Especialização em *Data Science* e Estatística Aplicada

Módulo IV - Análise de Sobrevivência

Prof. Dr. Eder Angelo Milani

Goiânia, 2025

IME
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
FEN
FACULDADE DE ENFERMAGEM
UFG
UNIVERSIDADE FEDERAL DE GOIÁS

Conteúdo Programático

- Conceitos básicos. *(Aula 1)*
- Técnicas não-paramétricas. *(Aula 1)*
- Modelos probabilísticos em análise de sobrevivência.
- Modelos de regressão paramétrico.
- Modelo semiparamétrico de riscos proporcionais de Cox.
- Métodos para verificação do modelo ajustado.
- Modelo de Cox estratificado.

Conteúdo - Aula 2

1. Modelos probabilísticos em análise de sobrevivência

- Introdução
- Modelos paramétricos
  - Modelo Exponencial
  - Modelo Weibull
  - Modelo log-normal
- Estimação dos modelos paramétricos
- Adequação do modelo probabilístico

2. Modelos de Regressão Paramétricos

- Modelo de regressão exponencial
- Modelo de regressão paramétrico - resumo
- Adequação do modelo ajustado
  - Resíduo de Cox-Snell
- Interpretação dos coeficientes estimados

Introdução

O objetivo agora é apresentar o uso de distribuições de probabilidade na análise estatística de dado de sobrevivência.

Em geral, os métodos paramétricos fazem suposições fortes sobre a distribuição dos dados (por exemplo, normalidade), o que permite estimativas mais precisas com menos dados. Isso significa que, para um mesmo tamanho amostral, um modelo paramétrico pode ter menor variância nas estimativas dos parâmetros.

A partir de agora vamos assumir que o tempo (T) até a falha segue uma distribuição conhecida de probabilidade.

Modelos paramétricos

Embora exista uma série de modelos probabilísticos utilizados em análise de dados de sobrevivência, alguns deles ocupam uma posição de destaque por sua comprovada adequação a várias situações práticas. Entre estes modelos, é possível citar o exponencial, o Weibull e o log-normal.

Modelo Exponencial

Em termos matemáticos, a **distribuição exponencial é um dos modelos probabilísticos mais simples** usados para descrever o tempo de falha.

Esta distribuição apresenta um único parâmetro e é a única que se caracteriza por ter uma **função taxa de falha constante**.

> **Principal propriedade do modelo exponencial**
> A distribuição exponencial apresenta taxa de falha constante. Isto significa que tanto uma unidade velha quanto uma nova, que ainda não falharam, apresentam a mesma taxa de falha em um intervalo futuro. Esta propriedade é chamada de falta de memória da distribuição exponencial.

Modelo Exponencial

- A função de densidade exponencial é definida como:

$$f(t) = \frac{1}{\alpha} \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}, \qquad t \geq 0 \quad \text{e} \quad \alpha > 0;$$

- A função de sobrevivência é:

$$S(t) = 1 - F(t) = 1 - \left[1 - \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}\right] = \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}, \qquad t \geq 0 \quad \text{e} \quad \alpha > 0;$$

Importante: $S(0) = 1$ e $S(\infty) = 0$.

Modelo Exponencial

- A função de risco é:

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

- MTTF $= \int_0^\infty S(t)dt = \int_0^\infty \exp\left\{-\left(\frac{t}{\alpha}\right)\right\} dt = -\alpha \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}\Big|_0^\infty = \alpha.$

Importante: Função de risco constante e MTTF igual ao parâmetro da distribuição.

Modelo Exponencial

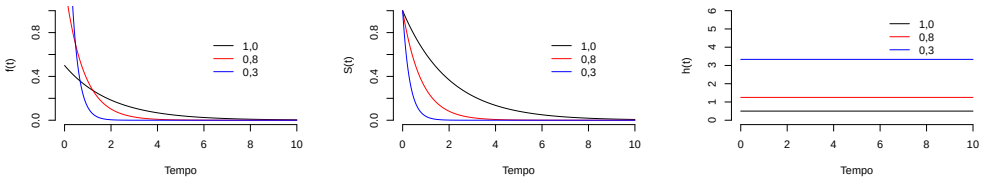

Figura 1: Forma típica das funções de densidade de probabilidade, de sobrevivência e de taxa de falha do modelo exponencial, para diferente valores do parâmetro α.

Modelo Weibull

Principal propriedade do modelo Weibull

A sua popularidade em aplicações práticas se deve ao fato dela apresentar uma grande variedade de formas, todas com uma propriedade básica: a sua função de taxa de falha é monótona, isto é, ela é crescente, decrescente ou constante.

