EPI10 - Análise de Sobrevivência

Riscos competitivos em análise de dados de sobrevivência

Rodrigo Citton P. dos Reis
citton.padilha@ufrgs.br

Universidade Federal do Rio Grande do Sul
Faculdade de Medicina
Programa de Pós-Graduação em Epidemiologia

Porto Alegre, 2021

PPGEPI
Programa de Pós-Graduação em Epidemiologia

Introdução

Análise de sobrevivência

Análise de sobrevivência tipicamente foca em dados de **tempo até o evento**. De maneira mais geral, consiste em técnicas para variáveis aleatórias positivas, tais como

- tempo até a morte
- tempo até o início (ou recidiva) da doença
- tempo de permanência no hospital
- tempo de duração de uma greve
- medições de carga viral

Tipicamente, dados de sobrevivência não são completamente observados, mas são **censurados**.

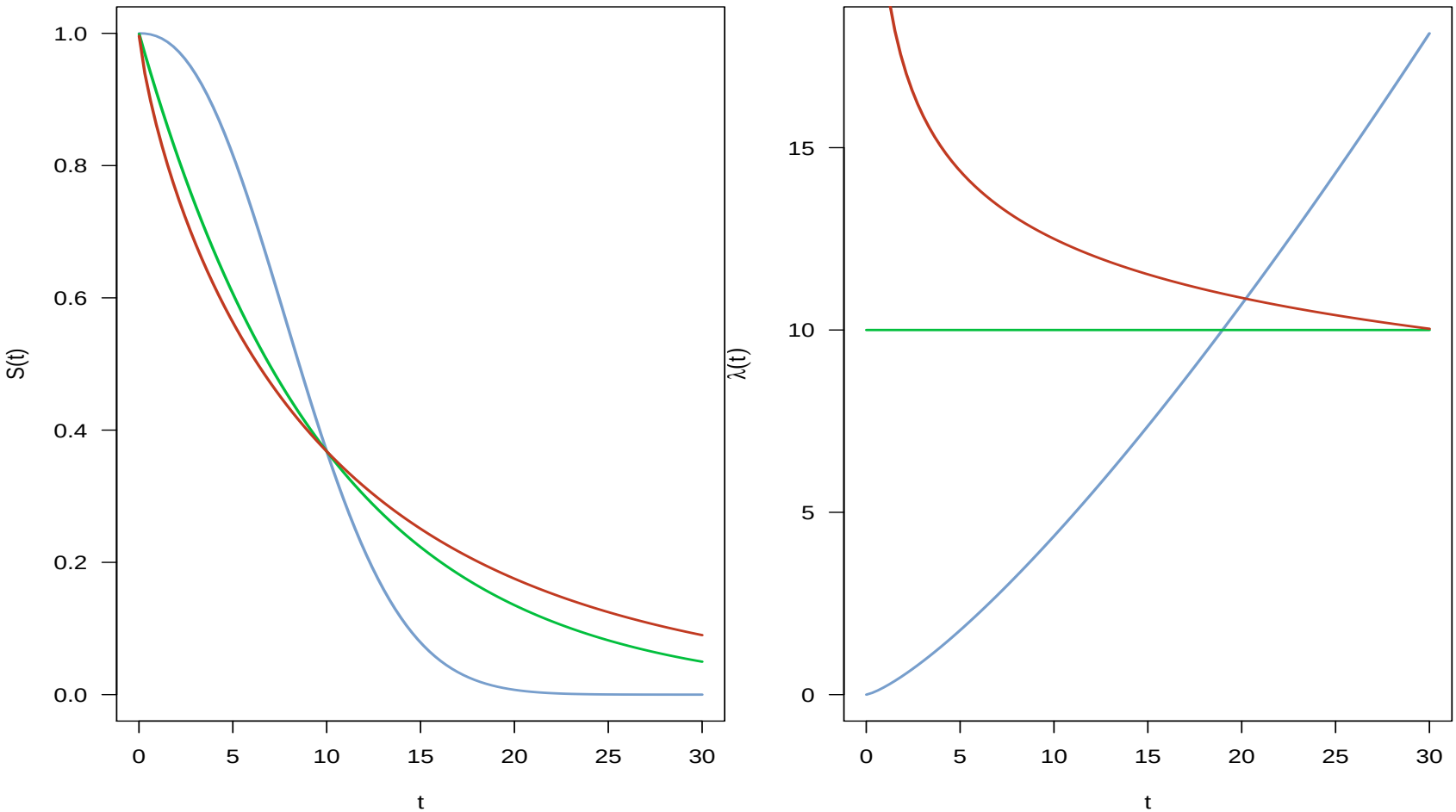
Análise de sobrevivência

Estaremos interessados em estimar certas características associadas com a variável aleatória tempo até o evento, denominada por T, como por exemplo

