

Análise de duração

Prof. Francis Petterini

20 de outubro de 2025

Objetivo e notação

Objetivo: modelar o tempo até um evento (saída do emprego, desemprego, reincidência, falha etc.).

Notação básica para indivíduo i :

- Duração $T_i \in (0, \infty)$, intervalo aberto
- Sobrevivência $S(t) = \Pr(T \geq t) \quad \Rightarrow \quad F(t) = 1 - S(t)$
- Densidade $f(t) = -\frac{d}{dt}S(t)$
- Risco (hazard) $\lambda(t) = \lim_{\Delta \rightarrow 0} \frac{\Pr(t \leq T < t + \Delta \mid T \geq t)}{\Delta} = \frac{f(t)}{S(t)}$, note que não é exatamente uma probabilidade, porque pode assumir valores maiores que 1

Relações fundamentais

$$f(t) = \lambda(t) S(t), \quad \Lambda(t) = \int_0^t \lambda(u) du$$

(risco acumulado)

$$\boxed{S(t) = \exp\{-\Lambda(t)\}} \iff \boxed{\lambda(t) = -\frac{d}{dt} \log S(t)}$$

Função de risco cumulativa (Nelson–Aalen): $\hat{\Lambda}(t) = \sum_{t_{(j)} \leq t} \frac{d_j}{n_j}$

onde $t_{(j)}$ são tempos distintos de evento, d_j eventos em $t_{(j)}$ e n_j em risco imediatamente antes.

Função de verossimilhança (censura à direita)

Para cada i , seja t_i o tempo observado e $\delta_i \in \{0, 1\}$ o indicador de evento (1=evento, 0=censura):

$$\boxed{L_i = [\lambda(t_i | X_i)]^{\delta_i} S(t_i | X_i)} \quad \text{com} \quad S(t | X) = \exp\left\{-\int_0^t \lambda(u | X) du\right\}.$$

Riscos proporcionais (PH) e hazard ratio (HR)

Modelo PH (Cox):

$$\lambda(t | X) = \lambda_0(t) \exp(\beta' X).$$

Hazard ratio (definição):

$$\text{HR}(t; X, \tilde{X}) = \frac{\lambda(t | X)}{\lambda(t | \tilde{X})}.$$

Sob PH,

$$\text{HR}(t; X, \tilde{X}) = \exp\{\beta'(X - \tilde{X})\} \quad (\text{constante em } t).$$

Modelo não-PH (efeito variante no tempo):

$$\lambda(t | X) = \lambda_0(t) \exp(\beta' X + f(t) \gamma' X) \Rightarrow \text{HR}(t; X, \tilde{X}) = \exp\{[\beta + f(t)\gamma]'(X - \tilde{X})\}.$$

Modelagem: Cox (PH) e AFT

Cox (PH): $\lambda(t | X) = \lambda_0(t) \exp(\beta' X)$

- Estima β por verossimilhança parcial; $\lambda_0(t)$ via Breslow/ef. equivalentes.
- Predição: $S(t | X) = S_0(t)^{\exp(\beta' X)}$.

AFT (Accelerated Failure Time): $\log T = \mu + \beta' X + \sigma W$

- Famílias: Weibull, log-normal, log-logística, Gompertz, etc.
- Interpretação em *time ratio*: $TR = \exp(\beta_k)$ (multiplica o tempo característico).

Escolha prática:

- PH plausível $\Rightarrow$ Cox; violações sistemáticas de PH $\Rightarrow$ AFT/PH com efeitos do tempo.

AFT: μ , W e interpretação

Modelo AFT:

$$\log T = \mu + \beta'X + \sigma W,$$

onde μ é **locação** (intercepto em $\log T$), $\sigma > 0$ é **escala**, e W é um erro **padronizado** com distribuição escolhida:

- $W \sim \mathcal{N}(0, 1) \Rightarrow \text{log-normal}$;
- W logístico padrão $\Rightarrow \text{log-logístico}$;
- W Gumbel (extreme value) $\Rightarrow \text{Weibull}$ (em AFT).

Interpretação (time ratio):

$TR = \exp(\beta_k)$ multiplica um tempo característico (mediana/quantis).

Sobrevivência e hazard

$$S(t | X) = 1 - F_W\left(\frac{\log t - \mu - \beta'X}{\sigma}\right), \quad \lambda(t | X) = \frac{f_W(z)}{\sigma t [1 - F_W(z)]}, \quad z = \frac{\log t - \mu - \beta'X}{\sigma}.$$

Resumo: μ desloca a distribuição de $\log T$; W define a *família* (forma); $\exp(\beta)$ fornece **time ratios** claros para comunicação substantiva.

Covariáveis variantes no tempo

- Estrutura: $\lambda(t \mid X_i(t)) = \lambda_0(t) \exp(\beta' X_i(t))$.
- Implementação: dividir o seguimento em intervalos $[t_{k-1}, t_k)$ com X constante em cada.
- Cuidados: evitar simultaneidade (*internal time-dependent covariates*); usar defasagens quando necessário.
- Interpretação: a HR é instantânea em t e depende de $X(t)$ naquele instante.

