Curso de Especialização em *Data Science* e Estatística Aplicada
Módulo IV - Análise de Sobrevivência

Prof. Dr. Eder Angelo Milani

26/04/2025

Instruções

- O desenvolvimento desta atividade deve ser realizada de forma individual ou em dupla.
- Deve-se completar o arquivo Rmd enviado na atividade.
- É necessário devolver o arquivo em Rmd e em pdf.
- Valor da atividade: 10 pontos.
- Use o código fornecido como base.

Descrição da atividade

O objetivo da atividade avaliativa é a análise do conjunto de dados de pacientes diagnosticados com neoplasia maligna do estômago (CID C16), com diagnóstico entre os anos de 2013 a 2016, com acompanhamento até o ano de 2021. Os dados foram obtidos da Fundação Oncocentro de São Paulo (FOSP).

As variáveis disponíveis para análise são:

- TOPOGRUP: grupo da topografia
- TEMPO: tempo em anos do diagnóstico até a falha ou censura
- CENSURA: variável que indica se o tempo é de falha (=1) ou de censura à direita (=0)
- ANODIAG: indica o ano do diagnóstico
- SEXO: 0 - Masculino ou 1 - Feminino
- CIRURGIA: 0 se não realizou cirurgia - 1 se realizou cirurgia
- RADIO: 0 se não realizou radioterapia - 1 se realizou radioterapia
- QUIMIO: 0 se não realizou quimioterapia - 1 se realizou quimioterapia

Questão 1 Faça a leitura do conjunto de dados *cancer_c16.csv* e formate as variáveis ANODIAG, SEXO, CIRURGIA, RADIO e QUIMIO para fator.

Questão 2 Faça as seguintes análises descritivas

(i) calcule a proporção e o valor absoluto para os possíveis valores das variáveis CENSURA, ANODIAG, SEXO, CIRURGIA, RADIO e QUIMIO. Comente os valores encontrados.

(ii) faça o gráfico com a estimativa de Kaplan-Meier para a função de sobrevivência sem considerar covariáveis. Faça comentários sobre o gráfico.

(iii) construa o gráfico com a estimativa de Kaplan-Meier para a função de sobrevivência considerando as covariáveis ANODIAG, SEXO, CIRURGIA, RADIO e QUIMIO, uma de cada vez. Faça comentários sobre os gráficos.

(iv) realize o teste logrank considerando as covariáveis ANODIAG, SEXO, CIRURGIA, RADIO e QUIMIO, uma de cada vez. Descreva a conclusão do teste utilizando o p-valor.

Questão 3 Utilizando as covariáveis ANODIAG, SEXO, CIRURGIA, RADIO e QUIMIO, e os modelos exponencial, Weibull e log-normal, responda:

(i) adotando o modelo exponencial, qual é o conjunto de covariáveis que apresenta melhor AIC, considerando a rotina *stepAIC* com *direction = "both"*?

(ii) adotando o modelo Weibull, qual é o conjunto de covariáveis que apresenta melhor AIC, considerando a rotina *stepAIC* com *direction = "both"*?

(iii) adotando o modelo log-normal, qual é o conjunto de covariáveis que apresenta melhor AIC, considerando a rotina *stepAIC* com *direction = "both"*?

(iv) usando o critério AIC, qual o melhor modelo?

(v) faça o gráfico dos resíduos de Cox-Snell. O que se pode afirmar sobre a qualidade do ajuste do modelo aos dados?

(vi) obtenha a estimativa da sobrevivência nos instantes de 1 e 10 anos, para dois pacientes com:

- paciente 1 - sexo masculino, não fez cirurgia, não fez radio, não fez quimio - x=(1, 0, 0, 0, 0)
- paciente 2 - sexo feminino, fez cirurgia, fez radio, fez quimio - x=(1, 1, 1, 1, 1)

(vii) calcule para os dois pacientes do item (vi) o MTTF.

Questão 4 Utilizando as covariáveis ANODIAG, SEXO, CIRURGIA, RADIO e QUIMIO, e o modelo de Cox, responda:

(i) faça o ajuste do modelo de Cox considerando todas as covariáveis.

