Gabarito da 5ª Lista de Exercícios - MAE514

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

Exercício 1

Vamos considerar 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 (alogênico) 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**. Vamos importar os dados no R:

```
# IMPORTANDO OS DADOS NO R
library(readr)
bmt <- read_csv("BMT_Data.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
bmt
```

```
## # A tibble: 137 x 22
##    Dgroup    T1    T2 Delta1 Delta2 Delta3    TA     A    TC     C    TP     P
##     <dbl> <dbl> <dbl>  <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1      1  2081  2081      0      0      0    67     1   121     1    13     1
##  2      1  1602  1602      0      0      0  1602     0   139     1    18     1
##  3      1  1496  1496      0      0      0  1496     0   307     1    12     1
##  4      1  1462  1462      0      0      0    70     1    95     1    13     1
##  5      1  1433  1433      0      0      0  1433     0   236     1    12     1
##  6      1  1377  1377      0      0      0  1377     0   123     1    12     1
##  7      1  1330  1330      0      0      0  1330     0    96     1    17     1
##  8      1   996   996      0      0      0    72     1   121     1    12     1
##  9      1   226   226      0      0      0   226     0   226     0    10     1
## 10      1  1199  1199      0      0      0  1199     0    91     1    29     1
## # ... with 127 more rows, and 10 more variables: Z1 <dbl>, Z2 <dbl>, Z3 <dbl>,
## #   Z4 <dbl>, Z5 <dbl>, Z6 <dbl>, Z7 <dbl>, Z8 <dbl>, Z9 <dbl>, Z10 <dbl>
```

a)

Vamos definir duas variáveis dependentes do tempo que indicam a ocorrência de DECH aguda e crônica:

$$X_A(t) = \begin{cases} 1, & \text{se } t < \text{ tempo em que ocorreu a DECH aguda} \\ 0, & \text{se } t \geq \text{ tempo em que ocorreu a DECH aguda} \end{cases}$$

$$X_C(t) = \begin{cases} 1, & \text{se } t < \text{ tempo em que ocorreu a DECH crônica} \\ 0, & \text{se } t \geq \text{ tempo em que ocorreu a DECH crônica} \end{cases}$$

Assim, considerando as variáveis idade do paciente, sexo do paciente, tempo até o transplante, classificação morfológica Franco-Americana-Britânica (FAB), tratamento profilático para DECH (MTX) e as indicadoras definidas anteriormente, temos o seguinte modelo semiparamétrico de Cox:

$$\alpha(t) = \alpha_0(t).exp\left(\beta_I X_I + \beta_S X_S + \beta_T X_T + \beta_M X_M + \beta_P X_P + \beta_A X_A(t) + \beta_C X_C(t)\right),$$

em que $\alpha(t)$ é a funcão de taxa de falha, $\alpha_0(t)$ é o risco basal e as variáveis X são:

- X_I: idade do paciente;
- X_S: sexo do paciente;
- X_T: tempo até o transplante;
- X_M: classificação morfológica FAB;
- X_P: tratamento profilático para DECH (MTX).

Além disso, os $\beta's$ são os coeficientes associados às variáveis listadas.

b)

Vamos ajustar um modelo de Cox com as variáveis do item **a)**.

Para tal, vamos adaptar o banco de dados para o formato "longo", para o ajuste no R. Primeiro, definimos até três intervalos de tempo para cada paciente, delimitados pelos seguintes tempos, ordenados do menor para o maior: 0, T_A (incluído caso tenha ocorrido DECH aguda e $T_A \leq T_2$), T_C (incluído caso tenha ocorrido DECH crônica e $T_C \leq T_2$) e T_2. Os dados de cada paciente para cada intervalo aparecerão em linhas distintas, ou seja, os dados de cada paciente podem constar em até três linhas do banco de dados transformado. Em cada linha, replicam-se as informações relativas às variáveis não dependentes do tempo (idade, sexo, tempo até o transplante, classificação morfológica FAB e tratamento profilático para DECH). Também em cada linha registram-se, de maneira condizente com o respectivo intervalo de tempo, os valores de $Z_A(t)$, $Z_C(t)$, e δ_2 (variável indicadora de recorrência ao fim do intervalo):

```
# AJUSTANDO O CONJUNTO DE DADOS NO R

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
```

```
library(survival)
library(ggplot2)
library(magrittr)
library(data.table)
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
library(survminer)
```

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

```
bmt$ID <- 1:nrow(bmt)
bmt <- bmt %>% mutate(A_corr = pmin(TA <= T2, A), C_corr = pmin(TC <= T2, C),
                      TA_corr = ifelse(A_corr == 1, TA, NA),
                      TC_corr = ifelse(C_corr == 1, TC, NA))
bmt[,c('T_0', 'T_1', 'T_2', 'T_3')] <- cbind(0, bmt[,c('TA_corr', 'TC_corr', 'T2')]) %>%
   apply(1, function(x) sort(x, na.last = T)) %>% t()
bmt <- bmt %>% mutate(ZA_T0_T1 = 0, ZA_T1_T2 = pmin(A_corr, T_1 >= TA),
                      ZA_T2_T3 = pmin(A_corr, T_2 >= TA), ZC_T0_T1 = 0,
                      ZC_T1_T2 = pmin(C_corr, T_1 >= TC), ZC_T2_T3 = pmin(C_corr, T_2 >= TC),
                      Delta2_T0_T1 = pmin(Delta2, T2 == T_1), Delta2_T1_T2 = pmin(Delta2, T2 == T_2),
                      Delta2_T2_T3 = pmin(Delta2, T2 == T_3))
bmt_start_stop <- setnames(bmt[,c('ID', 'T_0', 'T_1', 'ZA_T0_T1', 'ZC_T0_T1', 'Delta2_T0_T1')],
                           c('ID', 'inicio', 'fim', 'apos_DECH_aguda', 'apos_DECH_cron', 'recorrencia')) %>%
   rbind(setnames(bmt[,c('ID', 'T_1', 'T_2', 'ZA_T1_T2', 'ZC_T1_T2', 'Delta2_T1_T2')],
                  c('ID', 'inicio', 'fim', 'apos_DECH_aguda', 'apos_DECH_cron', 'recorrencia'))) %>%
   rbind(setnames(bmt[,c('ID', 'T_2', 'T_3', 'ZA_T2_T3', 'ZC_T2_T3', 'Delta2_T2_T3')],
                  c('ID', 'inicio', 'fim', 'apos_DECH_aguda', 'apos_DECH_cron', 'recorrencia'))) %>%
   filter(!is.na(fim))
bmt_start_stop <- bmt_start_stop %>%
   left_join(select(bmt, ID, Z1, Z3, Z7, Z8, Z10), by = 'ID') %>%
   rename(idade = Z1, sexo = Z3, tempo_transpl = Z7, clas_morfol_FAB = Z8, trat_profil = Z10)
```

Com os dados arranjados, fazemos o modelo de Cox completo:

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

modelo_cox_completo <- coxph(Surv(time = inicio, time2 = fim, recorrencia) ~ idade + sexo +
                               tempo_transpl + clas_morfol_FAB + trat_profil +
                               apos_DECH_aguda + apos_DECH_cron,
                             bmt_start_stop)
summary(modelo_cox_completo)
```

```
## Call:
## coxph(formula = Surv(time = inicio, time2 = fim, recorrencia) ~
##     idade + sexo + tempo_transpl + clas_morfol_FAB + trat_profil +
##         apos_DECH_aguda + apos_DECH_cron, data = bmt_start_stop)
##
##   n= 222, number of events= 42
##
##                       coef  exp(coef)   se(coef)      z Pr(>|z|)
## idade            0.0007762  1.0007765  0.0181277  0.043  0.96584
## sexo            -0.2157522  0.8059350  0.3210784 -0.672  0.50161
## tempo_transpl   -0.0003815  0.9996186  0.0006546 -0.583  0.56003
## clas_morfol_FAB  1.0360605  2.8180933  0.3163231  3.275  0.00106 **
## trat_profil      0.4141110  1.5130250  0.3553380  1.165  0.24386
## apos_DECH_aguda -0.4287630  0.6513143  0.4876202 -0.879  0.37924
## apos_DECH_cron   0.1159004  1.1228840  0.3940184  0.294  0.76864
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## idade              1.0008     0.9992    0.9658     1.037
## sexo               0.8059     1.2408    0.4295     1.512
## tempo_transpl      0.9996     1.0004    0.9983     1.001
## clas_morfol_FAB    2.8181     0.3548    1.5160     5.239
## trat_profil        1.5130     0.6609    0.7540     3.036
## apos_DECH_aguda    0.6513     1.5354    0.2505     1.694
## apos_DECH_cron     1.1229     0.8906    0.5187     2.431
##
## Concordance= 0.689  (se = 0.037 )
## Likelihood ratio test= 14.54  on 7 df,   p=0.04
## Wald test            = 14.48  on 7 df,   p=0.04
## Score (logrank) test = 15.73  on 7 df,   p=0.03
```

Analisando o ajuste do modelo, podemos concluir, a um nível de significância 5%, que a única variável significativa é a classificação morfológica FAB (se for retirando cada uma das variáveis, o resultado é o mesmo). Assim, o modelo apenas com a covariável significativa é dado por:

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

modelo_cox_final <- coxph(Surv(time = inicio, time2 = fim, recorrencia) ~ clas_morfol_FAB,
                          bmt_start_stop)
summary(modelo_cox_final)
```

```
## Call:
```

```
## coxph(formula = Surv(time = inicio, time2 = fim, recorrencia) ~
##     clas_morfol_FAB, data = bmt_start_stop)
##
##   n= 222, number of events= 42
##
##                   coef exp(coef) se(coef)     z Pr(>|z|)
## clas_morfol_FAB 1.0441    2.8408   0.3099 3.369 0.000755 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## clas_morfol_FAB     2.841      0.352     1.548     5.215
##
## Concordance= 0.625  (se = 0.038 )
## Likelihood ratio test= 11.05  on 1 df,   p=9e-04
## Wald test            = 11.35  on 1 df,   p=8e-04
## Score (logrank) test = 12.4  on 1 df,   p=4e-04
```

