MAE0514 - Introducão a Análise de Sobrevivência - Lista 4

Bruno de Castro Paul Schultze[1]
Rubens Santos Andrade Filho[2]

Julho de 2021

Sumário

[1]Número USP: 10736862
[2]Número USP: 10370336

Questão 1

Breve resumo do artigo:

(ii) Fabiani, M., Ramigni, M., Gobbetto, V., Mateo-Urdiales, A., Pezzotti, P., Piovesan, C. (2021). Effectiveness of the Comirnaty (BNT162b2, BioNTech/Pfizer) vaccine in preventing SARS-CoV-2 infection among healthcare workers, Treviso province, Veneto region, Italy, 27 December 2020 to 24 March 2021, *Euro Surveillance*, 26(17).

O objetivo do artigo foi estimar a eficácia da vacina Comirnaty (BNT162b2, BioNTech/Pfizer) na prevenção de infecção por SARS-CoV-2. Para isso, foi realizado um estudo de coorte retrospectivo no qual foram utilizados dados demográficos, características profissionais e datas de vacinção de 6.423 trabalhadores da área de saúde empregados na unidade local de saúde da província de Treviso em Veneto na Itália.

Primeiro, foram realizadas curvas de Kaplan-Meier para o tempo desde o inicio da campanha de vacinação até a infecção pelo vírus ou o fim do estudo para indivíduos não vacinados, que tomaram a primeira dose e ambas as doses da vacina. As curvas mostram que os pacientes que tomaram ambas as doses, e depois os que tomaram a primeira dose, apresentaram consistentemente uma menor probabilidade acumulada de infecção.

Também foi feita uma análise para o número de dias desde a administração da vacina para medir a duração do acompanhamento. A eficácia da vacina em diferentes intervalos de tempo foi estimada usando um modelo multivariável de risco proporcional de Cox, incluindo sexo, faixa etária, categoria profissional, contexto de trabalho e semana inicial de exposição como covariáveis.

A análise sugeriu que a vacina Comirnaty teve uma alta eficácia na prevenção da infecção por SARS-CoV-2 em HCW durante os intervalos de tempo após a administração em que a proteção pode ser esperada.

