Atividade 3

Guilherme Navarro NUSP: 8943160

03 de agosto de 2020

Atividade 3

Considere o conjunto de dados sobre câncer de mama, obtidos de um grupo de estudos de câncer de mama da Alemanha. Os dados são referentes a 686 mulheres diagnosticadas com câncer da mama entre julho de 1984 e dezembro de 1989. Algumas variáveis disponíveis são:

- Tempo (survtime): tempo de vida desde o diagnóstico até o óbito (em dias);
- Delta (censdead): indicadora de óbito ou censura à direita;
- Idade (age): idade na data do diagnóstico;
- Terapia hormonal (hormone): indicadora de terapia hormonal (2 = Sim / 1= Não);
- Estadiamento (grade): estadiamento do tumor (1, 2 e 3), que é uma classificação que indica a gravidade do tumor encontrado;
- Tamanho (size): tamanho do tumor primário, em milímetros;
- Linfonodos (nodes): número de linfonodos comprometidos;
- Menopausa (menopause): indicadora de ocorrência de menopausa no diagnóstico (2=Sim/1=Não).

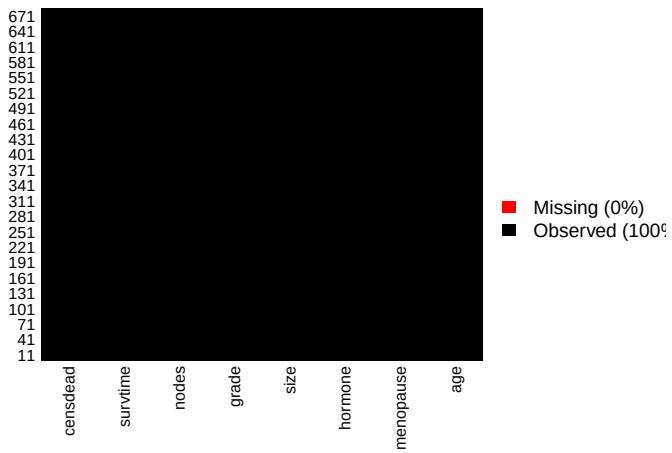
Os dados estão disponíveis no arquivo Dados-ex3-prova2.csv e a descrição está no arquivo Dados- ex3-prova2.txt. Utilizando esses dados, responda os itens descritos a seguir:

(a) Faça uma análise descritiva dos dados. Essa análise descritiva deve envolver curvas de Kaplan-Meier segundo as covariáveis descritas, bem como testes para comparação das curvas obtidas.

Resolução

Ao iniciar a análise irei fazer um mapa de calor com os missings da base de dados:

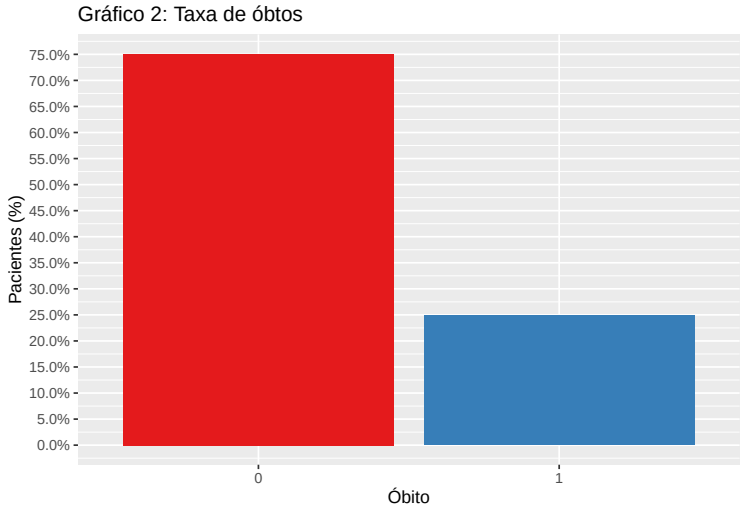
Gráfico 1: Heatmap missing



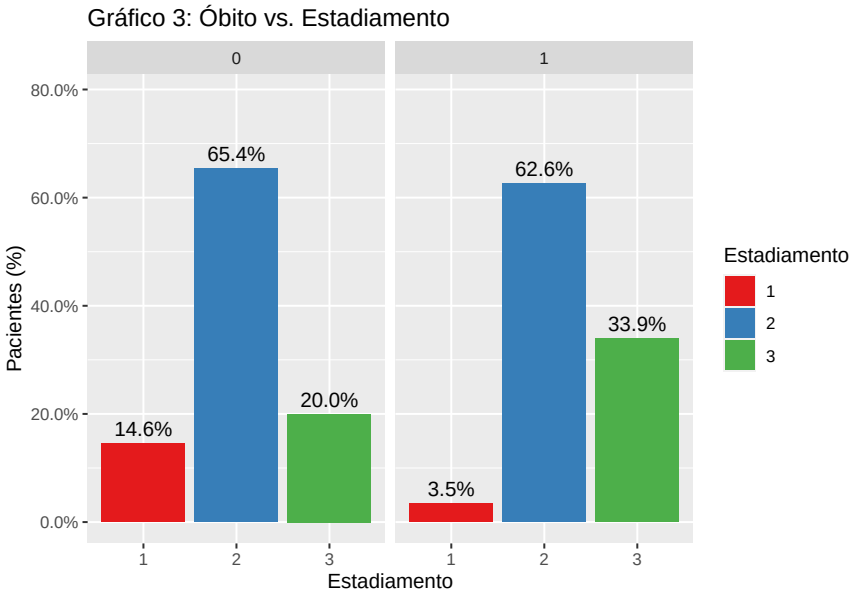
Como se pode verificar não há nenhum *missing value* (gráfico 1). Temos variáveis do tipo categórica (Terapia hormonal, Estadiamento e Menopausa) e variáveis contínuas (Idade,Tamanho tumor e Número de Linfonodos), para as contínuas, foi feita uma categorização apenas para fins descritivos, sendo as categorias definidas por seus quartis (quando viáveis), assim foi definido:

- Idade: idade_cat como sendo menor que 46 anos (1º quartil), entre 46 e 61 anos (3º quartil) e maior que 61 anos.
- Número de infonodos comprometidos: nodes_cat como sendo menor que 3 (mediana), entre 3 e 7 (3º quartil) e maior que 7.
- Tamanho do tumor: size_Cat como sendo menor que 20 mm (1º quartil), entre 20 e 35 mm (3º quartil) e maior que 35 mm.

Feita estas transformações, podemos partir para análise descritiva, sendo assim:

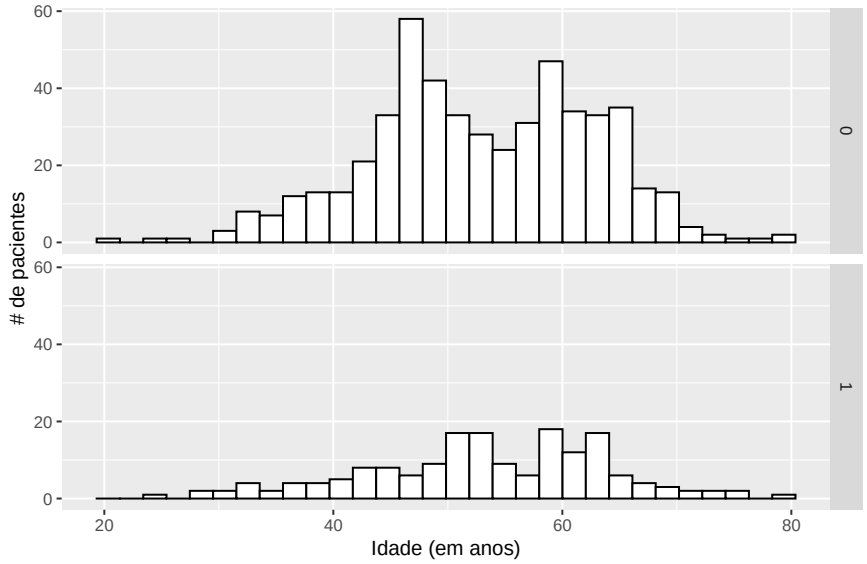


Em que podemos notar que cerca de 25% das pacientes morreram.



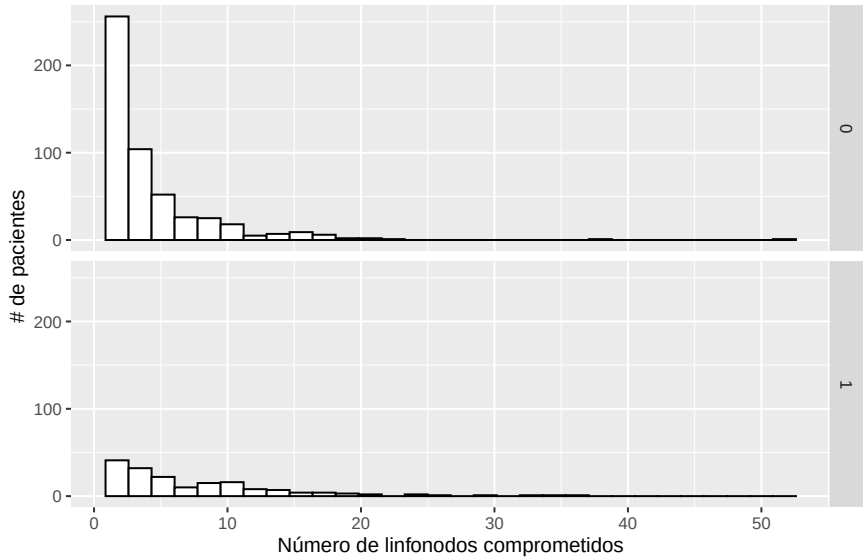
No gráfico 3, das pacientes que morreram cerca de 3,5% tiveram o estadiamento do tumor número 1, seguido de 39% com número 3, enquanto isso para as pacientes que sobreviveram no periodo do estudo, cerca de 14,6% o estadiamento do tumor número 1 enquanto 20% com número 3 e para o estadiamento do tumor número 2 não há muita diferença entre quem veio a óbito ou não com aproximadamente 63% das pacientes para cada caso.

