Atividade 3

Guilherme Navarro NUSP: 8943160

22 de junho de 2020

Atividade 3

Um estudo foi conduzido em uma comunidade de Tartu, segunda maior cidade da Estônia, para avaliar a sobrevida de pacientes que sofreram infarte no período de 1991 a 1993 (com seguimento até 1996). Os dados considerados são de 824 pacientes com 18 anos ou mais que tiveram infarto no período de 1991 a 1993 e algumas variáveis foram observadas:

- Tempo de vida, em meses, após o infarto (os dados estão sujeitos a censura à direita);
- Gênero (feminino ou masculino);
- Idade (em anos);
- Diagnóstico (Isquêmico/Hemorragia intracranial/Não identificado/hemorragia subaracnóide);
- Coma - variável binária (Não/Sim) indicando se o paciente entrou em coma após o infarto;
- Infarto prévio do miocárdio - variável binária (Não/Sim) indicando se o paciente tem histórico de infarto prévio do miocárdio.

Os dados estão disponíveis no arquivo **stroke-final.csv**, com as seguintes variáveis:

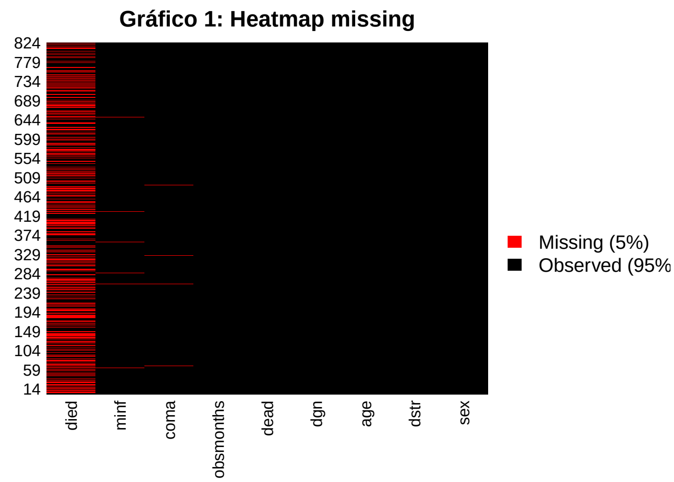
- sex: sexo do paciente
- died: data do óbito
- dstr: data o infarto
- age: idade na data do infarto
- dgn: diagnóstico
- coma: indicadora de coma após infarto
- minf: infarto prévio do miocárdio
- obsmonths: tempo, em meses, decorrido entre infarto e óbito ou censura (optou-se por imputar 0,1 para pacientes que morreram no mesmo dia do infarto)
- dead: indica ocorrência de óbito ou não.

Utilizando esses dados, responda os itens descritos a seguir:

(a) Faça uma análise descritiva dos dados. Essa análise descritiva deve envolver curvas de Kaplan-Meier segundo as covariáveis descritas, bem como testes para comparação das curvas obtidas.

Resolução

Ao iniciar a análise irei fazer um mapa de calor com os missings da base de dados

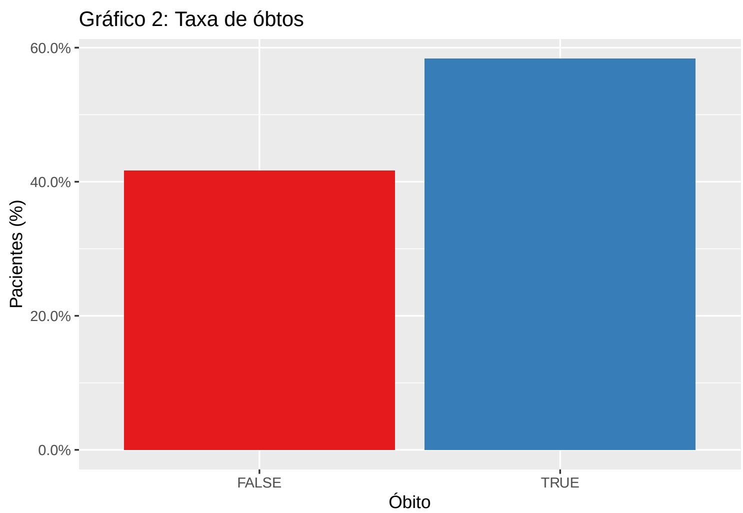


Ao analisar o gráfico 1, a variável "died" apresentou quase 5% de valores missings, por ser uma data, não entrará na análise assim como a variável "dstr", quanto as variáveis "minf" e "coma" foi optado por remover as observações missing (10 linhas removidas, cerca de 1,21% dos dados).

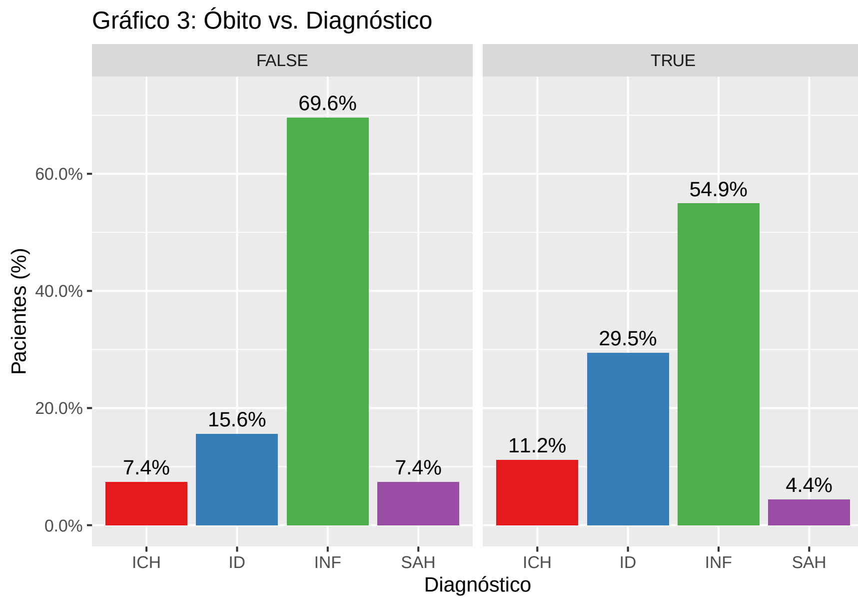
Além disso a variável "age" é do tipo contínua, com a finalidade facilitar a interpretação e a análise irei categorizar de forma binária divididas pela mediana (71 anos), pois matém um bom balanceamento de observações em cada categoria, como mostra a tabela abaixo

