MAE0514 - Introducão a Análise de Sobrevivência - Lista 3

Bruno de Castro Paul Schultze[1]
Rubens Santos Andrade Filho[2]

Junho de 2021

Sumário

[1]Número USP: 10736862
[2]Número USP: 10370336

Questão 1

Questão 2

Questão 2

2.

Temos que: $X \sim exp(\lambda)$

$Y \sim exp(\theta)$

X e Y independentes

$\delta = I(X \leq Y) \qquad Z = min(X, Y)$

2.a

Sem perda de generalidade, no lugar de $P(X \in [t, t+\Delta t])$ com $\Delta t \to 0$, vamos escrever simplesmente $P(X = t)$.

Assim:

$$P(\delta = 1) = \int_0^\infty P(X = t, Y \geq t)\, dt \overset{ind}{=} \int_0^\infty P(X = t)\, P(Y \geq t)\, dt$$

$$P(X = t) = \lambda e^{-\lambda t}$$

$$P(Y \geq t) = e^{-\theta t}$$

$$P(\delta = 1) = \int_0^\infty \lambda e^{-(\lambda+\theta)t}\, dt$$

$$P(\delta = 1) = \lambda \cdot \left[-\frac{e^{-(\lambda+\theta)t}}{(\lambda+\theta)} \right]_{t=0}^{t \to \infty}$$

Lembremos que: $\lim_{t\to\infty} \frac{e^{-at}}{a} = 0$, se $a > 0$

Assim:

$$P(\delta = 1) = \frac{\lambda}{\lambda+\theta}$$

2. b

$$P(Z > z) = P(X > z \; ; \; Y > z) \overset{ind}{=} P(X > z)\,P(Y > z)$$

$$P(Z > z) = e^{-(\lambda+\theta)z}, \quad z > 0$$

$$f_z(z) = \frac{d}{dz} F_z(z) = \frac{d}{dz}\left[1 - e^{-(\lambda+\theta)z}\right]$$

Assim:

$$F_z(z) = 1 - e^{-(\lambda+\theta)z}$$

$$f_z(z) = (\lambda+\theta)\, e^{-(\lambda+\theta)z}, \quad z > 0$$

$\therefore Z \sim \exp(\lambda+\theta)$

2.c

Se forem independentes:

$$P(\delta = k \mid Z = z) = P(\delta = k) \quad \forall \; k \in \{0,1\}, \; z > 0$$

Para $k = 1$:

$$P(\delta = 1 \mid Z = z) = \frac{P(X \leq Y \cap \min(X,Y) = z)}{P(Z = z)}$$

$$P(\delta = 1 \mid Z = z) = \frac{P(X = z)\, P(Y \geq z)}{P(Z = z)}$$

$$= \frac{\lambda e^{-\lambda t} e^{-\theta t}}{(\lambda+\theta)\, e^{-(\lambda+\theta)t}}$$

$$\therefore P(\delta = 1 \mid Z = z) = \frac{\lambda}{\lambda+\theta} = P(\delta = 1) \quad \forall z > 0$$

Analogamente:

$$P(\delta = 0 \mid Z = z) = \frac{\theta}{\lambda+\theta} = P(\delta = 0) \quad \forall z > 0$$

Portanto

δ e Z são independentes //

2.d

Lembremos que:

$$P(\delta=1) = \frac{\lambda}{\lambda+\theta} \qquad P(\delta=0) = \frac{\theta}{\lambda+\theta} = 1 - \frac{\lambda}{\lambda+\theta}$$

Assim, escrevendo $p = \frac{\lambda}{\lambda+\theta}$, temos que:

$$P(\delta=1) = p$$
$$P(\delta=0) = 1-p$$

$$f_\delta(x) = P(\delta=x) = \left[\frac{\lambda}{\lambda+\theta}\right]^x \left[\frac{\theta}{\lambda+\theta}\right]^{1-x} \mathbb{1}_{\{0,1\}}(x)$$

Logo, torna-se evidente que

$$\delta \sim \text{Bernoulli}\left(\frac{\lambda}{\lambda+\theta}\right)$$

De forma então que a distribuição da soma de n Bernoullis iid é:

$$D \sim \text{Binomial}\left(n, \frac{\lambda}{\lambda+\theta}\right)$$

2.e

$$E(\hat{\lambda}) = E\left[\frac{\sum \delta_i}{\sum z_i}\right] = E\left[\frac{\delta_1}{\sum z_i} + \frac{\delta_2}{\sum z_i} + \ldots + \frac{\delta_n}{\sum z_i}\right]$$