Assim, temos que o risco estimado de um paciente com leucemia mieloide aguda (LMA) e escore FAB 4 ou 5 ter recorrência da doença é 2,84 vezes o risco dos demais pacientes, em todos os instantes.

Agora, vamos fazer a verificação da proporcionalidade dos riscos:

```r
# VERIFICANDO A PROPORCIONALIDADE NO R

cox.zph(modelo_cox_final)
```

```
##                   chisq df    p
## clas_morfol_FAB 0.00991  1 0.92
## GLOBAL          0.00991  1 0.92
```

O teste para a proporcionalidade dos riscos em relação à variável classificação morfológica FAB nos indica um valor-p de 0,92, o que nos leva a não rejeitar a hipótese de proporcionalidade dos riscos.

c)

Agora, vamos repetir os itens **a)** e **b)** considerando como resposta o tempo até o óbito.

Primeiro, refazemos o procedimento do item **b)** para adaptar o banco de dados novamente, com a diferença de que desta vez não precisamos corrigir e (quando e/ou , respectivamente), pois e são sempre (tempo até o óbito ou fim do estudo):

```r
# AJEITANDO A BASE DE DADOS NO R

bmt_1c <- read.csv('BMT_Data.csv')

bmt_1c$ID <- 1:nrow(bmt_1c)

bmt_1c <- bmt_1c %>% mutate(TA_corr = ifelse(A == 1, TA, NA),
                            TC_corr = ifelse(C == 1, TC, NA))
bmt_1c[,c('T_0', 'T_1', 'T_2', 'T_3')] <- cbind(0, bmt_1c[,c('TA_corr', 'TC_corr', 'T1')]) %>%
   apply(1, function(x) sort(x, na.last = T)) %>%
   t()
bmt_1c <- bmt_1c %>% mutate(ZA_T0_T1 = 0, ZA_T1_T2 = pmin(A, T_1 >= TA),
                            ZA_T2_T3 = pmin(A, T_2 >= TA), ZC_T0_T1 = 0,
                            ZC_T1_T2 = pmin(C, T_1 >= TC),
                            ZC_T2_T3 = pmin(C, T_2 >= TC),
                            Delta1_T0_T1 = pmin(Delta1, T1 == T_1),
                            Delta1_T1_T2 = pmin(Delta1, T1 == T_2),
                            Delta1_T2_T3 = pmin(Delta1, T1 == T_3))
bmt_start_stop_1c <- setnames(bmt_1c[,c('ID', 'T_0', 'T_1', 'ZA_T0_T1', 'ZC_T0_T1', 'Delta1_T0_T1')],
                              c('ID', 'inicio', 'fim', 'apos_DECH_aguda', 'apos_DECH_cron', 'obito')) %>%
   rbind(setnames(bmt_1c[,c('ID', 'T_1', 'T_2', 'ZA_T1_T2', 'ZC_T1_T2', 'Delta1_T1_T2')],
                  c('ID', 'inicio', 'fim', 'apos_DECH_aguda', 'apos_DECH_cron', 'obito'))) %>%
   rbind(setnames(bmt_1c[,c('ID', 'T_2', 'T_3', 'ZA_T2_T3', 'ZC_T2_T3', 'Delta1_T2_T3')],
                  c('ID', 'inicio', 'fim', 'apos_DECH_aguda', 'apos_DECH_cron', 'obito'))) %>%
filter(!is.na(fim))
bmt_start_stop_1c <- bmt_start_stop_1c %>%
   left_join(select(bmt_1c, ID, Z1, Z3, Z7, Z8, Z10), by = 'ID') %>%
   rename(idade = Z1, sexo = Z3, tempo_transpl = Z7, clas_morfol_FAB = Z8, trat_profil = Z10)
```

Com os dados arranjados, fazemos o modelo de Cox completo:

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

modelo_cox_completo <- coxph(Surv(time = inicio, time2 = fim, obito) ~ idade + sexo + tempo_transpl +
                               clas_morfol_FAB + trat_profil + apos_DECH_aguda + apos_DECH_cron,
                             bmt_start_stop_1c)
summary(modelo_cox_completo)
```

```
## Call:
## coxph(formula = Surv(time = inicio, time2 = fim, obito) ~ idade +
##     sexo + tempo_transpl + clas_morfol_FAB + trat_profil + apos_DECH_aguda +
##     apos_DECH_cron, data = bmt_start_stop_1c)
##
##   n= 223, number of events= 81
##
##                       coef  exp(coef)   se(coef)      z Pr(>|z|)
## idade            0.0081134  1.0081464  0.0123202  0.659  0.51019
## sexo            -0.1078528  0.8977597  0.2334941 -0.462  0.64415
## tempo_transpl    0.0001395  1.0001395  0.0003440  0.405  0.68520
## clas_morfol_FAB  0.6184540  1.8560564  0.2321508  2.664  0.00772 **
## trat_profil      0.4087886  1.5049935  0.2554075  1.601  0.10948
## apos_DECH_aguda  0.3701644  1.4479726  0.2944715  1.257  0.20874
```

```
## apos_DECH_cron  -0.0390302  0.9617217  0.2720083 -0.143  0.88590
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## idade              1.0081     0.9919    0.9841     1.033
## sexo               0.8978     1.1139    0.5681     1.419
## tempo_transpl      1.0001     0.9999    0.9995     1.001
## clas_morfol_FAB    1.8561     0.5388    1.1776     2.925
## trat_profil        1.5050     0.6645    0.9123     2.483
## apos_DECH_aguda    1.4480     0.6906    0.8130     2.579
## apos_DECH_cron     0.9617     1.0398    0.5643     1.639
##
## Concordance= 0.624  (se = 0.03 )
## Likelihood ratio test= 11.57  on 7 df,   p=0.1
## Wald test            = 11.78  on 7 df,   p=0.1
## Score (logrank) test = 12.07  on 7 df,   p=0.1
```

Analisando o ajuste do modelo, podemos concluir, a um nível de significância 5%, que a única variável significativa é a classificação morfológica FAB (se for retirando cada uma das variáveis, o resultado é o mesmo). Assim, o modelo apenas com a covariável significativa é dado por:

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

modelo_cox_final <- coxph(Surv(time = inicio, time2 = fim, obito) ~ clas_morfol_FAB,
                          bmt_start_stop_1c)
summary(modelo_cox_final)
```

```
## Call:
## coxph(formula = Surv(time = inicio, time2 = fim, obito) ~ clas_morfol_FAB,
##     data = bmt_start_stop_1c)
##
##   n= 223, number of events= 81
##
##                    coef exp(coef) se(coef)     z Pr(>|z|)
## clas_morfol_FAB 0.5761    1.7791   0.2257 2.552   0.0107 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## clas_morfol_FAB     1.779     0.5621     1.143     2.769
##
## Concordance= 0.561  (se = 0.027 )
## Likelihood ratio test= 6.24  on 1 df,   p=0.01
## Wald test            = 6.51  on 1 df,   p=0.01
## Score (logrank) test = 6.69  on 1 df,   p=0.01
```

Assim, temos que o risco estimado de um paciente com leucemia mieloide aguda (LMA) e escore FAB 4 ou 5 ter recorrência da doença é 78% maior que o risco dos demais pacientes, em todos os instantes.

Agora, vamos fazer a verificação da proporcionalidade dos riscos:

```
# VERIFICANDO A PROPORCIONALIDADE NO R

cox.zph(modelo_cox_final)
```

```
##                 chisq df   p
## clas_morfol_FAB 0.705  1 0.4
## GLOBAL          0.705  1 0.4
```

O teste para a proporcionalidade dos riscos em relação à variável classificação morfológica FAB nos indica um valor-p de 0,40, o que nos leva a não rejeitar a hipótese de proporcionalidade dos riscos.