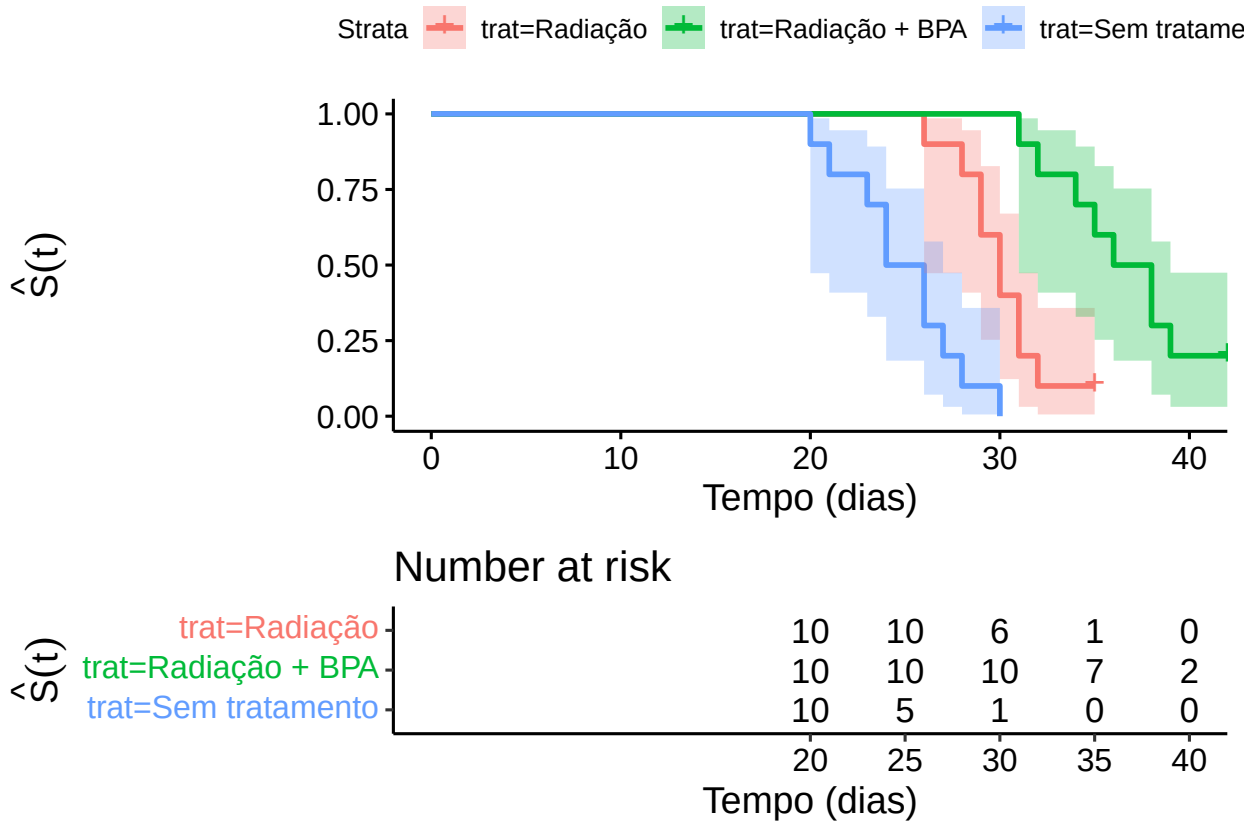
Questão 2

2.a

Primeiro vamos observar as curvas de Kaplan-Meier para cada grupo.

```
q2_km = survfit(Surv(tempo, falha)~trat, data = q2,
                    conf.type = 'log-log')

ggsurvplot(q2_km,q2, conf.int = T, risk.table = T, tables.height = 0.35)+
  labs(x="Tempo (dias)", y=expression(hat(S)(t)))
```



Notemos então que parece existir uma ordenação entre as curvas dos três curvas, onde o grupo sem tratamento tem uma função de sobrevivência menor nos tempos superiores a 20, e o grupo com Radiação + BPA parece ter os maiores valores da função de sobrevivência.

Aplicando então o teste de Log-Rank, que tem como hipótese nula a de que as funções de sobrevivência populacionais são iguais para os três grupos:

```
surv_pvalue(q2_km)
```

```
##   variable         pval   method   pval.txt
## 1     trat 5.643557e-08 Log-rank p < 0.0001
```

Vemos então que o p-valor leva à rejeição da hipótese nula sob qualquer nível de significância usual, de modo que há então evidências que a função de sobrevivência de ao menos um grupo é diferente das demais.

2.b

Criando as variáveis binárias:

```
# 2.b
q2$z1 = ifelse(q2$trat == 'Radiação', 1, 0)
q2$z2 = ifelse(q2$trat == 'Radiação + BPA', 1, 0)
```

Como existem empates, vamos testar os coeficientes e erros-padrão gerados a partir de três algoritmos diferentes: Efron, Breslow e Cox (exato).

Efron

```
q2_cox_efron = coxph(Surv(tempo, falha)~ z1 + z2, data = q2, ties = 'efron')
summary(q2_cox_efron)
```

```
## Call:
## coxph(formula = Surv(tempo, falha) ~ z1 + z2, data = q2, ties = "efron")
##
##   n= 30, number of events= 27
##
##        coef exp(coef) se(coef)      z Pr(>|z|)
## z1 -1.92238   0.14626  0.56670 -3.392 0.000693 ***
## z2 -3.76323   0.02321  0.76612 -4.912 9.01e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##    exp(coef) exp(-coef) lower .95 upper .95
## z1   0.14626      6.837   0.04817    0.4441
## z2   0.02321     43.087   0.00517    0.1042
##
## Concordance= 0.819  (se = 0.018 )
## Likelihood ratio test= 29.93  on 2 df,   p=3e-07
## Wald test            = 24.52  on 2 df,   p=5e-06
## Score (logrank) test = 35.2  on 2 df,   p=2e-08
```

Breslow

```
q2_cox_breslow = coxph(Surv(tempo, falha)~ z1 + z2, data = q2, ties = 'breslow')
summary(q2_cox_breslow)
```

```
## Call:
## coxph(formula = Surv(tempo, falha) ~ z1 + z2, data = q2, ties = "breslow")
##
##   n= 30, number of events= 27
##
##        coef exp(coef) se(coef)      z Pr(>|z|)
## z1 -1.81197   0.16333  0.55971 -3.237  0.00121 **
## z2 -3.55737   0.02851  0.75825 -4.692 2.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##    exp(coef) exp(-coef) lower .95 upper .95
## z1   0.16333      6.123  0.054531    0.4892
## z2   0.02851     35.071  0.006451    0.1260
```

```
##
## Concordance= 0.819  (se = 0.018 )
## Likelihood ratio test= 27.37  on 2 df,   p=1e-06
## Wald test            = 22.45  on 2 df,   p=1e-05
## Score (logrank) test = 31.74  on 2 df,   p=1e-07
```

Exato

```
q2_cox_exact = coxph(Surv(tempo, falha)~ z1 + z2, data = q2, ties = 'exact')
summary(q2_cox_exact)
```

```
## Call:
## coxph(formula = Surv(tempo, falha) ~ z1 + z2, data = q2, ties = "exact")
##
##   n= 30, number of events= 27
##
##        coef exp(coef) se(coef)      z Pr(>|z|)
## z1 -2.28322   0.10196  0.71469 -3.195   0.0014 **
## z2 -4.23235   0.01452  0.90716 -4.665 3.08e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##    exp(coef) exp(-coef) lower .95 upper .95
## z1   0.10196      9.808  0.025123   0.41376
## z2   0.01452     68.879  0.002453   0.08592
##
## Concordance= 0.819  (se = 0.018 )
## Likelihood ratio test= 30.78  on 2 df,   p=2e-07
## Wald test            = 21.77  on 2 df,   p=2e-05
## Score (logrank) test = 33.38  on 2 df,   p=6e-08
```

Conclusão

Notemos então que ao nível de 5% todos os coeficientes parecem significativos através dos testes de Wald individuais, independentemente do algoritmo usado para empates.