Onde $\sum z_i \sim \text{Gama}(n, \lambda+\theta)$, pois $z_i \sim \exp(\lambda+\theta)$

Além disso:

$$E\left[\frac{\delta_i}{\Sigma z_i}\right] = E\left[E\left(\frac{\delta_i}{\Sigma z_i} \mid \delta_i\right)\right]$$

$$= P(\delta_i=0) \cdot E\left[\frac{0}{\Sigma z_i}\right] + P(\delta_i=1)\, E\left[\frac{1}{\Sigma z_i}\right]$$

$$= 0 + \frac{\lambda}{\lambda+\theta} \cdot \int_0^\infty \frac{1}{x} \cdot \frac{[\lambda+\theta]^n}{\Gamma(n)} \cdot e^{-(\lambda+\theta)x} \cdot x^{n-1}\, dx$$

$$= \lambda \cdot \int_0^\infty \frac{1}{\Gamma(n)} \cdot [\lambda+\theta]^{n-1} \cdot e^{-(\lambda+\theta)x} \cdot x^{n-2}\, dx$$

Sabemos que, sendo n um inteiro positivo:

$$\Gamma(n) = (n+1)! = (n+1) \cdot n! = (n+1)\,\Gamma(n-1)$$

Assim:

$$E\left[\frac{\delta_i}{\Sigma z_i}\right] = \frac{\lambda}{(n+1)} \cdot \underbrace{\int_0^\infty \frac{[\lambda+\theta]^{n-1}}{\Gamma(n-1)} \cdot e^{-(\lambda+\theta)x} \cdot x^{n-2}\, dx}_{= \int_0^\infty \text{Gama}(n-1,\, \lambda+\theta)\, dx = 1}$$

Assim:

$$E\left[\frac{\delta_i}{\Sigma z_i}\right] = \frac{\lambda}{n+1} \qquad \therefore \quad E[\hat{\lambda}] = \frac{n}{n+1}\lambda$$

Questão 3

Questão 4

4.a

Primeiro vamos dividir a variável Idade segundo as faixas propostas pelo enunciado, e as outras variáveis contínuas vamos dividir entre:

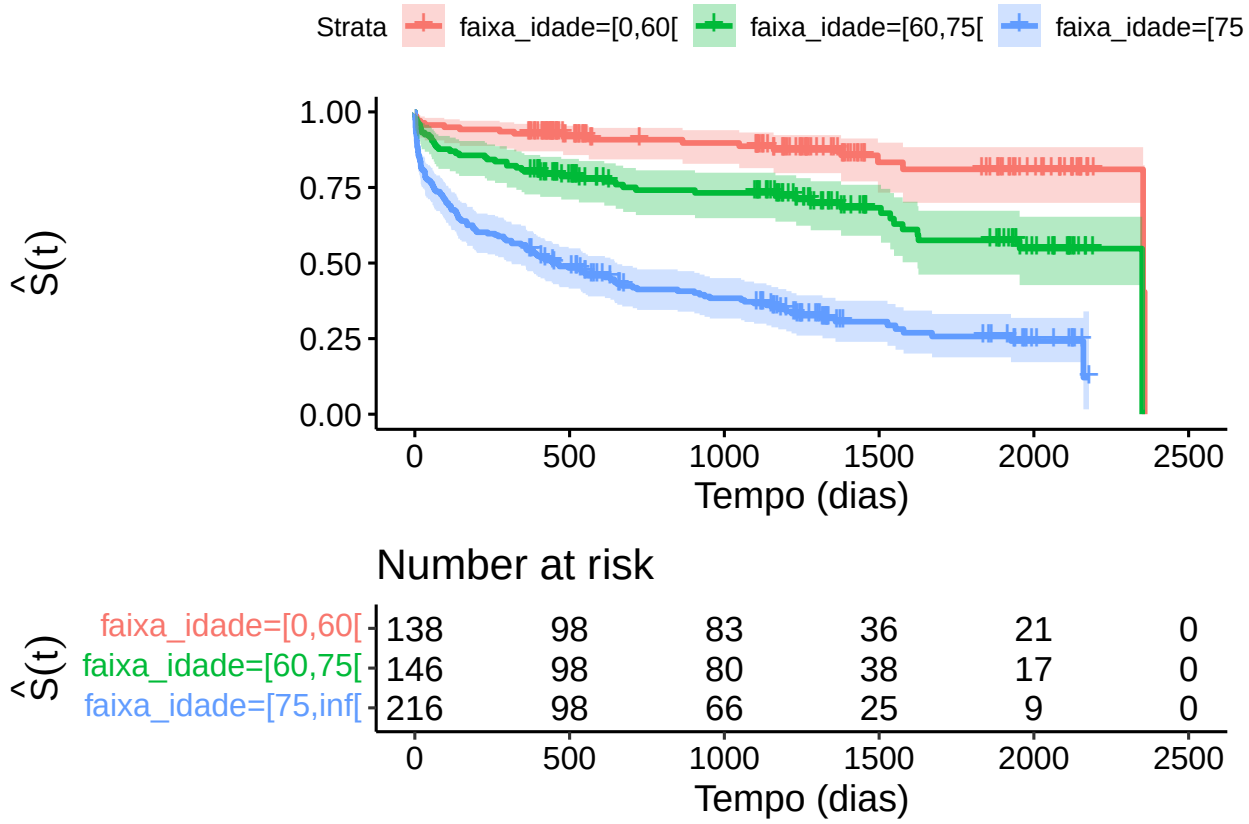
- Menor que o primeiro quartil
- Entre o primeiro e o segundo quartis
- Entre o segundo e o terceiro quartis
- Maior que o terceiro quartil

