IX ENCONTRO
PÓS-GRADUAÇÃO EM MATEMÁTICA
Área de concentração: Estatística

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA | UFBA

Inferência Bayesiana em Modelos de Sobrevivência


Maristela Dias de Oliveira

Departamento de Estatística - UFBA

QR code: Material do curso

Conteúdo do Curso

- Características;

Conteúdo do Curso

- Características;
- Descritiva;

Conteúdo do Curso

- Características;
- Descritiva;
- Funções/Quantidades de interesse;

Conteúdo do Curso

- Características;
- Descritiva;
- Funções/Quantidades de interesse;
- Estimação;

Conteúdo do Curso

- Características;
- Descritiva;
- Funções/Quantidades de interesse;
- Estimação;
- Modelos de Regressão;

Conteúdo do Curso

- Características;
- Descritiva;
- Funções/Quantidades de interesse;
- Estimação;
- Modelos de Regressão;
- Modelos de Probabilidade;

Conteúdo do Curso

- Características;
- Descritiva;
- Funções/Quantidades de interesse;
- Estimação;
- Modelos de Regressão;
- Modelos de Probabilidade;
- Abordagens:

Conteúdo do Curso

- Características;
- Descritiva;
- Funções/Quantidades de interesse;
- Estimação;
- Modelos de Regressão;
- Modelos de Probabilidade;
- Abordagens:
  1. Frequentista;

Conteúdo do Curso

- Características;
- Descritiva;
- Funções/Quantidades de interesse;
- Estimação;
- Modelos de Regressão;
- Modelos de Probabilidade;
- Abordagens:
  1. Frequentista;
  2. Bayesiana;

Conteúdo do Curso

- Características;
- Descritiva;
- Funções/Quantidades de interesse;
- Estimação;
- Modelos de Regressão;
- Modelos de Probabilidade;
- Abordagens:
  1. Frequentista;
  2. Bayesiana;
- Diagnóstico;

Conteúdo do Curso

- Características;
- Descritiva;
- Funções/Quantidades de interesse;
- Estimação;
- Modelos de Regressão;
- Modelos de Probabilidade;
- Abordagens:
  1. Frequentista;
  2. Bayesiana;
- Diagnóstico;
- Extensões.

Modelos de Sobrevivência

Definição

Modelos de sobrevivência podem ser definidos como uma classe de modelos quantitativos estocásticos utilizados para analisar características e fatores associados ao tempo até a ocorrência do desfecho ou evento de interesse.

Modelos de Sobrevivência

Definição

Modelos de sobrevivência podem ser definidos como uma classe de modelos quantitativos estocásticos utilizados para analisar características e fatores associados ao tempo até a ocorrência do desfecho ou evento de interesse.

O tempo é, portanto, a variável de interesse ou resposta do estudo.

Caracterização

Tempo de falha (T)

Uma falha representa a ocorrência do evento de interesse. O momento em que ele ocorre é chamado de *tempo de falha*.

Caracterização

Tempo de falha (T)

Uma falha representa a ocorrência do evento de interesse. O momento em que ele ocorre é chamado de *tempo de falha*.

Censura (C)

A censura representa a observação parcial da resposta. Por exemplo, o evento de interesse pode não ocorrer até o término do experimento. Diz-se, então, que o indivíduo sobreviveu ao experimento.

Mecanismos de censura

- Censura do tipo I: o estudo é terminado após um período pré-estabelecido de tempo;

Mecanismos de censura

- Censura do tipo I: o estudo é terminado após um período pré-estabelecido de tempo;
- Censura do tipo II: o estudo é terminado após um número pré-estabelecido de falhas;

Mecanismos de censura

- Censura do tipo I: o estudo é terminado após um período pré-estabelecido de tempo;
- Censura do tipo II: o estudo é terminado após um número pré-estabelecido de falhas;
- Censura do tipo aleatório: o indivíduo é excluído do estudo antes de experimentar a falha;

Mecanismos de censura

- Censura do tipo I: o estudo é terminado após um período pré-estabelecido de tempo;
- Censura do tipo II: o estudo é terminado após um número pré-estabelecido de falhas;
- Censura do tipo aleatório: o indivíduo é excluído do estudo antes de experimentar a falha;
  - Os dados observados são:

$$t = \min(T; C) \qquad \text{e} \qquad \delta = \begin{cases} 1, & T \leq C \\ 0, & T > C. \end{cases}$$

Mecanismos de censura

- Censura do tipo I: o estudo é terminado após um período pré-estabelecido de tempo;
- Censura do tipo II: o estudo é terminado após um número pré-estabelecido de falhas;
- Censura do tipo aleatório: o indivíduo é excluído do estudo antes de experimentar a falha;
  - Os dados observados são:

$$t = \min(T; C) \qquad \text{e} \qquad \delta = \begin{cases} 1, & T \leq C \\ 0, & T > C. \end{cases}$$

  - Se, numa amostra de n indivíduos, observa-se $C_i = C, \forall i = 1, \ldots, n$, tem-se também a censura do tipo I.

Ilustração dos mecanismos de censura (à direita)

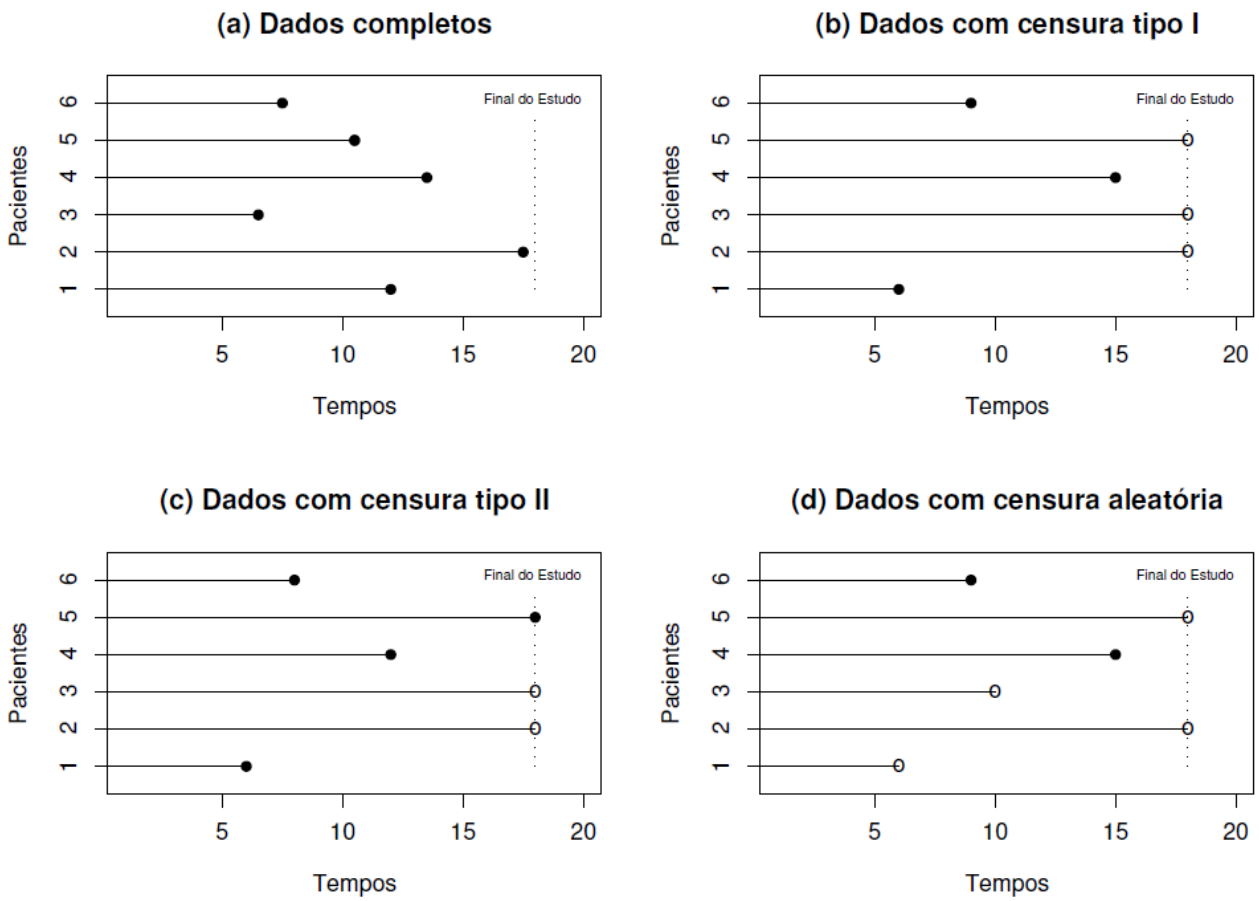


Figura: Mecanismos de censura à direita, em que •— falha e ○— censura.

Outras formas de dados incompletos

- Censura à esquerda;

Outras formas de dados incompletos

- Censura à esquerda;
- Censura Intervalar;

Outras formas de dados incompletos

- Censura à esquerda;
- Censura Intervalar;
- Truncamento à esquerda;

Outras formas de dados incompletos

- Censura à esquerda;
- Censura Intervalar;
- Truncamento à esquerda;
- Truncamento à direita.