- a função **sobrevivência** $S(t) = \Pr(T > t)$
- a função **de taxa de falha** $\lambda(t) = \lim_{\triangle t \to 0} \dfrac{\Pr(t \leq T < t + \triangle t | T \geq t)}{\triangle t}$
- Lembrando da relação entre estas duas funções:

$$S(t) = \exp\left\{ - \int_0^t \lambda(u) du \right\}$$

Análise de sobrevivência



Análise de sobrevivência

- Descrever dados de sobrevivência
  - Estimador **Kaplan-Meier**
- Comparar funções de sobrevivência entre grupos
  - Teste **logrank**
- Explicar a sobrevivência com covariáveis
  - Modelos de regressão $\Rightarrow$ **Modelo de Cox**

Definição de riscos competitivos

- **Riscos competitivos:** situação em que mais de uma causa de falha é possível.
- Se falhas são diferentes causas de morte, apenas a primeira destas a ocorrer é observada.
- Em outras situações, observações após a primeira falha podem ser observadas, mas não são de interesse.

Definição de riscos competitivos

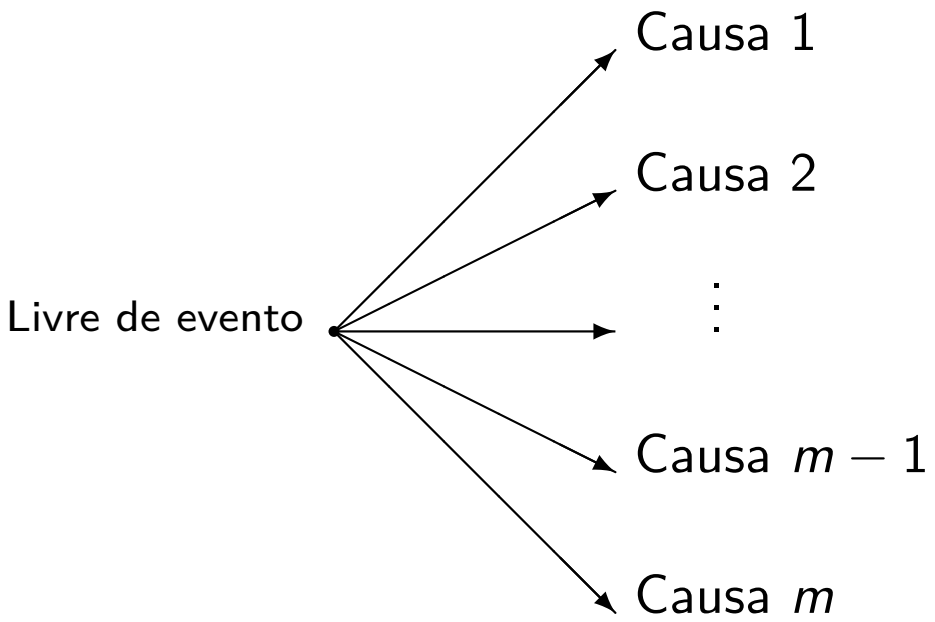


Figure 1: Situação de riscos competitivos com m causas de falhas.

