Análise de Sobrevivência Aplicada à Saúde

Lupércio França Bessegato
Dep. Estatística/UFJF

Roteiro

Introdução

Técnicas Não Paramétricas

- Técnicas não paramétricas:
  - √ Fáceis de entender
  - √ Não permitem inclusão de covariáveis na análise
- Estratificação:
  - √ Simplicidade dos cálculos e facilidade de entendimento
  - √ Análise com várias covariáveis gera um número muito grande de estratos que podem conter poucas observações

Modelos de Regressão

(Apropriados para dados de sobrevivência)

- Modelos Paramétricos:
  - √ São mais eficientes e menos flexíveis

$$Y = \log(T) = \beta_0 + \beta_1 + \cdots + \beta_p + \sigma W$$

    - W: erros com distribuição especificada
  - √ Classe denominada tempo de vida acelerado é muito utilizada
    - Tempo de vida obedece à seguinte relação:

$$Y = \log(T) = \mu + \sigma W$$

      - μ: parâmetro de locação
      - σ: parâmetro de escala

- Modelos Semi-paramétricos:
  - √ Modelo de riscos proporcionais de Cox
  - √ Formado por dois elementos: um paramétrico e outro não-paramétrico
  - √ Flexibilidade
  - √ Permite incorporação de covariáveis dependentes do tempo

Distribuições do Tempo de Sobrevivência

Distribuições

- Adotadas para se ajustar a distribuição do tempo de sobrevivência T:
  - √ Exponencial
  - √ Weibull
  - √ Log-normal
  - √ Log-logística
  - √ Etc.

Distribuição Exponencial

- Densidade: $f(t) = \alpha \exp\{-\alpha t\}$
  - √ $t \geq 0$ e $\alpha > 0$
- Função de sobrevivência: $S(t) = \exp\{-\alpha t\}$
- Parâmetro:
  - √ α: representa a queda da função de sobrevivência ao longo do tempo

- Função de risco:

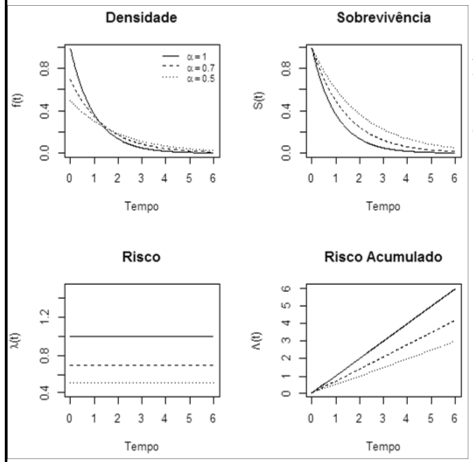
$$\lambda(t) = \frac{\alpha \exp\{-\alpha t\}}{\exp\{-\alpha t\}} = \alpha$$

  - √ Constante e igual a α
- Função de risco acumulado

$$\Lambda(t) = -\ln S(t) = \alpha t$$

  - √ É uma função linear no tempo

√ Exponencial – Forma típica das funções



Função de risco
- √ Paralela ao eixo do tempo

Quanto maior o risco α:
- √ Mais rapidamente decai a sobrevivência
- √ Maior a inclinação da reta do risco acumulado

- Exponencial – Esperança e Variância:

$$\mathrm{E}[T] = \frac{1}{\alpha}$$

$$\mathrm{Var}[T] = \frac{1}{\alpha^2}$$

- Função quantílica:

$$F(t_p) = p, \; 0 < p < 1$$

$$1 - \exp\{-\alpha t_p\} = p$$

$$t_p = -\frac{\ln(1-p)}{\alpha}$$

- Mediana:

$$t_{med} = -\frac{\ln(1-0,5)}{\alpha} = \frac{\ln(2)}{\alpha}$$

  - √ $F(t_{med}) = 0,5$

- Exponencial – Média e mediana:

$$\mathrm{E}[T] = \frac{1}{\alpha} \qquad t_{med} = \frac{\ln(2)}{\alpha}$$

√ Tempo de sobrevivência é assimétrico

– Tempo mediano é mais usual que tempo médio

√ t_{med} < E[T], para qualquer valor de α

√ E[T] e t_{med} são inversamente proporcionais ao risco

Exemplo

- Sobrevivência ao diagnóstico de Aids

√ Coorte completa entre 1986 e 2000

– Óbitos: 90

– Censuras: 103

√ Modelagem:

– Tempo entre diagnóstico de Aids e a morte segue uma distribuição exponencial

▪ Estimativa: α = 0,000497

- Encerramento de acompanhamento: t = 3.500 dias

√ Espera-se que cerca de 20% dos pacientes ainda estejam vivos

√ Sobrevivência teórica adotada – Exponencial:

$$\hat{S}_e(3500) = \exp\{-(0,000497)(1500)\} = 0,17561$$

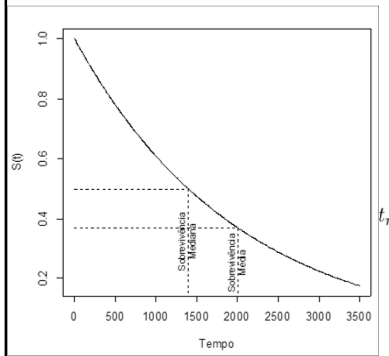
– Sobrevivência teórica tende a zero

– (Aconteceria se tempo de estudo se prolongasse até que todos os pacientes fossem acompanhados até o óbito)

- Função de risco é constante:

$$\lambda(t) = 0,000497$$

√ Coorte de Aids – Modelo paramétrico exponencial



Tempo médio de sobrevivência:

$$\mathrm{E}[T] = \frac{1}{\alpha} = \frac{1}{0,000497} = 2.012 \text{ dias}$$

Tempo mediano de sobrevivência:

$$t_{med} = \frac{\ln(2)}{\alpha} = \frac{\ln(2)}{0,000497} = 1.394 \text{ dias}$$

Sobrevivência ao tempo médio:

$$S(\mathrm{E}[T]) = \exp -1 = 0,3679$$

SEMPRE!!! (qualquer exponencial)

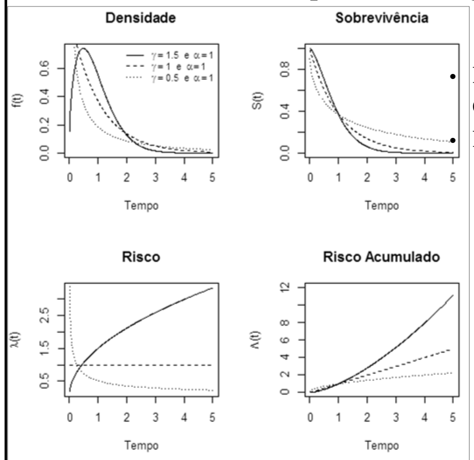
- Exponencial – Comentários:
  - √ Simplifica a modelagem
    - – É expressa por um único parâmetro
  - √ Suposição de risco constante é pouco plausível na maioria dos fenômenos da saúde
    - – Diz-se que a distribuição não tem memória
  - √ Pode ser uma aproximação válida quando o tempo de acompanhamento é curto o suficiente para se considerar constante o risco no período
    - – Risco de acidentes domésticos em crianças de 4 a 6 anos
  - √ Utilizada em alguns estudos de tempos de vida e em remissão de doenças crônicas e infecciosas

Distribuição de Weibull

- Densidade: $f(t) = \gamma\alpha^{\gamma}t^{\gamma-1}\exp\{-(\alpha t)^{\gamma}\}$
  - √ $t \geq 0$, α e $\gamma > 0$
- Função de sobrevivência: $S(t) = \exp\{-(\alpha t)^{\gamma}\}$
- Parâmetros:
  - √ α: parâmetro de escala
    (mesma unidade de t)
  - √ γ: parâmetro de forma
    (adimensional)
- A exponencial é uma caso especial da Weibull (γ=1)

- Função de risco: $\lambda(t) = \gamma\,\alpha^{\gamma}\,t^{\gamma-1}$
  - √ Função potência
  - √ Risco não é constante ao longo do tempo
    - – A função de risco é monótona
  - √ Parâmetro de forma da função de risco
    - – Modela a variação do risco no tempo
    - $\gamma < 1$: função de risco decresce
    - $\gamma > 1$: função de risco cresce
    - $\gamma = 1$: função de risco constante
- Função de risco acumulado: $\Lambda(t) = -\ln S(t) = (\alpha t)^{\gamma}$

- Weibull – Forma típica das funções



- Flexibilidade de formas da distribuição
- Parâmetro de forma γ:
  - $\gamma < 1$: função de risco decresce
  - $\gamma > 1$: função de risco cresce
  - $\gamma = 1$: função de risco constante

- Exponencial – Esperança e Variância:

$$E[T] = \frac{1}{\alpha}\Gamma\left(1+\frac{1}{\gamma}\right) \qquad \Gamma(k) = \int_0^\infty x^{k-1}e^{-x}dx$$

$$\text{Var}[T] = \frac{1}{\alpha^2}\left[\Gamma\left(1+\frac{2}{\gamma}\right) - \Gamma^2\left(1+\frac{1}{\gamma}\right)\right]$$

- Função quantílica: $F(t_p) = p,\ 0 < p < 1$

$$1 - \exp\{-(\alpha t_p)^\gamma\} = p$$

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

- Mediana: $t_{med} = \dfrac{[-\ln(1-0,5)]^{\frac{1}{\gamma}}}{\alpha} = \dfrac{[\ln(2)]^{\frac{1}{\gamma}}}{\alpha}$
  - √ $F(t_{med}) = 0,5$

Weibull – Comentários

- Para a maioria dos fenômenos de saúde
  - √ É plausível supor que o risco não é constante no tempo
- Função mais utilizada na modelagem paramétrica em saúde
- Muito flexível
  - √ O parâmetro γ modela a variação do risco no tempo

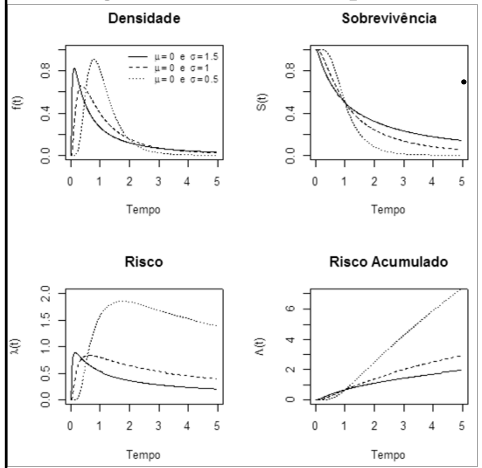
Distribuição Lognormal

- Densidade: $f(t) = \dfrac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\dfrac{1}{2}\left(\dfrac{\ln(t)-\mu}{\sigma}\right)\right\}$
  - √ $t > 0$, $-\infty < \mu < \infty$ e $\sigma > 0$
- Função de sobrevivência: $S(t) = \Phi\left(\dfrac{-\ln(t)+\mu}{\sigma}\right)$
  - √ Φ: função de distribuição acumulada da normal padrão
- Parâmetros:
  - √ O logaritmo dos dados de uma lognormal têm distribuição normal.

- Função de risco: $\lambda(t) = \dfrac{f(t)}{S(t)}$
  - √ A função de risco não é monótona como as da distribuição de Weibull
- Função de risco acumulado:

$$\Lambda(t) = -\ln S(t)$$
$$= -\ln\left[\Phi\left(\frac{-\ln(t)+\mu}{\sigma}\right)\right]$$

- Lognormal – Forma típica das funções



- Comportamento da função de risco:
  - Crescem, atingem um valor máximo e depois decrescem

Análise de Sobrevivência Aplicada à Saúde -- 2014

- Lognormal – Esperança e Variância:

$$E[T] = \exp\left\{\mu + \frac{\sigma^2}{2}\right\}$$
$$\mathrm{Var}[T] = \exp\left\{2\mu + \sigma^2\right\}\left(\exp\{\sigma^2\} - 1\right)$$

- Função quantílica:
  - √ z_p: 100p% percentil da normal padrão

$$F(t_p) = p,\ 0 < p < 1$$
$$\mathrm{P}\{T \leq t_p\} = p$$
$$\mathrm{P}\{\ln(T) \leq \ln(t_p)\} = \Phi\left(\frac{\ln(t_p) - \mu}{\sigma}\right)$$
$$t_p = \exp\{\mu + z_p\,\sigma\}$$

- Mediana: $t_{med} = \exp\{\mu\}$
  - √ $F(t_{med}) = 0,5$

Análise de Sobrevivência Aplicada à Saúde -- 2014

Lognormal – Comentários

- Muito utilizada para caracterizar tempo de vida de indivíduos
  - √ Tempo de vida de pacientes com leucemia
- $\ln(T) \sim N(\mu, \sigma^2)$
  - √ μ: média do logaritmo do tempo de sobrevivência
  - √ σ: desvio padrão do logaritmo do tempo de sobrevivência

Análise de Sobrevivência Aplicada à Saúde -- 2014

Estimação dos Parâmetros dos Modelos

Estimação de Parâmetros

- Parâmetros do modelo devem ser estimados a partir das observações amostrais
  - √ Determinação do modelo para responder as perguntas de interesse

Métodos de Estimação

- Método de Mínimos Quadrados:
  - √ Inadequado para estudo de tempo de sobrevida
  - √ Censura não é incorporada no processo de estimação
- Método de Máxima Verossimilhança:
  - √ Opção apropriada para dados censurados
  - √ Incorpora a censura
  - √ Relativamente simples de ser entendido
  - √ Possui boas propriedades para grandes amostras

Método da Máxima Verossimilhança

- Exemplo de Motivação:
  - √ Dados oriundos de população binomial com parâmetros 10 e p_0.
    - p_0: constante e desconhecido
    - Função de probabilidade de X:

$$f(x;p_0) = \mathrm{P}(X = x) = \binom{10}{x} p_0^x (1 - p_0)^{10-x}$$

    - Observa-se X = 3
  - √ Objetivo:
    - Basear-se no dado disponível para estimar o valor verdadeiro do parâmetro

- Não conhecemos p_0, mas podemos considerar o cenário em p_0 = ½.
  - √ Sob esta particular condição, a probabilidade de gerar o dado que realmente observamos (X = 3) é:

$$\mathrm{P}(X = 3) = f(3; 0,5) = \binom{10}{3} (0,5)^3 (0,5)^7 \approx 0,117$$

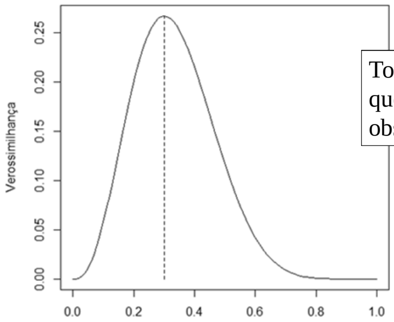
- Podemos calcular esta probabilidade sob a condição que p_0 = p

$$L(p;3) = \binom{10}{3} (p)^3 (1-p)^7,\ p \in [0,1]$$

  - √ Essa função é denominada função de verossimilhança e denotamos por $L(p;\,3)$

Princípio da Máxima Verossimilhança

- Devemos usar como nossa estimativa de p_0 o valor de p que faz $L(p; 3)$ o maior possível



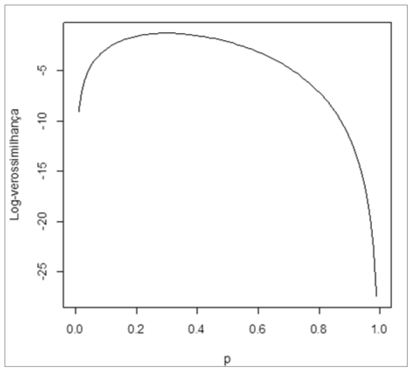
Toma-se o valor do parâmetro que torna mais provável o dado observado

$$\mathrm{P}(X=3) = f(3; 0,3) = \binom{10}{3}(0,3)^3(0,7)^7 \approx 0,267$$

Função de Log-Verossimilhança

- Como o log é uma função crescente, o valor de p que maximiza $L(p; 3)$ é o mesmo que maximiza log $L(p; 3)$
  - √ Em geral, é conveniente maximizar log $L(p; 3)$ ao invés de $L(p; 3)$
  - √ Assim, em nosso exemplo, a função de log-verossimilhança é definida como:

$$l(p;3) \stackrel{\text{def}}{=} \ln L(p;3)$$
$$= 3\ln p + 7\log(1-p) + \ln\binom{10}{3}$$

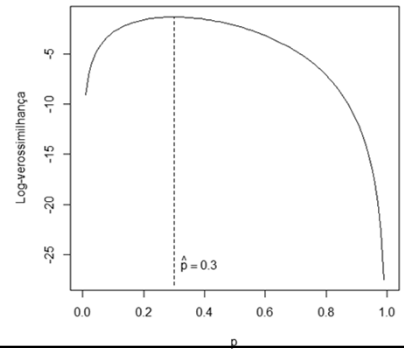


$$l(p;3) = 3\ln p + 7\ln(1-p) + \ln\binom{10}{3}$$

√ Pode-se calcular o ponto máximo da função

– ponto no domínio em que a derivada é zero

$$\hat{p} = \frac{3}{10}$$



- A estimativa é determinada pelo valor de X
  - √ Se tivéssemos observado X = k, teríamos a estimativa $\hat{p} = \frac{k}{10}$
- Em nosso exemplo, o estimador de máxima verossimilhança é?

$$\hat{p} = \frac{X}{10}$$

Máxima Verossimilhança – Procedimento de Estimação

- Baseado nos resultados observados, entre todas as distribuições definidas pelos possíveis valores de seus parâmetros, qual é aquela com maior probabilidade de ter gerado tal amostra?
  - √ A função de verossimilhança avalia o quanto os dados apoiam, concordam ou suportam cada valor possível do parâmetro a ser estimado

Função de Verossimilhança – Definição

- Suponha T uma variável aleatória com distribuição de probabilidades f(t; θ), em que θ é um único parâmetro desconhecido.
  - √ Sejam $t_1, t_2, ..., t_n$ os valores observados de uma certa população de interesse.
  - √ A função de verossimilhança para θ é expressa por:

$$L(\theta; t_1, t_2, \ldots t_n) = f(t_1;\theta) f(t_2;\theta) \ldots f(t_n;\theta)$$
$$= \prod_{i=1}^{n} f(t_i, \theta)$$

Estimador de Máxima Verossimilhança

- O estimador de máxima verossimilhança (EMV) de θ é o valor $\hat{\theta}$ que maximiza a função de verossimilhança L(θ; t)

$$L(\hat{\theta}; \mathbf{t}) = \max_{\theta \in \Theta} L(\theta; \mathbf{t})$$

  - √ Valor de θ que maximiza a probabilidade da amostra observada ocorrer.
  - √ No caso discreto, o EMV é um estimador que maximiza a probabilidade de ocorrência dos valores da amostra

Exemplo – Distribuição de Bernoulli

- Seja X uma variável aleatória de Bernoulli
  - √ Função de probabilidade:

$$f(x;p) = \begin{cases} p^x(1-p)^{1-x} & , x = 0,1 \\ 0 & , \text{caso contrário} \end{cases}$$

p é o parâmetro desconhecido a ser estimado

- Função de verossimilhança da amostra

$$\begin{aligned} L(p;\mathbf{x}) &= p^{x_1}(1-p)^{1-x_1}p^{x_2}(1-p)^{1-x_2}\dots p^{x_n}(1-p)^{1-x_n} \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{\sum_{i=1}^n (1-x_i)} \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \end{aligned}$$

- Função de log-verossimilhança

$$l(p;\mathbf{x}) = \sum_{i=1}^{n} x_i \ln(p) + \left(n - \sum_{i=1}^{n} x_i\right)\ln(1-p)$$

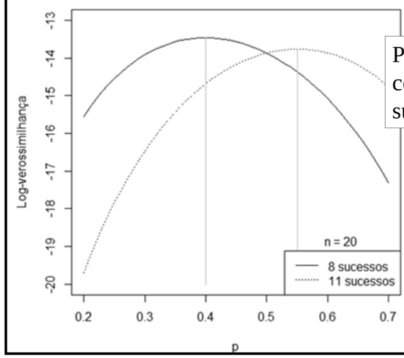
- Estimativa de máxima verossimilhança da amostra **x**:

$$\hat{p} = \frac{\sum_{i=1}^n x_i}{n}$$

- Estimador de máxima verossimilhança para amostras de Bernoulli $\hat{p} = \bar{X}$
  - √ Proporção de sucessos na amostra

Exemplo

- Bernoulli com parâmetro p desconhecido
  - √ Duas amostras de n= 20 com $\sum_{i=1}^{20} x_i = 8$ e $\sum_{i=1}^{20} x_i = 11$



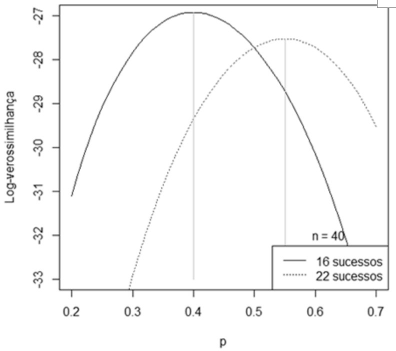
Pontos de máxima verossimilhança correspondentes à proporção de sucessos de cada amostra

$$\hat{p}_1 = \frac{8}{20} = 0,40$$

$$\hat{p}_2 = \frac{11}{20} = 0,55$$

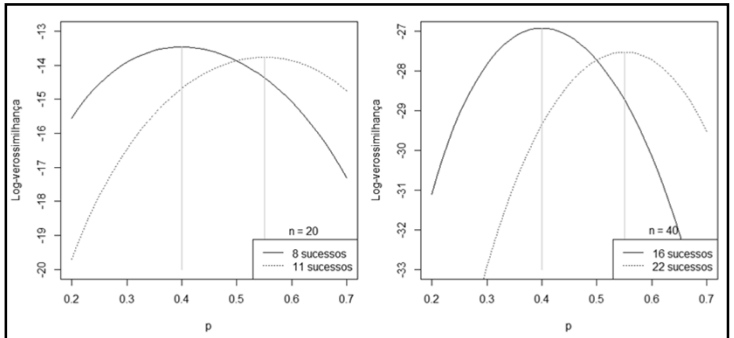
√ Duas amostras de n= 40 com $\sum_{i=1}^{20} x_i = 16$ e $\sum_{i=1}^{20} x_i = 22$

Pontos de máxima verossimilhança correspondentes à proporção de sucessos de cada amostra



$$\hat{p}_1 = \frac{16}{40} = 0,40$$

$$\hat{p}_2 = \frac{22}{40} = 0,55$$



√ As curvas crescem (decrescem) mais acentuadamente nas proximidades dos pontos de máximo na amostra maior (n = 40)

– Variância do EMV é menor para amostras maiores)

Exemplo

- Amostra para estimar prevalência de hipertensão
  - √ 10% dos indivíduos da amostra são hipertensos
  - √ É muita baixa a plausibilidade de a proporção de hipertensos na população ser 90%
  - √ Quanto mais próximo de 10% maior a verossimilhança
    - – Máxima verossimilhança

Análise de Sobrevivência Aplicada à Saúde -- 2014

Função de Verossimilhança

- Dados não censurados:
  - √ Contribuição de cada observação:
    - – Função de densidade de probabilidade
  - √ Função de verossimilhança:

$$L(\theta; t_1, t_2, \ldots t_n) \propto \prod_{i \in O} f(t_i, \theta)$$

O: conjunto das observações de desfecho

Análise de Sobrevivência Aplicada à Saúde -- 2014

- Dados censurados à direita:
  - √ Contribuição de cada observação:
    - Informa que o tempo de sobrevida é maior que o tempo observado
    - Contribuição para L(θ) é sua função de sobrevivência (não é sua função de densidade!)
  - √ Função de verossimilhança:

$$L(\theta) \propto \prod_{i \in O} f(t_i, \theta) \prod_{i \in D} S(t_i, \theta)$$

O: conjunto das observações de desfecho

D: conjunto das observações que sofreram censura à direita

- Dados censurados à esquerda:
  - √ Contribuição de cada observação:
    - Sabemos somente que o tempo de sobrevivência é menor que o tempo observado
    - Contribuição para L(θ) é sua função de distribuição acumulada:

$F(t; \theta) = 1 - S(t; \theta)$

  - √ Função de verossimilhança:

$$L(\theta) \propto \prod_{i \in O} f(t_i, \theta) \prod_{i \in D} S(t_i, \theta) \prod_{i \in E} [1 - S(t_i; \theta)]$$

O: conjunto das observações de desfecho

E: conjunto das observações que sofreram censura à esquerda

D: conjunto das observações que sofreram censura à direita

- Dados com censura intervalar:
  - √ Contribuição de cada observação:
    - Sabemos que o tempo de sobrevivência está dentro de um intervalo entre t^- e t^+.
    - Contribuição para L(θ) é a probabilidade de ocorrência de de desfecho nesse intervalo de tempo:

$S(t^-; \theta) - S(t^+; \theta)$

  - √ Função de verossimilhança:

$$L(\theta) \propto \prod_{i \in O} f(t_i, \theta) \prod_{i \in D} S(t_i, \theta) \prod_{i \in E} [1 - S(t_i; \theta)] \prod_{i \in I} \left[S(t_i^-; \theta) - (t_i^+; \theta)\right]$$

O: conjunto das observações de desfecho

D: conjunto das observações que sofreram censura à direita

E: conjunto das observações que sofreram censura à esquerda

I: conjunto das observações que sofreram censura intervalar

Estimadores de Máxima Verossimilhança – Dados Censurados

- Forma geral da verossimilhança
  - √ considerados os mecanismos de censura à direita:

$$L(\theta) = \prod_{i \in O} f(t_i, \theta) \prod_{i \in D} S(t_i, \theta)$$

  - √ ou

$$L(\theta) = \prod_{i=1}^{n} [f(t_i, \theta)]^{\delta_i} [S(t_i, \theta)]^{1-\delta_i}]$$

  - √ Estimadores de máxima verossimilhança são os valores de θ que maximizam L(θ) ou, equivalentemente, l(θ) = ln(L(θ)).

