Análise de Sobrevida

Teoria e Aplicações em Saúde

Caderno de Respostas

Marilia Sá Carvalho
Valeska Lima Andreozzi
Claudia Torres Codeço
Maria Tereza Serrano Barbosa
Silvia Emiko Shimakura

Sumário

2 O tempo na análise de sobrevida 3

3 Funções básicas de sobrevida 18

4 Estimação não-paramétrica 22

5 Estimação paramétrica 39

6 Modelos de regressão paramétricos 54

7 Modelos de regressão semiparamétricos 67

8 Análise de resíduos para modelos de Cox 86

9 Covariáveis tempo-dependentes 98

10 Eventos múltiplos 115

11 Fragilidade 131

2

O tempo na análise de sobrevida

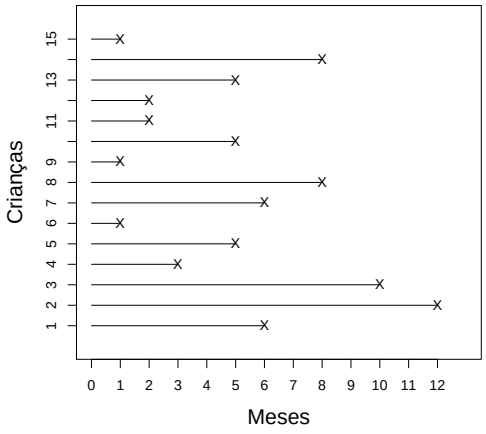
Exercícios

Exercício 2.1: O tempo de aleitamento, isto é, o tempo decorrido desde o nascimento até o desmame, pode ser considerado como uma variável tempo de sobrevida. Suponha que o tempo até o desmame, em meses, tenha sido registrado para 15 crianças:

