Análise de Sobrevivência - Modelos de Regressão Paramétricos

José Luiz Padilha

Abril de 2024

Exemplo 1: Pacientes com Câncer de Bexiga

Considere os tempos de reincidência, em meses, de um grupo de 20 pacientes com câncer de bexiga que foram submetidos a um procedimento cirúrgico realizado por laser.

Análise exploratória

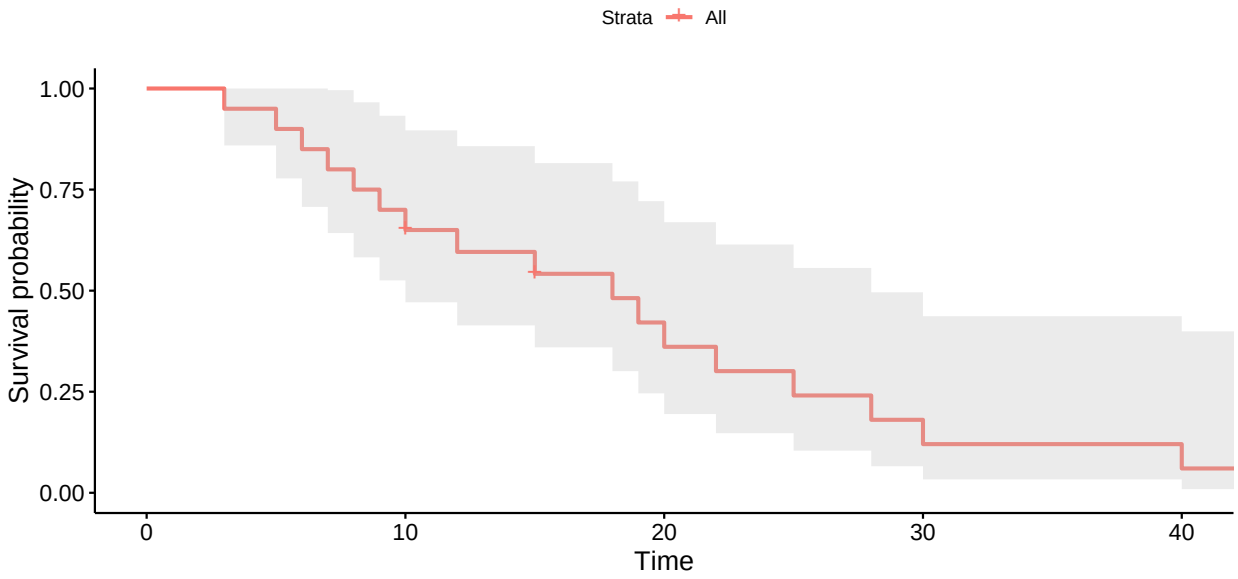
```
library(survival)
library(survminer)
tempos <- c(3, 5, 6, 7, 8, 9, 10, 10, 12, 15, 15, 18, 19, 20, 22, 25, 28, 30, 40, 45)
cens <- c(1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0)
bexiga <- data.frame(tempos, cens)
```

As estimativas de sobrevivência via Kaplan-Meier são mostradas a seguir.

```
ekm <- survfit(Surv(tempos, cens) ~ 1, data = bexiga)
summary(ekm)
```

```
## Call: survfit(formula = Surv(tempos, cens) ~ 1, data = bexiga)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     3     20       1   0.9500  0.0487      0.85913        1.000
##     5     19       1   0.9000  0.0671      0.77767        1.000
##     6     18       1   0.8500  0.0798      0.70707        1.000
##     7     17       1   0.8000  0.0894      0.64257        0.996
##     8     16       1   0.7500  0.0968      0.58233        0.966
##     9     15       1   0.7000  0.1025      0.52541        0.933
##    10     14       1   0.6500  0.1067      0.47124        0.897
##    12     12       1   0.5958  0.1107      0.41402        0.857
##    15     11       1   0.5417  0.1131      0.35976        0.816
##    18      9       1   0.4815  0.1154      0.30096        0.770
##    19      8       1   0.4213  0.1156      0.24601        0.721
##    20      7       1   0.3611  0.1137      0.19481        0.669
##    22      6       1   0.3009  0.1095      0.14745        0.614
##    25      5       1   0.2407  0.1028      0.10422        0.556
##    28      4       1   0.1806  0.0931      0.06573        0.496
##    30      3       1   0.1204  0.0792      0.03317        0.437
##    40      2       1   0.0602  0.0581      0.00907        0.399
```

```
ggsurvplot(ekm, conf.int = TRUE)
```



Suponha que o pesquisador tenha interesse em responder três questões:

- Qual é o tempo médio de vida?
- Qual é o tempo mediano de vida?
- Qual é a probabilidade de um paciente sobreviver a 20 meses?

```
print(ekm, print.rmean = TRUE)
```

```
## Call: survfit(formula = Surv(tempos, cens) ~ 1, data = bexiga)
##
##       n events rmean* se(rmean) median 0.95LCL 0.95UCL
## [1,] 20     17   18.7      2.73     18      10      28
##     * restricted mean with upper limit =  45
```

```
round(r <- summary(ekm)$table, 2)
```

```
##   records     n.max   n.start    events     rmean se(rmean)    median   0.95LCL
##     20.00     20.00     20.00     17.00     18.73      2.73     18.00     10.00
##   0.95UCL
##     28.00
```

```
round(r["rmean"] + c(-1, 1)*1.96*r["se(rmean)"], 2)
```

```
## [1] 13.37 24.08
```

O tempo médio ($E(T)$) é estimado em 18,73 meses com intervalo de confiança de 95% (13,37; 24,08). Note que este valor é mal estimado pois o último tempo observado é uma censura. Já o tempo mediano ($t_{0,50}$), obtido diretamente da curva de sobrevivência estimada, é dado por 18,00 meses com intervalo de confiança (10,00; 28,00). Estimativas para o tempo mediano poderiam também ser obtidas por interpolação.

```
summary(ekm, times = 20)
```

```
## Call: survfit(formula = Surv(tempos, cens) ~ 1, data = bexiga)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    20      7      12    0.361   0.114        0.195        0.669
```

A sobrevida aos 20 meses, $S(20)$, é estimada em 0,361 com intervalo de confiança de 0,195 a 0,669.

Ajuste dos modelos

Na sequência, faremos ajustes de modelos paramétricos. Com base no melhor modelo escolhido, daremos uma resposta a cada questão de interesse.

- **Modelo exponencial**

```
ajust1 <- survreg(Surv(tempos, cens) ~ 1, dist = 'exponential')
ajust1
```

```
## Call:
## survreg(formula = Surv(tempos, cens) ~ 1, dist = "exponential")
##
## Coefficients:
## (Intercept)
##    3.016111
##
## Scale fixed at 1
##
## Loglik(model)= -68.3   Loglik(intercept only)= -68.3
## n= 20
```

```
alpha <- exp(ajust1$coefficients[1])
alpha
```

```
## (Intercept)
##    20.41176
```

- **Modelo Weibull**

```
ajust2 <- survreg(Surv(tempos, cens) ~ 1, dist = 'weibull')
ajust2
```

```
## Call:
## survreg(formula = Surv(tempos, cens) ~ 1, dist = "weibull")
##
## Coefficients:
## (Intercept)
##    3.060529
##
## Scale= 0.647922
##
## Loglik(model)= -66.1   Loglik(intercept only)= -66.1
## n= 20
```

```
alpha <- exp(ajust2$coefficients[1])
gama <- 1/ajust2$scale
cbind(gama, alpha)
```

```
##                 gama    alpha
## (Intercept) 1.543396 21.33885
```

- **Modelo log-normal**

```
ajust3 <- survreg(Surv(tempos, cens) ~ 1, dist = 'lognorm')
ajust3
```

```
## Call:
## survreg(formula = Surv(tempos, cens) ~ 1, dist = "lognorm")
##
```

```
## Coefficients:
## (Intercept)
##    2.717176
##
## Scale= 0.7648167
##
## Loglik(model)= -65.7   Loglik(intercept only)= -65.7
## n= 20
```

As estimativas de sobrevivência para os modelos exponencial, Weibull e log-normal, são dadas, respectivamente, por:

$$\begin{aligned}\hat{S}_e(t) &= \exp\{-t/20,41\},\\ \hat{S}_w(t) &= \exp\left\{-(t/20,34)^{1,54}\right\},\\ \hat{S}_l(t) &= \Phi\{-(\log(t)-2,72)/0,76\}.\end{aligned}$$

Por exemplo, para o tempo $t = 10$, obtemos

$$\begin{aligned}\hat{S}_e(10) &= \exp\{-10/20,41\} = 0,613,\\ \hat{S}_w(10) &= \exp\left\{-(10/20,34)^{1,54}\right\} = 0,733,\\ \hat{S}_l(10) &= \Phi\{-(\log(10)-2,72)/0,76\} = 0,709.\end{aligned}$$

```
round(c(exp(-10/20.41), exp(-(10/21.34)^1.54), pnorm((-log(10)+ 2.72)/0.76)), 3)
```

```
## [1] 0.613 0.733 0.709
```

