

ANÁLISE DE SOBREVIDA DE PACIENTES COM CÂNCER DE PULMÃO

joseferson da silva barreto

.

Professor: Cleanderson Romualdo Fidelis

Projeto final de análise de sobrevivência apresentado na
última aula da disciplina .

julho,25,2022

Sumário

- Objetivo
- Metodos
- Introdução
- Resultados
- Conclusão

Objetivo

O objetivo é avaliar e observar o tempo até a morte dos indivíduos por câncer no pulmão. O estudo teve como objetivo estudar o tempo até a reincidência de morte por cancer no pulmão no estado de Minas Gerais, buscando a identificação de possíveis fatores que podem estar relacionados ao tempo até a reincidência desses indivíduos.

Metodo

Foi feito a análise exploratória e os tratamentos dos dados para buscar relações entre as variáveis , a limpeza dos dados foi feita no software Rstudio ,utilizando a linguagem R para as demais análises. Para essa análise foi utilizado um banco de dados contendo informações sobre pacientes com câncer no estado de Minas Gerais , após o tratamento dos dados foram adicionados duas novas variáveis **meses** e **falhas** , onde a variável **meses** é referente ao número de meses até a primeira falha acontecer ou seja , será nosso tempo observado, já a variável **falhas** contém a informação se nosso evento falhou ou foi censurado(variável binária, 0 ou 1) onde 0 é caso tenha ocorrido um censura e 1 caso tenha ocorrido uma falha. Foi utilizado o estimador de kaplan-meier , para fazer a comparação das curvas entre os grupos usamos neste projeto o teste log-rank .

Introdução

O câncer de pulmão é uma das maiores causas de morte no mundo, sendo a primeira causa em morte por câncer entre homens no Brasil. A história natural da doença inclui elevada letalidade e evolução agressiva, quase sempre com o paciente chegando ao médico quando a doença já se encontra em fase avançada. Nos últimos 40 anos a taxa de sobrevivência em neoplasias de pulmão melhorou em 8 %, nos Estados Unidos, apesar do avanço nas tecnologias de detecção precoce e no refinamento das técnicas de tratamento. Um desafio contínuo é descrever o papel das variáveis clínicas, assistenciais e demográficas dos pacientes com câncer de pulmão.

A análise de sobrevivência é uma das áreas da estatística que mais cresceu nas últimas duas décadas do século passado. A razão deste crescimento é o desenvolvimento e aprimoramento de técnicas estatísticas combinado com computadores cada vez mais velozes. Uma evidência quantitativa deste sucesso é o número de aplicações de análise de sobrevivência em medicina.

Representação dos Dados de Sobrevivência

Os dados de sobrevivência para o indivíduo i ($i = 1, \dots, n$) sob estudo, são representados, em geral, pelo par (t_i, δ_i) sendo t_i o tempo de falha ou de censura e δ_i a variável indicadora de falha ou censura, isto é

$$\delta_i = \begin{cases} 1 & \text{se } t_i \text{ é um tempo de falha} \\ 0 & \text{se } t_i \text{ é um tempo censurado.} \end{cases}$$

Carregando os pacotes

Vamos iniciar carregando os pacotes que iremos utilizar

```
library(tidyverse)

#rm(list = ls())

library(tidyverse)
```

```

library(plotly)
library(readxl)

dados<-read.csv("analise_de_sobrevivencia.csv",sep = ";",encoding = "latin1")

# table(dados$Descrição.da.Doenca)

# dados %>%
# # dplyr::select(where(is.factor)) %>%
# # glimpse()

pulmao<-dados %>%
  dplyr::filter(Descrição.da.Topografia == "PULMAO, SOE")

pulmao$falhas<-pulmao$Tipo.do.Obito
pulmao$falhas[pulmao$falhas == "CÂNCER"] <- "1"
pulmao$falhas[pulmao$falhas == "NÃO CÂNCER"] <- "0"
pulmao$falhas[pulmao$falhas == "SEM INFORMAÇÃO"] <- "0"
pulmao$falhas[pulmao$falhas == ""] <- "0"

#pulmao %>% select(Idade) %>% table()
#pulmao<-read.csv("pulmaonovot.csv",sep = ",")

```

Convertendo os possíveis fatores

```

# sexo
# grau de instrução
# Estado civil
# nome.profissão
# raca.cor
# nacionalidade
# cidade.endereço
# Descrição.da.Doenca.Infantil
# Código.da.Doenca.Infantil
# Indicador.de.Caso.Raro
# Meio.de.Diagnostico
# Status.Vital
# Metástase.à.distância

pulmao$Sexo<- as.factor(pulmao$Sexo)

# glimpse(pulmao)

#c(pulmao$Sexo,pulmao$Nacionalidade)<-lapply(c(pulmao$Sexo,pulmao$Nacionalidade), as.factor)

pulmao$Nacionalidade<- as.factor(pulmao$Nacionalidade)
pulmao$Raca.Cor<- as.factor(pulmao$Raca.Cor)

```

```

pulmao$Naturalidade<- as.factor(pulmao$Naturalidade)

pulmao$Grau.de.Instrução<- as.factor(pulmao$Grau.de.Instrução)

pulmao$Estado.Civil<- as.factor(pulmao$Estado.Civil)

# pulmao %>% select(
#   Descrição.da.Morfologia
# ) %>% table()

pulmao$Descrição.da.Morfologia<- as.factor(pulmao$Descrição.da.Morfologia)
#
#   Descrição.da.Doença
# ) %>% table()

pulmao$Descrição.da.Doença<- as.factor(pulmao$Descrição.da.Doença)
pulmao$Descrição.da.Doença.Adulto.Jovem<- as.factor(pulmao$Descrição.da.Doença.Adulto.Jovem)
pulmao$Descrição.da.Doença.Infantil<- as.factor(pulmao$Descrição.da.Doença.Infantil)
pulmao$Indicador.de.Caso.Raro<- as.factor(pulmao$Indicador.de.Caso.Raro)
pulmao$Meio.de.Diagnostico<- as.factor(pulmao$Meio.de.Diagnostico)
pulmao$Extensão<- as.factor(pulmao$Extensão)
pulmao$Tipo.do.Obito<- as.factor(pulmao$Tipo.do.Obito)
pulmao$Nome.Profissão<- as.factor(pulmao$Nome.Profissão)

pulmao <- pulmao %>% filter(pulmao$Extensão != 'SEM INFORMAÇÃO' )

```

Selecionando as Colunas que Iremos Utilizar

```

## Registered S3 methods overwritten by 'ggalt':
##   method                from
##   grid.draw.absoluteGrob ggplot2
##   grobHeight.absoluteGrob ggplot2
##   grobWidth.absoluteGrob  ggplot2
##   grobX.absoluteGrob      ggplot2
##   grobY.absoluteGrob      ggplot2

```

Calculando a taxa de censura

Antes de mais nada ,iremos calcular a taxa de censura presente em nossos dados

```
round(1 - mean(pulmao$falhas),3)
```

```
## [1] 0.209
```

```
#library(lubridate)
```

nossa taxa de censura é de aproximadamente 20,9% .

Estimador de Kaplan-Meier

O Esttimador de Kaple-Meier é um estimador não-paramétrico onde a função de sobrevivência será dada da forma

