

Lista 5 - MAE0514

Guilherme N^oUSP: 8943160 e Leonardo N^oUSP: 9793436

Exercício 1

Considere o conjunto de dados de pacientes com leucemia aguda que receberam transplante de medula óssea, apresentado na seção 1.3 de Klein e Moeschberger (2003). Os pacientes que recebem transplante de doador compatível (allogênico) podem desenvolver uma doença conhecida como DECH (doença do enxerto contra o hospedeiro) ou GVHD (graft-versus-host disease), que pode ser muito grave. No entanto, suspeita-se que DECH tenha um efeito anti-leucêmico nos pacientes. Para verificar essa hipótese, deseja-se ajustar um modelo de Cox aos dados, considerando-se como variável resposta o tempo até a recorrência da doença (relapse). Pacientes que apresentaram óbito antes da recorrência são considerados como observações censuradas. Os dados que devem ser utilizados estão disponíveis no arquivo BMT-Data.csv e a descrição das variáveis está no arquivo BMT-Data.des.

- (a) Defina duas variáveis dependentes do tempo indicadoras da ocorrência de DECH aguda e crônica. Escreva o modelo semiparamétrico de Cox incluindo as variáveis idade do paciente, sexo do paciente, tempo até o transplante, classificação morfológica Franco-AmericanaBritânica (FAB) e tratamento profilático para DECH (MTX), além das variáveis dependentes do tempo definidas.

Resolução

Inicialmente definindo as duas variáveis dependentes do tempo indicadoras da ocorrência de DECH aguda e crônica que no banco de dados são “A” e “C” respectivamente, sendo ambas covariáveis do tipo internas, ou seja, são medidas no indivíduo quando o paciente estava em observação, em relação o tempo até a recorrência da doença.

O modelo semi-paramétrico de cox pode ser escrito como:

$$\alpha(t|x(t)) = \alpha_0(t)e^{x'(t)\beta}$$

Em que $\alpha(t|x(t))$ é a função de taxa de falha a ser modelada, a matriz $x(t)$ é conhecida para todo instante no qual a observação “i” está sendo acompanhada (ao longo do tempo), $\alpha_0(t)$ é a função de taxa de falha de referência e por fim β é o vetor de parâmetros do modelo.

- (b) Ajuste um modelo de Cox com as variáveis do item (a). Selecione as variáveis significativas e faça a verificação da proporcionalidade dos riscos. O que pode ser concluído?

Resolução

Fazendo o ajuste do banco de dados para o formato “longo” e ajustando o modelo com as covariáveis idade do paciente (Z1), sexo do paciente (Z3), tempo até o transplante (Z7), classificação morfológica Franco-Americana Britânica (Z8) e tratamento profilático para DECH (Z10) e as covariáveis indicadoras da ocorrência de DECH aguda (A) e crônica (C) e com isso temos:

```
## Call:
## coxph(formula = Surv(Tinicio, Tfim, Delta2) ~ A + C + Z1 + Z3 +
##       Z7 + Z8 + Z10, data = BMT_DataRELAPSED)
```

```
##
## n= 223, number of events= 43
##
##      coef exp(coef) se(coef)      z Pr(>|z|)
## A   -0.4753906 0.6216422 0.4856246 -0.979 0.32762
## C    0.2505825 1.2847736 0.3869404 0.648 0.51724
## Z1   0.0004414 1.0004415 0.0180202 0.024 0.98046
## Z3  -0.2633559 0.7684684 0.3172550 -0.830 0.40648
## Z7  -0.0003670 0.9996330 0.0006393 -0.574 0.56587
## Z8   0.9989454 2.7154166 0.3131941 3.190 0.00142 **
## Z10  0.4787014 1.6139772 0.3460418 1.383 0.16655
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## A      0.6216    1.6086    0.2400    1.610
## C      1.2848    0.7783    0.6018    2.743
## Z1      1.0004    0.9996    0.9657    1.036
## Z3      0.7685    1.3013    0.4126    1.431
## Z7      0.9996    1.0004    0.9984    1.001
## Z8      2.7154    0.3683    1.4698    5.017
## Z10     1.6140    0.6196    0.8191    3.180
##
## Concordance= 0.689 (se = 0.037 )
## Likelihood ratio test= 14.94 on 7 df,  p=0.04
## Wald test              = 14.81 on 7 df,  p=0.04
## Score (logrank) test = 15.95 on 7 df,  p=0.03
```

Agora fazendo a verificação das variáveis mais significativas, temos:

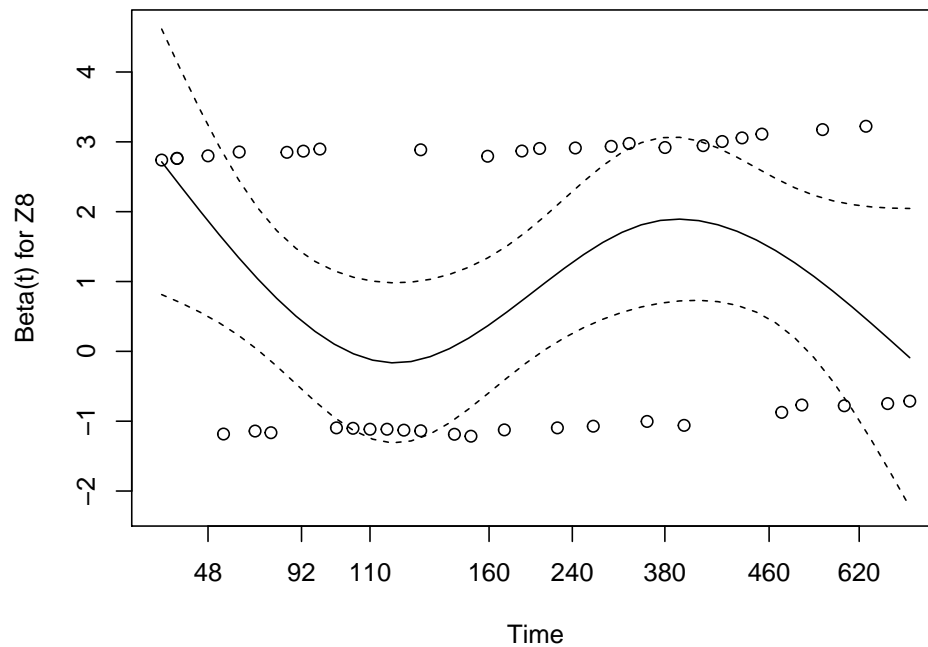
	Df	AIC	LRT	Pr(>Chi)
	NA	389.034	NA	NA
A	1	388.103	1.070	0.301
C	1	387.450	0.416	0.519
Z1	1	387.034	0.001	0.980
Z3	1	387.721	0.687	0.407
Z7	1	387.407	0.373	0.541
Z8	1	396.955	9.921	0.002
Z10	1	388.858	1.824	0.177

