Gabarito da 4ª Lista de Exercícios - MAE514

Professora: Gisela Tunes **Monitor:** Rodrigo Passos Martins

Exercício 1

Para essa questão, leremos dois artigos:

- Gonzalez, A. B. et al. (2010). Body-Mass Index and Mortality among 1.46 Million White Adults, *New England Journal of Medicine* **363**(23): 2211-2219.
- He, J., McGee, D. , Niu, X. e Choi, W. (2009). Examining the Dynamic Association of BMI and Mortality in the Framingham Heart Study, *International Journal of Environmental Research and Public Health* **6**: 3115-3126.

PRIMEIRO ARTIGO: ***Body-Mass Index and Mortality among 1.46 Million White Adults***

O primeiro artigo em questão analisa a associação do alto índice de massa corporal (IMC, o peso em quilogramas dividido pelo quadrado da altura em metros) com o aumento da mortalidade por doenças cardiovasculares e por certos tipos de câncer. Essa associação, contudo, ainda é muito incerta. Por conta dessas incertezas, este artigo tem o propósito de avaliar a faixa ideal de IMC fornecer estimativas estáveis dos riscos associados ao excesso de peso, obesidade e obesidade mórbida (IMC $>$ 40,0), com um mínimo de perturbação causado devido ao tabagismo ou à doença predominante.

As análises se limitaram à pacientes brancos não-hispânicos, pois a relação entre IMC e mortalidade pode diferir entre os grupos raciais e étnicos. Além disso, durante a coleta de dados foram excluídos os pacientes com idade $>$ 85 anos, os pacientes que possuíam menos de 1 ano de acompanhamento ou com um IMC inferior a 15.0 ou superior a 50, e os pacientes que tinham informações faltantes sobre a altura ou peso. As variáveis foram categorizadas conforme um padrão, desta forma temos as seguintes variáveis consideradas no estudo: situação de tabagismo (nunca fumou, ex-fumante ou fumante); número de anos desde que a pessoa parou de fumar (menos de 10, 10 a 19, ou 20 ou mais); consumo de álcool (em gramas por dia); nível de atividade física (baixa, média ou alta); nível educacional (menos que o ensino médio, ensino médio, curso técnico, graduação ou pós-graduação) e estado civil (casado, divorciado, viúvo ou solteiro). Os participantes foram acompanhados desde o início até a data da morte, fim do estudo ou perda do acompanhamento, o que ocorrer primeiro. A causa da morte foi averiguada em certidões de óbito ou registros médicos e foi codificada de acordo com a Classificação Internacional de Doenças.

Uma amostra de 1.46 milhões de adultos brancos não-hispânicos foi coletada e foi utilizado a regressão de Cox para estimar razões de risco e intervalos com 95% de confiança para avaliar a relação entre IMC e mortalidade por todas as causas (idade, estudo, atividade física, consumo de álcool, educação e estado civil). Além disso, as análises envolveram o uso de modelos de riscos proporcionais e foram realizadas com o uso do software estatístico SAS.

Definiu-se um IMC de 22.5 a 24.9 como a categoria de referência, baseando-se cm uma análise preliminar indicando que essa era geralmente a faixa de IMC associada à menor mortalidade. A relação entre o IMC e a mortalidade por todas as causas estudadas não foi linear quando avaliada em toda a faixa de IMC (15,0 a 50,0), desta forma foi testada a heterogeneidade entre coortes com o uso da estatística Q, onde o IMC foi analisado como variável contínua em duas categorias de IMC: (15.0 a 24.9) e (25.0 a 49.9).

Um total de 160.087 mortes foram relatadas durante um acompanhamento mediano de 10 anos, sendo que 35.369 dessas mortes foram de indivíduos saudáveis no início do estudo (não relataram histórico de câncer ou doença cardíaca) e nunca haviam fumado. Os fumantes representaram 25% dos participantes do estudo classificados na menor categoria de IMC (15,0 a 18,4), mas apenas 8% dos fumantes encontram-se na categoria mais alta de IMC (40,0 a 49,9). Câncer pré-existente e enfisema foram ligeiramente mais comuns nas categorias com IMC baixo, enquanto a prevalência de cardiopatia pré-existente aumentou com o aumento do IMC. A inatividade física e falta de um diploma universitário foram ambos associados com um maior IMC.

A estratificação ou exclusão de algumas variáveis, em vez de ajuste, é considerada necessária, uma vez que o tabagismo está fortemente relacionado à obesidade e mortalidade, dificultando assim evitar perturbações residuais por meio de ajustes típicos do tabagismo e do número de cigarros fumado por dia. Logo, foi concluído que há vantagens em concentrar o estudo em participantes saudáveis que nunca fumaram, pois o acompanhamento a longo prazo reforçou ao invés de enfraquecer a associação entre obesidade e mortalidade.

Pode-se citar também que a relação entre o IMC baixo e a mortalidade por todas as causas estudadas é mais forte entre os ex-fumantes que abandonaram o hábito de fumar a menos de 20 anos atrás do que entre os fumantes atuais.

SEGUNDO ARTIGO: ***Examining the Dynamic Association of BMI and Mortality in the Framing ham Heart Study***

O Framingham Heart Study (FHS) é um estudo epidemiológico que avaliou 5.209 homens e mulheres entre 30 e 62 anos de Framingham, Massachusetts. Exames extensivos foram realizados a cada dois anos desde 1948 e o exame de número 29 começou em abril de 2006. Neste estudo, as informações dos primeiros 20 exames (cerca de 40 anos de seguimento) foram utilizadas para as análise. O tempo exato de sobrevivência de cada indivíduo está disponível. O objetivo do estudo é examinar e identificar algumas evidências de que a associação entre o IMC e a mortalidade pode não ser estática, de forma que estudos com diferentes delineamentos captam imagens diferentes da associação entre IMC e mortalidade. A análise é feita separadamente para homens e mulheres, uma vez que é provável que exista uma diferença de gênero para a relação entre o IMC e a mortalidade.

Existem diversos métodos, como tabulações cruzadas, modelos de regressão logística e modelos de sobrevivência de Cox que fornecem um ou vários valores numéricos (razão de chances, risco relativo ou razão de risco) resumindo a relação entre o IMC (medido em um determinado ponto de tempo) e a mortalidade (dentro de um determinado período de acompanhamento). Primeiro, foi utilizado modelos de regressão logística para demonstrar que projetos diferentes na mesma população podem levar a resultados diferentes.

Em seguida, foi aplicado modelos dinâmicos de sobrevivência para capturar a relação variando no tempo entre o IMC e a mortalidade na FHS. A análise usando modelos de regressão logística mostra que o resultado da associação entre o IMC e a mortalidade pode diferir quando o design de um estudo muda, como por exemplo pelo uso do IMC medido em um tempo diferente ou pela mudança do tempo de seguimento.

Neste artigo foi demonstrado a partir de um único conjunto de dados, que a relação entre IMC e mortalidade é, para esses dados, uma relação dinâmica complexa onde os resultados obtidos dependem fortemente dos projetos do estudo e que a relação não parece satisfazer a suposição de riscos proporcionais. Também é observado a partir deste estudo que a característica dinâmica é mais forte para homens do que para mulheres. No entanto, a mesma análise pode não ser repetida em outros estudos se eles tiverem muitas medidas repetidas de IMC dentro de um longo período de acompanhamento. Nossa conclusão é limitada a este estudo e população em particular.

A análise sugere que s estratégia de eliminar mortes precoces leva à associações diretas mais fortes entre IMC e mortalidade, especialmente para homens. Contudo, existem limitações já que apenas a associação linear entre IMC e mortalidade está sendo considerada.

BREVE COMPARAÇÃO Os dois estudos tratam sobre a associação entre o IMC e a mortalidade, contudo as conclusões são opostas.

O primeiro artigo apresenta um foco nas diversas variáveis que podem pertubar as análises finais, já o segundo estudo leva em consideração o tempo com a maior influência nos dados, eliminando diversos dados que possuiam mortes precoces. O segundo artigo apresenta uma relação linear para a associação entre IMC e mortalidade e separa as análises de acordo com a variável sexo, pois é comprovado que há efeito da variável sexo.

Contudo, para o primeiro artigo, notamos que os resultados e conclusões das análises são parecidas para homens e mulheres, sendo uma variável de pouco destaque no estudo, além disso os dados não apresentam relação linear.

Portanto, as conclusões para ambos os artigos são diferentes, pois as análises levam em consideração variáveis distintas e diferentes modelos.