$$\hat{S}(t) = \frac{\text{no.deobservaçõesquenãofalharamatéotempo}(t)}{\text{no.totaldeobservaçõesnoestudo}}$$

```
#library(lubridate)

#pulmao$meses<-as.integer(pulmao$meses)

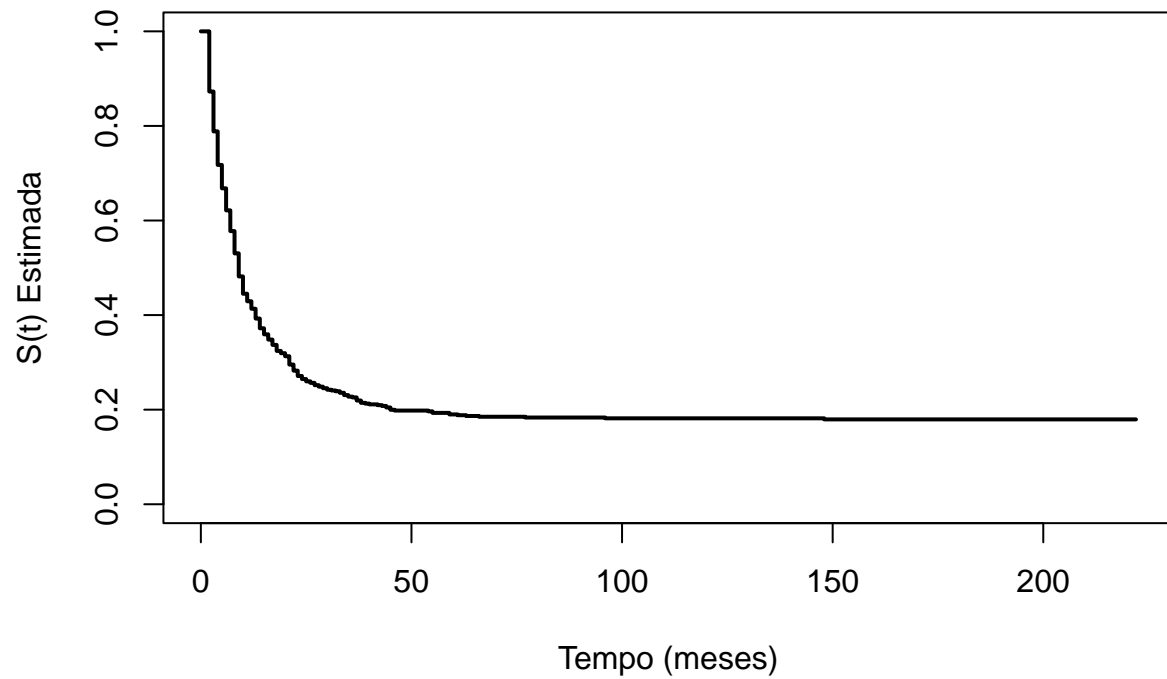
pulmao<- pulmao %>% filter(Extensão != "NÃO SE APLICA")
pulmao$falhas<-as.numeric(pulmao$falhas)
library(survival)

ek <- survfit(Surv(pulmao$meses,pulmao$falhas)~1)
#par(bg = '#1b1a3d')

# plot(ek, lty=1, xlab="Tempo (meses)",
#      # Symbol color
#      col.main = "white",      # Title color
#      col.sub = "white",      # Subtitle color
#      col.lab = "white",      # X and Y-axis labels color
#      col.axis = "white",     # Tick labels color
#      fg = "white",
#      ylab="S(t) Estimada", col = "white",
#      lwd = 2,main = "Curva de Sobrevivência ",conf.int=F)

plot(ek, lty=1, xlab="Tempo (meses)",
#     # Symbol color
#     col.main = "white",      # Title color
#     col.sub = "white",      # Subtitle color
#     col.lab = "white",      # X and Y-axis labels color
#     col.axis = "white",     # Tick labels color
#     fg = "white",
#     ylab="S(t) Estimada", col = 1,
#     lwd = 2,main = "Curva de Sobrevivência ",conf.int=F)
```

Curva de Sobrevivência



```
#plot(ek, lty=1, col = "blue", lwd = 2, conf.int = F)
```

```
#install.packages("ggfortify")
```

```
library(modelsummary)
```

```
##
```

```
## Attaching package: 'modelsummary'
```

```
## The following object is masked from 'package:gt':
```

```
##
```

```
##   escape_latex
```

```
library(tidyverse)
```

```
library(ggfortify)
```

```
## Registered S3 method overwritten by 'ggfortify':
```

```
##   method      from
```

```
##   fortify.table ggalt
```

```
#fortify(ek)
```

Fazendo com Base o fator Extensão

```
# Estimação da função de sobrevivência pelo método de Kaplan-Meier
library(survival)
pulmao$falhas<-as.numeric(pulmao$falhas)
ekm_exp <- survfit(Surv(pulmao$meses,pulmao$falhas)~pulmao$Extensão)

# pulmao[28]<-as.factor(pulmao[28])
# resumo do Kaplan-Meier

#summary(ekm_exp)

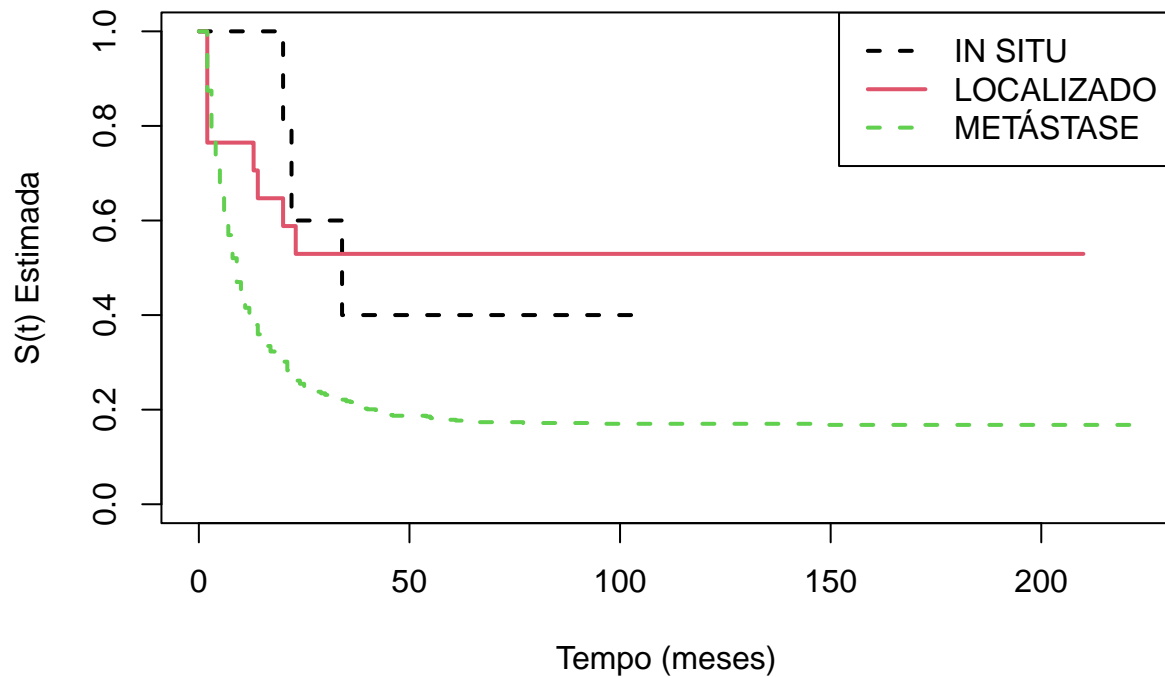
# Gráfico de Kaplan-Meier

# pulmao %>%
#   dplyr::select(where(is.factor)) %>%
#   glimpse()

# table(pulmao$Extensão)

plot(ekm_exp, lty=c(2,1), xlab="Tempo (meses)",
      # Symbol color
#   col.main = "white",      # Title color
#   col.sub = "white",      # Subtitle color
#   col.lab = "white",      # X and Y-axis labels color
#   col.axis = "white",     # Tick labels color
#   fg = "white",
  ylab="S(t) Estimada", col = 1:3,
  lwd = c(2,2),main = "Curva de Sobrevivência com Base no Estágio da Doença")
legend(x = "topright",      # Position
       c("IN SITU", "LOCALIZADO","METÁSTASE"), # Legend texts
       lty = c(2,1),        # Line types
       col = c(1:3),        # Line colors
       lwd = 2)
```


Curva de Sobrevivência com Base no Estágio da Doença



```
# plot(ek, lty=1, xlab="Tempo (meses)",
#      # Symbol color
#      col.main = "white",      # Title color
#      col.sub = "white",      # Subtitle color
#      col.lab = "white",      # X and Y-axis labels color
#      col.axis = "white",      # Tick labels color
#      fg = "white",
#      ylab="S(t) Estimada", col = "white",
#      lwd = 2, main = "Curva de Sobrevivência ", conf.int=F)

# d_km <- tidy(ekm_exp)
# p<-ggplot(d_km, aes(x = time, y = estimate, colour = strata)) +
#   geom_ribbon(aes(ymin = conf.low, ymax = conf.high, fill = strata),
#             stat = 'stepribbon', alpha = .1, size = .1) +
#   geom_step() +
#   theme_bw(16)
#
# library(plotly)
#
# ggplotly(p)
```

Neste Gráfico podemos perceber que após 34 meses após o diagnostico que o paciente possui câncer é melhor pertencer ao grupo onde o câncer no pulmão está localizado, pois após 34 meses é o grupo que possui a maior

probabilidade de sobrevivência, uma vez que quando o câncer é localizado isso significa que não existe sinal de disseminação da doença.