Note que as estimativas dos modelos Weibull e log-normal estão próximas, enquanto o modelo exponencial retorna um valor ligeiramente diferente. Na sequência, são calculadas as estimativas para os três modelos e também o Kaplan-Meier.

```
time <- ekm$time
st <- ekm$surv
ste <- exp(-time/20.41)
stw <- exp(-(time/21.34)^1.54)
stln <- pnorm((-log(time) + 2.72)/0.76)
round(cbind(time, st, ste, stw, stln), 3)
```

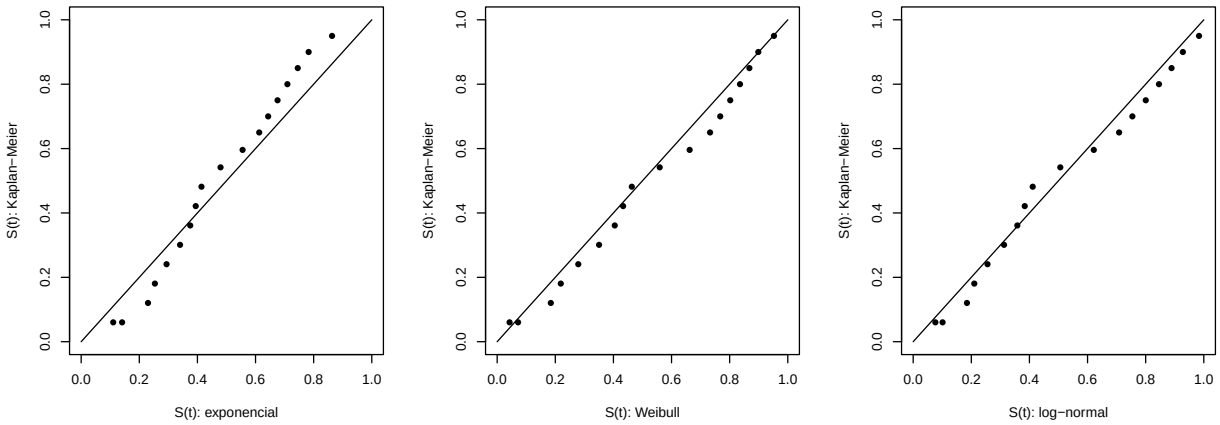
```
##       time    st   ste   stw  stln
##  [1,]    3 0.950 0.863 0.952 0.984
##  [2,]    5 0.900 0.783 0.899 0.928
##  [3,]    6 0.850 0.745 0.868 0.889
##  [4,]    7 0.800 0.710 0.836 0.846
##  [5,]    8 0.750 0.676 0.802 0.800
##  [6,]    9 0.700 0.643 0.768 0.754
##  [7,]   10 0.650 0.613 0.733 0.709
##  [8,]   12 0.596 0.555 0.662 0.621
##  [9,]   15 0.542 0.480 0.559 0.506
## [10,]   18 0.481 0.414 0.463 0.411
## [11,]   19 0.421 0.394 0.433 0.384
## [12,]   20 0.361 0.375 0.405 0.358
## [13,]   22 0.301 0.340 0.351 0.313
## [14,]   25 0.241 0.294 0.279 0.256
## [15,]   28 0.181 0.254 0.219 0.210
## [16,]   30 0.120 0.230 0.185 0.185
## [17,]   40 0.060 0.141 0.072 0.101
## [18,]   45 0.060 0.110 0.043 0.076
```

Diagnóstico

Método gráfico 1

Comparamos as estimativas das sobrevivências obtidas via Kaplan-Meier com aquelas obtidas pelos modelos.

```
par(mfrow = c(1, 3))
plot(ste, st, pch = 16, ylim = range(c(0.0, 1)), xlim = range(c(0, 1)),
     ylab = "S(t): Kaplan-Meier", xlab = "S(t): exponencial")
lines(c(0, 1), c(0, 1), type = "l", lty = 1)
plot(stw, st, pch = 16, ylim = range(c(0.0, 1)), xlim = range(c(0, 1)),
     ylab = "S(t): Kaplan-Meier", xlab = "S(t): Weibull")
lines(c(0, 1), c(0, 1), type = "l", lty = 1)
plot(stln, st, pch = 16, ylim = range(c(0.0, 1)), xlim = range(c(0, 1)),
     ylab = "S(t): Kaplan-Meier", xlab = "S(t): log-normal")
lines(c(0, 1), c(0, 1), type = "l", lty = 1)
```

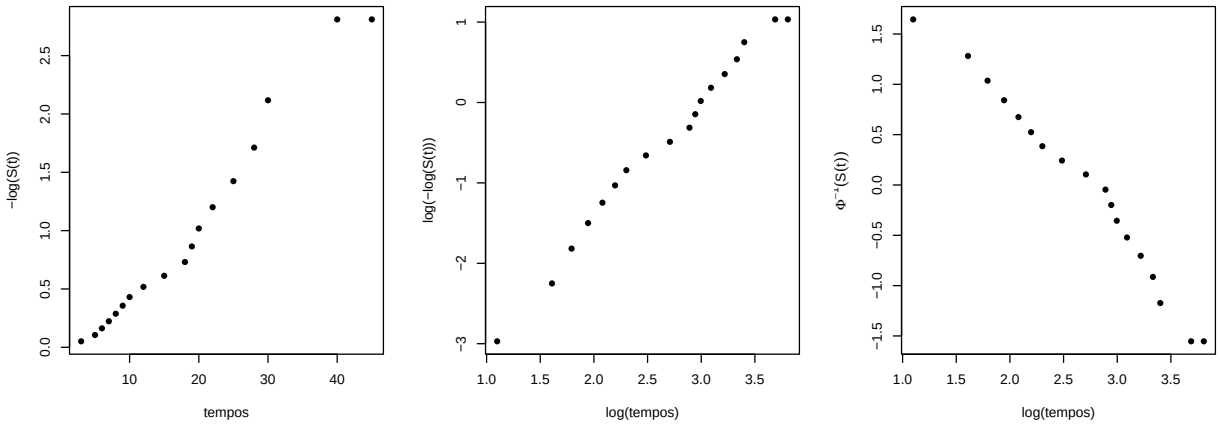


O modelo exponencial não parece ser adequado para estes dados, enquanto os modelos Weibull e log-normal apresentam melhor ajuste por estarem mais próximos da reta $y = x$.

Método gráfico 2

Construímos os gráficos linearizados para os três modelos.

```
par(mfrow = c(1, 3))
invst <- qnorm(st)
plot(time, -log(st), pch = 16, xlab = "tempos", ylab = "-log(S(t))")
plot(log(time), log(-log(st)), pch = 16, xlab = "log(tempos)", ylab = "log(-log(S(t)))")
plot(log(time), invst, pch = 16, xlab = "log(tempos)", ylab = expression(Phi^{-1}*(S(t))))
```



Diferente do modelo exponencial, os modelos Weibull e log-normal não mostram grandes desvios da reta.

Podemos comparar os modelos por meio do teste da razão de verossimilhanças. Para isso, vamos ajustar o modelo gama generalizado usando o pacote `flexsurv`.

```
library(flexsurv)
ajust4 <- flexsurvreg(Surv(tempos, cens) ~ 1, dist = 'gengamma')
ajust4
```

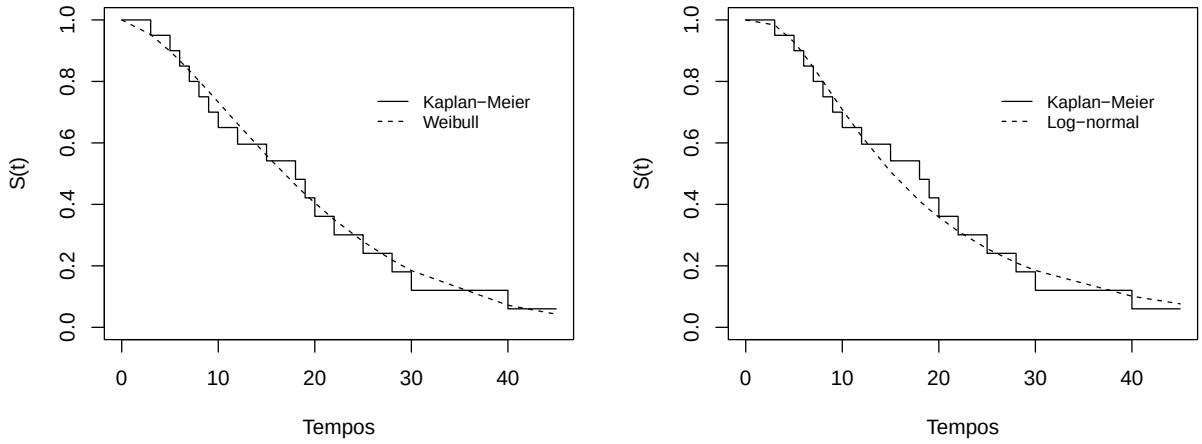
```
## Call:
## flexsurvreg(formula = Surv(tempos, cens) ~ 1, dist = "gengamma")
##
## Estimates:
##        est     L95%    U95%    se
## mu      2.805   2.168   3.442   0.325
## sigma   0.743   0.498   1.110   0.152
## Q       0.247  -1.291   1.786   0.785
##
## N = 20,  Events: 17,  Censored: 3
## Total time at risk: 347
## Log-likelihood = -65.69074, df = 3
## AIC = 137.3815
```

```
#
models <- c("Gama Generalizada", "Exponencial", "Weibull", "Log-normal")
loglik <- c(ajust4$loglik, ajust1$loglik[1], ajust2$loglik[1], ajust3$loglik[1])
TRV <- c(NA, 2*(ajust4$loglik - ajust1$loglik[1]), 2*(ajust4$loglik - ajust2$loglik[1]),
         2*(ajust4$loglik - ajust3$loglik[1]))
npar <- c(3, 1, 2, 2)
p <- c(NA, pchisq(TRV[2], df = 2, lower.tail = FALSE), pchisq(TRV[3:4], df = 1,
                                                              lower.tail = FALSE))
res_trv <- data.frame(models, npar, loglik, TRV, p)
print(res_trv, digits = 3)
```

```
##              models npar loglik    TRV      p
## 1 Gama Generalizada    3  -65.7     NA     NA
## 2       Exponencial    1  -68.3 5.1663 0.0755
## 3           Weibull    2  -66.1 0.8852 0.3468
## 4        Log-normal    2  -65.7 0.0983 0.7539
```