Exemplos

- O tempo até o diagnóstico de uma certa doença (**demência por Alzheimer**).
  - A morte antes da doença é um risco concorrente.
- Em estudos de câncer, a morte por câncer pode ser de interesse, e morte por outras causas (mortalidade cirúrgica, velhice) são riscos concorrentes.
  - Por outro lado, pode-se estar interessado em tempo até a recidiva, em que a morte por qualquer causa é um risco concorrente.
  - Outra possibilidade, é que estajamos interessados no tempo até a morte por um tipo de câncer em específico, e mortes por outros tipos de câncer são riscos concorrentes.
- O tempo até a morte por evento cardíaco, em que outras causas de morte estão presentes, e portanto são riscos concorrentes.
  - **Populações mais velhas.**

Eventos concorrentes ⇒ Riscos competitivos

Eventos concorrentes ⇒ Riscos competitivos

- A introdução de novas técnicas para análise de dados de sobrevivência sujeitos a riscos competitivos se faz necessária, pois as antigas técnicas consideram que eventos ocorridos pelas demais causas, que não a de interesse, são observações **censuradas**.
  - Por observação censurada (**a direita**) entendemos que o evento ocorrerá após o último tempo observado.
  - No exemplo de tempo até o diagnóstico de demência por Alzheimer, quando um indivíduo morre antes do evento de interesse, ele não pode mais experimentar o evento.

Formulação básica

- Seja T o tempo até o evento e D a causa do evento. A função de taxa de falha **de causa específica** da j-ésima causa é definida por

$$\lambda_j(t) = \lim_{\triangle t \to 0} \frac{\Pr(t \leq T < t + \triangle t, D = j | T \geq t)}{\triangle t}, \ j = 1, \ldots, m.$$

- A respectiva função de taxa de falha acumulada de causa específica é definida por

$$\Lambda_j(t) = \int_0^t \lambda_j(u) du.$$

  - A função $S_j(t) = \exp\{-\Lambda_j(t)\}$ não deve ser interpretada como uma função de sobrevivência marginal.

Formulação básica

- As funções de **taxa de falha geral** $\lambda(t)$ e **sobrevivência** $S(t)$ são definidas em termos das funções de taxa de falha de causa específica

$$\begin{aligned}
\lambda(t) &= \lim_{\triangle t \to 0} \frac{\Pr(t \leq T < t + \triangle t | T \geq t)}{\triangle t} \\
&= \lim_{\triangle t \to 0} \frac{1}{\triangle t} \Pr\left(\bigcup_{j=1}^{m} \{t \leq T < t + \triangle t, D = j\} | T \geq t \right) \\
&= \sum_{j=1}^{m} \lim_{\triangle t \to 0} \frac{1}{\triangle t} \Pr(t \leq T < t + \triangle t, D = j | T \geq t) = \sum_{j=1}^{m} \lambda_j(t), \\
S(t) &= \exp\left\{ - \int_0^t \lambda(u) du \right\} = \exp\left\{ - \sum_{j=1}^{m} \Lambda_j(t) \right\}.
\end{aligned}$$

  - A função de sobrevivência geral tem a seguinte interpretação: é a probabilidade de não ocorrência de evento por qualquer uma das causas até o tempo t.

Formulação básica

- A **função incidência acumulada** da causa j, $I_j(t)$, é definida por

$$I_j(t) = \Pr(T \leq t, D = j) = \int_0^t \lambda_j(u) S(u) du, \ j = 1, \ldots, m,$$

e representa a **probabilidade de um indivíduo falhar pela causa j na presença de todos os riscos concorrentes**.