Em que apenas a variável “Z8” que corresponde a classificação morfológica Franco-Americana Britânica (FAB) é significativa, assim ajustando o modelo novamente obtemos o seguinte resultado:

```
## Call:
## coxph(formula = Surv(Tinicio, Tfim, Delta2) ~ Z8, data = BMT_DataRELAPSED)
##
## n= 223, number of events= 43
##
##      coef exp(coef) se(coef)      z Pr(>|z|)
## Z8 0.9865    2.6818    0.3059 3.225 0.00126 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##      exp(coef) exp(-coef) lower .95 upper .95
## Z8      2.682      0.3729      1.473      4.884
##
## Concordance= 0.618 (se = 0.038 )
## Likelihood ratio test= 10.08 on 1 df,  p=0.001
## Wald test              = 10.4 on 1 df,  p=0.001
## Score (logrank) test = 11.26 on 1 df,  p=8e-04
```

Resíduos de Schoenfeld:



Em que se pode notar que a variável tratamento apresenta um comportamento próximo de linear, não sendo possível traçar uma reta ao longo do tempo.

Fazendo a tabela com o resultado do teste, em que:

$$\begin{cases} H_0 : O \text{ modelo e de riscos proporcionais} \\ H_1 : O \text{ modelo nao e de riscos proporcionais} \end{cases}$$

	chisq	df	p
Z8	0.009	1	0.925
GLOBAL	0.009	1	0.925

Queremos testar se o modelo de riscos proporcionais se adequa a esse caso, e ao observar a tabela acima assim como resultado do gráfico acima a variável tratamento não se rejeita a hipótese nula com um p-valor de

0.925 considerando um nível de significância de 5%, ou seja, ela se adequa no modelo de riscos proporcionais.

E com isso podemos concluir que as funções de sobrevivência para indivíduos com a variável Z8 igual a 0 e igual a 1 são razoavelmente paralelas ao longo de todo tempo. Ainda é possível dizer que indivíduos que assumem o valor 1 para a variável Z8 possuem 2.682 vezes de ter recorrência da doença (relapse) em relação a indivíduos que assumem o valor 0 para a variável Z8.

(c) Repita os itens (a) e (b) utilizando o tempo até o óbito como variável reposta.

Resolução

- a) Definindo as duas variáveis dependentes do tempo indicadoras da ocorrência de DECH aguda e crônica que no banco de dados são “A” e “C” respectivamente, sendo ambas covariáveis do tipo externa, ou seja, podem influenciar no risco de falha, mas não dependem do processo de falha, em relação ao tempo até o óbito

O modelo semi-paramétrico de cox pode ser escrito como:

$$\alpha(t|x(t)) = \alpha_0(t)e^{x'(t)\beta}$$

Em que $\alpha(t|x(t))$ é a função de taxa de falha a ser modelada, a matriz $x(t)$ é conhecida para todo instante no qual a observação “i” está sendo acompanhada (ao longo do tempo), $\alpha_0(t)$ é a função de taxa de falha de referência e por fim β é o vetor de parâmetros do modelo.

- b) Fazendo o ajuste do banco de dados para o formato “longo” e ajustando o modelo com as covariáveis idade do paciente (Z1), sexo do paciente (Z3), tempo até o transplante (Z7), classificação morfológica Franco-Americana Britânica (Z8) e tratamento profilático para DECH (Z10) e as covariáveis indicadoras da ocorrência de DECH aguda (A) e crônica (C) e com isso temos:

```
## Call:
## coxph(formula = Surv(Tinicio, Tfim, Delta1) ~ . - Id, data = BMT_Data0BITO)
##
##   n= 223, number of events= 91
##
##           coef exp(coef)    se(coef)      z Pr(>|z|)
## A      0.2794764  1.3224373  0.2902907  0.963   0.3357
## C     -0.0704853  0.9319415  0.2716315 -0.259   0.7953
## Z1     0.0176646  1.0178216  0.0115191  1.534   0.1252
## Z3    -0.1298272  0.8782471  0.2214194 -0.586   0.5576
## Z7     0.0003849  1.0003850  0.0002860  1.346   0.1783
## Z8     0.5587296  1.7484499  0.2204795  2.534   0.0113 *
## Z10    0.3160973  1.3717638  0.2410219  1.311   0.1897
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## A           1.3224      0.7562    0.7487    2.336
## C           0.9319      1.0730    0.5472    1.587
## Z1           1.0178      0.9825    0.9951    1.041
## Z3           0.8782      1.1386    0.5690    1.355
## Z7           1.0004      0.9996    0.9998    1.001
## Z8           1.7484      0.5719    1.1350    2.694
```

```
## Z10    1.3718    0.7290    0.8553    2.200
##
## Concordance= 0.621 (se = 0.03 )
## Likelihood ratio test= 12.65 on 7 df, p=0.08
## Wald test            = 12.73 on 7 df, p=0.08
## Score (logrank) test = 13.01 on 7 df, p=0.07
```

