

STIMA DI MODELLI DI SOPRAVVIVENZA PARAMETRICI

- Stima di modelli di sopravvivenza parametrici – uscite per morte e altra causa
- Modelli di sopravvivenza parametrici con componenti di regressione

STIMA DI MODELLI DI SOPRAVVIVENZA PARAMETRICI USCITE PER MORTE E PER ALTRA CAUSA

Supponiamo di avere osservato n individui e di disporre di dati individuali esatti, riassunti per ogni individuo i dal vettore delle età

$$(y_i, z_i, \theta_i, \phi_i,) \quad i = 1, \dots, n$$

essendo

y_i l'età esatta di ingresso in osservazione

z_i l'età esatta di uscita pianificata

θ_i l'età esatta di uscita per morte ($\theta_i = 0$ se l'individuo i non è uscito per morte)

ϕ_i l'età esatta di uscita per altra causa ($\phi_i = 0$ se l'individuo i non è uscito per altra causa)

Obiettivo: stimare la funzione di sopravvivenza di un **modello di sopravvivenza parametrico**

Modello di sopravvivenza a due cause di eliminazione: morte ed altra causa

Sia

T_x durata di permanenza nella collettività per un individuo presente all'età x

$$T_x = \min(T_x^{(d)}, T_x^{(w)})$$

essendo

$T_x^{(d)}$ durata di permanenza nella collettività finché non si ha l'uscita per morte, per un individuo presente nella collettività all'età x

$T_x^{(w)}$ durata di permanenza nella collettività finché non si ha l'uscita per altra causa, per un individuo presente nella collettività all'età x

Siano

$S^{(d)}(t) = P(T_0^{(d)} > t)$ funzione di sopravvivenza (parametrica) relativa al n.a. $T_0^{(d)}$

$S^{(w)}(t) = P(T_0^{(w)} > t)$ funzione di sopravvivenza (parametrica) relativa al n.a. $T_0^{(w)}$

Ipotesi: uscite non informative

$$\mu^{(d)}(t) = a\mu^{(d)}(t) \quad t \geq 0$$

$$\mu^{(w)}(t) = a\mu^{(w)}(t) \quad t \geq 0$$

essendo

$$\mu^{(d)}(x+t) = \lim_{\Delta t \rightarrow 0} \frac{P(T_x^{(d)} \leq t + \Delta t | T_x^{(d)} > t)}{\Delta t}$$

$$\mu^{(w)}(x+t) = \lim_{\Delta t \rightarrow 0} \frac{P(T_x^{(w)} \leq t + \Delta t | T_x^{(w)} > t)}{\Delta t}$$

le intensità marginali di eliminazione e

$$a\mu^{(d)}(x+t) = \lim_{\Delta t \rightarrow 0} \frac{P(T_x \leq t + \Delta t, C = 1 | T_x > t)}{\Delta t}$$

$$a\mu^{(w)}(x+t) = \lim_{\Delta t \rightarrow 0} \frac{P(T_x \leq t + \Delta t, C = 2 | T_x > t)}{\Delta t}$$

le intensità di uscita per le varie cause, dove $C = 1$ denota l'evento "uscita per morte" e $C = 2$ denota l'evento uscita per altra causa

Stima di modelli di sopravvivenza parametrici – uscite per morte e per altra causa

In ipotesi di uscite non informative sussiste la **relazione di Karup**

$${}_t p_x^{(\tau)} = {}_t p_x^{(d)} {}_t p_x^{(w)}$$

essendo

$${}_t p_x^{(d)} = \exp\left(-\int_0^t \mu^{(d)}(x+u) du\right) = \frac{S^{(d)}(x+t)}{S^{(d)}(x)}$$

$${}_t p_x^{(w)} = \exp\left(-\int_0^t \mu^{(w)}(x+u) du\right) = \frac{S^{(w)}(x+t)}{S^{(w)}(x)}$$

Stima di modelli di sopravvivenza parametrici – uscite per morte e per altra causa

Stima con il metodo della massima verosimiglianza

Per scrivere la verosimiglianza delle osservazioni

$$(y_i, z_i, \theta_i, \phi_i), \quad i = 1, \dots, n$$

definiamo, per ogni $i = 1, \dots, n$, i n.a.