Age<=71	Freq.
FALSE	408
TRUE	416

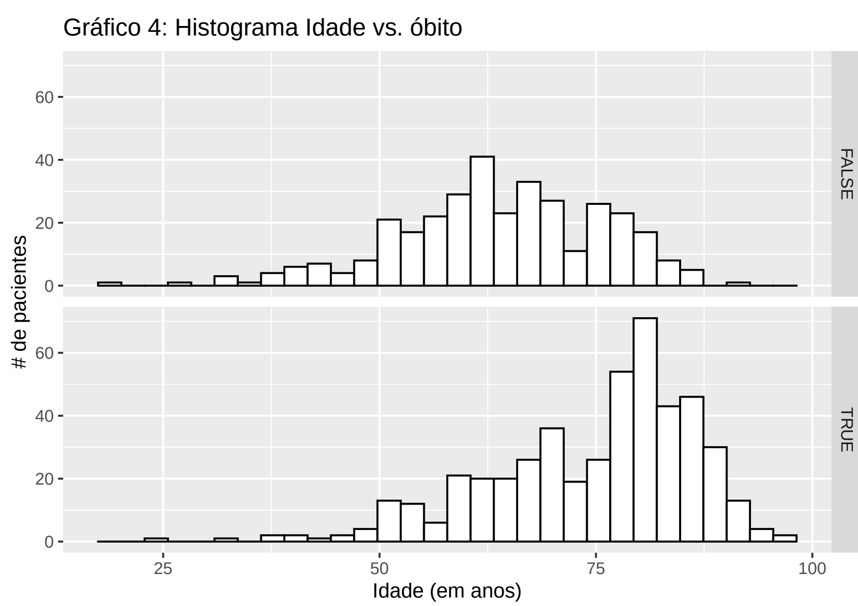
Removida as devidas variáveis e observações problemáticas podemos partir para análise descritiva, sendo assim:



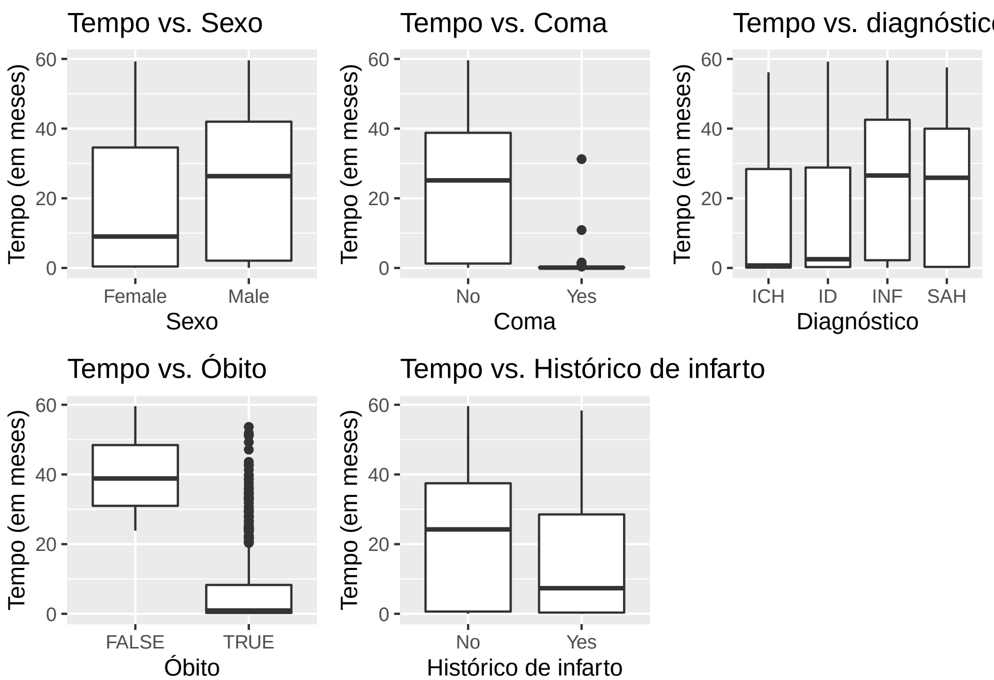
Em que podemos notar que cerca de 60% dos pacientes morreram.



No gráfico 3, dos pacientes que morreram cerca de 55% tiveram o diagnóstico “não identificado”, seguido de 30% com “hemorragia intracraniana”, enquanto isso para os pacientes que sobreviveram no periodo do estudo, cerca de 70% tiveram dignóstico “não identificado”.



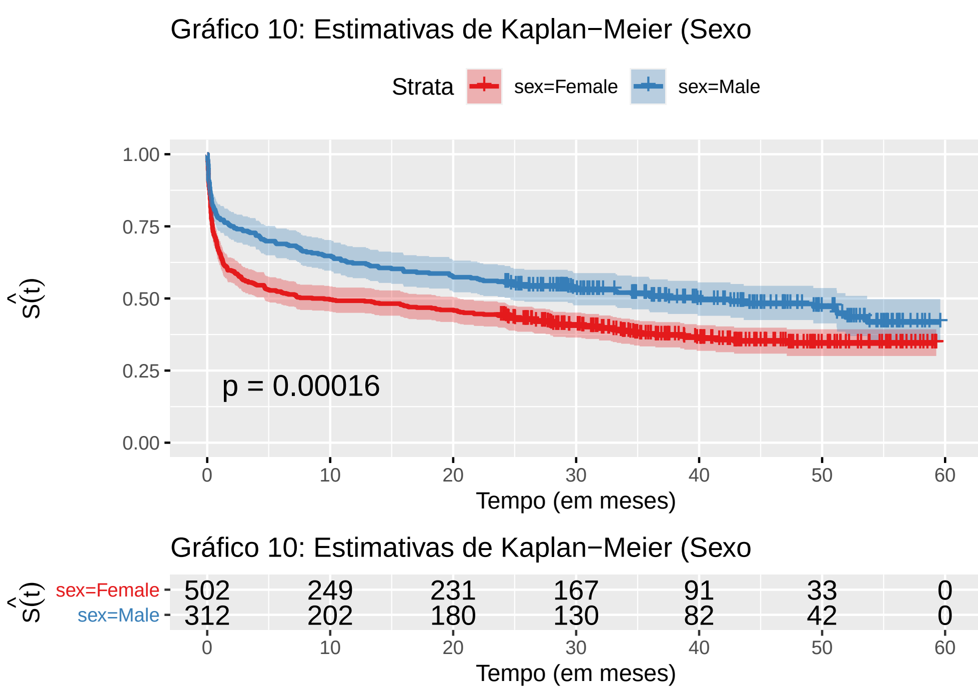
No gráfico 4, é possível ver que a massa de pacientes que vieram a óbito esta concentrada entre 75 e 90 anos assim como esperado, em contrapartida os pacientes que sobreviveram durante o estudo tiveram uma concentração de idade próximo dos 60 anos.



