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

$$\left(y_i, z_i, \theta_i, \phi_i, \boldsymbol{x}_i\right) \qquad\qquad i = 1, \ldots, n$$

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)

$\boldsymbol{x}_i$ 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

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

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

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

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

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

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

ovvero

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

Modello di Cox o proportional hazard model

Si considera una particolare ipotesi sulla funzione di rischio

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

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 $\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ dipendente dalle variabili concomitanti di ciascun individuo.

Sostituendo nella verosimiglianza si ottiene

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

Ponendo

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

la verosimiglianza diventa

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

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

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

Nota la funzione di intensità istantanea di mortalità base $\mu_0^{(d)}(t)$, per la stima di $\boldsymbol{\beta}$ 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 $i = 1, \ldots, n$
- previsore lineare della i-esima osservazione:

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

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

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.