2.c

O próprio método utilizado já nos retorna o p-valor do teste de razão de verossimilhanças comparando com o modelo nulo. Portanto, olhando para as saídas disponíveis em 2.b nós notamos que a hipótese nula desse teste é rejeitada ao nível de 5% independentemente da aproximação utilizada.

2.d

O próprio método utilizado também nos retorna o p-valor do teste de Wald comparando com o modelo nulo. Portanto, olhando para as saídas disponíveis em 2.b nós notamos que a hipótese nula desse teste é rejeitada ao nível de 5% independentemente da aproximação utilizada.

2.e

Para isso vamos comparar o modelo completo (os feitos em 2.b) com o modelo substituindo Z1 e Z2 por uma única variável (z3), que é 1 se Z1 ou Z2 forem 1, e 0 caso contrário (o que é o mesmo que falar que o efeito é o mesmo/são a mesma variável).

Para a comparação utilizaremos o teste de razão de verossimilhanças.

Efron

```
# 2.e
q2$z3 = ifelse(q2$z1 ==1 | q2$z2 ==1, 1, 0)
q2_cox_efron2 = coxph(Surv(tempo, falha)~ z3, data = q2, ties = 'efron')
anova(q2_cox_efron, q2_cox_efron2)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(tempo, falha)
##  Model 1: ~ z1 + z2
##  Model 2: ~ z3
##    loglik  Chisq Df P(>|Chi|)
## 1 -57.057
## 2 -61.755 9.3978  1  0.002173 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Breslow

```
q2_cox_breslow2 = coxph(Surv(tempo, falha)~ z3, data = q2, ties = 'breslow')
anova(q2_cox_breslow, q2_cox_breslow2)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(tempo, falha)
##  Model 1: ~ z1 + z2
##  Model 2: ~ z3
##    loglik  Chisq Df P(>|Chi|)
## 1 -59.331
## 2 -63.551 8.4396  1  0.003671 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exato

```
q2_cox_exact2 = coxph(Surv(tempo, falha)~ z3, data = q2, ties = 'exact')
anova(q2_cox_exact, q2_cox_exact2)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(tempo, falha)
##  Model 1: ~ z1 + z2
##  Model 2: ~ z3
##    loglik  Chisq Df P(>|Chi|)
## 1 -47.789
## 2 -52.531 9.4839  1  0.002073 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusão

Em todas as aproximações a hipótese nula, que os efeitos são iguais, é rejeitada ao nível de 5%.

Questão 3

Questão 4

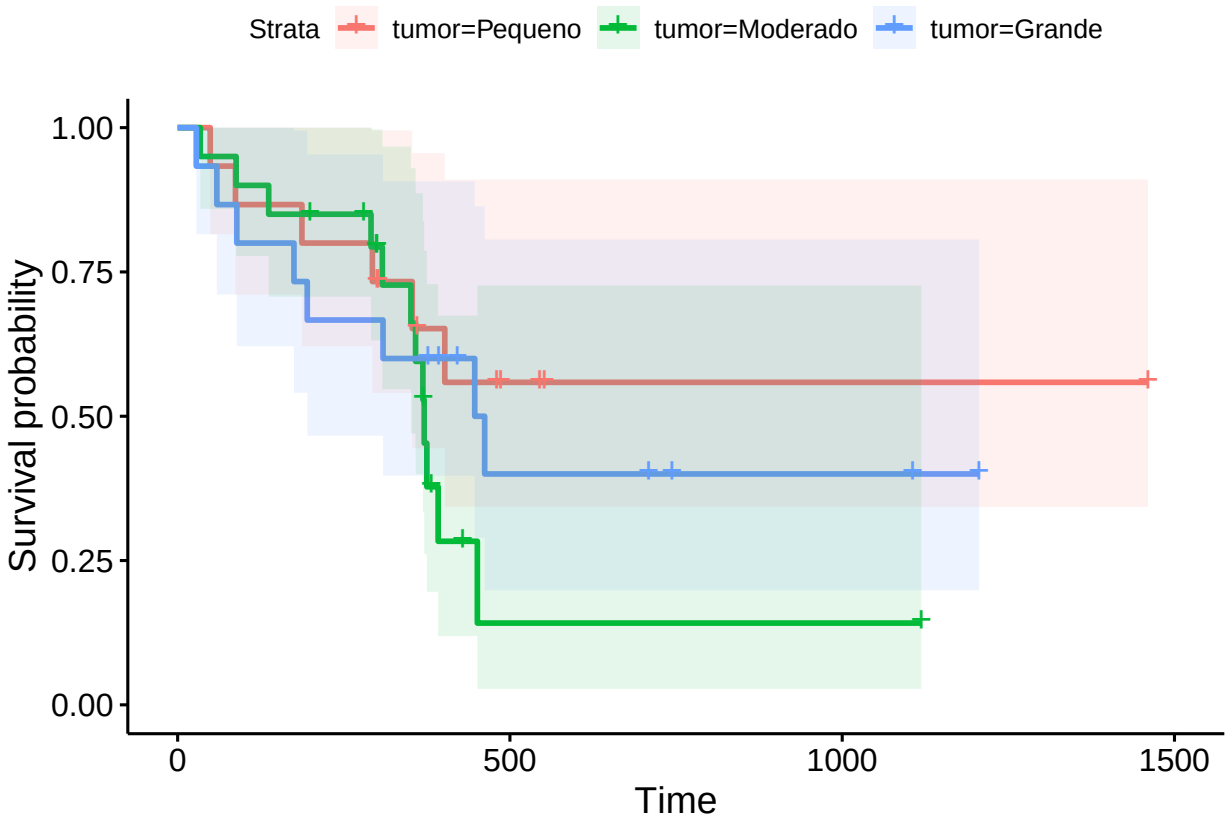
a)