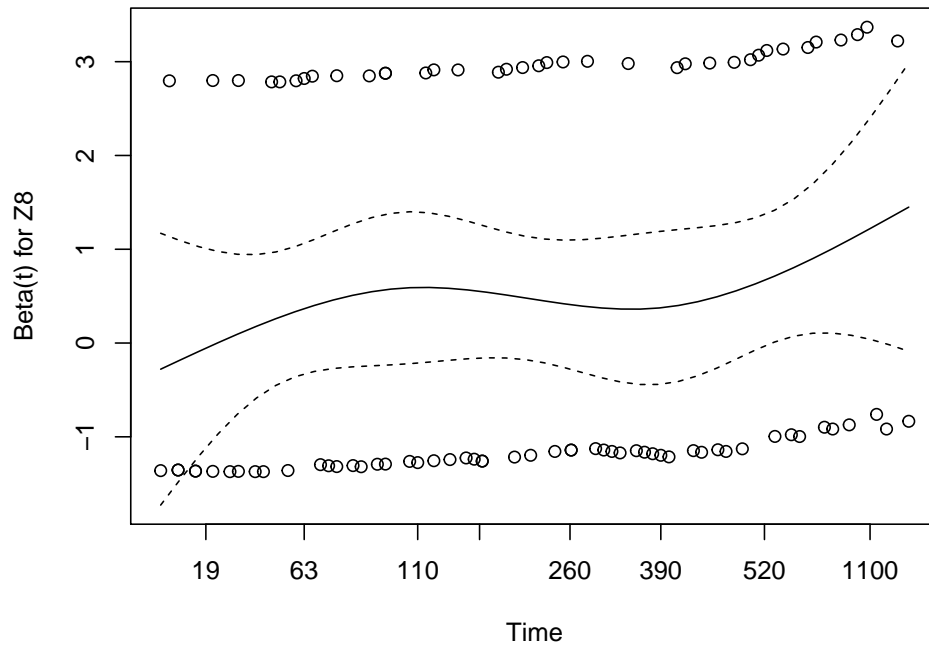
Agora fazendo a verificação das variáveis mais significativas, temos:

	Df	AIC	LRT	Pr(>Chi)
	NA	827.649	NA	NA
A	1	826.531	0.882	0.348
C	1	825.717	0.067	0.795
Z1	1	827.967	2.318	0.128
Z3	1	825.992	0.343	0.558
Z7	1	827.243	1.593	0.207
Z8	1	831.832	6.183	0.013
Z10	1	827.310	1.661	0.197

Assim como no modelo do item b apenas a variável “Z8” que corresponde a classificação morfológica Franco-Americana Britânica (FAB) é significativa, assim ajustando o modelo novamente obtemos o seguinte resultado:

```
## Call:
## coxph(formula = Surv(Tinicio, Tfim, Delta1) ~ Z8, data = BMT_Data0BIT0)
##
## n= 223, number of events= 91
##
##      coef exp(coef) se(coef)      z Pr(>|z|)
## Z8 0.5004    1.6493   0.2140 2.338   0.0194 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## Z8      1.649      0.6063    1.084    2.509
##
## Concordance= 0.55 (se = 0.026 )
## Likelihood ratio test= 5.25 on 1 df, p=0.02
## Wald test            = 5.47 on 1 df, p=0.02
## Score (logrank) test = 5.58 on 1 df, p=0.02
```

Resíduos de Schoenfeld:



Em que se pode notar que a variável tratamento apresenta um comportamento próximo de linear, não sendo possível traçar uma reta ao longo do tempo.

Fazendo a tabela com o resultado do teste, em que:

$$\begin{cases} H_0 : O \text{ modelo e de riscos proporcionais} \\ H_1 : O \text{ modelo nao e de riscos proporcionais} \end{cases}$$

	chisq	df	p
Z8	1.591	1	0.207
GLOBAL	1.591	1	0.207

Queremos testar se o modelo de riscos proporcionais se adequa a esse caso, e ao observar a tabela acima assim como resultado do gráfico acima a variável tratamento não se rejeita a hipótese nula com um p-valor de 0.207 considerando um nível de significância de 5%, ou seja, ela se adequa no modelo de riscos proporcionais.

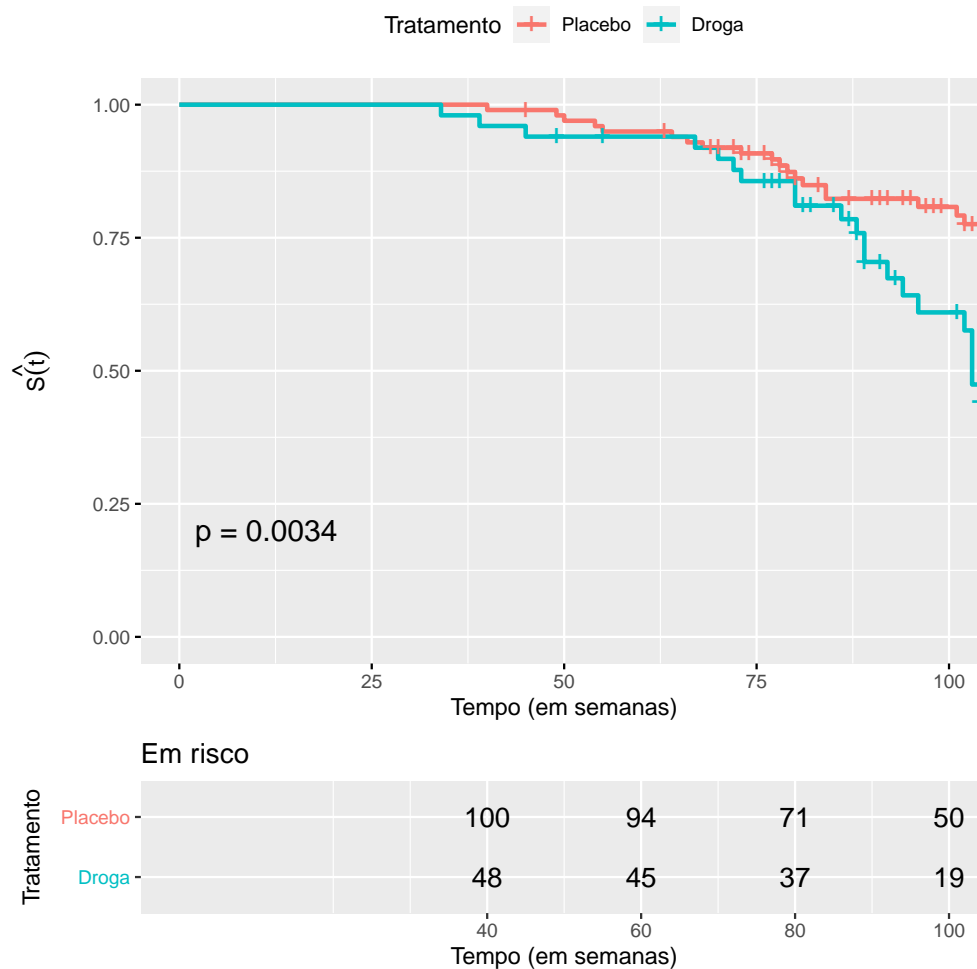
E com isso podemos concluir que as funções de sobrevivência para indivíduos com a variável Z8 igual a 0 e igual a 1 são razoavelmente paralelas ao longo de todo tempo. Ainda é possível dizer que indivíduos que assumem o valor 1 para a variável Z8 possuem 1.649 vezes de ir a óbito em relação a indivíduos que assumem o valor 0 para a variável Z8.