Exercício 2

Um estudo foi conduzido para determinar a eficiência de uma terapia conhecida como BNCT (*Boron Neutron Capture Therapy*), usando BPA (*boronophenylalanine*) como agente de captura, no tratamento de glioma F98, que é um tipo de tumor no sistema nervoso central, nas células gliais. Células com glioma F98 foram implantadas no cérebro de ratos, que foram divididos em três grupos. O primeiro grupo não recebeu tratamento, o segundo grupo foi tratado apenas com radioterapia e, por fim, o terceiro grupo recebeu radiação e também uma dose apropriada de BPA. Os dados disponíveis são os tempos de vida (em dias) dos ratos em cada um dos três grupos e estão apresentados na tabela abaixo (tempos censurados à direita estão denotados por um sinal +):

Sem tratamento	**Radiação**	**Radiação + BPA**
20	26	31
21	28	32
23	29	34
24	29	35
24	30	36
26	30	38
26	31	38
27	31	39
28	32	42^+
30	35^+	42^+

Vamos importar esses dados para o R:

```
# IMPORTANDO OS DADOS PARA O R
library(magrittr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
dados_ex2 <- c(20, 26, 31, 21, 28, 32, 23, 29, 34, 24, 29, 35, 24, 30, 36,
               26, 30, 38, 26, 31, 38, 27, 31, 39, 28, 32, 42, 30, 35, 42)

dados_ex2 %<>% as.data.frame()

names(dados_ex2) <- "TEMPOS"

dados_ex2$TRATAMENTO <- rep(c("Sem tratamento", "Radiação", "Radiação + BPA"), 10)

dados_ex2$FALHAS <- c(rep(1, 26), 0, 1, 0, 0)

dados_ex2$CENSURAS <- c(rep(0, 26), 1, 0, 1, 1)
```

a)

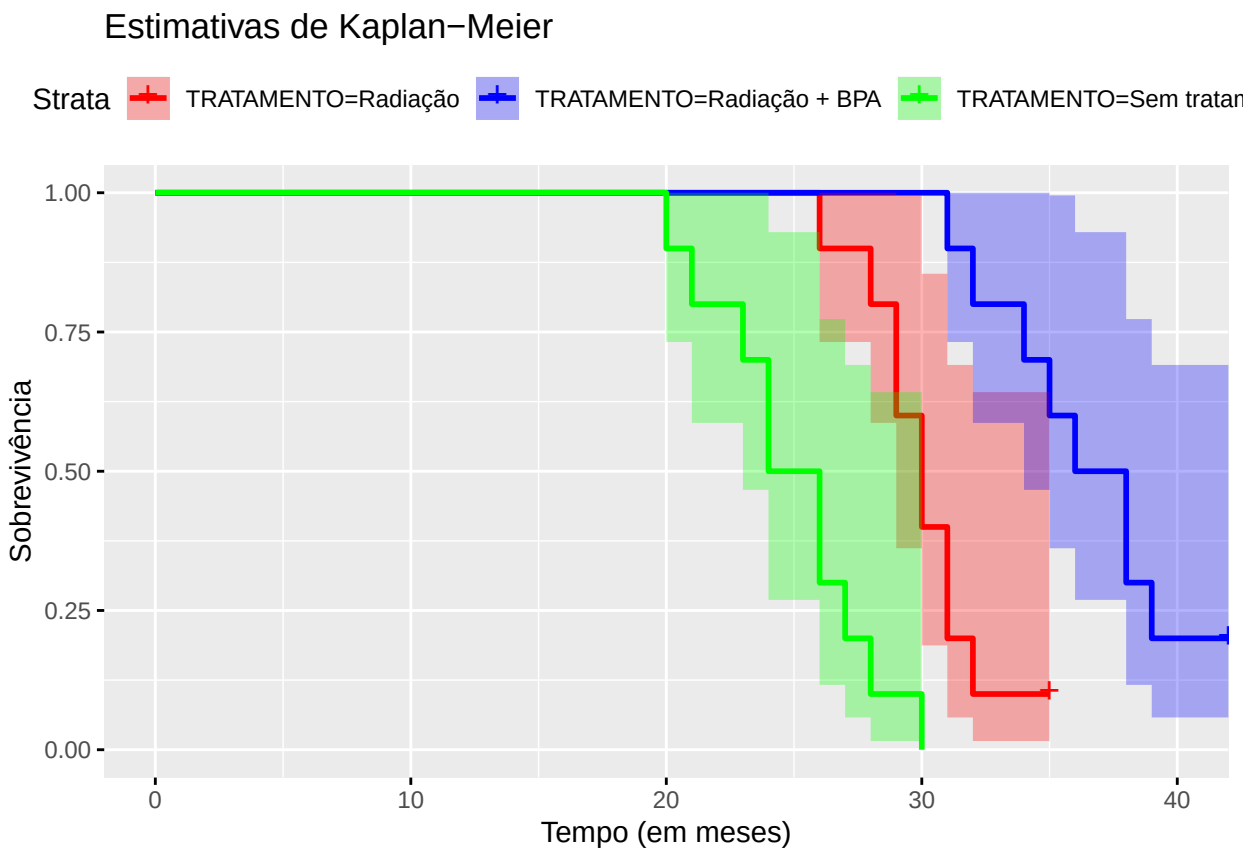
Vamos obter e comparar as curvas de sobrevivência dos três grupos, utilizando algum teste apropriado:

```
# FAZENDO AS CURVAS DE SOBREVIVÊNCIA NO R
library(survival)
library(survminer)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: ggpubr
```

```
S_KM <- survfit(Surv(TEMPOS, FALHAS) ~ TRATAMENTO, data = dados_ex2)

ggsurvplot(S_KM, data = dados_ex2, palette = c("red", "blue", "green"),
           conf.int = T, ggtheme = theme_gray()) +
  labs(x = "Tempo (em meses)", y = "Sobrevivência", title = "Estimativas de Kaplan-Meier")
```




```
survdiff(Surv(TEMPOS, FALHAS) ~ TRATAMENTO, data = dados_ex2)
```

```
## Call:
## survdiff(formula = Surv(TEMPOS, FALHAS) ~ TRATAMENTO, data = dados_ex2)
##
##                             N Observed Expected (O-E)^2/E (O-E)^2/V
## TRATAMENTO=Radiação        10        9     7.60     0.258     0.431
## TRATAMENTO=Radiação + BPA  10        8    16.74     4.566    16.587
## TRATAMENTO=Sem tratamento  10       10     2.66    20.300    27.365
##
##  Chisq= 33.4  on 2 degrees of freedom, p= 6e-08
```

Tendo em vista o gráfico de Sobrevivência de Kaplan Meier, vemos que os ratos que foram submetidos ao tratamento de Radiação + BPA apresentam maior taxa de sobrevivência comparado ao grupo Controle e o tratamento apenas com Radiação. Não há uma proximidade que justifique a igualdade entre as curvas, mas é possível supor uma proporcionalidade entre as curvas. O teste log-rank, com valor-p $< 0{,}001$, corrobora com essa impressão descritiva de diferença dos tratamentos.

b)

Vamos criar duas variáveis binárias Z_1 (igual a 1 se o animal recebeu radiação apenas e igual a 0, caso contrário) e Z_2 (igual a 1 se o animal recebeu radiação e BPA e igual a 0, caso contrário):

```
# CRIANDO AS VARIÁVEL Z1 E Z2 NO R

dados_ex2 %<>% mutate(Z1 = if_else(TRATAMENTO == "Radiação", 1, 0),
                      Z2 = if_else(TRATAMENTO == "Radiação + BPA", 1, 0))
```

Agora, considerando o modelo semiparamétrico de riscos proporcionais, obteremos as estimativas dos coeficientes associados com cada variável criada e dos respectivos erros padrão utilizando três métodos diferentes para empates:

Efron:

```
# MODELO DE COX COM EMPATE "EFRON" NO R

modelo_cox_efron <- coxph(Surv(TEMPOS, FALHAS) ~ Z1 + Z2,
                          ties = c("efron", "breslow", "exact")[1], data = dados_ex2)
summary(modelo_cox_efron)
```

```
## Call:
## coxph(formula = Surv(TEMPOS, FALHAS) ~ Z1 + Z2, data = dados_ex2,
##     ties = c("efron", "breslow", "exact")[1])
##
##   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:

```
# MODELO DE COX COM EMPATE "BRESLOW" NO R

modelo_cox_breslow <- coxph(Surv(TEMPOS, FALHAS) ~ Z1 + Z2,
                            ties = c("efron", "breslow", "exact")[2], data = dados_ex2)
summary(modelo_cox_breslow)
```

```
## Call:
## coxph(formula = Surv(TEMPOS, FALHAS) ~ Z1 + Z2, data = dados_ex2,
##     ties = c("efron", "breslow", "exact")[2])
##
##   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:

```
# MODELO DE COX COM EMPATE "EXATO" NO R

modelo_cox_exato <- coxph(Surv(TEMPOS, FALHAS) ~ Z1 + Z2,
                            ties = c("efron", "breslow", "exact")[3], data = dados_ex2)
summary(modelo_cox_exato)
```

```
## Call:
## coxph(formula = Surv(TEMPOS, FALHAS) ~ Z1 + Z2, data = dados_ex2,
##     ties = c("efron", "breslow", "exact")[3])
##
##   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
```