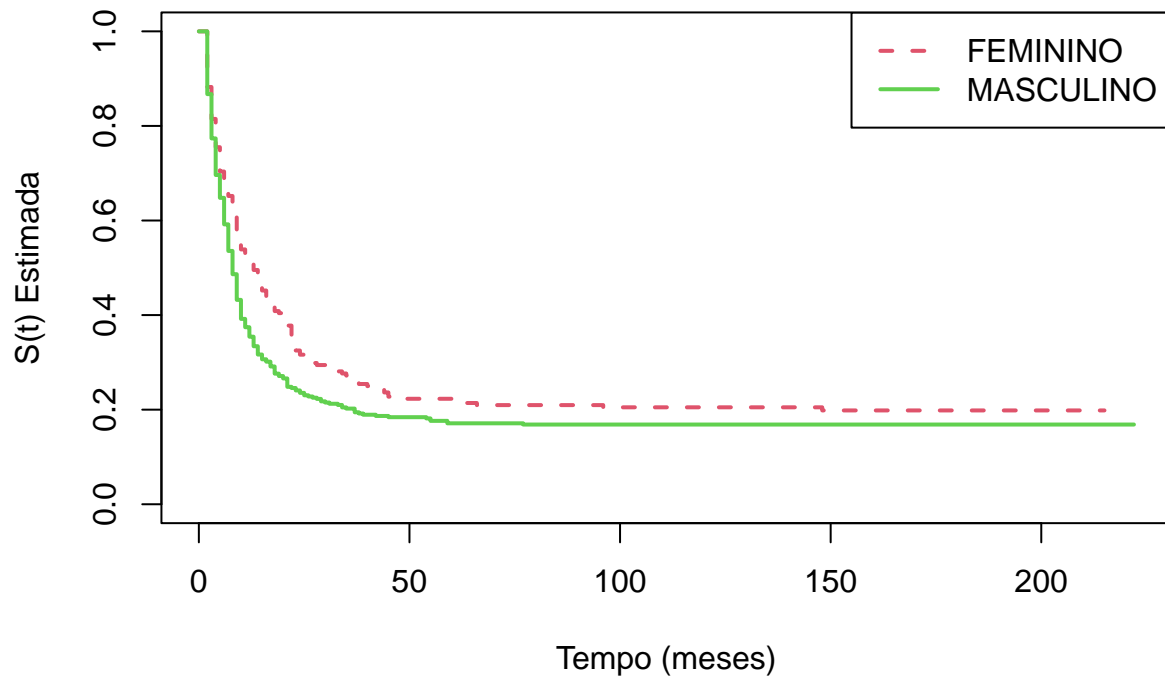
Além disso, os pacientes que possuem câncer em metáfase possuem a menor probabilidade de sobrevivência, tendo em vista que as células cancerosas se destacam do tumor original e viajam através da circulação sanguínea ou linfática, formando novos tumores. Esse processo é conhecido como metástase.

Fazendo o Kaplan-Meier para o variável Sexo

```
KPS <- survfit(Surv(pulmao$meses,pulmao$falhas)~pulmao$Sexo)
#par(bg = '#1b1a3d')

plot(KPS, lty=c(2,1), xlab="Tempo (meses)",
      ylab="S(t) Estimada", col = 2:4,
      # col.main = "white",      # Title color
      # col.sub = "white",      # Subtitle color
      # col.lab = "white",      # X and Y-axis labels color
      # col.axis = "white",      # Tick labels color
      # fg = "white",
      lwd = c(2,2), main = "Curva de Sobrevivência com Base no Sexo")
legend(x = "topright",      # Position
       c("FEMININO", "MASCULINO"), # Legend texts
       lty = c(2,1),      # Line types
       col = c(2:4),      # Line colors
       lwd = 2)#bg= "white")
```

Curva de Sobrevivência com Base no Sexo



```
#install.packages("broom")
library(ggplot2)
library(ggalt)
library(broom)

# d_km <- tidy(KPS)
# p<-ggplot(d_km, aes(x = time, y = estimate, colour = strata)) +
#   geom_ribbon(aes(ymin = conf.low, ymax = conf.high, fill = strata),
#             stat = 'stepribbon', alpha = .1, size = .1) +
#   geom_step() +
#   theme_bw(16)
#
#
# library(plotly)
#
#
# ggplotly(p)
```

Pelos Gráficos acima podemos observar que após O quinto mês após ter sido feito o diagnóstico o grupo feminino começa apresentar uma maior probabilidade de sobrevivência se comparado ao grupo masculino.

vamos comparar as curvas de sobrevivência dos grupos sexo

Teste log-rank

O teste log-rank é apropriado quando a razão das funções de risco dos grupos a serem comparados é, aproximadamente, constante. A estatística do teste log-rank é a diferença entre o número observado de falhas em cada grupo e o correspondente número esperado de falhas sob a hipótese nula

```
survdif(Surv(pulmao$meses,pulmao$falhas)~pulmao$Sexo,rho=0)

## Call:
## survdiff(formula = Surv(pulmao$meses, pulmao$falhas) ~ pulmao$Sexo,
##          rho = 0)
##
##               N Observed Expected (O-E)^2/E (O-E)^2/V
## pulmao$Sexo=FEMININO 238         185         211         3.26         5.91
## pulmao$Sexo=MASCULINO 422         339         313         2.20         5.91
##
##  Chisq= 5.9  on 1 degrees of freedom, p= 0.02
```

calculando o Teste Log Rank para os Grupos de Extensão

Adenocarcinoma. Os adenocarcinomas começam nas células que revestem os alvéolos e produzem substâncias como muco. Esse tipo de câncer de pulmão ocorre principalmente em fumantes e ex-fumantes, mas também é o tipo mais comum em não fumantes. É mais frequente em mulheres do que em homens, e é mais propenso a ocorrer em pessoas mais jovens do que outros tipos de câncer de pulmão. O adenocarcinoma é normalmente encontrado nas áreas externas do pulmão. Ele tende a crescer mais lentamente do que os outros tipos de câncer de pulmão, e tem mais chance de ser diagnosticado antes de se disseminar.

Logo, será nosso grupo de referência

```
#colnames(pulmao)[14]<-'Extensão'
library(tidyverse)
library(survival)
teste1<- pulmao %>% filter(Extensão != "METÁSTASE" )
grupo2<- pulmao %>% filter(Extensão != "LOCALIZADO" )
caso3<- pulmao %>% filter(Extensão != "IN SITU" )

#survdif(Surv(pulmao$meses,pulmao$falhas)~pulmao$Extensão,rho=0)
```

grupo 1 vs Grupo 2

```
survdif(Surv(teste1$meses,teste1$falhas)~teste1$Extensão,rho=0)

## Call:
## survdiff(formula = Surv(teste1$meses, teste1$falhas) ~ teste1$Extensão,
##          rho = 0)
##
```

```
##               N Observed Expected (O-E)^2/E (O-E)^2/V
## teste1$Extensão=IN SITU      5         3    2.87   0.00566   0.00821
## teste1$Extensão=LOCALIZADO 17         8    8.13   0.00200   0.00821
##
## Chisq= 0   on 1 degrees of freedom, p= 0.9
```

A 5% de significância rejeita-se a hipótese H_0 , logo, não existe diferença entre as curvas de sobrevivência dos grupos IN SITU e do grupo localizado.

grupo 1 vs Grupo 2

```
survdif(Surv(grupo2$meses, grupo2$falhas)~grupo2$Extensão, rho=0)
```

```
## Call:
## survdiff(formula = Surv(grupo2$meses, grupo2$falhas) ~ grupo2$Extensão,
##           rho = 0)
##
##               N Observed Expected (O-E)^2/E (O-E)^2/V
## grupo2$Extensão=IN SITU      5         3    7.26   2.4956   2.7
## grupo2$Extensão=METÁSTASE 638      513   508.74   0.0356   2.7
##
## Chisq= 2.7   on 1 degrees of freedom, p= 0.1
```

A 5% de significância rejeita-se a hipótese H_0 , logo, não existe diferença entre as curvas de sobrevivência dos grupos IN SITU e do grupo metástase.

grupo 2 vs Grupo 3

```
survdif(Surv(caso3$meses, caso3$falhas)~caso3$Extensão, rho=0)
```

```
## Call:
## survdiff(formula = Surv(caso3$meses, caso3$falhas) ~ caso3$Extensão,
##           rho = 0)
##
##               N Observed Expected (O-E)^2/E (O-E)^2/V
## caso3$Extensão=LOCALIZADO 17         8   19.8    7.055    7.85
## caso3$Extensão=METÁSTASE 638      513   501.2    0.279    7.85
##
## Chisq= 7.9   on 1 degrees of freedom, p= 0.005
```

Logo, rejeita-se a hipótese nula a 5% de significância, ou seja, as curvas de sobrevivência apresentaram diferença significativa.