Exercício 2

Um estudo foi conduzido para estudar se uma determinada droga era cancerígena ou não. Para isso, foram selecionadas 50 ninhadas de ratos e, de cada ninhada, foram selecionados 3 ratos. Dos ratos de cada ninhada, selecionou-se aleatoriamente um deles para receber a droga e os demais receberam placebo. Observou-se o tempo, em semanas, até o desenvolvimento de um tumor. Os dados estão apresentados no arquivo litter-data.csv.

- (a) Obtenha as curvas de Kaplan-Meier para os dois grupos e comente.

Resolução



Em que podemos notar que as curvas são muito próximas com alguns cruzamentos até o dia 88, depois disso há um distanciamento entre as curvas a partir de então para os ratos que receberam a droga a curva decresce mais rapidamente. Entretanto, nota-se muitas censuras para os ratos que receberam placebo.

- (b) Ajuste um modelo de riscos proporcionais de Cox, ignorando o fato de existirem ratos pertencentes a uma mesma ninhada. Obtenha as estimativas dos parâmetros e interprete os resultados. Obtenha os resíduos de Schoenfeld para testar a proporcionalidade dos riscos. Faça um teste estatístico para a proporcionalidade dos riscos, utilizando a transformação KM, ou seja, a curva de sobrevivência de Kaplan-Meier (versão contínua à esquerda).

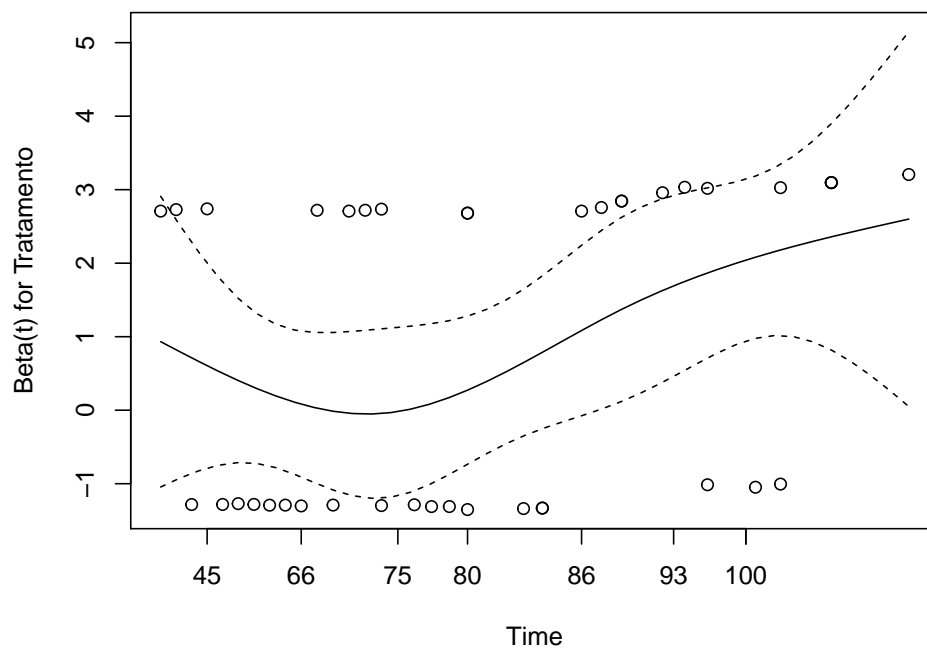
Resolução

Ajustando o modelo de riscos proporcionais de Cox, ignorando o fato de existirem ratos pertencentes a uma mesma ninhada, temos:

```
## Call:
## coxph(formula = Surv(Tempo, Delta) ~ Tratamento, data = litter_Data)
##
##      n= 150, number of events= 40
##
##              coef exp(coef) se(coef)      z Pr(>|z|)
## Tratamento 0.9047    2.4711   0.3175 2.849  0.00438 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## Tratamento    2.471    0.4047    1.326    4.604
##
## Concordance= 0.586 (se = 0.041 )
## Likelihood ratio test= 7.97 on 1 df,  p=0.005
## Wald test               = 8.12 on 1 df,  p=0.004
## Score (logrank) test = 8.68 on 1 df,  p=0.003
```

Como temos um único parâmetro β a partir do modelo ajustado, pode-se dizer que a taxa de ocorrência de tumores em ratos que recebem a droga é 2.4 vezes a taxa associada a ratos que recebem placebo.

Resíduos de Schoenfeld com a transformação KM:



Em que se pode notar que a variável tratamento não apresenta um comportamento linear, não sendo possível traçar uma reta ao longo do tempo.

Fazendo a tabela com o resultado do teste, em que:

$$\begin{cases} H_0 : O \text{ modelo e de riscos proporcionais} \\ H_1 : O \text{ modelo nao e de riscos proporcionais} \end{cases}$$

	chisq	df	p
Tratamento	5.02	1	0.025
GLOBAL	5.02	1	0.025

Queremos testar se o modelo de riscos proporcionais se adequa a esse caso, e ao observar a tabela acima assim como resultado do gráfico acima a variável tratamento rejeita a hipótese nula com um p-valor de 0.025 considerando um nível de significância de 5%, ou seja, ela não se adequa no modelo de riscos proporcionais.

- (c) Ajuste um modelo estratificado por ninhada, obtenha as estimativas e interprete os parâmetros. Obtenha os resíduos de Schoenfeld e faça o mesmo teste do item anterior para testar a proporcionalidade dos riscos. Compare os resultados.