Principais objetivos da estatística descritiva

- Sumarizar dados;

Principais objetivos da estatística descritiva

- Sumarizar dados;
- Comparar grupos;

Principais objetivos da estatística descritiva

- Sumarizar dados;
- Comparar grupos;
- Visualizar dados;

Principais objetivos da estatística descritiva

- Sumarizar dados;
- Comparar grupos;
- Visualizar dados;
- Identificar padrões;

Principais objetivos da estatística descritiva

- Sumarizar dados;
- Comparar grupos;
- Visualizar dados;
- Identificar padrões;
- Verificar se os dados foram coletados e processados corretamente;

Principais objetivos da estatística descritiva

- Sumarizar dados;
- Comparar grupos;
- Visualizar dados;
- Identificar padrões;
- Verificar se os dados foram coletados e processados corretamente;
- Adicionalmente:

Principais objetivos da estatística descritiva

- Sumarizar dados;
- Comparar grupos;
- Visualizar dados;
- Identificar padrões;
- Verificar se os dados foram coletados e processados corretamente;
- Adicionalmente:
  - Supor a distribuição dos dados;

Principais objetivos da estatística descritiva

- Sumarizar dados;
- Comparar grupos;
- Visualizar dados;
- Identificar padrões;
- Verificar se os dados foram coletados e processados corretamente;
- Adicionalmente:
  - Supor a distribuição dos dados;
  - Fazer inferências estatísticas.

Exemplo de aplicação

Objetivo do estudo

Analisar o tempo até cada reinternação de pacientes diagnosticados com câncer colorretal e submetidos à cirurgia de remoção do tumor.

Exemplo de aplicação

Objetivo do estudo

Analisar o tempo até cada reinternação de pacientes diagnosticados com câncer colorretal e submetidos à cirurgia de remoção do tumor.

- Estudo de coorte que analisa as reinternações hospitalares de pacientes com diagnóstico de câncer colorretal no Hospital de Bellvitge.

Exemplo de aplicação

Objetivo do estudo

Analisar o tempo até cada reinternação de pacientes diagnosticados com câncer colorretal e submetidos à cirurgia de remoção do tumor.

- Estudo de coorte que analisa as reinternações hospitalares de pacientes com diagnóstico de câncer colorretal no Hospital de Bellvitge.
- Acompanhou-se ativamente, até o ano de 2002, 403 pacientes diagnosticados entre janeiro de 1996 e dezembro de 1998.

Exemplo de aplicação

Objetivo do estudo

Analisar o tempo até cada reinternação de pacientes diagnosticados com câncer colorretal e submetidos à cirurgia de remoção do tumor.

- Estudo de coorte que analisa as reinternações hospitalares de pacientes com diagnóstico de câncer colorretal no Hospital de Bellvitge.
- Acompanhou-se ativamente, até o ano de 2002, 403 pacientes diagnosticados entre janeiro de 1996 e dezembro de 1998.
- A variável resposta é o tempo até as sucessivas reinternações (em dias) após a cirurgia para remoção do tumor.

Exemplo de aplicação

Objetivo do estudo

Analisar o tempo até cada reinternação de pacientes diagnosticados com câncer colorretal e submetidos à cirurgia de remoção do tumor.

- Estudo de coorte que analisa as reinternações hospitalares de pacientes com diagnóstico de câncer colorretal no Hospital de Bellvitge.
- Acompanhou-se ativamente, até o ano de 2002, 403 pacientes diagnosticados entre janeiro de 1996 e dezembro de 1998.
- A variável resposta é o tempo até as sucessivas reinternações (em dias) após a cirurgia para remoção do tumor.
- Disponível no pacote `'frailtypack'`.

Comandos R

```
dados <- read.csv("readmission.csv") |>
  filter(enum == 1) |>
  mutate(chemo = relevel(factor(chemo), "NonTreated"),
         sex = relevel(factor(sex), "Male"),
         dukes = relevel(factor(dukes), "A-B"),
         charlson = relevel(factor(charlson), "0"))|>
  select(id, time, event, chemo, sex, dukes, charlson, death)
#tabelas
round(prop.table(table(dados$event)),4)
round(prop.table(table(dados$sex)),4)
round(prop.table(table(dados$dukes)),4)
round(prop.table(table(dados$charlson)),4)

d1 <- dados %>% group_by(dukes) |>
  summarize(count = n()) |>
  mutate(pct = count/sum(count))

d2 <- dados %>% group_by(charlson) |>
  summarize(count = n()) |>
  mutate(pct = count/sum(count))
#Gráficos descritivos das variáveis
ggplot(d1, aes(dukes, pct)) +
  geom_bar(fill = "#001f3f",stat='identity') +
  scale_y_continuous(labels=scales::percent, breaks=seq(0,1,0.1))+
  labs(y="Porcentagem de pacientes", x="Estágio tumoral de Duke")+
  theme_bw() |
ggplot(d2, aes(charlson, pct)) +
  geom_bar(fill = "#001f3f",stat='identity') +
  scale_y_continuous(labels=scales::percent, breaks=seq(0,1,0.2))+
  labs(y="Porcentagem de pacientes", x="Índice de Comorbidade de Charlson")+
  theme_bw()
```

Objetivo da Análise de Sobrevivência

Questão

Como relacionar o tempo observado com as demais variáveis do estudo?

Funções básicas

Seja T uma variável aleatória não negativa (contínua) representando o tempo de falha de um indivíduo.

- Função de sobrevivência;
- Função de taxa de falha;
- Função taxa de falha acumulada.

Funções básicas

Seja T uma variável aleatória não negativa (contínua) representando o tempo de falha de um indivíduo.

- Função de sobrevivência;

$$S(t) = P(T \geq t), \quad t \geq 0$$

- Função de taxa de falha;
- Função taxa de falha acumulada.

Funções básicas

Seja T uma variável aleatória não negativa (contínua) representando o tempo de falha de um indivíduo.

- Função de sobrevivência;

Sendo F a função de distribuição acumulada de T, $S(t)$ pode ser expressa por $1-F(t)$.

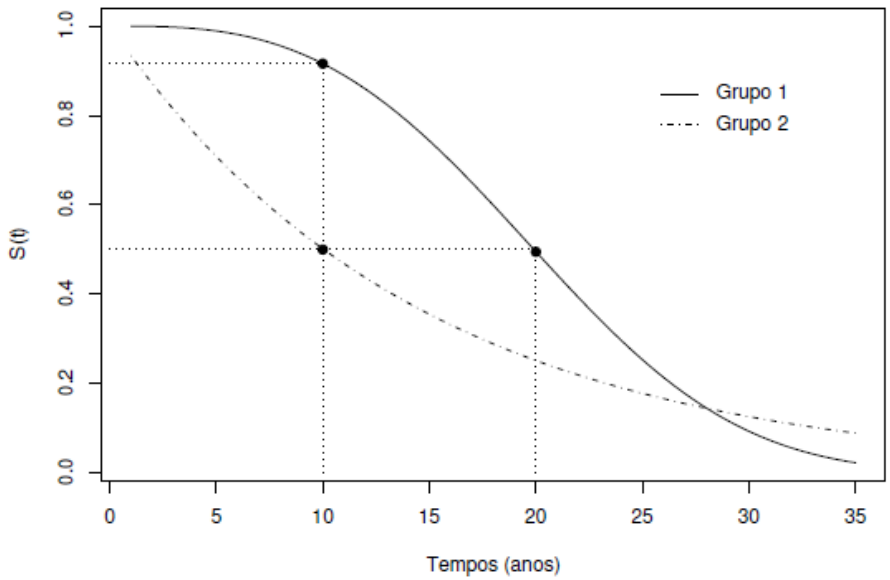


Figura: Comparação de duas curvas de sobrevivência.

Funções básicas

Seja T uma variável aleatória não negativa (contínua) representando o tempo de falha de um indivíduo.

- Função de sobrevivência;
- Função de taxa de falha;

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

- Função taxa de falha acumulada.

Funções básicas

Seja T uma variável aleatória não negativa (contínua) representando o tempo de falha de um indivíduo.

- Função de sobrevivência;
- Função de taxa de falha;

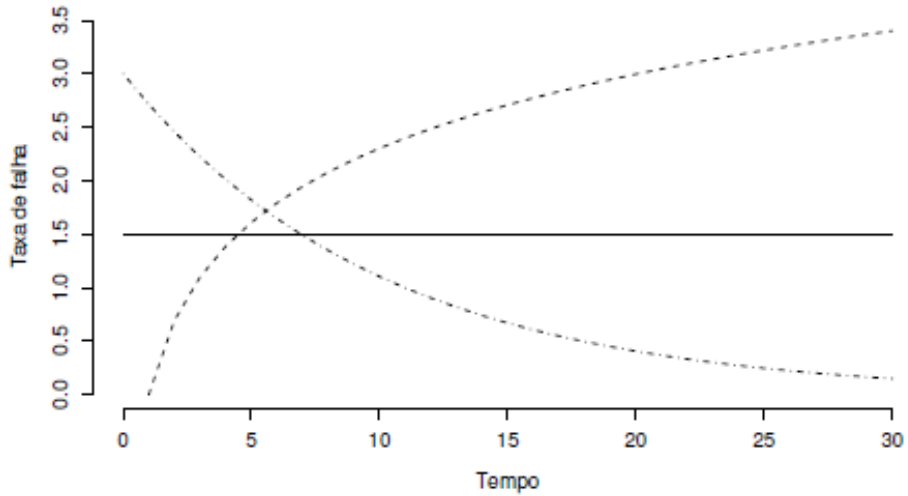


Figura: Exemplos de funções taxa de falha.

- Função taxa de falha acumulada.

Funções básicas

Seja T uma variável aleatória não negativa (contínua) representando o tempo de falha de um indivíduo.

- Função de sobrevivência;
- Função de taxa de falha;
- Função taxa de falha acumulada.

$$\Lambda(t) = \int_0^t \lambda(u)\, du, \quad t \geq 0.$$

Funções básicas

Seja T uma variável aleatória não negativa (contínua) representando o tempo de falha de um indivíduo.

- Função de sobrevivência;
- Função de taxa de falha;
- Função taxa de falha acumulada.