Feito isso, vamos agora observar as curvas de Kaplan-Meier para cada grupo de cada variável explicativa individualmente, com as estimativas dos intervalos de confiança ponto a ponto, criados usando a transformação "log-log" para que fiquem bem comportados.

Idade

```
q4_km_age = survfit(Surv(lenfol, fstat)~faixa_idade, type = 'kaplan-meier', data = df,
                    conf.type = 'log-log')

ggsurvplot(q4_km_age,df, conf.int = T, risk.table = T, tables.height = 0.35)+
  labs(x="Tempo (dias)", y=expression(hat(S)(t)))
```

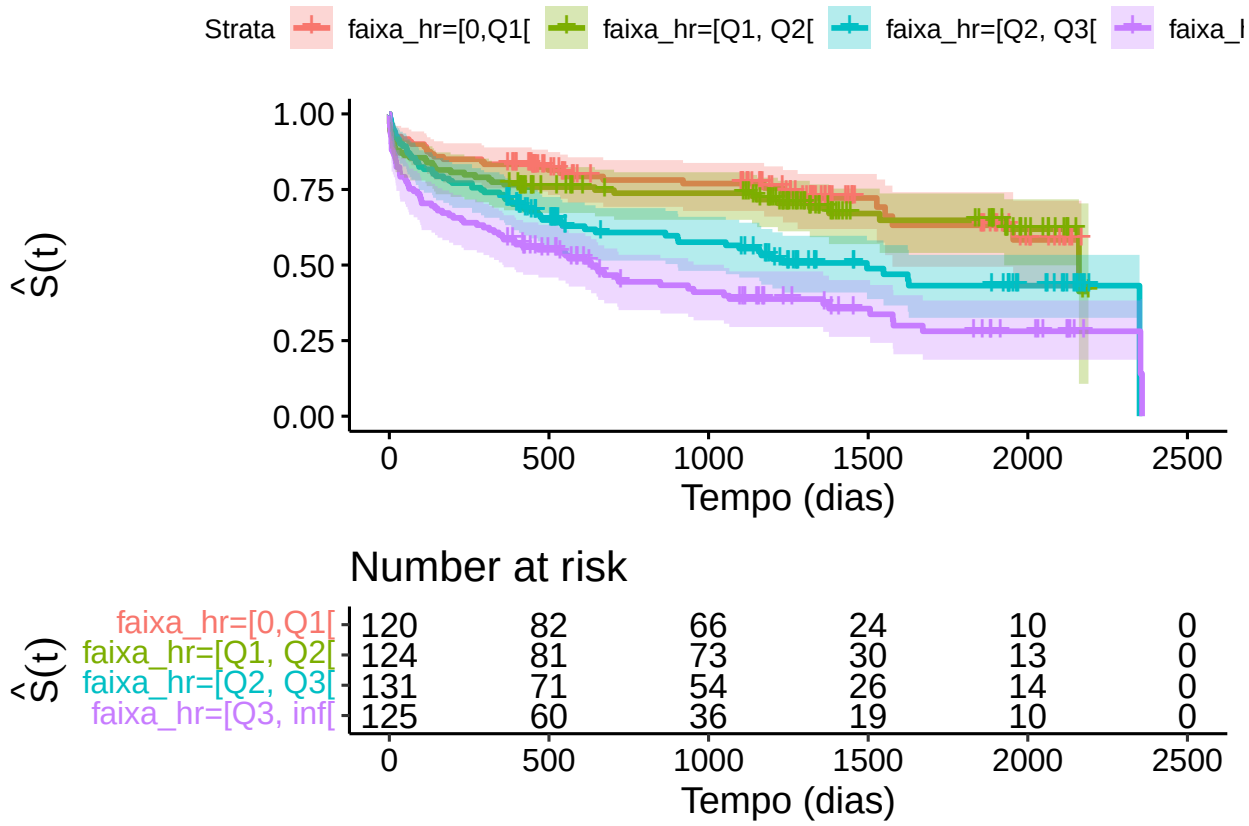


Aqui vemos que, fixado um t menor que 2000 (pois depois disso o número de observações é muito baixo), idades mais elevadas levam a estimativas menores da função de sobrevivência.

Frequência cardíaca inicial

```
q4_km_hr = survfit(Surv(lenfol, fstat)~faixa_hr, type = 'kaplan-meier', data = df,
                   conf.type = 'log-log')

ggsurvplot(q4_km_hr,df, conf.int = T, risk.table = T, tables.height = 0.35)+
  labs(x="Tempo (dias)", y=expression(hat(S)(t)))
```