```
# Questão 4 ----


# t<0 representa censura à direita
t = c(28,59,89,175,195,309,-377, -393, -421, 447, 462,-709, -744, -1106, -1206,
      34,88,137,-199, -280, 291,-299, -300, 308,351, 358,369,-370, 371,375,
      -382, 392,-429, 451,-1119, 49,87,187,293,-302, -300, 353,-360, 402,-480,
      -486, -545, -552, -1460, -1460)

dados <- tibble(
  tempo=abs(t),
  delta=as.numeric(t>0),
  tumor=c(rep("Grande",15), rep("Moderado",20), rep("Pequeno",15))) %>%
  mutate(tumor=factor(tumor, levels = c("Pequeno","Moderado","Grande") ))

## 4a  ----
library(survival)
library(survminer)
fit <- survfit(Surv(tempo, delta) ~ tumor, data = dados)
p=ggsurvplot(fit, data = dados, conf.int = T, conf.int.alpha=0.1)
p$plot
```



As curvas de Kaplan-Meier mostram que as estimativas da probabilidade de sobrevivência para o grupo com tumor moderado são menores a partir do dia 350. Entretanto, as curvas se cruzam por volta do dia 300 e todas as estimativas apresentam intervalos de confiança sobrepostos, o que indica que as curvas são próximas.

b)

```
## 4b  ----
# logrank
survdiff(formula(fit),data=dados)
```

```
## Call:
## survdiff(formula = formula(fit), data = dados)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## tumor=Pequeno   15        6     8.58    0.7773     1.177
## tumor=Moderado  20       12     8.87    1.1007     1.748
## tumor=Grande    15        8     8.54    0.0344     0.052
##
##  Chisq= 2  on 2 degrees of freedom, p= 0.4
```

```
# Harrington and Fleming com rho = 0.5
survdiff(formula(fit),data=dados, rho=0.5)
```

```
## Call:
## survdiff(formula = formula(fit), data = dados, rho = 0.5)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## tumor=Pequeno   15     5.25     7.19   0.52342    0.9139
## tumor=Moderado  20     9.96     7.78   0.60816    1.1100
## tumor=Grande    15     6.90     7.14   0.00779    0.0136
##
##  Chisq= 1.3  on 2 degrees of freedom, p= 0.5
```

Realizamos os testes de *log-rank* e o teste da familia Harrington and Fleming com $\rho = 0.5$ e obtemos os valores p 0.4 e 0.5 respectivamente. Em ambos os testes, não rejeitamos a hipótese nula de que as curvas de sobrevivência são iguais entre os pacientes com tumor pequeno, moderado e grande, a um nível de significância de 5%.

c)

```r
## 4c  ----
fit2 <- coxph(Surv(tempo, delta) ~ tumor, data = dados)
summary(fit2)
```