Nos boxplots (gráficos 5 à 9) acima, temos o tempo decorrido entre infarto e óbito ou censura contra cada covarável, e pode-se notar que os homens tem uma mediana de tempo um pouco maior que das mulheres, também que os pacientes que ficaram em coma tiveram o tempo muito menor contra os que não ficaram, e assim como no gráfico 3 os pacientes diagnosticados com hemorragia intracraniana ou não identificados tem um tempo maior, como esperado os pacientes que vieram a óbito tiveram um tempo menor do que os censurados e por fim os pacientes que tiveram um histórico de infarto tiveram um tempo menor.

Avaliando os gráficos de Kaplan-Meier para cada covariável, temos:

Para a covariável sexo:



Queremos testar a igualdade das curvas, assim:

$$\begin{cases} H_0 : S_1(t) = S_2(t),\ \forall\ t \in\ [0, \tau] \\ H_1 : S_1(t) \neq S_2(t)\ para\ algum\ t \in\ [0, \tau] \end{cases}$$

Em que τ é o maior instante observado tal que os dois grupos possuem pelo menos um indivíduo em risco.

Sob a hipótese nula, a estatística do teste Log-Rank é:

$$L_r = \frac{[\sum_{j=1}^{L}(d_{2j} - e_{2j})]^2}{\sum_{j=1}^{L} V_j^2}$$

Em que d_{2j} é o # de indivíduos observados no grupo 2, e_{2j} é o # de indivíduos esperadados no grupo 2 e V_j é a variância de d_{2j} que é dada por:

$$V_j = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_{1_j} - 1)}$$

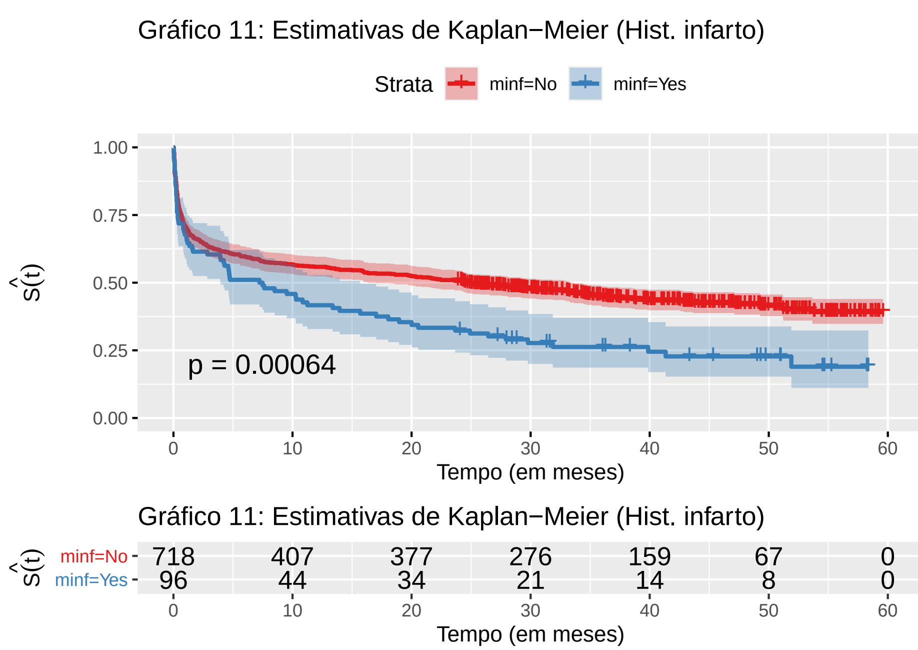
Em que n_{1j} e n_{2j} são o número de indivíduos nos grupo 1 e 2 respectivamente. Assim sendo, sob a hipótese nula,

$$L_r \overset{a}{\sim} \chi^2_{(1)}$$

Utilizando o teste log-rank, temos:

variable	pval	method
sex	0.0001563	Log-rank

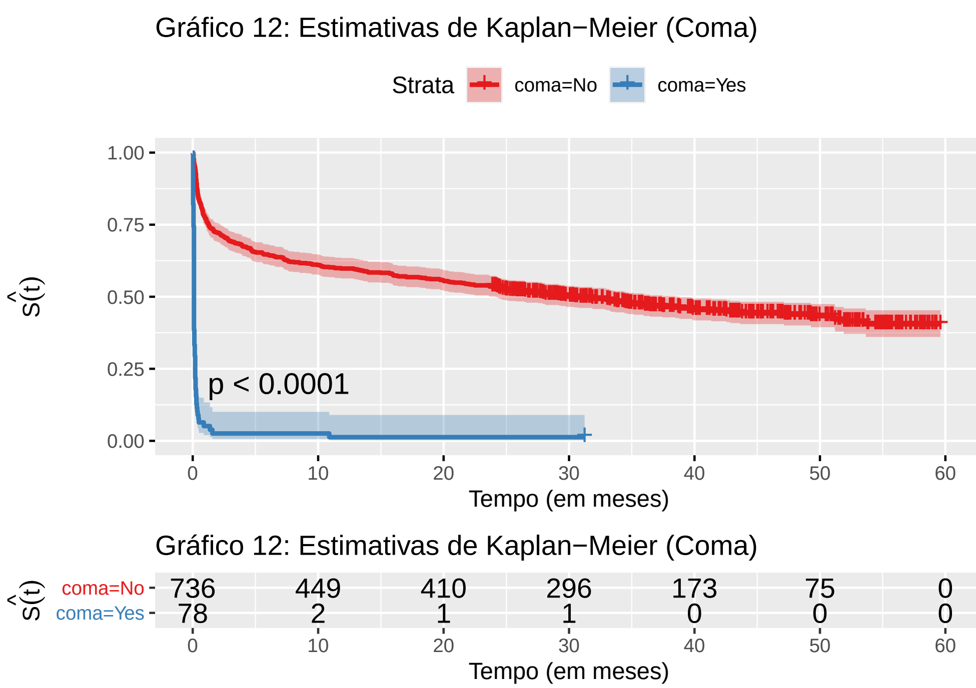
Pelo gráfico 10 o sexo feminino aparenta possuir uma probabilidade de sobrevivência um pouco menor do que a do sexo masculino, porém no teste de Log-Rank e o gráfico das estimavivas de Kaplan-Meier as curvas não são iguais a um nível de significância de 5%.