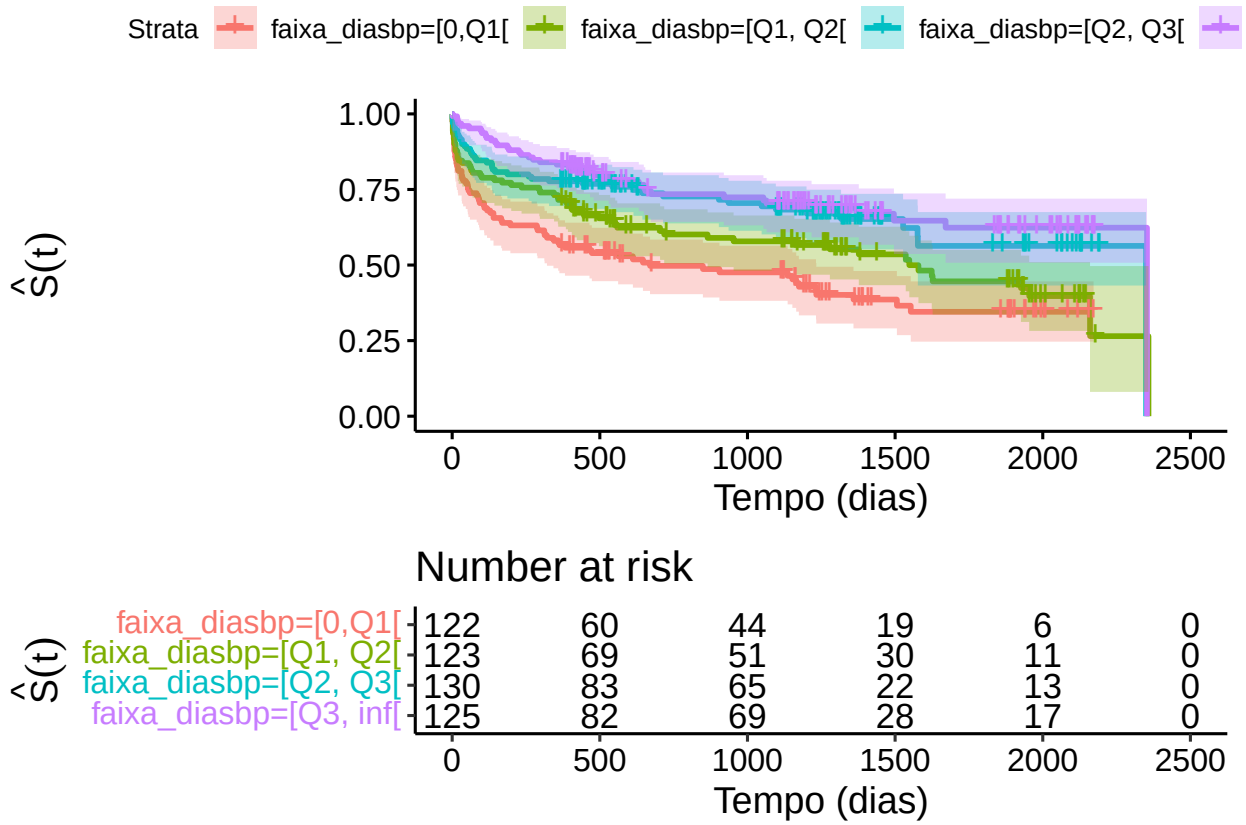


Notemos que frequência cardíaca inicial parece estar negativamente correlacionada com tempo de sobrevivência, já que faixas menores de frequência cardíaca tem maior S(t) estimado em praticamente todos os instantes de tempo, com excessão dos instantes finais onde existem pouquíssimas observações.

Pressão diastólica inicial

```
q4_km_diasbp = survfit(Surv(lenfol, fstat)~faixa_diasbp, type = 'kaplan-meier', data = df,
                       conf.type = 'log-log')

ggsurvplot(q4_km_diasbp,df, conf.int = T, risk.table = T, tables.height = 0.35)+
  labs(x="Tempo (dias)", y=expression(hat(S)(t)))
```