```
## Call:
## coxph(formula = Surv(tempo, delta) ~ tumor, data = dados)
##
##   n= 50, number of events= 26
##
##                  coef exp(coef) se(coef)     z Pr(>|z|)
## tumorModerado 0.6878    1.9893   0.5100 1.349    0.177
## tumorGrande   0.2938    1.3415   0.5408 0.543    0.587
##
##               exp(coef) exp(-coef) lower .95 upper .95
## tumorModerado     1.989     0.5027    0.7321     5.405
## tumorGrande       1.342     0.7454    0.4648     3.872
##
## Concordance= 0.54  (se = 0.056 )
## Likelihood ratio test= 1.98  on 2 df,   p=0.4
## Wald test            = 1.94  on 2 df,   p=0.4
## Score (logrank) test = 1.99  on 2 df,   p=0.4
```

A partir da saída do modelo, interpretamos que

- O risco de óbito no grupo com tumor moderado é $e^{\hat{\beta}_M} \approx 2$ vezes o risco no grupo com tumor pequeno.
- O risco de óbito no grupo com tumor moderado é $(e^{\hat{\beta}_M} - 1)100\% \approx 34\%$ maior que o risco no grupo com tumor pequeno.

Entretanto as estimativas para o erro padrão das estimativas dos coeficientes de tumor moderado e grande são altas, e isso reflete que as estimativas intervalares de 95% de confiança para e^{β_M} e e^{β_G} contém o 1, ou, de forma equivalente, os valores p não são significativos a um nível de 5%. Dessa forma, não rejeitamos as hipóteses de que $\beta_M = 0$ e $\beta_G = 0$.

d)

Testamos a hipótese $H_0 : \boldsymbol{\beta} = \mathbf{0}$ utilizando a estatística de teste da razão de verossimilhanças. A saída do teste se encontra no item c). O valor da estatística obtido foi de $1,98$ com dois graus de liberdade e um valor-p de $0,4$. Com isso, não rejeteitamos a hipótese nula a um nível de 5%.

Questão 5

a)

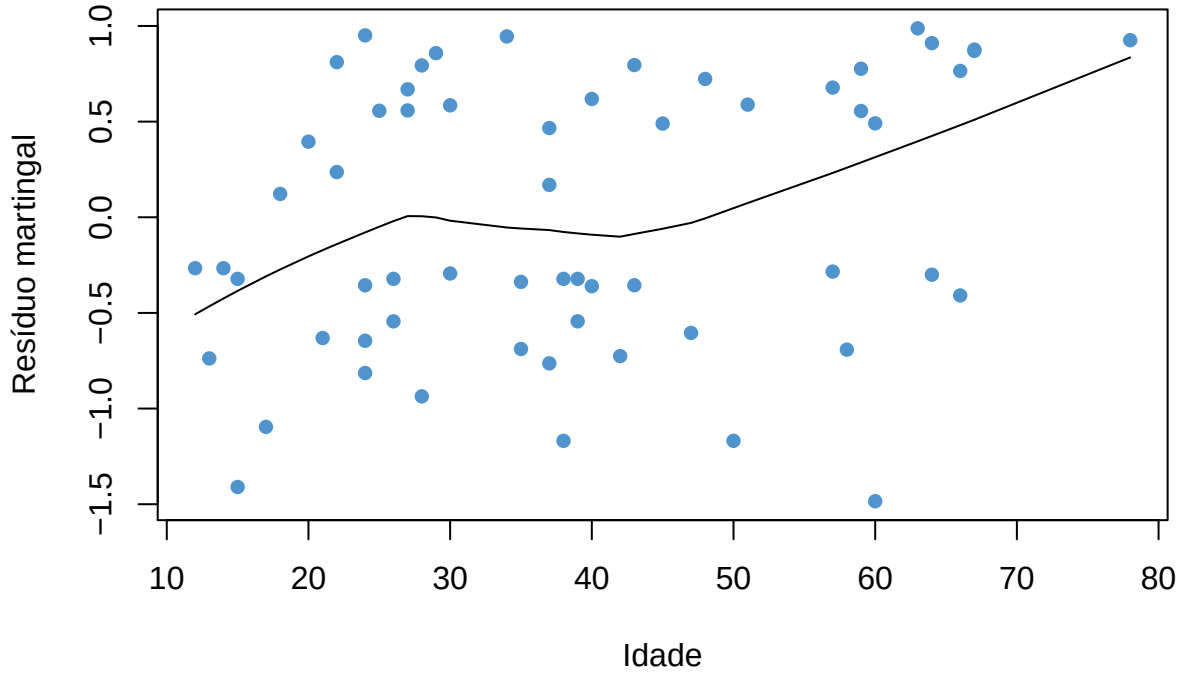
Primeiro, ajustando o modelo de Cox sem usar a idade:

```
q5_cox_a = coxph(Surv(survivaltime, dead)~sex+stage+hist, data = q5)
summary(q5_cox_a)
```

```
## Call:
## coxph(formula = Surv(survivaltime, dead) ~ sex + stage + hist,
##     data = q5)
##
##   n= 60, number of events= 30
##
##           coef exp(coef) se(coef)     z Pr(>|z|)
## sex1   0.19046   1.20981  0.45891 0.415  0.67812
## stage1 0.76467   2.14828  0.40826 1.873  0.06107 .
## hist2  0.06433   1.06644  0.42182 0.153  0.87879
## hist3  1.66692   5.29586  0.55367 3.011  0.00261 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## sex1       1.210     0.8266    0.4921     2.974
## stage1     2.148     0.4655    0.9651     4.782
## hist2      1.066     0.9377    0.4665     2.438
## hist3      5.296     0.1888    1.7892    15.675
##
## Concordance= 0.667  (se = 0.051 )
## Likelihood ratio test= 12.43  on 4 df,   p=0.01
## Wald test            = 14.51  on 4 df,   p=0.006
## Score (logrank) test = 17.01  on 4 df,   p=0.002
```

Em seguida os resíduos martingal:

```
plot(q5$age, resid(q5_cox_a), xlab = 'Idade',
     ylab = 'Resíduo martingal', pch = 16, col = 'steelblue3')
smooth = lowess(q5$age, resid(q5_cox_a), iter = 1)
lines(smooth)
```



Observemos então que parece ser realmente uma tendência linear, indicando que transformações não são necessárias para essa variável no modelo.

b)

```
# 5.b
q5_cox_b = coxph(Surv(survivaltime, dead)~sex+stage+hist+age, data = q5)
summary(q5_cox_b)
```

```
## Call:
## coxph(formula = Surv(survivaltime, dead) ~ sex + stage + hist +
##     age, data = q5)
##
##   n= 60, number of events= 30
##
##            coef exp(coef) se(coef)      z Pr(>|z|)
## sex1    0.24416   1.27654  0.45492  0.537   0.5915
## stage1  0.84294   2.32320  0.41313  2.040   0.0413 *
## hist2  -0.13859   0.87058  0.42942 -0.323   0.7469
## hist3   1.30686   3.69455  0.59721  2.188   0.0287 *
## age     0.03028   1.03075  0.01241  2.441   0.0147 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##        exp(coef) exp(-coef) lower .95 upper .95
## sex1      1.2765     0.7834    0.5234     3.114
## stage1    2.3232     0.4304    1.0338     5.221
## hist2     0.8706     1.1487    0.3752     2.020
## hist3     3.6946     0.2707    1.1461    11.910
## age       1.0307     0.9702    1.0060     1.056
##
## Concordance= 0.74  (se = 0.049 )
## Likelihood ratio test= 18.64  on 5 df,   p=0.002
## Wald test            = 21.75  on 5 df,   p=6e-04
## Score (logrank) test = 25.04  on 5 df,   p=1e-04
```

Notemos que pelos testes de Wald apenas *sex* não se mostra significativa ao nível de 5%.

c)

Aplicando o Teste de Razão de Verossimilhanças para avaliar a exclusão de cada variável individualmente temos, para Idade:

```
# 5.c
q5_cox_c_age = coxph(Surv(survivaltime, dead)~sex+stage+hist, data = q5)
anova(q5_cox_b, q5_cox_c_age)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(survivaltime, dead)
##  Model 1: ~ sex + stage + hist + age
##  Model 2: ~ sex + stage + hist
##    loglik  Chisq Df P(>|Chi|)
## 1 -101.29
## 2 -104.40 6.2124  1   0.01269 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para Sexo:

```
q5_cox_c_sex = coxph(Surv(survivaltime, dead)~age+stage+hist, data = q5)
anova(q5_cox_b, q5_cox_c_sex)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(survivaltime, dead)
##  Model 1: ~ sex + stage + hist + age
##  Model 2: ~ age + stage + hist
##    loglik  Chisq Df P(>|Chi|)
## 1 -101.29
## 2 -101.44 0.2972  1    0.5856
```

Para Estágio:

```
q5_cox_c_stage = coxph(Surv(survivaltime, dead)~age+sex+hist, data = q5)
anova(q5_cox_b, q5_cox_c_stage)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(survivaltime, dead)
##  Model 1: ~ sex + stage + hist + age
##  Model 2: ~ age + sex + hist
##    loglik  Chisq Df P(>|Chi|)
## 1 -101.29
## 2 -103.55 4.5201  1    0.0335 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para Histologia:

```
q5_cox_c_hist = coxph(Surv(survivaltime, dead)~age+stage+sex, data = q5)
anova(q5_cox_b, q5_cox_c_hist)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(survivaltime, dead)
##  Model 1: ~ sex + stage + hist + age
##  Model 2: ~ age + stage + sex
##    loglik Chisq Df P(>|Chi|)
## 1 -101.29
## 2 -104.03 5.478  2   0.06463 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notemos então que ao nível de 5% somente as variáveis Idade e Estágio serão utilizadas no modelo final.

```
q5_cox = coxph(Surv(survivaltime, dead)~age+stage, data = q5)
summary(q5_cox)
```