Gráfico 4: Histograma Idade vs. óbito



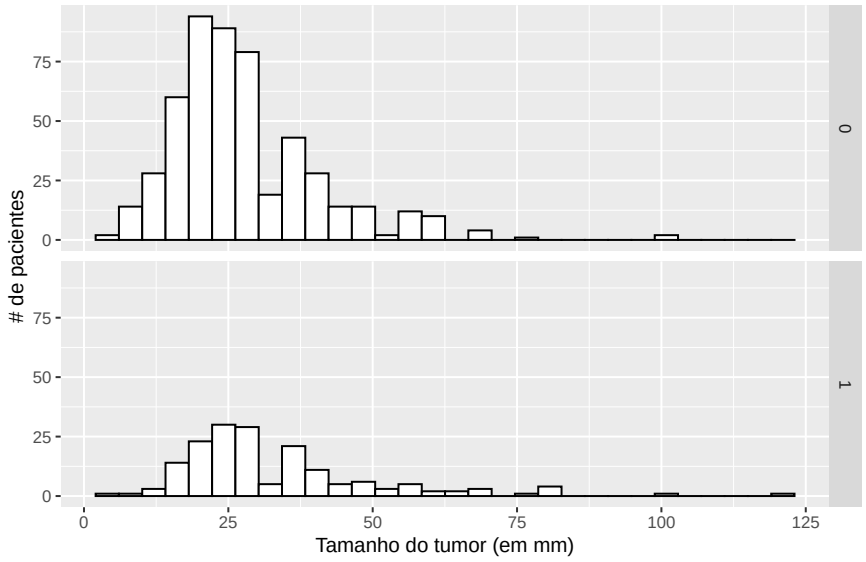
No gráfico 4, é possível ver que a idade das pacientes não importa muito, pois os comportamentos são muito próximos até mesmo para as caudas indicando que idade não execerce influencia no óbito das pacientes do estudo.

Gráfico 5: Número de linfonodos comprometidos vs. óbito



No gráfico 5, é possível ver que temos a maioria das pacientes que sobreviveram no periodo do estudo tem um número menor de linfonodos comprometidos, enquanto que para as pacientes que vieram a óbito observando as que tiveram mais de 10 de linfonodos comprometidos tem uma pequena diferença em relação as pacientes que sobreviveram (uma leve piora).

Gráfico 6: Histograma tamanho do tumor vs. óbito



No gráfico 6, é possível ver que a maior parte das pacientes que sobreviveram no período do estudo tinham um tumor médio de 25 mm, enquanto as pacientes que vieram a óbito tinha um tumor médio de 30 mm, além disso podemos ver que na cauda mais a direita para quem faleceu temos tumores de tamanho maiores.

Avaliando os gráficos de Kaplan-Meier para cada covariável, temos:

Para a covariável menopausa:

Gráfico 7: Estimativas de Kaplan−Meier (Menopausa)

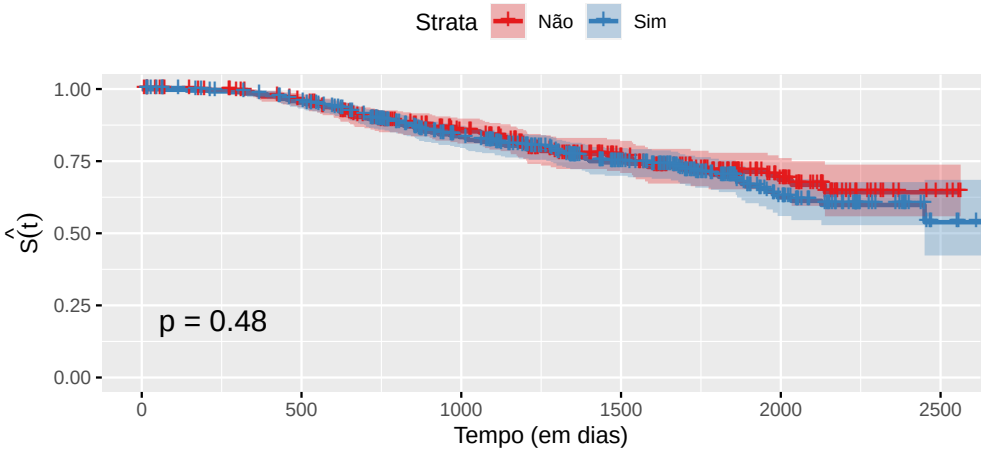
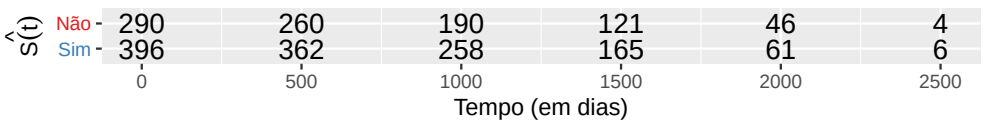


Gráfico 7: Estimativas de Kaplan−Meier (Menopausa)



Queremos testar a igualdade das curvas, assim:

$$\begin{cases} H_0 : S_1(t) = S_2(t), \ \forall \ t \in \ [0, \tau] \\ H_1 : S_1(t) \neq S_2(t) \ para \ algum \ t \in \ [0, \tau] \end{cases}$$

Em que τ é o maior instante observado tal que os dois grupos possuem pelo menos um indivíduo em risco.

Sob a hipótese nula, a estatística do teste Log-Rank é:

$$L_r = \frac{[\sum_{j=1}^{L}(d_{2j} - e_{2j})]^2}{\sum_{j=1}^{L} V_j^2}$$

Em que d_{2j} é o # de indivíduos observados no grupo 2, e_{2j} é o # de indivíduos esperadados no grupos 2 e V_j é a variância de d_{2j} que é dada por:

$$V_j = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_{1_j} - 1)}$$

Em que n_{1j} e n_{2j} são o número de indivíduos nos grupo 1 e 2 respectivamente. Assim sendo, sob a hipótese nula,

$$L_r \overset{a}{\sim} \chi^2_{(1)}$$

Utilizando o teste log-rank, temos:

variable	pval	method
menopause	0.484	Log-rank

Pelo gráfico 7 podemos notar que as estimavivas de Kaplan-Meier que compõem as curvas são muito próximas e podemos confirmar na tabela acima com o teste de Log-Rank que as curvas são iguais com um p-valor de 0.48 a nível de significância de 5% não rejeitamos H_0.

Gráfico 8: Estimativas de Kaplan−Meier (Terapia Hormonal)

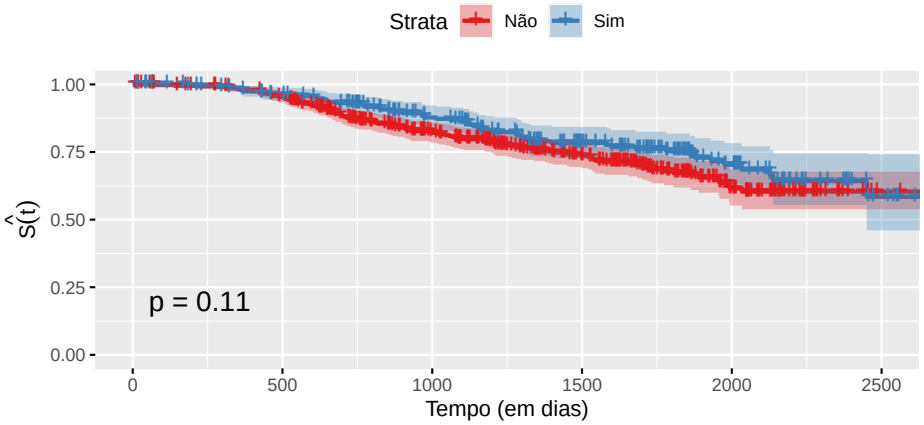
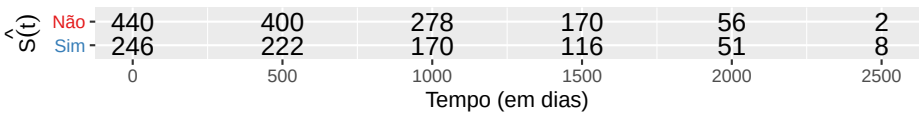


Gráfico 8: Estimativas de Kaplan−Meier (Terapia Hormonal)



Utilizando o teste log-rank, temos:

variable	pval	method
hormone	0.109	Log-rank

Pelo gráfico 8 podemos notar que as estimavivas de Kaplan-Meier que compõem as curvas são muito próximas e podemos confirmar na tabela acima com o teste de Log-Rank que as curvas são iguais com um p-valor de 0.11 a nível de significância de 5% não rejeitamos H_0.