Modelo Weibull

- A função de densidade da Distribuição Weibull é dada por:

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \exp\left[-\left(\frac{t}{\alpha}\right)^{\gamma}\right], \qquad t \geq 0 \quad \text{e} \quad \gamma, \alpha > 0;$$

- A função de sobrevivência é:

$$S(t) = \exp\left[-\left(\frac{t}{\alpha}\right)^{\gamma}\right], \qquad t \geq 0 \quad \text{e} \quad \gamma, \alpha > 0;$$

Importante: $S(0) = 1$ e $S(\infty) = 0$.

Modelo Weibull

- A função de risco é:

$$\lambda(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1}, \qquad t \geq 0 \quad \text{e} \quad \gamma, \alpha > 0;$$

- MTTF $= \alpha\Gamma\left(1 + \frac{1}{\gamma}\right)$.

Importante: Função de risco é monótona e o MTTF envolve os parâmetros do modelo e a função gama - $\Gamma(.)$.

Modelo Weibull

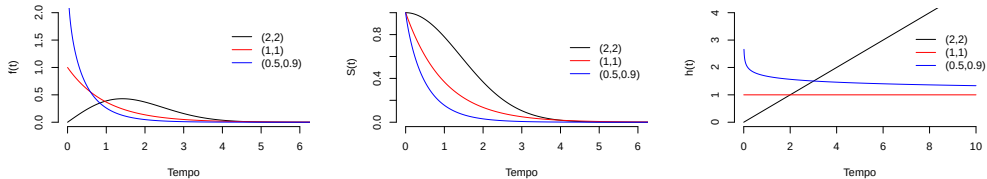

Figura 2: Forma típica das funções de densidade de probabilidade, de sobrevivência e de taxa de falha do modelo Weibull, para diferente valores dos parâmetros (α, γ).

Modelo Weibull

Observações

- Quando $\gamma = 1$, tem-se a distribuição exponencial, logo a distribuição exponencial é um caso particular da distribuição Weibull
- A função taxa de falha $\lambda(t)$ é estritamente crescente para $\gamma > 1$, estritamente decrescente para $\gamma < 1$ e constante para $\gamma = 1$.

Modelo log-normal

Assim como a distribuição Weibull, a distribuição log-normal é muito utilizada para caracterizar tempos de vida de produtos e indivíduos.

Como o nome sugere, o logaritmo de uma variável com distribuição log-normal de parâmetros μ e σ tem distribuição normal com média μ e desvio-padrão σ.

As funções de sobrevivência e de taxa de falha de uma variável log-normal não apresentam uma forma analítica explícita.

Modelo log-normal

- A função de densidade da distribuição log-normal é dada por:

$$f(t) = \frac{1}{\sqrt{2\pi}t\sigma} \exp\left\{-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right\} \qquad t \geq 0, \quad \mu \in \mathbb{R}, \quad \text{e} \quad \sigma > 0;$$

- A função de sobrevivência é:

$$S(t) = \Phi\left(\frac{-\log(t)+\mu}{\sigma}\right) \qquad t \geq 0, \quad \mu \in \mathbb{R}, \quad \text{e} \quad \sigma > 0;$$

Obs.: $\Phi(.)$ é a função distribuição acumulada da distribuição normal padrão e z_p é o $100p$% percentil da distribuição normal padrão.

Importante: $S(0) = 1$ e $S(\infty) = 0$.

Modelo log-normal

- A função de risco é:

$$\lambda(t) = \frac{f(t)}{S(t)};$$

- MTTF $= \exp(\mu + \sigma^2/2)$.

Importante: Função de risco não é monótona e o cálculo do MTTF é simples de ser feito.

Modelo log-normal

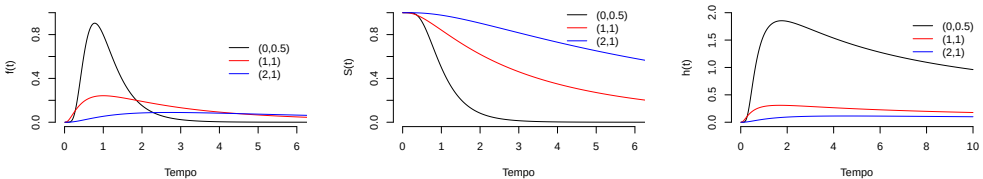

Figura 3: Forma típica das funções de densidade de probabilidade, de sobrevivência e de taxa de falha do modelo log-normal, para diferente valores dos parâmetros (μ, σ).

Modelo log-normal

Observação
Observe que as funções de taxa de falha não são monótonas como as da distribuição Weibull. Elas crescem, atingem um valor máximo e depois decrescem.