√ Os estimadores de máxima verossimilhança são encontrados resolvendo-se o sistema de equações

$$U(\theta) = \frac{\partial \ln(\theta)}{\partial \theta} = \frac{\partial l(\theta)}{\partial \theta} = 0$$

√ A função de verossimilhança pode também ser expressa em termos da função de risco:

$$L(\theta) = \prod_{i=1}^{n} [\lambda(t_i, \theta)]^{\delta_i} S(t_i, \theta)$$

Distribuição Exponencial

√ Suponha uma amostra de n itens de uma população exponencial, em que r ≤ n são desfechos e os demais (r – n) são censuras

$$\begin{aligned} L(\alpha) &= \prod_{i=1}^{n} [f(t_i, \alpha)]^{\delta_i} [S(t_i, \alpha)]^{1-\delta_i} \\ &= \prod_{i=1}^{n} [\alpha \exp\{-\alpha t_i\}]^{\delta_i} [\exp\{-\alpha t_i\}]^{1-\delta_i} \\ &= \prod_{i=1}^{n} (\alpha)^{\delta_i} \exp\{-\alpha t_i\} \\ &= \alpha^{\sum_{i=1}^{n} \delta_i} \exp\{-\alpha \textstyle\sum_{i=1}^{n} t_i\} \end{aligned}$$

√ Função da log-verossimilhança:

$$\begin{aligned} l(\alpha) &= \ln\left(\alpha^{\sum_{i=1}^{n} \delta_i} \exp\{-\alpha \textstyle\sum_{i=1}^{n} t_i\}\right) \\ &= \sum_{i=1}^{n} \delta_i \ln(\alpha) - \alpha \sum_{i=1}^{n} t_i \end{aligned}$$

√ Estimador de máxima verossimilhança:

$$\begin{aligned} \frac{\partial l(\alpha)}{\partial \alpha} &= \frac{\sum_{i=1}^{n} \delta_i}{\alpha} - \sum_{i=1}^{n} t_i \\ &= \frac{r}{\alpha} - \sum_{i=1}^{n} t_i \end{aligned}$$

$$\frac{\partial l(\alpha)}{\partial \alpha} = 0 \Rightarrow \hat{\alpha} = \frac{r}{\sum_{i=1}^{n} t_i}$$

$$\hat{\alpha} = \frac{\text{total desfechos}}{\text{tempo total sob teste}}$$

Se não houver censuras $\quad \hat{\alpha} = \frac{n}{\sum_{i=1}^{n} t_i} = \frac{1}{\bar{x}}$

Exemplo

- Sobrevivência ao diagnóstico de Aids
  - √ Pacientes com tempo de observação menor que 90 dias
    - Óbitos: 15
    - Censuras: 6
  - √ Suposição:
    - Tempo entre diagnóstico de Aids e a morte segue uma distribuição exponencial

- Estimação do parâmetro a da exponencial:

$$\hat{\alpha} = \frac{\text{total de desfechos}}{\sum_{i=1}^{n} t_i} = \frac{15}{1012} = 0,0148$$

√ Taxa de mortalidade por pessoa-tempo
  - 0,015 morte por pessoa-mês
  - 17,8 mortes por 100 pessoas/ano

- Tempo médio de sobrevivência:

$$\hat{\mathrm{E}}[T] = \frac{1}{\hat{\alpha}} = 67,5 \text{ meses}$$

- Função de Sobrevivência estimada

$$\hat{S}_e(t) = \exp\{-\hat{\alpha}t\} = \exp\{0,0148t\}$$

- Estimativa da probabilidade de um paciente estar vivo depois de 16 dias do diagnóstico de Aids

√ Paramétrica – Exponencial:

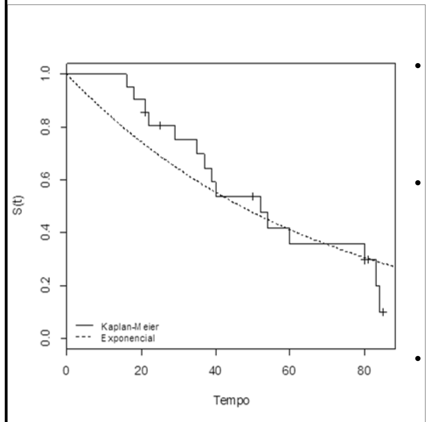
$$\hat{S}_e(16) = \exp\{-(0,0148)(16)\} = 0,7891$$

√ Não Paramétrica – Kaplan-Meier

$$\hat{S}_K M(16) = 0,9524$$

√ As estimativas são bastante diferentes
  - A suposição de distribuição exponencial é adequada aos dados?

- Estimativas paramétrica e não paramétrica



- Exponencial está afastada dos dados?
  - Ela é adequada à situação? Risco constante?
- As distribuições de Weibull ou Lognormal seriam mais próximas dos dados?
  - Como estimar seus parâmetros?
- O risco é constante?

Modelos Paramétricos de Sobrevivência

- Parametrização dos modelos no R

√ Exponencial:
  - Parâmetro α: $\alpha = \exp\{-\text{Intercept}\}$

√ Weibull:
  - Parâmetro α: $\alpha = \exp\{-\text{Intercept}\}$
  - Parâmetro γ: $\hat{\gamma} = \dfrac{1}{\text{scale}}$

Modelos Paramétricos de Regressão

- Parametrização dos modelos Weibull no R:
  - √ Na presença de covariáveis, os parâmetros de regressão β ficam com o sinal inverso ao que aparece na saída do pacote

√ Modelo Exponencial:

```
>          # Exponencial
>
> sobrev.e<-survreg(Surv(tempo, status)~1,data = ipec90, dist='exponential')
> sobrev.e
```

√ Saída:

```
> sobrev.e
Call:
survreg(formula = Surv(tempo, status) ~ 1, data = ipec90, dist = "exponential")
Coefficients:
(Intercept)
   4.211634
Scale fixed at 1
Loglik(model)= -78.2   Loglik(intercept only)= -78.2
n= 21
```

√ Estimação parâmetro:

```
> alfa.exp<-exp(-sobrev.e$coefficients[1]) #(na forma alpha t)
> alfa.exp
(Intercept)
 0.01482213
```

√ Sobrevivência estimada: $\hat{S}_e(t) = \exp\{-0,01482\,t\}$

Outras Distribuições

- Tempos de sobrevivência com distribuição de Weibull ou Lognormal:
  - √ A construção da verossimilhança é feita da mesma forma
  - √ (substitui-se funções de densidade e sobrevivência)
  - √ Equações não podem ser resolvidas analiticamente
  - √ Necessário utilizar métodos iterativos de solução
  - √ (por exemplo, método de Newton-Raphson)

√ Modelo Weibull:

```
          # Weibull

> sobrev.w<-survreg(Surv(tempo, status)~1,data = ipec90, dist='weibull')
> sobrev.w
```

√ Saída:

```
> sobrev.w<-survreg(Surv(tempo, status)~1,data = ipec90, dist='weibull')
> sobrev.w
Call:
survreg(formula = Surv(tempo, status) ~ 1, data = ipec90, dist = "weibull")
Coefficients:
(Intercept)
   4.157605
Scale= 0.5156146
Loglik(model)= -74.4   Loglik(intercept only)= -74.4
n= 21
```

√ Estimação parâmetro:

```
> alfa.weib<-exp(-sobrev.w$coefficients[1])
> gama.weib<-1/sobrev.w$scale
> cbind(gama.weib, alfa.weib)
             gama.weib  alfa.weib
(Intercept)   1.939433 0.01564499
```

√ Sobrevivência estimada: $\hat{S}_w(t) = \exp\left\{-(0,01564\,t)^{1,9394}\right\}$

√ Modelo Log-normal:

```
> # Lognormal
>
> sobrev.ln <- survreg(Surv(tempo, status)~1,data = ipec90, dist='lognorm')
> sobrev.ln
```

√ Saída:

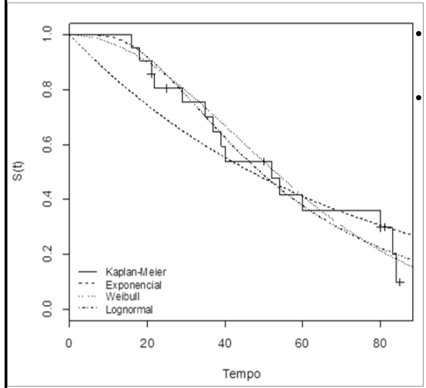
```
> sobrev.ln
Call:
survreg(formula = Surv(tempo, status) ~ 1, data = ipec90, dist = "lognorm")
Coefficients:
(Intercept)
   3.893372
Scale= 0.6425924
Loglik(model)= -73.8   Loglik(intercept only)= -73.8
n= 21
```

√ Estimação parâmetro:

```
> mi.ln <- sobrev.ln$coefficients[1]
> sigma.ln <- sobrev.ln$scale
> cbind(mi.ln, sigma.ln)
                mi.ln  sigma.ln
(Intercept) 3.893372 0.6425924
```

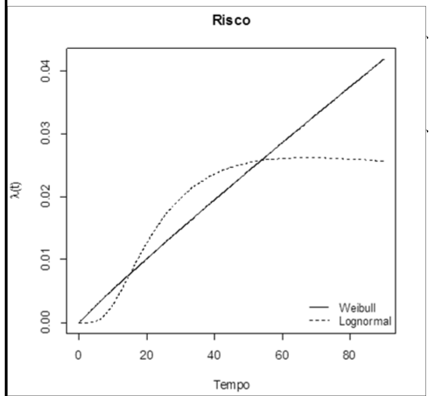
√ Sobrevivência estimada: $\hat{S}_{ln}(t) = \Phi\left[\frac{-\ln(t) - 3{,}8934}{0{,}6426}\right]$

- Curva de Sobrevivência
  - √ Estimativas paramétricas e não paramétrica



- Exponencial está afastada dos dados
- Weibull e Lognormal estão mais próximas
  - Weibull se afasta para tempos de sobrevivência mais longos

- Função de Risco
  - √ Modelos Weibull e Lognormal



- Qual das funções de risco representam o problema mais adequadamente?
- Que critérios adotar para escolher o modelo?

Intervalos de Confiança

- Intervalos de Confiança e Testes de hipóteses:
  - √ Método de máxima verossimilhança permite a construção de intervalos de confiança para os parâmetros e quantidades de interesse.
- Distribuição assintótica do EMV $\hat{\theta}$:
  - √ Para grandes amostras: $\hat{\theta} \overset{a.s.}{\sim} N(\theta, \mathrm{Var}(\hat{\theta}))$
  - √ Var($\hat{\theta}$): variância de $\hat{\theta}$:
    - Em geral, depende de θ.

√ Intervalo aproximado de (1 – a)100% de confiança para θ:

$$\hat{\theta} \pm z_{\alpha/2}\,\mathrm{ep}(\hat{\theta})$$

– ep($\hat{\theta}$): erro padrão de $\hat{\theta}$:

Testes de Hipóteses

- Seja um modelo com vetor de parâmetros

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$$

  √ Deseja-se testar hipóteses relacionadas a este vetor ou a algum de seus subconjuntos
- Testes utilizados:
  √ Teste da Razão de Verossimilhanças
  √ Teste de Wald
  √ Teste Escore

Teste da Razão de Verossimilhanças

- Compara os valores da log-verossimilhança maximizada sem restrição e sob H_0
  √ Hipótese:
  – H_0: $\theta = \theta_0$.
  √ Estatística de teste:

$$\mathrm{RV} = 2\ln\left[\frac{L(\hat{\theta}_0)}{L(\hat{\theta})}\right] = 2\left[\ln L(\hat{\theta}) - L(\hat{\theta}_0)\right]$$
$$= 2[l(\hat{\theta}) - l(\hat{\theta}_0)]$$

  √ Distribuição amostral: $\mathrm{RV} \overset{as.}{\sim} \chi^2_{gl}$
  – gl: diferença entre o número de parâmetros não especificados em H_0 e H_1

Teste de Wald

- Baseado na distribuição assintótica de θ:
  √ Em geral, usado para testar hipóteses relativas a um único parâmetro θ_j.
  √ Hipótese:
  – H_0: $\theta = \theta_0$.
  √ Estatística de teste: $\mathrm{W} = \frac{(\hat{\theta}-\theta_0)^2}{\widehat{\mathrm{Var}}(\hat{\theta})} \overset{as.}{\sim} \chi^2_1$
  – ou, equivalentemente $\frac{\hat{\theta}-\theta_0}{\widehat{\mathrm{ep}}(\hat{\theta})} \overset{as.}{\sim} N(0,1)$

Escolha do Modelo Probabilístico

- Especificação do modelo
  - √ É de extrema importância na análise paramétrica de dados de tempo de sobrevivência
- Em muitas situações:
  - √ Não há evidências de testes anteriores de que um modelo se ajusta bem aos dados
  - √ Solução para essas situações é basicamente empírica

- Proposta empírica:
  - √ Ajustar modelos probabilísticos
    - – Comparação entre valores estimados e observados para decidir qual deles "melhor explica os dados amostrais"
  - √ Técnicas gráficas:
    - – Forma simples e eficiente para selecionar o "melhor" modelo para ser usado por um conjunto de dados
    - – Ferramenta exploratória para escolha de distribuições candidatas
  - √ Teste de hipóteses:
    - – Pode ser usado teste de hipóteses para modelos encaixados

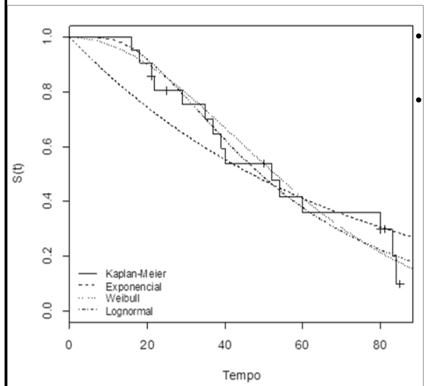
Análise de Sobrevivência Aplicada à Saúde -- 2014

Métodos Gráficos

1. Comparação do modelo proposto com o estimador de Kaplan-Meier:
   - √ Ajusta-se o modelo proposto ao conjunto de dados
   - √ Obter a estimativa de Kaplan-Meier para a sobrevivência
   - √ Comparam-se graficamente as funções de sobrevivência obtidas
   - √ Quanto mais próximo o modelo paramétrico estiver da curva do Kaplan-Meier, melhor o modelo

Análise de Sobrevivência Aplicada à Saúde -- 2014

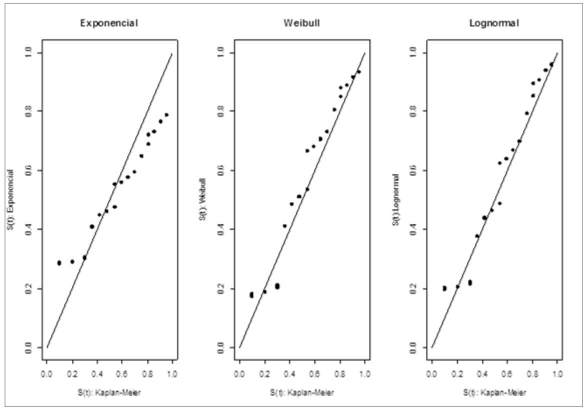
- Comparação das curvas de sobrevivência
  - √ Estimativas paramétricas e não paramétrica



- Exponencial está afastada dos dados
- Weibull e Lognormal estão mais próximas
  - Weibull se afasta para tempos de sobrevivência mais longos

Análise de Sobrevivência Aplicada à Saúde -- 2014

- Comparação entre os valores estimados
  - √ (S_{km}, $S_{paramétrico}$)



- Comentários:
  - √ Modelo exponencial parece não ser adequado aos dados:
    (Curva encontra-se um tanto afastada da reta $y = x$)
  - √ Modelo de Weibull ou log-normal possivelmente é adequado aos dados
    (acompanham mais de perto a reta $y = x$)

- Comparação do modelo proposto com o estimador de Nelson-Aalen:
  - √ Ajusta-se o modelo proposto ao conjunto de dados
  - √ Obter a estimativa de Nelson-Aalen para a a função de risco acumulado
  - √ Comparam-se graficamente as funções de sobrevivência ou de risco acumulado obtidas
  - √ Gráficos envolvendo funçãode sobrevivência ou a função de risco acumulado são úteis para discriminar modelos

- Funções de sobrevivência e de risco acumulado:

$$\Lambda(t) = -\ln[S(t)]$$

- Risco acumulado:
  - √ Modelo exponencial: $\hat{\Lambda}(t) = \hat{\alpha} t$
  - √ Modelo de Weibull: $\hat{\Lambda}(t) = (\hat{\alpha} t)^{\hat{\gamma}}$
  - √ Modelo Lognormal: $\hat{\Lambda}(t) = -\ln\left[\Phi\left(\frac{-\ln(t) + \hat{\mu}}{\hat{\sigma}}\right)\right]$

2. Linearização da função de sobrevivência:
   - √ Ideia básica:
     - Construção de gráficos que sejam aproximadamente lineares caso o modelo proposto seja apropriado
     - Violações da linearidade podem ser verificadas visualmente
   - √ Utiliza-se gráfico de transformação de linearização
     - Modelo exponencial $-\ln S(t) = \alpha t$
     - Modelo Weibull

- Linearização do modelo exponencial

$$S(t) = \exp\{-\alpha t\}$$
$$-\ln[S(t)] = \alpha t$$

√ Gráfico de $-\ln[\hat{S}_{KM}]$ vs. tempo deve ser aproximadamente linear

- Linearização do modelo de Weibull

$$S(t) = \exp\{(\alpha t)^\gamma\}$$
$$-\ln S(t) = (\alpha t)^\gamma$$
$$\ln[-\ln[S(t)]] = \gamma\ln(\alpha) + \gamma\ln(t)$$

√ Se o modelo for adequado, gráfico de $\ln[-\ln[\hat{S}_{KM}]]$ vs. tempo deve ser aproximadamente linear

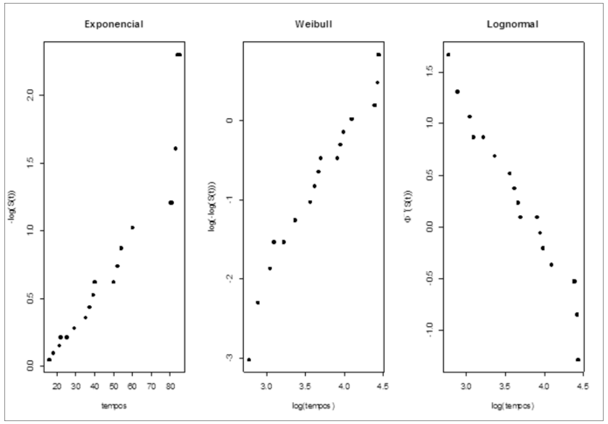
√ Indicação a favor do modelo exponencial
- Se além de linear o gráfico passar pela origem e tiver inclinação igual a 1.