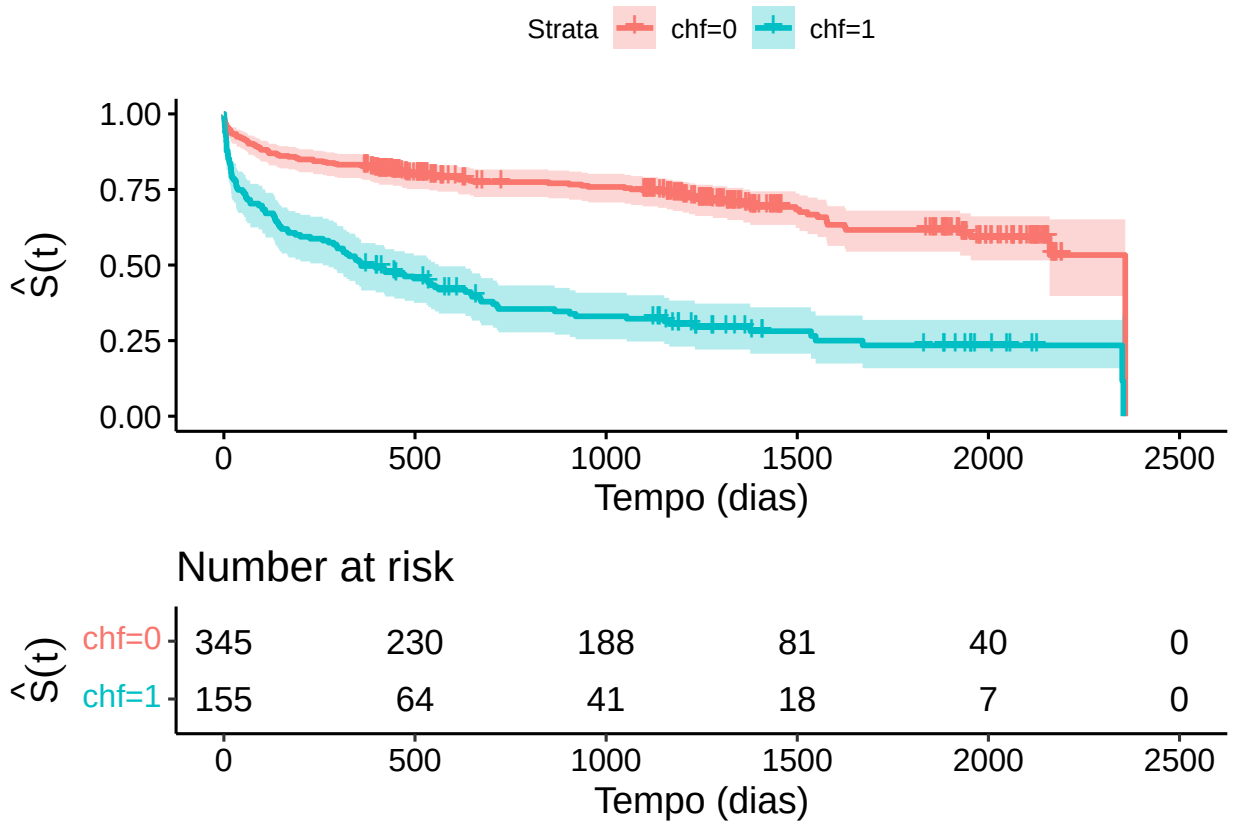


Notemos então que pressão distólica inicial parece estar positivamente correlacionada com tempo de sobrevivência, já que faixas de pressão menores tem, em todos os instantes até 2000 onde já existem poucas observações, estimativas de S(t) menores.

Complicações congestivas

```
q4_km_chf = survfit(Surv(lenfol, fstat)~chf, type = 'kaplan-meier', data = df,
                    conf.type = 'log-log')

ggsurvplot(q4_km_chf,df, conf.int = T, risk.table = T, tables.height = 0.35)+
  labs(x="Tempo (dias)", y=expression(hat(S)(t)))
```



Aqui fica evidente que não ter complicações congestivas está associado com ter uma maior probabilidade de sobrevivência em cada instante.

4.b

A ideia aqui é construir um modelo de regressão Weibull.

```
# 4.b
q4_w = survreg(Surv(lenfol, fstat) ~ faixa_idade + diasbp + hr + chf,
data = df, dist = "weibull")
summary(q4_w)
```

```
##
## Call:
## survreg(formula = Surv(lenfol, fstat) ~ faixa_idade + diasbp +
##     hr + chf, data = df, dist = "weibull")
##                          Value Std. Error     z       p
## (Intercept)          10.56895    0.76854 13.75 < 2e-16
## faixa_idade[60,75[  -1.35734    0.50966 -2.66 0.00774
## faixa_idade[75,inf[ -2.98376    0.48448 -6.16 7.3e-10
## diasbp                0.02559    0.00648  3.95 7.9e-05
## hr                   -0.02111    0.00555 -3.80 0.00014
## chf1                 -1.58670    0.27805 -5.71 1.2e-08
## Log(scale)            0.63585    0.05930 10.72 < 2e-16
```