Ajustes de modelos

Teste da Razão de Verossimilhança

É conhecido que existem técnicas gráficas que são bastante úteis na seleção de modelos. Entretanto, conclusões a partir delas podem diferir para diferentes analistas. Ou seja, existem nas técnicas gráficas um componente subjetivo na sua interpretação. Outra forma de selecionar modelos é através de testes de hipóteses, onde as hipóteses a serem testadas são:

H_0 : o modelo de interesse é adequado

H_1 : O modelo escolhido não é adequado.

```
## Carregando pacotes exigidos: zoo
```

```
##
```

```
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
```

```
##
```

```
##      as.Date, as.Date.numeric
```

```
# H_1: O modelo escolhido não é adequado.
```

```
ekm <- survfit(Surv(pulmao$meses,pulmao$falhas)~1)
```

```
ajust1<-survreg(Surv(pulmao$meses,pulmao$falhas)~1,dist='exponential')
```

```
fit_exp<- ajust1
```

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

```
alpha
```

```
## (Intercept)
```

```
##      46.84542
```

```
ajust2<-survreg(Surv(pulmao$meses,pulmao$falhas)~1,dist='weibull')
```

```
fit_wei <-ajust2
```

```
alpha2<-exp(ajust2$coefficients[1])
```

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

```
ajust3<-survreg(Surv(pulmao$meses,pulmao$falhas)~1,dist='lognorm')
```

```
# ajust7<-survreg(Surv(pulmao$meses,pulmao$falhas)~+factor(pulmao$Sexo)+factor(pulmao$Extensão),dist='
```

```
fit_log<- ajust3
```

```
mi<-ajust3$coefficients
```

```
sigmaa<-ajust3$scale
```

```
library(flexsurv)
```

```
ajust4<-flexsurvreg(Surv(as.integer(pulmao$dias),pulmao$falhas)~1,  
                    dist="gengamma")
```

```
fit_gammagen<-ajust4
```

```
ajust5<-flexsurvreg(Surv(as.integer(pulmao$dias),pulmao$falhas)~1,  
                    dist="gamma")
```

```
fit_gamma <-ajust5
```

```
lrtest(ajust1,ajust4)
```

```
## Warning in modelUpdate(objects[[i - 1]], objects[[i]]): original model was of  
## class "survreg", updated model is of class "flexsurvreg"
```

```
## Likelihood ratio test
```

```
##
```

```
## Model 1: Surv(pulmao$meses, pulmao$falhas) ~ 1
```

```
## Model 2: Surv(as.integer(pulmao$dias), pulmao$falhas) ~ 1
```

```
##   #Df LogLik Df  Chisq Pr(>Chisq)
```

```
## 1    1 -2539.8
```

```
## 2    3 -3830.4  2 2581.4  < 2.2e-16 ***
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(ajust2,ajust4)
```

```
## Warning in modelUpdate(objects[[i - 1]], objects[[i]]): original model was of  
## class "survreg", updated model is of class "flexsurvreg"
```

```
## Likelihood ratio test
```

```
##
```

```
## Model 1: Surv(pulmao$meses, pulmao$falhas) ~ 1
```

```
## Model 2: Surv(as.integer(pulmao$dias), pulmao$falhas) ~ 1
```

```
##   #Df LogLik Df  Chisq Pr(>Chisq)
```

```
## 1    2 -2321.7
```

```
## 2    3 -3830.4  1 3017.5  < 2.2e-16 ***
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(ajust3,ajust4)
```

```
## Warning in modelUpdate(objects[[i - 1]], objects[[i]]): original model was of  
## class "survreg", updated model is of class "flexsurvreg"
```

```
## Likelihood ratio test
```

```
##
```

```
## Model 1: Surv(pulmao$meses, pulmao$falhas) ~ 1
```

```
## Model 2: Surv(as.integer(pulmao$dias), pulmao$falhas) ~ 1
```

```
##   #Df LogLik Df  Chisq Pr(>Chisq)
```

```
## 1    2 -2205.9
```

```
## 2    3 -3830.4  1 3249.1  < 2.2e-16 ***
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```

time<-ekm$time
st<-ekm$surv
stexp<- exp(-time/alpha)
stwei<- exp(-(time/alpha2)^gama)
stlnoln<- pnorm((-log(time)+ mi)/sigmaa)
# cbind(time,st,stexp,stwei,stlnoln)

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

plot(ek, conf.int=F, xlab="Tempos", ylab="S(t)")
lines(c(0,time),c(1,stexp), lty=2)
legend(25,0.8,lty=c(1,2),
       c("Kaplan-Meier", "exponencial"),
       bty="n",cex=0.8)

#=====
# par(bg = '#1b1a3d')
#
# par(mfrow=c(2,2))
#
# plot(ek, conf.int=F, xlab="Tempos",
#      ylab="S(t)",
#      col = "white",
#      col.main = "white",      # Title color
#      col.sub = "white",      # Subtitle color
#      col.lab = "white",      # X and Y-axis labels color
#      col.axis = "white",      # Tick labels color
#      fg = "white")
# lines(c(0,time),c(1,stwei), lty=2,col = "white")
# legend(25,0.8,lty=c(1,2),
#      c("Kaplan-Meier", "Weibull"),
#      bty="n",cex=0.8,col = "white",
#      text.col = "white")
#
#
# plot(ek, conf.int=F, xlab="Tempos", ylab="S(t)",

```



```

#      col = "white",
#      col.main = "white",      # Title color
#      col.sub = "white",      # Subtitle color
#      col.lab = "white",      # X and Y-axis labels color
#      col.axis = "white",     # Tick labels color
#      fg = "white")
# lines(c(0,time),c(1,stlnoln), lty=2,col = "white")
# legend(25,0.8,lty=c(1,2),
#       c("Kaplan-Meier", "Log-normal"),
#       bty="n",cex=0.8,col = "white",text.col = "white")
#
#
#
# plot(ek, conf.int=F, xlab="Tempos", ylab="S(t)",
#      col = "white",
#      col.main = "white",      # Title color
#      col.sub = "white",      # Subtitle color
#      col.lab = "white",      # X and Y-axis labels color
#      col.axis = "white",     # Tick labels color
#      fg = "white")
# lines(c(0,time),c(1,steap), lty=2, col = "white")
# legend(25,0.8,lty=c(1,2),
#       c("Kaplan-Meier", "exponencial"),
#       bty="n",cex=0.8,col = "white",text.col = "white")

```

```
anova(ajust1,ajust3)
```

##	Terms	Resid. Df	-2*LL Test	Df	Deviance	Pr(>Chi)
## 1	1	659	5079.502	NA	NA	NA
## 2	1	658	4411.776	= 1	667.7267	3.118803e-147

```
anova(ajust2,ajust3)
```

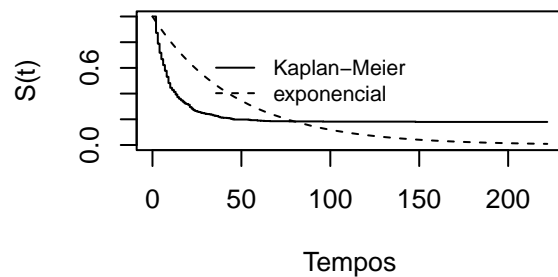
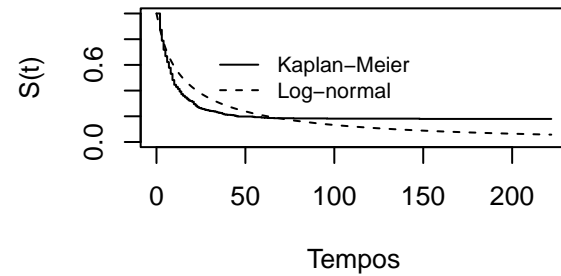
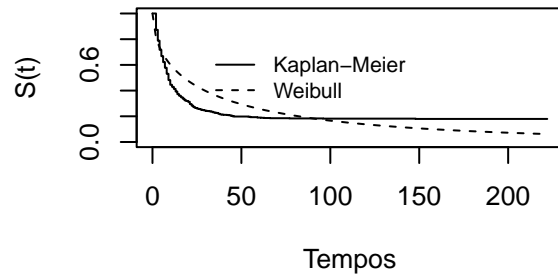
##	Terms	Resid. Df	-2*LL Test	Df	Deviance	Pr(>Chi)
## 1	1	658	4643.425	NA	NA	NA
## 2	1	658	4411.776	= 0	231.6495	NA

```

# plot(ek, conf.int=F, xlab="Tempos", ylab="S(t)")
# lines(c(0,time),c(1,steap), lty=2)
# legend(25,0.8,lty=c(1,2),
#       c("Kaplan-Meier", "exponencial"),
#       bty="n",cex=0.8)
#
# Modelo = c("Gama Generalizado", "Exponencial", "Log-Normal", "Weibull", "gamma")
# Verossimilhanca = c(fit_gammagen$loglik, unique(fit_exp$loglik), unique(fit_log$loglik), unique(fit_w
# TRV = 2*(fit_gammagen$loglik-Verossimilhanca)
# valor_p = pchisq(TRV,df=2,lower.tail=FALSE) ##>% round(2)
#
# resultado = data.frame(Modelo=Modelo,
#                          Verossimilhanca = Verossimilhanca,
#                          TRV=TRV,

