Data Science and Big Data

Silva, J.L.P.

21 de setembro, 2024

Objetivos da Aula

- Introduzir o problema da análise de tempo até um evento de interesse.
- Discutir a questão da censura e como ela é acomodada na análise.
- Apresentar métodos não paramétricos para descrição e comparação de curvas de sobrevivência.
- Apresentar alguns modelos paramétricos e semi-paramétricos adequados para dados censurados.

Parte I: Introdução

Considerações Iniciais

Em *Análise de Sobrevivência* a variável resposta é o tempo até a ocorrência de um evento de interesse.

Estamos interessados em variáveis aleatórias positivas, tais como:

- tempo até a morte de um paciente, ou até o início (ou recidiva) da doença;
- tempo de duração de uma greve;
- tempo até a falha de um equipamento;
- tempo entre a liberação de presos e a ocorrência de crimes;
- tempo até o pagamento de uma dívida; etc.

Este tempo é denominado **tempo de falha**.

Introdução

A principal característica destes dados é a presença de **censura**, que é a observação parcial da resposta.

Na presença de censura, a informação que temos sobre a resposta se resume ao conhecimento de que o tempo de falha é superior àquele observado.

Sem a presença de censura, técnicas clássicas como análise de regressão poderiam ser utilizadas na análise destes dados.

Introdução

Por exemplo, imagine que queremos comparar o tempo médio de vida de três grupos de pacientes:

- Se não houver censuras, pode-se usar as técnicas de análise de variância.
- Havendo censuras, o que é provável, tais técnicas não podem ser utilizadas, pois elas necessitam de todos os tempos de falha.
- Faz-se necessário o uso dos métodos de análise de sobrevivência que possibilitem incorporar na análise estatística a informação contida nos dados censurados.

Introdução

Mesmo censurados, todos os resultados provenientes de um estudo de sobrevivência devem ser usados na análise, por duas razões:

1. Mesmo sendo incompletas, as observações censuradas fornecem informações sobre o tempo de falha.
2. A omissão das censuras no cálculo das estatísticas de interesse pode acarretar conclusões viciadas.

Tipos de Censura

Censura do tipo I ocorre naqueles estudos em que, ao serem finalizados após um tempo pré-estabelecido, alguns indivíduos ainda não apresentaram o evento de interesse.

Censura do tipo II ocorre em estudos que são encerrados após a observação de um número pré-estabelecido de falhas.

Censura aleatória ocorre com frequência na área médica. Neste caso o indivíduo sai do estudo sem ter ocorrido a falha.

Tipos de Censura

Seja T uma v.a. representando o tempo de falha e seja C, uma outra v.a. independente de T, representando o tempo de censura para determinado indivíduo.

O que se observa é

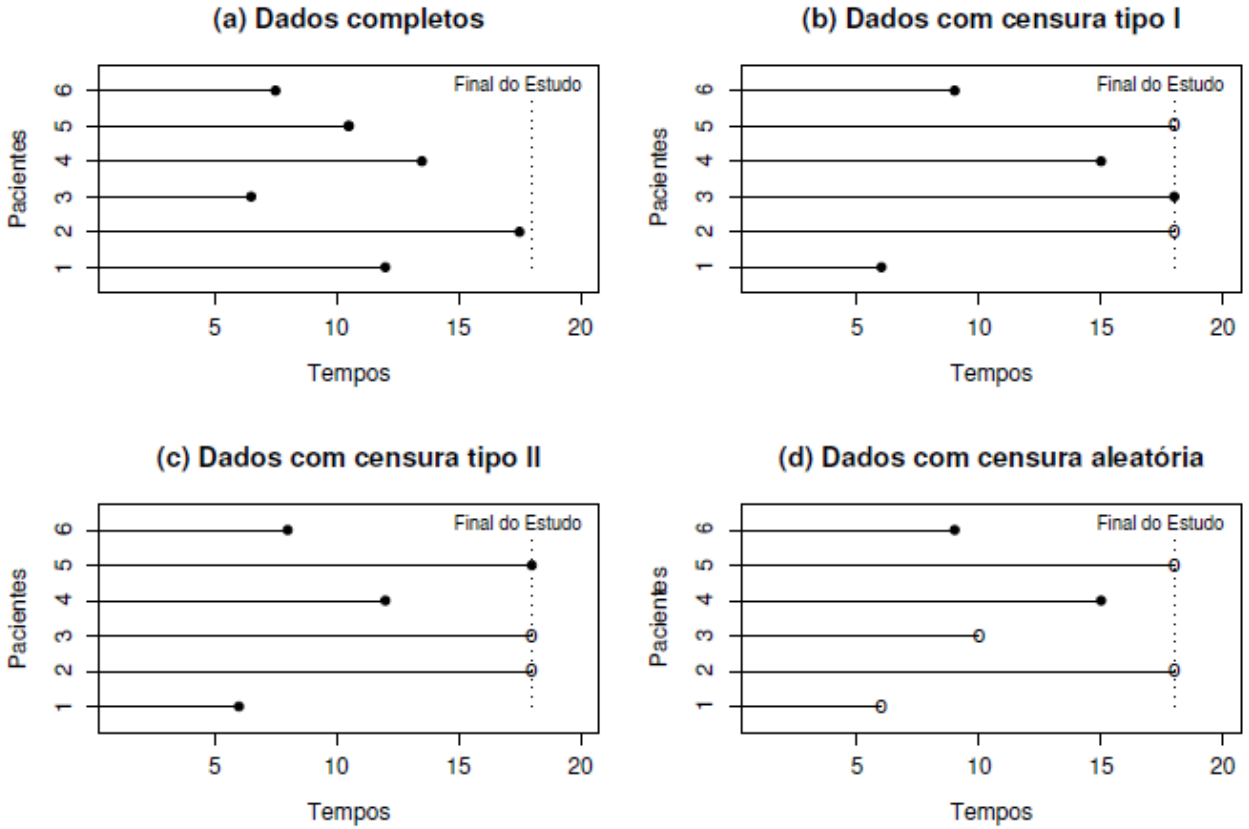
$$t = min(T, C)$$

e

$$\delta = \begin{cases} 1 & \text{se} \quad T \leq C \\ 0 & \text{se} \quad T > C. \end{cases}$$

Suponha que os pares (T_i, C_i), $i = 1, \ldots, n$ formam uma amostra aleatória de n indivíduos. Quando $C_i = C$, uma constante fixada, obtém-se a censura do tipo I.

Tipos de Censura



(a) todos falham antes do término do estudo, (b) censuras ocorrem devido ao término do estudo, (c) censuras ocorrem porque o término do estudo ocorre quando se observa um número pré-estabelecido de falhas e (d) censuras ocorrem por vários motivos.

Tipos de Censura

- O mecanismo típico de censura é *censura à direita*, mostrado na figura.
- *Censura à esquerda* surge quando o tempo registrado é maior que o tempo de falha. Ou seja, o evento já ocorreu quando o indivíduo foi observado.
- *Censura intervalar* acontece quando sabe-se que o evento ocorre em um intervalo de tempo, isto é $T \in (L, U]$.
  - Se $L = U$ temos tempos exatos de falha;
  - se $L = \infty$ temos censura à direta;
  - se $L = 0$ temos censura à esquerda.

Exemplo 1: Dados de Hepatite (Gregory et al., 1976)

- Estudo clínico aleatorizado envolvendo pacientes com Hepatite Viral Aguda.
- O objetivo é investigar o efeito da terapia com esteroide.
- Vinte e nove pacientes com esta doença foram aleatorizados para receber um placebo ou o tratamento com esteroide.
- Cada paciente foi acompanhado por 16 semanas ou até à morte (evento de interesse) ou até a perda de acompanhamento.

Exemplo 1: Dados de Hepatite (Gregory et al., 1976)

Os tempos de sobrevivência observados, em semanas, para os dois grupos (+ indica censura) foram:

Grupo	Tempo de sobrevivência em semanas
Controle	1+, 2+, 3, 3, 3+, 5+, 5+, 16+, 16+, 16+, 16+, 16+, 16+, 16+, 16+
Esteróide	1, 1, 1, 1+, 4+, 5, 7, 8, 10, 10+, 12+, 16+, 16+, 16+

A censura é do tipo aleatória.

Exemplo 2: Dados de Malária (pag. 14, Colosimo e Giolo, 2006)

- Estudo experimental com camundongos conduzido no Centro de Pesquisas René Rachou, FioCruz, MG.
- 44 camundongos foram infectados pela malaria (*Plasmodium berguei*).
- Os camundongos foram aleatoriamente alocados em três grupos:
  - Grupo 1: infectado também pela esquistossomose e imunizados 30 dias antes da infecção.
  - Grupo 2: controle.
  - Grupo 3: infectado também pela esquistossomose.

Exemplo 2: Dados de Malária (pag. 14, Colosimo e Giolo, 2006)

- A resposta foi o tempo decorrido desde a infecção pela malária até a morte do camundongo.
- O tempo foi medido em dias e o estudo foi acompanhado por 30 dias.
- Dados disponíveis na pag. 14 (Colosimo e Giolo, 2006).

A censura é, portanto, do tipo I.

Exemplo 3: Confiabilidade

O fabricante de um tipo de isolador elétrico quer conhecer o comportamento de seu produto funcionando a uma temperatura de 200^oC.

Um teste de vida foi realizado nestas condições usando-se 60 isoladores elétricos. O teste terminou quando 45 deles havia falhado (censura do tipo II).