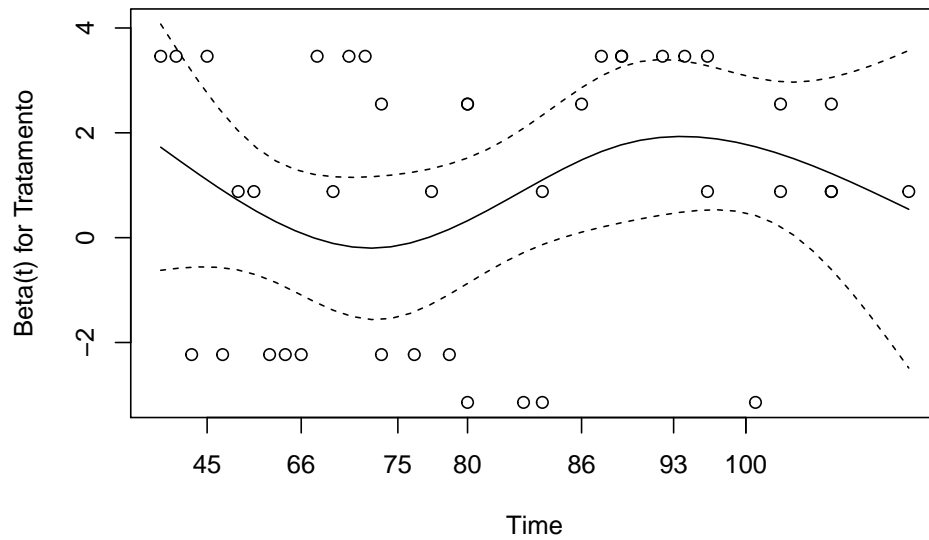
Resolução

Ajustando um modelo estratificado por ninhada, temos:

```
## Call:
## coxph(formula = Surv(Tempo, Delta) ~ Tratamento + strata(Ninhada),
##       data = litter_Data)
##
##      n= 150, number of events= 40
##
##              coef exp(coef) se(coef)      z Pr(>|z|)
## Tratamento 0.8800      2.4110   0.3772 2.333   0.0196 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## Tratamento      2.411      0.4148    1.151      5.05
##
## Concordance= 0.67 (se = 0.088 )
## Likelihood ratio test= 5.58 on 1 df,  p=0.02
## Wald test            = 5.44 on 1 df,  p=0.02
## Score (logrank) test = 5.78 on 1 df,  p=0.02
```

Como temos um único parâmetro β a partir do modelo ajustado, pode-se dizer que a taxa de ocorrência de tumores em ratos que recebem a droga é 2.5 vezes a taxa associada a ratos que recebem placebo.

Resíduos de Schoenfeld com a transformação KM:



Em que se pode notar que a variável tratamento apresenta um comportamento linear, sendo possível traçar uma reta ao longo do tempo.

Fazendo a tabela com o resultado do teste, em que:

$$\begin{cases} H_0 : O \text{ modelo e de riscos proporcionais} \\ H_1 : O \text{ modelo nao e de riscos proporcionais} \end{cases}$$

	chisq	df	p
Tratamento	1.054	1	0.305
GLOBAL	1.054	1	0.305

Queremos testar se o modelo de riscos proporcionais se adequa a esse caso, e ao observar a tabela acima assim como resultado do gráfico acima a variável tratamento não se rejeita a hipótese nula com um p-valor de 0.305 considerando um nível de significância de 5%, ou seja, ela se adequa no modelo de riscos proporcionais.

Assim comparando com o item anterior, adicionando o efeito de ninhada, assim como esperado a suposição de riscos proporcionais é satisfeita.

Exercício 3

Um estudo foi feito sobre o efeito da radiação na sobrevivência de ratos. Um grupo de ratos recebeu uma dose de 300 rad de radiação quando tinham entre 5 e 6 semanas de vida e foram acompanhados até o óbito. Quando morriam, os ratos eram necropsiados e a causa da morte determinada. Em particular, os pesquisadores tinham interesse em estudar as mortes por um tipo específico de linfoma um tipo de sarcoma.

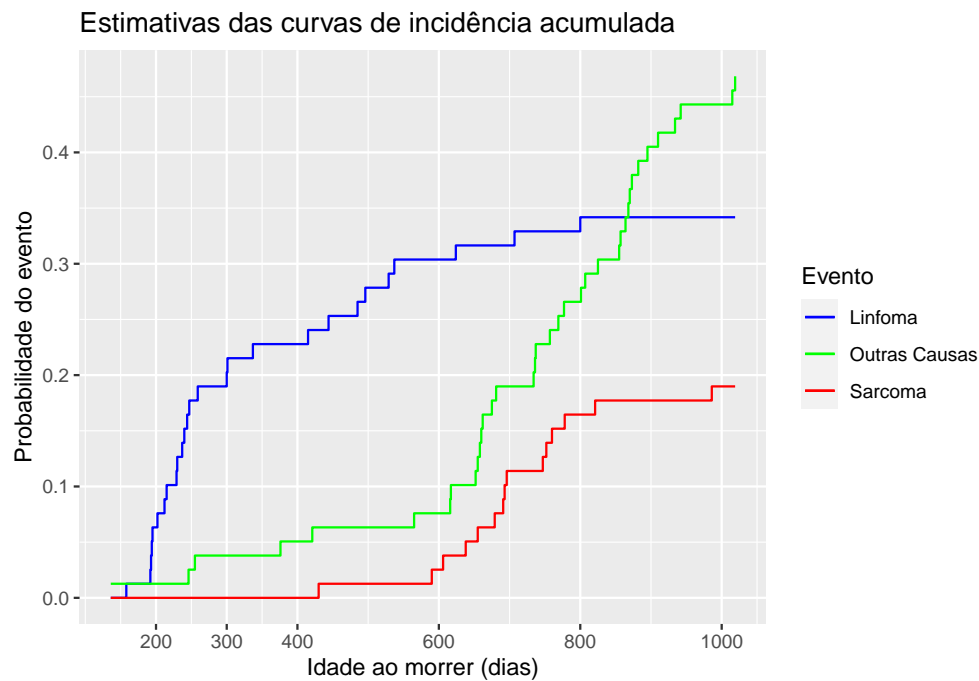
Os tempos de vida, em dias, dos ratos e as causas das mortes estão apresentadas na tabela abaixo.

Causa da morte	Idade ao morrer (dias)
Linfoma tímico	158, 192, 194, 195, 202, 215, 212, 215, 229, 230, 237,
	240, 244, 247, 259, 300, 301, 337, 415, 444, 485, 496,
Sarcoma de células reticulares	529, 537, 624, 707, 800
	430, 590, 606, 638, 655, 679, 691, 693, 696,
Outras causas	747, 752, 760, 778, 821, 986
	136, 246, 255, 376, 421, 565, 616, 617, 652, 655, 658,
	660, 662, 675, 681, 734, 736, 737, 757, 769, 777, 801,
	807, 825, 855, 857, 864, 868, 870, 873, 882, 895, 910,
	934, 942, 1015, 1019

(a) Obtenha as curvas de incidência acumulada dos três riscos competitivos. Comente.

