# Análise de Sobrevivência

0.4 - Aula Prática

Prof. Dr. Eder Angelo Milani

05/04/2025

## Leitura dos dados

As linhas de cógido a seguir executam as seguintes tarefas:

- leitura dos dados filtrados para CID C34;
- breve visualização do conjunto de dados;
- modificar a escala de tempo para ano, em vez de dia;
- excluir os tempos iguis a zero.

```r
# limpando o que tem na memoria
rm(list=ls())

# local onde esta o arquivo com os dados
setwd("G:\\Meu Drive\\UFG\\Especializacao\\Aulas Análise Sobrevivência\\Códigos")

### leitura
dados <- read.csv("cancer_c34.csv")
head(dados)
```

```
##   TOPOGRUP TEMPO CENSURA ANODIAG IDADE SEXO CIRURGIA RADIO QUIMIO ECGRUP
## 1      C34   292       1    2014    63    1        0     1      1    III
## 2      C34   132       1    2016    58    2        0     0      0      I
## 3      C34     3       0    2016    61    2        0     0      0     IV
## 4      C34    17       1    2016    67    1        0     0      0     IV
## 5      C34   182       1    2015    57    1        0     0      1    III
## 6      C34   287       1    2015    69    1        0     0      1     IV
```

```r
# mudança na variável tempo - de dias para ano
dados$TEMPO <- dados$TEMPO/365

# excluir os tempos iguais a zero
ind_tempo_zero <- which(dados$TEMPO == 0)
ind_tempo_zero
```

```
##  [1]  256  297  322  374  865  996 1010 1049 1083 1165 1514 1665 1754 2090 5830
## [16] 6012 6196 7079 8030 8049 8383 8495
```

```r
dados$TEMPO[256]
```

```
## [1] 0
```

```r
dados <- dados[-ind_tempo_zero,]

# outra forma de filtrar
# dados <- dados %>% filter(TEMPO != 0)
```

## Ajuste de modelos paramétricos

As linhas de cógido a seguir executam as seguintes tarefas:

- ajusta o modelo exponencial ao conjunto de dados;
- obtenção da estimativa do parâmetro $\alpha$ do modelo exponencial;
- ajusta o modelo Weibull ao conjunto de dados;
- obtenção da estimativa dos parâmetros $(\alpha, \gamma)$ do modelo Weibull;
- ajusta o modelo log-normal ao conjunto de dados;
- obtenção da estimativa dos parâmetros $(\mu, \sigma)$ do modelo log-normal.

```r
require(survival)
```

```
## Carregando pacotes exigidos: survival
```

```r
# modelo exponencial
ajuste_exp <- survreg(Surv(TEMPO, CENSURA) ~ 1,
                      dist='exponential', data = dados)

ajuste_exp
```

```
## Call:
## survreg(formula = Surv(TEMPO, CENSURA) ~ 1, data = dados, dist = "exponential")
##
## Coefficients:
## (Intercept)
##   0.6576528
##
## Scale fixed at 1
##
## Loglik(model)= -11424.5   Loglik(intercept only)= -11424.5
## n= 8909
```

```r
alpha <- exp(ajuste_exp$coefficients[1])

cat("a estimativa do parâmetro alpha é ", alpha, "\n")
```

```
## a estimativa do parâmetro alpha é  1.930256
```

```r
# modelo Weibull
ajuste_w <- survreg(Surv(TEMPO, CENSURA)~1,
                    dist='weibull', data = dados)

ajuste_w
```

```
## Call:
## survreg(formula = Surv(TEMPO, CENSURA) ~ 1, data = dados, dist = "weibull")
##
## Coefficients:
```

```
## (Intercept)
##   0.5454752
##
## Scale= 1.524059
##
## Loglik(model)= -10251.9   Loglik(intercept only)= -10251.9
## n= 8909
```

```r
alpha_w <- exp(ajuste_w$coefficients[1])
gama_w <- 1/ajuste_w$scale

# obs.: o R utiliza uma parametrizacao diferente da adotada no livro que estamos utilizando

cat("a estimativa dos parâmetro alpha e gama são ", alpha_w, ", ", gama_w, "\n")
```

```
## a estimativa dos parâmetro alpha e gama são  1.725428 ,  0.6561427
```

```r
# modelo lognormal

ajuste_ln <- survreg(Surv(TEMPO, CENSURA)~1,
                     dist='lognorm', data = dados)

ajuste_ln
```

```
## Call:
## survreg(formula = Surv(TEMPO, CENSURA) ~ 1, data = dados, dist = "lognorm")
##
## Coefficients:
## (Intercept)
##  -0.2013292
##
## Scale= 1.755354
##
## Loglik(model)= -9674   Loglik(intercept only)= -9674
## n= 8909
```

```r
mu <- ajuste_ln$coefficients[1]
sigma <- ajuste_ln$scale

cat("a estimativa dos parâmetro mu e sigma são ", mu, ", ", sigma, "\n")
```

```
## a estimativa dos parâmetro mu e sigma são  -0.2013292 ,  1.755354
```

# Adequação do modelo probabilístico

Para verificar a adequação dos modelos probabilísticos apresentamos a equação da função de sobrevivência estimada para cada um dos três modelos, que são:

- modelo exponencial

$$\widehat{S}_e(t) = \exp(-t/1.9302562)$$

- modelo Weibull

$$\widehat{S}_W(t) = \exp[-(t/1.7254281)^{0.6561427}]$$

- modelo log-normal

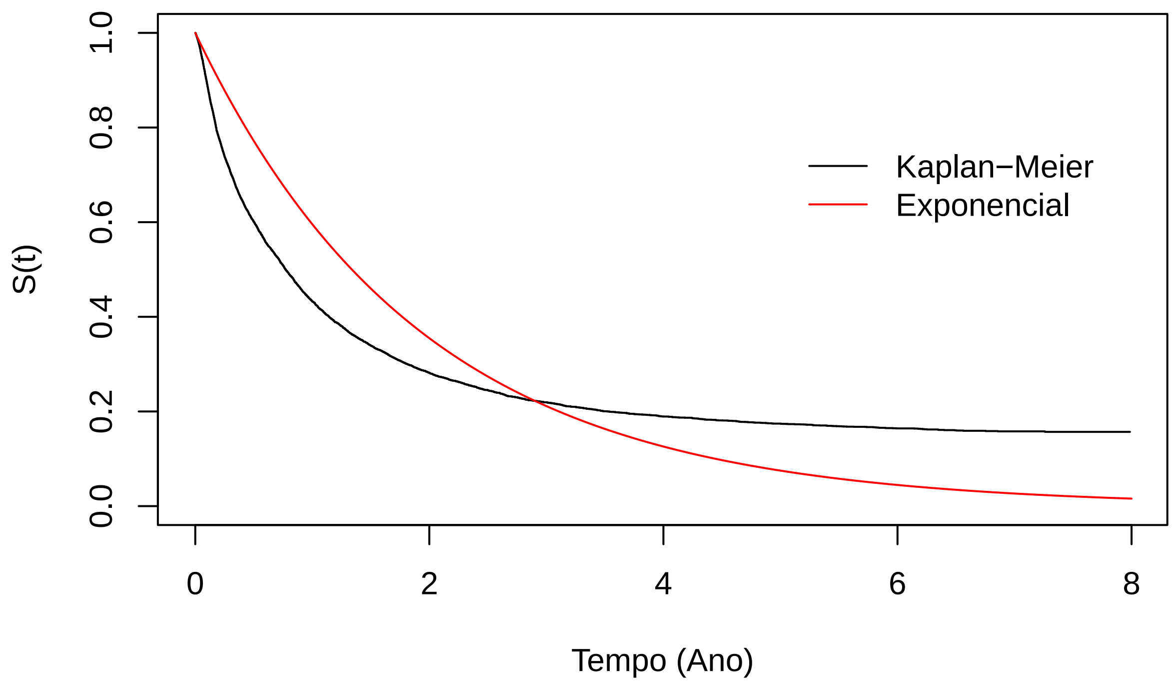
$$\widehat{S}_{ln}(t) = \Phi\Big(\frac{-log(t) - 0.2013292}{1.7553535}\Big)$$

As linhas de cógido a seguir executam as seguintes tarefas:

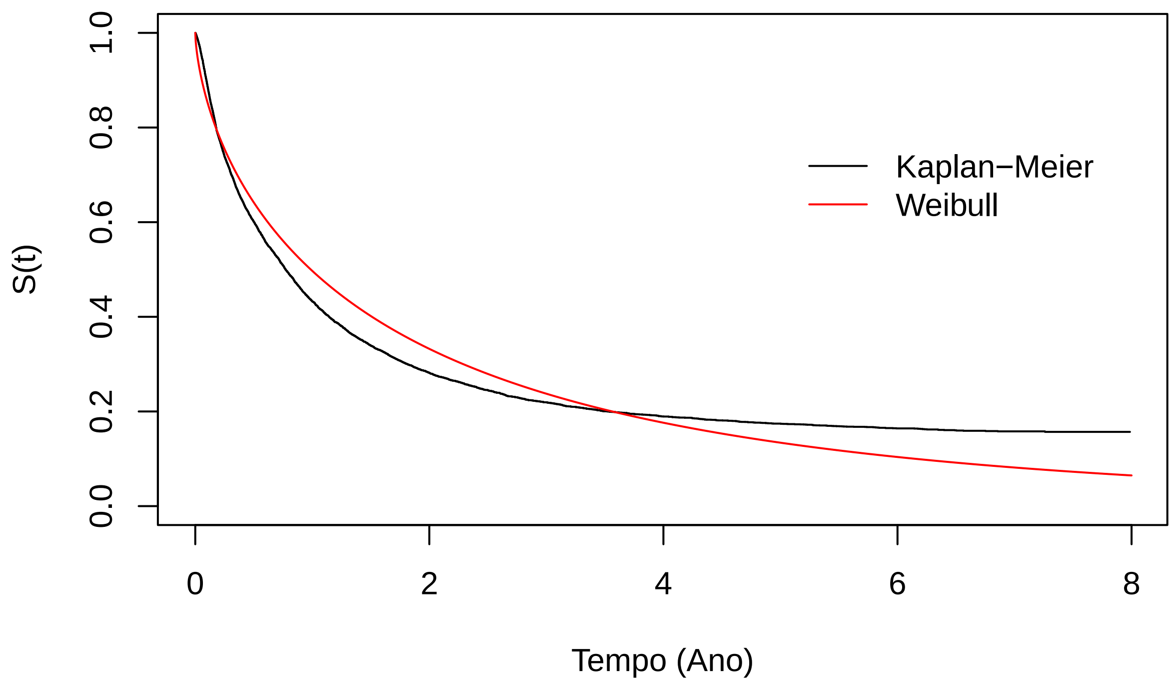
- estima a função de sobrevivência utilizando o estimador de Kaplan-Meier;
- estima a função de sobrevivência a partir do modelos exponencial, Weibull e log-normal;
- elabora os gráfico $\widehat{S}(t)$ versus $t$, juntamente com a curva $\widehat{S}_{Exp}(t)$ versus $t$. E para os modelos Weibull e log-normal
- elabora o gráfico $\widehat{S}(t)$ versus $\widehat{S}_{Exp}(t)$, e para os modelos Weibull e log-normal

Comparando os valores da sobrevivência estimado pelo Kaplan-Meier e os três modelos definidos anteriormente.