O teste aponta que o modelo exponencial pode ser descartado, mas não conseguimos diferenciar entre os modelos log-normal e Weibull. A provável razão é o pequeno tamanho de amostra. Seguimos as análises com os modelos Weibull e log-normal. As curvas de sobrevivência para os dois modelos versus as estimativas de Kaplan-Meier são dadas a seguir.

```
par(mfrow = c(1, 2))
plot(ekm, conf.int = F, xlab = "Tempos", ylab = "S(t)")
lines(c(0,time),c(1,stw), lty=2)
legend(25, 0.8, lty = c(1, 2), c("Kaplan-Meier", "Weibull"), bty = "n", cex = 0.8)
plot(ekm, conf.int = F, xlab = "Tempos", ylab = "S(t)")
lines(c(0, time), c(1, stln), lty = 2)
legend(25, 0.8, lty = c(1, 2), c("Kaplan-Meier", "Log-normal"), bty = "n", cex = 0.8)
```



Ambos os modelos apresentam ajustes satisfatórios.

Tempo médio

Estimativas para o tempo médio, com base nas distribuições Weibull e log-normal, são dadas, respectivamente, por:

$$\hat{E}(T) = \hat{\alpha}[\Gamma(1 + 1/\hat{\gamma})] = 21,34[\Gamma(1 + (1/1,54))] = 19,201 \text{ meses},$$
$$\hat{E}(T) = \exp\left\{\hat{\mu} + \hat{\sigma}^2/2)\right\} = \exp\left\{2,72 + 0,76^2/2)\right\} = 20,280 \text{ meses}.$$

```
round(cbind("Weibull" = alpha*gamma(1 + 1/gama),
            "Log-normal" = exp(coef(ajust3) + (ajust3$scale^2/2))), 3)
```

```
##             Weibull Log-normal
## (Intercept)  19.201      20.28
```

Intervalos de confiança para $E(T)$ podem ser obtidos a partir de estimativas para $Var(\hat{E}(T))$, o que faremos usando o método delta. Para o modelo log-normal, vimos anteriormente que:

$$\widehat{Var}(\hat{E}[T]) \approx \widehat{Var}(\hat{\mu})\left[\exp\left\{\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right\}\right]^2 + \widehat{Var}(\hat{\sigma})\left[\hat{\sigma}\exp\left\{\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right\}\right]^2$$
$$+ 2\widehat{Cov}(\hat{\mu}, \hat{\sigma})\left[\exp\left\{\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right\}\right]\left[\hat{\sigma}\exp\left\{\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right\}\right].$$

O modelo ajustado retorna a seguinte matriz de variância-covariância:

```
vcov <- ajust3$var
vcov
```

```
##             (Intercept)  Log(scale)
## (Intercept) 0.031061677 0.002706896
## Log(scale)  0.002706896 0.030119031
```

Note que temos as estimativas $\widehat{Var}(\hat{\mu})$ na posição $[1,1]$, $\widehat{Var}(\log(\hat{\sigma}))$ na posição $[2,2]$, e $\widehat{Cov}(\hat{\mu}, \log(\hat{\sigma}))$ na posição $[1,2]$. Precisamos ainda obter $\widehat{Var}(\hat{\sigma})$ e $\widehat{Cov}(\hat{\mu}, \hat{\sigma})$

Tomando $\theta = \log(\sigma)$, temos que $g(\theta) = \exp(\theta) = \sigma$. Logo, $\widehat{Var}(\hat{\sigma}) \approx \hat{\sigma}^2 \widehat{Var}(\log(\hat{\sigma}))$ e $\widehat{Cov}(\hat{\mu}, \hat{\sigma}) = \hat{\sigma}\widehat{Cov}(\hat{\mu}, \log(\hat{\sigma}))$. Logo, uma aproximação para $\widehat{Var}(\hat{E}[T])$ é obtida como

```
mu <- coef(ajust3)
var_mu <- vcov[1,1]
sigma <- ajust3$scale
sigma2 <- sigma^2
var_sigma <- sigma2*vcov[2,2]
cov_sigma_mu <- sigma*vcov[1,2]
var_et <- var_mu*((exp(mu+sigma2/2))^2) + var_sigma*((sigma*exp(mu+sigma2/2))^2) +
  2*cov_sigma_mu*(exp(mu+sigma2/2))*(sigma*exp(mu+sigma2/2))
var_et
```

```
## (Intercept)
##    18.31633
```

```
et <- exp(mu + sigma2/2)
```

O intervalo de confiança de 95% para $E(T)$ usando o método delta é

```
ic_boot_delta <- et + c(-1, 1)*1.96*sqrt(var_et)
round(ic_boot_delta, 2)
```

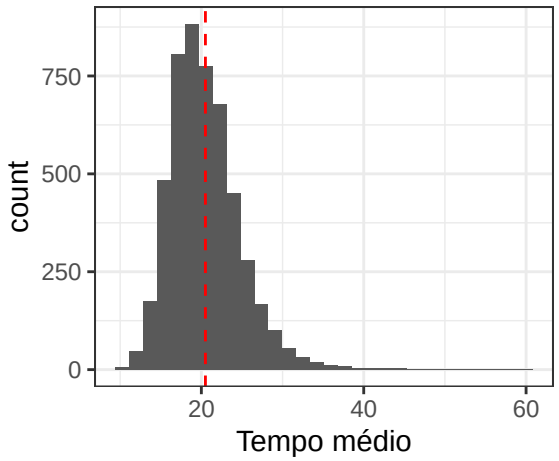
```
## [1] 11.89 28.67
```

Como alternativa ao método delta, podemos calcular intervalos de confiança via bootstrap, reamostrando B vezes com reposição o par (t_i, δ_i), $i = 1, \ldots, n$. Inicialmente, obtemos uma estimativa para a variância do tempo médio no modelo log-normal.

```
set.seed(123)
B <- 5000
tmb <- rep(NA, B)
for(i in 1:B){
  dat <- bexiga[sample(1:nrow(bexiga), replace = TRUE),]
  fit_boot <- survreg(Surv(tempos, cens) ~ 1, dist = 'lognorm', data = dat)
  tmb[i] <- exp(fit_boot$coef[1] + (fit_boot$scale^2/2))
}
c(mean(tmb), var(tmb))
```

```
## [1] 20.49809 19.51711
```

```
tmb %>% as.data.frame() %>% ggplot(aes(x = .)) + geom_histogram() + theme_bw() +
  labs(x = "Tempo médio", ylab = "Frequência") + geom_vline(xintercept = mean(tmb),
                                                            linetype = "dashed",
                                                            color = "red")
```



Os intervalo de confiança gaussiano de 95% para $E(T)$ usando a variância boostrap e o intervalo de confiança percentílico são calculados na sequência.

```
ic_boot_gauss <- et + c(-1, 1)*1.96*sqrt(var(tmb))
ic_boot_perc <- quantile(tmb, probs = c(0.025, 0.975))
```

Em resumo, temos as estimativas pontual e intervalar para o tempo médio

```
round(cbind(et, rbind(ic_boot_delta, ic_boot_gauss, ic_boot_perc)), 2)
```

```
##                   et  2.5% 97.5%
## ic_boot_delta 20.28 11.89 28.67
## ic_boot_gauss 20.28 11.62 28.94
## ic_boot_perc  20.28 13.87 30.48
```

As estimativas pelo método delta e assintótica usando a variância bootstrap são bastante similares. As diferenças para o intervalo percentílico podem ser explicadas pela assimetria observada nas estimativas bootstrap do tempo médio.