Para a covariável Estadiamento (grade), por ter mais categorias, apresentou o seguinte gráfico com as estimativas de Kaplan-Meier:

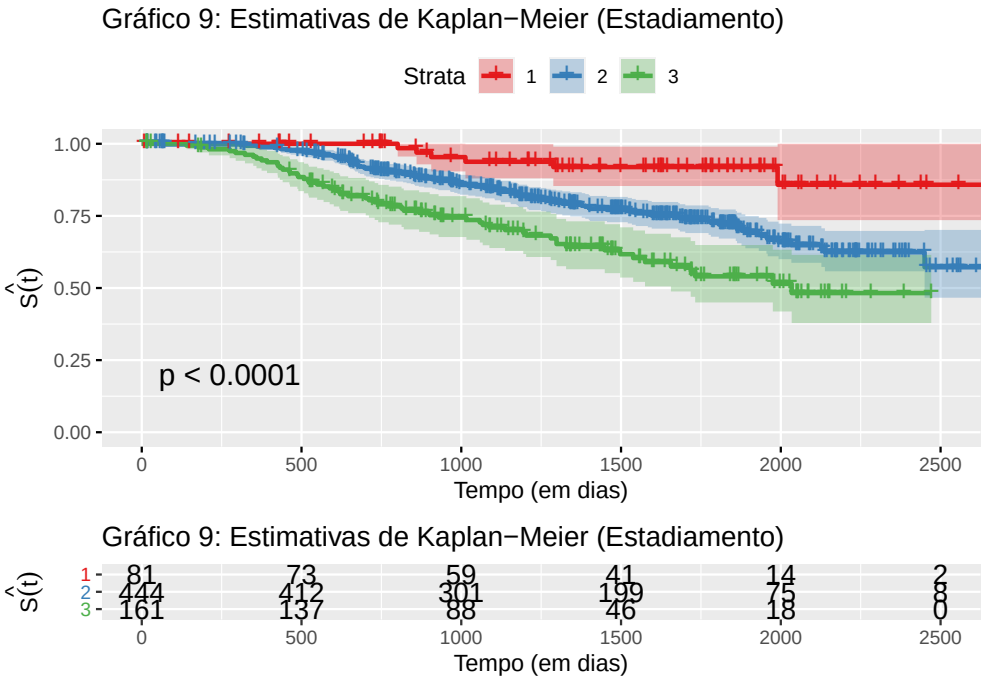


Gráfico 9: Estimativas de Kaplan−Meier (Estadiamento)

Queremos comparar se pelo menos uma das curvas são diferentes, assim utilizando o teste log-rank generalizado, com a seguinte hipóteses, temos:

Sob a hipótese:

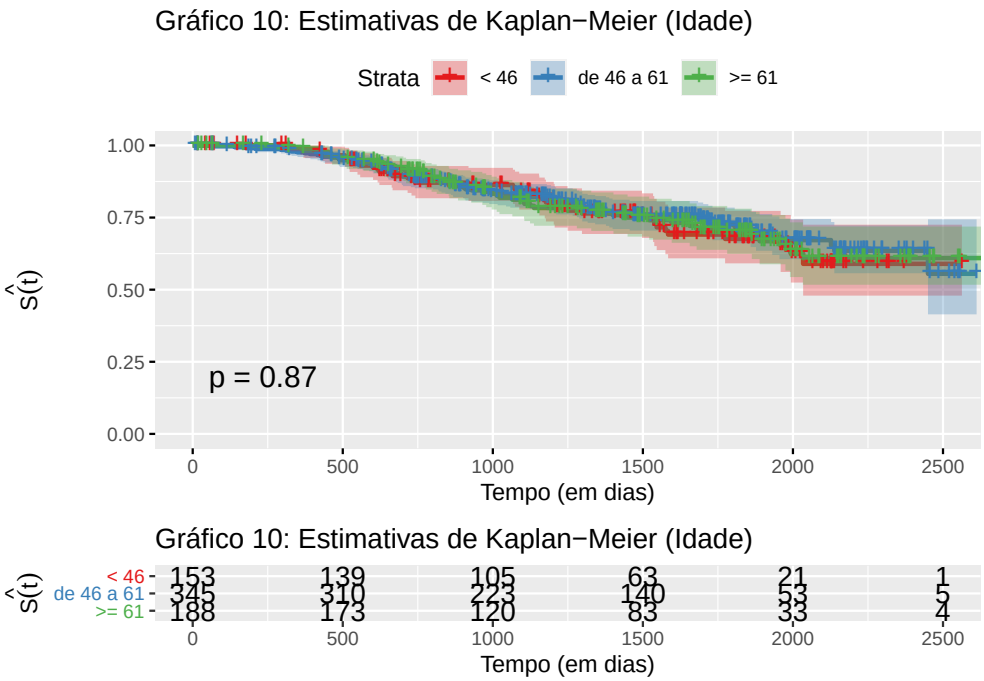
$$\begin{cases} H_0 : S_1(t) = S_2(t) = S_3(t), \ \forall \ t \in \ [0, \tau] \\ H_1 : pelo\ menos\ uma\ funcao\ diferente\ para\ algum\ t \in \ [0, \tau] \end{cases}$$

Utilizando o teste log-rank generalizado:

variable	pval	method
grade	0	Log-rank

Em que segundo o teste de Log-Rank e o gráfico das estimavivas de Kaplan-Meier pelo menos uma das curvas não são iguais (gráfico 9) a um nível de significância de 5%, sendo que o estadiamento número 3 tem um decaimento mais rápido do que os dois primeiros.

Para a covariável Idade categorizada temos uma situação idêntica a anterior, gerando um gráfico com as estimativas de Kaplan-Meier:

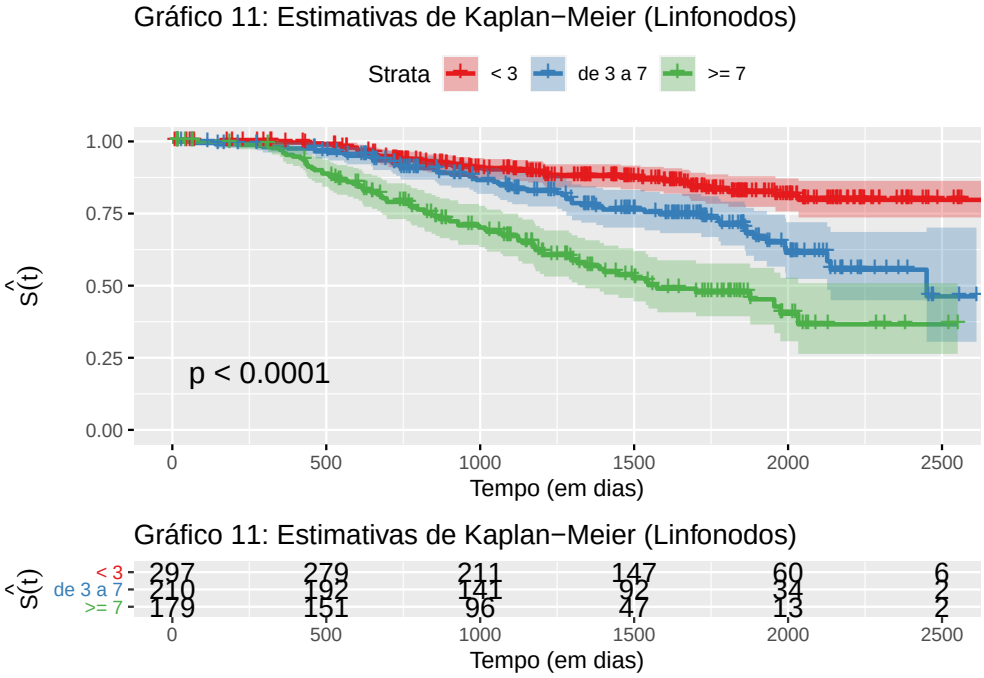


Utilizando o teste log-rank generalizado, temos:

variable	pval	method
age_cat	0.867	Log-rank

Pelo gráfico 10 podemos notar que as estimavivas de Kaplan-Meier que compõem as curvas são muito próximas, assim como no gráfico 4 que apresenta resultados similares, podemos confirmar na tabela acima com o teste de Log-Rank que as curvas são iguais com um p-valor de 0.87 a nível de significância de 5% não rejeitamos H_0.

Para a covariável número de linfonodos comprometidos categorizada temos uma situação idêntica a anterior, gerando um gráfico com as estimativas de Kaplan-Meier:

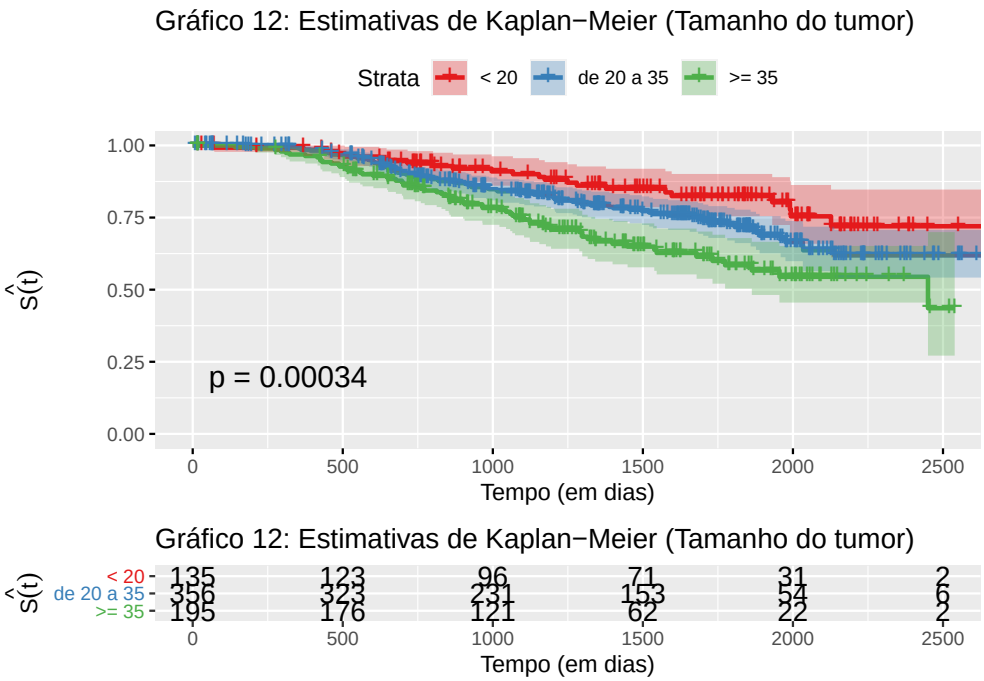


Utilizando o teste log-rank generalizado, temos:

variable	pval	method
nodes_cat	0	Log-rank

Em que segundo o teste de Log-Rank e o gráfico das estimavivas de Kaplan-Meier pelo menos uma das curvas não são iguais (gráfico 11) a um nível de significância de 5%, sendo que o para as pacientes com número de linfonodos comprometidos maiores que 7 tem um decaimento mais rápido do que as pacientes com número de linfonodos comprometidos entre 3 e 7, e para as pacientes com menos de 3 tem uma alta probabilidade de sobrevivência.