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 referentes a este estudo estão no arquivo **litter-data.csv**}. Vamos importá-los para o R:

```
# IMPORTANDO OS DADOS NO R
library(readr)
dados_ex2 <- read_csv("litter-data.csv")
```

```
## Parsed with column specification:
## cols(
##   Tempo = col_double(),
##   Delta = col_double(),
##   Tratamento = col_double(),
##   Ninhada = col_double()
## )
```

```
dados_ex2
```

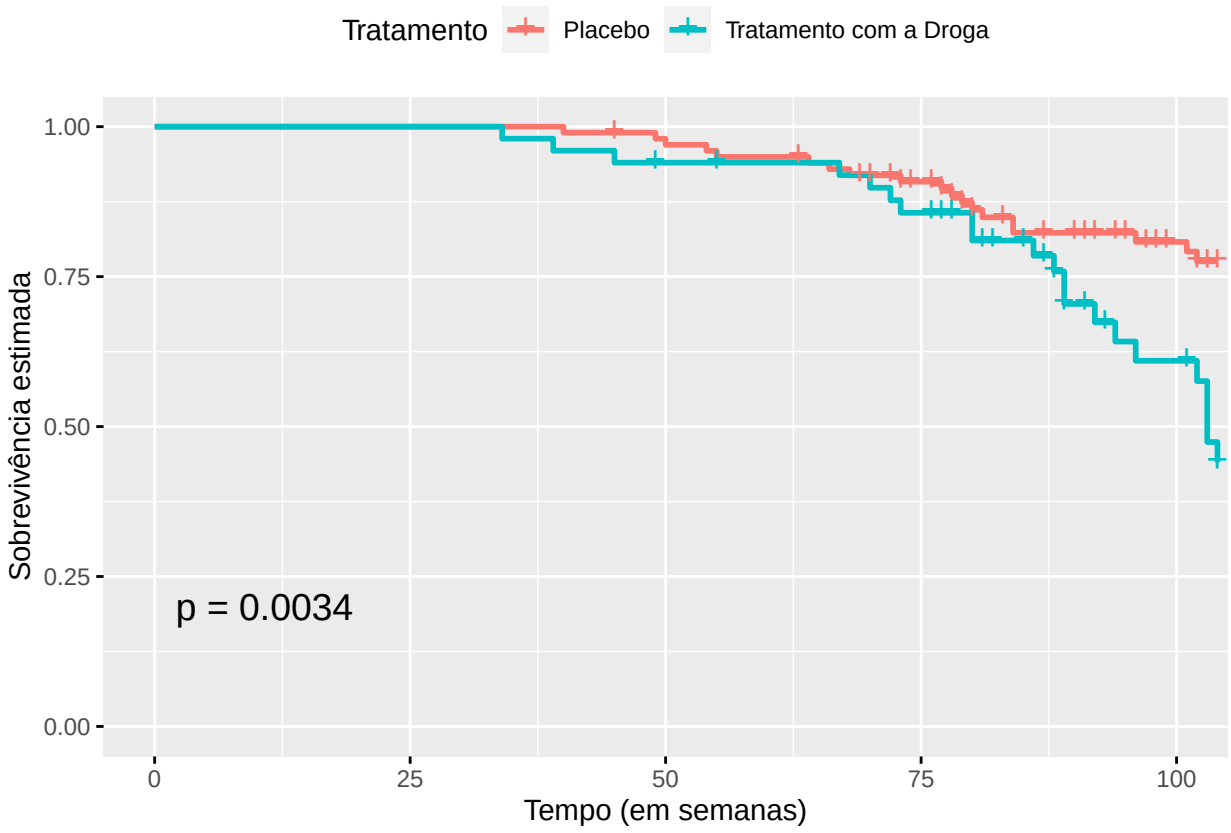
```
## # A tibble: 150 x 4
##    Tempo Delta Tratamento Ninhada
##    <dbl> <dbl>      <dbl>   <dbl>
##  1   101     0          1       1
##  2   104     0          1       2
##  3   104     0          1       3
##  4    77     0          1       4
##  5    89     0          1       5
##  6    88     1          1       6
##  7   104     1          1       7
##  8    96     1          1       8
##  9    82     0          1       9
## 10    70     1          1      10
## # ... with 140 more rows
```

a)

Vamos fazer as curvas de Kaplan-Meier para os dois grupos:

```
# IMPORTANDO OS DADOS NO R
library(ggplot2)
library(survival)
```

```r
library(survminer)

S_KM <- survfit(Surv(Tempo, Delta) ~ Tratamento, data = dados_ex2)

# Grafico Kaplan-Meier
ggsurvplot(S_KM, data = dados_ex2, linetype = 1, xlab="Tempo (em semanas)", ylab = "Sobrevivência estimada",
           legend.title = "Tratamento", ggtheme = theme_gray(),
           legend.labs = c("Placebo", "Tratamento com a Droga"), pval = TRUE)
```




```r
survdiff(Surv(Tempo, Delta) ~ Tratamento, data = dados_ex2) # log-rank
```

```
## Call:
## survdiff(formula = Surv(Tempo, Delta) ~ Tratamento, data = dados_ex2)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## Tratamento=0 100       19     27.5      2.65      8.59
## Tratamento=1  50       21     12.5      5.86      8.59
##
##  Chisq= 8.6  on 1 degrees of freedom, p= 0.003
```

Um pouco depois da 30ª semana, ocorrem as primeiras falhas. A partir daí, a curva de sobrevivência do tratamento está sempre abaixo ou igual a do placebo, indicando que a Droga do tratamento tem um potencial cancerígeno. Além disso, com o teste log-rank, podemos concluir que as curvas de sobrevivência são estatisticamente diferentes.

b)

Vamos ajustar um modelo de riscos proporcionais de Cox, ignorando o fato de existirem ratos pertencentes a uma mesma ninhada:

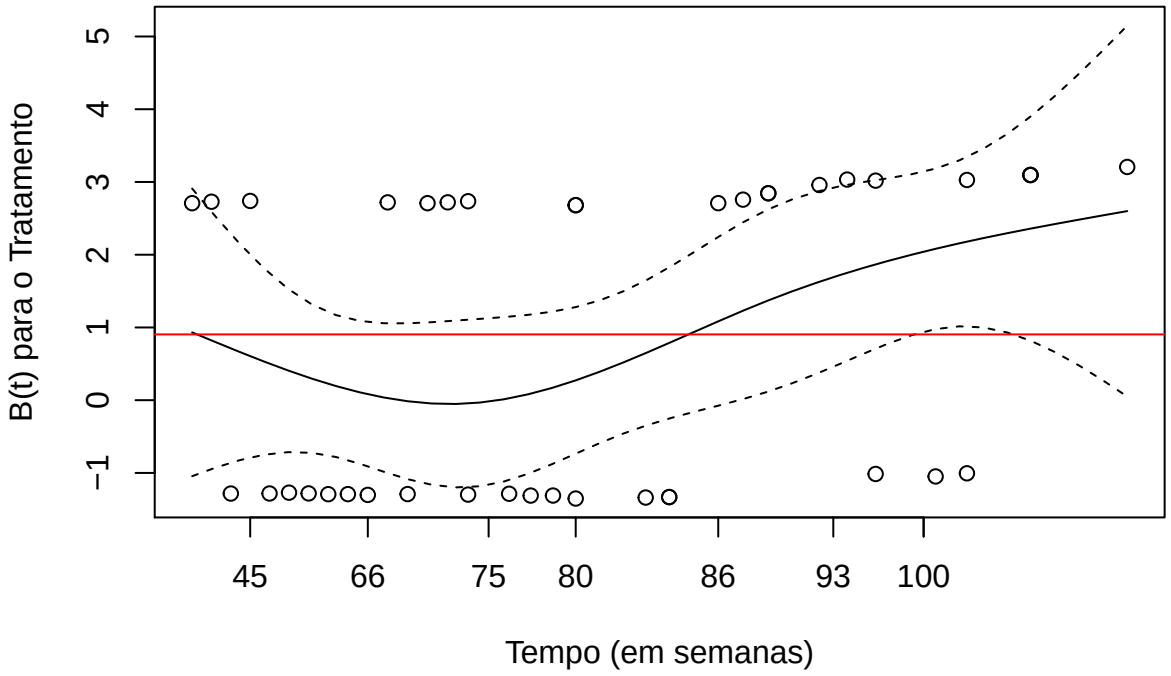
```
# AJUSTANDO UM MODELO DE COX NO R

modelo_cox <- coxph(Surv(Tempo, Delta) ~ Tratamento, data = dados_ex2)
summary(modelo_cox)
```

```
## Call:
## coxph(formula = Surv(Tempo, Delta) ~ Tratamento, data = dados_ex2)
##
##   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
```

O parâmetro associado à covariável Tratamento de valor 0,9047 indica que quando trocamos do placebo para a droga, a chance de se desenvolver câncer aumenta em $exp(0,9047) - 1 = 1,4711 = 147,11\%$. Ou seja, a chance de se ter câncer com a droga é mais do que o dobro da chance de se ter câncer com o placebo.