As 15 unidades que não haviam falhado ao final do teste foram, desta forma, censuradas no tempo $t = 2729$ horas.

O fabricante tem interesse em estimar o tempo médio e mediano de vida do isolador e o percentual de falhas após 500 horas de uso.

Exemplo 4: Ciências Sociais

Demógrafos e cientistas sociais estão interessados na duração de certos "estados'' de vida para humanos.

Considere, por exemplo, o casamento e, em particular, os casamentos formados durante o ano de 1980 em um determinado país.

Então o *tempo de vida* de um casamento seria a sua duração; um casamento pode terminar por razões como anulação, divórcio ou morte.

Representação Probabilística do Mecanismo de Censura Aleatória

- T: tempo de falha.
- C: tempo de censura.
- T e C independentes (mecanismo não informativo).
- Os valores observados são:

$$t = min(T, C)$$

e

$$\delta = \begin{cases} 1 & \text{se} \quad T \leq C \\ 0 & \text{se} \quad T > C. \end{cases}$$

Especificação da Resposta

- Função de Sobrevivência:

$$S(t) = P(T \geq t)$$

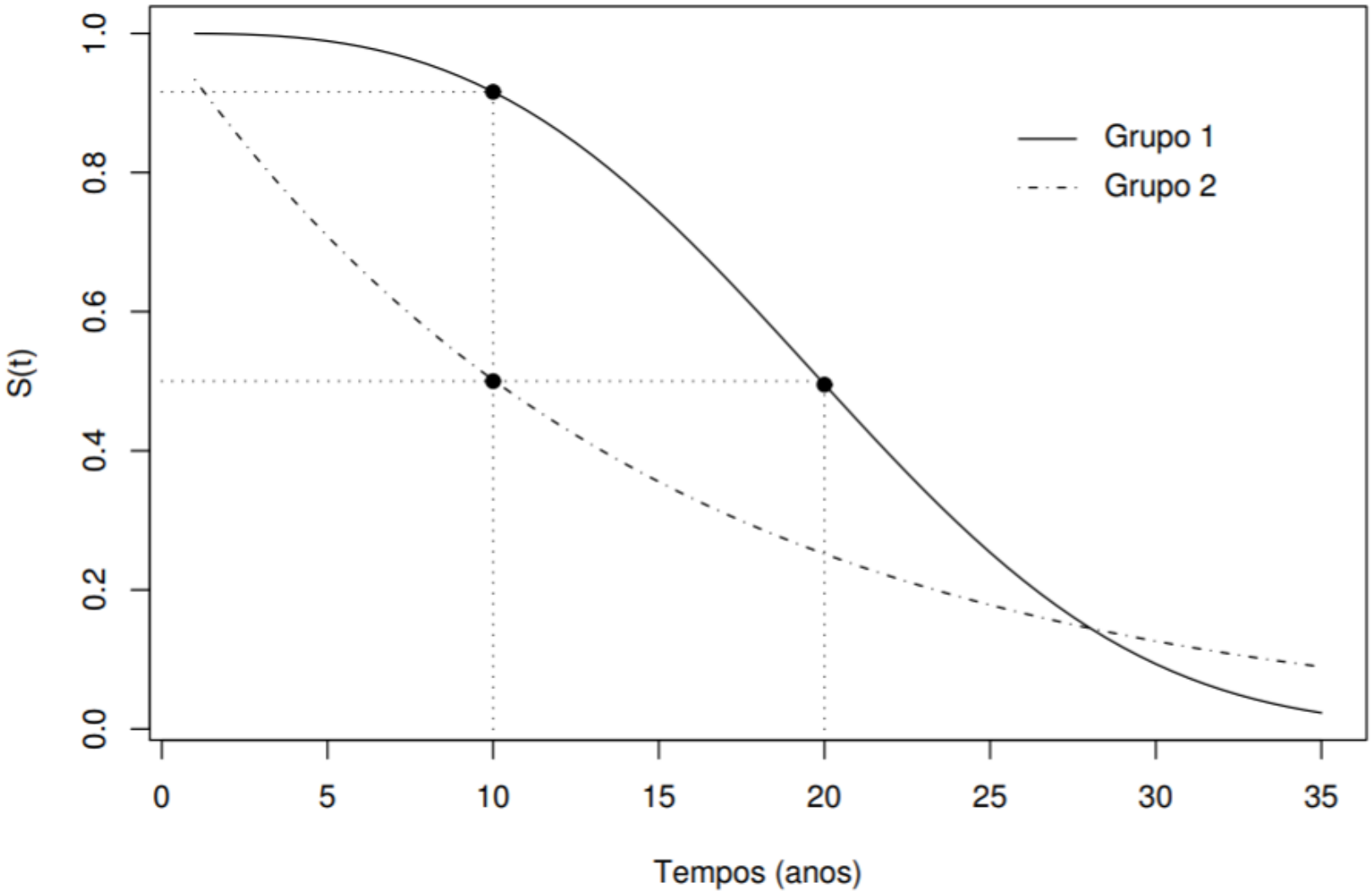
- Função de Taxa de Falha:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

