# EPI10 - Análise 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



Relembrando

#### Relembrando

#### Modelo de regressão de Cox

Assume-se, nesse modelo, que os tempos  $t_i$ ,  $i=1,\ldots,n$ , são independentes e que a taxa de falha (risco) tem a seguinte forma:

$$\lambda(t) = \lambda_0(t) \exp{\{\beta_1 x_1 + \ldots + \beta_p x_p\}}.$$

- ▶ O componente não-paramétrico,  $\lambda_0(t)$ , não é especificado e é uma função não-negativa do tempo.
  - Ele é usualmente chamado de função de taxa de falha basal.
- ▶ O componente paramétrico  $\exp\{x'\beta\} = \exp\{\beta_1x_1 + \ldots + \beta_px_p\}$  é o nosso interesse, em especial no vetor de parâmetros  $\beta$ , e  $x' = (x_1, x_2, \ldots, x_p)$  é um vetor de covariáveis observadas (como, por exemplo, sexo, idade, grupo de tratamento ou espoxição, etc.).

#### Modelo de regressão de Cox

- Estimação pelo método da máxima verossimilhança parcial.
  - Intervalos de confiança e testes de hipóteses podem ser construídos para cada componente  $\beta_r$  do vetor de coeficientes  $\beta$ .
- ► Hazard ratio ( $HR_r = e^{\beta_r}$ ) expressa a razão entre taxas de falha entre dois grupos (definidos por alguma variável de tratamento ou exposição).
  - Por conta da proporcionalidade dos riscos, HR não depende do tempo t (constante ao longo do tempo).
  - HR = 1 ( $\beta = 0$ ): a covariável não influencia na função de taxa de falha.
  - HR > 1 ( $\beta > 0$ ): a covariável acelera a função de taxa de falha.
  - $ightharpoonup HR < 1 \ (eta < 0)$ : a covariável desacelera a função de taxa de falha.

Exemplo

## Exemplo

- Neste exemplo, os dados considerados referem-se a um estudo realizado com 90 pacientes do sexo masculino diagnosticados no período de 1970 a 1978 com câncer de laringe e que foram acompanhados até 01/01/1983.
- Para cada paciente, foram registrados, no diagnóstico:
  - a idade (em anos);
  - o estágio da doença (ordenados por grau de severidade da doença):
    - I. tumor primário;
      - II. envolvimento de nódulos;
      - III. metástases;
      - IV. combinações dos 3 estágios anteriores.
  - tempos de óbito ou censura (em meses).

```
## id tempos cens idade estagio
## 1 1 0.6 1 77 1
## 2 2 1.3 1 53 1
## 3 3 2.4 1 45 1
## 4 4 3.2 1 58 1
## 5 5 3.3 1 76 1
## 6 6 3.5 1 43
```