Outros modelos

Observação

Existem outras distribuições de probabilidade apropriadas para modelar o tempo de falha de produtos, materiais e situações clínicas. Dentre elas, podem ser citadas as distribuições gama, log-gama, Rayleigh, normal inversa, Gompertz, Birnbaum-Saunders, entre outras.

Estimação dos modelos paramétricos

Os modelos probabilísticos apresentados são caracterizados por quantidades desconhecidas, denominadas parâmetros. Os modelos Weibull e log-nomral são caracterizados por dois parâmetros, e o exponencial por apenas um.

Estas quantidades conferem uma forma geral aos modelos probabilísticos. Entretanto, **em cada estudo envolvendo tempos de falha, os parâmetros devem ser estimados a partir das observações amostrais, para que o modelo fique determinado** e, assim, seja possível responder às perguntas de interesse.

Aplicação

Aplicação no *software* **R.**

Adequação do modelo probabilístico

A escolha do modelo a ser utilizado é um tópico importante na análise paramétrica de dados de tempo de vida.

Se um modelo for usado inadequadamente para um certo conjunto de dados, toda a análise estatística fica comprometida e consequentemente as respostas às perguntas de interesse ficam distorcidas.

Mas por que usar o modelo log-normal e não o de Weibull?

Adequação do modelo probabilístico

A escolha do modelo a ser utilizado é um tópico importante na análise paramétrica de dados de tempo de vida.

Se um modelo for usado inadequadamente para um certo conjunto de dados, toda a análise estatística fica comprometida e consequentemente as respostas às perguntas de interesse ficam distorcidas.

Mas por que usar o modelo log-normal e não o de Weibull? Algumas vezes existem evidências provenientes de teste realizados no passado de que um certo modelo se ajusta bem aos dados. No entanto, em muitas situações, este tipo de informação não se encontra disponível. A solução para esta situação é basicamente empírica.

Adequação do modelo probabilístico

A proposta empírica consiste em ajustar os modelos probabilísticos e, com base na **comparação entre valores estimados e observados, decidir qual deles "melhor" explica os dados amostrais**.

A forma mais simples e eficiente de selecionar o "melhor" modelo a ser utilizado para um conjunto de dados é por meio de técnicas gráficas.

Testes de hipóteses com modelos encaixados também podem ser utilizados para esta finalidade.

Adequação do modelo probabilístico

A proposta empírica consiste em ajustar os modelos probabilísticos e, com base na **comparação entre valores estimados e observados, decidir qual deles "melhor" explica os dados amostrais**.

A forma mais simples e eficiente de selecionar o "melhor" modelo a ser utilizado para um conjunto de dados é por meio de técnicas gráficas.

Testes de hipóteses com modelos encaixados também podem ser utilizados para esta finalidade.

Segundo George Box, *"Essencialmente, todos os modelos estão errados, mas alguns são úteis"*.

Adequação do modelo probabilístico

O primeiro método gráfico a ser utilizado consiste na comparação da função de sobrevivência do modelo proposto com o estimador Kaplan-Meier, seguindo os seguintes passos:

(i) Ajustam-se os modelos propostos ao conjunto de dados, por exemplo, os modelos log-normal e Weibull;

(ii) A partir da estimativa dos parâmetros dos modelos propostos, estimar a função de sobrevivência, por exemplo: $\widehat{S}_{ln}(t)$ e $\widehat{S}_{W}(t)$, respectivamente, para os modelos log-normal e Weibull;

Adequação do modelo probabilístico

(iii) Estimar a função de sobrevivência utilizando o estimador de Kaplan-Meier ($\widehat{S}(t)$);

(iv) Comparar graficamente as funções de sobrevivência estimadas para cada modelo proposto com $\widehat{S}(t)$. O(s) modelo(s) adequado é aquele em que sua curva de sobrevivência mais se aproximar daquela do estimador de Kaplan-Meier.

Adequação do modelo probabilístico

Na prática, pode ser feito os gráficos

- $\widehat{S}(t)$ versus $\widehat{S}_{ln}(t)$ e $\widehat{S}(t)$ versus $\widehat{S}_W h(t)$, assim o “melhor” modelo é aquele cujos pontos no gráfico estiverem mais próximos da reta $x = y$, com $x = \widehat{S}(t)$ e $y = \widehat{S}_{ln}(t)$, por exemplo.

- $\widehat{S}(t)$ versus t, juntamente com a curva $\widehat{S}_{ln}(t)$ versus t. O melhor “modelo” é o que apresentar curva mais próxima da função de sobrevivência estimada por Kaplan-Meier.