Utilizando o teste log-rank, temos:

variable	pval	method
minf	0.0006447	Log-rank

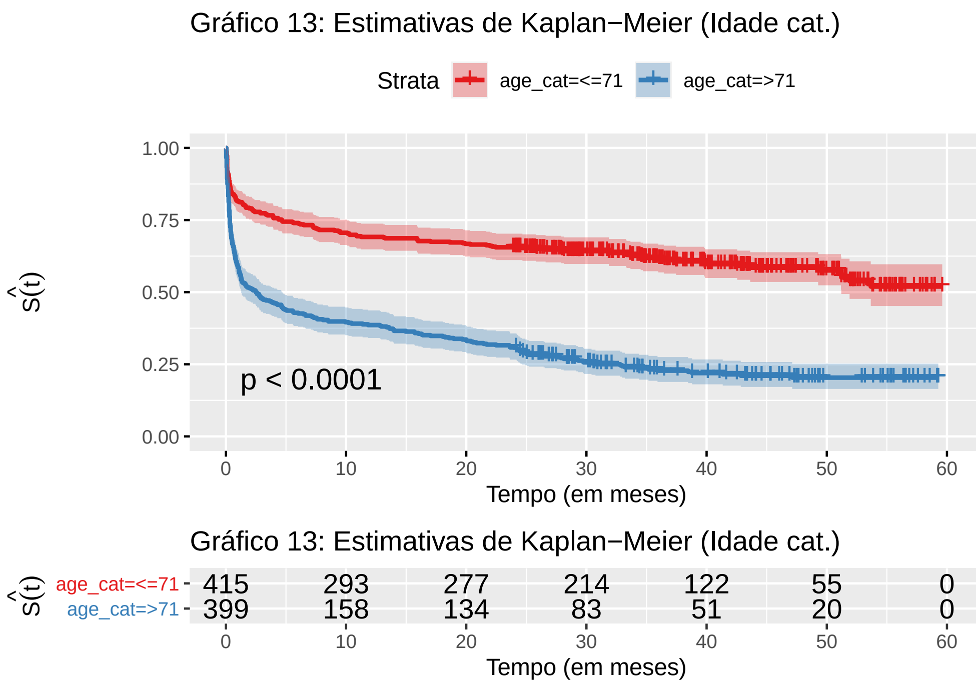
Pelo gráfico 11 os pacientes que tem histórico de infarto aparentam possuir uma probabilidade de sobrevivência e um tempo de vida um pouco menor do que a dos os pacientes que não tem histórico de infarto, porém no teste de Log-Rank e o gráfico das estimavivas de Kaplan-Meier as curvas não são iguais a um nível de significância de 5%.



Utilizando o teste log-rank, temos:

variable	pval	method
coma	0	Log-rank

Pelo gráfico 12 os pacientes que ficaram em coma aparentam possuir uma probabilidade de sobrevivência e um tempo de vida muito inferior do que a dos os pacientes que não ficaram em coma e isso se confima no teste de Log-Rank e o gráfico das estimavivas de Kaplan-Meier as curvas não são iguais a um nível de significância de 5%.

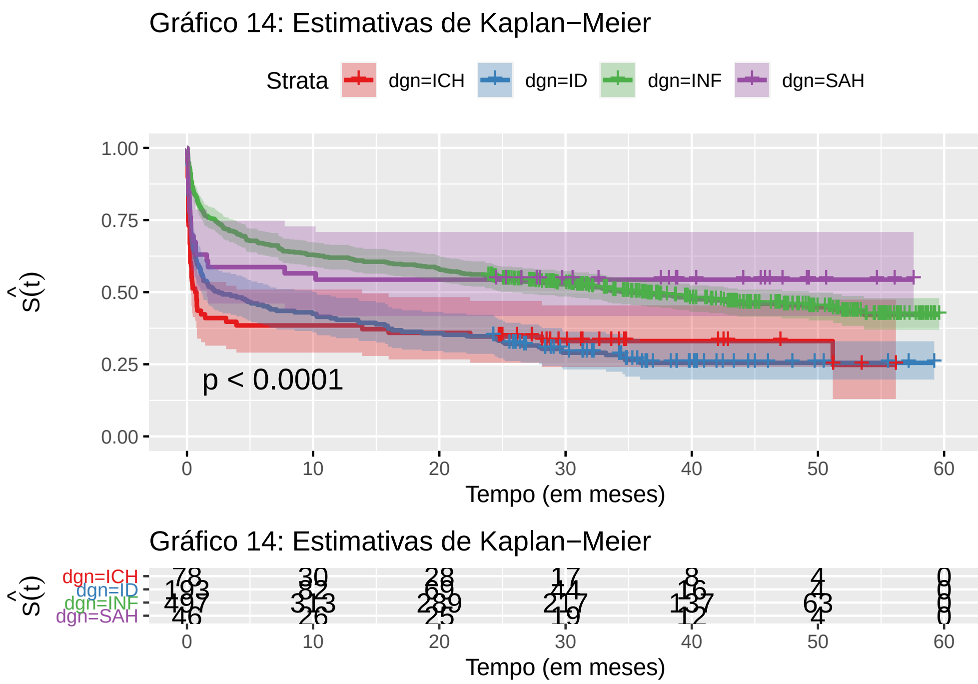


Utilizando o teste log-rank, temos:

variable	pval	method
coma	0	Log-rank

Pelo gráfico 13 os pacientes que tem idade superior a 71 anos aparentam possuir uma probabilidade de sobrevivência muito inferior do que a dos os pacientes que tem menos de 71 anos e isso se confima no teste de Log-Rank e o gráfico das estimavivas de Kaplan-Meier as curvas não são iguais a um nível de significância de 5%.

Para a covariável diagnóstico, por ter mais categorias, apresentou o seguinte gráfico com as estimativas de Kaplan-Meier:



Queremos comparar se pelo menos uma das curvas é diferente, assim utilizando o teste log-rank generalizado, com a seguinte hipóteses, temos:

Sob a hipótese:

$$\begin{cases} H_0 : S_1(t) = S_2(t) = S_3(t) = S_4(t), \ \forall \ t \in \ [0, \tau] \\ H_1 : pelo \ menos \ uma \ funcao \ diferente \ para \ algum \ t \in \ [0, \tau] \end{cases}$$

Utilizando o teste log-rank generalizado:

variable	pval	method
dgn	0	Log-rank

Em que segundo o teste de Log-Rank e o gráfico das estimavivas de Kaplan-Meier pelo menos uma das curvas não são iguais a um nível de significância de 5%.

(b) Ajuste um modelo Weibull aos dados. Apresente os resultados do modelo completo, com todas as covariáveis incluídas. Faça um processo de seleção de variáveis e apresente o resultado do modelo final obtido. Você precisa descrever claramente o processo de seleção das variáveis adotado, mas deve apresentar apenas as estimativas e resultados de dois modelos: modelo completo e modelo final. Você pode apresentar os resultados do modelo na parametrização de locação-escala.

