METODI STATISTICI PER LA BIOINGEGNERIA (B)

PARTE 15: IL MODELLO DI COX

A.A. 2025-2026

Prof. Martina Vettoretti

FORMULARIO DELL'ANALISI DI SOPRAVVIVENZA

- Funzione sopravvivenza: $S(t) = 1 - F(t) = \int_t^{+\infty} f(u)du$
- Densità del tempo di sopravvivenza: $f(t) = -\frac{dS(t)}{dt}$
- Hazard function: $h(t) = \frac{f(t)}{S(t)}$
- Cumulative hazard function: $\mathrm{H}(t) = \int_0^t h(u)du$
- $H(t) = -\log(S(t))$
- $h(t) = -\frac{d}{dt}(\log(S(t)))$
- $S(t) = e^{-H(t)}$
- $f(t) = h(t)\, e^{-H(t)}$

OBIETTIVI DELL'ANALISI DI SOPRAVVIVENZA

- Se vogliamo studiare il tempo ad un evento di interesse pertanto abbiamo bisogno di altri metodi statistici → metodi dell'analisi di sopravvivenza
- Tre principali obiettivi dell'analisi di sopravvivenza:

1. Stimare il tempo ad un evento per un gruppo di individui
2. Confrontare il tempo ad un evento per due o più gruppi di individui
3. Studiare la relazione tra una o più variabili esplicative e il tempo all'evento

PROPORTIONAL HAZARDS MODELS

Proportional hazards models (modelli di rischio proporzionale): descrivono la relazione tra un set di variabili indipendenti e la **hazard function** che caratterizza il tempo ad un evento di interesse.

$$h(t) = h_0(t) \cdot e^{\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m} = h_0(t) \cdot e^{\boldsymbol{\beta}^T \mathbf{X}}$$

- $\mathbf{X}^T = [X_1\ X_2\ \ldots\ X_m]$ → vettore contenente i valori delle variabili indipendenti
- $\boldsymbol{\beta}^T = [\beta_1\ \beta_2\ \ldots\ \beta_m]$ → vettore dei coefficienti
- $h_0(t)$ → **baseline hazard function** → la hazard function per gli individui per cui le variabili indipendenti sono tutte nulle.

PROPORTIONAL HAZARDS MODELS

- Formulazione logaritmica:

$$log\big(h(t)\big) = log\big(h_0(t)\big) + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m$$

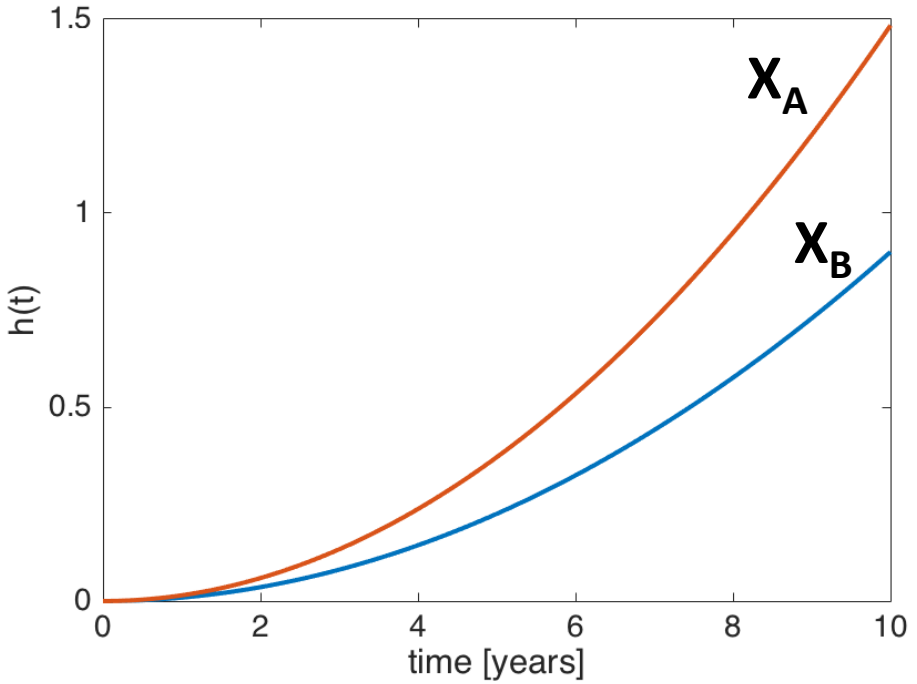
- Il modello può essere visto come una regressione lineare multipla avente come outcome il logaritmo della hazard function, come variabili indipendenti X_1, X_2, …, X_m, e come intercetta il logaritmo della baseline hazard function.