Aplicação

Aplicação no *software* **R.**

Introdução

Os estudos na área médica muitas vezes envolvem **covariáveis que podem estar relacionadas com o tempo de sobrevivência**.

Tais covariáveis devem ser incluídas na análise estatística dos dados.

Uma forma simples de usar as covariáveis é **dividir os dados em estratos de acordo com estas covariáveis e usar as técnicas não-paramétricas** apresentadas anteriormente.

A forma mais eficiente de acomodar o efeito dessas covariáveis é utilizar um modelo de regressão apropriado para dados censurados.

Introdução

Em análise de sobrevivência, existem duas classes de modelos propostos na literatura: os **modelos paramétricos e os semiparamétricos**.

Os modelos paramétricos, também denominados de tempo de vida acelerado, **são mais eficientes, porém menos flexíveis do que os modelos semiparamétricos**.

A segunda classe de modelos, também denominada simplesmente de **modelo de regressão de Cox, tem sido bastante utilizada em estudos clínicos**. Além da flexibilidade, este modelo permite incorporar facilmente covariáveis dependentes do tempo.

Introdução

Quando se tem uma resposta envolvendo o tempo até a ocorrência de um evento e a presença de censura, em geral, o que **se deseja é utilizar um modelo de regressão para estudar a relação entre as variáveis**.

No entanto, **o tipo de resposta e o comportamento das variáveis não permitem, em geral, a utilização direta do modelo de regressão linear (definido na disciplina de Modelo de Regressão Linear - Profa. Cynthia)**.

Junta-se a isto o fato de que **a distribuição da resposta tende também, em geral, a ser assimétrica na direção dos maiores tempos de sobrevivência**, o que torna inapropriado o uso da distribuição normal para o componente estocástico do modelo.

Modelo de regressão exponencial

A utilização da distribuição exponencial para o erro e um componente determinístico da forma $\exp(\beta_0 + \beta_1 x)$ é certamente o modelo de regressão mais simples.

Este modelo, envolvendo uma única covariável, será utilizado para introduzir a modelagem de uma situação simples em análise de sobrevivência.

Na linguagem de modelos lineares generalizados, tem-se uma função de ligação logarítmica e a resposta com distribuição exponencial.

Em alguns livros a introdução das covariáveis para explicar o tempo até a ocorrência de um evento é feita usando a seguinte relação

$$\alpha(x) = \exp(x\beta) = \exp(\beta_0 + \beta_1 x).$$

Modelo de regressão exponencial

Considerando o caso mais simples com x podendo assumir apenas os valores 0 ou 1, tem-se que

$$\alpha(x) = \begin{cases} \exp(\beta_0), & \text{se } x = 0, \\ \exp(\beta_0 + \beta_1), & \text{se } x = 1. \end{cases}$$

A generalização do modelo de regressão exponencial é no sentido que se **pode incluir mais do que uma covariável**, sendo que as covariáveis podem ser qualitativas ou quantitativas. Neste caso, a expressão matemática do modelo é dada por

$$\alpha(\boldsymbol{x}) = \exp(\boldsymbol{x}\boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p),$$

sendo que p é a quantidade de covariáveis no modelo.

Modelo de regressão exponencial

O passo seguinte, após a especificação do modelo, é a estimação dos seus parâmetros. No caso particular que estamos estudando (apenas uma covariável), é necessário estimar e fazer inferência sobre o vetor de parâmetros $\boldsymbol{\theta} = (\beta_0, \beta_1)$.

Modelo de regressão exponencial

O passo seguinte, após a especificação do modelo, é a estimação dos seus parâmetros. No caso particular que estamos estudando (apenas uma covariável), é necessário estimar e fazer inferência sobre o vetor de parâmetros $\boldsymbol{\theta} = (\beta_0, \beta_1)$.

Modelo de regressão paramétrico - resumo

A utilização de covariáveis no modelo de regressão paramétrico pode ser resumido na Tabela 1.

Modelo	Parâmetro	Função de ligação	Survreg
Exponencial	$\alpha > 0$	$\log(\alpha) = \beta_0 + \beta_1 x$	$\alpha = \exp(\beta_0 + \beta_1 x)$
Weibull	$\alpha > 0$	$\log(\alpha) = \beta_0 + \beta_1 x$	$\alpha = \exp(\beta_0 + \beta_1 x)$
Log-normal	$\mu \in \mathbb{R}$	$\mu = \beta_0 + \beta_1 x$	$\mu = \beta_0 + \beta_1 x$

Tabela 1: Resumo das funções de ligação dos modelos Exponencial, Weibull e log-normal e sua utilização no pacote *survival*.