Resolução

Ajustando o modelo completo com todas as variáveis:

```
##
## Call:
## survreg(formula = Surv(obsmonths, dead) ~ sex + dgn + coma +
##     minf + age_cat, data = data, dist = "weibull")
##               Value Std. Error      z       p
## (Intercept)  4.2803     0.3350  12.78 < 2e-16
## sexMale      0.1645     0.2122   0.78 0.43825
## dgnID        0.3139     0.3533   0.89 0.37422
## dgnINF       1.2861     0.3376   3.81 0.00014
## dgnSAH       1.1683     0.5453   2.14 0.03216
## comaYes     -4.9935     0.2767 -18.05 < 2e-16
## minfYes     -1.0284     0.2714  -3.79 0.00015
## age_cat>71  -2.2113     0.2216  -9.98 < 2e-16
## Log(scale)   0.7420     0.0389  19.07 < 2e-16
##
## Scale= 2.1
##
## Weibull distribution
## Loglik(model)= -1517.2   Loglik(intercept only)= -1704.9
##  Chisq= 375.3 on 7 degrees of freedom, p= 4.7e-77
## Number of Newton-Raphson Iterations: 5
## n= 814
```

Após o ajuste do modelo acima, nota-se que a varável sexo não é significativa a um nível de significância de 5%, opta-se por remove la do modelo, alternativamente testei o método de stepwise e obtive o mesmo resultado com apenas a remoção da variável sexo, otendo assim um modelo com menor AIC, ajustando novamente:

```
##
## Call:
## survreg(formula = Surv(obsmonths, dead) ~ dgn + coma + minf +
##     age_cat, data = data, dist = "weibull")
##               Value Std. Error      z       p
## (Intercept)  4.3577     0.3207  13.59 < 2e-16
## dgnID        0.3074     0.3528   0.87 0.38349
## dgnINF       1.2945     0.3368   3.84 0.00012
## dgnSAH       1.1599     0.5448   2.13 0.03326
## comaYes     -4.9980     0.2763 -18.09 < 2e-16
## minfYes     -1.0103     0.2703  -3.74 0.00019
## age_cat>71  -2.2500     0.2163 -10.40 < 2e-16
## Log(scale)   0.7417     0.0389  19.04 < 2e-16
##
## Scale= 2.1
##
## Weibull distribution
## Loglik(model)= -1517.5   Loglik(intercept only)= -1704.9
##  Chisq= 374.7 on 6 degrees of freedom, p= 7.7e-78
## Number of Newton-Raphson Iterations: 5
## n= 814
```

(c) Interprete os parâmetros do modelo final obtido em (b).

Resolução

Como os parâmetros que o R devolve não são usuais, deve se fazer uma pequena tranformação para a interpretação:

$$\rho = \frac{1}{\sigma} \Rightarrow \hat{\rho} = \frac{1}{\hat{\sigma}} = \frac{1}{2.1} = 0.476$$

E

$$\beta = -\frac{\gamma}{\sigma} \Rightarrow \hat{\beta} = -\frac{\hat{\gamma}}{\hat{\sigma}}$$

Em que γ são os parâmetros que o R devolve, sendo assim, escrevendo o modelo como riscos proporcionais:

$$\widehat{\alpha(t|x)} = \hat{\rho} t^{\hat{\rho}-1} e^{x'\hat{\beta}} = 0.476 t^{-0.524} e^{(-2.07-0.146x_1-0.616x_2-0.552x_3+2.379x_4+0.48x_5+1.071x_6)}$$

Assim comparando um indivíduo i com um j, temos:

$$\frac{\widehat{\alpha(t|x_i)}}{\widehat{\alpha(t|x_j)}} = \frac{\hat{\rho} t^{\hat{\rho}-1} e^{x_i'\hat{\beta}}}{\hat{\rho} t^{\hat{\rho}-1} e^{x_j'\hat{\beta}}} = e^{(x_i - x_j)'\hat{\beta}}$$

Então, fixando as outras covariáveis, pode-se dizer que para a covariável dignóstico Hemorragia intracranial=1 o risco de óbito é $e^{-0.146} = 0.86$ vezes o risco de óbito de um indivíduo com dignóstico Hemorragia intracranial=0.

Para a covariável dignóstico Não identificado=1 o risco de óbito é $e^{-0.616} = 0.54$ vezes o risco de óbito de um indivíduo com dignóstico Não identificado=0.

Para a covariável dignóstico hemorragia subaracnóide=1 o risco de óbito é $e^{-0.552} = 0.576$ vezes o risco de óbito de um indivíduo com dignóstico hemorragia subaracnóide=0.

Para a covariável coma="Yes" o risco de óbito é $(e^{2.379} - 1) * 100\% = 9.79\%$ maior do que risco de óbito de um indivíduo com coma="No"

Para a covariável minf="Yes" o risco de óbito é $(e^{0.48} - 1) * 100\% = 0.62\%$ maior do que risco de óbito de um indivíduo com minf="No"

Para a covariável age_cat=">71" o risco de óbito é $(e^{1.071} - 1) * 100\% = 1.92\%$ maior do que risco de óbito de um indivíduo com age="<71"

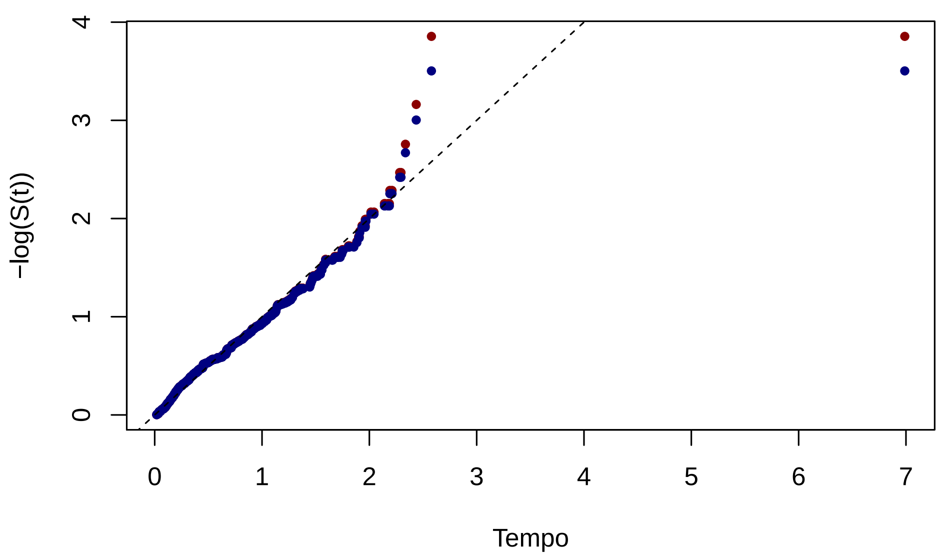
(d) Faça análise de resíduos do modelo final obtido em (b).