ASSUNZIONE DI PROPORZIONALITA' DEI RISCHI

- Assunzione: la hazard function di un qualsiasi individuo è data dalla baseline hazard function ($h_0(t)$) moltiplicata per una costante ($e^{\boldsymbol{\beta}^T \boldsymbol{X}}$).
- Il rapporto tra i valori di h(t) per due individui, A e B, aventi variabili indipendenti $\mathbf{X_A}$ e $\mathbf{X_B}$ è costante e pari a $e^{\boldsymbol{\beta}^T(\boldsymbol{X_A}-\boldsymbol{X_B})}$.

$$\frac{h_A(t)}{h_B(t)} = \frac{\cancel{h_0(t)} \cdot e^{\boldsymbol{\beta}^T \boldsymbol{X_A}}}{\cancel{h_0(t)} \cdot e^{\boldsymbol{\beta}^T \boldsymbol{X_B}}} = \underbrace{e^{\boldsymbol{\beta}^T(\boldsymbol{X_A}-\boldsymbol{X_B})}}_{\text{costante}}$$

Le funzioni di hazard per qualsiasi diverso valore di **X** non possono mai intersecarsi tra loro.

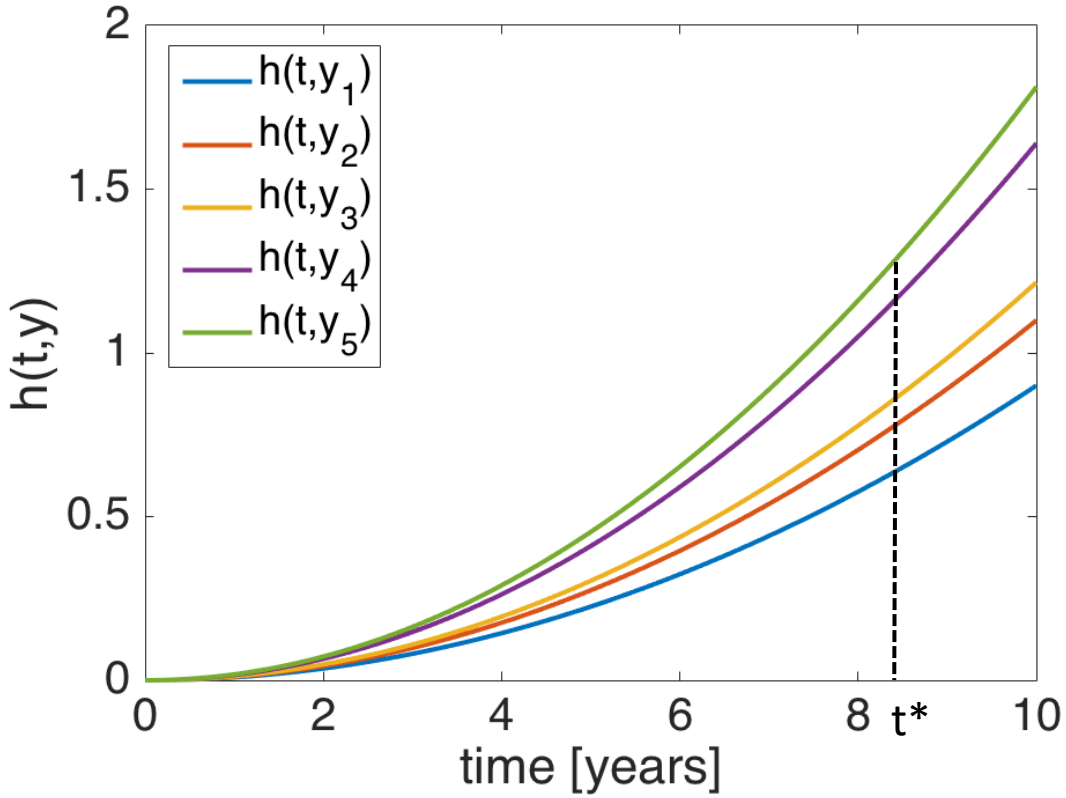


RISK SCORE

- Il valore $y = \boldsymbol{\beta}^T \boldsymbol{X}$ viene chiamato **risk score**, o score di rischio → quantità che consente di ordinare diversi individui in base al loro rischio di sperimentare l'evento di interesse.