$T^{(i)}$ durata aleatoria di permanenza dell'individuo i nella collettività

Nota: $T^{(i)}$ ha determinazioni $]0, z_i - y_i]$

Si definiscono

$$S = \{i \mid \text{l'individuo } i \text{ è presente nella collettività all'età } z_i\} \quad \textit{survival}$$

$$D = \{i \mid \text{l'individuo } i \text{ esce per morte all'età } \theta_i\} \quad \textit{death}$$

$$W = \{i \mid \text{l'individuo } i \text{ esce per altra causa all'età } \phi_i\} \quad \textit{withdrawal}$$

Stima di modelli di sopravvivenza parametrici – uscite per morte e per altra causa

Se l'individuo i è presente nella collettività all'età di uscita pianificata:

$$T^{(i)} = z_i - y_i \quad P(T^{(i)} = z_i - y_i) = {}_{z_i - y_i}P_{y_i}^{(\tau)} = {}_{z_i - y_i}P_{y_i}'^{(d)} \cdot {}_{z_i - y_i}P_{y_i}'^{(w)} = \frac{S^{(d)}(z_i)}{S^{(d)}(y_i)} \cdot \frac{S^{(w)}(z_i)}{S^{(w)}(y_i)}$$

Se l'individuo i esce per morte all'età esatta θ_i :

$$T^{(i)} = \theta_i - y_i \quad f_{T,C}(\theta_i - y_i, 1) = {}_{\theta_i - y_i}P_{y_i}^{(\tau)} \cdot \mu^{(d)}(\theta_i) = \frac{S^{(d)}(\theta_i)}{S^{(d)}(y_i)} \cdot \frac{S^{(w)}(\theta_i)}{S^{(w)}(y_i)} \cdot \mu^{(d)}(\theta_i)$$

Se l'individuo i esce per altra causa all'età esatta ϕ_i :

$$T^{(i)} = \phi_i - y_i \quad f_{T,C}(\phi_i - y_i, 2) = {}_{\phi_i - y_i}P_{y_i}^{(\tau)} \cdot \mu^{(w)}(\phi_i) = \frac{S^{(d)}(\phi_i)}{S^{(d)}(y_i)} \cdot \frac{S^{(w)}(\phi_i)}{S^{(w)}(y_i)} \cdot \mu^{(w)}(\phi_i)$$

In ipotesi di indipendenza stocastica dei n.a. $T^{(i)}$ la verosimiglianza delle osservazioni è

$$\begin{aligned}
 L &= \prod_{i \in S} P(T^{(i)} = z_i - y_i) \cdot \prod_{i \in D} f_{T,C}(\theta_i - y_i, 1) \cdot \prod_{i \in W} f_{T,C}(\phi_i - y_i, 2) \\
 &= \prod_{i \in S} p_{y_i}^{(z_i - y_i)} \cdot \prod_{i \in D} p_{y_i}^{(\theta_i - y_i)} \cdot \mu^{(d)}(\theta_i) \cdot \prod_{i \in W} p_{y_i}^{(\phi_i - y_i)} \cdot \mu^{(w)}(\phi_i) \\
 &= \prod_{i \in S} \frac{S^{(d)}(z_i)}{S^{(d)}(y_i)} \cdot \frac{S^{(w)}(z_i)}{S^{(w)}(y_i)} \\
 &\quad \cdot \prod_{i \in D} \frac{S^{(d)}(\theta_i)}{S^{(d)}(y_i)} \cdot \frac{S^{(w)}(\theta_i)}{S^{(w)}(y_i)} \cdot \mu^{(d)}(\theta_i) \\
 &\quad \cdot \prod_{i \in W} \frac{S^{(d)}(\phi_i)}{S^{(d)}(y_i)} \cdot \frac{S^{(w)}(\phi_i)}{S^{(w)}(y_i)} \cdot \mu^{(w)}(\phi_i)
 \end{aligned}$$

Stima di modelli di sopravvivenza parametrici – uscite per morte e per altra causa

Poiché

$$S^{(d)}(\theta_i) \cdot \mu^{(d)}(\theta_i) = f^{(d)}(\theta_i) \qquad S^{(w)}(\phi_i) \cdot \mu^{(w)}(\phi_i) = f^{(w)}(\phi_i)$$