Analisando as informações de cada um dos modelos ajustados, não há diferença interpretativa ou de significância para cada um dos parâmetros. Assim, podemos concluir que:

- O tratamento com **Radiação** se equivale a 14,6% (Efron), 16,3% (Breslow) ou 10,2% (Exato);
- O tratamento com **Radiação + BPA** se equivale a 2,3% (Efron), 2,8% (Breslow) ou 1,4% (Exato).

c)

Ainda considerando as variáveis criadas no item **b)**, vamos testar a hipótese global de que não há efeito de nenhum tratamento (ou seja, $H_0 : \beta_1 = \beta_2 = 0$), utilizando o teste da razão de verossimilhanças, para as três aproximações para empates usadas. No item **b)**, com os comandos **summary()**, obtemos cada um dos valores do teste da razão de verossimilhanças para cada modelo para essa hipótese:

- **Efron:** 3×10^{-7}
- **Breslow:** 1×10^{-6}
- **Exato:** 2×10^{-7}

Assim, temos que todas as aproximações levam a mesma interpretação sobre o teste de hipótese: há efeito de pelo menos um tratamento.

d)

Vamos repetir a ideia do item \texbf{c), mas usando o teste de Wald. No item **b)**, com os comandos **summary()**, obtemos cada um dos valores do teste de Wald para cada modelo para a hipótese de inexistência de efeitos dos tratamentos:

- **Efron:** 5×10^{-6}
- **Breslow:** 1×10^{-5}
- **Exato:** 2×10^{-5}

Assim, temos que todas as aproximações levam a mesma interpretação sobre o teste de hipótese: há efeito de pelo menos um tratamento.

e)

Vamos testar a hipótese de que o efeito da radiação e da radiação com BPA são iguais (ou seja, $H_0 : \beta_1 = \beta_2$), para as aproximações para empates consideradas, usando o teste da razão de verossimilhanças.

Vamos calcular isso testando se o modelo completo

$$\alpha(t) = \alpha_0(t).exp(\beta_1 Z_1 + \beta_2 Z_2)$$

pode ser substituído por:

$$\alpha(t) = \alpha_0(t).exp\left(\beta(Z_1 + Z_2)\right).$$

Assim, vamos utilizar o comando **anova()** do R para testar essa hipótese com esses dois modelos:

Efron:

```
# MODELO DE COX COM EMPATE "EFRON" NO R

modelo_cox_efron_igual <- coxph(Surv(TEMPOS, FALHAS) ~ I(Z1 + Z2),
                          ties = c("efron", "breslow", "exact")[1], data = dados_ex2)
anova(modelo_cox_efron_igual, modelo_cox_efron)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(TEMPOS, FALHAS)
##  Model 1: ~ I(Z1 + Z2)
##  Model 2: ~ Z1 + Z2
##    loglik  Chisq Df P(>|Chi|)
## 1 -61.755
## 2 -57.057 9.3978  1  0.002173 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Breslow:

```
# MODELO DE COX COM EMPATE "BRESLOW" NO R

modelo_cox_breslow_igual <- coxph(Surv(TEMPOS, FALHAS) ~ I(Z1 + Z2),
                          ties = c("efron", "breslow", "exact")[2], data = dados_ex2)
anova(modelo_cox_breslow_igual, modelo_cox_breslow)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(TEMPOS, FALHAS)
##  Model 1: ~ I(Z1 + Z2)
##  Model 2: ~ Z1 + Z2
##    loglik  Chisq Df P(>|Chi|)
## 1 -63.551
## 2 -59.331 8.4396  1  0.003671 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exato:

```
# MODELO DE COX COM EMPATE "EXATO" NO R

modelo_cox_exato_igual <- coxph(Surv(TEMPOS, FALHAS) ~ I(Z1 + Z2),
                          ties = c("efron", "breslow", "exact")[3], data = dados_ex2)
anova(modelo_cox_exato_igual, modelo_cox_exato)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(TEMPOS, FALHAS)
##  Model 1: ~ I(Z1 + Z2)
##  Model 2: ~ Z1 + Z2
##    loglik  Chisq Df P(>|Chi|)
```

```
## 1 -52.531
## 2 -47.789 9.4839  1  0.002073 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Agrupando os dados, temos que, o valor-p do teste é de:

- **Efron:** $0,002173 \cong 2,2 \times 10^{-3}$
- **Breslow:** $0,003671 \cong 3,7 \times 10^{-3}$
- **Exato:** $0,002073 \cong 2,1 \times 10^{-3}$

Para cada uma das aproximações, temos a mesma interpretação: rejeitamos a hipótese nula, a umn nível de significância de 5%, ou seja, os efeitos do tratamento com **Radiação** e com **Radiação + BPA** não são iguais. A **Radiação + BPA** é um tratamento melhor que só a **Radiação**.

Exercício 3

Vamos considerar um estudo com pacientes com câncer de ovário, em estágio mais avançado. Todos os pacientes foram submetidos a um tratamento padrão e observou-se, para cada paciente, o tempo até sua morte, em dias. No momento do início do tratamento, os pacientes foram separados em dois grupos: pacientes com tumor grande e pacientes com tumor moderado. Vamos avaliar o efeito do tamanho do tumor na sobrevida dos pacientes.

a)

Vamos indicar o grupo com a separação entre o tumor moderado e o tumor grande:

$$x_i = \begin{cases} 1, & \text{se o indivíduo } i \text{ tiver tumor grande} \\ 0, & \text{se o indivíduo } i \text{ tiver tumor moderado} \end{cases}$$

Note que, nesse caso, x_i é a função indicadora do indivíduo i ter tumor grande, ou seja, $\mathbb{1}$(Tumor é grande).

Sendo assim, a forma do modelo semiparamétrico é tal que:

$$\alpha_i(t_i|x_i) = \alpha(t_i).exp(x_i\beta) \blacksquare$$

b)

Assumindo que não há empates, podemos ordenar os tempos de forma que $t_{(1)} < t_{(2)} < \cdots < t_{(d)}$ para obtermos a verossimilhança parcial de Cox. Assim, para o indivíduo i, com tempo de falha em $t_{(i)}$, temos:

$$L(\beta) \propto \prod_{i=1}^{d} \frac{exp(x_i\beta)}{\sum_{j\in R_i} exp(x_j\beta)},$$

sendo R_i o conjunto de todos os indivíduos em risco em $t_{(i)}$ e d o número total de falhas. $\blacksquare$

c)

Com a verossimilhança parcial obtida no item **b)**, calculamos:

$$log(L(\beta)) = log\left(\prod_{i=1}^{d} \frac{exp(x_i\beta)}{\sum_{j\in R_i} exp(x_j\beta)}\right) = \sum_{i=1}^{d} log\left(\frac{exp(x_i\beta)}{\sum_{j\in R_i} exp(x_j\beta)}\right) = \sum_{i=1}^{d} x_i\beta - \sum_{i=1}^{d} log\left(\sum_{j\in R_i} exp(x_j\beta)\right)$$

Agora, vamos calcular o escore $U(\beta)$, que é dado por:

$$U(\beta) = \frac{\delta log(L(\beta))}{\delta\beta}$$

Assim,

$$U(\beta) = \sum_{i=1}^{d} x_i - \sum_{i=1}^{d} \frac{\frac{\delta \sum_{j \in R_i} exp(x_j\beta)}{\delta\beta}}{\sum_{j \in R_i} exp(x_j\beta)} = \sum_{i=1}^{d} x_i - \sum_{i=1}^{d} \frac{\sum_{j \in R_i} x_j.exp(x_j\beta)}{\sum_{j \in R_i} exp(x_j\beta)} \blacksquare$$

d)

A matriz de informação observada é dada por:

$$I(\beta) = -\frac{\delta^2 log(L(\beta))}{\delta\beta\delta\beta^T}$$

No caso, β é um escalar porque o problema é univariado. Assim,

$$I(\beta) = -\frac{\delta^2(U(\beta))}{\delta\beta^2} = -\frac{\delta(U(\beta))}{\delta\beta} = -\sum_{i=1}^{d} \left[\frac{\sum_{j \in R_i} x_j^2.exp(x_j\beta)}{\sum_{j \in R_i} exp(x_j\beta)} - \frac{\left(\sum_{j \in R_i} x_j.exp(x_j\beta)\right)^2}{\sum_{j \in R_i} exp(x_j\beta)} \right]$$

Podemos notar também que $x_j^2 = x_j$. Com isso, podemos simplificar a expressão anterior:

$$I(\beta) = -\sum_{i=1}^{d} \left[\frac{\sum_{j \in R_i} x_j.exp(x_j\beta)}{\sum_{j \in R_i} exp(x_j\beta)} - \frac{\left(\sum_{j \in R_i} x_j.exp(x_j\beta)\right)^2}{\sum_{j \in R_i} exp(x_j\beta)} \right] = -\sum_{i=1}^{d} \frac{\sum_{j \in R_i} x_j.exp(x_j\beta)}{\sum_{j \in R_i} exp(x_j\beta)} . \left[1 - \frac{\sum_{j \in R_i} x_j.exp(x_j\beta)}{\sum_{j \in R_i} exp(x_j\beta)} \right] \blacksquare$$

e)

Vamos supor que deseja-se testar a hipótese $H_0 : \beta = 0$, em que β é o parâmetro no modelo de Cox associado à covariável binária criada anteriormente. Para realizar este teste, podemos utilizar a estatística do teste do escore. Vamos escrever a expressão para a estatística do escore para os seguintes passos:

1º Passo: Com base no item **c)**, vamos obter $U(0)$, ou seja, a expressão para o escore avaliado no ponto $\beta = 0$. Simplificaremos a expressãoo obtida, escrevendo-a em função do número de indivíduos em risco em cada grupo:

$$U(0) = \sum_{i=1}^{d} x_i - \sum_{i=1}^{d} \frac{\sum_{j \in R_i} x_j.exp(x_j.0)}{\sum_{j \in R_i} exp(x_j.0)} = \sum_{i=1}^{d} x_i - \sum_{i=1}^{d} \frac{\sum_{j \in R_i} x_j}{\sum_{j \in R_i} 1}$$

2º Passo: Vamos obter $I(0)$, ou seja, a expressão para a informação avaliada no ponto $\beta = 0$. Simplificaremos a expressão obtida, escrevendo-a em função do número de indivíduos em risco em cada grupo:

$$I(0) = -\sum_{i=1}^{d} \frac{\sum_{j \in R_i} x_j.exp(x_j.0)}{\sum_{j \in R_i} exp(x_j.0)} . \left[1 - \frac{\sum_{j \in R_i} x_j.exp(x_j.0)}{\sum_{j \in R_i} exp(x_j.0)} \right] = -\sum_{i=1}^{d} \frac{\sum_{j \in R_i} x_j}{\sum_{j \in R_i} 1} . \left[1 - \frac{\sum_{j \in R_i} x_j}{\sum_{j \in R_i} 1} \right]$$

3º Passo: Vamos obter a estatística do teste do escore, dada por

$$S(0) = \frac{U^2(0)}{I(0)},$$

que, sob $H_0 : \beta = 0$, tem distribuição de qui-quadrado com 1 grau de liberdade:

$$S(0) = \frac{U^2(0)}{I(0)} = \frac{\left(\sum_{i=1}^{d} x_i - \sum_{i=1}^{d} \frac{\sum_{j \in R_i} x_j}{\sum_{j \in R_i} 1} \right)^2}{\left(- \sum_{i=1}^{d} \frac{\sum_{j \in R_i} x_j}{\sum_{j \in R_i} 1} . \left[1 - \frac{\sum_{j \in R_i} x_j}{\sum_{j \in R_i} 1} \right] \right)} \sim \chi^2_{(1)} \blacksquare$$

Exercício 4

Vamos considerar a situação descrita no **Exercício 3**. Esse problema corresponde a um estudo realizado com 35 mulheres, sendo observados os seguintes tempos (tempos censurados à direita estão denotados por um sinal +):

- **Tumor grande:** 28, 89, 175, 195, 309, 377^+, 393^+, 421^+, 447^+, 462, 709^+, 744^+, 770^+, 1106^+, 1206^+;
- **Tumor moderado:** 34, 88, 137, 199, 280, 291, 299^+, 300^+, 308, 351, 358, 369, 370, 371, 375, 382, 392, 429^+, 451, $1119+^+$.

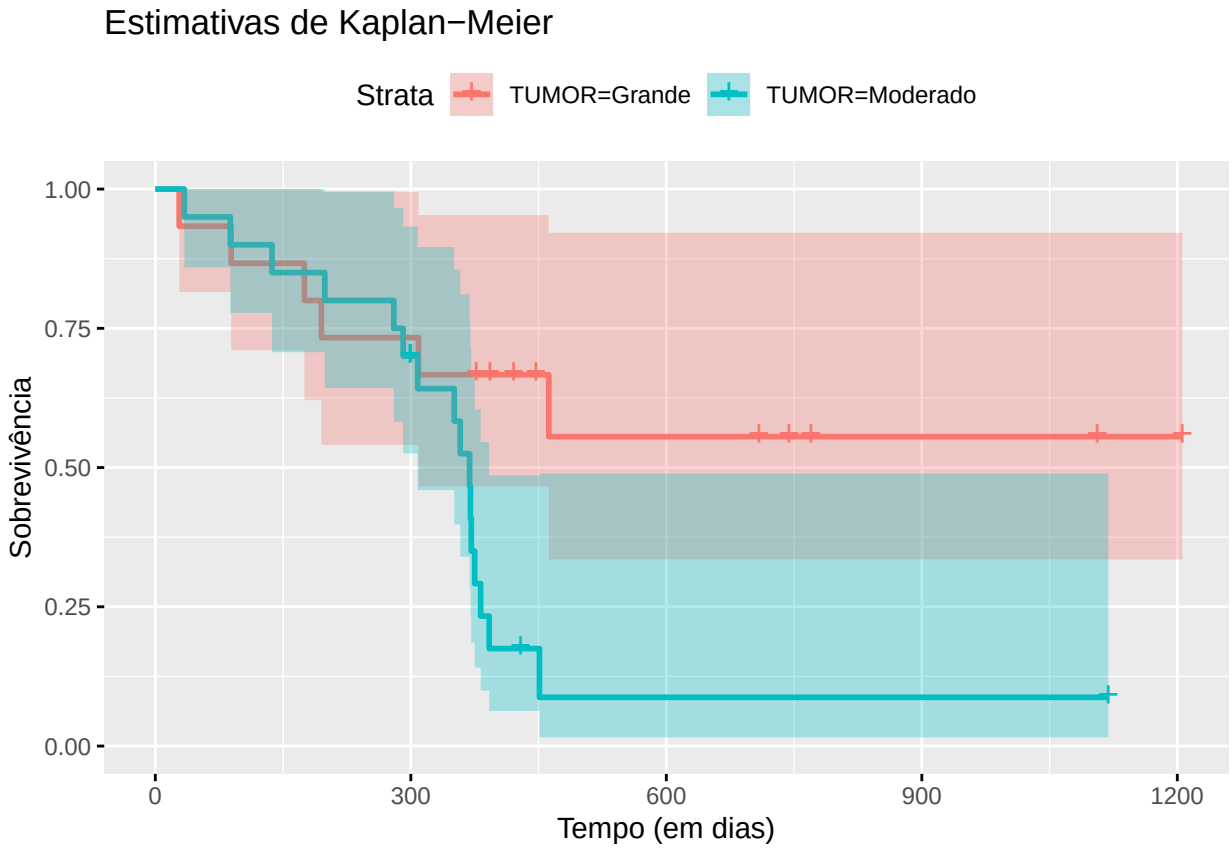
Vamos passar esses dados para o R:

```
# MONTANDO A BASE PARA O EXERCÍCIO 4 NO R

dados_ex4 <- data.frame(TEMPOS = c(28, 89, 175, 195, 309, 377, 393,
                                   421, 447, 462, 709, 744, 770, 1106,
                                   1206, 34, 88, 137, 199, 280, 291,
                                   299, 300, 308, 351, 358, 369, 370,
                                   371, 375, 382, 392, 429, 451, 1119),
                        TUMOR = c(rep("Grande", 15), rep("Moderado", 20)),
                        FALHAS = c(1, 1, 1, 1, 1, 0, 0,
                                   0, 0, 1, 0, 0, 0, 0,
                                   0, 1, 1, 1, 1, 1, 1,
                                   0, 0, 1, 1, 1, 1, 1,
                                   1, 1, 1, 1, 0, 1, 0))
```

a)

Vamos obter as curvas de Kaplan-Meier de cada um dos grupos:

```
# MONTANDO A BASE PARA O EXERCÍCIO 4 NO R

S_KM <- survfit(Surv(TEMPOS, FALHAS) ~ TUMOR, data = dados_ex4)
ggsurvplot(S_KM, data = dados_ex4, conf.int = T, ggtheme = theme_gray()) +
  labs(x = "Tempo (em dias)", y = "Sobrevivência", title = "Estimativas de Kaplan-Meier")
```



Analisando o gráfico, podemo perceber que para os tempos até 300 dias, as curvas estimadas de sobrevivência são próximas, cruzando-se em alguns pontos. Porém, a partir disso, o grupo com **tumor grande** apresenta uma sobrevivência estimada maior que o grupo com **tumor moderado**, ainda que seus intervalos de confiança se sobreponham.

b)

Vamos testar a igualdade das curvas utilizando o teste de log-rank e também testes da família de Fleming-Harington ($\rho = 0, 5$ e $\rho = 1$):

```
# FAZENDO O TESTES NO R

survdiff(Surv(TEMPOS, FALHAS) ~ TUMOR, data = dados_ex4) # log-rank
```

```
## Call:
## survdiff(formula = Surv(TEMPOS, FALHAS) ~ TUMOR, data = dados_ex4)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## TUMOR=Grande    15        6     11.4      2.55      5.62
## TUMOR=Moderado 20       16     10.6      2.73      5.62
##
##  Chisq= 5.6  on 1 degrees of freedom, p= 0.02
```

```r
survdiff(Surv(TEMPOS, FALHAS) ~ TUMOR, data = dados_ex4, rho = 0.5)  # família de Fleming-Harington (p = 0,5
```

```
## Call:
## survdiff(formula = Surv(TEMPOS, FALHAS) ~ TUMOR, data = dados_ex4,
##     rho = 0.5)
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## TUMOR=Grande   15     5.21     9.07      1.64       4.1
## TUMOR=Moderado 20    12.89     9.03      1.65       4.1
##
##  Chisq= 4.1  on 1 degrees of freedom, p= 0.04
```

```r
survdiff(Surv(TEMPOS, FALHAS) ~ TUMOR, data = dados_ex4, rho = 1)  # família de Fleming-Harington (p = 1)
```

```
## Call:
## survdiff(formula = Surv(TEMPOS, FALHAS) ~ TUMOR, data = dados_ex4,
##     rho = 1)
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## TUMOR=Grande   15     4.63      7.4     1.036       2.8
## TUMOR=Moderado 20    10.57      7.8     0.982       2.8
##
##  Chisq= 2.8  on 1 degrees of freedom, p= 0.09
```

A um nível de significância de 5%, o teste de log-rank nos leva a rejeitar a hipótese nula de igualdade das curvas de sobrevivência. Já para os teste da família de Fleming-Harington, com $\rho = 0,5$ nós rejeitamos a hipótese nula e com $\rho = 1$ nós não rejeitamos a igualdade das curvas de sobrevivência a um nível de significância de 5%.

c)

Vamos ajustar um modelo de riscos proporcionais de Cox:

```r
# AJUSTANDO O MODELO DE COX NO R

modelo_cox <- coxph(Surv(TEMPOS, FALHAS) ~ TUMOR, data = dados_ex4,
                    ties = c("efron", "breslow", "exact")[1])
modelo_cox
```

```
## Call:
## coxph(formula = Surv(TEMPOS, FALHAS) ~ TUMOR, data = dados_ex4,
##     ties = c("efron", "breslow", "exact")[1])
##
##                  coef exp(coef) se(coef)     z      p
## TUMORModerado 1.1283    3.0905   0.4969 2.271 0.0232
##
## Likelihood ratio test=5.83  on 1 df, p=0.01574
## n= 35, number of events= 22
```

Interpretando o coeficiente β estimado e tendo em vista que o grupo de referência é o com **tumor grande**, temos que o grupo com **tumor moderado** tem uma taxa de falha $exp(\hat{\beta}) = exp(1,13) = 3,09$ vezes a taxa de falha do grupo com **tumor grande**, com erro-padrão associado de, aproximadamente, 0,50.

d)

Vamos testar a hipótese $H_0 : \beta = 0$ utilizando a estatística do teste de Wald, calculada com base nas estimativas de β e de seu erro-padrão obtidos em no item **c)**. Para obter tal estatística, usaremos o comando **summary()** do R:

```
# INFORMAÇÕES DO MODELO DE COX NO R

summary(modelo_cox)
```

```
## Call:
## coxph(formula = Surv(TEMPOS, FALHAS) ~ TUMOR, data = dados_ex4,
##     ties = c("efron", "breslow", "exact")[1])
##
##   n= 35, number of events= 22
##
##                  coef exp(coef) se(coef)     z Pr(>|z|)
## TUMORModerado 1.1283    3.0905   0.4969 2.271   0.0232 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## TUMORModerado     3.091     0.3236     1.167     8.184
##
## Concordance= 0.588  (se = 0.064 )
## Likelihood ratio test= 5.83  on 1 df,   p=0.02
## Wald test            = 5.16  on 1 df,   p=0.02
## Score (logrank) test = 5.62  on 1 df,   p=0.02
```

O teste indica, a um nível de significância de 5%, a rejeição da hipótese nula de que $\beta = 0$, com valor-p de 0,02.

e)

Vamos testar a hipótese $H_0 : \beta = 0$ utilizando a estatística do teste do escore obtida no **Exercício 3**, na letra **c)**:

$$S(0) = \frac{U^2(0)}{I(0)} = \frac{\left(\sum_{i=1}^{d} x_i - \sum_{i=1}^{d} \frac{\sum_{j \in R_i} x_j}{\sum_{j \in R_i} 1}\right)^2}{\left(-\sum_{i=1}^{d} \frac{\sum_{j \in R_i} x_j}{\sum_{j \in R_i} 1} . \left[1 - \frac{\sum_{j \in R_i} x_j}{\sum_{j \in R_i} 1}\right]\right)}$$

Realizando o cálculo manual, temos que:

$$S(0) = \frac{(6 - 11,322)^2}{5,147} \cong 5,503$$

Vale lembrar que, sob H_0, $S(0) \sim \chi^2_{(1)}$. Logo, nível descritivo do teste será dado por $P(\chi^2_{(1)} > 5,503) \cong 0,019$. Assim, ao nível de significância de 5%, rejeitamos H_0.

Exercício 5

Vamos considerar o mesmo conjunto de dados da segunda lista de exercícios, disponíveis no arquivo "Lista2-hodgkins.xlsx", referentes a 60 pacientes com doença de Hodgkins que receberam tratamento padrão para a doença. O tempo de vida (em meses), bem como idade, sexo, histologia e estágio da doença de cada paciente foram observados.

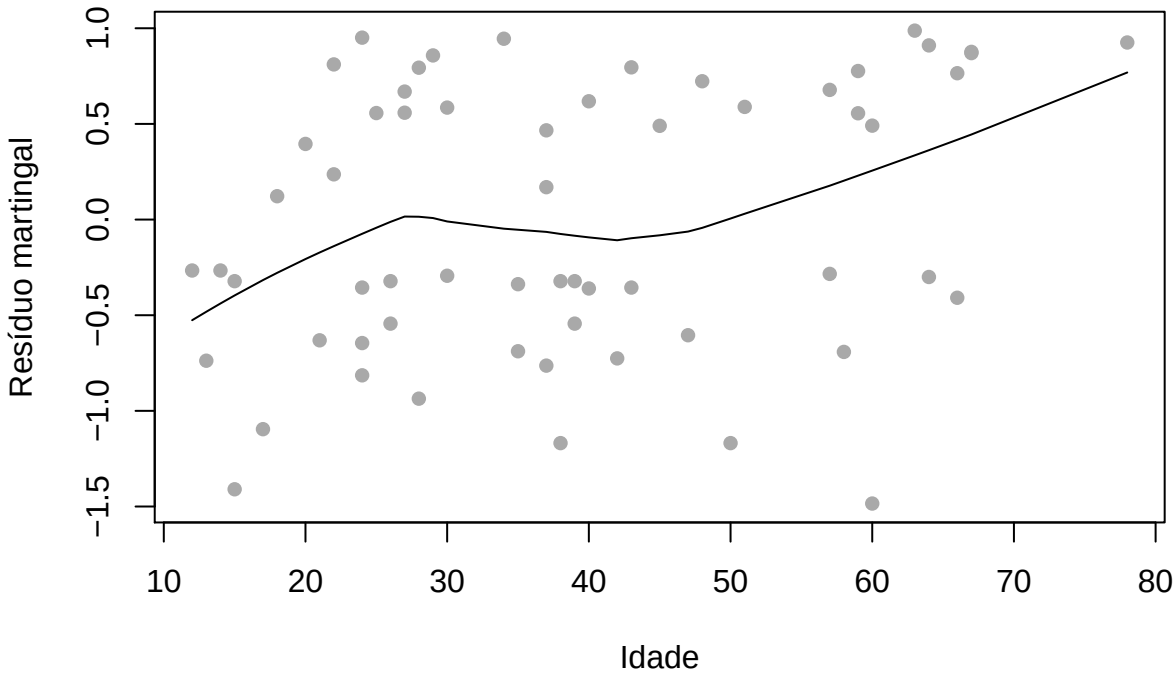
Vamos importar os dados para o R:

```
# IMPORTAR OS DADOS PARA O R

library(readxl)
dados_hodgkins <- read_excel("Lista2_Hodgkins.xlsx")
```

a)

Vamos ajustar um modelo de Cox sem a variável idade e, com o resíduo martingal, vamos avaliar a forma funcional dessa variável:

```
# AJUSTANDO O MODELO DE COX SEM IDADE NO R

modelo_hodgkins_sem_idade <- coxph(Surv(survivaltime, dead) ~ factor(sex) + factor(stage) + factor(hist),
                                   data = dados_hodgkins)
idade_transf <- ifelse(dados_hodgkins$age <= 40, 1, dados_hodgkins$age)
plot(dados_hodgkins$age, resid(modelo_hodgkins_sem_idade),
     xlab = 'Idade', ylab = 'Resíduo martingal', pch = 16, col = 'darkgray')
suavizacao <- lowess(dados_hodgkins$age, resid(modelo_hodgkins_sem_idade), iter = 0)
lines(suavizacao)
```



Analisando o gráfico, é possível perceber que os resíduos se distribuem de forma aproximadamente linear em relação à **idade**. Logo, é razoável que usemos no modelo a variável idade em seu formato original.

b)

Vamos ajustar um modelo de Cox incluindo todas as quatro covariáveis no modelo:

```
# AJUSTANDO O MODELO DE COX SEM IDADE NO R

modelo_hodgkins_completo <- coxph(Surv(survivaltime, dead) ~ factor(sex) + factor(stage) + factor(hist) + ag
                                  data = dados_hodgkins)
modelo_hodgkins_completo
```

```
## Call:
## coxph(formula = Surv(survivaltime, dead) ~ factor(sex) + factor(stage) +
##     factor(hist) + age, data = dados_hodgkins)
##
##                    coef exp(coef) se(coef)      z      p
## factor(sex)1    0.24416   1.27654  0.45492  0.537 0.5915
## factor(stage)1  0.84294   2.32320  0.41313  2.040 0.0413
## factor(hist)2  -0.13859   0.87058  0.42942 -0.323 0.7469
## factor(hist)3   1.30686   3.69455  0.59721  2.188 0.0287
## age             0.03028   1.03075  0.01241  2.441 0.0147
##
## Likelihood ratio test=18.64  on 5 df, p=0.002245
## n= 60, number of events= 30
```

c)

Considerando o modelo ajustado em **b)**, vamos selecionar as variáveis do modelo pelo teste da razão de verossimilhanças:

```
# ANALISANDO O MODELO COMPLETO NO R

modelo_hodgkins_sem_sexo <- coxph(Surv(survivaltime, dead) ~ factor(stage) + factor(hist) + age,
                                  data = dados_hodgkins)
modelo_hodgkins_sem_estagio <- coxph(Surv(survivaltime, dead) ~ factor(sex) + factor(hist) + age,
                                     data = dados_hodgkins)
modelo_hodgkins_sem_histologia <- coxph(Surv(survivaltime, dead) ~ factor(sex) + factor(stage) + age,
                                  data = dados_hodgkins)
modelo_hodgkins_sem_idade <- coxph(Surv(survivaltime, dead) ~ factor(sex) + factor(stage) + factor(hist),
                                   data = dados_hodgkins)

anova(modelo_hodgkins_sem_sexo, modelo_hodgkins_completo) # Testando sexo
```

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

```
anova(modelo_hodgkins_sem_estagio, modelo_hodgkins_completo) # Testando estágio
```

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

```
anova(modelo_hodgkins_sem_histologia, modelo_hodgkins_completo) # Testando histologia
```

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

```
anova(modelo_hodgkins_sem_idade, modelo_hodgkins_completo) # Testando idade
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(survivaltime, dead)
##  Model 1: ~ factor(sex) + factor(stage) + factor(hist)
##  Model 2: ~ factor(sex) + factor(stage) + factor(hist) + age
##    loglik  Chisq Df P(>|Chi|)
## 1 -104.40
## 2 -101.29 6.2124  1   0.01269 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A um nível de siginificância de 5%, selecionamos as seguintes variáveis:

- Estágio (*stage*);
- Idade (*age*).

Com isso, ajustamos o modelo final com essas duas variáveis:

```
# AJUSTANDO O MODELO DE COX SEM IDADE NO R

modelo_hodgkins_final <- coxph(Surv(survivaltime, dead) ~ factor(stage) + age,
                               data = dados_hodgkins)
modelo_hodgkins_final
```

```
## Call:
## coxph(formula = Surv(survivaltime, dead) ~ factor(stage) + age,
##     data = dados_hodgkins)
##
##                    coef exp(coef) se(coef)     z      p
## factor(stage)1 0.97067   2.63970  0.40653 2.388 0.0170
## age            0.03579   1.03644  0.01222 2.929 0.0034
##
## Likelihood ratio test=13.09  on 2 df, p=0.001434
## n= 60, number of events= 30
```

Analisando os coeficientes estimados pelo ajuste do modelo de Cox, temos que:

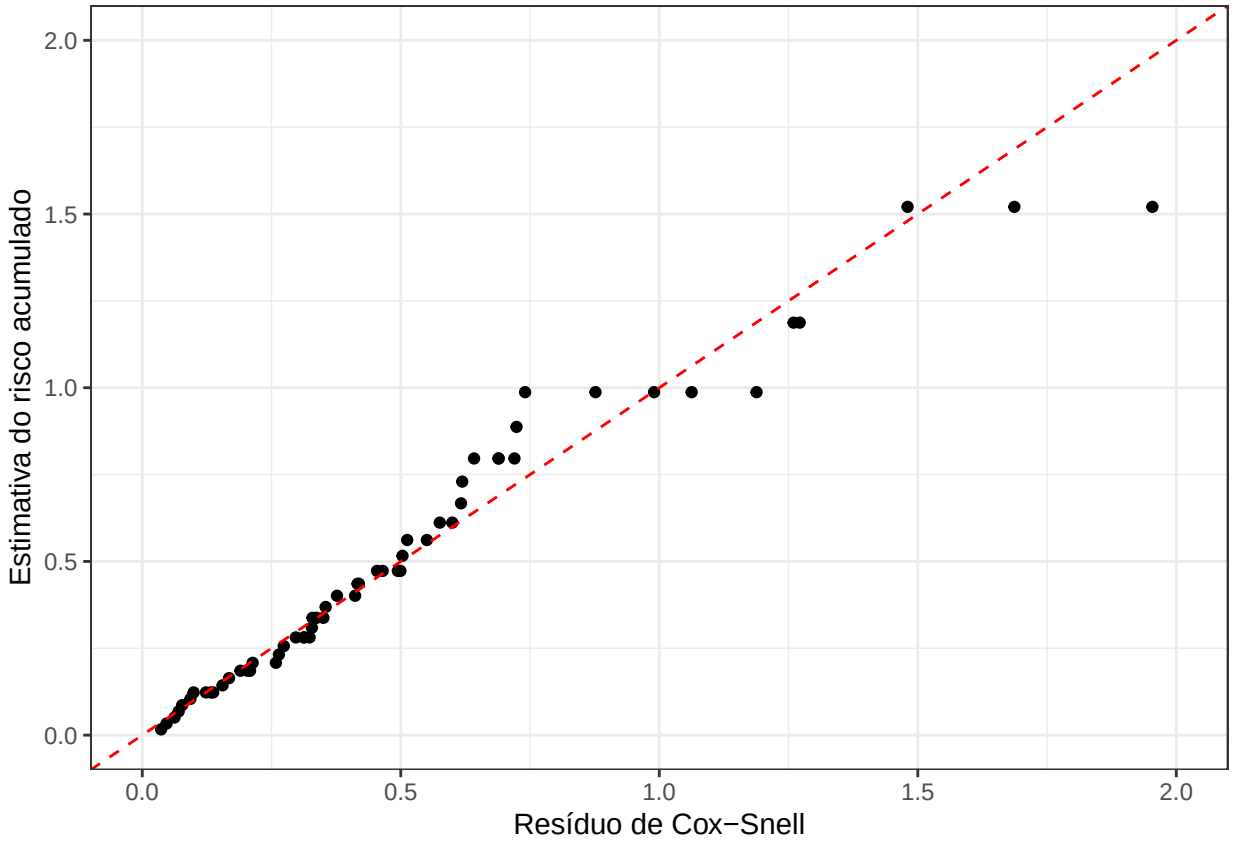
- **Estágio:** Os pacientes com estágio da doença **avançado** possuem uma taxa de falha 163,9% maior que os pacientes com estágio da doença **inicial**;
- **Idade:** Para cada acréscimo de 1 ano na idade do paciente, a taxa de falha, ou seja, a chance de morte desse paciente aumenta em 3,6%.