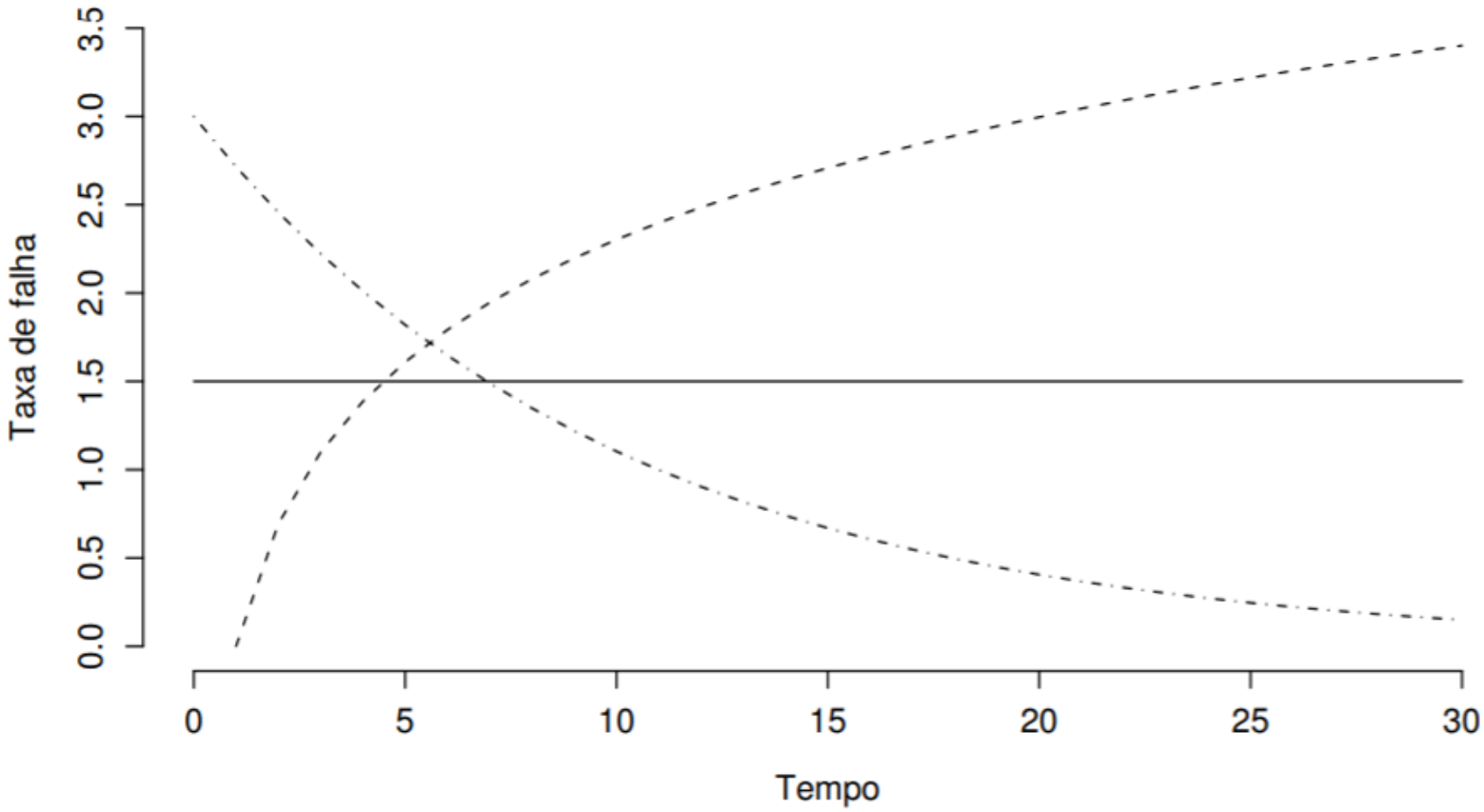
- Função de Taxa de Falha Acumulada

$$\Lambda(t) = \int_0^t \lambda(u) du$$

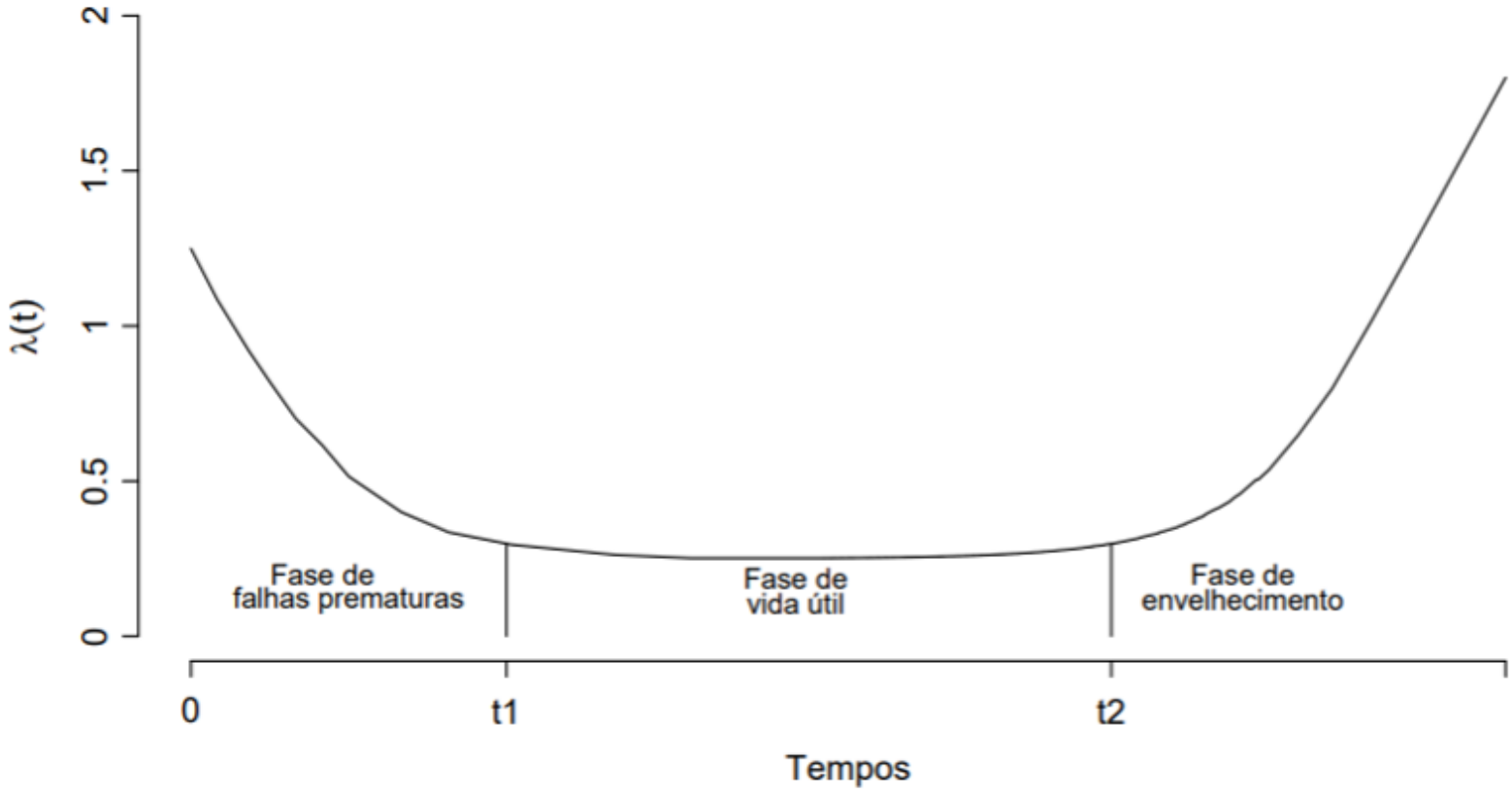
Exemplo: Funções de Sobrevivência



Exemplo: Funções de Taxa de Falha



Exemplo: Funções de Taxa de Falha Tipo Banheira



Relações Importantes Entre as Funções

- $$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\left(\log\ S(t)\right),$$
- $$\Lambda(t) = \int_0^t \lambda(u)du = -\log\ S(t),$$
- $$S(t) = \exp\left\{-\Lambda(t)\right\} = \exp\left\{-\int_0^t \lambda(u)du\right\}.$$

O conhecimento de uma das funções implica no conhecimento das demais.

Descrição de Dados de Sobrevivência

- Estimar a função de sobrevivência $S(t)$:
  - Estimador limite-produto de Kaplan-Meier (Kaplan e Meier, 1958).
- Comparar curvas de sobrevivência:
  - Teste log-rank;
  - Teste de Wilcoxon;
  - Outros testes: família de testes.

Estimador de Kaplan-Meier (EKM)

O estimador de Kaplan-Meier é uma adaptação da função de sobrevivência empírica que, na ausência de censuras, é definida como:

$$\hat{S}(t) = \frac{n^o \text{ de observações que não falharam até o tempo } t}{n^o \text{ de observações no estudo}}.$$

- $\hat{S}(t)$ é uma função escada com degraus nos tempos observados de falha de tamanho $1/n$, em que n é o tamanho da amostra.
- Se existirem empates no tempo t, o tamanho do degrau fica multiplicado pelo número de empates.
- O EKM considera tantos intervalos de tempo quantos forem o número de falhas distintas.

Estimador de Kaplan-Meier (EKM)

Construção do EKM em dados que envolvam censuras:

- Ordenar os tempos distintos de falha: $t_1 < t_2 < \ldots < t_k$.
- Obter d_j: número de falhas no tempo t_j.
- Obter n_j: número de observações sob risco (não falhou e não foi censurado) até o tempo t_j (exclusive);
- Obter $q_j = d_j/n_j$.
- A sobrevivência em t_j é estimada por

$$\hat{S}(t_j) = (1-q_1)(1-q_2)\ldots(1-q_j) = \prod_{i \leq j}\left(1 - \frac{d_i}{n_i}\right).$$

Estimador de Kaplan-Meier (EKM)

O EKM é então definido por:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right).$$

Principais propriedades do EKM:

- é não-viciado para amostras grandes;
- é fracamente consistente;
- converge assintoticamente para um processo gaussiano;
- é estimador de máxima verossimilhança de $S(t)$.

Estimador de Kaplan-Meier (EKM)

A variância assintótica do EKM é estimada pela fórmula de Greenwood:

$$\widehat{Var}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{j:t_j<t} \left(\frac{d_j}{n_j(n_j - d_j)} \right).$$

Como $\hat{S}(t)$, para t fixo, tem distribuição assintótica Normal, segue que um intervalo de confiança aproximado para $S(t)$ é dado por:

$$\hat{S}(t) \pm z_{\alpha/2} \sqrt{\widehat{Var}(\hat{S}(t))}.$$

Existem correções para valores extremos de t, em que o intervalo acima pode apresentar valores menores que zero e maiores que um.

Exemplo

Exemplo no R.

Parte III: Comparação de Curvas de Sobrevivência

Comparação de Curvas de Sobrevivência

Considere testar a hipótese $H_0 : S_1(t) = S_2(t)$.

Estatísticas comumente usadas para comparação de curvas de sobrevivência podem ser vistas como generalizações de testes não-paramétricos para dados de sobrevivência.

Dentre os destes, podemos destacar:

- Logrank (Mantel, 1966)
- Wilcoxon (Gehan, 1965)
- Tarone-Ware (1977)

Teste Logrank

Sejam $t_1 < t_2 < \ldots < t_k$ os tempos de falha distintos da amostra formada pela combinação das duas amostras individuais.

Suponha que no tempo t_j acontecem d_j falhas e n_j indivíduos estão sob risco em um tempo imediatamente inferior a t_j na amostra combinada, respectivamente, d_{ij} e n_{ij} na amostra i; $i = 1, 2$ e $j = 1, \ldots, k$.

Em cada tempo de falha, temos:

	Grupos		
	1	2	Totais
Falha	d_{1j}	d_{2j}	d_j
Não falha	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
Totais	n_{1j}	n_{2j}	n_j

Teste Logrank

- Condicional à experiência de falha e censura até o tempo t_j (fixando as marginais de coluna) e ao número de falhas no tempo t_j (fixando as marginais de linha), a distribuição de d_{2j} é hipergeométrica.
- A média de d_{2j} é $w_{2j} = n_{2j} d_j n_j^{-1}$ e a variância de d_{2j} é $(V_j)_2 = n_{2j}(n_j - n_{2j}) d_j (n_j - d_j) n_j^{-2} (n_j - 1)^{-1}$.
- Assim, a estatística $d_{2j} - w_{2j}$ tem média zero e variância $(V_j)_2$.

Teste Logrank

- Se as k tabelas de contingência forem independentes, um teste aproximado para a igualdade das duas funções de sobrevivência pode ser baseado na estatística:

$$T = \frac{\left[\sum_{j=1}^{k}(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k}(V_j)_2}.$$

- Sob a hipótese nula T tem uma distribuição qui-quadrado com 1 grau de liberdade para amostras grandes, ou seja, $T \sim \chi^2_{(1)}$.
- É possível generalizar o teste Logrank para testar a igualdade de $r > 2$ funções de sobrevivência $S_1(t), \ldots, S_r(t)$.

Família de Testes

- Família de Testes para comparação de duas funções $S(t)$:

$$S = \frac{\left[\sum_{j=1}^{k} u_j(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k} u_j^2(V_j)_2},$$

com u_j os pesos que especificam o tipo de diferença a ser detectada.

  - Logrank: $u_j = 1,\ j = 1, \ldots, k.$
  - Wilcoxon: $u_j = n_j.$
  - Tarone-Ware: $u_j = \sqrt{n_j}.$

Família de Testes

- O teste de Wilcoxon, que utiliza peso igual ao número de indivíduos sob risco, coloca mais peso na porção inicial do eixo do tempo.
- O teste Logrank coloca o mesmo peso para todo o eixo do tempo, o que reforça o enfoque nos tempos maiores, quando comparado ao teste de Wilcoxon.
- O teste de Tarone-Ware se localiza em uma situação intermediária.