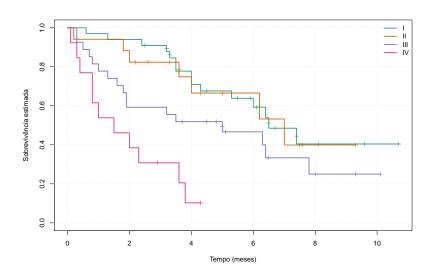
```
str(df.laringe)
## 'data.frame':
                   90 obs. of 5 variables:
   $ id
             : int 1 2 3 4 5 6 7 8 9 10 ...
    $ tempos : num 0.6 1.3 2.4 3.2 3.3 3.5 3.5 4 4 4.3 ...
   $ cens : int 1 1 1 1 1 1 1 1 1 1 ...
##
## $ idade : int 77 53 45 58 76 43 60 52 63 86 ...
   $ estagio: int 1 1 1 1 1 1 1 1 1 ...
summary(df.laringe)
##
         id
                                                         idade
                       tempos
                                                                        estagio
                                         cens
    Min.
           : 1.00
                   Min.
                          : 0.100
                                    Min.
                                           :0.0000
                                                     Min.
                                                            :41.00
                                                                     Min.
                                                                            :1.000
    1st Qu.:23.25
                   1st Qu.: 2.000
                                    1st Qu.:0.0000
                                                     1st Qu.:57.00
                                                                     1st Qu.:1.000
   Median :45.50
                   Median: 4.000
                                    Median :1.0000
                                                     Median :65.00
                                                                     Median :2.000
   Mean
        :45.50
                   Mean : 4.198
                                    Mean
                                           :0.5556
                                                     Mean :64.61
                                                                     Mean
                                                                            :2.222
##
   3rd Qu.:67.75
##
                   3rd Qu.: 6.200
                                    3rd Qu.:1.0000
                                                     3rd Qu.:72.00
                                                                     3rd Qu.:3.000
##
   Max.
           :90.00
                   Max.
                           :10.700
                                    Max.
                                           :1.0000
                                                     Max.
                                                            :86.00
                                                                     Max.
                                                                            :4.000
```

```
df.laringe$estagio <- factor(x = df.laringe$estagio,
                            levels = 1:4.
                            labels = c("I", "II", "III", "IV"))
str(df.laringe)
## 'data frame': 90 obs. of 5 variables:
   $ id : int 1 2 3 4 5 6 7 8 9 10 ...
##
##
   $ tempos : num 0.6 1.3 2.4 3.2 3.3 3.5 3.5 4 4 4.3 ...
   $ cens : int 1 1 1 1 1 1 1 1 1 ...
##
##
   $ idade : int 77 53 45 58 76 43 60 52 63 86 ...
   $ estagio: Factor w/ 4 levels "I","II","III",..: 1 1 1 1 1 1 1 1 1 1 ...
##
summary(df.laringe)
```

```
##
         id
                                                     idade
                      tempos
                                                                estagio
                                      cens
   Min.
        : 1.00
                                                        :41.00
                                                                T :33
##
                  Min.
                         : 0.100
                                  Min.
                                        :0.0000
                                                 Min.
##
   1st Qu.:23.25 1st Qu.: 2.000
                                  1st Qu.:0.0000
                                                 1st Qu.:57.00
                                                                II:17
   Median :45.50
                 Median : 4.000
                                  Median :1.0000
                                                 Median :65.00
                                                                TTT:27
##
   Mean :45.50 Mean : 4.198
                                  Mean : 0.5556
                                                 Mean :64.61
                                                                TV:13
##
##
   3rd Qu.:67.75 3rd Qu.: 6.200
                                  3rd Qu.:1.0000