Relações entre as funções:

$$S(t) = \int_t^\infty f(u)\,du, \qquad \lambda(t) = \frac{f(t)}{S(t)} \quad \text{e} \quad \Lambda(t) = -\log(S(t)),$$

em que f é a função densidade de T

Outras quantidades

- Tempo médio

$$t_m = \int_0^\infty S(t)dt$$

- Vida média residual

$$vmr(t) = \frac{\int_t^\infty S(u)du}{S(t)}$$

Aplicação numa amostra

Estimação

As respostas às perguntas de interesse são dadas a partir de um conjunto de dados de sobrevivência (amostra), com pretensões de se expandir para a população.

Estimação de $S(t)$

Sejam

- $t_1 < \cdots < t_k$ os k tempos ordenados de falha;
- d_j o número de falhas em t_j, $j = 1 \ldots, k$;
- n_j o número de indivíduos sob risco em t_j.

Estimação de $S(t)$

Sejam

- $t_1 < \cdots < t_k$ os k tempos ordenados de falha;
- d_j o número de falhas em t_j, $j = 1 \ldots, k$;
- n_j o número de indivíduos sob risco em t_j.

Estimador de Kaplan-Meier

O estimador de Kaplan-Meier considera tantos intervalos de tempo quantos forem o número de falhas distintas. Os limites dos intervalos de tempo são os tempos de falha da amostra.

Estimação de $S(t)$

Sejam

- $t_1 < \cdots < t_k$ os k tempos ordenados de falha;
- d_j o número de falhas em t_j, $j = 1 \ldots, k$;
- n_j o número de indivíduos sob risco em t_j.

Estimador de Kaplan-Meier

O estimador de Kaplan-Meier considera tantos intervalos de tempo quantos forem o número de falhas distintas. Os limites dos intervalos de tempo são os tempos de falha da amostra.

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

Comandos R

```
library(survival)

## Estimação de S(t)

ekm <- survfit(Surv(time,event)~1, data=dados)
summary(ekm)
plot(ekm,xlab ="Tempo até a reiternação (dias)",ylab="S(t)")#, conf.int = F)

## Estimação de S(t) por grupos de variáveis
#1. Quimioterapia

ekm.q <- survfit(Surv(time,event)~chemo, data=dados)
summary(ekm.q)
plot(ekm.q,xlab ="Tempo até a reiternação (dias)",ylab="S(t)", col=c(3,2))

tq1 <- ekm.q$time[1:161]
tq2 <- ekm.q$time[162:(length(ekm.q$time))]
sq1 <- ekm.q$surv[1:161]
sq2 <- ekm.q$surv[162:(length(ekm.q$time))]

d.quimio <- data.frame(t = c(tq1,tq2), s = c(sq1,sq2),
            Quimioterapia=c(rep("Não tratado",length(tq1)),rep("Tratado",length(tq2))))
ggplot(d.quimio)+
  geom_step(aes(x=t, y=s, color=Quimioterapia))+
  labs(x="Tempo até a reiternação (dias)", y="S(t)")+
  theme_bw()

(...)
```

$$H_0 : S_1(t) = S_2(t)$$

Teste de Logrank

É o teste mais utilizado para testar a hipótese de igualdade de curvas de sobrevivência e comparar grupos de indivíduos.

$$H_0 : S_1(t) = S_2(t)$$

Teste de Logrank

É o teste mais utilizado para testar a hipótese de igualdade de curvas de sobrevivência e comparar grupos de indivíduos.

Tabela de contingência gerada no tempo t_j

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

$H_0 : S_1(t) = S_2(t)$

Teste de Logrank

É o teste mais utilizado para testar a hipótese de igualdade de curvas de sobrevivência e comparar grupos de indivíduos.

Tabela de contingência gerada no tempo t_j

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

$$\text{Sob } H_0 \quad T = \frac{\left[\sum_{j=1}^{k}(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k}(V_j)_2} \sim \chi^2_1,$$

$H_0 : S_1(t) = S_2(t)$

Teste de Logrank

É o teste mais utilizado para testar a hipótese de igualdade de curvas de sobrevivência e comparar grupos de indivíduos.

Tabela de contingência gerada no tempo t_j

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

$$\text{Sob } H_0 \quad T = \frac{\left[\sum_{j=1}^{k}(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k}(V_j)_2} \sim \chi^2_1,$$

com $(V_j)_2 = \frac{n_{2j}(n_j - n_{2j})d_j(n_j - d_j)}{n_j^2(n_j - 1)}$ e $w_{2j} = \frac{n_{2j}d_j}{n_j}$

Características do Kaplan-Meier

Vantagens

- É não viesado em grandes amostras;

Características do Kaplan-Meier

Vantagens

- É não viesado em grandes amostras;
- É consistente;

Características do Kaplan-Meier

Vantagens

- É não viesado em grandes amostras;
- É consistente;
- É o estimador de máxima verossimilhança de $S(t)$.

Características do Kaplan-Meier

Vantagens

- É não viesado em grandes amostras;
- É consistente;
- É o estimador de máxima verossimilhança de $S(t)$.

Características do Kaplan-Meier

Vantagens

- É não viesado em grandes amostras;
- É consistente;
- É o estimador de máxima verossimilhança de $S(t)$.

Função de verossimilhança para dados de sobrevivência:

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

com $S(t) = \exp\{-\Lambda(t_i)\}$

Características do Kaplan-Meier

Desvantagem

Não é capaz de medir o efeito de covariáveis sobre o tempo de sobrevivência.

Como incluir covariáveis?

Dados de sobrevivência geralmente contêm informações de covariáveis que podem influenciar o tempo de sobrevivência do indivíduo estudado.

Como incluir covariáveis?

Dados de sobrevivência geralmente contêm informações de covariáveis que podem influenciar o tempo de sobrevivência do indivíduo estudado.

A inclusão de covariáveis pode ser feita considerando duas classes de modelos de regressão:

- Modelos semi-paramétricos;
- Modelos paramétricos.

Como incluir covariáveis?

Dados de sobrevivência geralmente contêm informações de covariáveis que podem influenciar o tempo de sobrevivência do indivíduo estudado.

A inclusão de covariáveis pode ser feita considerando duas classes de modelos de regressão:

- **Modelos semi-paramétricos;**
  Também conhecido como modelo de regressão de Cox (ou de **riscos proporcionais**), é uma classe flexível de modelos de regressão, pois não supõe nenhuma forma funcional para a função taxa de falha.
  Tem como formulação geral: $\lambda(t) = \lambda_0(t)g(x'\beta)$.
  A suposição básica para o uso do modelo de regressão de Cox é que as taxas de falha sejam proporcionais.
- **Modelos paramétricos.**

Como incluir covariáveis?

Dados de sobrevivência geralmente contêm informações de covariáveis que podem influenciar o tempo de sobrevivência do indivíduo estudado.

A inclusão de covariáveis pode ser feita considerando duas classes de modelos de regressão:

- **Modelos semi-paramétricos;**
- **Modelos paramétricos.**
  Uma distribuição de probabilidade é proposta para o tempo de falha.

Como incluir covariáveis?

Dados de sobrevivência geralmente contêm informações de covariáveis que podem influenciar o tempo de sobrevivência do indivíduo estudado.

A inclusão de covariáveis pode ser feita considerando duas classes de modelos de regressão:

- **Modelos semi-paramétricos;**
- **Modelos paramétricos.**
  Uma distribuição de probabilidade é proposta para o tempo de falha. As estimativas dos parâmetros são usadas para caracterizar a relação entre as covariáveis e o tempo de falha.

Como incluir covariáveis?

Dados de sobrevivência geralmente contêm informações de covariáveis que podem influenciar o tempo de sobrevivência do indivíduo estudado.

A inclusão de covariáveis pode ser feita considerando duas classes de modelos de regressão:

- **Modelos semi-paramétricos;**
- **Modelos paramétricos.**
  Uma distribuição de probabilidade é proposta para o tempo de falha. As estimativas dos parâmetros são usadas para caracterizar a relação entre as covariáveis e o tempo de falha. Quando a distribuição é adequada, as estimativas obtidas são mais precisas.

Principais distribuições de probabilidade

Caracterização a partir das funções taxa de falha:

- Exponencial
- Weibull
- Log-normal
- Gama
- Gama Generalizada

Principais distribuições de probabilidade

Caracterização a partir das funções taxa de falha:

- Exponencial

$$\lambda(t) = \frac{1}{\alpha}, \quad \text{com } \alpha > 0$$

- Weibull
- Log-normal
- Gama
- Gama Generalizada

Principais distribuições de probabilidade

Caracterização a partir das funções taxa de falha:

- Exponencial

$$\lambda(t) = \frac{1}{\alpha}, \quad \text{com } \alpha > 0$$

  λ é constante.

- Weibull
- Log-normal
- Gama
- Gama Generalizada

Principais distribuições de probabilidade

Caracterização a partir das funções taxa de falha:

- Exponencial

$$\lambda(t) = \frac{1}{\alpha}, \quad \text{com } \alpha > 0$$

λ é constante. Para acomodar covariáveis, $\alpha = \exp(x'\beta)$.

- Weibull
- Log-normal
- Gama
- Gama Generalizada

Principais distribuições de probabilidade

Caracterização a partir das funções taxa de falha:

- Exponencial
- Weibull

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

- Log-normal
- Gama
- Gama Generalizada

Principais distribuições de probabilidade

Caracterização a partir das funções taxa de falha:

- Exponencial
- Weibull

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

λ é monótona.

- Log-normal
- Gama
- Gama Generalizada

Principais distribuições de probabilidade

Caracterização a partir das funções taxa de falha:

- Exponencial
- Weibull

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

λ é monótona. Para acomodar covariáveis, $\alpha = \exp(x'\beta)$.