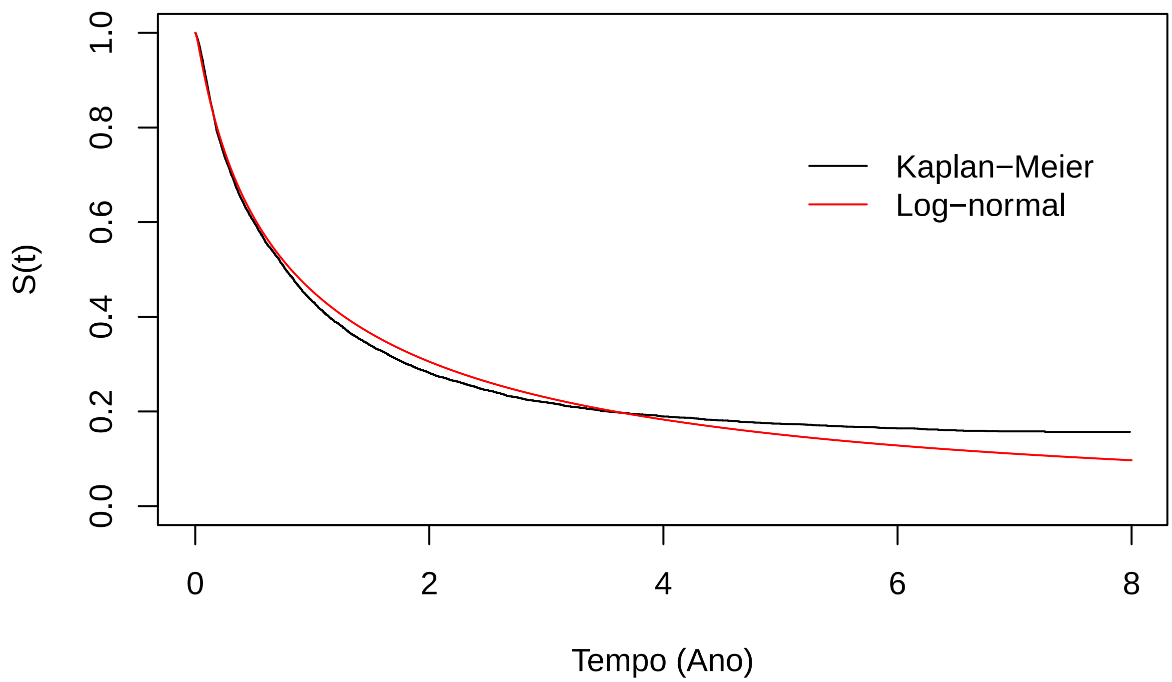
```r
# funcao de sobrevivência estimada via Kaplan-Meier
ekm <- survfit(Surv(TEMPO, CENSURA) ~ 1, data = dados)

# criando um vetor de tempo onde as funcoes de sobrevivencia serao calculadas
t <- seq(0, 8, length.out = length(ekm$time))

# funcao de sobrevivência estimada pelo modelo exponencial
s_exp <- exp(-t/alpha)

# funcao de sobrevivência estimada pelo modelo weibull
s_w <- exp(-(t/alpha_w)^gama_w)

# funcao de sobrevivência estimada pelo modelo log-normal
s_ln <- pnorm((-log(t)+mu)/sigma)

# graficos

# para o modelo exponencial

plot(ekm, lty=1, xlab= "Tempo (Ano)", ylab="S(t)", mark.time = F,  conf.int = F)
lines(t, s_exp, col="red")
legend(5, 0.8, lty=c(1,1), c("Kaplan-Meier", "Exponencial"), col=c("black", "red"), bty="n")
```



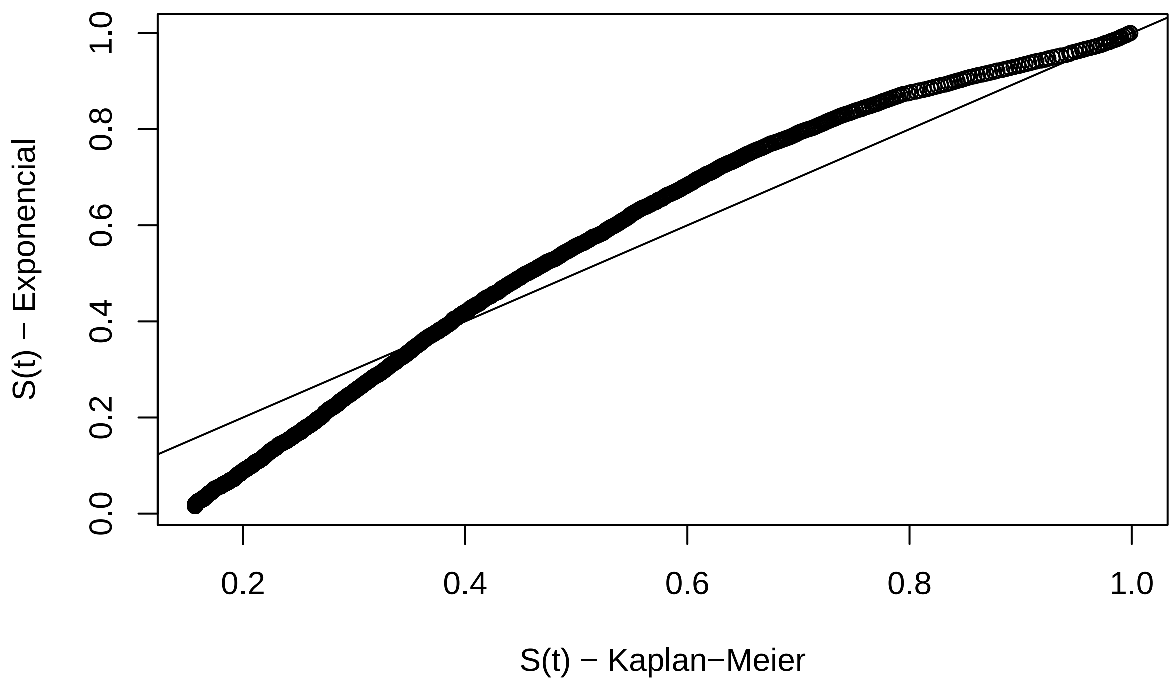

```r
# para o modelo weibull

plot(ekm, lty=1, xlab= "Tempo (Ano)", ylab="S(t)", mark.time = F,  conf.int = F)
lines(t, s_w, col="red")
legend(5, 0.8, lty=c(1,1), c("Kaplan-Meier", "Weibull"), col=c("black", "red"), bty="n")
```