```

```
#                               valor_p=valor_p)
#
# ggpubr::ggtexttable(resultado, rows=NULL)
```



```
ekm <- survfit(Surv(pulmao$meses,pulmao$falhas)~1)
ajust1<-survreg(Surv(pulmao$meses,pulmao$falhas)~1,dist='exponential')
alpha<-exp(ajust1$coefficients[1])
alpha
```

```
## (Intercept)
##      46.84542
```

```
ajust2<-survreg(Surv(pulmao$meses,pulmao$falhas)~1,dist='weibull')
alpha2<-exp(ajust2$coefficients[1])
gama<-1/ajust2$scale

ajust3<-survreg(Surv(pulmao$meses,pulmao$falhas)~1,dist='lognorm')
mi<-ajust3$coefficients
sigmaa<-ajust3$scale

time<-ekm$time
st<-ekm$surv
stexp<- exp(-time/alpha)
stwei<- exp(-(time/alpha2)^gama)
```

```

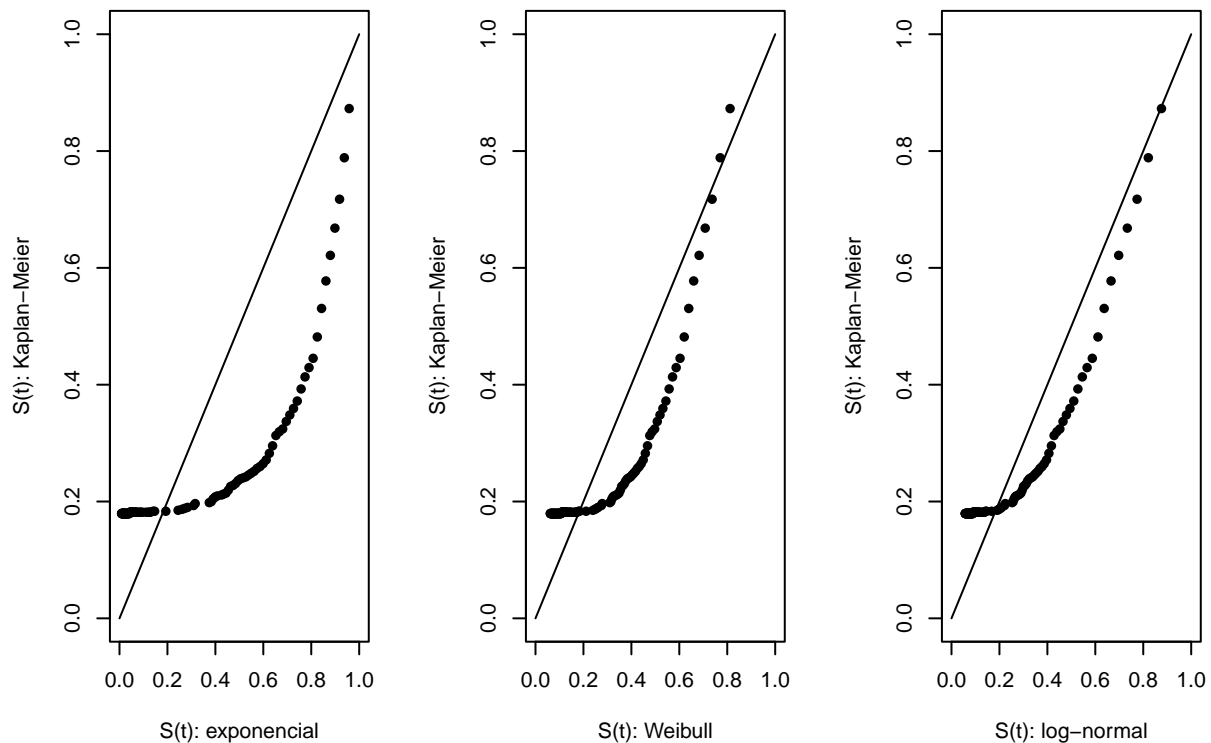
stlnoln<- pnorm((-log(time)+ mi)/sigmaa)
#cbind(time,st,stexp,stwei,stlnoln)

#gráficos
par(mfrow=c(1,3))
plot(stexp,st,pch=16,ylim = (c(0,1)), xlim = (c(0,1)),
     ylab = "S(t): Kaplan-Meier", xlab="S(t): exponencial")
lines(c(0,1), c(0,1), type="l", lty=1)

plot(stwei,st,pch=16,ylim = (c(0,1)), xlim = (c(0,1)),
     ylab = "S(t): Kaplan-Meier", xlab="S(t): Weibull")
lines(c(0,1), c(0,1), type="l", lty=1)

plot(stlnoln,st,pch=16,ylim = (c(0,1)), xlim = (c(0,1)),
     ylab = "S(t): Kaplan-Meier", xlab="S(t): log-normal")
lines(c(0,1), c(0,1), type="l", lty=1)

```



```

# par(mfrow=c(1,3))
# par(bg = '#1b1a3d')
# plot(stexp,st,pch=16,ylim = (c(0,1)), xlim = (c(0,1)),
#      ylab = "S(t): Kaplan-Meier", xlab="S(t): exponencial",
#      col = "white",
#      col.main = "white",      # Title color
#      col.sub = "white",      # Subtitle color

```

```

# col.lab = "white", # X and Y-axis labels color
# col.axis = "white", # Tick labels color
# fg = "white")
# lines(c(0,1), c(0,1), type="l", lty=1,col= "white")
#
#
# plot(stwei,st,pch=16,ylim = (c(0,1)), xlim = (c(0,1)),
# ylab = "S(t): Kaplan-Meier", xlab="S(t): Weibull",
# col="white",
# col.main = "white", # Title color
# col.sub = "white", # Subtitle color
# col.lab = "white", # X and Y-axis labels color
# col.axis = "white", # Tick labels color
# fg = "white")
# lines(c(0,1), c(0,1), type="l", lty=1, col ="white")
#
# plot(stlnoln,st,pch=16,ylim = (c(0,1)), xlim = (c(0,1)),
# ylab = "S(t): Kaplan-Meier", xlab="S(t): log-normal",
# col="white",
# col.main = "white", # Title color
# col.sub = "white", # Subtitle color
# col.lab = "white", # X and Y-axis labels color
# col.axis = "white", # Tick labels color
# fg = "white")
# lines(c(0,1), c(0,1), type="l", lty=1, col ="white")

# require(survival)
# ekm <- survfit(Surv(tempo,cens)~1)
# tempo<-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)
#
#
#
# ajust1<-survreg(Surv(tempo,cens)~1,dist='exponential')
# alpha<-exp(ajust1$coefficients[1])
# alpha
#
# ajust2<-survreg(Surv(tempo,cens)~1,dist='weibull')
# alpha2<-exp(ajust2$coefficients[1])
# gama<-1/ajust2$scale
#
# ajust3<-survreg(Surv(tempo,cens)~1,dist='lognorm')
# mi<-ajust3$coefficients
# sigmaa<-ajust3$scale
#
# time<-ekm$time
# st<-ekm$surv
# stexp<- exp(-time/alpha)
# stwei<- exp(-(time/alpha2)^gama)
# stlnoln<- pnorm((-log(time)+ mi)/sigmaa)
# cbind(time,st,stexp,stwei,stlnoln)
#
#

```

```
# #gráficos
# par(mfrow=c(1,3))
# plot(stexp,st,pch=16,ylim = (c(0,1)), xlim = (c(0,1)),
#      ylab = "S(t): Kaplan-Meier", xlab="S(t): exponencial")
# lines(c(0,1), c(0,1), type="l", lty=1)
#
# plot(stwei,st,pch=16,ylim = (c(0,1)), xlim = (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 = (c(0,1)), xlim = (c(0,1)),
#      ylab = "S(t): Kaplan-Meier", xlab="S(t): log-normal")
# lines(c(0,1), c(0,1), type="l", lty=1)
```

Modelo de Cox

um dos pressupostos do modelo de cox é a independência das curvas de sobrevivência , ou seja , as curvas não podem se cruzar , logo , apenas a variável sexo pode ser uma covariável do modelo .

```
# cox <- coxph(Surv(meses,falhas)~Sexo,
#              data = pulmao)
#
# summary(cox)
```

```
cox <- coxph(Surv(meses,falhas)~Sexo,
             data = pulmao)

summary(cox)
```

```
## Call:
## coxph(formula = Surv(meses, falhas) ~ Sexo, data = pulmao)
##
##    n= 660, number of events= 524
##
##              coef exp(coef) se(coef)      z Pr(>|z|)
## SexoMASCULINO 0.22466   1.25190  0.09164  2.452   0.0142 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## SexoMASCULINO      1.252      0.7988      1.046      1.498
##
## Concordance= 0.534 (se = 0.012 )
## Likelihood ratio test= 6.13  on 1 df,  p=0.01
## Wald test              = 6.01  on 1 df,  p=0.01
## Score (logrank) test = 6.04  on 1 df,  p=0.01
```

Após as análises verificamos que a variável extensão não é significativa para o modelo , segundo o nosso modelo , os pacientes do sexo masculino possui 0.22 vezes maior a probabilidade por morte por cancer no pulmão

```
library(tidyverse)
# residuals(cox,type="scaledsch")
teste_ph <- cox.zph(cox)
par(mfrow=c(2,3))
teste_ph$table %>% round(3) %>% ggpubr::ggtexttable()
```

	chisq	df	p
Sexo	2.172	1	0.141
GLOBAL	2.172	1	0.141

```
#install.packages("survminer")
library("survminer")
```

```
## Carregando pacotes exigidos: ggpubr
```

```
##
```

```
## Attaching package: 'survminer'
```