Tempo mediano

Uma estimativa para o tempo mediano usando o modelo log-normal é

```
exp(qnorm(0.5)*ajust3$scale + coef(ajust3))
```

```
## (Intercept)
##    15.13751
```

Novamente, um procedimento bootstrap poderia ser aplicado para obtenção dos intervalos de confiança

Sobrevida estimada em 20 meses

Para o modelo log-normal, a probabilidade estimada de um indivíduo estar livre da doença aos 20 meses é 35,8%.

```
pnorm((-log(20) + ajust3$coefficients)/ajust3$scale)
```

```
## (Intercept)
##   0.3578492
```

Este valor é bastante próximo da estimativa de Kaplan-Meier que vale 36,1% (veja tabela logo após ajuste dos modelos).

Exemplo 2: Sobrevida de Pacientes com Leucemia Aguda

Considere os seguintes tempos de sobrevivência de 17 pacientes com leucemia aguda (Lawless, 2003) juntamente com suas contagens de glóbulos brancos (WBC) e seus correspondentes logaritmos, na base 10.

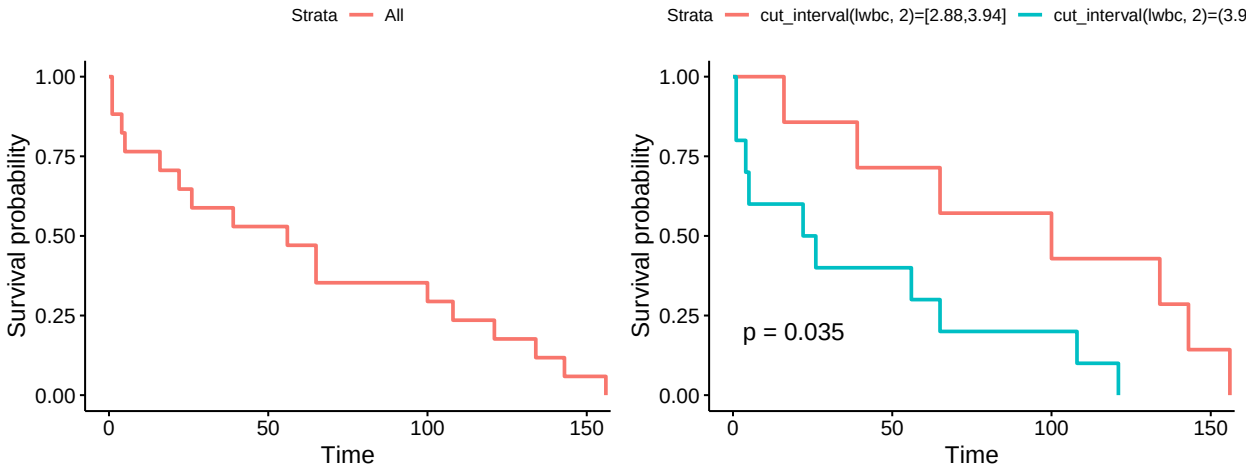
Tempos	WBC	$\log_{10}$(WBC)	Tempos	WBC	$\log_{10}$(WBC)
65	2300	3,36	143	7000	3,85
156	750	2,88	56	9400	3,97
100	4300	3,63	26	32000	4,51
134	2600	3,41	22	35000	4,54
16	6000	3,78	1	100000	5,00
108	10000	4,02	1	100000	5,00
121	10000	4,00	5	52000	4,72
4	17000	4,23	65	100000	5,00
39	5400	3,73			

```
library(survival)
library(survminer)
temp <- c(65, 156, 100, 134, 16, 108, 121, 4, 39, 143, 56, 26, 22, 1, 1, 5, 65)
cens <- rep(1, 17)
lwbc <- c(3.36, 2.88, 3.63, 3.41, 3.78, 4.02, 4.00, 4.23, 3.73, 3.85, 3.97, 4.51, 4.54,
          5.00, 5.00, 4.72, 5.00)
dados <- as.data.frame(cbind(temp, cens, lwbc))
```

Análise exploratória

Como WBC é contínua, a menos que esta seja categorizada, é inviável a obtenção das curvas de sobrevivência por meio do estimador de Kaplan-Meier. Uma possibilidade para avaliação do efeito de tal covariável por meio das técnicas não paramétricas discutidas anteriormente é a dividirmos em dois grupos com igual amplitude e compararmos as curvas por meio do teste logrank.

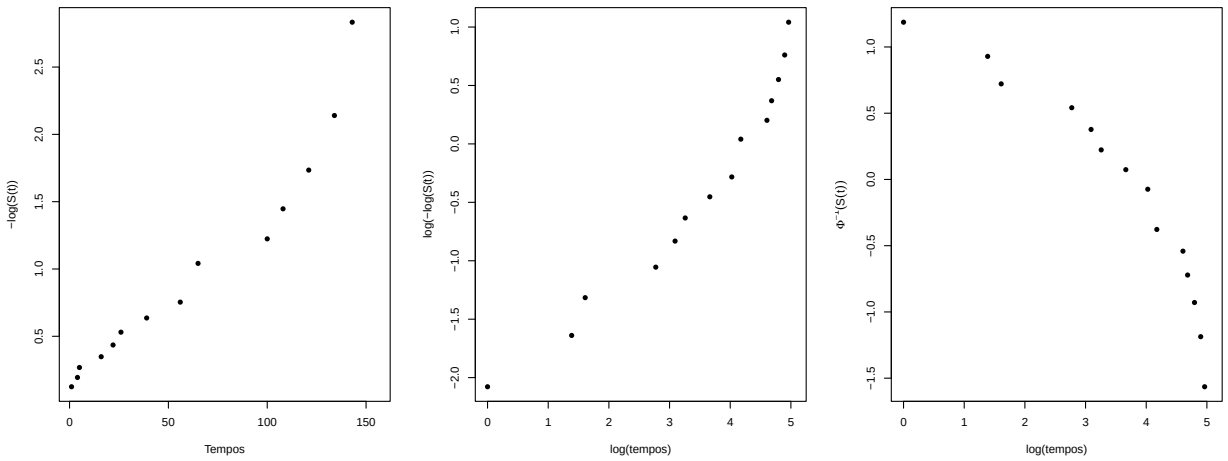
```
ekm <- survfit(Surv(temp, cens) ~ 1, data = dados)
ekm2 <- survfit(Surv(temp, cens) ~ cut_interval(lwbc, 2), data = dados)
splots <- list()
splots[[1]] <- ggsurvplot(ekm, conf.int = FALSE)
splots[[2]] <- ggsurvplot(ekm2, pval = TRUE, conf.int = FALSE)
arrange_ggsurvplots(splots)
```



Ajuste dos modelos de regressão paramétricos

Para escolha do modelo de regressão, inicialmente ignoraremos a covariável WBC e construiremos os gráficos das linearizações para os modelos exponencial, Weibull e log-normal.

```
st <- ekm$surv
temp <- ekm$time
invst <- qnorm(st)
par(mfrow = c(1, 3))
plot(temp, -log(st), pch = 16, xlab = "Tempos", ylab = "-log(S(t))")
plot(log(temp), log(-log(st)), pch = 16, xlab = "log(tempos)", ylab = "log(-log(S(t))")
plot(log(temp), invst, pch = 16, xlab = "log(tempos)", ylab = expression(Phi^-1 * (S(t))))
```



As distribuições exponencial e Weibull apresentam-se visualmente como as melhores candidatas. Considerando estes modelos com a covariável $X_1 = \log(\text{WBC})$, temos os seguintes resultados:

```
# Ajuste exponencial
ajust1 <- survreg(Surv(dados$temp, dados$cens) ~ dados$lwbc, dist = 'exponential')
ajust1
```

```
## Call:
## survreg(formula = Surv(dados$temp, dados$cens) ~ dados$lwbc,
```

```
##     dist = "exponential")
##
## Coefficients:
## (Intercept)  dados$lwbc
##    8.477498   -1.109298
##
## Scale fixed at 1
##
## Loglik(model)= -83.9   Loglik(intercept only)= -87.3
##  Chisq= 6.83 on 1 degrees of freedom, p= 0.00899
## n= 17
```

```
# Ajuste Weibull
ajust2 <- survreg(Surv(dados$temp, dados$cens) ~ dados$lwbc, dist = 'weibull')
ajust2
```

```
## Call:
## survreg(formula = Surv(dados$temp, dados$cens) ~ dados$lwbc,
##     dist = "weibull")
##
## Coefficients:
## (Intercept)  dados$lwbc
##    8.440773   -1.098237
##
## Scale= 0.9786422
##
## Loglik(model)= -83.9   Loglik(intercept only)= -87.1
##  Chisq= 6.48 on 1 degrees of freedom, p= 0.0109
## n= 17
```

O valor estimado de $\gamma = 1/\sigma$ é muito próximo de 1.

```
gama <- 1/ajust2$scale
gama
```

```
## [1] 1.021824
```