Resolução

As curvas de incidência acumulada dos três riscos competitivos são:



No gráfico acima apresentam-se as curvas referentes as estimativas de probabilidade de falha por cada uma dos desfechos (competitivos entre si) ao longo do tempo. Assim podemos observar que há um maior risco do rato vir a sofrer primeiro uma morte por linfoma tímico, sendo seguido por outras causas e com menor risco sarcoma de células reticulares até 800 dias. No final do estudo há uma inversão das curvas, logo há um maior risco do rato vir a sofrer primeiro uma morte por outras causas do que por linfoma tímico.

(b) Obtenha os valores estimados das funções de incidência acumulada dos três riscos competitivos nos instantes $t = 200, 300, 400, 600, 800$ e 1000 .

Resolução

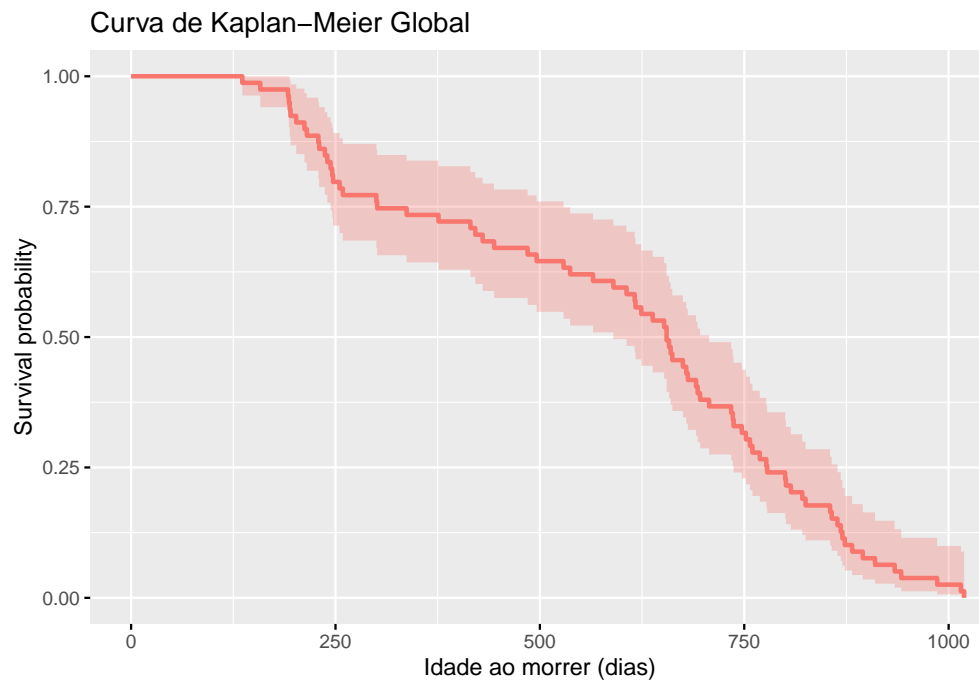
Os valores estimados das funções de incidência acumulada dos três riscos competitivos nos instantes $t = 200, 300, 400, 600, 800$ e 1000 são:

time	Linfoma	Outras Causas	Sarcoma
200	0.063	0.013	0.000
300	0.203	0.038	0.000
400	0.228	0.051	0.000
600	0.304	0.076	0.025
800	0.342	0.266	0.165
1000	0.342	0.443	0.190

- (c) Obtenha a curva de Kaplan Meier da sobrevivência global dos dados. Calcule o valor da função de sobrevivência nos instantes $t = 200, 300, 400, 600, 800$ e 1000 . Observe que a soma dos valores das funções do incidência acumulada em cada instante é igual a 1 menos o valor da curva de Kaplan-Meier naquele ponto.

Resolução

A curva de Kaplan Meier da sobrevivência global dos dados:



E os valores da função de sobrevivência nos instantes $t = 200, 300, 400, 600, 800$ e 1000 são:

time	km	1-km
200	0.924	0.076
300	0.759	0.241
400	0.722	0.278
600	0.595	0.405
800	0.228	0.772
1000	0.025	0.975

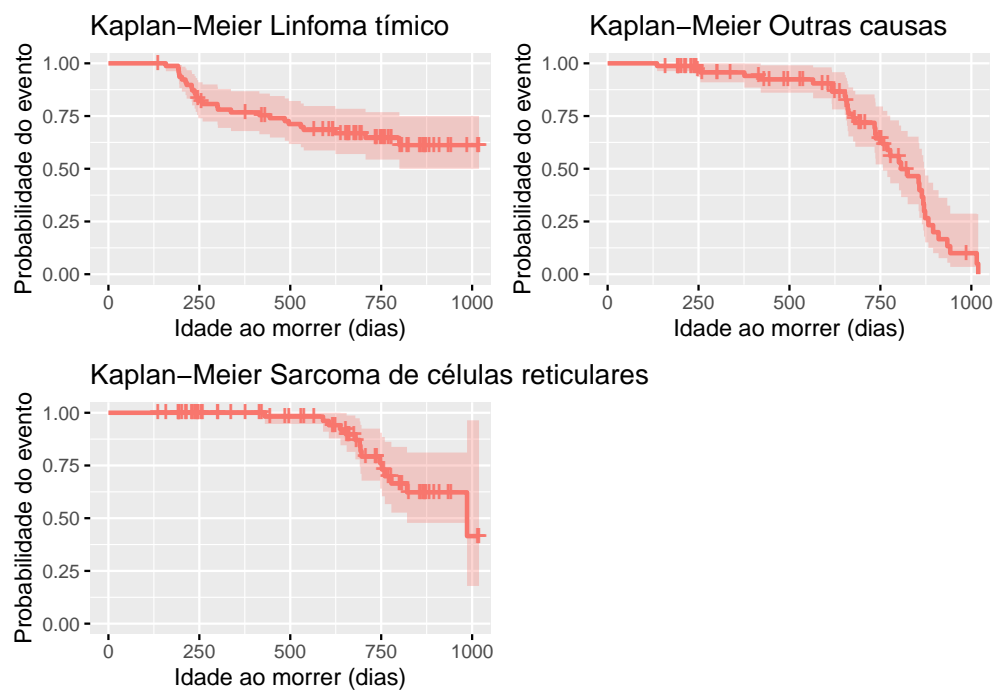
Em que podemos observar que a soma dos valores das funções de incidência acumulada em cada instante é igual a 1 menos o valor da curva de Kaplan-Meier naquele ponto.