- Per ogni istante temporale t^* si ha che:

$y_1 < y_2 < y_3 < y_4 < y_5$

$h(t^*,y_1) < h(t^*,y_2) < h(t^*,y_3) < h(t^*,y_4) < h(t^*,y_5)$



EQUAZIONE ALTERNATIVA DEL MODELLO

- Sfruttando la relazione tra S(t) e h(t) si può derivare anche questa formulazione alternativa del modello:

$$S(t) = S_0(t)^{e^{\boldsymbol{\beta}^T \boldsymbol{X}}}$$

- Dimostrazione (bonus):

$$h(t) = -\frac{d\left(\log\left(S(t)\right)\right)}{dt}$$

$$= -\frac{d}{dt}\left(e^{\boldsymbol{\beta}^T \boldsymbol{X}} \log\left(S_0(t)\right)\right)$$

$$= -\frac{d}{dt}\left(\log\left(S_0(t)\right)\right) e^{\boldsymbol{\beta}^T \boldsymbol{X}}$$

$$= h_0(t) e^{\boldsymbol{\beta}^T \boldsymbol{X}}$$

MODELLO DI COX

- Il modello di Cox (Cox 1972) è un modello di tipo proportional hazard **semi-parametrico:**

$$h(t) = h_0(t) \cdot e^{\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m}$$