Família de Pesos de Harrington-Fleming

Uma outra classe de pesos é dada por:

$$u_j = \left[\hat{S}(t_{j-1})\right]^{\rho}.$$

- Se $\rho = 0$, obtém-se $u_j = 1$ e tem-se o teste Logrank.
- se $\rho = 1$, o peso é o Kaplan-Meier no tempo de falha anterior e, neste caso, tem-se um teste similar ao de Wilcoxon.

O `R` utiliza esta família de testes no seu comando `survdiff`. Por exemplo: `survdiff(Surv(tempos,cens)~grupos,rho=0)`

Exemplo

Exemplo no R.

Técnicas Não-paramétricas

- Vantagens:
  - Fácil de entender;
  - Suposições fracas (não impõe distribuição para T).
- Desvantagens:
  - Pouco eficientes;
  - Difícil de incluir covariáveis na análise.

Parte III: Modelos Probabilísticos

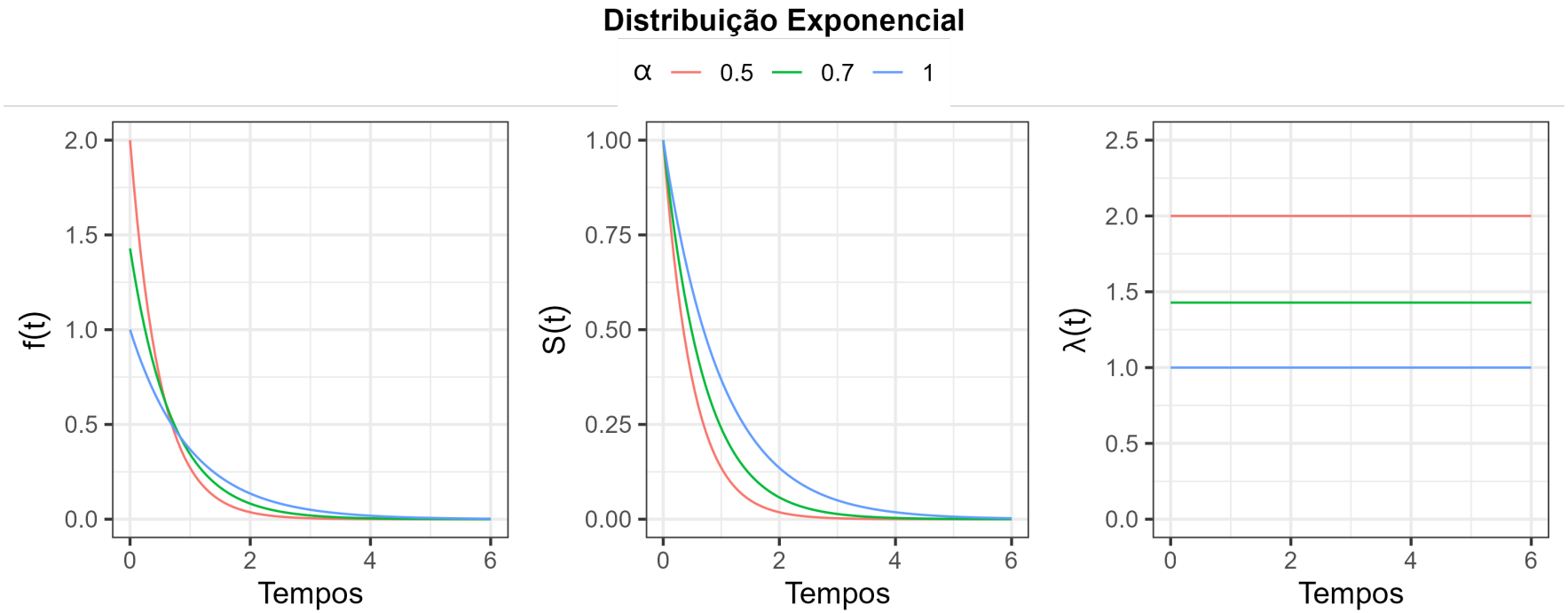
Modelos Probabilísticos

Há um grande número de modelos probabilísticos que podem ser usados na análise de tempo até o evento.

Destacamos alguns deles, por sua grande aplicação em situações práticas:

- Modelo exponencial;
- Modelo Weibull;
- Modelo Log-Normal;
- Modelo Log-logístico;
- Modelo Gama;
- Modelo Gama generalizada.

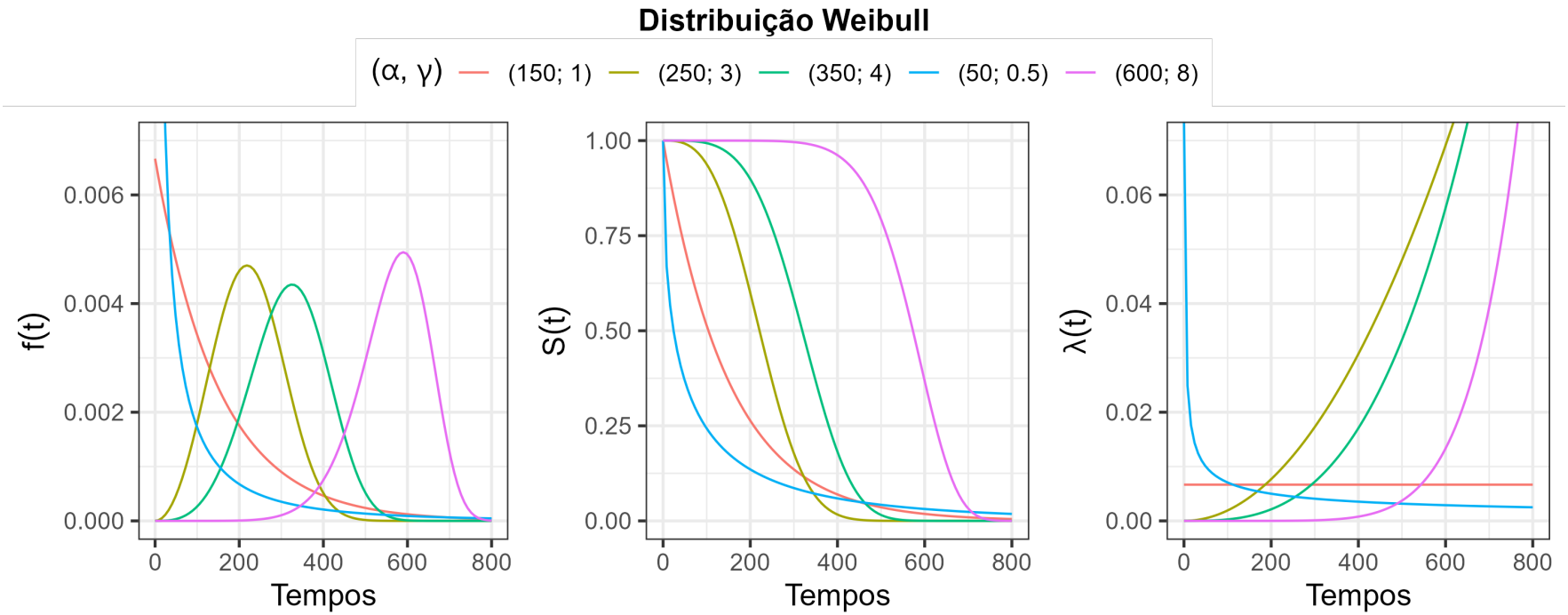
Modelo Exponencial



$$f(t) = \frac{1}{\alpha}\exp\left\{-\left(\frac{t}{\alpha}\right)\right\};\ S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)\right\};\ \lambda(t) = \frac{1}{\alpha},\ \alpha > 0$$

$$\text{Média} = \alpha;\quad \text{Variância} = \alpha^2;\quad \text{Percentil } 100p\% = t_p = -\alpha\log(1-p)$$

Modelo Weibull



$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\};\ S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\};\ \lambda(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1},$$

$$\alpha, \gamma > 0.$$

Modelo Weibull

- Quando $\gamma = 1$, tem-se a distribuição exponencial.
- A função $\lambda(t)$ é monótona: estritamente crescente para $\gamma > 1$, estritamente decrescente para $\gamma < 1$ e constante para $\gamma = 1$.
- $E[T] = \alpha\Gamma[1 + (1/\gamma)]$, sendo $\Gamma(k)$ a função gama, definida por $\Gamma(k) = \int_0^\infty x^{k-1}e^{-x}dx$.
- $Var[T] = \alpha^2 \left[\Gamma[1 + (2/\gamma)] - \Gamma[1 + (1/\gamma)]^2\right]$
- Percentis: $t_p = \alpha[-\log(1-p)]^{1/\gamma}$

Modelo Weibull

É conveniente trabalhar com o logaritmo dos tempos observados, o que leva à distribuição do valor extremo ou de Gumbel.

Assim, se T seja uma distribuição Weibull com parâmetros α e γ, então $Y = \log(T)$ segue uma distribuição do valor extremo com densidade dada por

$$f(y) = \frac{1}{\sigma} \exp\left\{\left(\frac{y-\mu}{\sigma}\right) - \exp\left\{\frac{y-\mu}{\sigma}\right\}\right\},$$

com $-\infty < \mu < \infty$ e $\sigma > 0$, os parâmetros de locação e escala, respectivamente. Se $\mu = 0$ e $\sigma = 1$ tem-se a distribuição do valor extremo padrão.