- Linearização do modelo Lognormal

$$S(t) = \Phi\left(\frac{-\ln(t) + \mu}{\sigma}\right)$$
$$\Phi^{-1}[S(t)] = \frac{-\ln(t) + \mu}{\sigma}$$

√ Se o modelo for adequado, gráfico de $\Phi^{-1}[\hat{S}_{KM}]$ vs. ln(tempo) deve ser aproximadamente linear
- Intercepto: μ/σ
- Inclinação: 1/σ

- Gráficos linearizados da sobrevivência
  - √ Modelos exponencial, Weibull e lognormal



- Comentários:
  - √ Modelos de Weibull e Log-normal não mostram afastamentos marcantes de uma reta
  - √ Observa-se desvio para modelo Exponencial
  - √ Não há discriminação entre os modelos de Weibull e Lognormal
    - – Tamanho amostral pequeno pode ser a principal razão

Análise Gráfica – Comentários

- Avaliação da qualidade de ajuste do modelo
  - √ Análises gráficas são mais indicadas que testes formais de qualidade global de ajuste
  - √ Testes formais
    - – Para amostras pequenas: tendem a ter baixo poder
    - – Para amostras grandes: tendem a rejeitar o modelo
  - √ Se nenhum desses modelos for adequado
    - – Podem ser necessários modelos mais flexíveis envolvendo mais de dois parâmetros
      - ▪ Ex.: gama generalizada

- Métodos gráficos devem ser usados para descartar modelos claramente inadequados
  - √ Há situações em que os gráficos apresentados não discriminam modelos, mas indicam que eles são igualmente bons
    - – Tamanho amostral pequeno ou pequena quantidade de desfechos
  - √ Diferentes modelos com ajustes razoáveis para o mesmo conjunto de dados
    - – Em especial, nos intervalos de tempo que concentram maior número de observações
    - – Apresentam resultados semelhantes das quantidade de interesse

√ Análise de resíduos é mais adequada para demonstrar que um modelo paramétrico é melhor

- Presença de componente subjetivo na interpretação de técnicas gráficas
  √ Conclusões podem diferir para diferentes analistas

Comparação de Modelos – Teste de Hipóteses

- Outra forma de discriminar modelos
- Hipóteses
  √ H_0: modelo de interesse é adequado
  √ H_1: modelo de interesse não é adequado
- Comparação de modelos encaixados:
  √ Modelo com maior número de parâmetros contém todos os parâmetros do modelo menor
  √ Deve-se identificar um modelo generalizado tal que o modelo de interesse seja caso particular

- Ajustes para realizar o teste:
  √ Obtenção da log-verossimilhança do modelo generalizado (modelo maior): $l(\theta_G) = \ln[L(\theta_G)]$
  √ Obtenção da log-verossimilhança do modelo de interesse (modelo menor): $l(\theta_M) = \ln[L(\theta_M)]$
- Estatística de teste:
  √ Razão de verossimilhanças: $\mathrm{RV} = 2\left[l(\hat{\theta}_G) - l(\hat{\theta}_M)\right])$
  √ Distribuição amostral:

$$\mathrm{RV} \sim \chi^2_{gl}$$

  – gl: diferença entre o número de parâmetros dos dois modelos

- Comparação modelos
  √ Weibull (maior) vs. Exponencial(menor)
  √ Logaritmos da verossimilhança:

```
> sobrev.e$loglik[2]
[1] -78.1745
> sobrev.w$loglik[2]
[1] -74.39242
```

  √ Razão de Verossimilhanças:

$$\mathrm{RV} = 2\left[l(\hat{\theta}_G) - l(\hat{\theta}_M)\right]) = 2[78,1745 - 74,3942] = 7,5642$$

  √ p-valor: $\mathrm{P}\left\{\chi^2_1 > 7,5642\right\} = 0,0059$

```
> 1 - pchisq(7.5642, 1)
[1] 0.005953908
```

  – Rejeita-se H0: modelo exponencial é adequado
- Condução do teste – comando **anova**:

```
> anova(sobrev.e, sobrev.w)
  Terms Resid. Df    -2*LL Test Df Deviance     Pr(>Chi)
1     1        20 156.3490      NA       NA           NA
2     1        19 148.7848    =  1 7.564177 0.005953983
```

  √ Resultado indica inadequação do modelo exponencial

- No contexto de Análise de Sobrevivência este teste é usualmente realizado utilizando-se a distribuição gama generalizada.
  - √ Apresenta os modelos exponencial, de Weibull, lognormal e gama como modelos encaixados
- Para ajuste do modelo gama generalizado no R:
  - > **library(flexsurv)**
  - > **flexsurvreg(formula, data, dist='gengamma')**

- No exemplo:
  - √ Ajuste do modelo gama generalizado (3 parâmetros)

```
> library(flexsurv)
> sobrev.gama<-flexsurvreg(Surv(tempo, status)~1,data = ipec90,dist='gengamma')
> sobrev.gama
Call:
flexsurvreg(formula = Surv(tempo, status) ~ 1, data = ipec90,     dist =
"gengamma")
Estimates:
        est     L95%    U95%    se
mu      3.802   3.055   4.548   0.381
sigma   0.656   0.453   0.950   0.124
Q      -0.318  -2.642   2.006   1.186
N = 21,  Events: 15,  Censored: 6
Total time at risk: 1012
Log-likelihood = -73.8024, df = 3
AIC = 153.6048
```


- Logaritmos da verossimilhança:

```
> sobrev.gama$loglik
[1] -73.8024
> sobrev.e$loglik[2]
[1] -78.1745
> sobrev.w$loglik[2]
[1] -74.39242
> sobrev.ln$loglik[2]
[1] -73.83833
```

- Teste da Razão da Verossimilhança – Resultados

	l(θ)	Razão	GL	p-valor
Gama generalizada	–73,80			
Exponencial	–78,17	2(78,17–73,80)=8,74	2	0,013
Weibull	–74,39	2(74,39–73,80)=1,18	1	0,277
Log-normal	–73,84	2(73,84–73,80)=0,08	1	0,777

```
> 1-pchisq(8.74,2)
[1] 0.01265124
```

  - √ Resultados indicam adequação dos modelos de Weibull e Log-normal

- Comandos em R:
  - √ A função **anova** não funciona com objeto **flexsurvreg**

AIC – Akaike Information Criterion

- Medida de qualidade relativa de um modelo estatístico para um conjunto de dados
  - √ Estimativa da informação perdida ao se usar um modelo para representar o processo de geração dos dados
  - √ Considera o 'trade-off' entre a bondade do ajuste e a complexidade do modelo
  - √ Não informa nada no sentido absoluto
    - – O AIC não informa nada se todos os modelos candidatos não se ajustarem bem aos dados

- Critério de Informação de Akaike:

$$\mathrm{AIC} = 2\left[p - l(\hat{\theta}_M)\right]$$

  - √ Correção para amostras finitas

$$\mathrm{AIC}_c = \mathrm{AIC} + \tfrac{2p(p+1)}{n-p-1}$$

  - √ Decisão:
    - – Prefere-se o modelo com o menor valor de AIC
  - √ No exemplo

Modelo	l(θ)	AIC
Gama generalizada	–73,80	2[3 – (–73,80)] = 153,6
Exponencial	–78,17	2[1 – (–78,17)] = 158,3
Weibull	–74,39	2[2 – (–74,39)] = 152,8
Log-normal	–73,84	2[2 – (–73,84)] = 151,7

- Comentários:
  - √ Recomenda-se usar o AIC para selecionar modelos quando o número de observações (n) é maior que pelo menos 40 vezes o número de parâmetros (p)
    - – (Burham e Anderson, 2002)
  - √ O AICc é recomendado para pequenas amostras e respostas com distribuição normal
    - – Aumenta consideravelmente a probabilidade de escolha do modelo adequado. (Davison, 2001)
    - – Deve-se considerar arriscado usar para distribuições não normais um critério que foi desenvolvido para distribuição normal

Exemplo

- Pacientes com câncer de bexiga
  - √ Tempo de reincidência de pacientes com câncer de bexiga submetidos a procedimento cirúrgico com laser
  - √ 20 pacientes
  - √ Modelos analisados:
    - – Exponencial
    - – Weibull
    - – Log-normal

√ Modelo Exponencial:

```
          Exponencial

 ajuste.e<-survreg( sobrevida.cancer ~1,dist='exponential')
 ajuste.e
```

√ Saída:

```
>  ajuste.e
Call:
survreg(formula = sobrevida.cancer ~ 1, dist = "exponential")
Coefficients:
(Intercept)
   3.016111
Scale fixed at 1
Loglik(model)= -68.3   Loglik(intercept only)= -68.3
n= 20
```

√ Estimação parâmetro:

```
>  alfa.exp<-exp(-ajuste.e$coefficients[1])
>  alfa.exp
(Intercept)
0.04899135
```

√ Sobrevivência estimada: $\hat{S}_e(t) = \exp\{-0,04899t\}$

√ Modelo Weibull:

```
#Weibull

 ajuste.w<-survreg(sobrevida.cancer~1,dist='weibull')
 ajuste.w
```

√ Saída:

```
>  ajuste.w
Call:
survreg(formula = sobrevida.cancer ~ 1, dist = "weibull")
Coefficients:
(Intercept)
   3.060529
Scale= 0.647922
Loglik(model)= -66.1   Loglik(intercept only)= -66.1
n= 20
```

√ Estimação parâmetro:

```
> alfa.weib<-exp(-ajuste.w$coefficients[1])
>  gama.weib<-1/ajuste.w$scale
>  cbind(gama.weib, alfa.weib)
            gama.weib  alfa.weib
(Intercept)  1.543396 0.04686288
```

√ Sobrevivência estimada: $\hat{S}_w(t) = \exp\left\{-(0,04686t)^{1,5433}\right\}$

√ Modelo Log-normal:

```
#Log-normal

 ajuste.ln<-survreg(sobrevida.cancer~1,dist='lognorm')
 ajuste.ln
```

√ Saída:

```
>  ajuste.ln
Call:
survreg(formula = sobrevida.cancer ~ 1, dist = "lognorm")
Coefficients:
(Intercept)
   2.717176
Scale= 0.7648167
Loglik(model)= -65.7   Loglik(intercept only)= -65.7
n= 20
```

√ Estimação parâmetro:

```
>  mi.ln<- ajuste.ln$coefficients[1]
>  sigma.ln <- ajuste.ln$scale
>  cbind(mi.ln, sigma.ln)
               mi.ln  sigma.ln
(Intercept) 2.717176 0.7648167
```

√ Sobrevivência estimada: $\hat{S}_{ln}(t) = \Phi\left[\frac{-\ln(t)-2,7172}{0,7648}\right]$

- Estimativas da função de sobrevivência:
  - √ Modelo exponencial: $\hat{S}_e(t) = \exp\{-0,04899t\}$
  - √ Modelo de Weibull: $\hat{S}_w(t) = \exp\left\{-(0,04686t)^{1,5433}\right\}$
  - √ Modelo Log-normal: $\hat{S}_{ln}(t) = \Phi\left[\frac{-\ln(t)-2,7172}{0,7648}\right)$
- Valores estimados para $t = 10$:
  - √ Exponencial: $\hat{S}_e(t) = \exp\{-(0,04899)(10)\} = 0,612$
  - √ Weibull: $\hat{S}_w(t) = \exp\left\{-(0,04686 \times 10)^{1,5433}\right\} = 0,732$
  - √ Log-normal: $\hat{S}_{ln}(t) = \Phi\left[\frac{-\ln(10)-2,7172}{0,7648}\right) = 0,708$
  - √ São bem próximas as estimativas obtidas pelos modelos de Weibull e Log-normal

- Comparação das estimativas da sobrevivência:
  - √ Estimador Kaplan-Meier e modelos Exponencial, Weibull e log-normal
    - – Comandos em R:

```
> # Estimativa de Kaplan-Meier
>
>  e.km<-survfit(sobrevida.cancer~1)
>  tempo.km<-e.km$time
>  s.km<-e.km$surv
>
> # Comparação de Estimativas
>
>  s.e<- exp(-alfa.exp*tempo.km)
>  s.w<- exp(-(alfa.weib*tempo.km)^gama.weib)
>  s.ln<- pnorm((-log(tempo.km)+ mi.ln)/sigma.ln)
>  cbind(tempo.km,s.km,s.e,s.w,s.ln)
```

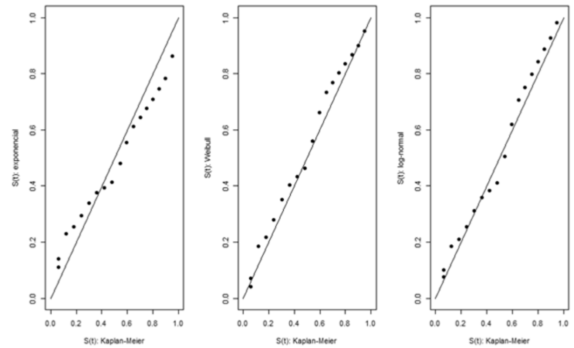

- Estimativas da sobrevivência para os tempos de reincidência:

	tempo.km	s.km	s.e	s.w	s.ln
[1,]	3	0.95000000	0.8633164	0.95274148	0.98283934
[2,]	5	0.90000000	0.7827384	0.89897484	0.92624322
[3,]	6	0.85000000	0.7453152	0.86839357	0.88685752
[4,]	7	0.80000000	0.7096812	0.83609525	0.84337638
[5,]	8	0.75000000	0.6757509	0.80253272	0.79781416
[6,]	9	0.70000000	0.6434428	0.76809812	0.75169629
[7,]	10	0.65000000	0.6126794	0.73313414	0.70611769
[8,]	12	0.59583333	0.5554947	0.66278292	0.61931883
[9,]	15	0.54166667	0.4795676	0.55966698	0.50475984
[10,]	18	0.48148148	0.4140186	0.46346069	0.41042396
[11,]	19	0.42129630	0.3942241	0.43345774	0.38317692
[12,]	20	0.36111111	0.3753760	0.40461041	0.35784922
[13,]	22	0.30092593	0.3403401	0.35056108	0.31248031
[14,]	25	0.24074074	0.2938212	0.27891623	0.25592025
[15,]	28	0.18055556	0.2536607	0.21851514	0.21065456
[16,]	30	0.12037037	0.2299851	0.18419003	0.18556452
[17,]	40	0.06018519	0.1409071	0.07154720	0.10195229
[18,]	45	0.06018519	0.1102934	0.04229094	0.07714984

- Escolha do modelo – Método Gráfico:
  - √ Gráfico das estimativas da sobrevivência obtidas pelo estimador de Kaplan-Meier vs. Estimativas das sobrevivências a partir dos modelos paramétricos
  - √ Comandos:

```
par(mfrow=c(1,3))

          # Exponencial

plot(s.km,s.e,pch=16,ylim=range(c(0.0,1)), xlim=range(c(0,1)),
         xlab = "S(t): Kaplan-Meier", ylab="S(t): exponencial")
lines(c(0,1), c(0,1), type="l", lty=1)

          # Weibull

plot(s.km,s.w,pch=16,ylim=range(c(0.0,1)), xlim=range(c(0,1)),
            xlab = "S(t): Kaplan-Meier", ylab="S(t): Weibull")
lines(c(0,1), c(0,1), type="l", lty=1)

          # Log-normal

plot(s.km,s.ln,pch=16,ylim=range(c(0.0,1)), xlim=range(c(0,1)),
          xlab = "S(t): Kaplan-Meier", ylab="S(t): log-normal")
lines(c(0,1), c(0,1), type="l", lty=1)lty=1)
```


- Estimativa Kaplan-Meier vs. Modelos

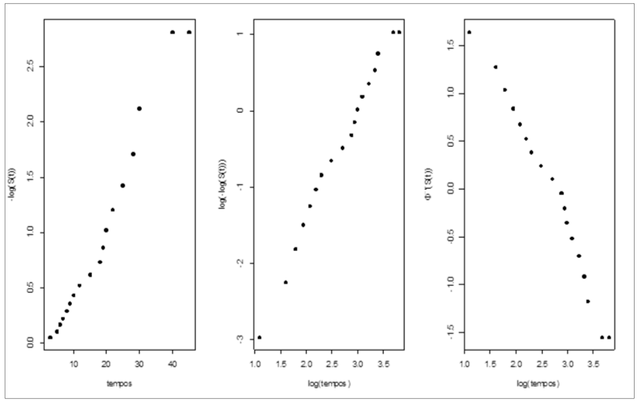


- Comentários:
  - √ Modelo exponencial parece não ser adequado aos dados:
    (Curva encontra-se um tanto afastada da reta $y = x$)
  - √ Indicação de que os modelos de Weibull ou log-normal é possivelmente adequado aos dados
    (acompanham mais de perto a reta $y = x$)

- Gráficos linearizados para os modelos:
  - √ Comandos:

```
inv.km<-qnorm(s.km)
 plot(tempo.km, -log(s.km),pch=16,xlab="tempos",ylab="-log(S(t))")
 plot(log(tempo.km),log(-log(s.km)),pch=16,xlab="log(tempos)",
          ylab="log(-log(S(t)))")
 plot(log(tempo.km),inv.km, pch=16,xlab="log(tempos)",
ylab=expression(Phi^-1 * (S(t))))
```


- Gráficos linearizados para os modelos Exponencial, Weibull e Log-normal



- Comentários:
  - √ Modelos de Weibull e Log-normal não mostram afastamentos marcantes de uma reta
  - √ Observa-se certo desvio para modelo Exponencial
  - √ Gráficos confirmam resultados observados anteriormente
  - √ Tamanho amostral pequeno provavelmente é principal razão de não ter ocorrido discriminação entre os modelos de Weibull e Log-normal

- Teste da razão de verossimilhança:
  - √ H_0: modelo de interesse é adequado
  - √ Modelo generalizado (modelos de interesse são casos particulares: gama generalizada (3 parâmetros)
  - √ Ajuste do modelo gama generalizado

```
>  require(flexsurv)
> ajuste.gama<-flexsurvreg(sobrevida.cancer ~1,dist='gengamma')
> ajuste.gama
Call:
flexsurvreg(formula = sobrevida.cancer ~ 1, dist = "gengamma.orig")
Estimates:
       est       L95%      U95%      se
shape  4.21e-01  2.97e-03  5.97e+01  1.06e+00
scale  6.46e-02  7.20e-24  5.80e+20  1.67e+00
k      1.04e+01  8.06e-04  1.34e+05  5.03e+01
N = 20,  Events: 17,  Censored: 3
Total time at risk: 347
Log-likelihood = -65.69393, df = 3
AIC = 137.3879
```


- Logaritmos da verossimilhança:

```
> ajuste.gama$loglik
[1] -65.69074
>  ajuste.e$loglik[2]
[1] -68.27389
>  ajuste.w$loglik[2]
[1] -66.13336
>  ajuste.ln$loglik[2]
[1] -65.7399
```

- Teste da Razão de Verossimilhanças–Resultados

	ln(θ)	**Razão**	**GL**	**p-valor**
Gama generalizada	–65,69			
Exponencial	–68,27	2(68,27–65,69)=5,16	2	0,075
Weibull	–66,13	2(66,13–65,69)=0,88	1	0,348
Log-normal	–65,74	2(65,74–65,69)=0,10	1	0,752

```
> 1-pchisq(5.16,2)
[1] 0.075774
```

  - √ Resultados indicam adequação dos modelos de Weibull e Log-normal

- Critério de Informação de Akaike:

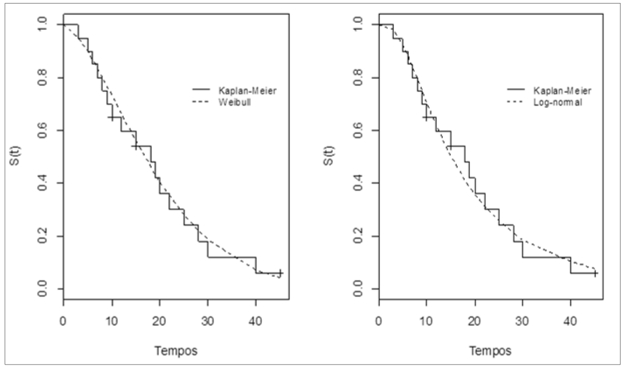
Modelo	**ln(θ)**	**AIC**
Gama generalizada	–65,69	2[3 – (–65,69)] = 137,4
Exponencial	–68,27	2[1 – (–68,27)] = 138,5
Weibull	–66,13	2[2 – (–66,13)] = 136,3
Log-normal	–65,74	2[2 – (–65,74)] = 135,5

  - √ Confirmam o descarte do modelo exponencial

- Curvas de sobrevivência estimadas:
  - – Modelos Weibull e log-normal
  - – Estimador de Kaplan-Meier
  - √ Comandos:

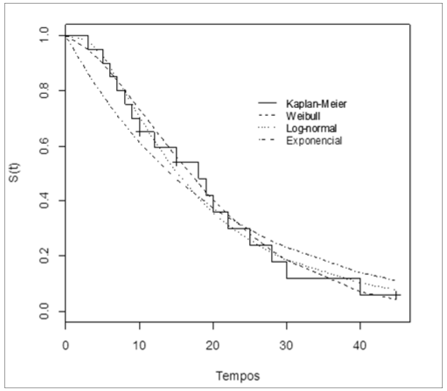
```
          # Curvas de Sobrevivência Estimadas

par(mfrow=c(1,2))
plot(e.km, conf.int=F, xlab="Tempos", ylab="S(t)")
lines(c(0,tempo.km),c(1,s.w), lty=2)
legend(25,0.8,lty=c(1,2),c("Kaplan-Meier", "Weibull"),bty="n",cex=0.8)
plot(e.km, conf.int=F, xlab="Tempos", ylab="S(t)")
lines(c(0,tempo.km),c(1,s.ln), lty=2)
legend(25,0.8,lty=c(1,2),c("Kaplan-Meier", "Log-normal"),bty="n",cex=0.8)
```

- Curvas de sobrevivência estimadas:



√ Ambos os modelos apresentam ajustes satisfatórios

- Comparação das curvas estimadas de sobrevivência



- Estimativa para o tempo médio:
  - √ Modelo Log-normal:

$$\hat{E}(T) = \exp\left\{\mu + \frac{\sigma^2}{2}\right\}$$
$$= \exp\left\{2,72 + \frac{0,76^2}{2}\right\}$$
$$= 20,263 \text{ meses}$$

  - √ Modelo de Weibull

$$\hat{E}(T) = \frac{1}{\alpha}\Gamma\left(1 + \frac{1}{\gamma}\right)$$
$$= \frac{1}{0,0468}\Gamma\left(1 + \frac{1}{1,54}\right)$$
$$= 19,206 \text{ meses}$$

- Estimativa de Var[$\hat{\text{E}}$(T)]:
  - √ Modelo Log-normal:

```
> ajuste.ln$var
            (Intercept)  Log(scale)
(Intercept) 0.031061677 0.002706896
Log(scale)  0.002706896 0.030119031
```

$$\widehat{\text{Var}}(\hat{\mu}) = 0,0311$$
$$\widehat{\text{Var}}(\hat{\sigma}) = \widehat{\text{Var}}(\ln\hat{\sigma})(\hat{\sigma})^2$$
$$= (0,0311)(0,7648)^2$$
$$= 0,0176$$
$$\widehat{\text{Cov}}(\hat{\mu},\hat{\sigma}) = 0,0027$$

  - √ Estimativa :

$$\widehat{\text{Var}}[\hat{E}(T)] = \widehat{\text{Var}}(\hat{\mu})\left[\exp\left\{\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right\}\right]^2 + \widehat{\text{Var}}(\hat{\sigma})\left[\hat{\sigma}\exp\left\{\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right\}\right]^2$$
$$+ 2\widehat{\text{Cov}}(\hat{\mu},\hat{\sigma})\left[\exp\left\{\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right\}\right]\left[\hat{\sigma}\exp\left\{\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right\}\right]$$
$$= (0,031)(20,263)^2 + (0,0176)[(0,76)(20,263)]^2$$
$$+ 2(0,00207)(0,76)(20,263)^2 = 18,2$$

- Intervalo de 95% de confiança para E[T]:

$$\hat{E}[T] \pm z_{\alpha/2}\sqrt{\widehat{\text{Var}}[\hat{E}(T)]} = 20,263 \pm 1,96\sqrt{18,2}$$

$$[11,901; 28,625]$$

- Estimativa de Var[$\hat{E}$(T)]:
  - √ Modelo de Weibull:
    - Nesse caso, a estimativa é mais complicada pois aparece a derivada da função gama
    - Envolve a função digama

- Estimativa para o tempo mediano:
  - √ Modelo Log-normal: $\hat{t}_{0,5} = \exp\{\hat{\mu}\} = \exp\{2,72\}$ $= 15,18$ meses
  - √ Variância do estimador do tempo mediano

$$\text{Var}(\hat{t}_{0,5}) = \text{Var}(\mu)\,[\exp\{\hat{\mu}\}]^2$$
$$\widehat{\text{ep}}(\hat{t}_{0,5}) = \sqrt{0,031}\exp\{2,72\}$$
$$= 2,675 \text{ meses}$$

  - √ Intervalo com 95% de confiança para o tempo mediano

$$\hat{t}_{0,5} \pm z_{\alpha/2}\widehat{\text{ep}}(\hat{t}_{0,5}) = 15,18 \pm (1,96)(2,675)$$

$$[9,94; 20,42]$$

- Estimativa para o tempo mediano:
  - √ Modelo de Weibull: $\hat{t}_{0,5} = \dfrac{[\ln(2)]^{\frac{1}{\gamma}}}{\alpha} = \dfrac{[\ln(2)]^{\frac{1}{1,54}}}{0,0468}$ $= 16,84$ meses
  - √ Interpolação linear da função de sobrevivência

tempo.km		s.km
[1,]	3	0.95000000
[2,]	5	0.90000000
[3,]	6	0.85000000
[4,]	7	0.80000000
[5,]	8	0.75000000
[6,]	9	0.70000000
[7,]	10	0.65000000
[8,]	12	0.59583333
[9,]	15	0.54166667
[10,]	18	0.48148148
[11,]	19	0.42129630

$$\hat{t}_{0,5} = 17,05 \text{ meses}$$

- Estimativa para S(20):
  - √ Modelo Log-normal:

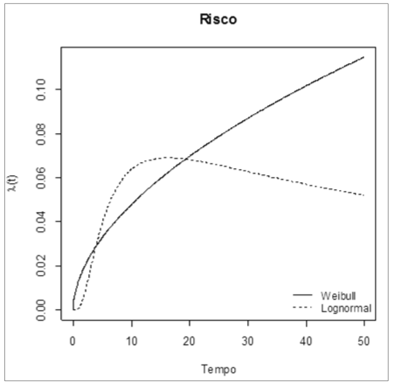
$$\hat{S}_{ln}(20) = \Phi\left(\frac{-\ln(t)+\hat{\mu}}{\hat{\sigma}}\right)$$
$$= \Phi\left(\frac{-\ln(20)+2,72}{0,76}\right)$$
$$= 0,3583$$

  - √ Estimador Kaplan-Meier: $\hat{S}_{KM}(20) = 0,361$
  - √ Paciente tem uma probabilidade de cerca de 36% de estar livre de reincidência 20 meses após da realização do procedimento cirúrgico

Comentário

- Utilizado o modelo lognormal para ilustrar o cálculo das estimativas intervalares de E(T) e t_{med}.