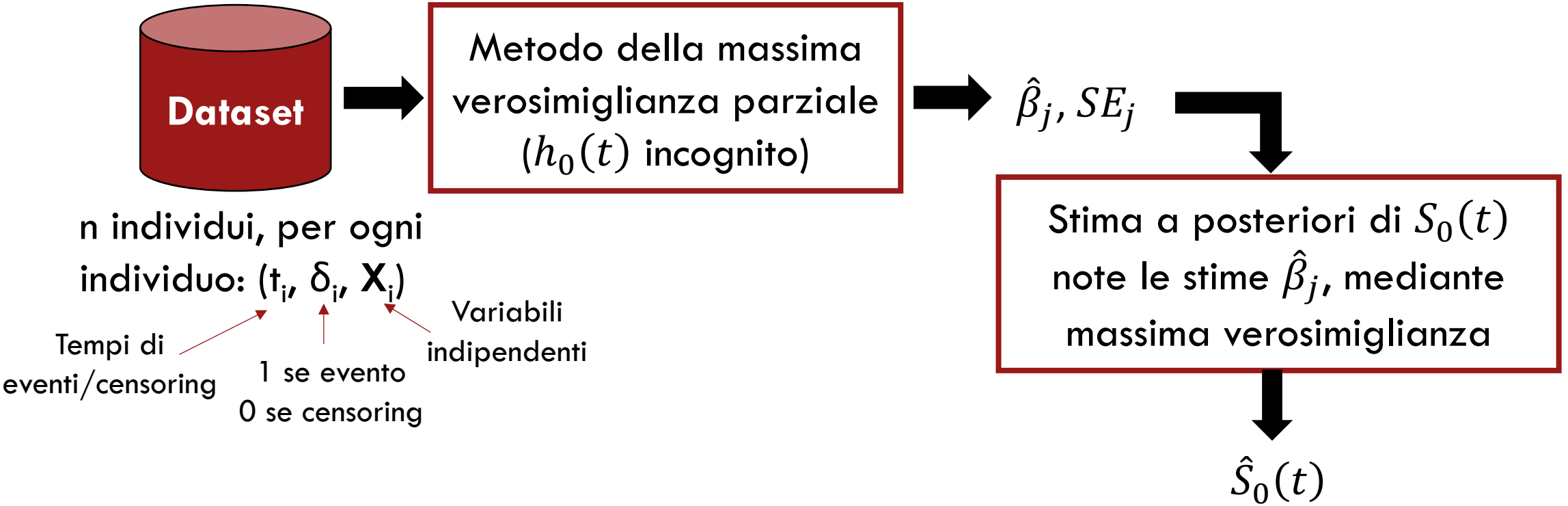
Parte non parametrica
Non si fanno assunzioni sulla forma della baseline hazard function

Parte parametrica
Funzione parametrica delle variabili indipendenti

- Incognite del modello: $h_0(t), \beta_1, \ldots, \beta_m$ → come le stimiamo a partire dai dati?

STIMA DEGLI ELEMENTI INCOGNITI DEL MODELLO

- **Metodo della massima verosimiglianza parziale** (*partial maximum likelihood*) → consente di stimare i coefficienti $\beta_j, j = 1, \dots, m$ senza conoscere $h_0(t)$



INTERPRETAZIONE DEI COEFFICIENTI

- **Hazard ratio:** $e^{\hat{\beta}_j}$ → indica di quanto un aumento di una unità della variabile X_j, tenendo costanti tutte le altre variabili indipendenti, amplifica o attenua $h_0(t)$.
- Se $\hat{\beta}_j > 0$ o $e^{\hat{\beta}_j} > 1$ → se X_j aumenta, anche $h(t)$ aumenta
- Se $\hat{\beta}_j < 0$ o $e^{\hat{\beta}_j} < 1$ → se X_j aumenta, $h(t)$ diminuisce
- Se $\hat{\beta}_j = 0$ → la variabile X_j non ha un impatto su $h(t)$
- **Wald test** per determinare se i coefficienti β_j sono significativamente diversi da 0 (analogamente alla regressione logistica).

ESEMPIO

- Modello di Cox per la predizione del tempo all'insorgenza di diabete di tipo 2 negli adulti.

Variabile	Stime dei coefficienti	Hazard ratio	Standard error delle stime dei coefficienti	P-value (Wald test)
Sesso maschile	-0.3453	0.71	0.0646	<0.0001
BMI	0.0971	1.10	0.0051	<0.0001
Ipertensione	0.3232	1.38	0.0665	<0.0001
Malattia cardiaca	0.1968	1.22	0.0969	0.0422
Fumatore	0.0101	1.11	0.0789	0.2001
Non caucasico	0.2207	1.27	0.0739	0.0011
Livello di istruzione medio	-0.2323	0.79	0.0795	0.0035
Livello di istruzione alto	-0.3344	0.72	0.0803	<0.0001

DOMANDE SUI RISULTATI DELL'ESEMPIO

1. Quali variabili hanno un impatto significativo sul rischio di insorgenza di diabete considerando un livello di significatività al 5%?

