EPI10 - Análise de Sobrevivência

Modelo de Cox

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, 2022

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 violação 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

Método gráfico descritivo

- 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.

Método gráfico descritivo

- 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 gráfico para cada covariável incluída no modelo.
  - 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.

Método gráfico descritivo

No R, utilizando os dados do estudo sobre câncer de laringe, temos:

```
library(survival)

ekm <- survfit(Surv(time = tempos, event = cens) ~ estagio,
               data = df.laringe,
               conf.type = "log-log")

ekm
```

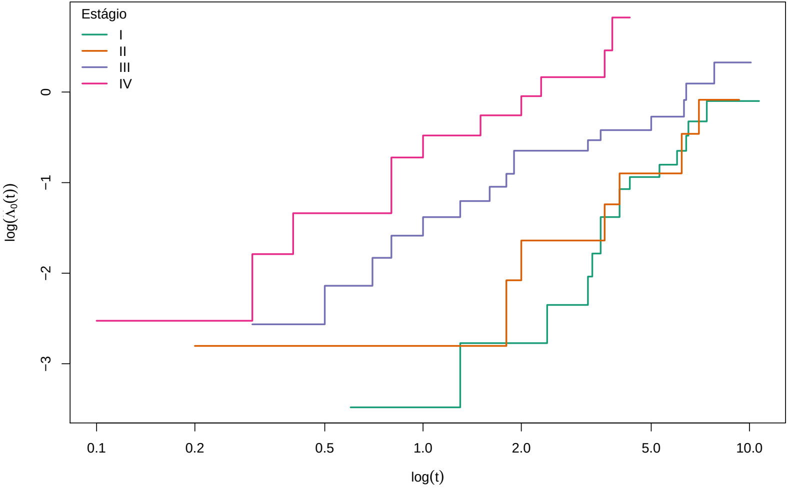
```
## Call: survfit(formula = Surv(time = tempos, event = cens) ~ estagio,
##     data = df.laringe, conf.type = "log-log")
##
##               n events median 0.95LCL 0.95UCL
## estagio=I    33     15    6.5     4.3      NA
## estagio=II   17      7    7.0     3.6      NA
## estagio=III  27     17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

Método gráfico descritivo

```
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))))

legend("topleft",
       c("I", "II", "III", "IV"),
       title = "Estágio",
       col = c("#1B9E77", "#D95F02",
             "#7570B3", "#E7298A"),
       lwd = 2, bty = "n")
```

Método gráfico descritivo



Método com coeficiente dependente do tempo

- 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 apresenta evidência a favor da proporcionalidade dos riscos.

Método com coeficiente dependente do tempo

No R:

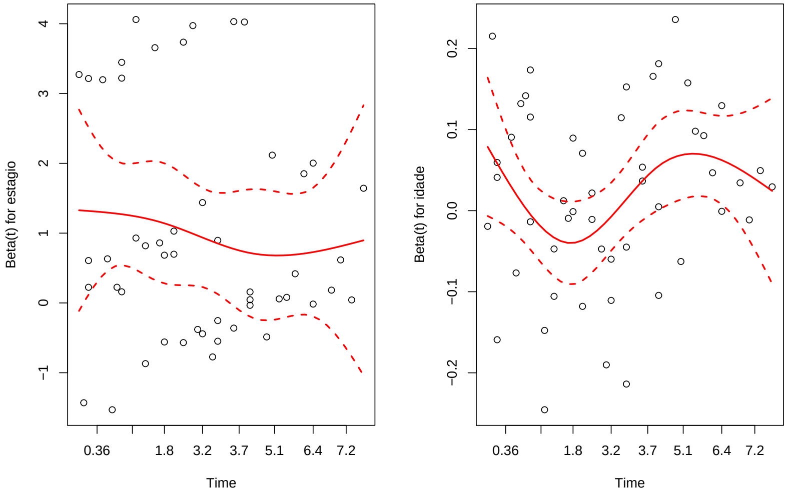
```
mod1 <- coxph(Surv(time = tempos, event = cens) ~ estagio,
              data = df.laringe, method = "breslow")

mod2 <- coxph(Surv(time = tempos, event = cens) ~ idade,
              data = df.laringe, method = "breslow")

mod3 <- coxph(Surv(time = tempos, event = cens) ~ estagio + idade,
              data = df.laringe, method = "breslow")

par(mfrow = c(1,2))
plot(cox.zph(mod3),
     col = "red",
     lwd = 2)
```

Método com coeficiente dependente do tempo



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 $g(t)$[1] para cada covariável é uma dessas medidas.
- Valores de ρ 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.

[1] A função $g(\cdot)$ pode ser especificada tal como $g(t) = t$ ou $g(t) = \log(t)$, por exemplo.

Medidas estatísticas e testes de hipóteses

No R:

```
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.