Já o teste de razão de verossimilhanças para as hipóteses $H_0 : \gamma = 1$ versus $H_1 : \gamma \neq 1$ fornece indicações favoráveis ao modelo exponencial.

```
anova(ajust1, ajust2)
```

```
##        Terms Resid. Df    -2*LL Test Df   Deviance  Pr(>Chi)
## 1 dados$lwbc        15 167.7541      NA         NA        NA
## 2 dados$lwbc        14 167.7427    =  1 0.01138216 0.9150372
```

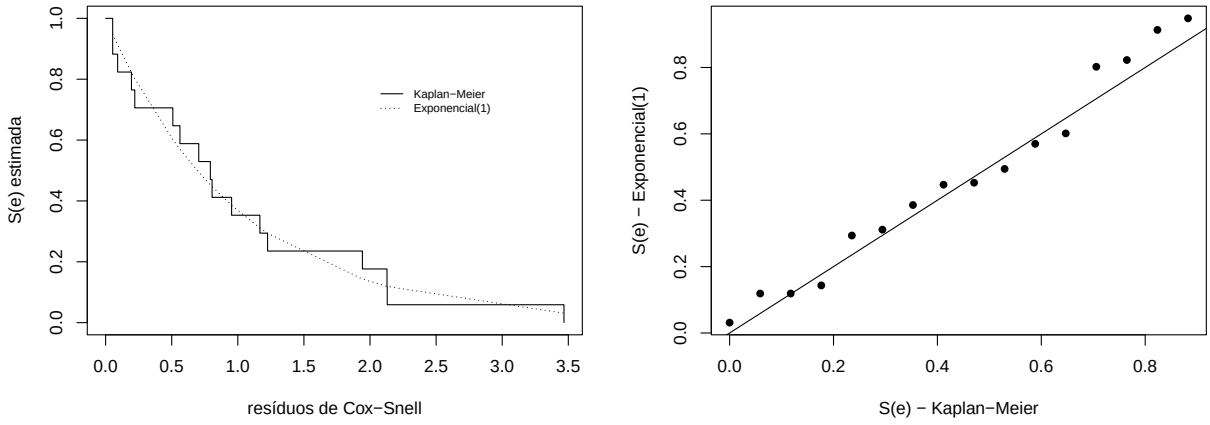
A escolha pelo modelo exponencial foi realizada antes da inclusão da covariável. Para avaliar a adequação do ajuste foram utilizados os resíduos de Cox-Snell, definidos por:

$$\hat{e}_i = t_i \exp(-\hat{\beta}_0 - \hat{\beta}_1 x_{i1}),\ i = 1, \ldots, n.$$

Se o modelo for adequado, tais resíduos devem ser vistos como provenientes de uma amostra da distribuição exponencial padrão:

- as estimativas das curvas de sobrevivência desses resíduos obtidas por Kaplan-Meier ($\hat{S}(\hat{e}_i)_{KM}$) e pelo modelo exponencial padrão ($\hat{S}(\hat{e}_i)_{Exp}$) devem ser próximas;
- o gráfico dos pares de pontos $(\hat{S}(\hat{e}_i)_{KM}, \hat{S}(\hat{e}_i)_{Exp})$ deve ser aproximadamente uma reta para que o modelo ajustado possa ser considerado satisfatório.

```r
t <- dados$temp
x <- dados$lwbc
bo <- ajust1$coefficients[1]
b1 <- ajust1$coefficients[2]
res <- t*exp(-bo-b1*x) # residuos de Cox-Snell
ekm <- survfit(Surv(res, dados$cens) ~ 1)
par(mfrow = c(1, 2))
plot(ekm, conf.int = F, lty = c(1, 1), xlab = "resíduos de Cox-Snell",
     ylab = "S(e) estimada")
res <- sort(res)
exp1 <- exp(-res)
lines(res, exp1, lty = 3)
legend(2, 0.8, lty = c(1, 3), c("Kaplan-Meier", "Exponencial(1)"), lwd = 1, bty = "n",
       cex = 0.7)
st <- ekm$surv
t <- ekm$time
sexp1 <- exp(-t)
plot(st, sexp1, xlab = "S(e) - Kaplan-Meier", ylab = "S(e) - Exponencial(1)", pch = 16)
abline(coef = c(0, 1))
```



Os gráficos indicam a adequação do modelo exponencial.

```r
summary(ajust1)
```

```
##
## Call:
## survreg(formula = Surv(dados$temp, dados$cens) ~ dados$lwbc,
##     dist = "exponential")
##               Value Std. Error     z       p
## (Intercept)  8.477      1.711  4.95 7.3e-07
## dados$lwbc  -1.109      0.414 -2.68  0.0073
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -83.9   Loglik(intercept only)= -87.3
##  Chisq= 6.83 on 1 degrees of freedom, p= 0.009
## Number of Newton-Raphson Iterations: 5
```

```
## n= 17
```

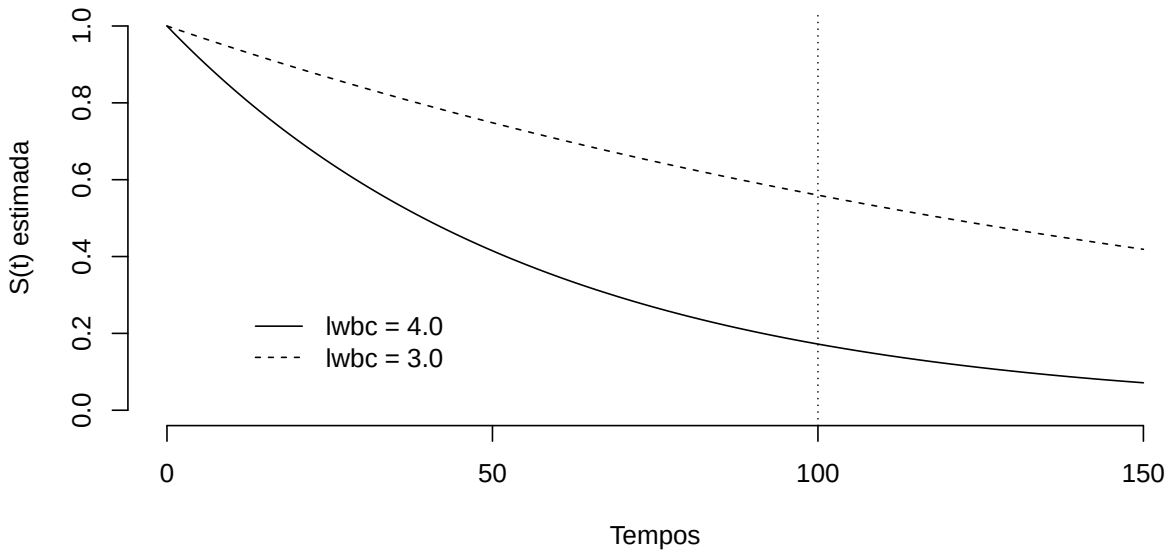
O teste de razão de verossimilhanças para as hipóteses $H_0 : \beta_1 = 1$ versus $H_1 : \beta_1 \neq 1$ conclui pela rejeição da hipótese nula. Podemos dizer que parte da variação observada nos tempos de sobrevivência pode ser explicada pela contagem de glóbulos brancos.

A função de sobrevivência ajustada é dada por:

$$\hat{S}(t|x_1) = \exp\left\{-\left(\frac{t}{\exp\{8,4775 - 1,1093x_1\}}\right)\right\}, \quad t > 0.$$

Como $\hat{\beta}_1$ é negativo, quanto maior o valor de x_1, menor a probabilidade de sobrevivência estimada. A seguir temos as curvas de sobrevivência estimadas para dois pacientes, um com $x_1 = 4$ e outro com $x_1 = 3$.

```
x1 <- 4.0; temp1 <- 0:150
ax1 <- exp(ajust1$coefficients[1] + ajust1$coefficients[2]*x1)
ste1 <- exp(-(temp1/ax1))
x1 <- 3.0; temp2 <- 0:150
ax2 <- exp(ajust1$coefficients[1] + ajust1$coefficients[2]*x1)
ste2 <- exp(-(temp2/ax2))
par(mfrow = c(1, 1))
plot(temp1, temp1*0, pch = "", ylim = range(c(0, 1)), xlim = range(c(0, 150)),
     xlab = "Tempos", ylab = "S(t) estimada", bty = "n")
lines(temp1, ste1, lty = 1)
lines(temp2, ste2, lty = 2)
abline(v = 100, lty = 3)
legend(10, 0.3, lty = c(1, 2), c("lwbc = 4.0", "lwbc = 3.0"), lwd = 1, bty = "n")
```



Para o tempo $t = 100$ semanas (linha vertical tracejada), temos:

- $\hat{S}(100|x_1 = 4) = 0,172$ que significa que em torno de 17% dos pacientes que apresentam, no diagnóstico, logaritmo da contagem de glóbulos brancos igual a 4 estarão vivos no tempo $t = 100$ semanas;
- $\hat{S}(100|x_1 = 3) = 0,559$ indicando que aqueles pacientes com logaritmo da contagem de glóbulos brancos igual a 3, cerca de 56% estarão vivos na centésima semana.