2. Quali variabili influenzano positivamente il rischio di insorgenza di diabete?

3. Quali variabili influenzano negativamente il rischio di insorgenza di diabete?

RISPOSTE

1. Quali variabili hanno un impatto significativo sul rischio di insorgenza di diabete considerando un livello di significatività al 5%?
   - Tutte tranne fumatore.

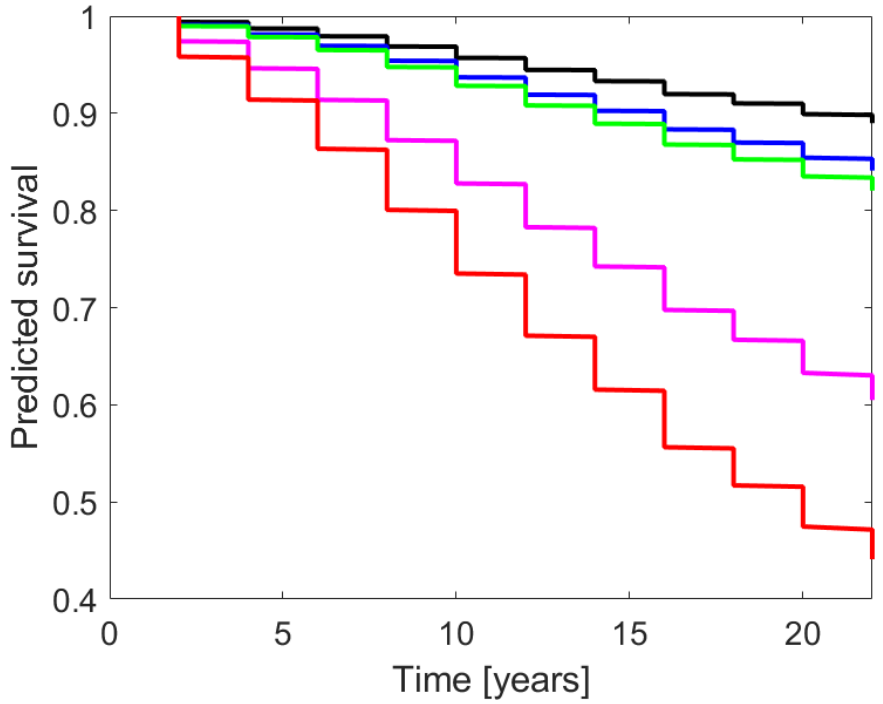
2. Quali variabili influenzano positivamente il rischio di insorgenza di diabete?
   - BMI, ipertensione, malattia cardiaca, non caucasico

3. Quali variabili influenzano negativamente il rischio di insorgenza di diabete?
   - Sesso maschile, livello di istruzione medio, livello di istruzione alto.

CURVE DI SOPRAVVIVENZA PREDETTE

- Una volta ottenute le stime $\hat{\beta}_j$ e $\hat{S}_0(t)$, è possibile predire la curva di sopravvivenza di un individuo noti i suoi valori delle variabili indipendenti.

$$\hat{S}(t) = \hat{S}_0(t)^{e^{\boldsymbol{\beta}^T \boldsymbol{X}}}$$



Curve di sopravvivenza stimate per diversi individui caratterizzati da diversi valori delle variabili indipendenti.

VALUTAZIONE DELLE PERFORMANCE DEL MODELLO DI COX

- **Concordance index** o **C-index**: principale metrica per valutare la capacità predittiva del modello di Cox.
- Definito come la probabilità che il modello assegni risk score maggiori agli individui aventi tempi di sopravvivenza minori:

$$C := P\left(Y_j > Y_i \middle| T_j < T_i\right)$$

  - Y_i, Y_j: variabili aleatorie rappresentanti gli score di rischio degli individui i e j
  - T_i, T_j: variabili aleatorie rappresentanti i tempi di sopravvivenza degli individui i e j

STIMATORE DI HARRELL DEL C-INDEX (1/2)