Para a covariável tamanho do tumor categorizada temos uma situação idêntica a anterior, gerando um gráfico com as estimativas de Kaplan-Meier:



Utilizando o teste log-rank generalizado, temos:

variable	pval	method
size_cat	3e-04	Log-rank

Em que segundo o teste de Log-Rank e o gráfico das estimavivas de Kaplan-Meier pelo menos uma das curvas não são iguais (gráfico 12) a um nível de significância de 5%, sendo que o para as pacientes com tumores maiores que 35 mm tem um decaimento mais rápido do que as pacientes tumores entre 20 e 35 mm, e para as pacientes com tumores menores (<20 mm) tem uma probabilidade maior de sobrevivência.

(b) Ajuste o modelo de riscos proporcionais de Cox aos dados. Observe que as variáveis idade, linfonodos e tamanho podem ser consideradas como variáveis contínuas. Apresente os resultados do modelo completo, com todas as covariáveis incluídas. Faça um processo de seleção de variáveis e apresente o resultado do modelo final obtido. Você precisa descrever claramente o processo de seleção das variáveis adotado, mas deve apresentar apenas as estimativas e resultados de dois modelos: modelo completo e modelo final.

Resolução

Ajustando o modelo completo com todas as variáveis:

```
## Call:
## coxph(formula = Surv(survtime, censdead) ~ ., data = data, ties = "breslow")
##
##   n= 686, number of events= 171
##
##                coef exp(coef)  se(coef)      z Pr(>|z|)
## age       -0.001969  0.998033  0.011530 -0.171 0.864386
## menopause  0.258348  1.294789  0.247916  1.042 0.297376
## hormone   -0.270758  0.762801  0.167942 -1.612 0.106916
## size       0.012505  1.012584  0.004712  2.654 0.007955 **
## grade2     1.093277  2.984036  0.420824  2.598 0.009378 **
## grade3     1.633708  5.122835  0.431138  3.789 0.000151 ***
## nodes      0.054695  1.056219  0.009296  5.884    4e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## age          0.9980     1.0020    0.9757     1.021
## menopause    1.2948     0.7723    0.7965     2.105
## hormone      0.7628     1.3110    0.5489     1.060
## size         1.0126     0.9876    1.0033     1.022
## grade2       2.9840     0.3351    1.3080     6.808
## grade3       5.1228     0.1952    2.2005    11.926
## nodes        1.0562     0.9468    1.0371     1.076
##
## Concordance= 0.705  (se = 0.02 )
## Likelihood ratio test= 78.24  on 7 df,   p=3e-14
## Wald test            = 90.41  on 7 df,   p=<2e-16
## Score (logrank) test = 101.1  on 7 df,   p=<2e-16
```

Com o modelo ajustado, utilizando o teste da razão de verossimilhanças, obtemos a seguinte tabela:

	Df	AIC	LRT	Pr(>Chi)
	NA	1991.779	NA	NA
age	1	1989.808	0.029	0.864
menopause	1	1990.864	1.085	0.298
hormone	1	1992.443	2.664	0.103
size	1	1996.207	6.428	0.011
grade	2	2011.289	23.510	0.000
nodes	1	2015.272	25.493	0.000

E como podemos ver, temos algumas variáveis que não são significativas (a um nível de significância fixado de 5%) para o modelo, assim removendo uma por vez e ajustando o modelo para cada remoção, chega-se no seguinte modelo final:

```
## Call:
## coxph(formula = Surv(survtime, censdead) ~ size + grade + nodes,
##     data = data, ties = "breslow")
##
##   n= 686, number of events= 171
##
##            coef exp(coef) se(coef)     z Pr(>|z|)
## size   0.012156  1.012230 0.004632 2.624 0.008685 **
## grade2 1.105068  3.019428 0.420736 2.627 0.008627 **
## grade3 1.651139  5.212912 0.431053 3.830 0.000128 ***
## nodes  0.054688  1.056211 0.009389 5.825 5.72e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## size       1.012     0.9879     1.003     1.021
## grade2     3.019     0.3312     1.324     6.887
## grade3     5.213     0.1918     2.240    12.134
## nodes      1.056     0.9468     1.037     1.076
##
## Concordance= 0.704  (se = 0.02 )
## Likelihood ratio test= 74.27  on 4 df,   p=3e-15
## Wald test            = 85.76  on 4 df,   p=<2e-16
## Score (logrank) test = 96.69  on 4 df,   p=<2e-16
```

(c) Interprete os parâmetros do modelo final obtido em (b).

Resolução

O modelo de cox é interpretado por meio de riscos relativos, que são dados por:

$$\frac{\alpha(t|x_1)}{\alpha(t|x_2)} = \frac{e^{x_1'\beta}}{e^{x_2'\beta}} = e^{\beta(x_1'-x_2')}$$

Assim para o modelo final do item anterior, temos:

Fixando as outras covariáveis, pode-se dizer que o acréscimo de uma unidade no tamanho do tumor (size) aumenta em $e^{0.012} = 1.012$ (1,2%) o risco de óbito por câncer de mama.

Assim, fixando as outras covariáveis, pode-se dizer que o acréscimo de uma unidade no número de linfonodos comprometidos (nodes) aumenta em $e^{0.054} = 1.056$ (5,6%) o risco de óbito por câncer de mama.

Para a covariável estadiamento = 2 (grade=2), fixando as demais covariáveis o risco de óbito é 3 vezes maior do que à paciente com estadiamento = 1.

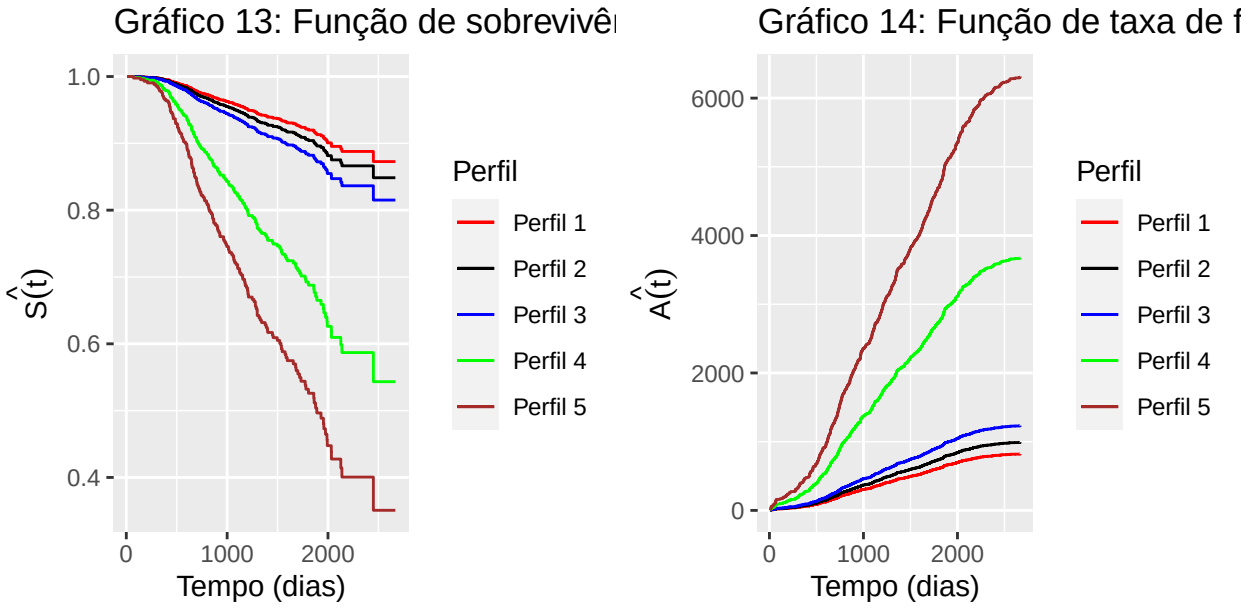
Para a covariável estadiamento = 3 (grade=3), fixando as demais covariáveis o risco de óbito é 5 vezes maior do que à paciente com estadiamento = 1.

(d) Utilizando o estimador de Breslow para função de taxa de falha acumulada do modelo final em (b), obtenha gráficos da função de taxa de falha acumulada e da função de sobrevivência parà pacientes nas seguintes situações:

- Idade de 53 anos, sem terapia hormonal, na menopausa, tumor de 15 milímetros, com estadiamento 1 e apenas 1 linfonodo comprometido;
- Idade de 53 anos, sem terapia hormonal, na menopausa, tumor de 30 milímetros, com estadiamento 1 e apenas 1 linfonodo comprometido;
- Idade de 53 anos, sem terapia hormonal, na menopausa, tumor de 30 milímetros, com estadiamento 1 e 5 linfonodos comprometidos;
- Idade de 53 anos, sem terapia hormonal, na menopausa, tumor de 30 milímetros, com estadiamento 2 e 5 linfonodos comprometidos;
- Idade de 53 anos, sem terapia hormonal, na menopausa, tumor de 30 milímetros, com estadiamento 3 e 5 linfonodos comprometidos.

Você deve fazer um gráfico para função de taxa de falha acumulada, comparando as 5 curvas referentes a cadà paciente, e um gráfico com as 5 curvas para a função de sobrevivência.