                                                 3rd Qu.:72.00
                         :10.700
                                                        :86.00
##
   Max.
          :90.00
                  Max.
                                  Max.
                                        :1.0000
                                                 Max.
```

```
library(survival)
ekm <- survfit(Surv(time = tempos, event = cens) ~ estagio,</pre>
             data = df.laringe,
             conf.type = "log-log")
ekm
## Call: survfit(formula = Surv(time = tempos, event = cens) ~ estagio,
      data = df.laringe, conf.type = "log-log")
##
##
##
            n events median 0.95LCL 0.95UCL
## estagio=I 33 15 6.5 4.3
                                       NΑ
## estagio=II 17 7 7.0 3.6 NA
## estagio=III 27 17 5.0 1.6 7.8
## estagio=IV 13 11 1.5 0.4 3.6
```

```
plot(ekm, conf.int = FALSE,
     mark.time = TRUE,
     col = c("#1B9E77", "#D95F02",
             "#7570B3", "#E7298A"),
    lwd = 2, xlab = "Tempo (meses)",
     ylab = "Sobrevivência estimada")
abline(h = seq(0, 1, by = 0.2),
       v = seq(0, 10, by = 2),
       col = "lightgrey", lty = 3)
legend("topright",
       c("I", "II", "III", "IV"),
       col = c("#1B9E77", "#D95F02",
             "#7570B3", "#E7298A"),
       lwd = 2, bty = "n")
```



```
survdiff(Surv(time = tempos, event = cens) ~ estagio,
             data = df.laringe)
## Call:
## survdiff(formula = Surv(time = tempos, event = cens) ~ estagio,
      data = df.laringe)
##
##
##
             N Observed Expected (0-E)^2/E (0-E)^2/V
## estagio=I 33
                    15
                          22.57
                                  2.537 4.741
               7 10.01 0.906 1.152
## estagio=II 17
## estagio=III 27 17 14.08 0.603 0.856
## estagio=IV 13 11 3.34 17.590 19.827
##
##
   Chisq= 22.8 on 3 degrees of freedom, p= 5e-05
```

```
mod1 <- coxph(Surv(time = tempos, event = cens) ~ estagio,</pre>
            data = df.laringe, method = "breslow")
summary(mod1)
## Call:
## coxph(formula = Surv(time = tempos, event = cens) ~ estagio,
      data = df.laringe. method = "breslow")
##
##
##
   n= 90, number of events= 50
##
             coef exp(coef) se(coef) z Pr(>|z|)
##
## estagioII 0.06576 1.06797 0.45844 0.143 0.8859
## estagioIV 1.72284 5.60040 0.41966 4.105 4.04e-05 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.05 '.' 0.1 ' ' 1
##
##
           exp(coef) exp(-coef) lower .95 upper .95
## estagioII 1.068 0.9364 0.4348 2.623
## estagioIII 1.844 0.5422 0.9193 3.700
## estagioIV 5.600 0.1786 2.4604 12.748
```

```
##
## Concordance= 0.668 (se = 0.037)
## Likelihood ratio test= 16.26 on 3 df, p=0.001
## Wald test = 18.95 on 3 df, p=3e-04
## Score (logrank) test = 22.46 on 3 df, p=5e-05
```

```
mod2 <- coxph(Surv(time = tempos, event = cens) ~ idade,
              data = df.laringe, method = "breslow")
summary(mod2)
## Call:
## coxph(formula = Surv(time = tempos, event = cens) ~ idade, data = df.laringe
      method = "breslow")
##
##
    n= 90. number of events= 50
##
##
##
           coef exp(coef) se(coef) z Pr(>|z|)
## idade 0.02318    1.02345    0.01447    1.602    0.109
##
        exp(coef) exp(-coef) lower .95 upper .95
##
## idade
            1.023
                     0.9771 0.9948
                                          1.053
##
## Concordance= 0.555 (se = 0.046)
## Likelihood ratio test= 2.61 on 1 df. p=0.1
## Wald test = 2.57 on 1 df, p=0.1
## Score (logrank) test = 2.58 on 1 df, p=0.1
```

```
mod3 <- coxph(Surv(time = tempos, event = cens) ~ estagio + idade,</pre>
             data = df.laringe, method = "breslow")
summary(mod3)
## Call:
## coxph(formula = Surv(time = tempos, event = cens) ~ estagio +
      idade, data = df.laringe, method = "breslow")
##
##
##
    n= 90. number of events= 50
##
              coef exp(coef) se(coef) z Pr(>|z|)
##
## estagioII 0.13856 1.14862 0.46231 0.300 0.764
## estagioIII 0.63835 1.89335 0.35608 1.793 0.073 .
## estagioIV 1.69306 5.43607 0.42221 4.010 6.07e-05 ***
## idade 0.01890 1.01908 0.01425 1.326 0.185
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.05 '.' 0.1 ' ' 1
##
##
            exp(coef) exp(-coef) lower .95 upper .95
## estagioII 1.149 0.8706 0.4642 2.842
## estagioIII 1.893 0.5282 0.9422 3.805
```

```
## estagioIV 5.436 0.1840 2.3763 12.436

## idade 1.019 0.9813 0.9910 1.048

##

## Concordance= 0.682 (se = 0.039)

## Likelihood ratio test= 18.07 on 4 df, p=0.001

## Wald test = 20.82 on 4 df, p=3e-04

## Score (logrank) test = 24.33 on 4 df, p=7e-05
```