- Funções de risco estimadas dos modelos



  - √ Qual função de risco expressa melhor o problema?

Modelos de Regressão Paramétricos

Covariáveis

- Variáveis que podem estar relacionadas com o tempo de sobrevivência
  - √ Ex.: Tempo até a ocorrência de Aids em pacientes infectados pelo HIV
    - Contagem de células CD4 e CD8
    - Fatores de prognóstico importantes

Técnicas Não Paramétricas

- Não envolvem nenhuma estrutura paramétrica
- Vantagens:
  - √ Simplicidade e facilidade de aplicação
- Limitação:
  - √ Análise mais detalhada é inviável
    - Não permitem a inclusão direta de covariáveis
  - √ Importância:
    - Descrição dos dados de sobrevivência

Estratificação

- Forma simples de fazer análise mais elaborada incluindo covariáveis
  - √ Dividir dados em estratos de acordo com covariáveis
    - Usar as técnicas não paramétricas apropriadas
  - √ Vantagem:
    - Facilidade de cálculo e facilidade de entendimento
  - √ Limitação:
    - Difícil análise com muitas covariáveis (gera um número muito grande de estratos)

Modelos de Regressão para Dados de Sobrevivência

  - √ Forma eficiente de acomodar o efeito de covariáveis
- Classes de modelos:
  - √ Modelos paramétricos:
    (modelos de tempo de vida acelerado)
    - Mais eficientes porém menos flexíveis
  - √ Modelos semiparamétricos:
    (modelo de regressão de Cox)
    - São mais flexíveis
    - Fácil a incorporação de covariáveis dependentes do tempo

Modelo de Regressão Linear

√ Modelo mais conhecido em Estatística

√ Resposta é associada com as variáveis explicativas (covariáveis) por meio de modelo linear

$$Y = \overbrace{\beta_0 + \beta_1 x}^{\text{Componente determinístico}} + \overbrace{\epsilon}^{\text{Componente estocástico}}$$

- Y: variável resposta
- x: covariável (variável explicativa, variável independente)
- β_0, β_1: parâmetros a serem estimados
- ε: erro aleatório (com distribuição normal)

- Estrutura de um modelo de regressão:
  - √ Componente aleatório:
    - Descreve probabilisticamente o comportamento da resposta
  - √ Componente determinístico ou estrutural:
    - Descreve a relação entre os parâmetros da distribuição e as covariáveis

Modelo de Regressão de Sobrevivência

- Estima efeito das covariáveis sobre o tempo de sobrevivência
  - √ Variável resposta: tempo até desfecho e censura
    - Em geral, o tipo de resposta e o comportamento da variáveis não permitem a utilização direta do modelo linear
  - √ Distribuição da resposta tende a ser assimétrica na direção dos maiores tempos de sobrevivência
    - Inapropriado o uso da distribuição normal para o comportamento estocástico

- Formas de modelagem estatística em análise de sobrevivência:
  - √ Transformar a resposta para tentar retornar ao modelo linear, ou
  - √ Modificar o modelo, usando
    - Componente determinístico não linear nos parâmetros
    - Componente estocástico com distribuição assimétrica
  - √ As duas formas são equivalentes:
    - Usar como componente determinístico $\exp\{\beta_0 + \beta_1 x\}$ (transformação logarítmica da resposta)
    - Distribuição lognormal para os erros (há outras distribuições assimétricas disponíveis)

Modelo de Regressão do Risco

- Função de risco de um indivíduo no tempo t, dado o vetor de covariáveis fixas ***x***:

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\, g(\boldsymbol{x}'\boldsymbol{\beta})$$

√ **β**: vetor de coeficientes a estimar

√ g(.): função positiva e contínua das covariáveis

√ $\lambda_0(t)$: risco basal de um indivíduo que possui g(**x'β**)=1

√ Efeito das covariáveis é aumentar ou diminuir o risco por uma quantidade proporcional em todas as durações

- Razão de risco de dois indivíduos:

$$\frac{\lambda_i(t|x_{i1})}{\lambda_j(t|x_{j1})} = \frac{\lambda_0(t)\, g(\beta_1 x_1)}{\lambda_0(t)\, g(\beta_1 x_1^*)} = \frac{g(\beta_1 x_1)}{g(\beta_1 x_1^*)}$$

√ Função das covariáveis e não depende do tempo

√ Diferentes indivíduos com covariáveis x_1 e x_1^* possuem riscos proporcionais

- Ajuste de modelo de regressão paramétrico

√ Necessário supor uma distribuição de probabilidade para o tempo de sobrevivência

√ Parâmetros da distribuição adotada definem forma dessa distribuição

– Não confundir com os parâmetros da regressão

- Modelo de riscos proporcionais

√ Efeito multiplicativo exponencial

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\, \exp\{\boldsymbol{x}'\boldsymbol{\beta}\}$$
$$= \lambda_0(t)\, e^{\beta_0}\, e^{\beta_1 x_1} \ldots e^{\beta_p x_p}$$

– Efeito das covariáveis é aumentar ou diminuir o risco por uma quantidade proporcional em todas as durações

√ O risco acumulado segue a mesma relação:

$$\Lambda(t|\mathbf{x}) = \Lambda_0(t)\, g(\boldsymbol{x}'\boldsymbol{\beta})$$

√ A relação da sobrevivência é:

$$\exp\{-\Lambda(t|\boldsymbol{x})\} = \exp\{-\Lambda_0(t)\, g(\boldsymbol{x}'\boldsymbol{\beta})\} = \exp\{-\Lambda_0(t)\}^{g(\boldsymbol{x}'\boldsymbol{\beta})}$$
$$S(t|\boldsymbol{x}) = [S_0(t)]^{g(\boldsymbol{x}'\boldsymbol{\beta})}$$

– Sobrevivência para as covariáveis ***x*** é a sobrevivência de referência elevada a uma potência

– Ex: se o indivíduo está exposto ao dobro do risco de um indivíduo de referência, em um determinado valor de covariáveis, então a probabilidade de que o indivíduo estará vivo em qualquer tempo é o quadrado da probabilidade de o indivíduo de referência estar vivo na mesma idade

Estimação dos Parâmetros do Modelo

- Estimação e inferência dos efeitos das covariáveis (β's da regressão)
- Método dos mínimos quadrados
  - √ Tem propriedades desejáveis na presença de erros com distribuição normal
    - – Inadequado na ausência de normalidade e na presença de censuras
- Método da máxima verossimilhança
  - √ Adequado na presença de censura e ausência de normalidade

Modelo de Regressão Exponencial

√ Usado quando se assume que o risco é constante

√ Parâmetro α depende das covariáveis

$$\alpha(\boldsymbol{x}) = \exp\{\boldsymbol{x}'\boldsymbol{\beta}\}$$

- – β: estimativa dos efeitos das covariáveis
- – α: parâmetro que define o risco exponencial

√ Modela o risco como uma função das covariáveis

- – Idade aumenta o risco de óbito por insuficiência renal
- – Categoria "usuário de drogas injetáveis" tem risco maior do que a categoria "transmissão sexual" para o tempo de sobrevida com Aids

- Funções para o modelo de regressão exponencial
  - √ Função de risco

$$\lambda(t|\boldsymbol{x}) = \alpha(\boldsymbol{x}) = \exp\{\boldsymbol{x}'\boldsymbol{\beta}\}$$

  - √ Função de sobrevivência:

$$S(t|\boldsymbol{x}) = \exp\{-\alpha(\boldsymbol{x})t\} = \exp\{-\exp\{\boldsymbol{x}'\boldsymbol{\beta}\}t\}$$

Exemplo

- Pacientes com insuficiência renal submetidos à hemodiálise
  - √ Período: janeiro/1988 a outubro/2001
  - √ Dados: Sistema Apac
  - √ Coorte com 6.805 pacientes, no Rio de Janeiro
    - – 1.603 óbitos
  - √ Objetivo:
    - – Estimar o efeito da idade no risco de morte

- Hipótese:
  - √ Risco de morrer é constante ao longo do tempo (modelo exponencial)
- Modelo: $\lambda(t|idade) = \exp\{\beta_0 + \beta_1 \times idade\}$
- Ajuste do modelo

```
> library(survival)
>
> dialise <- read.csv("dados/dialise.csv", header = T)
> reg.exp <- survreg(Surv(tempo, status)~idade, data = dialise,
+  dist='exponential')
> reg.exp$coefficients
(Intercept)       idade
 6.13552416 -0.03700427
```

  – Importante: sinal do parâmetro está invertido em relação ao utilizado no texto

$$\boxed{\begin{array}{l}\hat{\beta}_0 = -6,136 \\ \hat{\beta}_1 = 0,037\end{array}}$$

- Estimativa do risco no tempo t:

$$\hat{\lambda}(t|idade) = \exp\{-6,136 + 0,037 \times idade\}$$

√ Para cada ano de vida o risco aumenta

exp{0,037} = 1,037 vezes

√ Comparação entre indivíduos com 30 e 70 anos,

$$\frac{\hat{\lambda}(t|idade = 70)}{\hat{\lambda}(t|idade = 30)} = \frac{\exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \times 70\right\}}{\exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \times 30\right\}}$$

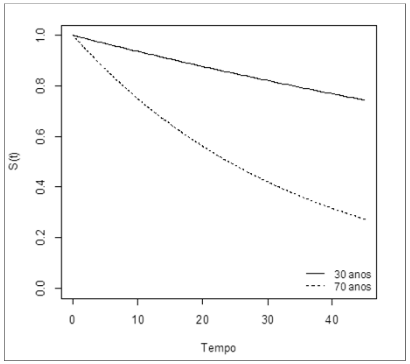
$$= \exp\{0,037 \times 40\} = 4,4$$

– Risco de indivíduo em diálise morrer aos 70 anos é 4,4 vezes maior do que o de morrer aos 30

– Quaisquer dois indivíduos com diferença de 40 anos entre si gera um risco relativo estimado em 4,4

(modelo de riscos proporcionais)

- Curva de sobrevivência ajustada para 30 e 70 anos



√ Indivíduo com 70 anos progride 4,4 vezes mais rápido ao longo do tempo, acelerando a uma taxa constante

Modelo de Regressão de Weibull

√ T ~ Weibull e o parâmetro α é modelado pelas covariáveis

$$\alpha(\boldsymbol{x}) = \exp\{\boldsymbol{x}'\boldsymbol{\beta}\}$$

– **B**’: estimativa dos efeitos das covariáveis

– **x**’: valores das covariáveis

√ Modela o risco e sobrevivência como funções das covariáveis

- Funções para o modelo de regressão exponencial
  - √ Função de risco $\lambda(t|\boldsymbol{x}) = \gamma t^{\gamma-1}\alpha(\boldsymbol{x})^{\gamma}$
    $= \gamma t^{\gamma-1}[\exp\{(\boldsymbol{x}'\boldsymbol{\beta})\}]^{\gamma}$
  - √ Função de sobrevivência:
    $$S(t|\boldsymbol{x}) = \exp\{-[\alpha(\boldsymbol{x})\, t]^{\gamma}\} = \exp\{-[\exp\{\boldsymbol{x}'\boldsymbol{\beta}\}\, t]^{\gamma}\}$$
  - √ Caso α = g(***x***) (modelo em questão)
    - Os riscos para diferentes indivíduos são proporcionais
  - √ Caso γ também varie com as covariáveis [γ= h(***x***)]
    - Modelo de regressão deixa de ser de riscos proporcionais

- Parâmetro γ:
  - √ γ > 1: o risco aumenta no tempo
  - √ γ < 1: o risco diminui no tempo
  - √ γ = 1: a função de risco é constante (exponencial)

Exemplo

- Pacientes com insuficiência renal submetidos à hemodiálise
  - √ Período: janeiro/1988 a outubro/2001
  - √ Dados: Sistema Apac
  - √ Coorte com 6.805 pacientes, no Rio de Janeiro
    - 1.603 óbitos
  - √ Objetivo:
    - Estimar o efeito da idade no risco de morte

- Hipótese:
  - √ Tempo até óbito tem distribuição de Weibull
- Modelo: $\lambda(t|idade) = \gamma t^{\gamma-1}[\exp\{\beta_0 + \beta_1 \times idade\}]^{\gamma}$
- Ajuste do modelo

```
> dialise <- read.csv("dados/dialise.csv", header = T)
> reg.weib <- survreg(Surv(tempo, status)~idade, data = dialise,
+  dist='weibull')
> summary(reg.weib)
Call:
survreg(formula = Surv(tempo, status) ~ idade, data = dialise,
    dist = "weibull")
              Value Std. Error      z        p
(Intercept)  6.7512    0.14693  45.95 0.00e+00
idade       -0.0436    0.00224 -19.50 1.12e-84
Log(scale)   0.1987    0.02083   9.54 1.49e-21
Scale= 1.22
Weibull distribution
Loglik(model)= -7877.7   Loglik(intercept only)= -8104.2
        Chisq= 453.16 on 1 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 7
n= 6805
```

$\hat{\beta}_0 = -6,751$

$\hat{\beta}_1 = 0,0436$

$\hat{\sigma} = \exp\{0,1987\} = 1,22$

- Importante: inverter o sinal do parâmetro

- Análise resultados

```
> summary(reg.weib)
Call:
survreg(formula = Surv(tempo, status) ~ idade, data = dialise,
    dist = "weibull")
              Value Std. Error      z        p
(Intercept)  6.7512    0.14693  45.95 0.00e+00
idade       -0.0436    0.00224 -19.50 1.12e-84
Log(scale)   0.1987    0.02083   9.54 1.49e-21
Scale= 1.22
Weibull distribution
Loglik(model)= -7877.7   Loglik(intercept only)= -8104.2
        Chisq= 453.16 on 1 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 7
n= 6805
```

$$\hat\beta_0 = -6,751 \qquad \hat\beta_1 = 0,0436 \qquad \hat\sigma = \exp\{0,1987\} = 1,22$$

√ Teste de significância do parâmetro

- Efeito da idade (β_1) é significativo
- Parâmetro de forma (γ) é significativamente diferente de 1

√ Parâmetros da distribuição do tempo:

$$\hat\alpha = \exp\{-\boldsymbol{x}'\boldsymbol{\beta}\} \qquad \hat\gamma = \frac{1}{\hat\sigma} = 0,820$$

√ Parâmetros da regressão:

$$\hat\beta_0 = -6,751 \qquad \hat\beta_1 = 0,0436 \qquad \hat\sigma = \exp\{0,1987\} = 1,22$$

√ Parâmetros da distribuição de T:

$$\hat\alpha = \exp\{-\boldsymbol{x}'\boldsymbol{\beta}\} \qquad \hat\gamma = \frac{1}{\hat\sigma} = 0,820$$

- Estimativa do risco no tempo t:

$$\hat\lambda(t|idade) = (0,82)t^{-0,18}[\exp\{-6,751 + 0,044 \times idade\}]^{0,82}$$

√ Para cada ano de vida o risco aumenta

$(\exp\{0{,}0436\})^{0,82} = 1{,}036$ vezes

√ Parâmetro de forma ($\gamma = 0{,}82$)

- O risco de morrer diminui à medida que aumenta o tempo do paciente em hemodiálise ($\gamma < 1$)
- Possível motivo:
  - Baixo preparo inicial dos pacientes (hemodiálise em emergência) (excesso de óbitos no início do processo)

√ Comparação entre indivíduos com 30 e 70 anos,

$$\frac{\hat\lambda(t|idade=70)}{\hat\lambda(t|idade=30)} = \frac{\gamma t^{\gamma-1}[\exp\left\{\hat\beta_0 + \hat\beta_1 \times 70\right\}]^\gamma}{\gamma t^{\gamma-1}[\exp\left\{\hat\beta_0 + \hat\beta_1 \times 30\right\}]^\gamma}$$

$$= \left[\frac{\exp\{-6,7512 + 0,0436 \times 70\}}{\exp\{-6,7512 + 0,0436 \times 30\}}\right]^{0,82}$$

$$= [\exp\{0,0436 \times 40\}]^{0,82} = 4,11$$

- Risco de indivíduo em diálise morrer aos 70 anos é 4,11 vezes maior do que o de morrer aos 30
- Quaisquer dois indivíduos com diferença de 40 anos entre si gera um risco relativo estimado em 4,11

  (modelo de riscos proporcionais)

Estratificação

- Podem-se utilizar estratos na modelagem

√ Diferentes parâmetros γ da distribuição do tempo dos subgrupos populacionais

√ Parâmetros da regressão comuns para todo o conjunto

- Mesmo efeito das covariáveis

√ Permite-se a variação da curva de risco basal entre os estratos

- Melhora na qualidade da estimativa do efeito das covariáveis

√ Caso indicado:

- Pode-se estimar o efeito das covariáveis para cada estrato

- Causa da insuficiência renal: doença congênita
  - √ Padrão de sobrevivência pode ser muito diferente (pacientes mais jovens são mais transplantados)
- Análise dessa hipótese
  - √ Utilizar doença congênita no modelo
    - – Aditivamente, como estrato e com interação entre o estrato e a idade

```
> congenita.simples <- survreg(Surv(tempo, status)~idade + congenita,
+ data = dialise, dist='weibull')
> congenita.strat <- survreg(Surv(tempo, status)~idade + strata(congenita),
+     data = dialise, dist='weibull')
> congenita.inter <- survreg(Surv(tempo, status)~idade*strata(congenita),
+ data = dialise, dist='weibull')
```


- Doença congênita incluída aditivamente

```
> summary(congenita.simples)
Call:
survreg(formula = Surv(tempo, status) ~ idade + congenita, data = dialise,
    dist = "weibull")
              Value Std. Error      z        p
(Intercept)  6.7357    0.14647  45.99 0.00e+00
idade       -0.0437    0.00223 -19.58 2.20e-85
congenita    0.9781    0.27457   3.56 3.67e-04
Log(scale)   0.1971    0.02082   9.47 2.82e-21
Scale= 1.22
Weibull distribution
Loglik(model)= -7869.3   Loglik(intercept only)= -8104.2
        Chisq= 469.95 on 2 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 7
n= 6805
```

Coeficientes significativos

$$\hat{\beta}_0 = -6,7357$$
$$\hat{\beta}_{\text{idade}} = 0,0437$$
$$\hat{\beta}_{\text{cong}} = -0,9781$$
$$\hat{\sigma}_{\text{global}} = 1,22$$

$$\hat{\gamma}_{\text{global}} = \frac{1}{\hat{\sigma}} = 0,820$$

Modelo Simples – Comentários

- Adaptação para compatibilização entre saída do R e o texto
  - √ Parâmetro α: $\hat{\alpha} = \exp\{-\boldsymbol{x}'\boldsymbol{\beta}\}$
  - √ Parâmetro γ: $\hat{\gamma} = \frac{1}{\hat{\sigma}} = \frac{1}{\text{scale}}$
- Erro padrão e teste de significância (γ=1)

$$\ln(\text{scale}) = \ln\left(\frac{1}{\hat{\gamma}}\right)$$

- O risco de morrer diminui à medida que aumenta o tempo do paciente em hemodiálise (γ < 1)
  - √ Sem distinção entre os dois grupos

- Estratificação por doença congênita

```
> summary(cong.strat)
Call:
survreg(formula = Surv(tempo, status) ~ idade + strata(congenita),
    data = dialise, dist = "weibull")
               Value Std. Error      z        p
(Intercept)   6.7385    0.14645  46.01 0.00e+00
idade        -0.0434    0.00223 -19.50 1.17e-84
congenita=0   0.2047    0.02087   9.81 1.02e-22
congenita=1  -0.2716    0.11620  -2.34 1.94e-02

Scale:
congenita=0 congenita=1
      1.227       0.762

Weibull distribution
Loglik(model)= -7870.3   Loglik(intercept only)= -8097
        Chisq= 453.49 on 1 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 7
n= 6805
```

Log scale

Coeficientes significativos

$$\hat{\beta}_0 = -6,7385$$
$$\hat{\beta}_{\text{idade}} = 0,0434$$

$$\hat{\gamma}_{\text{cong}=0} = \frac{1}{1,227} = 0,814$$
$$\hat{\gamma}_{\text{cong}=1} = \frac{1}{0,762} = 1,312$$

Modelo Estratificado – Comentários

- Efeito da doença congênita não é estimado quando a variável é usada para estratificação
- Há parâmetros de escala diferentes para cada categoria
  - √ Congênita = 0
    - O risco de morrer diminui com o tempo de hemodiálise
  - √ Congênita = 1
    - O risco de morrer aumenta com o tempo de hemodiálise

- Interação entre doença congênita e idade

```
> summary(congenita.inter)
Call:
survreg(formula = Surv(tempo, status) ~ idade * strata(congenita),
    data = dialise, dist = "weibull")
                                    Value Std. Error       z        p
(Intercept)                       6.74462    0.14648  46.046 0.00e+00
idade                            -0.04375    0.00223 -19.602 1.48e-85
idade:strata(congenita)congenita=1  0.00735    0.00446   1.649 9.92e-02
congenita=0                       0.20041    0.02091   9.583 9.45e-22
congenita=1                      -0.11184    0.15546  -0.719 4.72e-01

Scale:
congenita=0 congenita=1
      1.222       0.894

Weibull distribution
Loglik(model)= -7868.1   Loglik(intercept only)= -8097
        Chisq= 457.81 on 2 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 17
n= 6805
```