- Log-normal
- Gama
- Gama Generalizada

Principais distribuições de probabilidade

Caracterização a partir das funções taxa de falha:

- Exponencial
- Weibull
- Log-normal
- Gama
- Gama Generalizada

Principais distribuições de probabilidade

Caracterização a partir das funções taxa de falha:

- Exponencial
- Weibull
- Log-normal
- Gama
- Gama Generalizada

$$\lambda(t) = \frac{t^{k\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}}{1 - \Gamma_1^*[k, -\left(\frac{t}{\alpha}\right)^{\gamma}]}, \quad \text{com } k,\ \alpha, \gamma > 0$$

-

Principais distribuições de probabilidade

Caracterização a partir das funções taxa de falha:

- Exponencial
- Weibull
- Log-normal
- Gama
- Gama Generalizada

$$\lambda(t) = \frac{t^{k\gamma-1}\exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}}{1-\Gamma_1^*[k,-\left(\frac{t}{\alpha}\right)^{\gamma}]}, \quad \text{com } k,\ \alpha,\gamma > 0$$

  - Tem a Gama e a Weibull como casos particulares e a Log-normal como distribuição limite;
  -

Principais distribuições de probabilidade

Caracterização a partir das funções taxa de falha:

- Exponencial
- Weibull
- Log-normal
- Gama
- Gama Generalizada

$$\lambda(t) = \frac{t^{k\gamma-1}\exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}}{1-\Gamma_1^*[k,-\left(\frac{t}{\alpha}\right)^{\gamma}]}, \quad \text{com } k,\ \alpha,\gamma > 0$$

- Tem a Gama e a Weibull como casos particulares e a Log-normal como distribuição limite;
- Acomoda diferentes formas de λ;
-

Principais distribuições de probabilidade

Caracterização a partir das funções taxa de falha:

- Exponencial
- Weibull
- Log-normal
- Gama
- Gama Generalizada

$$\lambda(t) = \frac{t^{k\gamma-1}\exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}}{1-\Gamma_1^*[k,-\left(\frac{t}{\alpha}\right)^{\gamma}]}, \quad \text{com } k,\ \alpha,\gamma > 0$$

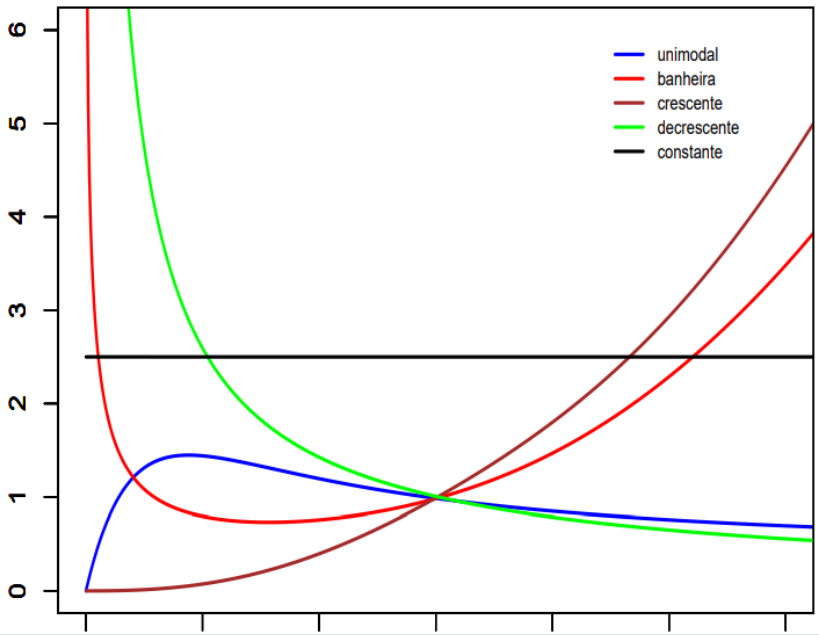
- Tem a Gama e a Weibull como casos particulares e a Log-normal como distribuição limite;
- Acomoda diferentes formas de λ;
- Está implementada no R.

Principais distribuições de probabilidade

Caracterização a partir das funções taxa de falha:

- Exponencial
- Weibull
- Log-normal
- Gama
- Gama Generalizada

Algumas formas de $\lambda(t)$



Estimação por máxima verossimilhança

Estimação por máxima verossimilhança

- Consiste em encontrar os parâmetros da distribuição proposta que maximizem a verossimilhança

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

em que $\delta_i = 1$ quando o $i-$ésimo tempo é uma falha.

Estimação por máxima verossimilhança

- Consiste em encontrar os parâmetros da distribuição proposta que maximizem a verossimilhança

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

em que $\delta_i = 1$ quando o $i-$ésimo tempo é uma falha.

- No caso do modelo Log-Normal temos:

$$f(t_i) = \frac{1}{\sqrt{2\pi}\sigma t_i} e^{-\frac{(\log(t_i)-\mu)^2}{2\sigma^2}} \quad \text{e} \quad \Lambda(t_i) = -\log\left[\Phi\left(\frac{-\log(t_i)+\mu}{\sigma}\right)\right]$$

Para acomodar covariáveis, $\mu = \exp(x'\beta)$.

Estimação por máxima verossimilhança no R

Funções disponíveis para análise de dados de sobrevivência:

Estimação por máxima verossimilhança no R

Funções disponíveis para análise de dados de sobrevivência:

- 'optim'
- 'survreg'
- 'flexsurvreg'

Estimação por máxima verossimilhança no R

Funções disponíveis para análise de dados de sobrevivência:

- 'optim'
  Precisa escrever toda a função de logverossimilhança para ser maximizada;
- 'survreg'
- 'flexsurvreg'

Estimação por máxima verossimilhança no R

Funções disponíveis para análise de dados de sobrevivência:

- 'optim'
- 'survreg'
  Pacote 'survival'. Contém uma variedade de modelos, inclusive Cox (função 'coxph')
- 'flexsurvreg'

Estimação por máxima verossimilhança no R

Funções disponíveis para análise de dados de sobrevivência:

- 'optim'
- 'survreg'
- 'flexsurvreg'
  Pacote 'flexsurv'. Generaliza o pacote 'survival'

Comandos R: seleção do modelo de probabilidade

```
library(flexsurv)

modelo.wei<-flexsurvreg(Surv(time,event)~1, dist="weibull", data = dados)
plot(modelo.wei, type="survival", ci = FALSE,conf.int=F)

modelo.gam<-flexsurvreg(Surv(time,event)~1, dist="gamma", data = dados)
plot(modelo.gam, col="green", type="survival", ci = FALSE,conf.int=F)

modelo.ggam<-flexsurvreg(Surv(time,event)~1, dist="gengamma", data = dados)
plot(modelo.ggam, col="blue", type="survival", ci = FALSE,conf.int=F)

modelo.ln<-flexsurvreg(Surv(time,event)~1, dist="lnorm", data = dados)
plot(modelo.ln, col="orange", type="survival", ci = FALSE,conf.int=F)

modelo.e<-flexsurvreg(Surv(time,event)~1, dist="exp", data = dados)
plot(modelo.e, col="pink", type="survival", ci = FALSE,conf.int=F)

#### AIC
AICs<-data.frame(Exponencial=AIC(modelo.e),Weibull=AIC(modelo.wei),Gama=AIC(modelo.gam),
                 Log_Normal=AIC(modelo.ln),Gen_Gama=AIC(modelo.ggam))

#### TRV - H0: O modelo j é melhor do que o Gama Generalizado
TRV<-data.frame(Exponencial=2*(logLik(modelo.ggam)-logLik(modelo.e)),
                 Weibull=2*(logLik(modelo.ggam)-logLik(modelo.wei)),
                 Gama=2*(logLik(modelo.ggam)-logLik(modelo.gam)),
                 Log_Normal=2*(logLik(modelo.ggam)-logLik(modelo.ln)))
p.val<-pchisq(as.numeric(TRV[1,1:4]),c(2,1,1,1),lower.tail = F)
TRV[1,]<-p.val
```

Interpretação dos coeficientes do modelo de regressão

```
> modelo.reg<-flexsurvreg(Surv(time,event)~1+chemo+sex+dukes+charlson, dist="lnorm", data = dados)
> modelo.reg
Call:
flexsurvreg(formula = Surv(time, event) ~ 1 + chemo + sex + dukes +
    charlson, data = dados, dist = "lnorm")

Estimates:
              data mean  est       L95%      U95%      se        exp(est)  L95%      U95%
meanlog             NA   7.2586    6.7268    7.7903    0.2713        NA        NA        NA
sdlog               NA   1.9373    1.7340    2.1643    0.1095        NA        NA        NA
chemoTreated    0.5260   0.2414   -0.2444    0.7272    0.2479    1.2730    0.7832    2.0693
sexFemale       0.4110   0.2909   -0.1686    0.7503    0.2344    1.3376    0.8448    2.1177
dukesC          0.3890  -0.7563   -1.2970   -0.2156    0.2758    0.4694    0.2734    0.8060
dukesD          0.1753  -1.9545   -2.6806   -1.2285    0.3704    0.1416    0.0685    0.2927
charlson1-2     0.0438  -1.0644   -2.1636    0.0349    0.5608    0.3449    0.1149    1.0355
charlson3       0.2466  -0.1570   -0.7513    0.4372    0.3032    0.8547    0.4718    1.5484

N = 365,  Events: 184,  Censored: 181
Total time at risk: 262013
Log-likelihood = -1462.73, df = 8
AIC = 2941.459
```

Interpretação dos coeficientes do modelo de regressão

```
> modelo.reg<-flexsurvreg(Surv(time,event)~1+chemo+sex+dukes+charlson, dist="lnorm", data = dados)
> modelo.reg
Call:
flexsurvreg(formula = Surv(time, event) ~ 1 + chemo + sex + dukes +
    charlson, data = dados, dist = "lnorm")

Estimates:
              data mean  est       L95%      U95%      se        exp(est)  L95%      U95%
meanlog             NA   7.2586    6.7268    7.7903    0.2713        NA        NA        NA
sdlog               NA   1.9373    1.7340    2.1643    0.1095        NA        NA        NA
chemoTreated    0.5260   0.2414   -0.2444    0.7272    0.2479    1.2730    0.7832    2.0693
sexFemale       0.4110   0.2909   -0.1686    0.7503    0.2344    1.3376    0.8448    2.1177
dukesC          0.3890  -0.7563   -1.2970   -0.2156    0.2758    0.4694    0.2734    0.8060
dukesD          0.1753  -1.9545   -2.6806   -1.2285    0.3704    0.1416    0.0685    0.2927
charlson1-2     0.0438  -1.0644   -2.1636    0.0349    0.5608    0.3449    0.1149    1.0355
charlson3       0.2466  -0.1570   -0.7513    0.4372    0.3032    0.8547    0.4718    1.5484

N = 365,  Events: 184,  Censored: 181
Total time at risk: 262013
Log-likelihood = -1462.73, df = 8
AIC = 2941.459
```

- O risco de reinternação dos indivíduos com estágio tumoral de Dukes A-B é $\approx 0,47$ vezes o risco dos indivíduos que apresentam estágio C.