- $\triangleright$  Os coeficientes de regressão  $\beta$  são as quantidades de maior interesse na modelagem estatística de dados de sobrevivência.
- Entretanto, funções relacionadas a  $\lambda_0(t)$  são também importantes no modelo de Cox.
- Estas funções referem-se referem-se basicamente à função de taxa de falha acumulada de base

$$\Lambda_0(t) = \int_0^t \lambda_0(u) du$$

e à correspondente função de sobrevivência

$$S_0(t) = \exp\left\{-\Lambda_0(t)\right\}.$$

- A maior importância destas funõees diz respeito ao uso delas em técnicas gráficas para avaliar a adequação do modelo ajustado.
- A função de sobrevivência

$$S(t) = [S_0(t)]^{\exp\{x'\beta\}}$$

é também útil quando se deseja concluir a análise em termos de percentis associados a grupos de indivíduos.

Ou, como veremos na próxima aula, quando se deseja estimar a função de sobrevivência em um certo tempo t especificado.

- Se  $\lambda_0(t)$  fosse especificado parametricamente, poderia ser estimado usando a função de verossimilhança.
- Entretanto, na verossimilhança parcial, o argumento condicional elimina completamente esta função.
- Desta forma, os estimadores para estas quantidades serão de natureza não-paramétrica.
- ▶ Uma estimativa simples para  $\Lambda_0(t)$ , proposta por Breslow (1972)<sup>1</sup>, é uma função escada com saltos nos tempos distintos de falha e expressa por

$$\widehat{\Lambda}_0(t) = \sum_{i:t:\leq t} \frac{d_i}{\sum_{l\in R_i} \exp\{x'\widehat{\beta}\}},$$

em que  $d_i$  é o número de falhas em  $t_i$ .

<sup>&</sup>lt;sup>1</sup>Breslow, N. (1972), Discussion on Professor Cox's Paper. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34: 202-220.

Consequentemente, as funções de sobrevivência  $S_0(t)$  e S(t) podem ser estimada a partir da expressão acima por

$$\widehat{S}_0(t) = \exp\left\{-\widehat{\Lambda}_0(t)
ight\}.$$

e

$$\widehat{S}(t) = [\widehat{S}_0(t)]^{\exp\{x'\widehat{\beta}\}}.$$

#### Comentários

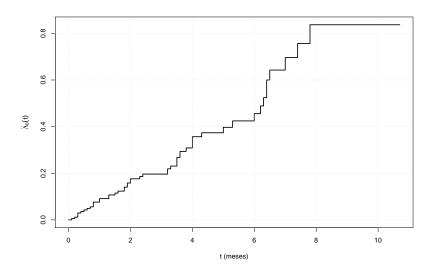
- ▶ Tanto  $\widehat{S}_0(t)$  quanto  $\widehat{S}(t)$  são funções escada decrescentes com o tempo.
- Na ausência de covariáveis (x'=0), a expressão de  $\widehat{\Lambda}_0(t)$  reduz-se a

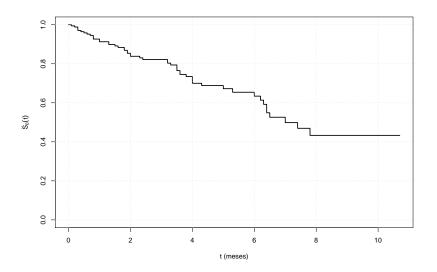
$$\widehat{\Lambda}_0(t) = \sum_{j:t_i < t} \left( \frac{d_j}{n_j} \right).$$

- Esta expressão acima é conhecida como o estimador de Nelson-Aalen.
  - $\tilde{S}(t) = \exp\{-\widehat{\Lambda}(t)\}$ , em que  $\widehat{\Lambda}(t) = \sum_{j:t_j < t} (d_j/n_j)$ , é uma estimativa da função de sobrevivência com base no estimador de Nelson-Aalen, e é um estimador alternativo ao estimador de Kaplan-Meier.

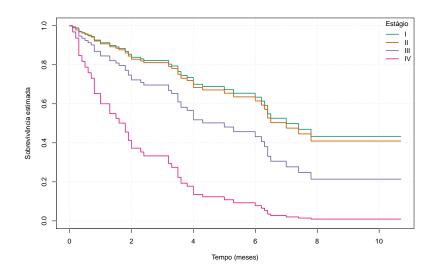
```
plot(survfit(mod1),
    cumhaz = TRUE,
    conf.int = FALSE,
    lwd = 2, xlab = "t (meses)",
    ylab = expression(hat(Lambda)[0](t)))

abline(h = seq(0, 1, by = 0.2),
    v = seq(0, 10, by = 2),
    col = "lightgrey", lty = 3)
```