```
##
## Scale= 1.89
##
## Weibull distribution
## Loglik(model)= -1660.3   Loglik(intercept only)= -1752.5
##  Chisq= 184.26 on 5 degrees of freedom, p= 6.6e-38
## Number of Newton-Raphson Iterations: 5
## n= 500
```

Notemos que ao nível de 5% todos os coeficientes foram significativos.

Primeiro lembremos que o algoritmo não solta os parâmetros usuais, de forma que temos que fazer as seguintes transformações sendo $\hat{\sigma}$ o parâmetro de escala retornado e $\hat{\gamma_j}$ o j-ésimo coeficiente retornado, onde $\hat{\gamma_0}$ é o intercepto.

$$\hat{\rho} = \frac{1}{\hat{\sigma}} = \frac{1}{1.89}$$

$$\hat{\beta_0} = exp\{-\frac{\hat{\gamma_0}}{\hat{\sigma}}\}$$

$$\hat{\beta_j} = -\frac{\hat{\gamma_j}}{\hat{\sigma}}$$

Assim, temos que:

- $\hat{\rho} \approx 0.519$
- $\hat{\beta_0} = exp\{-\frac{10.56895}{1.89}\} \approx 0.0037$
- $\hat{\beta}_{idade \geq 60 < 75} = 0.7187$
- $\hat{\beta}_{idade \geq 75} = 1.5799$
- $\hat{\beta}_{diasbp} = -0.0135$
- $\hat{\beta}_{hf} = 0.0112$
- $\hat{\beta}_{chf=1} = 0.8401$

Além disso, utilizando o ponto de vista dos modelos de taxa de risco proporcionais, temos que, sendo x_i e x_j duas observações, então:

$$\frac{\hat{\alpha}(t|x_i)}{\hat{\alpha}(t|x_j)} = exp\{(x_i^T - x_j^T)\hat{\beta}\}$$

Onde os parâmetros que compõem o vetor $\hat{\beta}$ são os mesmos de antes. Com isso é possível facilitar a interpretação dos parâmetros.