Interpretação dos coeficientes do modelo de regressão

```
> modelo.reg<-flexsurvreg(Surv(time,event)~1+chemo+sex+dukes+charlson, dist="lnorm", data = dados)
> modelo.reg
Call:
flexsurvreg(formula = Surv(time, event) ~ 1 + chemo + sex + dukes +
    charlson, data = dados, dist = "lnorm")

Estimates:
               data mean  est       L95%      U95%      se        exp(est)  L95%      U95%
meanlog              NA    7.2586    6.7268    7.7903    0.2713        NA        NA        NA
sdlog                NA    1.9373    1.7340    2.1643    0.1095        NA        NA        NA
chemoTreated     0.5260    0.2414   -0.2444    0.7272    0.2479    1.2730    0.7832    2.0693
sexFemale        0.4110    0.2909   -0.1686    0.7503    0.2344    1.3376    0.8448    2.1177
dukesC           0.3890   -0.7563   -1.2970   -0.2156    0.2758    0.4694    0.2734    0.8060
dukesD           0.1753   -1.9545   -2.6806   -1.2285    0.3704    0.1416    0.0685    0.2927
charlson1-2      0.0438   -1.0644   -2.1636    0.0349    0.5608    0.3449    0.1149    1.0355
charlson3        0.2466   -0.1570   -0.7513    0.4372    0.3032    0.8547    0.4718    1.5484

N = 365,  Events: 184,  Censored: 181
Total time at risk: 262013
Log-likelihood = -1462.73, df = 8
AIC = 2941.459
```

- O risco de reinternação dos indivíduos com estágio tumoral de Dukes A-B é $\approx 0,47$ vezes o risco dos indivíduos que apresentam estágio C.
- O risco de reinternação dos indivíduos com estágio tumoral de Dukes A-B é $\approx 0,14$ vezes o risco dos indivíduos que apresentam estágio D.

Diagnósticos básicos

Resíduos de Cox-Snell também estão implementados

Diagnósticos básicos

Resíduos de Cox-Snell também estão implementados

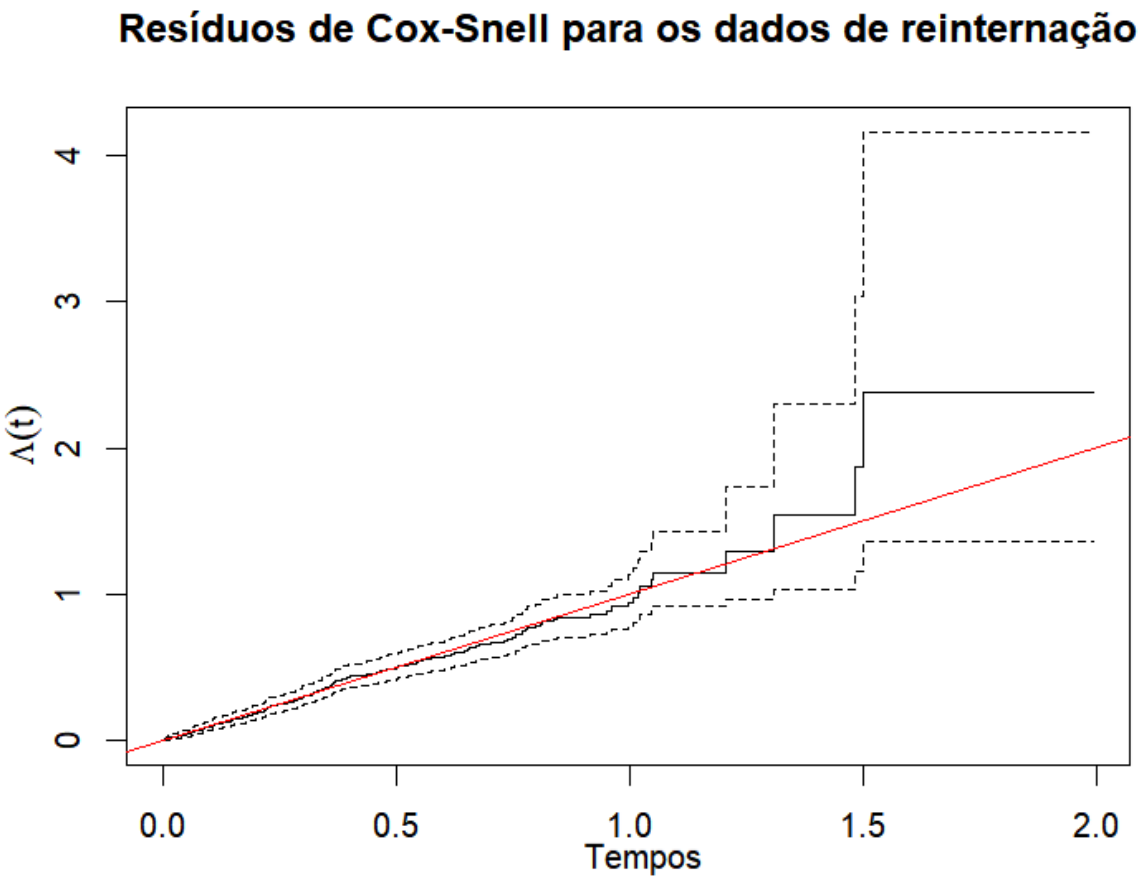
```
res<-residuals(modelo.reg, type="coxsnell")
surv <- survfit(Surv(res, dados$event) ~ 1)

#ou

res <- coxsnell_flexsurvreg(modelo.reg)
surv <- survfit(Surv(cs$est, dados$event) ~ 1)

#Em ambos os casos

plot(surv, fun="cumhaz")
abline(0, 1, col="red")
```



Diagnósticos básicos

Resíduos de Cox-Snell também estão implementados

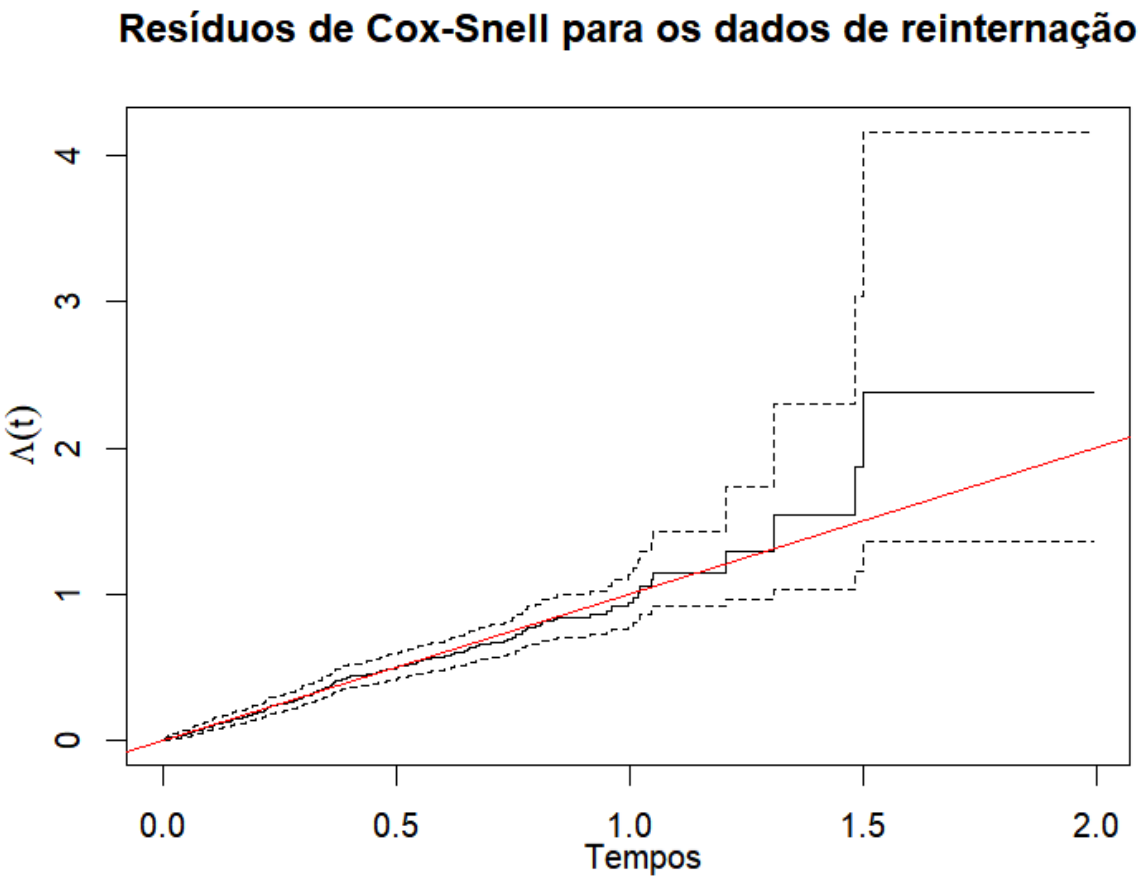
```
res<-residuals(modelo.reg, type="coxsnell")
surv <- survfit(Surv(res, dados$event) ~ 1)

#ou

res <- coxsnell_flexsurvreg(modelo.reg)
surv <- survfit(Surv(cs$est, dados$event) ~ 1)

#Em ambos os casos

plot(surv, fun="cumhaz")
abline(0, 1, col="red")
```



Se o modelo está bem ajustado, o gráfico deve acompanhar a diagonal.