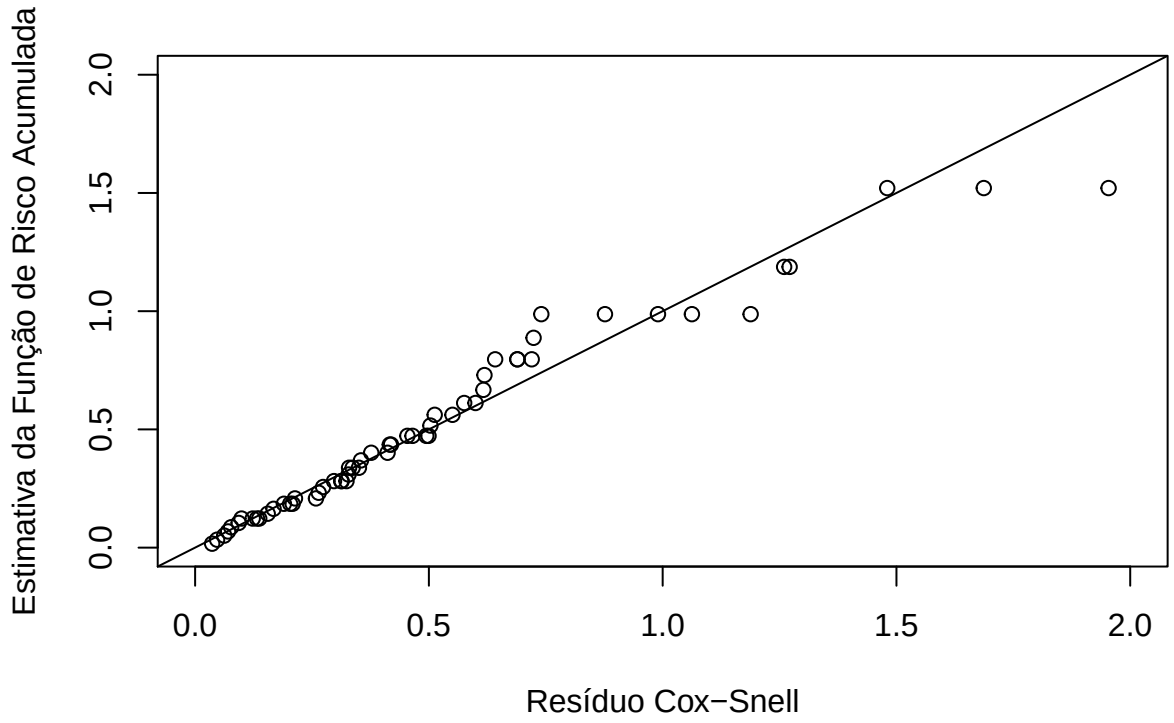
```
## Call:
## coxph(formula = Surv(survivaltime, dead) ~ age + stage, data = q5)
##
##   n= 60, number of events= 30
##
##           coef exp(coef) se(coef)     z Pr(>|z|)
## age    0.03579   1.03644  0.01222 2.929   0.0034 **
## stage1 0.97067   2.63970  0.40653 2.388   0.0170 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## age        1.036     0.9648     1.012     1.062
## stage1     2.640     0.3788     1.190     5.856
##
## Concordance= 0.711  (se = 0.047 )
```

```
## Likelihood ratio test= 13.09  on 2 df,   p=0.001
## Wald test            = 12.1  on 2 df,   p=0.002
## Score (logrank) test = 12.71  on 2 df,   p=0.002
```

Notemos que ter Estágio igual a 1 leva a ter uma taxa de falha 164% maior do que pacientes com Estágio igual a 0. E a cada ano a mais na idade a taxa de falha é 3.6% maior comparada com o ano anterior.

d)

```
# 5.d
q5$cox_snell = - (resid(q5_cox) - q5$dead)
q5_cox_snell = coxph(Surv(cox_snell, dead)~1, data = q5)
q5_base = basehaz(q5_cox_snell, centered = FALSE)
plot(q5_base$time, q5_base$hazard, xlab = 'Resíduo Cox-Snell' ,
     ylab = 'Estimativa da Função de Risco Acumulada', xlim = c(0,2), ylim = c(0,2))
abline(0,1)
```



Notemos então que, com excessão dos último ponto à direita, os pontos se mantém de forma considerável sobre a reta com intercepto 0 e inclinação 1 (x=y), mostrando que o ajuste parece estar razoável.