- O estimador Kaplan-Meier da probabilidade de falhar devido a causa j até o tempo t satisfaz

$$1 - S_j(t) = \int_0^t \lambda_j(u) S_j(u) du.$$

Formulação básica

- Note que $\Lambda_j(t) \geq 0,\ j = 1, \ldots, m$, e assim

$$\begin{aligned} \Lambda_j(t) &\leq \sum_{j=1}^{m} \Lambda_j(t), \\ S_j(t) = \exp\{-\Lambda_j(t)\} &\geq \exp\left\{-\sum_{j=1}^{m} \Lambda_j(t)\right\} = S(t), \end{aligned}$$

logo

$$I_j(t) = \int_0^t \lambda_j(u) S(u) du \leq \int_0^t \lambda_j(u) S_j(u) du = 1 - S_j(t).$$

- Isto mostra o viés do estimador de Kaplan-Meier, se este é usado para estimar $I_j(t)$.

Métodos não paramétricos

Estimação da função incidência acumulada

- Seja $0 < t_1 < t_2 < \ldots < t_n$ os tempos distintos observados de falha por qualquer causa.
  - d_{jk} é o número de indivíduos que falharam da causa j no tempo t_k.
  - $d_k = \sum_{j=1}^{m} d_{jk}$ é o número total de falhas (**qualquer causa**) no tempo t_k.
  - n_k é o número de indivíduos em risco (**indivíduos ainda presentes no estudo que não falharam por qualquer causa**) no tempo t_k.
- A função incidência acumulada da causa j no tempo t pode ser estimada por

$$\hat{I}_j(t) = \sum_{k:t_k \leq t} \hat{\lambda}_j(t_k)\hat{S}(t_{k-1}),$$

em que $\hat{\lambda}_j(t_k) = \dfrac{d_{jk}}{n_k}$ e $\hat{S}(t) = \prod_{k:t_k \leq t} \left(1 - \sum_{j=1}^{m} \hat{\lambda}_j(t_k)\right)$.

Dados de gamopatia monoclonal

História natural de 241 indivíduos com gamopatia monoclonal de significado indeterminado (MGUS).

- `mgus`: A data frame with 241 observations on the following 12 variables.
  - `id`: subject id
  - `age`: age in years at the detection of MGUS
  - `sex`: male or female
  - `dxyr`: year of diagnosis
  - `pcdx`: for subjects who progress to a plasma cell malignancy the subtype of malignancy: multiple myeloma (MM) is the most common, followed by amyloidosis (AM), macroglobulinemia (MA), and other lymphprolifative disorders (LP)
  - `pctime`: days from MGUS until diagnosis of a plasma cell malignancy
  - `futime`: days from diagnosis to last follow-up
  - `death`: 1= follow-up is until death
  - `alb`: albumin level at MGUS diagnosis
  - `creat`: creatinine at MGUS diagnosis
  - `hgb`: hemoglobin at MGUS diagnosis

Dados de gamopatia monoclonal

- `mspike`: size of the monoclonal protein spike at diagnosis

- `mgus1`: The same data set in start,stop format. Contains the id, age, sex, and laboratory variable described above along with
  - `start`, `stop`: sequential intervals of time for each subject
  - `status`: =1 if the interval ends in an event
  - `event`: a factor containing the event type: censor, death, or plasma cell malignancy
  - `enum`: event number for each subject: 1 or 2

Dados de gamopatia monoclonal

```
library(survival)

head(mgus1[c("id","sex","start","stop","status","event")])
```

```
##   id    sex start stop status event
## 1  1 female     0  748      1 death
## 2  2 female     0 1310      1   pcm
## 4  3   male     0  277      1 death
## 5  4   male     0 1815      1 death
## 6  5 female     0 2587      1 death
## 7  6   male     0  563      1 death
```