Inferência Bayesiana

Inferência Bayesiana

- O parâmetro do modelo é uma quantidade cuja variabilidade pode ser descrita por uma distribuição de probabilidade: distribuição a *priori*;

Inferência Bayesiana

- O parâmetro do modelo é uma quantidade cuja variabilidade pode ser descrita por uma distribuição de probabilidade: distribuição a *priori*;
- A partir das informações da amostra (verossimilhança), a distribuição a *priori* é atualizada, gerando a distribuição a *posteriori*.

Inferência Bayesiana

- O parâmetro do modelo é uma quantidade cuja variabilidade pode ser descrita por uma distribuição de probabilidade: distribuição a *priori*;
- A partir das informações da amostra (verossimilhança), a distribuição a *priori* é atualizada, gerando a distribuição a *posteriori*.

Permite incorporar informações prévias, ou conhecimento do especialista

Inferência Bayesiana

- O parâmetro do modelo é uma quantidade cuja variabilidade pode ser descrita por uma distribuição de probabilidade: distribuição a *priori*;
- A partir das informações da amostra (verossimilhança), a distribuição a *priori* é atualizada, gerando a distribuição a *posteriori*.

Inferência Bayesiana

- O parâmetro do modelo é uma quantidade cuja variabilidade pode ser descrita por uma distribuição de probabilidade: distribuição a *priori*;
- A partir das informações da amostra (verossimilhança), a distribuição a *priori* é atualizada, gerando a distribuição a *posteriori*.

  - A distribuição de probabilidade a *posteriori* dos parâmetros costuma ser o resultado de uma abordagem bayesiana;

Inferência Bayesiana

- O parâmetro do modelo é uma quantidade cuja variabilidade pode ser descrita por uma distribuição de probabilidade: distribuição a *priori*;
- A partir das informações da amostra (verossimilhança), a distribuição a *priori* é atualizada, gerando a distribuição a *posteriori*.

  - A distribuição de probabilidade a *posteriori* dos parâmetros costuma ser o resultado de uma abordagem bayesiana;
  - Essa distribuição é que deve ser estudada para responder os questionamentos do pesquisador.

Inferência Bayesiana

Suponha que queremos estimar um parâmetro θ

Inferência Bayesiana

Suponha que queremos estimar um parâmetro θ

O Teorema de Bayes nos mostra de forma matemática como podemos incluir informações prévias (sobre θ) no mdelo, bem como atualizá-lo de acordo com a chegada de novas informações relevantes ao problema (amostra).

Inferência Bayesiana

Suponha que queremos estimar um parâmetro θ

O Teorema de Bayes nos mostra de forma matemática como podemos incluir informações prévias (sobre θ) no mdelo, bem como atualizá-lo de acordo com a chegada de novas informações relevantes ao problema (amostra).

Se denotamos a distribuição a *priori* por $\pi(\theta)$, e a função de verossimilhança por $L(\mathrm{x}|\theta)$, a distribuição de θ a *posteriori* é

$$\pi(\theta|\mathrm{x}) \propto L(\mathrm{x}|\theta)\pi(\theta).$$

Inferência Bayesiana

Suponha que queremos estimar um parâmetro θ

O Teorema de Bayes nos mostra de forma matemática como podemos incluir informações prévias (sobre θ) no mdelo, bem como atualizá-lo de acordo com a chegada de novas informações relevantes ao problema (amostra).

Se denotamos a distribuição a *priori* por $\pi(\theta)$, e a função de verossimilhança por $L(\mathrm{x}|\theta)$, a distribuição de θ a *posteriori* é

$$\pi(\theta|\mathrm{x}) \propto L(\mathrm{x}|\theta)\pi(\theta).$$

Problema

Como resumir a informação descrita na distribuição a *posteriori*?

Teoria das decisões e o Princípio de Bayes

Toda decisão, carrega consigo uma perda

Teoria das decisões e o Princípio de Bayes

Toda decisão, carrega consigo uma perda

- Seja $\ell(\theta, d)$ a função que representa a perda incorrida pelo pesquisador ao tomar a decisão d quando θ é a escolha feita pela natureza;

Teoria das decisões e o Princípio de Bayes

Toda decisão, carrega consigo uma perda

- Seja $\ell(\theta, d)$ a função que representa a perda incorrida pelo pesquisador ao tomar a decisão d quando θ é a escolha feita pela natureza;
- Queremos tomar decisões em que as perdas sejam mínimas;

Teoria das decisões e o Princípio de Bayes

Toda decisão, carrega consigo uma perda

- Seja $\ell(\theta, d)$ a função que representa a perda incorrida pelo pesquisador ao tomar a decisão d quando θ é a escolha feita pela natureza;
- Queremos tomar decisões em que as perdas sejam mínimas;
- Quando $\ell(\theta, d) = (\theta - d)^2$, conhecida como perda quadrática, é possível mostrar que (ver Bolfarine e Sandoval (2010), dentre outros)

$$\mathrm{E}\left[\theta|\mathrm{X}\right] = \int_{\theta\in\Theta} \theta\pi(\theta|\mathrm{x})\, d\theta. \tag{2}$$

minimiza o risco de se tomar uma decisão errada, e portanto, minimiza ℓ;

Teoria das decisões e o Princípio de Bayes

Toda decisão, carrega consigo uma perda

- Seja $\ell(\theta, d)$ a função que representa a perda incorrida pelo pesquisador ao tomar a decisão d quando θ é a escolha feita pela natureza;
- Queremos tomar decisões em que as perdas sejam mínimas;
- Quando $\ell(\theta, d) = (\theta - d)^2$, conhecida como perda quadrática, é possível mostrar que (ver Bolfarine e Sandoval (2010), dentre outros)

$$\mathrm{E}\left[\theta|\mathrm{X}\right] = \int_{\theta\in\Theta} \theta\pi(\theta|\mathrm{x})\, d\theta. \tag{2}$$

minimiza o risco de se tomar uma decisão errada, e portanto, minimiza ℓ;

- Note que (2) é o valor esperado a *posteriori* de θ, e é chamado de Estimador de Bayes sob perda quadrática.

Exemplo

Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma população $N(\mu; \sigma_0^2)$, com σ_0^2 conhecido. Consideremos para μ a distribuição a *priori* $N(a\,; b^2)$, ou seja,

$$L(\mathrm{x}|\mu) = \left(\frac{1}{\sqrt{2\pi}\sigma_0}\right)^n e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma_0^2}} \quad \text{e} \quad \pi(\mu) = \frac{1}{\sqrt{2\pi}b} e^{-\frac{(\mu-a)^2}{2b^2}}$$

Exemplo

Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma população $N(\mu; \sigma_0^2)$, com σ_0^2 conhecido. Consideremos para μ a distribuição a *priori* $N(a\,; b^2)$, ou seja,

$$L(\mathrm{x}|\mu) = \left(\frac{1}{\sqrt{2\pi}\sigma_0}\right)^n e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma_0^2}} \quad \text{e} \quad \pi(\mu) = \frac{1}{\sqrt{2\pi}b} e^{-\frac{(\mu-a)^2}{2b^2}}$$

(Atenção para os valores de a e b)

Exemplo

Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma população $N(\mu; \sigma_0^2)$, com σ_0^2 conhecido. Consideremos para μ a distribuição a *priori* $N(a\,; b^2)$, ou seja,

$$L(\mathrm{x}|\mu) = \left(\frac{1}{\sqrt{2\pi}\sigma_0}\right)^n e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma_0^2}} \quad \text{e} \quad \pi(\mu) = \frac{1}{\sqrt{2\pi}b} e^{-\frac{(\mu-a)^2}{2b^2}}$$

(Atenção para os valores de a e b)

Após receber as informações da amostra, tem-se:

$$\mu|\mathrm{X} \sim N\left(\frac{\frac{n}{\sigma_0^2}}{\frac{n}{\sigma_0^2}+\frac{1}{b^2}}\overline{X} + \frac{\frac{1}{b^2}}{\frac{n}{\sigma_0^2}+\frac{1}{b^2}}a \;\; ; \;\; \frac{1}{\frac{n}{\sigma_0^2}+\frac{1}{b^2}}\right)$$

Destaque para os valores **atualizados** de a e b *posteriori*!

Exemplo

Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma população $N(\mu; \sigma_0^2)$, com σ_0^2 conhecido. Consideremos para μ a distribuição a *priori* $N(a\,; b^2)$, ou seja,

$$L(\mathrm{x}|\mu) = \left(\frac{1}{\sqrt{2\pi}\sigma_0}\right)^n e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma_0^2}} \quad \text{e} \quad \pi(\mu) = \frac{1}{\sqrt{2\pi}b} e^{-\frac{(\mu-a)^2}{2b^2}}$$

(Atenção para os valores de a e b)

Após receber as informações da amostra, tem-se:

$$\mu|\mathrm{X} \sim N\left(\frac{\frac{n}{\sigma_0^2}}{\frac{n}{\sigma_0^2}+\frac{1}{b^2}}\overline{X} + \frac{\frac{1}{b^2}}{\frac{n}{\sigma_0^2}+\frac{1}{b^2}}a \;\; ; \;\; \frac{1}{\frac{n}{\sigma_0^2}+\frac{1}{b^2}}\right)$$

Destaque para os valores **atualizados** de a e b *posteriori*!

