stima di Kaplan-Meier per la funzione di sopravvivenza in un modello discreto

La stima di Kaplan-Meier per la funzione di sopravvivenza è allora

$$\hat{S}(t) = \prod_{t_j \leq t} (1 - \hat{\lambda}_j) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

Osservazione: poiché

$$S(t) = \prod_{t_j \leq t} (1 - \lambda(t_j)) = \prod_{t_j \leq t} P(T > t_j | T > t_{j-1})$$

si può allora interpretare la formula

$$\hat{S}(t) = \prod_{t_j \leq t} (1 - \hat{\lambda}_j) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

come prodotto delle stime delle probabilità condizionate di sopravvivenza.

La stima di massima verosimiglianza, di Kaplan-Meier, della funzione di ripartizione $F(t)$ è allora

$$\hat{F}(t) = 1 - \hat{S}(t) = 1 - \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

La stima di Kaplan-Meier per la funzione di sopravvivenza in un modello discreto

Esempio

j	t_j	n_j	d_j	c_j	$1 - d_j/n_j$	$\hat{S}(t_j)$
0	0	8	0	1	1	1
1	1,5	7	1	0	0,857143	0,857143
2	2	6	2	1	0,666667	0,571429
3	3,5	3	1	2	0,666667	0,380952

Osservazioni

- La stima di Kaplan-Meier $\hat{F}(t)$ è definita per $t \leq t_k$
- Se l'età massima osservata è un dato censurato, cioè se $c_k > 0$, la $\hat{F}(t)$ non è definita per $t > t_k$ ed è $\hat{F}(t_k) < 1$; una possibilità è definire $\hat{F}(t) = \hat{F}(t_k)$ per $t > t_k$ ma ciò comporta la presenza di masse aderenti a $+\infty$
- La stima di Kaplan-Meier $\hat{F}(t)$ può essere definita anche in presenza di osservazioni troncate a sinistra, cioè di osservazioni per le quali non si è osservato l'istante iniziale essendo entrate in osservazione ad un'età maggiore di t_0 . In tal caso $\hat{S}(t)$ è la stima della funzione di sopravvivenza condizionata alla età minima di ingresso osservata t_{\min} , essendo $t_0 \leq t_{\min} < t_1$; quindi è una stima di $P(T > t | T > t_{\min}) = S(t) / S(t_{\min})$

La stima di Kaplan-Meier per la funzione di sopravvivenza in un modello discreto

Proprietà dello stimatore di Kaplan-Meier

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j} \right) \quad \text{è la stima di Kaplan-Meier di } S(t), t \geq 0$$

Sia $S(t) = \prod_{t_j \leq t} p_j$ con $p_j = 1 - \lambda(t_j) = P(T > t_j | T > t_{j-1})$

si ha

$$\hat{S}(t) = \prod_{t_j \leq t} \hat{p}_j \quad \text{essendo} \quad \hat{p}_j = 1 - \frac{d_j}{n_j}$$

dove \hat{p}_j è una stima della probabilità $P(T > t_j | T > t_{j-1})$, $j = 1, \dots, k$

Indichiamo con

$$\tilde{S}(t) = \prod_{t_j \leq t} \tilde{p}_j, \quad t \geq 0$$

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

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

La stima di Kaplan-Meier per la funzione di sopravvivenza in un modello discreto

Osservazione

Sia

D_j il n.a. dei decessi da una popolazione di n_j individui in vita all'età t_j

nell'ipotesi che D_j abbia distribuzione Binomiale(n_j, q_j) si ha

$$E(D_j) = n_j q_j \qquad \text{Var}(D_j) = n_j q_j (1 - q_j)$$

si ha inoltre che il n.a.

$\frac{D_j}{n_j}$ ha distribuzione Binomiale scalata ed è

$$E\left(\frac{D_j}{n_j}\right) = q_j \qquad \text{Var}\left(\frac{D_j}{n_j}\right) = \frac{q_j (1 - q_j)}{n_j}$$

La stima di Kaplan-Meier per la funzione di sopravvivenza in un modello discreto

Sia $\mathcal{J} = \{n_0, n_1, \dots, n_k\}$ lo stato di informazione sugli individui presenti alle età $t_j, j = 1, \dots, k$ subito prima che si osservino i decessi $d_j, j = 1, \dots, k$

Si formulano le seguenti ipotesi condizionate a $\mathcal{J} = \{n_0, n_1, \dots, n_k\}$ sui n.a.

$$\tilde{p}_j = 1 - \frac{D_j}{n_j}, \quad j = 1, \dots, k$$

Ipotesi

Condizionatamente a $\mathcal{J} = \{n_0, n_1, \dots, n_k\}$, i n.a.

$$\tilde{p}_j = 1 - \frac{D_j}{n_j}, \quad j = 1, \dots, k \quad \text{siano stocasticamente indipendenti}$$

e siano

$$E(\tilde{p}_j | \mathcal{I}) = p_j \quad \text{Var}(\tilde{p}_j | \mathcal{I}) = \frac{p_j (1 - p_j)}{n_j} \quad j = 1, \dots, k$$

La stima di Kaplan-Meier per la funzione di sopravvivenza in un modello discreto

Risulta allora che $\tilde{S}(t) = \prod_{t_j \leq t} \tilde{p}_j$ è uno stimatore non distorto, infatti

$$E(\tilde{S}(t)|I) = E\left(\prod_{t_j \leq t} \tilde{p}_j | I\right) = \prod_{t_j \leq t} p_j = S(t) \quad t \geq 0$$

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

$$\text{Var}(\tilde{S}(t)|I) = [S(t)]^2 \left[\prod_{t_j \leq t} \left(1 + \frac{1-p_j}{p_j n_j}\right) - 1 \right]$$

e può essere approssimata da

$$\text{Var}(\tilde{S}(t)|I) \cong [S(t)]^2 \sum_{t_j \leq t} \frac{1-p_j}{p_j n_j}$$

dalla quale si ottiene la **formula di Greenwood**, che fornisce una stima della varianza dello stimatore di Kaplan-Meier

$$\hat{\text{Var}}(\tilde{S}(t)|I) = \left[\prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right) \right]^2 \sum_{t_j \leq t} \frac{d_j}{(n_j - d_j)n_j}$$

LA STIMA DI NELSON-AALEN PER LA FUNZIONE DI SOPRAVVIVENZA

Fornisce una formula approssimata per la stima della funzione di sopravvivenza di Kaplan-Meier.

Dalla stima di Kaplan-Meier della funzione di sopravvivenza $\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$ $t \geq 0$

si ottiene

$$\log(\hat{S}(t)) = \sum_{t_j \leq t} \log\left(1 - \frac{d_j}{n_j}\right) \quad t \geq 0$$

Considerando l'approssimazione lineare della funzione $f(x) = \log(1 - x)$ si ottengono le

seguenti approssimazioni: $\log\left(1 - \frac{d_j}{n_j}\right) \cong -\frac{d_j}{n_j}$

Si ottiene allora la seguente stima della funzione di sopravvivenza, detta **stima di Nelson-Aalen**

$$\hat{S}(t) = \exp\left(-\sum_{t_j \leq t} \frac{d_j}{n_j}\right) \quad t \geq 0$$