Agora, vamos obter os resíduos de Schoenfeld para testar a proporcionalidade dos riscos utilizando a transformação KM:

```
# RESÍDUOS DE SCHOENFELD NO R

library(KMsurv)
sch <- cox.zph(modelo_cox, transform = "km")
plot(sch, main = "Resíduos de Schoenfeld",
     ylab = "B(t) para o Tratamento", xlab = "Tempo (em semanas)")
abline(h = modelo_cox$coefficients, col = "red")
```



Podemos perceber com o gráfico que a variável Tratamento não apresenta um comportamento linear, não sendo possível traçar uma reta ao longo do tempo com o valor de $\hat{\beta}$ dentro das bandas de confiança.

Faremos agora o teste de hipóteses em que:

$$\begin{cases} H_0 : \text{O modelo é de riscos proporcionais} \\ H_A : \text{O modelo não é de riscos proporcionais} \end{cases}$$

```
# CONSULTANDO OS VALORES DO TESTE NO R

sch$table
```

```
##                chisq df          p
## Tratamento 5.019671  1 0.02506092
## GLOBAL     5.019671  1 0.02506092
```

Como só temos uma covariável, o teste local para o Tratamento é equivalente ao Global. Em ambos, rejeitamos a hipótese nula de que o modelo de riscos proporcionais é adequado ao nível de significância de 5%, com valor-p de 0,0251.

c)

Agora, vamos ajustar um modelo estratificado por ninhada:

```
# AJUSTANDO UM MODELO DE COX ESTRATIFICADO NO R

modelo_cox_ninhada <- coxph(Surv(Tempo, Delta) ~ Tratamento + strata(Ninhada),
                            data = dados_ex2)
summary(modelo_cox_ninhada)
```

```
## Call:
## coxph(formula = Surv(Tempo, Delta) ~ Tratamento + strata(Ninhada),
##     data = dados_ex2)
##
##   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
```

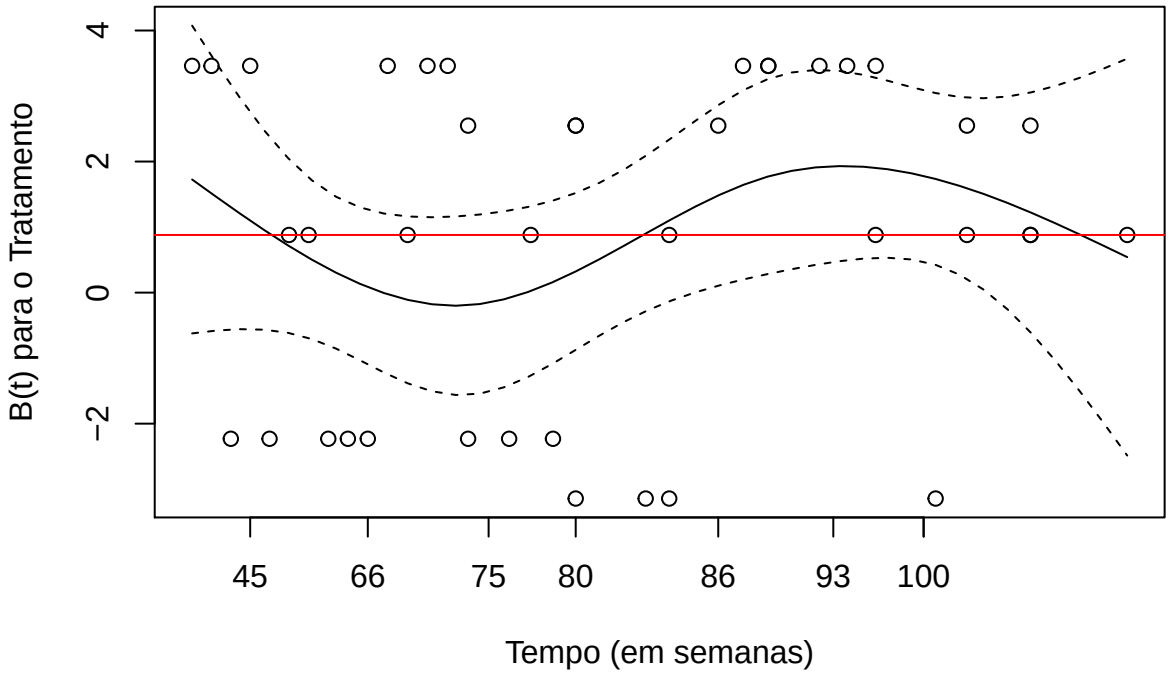
O parâmetro associado à covariável Tratamento de valor 0,8800 indica que quando trocamos do placebo para a droga, a chance de se desenvolver câncer aumenta em $exp(0,8800) - 1 = 1,4110 = 141,10\%$. Ou seja, a chance de se ter câncer com a droga é mais do que o dobro da chance de se ter câncer com o placebo.

Agora, vamos obter os resíduos de Schoenfeld para testar a proporcionalidade dos riscos utilizando a transformação KM:

```
# RESÍDUOS DE SCHOENFELD NO R

sch_ninhada <- cox.zph(modelo_cox_ninhada, transform = "km")
sch_ninhada
```

```
##            chisq df   p
## Tratamento  1.05  1 0.3
## GLOBAL      1.05  1 0.3
```

```
plot(sch_ninhada, main = "Resíduos de Schoenfeld",
     ylab = "B(t) para o Tratamento", xlab = "Tempo (em semanas)")
abline(h = modelo_cox_ninhada$coefficients, col = "red")
```



Podemos perceber com o gráfico que a variável Tratamento apresenta um comportamento linear, já que é possível traçar uma reta ao longo do tempo com o valor de $\hat{\beta}$ dentro das bandas de confiança.

Faremos agora o teste de hipóteses análogo ao anterior, em que:

$$\begin{cases} H_0 : \text{O modelo é de riscos proporcionais} \\ H_A : \text{O modelo não é de riscos proporcionais} \end{cases}$$

```
# CONSULTANDO OS VALORES DO TESTE NO R

sch_ninhada$table
```

```
##                chisq df         p
## Tratamento 1.054266  1 0.3045266
## GLOBAL     1.054266  1 0.3045266
```

Como só temos uma covariável, o teste local para o Tratamento é equivalente ao Global. Em ambos, não rejeitamos a hipótese nula de que o modelo de riscos proporcionais é adequado ao nível de significância de 5%, com valor-p de 0,3045.

Assim, consideramos que o remédio é realmente cancerígino e conseguimos averiguar isso com o ajuste do modelo de taxas de falha proporcionais de Cox que leva em conta a mesma origens de alguns ratos, já que uma ninhada pode ser mais propensa ao câncer do que outras.

Exercício 3

Um estudo foi feito sobre o efeito da radiaÇÃO na sobrevida 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 abaixo:

- **Linfoma tímico:** 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;
- **Sarcoma de células reticulares:** 430, 590, 606, 638, 655, 679, 691, 693, 696, 747, 752, 760, 778, 821, 986;
- **Outras causas:** 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.

Primeiro, vamos passar esses dados para o R:

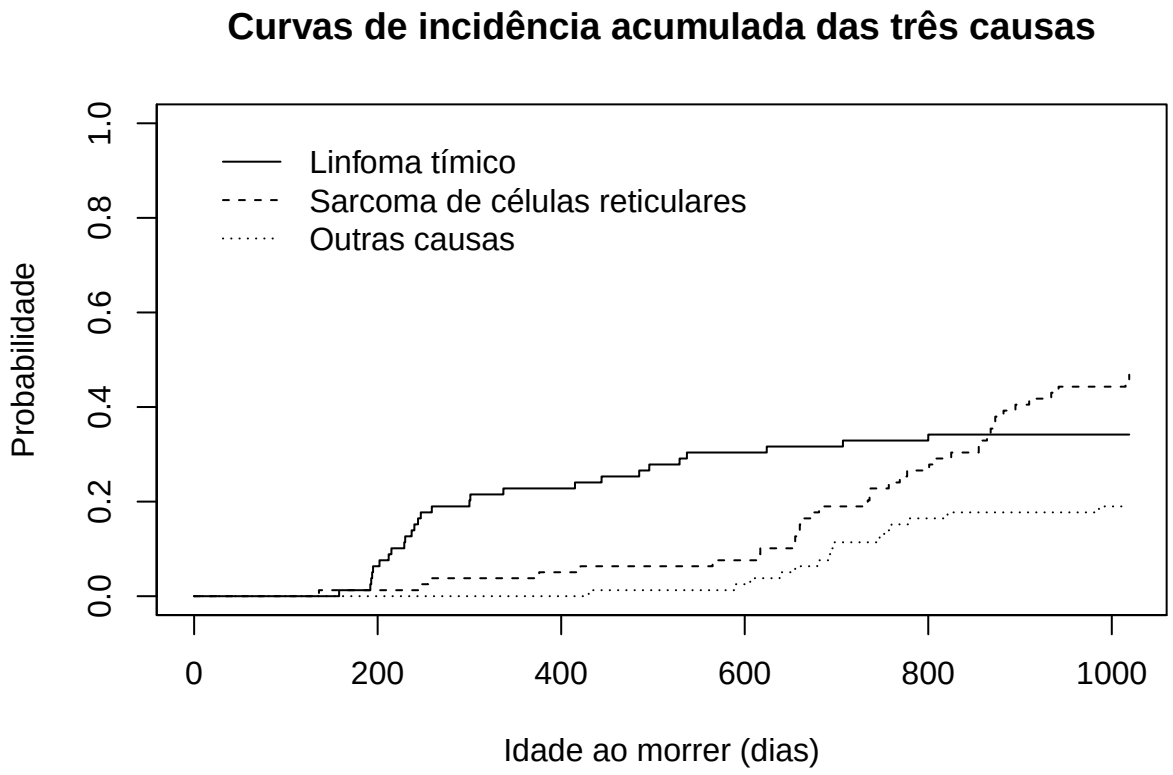
```
# COLOCANDO OS DADOS NO R

dados_ex3 <- data.frame(TEMPOS = 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, 430, 590, 606, 638, 655,
                                   679, 691, 693, 696, 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),
                        CAUSA = c(rep("Linfoma tímico", 27),
                                  rep("Sarcoma de células reticulares", 15),
                                  rep("Outras causas", 37)))

head(dados_ex3)
```

```
##   TEMPOS          CAUSA
## 1    158 Linfoma tímico
## 2    192 Linfoma tímico
## 3    193 Linfoma tímico
## 4    194 Linfoma tímico
## 5    195 Linfoma tímico
## 6    202 Linfoma tímico
```

a)

Vamos obter as curvas de incidência acumulada dos três riscos competitivos:

```r
# CURVAS DE INCIDÊNCIA NO R

library(cmprsk)
radiacao_incid_acum <- cuminc(dados_ex3$TEMPOS, dados_ex3$CAUSA)
plot(radiacao_incid_acum, xlab = 'Idade ao morrer (dias)', ylab = 'Probabilidade',
     main = "Curvas de incidência acumulada das três causas",
     curvlab = unique(dados_ex3$CAUSA))
box()
```



Até aproximadamente os 850 dias, a principal causa de morte foi o linfoma tímico. COntudo, quando consideramos mortes ocorridas após 850 dias, outras causas de morte foram mais comuns que o linfoma tímico e o sarcoma de células reticulares. Esse, por sua vez, foi a causa de morte menos frequente em todos os tempos.

b)

Vamos obter os valores estimados das funções de incidência acumulada dos três riscos competitivos nos instantes $t =$ 200; 300; 400; 600; 800 e 1000:

```r
# VALORES ESTIMADOS DAS FUNÇÕES DE INCIDÊNCIA NO R

timepoints(radiacao_incid_acum, c(200, 300, 400, 600, 800, 1000))$est
```

```
##                                       200        300        400        600
## 1 Linfoma tímico                0.06329114 0.20253165 0.22784810 0.30379747
## 1 Outras causas                 0.01265823 0.03797468 0.05063291 0.07594937
## 1 Sarcoma de células reticulares 0.00000000 0.00000000 0.00000000 0.02531646
##                                       800       1000
## 1 Linfoma tímico                0.3417722 0.3417722
## 1 Outras causas                 0.2658228 0.4430380
## 1 Sarcoma de células reticulares 0.1645570 0.1898734
```

Organizando em uma tabela mais legível, ficamos com:

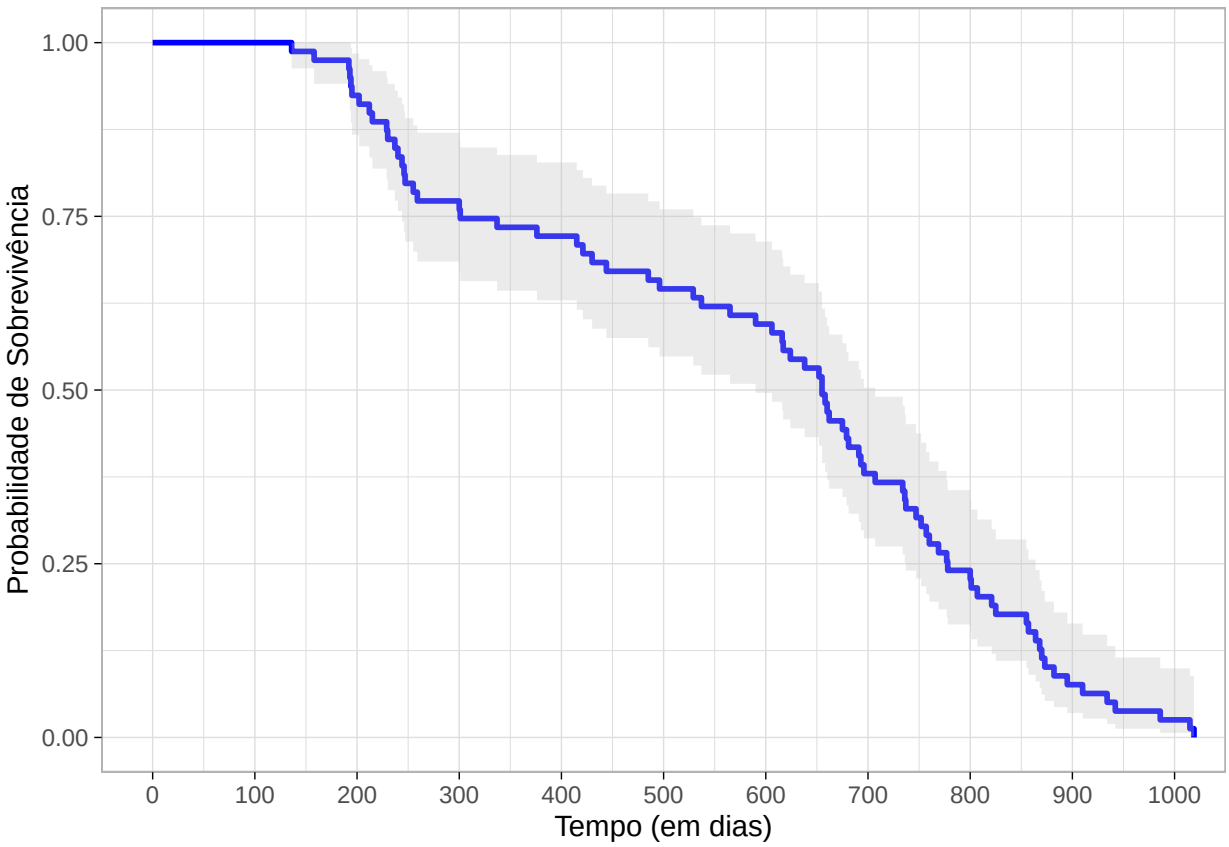
Causa da morte	**t = 200**	**t = 300**	**t = 400**	**t = 600**	**t = 800**	**t = 1000**
Linfoma tímico	0,0633	0,2025	0,2278	0,3038	0,3418	0,3418
Sarcoma de células reticulares	0,0000	0,0000	0,0000	0,0253	0,1646	0,1899
Outras causas	0,0127	0,0380	0,0506	0,0759	0,2658	0,4430

Numericamente, temos a mesma conclusão do gráfico do item **a)**, em que o sarcoma tem sempre a menor representividade nos tempos e que há um revezamento entre o línfoma (até $t = 800$) e outras causas de morte.

c)

Agora, vamos obter a curva de Kaplan Meier da sobrevivência global dos dados:

```r
# CURVA DE KM GLOBAL NO R

ajuste_km_radiacao <- survfit(Surv(TEMPOS, rep(1, nrow(dados_ex3))) ~ 1,
                              data = dados_ex3)
ggsurvplot(ajuste_km_radiacao, pval = TRUE, conf.int = TRUE,
           xlab = "Tempo (em dias)", ylab = "Probabilidade de Sobrevivência",
           break.time.by = 100, ggtheme = theme_light(), legend = 'none', col = "blue")
```



A curva decai ao longo do tempo de maneira quase que constante, sendo que entre 100 e 200 há um decaimento mais brusco da sobrevivência estimada e, após, há um queda menos rápida no tempo.