Obs: Se σ é desconhecido, temos uma distribuição a *posteriori* bivariada.

E quando não identificamos a distribuição a *posteriori*?

- A distribuição a *posteriori* é frequentemente uma distribuição altamente multivariada e que só está disponível em forma fechada para alguns modelos;

E quando não identificamos a distribuição a *posteriori*?

- A distribuição a *posteriori* é frequentemente uma distribuição altamente multivariada e que só está disponível em forma fechada para alguns modelos;
- Quando a distribuição a *posteriori* não está disponível em uma forma fechada, é necessário recorrer a outros métodos para estimá-la ou, alternativamente, extrair amostras dela.

E quando não identificamos a distribuição a *posteriori*?

- A distribuição a *posteriori* é frequentemente uma distribuição altamente multivariada e que só está disponível em forma fechada para alguns modelos;
- Quando a distribuição a *posteriori* não está disponível em uma forma fechada, é necessário recorrer a outros métodos para estimá-la ou, alternativamente, extrair amostras dela.
- Em geral, os métodos computacionais visam estimar as integrais que aparecem na inferência bayesiana:

E quando não identificamos a distribuição a *posteriori*?

- A distribuição a *posteriori* é frequentemente uma distribuição altamente multivariada e que só está disponível em forma fechada para alguns modelos;
- Quando a distribuição a *posteriori* não está disponível em uma forma fechada, é necessário recorrer a outros métodos para estimá-la ou, alternativamente, extrair amostras dela.
- Em geral, os métodos computacionais visam estimar as integrais que aparecem na inferência bayesiana:
  - A distribuição marginal a *posteriori* de cada elemento de θ pode ser obtida integrando a distribuição conjunta *posteriori* sobre o restante dos parâmetros:

$$\pi(\theta_i|\mathrm{x}) = \int \pi(\boldsymbol{\theta}|\mathrm{x}) d\boldsymbol{\theta}_{-i}$$

No exemplo anterior, $\boldsymbol{\theta} = (\mu, \sigma)$.

Métodos aproximados

- Métodos de Monte Carlo.
- Monte Carlo via cadeias de Markov (MCMC).
  - Metropolis-Hastings
  - Amostrador de Gibbs
  - Algoritmo Adaptative Metropolis-within-Gibbs.
- Métodos de integração numérica;
- Algoritmo EM;
- Aproximação de Laplace.

Métodos aproximados

- Métodos de Monte Carlo.
- Monte Carlo via cadeias de Markov (MCMC).
  - Metropolis-Hastings
  - Amostrador de Gibbs
  - Algoritmo Adaptative Metropolis-within-Gibbs.
- Métodos de integração numérica;
- Algoritmo EM;
- Aproximação de Laplace.

Os métodos MCMC são uma classe de métodos computacionais para extrair amostras da distribuição a *posteriori* conjunta.

Estimando a distribuição a *posteriori* em Sobrevivência

Geralmente um modelo de regressão em sobrevivência tem uma distribuição a *posteriori* multivariada.

Estimando a distribuição a *posteriori* em Sobrevivência

Geralmente um modelo de regressão em sobrevivência tem uma distribuição a *posteriori* multivariada.

No caso do modelo Log-Normal, por exemplo:

$$L = \prod_{i=1}^{n} \left\{ \left[\frac{1}{\sqrt{2\pi}\sigma t_i} e^{-\frac{(\log(t_i)-\mu)^2}{2\sigma^2}} \right]^{\delta_i} \left[\Phi\left(\frac{-\log(t_i)+\mu}{\sigma} \right) \right]^{1-\delta_i} \right\}$$

Estimando a distribuição a *posteriori* em Sobrevivência

Geralmente um modelo de regressão em sobrevivência tem uma distribuição a *posteriori* multivariada.

No caso do modelo Log-Normal, por exemplo:

$$L = \prod_{i=1}^{n} \left\{ \left[\frac{1}{\sqrt{2\pi}\sigma t_i} e^{-\frac{(\log(t_i)-\mu)^2}{2\sigma^2}} \right]^{\delta_i} \left[\Phi\left(\frac{-\log(t_i)+\mu}{\sigma} \right) \right]^{1-\delta_i} \right\}$$

$$\pi(\mu) = \frac{1}{\sqrt{2\pi}b_1} e^{-\frac{(\mu-a_1)^2}{2b_1^2}} \quad (\textit{priori} \text{ normal})$$

$$\pi(\sigma) = \frac{b_2^{a_2}}{\Gamma(a_2)} \sigma^{a_2-1} \exp\{-b_2\sigma\} \quad (\textit{priori} \text{ gama})$$

$$\beta \sim N_p(\mathrm{m}\ ;\ \mathbf{\Sigma}),\ \text{a } \textit{priori} \quad (\text{ sendo } \mathbf{\Sigma} \text{ uma matriz diagonal})$$

Estimando a distribuição a *posteriori* em Sobrevivência

Geralmente um modelo de regressão em sobrevivência tem uma distribuição a *posteriori* multivariada.

No caso do modelo Log-Normal, por exemplo:

$$L = \prod_{i=1}^{n} \left\{ \left[\frac{1}{\sqrt{2\pi}\sigma t_i} e^{-\frac{(\log(t_i)-\mu)^2}{2\sigma^2}} \right]^{\delta_i} \left[\Phi\left(\frac{-\log(t_i)+\mu}{\sigma} \right) \right]^{1-\delta_i} \right\}$$

$$\pi(\mu) = \frac{1}{\sqrt{2\pi}b_1} e^{-\frac{(\mu-a_1)^2}{2b_1^2}} \qquad (\textit{priori} \text{ normal})$$

$$\pi(\sigma) = \frac{b_2^{a_2}}{\Gamma(a_2)} \sigma^{a_2-1} \exp\{-b_2\sigma\} \qquad (\textit{priori} \text{ gama})$$

$$\boldsymbol{\beta} \sim N_p(\mathrm{m}\ ;\ \boldsymbol{\Sigma}),\ \text{a } \textit{priori} \qquad (\text{ sendo } \boldsymbol{\Sigma} \text{ uma matriz diagonal})$$

Note que, na presença de covariáveis, a distribuição a *priori* para μ é estabelecida apenas através da distribuição a *priori* para o vetor $\boldsymbol{\beta}$.

Distribuição a *posteriori* para os dados de reinternação

$$
\begin{aligned}
\pi(\sigma, \boldsymbol{\beta}|\mathrm{t}) &= \prod_{i=1}^{n} \left\{ \left[\frac{1}{\sqrt{2\pi}\sigma t_i} e^{-\frac{(\log(t_i)-\exp(\mathrm{x}'\boldsymbol{\beta}))^2}{2\sigma^2}} \right]^{\delta_i} \right. \\
&\quad \left. \left[\Phi\left(\frac{-\log(t_i) + \exp(\mathrm{x}'\boldsymbol{\beta})}{\sigma} \right) \right]^{1-\delta_i} \right\} \times \\
&\times \frac{b_2^{a_2}}{\Gamma(a_2)} \sigma^{a_2-1} \exp\{-b_2\sigma\} \times \\
&\times |\boldsymbol{\Sigma}|^{-1/2} (2\pi)^{-6/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - \mathrm{m})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\beta} - \mathrm{m}) \right\},
\end{aligned}
$$

com $\mathrm{m} = (m_1, m_2, m_3, m_4, m_5, m_6)$ e $\boldsymbol{\Sigma} = \begin{bmatrix} d_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & d_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & d_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & d_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & d_6 \end{bmatrix}$

Estimação das marginais a *posteriori* no **R**

Pacotes úteis para estimação bayesiana em modelos de sobrevivência:

- Laplaces Demon
- INLA

Estimação das marginais a *posteriori* no **R**

Pacotes úteis para estimação bayesiana em modelos de sobrevivência:

- Laplaces Demon
  - O usuário pode construir qualquer tipo de modelo de probabilidade com uma função de modelo especificada por ele.
  - O modelo pode ser atualizado com quadratura iterativa, Aproximação de Laplace, MCMC, etc.
  - Disponibiliza uma variedade de recursos, incluindo diagnósticos MCMC, dentre outros.
  - O usuário precisa escrever explictamente as expressões da verossimilhança.
  - Possui um tempo computacional longo.
- INLA

Estimação das marginais a *posteriori* no **R**

Pacotes úteis para estimação bayesiana em modelos de sobrevivência:

- Laplaces Demon
- INLA

Estimação das marginais a *posteriori* no R

Pacotes úteis para estimação bayesiana em modelos de sobrevivência:

- Laplaces Demon
- INLA
  - Realiza uma análise bayesiana completa de modelos aditivos usando aproximação de Laplace;
  - Contém uma variedade de modelos de sobrevivência implementados (se comunica com o `'survival'`);
  - Contém uma variedade de distribuições a *piori* já implementadas;
  - Fornece as marginais a *posteriori* de cada parâmetro;
  - É extremamente rápido computacionalmente;
  - Não permite que o usuário implemente seu próprio modelo.