Resolução

Os resíduos de Cox-Snell para o modelo Weibull são obtidos a seguir:

É possível elaborar gráficos desses resíduos para a análise da escolha do modelo. Uma opção é realizar um gráfico da função de risco acumulada para os resíduos de Cox-Snell, utilizando os estimadores de Kaplan-Meier (em vermelho) e Nelson_Aalen (em azul), primeiramente para o modelo Weibull:

Gráfico 15: Função de risco acumulada − Weibull



O esperado é que os resíduos acompanhem a linha pontilhada, porém o que vemos no gráfico 15 a partir do tempo (2,5 meses) vemos um distanciamento da linha pontilhada

(e) De forma semelhante ao item (b), ajuste um modelo log-logístico aos dados. Faça da mesma forma (porém utiliazndo a distribuição log-logística) e apresente os resultados do modelo completo e do modelo final.

Resolução

Ajustando o modelo completo com todas as variáveis:

```
##
## Call:
## survreg(formula = Surv(obsmonths, dead) ~ sex + dgn + coma +
##     minf + age_cat, data = data, dist = "loglogistic")
##               Value Std. Error      z       p
## (Intercept)  2.7301     0.3726   7.33 2.4e-13
## sexMale      0.3507     0.2301   1.52   0.127
## dgnID        0.8733     0.3975   2.20   0.028
## dgnINF       1.8877     0.3691   5.11 3.2e-07
## dgnSAH       1.4065     0.5850   2.40   0.016
## comaYes     -4.4449     0.3183 -13.96 < 2e-16
## minfYes     -1.0319     0.3137  -3.29   0.001
## age_cat>71  -2.3137     0.2311 -10.01 < 2e-16
## Log(scale)   0.4768     0.0388  12.29 < 2e-16
##
## Scale= 1.61
##
## Log logistic distribution
## Loglik(model)= -1509.5   Loglik(intercept only)= -1685.4
##  Chisq= 351.78 on 7 degrees of freedom, p= 5.1e-72
## Number of Newton-Raphson Iterations: 4
## n= 814
```

Após o ajuste do modelo acima, analogamente ao modelo weibull nota-se que a varável sexo não é significativa a um nível de significância de 5%, porém opta-se por remove la do modelo, alternativamente testei o método de stepwise e obtive o mesmo resultado com apenas a remoção da variável sexo, otendo assim um modelo com menor AIC, ajustando novamente:

```
##
## Call:
## survreg(formula = Surv(obsmonths, dead) ~ dgn + coma + minf +
##     age_cat, data = data, dist = "loglogistic")
##               Value Std. Error      z       p
## (Intercept)  2.9072     0.3566   8.15 3.5e-16
## dgnID        0.8705     0.3988   2.18  0.0290
## dgnINF       1.9000     0.3706   5.13 3.0e-07
## dgnSAH       1.4184     0.5850   2.42  0.0153
## comaYes     -4.4465     0.3190 -13.94 < 2e-16
## minfYes     -0.9878     0.3134  -3.15  0.0016
## age_cat>71  -2.4172     0.2224 -10.87 < 2e-16
## Log(scale)   0.4790     0.0388  12.34 < 2e-16
##
## Scale= 1.61
##
## Log logistic distribution
## Loglik(model)= -1510.7   Loglik(intercept only)= -1685.4
##  Chisq= 349.46 on 6 degrees of freedom, p= 2e-72
## Number of Newton-Raphson Iterations: 4
## n= 814
```

(f) Interprete os parâmetros do modelo final obtido em (e).

Resolução

Como os parâmetros que o R devolve não são usuais, deve se fazer uma pequena tranformação para a interpretação:

$$\beta = -\frac{\gamma}{\sigma} \Rightarrow \hat{\beta} = -\frac{\hat{\gamma}}{\hat{\sigma}}$$

Com $\hat{\sigma} = 1.61$

Sendo assim, escrevendo o modelo como razão de chances:

$$\frac{\widehat{S(t|x)}}{1-\widehat{S(t|x)}} = \frac{\widehat{S(t|x=0)}}{1-\widehat{S(t|x=0)}} e^{-x'\hat{\beta}} = t^{1/\hat{\sigma}} e^{-\hat{\mu}/\hat{\sigma}} e^{-x'\hat{\beta}} = t^{0.621} e^{-1.80} e^{(-0.54x_1 - 1.18x_2 - 0.881x_3 + 2.762x_4 + 0.613x_5 + 1.501x_6)}$$

Em que $\hat{\sigma}$ é o praêmtro de escala, $\hat{\mu}$ é o intercepto, x' é a matriz de dados sem intercepto e $\hat{\beta}$ é o vetor de parâmetros. É usual interpretar os parâmetros utilizando a razão de chances proporcionais entre um indivíduo i e outro j com seguinte expressão:

Logo

$$\frac{\frac{\widehat{S(t|x_i)}}{1-\widehat{S(t|x_i)}}}{\frac{\widehat{S(t|x_j)}}{1-\widehat{S(t|x_j)}}} = \frac{t^{1/\hat{\sigma}} e^{-\hat{\mu}/\hat{\sigma}} e^{-x_i'\hat{\beta}}}{t^{1/\hat{\sigma}} e^{-\hat{\mu}/\hat{\sigma}} e^{-x_j'\hat{\beta}}} = e^{(x_j - x_i)'\hat{\beta}}$$

Então, fixando as outras covariáveis, pode-se dizer que para a covariável dignóstico Hemorragia intracranial=1 apresentam chance de óbito de $e^{-0.54} = 0.16$ vezes do que a de um indivíduo com dignóstico Hemorragia intracranial=0.

Para a covariável dignóstico Não identificado=1 apresentam chance de óbito de $e^{-1.18} = 0.37$ vezes do que a de um indivíduo com dignóstico Não identificado=0.

Para a covariável dignóstico hemorragia subaracnóide=1 apresentam chance de óbito é $e^{-0.881} = 0.414$ vezes do que a de um indivíduo com dignóstico hemorragia subaracnóide=0.