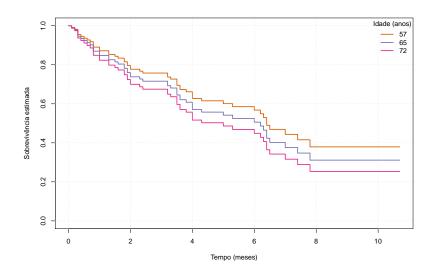




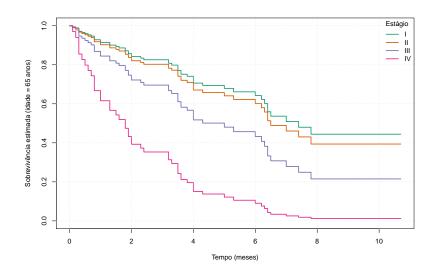
```
df.novo <- data.frame(
 estagio = levels(df.laringe$estagio))
plot(survfit(mod1, newdata = df.novo),
     col = c("#1B9E77", "#D95F02",
             "#7570B3", "#E7298A").
     lwd = 2, xlab = "Tempo (meses)",
     ylab = "Sobrevivência estimada")
abline(h = seq(0, 1, by = 0.2),
       v = seq(0, 10, by = 2),
       col = "lightgrey", lty = 3)
legend("topright",
       c("I", "II", "III", "IV"),
       title = "Estágio",
       col = c("#1B9E77", "#D95F02",
            "#7570B3", "#E7298A"),
       lwd = 2, bty = "n")
```



```
df.novo <- data.frame(</pre>
  idade = c(57, 65, 72))
plot(survfit(mod2, newdata = df.novo),
     col = c("#D95F02".
             "#7570B3", "#E7298A").
     lwd = 2, xlab = "Tempo (meses)",
     vlab = "Sobrevivência estimada")
abline(h = seq(0, 1, by = 0.2),
       v = seq(0, 10, by = 2),
       col = "lightgrey", lty = 3)
legend("topright",
       legend = c(57, 65, 72),
       col = c("#D95F02",
             "#7570B3". "#E7298A").
       title = "Idade (anos)",
       lwd = 2, bty = "n")
```



```
df.novo <- data.frame(
 idade = 65,
 estagio = levels(df.laringe$estagio))
plot(survfit(mod3, newdata = df.novo),
     col = c("#1B9E77", "#D95F02",
             "#7570B3", "#E7298A"),
     lwd = 2. xlab = "Tempo (meses)".
     ylab = "Sobrevivência estimada (idade = 65 anos)")
abline(h = seq(0, 1, by = 0.2),
       v = seq(0, 10, by = 2),
       col = "lightgrey", lty = 3)
legend("topright",
       c("I", "II", "III", "IV"),
       title = "Estágio".
       col = c("#1B9E77", "#D95F02",
             "#7570B3", "#E7298A"),
      lwd = 2, bty = "n")
```



Adequação do Modelo de Cox

#### Adequação do Modelo de Cox

#### Adequação do Modelo de Cox

- O modelo de regressão de Cox é bastante flexível devido à presença do componente não-paramétrico.
- Mesmo assim, ele não se ajusta a qualquer situação e como qualquer outro modelo estatéstico, requer o uso de técnicas para avaliar a sua adequação.
- Em particular, a suposição de riscos proporcionais.
  - A violaçãao desta suposição pode acarretar sérios viéses na estimação dos coeficientes do modelo.

### Adequação do Modelo de Cox

- Diversos métodos para avaliar a adequação deste modelo encontram-se disponíveis na literatura.
- Estes baseiam-se, essencialmente, em análise de resíduos.
- Alguns desses métodos são apresentados a seguir.

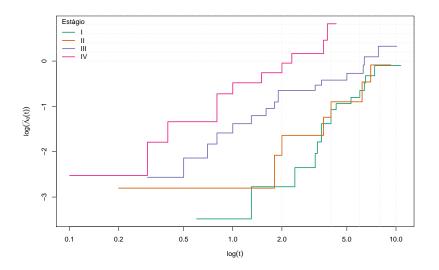
Avaliação da proporcionalidade dos riscos

Avaliação da proporcionalidade dos riscos

- Para verificar a suposição de riscos proporcionais no modelo de Cox, um gráfico simples e bastante usado é obtido, inicialmente, dividindo os dados em m estratos, usualmente de acordo com alguma covariável.
  - Por exemplo, dividir os dados em dois estratos de acordo com a covariável sexo.
- ► Em seguida, deve-se estimar  $\widehat{\Lambda}_0(t)$  para cada estrato.