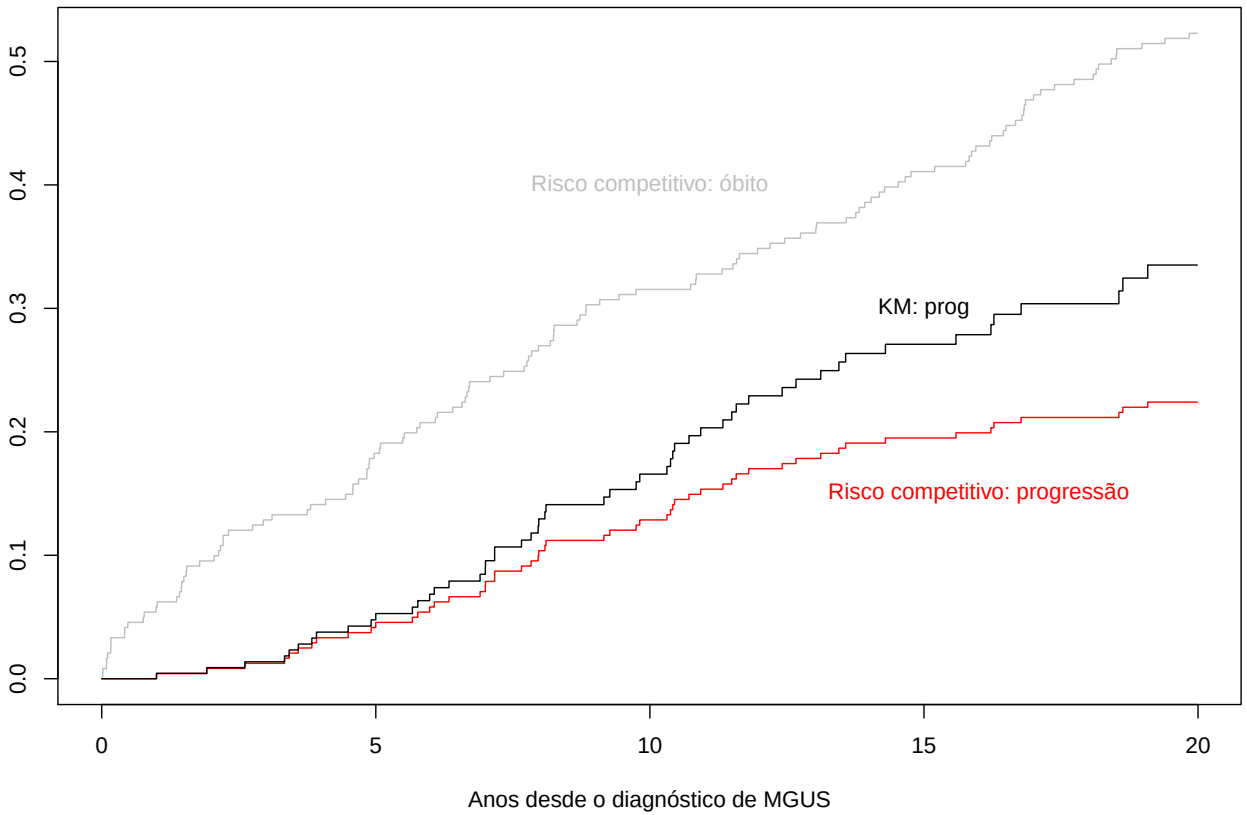
Dados de gamopatia monoclonal

```
fitKM <- survfit(Surv(time = stop, event == 'pcm') ~ 1,
                 data = mgus1, subset = (start == 0))

fitCI <- survfit(Surv(time = stop, event = status*as.numeric(event),
                      type = "mstate") ~ 1,
                   data = mgus1, subset = (start == 0))

plot(fitCI, xscale = 365.25, xmax = 7300,
     mark.time = FALSE, col = c("red", "grey"),
     xlab = "Anos desde o diagnóstico de MGUS")
lines(fitKM, fun = 'event', xscale = 365.25,
      xmax = 7300, mark.time = FALSE, conf.int = FALSE)
text(x = 10*365.25, y = .4, "Risco competitivo: óbito",
     col = "grey")
text(x = 16*365.25, .15,"Risco competitivo: progressão",
     col = "red")
text(x = 15*365.25, .30, "KM: prog")
```