Resolução



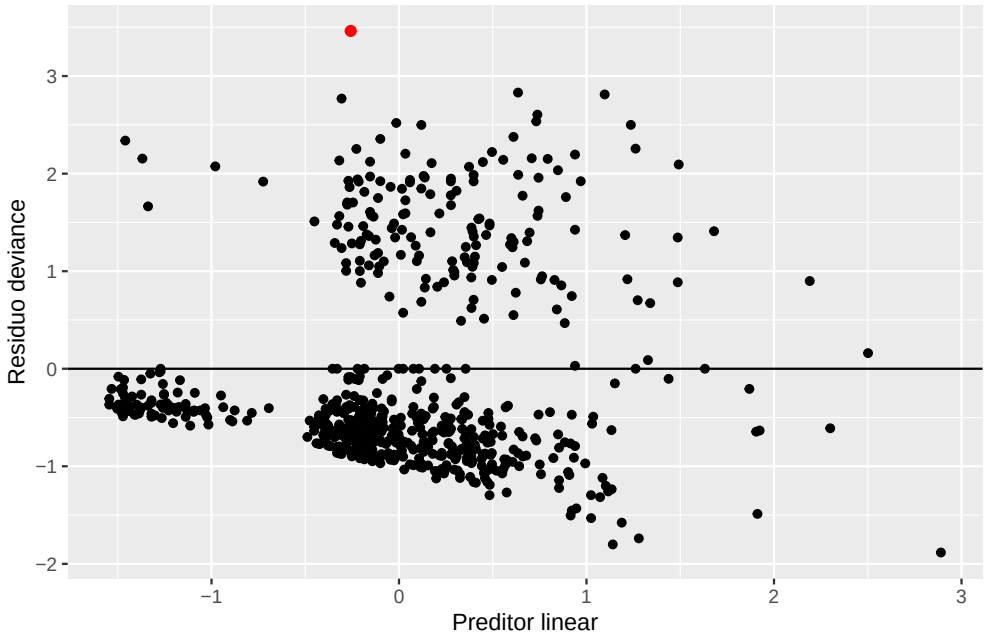
No gráfico 13 temos a função de sobrevivência para cada um dos perfis solicitados e assim como esperado e analisado no modelo o pior prefil é o 5 em que à paciente tem idade de 53 anos, sem terapia hormonal, na menopausa, tumor de 30 milímetros, com estadiamento 3 e 5 linfonodos comprometidos, onde temos um decaimento da probabilidade de sobrevivência muito rápida, sendo o pior caso, seguido do perfil 4 com dade de 53 anos, sem terapia hormonal, na menopausa, tumor de 30 milímetros, com estadiamento 2 e 5 linfonodos comprometidos, agora os 3 primeiros perfil são muito próximos e se enquadram em uma situação melhor com uma probabilidade de sobrevência maior e um decaimento mais lento, a mesmo resultado pode ser interpretado através do gráfico 14.

(e) Faça análise de resíduos do modelo final obtido em (b). Obtenha todos os resíduos discutidos em aula, bem como os testes de Schoenfeld para a proporcionalidade das taxas de falha.

Resolução

Avaliando os Resíduos Deviance, temos:

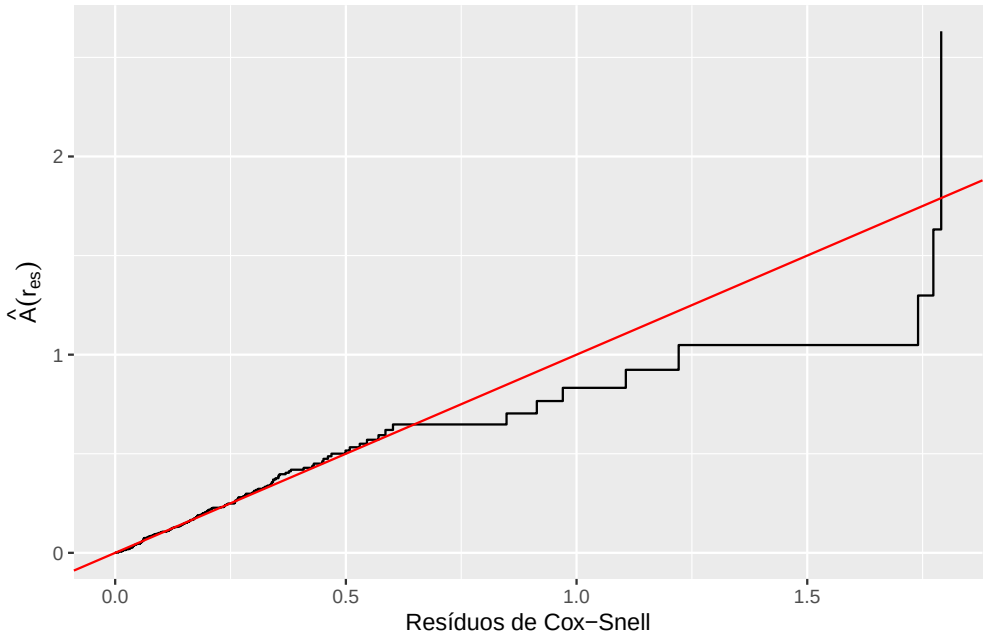
Gráfico 15: Resíduo Deviance x preditor linear



Em que podemos notar no gráfico 15 uma concentração elevada em torno do zero, o que pode indicar uma alta porporção de censuras e também se pode ver um ponto atípico com resíduo deviance maior que 3, que se trata da observação 56.

Avaliando os Resíduos de Cox-Snell, temos:

Gráfico 16:Resíduos de Cox−Snell



Em que podemos notar no gráfico 16 um bom ajuste, mesmo com uma pequena distoação com os resíduos de Cox-Snell à partir de 0.7, mas a maior parte das obervações acompanham a reta "y=x", o que indica um bom ajuste global, pois os resíduos de Cox-Snell se comportam como uma amostra censurada de uma variável com distribuição exponecial(1).

Avaliando os Resíduos Martingal, temos:

Gráfico 17: Resíduo Martingal

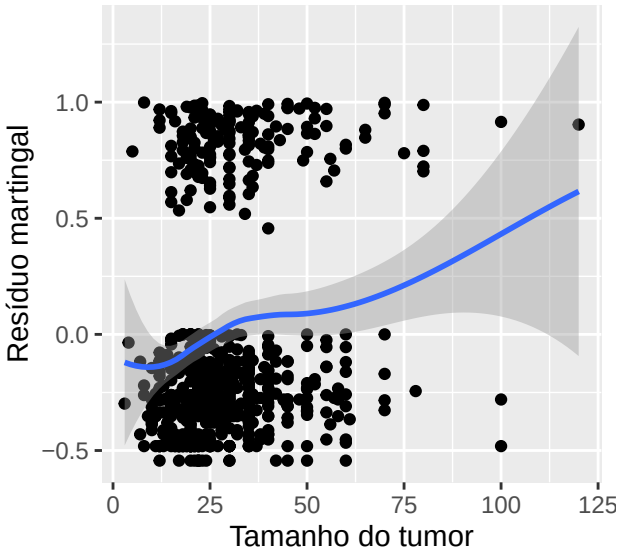
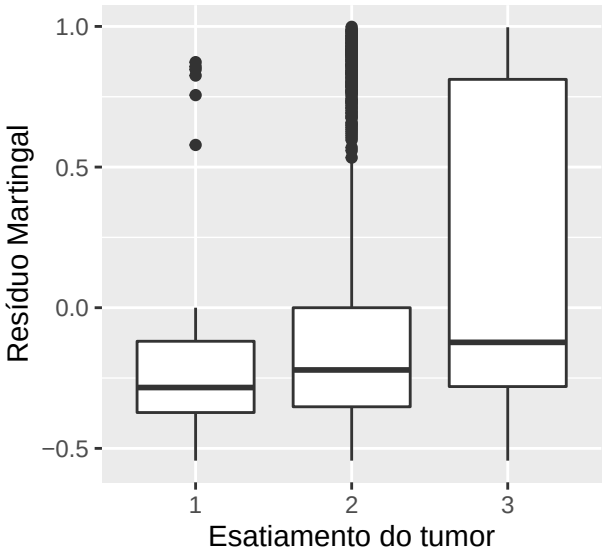


Gráfico 18: Resíduo Martingal x E



O objetivo do resíduo martingal é detectar a forma funcional das covariáveis ajustando um modelo sem elas e fazendo um plot dos resíduos martingal desse modelo contra os valores da covariável, assim podemos notar no gráfico 17 para a covariável Tamanho do tumor, apresenta uma tendência linear, indicando que estamos utilizando a forma funcional correta. Para a covariável Esatiamento do tumor (gráfico 18) temos que para cada categoria a um aumento na mediana dos resíduos o que tráz indícios que existe um comportameno linear em relação ao estadiamento do tumor, logo estamos utilizando a variável de forma correta.

Gráfico 19: Resíduo Martingal x n

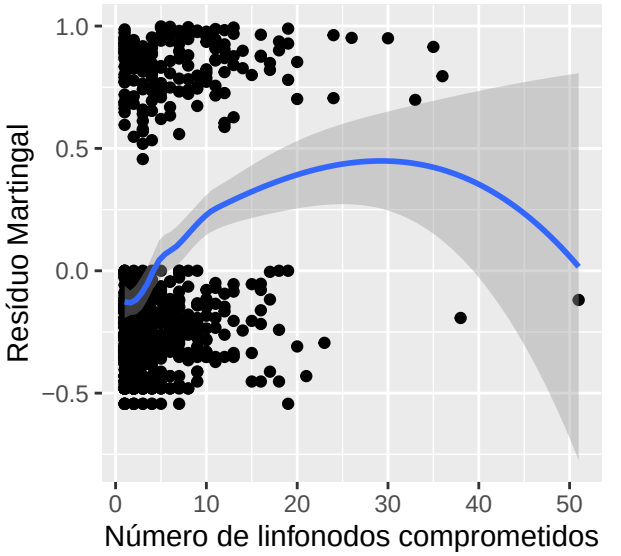
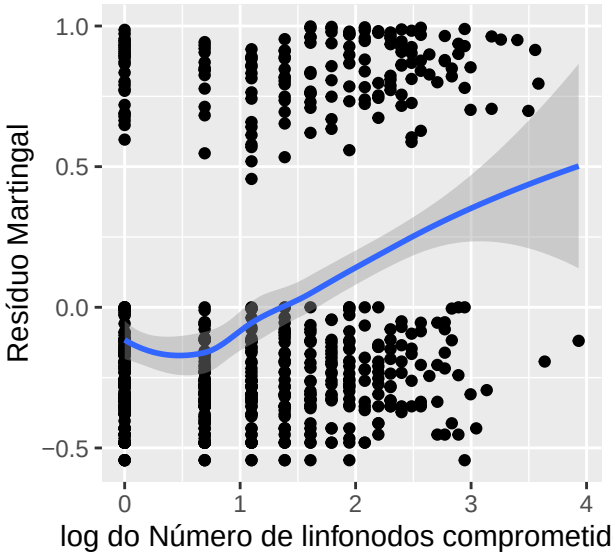


Gráfico 20: Resíduo Martingal x lo



Em que podemos notar no gráfico 19 que para Número de linfonodos comprometidos não temos uma tendência linear, neste caso se sugere uma transformação e ao avaliar a situação a transformação log seria mais adequada pois obsevando o gráfico 20 podemos encontrar a forma mais linear que é desejada.