Código Completo

```
knitr::opts_chunk$set(warning=FALSE,
                      # fig.dim = c(5,5),
                      # out.height = '40%',
                      # fig.align = 'center',
                      message=FALSE
                      )

library(tidyverse)
library(ggplot2)
library(knitr)
library(readr)
library(dplyr)


# Questão 2
library(survival)
library(survminer)

q2 <- data.frame(trat = c(rep("Sem tratamento",10),rep("Radiação",10),
rep("Radiação + BPA",10)),
tempo = c(20,21,23,24,24,26,26,27,28,30,26,28,29,29,30,
30,31,31,32,35,31,32,34,35,36,38,38,39,42,42),
falha = c(rep(1,19),0,rep(1,8),0,0))
q2$trat = as.factor(q2$trat)
q2_km = survfit(Surv(tempo, falha)~trat, data = q2,
                    conf.type = 'log-log')

ggsurvplot(q2_km,q2, conf.int = T, risk.table = T, tables.height = 0.35)+
  labs(x="Tempo (dias)", y=expression(hat(S)(t)))
surv_pvalue(q2_km)
# 2.b
q2$z1 = ifelse(q2$trat == 'Radiação', 1, 0)
q2$z2 = ifelse(q2$trat == 'Radiação + BPA', 1, 0)
q2_cox_efron = coxph(Surv(tempo, falha)~ z1 + z2, data = q2, ties = 'efron')
summary(q2_cox_efron)
q2_cox_breslow = coxph(Surv(tempo, falha)~ z1 + z2, data = q2, ties = 'breslow')
summary(q2_cox_breslow)
q2_cox_exact = coxph(Surv(tempo, falha)~ z1 + z2, data = q2, ties = 'exact')
summary(q2_cox_exact)
# 2.e
q2$z3 = ifelse(q2$z1 ==1 | q2$z2 ==1, 1, 0)
q2_cox_efron2 = coxph(Surv(tempo, falha)~ z3, data = q2, ties = 'efron')
anova(q2_cox_efron, q2_cox_efron2)
q2_cox_breslow2 = coxph(Surv(tempo, falha)~ z3, data = q2, ties = 'breslow')
anova(q2_cox_breslow, q2_cox_breslow2)
q2_cox_exact2 = coxph(Surv(tempo, falha)~ z3, data = q2, ties = 'exact')
anova(q2_cox_exact, q2_cox_exact2)
# Questão 4 ----
```

```r
# t<0 representa censura à direita
t = c(28,59,89,175,195,309,-377, -393, -421, 447, 462,-709, -744, -1106, -1206,
      34,88,137,-199, -280, 291,-299, -300, 308,351, 358,369,-370, 371,375,
      -382, 392,-429, 451,-1119, 49,87,187,293,-302, -300, 353,-360, 402,-480,
      -486, -545, -552, -1460, -1460)

dados <- tibble(
  tempo=abs(t),
  delta=as.numeric(t>0),
  tumor=c(rep("Grande",15), rep("Moderado",20), rep("Pequeno",15))) %>%
  mutate(tumor=factor(tumor, levels = c("Pequeno","Moderado","Grande") ))

## 4a  ----
library(survival)
library(survminer)
fit <- survfit(Surv(tempo, delta) ~ tumor, data = dados)
p=ggsurvplot(fit, data = dados, conf.int = T, conf.int.alpha=0.1)
p$plot
## 4b  ----
# logrank
survdiff(formula(fit),data=dados)
# Harrington and Fleming com rho = 0.5
survdiff(formula(fit),data=dados, rho=0.5)
## 4c  ----
fit2 <- coxph(Surv(tempo, delta) ~ tumor, data = dados)
summary(fit2)
# QUESTAO 5 ----
# QUESTAO 5a ----
library(survival)
library(survminer)
library(readxl)

q5 = read_excel('data/Lista4_Hodgkins.xlsx')
q5$sex = as.factor(q5$sex)
q5$stage = as.factor(q5$stage)
q5$hist = as.factor(q5$hist)
q5
q5_cox_a = coxph(Surv(survivaltime, dead)~sex+stage+hist, data = q5)
summary(q5_cox_a)
plot(q5$age, resid(q5_cox_a), xlab = 'Idade',
     ylab = 'Resíduo martingal', pch = 16, col = 'steelblue3')
smooth = lowess(q5$age, resid(q5_cox_a), iter = 1)
lines(smooth)
# 5.b
q5_cox_b = coxph(Surv(survivaltime, dead)~sex+stage+hist+age, data = q5)
summary(q5_cox_b)
# 5.c
q5_cox_c_age = coxph(Surv(survivaltime, dead)~sex+stage+hist, data = q5)
anova(q5_cox_b, q5_cox_c_age)
q5_cox_c_sex = coxph(Surv(survivaltime, dead)~age+stage+hist, data = q5)
anova(q5_cox_b, q5_cox_c_sex)
q5_cox_c_stage = coxph(Surv(survivaltime, dead)~age+sex+hist, data = q5)
anova(q5_cox_b, q5_cox_c_stage)
```

```r
q5_cox_c_hist = coxph(Surv(survivaltime, dead)~age+stage+sex, data = q5)
anova(q5_cox_b, q5_cox_c_hist)
q5_cox = coxph(Surv(survivaltime, dead)~age+stage, data = q5)
summary(q5_cox)
# 5.d
q5$cox_snell = - (resid(q5_cox) - q5$dead)
q5_cox_snell = coxph(Surv(cox_snell, dead)~1, data = q5)
q5_base = basehaz(q5_cox_snell, centered = FALSE)
plot(q5_base$time, q5_base$hazard, xlab = 'Resíduo Cox-Snell' ,
     ylab = 'Estimativa da Função de Risco Acumulada', xlim = c(0,2), ylim = c(0,2))
abline(0,1)
```