Dados de gamopatia monoclonal



Comparação de funções incidência acumulada entre grupos

- **Gray (1988)**[1] desenvolveu um tipo de teste *log-rank* para testar a igualdade de curvas de incidência acumulada.
  - Pacote `cmprsk` do `R`.

```
library(cmprsk)

mgus1 <- subset(mgus1, start == 0)
mgus1$evtype <- mgus1$status * as.numeric(mgus1$event)

fitCI <- cuminc(ftime = mgus1$stop,
                fstatus = mgus1$evtype,
                group = mgus1$sex, cencode = 0)
```

[1] Robert J. Gray. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, 16:1141–1154, 1988.

Comparação de funções incidência acumulada entre grupos

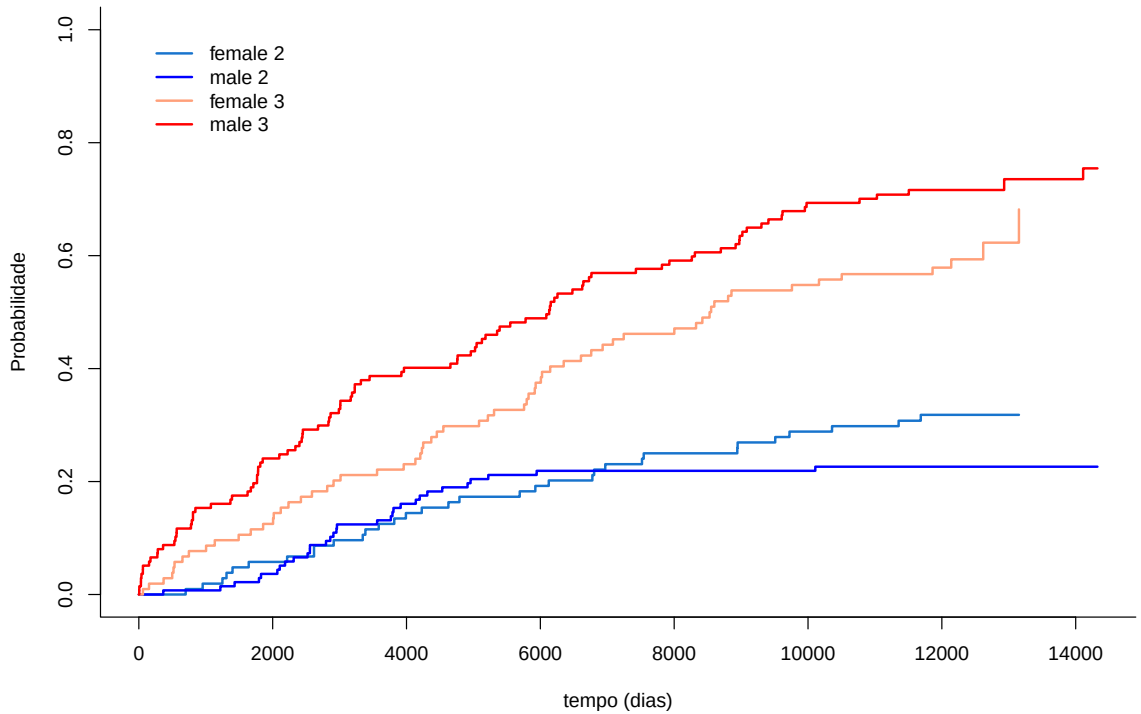
- Óbito = 3;
- Progressão = 2.

```
fitCI$Tests
```

```
##       stat        pv df
## 2 1.833497 0.1757150  1
## 3 4.791115 0.0286069  1
```

```
plot(fitCI,
     col = c("dodgerblue3", "blue",
             "lightsalmon1", "red"),
     lwd = 2, lty = 1,
     xlab = "tempo (dias)", ylab = "Probabilidade")
```