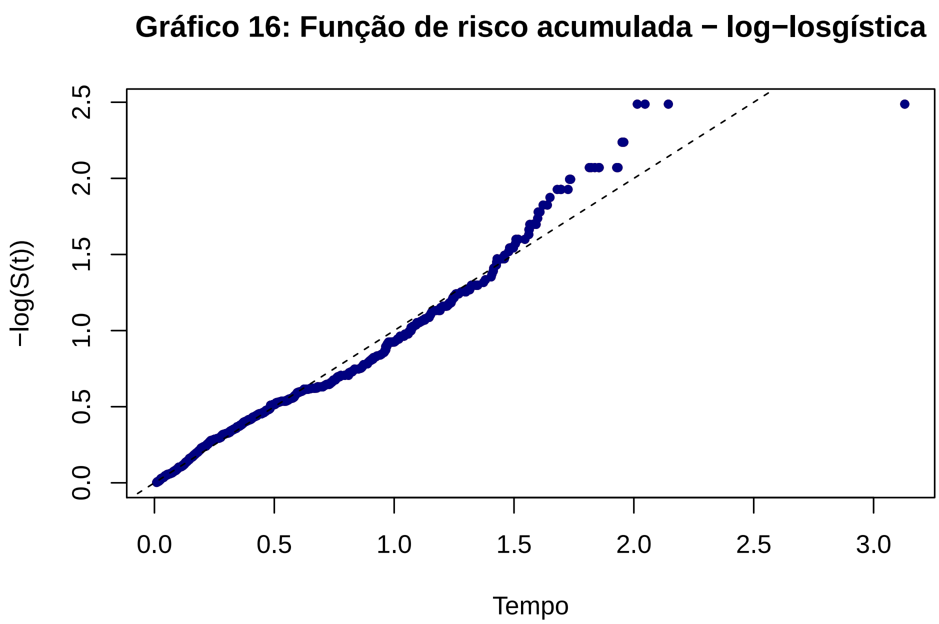
Para a covariável coma="Yes" apresentam chance óbito é $e^{2.761} = 15.823$ vezes do que a de um indivíduo com coma="No"

Para a covariável minf="Yes" apresentam chance de óbito é $e^{0.635} = 1.847$ vezes do que a de um indivíduo com minf="No"

Para a covariável age_cat=">71" apresentam chance de óbito é $e^{1.5} = 4.49$ vezes do que a de um indivíduo com age="<71"

(g) Faça análise de resíduos do modelo final obtido em (e).

Resolução



O esperado é que os resíduos acompanhem a linha pontilhada, e o que o gráfico 16 mostra os pontos muito próximos da linha pontilhada o que indica um bom ajuste do modelo.

(h) Compare os ajustes e os gráficos de resíduos dois modelos finais obtidos (com a distribuição Weibull e log-logística). Escolha um dos modelos para apresentar ao pesquisador como modelo final e justifique sua resposta.

Resolução

Comparando os gráficos 15 e 16, quanto mais os pontos estiverem próximos a linha pontilhada melhor é o ajuste do modelo. Logo, nota-se que pelos gráficos do resíduos de Cox-Snell, o modelo Log-logístico possuem os pontos mais próximos da reta pontilhada do que o modelo Weibull. Pode-se fazer os cretérios AIC e BIC para confirmar a escolha, pela tabela abaixo, temos:

	AIC	BIC
Weibull	3053.038	3095.356
Log-logistico	3039.323	3081.641

Para o modelo log-losgístico temos um menor valor de AIC e BIC em comparação com o modelo weibull, e associado a análise de resíduos pode-se concluir que o modelo log-logístico esta melhor ajustado aos dados.