A relação entre os parâmetros é dada por $\gamma = 1/\sigma$ e $\alpha = \exp\{\mu\}$.

Modelo Weibull

As funções de sobrevivência e de taxa de falha são dadas, respectivamente, por:

$$S(y) = \exp\left\{-\exp\left\{\frac{y-\mu}{\sigma}\right\}\right\},$$

e

$$\lambda(y) = \frac{1}{\sigma}\exp\left\{\frac{y-\mu}{\sigma}\right\}.$$

Modelo Weibull

A média e variância são, respectivamente, $E(Y) = \mu - \nu\sigma$ e $Var(Y) = (\pi^2/6)\sigma^2$, em que ν é a constante de Euler-Mascheroni:

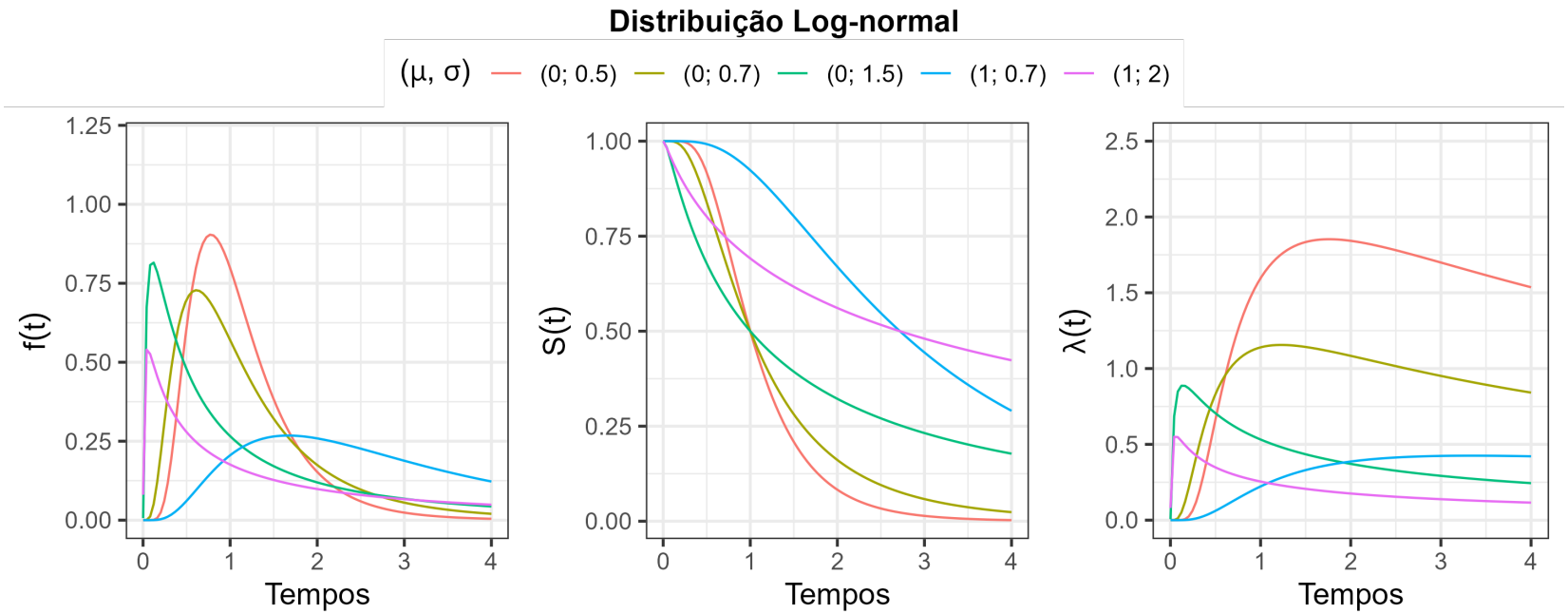
```
- digamma(1)# negativo da derivada da função gama em x=1
```

```
## [1] 0.5772157
```

O percentil $100p\%$ é dado por

$$t_p = \mu + \sigma \log\left[-\log(1-p)\right].$$

Modelo Log-normal



$$f(t) = \frac{1}{\sqrt{2\pi}t\sigma} \exp\left\{-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right\};\ S(t) = \Phi\left(\frac{-log(t)+\mu}{\sigma}\right),$$

em que $\Phi(\cdot)$ é a função de distribuição acumulada de uma normal padrão.

$$\lambda(t) = \frac{f(t)}{S(t)};\quad -\infty < \mu < \infty,\ \sigma > 0.$$

Modelo Log-normal

- $\lambda(t)$ não é monótona.
- Percentis: $t_p = \exp\{z_p\sigma + \mu\}$, com z_p o $100p\%$ percentil da distribuição normal padrão.
- $E[T] = \exp\{\mu + \sigma^2/2\}$.
- $Var[T] = \exp\{\mu + \sigma^2/2\}(\exp(\sigma^2) - 1)$.
- Se T tem distribuição log-normal, então $Y = \log(T)$ tem distribuição normal ou Gaussiana.

Modelo Log-logístico

O modelo log-logístico é uma alternativa aos modelos Weibull e log-normal.

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \left(1 + (t/\alpha)^\gamma\right)^{-2}, \ t > 0,$$

em que $\alpha > 0$ é o parâmetro de escala e $\gamma > 0$ é o parâmetro de forma.

$$S(t) = \frac{1}{1 + (t/\alpha)^\gamma},$$

e

$$\lambda(t) = \frac{\gamma(t/\alpha)^{\gamma-1}}{\alpha\left[1 + (t/\alpha)^\gamma\right]}.$$

Modelo Log-logístico

As expressões para a esperança e variância são as seguintes.

$$E(T) = \frac{\pi\alpha Csc(\pi/\gamma)}{\gamma},\ \gamma > 1$$

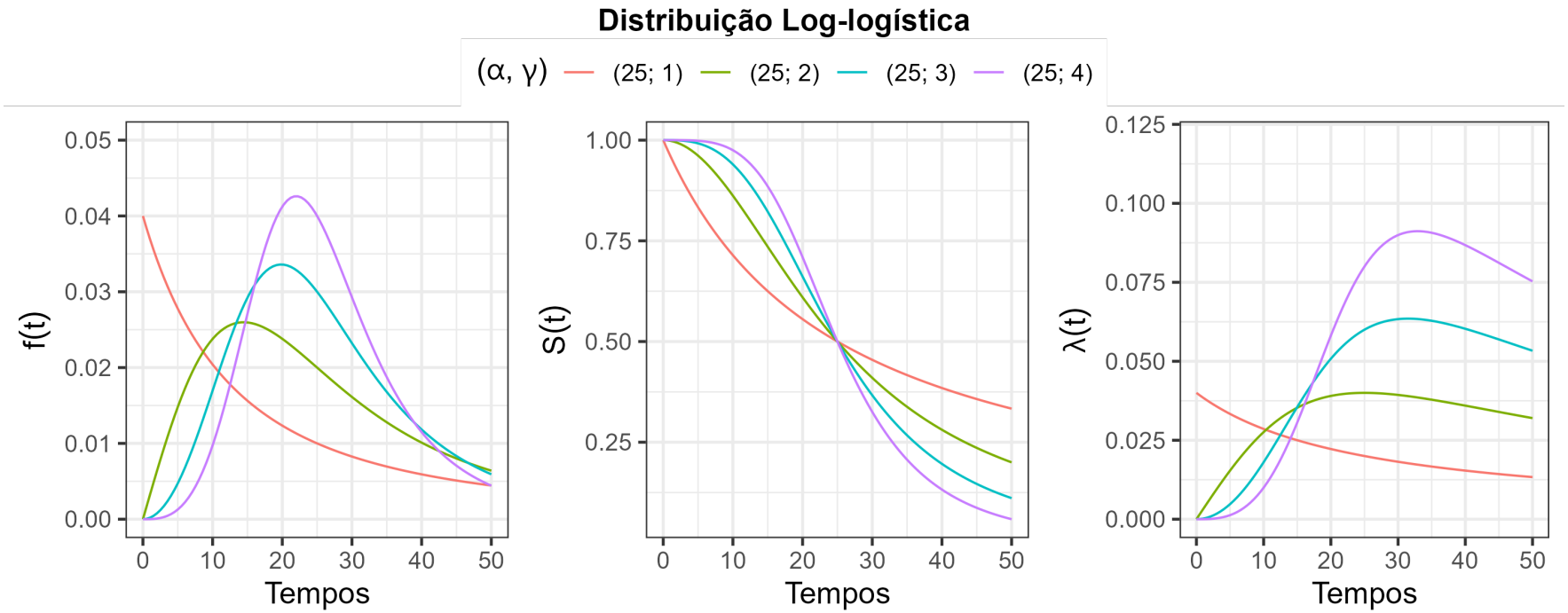
e

$$Var(T) = \frac{2\pi\alpha^2 Csc(2\pi/\gamma)}{\gamma} - E(T)^2,$$

em que $Csc(x) = 1/seno(x)$ é a função cossecante.

$$t_p = \alpha\left[\frac{p}{(1-p)}\right]^{1/\gamma}.$$

Modelo Log-logístico



Modelo Log-logístico

Assim como comentado para a distribuição Weibull, é conveniente trabalhar com o logaritmo dos tempos observados.