Comparação de funções incidência acumulada entre grupos



Modelos de riscos competitivos

Modelo de Cox de causa específica

- Para identificar potenciais fatores de risco para uma falha de **causa específica**, podemos utilizar o modelo de Cox, onde este modela a função de taxa de falha de causa específica

$$\lambda_j(t) = \lambda_{0j}(t) \exp\{x_1\beta_1 + \ldots + x_p\beta_p\}.$$

```
fitcph1 <- coxph(Surv(time = stop,
                      event = evtype == 2) ~ sex + age,
                 data=mgus1)
summary(fitcph1)$coef
```

```
##                coef exp(coef)   se(coef)          z  Pr(>|z|)
## sexmale -0.04175948 0.9591004 0.25207587 -0.1656623 0.8684227
## age     -0.00383109 0.9961762 0.01169479 -0.3275895 0.7432220
```

Modelo de Cox de causa específica

```
fitcph2 <- coxph(Surv(time = stop, event = evtype == 3) ~ sex + age,
                 data = mgus1)
summary(fitcph2)$coef
```

```
##                coef exp(coef)    se(coef)        z     Pr(>|z|)
## sexmale 0.27112927  1.311445 0.163199299 1.661338 9.664550e-02
## age     0.08475353  1.088449 0.008551291 9.911198 3.721572e-23
```

Modelo de Cox de causa específica

- O pacote `riskRegression`, por meio da função `CSC`, "envelopa" os ajustes dos modelos de Cox de causa específica:

```
library(riskRegression)

fitcox <- CSC(Hist(stop, evtype) ~ sex + age,
              data = mgus1)
```

Modelo de Cox de causa específica

```
print(fitcox)
```

```
## CSC(formula = Hist(stop, evtype) ~ sex + age, data = mgus1)
##
## Right-censored response of a competing.risks model
##
## No.Observations: 241
##
## Pattern:
##
## Cause      event right.censored
##   2           64              0
##   3          163              0
##   unknown      0             14
##
##
## ----------> Cause:  2
##
## Call:
## survival::coxph(formula = survival::Surv(time, status) ~ sex +
##     age, x = TRUE, y = TRUE)
##
```

Modelo de Cox de causa específica

```
##   n= 241, number of events= 64
##
##               coef exp(coef)  se(coef)      z Pr(>|z|)
## sexmale -0.041759  0.959100  0.252076 -0.166    0.868
## age     -0.003831  0.996176  0.011695 -0.328    0.743
##
##         exp(coef) exp(-coef) lower .95 upper .95
## sexmale    0.9591      1.043    0.5852     1.572
## age        0.9962      1.004    0.9736     1.019
##
## Concordance= 0.508  (se = 0.038 )
## Likelihood ratio test= 0.15  on 2 df,   p=0.9
## Wald test            = 0.15  on 2 df,   p=0.9
## Score (logrank) test = 0.15  on 2 df,   p=0.9
##
##
##
## ----------> Cause:  3
##
## Call:
## survival::coxph(formula = survival::Surv(time, status) ~ sex +
##     age, x = TRUE, y = TRUE)
```

Modelo de Cox de causa específica

```
##
##   n= 241, number of events= 163
##
##              coef exp(coef) se(coef)     z Pr(>|z|)
## sexmale 0.271129  1.311445 0.163199 1.661   0.0966 .
## age     0.084754  1.088449 0.008551 9.911   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## sexmale     1.311     0.7625    0.9524     1.806
## age         1.088     0.9187    1.0704     1.107
##
## Concordance= 0.731  (se = 0.022 )
## Likelihood ratio test= 119.6  on 2 df,   p=<2e-16
## Wald test            = 105.6  on 2 df,   p=<2e-16
## Score (logrank) test = 113  on 2 df,   p=<2e-16
```

Modelo de Cox de causa específica

- A função de taxa de falha de causa específica não possui uma interpretação direta em termos de probabilidade de sobrevivência para um tipo de falha em particular.
- O efeito da covariável na função de taxa de falha pode ser muito diferente do efeito da covariável na função incidência acumulada.