Por fim, passamos à interpretação do coeficiente de regressão estimado. A cada aumento de uma unidade no logaritmo de WBC, o tempo mediano de vida dos pacientes fica reduzido para um terço ($e^{\hat{\beta}_1} = e^{-1,1093} \approx 0,33$):

```
exp(ajust1$coef[2])
```

```
## dados$lwbc
##  0.3297904
```

Uma propriedade importante do modelo de regressão exponencial é que ele pertence à classe dos modelos de tempo de vida acelerados e à de taxas de falha proporcionais. Assim, a interpretação acima também poderia ser feita em termos de taxas de falha proporcionais.

Exemplo 3: Sobrevida de Pacientes com Câncer de Encéfalo

Os dados utilizados provêm do Registro Hospitalar do Câncer (RHC) do Hospital Erasto Gaertner. O RHC foi implantado em novembro de 1992. A amostra é formada por pacientes com câncer maligno de localização topográfica C71 (encéfalo). Os registros vão de 17/05/1990 a 30/12/2001.

A amostra é composta por 397 pacientes. As variáveis disponíveis são:

- Sexo: 246 do sexo masculino e 151 do sexo feminino.
- Idade: Varia de 0 a 77 anos, sem grandes concentrações
- Tratamento realizado:
  - Radioterapia: 257
  - Radioterapia+Cirurgia: 71
  - Cirurgia: 21
  - Outros: 48 (Quimioterapia, Hormonioterapia ou combinações de tratamentos).
- Estadiamento da doença:
  - I:6 II: 40 III: 28 IV: 12
  - Não pode ser aplicado: 2
  - Não codificado: 308
- AED: Avaliação da extensão da doença:
  - Localizado: 350
  - Extensão direta: 33
  - Metástase: 5
  - Não aplicável: 2
  - Ignorado: 8

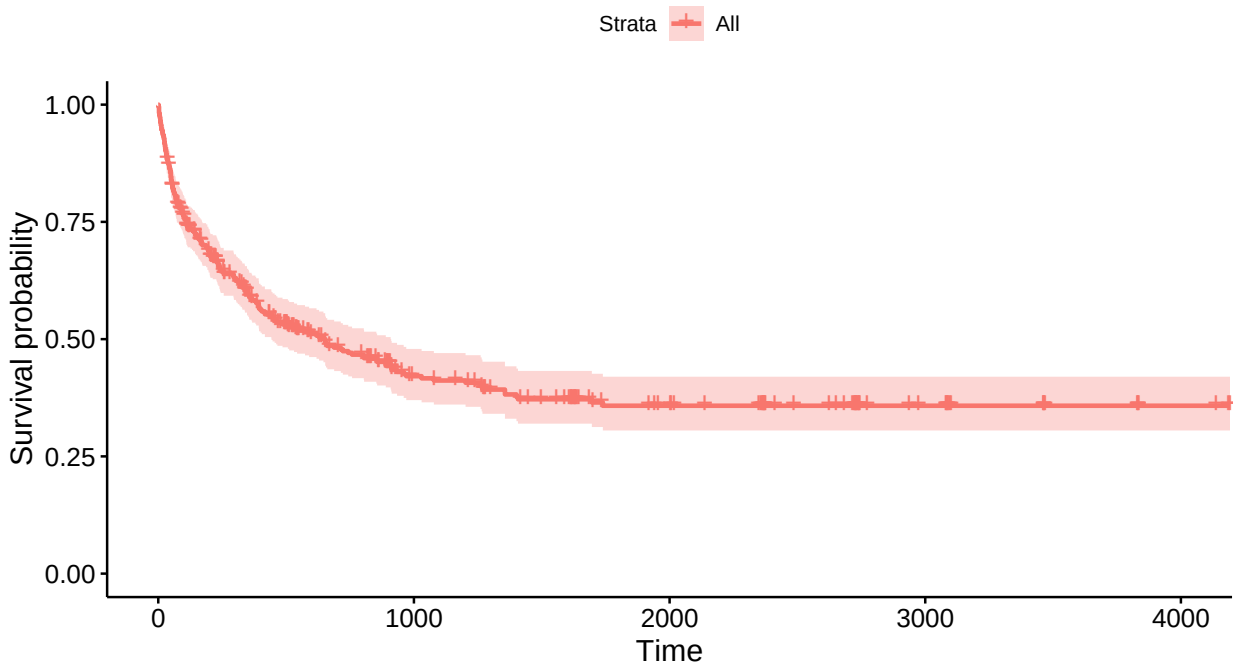
Foram observadas 216 falhas e 181 censuras. As covariáveis utilizadas nesta ilustração foram: idade do paciente, sexo e tipo de tratamento realizado.

```
encefalo <- read.table("enc.txt", header = TRUE)
```

Gráfico de Kaplan-Meier e testes logrank

A seguir o gráfico marginal de sobrevivência estimada.

```
ekm <- survfit(Surv(tempo, cens.1) ~ 1, data = encefalo)
ggsurvplot(ekm)
```



Uma estimativa do tempo mediano é dada por

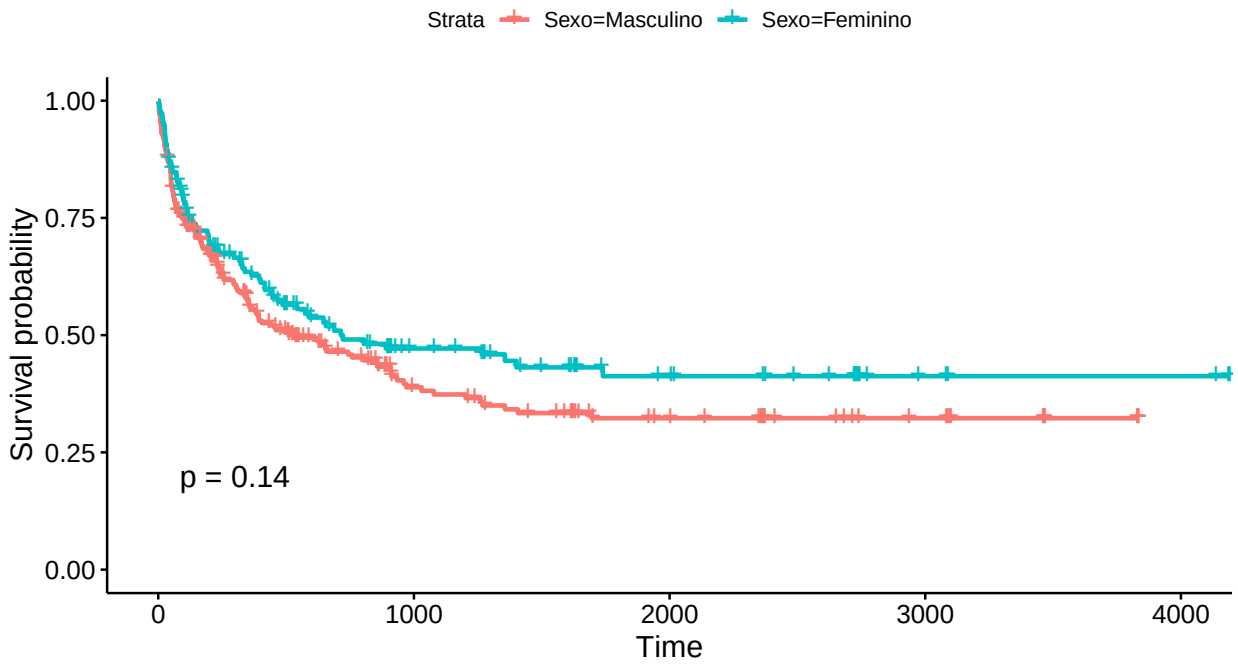
```
ekm
```

```
## Call: survfit(formula = Surv(tempo, cens.1) ~ 1, data = encefalo)
##
##         n events median 0.95LCL 0.95UCL
## [1,] 397    216    653     443     911
```

Vamos agora comparar as curvas de sobrevivência por meio do teste logrank. As covariáveis idade e tratamento foram dicotomizadas para aplicação do teste.