si ha

$$\begin{aligned} L &= \prod_{i \in S} \frac{S^{(d)}(z_i)}{S^{(d)}(y_i)} \cdot \prod_{i \in D} \frac{S^{(d)}(\theta_i)}{S^{(d)}(y_i)} \cdot \mu^{(d)}(\theta_i) \cdot \prod_{i \in W} \frac{S^{(d)}(\phi_i)}{S^{(d)}(y_i)} \cdot \\ &\quad \cdot \prod_{i \in S} \frac{S^{(w)}(z_i)}{S^{(w)}(y_i)} \cdot \prod_{i \in D} \frac{S^{(w)}(\theta_i)}{S^{(w)}(y_i)} \cdot \prod_{i \in W} \frac{S^{(w)}(\phi_i)}{S^{(w)}(y_i)} \cdot \mu^{(w)}(\phi_i) \\ &= \prod_{i \in S} \frac{S^{(d)}(z_i)}{S^{(d)}(y_i)} \cdot \prod_{i \in D} \frac{f^{(d)}(\theta_i)}{S^{(d)}(y_i)} \cdot \prod_{i \in W} \frac{S^{(d)}(\phi_i)}{S^{(d)}(y_i)} \cdot \\ &\quad \cdot \prod_{i \in S} \frac{S^{(w)}(z_i)}{S^{(w)}(y_i)} \cdot \prod_{i \in D} \frac{S^{(w)}(\theta_i)}{S^{(w)}(y_i)} \cdot \prod_{i \in W} \frac{f^{(w)}(\phi_i)}{S^{(w)}(y_i)} \end{aligned}$$

Stima di modelli di sopravvivenza parametrici – uscite per morte e per altra causa

Quindi

$$L = L^{(d)} \cdot L^{(w)}$$

con

$$L^{(d)} = \prod_{i \in S} \frac{S^{(d)}(z_i)}{S^{(d)}(y_i)} \cdot \prod_{i \in D} \frac{f^{(d)}(\theta_i)}{S^{(d)}(y_i)} \cdot \prod_{i \in W} \frac{S^{(d)}(\phi_i)}{S^{(d)}(y_i)}$$

e

$$L^{(w)} = \prod_{i \in S} \frac{S^{(w)}(z_i)}{S^{(w)}(y_i)} \cdot \prod_{i \in D} \frac{S^{(w)}(\theta_i)}{S^{(w)}(y_i)} \cdot \prod_{i \in W} \frac{f^{(w)}(\phi_i)}{S^{(w)}(y_i)}$$

Quindi per stimare la funzione di sopravvivenza $S^{(d)}(t)$ si porrà

$$\max L^{(d)}$$

mentre per stimare la funzione di sopravvivenza $S^{(w)}(t)$ si porrà

$$\max L^{(w)}$$

Stima di modelli di sopravvivenza parametrici – uscite per morte e per altra causa

Stima di massima verosimiglianza della funzione di sopravvivenza $S^{(d)}(t)$

$$L^{(d)} = \prod_{i \in S} \frac{S^{(d)}(z_i)}{S^{(d)}(y_i)} \cdot \prod_{i \in W} \frac{S^{(d)}(\phi_i)}{S^{(d)}(y_i)} \cdot \prod_{i \in D} \frac{f^{(d)}(\theta_i)}{S^{(d)}(y_i)}$$

dove

$$f^{(d)}(t) = -\frac{d}{dt} S^{(d)}(t)$$

Si nota che le informazioni sulle uscite per altra causa sono trattate come le informazioni sulla sopravvivenza all'età di uscita pianificata

Se il modello di sopravvivenza parametrico per la durata aleatoria di vita $T_0^{(d)}$ è assegnato mediante l'intensità istantanea di mortalità $\mu^{(d)}(t)$ la verosimiglianza può essere scritta nel modo seguente

$$L^{(d)} = \prod_{i \in S} \exp\left(-\int_{y_i}^{z_i} \mu^{(d)}(t) dt\right) \cdot \prod_{i \in W} \exp\left(-\int_{y_i}^{\phi_i} \mu^{(d)}(t) dt\right) \cdot \prod_{i \in D} \left[\exp\left(-\int_{y_i}^{\theta_i} \mu^{(d)}(t) dt\right) \mu^{(d)}(\theta_i) \right]$$

Stima di modelli di sopravvivenza parametrici – uscite per morte e per altra causa

Con riferimento all' i -esimo individuo osservato, indichiamo con t_i la determinazione del n.a. $T^{(i)}$, si ha

$$t_i = \begin{cases} z_i - y_i & \text{se } i \in S \\ \phi_i - y_i & \text{se } i \in W \\ \theta_i - y_i & \text{se } i \in D \end{cases}$$

Consideriamo inoltre l'indicatore di evento

$$d_i = \begin{cases} 0 & \text{se } i \in S \cup W \\ 1 & \text{se } i \in D \end{cases}$$

la verosimiglianza $L^{(d)}$ relativa agli n individui osservati può allora essere scritta come segue

$$L^{(d)} = \prod_{i=1}^n \left[\exp\left(-\int_0^{t_i} \mu^{(d)}(y_i + s) ds\right) \left(\mu^{(d)}(y_i + t_i)\right)^{d_i} \right]$$