Interação não significativa

Log scale

$$\hat{\beta}_0 = -6,7446$$
$$\hat{\beta}_{\text{idade}} = 0,04375$$
$$\hat{\beta}_{(\text{idade})\times(\text{congenita})} = 0,040735 = 0$$

$$\hat{\gamma}_{\text{cong}=0} = \frac{1}{1,222} = 0,818$$
$$\hat{\gamma}_{\text{cong}=1} = \frac{1}{0,894} = 1,118$$

Modelo Estratificado – Comentários

- Efeito da interação entre idade e a variável doença congênita não é significativa
- Há parâmetros de escala diferentes para cada categoria
  - √ Congênita = 0
    - O risco de morrer diminui com o tempo de hemodiálise
  - √ Congênita = 1
    - O risco de morrer aumenta com o tempo de hemodiálise

- Resumos dos modelos – Saídas do R

	Modelos		
	cong.simples	cong.strat	cong.inter
Coeficientes (β)			
Intercept	6.7357	6.7385	6.74462
idade	−0.0437	−0.0434	−0.04375
congenita	0.9781		
idade:strata(congenita)=1			0.00735
Scale			
Global	1,22		
congenita=0		1.227	1.222
Congenita=1		0.762	0.894

- Estimação dos parâmetros de forma

$$\hat{\gamma} = \frac{1}{\hat{\sigma}} = \frac{1}{\text{scale}}$$

```
> # Parametros de Forma
> # Modelo simples
> 1/cong.simples$scale
[1] 0.8210882
> # Modelo estratificado
> 1/cong.strat$scale
congenita=0 congenita=1
  0.8148844   1.3121269
> # Modelo estratificado c/ interação
> 1/cong.inter$scale
congenita=0 congenita=1
  0.8183929   1.1183291
```


- Estimação dos riscos relativos

```
> # Risco Relativo
> # Modelo simples
> exp(-cong.simples$coeff[-1])^(1/cong.simples$scale)
    idade congenita
1.0365145 0.4479174
> # Modelo estratificado para congenita=0
> exp(-cong.strat$coeff[-1])^(1/cong.strat$scale[1])
   idade
1.035988
> # Modelo estratificado para congenita=1
> exp(-cong.strat$coeff[-1])^(1/cong.strat$scale[2])
   idade
1.058581
> # Modelo estratificado com interação
> # congenita = 0
> exp(-cong.inter$coeff[2])^(1/cong.inter$scale[1])
   idade
1.036457
> # Modelo estratificado com interação
> # congenita = 1
> exp(-sum(cong.inter$coeff[2:3]))^(1/cong.inter$scale[2])
congenita=1
   1.041548
```

$[\exp\{\beta\}]^{\gamma}$

- Estimativas do risco relativo ($(\exp(\beta)^{\gamma})$)

	Modelos		
Variáveis	Simples	Estratificado	Interação
Idade	1,036		
Congênita = Não		1,034	1,036
Congênita = Sim		1,058	1,041
Congênita	0,448		
Forma (g)	0,821		
Congênita = Não		0,815	0,818
Congênita = Sim		1,312	1,118

Comentários

- Parâmetro de forma (γ)
  - √ É bastante diferente para o estrato de pacientes cuja insuficiência renal foi causada por doença congênita
  - √ $\gamma_{global} \approx \gamma_{congenita=0}$
    - Somente 2% dos pacientes possuem doença congênita
- Aumento do risco de morrer para cada ano de vida
  - √ Maior no estrato dos pacientes com doença congênita
  - √ Mantém-se próximos em todos os modelos para o estrato dos pacientes sem doença congênita

- Equações de risco dos modelos estimados
  - √ Modelo simples

$$\hat{\lambda}(t|\boldsymbol{x}) = (0,821)t^{0,821-1}[\exp\{-6,735 + 0,044 \times idade - 0,978 \times congenita\}]^{0,821}$$

  - √ Modelo estratificado
    - Congênita = Não

$$\hat{\lambda}(t|\boldsymbol{x}) = (0,815)t^{0,815-1}[\exp\{-6,739 + 0,0433 \times idade\}]^{0,815}$$

    - Congênita = Sim

$$\hat{\lambda}(t|\boldsymbol{x}) = (1,312)t^{1,312-1}[\exp\{-6,739 + 0,0433 \times idade\}]^{1,312}$$

√ Modelo estratificado com interação
- Congênita = Não

$$\hat{\lambda}(t|\boldsymbol{x}) = (0,818)t^{0,818-1}[\exp\{-6,745 + 0,044 \times idade\}]^{0,818}$$

- Congênita = Sim

$$\hat{\lambda}(t|\boldsymbol{x}) = (1,118)t^{1,118-1}[\exp\{-6,745 + (0,044 - 0,007) \times idade\}]^{1,118}$$

Exemplo

- Sobrevivência ao diagnóstico de Aids
  - √ Pacientes com tempo de observação menor que 193 dias
    - Óbitos: 90
    - Censuras: 103
  - √ Objetivo:
    - Estimar o efeito do tratamento controlado por idade e sexo

- Modelo de regressão paramétrico – Exponencial:

```
> modhiv.exp <- survreg(Surv(tempo, status)~idade, data = hiv,
+  dist='exponential')
> modhiv.exp <- survreg(Surv(tempo, status)~idade + sexo + tratam,
+     data = hiv, dist='exponential')
> summary(modhiv.exp)

Call:
survreg(formula = Surv(tempo, status) ~ idade + sexo + tratam,
    data = hiv, dist = "exponential")
              Value Std. Error      z        p
(Intercept)  6.11597     0.4921 12.429 1.83e-35
idade        0.00829     0.0112  0.739 4.60e-01
sexoM       -0.19455     0.2841 -0.685 4.93e-01
tratam       1.37541     0.1920  7.164 7.83e-13

Scale fixed at 1

Exponential distribution
Loglik(model)= -743.5   Loglik(intercept only)= -774.6
        Chisq= 62.3 on 3 degrees of freedom, p= 1.9e-13
Number of Newton-Raphson Iterations: 5
n= 193
```

- Modelo de regressão paramétrico – Weibull:

```
> modhiv.wei <- survreg(Surv(tempo, status)~idade + sexo + tratam
+     data = hiv, dist='weibull')
> summary(modhiv.wei)

Call:
survreg(formula = Surv(tempo, status) ~ idade + sexo + tratam,
    data = hiv, dist = "weibull")
               Value Std. Error      z        p
(Intercept)  6.06842     0.5674 10.695 1.07e-26
idade        0.00951     0.0130  0.731 4.65e-01
sexoM       -0.23627     0.3277 -0.721 4.71e-01
tratam       1.48608     0.2273  6.538 6.25e-11
Log(scale)   0.14185     0.0862  1.647 9.97e-02

Scale= 1.15

Weibull distribution
Loglik(model)= -742   Loglik(intercept only)= -770.3
        Chisq= 56.64 on 3 degrees of freedom, p= 3.1e-12
Number of Newton-Raphson Iterations: 5
n= 193
```


Modelo de Tempo de Vida Acelerado

$$Y = \ln(T) = \boldsymbol{x}'\boldsymbol{\beta} + \sigma\nu$$

√ Bastante utilizado na prática

√ Covariáveis aceleram (desaceleram) o tempo de vida

- Distribuição para υ:

√ Valor extremo (T é Weibull)

√ Normal (T é lognormal)

√ Log-gama (T é gama)

√ Logística (T é log-logística)

- Modelo de tempo de vida acelerado

$$Y = \ln(T) = \boldsymbol{x}'\boldsymbol{\beta} + \sigma\nu$$

- Modelo na escala original

$$T = \exp\{\boldsymbol{x}'\boldsymbol{\beta}\}\exp\{\sigma\nu\}$$

√ Generalização em termos paramétricos

– Acréscimo de mais um parâmetro de forma (Ex.: gama generalizada)

√ Generalização mais utilizada:

– Modelo semiparamétrico de Cox (modelos acelerados aparecem como casos particulares)

- Modelo de Vida Acelerada – Interpretação:

√ Vida acelerada:

συ : distribuição de referência quando ***x*** = 0

√ Escala original

T_0 = exp{συ}: distribuição de referência na escala original

√ Probabilidade de indivíduo de referência estar vivo no tempo t:

$$S_0(t) = \mathrm{P}\{T_0 > t\} = \mathrm{P}\left\{Y > \left(\frac{t}{\sigma}\right)\right\}$$

√ Considerando o efeito das covariáveis

$$T \sim T_0\, \mathrm{e}^{\boldsymbol{x}'\boldsymbol{\beta}}$$

– Covariáveis agem multiplicativamente no tempo de sobrevivência

- Probabilidade de indivíduo com valores de covariáveis $\boldsymbol{x}$ estar vivo no tempo t:

$$\begin{aligned} S(t|\boldsymbol{x}) &= \mathrm{P}\{T > t|\boldsymbol{x}\} = \mathrm{P}\left\{T_0\, \mathrm{e}^{\boldsymbol{x}'\boldsymbol{\beta}} > t\right\} \\ &= \mathrm{P}\left\{T_0 > t\,\mathrm{e}^{-\boldsymbol{x}'\boldsymbol{\beta}}\right\} \\ &= S_0\left(t\,\mathrm{e}^{-\boldsymbol{x}'\boldsymbol{\beta}}\right) \end{aligned}$$

  √ Probabilidade de indivíduo com valor $\boldsymbol{x}$ de covariáveis estar vivo no tempo t é a mesma probabilidade de indivíduo de referência estar vivo no tempo t $\mathrm{e}^{-\boldsymbol{x}'\boldsymbol{\beta}}$

  √ O tempo passa mais rapidamente por um fator $\mathrm{e}^{-\boldsymbol{x}'\boldsymbol{\beta}}$

  – (duas vezes mais rápido, metade mais rápido)

- Modelos de tempo de vida acelerados–Inferência:

  √ Propriedades assintóticas dos estimadores de máxima verossimilhança

  √ Forma geral da função de verossimilhança

$$L(\theta) = \prod_{i=1}^{n} [f(t_i, \theta)]^{\delta_i} [S(t_i, \theta)]^{1-\delta_i}$$

  – Especificação de distribuição para T (ou para Y) determinam completamente a função de verossimilhança

  – Estimadores de máxima verossimilhança e suas propriedades (construção de intervalos de confiança e testes de hipóteses)

Comandos em R

`m`: objeto que contém resultado do ajuste

> `summary(m)`

√ `Call`: modelo ajustado

√ `Value`: parâmetros estimados

√ `Std. Error`: erro padrão do parâmetro estimado

√ `z`: estatística para estimativa dos parâmetros

√ `p`: p-valor da estatística z

√ `Scale`: Weibull = 1/g; Exponencial = 1

√ `Loglik(model)`: log-verossimilhança modelo completo

√ `Loglik(intercept only)`: log-verossimilhança modelo nulo

√ `Number of Newton-Raphson Iterations`: estimação

√ `n`: número de observações

- Ajuste dos modelos paramétricos de regressão

  √ Formula = Surv(temp, status) ~

  ~ var1 + var2 + ... (covariáveis)

  ~ 1 (sem covariáveis)

  ~ var1*var2 (interação entre variáveis)

  ~ (var1:var2) (somente inclui o termo de interação)

- Parametrização dos modelos Weibull no R:
  - √ Na presença de covariáveis, os parâmetros de regressão β ficam com o sinal inverso ao que aparece na saída do pacote

Seleção de Modelos

Seleção de Modelos Paramétricos

- Para comparar modelos paramétricos aninhados
  - √ Teste da razão de verossimilhanças
  - √ Teste de Wald

Teste da Razão de Verossimilhanças

- Pode ser usado para comparar modelos aninhados

  (modelo com maior número de parâmetros contém todos os parâmetros do modelo menor)
  - √ Estatística de teste: $\mathrm{RV} = 2\,(l_{maior} - l_{menor})$
  - √ Distribuição amostral: $\mathrm{RV} \sim \chi^2_{gl}$
    - gl: diferença entre o número de parâmetros dos modelos

Exemplo

- Dados de Aids
  - √ Aplicação:
    - Comparar modelos com distribuição exponencial e Weibull
  - √ Modelos:
    - H_0: modhiv.exp = modhiv.wei ($\gamma = 1$)
  - √ Estatística de teste:

$$\text{RV} = 2\,(l_{weibull} - l_{exponencial}) = 2[(-742) - (-743,5)] = 3$$

  - √ Comando R:

```
> anova(modhiv.exp, modhiv.wei)
                   Terms Resid. Df    -2*LL Test Df Deviance   Pr(>Chi)
1 idade + sexo + tratam       189 1486.942      NA       NA         NA
2 idade + sexo + tratam       188 1484.049    =  1 2.893647 0.08892942
```


- Saída do modelo Weibull

```
> modhiv.wei
Call:
survreg(formula = Surv(tempo, status) ~ idade + sexo + tratam,
    data = hiv, dist = "weibull")
Coefficients:
 (Intercept)        idade        sexoM       tratam
 6.068424160  0.009505648 -0.236267819  1.486083990
Scale= 1.152408

Loglik(model)= -742   Loglik(intercept only)= -770.3
        Chisq= 56.64 on 3 degrees of freedom, p= 3.1e-12
n= 193
```

$$\text{RV} = 2\,[l_{estimado} - l_{nulo}] = 2[(-742) - (-770,3)] = 56,6$$

Teste de Wald

- Testa a significância individual de covariáveis
  - √ H_0: $\beta_j = 0$ vs. $\beta_j \neq 0$
  - √ Estatística de teste: $z_j = \dfrac{\hat{\beta}_j}{\text{ep}(\hat{\beta}_j)}$
  - √ Distribuição amostral (sob H_0): $Z_j \sim N(0, 1)$

- Modelo de regressão paramétrico – Weibull:

```
> modhiv.wei <- survreg(Surv(tempo, status)~idade + sexo + tratam,
+     data = hiv, dist='weibull')
> summary(modhiv.wei)

Call:
survreg(formula = Surv(tempo, status) ~ idade + sexo + tratam,
    data = hiv, dist = "weibull")
              Value Std. Error      z        p
(Intercept)  6.06842     0.5674 10.695 1.07e-26
idade        0.00951     0.0130  0.731 4.65e-01
sexoM       -0.23627     0.3277 -0.721 4.71e-01
tratam       1.48608     0.2273  6.538 6.25e-11
Log(scale)   0.14185     0.0862  1.647 9.97e-02

Scale= 1.15

Weibull distribution
Loglik(model)= -742   Loglik(intercept only)= -770.3
        Chisq= 56.64 on 3 degrees of freedom, p= 3.1e-12
```

Único efeito significativo:
- tratamento

  - Número de medicamentos antirretrovirais tem efeito protetor no risco de morte dos pacientes da coorte

Qualidade do Ajuste

- Avaliação da qualidade do ajuste de um modelo paramétrico
  - √ Medida global de qualidade:
    - Função desvio (ou *deviance*)
    - Ferramenta gráfica

Deviance

- Pode ser utilizada para obter medida global de qualidade de ajuste do modelo:
  - √ H_0: o modelo se ajusta aos dados
  - √ Estatística de teste: $D = 2\,[l_{saturado} - l_{modelo}]$
    - Modelo saturado: ajusta n parâmetros para as n observações
    - Modelo estimado
  - √ Parâmetros de perturbação são mantidos fixos
    - Parâmetros comuns aos dois ajustes e nos quais não há interesse no estimador
    - Ex. Parâmetro de forma no modelo de Weibull

- √ Quanto maior a *deviance* pior a qualidade do ajuste
- √ Distribuição amostral (sob H_0): $D \sim \chi^2_{(n-p-1)}$
  - Quanto menor a estatística, maior o p-valor e melhor a qualidade de ajuste
- √ Deviance – Comando no R

```
> # Deviance
>
> modhivwei.dev = sum(resid(modhiv.wei, type = "deviance")^2)
> modhivwei.dev
[1] 237.1074
> gl = 193 - 3 - 1
> 1-pchisq(modhivwei.dev, gl)
[1] 0.01004656
```

Teste de qualidade global de ajuste também rejeita H_0

- Conclusão – Modelo Weibull

```
                Value Std. Error      z        p
(Intercept)   6.06842     0.5674 10.695 1.07e-26
idade         0.00951     0.0130  0.731 4.65e-01
sexoM        -0.23627     0.3277 -0.721 4.71e-01
tratam        1.48608     0.2273  6.538 6.25e-11
Log(scale)    0.14185     0.0862  1.647 9.97e-02

Scale= 1.15
```

- √ Para cada medicamento acrescido no tratamento o risco diminui em 72,5%

$$\exp\{-1,486\}]^{1/1,15} = 0,275$$

- Conclusão – Modelo Exponencial

```
> summary(modhiv.tratam)
Call:
survreg(formula = Surv(tempo, status) ~ tratam, data = hiv, dist = "exponential")
            Value Std. Error     z         p
(Intercept)  6.23      0.173 36.02 3.41e-284
tratam       1.41      0.187  7.57  3.88e-14
Scale fixed at 1

Exponential distribution
Loglik(model)= -744   Loglik(intercept only)= -774.6
        Chisq= 61.32 on 1 degrees of freedom, p= 4.9e-15
```

√ Para cada medicamento acrescido no tratamento o risco diminui em 75,6%

$$\exp\{-1,41\} = 0,244$$

- Critério AIC
  - √ Modelos exponencial (tratam)
  - √ Modelo Weibull (idade + sexo + tratam)
  - √ Modelo exponencial (idade + sexo + tratam)

```
> AIC(modhiv.tratam, modhiv.wei, modhiv.exp)
              df      AIC
modhiv.tratam  2 1491.924
modhiv.wei     5 1494.049
modhiv.exp     4 1494.942
```

$$\begin{aligned}\text{AIC} &= 2\left[p - l(\hat{\beta}_0, \hat{\beta}_{\text{tratam}})\right] \\ &= 2[2-(-743.9620)] \\ &= 1.491,924\end{aligned}$$

– Quanto menor o valor do Critério de Informação de Akaike (AIC), melhor o ajuste do modelo aos dados

Análise Gráfica do Ajuste

- Ferramenta exploratória
  - √ Auxiliar na escolha de distribuições candidatas
- Gráficos:
  - √ log(Λ(t)) vs. log(t)
  - √ Comparação de curvas de sobrevivência
    - – Estimador de Kaplan-Meier vs. Estimação paramétrica

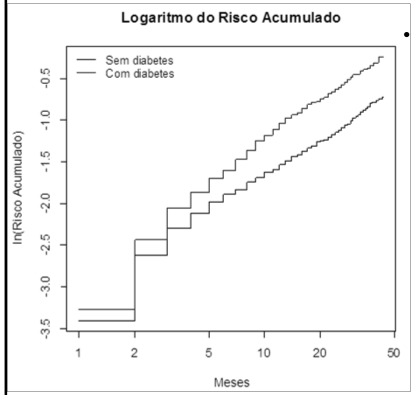
1. Gráfico de log(Λ(t)) vs. log(t)
   - √ log(Λ(t)) = log(–log(S(t))
   - √ Avaliação suposição de proporcionalidade dos riscos
   - √ Forma d risco
     - – Weibull: esperam-se retas paralelas
     - – Exponencial: inclinação igual a 1
     - – Curvas coincidentes: parâmetro referente à covariável provavelmente não será significativo

Exemplo

- Pacientes com insuficiência renal submetidos à hemodiálise
  - √ Período: janeiro/1988 a outubro/2001
  - √ Dados: Sistema Apac
  - √ Coorte com 6.805 pacientes, no Rio de Janeiro
    - 1.603 óbitos
  - √ Covariável:
    - Presença de diabetes
  - √ Modelo de regressão paramétrico de Weibull é candidato?