Aplicação

Aplicação no *software* **R.**

Adequação do modelo ajustado

Uma avaliação da adequação do modelo ajustado é parte fundamental da análise dos dados.

Técnicas gráficas, que fazem uso dos diferentes resíduos propostos, são, em particular, bastante utilizadas para examinar diferentes aspectos do modelo.

A seguir, os resíduos de Cox-Snell é descrito (úteis para examinar o ajuste global do modelo).

Adequação do modelo ajustado

CALMA!
Por que preciso aprender um outro método para análise a adequação do modelo? Não basta fazer um dos gráficos que já aprendemos?

Adequação do modelo ajustado

CALMA!
Por que preciso aprender um outro método para análise a adequação do modelo? Não basta fazer um dos gráficos que já aprendemos?

R.: Imagina que você tenha uma covariável do tipo dicotômica (0 ou 1), logo, para utilizar as técnicas que já aprendemos, temos que dividir o conjunto de dados em duas partes, $x = 0$ e $x = 1$, e então estimar duas curvas de sobrevivência, tanto via Kaplan-Meier quanto via modelo paramétrico (exponencial, Weibull ou log-normal).
Agora imagina que tenha 10 covariáveis, todas elas do tipo dicotômica (0 ou 1), logo tem-se $2^{10} = 1024$ possíveis combinações. Neste caso, para realizar as análises que já aprendemos, precisamos dividir nosso conjunto de dados considerando todas as possíveis combinações, o que torna tal análise inviável!

Resíduo de Cox-Snell

Os resíduos de Cox-Snell (1968) auxiliam examinar o ajuste global do modelo. Esses resíduos são quantidades determinadas por:

$$\hat{e}_i = \hat{\Lambda}(t_i|\mathbf{x_i}),$$

em que $\hat{\Lambda}(.)$ é a função de risco acumulado obtida no modelo ajustado.

Para os modelos de regressão exponencial, Weibull e log-normal, os resíduos de Cox-Snell são dados, respectivamente, por

$$\begin{aligned}
\text{Exponencial - } : \hat{e}_i &= [t_i \exp(-\boldsymbol{x_i'}\widehat{\boldsymbol{\beta}})] \\
\text{Weibull - } : \hat{e}_i &= [t_i \exp(-\boldsymbol{x_i'}\widehat{\boldsymbol{\beta}})]^{\widehat{\gamma}} \\
\text{Log-normal - } : \hat{e}_i &= -\log\left[1 - \Phi\left(\frac{log(t_i) - \boldsymbol{x}_i\widehat{\beta}}{\widehat{\sigma}}\right)\right]
\end{aligned}$$

Resíduo de Cox-Snell

Os resíduos $\hat{e}_i$ vêm de uma população homogênea e devem seguir uma distribuição exponencial padrão ($\alpha = 1$) se o modelo for adequado.

Embora os resíduos de Cox-Snell sejam úteis para examinar o ajuste global do modelo, eles não indicam o tipo de falha.

O uso de técnicas gráficas:

- O gráfico das curvas de sobrevivência dos resíduos, obtidas por Kaplan-Meier e pelo modelo exponencial padrão, podem ser utilizadas. Sendo que quanto mais próximas melhor, indicando assim o melhor modelo;
- O gráfico $\hat{e}_i$ versus $\hat{\Lambda}(\hat{e}_i)$ deve ser aproximadamente uma reta com inclinação 1, quando o modelo exponencial for adequado.

Interpretação dos coeficientes estimados

A interpretação dos coeficientes estimados do modelo não é simples, pois em geral, a escala da resposta foi transformada para a logarítmica.

Uma proposta razoável de interpretação é a de se fazer uso da razão de tempos medianos.

A razão dos tempos medianos de dois pacientes, 1 e 2, é dada por

$$\frac{t_{0,5}(\boldsymbol{x_1})}{t_{0,5}(\boldsymbol{x_2})} = \exp(\boldsymbol{x_1'}\widehat{\boldsymbol{\beta}} - \boldsymbol{x_2'}\widehat{\boldsymbol{\beta}}),$$

sendo que $\boldsymbol{x_1}$ e $\boldsymbol{x_2}$ são as covariáveis dos pacientes 1 e 2, respectivamente.

Aplicação

Aplicação no *software* **R.**

Especialização em *Data Science* e Estatística Aplicada

Módulo IV - Análise de Sobrevivência

Prof. Dr. Eder Angelo Milani

edermilani@ufg.br

IME INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

FEN FACULDADE DE ENFERMAGEM

UFG UNIVERSIDADE FEDERAL DE GOIÁS