Avaliando os Resíduos Escore, temos:

Gráfico 21:Resíduo Escore x índ

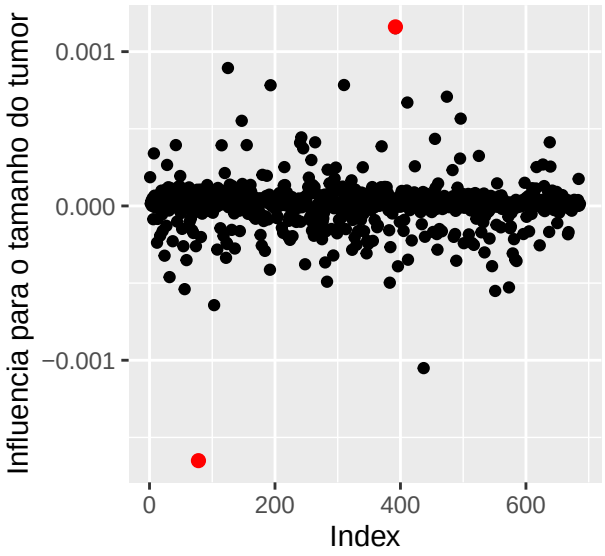
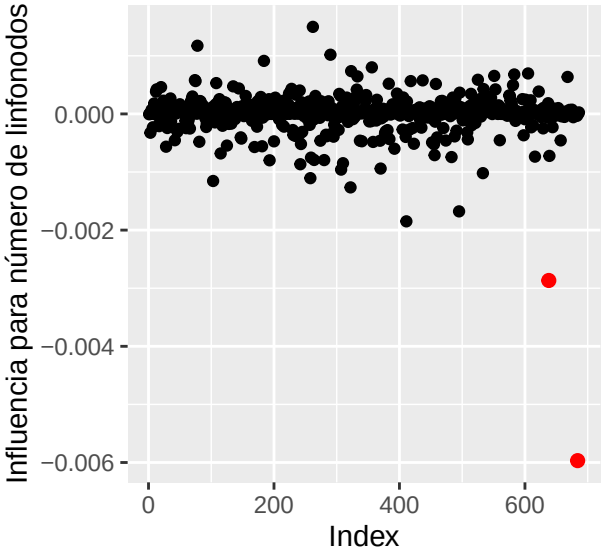


Gráfico 22: Resíduo Escore x ín



Com os resíduos escore o objetivo é avaliar se existem pontos influentes e observando os gráficos 21 e 22 temos algumas observações influentes para a variável tamanho do tumor temos 2 pontos influentes que correspondem as observações 78 e 392, mesmo aprentando uma diferença muito pequena em relação a zero esses pontos se destacam entre os demais, agora para a variável número de linfonodos comprometidos também encontramos dois pontos infuêntes que correspondem as observações 638 e 694.

Gráfico 23:Resíduo Escore x índ

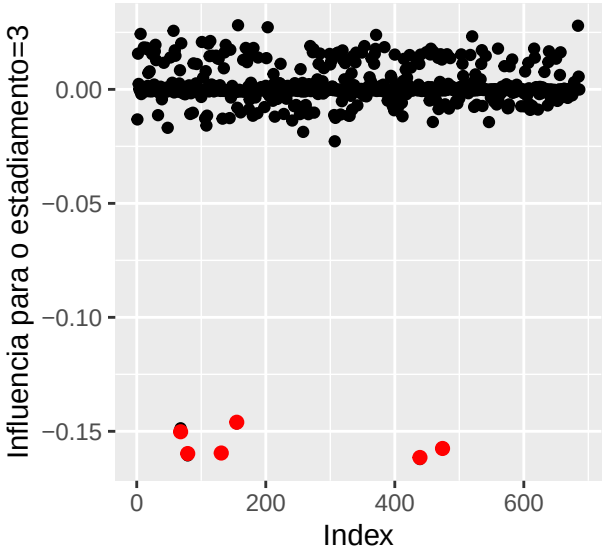
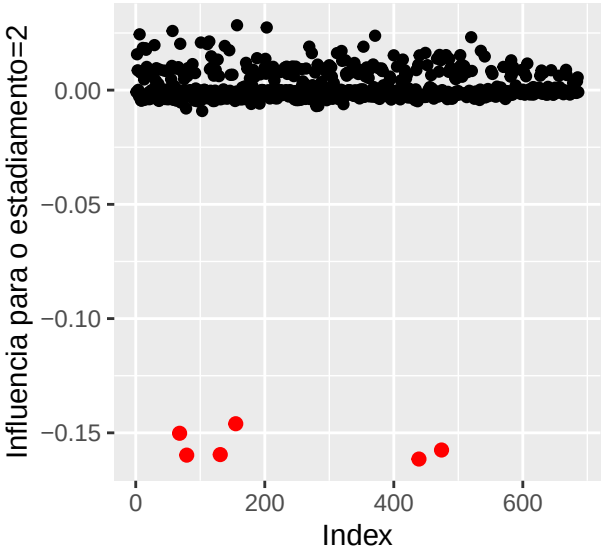
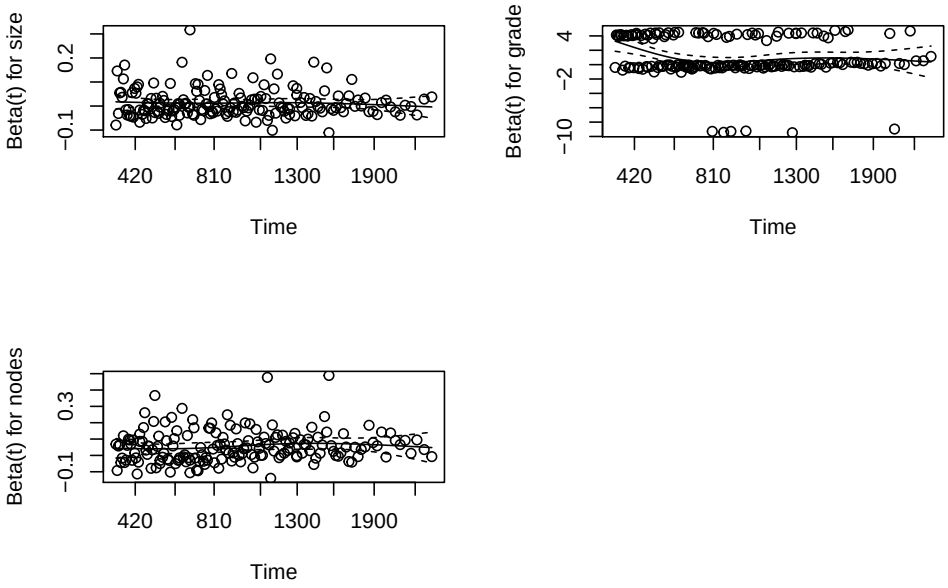


Gráfico 24:Resíduo Escore x índ



Análisando os resíduos escore para a variável estadiamento do tumor por ser uma variável categórica com 3 níveis, apresenta 2 gráficos de resíduos muito similares, assim pode-se notar que existem 6 pontos influêntes para esta covariável sendo as observações 68,79,131,155,439 e 474, assim como nos outros casos, o que se sugere ao encontrar pontos influêntes é ajustar o modelo sem os mesmos e verificar se obtém um ajuste com mais qualidade aos dados.

E por fim os resíduos de Schoenfeld:



Em que para todas as variáveis, podemos notar uma tendência linear ao longo do tempo, indicando que o a suposição de riscos proporcionais esta satisfeita, porém ao observar a variável estadiamento do tumor (grade) apresenta uma leve distoação, uma sugetão para caso a mesma não venha a satisfazer a suposição de riscos proporcionais é fazer um modelo estratificado para esta covariável. Agora fazendo o teste para verificar a proporcionalidade dos riscos, logo estabelecendo a hipótese, temos:

$$\begin{cases} H_0 : \beta_{(t)} = \beta \\ H_1 : \beta_{(t)} \neq \beta \end{cases}$$

E também Sob H_0, a estatística de teste tem distribuição de qui-quadrado, assim resultando na seguinte tabela:

	chisq	df	p
size	0.054	1	0.816
grade	4.067	2	0.131
nodes	0.878	1	0.349
GLOBAL	5.770	4	0.217

E com o resultado da tabela acima para todas as variáveis e também para o ajuste global, não rejeitamos H_0 com um nível de significância fixado de 5%, ou seja, a suposição de riscos proporcionais esta satisfeita, assim como os gráficos acima indicam.