Vamos calcular o valor da função de sobrevivência nos instantes $t = 200; 300; 400; 600; 800 e 1000$, considerando 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. Tomando como exemplo $t = 200$, temos que:

$$1 - (0,0633 + 0,0000 + 0,0127) = 0,9240$$

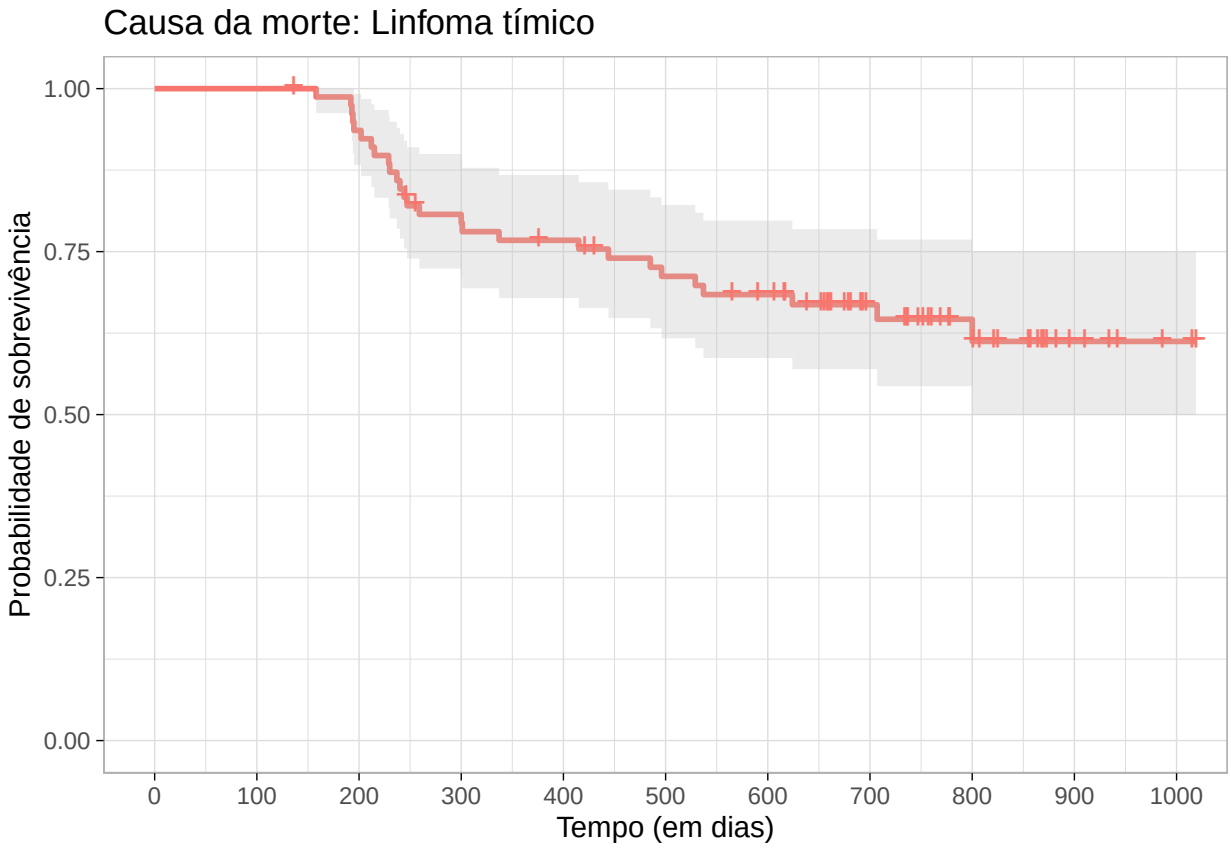
```
# CALCULANDO O VALOR DAS FUNÇÕES NOS INSTANTES NO R

estimativas_km_radiacao <- 1 - apply(timepoints(radiacao_incid_acum,
                                                c(200, 300, 400, 600, 800, 1000))$est, 2, sum)
estimativas_km_radiacao
```

```
##         200         300         400         600         800        1000
## 0.92405063 0.75949367 0.72151899 0.59493671 0.22784810 0.02531646
```

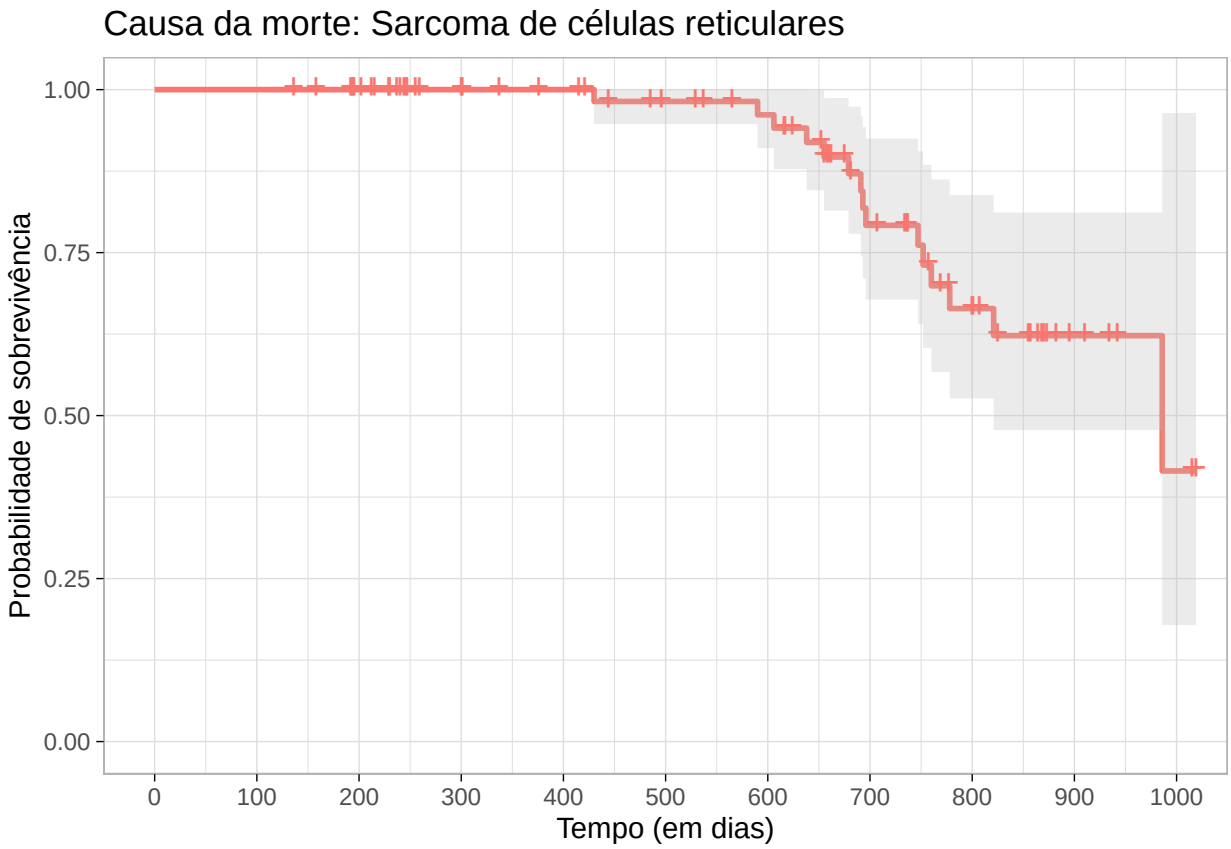
d)

Vamos obter 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:

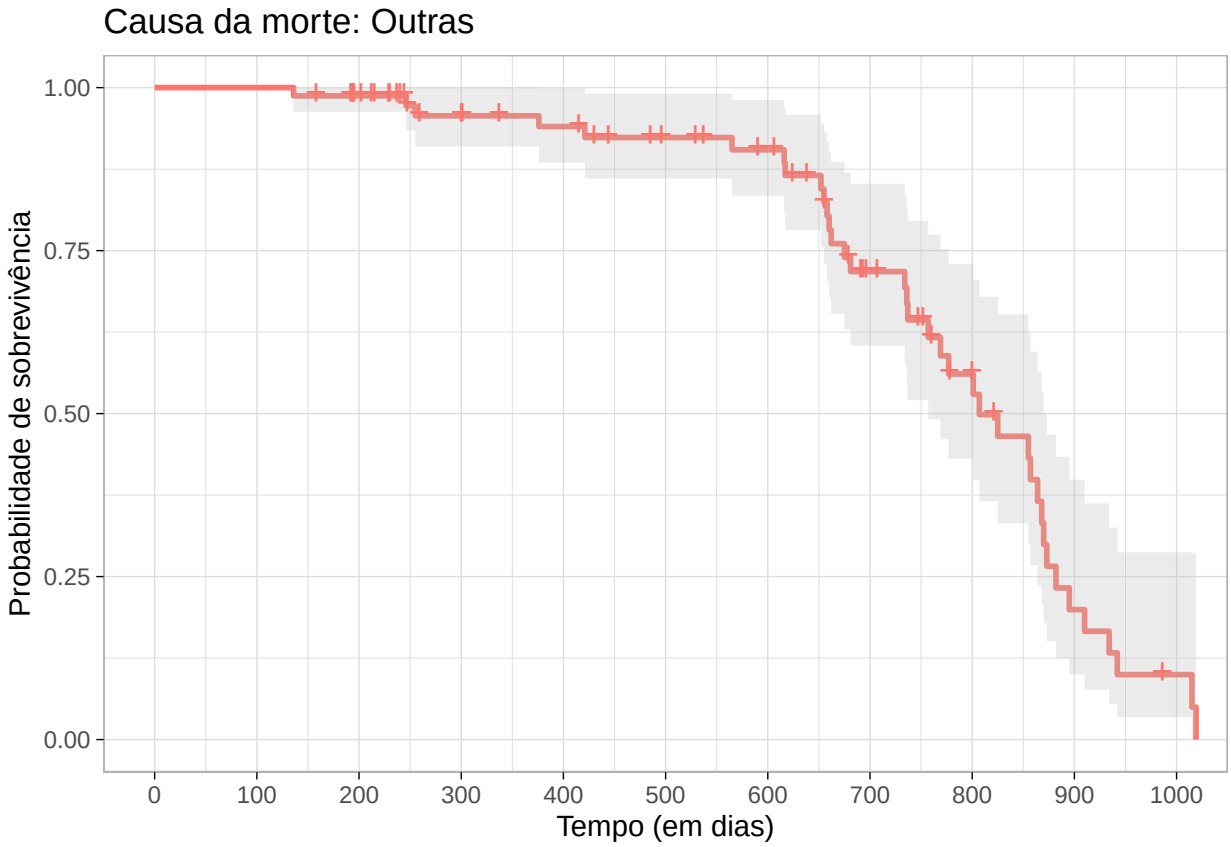
```r
# FAZENDO AS CURVAS DE KM NO R

dados_ex3$LINFOMA <- 1*(dados_ex3$CAUSA == "Linfoma tímico")
dados_ex3$SARCOMA <- 1*(dados_ex3$CAUSA == "Sarcoma de células reticulares")
dados_ex3$OUTRAS <- 1*(dados_ex3$CAUSA == "Outras causas")

ajuste_km_radiacao_linfoma <- survfit(Surv(TEMPOS, LINFOMA) ~ 1, data = dados_ex3)
ajuste_km_radiacao_sarcoma <- survfit(Surv(TEMPOS, SARCOMA) ~ 1, data = dados_ex3)
ajuste_km_radiacao_outras <- survfit(Surv(TEMPOS, OUTRAS) ~ 1, data = dados_ex3)

ggsurvplot(ajuste_km_radiacao_linfoma, pval = TRUE, conf.int = TRUE,
           xlab = "Tempo (em dias)", ylab = "Probabilidade de sobrevivência",
           break.time.by = 100, ggtheme = theme_light(), legend = "none",
           title = "Causa da morte: Linfoma tímico")
```




```r
ggsurvplot(ajuste_km_radiacao_sarcoma, pval = TRUE, conf.int = TRUE,
           xlab = "Tempo (em dias)", ylab = "Probabilidade de sobrevivência",
           break.time.by = 100, ggtheme = theme_light(), legend = "none",
           title = "Causa da morte: Sarcoma de células reticulares")
```




```
ggsurvplot(ajuste_km_radiacao_outras, pval = TRUE, conf.int = TRUE,
           xlab = "Tempo (em dias)", ylab = "Probabilidade de sobrevivência",
           break.time.by = 100, ggtheme = theme_light(), legend = "none",
           title = "Causa da morte: Outras")
```