Por exemplo, sendo x_i um indivíduo com idade acima de 75 anos e x_j um indivíduo com idade menor que 60 anos, mantendo-se as outras covariáveis iguais, temos que:

$$\frac{\hat{\alpha}(t|x_i)}{\hat{\alpha}(t|x_j)} = exp\{1.5799\} \approx 4.85$$

4.c

Mantendo-se todas as covariáveis iguais exceto chf, e sendo x_i um indivíduo com complicações congestivas e x_j um indivíduo sem, então:

$$\frac{\hat{\alpha}(t|x_i)}{\hat{\alpha}(t|x_j)} = exp\{0.8401\} \approx 2.317$$

Assim, a taxa de risco de um indivíduo com complicações congestiva é 1.317 vezes maior a de um indivíduo sem complicações congestivas.

Sendo x_i um indivíduo com idade acima de 75 anos e x_j um indivíduo com idade menor que 60 anos, mantendo-se as outras covariáveis iguais, temos que:

$$\frac{\hat{\alpha}(t|x_i)}{\hat{\alpha}(t|x_j)} = exp\{1.5799\} \approx 4.85$$

Assim, o indivíduo com idade maior que 75 anos tem um risco 3.85 vezes maior do que um com idade menor que 60 anos.

Além disso, sendo x_i um indivíduo com idade entre 60 e 75 anos e x_j um indivíduo com idade menor que 60 anos, mantendo-se as outras covariáveis iguais, temos que:

$$\frac{\hat{\alpha}(t|x_i)}{\hat{\alpha}(t|x_j)} = exp\{0.7187\} \approx 2.07$$

4.d

Sendo x_j um indivíduo com pressão diastólica de x mmHg e x_i um indivíduo com pressão diastólica $(x+1)$ mmHg, mantendo-se as outras covariáveis iguais temos que:

$$\frac{\hat{\alpha}(t|x_i)}{\hat{\alpha}(t|x_j)} = \frac{exp\{-0.0135(x+1)\}}{exp\{-0.0135(x)\}} = exp\{-0.0135\} \approx 0.987$$

Sendo x_j um indivíduo com pressão diastólica de x mmHg e x_i um indivíduo com pressão diastólica $(x+10)$ mmHg, mantendo-se as outras covariáveis iguais temos que:

$$\frac{\hat{\alpha}(t|x_i)}{\hat{\alpha}(t|x_j)} = \frac{exp\{-0.0135(x+10)\}}{exp\{-0.0135(x)\}} = exp\{-0.135\} \approx 0.874$$

Sendo x_j um indivíduo com frequência cardíaca de x bpm e x_i um indivíduo com frequência cardíaca $(x+1)$ bpm, mantendo-se as outras covariáveis iguais temos que:

$$\frac{\hat{\alpha}(t|x_i)}{\hat{\alpha}(t|x_j)} = \frac{exp\{0.0112(x+1)\}}{exp\{0.0112(x)\}} = exp\{0.0112\} \approx 1.011$$

Sendo x_j um indivíduo com frequência cardíaca de x bpm e x_i um indivíduo com frequência cardíaca $(x+5)$ bpm, mantendo-se as outras covariáveis iguais temos que:

$$\frac{\hat{\alpha}(t|x_i)}{\hat{\alpha}(t|x_j)} = \frac{exp\{0.0112(x+5)\}}{exp\{0.0112(x)\}} = exp\{0.056\} \approx 1.058$$

Sendo x_j um indivíduo com frequência cardíaca de x bpm e x_i um indivíduo com frequência cardíaca $(x + 10)$ bpm, mantendo-se as outras covariáveis iguais temos que:

$$\frac{\hat{\alpha}(t|x_i)}{\hat{\alpha}(t|x_j)} = \frac{exp\{0.0112(x + 10)\}}{exp\{0.0112(x)\}} = exp\{0.112\} \approx 1.119$$

Com isso nós confirmamos tudo o que foi observado na análise descritiva.