√ Gráfico do log(Risco acumulado)

- Pacientes com e sem diabetes



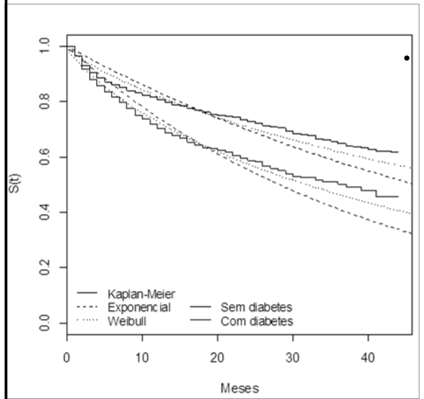
- Retas paralelas
  - Modelo de Weibull aparenta ser adequado

2. Comparação de curvas de sobrevivências
   - √ Estimativa por Kaplan-Meier vs. curvas estimadas parametricamente
   - √ Quanto mais próximo o modelo paramétrico estiver da curva de sobrevivência do estimador e Kaplan-Meier, melhor o modelo

Exemplo – Continuação

- Pacientes com insuficiência renal submetidos à hemodiálise
  - √ Covariável:
    - Presença ou ausência de diabetes
  - √ Objetivo:
    - Comparação de curvas de sobrevivência:
      - Estimativa não paramétrica de Kaplan-Meier
      - Estimativa do modelo de regressão exponencial
      - Estimativa do modelo de regressão de Weibull

√ Comparação de curvas de sobrevivência



- Modelo que mais se aproxima da estimação não paramétrica
  - Modelo de regressão de Weibull

Comentários

√ Análises gráficas de avaliação de qualidade do ajuste do modelo são mais indicadas do que testes estatísticos formais

– Tendem a ter baixo poder para amostras pequenas e rejeitar o modelo para amostras grandes

√ Os métodos apresentados devem ser usados para descartar modelos claramente inapropriados

– E não para demonstrar que um modelo paramétrico é melhor

– (nesse caso o melhor é análise de resíduos)

√ É comum diferentes modelos paramétricos terem ajustes razoáveis para um mesmo conjunto de dados

– Especialmente nos intervalos que concentram um maior número de observações, apresentando resultados semelhantes das quantidades de interesse

Análise de Resíduos

- No contexto de sobrevivência, a definição de uma medida de resíduo não clara
- Resíduos para modelos paramétricos:
  - √ Resíduos de perturbação para modelos tempos de vida acelerados
- Resíduos para modelos de Cox
  - √ Resíduo de Schoenfeld
    - – Avaliação do pressuposto de proporcionalidade

- Resíduos para modelos paramétricos e de Cox:
  - √ Resíduo de Cox-Snell
  - √ Resíduo martingale
  - √ Resíduo *deviance*
  - √ Resíduo de escore

- Resíduos de perturbação para modelos tempos de vida acelerados
  - √ Avaliam o impacto da retirada de uma observação no ajuste global do modelo
  - √ Tipos:
    - Mede mudança no vetor estimado de parâmetros
    - Mede mudança na resposta predita (em desvios padrão)
    - Avalia efeito sobre parâmetro de forma

- Objetivo
  - √ Identificar pontos influentes do modelo
  - √ Verificar seu impacto sobre
    - Conjunto de parâmetros de regressão: **`ldcase`**
    - Valores preditos: **`ldresp`**
    - Parâmetro de forma: **`ldshape`**
- Comandos em R:
  - √ Para identificar indivíduos muito afastados
    - Usar função **`identify`**

Exemplo – Continuação

- Sobrevivência ao diagnóstico de Aids
  - √ Pacientes com tempo de observação menor que 193 dias
    - Óbitos: 90
    - Censuras: 103
  - √ Objetivo:
    - Determinar os resíduos do modelo paramétrico de Weibull
    - Identificar e analisar possíveis pontos influentes

- Modelo de regressão paramétrico – Weibull:

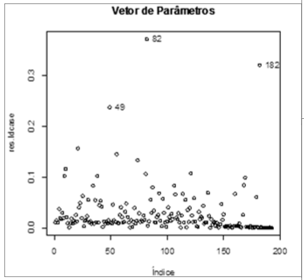
```
> modhiv.wei <- survreg(Surv(tempo, status)~idade + sexo + tratam
+     data = hiv, dist='weibull')
> summary(modhiv.wei)

Call:
survreg(formula = Surv(tempo, status) ~ idade + sexo + tratam,
    data = hiv, dist = "weibull")
               Value Std. Error      z        p
(Intercept)  6.06842     0.5674 10.695 1.07e-26
idade        0.00951     0.0130  0.731 4.65e-01
sexoM       -0.23627     0.3277 -0.721 4.71e-01
tratam       1.48608     0.2273  6.538 6.25e-11
Log(scale)   0.14185     0.0862  1.647 9.97e-02

Scale= 1.15

Weibull distribution
Loglik(model)= -742   Loglik(intercept only)= -770.3
        Chisq= 56.64 on 3 degrees of freedom, p= 3.1e-12
Number of Newton-Raphson Iterations: 5
n= 193
```


- Identificação dos resíduos

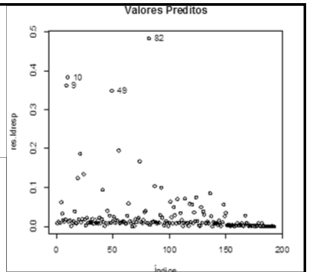
```
> # Análise de Resíduos de Perturbação - Modelo de Weibull
>
> res.ldcase <- residuals(modhiv.wei, type = "ldcase")
> res.x = 1:length(res.ldcase)
> plot(res.ldcase, xlab = "Índice")
> title("Vetor de Parâmetros")
> idx.ldcase<-identify(x = res.x, y = res.ldcase, n = 5 )
>
> res.ldresp <- residuals(modhiv.wei, type = "ldresp")
> plot(res.ldresp, xlab = "Índice")
> title("Valores Preditos")
> idx.ldresp<-identify(x = res.x, y = res.ldresp, n = 5)
>
> res.ldshape <- residuals(modhiv.wei, type = "ldshape")
> plot(res.ldshape, xlab = "Índice")
> title("Parâmetro de Forma")
> idx.ldshape<-identify(x = res.x, y = res.ldshape, n = 5)
>
> idx<-as.numeric(levels(as.factor(c(idx.ldcase, idx.ldresp, idx.ldshape))))
> idx
[1]   9  10  49  82 182
```

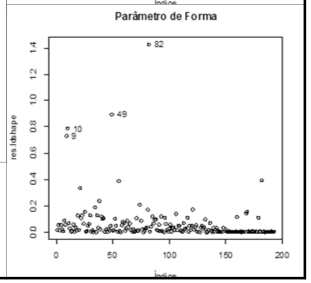

√ Resíduos de perturbação de verossimilhança

- Coorte de Aids – modelo de regressão de Weibull



- Casos identificados:
  - 9, 10, 49, 82, 182





- Casos identificados nos gráficos:

```
> idx
[1]   9  10  49  82 182
> hiv[idx,c("tempo", "status", "sexo", "idade", "tratam")]
    tempo status sexo idade tratam
9    1563      1    M    44      0
10   1247      1    M    23      0
49   1344      0    M    30      0
82   1272      0    M    22      0
182    16      1    M    42      3
```

√ Casos 9 e 10:

- Sobreviveram um longo tempo

√ Casos 49 e 82:

- Censurados sem receber antirretroviral

√ Caso 182:

- Morreu rapidamente e recebeu 3 medicamentos

- Modelo de regressão de Weibull, sem o caso 82
  - √ Aparece nos três gráficos

```
> summary(survreg( formula = Surv(tempo, status) ~ idade + sexo + tratam,
+    data = hiv, dist = "weibull", subset = -82))
Call:
survreg(formula = Surv(tempo, status) ~ idade + sexo + tratam,
    data = hiv, subset = -82, dist = "weibull")
              Value Std. Error      z        p
(Intercept)  5.7996     0.5760 10.069 7.60e-24
idade        0.0151     0.0133  1.137 2.55e-01
sexoM       -0.2603     0.3231 -0.806 4.20e-01
tratam       1.5490     0.2266  6.836 8.16e-12
Log(scale)   0.1281     0.0857  1.496 1.35e-01
Scale= 1.14
Weibull distribution
Loglik(model)= -739.2   Loglik(intercept only)= -769.7
        Chisq= 61.03 on 3 degrees of freedom, p= 3.5e-13
Number of Newton-Raphson Iterations: 5
n= 192
```


- Modelo de regressão paramétrico – Weibull:

```
> modhiv.wei <- survreg(Surv(tempo, status)~idade + sexo + tratam
+     data = hiv, dist='weibull')
> summary(modhiv.wei)

Call:
survreg(formula = Surv(tempo, status) ~ idade + sexo + tratam,
    data = hiv, dist = "weibull")
               Value Std. Error      z        p
(Intercept)  6.06842     0.5674 10.695 1.07e-26
idade        0.00951     0.0130  0.731 4.65e-01
sexoM       -0.23627     0.3277 -0.721 4.71e-01
tratam       1.48608     0.2273  6.538 6.25e-11
Log(scale)   0.14185     0.0862  1.647 9.97e-02

Scale= 1.15

Weibull distribution
Loglik(model)= -742   Loglik(intercept only)= -770.3
        Chisq= 56.64 on 3 degrees of freedom, p= 3.1e-12
Number of Newton-Raphson Iterations: 5
n= 193
```


√ Modelo de Weibull – Comparação das estimativas dos parâmetros e da log-verossimilhança

	Completo	**Sem indivíduo 82**
Intercept	6,0684	5,7996
Idade	0,0095	0,0151
sexoM	–0,2363	–0,2603
tratam	1,4861	1,5490
Scale	1,1524	1,1367
Shape	0,868	0,880
Loglik(model)	–742,02	–739,21

√ Alteração das estimativas dos parâmetros (**`ldcase`**)

√ Aumento da log-verossimilhança (**`ldresp`**)

√ Mudança na estimativa do parâmetro de forma (**`ldshape`**)

Exemplo

- Sobrevida de pacientes com leucemia aguda
  - √ Tempo de sobrevivência a diagnóstico (semanas)
  - √ 17 pacientes com leucemia aguda
  - √ Dados sem censura
  - √ Covariável
    - wbc: contagem de glóbulos brancos na data do diagnóstico
    - lwbc: $\log_{10}$(WBC)

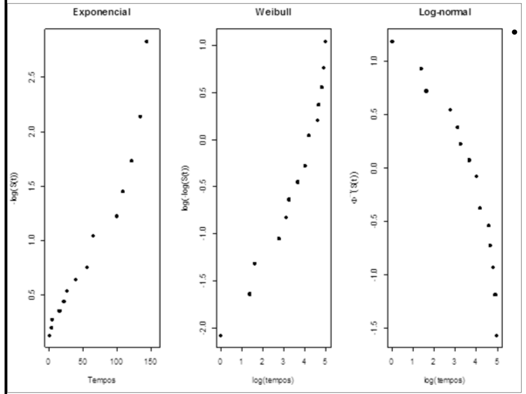
√ Pacientes com leucemia – Conjunto de dados

```
> temp<-c(65,156,100,134,16,108,121,4,39,143,56,26,22,1,1,5,65)
> cens<-rep(1,17)
> lwbc<-c(3.36,2.88,3.63,3.41,3.78,4.02,4.00,4.23,3.73,3.85,3.97,
+          4.51,4.54,5.00,5.00,4.72,5.00)
> dados<-cbind(temp,cens,lwbc)
> dados<-as.data.frame(dados)
> dados
  temp cens lwbc
1    65    1 3.36
2   156    1 2.88
3   100    1 3.63
4   134    1 3.41
5    16    1 3.78
6   108    1 4.02
7   121    1 4.00
8     4    1 4.23
9    39    1 3.73
10  143    1 3.85
11   56    1 3.97
12   26    1 4.51
13   22    1 4.54
14    1    1 5.00
15    1    1 5.00
16    5    1 4.72
17   65    1 5.00
```

Dados sem censura

– Inviável estratificação pela covariável lwbc

- Modelo de regressão para análise dos dados
  √ Gráficos de linearização - sem a covariável: lwbc



- Candidatas:
  - Exponencial
  - Weibull

√ Modelo de regressão linear

– Covariável: X_1 = log(WBC)

– Exponencial e Weibull

```
> leuk.exp<-survreg(Surv(temp, cens)~ lwbc,
+ data = dados, dist='exponential')
> summary(leuk.exp)
Call:
survreg(formula = Surv(temp, cens) ~ lwbc,
        data = dados, dist = "exponential")
              Value Std. Error     z        p
(Intercept)    8.48      1.711  4.95 7.27e-07
lwbc          -1.11      0.414 -2.68 7.31e-03
Scale fixed at 1
Exponential distribution
Loglik(model)= -83.9   Loglik(intercept only)= -87
        Chisq= 6.83 on 1 degrees of freedom, p= 0.
Number of Newton-Raphson Iterations: 5
n= 17
> leuk.exp$loglik
[1] -87.28983 -83.87705
```

$$\hat\beta_0 = -8,4775$$
$$\hat\beta_{\text{lwbc}} = 1,1093$$

```
> leuk.wei<-survreg(Surv(temp, cens)~ lwbc,
+ data = dados, dist='weibull')
> summary(leuk.wei)
Call:
survreg(formula = Surv(temp, cens) ~ lwbc,
               data = dados, dist = "weibull")
              Value Std. Error      z        p
(Intercept)  8.4408      1.709  4.940 7.81e-07
lwbc        -1.0982      0.418 -2.630 8.53e-03
Log(scale)  -0.0216      0.202 -0.107 9.15e-01
Scale= 0.979
Weibull distribution
Loglik(model)= -83.9   Loglik(intercept only)= -87.1
        Chisq= 6.48 on 1 degrees of freedom, p= 0.011
Number of Newton-Raphson Iterations: 6
n= 17
> leuk.wei$loglik
[1] -87.10948 -83.87136
```

Log scale não significativo

$$\hat\beta_0 = -8,4408$$
$$\hat\beta_{\text{lwbc}} = 1,0982$$
$$\hat\gamma = \frac{1}{\text{scale}} = \frac{1}{0,979} = 1,0218$$

- Estimativas para os dados de leucemia aguda

Regressão

Exponencial	Weibull
β_0 = –8,4775	β_0 = –8,4408
β_1 = 1,1093	β_1 = 1,0982
γ = 1 (fixo)	γ = 1,0218

√ Estimativa de γ muito próxima de 1

– (log scale é não significativo)

– Indicativo de que o modelo é exponencial (γ=1)

√ Teste da razão da verossimilhança

– Weibull e exponencial são aninhados

$$\boxed{\text{RV} = 2(83,8771 - 83,8714) = 0,0113} \qquad \boxed{p = \text{P}\{\chi^2_1 > 0,0113\} = 0,915}$$

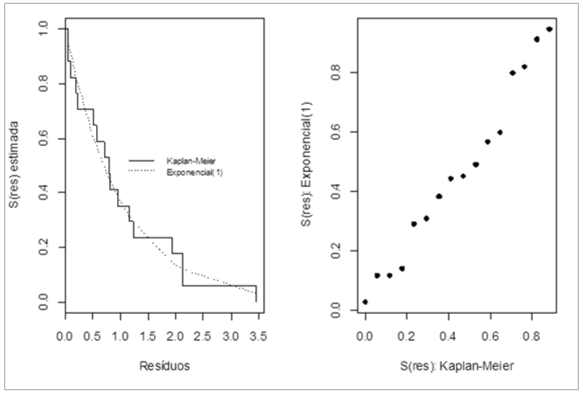
– Resultado fornece indicações favoráveis ao modelo de regressão exponencial

- Avaliação do ajuste
  - √ Resíduos de Cox-Snell – modelo exponencial

$$\boxed{\hat{e}_i = \hat{\Lambda}(t_i|x_1) = -\ln[S(t_i|x_1)] = \left[t_i \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_1\}\right]}$$

  - √ Se o modelo for adequado
    - Resíduos seguem distribuição exponencial padrão
    - Curvas de sobrevivência dos resíduos, por Kaplan-Meier e exponencial padrão devem estar próximas
    - Gráfico dos pontos (S_{KM}(res), S_{exp}(res)) devem se aproximar de uma reta

√ Análise gráfica dos resíduos Cox-Snell – Exponencial



- Exponencial padrão parece aceitável (modelo adequado)

- Teste da razão de verossimilhanças
  - √ H_0: $\beta_1 = 0$
  - √ Estatística de teste

$$\boxed{\mathrm{RV} = 2(87.2898 - 83.8771) = 6,825}$$

  - √ Probabilidade de significância

$$\boxed{p = \mathrm{P}\{\chi_1^2 > 6,825\} = 0.0090}$$

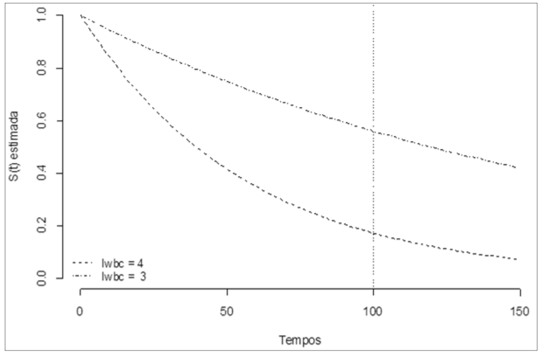
  - √ Rejeita-se H_0
    - É possível que parte da variação observada nos tempos de sobrevivência dos pacientes pode ser explicada pela contagem de glóbulos brancos

- Modelo de regressão exponencial
  - √ Função de sobrevivência estimada

$$S(t|x_1) = \exp\{-\exp\{-8,4775 + 1.1093x_1\}\, t\},\ t \geq 0$$

  - √ Na expressão, β_1 é positivo
    - Quanto maior o valor de x1, menor a probabilidade de sobrevivência estimada

√ Curvas de sobrevivência para valores de lwbc



- $S(100 \mid x_1 = 4) = 17\%$ (estarão vivos em 100 semanas)
- $S(100 \mid x_2 = 3) = 56\%$

- Comentários
  - √ Modelo de regressão exponencial ajustou-se satisfatoriamente aos dados
  - √ Conclusão:
    - Tempo estimado diminui à medida que são observadas contagens crescentes de glóbulos brancos, no diagnóstico
  - √ Probabilidade de sobrevivência dado valor de x_1:

$$S(t|x_1) = \exp\{-\exp\{-8,4775 + 1.1093x_1\}\, t\},\ t \geq 0$$

Exemplo

- Sobrevida de pacientes com leucemia aguda
  - √ Tempo de sobrevivência a diagnóstico (semanas)
    - 17 pacientes que expressam antígeno Calla (Ag+)
    - 16 paciente que não expressam antígeno Calla (Ag–)
  - √ Dados sem censura
  - √ Covariáveis
    - X_1: lwbc: $\log_{10}$(WBC)
    - X_2: grupo (0, se grupo Ag+; 1, se grupo Ag–).
  - √ Foi escolhido modelo de regressão exponencial para o grupo Ag+.
    - Investigar se o mesmo se aplica ao grupo Ag–

√ Pacientes com leucemia – Conjunto de dados

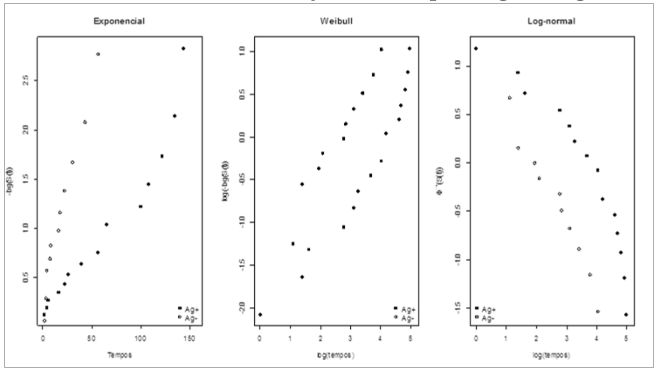
```
> tempo <- c(65,156,100,134,16,108,121,4,39,143,56,26,22,1,1,5,65,
+            56,65,17,7,16,22,3,4,2,3,8,4,3,30,4,43)
> status <- c(rep(1,17),rep(1,16))
> lwbc <-
c(3.36,2.88,3.63,3.41,3.78,4.02,4,4.23,3.73,3.85,3.97,4.51,4.54,5,5,4.72,
+
5,3.64,3.48,3.6,3.18,3.95,3.72,4,4.28,4.43,4.45,4.49,4.41,4.32,4.90,5,5)
>  grupo <- c(rep(0,17),rep(1,16))
> dados<-cbind(tempo, status, lwbc, grupo)
> dados<-as.data.frame(dados)
> dados
```

Dados sem censura

```
   tempo status lwbc grupo
1     65      1 3.36     0
2    156      1 2.88     0
3    100      1 3.63     0
4    134      1 3.41     0
5     16      1 3.78     0
6    108      1 4.02     0
7    121      1 4.00     0
8      4      1 4.23     0
9     39      1 3.73     0
10   143      1 3.85     0
11    56      1 3.97     0
12    26      1 4.51     0
13    22      1 4.54     0
14     1      1 5.00     0
15     1      1 5.00     0
16     5      1 4.72     0
17    65      1 5.00     0
```

```
   tempo status lwbc grupo
18    56      1 3.64     1
19    65      1 3.48     1
20    17      1 3.60     1
21     7      1 3.18     1
22    16      1 3.95     1
23    22      1 3.72     1
24     3      1 4.00     1
25     4      1 4.28     1
26     2      1 4.43     1
27     3      1 4.45     1
28     8      1 4.49     1
29     4      1 4.41     1
30     3      1 4.32     1
31    30      1 4.90     1
32     4      1 5.00     1
33    43      1 5.00     1
```

√ Gráficos das linearizações – Grupos Ag+ e Ag–



√ Possível observar indicações favoráveis ao modelo de regressão exponencial também para o grupo Ag–

- Ajuste de modelos de regressão exponencial
  - √ Modelo 1:
    - Nenhuma covariável incluída
  - √ Modelo 2:
    - Incluída apenas a covariável X_1: lwbc
  - √ Modelo 3:
    - Incluída apenas a covariável X_2: grupos
  - √ Modelo 4:
    - Incluída as covariáveis X_1 e X_2
  - √ Modelo 5:
    - Incluídas variáveis X1, X2 e X1*X2


```
> leuk.mod1 <-survreg(Surv(temp, status)~ 1,
           data = dados, dist = 'exponential')
> leuk.mod1
Call:
survreg(formula = Surv(temp, status) ~ 1,
           data = dados, dist = "exponential")
Coefficients:
(Intercept)
   3.710611
Scale fixed at 1
Loglik(model)= -155.5   Loglik(intercept only)= -155.5
n= 33
```

Modelo 1:
nenhuma covariável

$\hat{\beta}_0 = -3,7106$

$l(\hat{\boldsymbol{\beta}}) = -155,5$

```
> leuk.mod2 <- survreg(Surv(temp, status)~ lwbc,
                 data = dados, dist = 'exponential')
> leuk.mod2
Call:
survreg(formula = Surv(temp, status) ~ lwbc,
        data = dados, dist = "exponential")
Coefficients:
(Intercept)        lwbc
  7.3712449  -0.9229052
Scale fixed at 1
Loglik(model)= -150.3   Loglik(intercept only)= -155.5
        Chisq= 10.31 on 1 degrees of freedom, p= 0.0013
n= 33
```

Modelo 2:
X_1

$\hat{\beta}_0 = -7,3712$

$\hat{\beta}_1 = 0,9229$

$l(\hat{\boldsymbol{\beta}}) = -150,3$


```
> leuk.mod3 <- survreg(Surv(temp, status)~ grupo,
                 data = dados, dist = 'exponential')
> leuk.mod3
Call:
survreg(formula = Surv(temp, status) ~ grupo,
        data = dados, dist = "exponential")
Coefficients:
(Intercept)       grupo
   4.134696   -1.247802
Scale fixed at 1
Loglik(model)= -149.5   Loglik(intercept only)= -155.5
        Chisq= 11.94 on 1 degrees of freedom, p= 0.00055
n= 33
```

Modelo 3:
X_2

$\hat{\beta}_0 = -4,1347$

$\hat{\beta}_2 = 1,2478$

$l(\hat{\boldsymbol{\beta}}) = -149,5$

```
> leuk.mod4 <- survreg(Surv(temp, status)~ lwbc + grupo,
                 data = dados, dist = 'exponential')
> leuk.mod4
Call:
survreg(formula = Surv(temp, status) ~ lwbc + grupo,
        data = dados, dist = "exponential")
Coefficients:
(Intercept)        lwbc       grupo
  6.8295300  -0.7000326  -1.0180633
Scale fixed at 1
Loglik(model)= -146.5   Loglik(intercept only)= -155.5
        Chisq= 17.81 on 2 degrees of freedom, p= 0.00014
n= 33
```

Modelo 4:
X_1

$\hat{\beta}_0 = -6,8295$

$\hat{\beta}_1 = 0,7000$

$\hat{\beta}_2 = 1,0181$

$l(\hat{\boldsymbol{\beta}}) = -146,5$

```
> leuk.mod5 <- survreg(Surv(temp, status)~ lwbc + grupo + lwbc*grupo,
                       data = dados, dist = 'exponential')
> leuk.mod5
Call:
survreg(formula = Surv(temp, status)~lwbc + grupo + lwbc*grupo,
                       data = dados, dist = "exponential")
Coefficients:
(Intercept)        lwbc       grupo  lwbc:grupo
   8.477498   -1.109298   -4.138509    0.755652
Scale fixed at 1
Loglik(model)= -145.7   Loglik(intercept only)= -155.5
        Chisq= 19.58 on 3 degrees of freedom, p= 0.00021
n= 33
```

Modelo 5:
X_1 , X_2 e X_1 *X_2

$\hat{\beta}_0 = -8,4775$
$\hat{\beta}_1 = 1,1093$
$\hat{\beta}_2 = 4,1385$
$\hat{\beta}_3 = -0,7557$
$l(\hat{\boldsymbol{\beta}}) = -145,7$

√ Estimativas dos parâmetros e da log-verossimilhança

Modelo	Covariáveis	Estimativas	Log-verossimilhança
1	Nenhuma	$\beta_0 = -3,7106$	$l_1 = -155,5$
2	X_1	$\beta_0 = -7,3712$	$l_2 = -150,3$
		$\beta_1 = 0,9229$	
3	X_2	$\beta_0 = -4,1347$	$l_3 = -149,5$
		$\beta_1 = 1,2478$	
4	X_1 e X_2	$\beta_0 = -6,8295$	$l_4 = -146,5$
		$\beta_1 = 0,7000$	
		$\beta_2 = 1,0181$	
5	X_1, X_2 e X_1*X_2	$\beta_0 = -8,4775$	$l_5 = -145,7$
		$\beta_1 = 1,1093$	
		$\beta_2 = 4,1385$	
		$\beta_3 = -0,7557$	