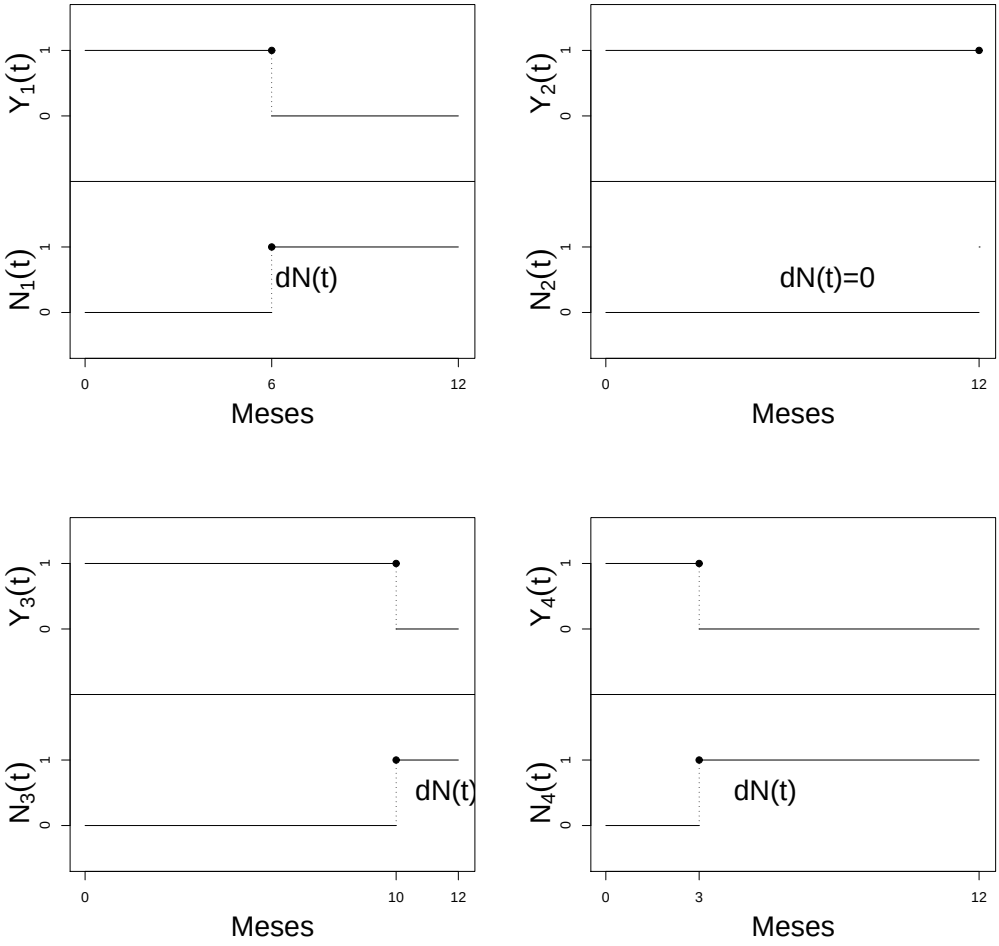
6 12 10 3 5 1 6 8 1 5 2 2 5 8 1

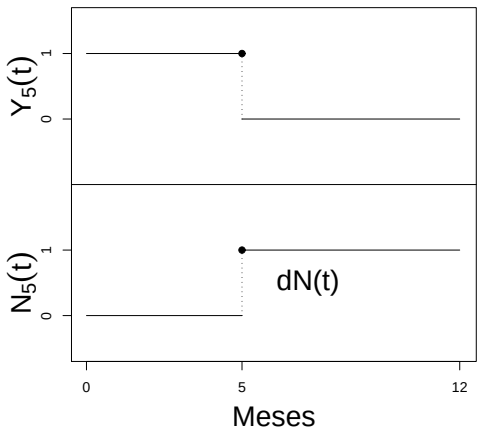
Considerando que não houve censura:

1. Represente graficamente os tempos de observação das 15 crianças.



2. Represente as trajetórias dos primeiros cinco indivíduos utilizando as variáveis $N(t)$ e $Y(t)$ do processo de contagem.





3. Como você construiria um banco de dados para analisar estes dados pelo processo clássico?

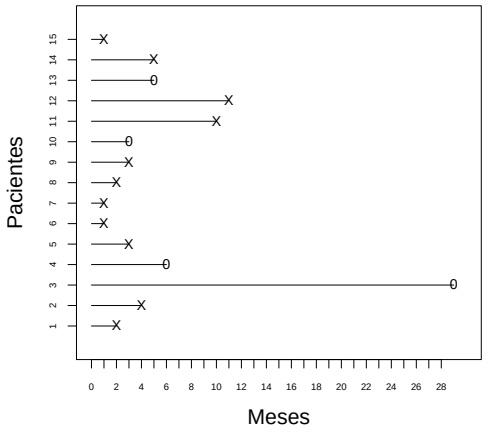
crianca	tempo	status
1	6	1
2	12	1
3	10	1
4	3	1
5	5	1
6	1	1
7	6	1
8	8	1
9	1	1
10	5	1
11	2	1
12	2	1
13	5	1
14	8	1
15	1	1

Exercício 2.2: Considere agora o tempo de sobrevida de 15 pacientes submetidos a hemodiálise. Neste caso, a variável de interesse é o tempo desde a primeira diálise até o óbito.