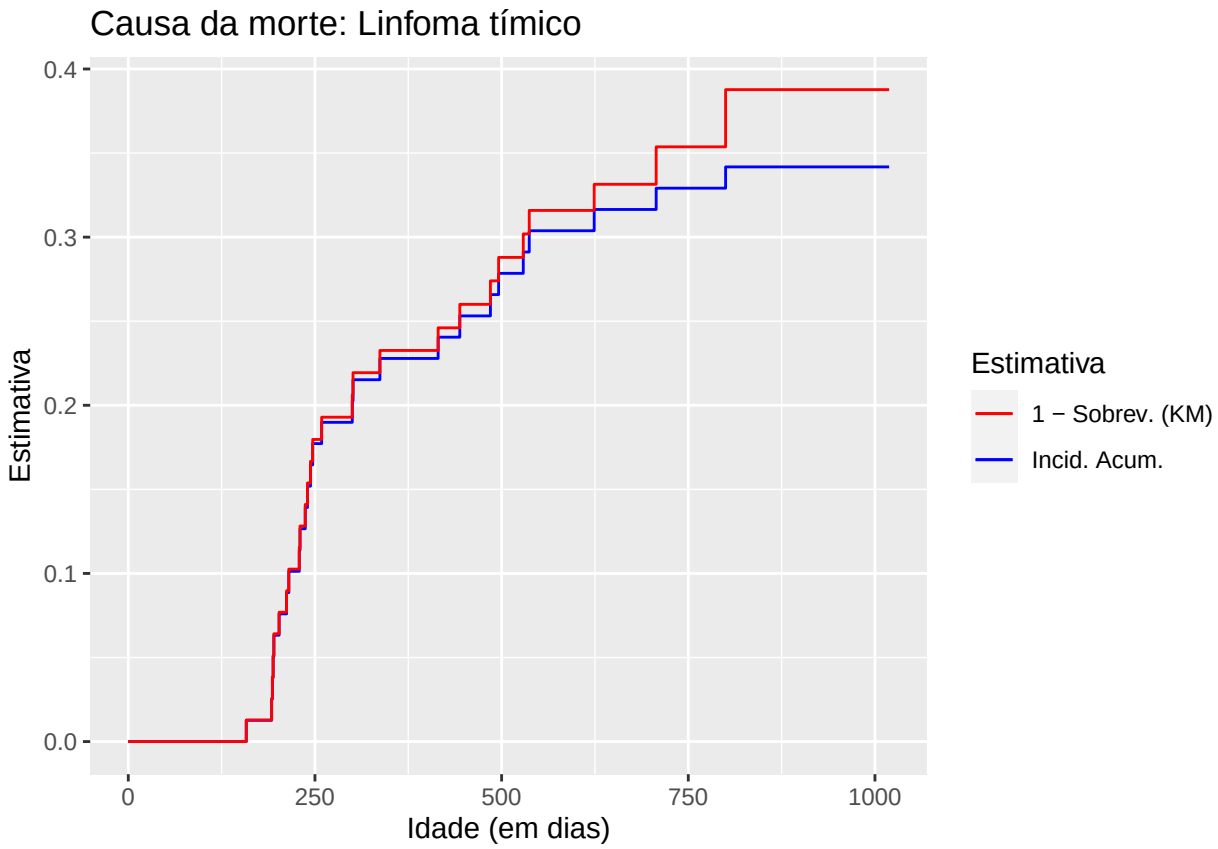


Colocamos nos mesmos gráficos as curvas de incidência acumulada e 1 menos a curva de Kaplan-Meier, sendo um gráfico para cada causa de morte:

```
# CURVAS DE INCIDÊNCIA NO R

incid_km_linfoma <- data.frame(tempo = radiacao_incid_acum$`1 Linfoma tímico`$time,
                               incid_acum = radiacao_incid_acum$`1 Linfoma tímico`$est) %>%
    group_by(tempo) %>%
    summarise(incid_acum = max(incid_acum)) %>%
    left_join(data.frame(tempo = ajuste_km_radiacao_linfoma$time,
                         km = ajuste_km_radiacao_linfoma$surv), by = "tempo")
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

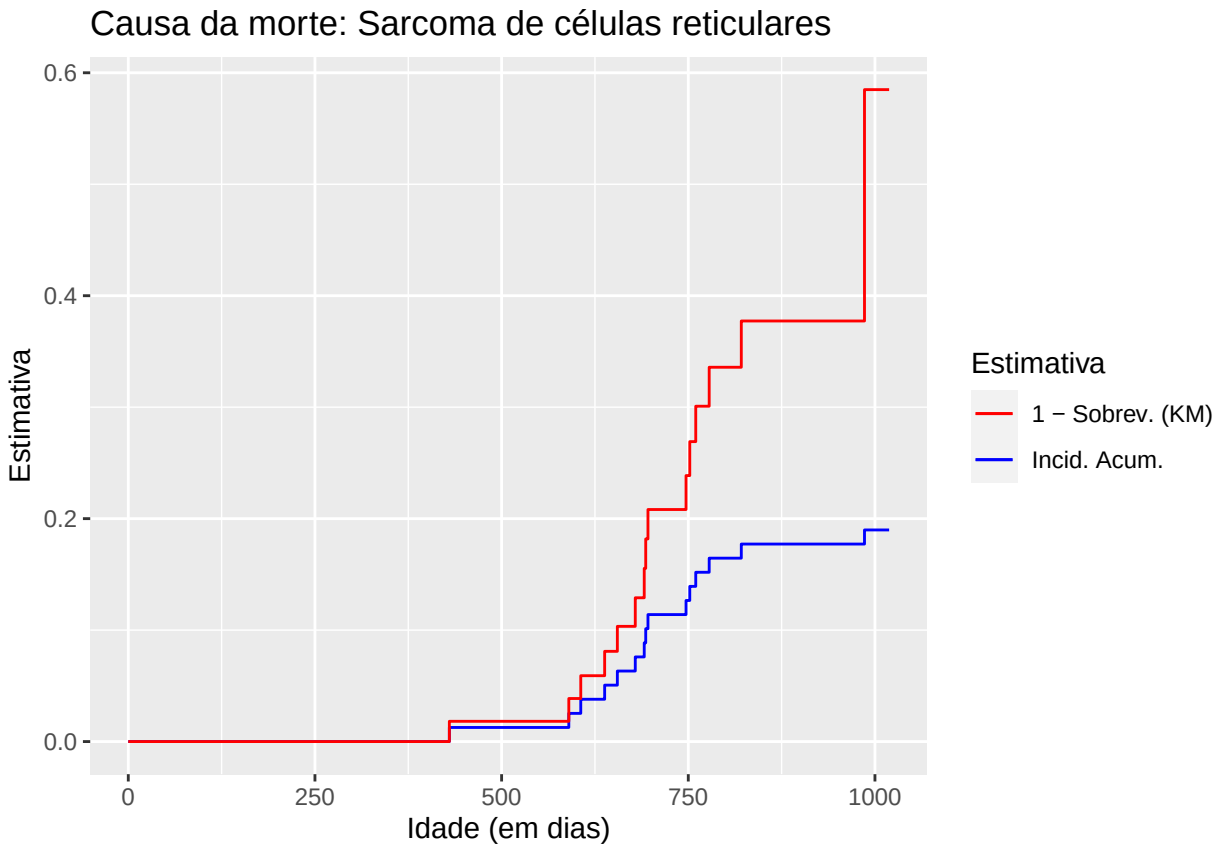
```
incid_km_linfoma[1, 3] <- 1

ggplot(incid_km_linfoma) +
   geom_step(aes(x = tempo, y = incid_acum, col = 'Incid. Acum.')) +
   geom_step(aes(x = tempo, y = 1-km, col = '1 - Sobrev. (KM)')) +
   labs(x = 'Idade (em dias)', y = 'Estimativa', title = 'Causa da morte: Linfoma tímico') +
   scale_color_manual(name = 'Estimativa', values = c('Incid. Acum.' = 'blue', '1 - Sobrev. (KM)' = 'red'))
```




```
incid_km_sarcoma <- data.frame(tempo = radiacao_incid_acum$`1 Sarcoma de células reticulares`$time,
                               incid_acum = radiacao_incid_acum$`1 Sarcoma de células reticulares`$est) %>%
   group_by(tempo) %>%
   summarise(incid_acum = max(incid_acum)) %>%
   left_join(data.frame(tempo = ajuste_km_radiacao_sarcoma$time,
                        km = ajuste_km_radiacao_sarcoma$surv), by = 'tempo')
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

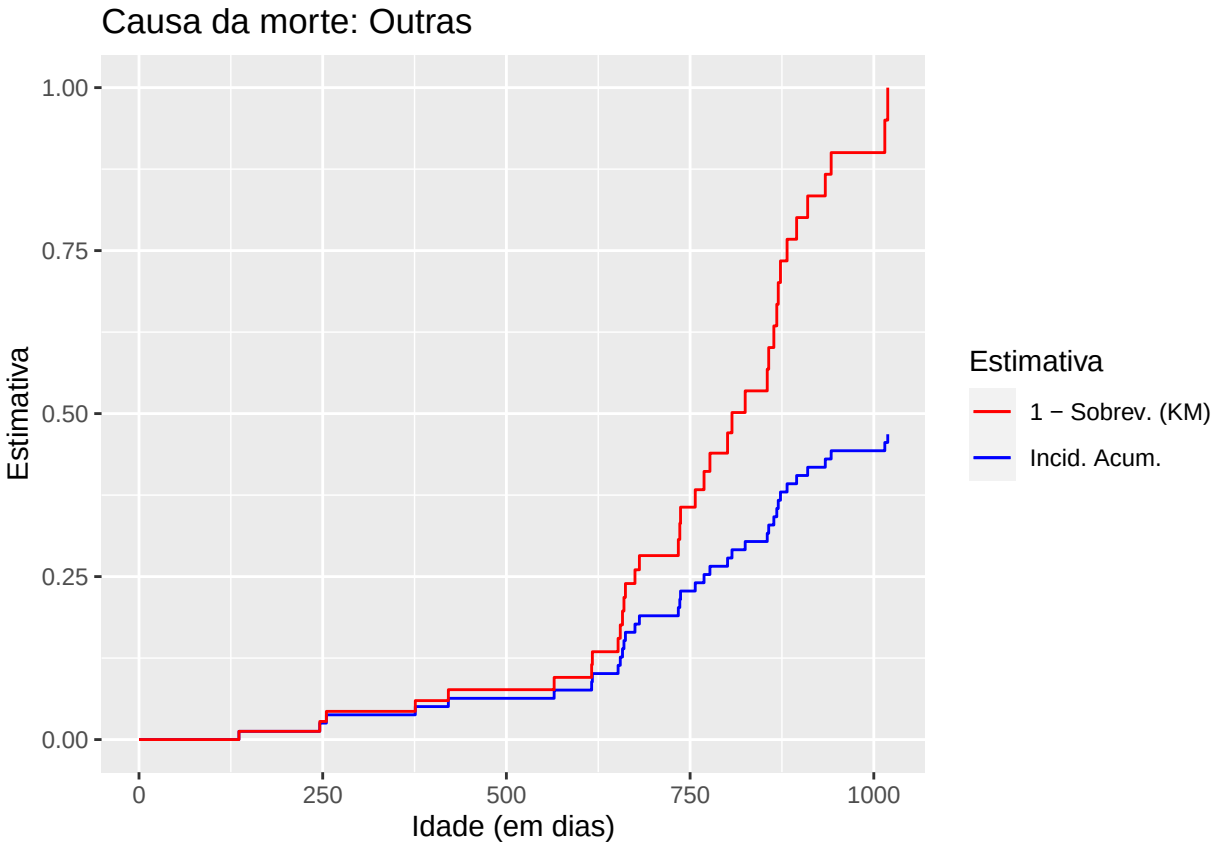
```
incid_km_sarcoma[1,3] <- 1

ggplot(incid_km_sarcoma) +
   geom_step(aes(x = tempo, y = incid_acum, col = 'Incid. Acum.')) +
   geom_step(aes(x = tempo, y = 1-km, col = '1 - Sobrev. (KM)')) +
   labs(x = 'Idade (em dias)', y = 'Estimativa', title = 'Causa da morte: Sarcoma de células reticulares') +
   scale_color_manual(name = 'Estimativa', values = c('Incid. Acum.' = 'blue', '1 - Sobrev. (KM)' = 'red'))
```




```
incid_km_outras <- data.frame(tempo = radiacao_incid_acum$`1 Outras causas`$time,
                              incid_acum = radiacao_incid_acum$`1 Outras causas`$est) %>%
   group_by(tempo) %>%
   summarise(incid_acum = max(incid_acum)) %>%
   left_join(data.frame(tempo = ajuste_km_radiacao_outras$time,
                        km = ajuste_km_radiacao_outras$surv), by = 'tempo')
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
incid_km_outras[1,3] <- 1

ggplot(incid_km_outras) +
   geom_step(aes(x = tempo, y = incid_acum, col = 'Incid. Acum.')) +
   geom_step(aes(x = tempo, y = 1-km, col = '1 - Sobrev. (KM)')) +
   labs(x = 'Idade (em dias)', y = 'Estimativa', title = 'Causa da morte: Outras') +
   scale_color_manual(name = 'Estimativa',
                      values = c('Incid. Acum.' = 'blue', '1 - Sobrev. (KM)' = 'red'))
```



Analisando os gráficos, vemos que a probabilidade de sobrevivência complementar estimada pelo método de Kaplan-Meier é, em geral, maior que a incidência acumulada na presença de riscos competitivos. A diferença principal é que o estimador de KM leva em conta somente as falhas pelo evento de interesse, tratando como censura não informativa as demais falhas por eventos que, na realidade, podem ser os competitivos. Isso fica mais evidente conforme aumentam o tempo e a quantidade de eventos competitivos.