- 1982: Harrell propose uno stimatore di C che tiene conto della presenza di dati censurati.
- n individui con score di rischio predetto dal modello $y_1, y_2, \ldots, y_n$, tempi all'evento o di censoring $t_1, t_2, \ldots, t_n$, indicatori di evento o censoring d $\delta_1, \delta_2, \ldots, \delta_n$ (se evento $\delta_i=1$, se censoring $\delta_i=0$).
- La coppia di valori (y_i, y_j) si dice:
  - **non confrontabile** se vale una delle seguenti condizioni:
    - $t_i = t_j$ e $\delta_i = \delta_j = 1$→ parità nei tempi di sopravvivenza
    - $\delta_i = 0$ e $\delta_j = 0$ → entrambi censurati
    - $\delta_i = 1$, $\delta_j = 0$ e $t_j < t_i$ → individuo j censurato prima dell'evento dell'individuo i
  - viceversa si dice **confrontabile.**

STIMATORE DI HARRELL DEL C-INDEX

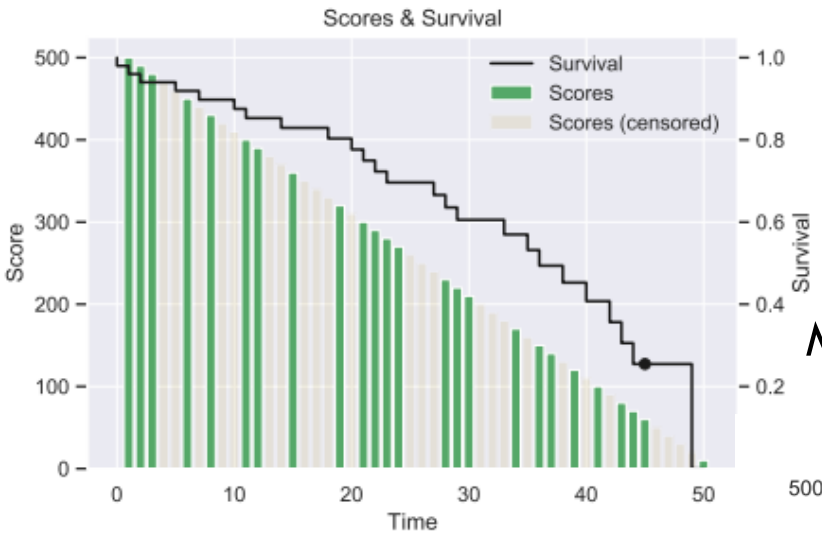
- Una coppia di valori confrontabili (y_i, y_j) si dice:
  - **concorde** se $y_i < y_j$ e $t_j < t_i$
  - **discorde** se $y_i > y_j$ e $t_j < t_i$
  - **pari** (tie) se $y_i = y_j$

$$\hat{C} = \frac{\mathrm{n}_{\mathrm{concordi}} + 0.5 \cdot \mathrm{n}_{\mathrm{pari}}}{\mathrm{n}_{\mathrm{confrontabili}}}$$

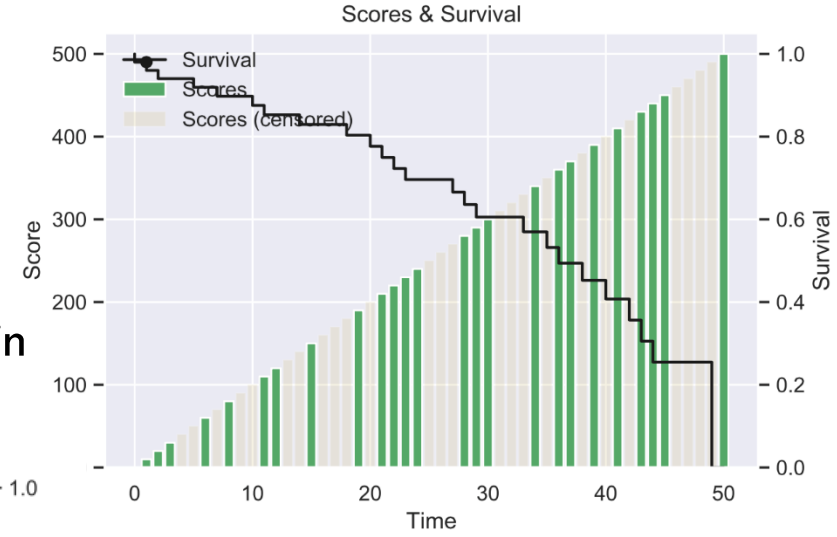
- $\mathrm{n}_{\mathrm{concordi}}$ = numero di coppie concordi
- $\mathrm{n}_{\mathrm{pari}}$ = numero di coppie pari
- $\mathrm{n}_{\mathrm{confrontabili}}$ = numero di coppie confrontabili