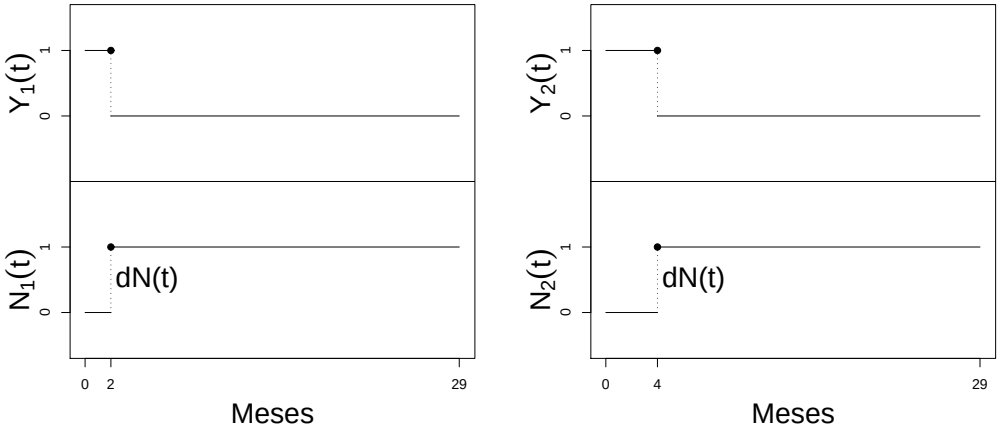
2 4 29+ 6+ 3 1 1 2 3 9+ 10 11 5+ 5 1

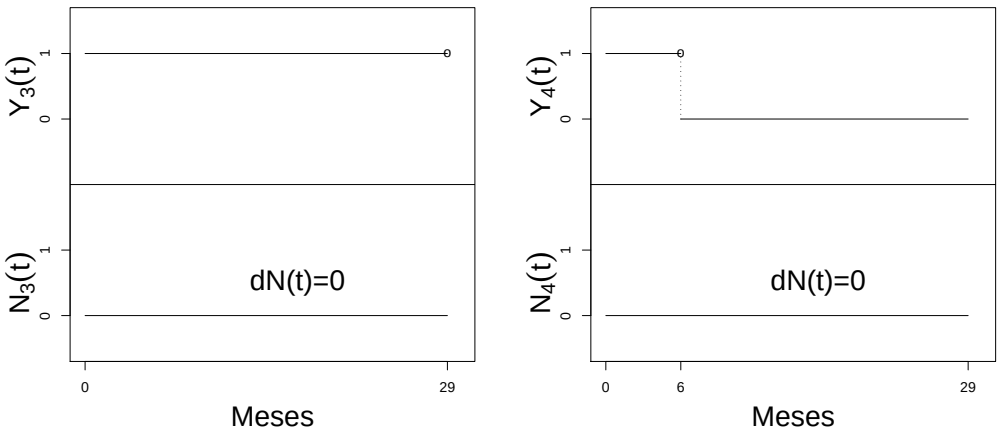
Os tempos censurados (censura à direita) estão indicados pelo sinal de +. Considere que todos os pacientes entraram juntos no início do estudo.

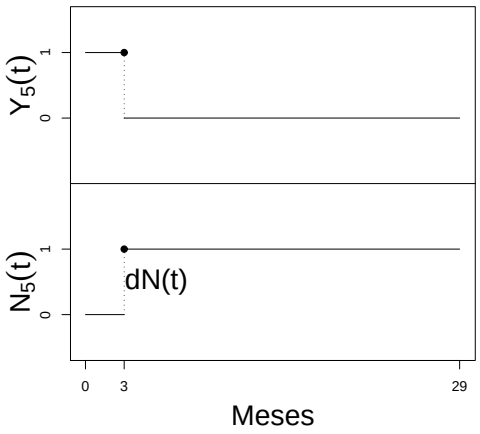
1. Represente graficamente os tempos de observação dos pacientes, utilizando abordagem clássica.



2. Represente as trajetórias dos primeiros cinco indivíduos utilizando as variáveis $N(t)$ e $Y(t)$ do processo de contagem.