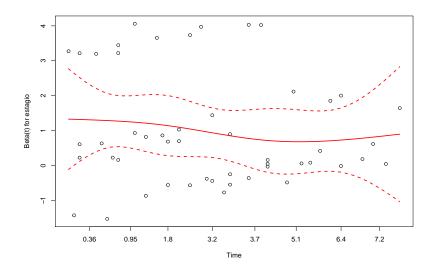
- Se a suposião for válida, as curvas do logaritmo de  $\widehat{\Lambda}_0(t)$  versus t, ou  $\log(t)$ , devem apresentar diferenças aproximadamente constantes no tempo.
  - Curvas não paralelas significam desvios da suposição de riscos proporcionais.
- ► É razoável construir este gra´fico para cada covariável incluída no estudo.
  - Se a covariável for de natureza contínua, uma sugestão é agrupá-la em um pequeno número de categorias.
- Situações extremas de violação da suposição ocorrem quando as curvas se cruzam.

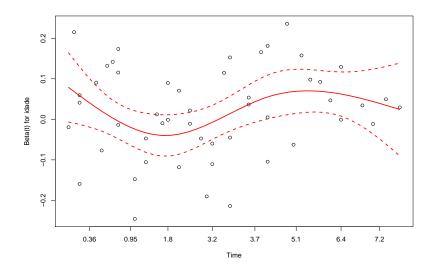
```
plot(ekm,
    fun = "cloglog",
     conf.int = FALSE.
     col = c("#1B9E77", "#D95F02",
             "#7570B3", "#E7298A"),
    lwd = 2, xlab = expression(log*(t)),
     ylab = expression(log*(hat(Lambda)[0]*(t))))
abline(h = seq(0, 1, by = 0.2),
       v = seq(0, 10, by = 2),
       col = "lightgrey", lty = 3)
legend("topleft",
       c("I", "II", "III", "IV"),
       title = "Estágio",
       col = c("#1B9E77", "#D95F02",
             "#7570B3", "#E7298A"),
      lwd = 2, bty = "n")
```



- Uma proposta adicional de análise da suposição de riscos proporcionais é fazer uso dos resíduos de Schoenfeld.
- Existe um conjunto de resíduos para cada covariável.
- Usar o gráfico dos resíduos padronizados contra o tempo para cada covariável.
- Inclinação zero mostra evidência a favor da proporcionalidade dos riscos.

```
plot(cox.zph(mod3),
    col = "red",
    lwd = 2)
```





## Medidas estatísticas e testes de hipóteses

- As técnicas gráficas envolvem uma interpretação com carácter subjetivo.
- ► Testes de hipóteses podem auxiliar neste processo de decisão.
- O coeficiente de correlação de Pearson (ρ) entre os resíduos padronizados de Schoenfeld e g(t) para cada covariável é uma dessas medidas.
- Valores de  $\rho$  próximos de zero mostram evidências em favor da suposição de riscos proporcionais.
- Um teste hipóteses global de proporcionalidade de riscos sobre todas as covariávies no modelo pode ser realizado.

### Medidas estatísticas e testes de hipóteses

```
cox.zph(mod3)
```

```
## chisq df p
## estagio 3.67 3 0.30
## idade 1.12 1 0.29
## GLOBAL 5.07 4 0.28
```

# Considerações finais

- Resíduos martingale e deviance podem ser obtidos para a avaliação de outros aspectos do modelo de Cox, tais como:
  - pontos atípicos
  - forma funcional da relação das covariáveis (não linearidade, por exemplo)
  - pontos influentes
- Retornaremos a estas técnicas nas próximas aulas, quando também discutiremos alternativas ao modelo de Cox quando a suposição de riscos proporcionais é violada.

#### Para casa

- 1. Leia o capítulo 5 do livro **Análise de sobrevivência aplicada**<sup>2</sup>.
- Leia os capítulo 6 e 7 do livro Análise de sobrevivência: teoria e aplicações em saúde<sup>3</sup>.

<sup>&</sup>lt;sup>2</sup>Colosimo, E. A. e Giolo, S. R. **Análise de sobrevivência aplicada**, Blucher, 2006.

<sup>&</sup>lt;sup>3</sup>Carvalho, M. S., Andreozzi, V. L., Codeço, C. T., Campos, D. P., Barbosa, M. T. S. e Shimakura, E. S. **Análise de sobrevivência: teoria e aplicações em saúde**, 2ª ed. Editora Fiocruz, 2011.

#### Próxima aula

► Aplicações com o modelo de Cox.

## Por hoje é só!

#### Bons estudos!