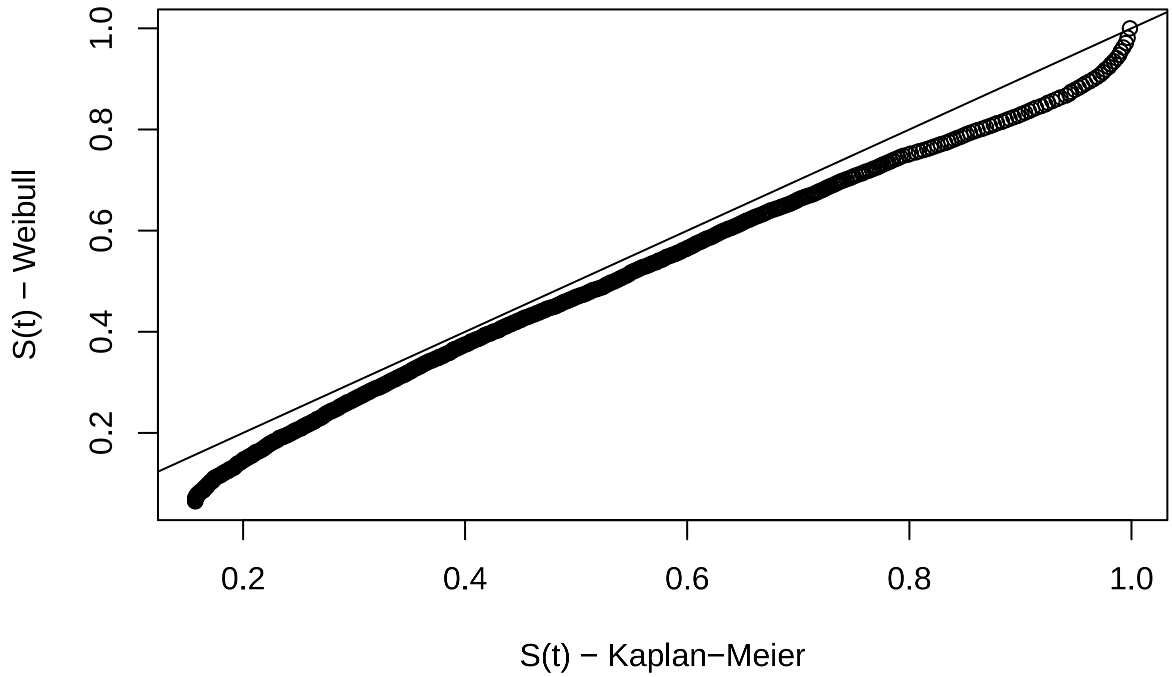

```r
# para o modelo log-normal

plot(ekm, lty=1, xlab= "Tempo (Ano)", ylab="S(t)", mark.time = F,  conf.int = F)
lines(t, s_ln, col="red")
legend(5, 0.8, lty=c(1,1), c("Kaplan-Meier", "Log-normal"), col=c("black", "red"), bty="n")
```



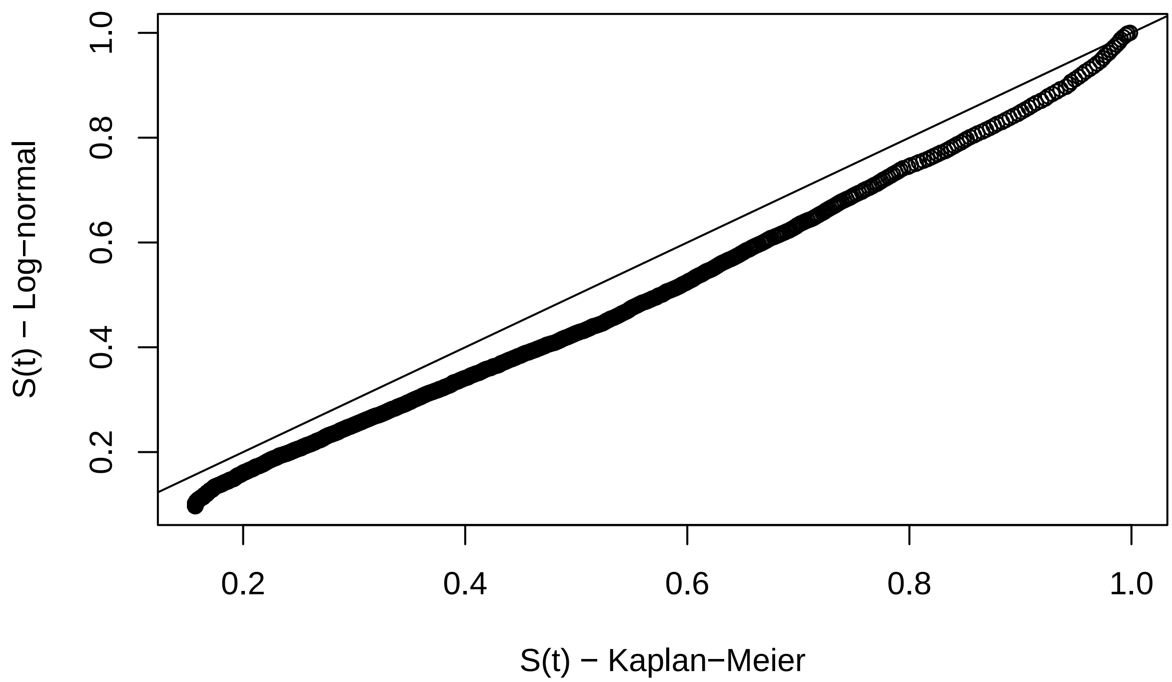

```r
# outro modelo de grafico

# para o modelo exponencial

plot(ekm$surv, s_exp, lty=1, xlab= "S(t) - Kaplan-Meier", ylab="S(t) - Exponencial")
abline(a=0, b=1)
```




```
# para o modelo Weibull

plot(ekm$surv, s_w, lty=1, xlab= "S(t) - Kaplan-Meier", ylab="S(t) - Weibull")
abline(a=0, b=1)
```




```
# para o modelo log-normal

plot(ekm$surv, s_ln, lty=1, xlab= "S(t) - Kaplan-Meier", ylab="S(t) - Log-normal")
abline(a=0, b=1)
```




## Modelo de regressão paramétrico

As linhas de cógido a seguir executam as seguintes tarefas:

- modifica a variável SEXO de 1-Masculino e 2-Feminino para 0-Masculino e 1-Feminino
- estima os parâmetros $\beta_0$ e $\beta_1$ dos modelos exponencial, Weibull e log-normal, considerando a variável SEXO como explicativa para o evento de interesse

```
require(survival)

table(dados$SEXO)
```

```
##
##    1    2
## 5042 3867
```

```
dados$SEXO <- dados$SEXO -1

table(dados$SEXO)
```

```
##
##    0    1
## 5042 3867
```

```
# modelo exponencial
ajuste_exp <- survreg(Surv(TEMPO, CENSURA) ~ SEXO,
                      dist='exponential', data = dados)
```

```r
ajuste_exp
```

```
## Call:
## survreg(formula = Surv(TEMPO, CENSURA) ~ SEXO, data = dados,
##     dist = "exponential")
##
## Coefficients:
## (Intercept)        SEXO
##   0.4691185   0.4104071
##
## Scale fixed at 1
##
## Loglik(model)= -11281.7   Loglik(intercept only)= -11424.5
##  Chisq= 285.7 on 1 degrees of freedom, p= <2e-16
## n= 8909
```

```r
alpha_0 <- exp(ajuste_exp$coefficients[1])
alpha_1 <- exp(ajuste_exp$coefficients[1] + ajuste_exp$coefficients[2])

cat("a estimativa do parâmetro alpha é ", c(alpha_0, alpha_1), "\n")
```

```
## a estimativa do parâmetro alpha é  1.598584 2.409756
```

```r
# modelo Weibull
ajuste_w <- survreg(Surv(TEMPO, CENSURA) ~ SEXO,
                    dist='weibull', data = dados)

ajuste_w
```

```
## Call:
## survreg(formula = Surv(TEMPO, CENSURA) ~ SEXO, data = dados,
##     dist = "weibull")
##
## Coefficients:
## (Intercept)        SEXO
##   0.3304047   0.4828715
##
## Scale= 1.508992
##
## Loglik(model)= -10165.9   Loglik(intercept only)= -10251.9
##  Chisq= 172.14 on 1 degrees of freedom, p= <2e-16
## n= 8909
```

```r
alpha_w_0 <- exp(ajuste_w$coefficients[1])
alpha_w_1 <- exp(ajuste_w$coefficients[1] + ajuste_w$coefficients[2])
gama_w <- 1/ajuste_w$scale

# obs.: o R utiliza uma parametrizacao diferente da adotada no livro que estamos utilizando

cat("a estimativa dos parâmetro alpha e gama são ", c(alpha_w_0, alpha_w_1, gama_w), "\n")
```

```
## a estimativa dos parâmetro alpha e gama são  1.391531 2.255285 0.6626941
```

```r
# modelo lognormal

ajuste_ln <- survreg(Surv(TEMPO, CENSURA) ~ SEXO,
```

```r
                     dist='lognorm', data = dados)

ajuste_ln
```

```
## Call:
## survreg(formula = Surv(TEMPO, CENSURA) ~ SEXO, data = dados,
##     dist = "lognorm")
##
## Coefficients:
## (Intercept)        SEXO
##  -0.3811368   0.4136815
##
## Scale= 1.744013
##
## Loglik(model)= -9616.7   Loglik(intercept only)= -9674
##  Chisq= 114.52 on 1 degrees of freedom, p= <2e-16
## n= 8909
```

```r
mu_0 <- ajuste_ln$coefficients[1]
mu_1 <- ajuste_ln$coefficients[1] + ajuste_ln$coefficients[2]
sigma <- ajuste_ln$scale

cat("a estimativa dos parâmetro mu e sigma são ", c(mu_0, mu_1, sigma), "\n")
```

```
## a estimativa dos parâmetro mu e sigma são  -0.3811368 0.03254469 1.744013
```

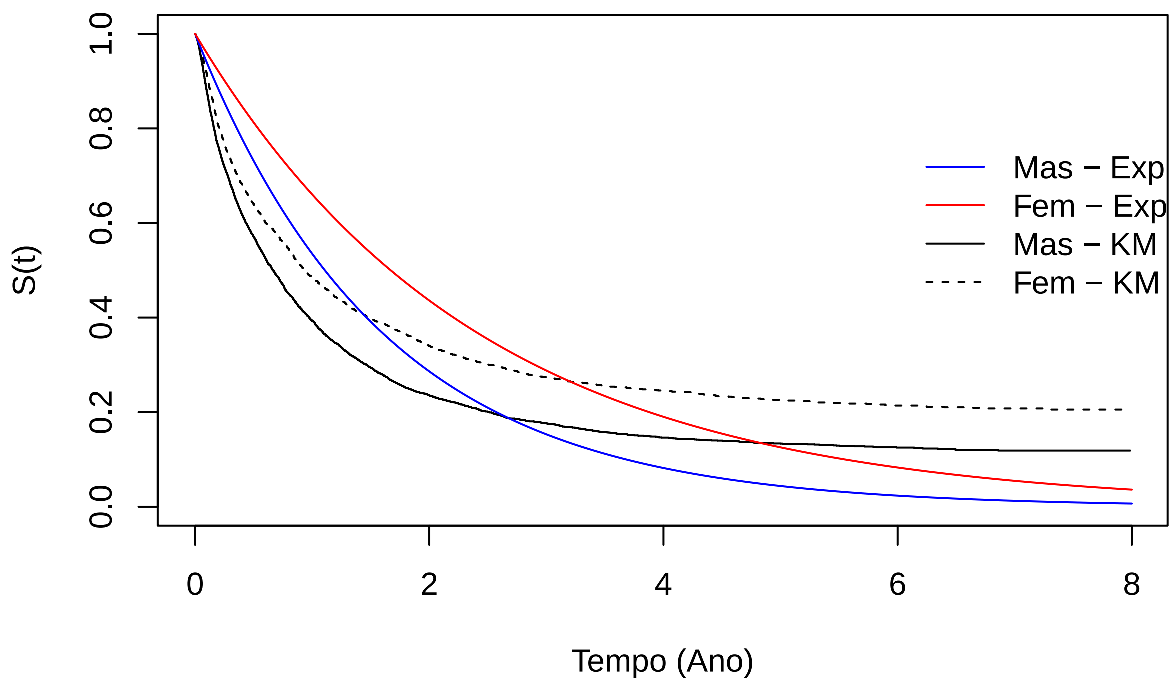
# Adequação do modelo ajustado

As linhas de cógido a seguir executam as seguintes tarefas:

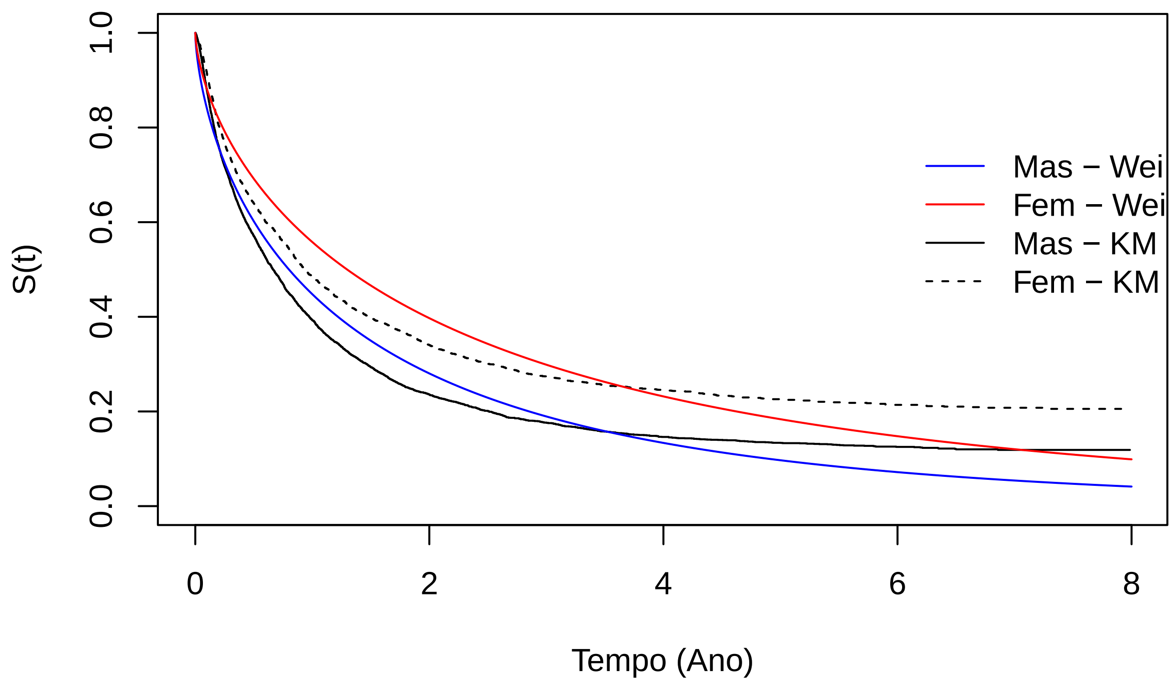
- calcula a função de sobrevivência estimada para ambos os sexos e considerando os três modelos
- elabora o gráfico $\widehat{S}(t)$ versus $t$, juntamente com a curva $\widehat{S}_{Exp}(t)$ versus $t$, para ambos os sexos, e para os modelos Weibull e log-normal
- calcula o resíduo de Cox-Snell para os três modelos

```r
# funcao de sobrevivência estimada via Kaplan-Meier
ekm <- survfit(Surv(TEMPO, CENSURA) ~ SEXO, data = dados)

# criando um vetor de tempo onde as funcoes de sobrevivencia serao calculadas
t <- seq(0, 8, length.out = length(ekm$time))

# funcao de sobrevivência estimada pelo modelo exponencial
s_exp_0 <- exp(-t/alpha_0) # para o sexo masculino
s_exp_1 <- exp(-t/alpha_1) # para o sexo feminino

# funcao de sobrevivência estimada pelo modelo weibull
s_w_0 <- exp(-(t/alpha_w_0)^gama_w) # para o sexo masculino
s_w_1 <- exp(-(t/alpha_w_1)^gama_w) # para o sexo feminino

# funcao de sobrevivência estimada pelo modelo log-normal
s_ln_0 <- pnorm(-(log(t)-mu_0)/sigma) # para o sexo masculino
s_ln_1 <- pnorm(-(log(t)-mu_1)/sigma) # para o sexo feminino

# graficos
```

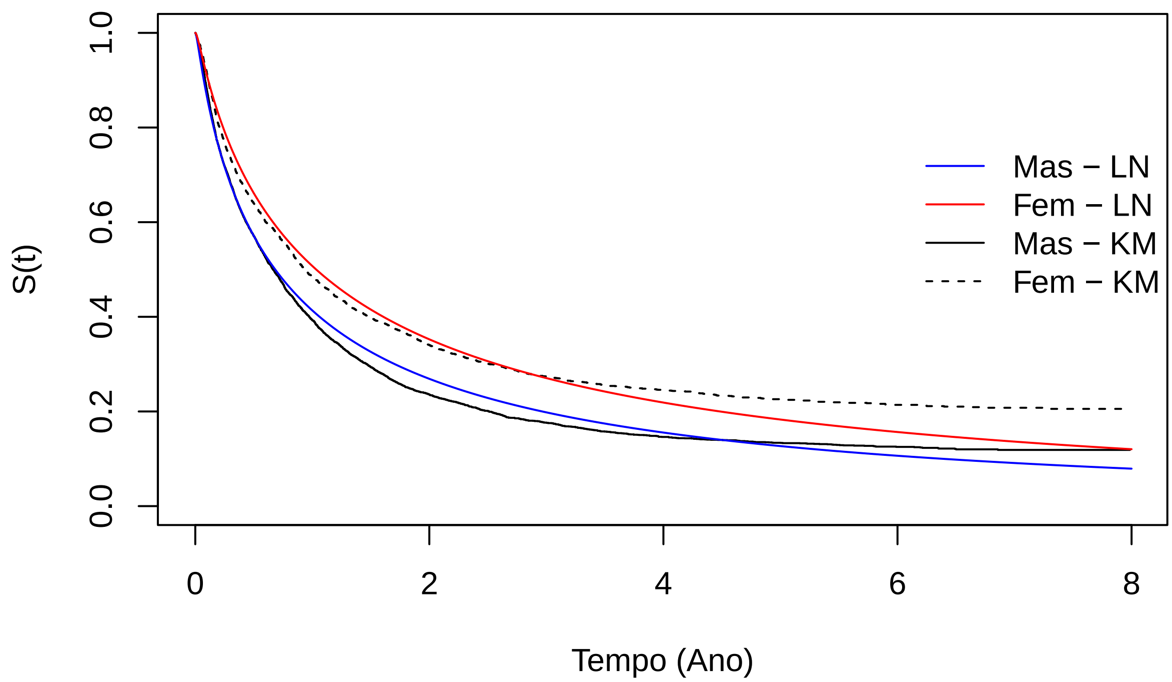
```r
# para o modelo exponencial

plot(ekm, lty=c(1, 2), xlab= "Tempo (Ano)", ylab="S(t)", mark.time = F,  conf.int = F)
lines(t, s_exp_0, col="blue")
lines(t, s_exp_1, col="red")
legend(6, 0.8, col=c("blue", "red", "black", "black"),
       c("Mas - Exp", "Fem - Exp", "Mas - KM", "Fem - KM"), lty=c(1,1, 1, 2), bty ="n")
```




```r
# para o modelo weibull

plot(ekm, lty=c(1, 2), xlab= "Tempo (Ano)", ylab="S(t)", mark.time = F,  conf.int = F)
lines(t, s_w_0, col="blue")
lines(t, s_w_1, col="red")
legend(6, 0.8, col=c("blue", "red", "black", "black"),
       c("Mas - Wei", "Fem - Wei", "Mas - KM", "Fem - KM"), lty=c(1,1, 1, 2), bty ="n")
```