Heterogeneidade não observada (fragilidade)

- **Fragilidade multiplicativa:** $\lambda(t \mid X, u) = u \lambda_0(t) \exp(\beta' X)$, tipicamente $u \sim \Gamma(\kappa, \kappa)$ (média 1).
- Interpretação: propensão individual persistente afeta a taxa em todos os tempos.
- Implicação: seleção dos mais frágeis no início pode induzir *decaimento* aparente da hazard (duração dependente).

Curva de Kaplan–Meier $\hat{S}(t)$: definição e estimação

Cenário: censura à direita. Seja $t_{(1)} < \dots < t_{(J)}$ os tempos *distintos* de evento.

$n_j = \#$ em risco imediatamente antes de $t_{(j)}$, $d_j = \#$ eventos em $t_{(j)}$.

Produto-limite (Kaplan–Meier):

$$\hat{S}(t) = \prod_{t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right), \quad \hat{S}(0) = 1.$$

- $\hat{S}(t)$ é **escada**: muda em tempos de evento.
- **Greenwood (variância):** $\widehat{\text{Var}}\{\hat{S}(t)\} = \hat{S}(t)^2 \sum_{t_{(j)} \leq t} \frac{d_j}{n_j(n_j - d_j)}.$

Curvas de Kaplan-Meier por grupo

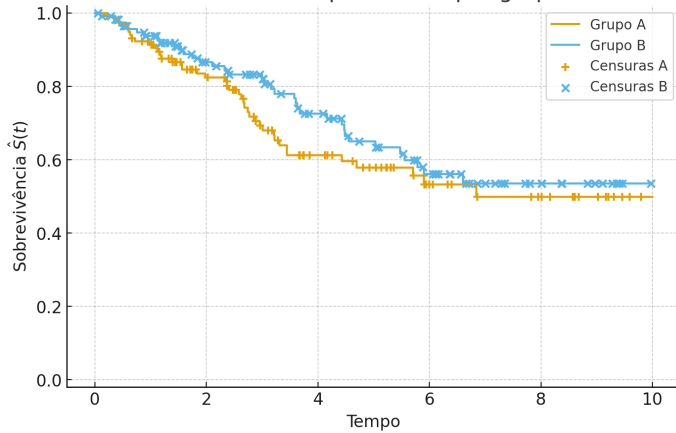


Tabela *at risk*

Tabela "at risk" por tempos de referência

Tempo	Grupo A: em risco	Grupo B: em risco
0	120	120
2	77	77
4	40	51
6	19	29
8	12	12
10	0	0

Cumulante de risco (Nelson–Aalen) e suavização

$$\hat{\Lambda}_{\text{NA}}(t) = \sum_{t_{(j)} \leq t} \frac{d_j}{n_j}, \quad \hat{S}_{\text{NA}}(t) = \exp\{-\hat{\Lambda}_{\text{NA}}(t)\}.$$

Hazard suavizada: aplicar um kernel K_h aos incrementos em $t_{(j)}$:

$$\hat{\lambda}_K(t) \approx \sum_j K_h(t - t_{(j)}) \frac{d_j}{n_j}.$$

Uso: exibir $\hat{\Lambda}(t)$ (passos) e $\hat{\lambda}_K(t)$ (suave) por grupo; ICs para $\hat{\Lambda}$ via variância padrão; bandas para $\hat{\lambda}_K$ por bootstrap.

Outputs fundamentais e comparações

Descritivos não paramétricos

- Curva de Kaplan–Meier $\hat{S}(t)$ + tabela *at risk*.
- Mediana (e percentis) do tempo até o evento, com ICs.
- Cumulante de risco (Nelson–Aalen) $\hat{\Lambda}(t)$ e hazard suavizada $\hat{\lambda}(t)$.

Modelos

- Cox (PH): HRs $\exp(\beta)$, baseline $\hat{\Lambda}_0(t)$, $\hat{S}_0(t)$.
- AFT paramétricos (Weibull, log-normal, log-logística): time ratios e parâmetros de forma/escala.

Diagnóstico: resíduos de ajuste e influência

Objetivos: checar forma funcional e detectar observações influentes. **Notação** δ_i : indicador de evento.

Martingale: $M_i = \delta_i - \hat{\Lambda}_0(t_i) \exp(\hat{\beta}' X_i)$, Deviance: $D_i = \text{sign}(M_i) \sqrt{-2\{M_i + \delta_i \log(\delta_i - M_i)\}}$.

$\text{dfbeta}_{ik} = \hat{\beta}_k - \hat{\beta}_{k(-i)}$ (estimação sem a obs. i).

Sinais de alerta: $|D_i|$ grande (ajuste ruim) e $|\text{dfbeta}_{ik}|$ grande (alta influência).

Diagnóstico: desempenho preditivo

Discriminação (C-index): ranking correto do risco.

$$C = \Pr(\hat{\eta}_i > \hat{\eta}_j \mid t_i < t_j, \delta_i = 1) \text{ (ajuste por censura: IPCW/Uno).}$$

Acurácia (Brier score)

$$BS(t) = \frac{1}{n} \sum_{i=1}^n w_i(t) (Y_i(t) - \hat{S}(t \mid X_i))^2, \quad Y_i(t) = \mathbf{1}\{T_i > t\}.$$

Ajuste global

$$AIC = -2\ell + 2k, \quad BIC = -2\ell + k \log n.$$