Se T seja uma distribuição log-logística com parâmetros α e $\gamma > 0$, então $Y = \log(T)$ segue uma distribuição logística com densidade dada por

$$f(y) = \frac{1}{\sigma} \exp\left\{\frac{y-\mu}{\sigma}\right\} \left(1 + \exp\left\{\frac{y-\mu}{\sigma}\right\}\right)^{-2},$$

com $-\infty < \mu < \infty$ e $\sigma > 0$, os parâmetros de locação e escala, respectivamente.

Modelo Log-logístico

Para este modelo, temos

$$S(y) = \frac{1}{1 + \exp\left\{\frac{y-\mu}{\sigma}\right\}},$$

e

$$\lambda(y) = \frac{1}{\sigma} \exp\left\{\frac{y-\mu}{\sigma}\right\} \left(1 + \exp\left\{\frac{y-\mu}{\sigma}\right\}\right)^{-1}.$$

A relação entre os parâmetros é dada por $\gamma = 1/\sigma$ e $\alpha = \exp\{\mu\}$.

Modelo Gama

O modelo gama é bastante utilizado em análise de sobrevivência.

A função de densidade, caracterizada pelos parâmetros de forma k ($k > 0$) e escala α ($\alpha > 0$), é dada por

$$f(t) = \frac{1}{\Gamma(k)\alpha^k} t^{k-1} \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}, \ t > 0.$$

A função de sobrevivência é dada por

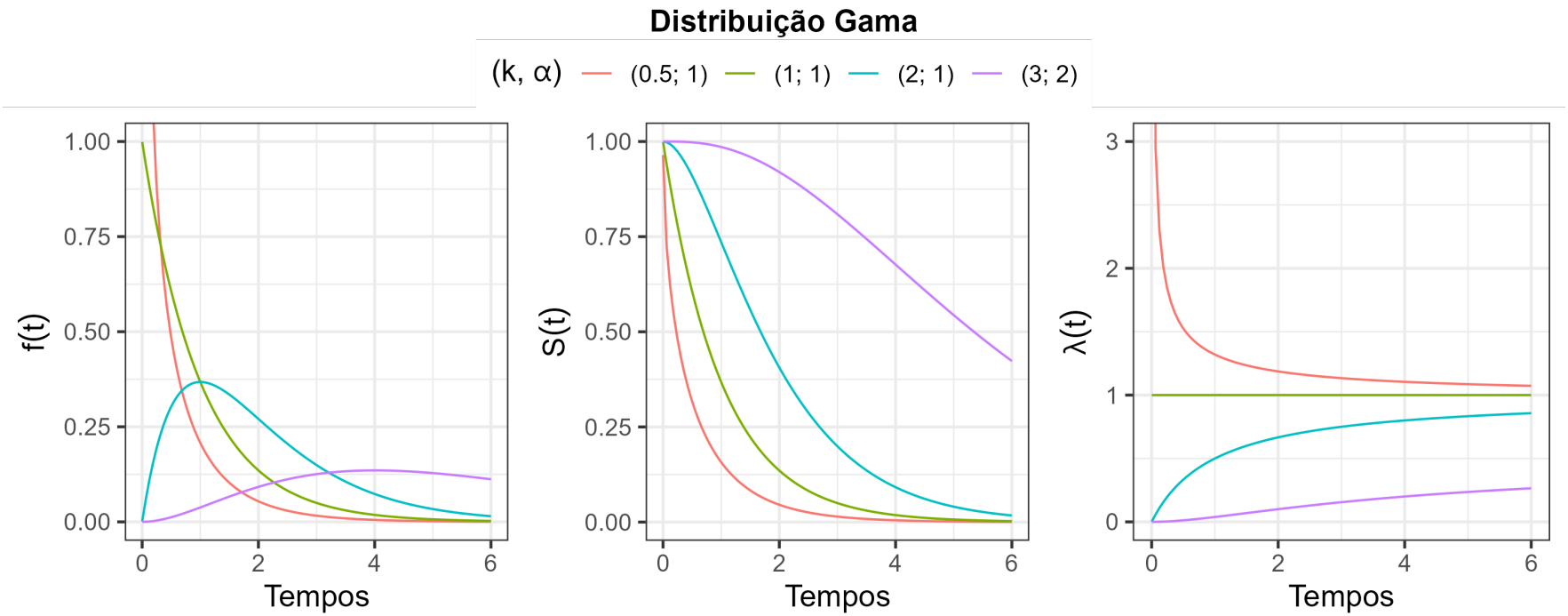
$$S(t) = \int_t^\infty \frac{1}{\Gamma(k)\alpha^k} u^{k-1} \exp\left\{-\left(\frac{u}{\alpha}\right)\right\} du.$$

Para este modelo temos que $E(T) = k\alpha$ e $Var(T) = k\alpha^2$.

Modelo Gama

A função taxa de falha, $\lambda(t) = f(t)/S(t)$, é crescente ou decrescente mas convergindo para um valor constante quanto t vai de 0 a infinito.

Mostramos a seguir as formas de $f(t)$, $S(t)$ e $\lambda(t)$ para o modelo gama.



Modelo Gama Generalizado

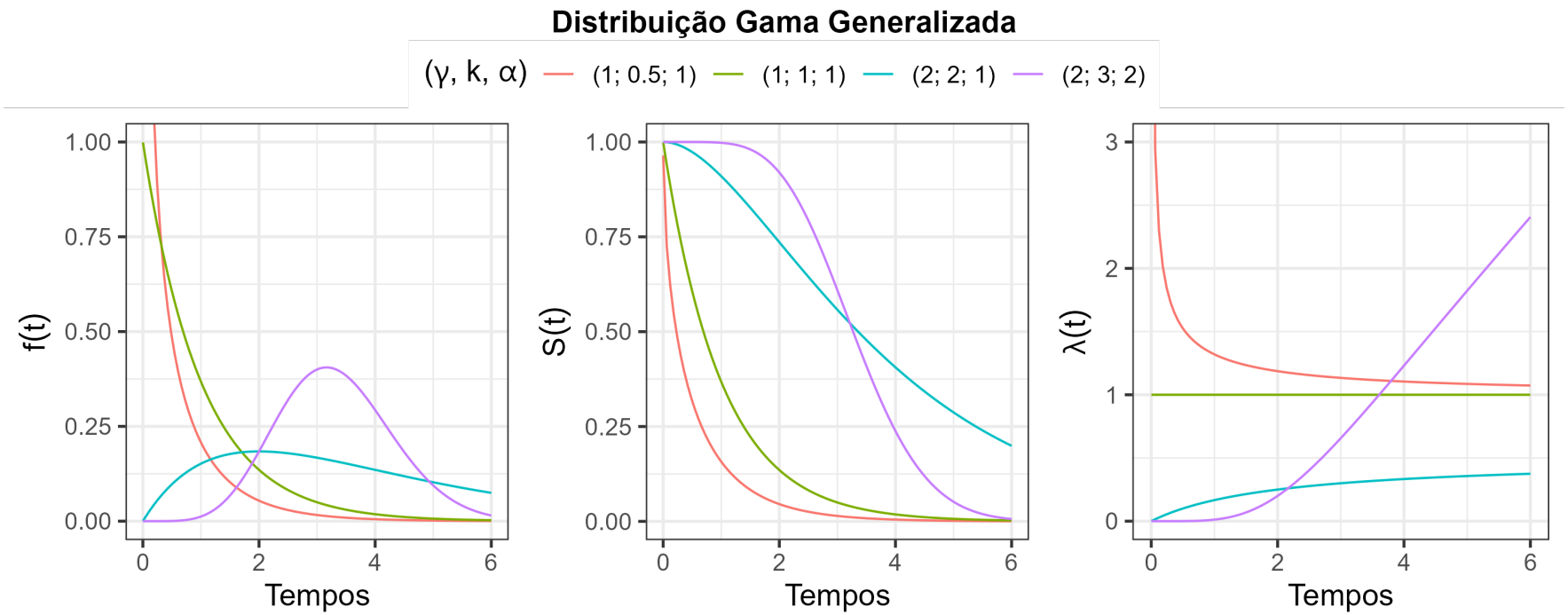
O modelo gama generalizado tem um parâmetro de escala, α, e dois de forma, γ e k, todos positivos:

$$f(t) = \frac{\gamma}{\Gamma(k)\alpha^{\gamma k}} t^{\gamma k - 1} \exp\left\{ -\left(\frac{t}{\alpha}\right)^{\gamma} \right\}.$$

Os principais modelos em análise de sobrevivência são casos particulares da gama generalizada:

- para $\gamma = k = 1$ tem-se $T \sim Exp(\alpha)$.
- para $k = 1$ tem-se $T \sim Weibull(\gamma, \alpha)$.
- para $\gamma = 1$ tem-se $T \sim Gama(k, \alpha)$.
- para $k \to \infty$ tem-se como caso limite a distribuição log-normal.

Modelo Gama Generalizado



Modelos Probabilísticos

Várias outras distribuições – como a log-gama, Rayleigh, normal inversa e Gompertz – podem ser apropriadas para modelar o tempo de falha de produtos, materiais e situações clínicas.

Se corretamente especificados, os modelos paramétricos são bastante eficientes.

A inferência para as quantidades desconhecidas dos modelos é baseada na função de verossimilhança e suas propriedades assintóticas.

Técnicas de adequação, via resíduos, são fundamentais para verificar a adequação dos modelos paramétricos.