Anexo

Códigos

```
# Prova 2 - MAE0514

# local de trabalho
setwd("~/Área de Trabalho/P2 - MAE 514")

# Pacotes
library(ggplot2)
library(survival)
library(gridExtra)
library(survminer)
library(KMsurv)
library(cmprsk)
library(Amelia)

# Leitura dos dados
data <- read.csv("Dados-eind3-prova2.csv",header = T)

# Removendo a primeira coluna (desnecessária)
data$id <- NULL

# transformando estadiamento em fator
data$grade <- as.factor(data$grade)

attach(data)

# item a
missmap(data,col = c('red','black'),main = "Gráfico 1: Heatmap missing")

# categorização das variáveis

# summary(age)
# summary(nodes)
# summary(size)
data$age_cat <- as.factor(findInterval(age,c(quantile(age)[2],quantile(age)[4])))
data$nodes_cat <- as.factor(findInterval(nodes,c(quantile(nodes)[3],quantile(nodes)[4])))
data$size_cat <- as.factor(findInterval(size,c(quantile(size)[2],quantile(size)[4])))

# Taxa de óbitos
ggplot(data, aes(x=factor(censdead),fill=factor(censdead)))+
  geom_bar(aes(y = (..count..)/sum(..count..))) +
  scale_y_continuous(labels=scales::percent,breaks = c(seq(0,0.75,0.05))) +
  theme(legend.position = "none") +
  scale_fill_brewer(palette="Set1") +
  labs(y = "Pacientes (%)",
       x="Óbito",
       title="Gráfico 2: Taxa de óbtos")

# Óbito vs. Estadiamento
ggplot(data, aes(x= grade,  group=factor(censdead))) +
```

```
  geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count") +
  geom_text(aes( label = scales::percent(..prop..),
                 y= ..prop.. ), stat= "count", vjust = -.5) +
  labs(y = "Pacientes (%)",
       x="Estadiamento",
       fill=c("Estadiamento"),
       title="Gráfico 3: Óbito vs. Estadiamento") +
  facet_grid(~factor(censdead)) +
  scale_fill_brewer(palette="Set1") +
  scale_y_continuous(labels = scales::percent, limits = c(0,0.79))

# Histograma idade  vs. óbito
ggplot(data, aes(x=age,  color=factor(censdead),fill=factor(censdead))) +
  geom_histogram(color="black", fill="white")+
  facet_grid(factor(censdead) ~ .) +
  labs(x="Idade (em anos)",
       y="# de pacientes",
       title="Gráfico 4: Histograma Idade vs. óbito")

# Histograma Número de linfonodos comprometidos vs. óbito
ggplot(data, aes(x=nodes,  color=factor(censdead),fill=factor(censdead))) +
  geom_histogram(color="black", fill="white")+
  facet_grid(factor(censdead) ~ .) +
  labs(x="Número de linfonodos comprometidos",
       y="# de pacientes",
       title="Gráfico 5: Número de linfonodos comprometidos vs. óbito")

# Histograma tamanho do tumor vs. óbito
ggplot(data, aes(x=size,  color=factor(censdead),fill=factor(censdead))) +
  geom_histogram(color="black", fill="white")+
  facet_grid(factor(censdead) ~ .) +
  labs(x="Tamanho do tumor (em mm)",
       y="# de pacientes",
       title="Gráfico 6: Histograma tamanho do tumor vs. óbito")

ekm_meno <- survfit(Surv(survtime, censdead)~ menopause,data = data)

# Grafico Kaplan-Meier Menopausa
ggsurvplot(ekm_meno, data = data,conf.int = T,palette="Set1",pval = T,risk.table = T,
           ggtheme=theme_gray(),
           legend.labs=c("Não","Sim")) +
  labs(x="Tempo (em dias)",
       y=expression(hat(S(t))),
       title = "Gráfico 7: Estimativas de Kaplan-Meier (Menopausa)")

knitr::kable(surv_pvalue(ekm_meno,data, method = c("1"))[,1:3],digits = 3)

ekm_horm <- survfit(Surv(survtime, censdead)~ hormone,data = data)

# Grafico Kaplan-Meier Hormonio
ggsurvplot(ekm_horm, data = data,conf.int = T,palette="Set1",pval = T,risk.table = T,
           ggtheme=theme_gray(),
           legend.labs=c("Não","Sim")) +
```

```
  labs(x="Tempo (em dias)",
       y=expression(hat(S(t))),
       title = "Gráfico 8: Estimativas de Kaplan-Meier (Terapia Hormonal)")

knitr::kable(surv_pvalue(ekm_horm,data, method = c("1"))[,1:3],digits = 3)

ekm_est <- survfit(Surv(survtime, censdead)~ grade,data = data)

# Grafico Kaplan-Meier Estadiamento
ggsurvplot(ekm_est, data = data,conf.int = T,palette="Set1",pval = T,risk.table = T,
           ggtheme=theme_gray(),
           legend.labs=c("1","2","3")) +
  labs(x="Tempo (em dias)",
       y=expression(hat(S(t))),
       title = "Gráfico 9: Estimativas de Kaplan-Meier (Estadiamento)")

knitr::kable(surv_pvalue(ekm_est,data, method = c("1"))[,1:3],digits = 3)

ekm_age_cat <- survfit(Surv(survtime, censdead)~age_cat,data = data)

# Grafico Kaplan-Meier Idade categorizada
ggsurvplot(ekm_age_cat, data = data,conf.int = T,palette="Set1",pval = T,risk.table = T,
           ggtheme=theme_gray(),
           legend.labs=c("< 46","de 46 a 61",">= 61")) +
  labs(x="Tempo (em dias)",
       y=expression(hat(S(t))),
       title = "Gráfico 10: Estimativas de Kaplan-Meier (Idade)")

knitr::kable(surv_pvalue(ekm_age_cat,data, method = c("1"))[,1:3],digits = 3)

ekm_nodes_cat <- survfit(Surv(survtime, censdead)~nodes_cat,data = data)

# Grafico Kaplan-Meier numero de Linfonodos
ggsurvplot(ekm_nodes_cat, data = data,conf.int = T,palette="Set1",pval = T,risk.table = T,
           ggtheme=theme_gray(),
           legend.labs=c("< 3","de 3 a 7",">= 7")) +
  labs(x="Tempo (em dias)",
       y=expression(hat(S(t))),
       title = "Gráfico 11: Estimativas de Kaplan-Meier (Linfonodos)")

knitr::kable(surv_pvalue(ekm_nodes_cat,data, method = c("1"))[,1:3],digits = 3)

ekm_size_cat <- survfit(Surv(survtime, censdead)~size_cat,data = data)

# Grafico Kaplan-Meier tamanho tumor
ggsurvplot(ekm_size_cat, data = data,conf.int = T,palette="Set1",pval = T,risk.table = T,
           ggtheme=theme_gray(),
           legend.labs=c("< 20","de 20 a 35",">= 35")) +
  labs(x="Tempo (em dias)",
       y=expression(hat(S(t))),
       title = "Gráfico 12: Estimativas de Kaplan-Meier (Tamanho do tumor)")

knitr::kable(surv_pvalue(ekm_size_cat,data, method = c("1"))[,1:3],digits = 4)
```

```r
# item b

# removendo as variáveis categorizadas
data$age_cat <- NULL
data$nodes_cat <- NULL
data$size_cat <- NULL

# modeo inicial
mod.ini <- coxph(Surv(survtime, censdead)~.,ties="breslow",data = data)
summary(mod.ini)

knitr::kable(drop1(mod.ini, test="Chisq"),digits = 3)

#mod1 <- coxph(Surv(survtime, censdead)~menopause+hormone+size+grade+nodes,data=data)
#drop1(mod1, test="Chisq")
#mod2 <- coxph(Surv(survtime, censdead)~hormone+size+grade+nodes,data=data)
#drop1(mod2, test="Chisq")
#mod3 <- coxph(Surv(survtime, censdead)~size+grade+nodes,data=data)
#drop1(mod3, test="Chisq")

mod.fim <- coxph(Surv(survtime, censdead)~size + grade + nodes,ties="breslow",
                 data = data)
summary(mod.fim)

# item d

# Nelson-Aalen
H_0 <- basehaz(mod.fim,centered=F)
S_0 <- exp(-H_0$hazard)

# indivíduo 1
ind1 <- c(53,1,0,15,0,0,1)

# perfil 1
p1 <- S_0^(exp(ind1[1]*mod.ini$coefficients[1]+ind1[2]*mod.ini$coefficients[2]+
              ind1[3]*mod.ini$coefficients[3]+ind1[4]*mod.ini$coefficients[4]+
              ind1[5]*mod.ini$coefficients[5]+ind1[6]*mod.ini$coefficients[6]+
              ind1[7]*mod.ini$coefficients[7]))

# indivíduo 2
ind2 <- c(53,1,0,30,0,0,1)

# perfil 2
p2 <- S_0^(exp(ind2[1]*mod.ini$coefficients[1]+ind2[2]*mod.ini$coefficients[2]+
              ind2[3]*mod.ini$coefficients[3]+ind2[4]*mod.ini$coefficients[4]+
              ind2[5]*mod.ini$coefficients[5]+ind2[6]*mod.ini$coefficients[6]+
              ind2[7]*mod.ini$coefficients[7]))

# indivíduo 3
ind3 <- c(53,1,0,30,0,0,5)

# perfil 3
p3 <- S_0^(exp(ind3[1]*mod.ini$coefficients[1]+ind3[2]*mod.ini$coefficients[2]+
```

```
              ind3[3]*mod.ini$coefficients[3]+ind3[4]*mod.ini$coefficients[4]+
              ind3[5]*mod.ini$coefficients[5]+ind3[6]*mod.ini$coefficients[6]+
              ind3[7]*mod.ini$coefficients[7]))

# indivíduo 4
ind4 <- c(53,1,0,30,1,0,5)