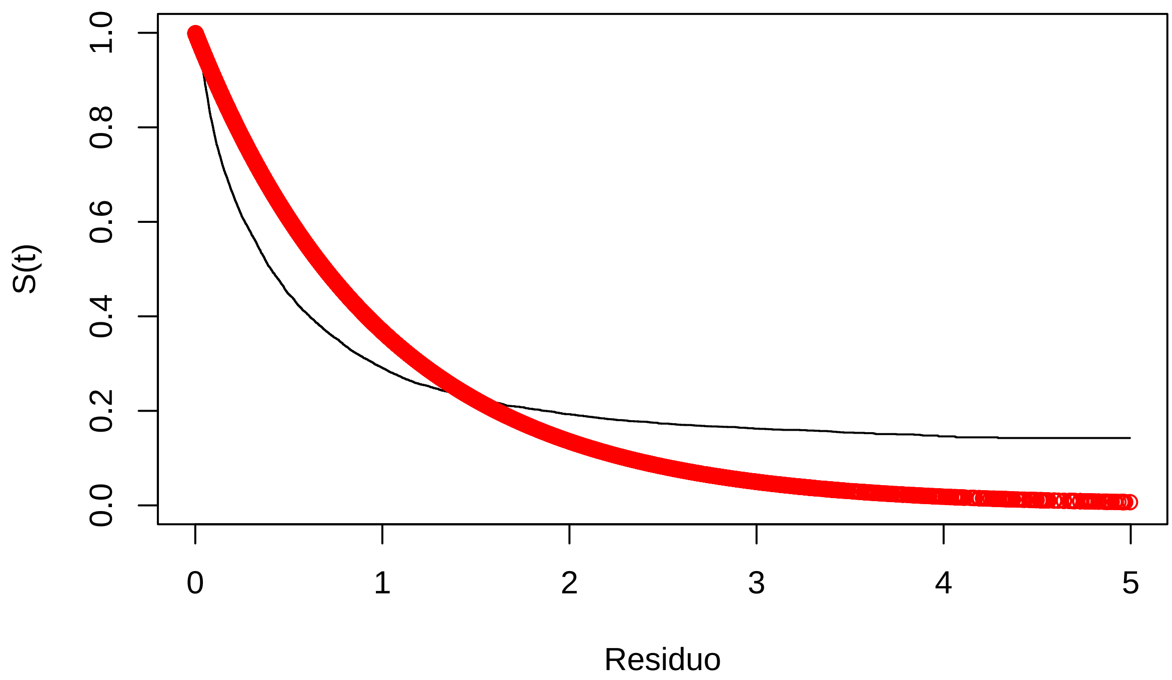

```
# para o modelo log-normal

plot(ekm, lty=c(1, 2), xlab= "Tempo (Ano)", ylab="S(t)", mark.time = F,  conf.int = F)
lines(t, s_ln_0, col="blue")
lines(t, s_ln_1, col="red")
legend(6, 0.8, col=c("blue", "red", "black", "black"),
       c("Mas - LN", "Fem - LN", "Mas - KM", "Fem - KM"), lty=c(1,1, 1, 2), bty ="n")
```




```r
# residuos de Cox_Snell

# modelo exponencial

residuo_exp <- dados$TEMPO*exp(-ajuste_exp$coefficients[1]
                          - ajuste_exp$coefficients[2]*dados$SEXO)

summary(residuo_exp)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.001137 0.114828 0.366092 0.773600 0.988889 4.995858
```

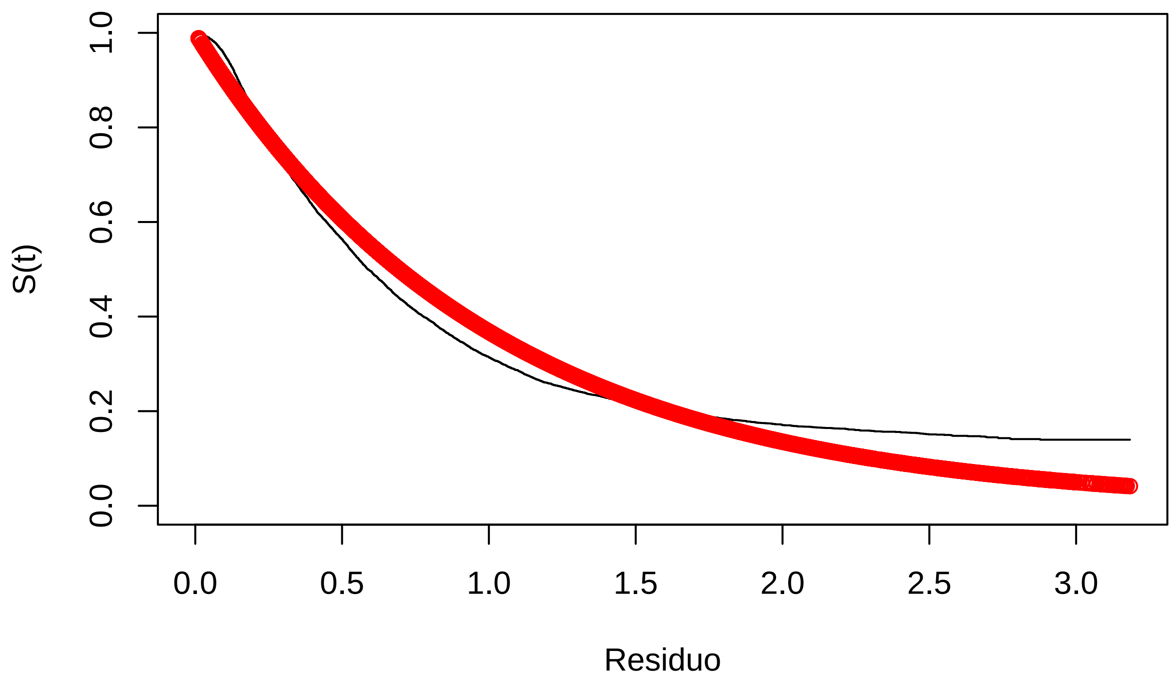
```r
ekm_residuo_exp <- survfit(Surv(residuo_exp, dados$CENSURA) ~ 1)

# grafico de curvas de sobrevivencia sobrepostas
plot(ekm_residuo_exp, lty=1, xlab= "Residuo", ylab="S(t)", mark.time = F,  conf.int = F)
points(residuo_exp, exp(-residuo_exp), col="red")
```




```r
# exp(-residuo_exp) e o calculo da sobrevivencia nos residuos de Cox-Snell


# modelo Weibull

residuo_w <- (dados$TEMPO*exp(-ajuste_w$coefficients[1]
                            - ajuste_w$coefficients[2]*dados$SEXO))^(1/ajuste_w$scale)

summary(residuo_w)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01169 0.25709 0.54812 0.77360 1.06684 3.18337
```