√ Testes da razão de verossimilhanças – Resultados:

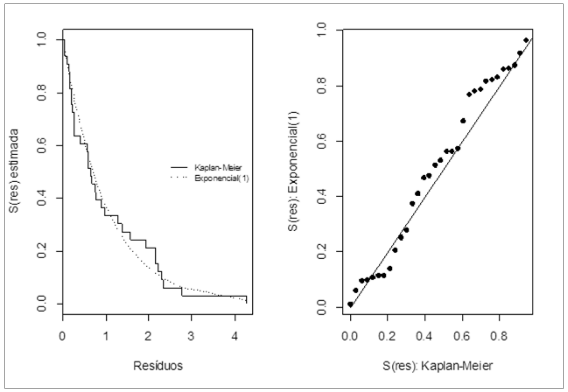
Efeito	H_0	RV	GL	p-valor
Interação: X_1*X_2	$\beta_3 = 0$	2(146,5–145,7)=1,6	1	0,2059
de X_2 | X_1	$\beta_2 = 0$	2(150,3–146,5)=7,6	1	0,0058
de X_1 | X_2	$\beta_1 = 0$	2(149,5–146,5)=8,0	1	0,0047

√ Não há evidências estatísticas de que a interação entre X_1 e X_2 seja significativa

√ Evidências estatísticas do efeito da covariável X_1
(p = 0,0047)

√ Evidências estatísticas do efeito da covariável X_2
(p = 0,0058)

√ Análise gráfica dos resíduos Cox-Snell – Exponencial



– Modelo exponencial apresenta ajuste razoável

√ Resultados do modelo de regressão exponencial final

– Modelo 4

```
> summary(leuk.mod4)
Call:
survreg(formula = Surv(temp, status) ~ lwbc + grupo, data = dados,
    dist = "exponential")
              Value Std. Error     z        p
(Intercept)  6.83      1.158  5.90 3.73e-09
lwbc        -0.70      0.286 -2.45 1.44e-02
grupo       -1.02      0.364 -2.80 5.12e-03
Scale fixed at 1
Exponential distribution
Loglik(model)= -146.5   Loglik(intercept only)= -155.5
        Chisq= 17.81 on 2 degrees of freedom, p= 0.00014
Number of Newton-Raphson Iterations: 5
n= 33
```

– Rejeita-se H_0: $\beta_1 = \beta_2 = 0$ (p-valor = 0,00014)

– É possível que parte da variação observada na resposta possa ser explicada pela contagem de glóbulos brancos e pelo grupo de antígeno

- Modelo de regressão exponencial – Modelo 4

√ Função de sobrevivência estimada

$$S(t|x_1, x_2) = \exp\{-\exp\{-6,8295 + 0,7000x_1 + 1,0181x_2\}\, t\},\ t \geq 0$$

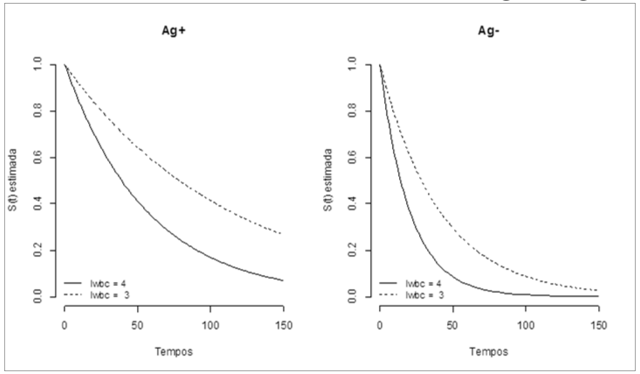
√ $\beta_1 > 0$

– Quanto maior o valor de x_1 (contagem de glóbulos brancos) menor a probabilidade de sobrevivência estimada

√ $\beta_2 > 0$

– Pacientes do grupo Ag– (x_2 = 1) apresentam probabilidades de sobrevivência menores do que os pacientes do grupo Ag+ (x_2 = 0)

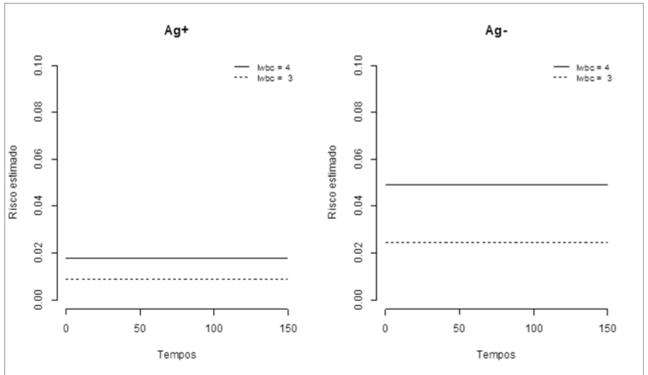
√ Curvas de sobrevivências estimadas – Ag+ e Ag–



√ Pacientes do grupo Ag– com menor sobrevivência

– Menor sobrevivência para pacientes com x_1 maior

√ Riscos estimados pelo modelo – Ag+ e Ag–



√ Riscos constantes ao longo do tempo

– Quanto maior a contagem, maior o risco

- Conclusões:
  - √ Quanto maior a contagem de glóbulos brancos no diagnóstico, maior o risco
    - Nos 2 grupos
  - √ Pacientes que apresentam o antígeno Calla ($x_2=0$) têm melhor prognóstico do que os que ainda não experimentaram

Exemplo

- Análise de dados de aleitamento materno
  - √ Estudo ambulatorial sobre amamentação
  - √ Amostra:
    - 150 mães de crianças com menos de 2 anos
  - √ Resposta:
    - Tempo máximo de aleitamento materno
    - Censura: crianças não acompanhadas até o desmame
  - √ Covariáveis: 11
  - √ Objetivo:
    - Identificar fatores de risco ou de proteção para desmame precoce

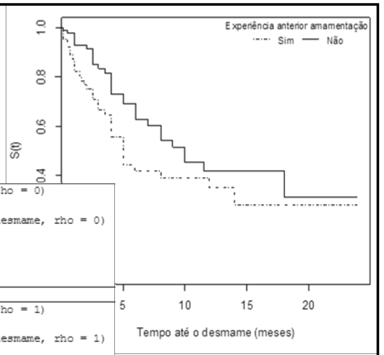
- Descrição das covariáveis

Código	Descrição	Categorias
V1	Experiência anterior amamentação	0: sim 1: não
V2	Número de filhos vivos	0: ≥ 2 1: < 2
V3	Conceito materno sobre o tempo ideal de amamentação	0: > 6 meses 1: ≤ 6 meses
V4	Dificuldade para amamentar nos primeiros dias pós-parto	0: não 1: sim
V5	Tipo de serviço em que realizou o pré-natal	0: público 1: privado/convênios
V6	Recebeu exclusivamente leite materno na maternidade	0: sim 1: não
V7	A criança teve contato com o pai	0: sim 1: não
V8	Renda per capita (em SM/mês)	0: ≥ 1 SM 1: < 1 SM
V9	Peso ao nascimento	0: ≥ 2,5 kg 1: < 2,5 kg
V10	Tempo de separação mãe-filho pós-parto	0: ≤ 6 horas 1: > 6 horas
V11	Permanência no berçário	0: não 1: sim

  - √ Todas as variáveis são dicotômicas
    (possível comparar categorias pelo Kaplan-Meier)

- Análise Descritiva e Exploratória
  - √ Construir estimativas Kaplan-Meier para comparar as duas categorias
  - √ Conduzir testes logrank e Peto para testar
    - H_0: $S_1(t) = S_2(t)$
  - √ Todas as covariáveis com p-valor < 0,25 devem ser incluídas na etapa de modelagem estatística
    - Escolha de nível de significância modesto é recomendado por:
      - Bendel e Afifi (1977) – para regressão linear
      - Constanza e Afifi (1979) – para análise discriminante
      - Mickey e Greenland (1989) – para mudança nos coeficientes de modelo de regressão logística

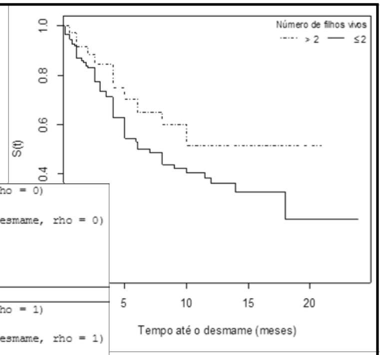
√ V1: Experiência anterior de amamentação




```
> survdiff(Surv(tempo, cens)~ V1, data = desmame, rho = 0)
Call:
survdiff(formula = Surv(tempo, cens) ~ V1, data = desmame, rho = 0)
       N Observed Expected (O-E)^2/E (O-E)^2/V
V1=0 87       32     39.6      1.46      3.95
V1=1 63       33     25.4      2.28      3.95
 Chisq= 3.9  on 1 degrees of freedom, p= 0.0469
```

```
> survdiff(Surv(tempo, cens)~ V1, data = desmame, rho = 1)
Call:
survdiff(formula = Surv(tempo, cens) ~ V1, data = desmame, rho = 1)
       N Observed Expected (O-E)^2/E (O-E)^2/V
V1=0 87     22.5     29.6      1.69      5.67
V1=1 63     26.3     19.2      2.61      5.67
 Chisq= 5.7  on 1 degrees of freedom, p= 0.0173
```

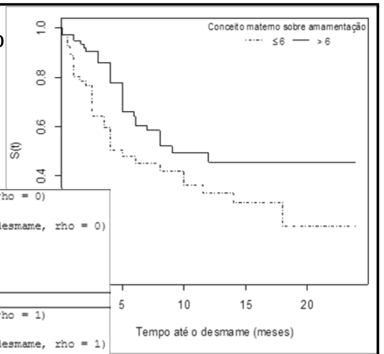

√ V2: Número de filhos vivos




```
> survdiff(Surv(tempo, cens)~ V2, data = desmame, rho = 0)
Call:
survdiff(formula = Surv(tempo, cens) ~ V2, data = desmame, rho = 0)
        N Observed Expected (O-E)^2/E (O-E)^2/V
V2=0 113       54     48.5     0.628       2.6
V2=1  37       11     16.5     1.843       2.6
 Chisq= 2.6  on 1 degrees of freedom, p= 0.107
```

```
> survdiff(Surv(tempo, cens)~ V2, data = desmame, rho = 1)
Call:
survdiff(formula = Surv(tempo, cens) ~ V2, data = desmame, rho = 1)
        N Observed Expected (O-E)^2/E (O-E)^2/V
V2=0 113    40.58     36.5     0.457      2.39
V2=1  37     8.24     12.3     1.353      2.39
 Chisq= 2.4  on 1 degrees of freedom, p= 0.122
```

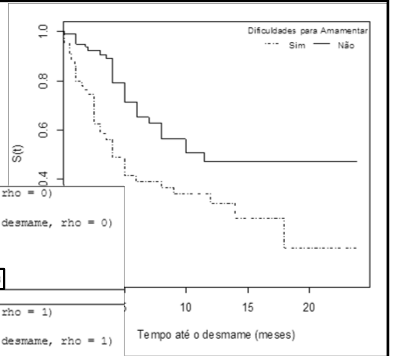

√ V3:Tempo ideal amamentação




```
> survdiff(Surv(tempo, cens)~ V3, data = desmame, rho = 0)
Call:
survdiff(formula = Surv(tempo, cens) ~ V3, data = desmame, rho = 0)
       N Observed Expected (O-E)^2/E (O-E)^2/V
V3=0 80       30     39.5      2.27      6.15
V3=1 70       35     25.5      3.51      6.15
 Chisq= 6.1  on 1 degrees of freedom, p= 0.0132
```

```
> survdiff(Surv(tempo, cens)~ V3, data = desmame, rho = 1)
Call:
survdiff(formula = Surv(tempo, cens) ~ V3, data = desmame, rho = 1)
       N Observed Expected (O-E)^2/E (O-E)^2/V
V3=0 80     21.2     29.6      2.36      7.93
V3=1 70     27.6     19.3      3.62      7.93
 Chisq= 7.9  on 1 degrees of freedom, p= 0.00485
```


√ V4: Apresentou dificuldades amamentar

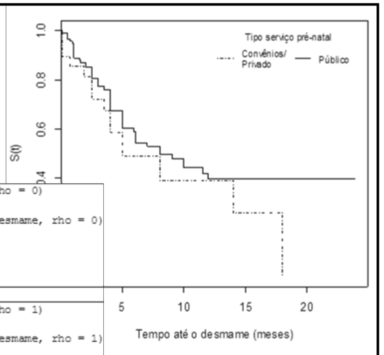



```
> survdiff(Surv(tempo, cens)~ V4, data = desmame, rho = 0)
Call:
survdiff(formula = Surv(tempo, cens) ~ V4, data = desmame, rho = 0)
       N Observed Expected (O-E)^2/E (O-E)^2/V
V4=0 84       28     41.2      4.24      12.3
V4=1 66       37     23.8      7.36      12.3
 Chisq= 12.3  on 1 degrees of freedom, p= 0.000463
```

```
> survdiff(Surv(tempo, cens)~ V4, data = desmame, rho = 1)
Call:
survdiff(formula = Surv(tempo, cens) ~ V4, data = desmame, rho = 1)
       N Observed Expected (O-E)^2/E (O-E)^2/V
V4=0 84     19.3     30.7      4.19      14.9
V4=1 66     29.5     18.1      7.09      14.9
 Chisq= 14.9  on 1 degrees of freedom, p= 0.000114
```

- Mães que não tiveram dificuldade para amamentar apresentam tempo até desmame maior

√ V5: Tipo de serviço pré-natal

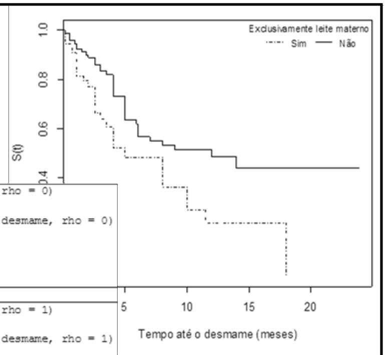



```
> survdiff(Surv(tempo, cens)~ V5, data = desmame, rho = 0)
Call:
survdiff(formula = Surv(tempo, cens) ~ V5, data = desmame, rho = 0)
       N Observed Expected (O-E)^2/E (O-E)^2/V
V5=0 122       49     52.6     0.249      1.38
V5=1  28       16     12.4     1.059      1.38
 Chisq= 1.4  on 1 degrees of freedom, p= 0.241
```

```
> survdiff(Surv(tempo, cens)~ V5, data = desmame, rho = 1)
Call:
survdiff(formula = Surv(tempo, cens) ~ V5, data = desmame, rho = 1)
       N Observed Expected (O-E)^2/E (O-E)^2/V
V5=0 122     36.9     39.6     0.182      1.29
V5=1  28     11.9      9.2     0.784      1.29
 Chisq= 1.3  on 1 degrees of freedom, p= 0.257
```

Análise de Sobrevivência Aplicada à Saúde -- 2014

√ V6: Recebeu somente leite materno

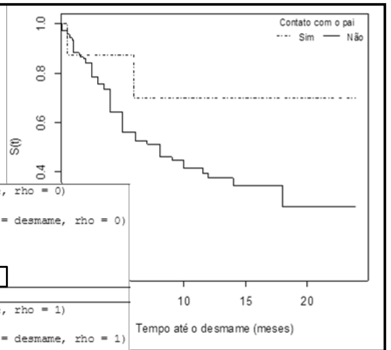



```
> survdiff(Surv(tempo, cens)~ V6, data = desmame, rho = 0)
Call:
survdiff(formula = Surv(tempo, cens) ~ V6, data = desmame, rho = 0)
      N Observed Expected (O-E)^2/E (O-E)^2/V
V6=0 94       36     45.8      2.08      7.47
V6=1 56       29     19.2      4.96      7.47
 Chisq= 7.5  on 1 degrees of freedom, p= 0.00629
```

```
> survdiff(Surv(tempo, cens)~ V6, data = desmame, rho = 1)
Call:
survdiff(formula = Surv(tempo, cens) ~ V6, data = desmame, rho = 1)
      N Observed Expected (O-E)^2/E (O-E)^2/V
V6=0 94     26.4     33.9      1.64      7.02
V6=1 56     22.4     15.0      3.72      7.02
 Chisq= 7  on 1 degrees of freedom, p= 0.00806
```

Análise de Sobrevivência Aplicada à Saúde -- 2014

√ V7: Contato com o pai

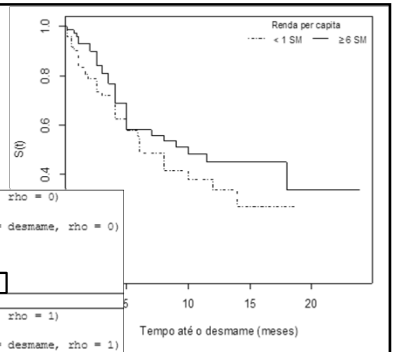



```
> survdiff(Surv(tempo, cens)~ V7, data = desmame, rho = 0)
Call:
survdiff(formula = Surv(tempo, cens) ~ V7, data = desmame, rho = 0)
       N Observed Expected (O-E)^2/E (O-E)^2/V
V7=0 141       63    60.23     0.128      1.84
V7=1   9        2     4.77     1.611      1.84
 Chisq= 1.8  on 1 degrees of freedom, p= 0.175
```

```
> survdiff(Surv(tempo, cens)~ V7, data = desmame, rho = 1)
Call:
survdiff(formula = Surv(tempo, cens) ~ V7, data = desmame, rho = 1)
       N Observed Expected (O-E)^2/E (O-E)^2/V
V7=0 141    47.27    45.46    0.0715      1.45
V7=1   9     1.55     3.36    0.9681      1.45
 Chisq= 1.5  on 1 degrees of freedom, p= 0.228
```

Análise de Sobrevivência Aplicada à Saúde -- 2014

√ V8: Renda per capita

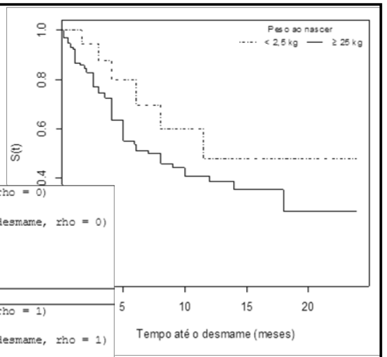



```
> survdiff(Surv(tempo, cens)~ V8, data = desmame, rho = 0)
Call:
survdiff(formula = Surv(tempo, cens) ~ V8, data = desmame, rho = 0)
      N Observed Expected (O-E)^2/E (O-E)^2/V
V8=0 77       29     34.7     0.932      2.11
V8=1 73       36     30.3     1.066      2.11
 Chisq= 2.1  on 1 degrees of freedom, p= 0.146
```

```
> survdiff(Surv(tempo, cens)~ V8, data = desmame, rho = 1)
Call:
survdiff(formula = Surv(tempo, cens) ~ V8, data = desmame, rho = 1)
      N Observed Expected (O-E)^2/E (O-E)^2/V
V8=0 77     21.2     25.7     0.796      2.22
V8=1 73     27.6     23.1     0.884      2.22
 Chisq= 2.2  on 1 degrees of freedom, p= 0.136
```

Análise de Sobrevivência Aplicada à Saúde -- 2014

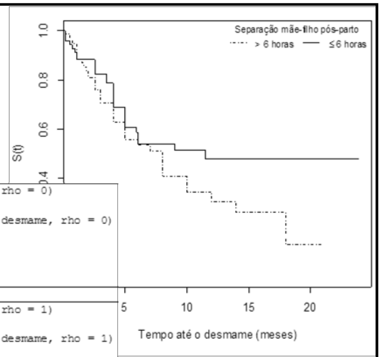
√ V9:Peso ao nascer




```
> survdiff(Surv(tempo, cens)~ V9, data = desmame, rho = 0)
Call:
survdiff(formula = Surv(tempo, cens) ~ V9, data = desmame, rho = 0)
        N Observed Expected (O-E)^2/E (O-E)^2/V
V9=0 131       59    55.14     0.271      1.87
V9=1  19        6     9.86     1.513      1.87
 Chisq= 1.9  on 1 degrees of freedom, p= 0.171
```

```
> survdiff(Surv(tempo, cens)~ V9, data = desmame, rho = 1)
Call:
survdiff(formula = Surv(tempo, cens) ~ V9, data = desmame, rho = 1)
        N Observed Expected (O-E)^2/E (O-E)^2/V
V9=0 131    44.88    41.57     0.264      2.38
V9=1  19     3.93     7.25     1.515      2.38
 Chisq= 2.4  on 1 degrees of freedom, p= 0.123
```


√ V10: Tempo de separação mãe-filho

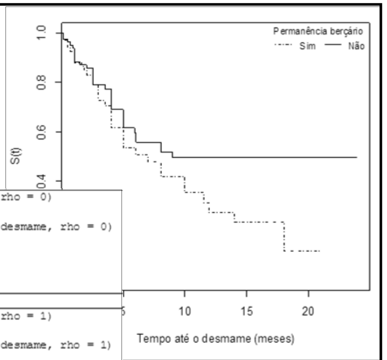



```
> survdiff(Surv(tempo, cens)~ V10, data = desmame, rho = 0)
Call:
survdiff(formula = Surv(tempo, cens) ~ V10, data = desmame, rho = 0)
        N Observed Expected (O-E)^2/E (O-E)^2/V
V10=0  69       27     33.3      1.20       2.6
V10=1  81       38     31.7      1.26       2.6
 Chisq= 2.6  on 1 degrees of freedom, p= 0.107
```

```
> survdiff(Surv(tempo, cens)~ V10, data = desmame, rho = 1)
Call:
survdiff(formula = Surv(tempo, cens) ~ V10, data = desmame, rho = 1)
        N Observed Expected (O-E)^2/E (O-E)^2/V
V10=0  69     20.8     24.6     0.592      1.58
V10=1  81     28.1     24.2     0.601      1.58
 Chisq= 1.6  on 1 degrees of freedom, p= 0.208
```

Análise de Sobrevivência Aplicada à Saúde -- 2014

√ V11: Permanência no berçário




```
> survdiff(Surv(tempo, cens)~ V11, data = desmame, rho = 0)
Call:
survdiff(formula = Surv(tempo, cens) ~ V11, data = desmame, rho = 0)
        N Observed Expected (O-E)^2/E (O-E)^2/V
V11=0  81       31     37.6      1.17      2.93
V11=1  69       34     27.4      1.61      2.93
 Chisq= 2.9  on 1 degrees of freedom, p= 0.0868
```

```
> survdiff(Surv(tempo, cens)~ V11, data = desmame, rho = 1)
Call:
survdiff(formula = Surv(tempo, cens) ~ V11, data = desmame, rho = 1)
        N Observed Expected (O-E)^2/E (O-E)^2/V
V11=0  81     24.1     28.0     0.542      1.67
V11=1  69     24.8     20.9     0.726      1.67
 Chisq= 1.7  on 1 degrees of freedom, p= 0.196
```