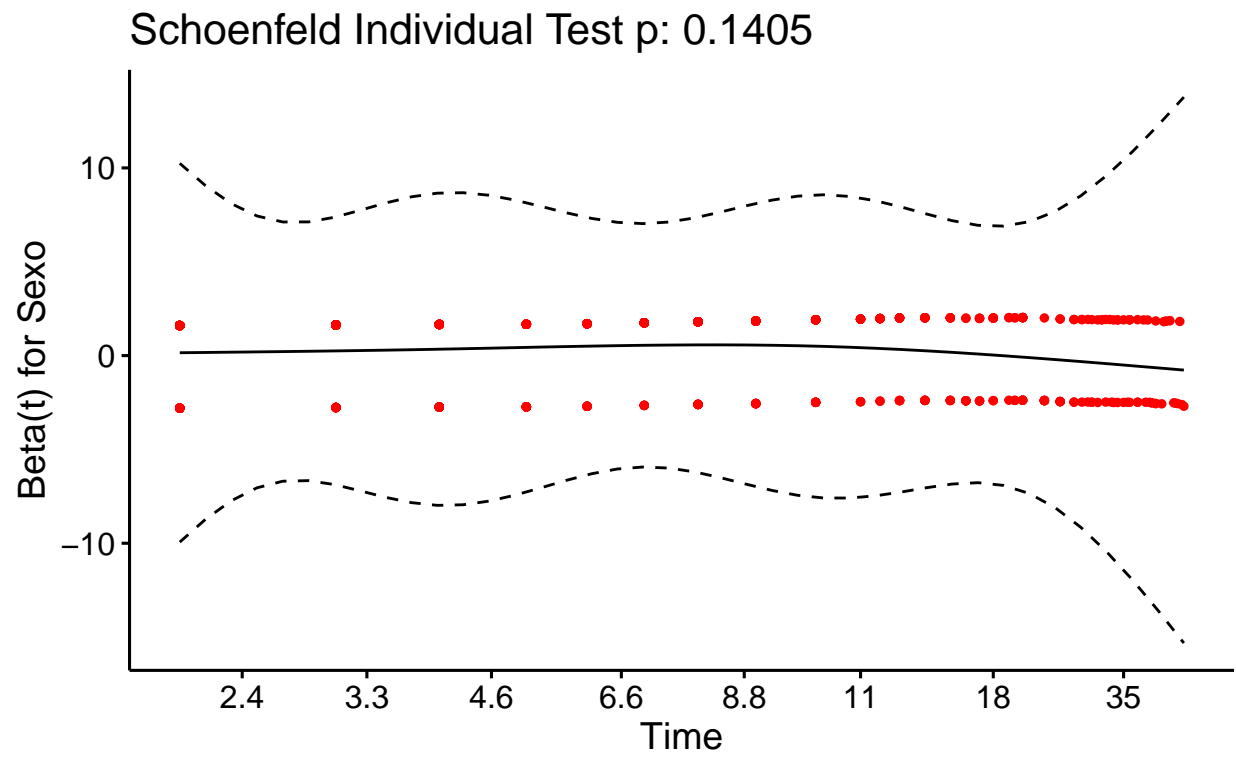
```
## The following object is masked from 'package:survival':
```

```
##
```

```
## myeloma
```

```
testph<-cox.zph(cox)
ggcoxzph(testph)
```

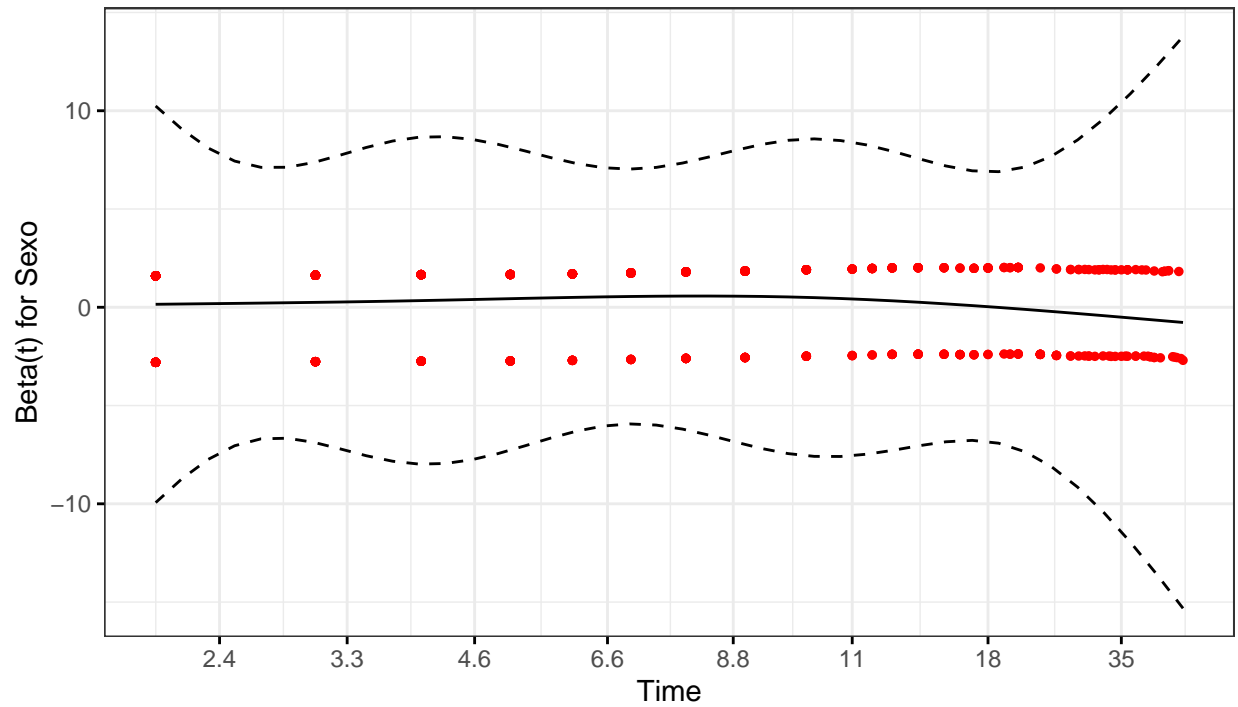
Global Schoenfeld Test p: 0.1405



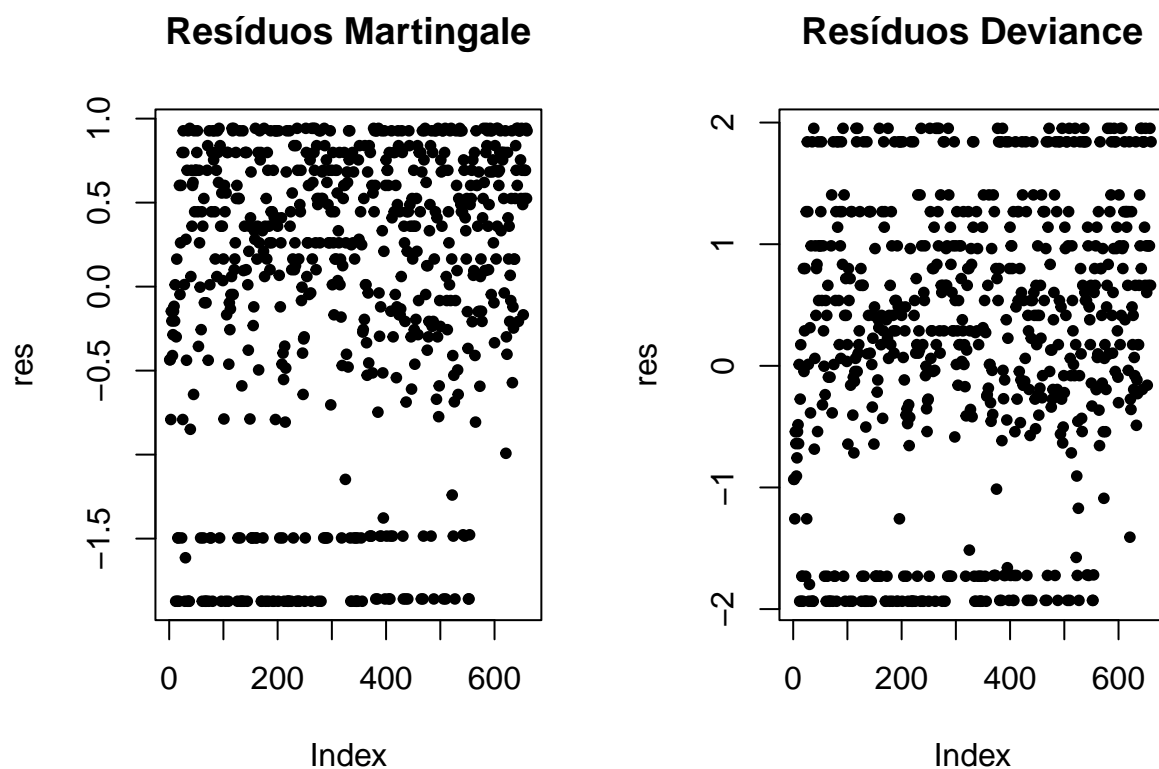
```
ggcoxzph(teste_ph, ggtheme = theme_bw())
```