- (d) Obtenha as curvas de sobrevivência de Kaplan Meier marginal dos dados, para cada tipo de evento, considerando a ocorrência dos outros eventos como censuras à direita. Coloque em um mesmo gráfico as curvas de incidência acumulada e 1 menos a curva de Kaplan-Meier.

Compare e comente as diferenças, enfatizando o que cada curva de fato estima.

Resolução

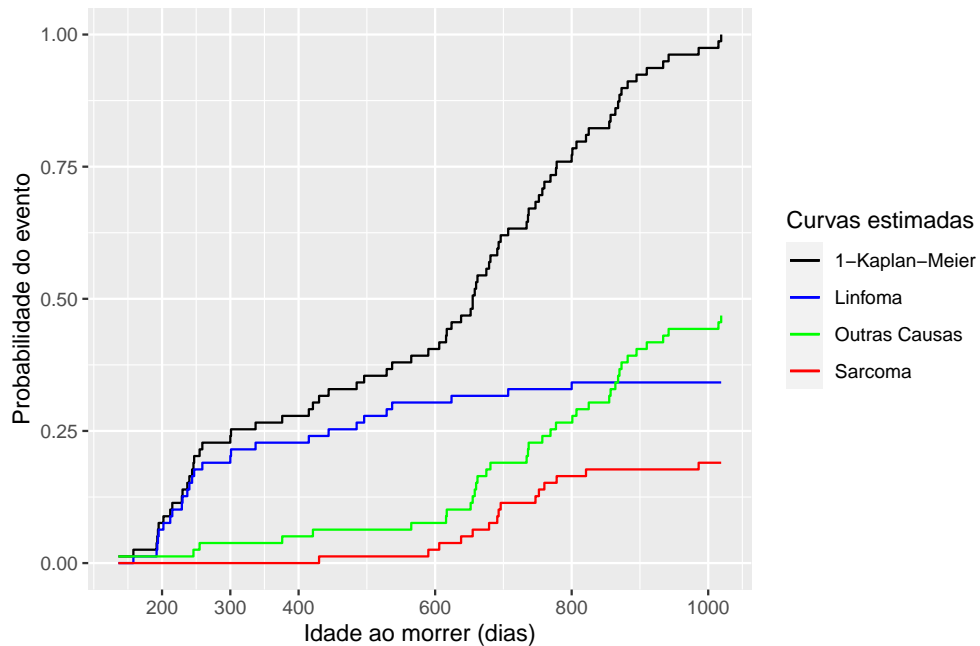
As curvas de sobrevivência de Kaplan Meier marginal dos dados, para cada tipo de evento, considerando a ocorrência dos outros eventos como censuras à direita:



Em que se pode notar que a curva por linfoma tímico decai mais suavemente a partir de 200 dias, a curva por outras causas, por sua vez, decai mais rapidamente a partir de 600 dias e a curva por sarcoma de células reticulares decai de forma suave a partir de 600 dias de estudo.

E as curvas de incidência acumulada e de 1-km, temos:

Estimativas de 1-Kaplan-Meier e incidência acumulada



Nota-se que a curva 1-Kaplan-Meier seria a soma das curvas azul, verde e vermelha, ou seja a probabilidade de qualquer um dos 3 eventos ocorrer.

Códigos

```
# local de trabalho
setwd("~/Área de Trabalho/Lista 5")

# Pacotes
library(ggplot2)
library(survival)
library(survminer)
library(KMsurv)
library(cmprsk)
library(readr)

# Exercício 1

# Dados
BMT_DataOBITO <- read_delim("BMT_DataOBITO.csv", ";",
                           escape_double = FALSE, trim_ws = TRUE)
BMT_DataRELAPSED <- read_delim("BMT_DataRELAPSED.csv", ";",
                               escape_double = FALSE, trim_ws = TRUE)

# item b

# modelo inicial
modelo1.cox <- coxph(Surv(Tinicio, Tfim, Delta2)~A+C+Z1+Z3+Z7+Z8+Z10,
```

```

                                data=BMT_DataRELAPSED)
summary(modelo1.cox)

knitr::kable(drop1(modelo1.cox, test="Chisq"),digits = 3)

# modelo final
modelo2.cox <- coxph(Surv(Tinicio, Tfim, Delta2)~Z8, data=BMT_DataRELAPSED)
summary(modelo2.cox)

drop1(modelo2.cox, test="Chisq")

# resíduos de Schoenfeld
zph <- cox.zph(modelo2.cox)
plot(zph)
knitr::kable(zph$table,digits = 3)

# item c
# modelo inicial
modelo3.cox <- coxph(Surv(Tinicio, Tfim, Delta1)~.-Id, data=BMT_DataOBITO)
summary(modelo3.cox)

drop1(modelo3.cox, test="Chisq")

# modelo final
modelo4.cox <- coxph(Surv(Tinicio, Tfim, Delta1)~Z8, data=BMT_DataOBITO)
summary(modelo4.cox)

knitr::kable(drop1(modelo3.cox, test="Chisq"),digits = 3)

# resíduos de Schoenfeld
zph2 <- cox.zph(modelo4.cox)
plot(zph2)
knitr::kable(zph2$table,digits = 3)

# Exercício 2

# dados
litter_Data <- read.csv("litter-data.csv")

# Kaplan Meier
KM <- survfit(Surv(Tempo,Delta)~ Tratamento,data = litter_Data)

# Grafico Kaplan-Meier
ggsurvplot(KM, data =litter_Data, linetype = 1, risk.table = TRUE,
            xlab="Tempo (em semanas)",
            ylab=expression(hat(S(t))), risk.table.title="Em risco",
            legend.title ="Tratamento",ggtheme=theme_gray(),
            legend.labs=c("Placebo", "Droga"), pval=TRUE)

# item b

# modelo de cox ignorando ninhada
cox_fit1 <- coxph(Surv(Tempo, Delta)~ Tratamento,data=litter_Data)