Questão 5

Questão 6

Código Completo

```
knitr::opts_chunk$set(warning=FALSE,
                      # fig.dim = c(5,5),
                      # out.height = '40%',
                      # fig.align = 'center',
                      message=FALSE
                      )

library(tidyverse)
library(ggplot2)
library(knitr)
library(readr)
library(dplyr)


# QUESTAO 4 ----
# QUESTAO 4a ----
library(survival)
library(survminer)
df = read.table('data/Lista3_whas500.dat')
colnames(df) = c('id','age','gender', 'hr', 'sysbp', 'diasbp', 'bmi','cvd',
                 'afb','sho','chf','av3','miord', 'mitype', 'year',
                 'admitdate', 'disdate',
                 'fdate', 'los', 'dstat', 'lenfol', 'fstat')
df =  df[,c('lenfol', 'fstat', 'age', 'hr', 'diasbp', 'chf')]
df$chf = as.factor(df$chf)

# criando faixas de idade
df$faixa_idade <- as.factor(sapply(df$age,
  function(x){
  if (x < 60) x = '[0,60['
  else if (x >= 60 & x < 75) x = '[60,75['
  else if (x >= 75 ) x = '[75,inf['
  }))
```

```r
df$faixa_idade = relevel(df$faixa_idade, ref = '[0,60[')

df$faixa_hr <- as.factor(sapply(df$hr,
  function(x){
  if (x < quantile(df$hr, probs = 0.25)) x = '[0,Q1['
  else if (x >= quantile(df$hr, probs = 0.25) &
           x < quantile(df$hr, probs = 0.50)) x = '[Q1, Q2['
  else if (x >= quantile(df$hr, probs = 0.50) &
           x < quantile(df$hr, probs = 0.75)) x = '[Q2, Q3['
  else if (x>=quantile(df$hr, probs = 0.75)) x = "[Q3, inf["
  }))


df$faixa_diasbp <- as.factor(sapply(df$diasbp,
  function(x){
  if (x < quantile(df$diasbp, probs = 0.25)) x = '[0,Q1['
  else if (x >= quantile(df$diasbp, probs = 0.25) &
           x < quantile(df$diasbp, probs = 0.50)) x = '[Q1, Q2['
  else if (x >= quantile(df$diasbp, probs = 0.50) &
           x < quantile(df$diasbp, probs = 0.75)) x = '[Q2, Q3['
  else if (x>=quantile(df$diasbp, probs = 0.75)) x = "[Q3, inf["
  }))
q4_km_age = survfit(Surv(lenfol, fstat)~faixa_idade, type = 'kaplan-meier', data = df,
                    conf.type = 'log-log')

ggsurvplot(q4_km_age,df, conf.int = T, risk.table = T, tables.height = 0.35)+
  labs(x="Tempo (dias)", y=expression(hat(S)(t)))
q4_km_hr = survfit(Surv(lenfol, fstat)~faixa_hr, type = 'kaplan-meier', data = df,
                    conf.type = 'log-log')

ggsurvplot(q4_km_hr,df, conf.int = T, risk.table = T, tables.height = 0.35)+
  labs(x="Tempo (dias)", y=expression(hat(S)(t)))
q4_km_diasbp = survfit(Surv(lenfol, fstat)~faixa_diasbp, type = 'kaplan-meier', data = df,
                    conf.type = 'log-log')

ggsurvplot(q4_km_diasbp,df, conf.int = T, risk.table = T, tables.height = 0.35)+
  labs(x="Tempo (dias)", y=expression(hat(S)(t)))
q4_km_chf = survfit(Surv(lenfol, fstat)~chf, type = 'kaplan-meier', data = df,
                    conf.type = 'log-log')

ggsurvplot(q4_km_chf,df, conf.int = T, risk.table = T, tables.height = 0.35)+
  labs(x="Tempo (dias)", y=expression(hat(S)(t)))

# 4.b
q4_w = survreg(Surv(lenfol, fstat) ~ faixa_idade + diasbp + hr + chf,
data = df, dist = "weibull")
summary(q4_w)
```