Anexo

Códigos

```
# Pacotes
library(ggplot2)
library(survival)
library(survminer)
library(KMsurv)
library(gridExtra)
library(Amelia)
library(RColorBrewer)

# Local de trabalho
setwd("~/Área de Trabalho/P1 - MAE 514")

# Leitura dos dados
data <- read.csv("stroke_final.csv",header = T)

# Removendo a primeira coluna (desnecessária)
data$X <- NULL

attach(data)

# item a
missmap(data,col = c('red','black'),main = "Gráfico 1: Heatmap missing")

knitr::kable(table(age<=71),col.names=c('Age<=71','Freq.'))

# categorizando variável age

data$age_cat <- sapply(data$age,
                       function(x){
                         if (x <= 71) x = '<=71'
                         else x = '>71'
                       })
data$age_cat <- as.factor(data$age_cat)

data$died <- NULL
data$dstr <- NULL

# removendo observações missing (7 observações)
data <- subset(subset(data, !is.na(minf)))

# removendo observações missing (3 observações)
data <- subset(subset(data, !is.na(coma)))

# Taxa de óbtos
ggplot(data, aes(x= dead,fill=dead))+
          geom_bar(aes(y = (..count..)/sum(..count..))) +
          scale_y_continuous(labels=scales::percent) +
  theme(legend.position = "none") +
  scale_fill_brewer(palette="Set1") +
```

```r
  labs(y = "Pacientes (%)",
       x="Óbito",
       title="Gráfico 2: Taxa de óbtos")

# Óbito vs. Diagnóstico
ggplot(data, aes(x= dgn,  group=dead)) +
    geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count") +
    geom_text(aes( label = scales::percent(..prop..),
                   y= ..prop.. ), stat= "count", vjust = -.5) +
    labs(y = "Pacientes (%)",
        x="Diagnóstico",
        title="Gráfico 3: Óbito vs. Diagnóstico") +
    facet_grid(~dead) +
    scale_fill_brewer(palette="Set1") +
    theme(legend.position = "none") +
    scale_y_continuous(labels = scales::percent, limits = c(0,0.73))

# Histograma Idade vs. óbito
ggplot(data, aes(x=age,  color=dead,fill=dead)) +
  geom_histogram(color="black", fill="white")+
  facet_grid(dead ~ .) +
  labs(x="Idade (em anos)",
       y="# de pacientes",
       title="Gráfico 4: Histograma Idade vs. óbito")

# covariáveis vs. tempo de sobrev.
plot1 <- ggplot(data, aes(x=sex, y=obsmonths)) +
  geom_boxplot() +
  labs(y = "Tempo (em meses)",
       x = "Sexo",
       title = "Tempo vs. Sexo")

plot2 <- ggplot(data, aes(x=dgn, y=obsmonths)) +
  geom_boxplot() +
  labs(y = "Tempo (em meses)",
       x = "Diagnóstico",
       title = "Tempo vs. diagnóstico")

plot3 <- ggplot(data, aes(x=coma, y=obsmonths)) +
  geom_boxplot() +
  labs(y = "Tempo (em meses)",
       x = "Coma",
       title = "Tempo vs. Coma")

plot4 <- ggplot(data, aes(x=minf, y=obsmonths)) +
  geom_boxplot() +
  labs(y = "Tempo (em meses)",
       x = "Histórico de infarto",
       title = "Tempo vs. Histórico de infarto")

plot5 <- ggplot(data, aes(x=dead, y=obsmonths)) +
  geom_boxplot() +
  labs(y = "Tempo (em meses)",
```

```r
       x = "Óbito",
       title = "Tempo vs. Óbito")

grid.arrange(plot1, plot3, plot2,
             plot5, plot4, ncol=3,nrow=2)

ekm_sex <- survfit(Surv(obsmonths, dead)~ sex,data = data)

# Grafico Kaplan-Meier sex
ggsurvplot(ekm_sex, data = data,conf.int = T,palette="Set1",pval = T,risk.table = T,
           ggtheme=theme_gray()) +
  labs(x="Tempo (em meses)",
       y=expression(hat(S(t))),
       title = "Gráfico 10: Estimativas de Kaplan-Meier (Sexo")

# log-rank sex
knitr::kable( surv_pvalue(ekm_sex,data, method = c("1"))[,1:3])

ekm_minf <- survfit(Surv(obsmonths, dead)~ minf,data = data)

# Grafico Kaplan-Meier hist inf
ggsurvplot(ekm_minf, data = data,conf.int = T,palette="Set1",pval = T,risk.table = T,
           ggtheme=theme_gray()) +
  labs(x="Tempo (em meses)",
       y=expression(hat(S(t))),
       title = "Gráfico 11: Estimativas de Kaplan-Meier (Hist. infarto)")

# log-rank hist inf
knitr::kable( surv_pvalue(ekm_minf,data, method = c("1"))[,1:3])

ekm_coma <- survfit(Surv(obsmonths, dead)~ coma,data = data)

# Grafico Kaplan-Meier coma
ggsurvplot(ekm_coma, data = data,conf.int = T,palette="Set1",pval = T,risk.table = T,
           ggtheme=theme_gray()) +
  labs(x="Tempo (em meses)",
       y=expression(hat(S(t))),
       title = "Gráfico 12: Estimativas de Kaplan-Meier (Coma)")

# log-rank coma
knitr::kable( surv_pvalue(ekm_coma,data, method = c("1"))[,1:3])

ekm_age <- survfit(Surv(obsmonths, dead)~ age_cat,data = data)

# Grafico Kaplan-Meier age cat
ggsurvplot(ekm_age, data = data,conf.int = T,palette="Set1",pval = T,risk.table = T,
           ggtheme=theme_gray()) +
  labs(x="Tempo (em meses)",
       y=expression(hat(S(t))),
       title = "Gráfico 13: Estimativas de Kaplan-Meier (Idade cat.)")

# log-rank age
knitr::kable( surv_pvalue(ekm_age,data, method = c("1"))[,1:3])
```

```r
ekm_dgn <- survfit(Surv(obsmonths, dead)~ dgn,data = data)

# Grafico Kaplan-Meier diag
ggsurvplot(ekm_dgn, data = data,conf.int = T,palette="Set1",pval = T,risk.table = T,
           ggtheme=theme_gray()) +
  labs(x="Tempo (em meses)",
       y=expression(hat(S(t))),
       title = "Gráfico 14: Estimativas de Kaplan-Meier")

# log-rank diag
knitr::kable( surv_pvalue(ekm_dgn,data, method = c("1"))[,1:3])

## item b

# modelo completo
mod.w <- survreg(Surv(obsmonths, dead)~ sex+dgn+coma+minf+age_cat, dist='weibull',
                 data = data)
summary(mod.w)

#step(mod.w)

#modelo reduzido
mod.w1 <- survreg(Surv(obsmonths, dead)~ dgn+coma+minf+age_cat, dist='weibull',
                  data = data)
summary(mod.w1)

## item d

#v2 <- ifelse(data$sex=="Male",1,0)
v2 <- ifelse(data$dgn=="ID",1,0)
v3 <- ifelse(data$dgn=="INF",1,0)
v4 <- ifelse(data$dgn=="SAH",1,0)
v5 <- ifelse(data$coma=="Yes",1,0)
v6 <- ifelse(data$minf=="Yes",1,0)
v7 <- ifelse(data$age_cat==">71",1,0)

xb_wei <- mod.w1$coef[1]+mod.w1$coef[2]*v2+mod.w1$coef[3]*v3+mod.w1$coef[4]*v4+
  mod.w1$coef[5]*v5+mod.w1$coef[6]*v6+mod.w1$coef[7]*v7

coxsnell_wei<- (data$obsmonths^(1/mod.w1$scale))*exp(-xb_wei/mod.w1$scale)

# Curva de Kaplan-Meier
KM_wei <-  survfit(Surv(coxsnell_wei, data$dead)~1)
TFAcum_KM_wei <- -log(KM_wei$surv)

# Estimador de Nelson_Aalen
Surv_Aa_wei <- survfit(coxph(Surv(coxsnell_wei, data$dead)~1,method='breslow'))
TFAcum_Aa_wei <- -log(Surv_Aa_wei$surv)

#Gráfico
plot(KM_wei$time,TFAcum_KM_wei, col="dark red", pch=16,
     main="Gráfico 15: Função de risco acumulada - Weibull", xlab="Tempo", ylab="-log(S(t))", cex=0.8 )
points(Surv_Aa_wei$time,TFAcum_Aa_wei, col="navy blue", pch=16, cex=0.8)
```

```r
abline(0,1,lty=2)

## item e

# modelo completo
mod.ll <- survreg(Surv(obsmonths, dead)~ sex+dgn+coma+minf+age_cat, dist='loglogistic',data = data)
summary(mod.ll)

#step(mod.w)

#modelo reduzido
mod.ll1 <- survreg(Surv(obsmonths, dead)~ dgn+coma+minf+age_cat, dist='loglogistic',data = data)
summary(mod.ll1)

## item g

xb_llog <- mod.ll1$coef[1]+mod.ll1$coef[2]*v2+mod.ll1$coef[3]*v3+mod.ll1$coef[4]*v4+mod.ll1$coef[5]*v5+
  mod.ll1$coef[6]*v6+mod.ll1$coef[7]*v7

coxsnell_llog <- log(1+(data$obsmonths^(1/mod.ll1$scale))*exp(-xb_llog/mod.ll1$scale))

# Curva de Kaplan-Meier
KM_llog <- survfit(Surv(coxsnell_llog, data$dead)~1)
TFAcum_KM_llog <- -log(KM_llog$surv)

# Estimador de Nelson_Aalen
Surv_Aa_llog <- survfit(coxph(Surv(coxsnell_llog, data$dead)~1,method='breslow'))
TFAcum_Aa_llog <- -log(Surv_Aa_llog$surv)

#Gráfico
plot(KM_llog$time,TFAcum_Aa_llog, col="dark red", pch=16,
     main="Gráfico 16: Função de risco acumulada - log-losgística", xlab="Tempo", ylab="-log(S(t))", ce
points(Surv_Aa_llog$time,TFAcum_Aa_llog, col="navy blue", pch=16, cex=0.8)
abline(0,1,lty=2)

## item h

AIC_llog <- -2*mod.ll1$loglik[2]+2*9
AIC_wei <- -2*mod.w1$loglik[2]+2*9

n <- dim(data)[1]

BIC_wei <- -2*mod.w1$loglik[2]+9*log(n)
BIC_llog <- -2*mod.ll1$loglik[2]+9*log(n)

df <- data.frame(cbind(c(AIC_wei,AIC_llog),
c(BIC_wei,BIC_llog)),row.names = c("Weibull","Log-logistico"))
colnames(df) <- c('AIC','BIC')

knitr::kable(df)
```