d)

Vamos obter os resíduos de Cox-Snell para o modelo final ajustado no item **c)**:

```
# RESÍDUOS DE COX-SNELL NO R

resid_mart <- residuals(modelo_hodgkins_final, type = "martingale")
resid_coxsnell <- -(resid_mart - dados_hodgkins$dead)
fit_coxsnell <- coxph(formula = Surv(resid_coxsnell, dead) ~ 1,
                      data = dados_hodgkins, ties = c("efron","breslow","exact")[1])

df_base_haz <- basehaz(fit_coxsnell, centered = FALSE)
```

```r
ggplot(data = df_base_haz, mapping = aes(x = time, y = hazard)) +
   geom_point() + labs(x = "Resíduo de Cox-Snell", y = "Estimativa do risco acumulado") +
   scale_x_continuous(limit = c(0, 2)) + scale_y_continuous(limit = c(0, 2)) +
   theme_bw() + theme(legend.key = element_blank()) +
   geom_abline(intercept = 0, slope = 1, lty = 'dashed', col = 'red')
```

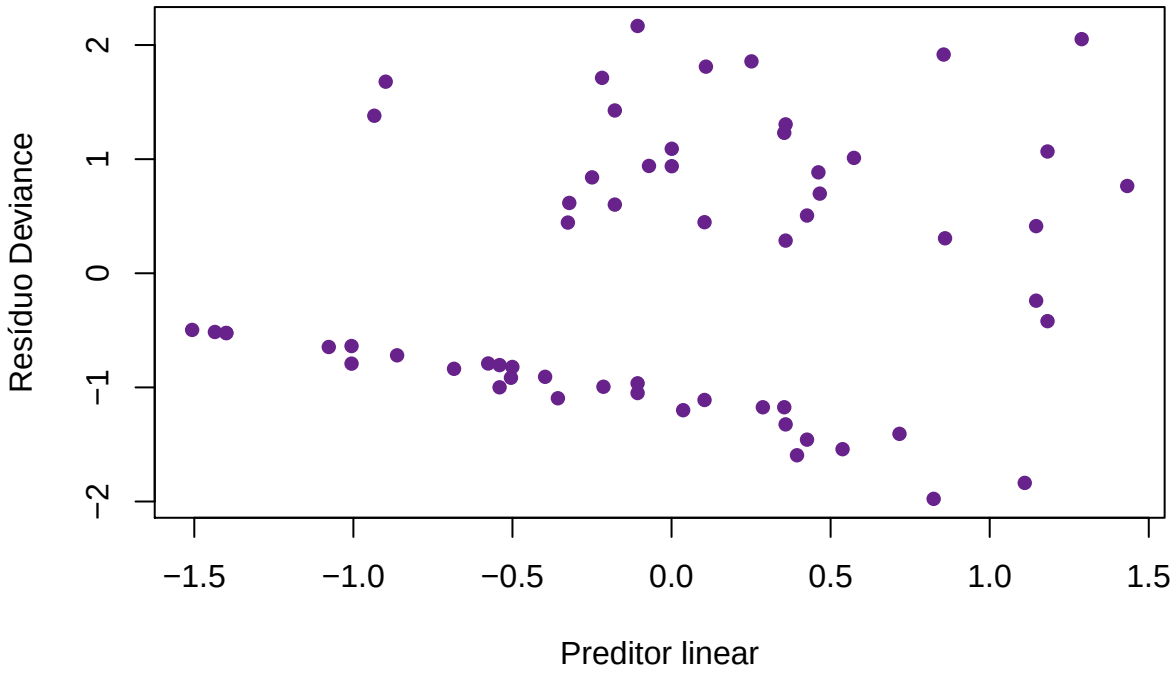


O gráfico de resíduos de Cox-Snell versus as estimativas do risco acumulado mostra que os pontos se distribuem aproximadamente em torno da reta $x = y$, o que nos leva a concluir que o modelo ajustado é razoável para os dados.

e)

Vamos obter os resíduos *deviance* a fim de encontrar pontos atípicos através do gráfico:

```r
# RESÍDUOS DEVIANCE NO R

plot(modelo_hodgkins_final$linear.predictors, resid(modelo_hodgkins_final, type='deviance'),
     xlab = "Preditor linear", ylab = "Resíduo Deviance", pch = 16, col = "darkorchid4")
```



Os resíduos estão dispostos aleatoriamente ao redor do zero (entre -2 e 2), o que indica que não há motivos para classificar algum ponto atípico ou desconfiar da adequabilidade do modelo.

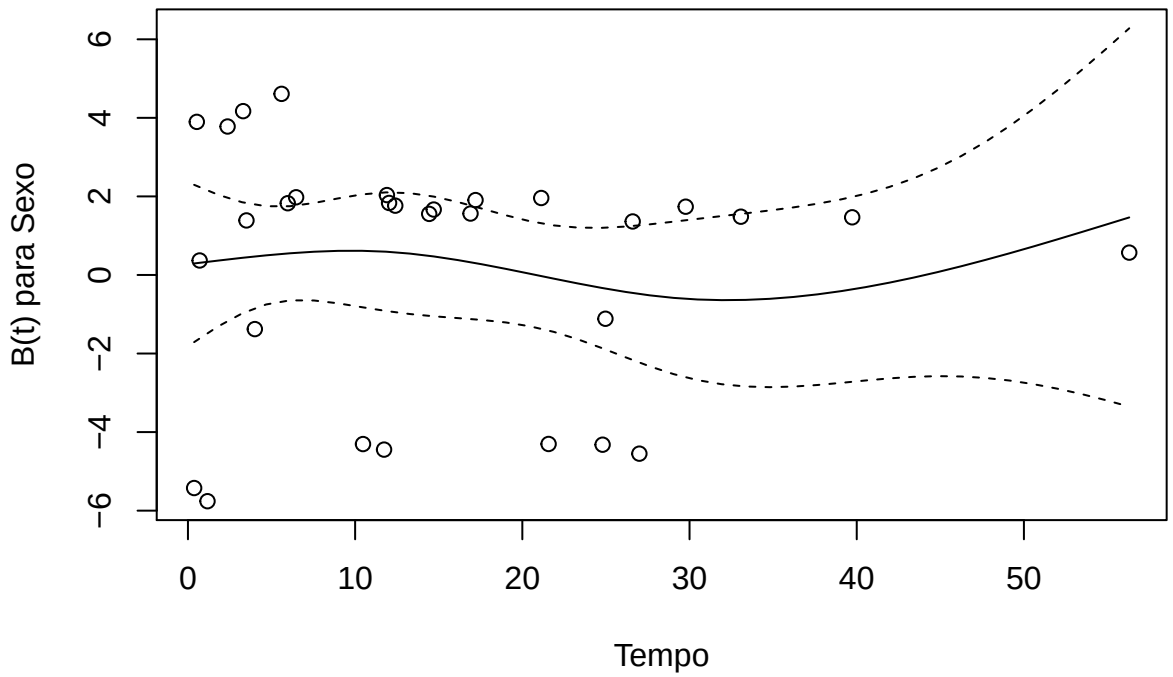
f)

Agora, vamos obter os resíduos de Schoenfeld do modelo no item **b)** e o teste para proporcionalidade dos riscos:

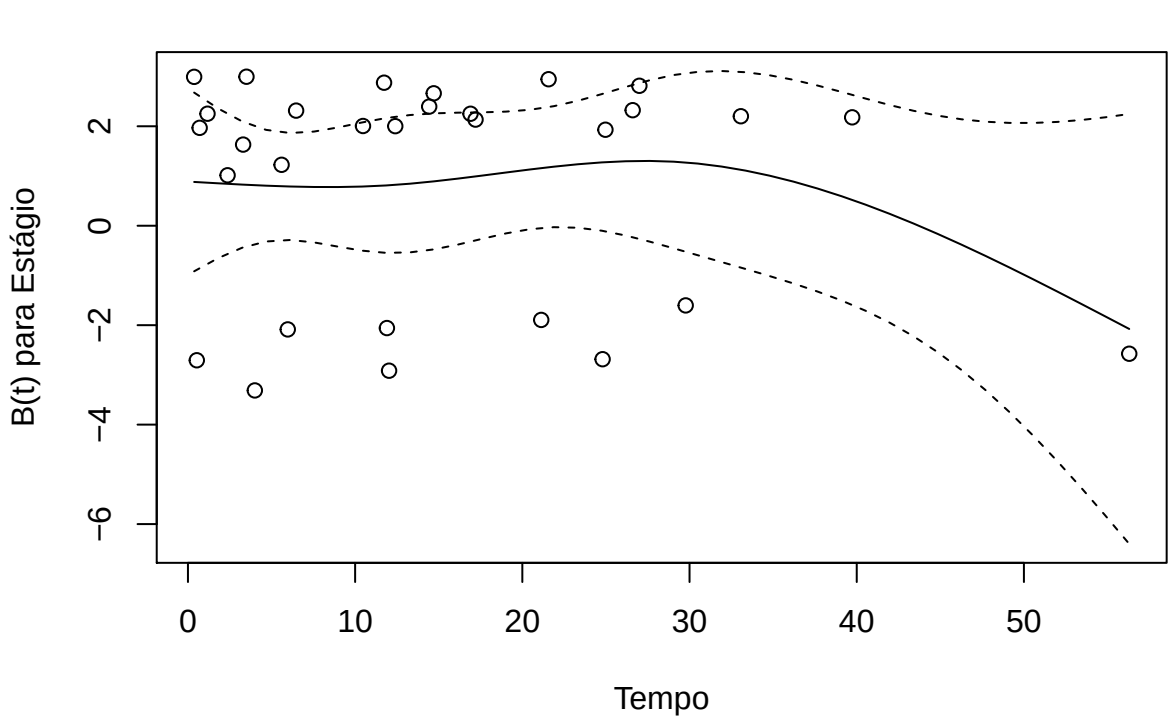
```
# RESÍDUOS SCHOENFELD NO R

sch <- cox.zph(modelo_hodgkins_completo, transform = 'identity')

sch
```

```
##                chisq df     p
## factor(sex)   0.0325  1 0.857
## factor(stage) 0.2572  1 0.612
## factor(hist)  5.3646  2 0.068
## age           0.6176  1 0.432
## GLOBAL        5.5272  5 0.355
```

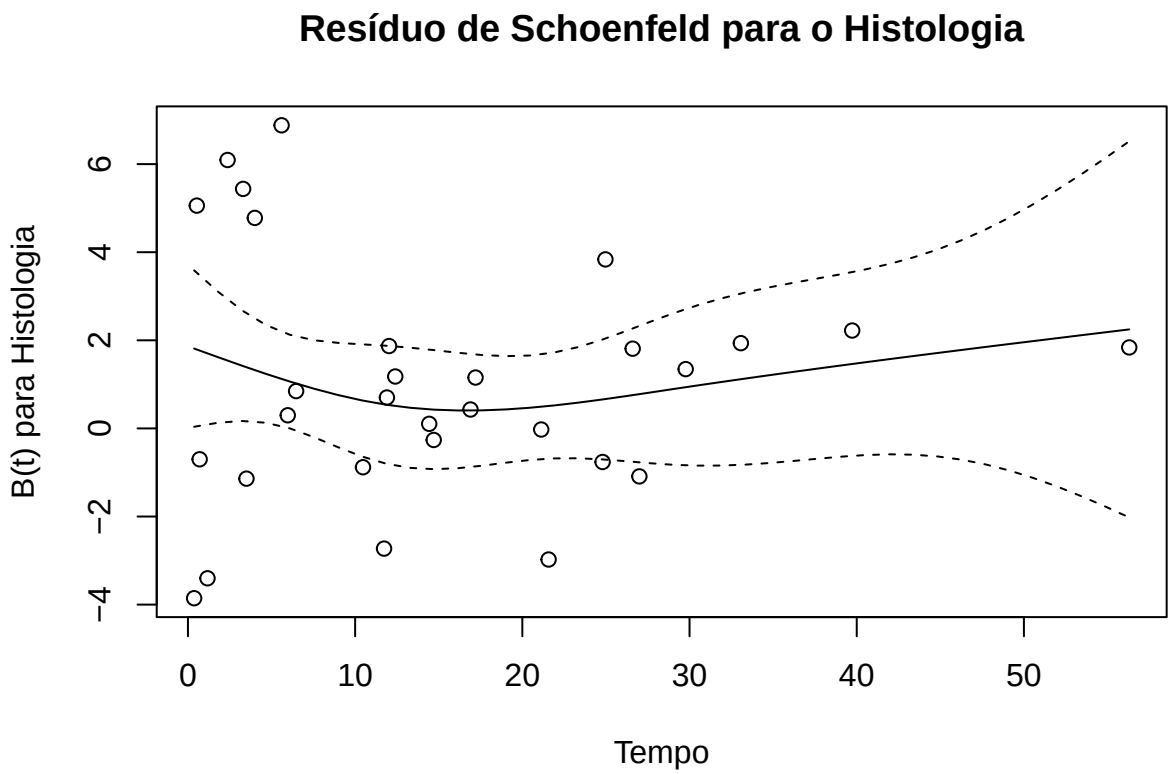
```
plot(sch[1,], main = "Resíduo de Schoenfeld para o Sexo",
     xlab = "Tempo", ylab = "B(t) para Sexo")
```



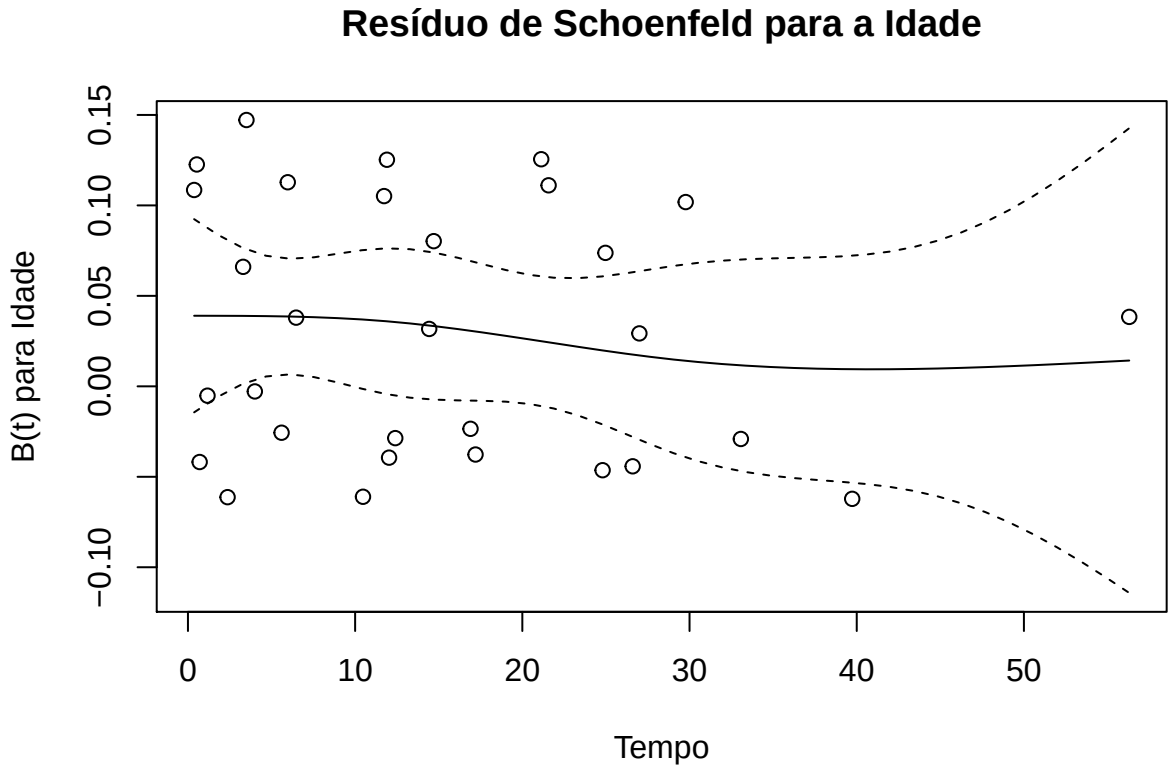

```
plot(sch[2,], main = "Resíduo de Schoenfeld para o Estágio",
     xlab = "Tempo", ylab = "B(t) para Estágio")
```




```
plot(sch[3,], main = "Resíduo de Schoenfeld para o Histologia",
     xlab = "Tempo", ylab = "B(t) para Histologia")
```




```
plot(sch[4,], main = "Resíduo de Schoenfeld para a Idade",
     xlab = "Tempo", ylab = "B(t) para Idade")
```



Em todos os gráficos de $\beta(t)$ contra t, é possível observar linhas paralelas ao eixo do tempo de modo que estejam completamente dentro da região de confiança, o que valida a suposição de riscos proporcionais. Além disso, os valores-p dos testes individuais evidenciam essa conclusão gráfica, a um nível de significância de 5% (com a Histologia com um

valor-p mais baixo):

- **Valor-p do Sexo:** 0,86
- **Valor-p do Estágio:** 0,61
- **Valor-p do Histologia:** 0,07
- **Valor-p da Idade:** 0,43