Comparação de modelos

Teste da razão de verossimilhanças

- **Testes da razão de verossimilhanças**, são comuns e amplamente usados, especialmente ao comparar **modelos aninhados** que diferem em relação a múltiplos parâmetros.
- Por exemplo, nos dados do exemplo do estudo de câncer de laringe, suponha que desejamos comparar o modelo 1 (`estagio`) versus o modelo 3 (`estagio + idade`).
  - Ou ainda, o modelo 3 como um modelo com efeito de interação entre `estagio + idade`.

Teste da razão de verossimilhanças

- Denote por $\widehat{\beta}_0$ o vetor de parâmetros estimados do primeiro modelo **("modelo reduzido")** e $\widehat{\beta}_1$ o vetor de parâmetros estimados do segundo modelo **("modelo cheio")**, o teste da razão de verossimilhança (que requer apenas o ajuste de dois modelos) é baseado em

$$2[\ell(\widehat{\beta}_1) - \ell(\widehat{\beta}_0)] \overset{a}{\sim} \chi^2_p,$$

em que $\ell(\beta)$ é função de log-verossimilhança do modelo, e p é o número de graus de liberdade da distribuição qui-quadrado, e é definido pela diferença do número de parâmetros dos dois modelos.

Teste da razão de verossimilhanças

No R:

```
anova(mod1, mod3)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time = tempos, event = cens)
##  Model 1: ~ estagio
##  Model 2: ~ estagio + idade
##    loglik  Chisq Df P(>|Chi|)
## 1 -189.08
## 2 -188.18 1.8036  1    0.1793
```

Teste da razão de verossimilhanças

```
mod4 <- coxph(Surv(time = tempos, event = cens) ~ estagio*idade,
              data = df.laringe, method = "breslow")

summary(mod4)
```

```
## 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        -7.946142  0.000354  3.678209 -2.160   0.0307 *
## estagioIII       -0.122500  0.884706  2.468331 -0.050   0.9604
## estagioIV         0.846986  2.332605  2.425717  0.349   0.7270
## idade            -0.002559  0.997444  0.026051 -0.098   0.9218
## estagioII:idade   0.120254  1.127783  0.052307  2.299   0.0215 *
## estagioIII:idade  0.011351  1.011416  0.037449  0.303   0.7618
## estagioIV:idade   0.013673  1.013767  0.035967  0.380   0.7038
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Teste da razão de verossimilhanças

```
##                    exp(coef) exp(-coef) lower .95 upper .95
## estagioII           0.000354  2824.6562 2.619e-07    0.4786
## estagioIII          0.884705     1.1303 7.011e-03  111.6467
## estagioIV           2.332605     0.4287 2.009e-02  270.7790
## idade               0.997444     1.0026 9.478e-01    1.0497
## estagioII:idade     1.127783     0.8867 1.018e+00    1.2495
## estagioIII:idade    1.011416     0.9887 9.398e-01    1.0884
## estagioIV:idade     1.013767     0.9864 9.448e-01    1.0878
##
## Concordance= 0.694  (se = 0.039 )
## Likelihood ratio test= 24.27  on 7 df,   p=0.001
## Wald test            = 24.11  on 7 df,   p=0.001
## Score (logrank) test = 28.59  on 7 df,   p=2e-04
```

```
anova(mod3, mod4)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time = tempos, event = cens)
##  Model 1: ~ estagio + idade
##  Model 2: ~ estagio * idade
##    loglik  Chisq Df P(>|Chi|)
## 1 -188.18
## 2 -185.08 6.2039  3    0.1021
```

Teste da razão de verossimilhanças

(Um parênteses sobre interações)

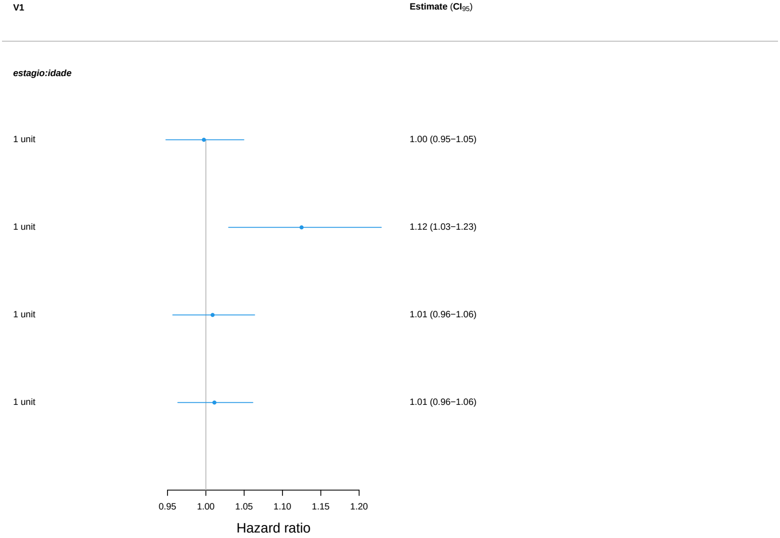
```
# install.packages(Publish)
library(Publish)