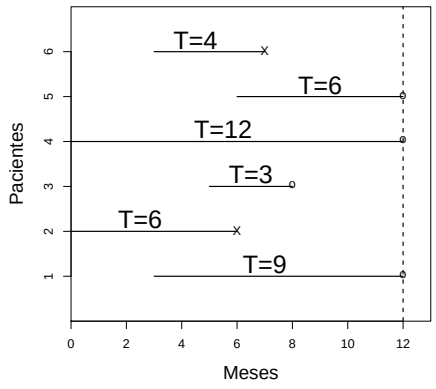


3. Como você construiria um banco de dados para analisar esses dados pelo processo clássico?

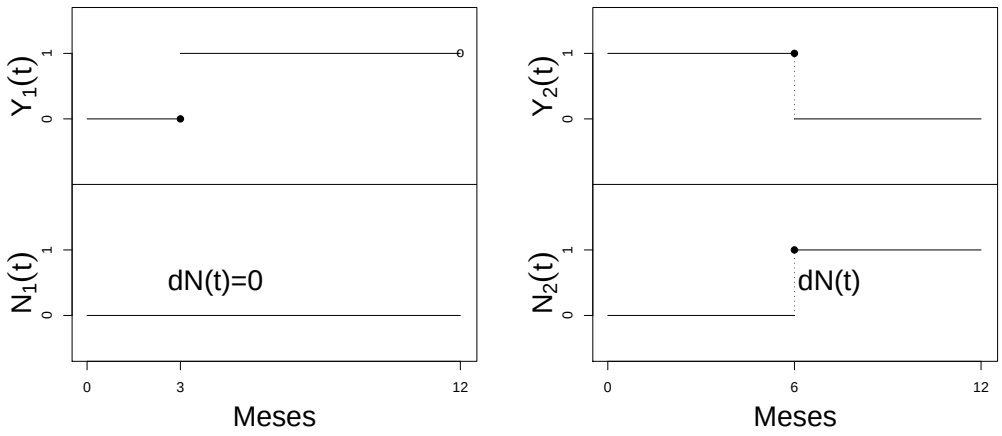
paciente	tempo	status
1	2	1
2	4	1
3	29	0
4	6	0
5	3	1
6	1	1
7	1	1
8	2	1
9	3	1
10	9	0