Stima di modelli di sopravvivenza parametrici – uscite per morte e per altra causa

Indicata con

$$H_y(t) = \int_0^t \mu^{(d)}(y+s) ds$$

la funzione di rischio integrata, la verosimiglianza $L^{(d)}$ diventa

$$L^{(d)} = \prod_{i=1}^n \left[\exp(-H_{y_i}(t_i)) \left(\mu^{(d)}(y_i + t_i) \right)^{d_i} \right]$$

e quindi si ottiene la seguente espressione per la log-verosimiglianza

$$\log L^{(d)} = \sum_{i=1}^n -H_{y_i}(t_i) + \sum_{i=1}^n d_i \log \left(\mu^{(d)}(y_i + t_i) \right)$$

Per ottenere quindi la stima di massima verosimiglianza di un modello di sopravvivenza parametrico, disponendo di dati individuali esatti, è sufficiente specificare la struttura della intensità istantanea di mortalità $\mu^{(d)}(t)$ e della relativa funzione di rischio integrato $H_y(t)$

Stima di modelli di sopravvivenza parametrici – uscite per morte e per altra causa

S.J. Richards (2012), “A handbook of parametric survival models for actuarial use”,
Scandinavian Actuarial Journal

V. Table 1. Some actuarial mortality laws and their corresponding integrated hazard functions.

MODELLI DI SOPRAVVIVENZA PARAMETRICI CON COMPONENTI DI REGRESSIONE

Supponiamo di avere osservato n individui e di disporre di dati individuali esatti.

Per ogni individuo i sono inoltre note le determinazioni di un insieme di variabili esplicative invarianti nel tempo (variabili concomitanti). I dati osservati sono allora

$$(y_i, z_i, \theta_i, \phi_i, \mathbf{x}_i) \quad i = 1, \dots, n$$

essendo

y_i l'età esatta di ingresso in osservazione

z_i l'età esatta di uscita pianificata

θ_i l'età esatta di uscita per morte ($\theta_i = 0$ se l'individuo i non è uscito per morte)

ϕ_i l'età esatta di uscita per altra causa ($\phi_i = 0$ se l'individuo i non è uscito per altra causa)

\mathbf{x}_i determinazioni di un vettore di variabili esplicative (es. sesso, stato di salute, ecc.)

Obiettivo: stimare un **modello di sopravvivenza parametrico** con delle variabili esplicative introdotte nel modello per mezzo di una **struttura di regressione**

Con riferimento al n.a.

$T^{(i)}$ durata aleatoria di permanenza dell'individuo i nella collettività

indichiamo con

$S^{(d)}(t; \mathbf{x}_i)$ la funzione di sopravvivenza

$\mu^{(d)}(t; \mathbf{x}_i)$ la funzione di rischio (intensità istantanea di mortalità)

$f^{(d)}(t; \mathbf{x}_i) = S^{(d)}(t; \mathbf{x}_i) \cdot \mu^{(d)}(t; \mathbf{x}_i)$ la funzione di densità

relative alla durata aleatoria di vita dell'individuo i , caratterizzato dal vettore delle variabili esplicative (variabili concomitanti) \mathbf{x}_i

Modelli di sopravvivenza parametrici con componenti di regressione

Indicata con t_i la determinazione del n.a. $T^{(i)}$, relativo all' i -esimo individuo osservato

$$t_i = \begin{cases} z_i - y_i & \text{se } i \in S \\ \phi_i - y_i & \text{se } i \in W \\ \theta_i - y_i & \text{se } i \in D \end{cases}$$

e dato l'indicatore di evento $d_i = \begin{cases} 0 & \text{se } i \in S \cup W \\ 1 & \text{se } i \in D \end{cases}$

si ottiene la seguente espressione della verosimiglianza $L^{(d)}$ relativa agli n individui osservati

$$L^{(d)} = \prod_{i=1}^n \left[\frac{S^{(d)}(y_i + t_i; \mathbf{x}_i)}{S^{(d)}(y_i; \mathbf{x}_i)} \left(\mu^{(d)}(y_i + t_i; \mathbf{x}_i) \right)^{d_i} \right]$$

ovvero

$$L^{(d)} = \prod_{i=1}^n \left[\exp \left(- \int_0^{t_i} \mu^{(d)}(y_i + s; \mathbf{x}_i) ds \right) \left(\mu^{(d)}(y_i + t_i; \mathbf{x}_i) \right)^{d_i} \right]$$

Modello di Cox o proportional hazard model

Si considera una particolare ipotesi sulla funzione di rischio

$$\mu^{(d)}(t; \mathbf{x}_i) = \mu_0^{(d)}(t) \exp(\mathbf{x}'_i \boldsymbol{\beta})$$

in cui si ha una componente $\mu_0^{(d)}(t)$ che esprime l'intensità istantanea di mortalità base, comune a tutti gli individui osservati, ed una componente $\exp(\mathbf{x}'_i \boldsymbol{\beta})$ dipendente dalle variabili concomitanti di ciascun individuo.