# publish(mod4)

plot(publish(mod4))
```

```
##              Variable Units HazardRatio        CI.95   p-value
##    idade: estagio(I)              1.00 [0.95;1.05]   0.92175
##   idade: estagio(II)              1.12 [1.03;1.23]   0.00915
##  idade: estagio(III)              1.01 [0.96;1.06]   0.74503
##   idade: estagio(IV)              1.01 [0.96;1.06]   0.65303
```

Teste da razão de verossimilhanças



Teste da razão de verossimilhanças

- Modelos não aninhados podem ser comparados pelos **critérios de informação de Akaike (AIC)** ou **Bayesiano (BIC)**.
  - Ao comparar modelos ajustados por máxima verossimilhança para os mesmos dados, quanto menor o AIC (ou BIC), melhor é o ajuste.

```
AIC(mod1)
```

```
## [1] 384.1625
```

```
AIC(mod2)
```

```
## [1] 393.8118
```

```
BIC(mod1)
```

```
## [1] 389.8986
```

```
BIC(mod2)
```

```
## [1] 395.7238
```

Efeitos não-lineares

Splines

- Existem diversas formas de introduzir efeitos não-lineares em modelos de regressão.
- As funções *spline* são uma forma elegante e eficiente de incorporar não-linearidades nos modelos de regressão.
  - As funções *spline* não são indicadas apenas para o modelo de Cox.
- Neste breve exemplo, vamos apenas considerar dois tipos de funções *spline* e suas respectivas implementações em `R`:
  - **P-splines**;
  - **Natural splines** (ou splines cúbicos restritos).

Splines

```
mod5 <- coxph(Surv(time = tempos, event = cens) ~ pspline(idade),
              data = df.laringe, method = "breslow")

summary(mod5)
```

```
## Call:
## coxph(formula = Surv(time = tempos, event = cens) ~ pspline(idade),
##     data = df.laringe, method = "breslow")
##
##   n= 90, number of events= 50
##
##                          coef    se(coef) se2     Chisq DF   p
## pspline(idade), linear 0.02146 0.01271  0.01271 2.85  1.00 0.091
## pspline(idade), nonlin                          5.24  3.04 0.160
##
##              exp(coef) exp(-coef) lower .95 upper .95
## ps(idade)3    0.39414      2.537  0.060231     2.579
## ps(idade)4    0.16349      6.117  0.008797     3.038
## ps(idade)5    0.09910     10.090  0.003650     2.691
## ps(idade)6    0.09968     10.032  0.003554     2.795
## ps(idade)7    0.11821      8.460  0.004589     3.045
## ps(idade)8    0.12668      7.894  0.005274     3.042
```

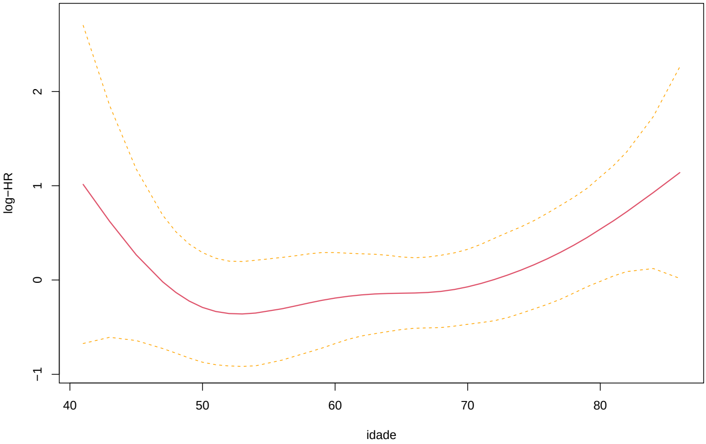
Splines

```
## ps(idade)9    0.12397      8.067  0.005253     2.925
## ps(idade)10   0.14586      6.856  0.006109     3.482
## ps(idade)11   0.18937      5.281  0.007888     4.546
## ps(idade)12   0.27940      3.579  0.011526     6.773
## ps(idade)13   0.44993      2.223  0.016728    12.102
## ps(idade)14   0.73136      1.367  0.015323    34.908
##
## Iterations: 3 outer, 10 Newton-Raphson
##      Theta= 0.6691749
## Degrees of freedom for terms= 4
## Concordance= 0.572  (se = 0.044 )
## Likelihood ratio test= 8.11  on 4.04 df,   p=0.09
```

Splines