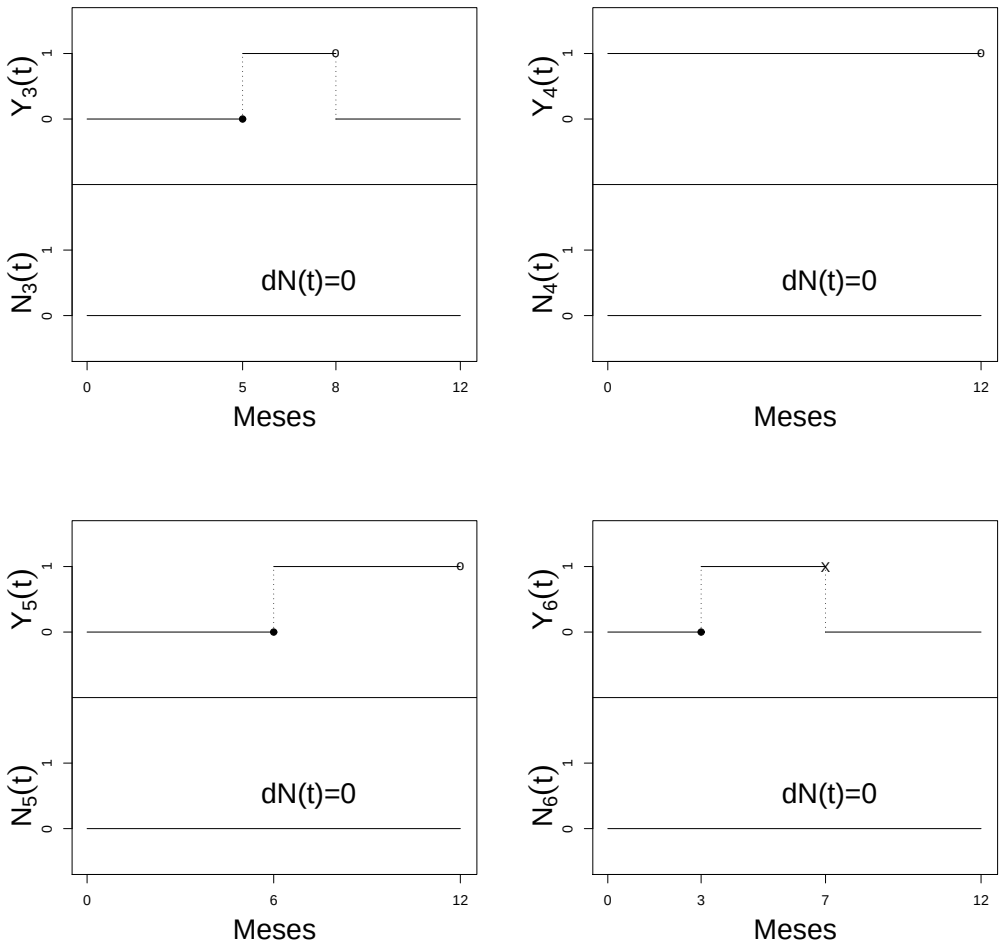
11	10	1
12	11	1
13	5	0
14	5	1
15	1	1

Exercício 2.3: Suponha que, em um hospital, 6 pacientes HIV positivo são acompanhados ao longo de um ano. No gráfico abaixo, as linhas horizontais representam o tempo de acompanhamento de cada paciente (linhas terminadas em $\times$ indicam a ocorrência do desfecho (óbito), linhas terminadas com $\circ$ indicam censuras). Represente, utilizando as variáveis $N(t)$ e $Y(t)$, as trajetórias de cada um dos 6 pacientes.

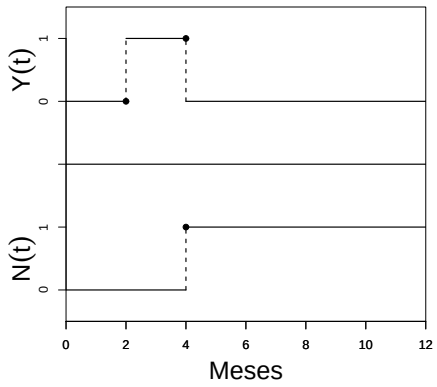


A seguir os gráficos dos 6 pacientes utilizando o processo de contagem





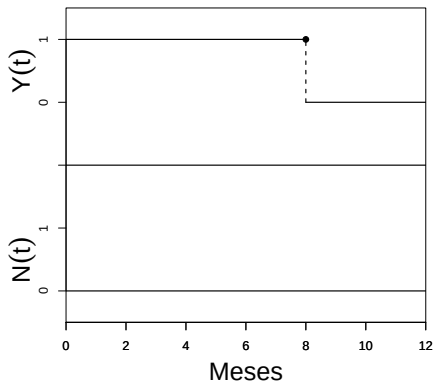
Exercício 2.4: Um paciente tem a seguinte trajetória de observação, segundo o processo de contagem:



Com base neste gráfico, responda:

1. Qual foi o mês de entrada do paciente no estudo?
   Resposta: Mês dois
2. Em que mês ocorreu o desfecho?
   Resposta: Mês quatro
3. Quais os meses em que o paciente estava sob risco de óbito?
   Resposta: Meses dois, três e quatro.

Exercício 2.5: Um outro paciente tem a seguinte trajetória de observação:



Com base neste gráfico, responda:

1. Qual foi a data de entrada do paciente no estudo?

   Resposta: Mês zero

2. Em que data ocorreu o desfecho?

   Resposta: Mês oito

Exercício 2.6: No R, abra o banco de dados dos pacientes com Aids, atendidos no Ipec (no Apêndice C1 você encontra uma descrição deste banco de dados). Este banco está em formato *csv* e também pode ser visualizado em um programa de planilha. Lembre-se de indicar ao R em que diretório está o arquivo com o banco de dados usando o menu *File > Change Dir*.