- Sexo

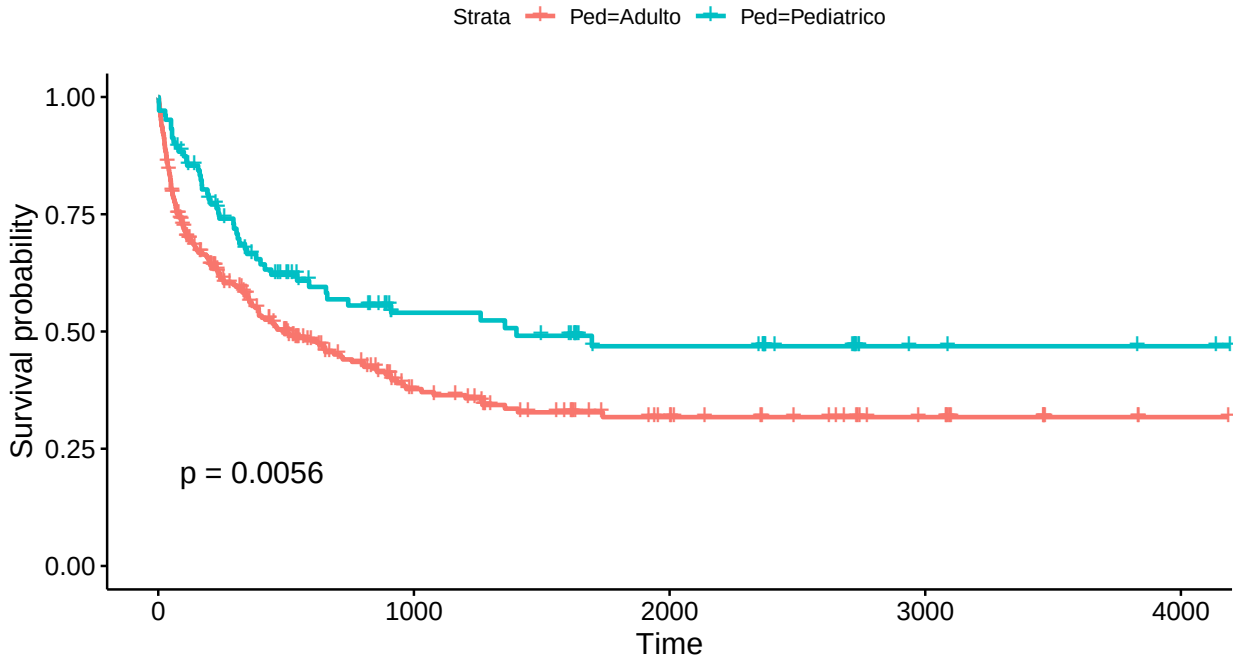
```
levels(encefalo$Sexo) <- c("Masculino", "Feminino")
ekm1 <- survfit(Surv(tempo, cens.1) ~ Sexo, data = encefalo)
ggsurvplot(ekm1, pval = TRUE)
```



O teste não aponta diferença entre os sexos.

- Idade

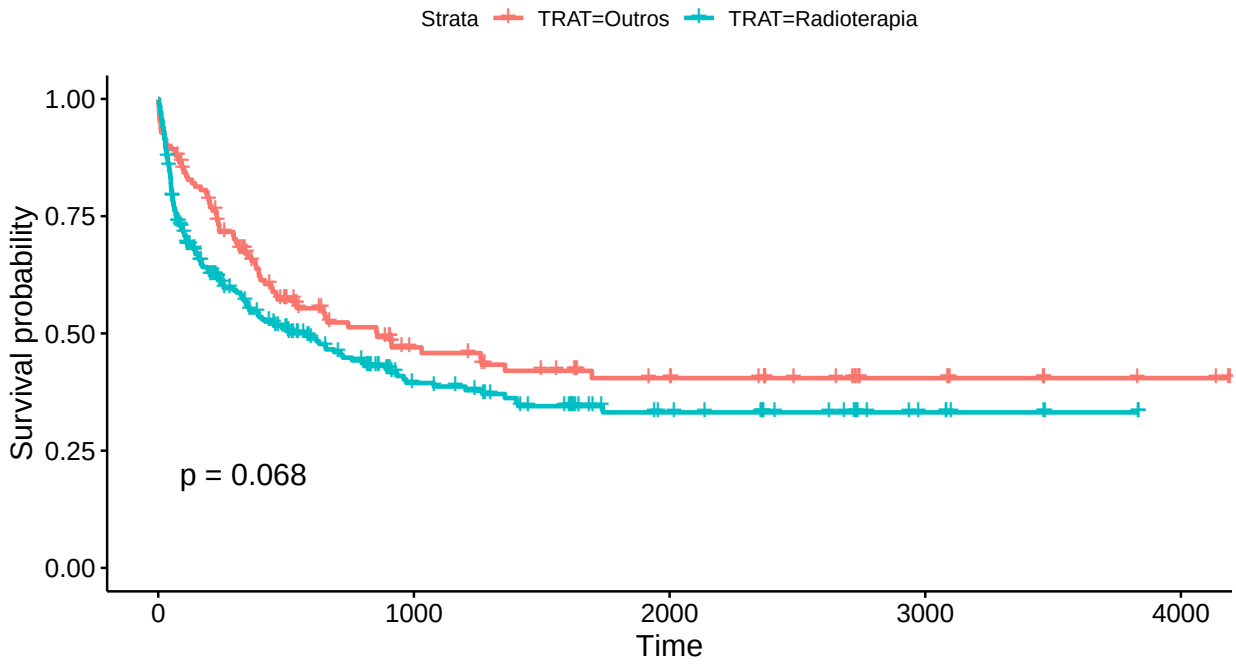
```
ekm2 <- survfit(Surv(tempo, cens.1) ~ Ped, data = encefalo)
ggsurvplot(ekm2, pval = TRUE)
```



Uma diferença estatisticamente significativa foi encontrada entre os grupos de idade. Os pacientes mais velhos apresentaram sobrevida pior.

- Tratamento

```
ekm3 <- survfit(Surv(tempo, cens.1) ~ TRAT, data = encefalo)
ggsurvplot(ekm3, pval = TRUE)
```



Com relação ao grupo tratamento dicotomizado, uma diferença marginal foi obtida. Os pacientes do grupo radioterapia apresentaram sobrevida estimada menor.

Ajuste dos modelos de regressão paramétricos

Vamos avaliar o ajuste dos modelos paramétricos exponencial, Weibull e log-normal.

```
m0e <- survreg(Surv(tempo, cens.1) ~ Idade + Sexo + TRAT, dis = "exponential",
               data = encefalo)
summary(m0e)
```

```
##
## Call:
## survreg(formula = Surv(tempo, cens.1) ~ Idade + Sexo + TRAT,
##     data = encefalo, dist = "exponential")
##                     Value Std. Error     z       p
## (Intercept)       7.87033    0.25911 30.37 < 2e-16
## Idade            -0.02664    0.00352 -7.56 4.1e-14
## Sexo              0.21937    0.14319  1.53    0.13
## TRATRadioterapia -0.14262    0.14905 -0.96    0.34
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -1741   Loglik(intercept only)= -1775.3
##  Chisq= 68.49 on 3 degrees of freedom, p= 9e-15
## Number of Newton-Raphson Iterations: 5
## n= 397
```

```r
m0w <- survreg(Surv(tempo, cens.1) ~ Idade + Sexo + TRAT, dis = "weibull",
               data = encefalo)
summary(m0w)
```

```
##
## Call:
## survreg(formula = Surv(tempo, cens.1) ~ Idade + Sexo + TRAT,
##     data = encefalo, dist = "weibull")
##                      Value Std. Error     z       p
## (Intercept)        8.15448    0.47224 17.27 < 2e-16
## Idade             -0.03466    0.00638 -5.44 5.5e-08
## Sexo               0.34607    0.26159  1.32    0.19
## TRATRadioterapia -0.13564    0.27367 -0.50    0.62
## Log(scale)         0.60088    0.05714 10.52 < 2e-16
##
## Scale= 1.82
##
## Weibull distribution
## Loglik(model)= -1666.8   Loglik(intercept only)= -1684.7
##  Chisq= 35.73 on 3 degrees of freedom, p= 8.5e-08
## Number of Newton-Raphson Iterations: 5
## n= 397
```

```r
m0l <- survreg(Surv(tempo, cens.1) ~ Idade + Sexo + TRAT, dis = "lognormal",
               data = encefalo)
summary(m0l)
```

```
##
## Call:
## survreg(formula = Surv(tempo, cens.1) ~ Idade + Sexo + TRAT,
##     data = encefalo, dist = "lognormal")
##                      Value Std. Error     z       p
## (Intercept)        7.12391    0.48611 14.65 < 2e-16
## Idade             -0.03133    0.00656 -4.78 1.8e-06
## Sexo               0.42578    0.27695  1.54    0.12
## TRATRadioterapia -0.12610    0.29305 -0.43    0.67
## Log(scale)         0.88086    0.05207 16.92 < 2e-16
##
## Scale= 2.41
##
## Log Normal distribution
## Loglik(model)= -1654   Loglik(intercept only)= -1668.2
##  Chisq= 28.33 on 3 degrees of freedom, p= 3.1e-06
## Number of Newton-Raphson Iterations: 3
## n= 397
```

Como visto, apenas a variável idade foi significativa. Vamos considerar então modelos apenas com este preditor.

```r
m1e <- survreg(Surv(tempo, cens.1) ~ Idade, dis = "exponential", data = encefalo)
m1w <- survreg(Surv(tempo, cens.1) ~ Idade, dis = "weibull", data = encefalo)
m1l <- survreg(Surv(tempo, cens.1) ~ Idade, dis = "lognormal", data = encefalo)
```

A seguir a comparação dos modelos via AIC:

```r
extractAIC(m1e)[2]
```

```
## [1] 3489.708
```

```r
extractAIC(m1w)[2]
```

```
## [1] 3341.762
```

```r
extractAIC(m1l)[2]
```

```
## [1] 3316.697
```