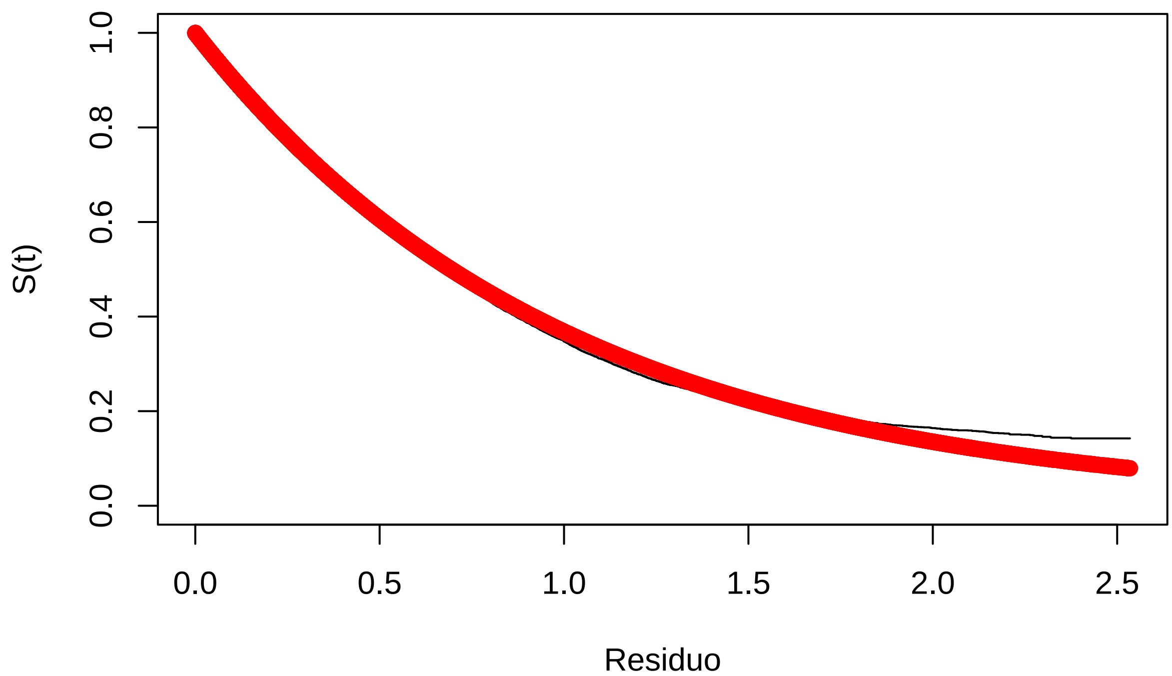
```r
ekm_residuo_w <- survfit(Surv(residuo_w, dados$CENSURA) ~ 1)

# grafico de curvas de sobrevivencia sobrepostas
plot(ekm_residuo_w, lty=1, xlab= "Residuo", ylab="S(t)", mark.time = F,  conf.int = F)
points(residuo_w, exp(-residuo_w), col="red")
```




```r
# exp(-residuo_w) e o calculo da sobrevivencia nos residuos de Cox-Snell


# modelo log-normal

residuo_ln <- -log(1 - pnorm((log(dados$TEMPO) - ajuste_ln$coefficients[1]
                           - ajuste_ln$coefficients[2]*dados$SEXO) / ajuste_ln$scale))

summary(residuo_ln)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000335 0.255644 0.623673 0.778851 1.152542 2.534690
```

```r
ekm_residuo_ln <- survfit(Surv(residuo_ln, dados$CENSURA) ~ 1)

# grafico de curvas de sobrevivencia sobrepostas
plot(ekm_residuo_ln, lty=1, xlab= "Residuo", ylab="S(t)", mark.time = F,  conf.int = F)
points(residuo_ln, exp(-residuo_ln), col="red")
```




```
# exp(-residuo_ln) e o calculo da sobrevivencia nos residuos de Cox-Snell
```

## Interpretação

A interpretação pode ser feita utilizando o tempo mediano ou o cálculo da função de sobrevivência em diferentes tempos. As linhas de códigos a seguir apresenta as duas formas de interpretação dos resultados.

A razão dos tempos medianos de dois pacientes, do sexo feminino e do masculino, é dada por

$$\frac{t_{0,5}(\boldsymbol{x} = \mathbf{1})}{t_{0,5}(\boldsymbol{x} = \mathbf{0})} = \exp(\widehat{\beta_1}).$$

```
# tempo mediano
exp(ajuste_ln$coefficients[2])
```

```
##     SEXO
## 1.512375
```

```
# calculo da funcao de sobrevivencia

# sexo masculino e tempo = 2 anos
pnorm((-log(2)+ajuste_ln$coefficients[1]) / ajuste_ln$scale)
```

```
## (Intercept)
##   0.2689526
```

```r
# sexo feminino e tempo = 2 anos
pnorm((-log(2) + (ajuste_ln$coefficients[1] + ajuste_ln$coefficients[2])) / ajuste_ln$scale)
```

```
## (Intercept)
##   0.3524245
```

Obtemos que a razão do tempo mediano de paciente do sexo feminino pelo sexo masculino é aproximadamente igual a 1,5, ou seja, o tempo mediano de pacientes do sexo feminino é 1,5 vezes o tempo mediano de paciente do sexo masculino.

Na segunda análise, obtemos que a função de sobrevivência para pacientes do sexo masculino calculado no tempo de 2 anos é igual a 0,27, enquanto que do sexo feminino é igual a 0,35. Ou seja, espera-se que 27% dos pacientes do sexo masculino estejam vivos após 2 anos do diagnóstico, já para pacientes do sexo feminino é esperado 35%.