Modelo de Fine e Gray

- **Fine e Gray (1999)**[2] introduziram um modelo que relaciona diretamente as covariáveis a função incidência acumulada. Seja

$$\bar{\lambda}_1(t) = -\frac{d}{dt}\left[\log(1 - I_1(t))\right],$$

a função ***subdistribution hazard***. Dado um conjunto de covariáveis, o modelo assume a seguinte forma

$$\bar{\lambda}_1(t) = \bar{\lambda}_{1,0}(t)\exp\{x_1\beta_1 + \ldots + x_p\beta_p\}.$$

- Desta forma, temos

$$I_1(t) = 1 - \exp\left\{-\exp\{x_1\beta_1 + \ldots + x_p\beta_p\}\int_0^t \bar{\lambda}_{1,0}(u)du\right\}.$$

[2] Jason P. Fine and Robert J. Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94:496–509, 1999.

Modelos de riscos competitivos: Modelo de Fine e Gray

- A função `FGR` do pacote `riskRegression` estima os coeficientes de regressão do modelo de Fine e Gray.

```
fitfg <- FGR(prodlim::Hist(stop, evtype) ~ sex + age,
             data = mgus1, cause = 2)
```

Modelos de riscos competitivos: Modelo de Fine e Gray

```
print(fitfg)
```

```
##
## Right-censored response of a competing.risks model
##
## No.Observations: 241
##
## Pattern:
##
## Cause      event right.censored
##   2           64              0
##   3          163              0
##   unknown      0             14
##
##
## Fine-Gray model: analysis of cause 2
##
## Competing Risks Regression
##
## Call:
```

Modelos de riscos competitivos: Modelo de Fine e Gray

```
## FGR(formula = prodlim::Hist(stop, evtype) ~ sex + age, data = mgus1,
##     cause = 2)
##
##             coef exp(coef) se(coef)      z p-value
## sexmale -0.1900     0.827  0.25449 -0.747 4.6e-01
## age     -0.0387     0.962  0.00924 -4.192 2.8e-05
##
##         exp(coef) exp(-coef)  2.5% 97.5%
## sexmale     0.827       1.21 0.502  1.36
## age         0.962       1.04 0.945  0.98
##
## Num. cases = 241
## Pseudo Log-likelihood = -334
## Pseudo likelihood ratio test = 15.4  on 2 df,
##
## Convergence: TRUE
```

Considerações finais do curso

Considerações finais do curso

- A análise de sobrevivência é uma grande área de estudo na (bio)estatística.
- Neste curso enfocamos nas técnicas e modelos mais utilizados, bem como em suas aplicações.
- Nem sempre estas técnicas e modelos serão adequados para o problema de estudo:
  - Não responde a questão de investigação;
  - As suposições não são razoáveis;
  - Os modelos não se ajustam aos dados.

Considerações finais do curso

- Nestes casos, técnicas e modelos avançados em análise de sobrevivência existem (e estão a se desenvolver) na literatura da área.
- Algumas destas técnicas envolvem:
  - Modelos aditivos;
  - Censura intervalar;
  - Modelos discretos;
  - Eventos recorrentes;
  - Processos de contagem;
  - Modelos de efeitos aleatórios;
  - Censura dependente.

Para casa

- Atividade de avaliação II.
  - Será postada no Moodle logo em seguida.
  - O professor está a disposição para esclarecimento de dúvidas com relação à atividade.

Por hoje é só!

Bons estudos! Bom final de ano! Até a próxima!