Global Schoenfeld Test p: 0.1405

Schoenfeld Individual Test p: 0.1405



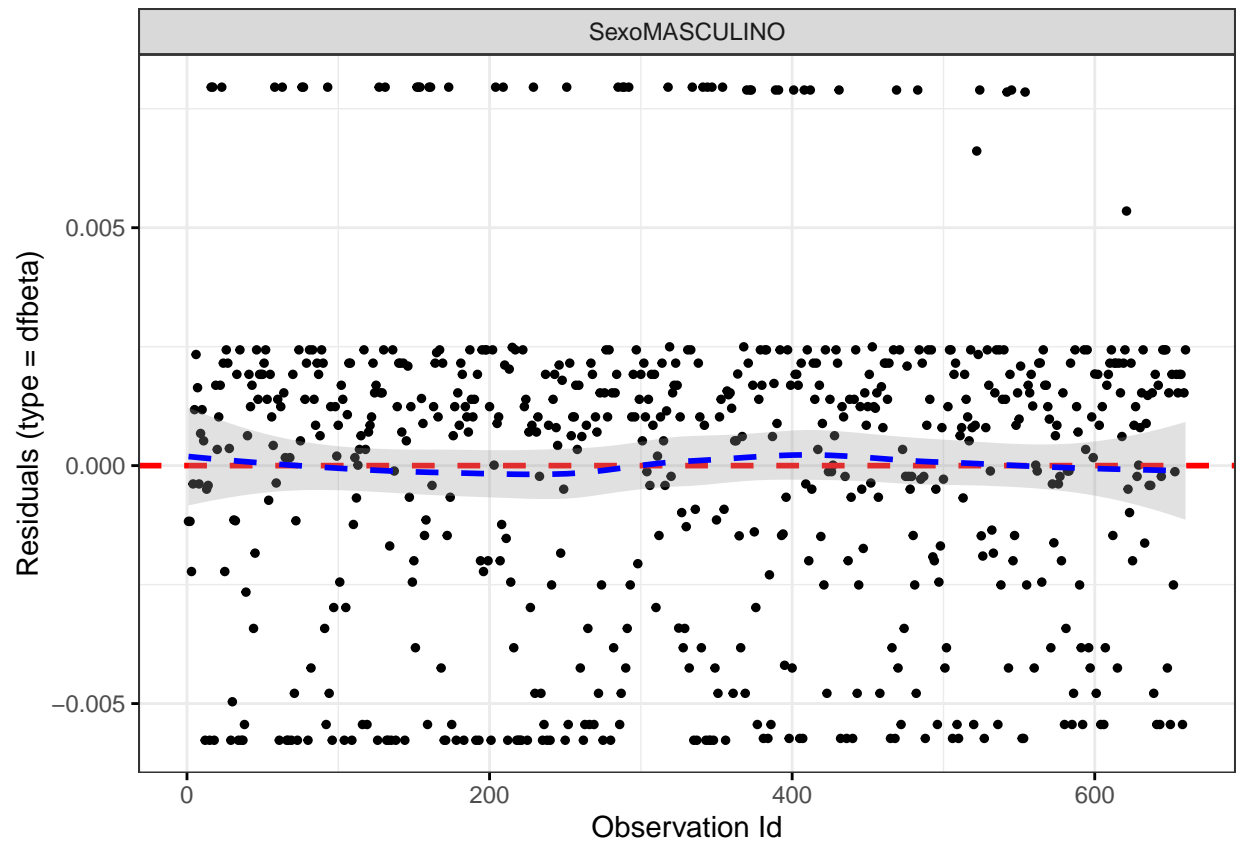
```
par(mfrow=c(1,2))
mart<-residuals(cox,type="martingale")
dev<-residuals(cox,type="deviance")
plot(mart,ylab="res",pch=20)
title("Resíduos Martingale")
plot(dev,ylab="res",pch=20)
title("Resíduos Deviance")
```

```
ggcoxdiagnostics(cox, type = "dfbeta",
                 linear.predictions = FALSE, ggtheme = theme_bw())
```

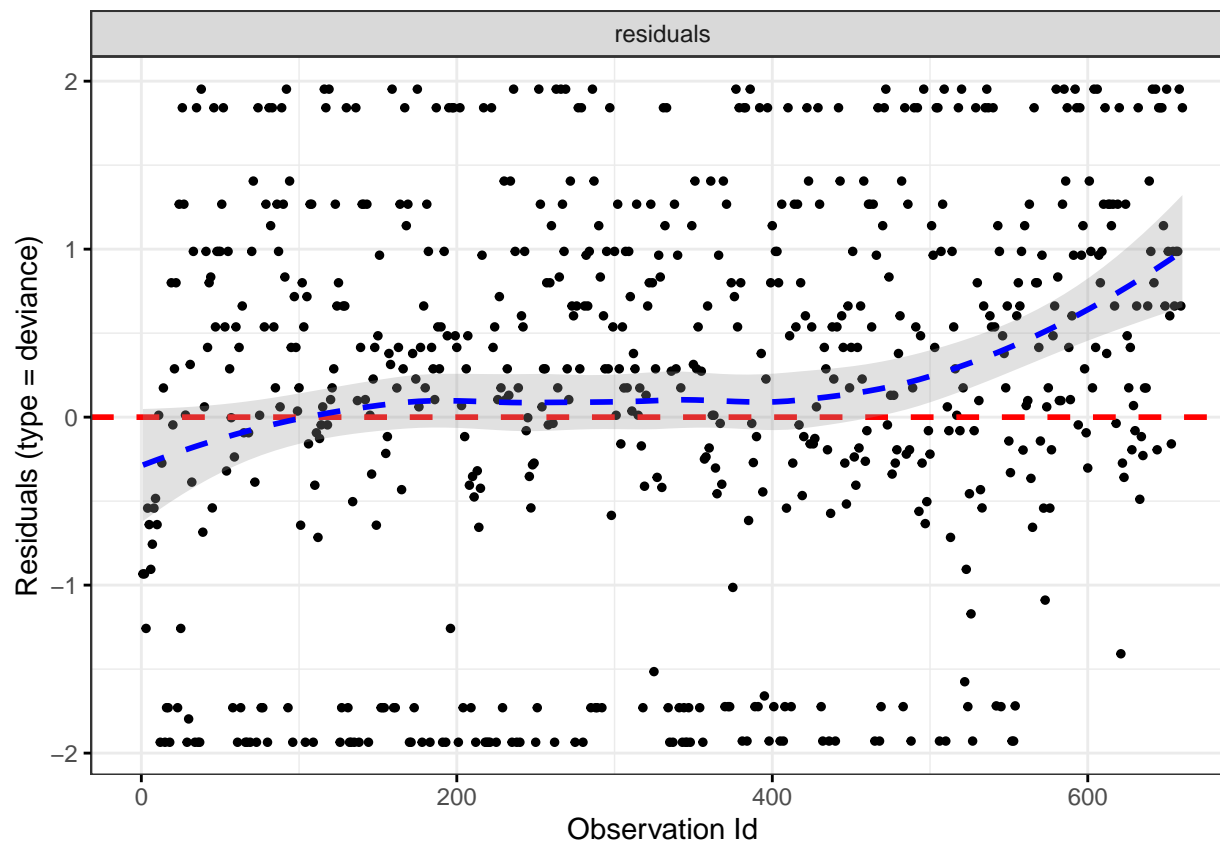
```
## Warning: 'gather_()' was deprecated in tidyr 1.2.0.
## Please use 'gather()' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was generated.
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



```
ggcoxdiagnostics(cox, type = "deviance",
  linear.predictions = FALSE, ggtheme = theme_bw())
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