```
termplot(model = mod5, terms = "pspline(idade)",
         se = TRUE,
         xlabs = "idade",
         ylabs = "log-HR",
         data = df.laringe)
```

Splines



Splines

```
library(splines)

mod6 <- coxph(Surv(time = tempos, event = cens) ~ ns(idade, df = 3),
              data = df.laringe, method = "breslow")

summary(mod6)
```

```
## Call:
## coxph(formula = Surv(time = tempos, event = cens) ~ ns(idade,
##     df = 3), data = df.laringe, method = "breslow")
##
##   n= 90, number of events= 50
##
##                         coef exp(coef) se(coef)      z Pr(>|z|)
## ns(idade, df = 3)1 -0.1101    0.8958   0.5696 -0.193   0.8468
## ns(idade, df = 3)2 -0.4264    0.6529   1.8140 -0.235   0.8142
## ns(idade, df = 3)3  1.1490    3.1550   0.5516  2.083   0.0373 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                    exp(coef) exp(-coef) lower .95 upper .95
## ns(idade, df = 3)1    0.8958      1.116   0.29334     2.735
```

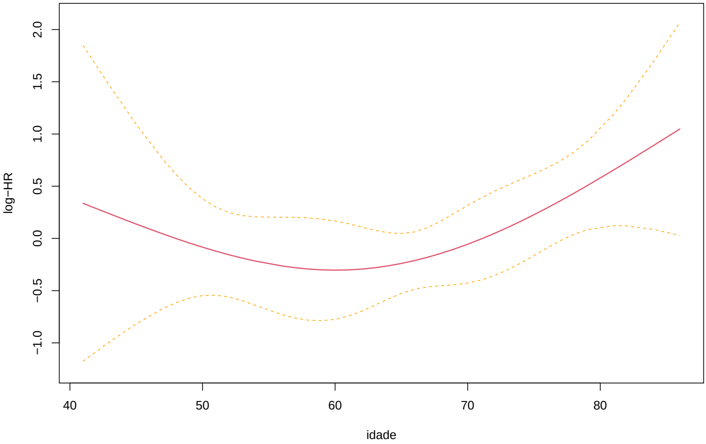
Splines

```
## ns(idade, df = 3)2    0.6529      1.532   0.01865    22.849
## ns(idade, df = 3)3    3.1550      0.317   1.07022     9.301
##
## Concordance= 0.569  (se = 0.043 )
## Likelihood ratio test= 5.62  on 3 df,   p=0.1
## Wald test            = 6.38  on 3 df,   p=0.09
## Score (logrank) test = 6.73  on 3 df,   p=0.08
```

Splines

```
termplot(model = mod6, terms = "ns(idade, df = 3)",
         se = TRUE,
         xlabs = "idade",
         ylabs = "log-HR",
         data = df.laringe)
```

Splines

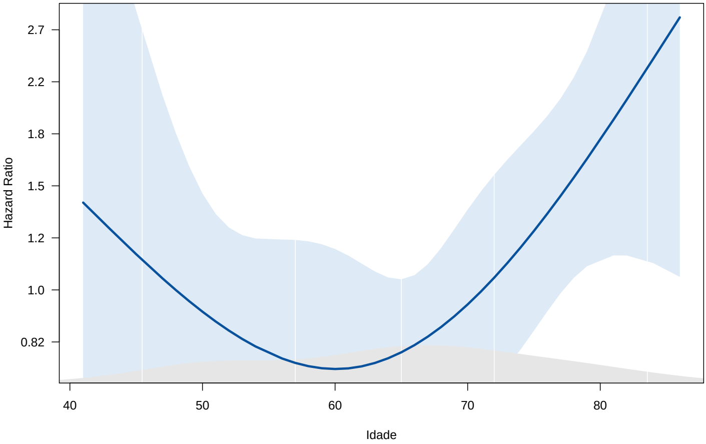


Splines

```
library(Greg)

plotHR(mod6,
       term = "idade",
       plot.bty = "o",
       xlim = c(41, 86),
       xlab = "Idade")
```

Splines



Para casa

1. Leia o capítulo 5 do livro **Análise de sobrevivência aplicada**[2].
2. Leia os capítulo 6 e 7 do livro **Análise de sobrevivência: teoria e aplicações em saúde**[3].

[2]Colosimo, E. A. e Giolo, S. R. **Análise de sobrevivência aplicada**, Blucher, 2006.

[3]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!