Comandos R - INLA

```
  ## Escolhendo uma normal p-variada para os parâmetros da regressão

  p<-6                                              #nro de parâmetros
  beta0 <- rep(0,p)                                 #hiper parâmetros com média 0 a priori
  V0 <- matrix(c(rep(0,p^2)),ncol=p,nrow=p)         #matriz de covariâncias a priori  6x6
  diag(V0) <- rep(1,p)                              #determina os elementos da diagonal

hiper.priori<-list(prior="mvnorm",param=c(beta0=beta0,V0=V0))      #indica a priori
#-------------------------------------------------
## Escolhe piori gama para o sigma da lognormal e estabelece a sua precisão

prec.prior <- list(prec = list(prior = "loggamma", param = c(0.01, 0.01)),
                   initial = 4, fixed = FALSE)
#---------------------------------------
##  Criando uma estrutura de dados com o formato necessário para análise de dados com INLA

sinla.dados<-inla.surv(dados$time,dados$event)
#-----------------------------------------------
## Modelo de regressão

form.ln<-sinla.dados ~ -1 +dados$chemo+dados$sex+dados$dukes+dados$charlson
ln.dados <- inla(form.ln, data = dados,family = "lognormal.surv", control.family(prec.prior),
            control.predictor =hiper.priori)
#---------------------------------------------------

#Quais fatores aumentam o tempo de sobrevivência (e portanto, reduzem o risco)
ef.post<-lapply(ln.dados$marginals.fixed, function(X){1 - inla.pmarginal(0, X)})
efeitos<-unlist(ef.post)

plot(ln.dados,single=T)
```

Seleção de covariáveis

- Abordagem frequentista
  - Teste da razão de verossimilhanças
  - Critério de Informação de Akaike (AIC)

Seleção de covariáveis

- Abordagem frequentista
- Abordagem Bayesiana
  - Fator de Bayes
  - Critério de Informação Deviance (DIC)
  - Critério de informação Watanabe-Akaike (WAIC)
  - Ordenadas preditivas condicionais (CPO)

Adequação do modelo ajustado

Os resíduos quantílicos (DUNN; SMYTH, 1996), para uma variável resposta contínua censurada à direita, são quantidades calculadas por:

$$r = \Phi^{-1}\{\hat{u}\}$$

em que $\Phi(\cdot)$ é a função de distribuição acumulada de uma distribuição normal padrão, $\hat{u}$ é definido como um valor aleatório de uma distribuição uniforme no intervalo $[\hat{\Lambda}_{ij}(t), 1]$ se o tempo for um tempo de censura e $\hat{u} = \hat{\Lambda}_{ij}(t)$ se o tempo for um tempo de falha.

Resultados

Seleção de Covariáveis - abordagem frequentista

- Valor-p do TRV: 0,92.
- AIC do modelo completo: 7409,45.
- AIC do modelo reduzido: 7407,46 (sem a variável Quimioterapia).

Estimativas frequentistas

Tabela: Estimativas dos parâmetros do modelo Weibull de fragilidade gama, obtidas a partir da abordagem frequentista, para os dados de pacientes diagnosticados com câncer colorretal no Hospital de Bellvitge entre 1996 e 1998.

	Parâmetro	Estimativa	E.P.	Valor-p	IC (2,5%)	IC (97,5%)
Intercepto	β_0	-6,895	0,290	<0,001	-7,464	-6,327
Sexo (Feminino)	β_1	-0,379	0,114	0,001	-0,603	-0,155
Estágio tumoral						
A-B (ref.)						
C	β_2	0,315	0,126	0,012	0,069	0,562
D	β_3	1,155	0,162	<0,001	0,837	1,473
Índ. de comorbidade						
0 (ref.)						
1-2	β_4	0,602	0,219	0,006	0,174	1,031
≥ 3	β_5	0,176	0,124	0,154	-0,066	0,419
*Weibull	γ	0,906	0,039	0,016	0,830	0,982
Fragilidade	ξ	0,145	0,070	0,037	0,009	0,281

*Em todos os testes a hipótese nula é de igualdade do coeficiente a zero, mas esse testa $H_0 : \gamma = 1$.

Seleção de Covariáveis - abordagem Bayesiana

- Fator de Bayes: 0,73.

Seleção de Covariáveis - abordagem Bayesiana

- Fator de Bayes: 0,73.
- DIC e tempo computacional.

	Completo	Reduzido
DIC	7416,04	7411,79
Tempo Computacional	17,02h	14,55h

Estimativas Bayesianas

Tabela: Estimativas dos parâmetros do modelo Weibull de fragilidade gama, obtidas a partir da abordagem Bayesiana para os dados de pacientes diagnosticados com câncer colorretal no Hospital de Bellvitge entre 1996 e 1998.

	Parâmetro	Estimativa (moda)	E.P.	LI (2,5%)	LS (97,5%)
Intercepto	β_0	-6,899	0,295	-7,573	-6,403
Sexo (Feminino)	β_1	-0,367	0,118	-0,619	-0,155
Estágio tumoral					
A-B (ref.)					
C	β_2	0,328	0,131	0,060	0,574
D	β_3	1,172	0,170	0,846	1,506
Índ. de comorbidade					
0 (ref.)					
1-2	β_4	0,624	0,223	0,135	1,007
≥ 3	β_5	0,178	0,124	-0,067	0,422
Weibull	γ	0,915	0,040	0,839	0,996
Fragilidade	ξ	0,162	0,076	0,067	0,353

Interpretações - abordagem Bayesiana

- Sexo
  - O risco de reinternação dos pacientes do sexo feminino é 0,69 vezes o risco dos pacientes do sexo masculino.

Interpretações - abordagem Bayesiana

- Sexo
- Estágio tumoral de Dukes

  - O risco de reinternação dos pacientes que apresentam estágio tumoral de Dukes A-B é 0,72 vezes o risco dos pacientes que apresentam estágio C;

  - O risco de reinternação dos pacientes que apresentam estágio tumoral de Dukes A-B é 0,31 vezes o risco dos pacientes que apresentam estágio D.

Interpretações - abordagem Bayesiana

- Sexo
- Estágio tumoral de Dukes
- Índice de comorbidade Charlson

  - O risco de reinternação dos pacientes que apresentam índice de comorbidade de Charlson 0 é 0,54 vezes o risco dos pacientes que apresentam índice de comorbidade 1-2;

  - O risco de reinternação dos pacientes que apresentam índice de comorbidade de Charlson 0 é 0,84 vezes o risco dos pacientes que apresentam índice de comorbidade $\geq$3.

Conclusão

- A proposta do minicurso foi apresentar como funciona a estimação Bayesiana para dados de sobrevivência;

Conclusão

- A proposta do minicurso foi apresentar como funciona a estimação Bayesiana para dados de sobrevivência;
- Aos mais habilidosos, esperamos que tenham conseguido acompanhar os ajustes dos modelos;

Conclusão

- A proposta do minicurso foi apresentar como funciona a estimação Bayesiana para dados de sobrevivência;
- Aos mais habilidosos, esperamos que tenham conseguido acompanhar os ajustes dos modelos;
- É importante avaliar a relação: ganho no processo de estimação $\times$ dificuldade de imeplemntação dos ajustes;

Conclusão

- A proposta do minicurso foi apresentar como funciona a estimação Bayesiana para dados de sobrevivência;
- Aos mais habilidosos, esperamos que tenham conseguido acompanhar os ajustes dos modelos;
- É importante avaliar a relação: ganho no processo de estimação $\times$ dificuldade de imeplemntação dos ajustes;
- No modelo de Cox, a estimação da quantidade não paramétrica deve ser feita com métodos especiais

Conclusão

- A proposta do minicurso foi apresentar como funciona a estimação Bayesiana para dados de sobrevivência;
- Aos mais habilidosos, esperamos que tenham conseguido acompanhar os ajustes dos modelos;
- É importante avaliar a relação: ganho no processo de estimação $\times$ dificuldade de imeplemntação dos ajustes;
- No modelo de Cox, a estimação da quantidade não paramétrica deve ser feita com métodos especiais
    - Splines;

Conclusão

- A proposta do minicurso foi apresentar como funciona a estimação Bayesiana para dados de sobrevivência;
- Aos mais habilidosos, esperamos que tenham conseguido acompanhar os ajustes dos modelos;
- É importante avaliar a relação: ganho no processo de estimação $\times$ dificuldade de imeplemntação dos ajustes;
- No modelo de Cox, a estimação da quantidade não paramétrica deve ser feita com métodos especiais
  - Splines;
  - Polinomios de Bernstein;

Conclusão

- A proposta do minicurso foi apresentar como funciona a estimação Bayesiana para dados de sobrevivência;
- Aos mais habilidosos, esperamos que tenham conseguido acompanhar os ajustes dos modelos;
- É importante avaliar a relação: ganho no processo de estimação $\times$ dificuldade de imeplemntação dos ajustes;
- No modelo de Cox, a estimação da quantidade não paramétrica deve ser feita com métodos especiais
  - Splines;
  - Polinomios de Bernstein;
  - Etc.

Referências


DUNN, P. K.; SMYTH, G. K., *Randomized quantile residuals*. Journal of Computational and graphical statistics, v. 5, n. 3, p. 236–244, 1996. Taylor & Francis


COLOSIMO, E. A.; GIOLO, S. R, *Análise de sobrevivência aplicada*. 1. ed.: Editora Blucher, 2006.


GELMAN, A.; CARLIN, J. B. Carlin; STERN, H S.; DUNSON, D. B.; VETHTARI, A. and RUBIN, D. B. 2013. Bayesian Data Analysis. 3rd ed. Boca Raton, FL: Chapman & Hall/CRC Press.


GONZÁLEZ, J. R. et al., *ex differences in hospital readmission among colorectal cancer patients*. Journal of Epidemiology & Community Health, BMJ Publishing Group Ltd, v. 59, n. 6, p. 506–511, 2005.


HOUGAARD, P, *Analysis of multivariate survival data*. Springer, 2000. v. 564.

Obrigada!