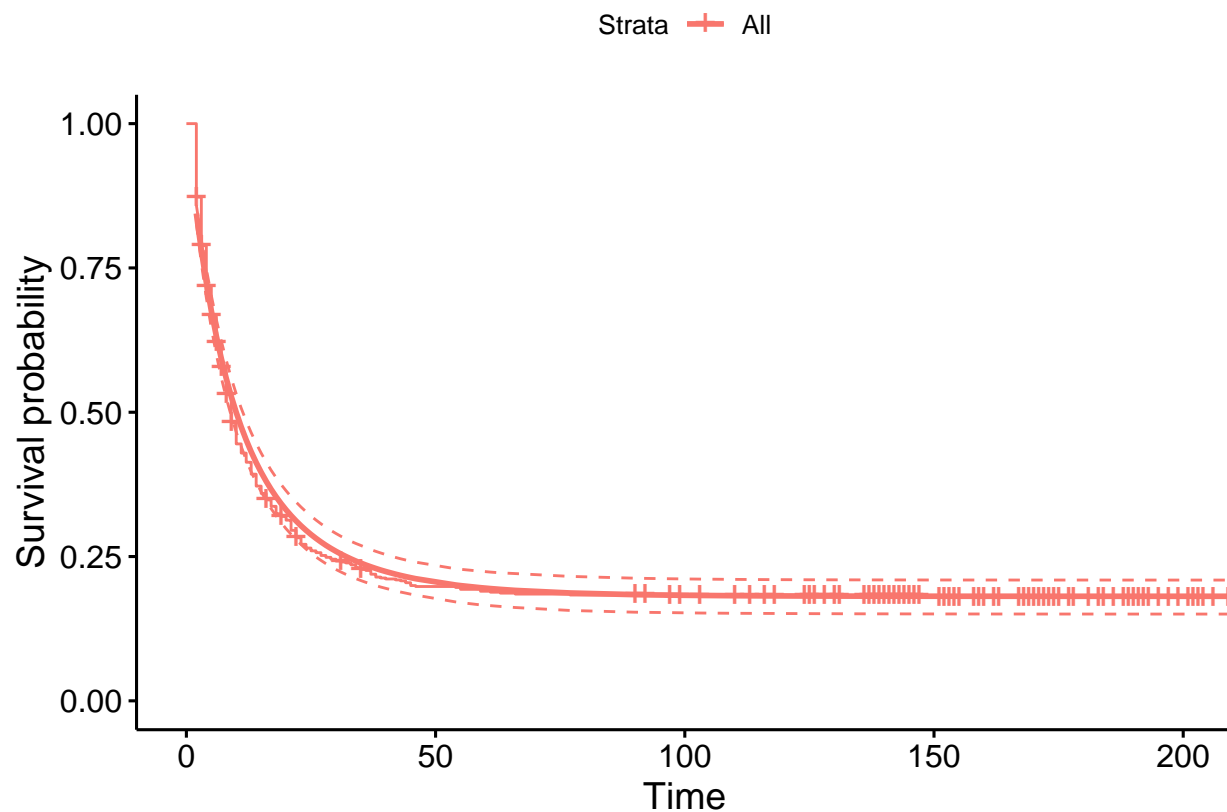
Podemos observar que nossos resíduos não seguem o comportamento esperado, logo o modelo de cox aparentemente não é auto explicativo para os nossos dados, entretanto, após algumas pesquisas encontramos uma distribuição para dados não-lineares que aparentemente apresenta um bom ajuste aos dados.

Distribuição de Gompertz

A distribuição de Gompertz é uma distribuição de probabilidade contínua, nomeada em homenagem a Benjamin Gompertz. A distribuição de Gompertz é frequentemente aplicada para descrever a distribuição da expectativa de vida de adultos por demógrafos e atuários. Campos da ciência relacionados, como biologia e gerontologia, também consideraram a distribuição de Gompertz para a análise de sobrevivência. Mais recentemente, os cientistas da computação também começaram a modelar as taxas de falha do código de computador pela distribuição Gompertz.

```
library(survminer)
library(flexsurv)
gomp<-flexsurvreg(Surv(pulmao$meses,pulmao$falhas)~1, data=pulmao, dist= "gompertz")

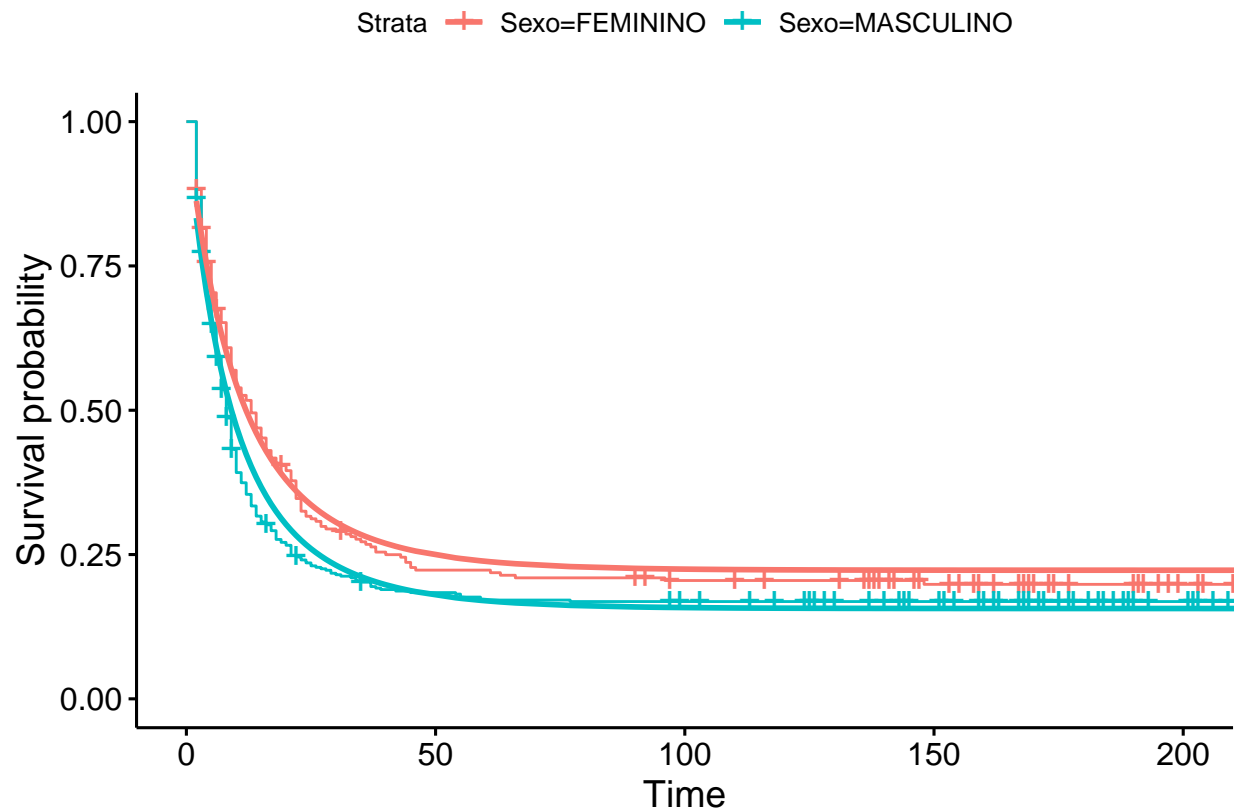
survminer::ggflexsurvplot(gomp)
```



Aparentemente observando pelo gráfico acima o modelo de Gompertz parece se ajustar de forma mais adequada aos nossos dados, vamos observar como ficaria um possível modelo final levando em consideração fatores que foram significativos anteriormente ou ficaram na região de aceitação ou rejeição pelo motivo do pvalor ter dado muito próximo a 0.05.

```
library(survminer)
library(flexsurv)
gomp<-flexsurvreg(Surv(pulmao$meses,pulmao$falhas)~Sexo
                  , data=pulmao, dist= "gompertz")

survminer::ggflexsurvplot(gomp)
```



Pelo gráfico podemos observar que o modelo se ajusta melhor aos dados referente ao sexo masculino do que os dados referente ao sexo feminino, talvez isso aconteça por termos uma maior quantidade de informações sobre o sexo masculino em nossos dados em relação ao sexo feminino

```
table(pulmao$Sexo)
```

```
##
##  FEMININO MASCULINO
##      238      422
```

Conclusão

Após varias análise concluímos que os modelos de distribuição de probabilidades , Exponencial, Weibull e Log Normal não se ajustaram adequadamente aos dados, o mesmo aconteceu com o modelo de cox , a opção mais viável seria o modelo para dados não lineares baseado na distribuição de Gompertz, mas isso necessitaria de uma análise mais profunda. Contudo, observamos que o grupo masculino tem uma menor probabilidade de sobrevivência se comparado ao grupo feminino ao longo do tempo , o mesmo acontece para os pacientes que estão com câncer no pulmão em metástase, que tem uma menor probabilidade de sobrevivência se comparado aos pacientes que possui câncer no pulmao IN SITU e os pacientes com câncer no pulmão Localizado.

Referências

Bemmaor, Albert C .; Glady, Nicolas (2011). “Implementando o modelo Gamma / Gompertz / NBD no MATLAB” (PDF) . Cergy-Pontoise: ESSEC Business School.

Colosimo, Enrico Antônio.; Giolo, Suely Ruiz. “Análise de Sobrevivência Aplicada” (PDF) . 2006

Gompertz, B. (1825). “Sobre a natureza da função expressiva da lei da mortalidade humana e sobre um novo modo de determinar o valor das contingências da vida” . Philosophical Transactions of the Royal Society of London . 115 : 513–583. doi : 10.1098 / rstl.1825.0026 . JSTOR 107756

Johnson, Norman L .; Kotz, Samuel; Balakrishnan, N. (1995). Distribuições univariadas contínuas . 2 (2^a ed.). Nova York: John Wiley & Sons. pp. 25–26. ISBN 0-471-58494-0.