Estimação de Parâmetros

Supondo r falhas e $n - r$ censuras, a função de verossimilhança é

$$L(\theta) = \prod_{i=1}^{r} f(t_i; \theta) \prod_{i=r+1}^{n} S(t_i; \theta),$$

ou, equivalentemente,

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} [f(t_i; \theta)]^{\delta_i} [S(t_i; \theta)]^{1-\delta_i} \\ &= \prod_{i=1}^{n} [\lambda(t_i; \theta)]^{\delta_i} S(t_i; \theta), \end{aligned}$$

em que δ_i é o indicador de falha para a i-ésima observação e $i = 1, \ldots, n$.

Estimação de Parâmetros

Os estimadores de máxima verossimilhança são os valores de θ que maximizam $L(\theta)$ ou equivalentemente o logaritmo de $L(\theta)$.

Eles são encontrados resolvendo-se o sistema de equações:

$$U(\theta) = \frac{\partial \log L(\theta)}{\partial \theta} = 0.$$

Geralmente são necessários métodos numéricos para solução do sistema de equações.

Propriedades Importantes

- Assintoticamente, temos $\hat{\theta} \sim N_k\left(\theta, Var(\hat{\theta})\right)$.
- $Var(\hat{\theta}) \approx -\left[E(\mathcal{F}(\theta))\right]^{-1}$.
- Em situações em que a esperança é impossível ou difícil de ser calculada, usa-se simplesmente $\left[\mathcal{F}(\theta)\right]^{-1}$.
- Geralmente, $Var(\hat{\theta})$ depende de θ.
- Uma estimativa para $Var(\hat{\theta})$ é obtida substituindo-se θ por $\hat{\theta}$.

Testes de Hipóteses

Dado um modelo com um vetor $\theta = (\theta_1, \ldots, \theta_p)^T$, podemos estar interessados em testar hipóteses relacionadas a este vetor ou a um subconjunto deles.

Três testes são em geral utilizados para esta finalidade:

- Teste de Wald: baseado na distribuição assintótica de $\hat{\theta}$, é uma generalização do teste t de Student.
- Teste da Razão da Verossimilhança: baseado na função de verossimilhança, envolve a comparação dos valores do logaritmo da função de verossimilhança maximizada sem restrição e sob H_0.
- Teste Escore: obtido através da função escore.

Seleção/Adequação de Modelos

- Método gráfico:
  - Modelo candidato *versus* Kaplan-Meier.
  - Linearização do modelo.
- Teste de hipóteses: utilizar o TRV para comparar o modelo proposto com a gama generalizada.

Método Gráfico

- Comparação direta da $S(t)$ estimada do modelo proposto com o Kaplan-Meier (no mesmo gráfico). Duas formas:
  - $\hat{S}_{KM}$ vs $\hat{S}_{MP}$;
  - $\hat{S}_{KM}$ vs o tempo e $\hat{S}_{MP}$ vs o tempo.
- Linearização da $S(t)$ para comparação com uma reta. Exemplos:
  - Exponencial: gráfico de $-\log[\hat{S}(t)]$ versus t deve ser aproximadamente linear, passando pela origem, se este modelo for apropriado. $\hat{S}(t)$ é o estimador de Kaplan-Meier.
  - Weibull: gráfico de $\log\left[-\log[\hat{S}(t)]\right]$ versus $\log(t)$ deve ser aproximadamente linear, passando pela origem, se este modelo for apropriado.

Teste de Hipóteses: modelo gama generalizado

As hipóteses a serem testadas são:

$$H_0: \text{ O modelo de interesse é adequado}$$

versus uma hipótese alternativa vaga, de que o modelo não é adequado.

Usamos a estatística da razão de verossimilhanças em modelos encaixados:

$$TRV = -2 \log \left[\frac{L(\hat{\theta}_M)}{L(\hat{\theta}_G)} \right] = 2 \left[\log L(\hat{\theta}_G) - \log L\hat{\theta}_M) \right].$$

Sob H_0, $TRV \sim \chi^2$ com g.l. igual a diferença do número de parâmetros ($\hat{\theta}_G$ e $\hat{\theta}_M$) dos modelos sendo comparados.

Exemplo

Exemplo no R.

Modelos de Regressão Paramétrico

Estudos em geral envolvem covariáveis que podem estar relacionadas com o tempo de sobrevivência, que devem ser incluídas na análise.

A forma mais eficiente de acomodar covariáveis é através de modelos de regressão.

Duas classes de modelos frequentemente utilizados são os modelos paramétricos e os modelos semi-paramétricos.

Modelos de Regressão Paramétrico

- Os modelos paramétricos (ou de tempo de vida acelerado) são utilizados com mais frequência na área industrial do que na médica, pois os estudos industriais podem ser planejados e as fontes de perturbação controladas.

- A segunda classe de modelos, também chamada simplesmente de modelo de regressão de Cox (Cox, 1972):
  - abriu uma nova fase na modelagem de dados clínicos;
  - é um dos mais populares na análise de sobrevivência;
  - permite incorporar covariáveis dependentes do tempo, que ocorrem com frequência em várias áreas de aplicação.

Modelos Paramétricos

- Modelo de Regressão Exponencial:

$$S(t|x) = \exp\left\{-\left(\frac{t}{\exp\{x'\beta\}}\right)\right\}.$$

- Modelo de Regressão Weibull:

$$S(t|x) = \exp\left\{-\left(\frac{t}{\exp\{x'\beta\}}\right)^{1/\sigma}\right\}.$$

A inferência é realizada por meio das propriedades assintóticas dos EMV, conforme discutido anteriormente.

Adequação do Modelo Ajustado

Diversos resíduos têm sido propostos na literatura para avaliar o ajuste dos modelos de regressão paramétricos em análise de sobrevivência:

- Resíduos de Cox-Snell (1968) e os resíduos padronizados: úteis para examinar o ajuste global do modelo;
- Resíduos martingal: úteis para determinar a forma funcional (linear, quadrática etc.) de uma covariável, em geral contínua, sendo incluída no modelo de regressão;
- Resíduos deviance: que auxiliam a examinar a acurácia do modelo para cada indivíduo sob estudo.

Interpretação dos Coeficientes

A interpretação direta dos coeficientes não é uma tarefa simples.

Uma proposta razoável é de se fazer uso da razão de tempos medianos (Hosmer e Lemeshow, 1999). Considerando uma covariável binária, a razão de tempos medianos é:

$$\frac{t_{0,5}(x=1,\hat{\beta})}{t_{0,5}(x=0,\hat{\beta})} = e^{\hat{\beta}}.$$

Exemplo $e^{\hat{\beta}} = 2$. Isto significa que o tempo mediano (estimado) de um grupo é duas vezes o do outro grupo (mantendo fixas as demais covariáveis).

Exemplo

Exemplo no R.

Parte IV: Modelo de Regressão de Cox

Modelo de Regressão de Cox

- O modelo de Cox assume a seguinte forma para a função de taxa de falha:

$$\lambda(t) = \lambda_0(t)g(x'\beta),$$

em que g é uma função não-negativa tal que $g(0) = 1$.

- O componente não-paramétrico, $\lambda_0(t)$, não é especificado e é uma função não-negativa do tempo.
- $\lambda(t) = \lambda_0(t)$, função de base, pode ser obtido para $x = 0$.
- O componente paramétrico é frequentemente usado na seguinte forma multiplicativa:

$$g(x'\beta) = \exp(x'\beta) = \exp(\beta_1 x_1 + \ldots + \beta_p x_p),$$

em que β é o vetor de parâmetros associado às covariáveis.

Modelo de Regressão de Cox

- O modelo semi-paramétrico de Cox assume que:
  - as covariáveis atuam multiplicativamente na taxa de falha pela relação $g(x'\beta) = exp(x'\beta)$, e
  - nenhuma forma paramétrica para $\lambda_0(t)$.
- A suposição de taxas de falhas proporcionais significa que, para dois indivíduos diferentes i e j, temos que,

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0 \exp(x_i'\beta)}{\lambda_0 \exp(x_j'\beta)} = \exp\left\{x_i'\beta - x_j'\beta\right\}$$

  não depende do tempo.
- Ou seja, se um indivíduo no início do estudo tem taxa de morte igual a duas vezes a taxa de um segundo indivíduo, então esta razão é a mesma para todo o período de acompanhamento.

Modelo de Regressão de Cox

Modelo de Cox:

$$\lambda(t|x) = \lambda_0(t) \exp\left\{x'\beta\right\}.$$

- O modelo de regressão de Cox é caracterizado pelos coeficientes β's, que medem os efeitos das covariáveis sobre a função de taxa de falha.
- Queremos fazer inferência nos coeficientes β a partir das observações amostrais.
- A presença do componente não-paramétrico $\lambda_0(t)$ na função de verossimilhança, traz dificuldades ao processo inferencial.

Verossimilhança Parcial

A função de verossimilhança usual do modelo de Cox envolve o componente não paramétrico $\lambda_0(t)$.

Cox (1975) desenvolveu uma alternativa que condiciona no conhecimento da história passada de falhas e censuras, eliminando o componente não paramétrico da função de verossimilhança.

Esta função foi denominada de *verossimilhança parcial* e é dada por:

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp\{x_i'\beta\}}{\sum_{j \in R(t_i)} \exp\left\{x_j'\beta\right\}} = \prod_{i=1}^{n} \left(\frac{\exp\{x_i'\beta\}}{\sum_{j \in R(t_i)} \exp\left\{x_j'\beta\right\}} \right)^{\delta_i},$$

em que $R(t_i)$ é o conjunto dos índices das observações sob risco no tempo t_i.