```

```

summary(cox_fit1)

# Resíduos de Schoenfeld
zph_km <- cox.zph(cox_fit1,transform='km')

plot(zph_km)

# tabela do testeS
knitr::kable(zph_km$table,digits = 3)

# item c

# modelo de cox estratificado por ninhada
cox_strat_fit2 <-coxph(Surv(Tempo, Delta)~ Tratamento+strata(Ninhada),
                      data = litter_Data)
summary(cox_strat_fit2)

# Resíduos de Schoenfeld
zph_km2 <- cox.zph(cox_strat_fit2,transform='km')

plot(zph_km2)

# tabela do teste
knitr::kable(zph_km2$table,digits = 3)

# Exercício 3

# Dados
lin <- c(158, 192, 193, 194, 195, 202, 212, 215, 229, 230, 237,
        240, 244, 247, 259, 300, 301, 337, 415, 444, 485, 496,
        529, 537, 624, 707, 800)
sar <- c(430, 590, 606, 638, 655, 679, 691, 693, 696,
        747, 752, 760, 778, 821, 986)
out <- c(136, 246, 255, 376, 421, 565, 616, 617, 652, 655, 658,
        660, 662, 675, 681, 734, 736, 737, 757, 769, 777, 801,
        807, 825, 855, 857, 864, 868, 870, 873, 882, 895, 910,
        934, 942, 1015, 1019)

causa <- factor(c(rep('Linfoma',27),rep('Sarcoma',15),rep('Outras_Causas',37)))
tempo <- c(lin,sar,out)
data3 <- data.frame(tempo,causa)

# item a

# estimativas das curvas de incidência acumulada
CIF <- cuminc(tempo,factor(causa), cencode = 0)

# Estimativas de km globais
KM3 <- survfit(Surv(tempo)~1,data = data3)

# dados para 1-km e cuvras de incidência acumulada
df3d <- data.frame(KM3$time,1-KM3$surv,
                   t(timepoints(CIF, times = data3$tempo)$est))

```



```

colnames(df3d) <- c("time", "F", "Linfoma", "Outras_Causas", "Sarcoma")

# cores
colors <- c("Linfoma" = "Blue",
            "Sarcoma" = "Red",
            "Outras Causas" = "Green")

# curvas das estimativas das curvas de incidência acumulada
ggplot(data = df3d) +
  geom_step(aes(x = time, y = Linfoma, color = "Linfoma")) +
  geom_step(aes(x = time, y = Sarcoma, color = "Sarcoma")) +
  geom_step(aes(x = time, y = Outras_Causas, color = "Outras Causas")) +
  scale_x_continuous(breaks = c(200, 300, 400, 600, 800, 1000)) +
  scale_color_manual(values = colors) +
  labs(x = "Idade ao morrer (dias)",
       y = "Probabilidade do evento",
       title = "Estimativas das curvas de incidência acumulada",
       color = "Evento")

# item b
knitr::kable(data.frame(time = c(200, 300, 400, 600, 800, 1000),
                          t(timepoints(CIF, times = c(200, 300, 400, 600, 800, 1000))$est)),
              digits = 3, row.names = F, col.names = c("time", "Linfoma", "Outras Causas", "Sarcoma"))

# item c

# Grafico Kaplan-Meier global
ggsurvplot(KM3, data = data3, linetype = 1, xlab = "Idade ao morrer (dias)", legend = 'none',
            title = 'Curva de Kaplan-Meier Global',
            ggtheme = theme_gray())

# tabela com as estimativas de km para o t=200, 300, 400, 600, 800 e 1000
table <- summary(KM3, time = c(200, 300, 400, 600, 800, 1000))
table <- data.frame(table$time, table$surv, 1 - table$surv)
colnames(table) <- c("time", "km", "1-km")
knitr::kable(table, digits = 3)

# item d
linf <- ifelse(causa == 'Linfoma', 1, 0)
sarc <- ifelse(causa == 'Sarcoma', 1, 0)
oc <- ifelse(causa == 'Outras_Causas', 1, 0)

# Km evento linfoma
KM4 <- survfit(Surv(tempo, linf) ~ 1, data = data3)

# Grafico Kaplan-Meier linfoma
plot1 <- ggsurvplot(KM4, legend = 'none',
                    title = 'Kaplan-Meier Linfoma tímico',
                    ggtheme = theme_gray()) +
  labs(y = 'Probabilidade do evento',
       x = 'Idade ao morrer (dias)')

# Km evento sarcoma

```

```

KM5 <- survfit(Surv(tempo,sarc)~1,data = data3)

# Grafico Kaplan-Meier sarcoma
plot2 <- ggsurvplot(KM5, legend='none',
                    title='Kaplan-Meier Sarcoma de células reticulares',
                    ggtheme=theme_gray()) +
  labs(y='Probabilidade do evento',
       x='Idade ao morrer (dias)')

# Km evento outras causas
KM6 <- survfit(Surv(tempo,oc)~1,data = data3)

# Grafico Kaplan-Meier outras causas
plot3 <- ggsurvplot(KM6,legend='none',
                    title='Kaplan-Meier Outras causas',
                    ggtheme=theme_gray()) +
  labs(y='Probabilidade do evento',
       x='Idade ao morrer (dias)')

# PLOT das 3 curvas
arrange_ggsurvplots(list(plot1, plot2, plot3),
                     print = TRUE,ncol = 2, nrow = 2)

# cores
colors <- c("1-Kaplan-Meier" = "Black",
            "Linfoma" = "Blue",
            "Sarcoma" = "Red",
            "Outras Causas" = "Green")

# Gráficos das estimativas de 1-Kaplan-Meier e incidência acumulada
ggplot(data = df3d) +
  geom_step(aes(x = time, y = F,color = "1-Kaplan-Meier")) +
  geom_step(aes(x = time, y = Linfoma,color = "Linfoma")) +
  geom_step(aes(x = time, y = Sarcoma,color = "Sarcoma")) +
  geom_step(aes(x = time, y = Outras_Causas,color = "Outras Causas")) +
  scale_x_continuous(breaks = c(200,300,400,600,800,1000)) +
  scale_color_manual(values = colors) +
  labs(x="Idade ao morrer (dias)",
       y="Probabilidade do evento",
       title = "Estimativas de 1-Kaplan-Meier e incidência acumulada",
       color = "Curvas estimadas")

```