Análise de Sobrevivência Aplicada à Saúde -- 2014

√ Tabela de valores dos testes logrank e Peto

Variável				*logrank*		Peto	
Código	Descrição	Valor	Qte.	χ^2	p-valor	χ^2	p-valor
V1	Experiência anterior amamentação	0	87	3,9	0,047	5,7	0,017
		1	63				
V2	Número de filhos vivos	0	113	2,6	0,107	2,4	0,122
		1	37				
V3	Conceito materno sobre o tempo ideal de amamentação	0	80	6,1	0,013	7,9	0,005
		1	70				
V4	Dificuldade para amamentar nos primeiros dias pós-parto	0	84	12,3	0,000	14,9	0,000
		1	66				
V5	Tipo de serviço em que realizou o pré-natal	0	122	1,4	0,241	1,3	0,257
		1	28				
V6	Recebeu exclusivamente leite materno na matemidade	0	94	7,5	0,006	7,0	0,008
		1	56				
V7	A criança teve contato com o pai	0	141	1,8	0,175	1,5	0,228
		1	9				
V8	Renda per capita (em SM/mês)	0	77	2,1	0,146	2,2	0,136
		1	73				
V9	Peso ao nascimento	0	131	1,9	0,171	2,4	0,123
		1	19				
V10	Tempo de separação mãe-filho pós-parto	0	69	2,6	0,107	1,6	0,208
		1	81				
V11	Permanência no berçário	0	81	2,9	0,087	1,7	0,196
		1	69				

– p-valores menores que 0,25

Análise de Sobrevivência Aplicada à Saúde -- 2014

- Comentários:
  - √ Há 11 covariáveis potencialmente importantes
    - $2^{11} = 2048$ modelos possíveis
  - √ É impraticável ajustar todos esses possíveis modelos para selecionar aquele que explica melhor a resposta
  - √ Há procedimentos automáticos disponíveis em pacotes estatísticos:

- Procedimentos automáticos:
  - √ Tipos:
    - Forward
    - Backward
    - Stepwise
  - √ Vantagem
    - Métodos implementados e disponíveis em pacotes
  - √ Desvantagens:
    - Tendem a identificar um particular conjunto de covariáveis em vez de possíveis conjuntos igualmente bons
    - Impede que dois ou mais conjuntos igualmente bons sejam apresentados ao pesquisador para a escolha do mais relevante em sua área de aplicação

- Seleção de covariáveis:
  - √ Covariáveis com significância clínica devem ser incluídas independente de significância estatística
  - √ Importância clínica deve ser considerada em cada passo de inclusão ou exclusão no processo de seleção de covariáveis
  - √ Evitar ser muito rigoroso ao testar cada nível individual de significância
    - Recomenda-se um valor próximo de 0,10

- Modelagem separada de cada covariável com a resposta
  - √ Objetivo:
    - selecionar quais covariáveis prosseguirão na análise
  - √ Critério utilizado:
    - Permanecer com covariáveis que apresentarem p-valor inferior a 0,25
  - √ Todas as covariáveis passam por esse critério

- Passos da estratégia para seleção de covariáveis:

  (Colosimo e Giolo, 2006)

  1. Ajustar todos os modelos contendo uma única covariável
     - Incluir todas as covariáveis que forem significativas ao nível de 0,10
     - Aconselhável usar o Teste da Razão de Verossimilhanças

- Teste de Razão de Verossimilhanças é usualmente realizado utilizando-se a distribuição gama generalizada.
  - √ Tem aninhados os modelos exponencial, de Weibull, lognormal e gama
- Para ajuste do modelo gama generalizado no R:

```
> library(flexsurv)
> flexsurvreg(formula, data,
              dist='gengamma')
```


- Ajuste de modelos com uma única covariável

Passo 1

Modelo		$l(\theta)$	$-2l(\theta)$	TRV	p-valor
	Nulo	-228,226	456,452		
V1	Experiência anterior amamentação	-225,579	451,158	5,294	0,0214
V2	Número de filhos vivos	-226,819	453,638	2,814	0,0935
V3	Conceito materno sobre o tempo ideal de amamentação	-225,330	450,659	5,793	0,0161
V4	Dificuldade para amamentar nos primeiros dias pós-parto	-221,811	443,621	12,831	0,0003
V5	Tipo de serviço em que realizou o pré-natal	-227,335	454,670	1,782	0,1819
V6	Recebeu exclusivamente leite materno na maternidade	-224,718	449,436	7,016	0,0081
V7	A criança teve contato com o pai	-227,101	454,202	2,250	0,1336
V8	Renda per capita (em SM/mês)	-226,736	453,473	2,979	0,0843
V9	Peso ao nascimento	-226,753	453,506	2,946	0,0861
V10	Tempo de separação mãe-filho pós-parto	-227,255	454,511	1,942	0,1635
V11	Permanência no berçário	-227,181	454,362	2,090	0,1483

Covariáveis retidas:
V1, V2, V3, V4, V6, V8, V9

Covariáveis excluídas:
V5, V7, V10, V11

2. Ajustar conjuntamente as covariáveis significativas no passo 1
   - Na presença de certas covariáveis , outras podem deixar de ser significativas
   - Ajustam-se modelos reduzidos, excuindo uma única covariável de cada vez
   - Permanecem no modelo somente aquelas que atingirem a significância

- Ajuste conjunto das covariáveis significativas
  - √ Exclusão de uma variável por vez
  - √ Permanecem apenas aquelas que atingirem significância

Passo 2

Modelo		$l(\theta)$	$-2l(\theta)$	TRV	p-valor
V1 + V2 + V3 + V4 + V6 + V8 + V9		-212,475	424,950		
V1+V2+V3+V4+V6+V8		-213,129	426,258	1,307	0,2529
V1+V2+V3+V4+V6+V9	V8	-214,213	428,426	3,475	0,0623
V1+V2+V3+V4+V8+V9	V6	-215,196	430,391	5,441	0,0197
V1+V2+V3+V6+V8+V9	V4	-216,698	433,397	8,447	0,0037
V1+V2+V4+V6+V8+V9	V3	-214,155	428,310	3,360	0,0668
V1+V3+V4+V6+V8+V9		-212,539	425,077	0,127	0,7219
V2+V3+V4+V6+V8+V9		-213,100	426,199	1,249	0,2638

Covariáveis retidas:
V3, V4, V6, V8

Covariáveis excluídas:
V1, V2, V9

3. Ajustar modelo com covariáveis retidas no passo 2
   - As covariáveis excluídas no passo 2 retornam ao modelo para confirmar sua significância estatística

- Ajuste de modelo com covariáveis retidas
  - √ Retorno de covariável excluída uma a uma

Passo 3

Modelo	$l(\theta)$	$-2l(\theta)$	TRV	p-valor
V3+V4+V6+V8	-212,475	428,397		
V3+V4+V6+V8+V1	-213,221	426,442	1,955	0,1620
V3+V4+V6+V8+V2	-213,635	427,269	1,127	0,2884
V3+V4+V6+V8+V9	-213,647	427,294	1,103	0,2936

Covariáveis retidas:
V3, V4, V6, V8

  - √ As covariáveis excluídas no passo anterior não são significativas

4. Inclusão no modelo de eventuais covariáveis significativas no passo 3
   - As covariáveis excluídas no passo 1 retornam ao modelo para confirmar sua significância estatística

- Ajuste de modelo com covariáveis retidas nos passos 2 e 3
  - √ Retorno de covariáveis excluídas no passo 1

Passo 4

Modelo	$l(\theta)$	$-2l(\theta)$	TRV	p-valor
V3+V4+V6+V8	-212,475	428,397		
V3+V4+V6+V8+V5	-214,198	428,397	0,000	0,9989
V3+V4+V6+V8+V7	-213,319	426,637	1,759	0,1847
V3+V4+V6+V8+V10	-214,072	428,143	0,253	0,6146
V3+V4+V6+V8+V11	-214,117	428,234	0,163	0,6863

Covariáveis retidas: V3, V4, V6, V8

  - √ As covariáveis excluídas no passo 1 não são significativas

5. Inclusão no modelo de eventuais covariáveis significativas no passo 4
   – Testa-se se alguma dela pode ser retirada do modelo

- Inclusão de covariáveis significativas no passo 4
  - √ Testa-se se alguma delas pode ser retirada do modelo

Passo 5

Modelo	$l(\theta)$	$-2l(\theta)$	TRV	p-valor
V3+V4+V6+V8	-212,475	428,397		
V3+V4+V6	-216,955	433,911	5,514	0,0189
V3+V4+V8	-216,747	433,493	5,097	0,0240
V3+V6+V8	-220,753	441,506	13,109	0,0003
V4+V6+V8	-216,109	432,218	3,822	0,0506

Covariáveis retidas: V3, V4, V6, V8

  - √ Todas as covariáveis incluídas nos passos anteriores são significativas

6. Ajusta-se modelo final para os efeitos principais utilizando as covariáveis que não foram excluídas no passo 5
   – Verifica-se a possibilidade de inclusão de termos de interação dupla entre as covariáveis incluídas no modelo

- Modelo final para os efeitos principais
  - √ Verificação inclusão de interações duplas

Passo 6

Modelo	$l(\theta)$	$-2l(\theta)$	TRV	p-valor
V3+V4+V6+V8	-212,475	428,397		
V3+V4+V6+V8+V3*V4	-213,844	427,689	0,708	0,4002
V3+V4+V6+V8+V3*V6	-213,295	426,590	1,807	0,1789
V3+V4+V6+V8+V3*V8	-214,059	428,118	0,279	0,5973
V3+V4+V6+V8+V4*V6	-213,823	427,647	0,750	0,3865
V3+V4+V6+V8+V4*V8	-213,826	427,652	0,744	0,3883
V3+V4+V6+V8+V6*V8	-214,056	428,112	0,284	0,5939

Covariáveis retidas:
V3, V4, V6, V8

  - √ Nenhum termo de interação dupla foi significativo

- Modelo Final:
  - √ Efeitos principais identificados no passo 5
    - V3, V4, V6, V8
  - √ Termos de interação identificados no passo 6
    - Nenhum é significativo

- Modelo Final:

	l(θ)	–2l(θ)	TRV	gl	p-valor
Nulo	–228,226	456,452			
V3+V4+V6+V8	–214,198	428,397	28,056	4	0,000
V3+V4+V6+V1	–215,228	430,456	25,996	4	0,000

  - √ Possível efeito de multicolienaridade entre:
    - V1: experiência anterior de amamentação
    - V8: renda per capita

    (estatísticas de teste não são muito diferentes)
  - √ Decisão sobre qual variável deveria permanecer foi baseada em evidências clínicas

- Variáveis prognósticas:
  - √ V1:
    - Experiência anterior de amamentação
  - √ V3:
    - Conceito materno sobre o tempo ideal de amamentação
  - √ V4:
    - Dificuldades de amamentar nos primeiros dias pós-parto
  - √ V6:
    - Recebimento exclusivo de leite materno na maternidade

- Saídas dos modelos Weibull e Lognormal

```
> desm.wei <- survreg(Surv(desmame$tempo, desmame$cens)~ V3 + V4 + V6 + V1,
+ dist = 'weibull')
> desm.wei
Call:
survreg(formula = Surv(desmame$tempo, desmame$cens) ~ V3 + V4 +
    V6 + V1, dist = "weibull")
Coefficients:
(Intercept)          V3          V4          V6          V1
  3.5871202  -0.4926820  -0.6820891  -0.6521603  -0.5336577
Scale= 1.048537
Loglik(model)= -217.9   Loglik(intercept only)= -229.2
        Chisq= 22.57 on 4 degrees of freedom, p= 0.00015
n= 150
```

```
> desm.ln <- survreg(Surv(desmame$tempo, desmame$cens)~ V3 + V4 + V6 + V1,
+ dist = 'lognorm')
> desm.ln
Call:
survreg(formula = Surv(desmame$tempo, desmame$cens) ~ V3 + V4 +
    V6 + V1, dist = "lognorm")
Coefficients:
(Intercept)          V3          V4          V6          V1
  3.2925917  -0.6314176  -0.8242242  -0.6804915  -0.5721767
Scale= 1.438643
Loglik(model)= -215.3   Loglik(intercept only)= -228.9
        Chisq= 27.21 on 4 degrees of freedom, p= 1.8e-05
n= 150
```


- Busca-se modelo mais simples que o gama generalizada

	l(θ)	–2l(θ)	TRV	gl	p-valor
Gama generalizada	–215,228	430,456			
Weibull	–217,902	435,803	5,347	1	0,0208
Lognormal	–215,337	430,673	0,218	1	0,6409

√ H_0: o modelo de interesse é adequado

√ Conclusão:

- Modelo de regressão lognormal é adequado para ajustar os tempos até o desmame
- Variáveis prognósticas: V1, V3, V4, V6

- Estimativas dos parâmetros do modelo de regressão lognormal

√ Coeficientes estimados expressos na escala logarítmica dos tempos

$$Y = \ln(T) = \boldsymbol{x}'\boldsymbol{\beta} + \sigma\nu$$

```
> desm.ln <- survreg(Surv(desmame$tempo, desmame$cens)~ V3 + V4 + V6 + V1,
+ dist = 'lognorm')
> summary(desm.ln)
Call:
survreg(formula = Surv(desmame$tempo, desmame$cens) ~ V3 + V4 +
    V6 + V1, dist = "lognorm")
             Value Std. Error     z        p
(Intercept)  3.293      0.304 10.84 2.29e-27
V3          -0.631      0.290 -2.18 2.93e-02
V4          -0.824      0.302 -2.73 6.33e-03
V6          -0.680      0.293 -2.33 2.00e-02
V1          -0.572      0.301 -1.90 5.72e-02
Log(scale)   0.364      0.090  4.04 5.36e-05
Scale= 1.44
Log Normal distribution
Loglik(model)= -215.3   Loglik(intercept only)= -228.9
        Chisq= 27.21 on 4 degrees of freedom, p= 1.8e-05
Number of Newton-Raphson Iterations: 4
```

$$\begin{aligned} \mathrm{ep}(\hat{\sigma}) &= \mathrm{ep}[\ln(\hat{\sigma})]\hat{\sigma} \\ &= (0,090)(1,44) \\ &= 0,1295 \end{aligned}$$

- Estimativa dos parâmetros:

	Covariável	Estimativa	Erro Padrão	p-valor
	Intercepto	3,293	0,304	0,000
V1	Experiência anterior de amamentação	–0,572	0,301	0,0572
V3	Conceito sobre o tempo de amamentação	–0,631	0,290	0,0293
V4	Dificuldades de amamentação pós-parto	–0,824	0,302	0,0063
V6	Recebimento exclusivo de leite materno	–0,680	0,293	0,0020
	Parâmetro de forma	1,439	0,129	

Adequação do Modelo Ajustado

- Métodos gráficos:
  - √ Utilização dos resíduos para confirmar a adequação do modelo lognormal
- Tipos de resíduos:
  - √ Resíduos padronizados
  - √ Resíduos de Cox-Snell

- Resíduos padronizados
  - √ Baseado na representação dos modelos log-lineares

$$\hat{\nu}_i = \frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\hat{\sigma}}, \text{ com } y_i = \ln(t_i)$$

  - √ Modelo lognormal bem ajustado para os dados:
    - Resíduos têm distribuição próxima à normal padrão
    - Exponencial dos resíduos tem distribuição lognormal
  - √ Procedimento:
    - Estimar sobrevivência dos resíduos com Kaplan-Meier
    - Comparar estimação de Kaplan-Meier com sobrevivência da lognormal padrão

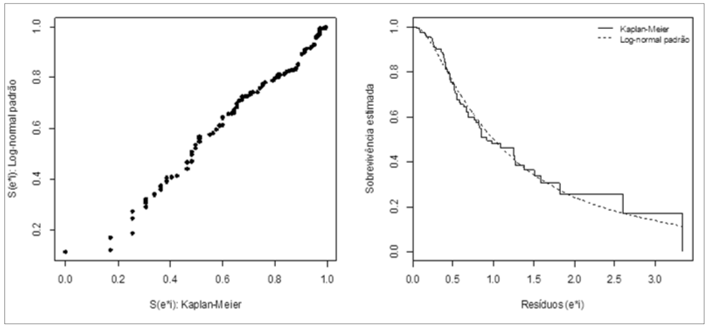
- Cálculo e transformação dos resíduos:

```
> # Determinacao Residuos e*i (log-normal padrão)
>
> residuos <- predict(desm.ln, type = "response")    # Escala original
> padr.log <-(log(tempo)-log(residuos))/desm.ln$scale # Padroniza na escala log
> padr.orig<-exp(padr.log)    # Escala original
> padr.km <- survfit(Surv(padr.orig, cens)~ 1)
> res <- padr.km$time
> sln <- pnorm(-log(res))
```

- Gráficos dos resíduos padronizados:

```
> # Gráficos Residuos e*i
>
> win.graph(width=12, height=6) # Default width = 7, height = 7
> par(mfrow = c(1,2))
> plot(padr.km$surv, sln, pch = 16, ylab = "S(e*i): Log-normal padrão",
+ xlab = "S(e*i): Kaplan-Meier")
> plot(padr.km, conf.int = F, mark.time = F, pch = 16,
+ xlab = "Resíduos (e*i)", ylab = "Sobrevivência estimada")
> lines(res, sln, lty = 2)
> legenda = c("Kaplan-Meier", "Log-normal padrão")
> legend(x = "topright", legend = legenda, lty = 1:2, cex = 0.8, bty = "n")
> par(mfrow = c(1,1))
```


- √ Sobrevivência dos resíduos e^*_i:
  - Estimativa Kaplan-Meier e Log-normal padrão



  - Pode-se acreditar que o modelo de regressão lognormal se encontra bem ajustado aos dados sob análise

√ Resíduos de Cox-Snell

$$\boxed{\hat{e}_i = \hat{\Lambda}(t_i|x_1) = -\ln[\hat{S}(t_i|x_1)]}$$

√ Modelos:

- Exponencial: $\boxed{\hat{e}_i = \left[t_i \exp\{\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\}\right]}$
- Weibull: $\boxed{\hat{e}_i = \left[t_i \exp\{\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\}\right]^{\hat{\gamma}}}$
- Lognormal: $\boxed{\hat{e}_i = -\ln\left[1 - \Phi\left(\frac{\ln(t_i) - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)\right]}$

- Se o modelo for adequado:
  √ Resíduos seguem distribuição exponencial padrão
  - Curvas de sobrevivência dos resíduos, por Kaplan-Meier e exponencial padrão devem estar próximas
  - Gráfico dos pontos (S_{KM}(res), S_{exp}(res)) devem se aproximar de uma reta
  - Gráfico e_i vs. $\Lambda(e_i)$ deve ser aproximadamente uma reta

- Cálculo dos resíduos de Cox-Snell:

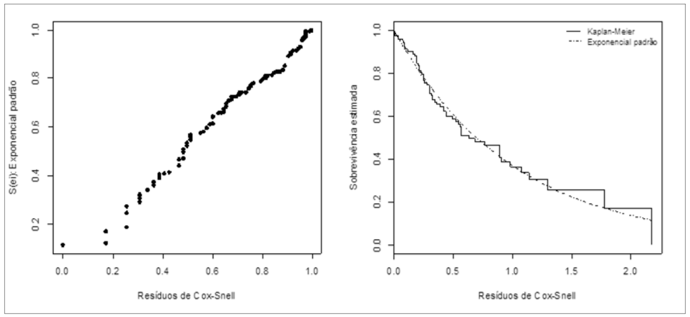
```
> # Determinação Resíduos Cox-Snell
>
> snell <- - log(1 - pnorm(padr.log))
> snell.km <- survfit(Surv(snell, cens)~ 1)
> t <- snell.km$time
> s.t <- snell.km$surv
> s.exp <- exp(-t)
```

- Gráficos dos resíduos padronizados:

```
> # Gráficos Resíduos Cox-Snell
>
> win.graph(width=12, height=6) # Default width = 7, height = 7
> par(mfrow = c(1,2))
> plot(s.t, s.exp, pch = 16, ylab = "S(ei): Exponencial padrão",
+ xlab = "Resíduos de Cox-Snell")
> plot(snell.km, conf.int = F, mark.time = F, pch = 16,
+ xlab = "Resíduos de Cox-Snell", ylab = "Sobrevivência estimada")
> lines(t, s.exp, lty = 4)
> legenda = c("Kaplan-Meier", "Exponencial padrão")
> legend(x = "topright", legend = legenda, lty = c(1, 4), cex = 0.8, bty = "n")
> par(mfrow = c(1,1))
```


√ Sobrevivência dos resíduos de Cox-Snell (e_i):

- Estimativa Kaplan-Meier e Exponencial padrão



- Evidências de que o modelo de regressão lognormal se encontra bem ajustado aos dados sob análise

Interpretação dos Coeficientes

- Para variável binária
  - √ $\exp\{\beta_i\}$ é a razão dos tempos medianos:

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

  - √ Válido para os modelos Weibull e lognormal

- No exemplo
  - √ Efeito da experiência anterior de amamentação

0: sim
1: não

$$\boxed{e^{\hat{\beta}_1} = e^{-0,572} = 0,564}$$

    - Tempo mediano até o desmame de mães que não tiveram experiência anterior de amamentação é aproximadamente a metade daquele das mães que já tiveram esta experiência
  - √ Efeito do conceito sobre o tempo de amamentação:

0: > 6 meses
1: ≤ 6 meses

$$\boxed{e^{\hat{\beta}_2} = e^{-0,631} = 0,532}$$

    - Mães que acreditam que o tempo ideal de desmame é superior a 6 meses apresentam tempo mediano até o desmame de aproximadamente 2 vezes maior do que as mães que pensam ser este tempo igual ou inferior a 6 meses

  - √ Efeito da dificuldade de amamentação pós-parto

0: sim
1: não

$$\boxed{e^{\hat{\beta}_3} = e^{-0,824} = 0,439}$$

    - Tempo mediano até o desmame de mães que não apresentaram dificuldades de amamentar nos primeiros dias até o parto é 2,3 vezes maior que o tempo das que sofreram este tipo de dificuldade.
  - √ Efeito do recebimento exclusivo de leite materno:

0: sim
1: não

$$\boxed{e^{\hat{\beta}_4} = e^{-0,680} = 0,507}$$

    - Crianças que receberam exclusivamente leite materno na maternidade têm um tempo mediano de amamentação duas vezes maior do que o tempo daqueles que receberam outro tipo de alimentação juntamente com o leite materno

Referências

Bibliografia

- Carvalho, M. S. et al. *Análise de Sobrevivência: Teoria e Aplicações em Saúde*. (Fiocruz)
- Colosimo, E. A. e Giolo, S. R. *Análise de Sobrevivência Aplicada*. (Edgard Blucher)
- Klein, J. P. e Moeschberger, M. L. *Survival Analysis: Techniques for Censored and Truncated Data*. (Springer)
- Kleinbaum, D. G. *Survival Analysis: a Self-Learning Text*

Material de Apoio

- R:
  - √ www.r-project.org
- Tutorial online do R:
  - √ http://www.leg.ufpr.br/Rtutorial
  - √ http://www.leg.ufpr.br/~paulojus/embrapa/Rembrapa
- Conjuntos de dados e material Análise de Sobrevivência – Carvalho et al.
  - √ http://sobrevida.fiocruz.br