(ii) utilizando da rotina *stepAIC* com *direction = "both"*, obtenha o conjunto de covariáveis que apresenta melhor AIC.

(iii) a partir do conjunto de covariáveis obtido do item (ii), faça o gráfico com os resíduos de Cox-Snell. Comente o resultado.

(iv) também utilizando do conjunto de covariáveis obtido do item (ii), verifique a adequação do modelo utilizando o teste de hipóteses do Resíduo de Schoenfeld. Comente o resultado.

Códigos de apoio

Questão 1

```
# limpando o que tem na memoria
rm(list=ls())

# local onde esta o arquivo com os dados
setwd("C:\\caminho\\do\\seu\\computador")

### leitura
dados <- read.csv("cancer_c16.csv")

# formatar as variáveis ANODIAG, SEXO, CIRURGIA, RADIO e QUIMIO para fatores
dados$ANODIAG <- factor(dados$ANODIAG)
```

Questão 2

```
# tabela
table( )
prop.table(table( ))

#install.packages("survival")
require(survival)

# Kaplan-Meier
ekm <- survfit(Surv(TEMPO, CENSURA) ~ 1, data=dados)
plot(ekm, ylab="S(t)", xlab="Tempo(ano)", main="", mark.time = F, conf.int = F)


# Sexo
# Kaplan-Meier
ekm2 <- survfit(Surv(TEMPO, CENSURA) ~ SEXO, data=dados)
plot(ekm2, lty=c(1,1), xlab= "Tempo (ano)", ylab="S(t)", mark.time = F,
     conf.int = F, col=c("black", "red"))
legend(6, 0.8, lty=c(1,1), c("Fem", "Mas"), col=c("red", "black"), bty="n")
# Teste logrank
survdiff(Surv(TEMPO, CENSURA) ~ SEXO, data = dados)
```

Questão 3

```
library(MASS)
# modelo inicial com todas as variaveis usando a distribuicao exponencial
modelo_inicial_exp <- survreg(Surv(TEMPO, CENSURA) ~ COVARIAVEIS,
                              data = dados, dist = "exponential")

summary(modelo_inicial_exp)

# vamos agora aplicar a selecao stepwise baseada no AIC
modelo_final_exp <- stepAIC(modelo_inicial_exp, direction = "both")

# Resumo do modelo final
summary(modelo_final_exp)
```

```r
AIC(modelo_final_exp)


# Adequacao do modelo ajustado
matriz_modelo <- model.matrix(~ COVARIAVEIS, data = dados)
head(matriz_modelo)
x_beta <- matriz_modelo %*% modelo_final$coefficients
cox_snell_modelo_final <- - log(1 - pnorm((log(dados$TEMPO)-x_beta) / modelo_final$scale) )

# grafico
ekm_cos_snell_final <- survfit(Surv(cox_snell_modelo_final, dados$CENSURA) ~ 1)
exp_cox_snell_modelo_final <- exp(- ekm_cos_snell_final$time)
plot(ekm_cos_snell_final$surv, exp_cox_snell_modelo_final,
     main= "Resíduos de Cox-Snell", ylab="S(t) - exponencial", xlab="S(t) - EKM")
abline(a=0, b=1)


# Para o paciente 1
x_beta_p1 <- c(1, 0, 0, 0, 0) %*% modelo_final$coefficients

# Para o paciente 2
x_beta_p2 <- c(1, 1, 1, 1, 1) %*% modelo_final$coefficients
```

Questão 4

```r
#install.packages("MASS")
library(MASS)

# modelo inicial com todas as variaveis
modelo_inicial_cox <- coxph(Surv(TEMPO, CENSURA) ~  COVARIAVEIS,
                            data = dados)

# vamos agora aplicar a selecao stepwise baseada no AIC
modelo_cox <- stepAIC(modelo_inicial_cox, direction = "both")

# Resumo do modelo final
summary(modelo_cox)

# Adequacao do modelo ajustado

# verificacao de proporcionalidade
cox.zph(modelo_cox, transform = "identity")
```