O modelo log-normal resulta no menor valor de AIC. Para o modelo log-normal, temos as estimativas:

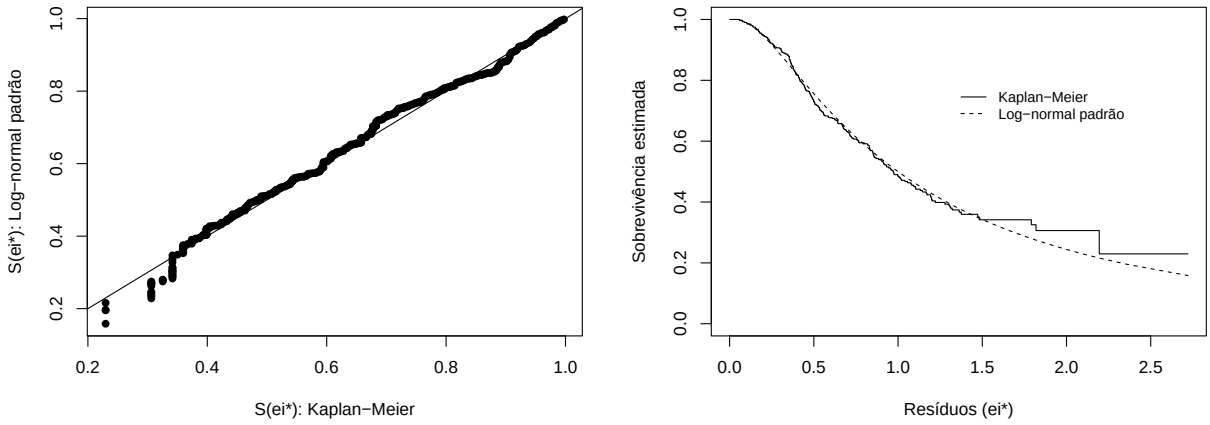
```r
summary(m1l)
```

```
##
## Call:
## survreg(formula = Surv(tempo, cens.1) ~ Idade, data = encefalo,
##     dist = "lognormal")
##               Value Std. Error    z       p
## (Intercept)  7.6594     0.2715 28.2 < 2e-16
## Idade       -0.0321     0.0063 -5.1 3.4e-07
## Log(scale)   0.8853     0.0521 17.0 < 2e-16
##
## Scale= 2.42
##
## Log Normal distribution
## Loglik(model)= -1655.3   Loglik(intercept only)= -1668.2
##  Chisq= 25.66 on 1 degrees of freedom, p= 4.1e-07
## Number of Newton-Raphson Iterations: 3
## n= 397
```

A razão de tempos medianos entre dois indivíduos com diferença de um ano (pacientes com 26 e 25 anos de idade, por exemplo) é dada por $e^{-0,0321} = 0,968$. Isso significa que o tempo mediano de vida vai diminuindo com a idade: pacientes mais jovens apresentam sobrevida superior àquela de pacientes mais velhos.

Na sequência são construídos gráficos dos resíduos para verificar a adequação do modelo log-normal. O gráfico das sobrevivências dos resíduos estimadas por Kaplan-Meier e pelo modelo log-normal padrão encontra-se a seguir. Foi aplicada a transformação exponencial nos resíduos $\hat{\nu}_i$, isto é, $\hat{e}_i^* = \exp\{\hat{\nu}_i\}$.

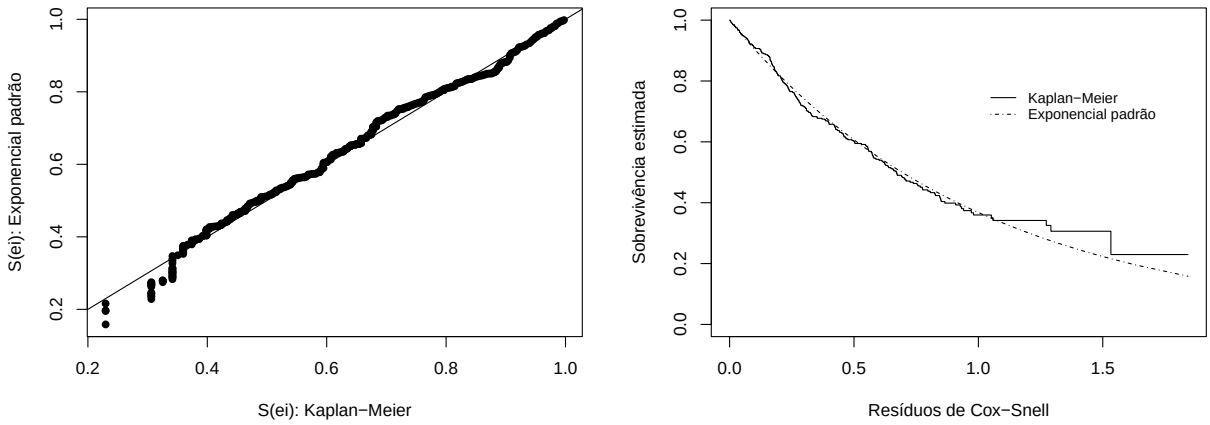
```r
xb <- m1l$coefficients[1] + m1l$coefficients[2]*encefalo$Idade
sigma <- m1l$scale
res <- (log(encefalo$tempo)-(xb))/sigma # residuos padronizados
resid <- exp(res) # exponencial dos residuos padronizados
ekm <- survfit(Surv(resid, encefalo$cens.1) ~ 1)
resid <- ekm$time
sln <- pnorm(-log(resid))
par(mfrow = c(1, 2))
plot(ekm$surv, sln, xlab = "S(ei*): Kaplan-Meier", ylab = "S(ei*): Log-normal padrão",
     pch = 16)
abline(coef = c(0, 1))
plot(ekm, conf.int = F, mark.time = F, xlab = "Resíduos (ei*)",
     ylab = "Sobrevivência estimada", pch = 16)
lines(resid, sln, lty = 2)
legend(1.3, 0.8, lty = c(1, 2), c("Kaplan-Meier", "Log-normal padrão"), cex = 0.8,
       bty = "n")
```



O modelo de regressão log-normal encontra-se relativamente bem ajustado aos dados, já que os valores dos resíduos do modelo proposto são bem próximas àqueles obtidas pelos resíduos obtidos pelo estimador não-paramétrico de Kaplan-Meier.

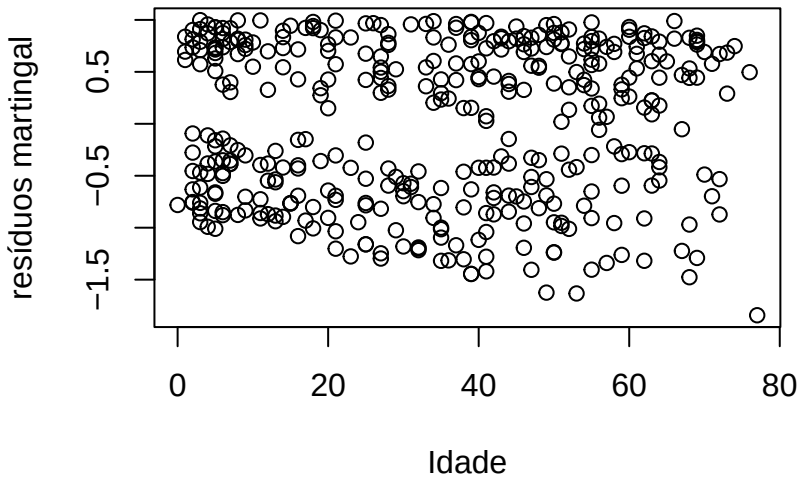
Vamos verificar também os resíduos de resíduos de Cox-Snell que devem ter uma distribuição exponencial padrão caso o modelo log-normal seja adequado..

```
ei <- -log(1-pnorm(res)) # residuos de Cox-Snell
ekm1 <- survfit(Surv(ei, encefalo$cens.1) ~ 1)
t <- ekm1$time
st <- ekm1$surv
sexp <- exp(-t)
par(mfrow = c(1, 2))
plot(st, sexp, xlab = "S(ei): Kaplan-Meier", ylab = "S(ei): Exponencial padrão", pch = 16)
abline(coef = c(0, 1))
plot(ekm1, conf.int = F, mark.time = F, xlab = "Resíduos de Cox-Snell",
     ylab = "Sobrevivência estimada")
lines(t, sexp, lty = 4)
legend(1.0, 0.8, lty = c(1, 4), c("Kaplan-Meier", "Exponencial padrão"), cex = 0.8,
       bty = "n")
```



Uma outra forma de verificação do ajuste do modelo é através do resíduo martingal. O gráfico desse resíduo para o modelo log-normal, com a covariável idade é apresentado na figura a seguir.

```
m <- encefalo$cens.1 - ei
plot(encefalo$Idade, m, xlab = "Idade", ylab = "resíduos martingal")
```



Como não percebemos nenhum padrão nos resíduos concluímos que o modelo está adequado e não se faz necessária nenhuma transformação na covariável idade.