Verossimilhança Parcial

O EMVP (Estimador de Máxima Verossimilhança Parcial) é o valor de $\hat{\beta}$ que maximiza $L(\beta)$.

EMVP é obtido resolvendo-se o sistema de equações definido por $U(\beta) = 0$, em que $U(\beta)$ é o vetor escore de primeiras derivadas da função $l(\beta) = \log\ L(\beta)$.

Isto é,

$$U(\hat{\beta}) = \sum_{i=1}^{n} \delta_i \left[x_i - \frac{\sum_{j \in R(t_i)} x_j \exp\left\{x_j' \beta\right\}}{\sum_{j \in R(t_i)} \exp\left\{x_j' \beta\right\}} \right]^{\delta_i} = 0.$$

Verossimilhança Parcial

- A função de verossimilhança parcial, $L(\beta)$ estabelecida anteriormente, não pressupõe a possibilidade de empates nos tempos observados de falha.
- Empates podem, contudo, ocorrer nos tempos de falhas devido a medições imprecisas.
- Aproximações para $L(\beta)$, quando ocorrem empates, foram propostas, dentre outros, por Breslow (1972), Peto (1972), Efron (1977).
- A função `coxph` do pacote `survival` do `R` assume de Efron (1977) como *default*.

Verossimilhança Parcial

Sob certas condições de regularidade, os estimadores de máxima verossimilhança parcial são consistentes e assintoticamente normais.

Para fazer inferências no modelo de Cox é possível, então, usar as estatísticas de Wald, da Razão de Verossimilhança e Escore.

O teste de Wald é o mais usado para testar hipóteses relativas a um único parâmetro, isto é

$$H_0 : \beta_j = \beta_{0j},\ j = 1, \ldots, p.$$

Interpretação dos Parâmetros

- No modelo de Cox, o efeito das covariáveis é de acelerar ou desacelerar a função de taxa de falha. A propriedade de taxas de falhas proporcionais do modelo é utilizada para interpretar os coeficientes estimados.
- Tomando a razão das taxas de falha de dois indivíduos i e j que têm os mesmos valores para as covariáveis com exceção da l-ésima, tem-se

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \exp\left\{\beta_l(x_{il} - x_{jl})\right\},$$

que é interpretado como a razão de taxas de falhas.

Interpretação dos Parâmetros

- Por exemplo, suponha que x_l seja uma covariável dicotômica indicando pacientes hipertensos. A taxa de morte entre os hipertensos é $\exp(\beta_l)$ vezes a taxa daqueles com pressão normal, mantidas fixas as outras covariáveis.
- Uma interpretação similar é obtida para covariáveis contínuas. Se, por ex., o efeito de idade é significativo e $e^{\hat{\beta}} = 1,05$ para este termo, tem-se com o aumento de 1 ano na idade, que a taxa de morte fica aumentada em 5%.
- Estimativa para $\exp(\beta_l)$ é obtida utilizando a propriedade de invariância do estimador de máxima verossimilhança parcial. O intervalo de 95% de confiança é dado por: $\exp\left\{\hat{\beta} \pm 1,96 \times \widehat{EP}(\hat{\beta})\right\}$.

Adequação do Modelo de Cox

- O modelo de Cox não se ajusta a qualquer situação e, como qualquer outro modelo estatístico, requer o uso de técnicas para avaliar a sua adequação.
- A violação da suposição básica de proporcionalidade das taxas de falhas pode acarretar em sérios vícios na estimação dos coeficientes do modelo (Struthers e Kalbfleisch, 1986).
- As técnicas de avaliação do modelo são baseadas em resíduos, como em outros modelos.

Adequação do Modelo de Cox

- Definir resíduo para o modelo de Cox não foi tarefa simples.
- Cox e Snell (1968) apresentam uma definição geral de resíduos.
- No entanto, os resíduos de Schoenfeld (1982) são atualmente os mais utilizados para verificar a adequação do modelo de Cox, em especial, a suposição de proporcionalidade das taxas de falhas.

Resíduos de Schoenfeld (1982)

- Para o i-ésimo indivíduo, correspondente a um evento, com covariáveis $x_i = (x_{i1}, \ldots, x_{ip})'$, o vetor de resíduos de Schoenfeld $r_i = (r_{i1}, \ldots, r_{ip})'$ e definido para cada componente r_{iq}, $q = 1, \ldots, p$, por:

$$r_{iq} = x_{iq} - \frac{\sum_{j \in R(t_i)} x_{jq} \exp\left\{x_j' \hat{\beta}\right\}}{\sum_{j \in R(t_i)} \exp\left\{x_j' \hat{\beta}\right\}}.$$

- Os resíduos padronizados de Schoenfeld são dados por:

$$s_i^* = [\mathcal{I}(\hat{\beta})]^{-1} \times r_i,$$

em que $\mathcal{I}(\hat{\beta})$ é a matriz de informação observada.

Avaliação da Proporcionalidade das Taxas – Gráfico

- Grambsch e Therneau (1994) sugerem a utilização de s_i^* para avaliar a suposição de proporcionalidade dos riscos.

- Considere o modelo de Cox dinâmico:

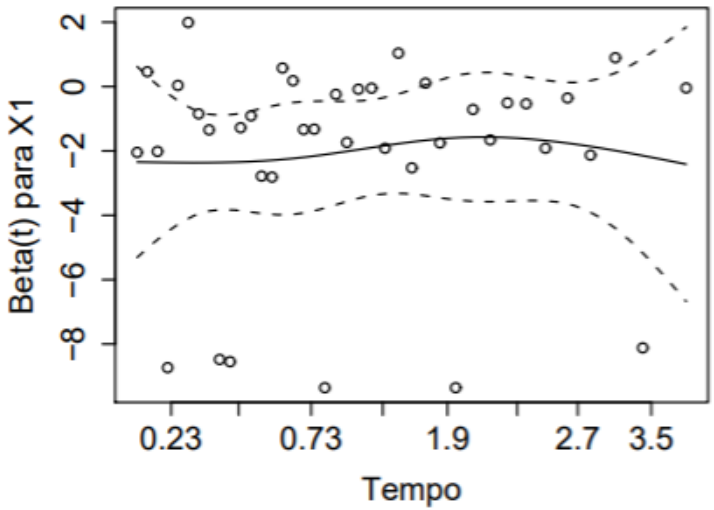
$$\lambda(t) = \lambda_0(t) \exp\left\{x'\beta(t)\right\},$$

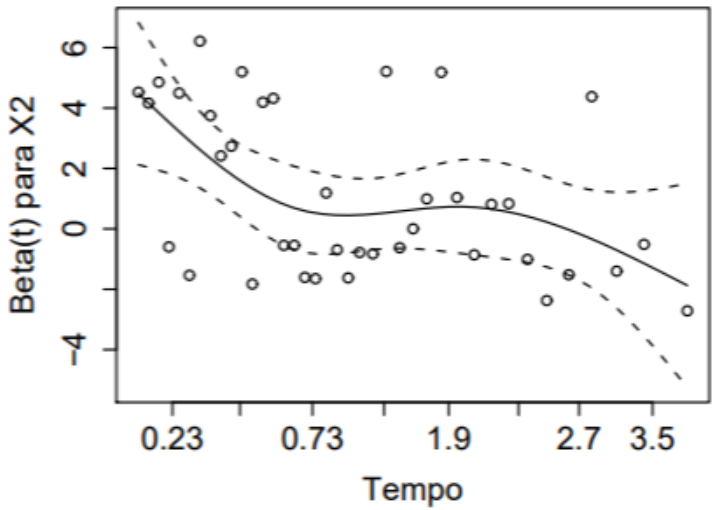
  a restrição $\beta(t) = \beta$ corresponde à proporcionalidade das taxas.

  - Se a suposição de proporcionalidade é válida, o gráfico de $\beta_q(t) \times t$ deve ser uma linha horizontal.
  - Sugestão: usar o gráfico de $(s_{iq}^* + \hat{\beta}) \times t$ ou alguma função do tempo, $g(t)$, $(q = 1, \ldots, p)$.

Avaliação da Proporcionalidade das Taxas – Gráfico

- Para auxiliar na detecção de uma possível falha da suposição de riscos proporcionais, uma curva suavizada, com bandas de confiança, é adicionada a este gráfico.
- As figuras abaixo ilustram estes gráficos (primeira, adequada e a segunda, inadequada).





Avaliação da Proporcionalidade das Taxas - Testes

Testes de hipóteses associado aos resíduos de Schoenfeld:

- O coeficiente de correlação de Pearson (ρ) entre os resíduos padronizados de Schoenfeld e $g(t)$ para cada covariável é uma dessas medidas.
- Valores de ρ próximos de zero mostram evidências a favor da suposição de riscos proporcionais.
- Estão disponíveis testes globais e locais para avaliação da proporcionalidade dos riscos.

Extensões do Modelo de Cox

- Algumas situações práticas envolvem covariáveis que são monitoradas durante o estudo, e seus valores podem mudar ao longo desse período. Tais covariáveis são chamadas de dependentes do tempo e o modelo de Cox pode ser estendido para incorporá-las.
- Em outras situações, a suposição de RP é violada e o modelo de Cox não é adequado. Modelos alternativos existem para enfrentar esta situação. Um deles é uma extensão do próprio modelo de Cox chamado de modelo de riscos proporcionais estratificado.

Exemplo

Exemplo no R.

Referências

Referência Principal

 Colosimo, E. A; Giolo, S. R. **Análise de Sobrevivência Aplicada**, *Ed. Blücher*, 2006.