INTERPRETAZIONE DEL C-INDEX (1/2)

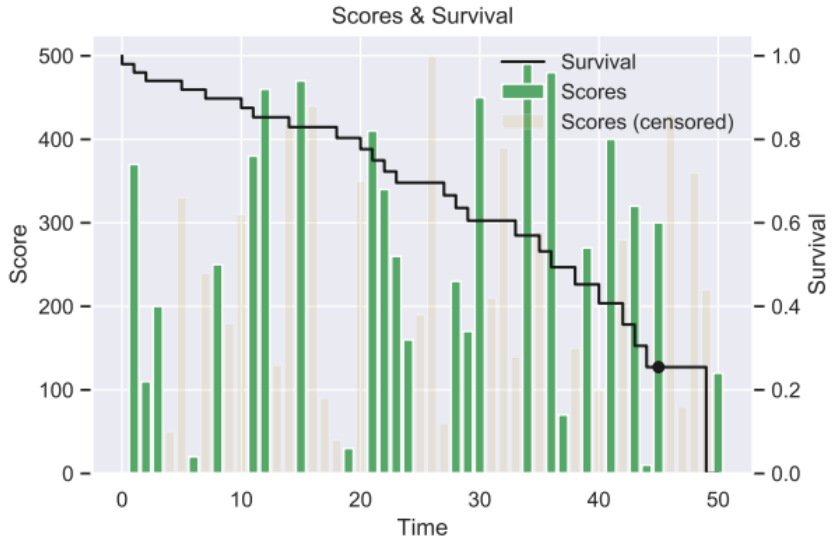
Concordanza perfetta tra risk score e tempi degli eventi→ C=1



Perfetta discordanza tra risk score e tempi degli eventi→ C=0



Modello che assegna i risk score in modo casuale→ C=0.5



INTERPRETAZIONE DEL C-INDEX (2/2)

- Il C-index varia tra 0 e 1.
- Tanto più è vicino a 1 tanto più il modello tenderà ad assegnare risk score elevati ai soggetti per cui l'evento si verifica prima nel tempo.
- Un modello che assegna i risk score in maniera random ha C-index pari a 0.5.
- C-index < 0.5 → il modello fa il contrario di quello che dovrebbe: assegna risk score più alti ai soggetti per cui l'evento si verifica più in là nel tempo.

- Nota: quando tutti i tempi degli eventi sono uguali, il C-index diventa equivalente all'area sotto la curva ROC (spesso chiamata C-statistic).

ESEMPIO

- Modello di Cox per la predizione del tempo all'insorgenza di diabete di tipo 2 negli adulti.

MODELLO	C-INDEX
Modello 1 – modello completo (slide 12)	0.7235
Modello 2 – senza fumatore	0.7207
Modello 3 – senza malattia cardiaca	0.7225
Modello 4 – senza fumatore e malattia cardiaca	0.7201

Le performance dei modelli in termini di C-index sono confrontabili. Tra questi sceglieremmo il modello con meno variabili (modello 4).