Sostituendo nella verosimiglianza si ottiene

$$L^{(d)} = \prod_{i=1}^n \left[\exp \left(- \exp(\mathbf{x}'_i \boldsymbol{\beta}) \int_0^{t_i} \mu_0^{(d)}(y_i + s) ds \right) \left(\mu_0^{(d)}(y_i + t_i) \exp(\mathbf{x}'_i \boldsymbol{\beta}) \right)^{d_i} \right]$$

Modelli di sopravvivenza parametrici con componenti di regressione

Ponendo

$$\mu_i = \exp(\mathbf{x}'_i \boldsymbol{\beta}) \int_0^{t_i} \mu_0^{(d)}(y_i + s) ds$$

la verosimiglianza diventa

$$L^{(d)} = \prod_{i=1}^n \left[\exp\left(-\exp(\mathbf{x}'_i \boldsymbol{\beta}) \int_0^{t_i} \mu_0^{(d)}(y_i + s) ds\right) \left(\exp(\mathbf{x}'_i \boldsymbol{\beta}) \int_0^{t_i} \mu_0^{(d)}(y_i + s) ds \right)^{d_i} \left(\frac{\mu_0^{(d)}(y_i + t_i)}{\int_0^{t_i} \mu_0^{(d)}(y_i + s) ds} \right)^{d_i} \right]$$
$$= \prod_{i=1}^n \left[\exp(-\mu_i) (\mu_i)^{d_i} \left(\frac{\mu_0^{(d)}(y_i + t_i)}{\int_0^{t_i} \mu_0^{(d)}(y_i + s) ds} \right)^{d_i} \right]$$

Nella verosimiglianza

$$L^{(d)} = \prod_{i=1}^n e^{-\mu_i} (\mu_i)^{d_i} \left(\frac{\mu_0^{(d)}(y_i + t_i)}{\int_0^{t_i} \mu_0^{(d)}(y_i + s) ds} \right)^{d_i}$$

il fattore

$$e^{-\mu_i} (\mu_i)^{d_i}$$

può essere visto come il nucleo di una distribuzione di Poisson di parametro μ_i che dipende dai parametri di regressione β ,

invece, il fattore

$$\left(\frac{\mu_0^{(d)}(y_i + t_i)}{\int_0^{t_i} \mu_0^{(d)}(y_i + s) ds} \right)^{d_i}$$

non dipende dai parametri di regressione

Modelli di sopravvivenza parametrici con componenti di regressione

Nota la funzione di intensità istantanea di mortalità base $\mu_0^{(d)}(t)$, per la stima di β la $L^{(d)}$ è equivalente alla verosimiglianza di un GLM con

- variabili risposta con distribuzioni di Poisson,
- funzione di collegamento logaritmo
- valori osservati delle variabili risposta: d_i con $i = 1, \dots, n$
- previsore lineare della i -esima osservazione:

$$\mathbf{x}'_i \boldsymbol{\beta} + \log \left(\int_0^{t_i} \mu_0^{(d)}(y_i + s) ds \right)$$

con $\log \left(\int_0^{t_i} \mu_0^{(d)}(y_i + s) ds \right)$ un termine offset

Modelli di sopravvivenza parametrici con componenti di regressione

Tipicamente per la funzione di intensità istantanea di mortalità base $\mu_0^{(d)}(t)$ si assume un modello di sopravvivenza parametrico, per esempio il modello di Weibull

Per stimare allora tutti i parametri sono proposti in letteratura algoritmi iterativi che alternano al passo di stima dei parametri di regressione, un passo per la stima del modello di sopravvivenza base.

RIFERIMENTI BIBLIOGRAFICI

L. Fahrmeir, G. Tutz, Multivariate statistical modelling based on Generalized Linear Models, Springer, 1994 (Cap. 9)

D. London, Survival models and their estimation, Actex publications, 1997 (Cap. 8)

S. J. Richards (2012), A handbook of parametric survival model for actuarial use, Scandinavian Actuarial Journal, 2012:4, 233-257