# perfil 4
p4 <- S_0^(exp(ind4[1]*mod.ini$coefficients[1]+ind4[2]*mod.ini$coefficients[2]+
              ind4[3]*mod.ini$coefficients[3]+ind4[4]*mod.ini$coefficients[4]+
              ind4[5]*mod.ini$coefficients[5]+ind4[6]*mod.ini$coefficients[6]+
              ind4[7]*mod.ini$coefficients[7]))

# indivíduo 5
ind5 <- c(53,1,0,30,0,1,5)

# perfil 5
p5 <- S_0^(exp(ind5[1]*mod.ini$coefficients[1]+ind5[2]*mod.ini$coefficients[2]+
              ind5[3]*mod.ini$coefficients[3]+ind5[4]*mod.ini$coefficients[4]+
              ind5[5]*mod.ini$coefficients[5]+ind5[6]*mod.ini$coefficients[6]+
              ind5[7]*mod.ini$coefficients[7]))

pf_time <- data.frame(time=H_0$time,p1,p2,p3,p4,p5)

col <- c('Perfil 1' ='red','Perfil 2' ='black','Perfil 3' ='blue',
         'Perfil 4' ='green','Perfil 5' ='brown')

pt1 <- ggplot(pf_time,aes(x=time)) +
  geom_step(aes(y=p1,colour="Perfil 1")) +
  geom_step(aes(y=p2,colour="Perfil 2")) +
  geom_step(aes(y=p3,colour="Perfil 3")) +
  geom_step(aes(y=p4,colour="Perfil 4")) +
  geom_step(aes(y=p5,colour="Perfil 5")) +
  scale_colour_manual(name="Perfil",values=col) +
  labs(x="Tempo (dias)",
       y=expression(hat(S(t))),
       title="Gráfico 13: Função de sobrevivência")

# para função de taxa de falha acumulda

p1 <- S_0*(exp(ind1[1]*mod.ini$coefficients[1]+ind1[2]*mod.ini$coefficients[2]+
              ind1[3]*mod.ini$coefficients[3]+ind1[4]*mod.ini$coefficients[4]+
              ind1[5]*mod.ini$coefficients[5]+ind1[6]*mod.ini$coefficients[6]+
              ind1[7]*mod.ini$coefficients[7]))


p2 <- S_0*(exp(ind2[1]*mod.ini$coefficients[1]+ind2[2]*mod.ini$coefficients[2]+
              ind2[3]*mod.ini$coefficients[3]+ind2[4]*mod.ini$coefficients[4]+
              ind2[5]*mod.ini$coefficients[5]+ind2[6]*mod.ini$coefficients[6]+
              ind2[7]*mod.ini$coefficients[7]))


p3 <- S_0*(exp(ind3[1]*mod.ini$coefficients[1]+ind3[2]*mod.ini$coefficients[2]+
```

```
              ind3[3]*mod.ini$coefficients[3]+ind3[4]*mod.ini$coefficients[4]+
              ind3[5]*mod.ini$coefficients[5]+ind3[6]*mod.ini$coefficients[6]+
              ind3[7]*mod.ini$coefficients[7]))


p4 <- S_0*(exp(ind4[1]*mod.ini$coefficients[1]+ind4[2]*mod.ini$coefficients[2]+
              ind4[3]*mod.ini$coefficients[3]+ind4[4]*mod.ini$coefficients[4]+
              ind4[5]*mod.ini$coefficients[5]+ind4[6]*mod.ini$coefficients[6]+
              ind4[7]*mod.ini$coefficients[7]))


p5 <- S_0*(exp(ind5[1]*mod.ini$coefficients[1]+ind5[2]*mod.ini$coefficients[2]+
              ind5[3]*mod.ini$coefficients[3]+ind5[4]*mod.ini$coefficients[4]+
              ind5[5]*mod.ini$coefficients[5]+ind5[6]*mod.ini$coefficients[6]+
              ind5[7]*mod.ini$coefficients[7]))

pf_time2 <- data.frame(time=H_0$time,cum1=cumsum(p1),
                       cum2=cumsum(p2),cum3=cumsum(p3),
                       cum4=cumsum(p4),cum5=cumsum(p5))

pt2 <- ggplot(pf_time2,aes(x=time)) +
  geom_step(aes(y=cum1,colour="Perfil 1")) +
  geom_step(aes(y=cum2,colour="Perfil 2")) +
  geom_step(aes(y=cum3,colour="Perfil 3")) +
  geom_step(aes(y=cum4,colour="Perfil 4")) +
  geom_step(aes(y=cum5,colour="Perfil 5")) +
  scale_colour_manual(name="Perfil",values=col) +
  labs(x="Tempo (dias)",
       y=expression(hat(A(t))),
       title="Gráfico 14: Função de taxa de falha acumulada")

grid.arrange(pt1, pt2, ncol=2,nrow=1)

# item e

# Resíduo deviance
dev <- data.frame(pred_lin=mod.fim$linear.predictors,
                  res_dev=resid(mod.fim,type='deviance'))

ggplot(dev, aes(x=pred_lin,y=res_dev))+
         geom_point() +
         scale_y_continuous(breaks = seq(-2,3,1)) +
         annotate(geom = "point", x = dev[56,1], y = dev[56,2], colour='red',
                  size = 2) +
         geom_abline(slope = 0) +
         labs(x="Preditor linear",
              y="Residuo deviance",
              title = "Gráfico 15: Resíduo Deviance x preditor linear")

# Resíduos de cox-snell
data$resid_mart <- residuals(mod.fim, type = "martingale")
data$resid_coxsnell <- -(data$resid_mart - data$censdead)
```

```r
# Modelo com os Resíduos de cox-snell
fit_coxsnell <- coxph(formula = Surv(resid_coxsnell, censdead) ~ 1,data = data)

# Nelson-Aalen
df_base_haz <- basehaz(fit_coxsnell, centered = FALSE)

ggplot(data = df_base_haz, mapping = aes(x = time, y = hazard)) +
  geom_step() +
  geom_abline(intercept = 0, slope = 1, color = "red") +
  labs(x = "Resíduos de Cox-Snell",
       y = expression(hat(A)(r[es])),
       title = "Gráfico 16: Resíduos de Cox-Snell ")

# Resíduo martingal
mod.0 <- coxph(Surv(survtime, censdead)~ 1)
data$resid_mart <- resid(mod.0)

# martingal tamanho do tumor
plot1 <- ggplot(data = data, mapping = aes(x = size, y = resid_mart)) +
  geom_point() +
  geom_smooth() +
  labs(x = "Tamanho do tumor",
       y = "Resíduo martingal",
       title = "Gráfico 17: Resíduo Martingal")

# martingal do estadiamento
plot2 <- ggplot(data = data, mapping = aes(x = grade, y = resid_mart)) +
  geom_boxplot() +
  labs(x = "Esatiamento do tumor",
       y = "Resíduo Martingal",
       title = "Gráfico 18: Resíduo Martingal x Esatiamento do tumor")

# martingal numero de linfonodos
plot3 <- ggplot(data = data, mapping = aes(x = nodes, y = resid_mart)) +
  geom_point() +
  geom_smooth() +
  labs(x = "Número de linfonodos comprometidos",
       y = "Resíduo Martingal",
       title = "Gráfico 19: Resíduo Martingal x número de linfonodos")

# martingal numero de linfonodos trsnformado em log
plot4 <- ggplot(data = data, mapping = aes(x = log(nodes), y = resid_mart)) +
  geom_point() +
  geom_smooth() +
  labs(x = "log do Número de linfonodos comprometidos",
       y = "Resíduo Martingal",
       title = "Gráfico 20: Resíduo Martingal x log do número de linfonodos")

grid.arrange(plot1, plot2, ncol=2,nrow=1)
grid.arrange(plot3, plot4, ncol=2,nrow=1)

# Resíduo escore
res_escore <- data.frame(resid(mod.fim,type='dfbeta'),
```

```
                        index=seq(1,dim(data)[1]))

p1 <- ggplot(data = res_escore, mapping = aes(x = index, y = ind1)) +
  geom_point() +
  annotate(geom = "point", x = c(392,78),
           y = c(1.160555e-03,-0.0016507054), colour='red',size = 2) +
  labs(x = "Index",
       y = "Influencia para o tamanho do tumor",
       title = "Gráfico 21:Resíduo Escore x índices")

p2 <- ggplot(data = res_escore, mapping = aes(x = index, y = ind4)) +
  geom_point() +
  annotate(geom = "point", x = c(684,638),
           y = c(-0.0059678128,-0.0028668434), colour='red',size = 2) +
  labs(x = "Index",
       y = "Influencia para número de linfonodos",
       title = "Gráfico 22: Resíduo Escore x índices")

p3 <- ggplot(data = res_escore, mapping = aes(x = index, y = ind3)) +
  geom_point() +
  annotate(geom = "point", x = c(439,79,131,474,68,155),
           y = c(-0.161467776,-0.159754032,-0.159516346,
                 -0.157515569,-0.150190617,-0.146015240),
           colour='red',size = 2) +
  labs(x = "Index",
       y = "Influencia para o estadiamento=3",
       title = "Gráfico 23: Resíduo Escore x índices")

p4 <- ggplot(data = res_escore, mapping = aes(x = index, y = ind2)) +
  geom_point() +
  annotate(geom = "point", x = c(439,79,131,474,68,155),
           y = c(-0.161467776,-0.159754032,-0.159516346,
                 -0.157515569,-0.150190617,-0.146015240),
           colour='red',size = 2) +
  labs(x = "Index",
       y = "Influencia para o estadiamento=2",
       title = "Gráfico 24:Resíduo Escore x índices")

grid.arrange(p1, p2, ncol=2,nrow=1)

grid.arrange(p3, p4, ncol=2,nrow=1)

# Resíduo de Schoenfeld
res_scho <- cox.zph(mod.fim)

# gráfico dos resíduos
plot(res_scho)

# teste de hipotese rp
knitr::kable(res_scho$table,digits=3)
```