Os dados estão disponíveis em http://dengue.procc.fiocruz.br/∼sobrevida/dados/. Copie o banco *ipec.csv* para este mesmo diretório de trabalho.

```
> ipec <- read.table("ipec.csv", header = T, sep = ";")
> ipec[1:5, ]
```

```
  id  ini  fim tempo status sexo escola idade risco acompan obito anotrat
1  1 1243 2095   852      1    M      3    34     0       1     S    1991
2  2 2800 2923   123      1    M      2    38     6       1     S       9
3  3 1250 2395  1145      1    M     NA    32     0       1     S    1992
4  4 1915 4670  2755      0    M     NA    43     6       0     N    1992
5  5 2653 4770  2117      0    M     NA    40     0       1     N    1992
  tratam doenca propcp
1      1      4      3
2      0      7      4
3      1      3      4
4      1     10      4
5      1      5      4
```

Crie a variável resposta da análise de sobrevida, combinando as informações de tempo e status no formato clássico e no formato de processo de contagem. Não esqueça de chamar a biblioteca *survival*.

```
> require(survival)
```

Formato clássico:

```
> Surv(ipec$tempo, ipec$status)
```

```
  [1]  852   123  1145  2755+ 2117+  329+   60   151  1563  1247    84   214
 [13]   25+ 1348   158   555   408  1116   998  1125   944+   54   151   855
 [25]  116    80+ 1757+  194   183    37   237+ 1506   168+  134   803+   18
 [37]  371   173   688   163  3178+   29    50+  887   516   645   310   204
 [49] 1344+ 1261   285    83   150  1307+ 1076+ 1226   865+  811  2898    80
 [61]  967   618   235  2236+  152   892    81+ 1085  1073+ 1615+   35   290
 [73] 1780+ 3228+   52   733  3213+ 1983  2304+  572    21  1272+ 1646+  304
 [85]  418   854  2973+   40   850  1139   323  1507+ 2717+ 1735+  388+  145
 [97]  905   927  1027+  631  2495+ 1331+  623  2568+ 2013+  721  1952+  397
[109]  254  1630+ 1523+  146+  108  1835+  499   333   202+ 2437+ 1015  2138+
[121]   22  2090+  179  2439+ 1063+   85+  343+ 2215+  259  2258+ 1371    39
[133] 2371+  975+  952  2492+ 1478+  295+  992  1011+  644   426   537+ 1454+
[145] 1869+  714+ 1310+ 2084+ 1918+ 1649+  290+ 1685+ 1348+  652+ 1384+ 1471+
[157] 1512+  378+ 1352+  419  1426+ 1488+ 1315+  643+ 1197+ 1343+ 1176+  944
[169]  340   881+  915+  948+  985+ 1242+  955+  987+  899+ 1056+  775   785+
[181]  731+   16   680+   21+  444+  524+  217+  440+  470+  390+  344+  578+
[193]  504+
```

Formato de processo de contagem:

```
> Surv(ipec$ini, ipec$fim, ipec$status)

  [1] (1243,2095 ] (2800,2923 ] (1250,2395 ] (1915,4670+] (2653,4770+]
  [6] (   3, 332+] (  36,  96 ] (   1, 152 ] ( 544,2107 ] (  71,1318 ]
 [11] ( 946,1030 ] ( 802,1016 ] ( 266, 291+] (1544,2892 ] (  57, 215 ]
 [16] (1270,1825 ] (2753,3161 ] ( 940,2056 ] ( 393,1391 ] (1000,2125 ]
 [21] ( 238,1182+] ( 423, 477 ] ( 206, 357 ] ( 480,1335 ] ( 226, 342 ]
 [26] ( 249, 329+] (3052,4809+] (1802,1996 ] (1395,1578 ] ( 354, 391 ]
 [31] ( 493, 730+] (1113,2619 ] ( 638, 806+] ( 655, 789 ] (1189,1992+]
 [36] ( 943, 961 ] (1715,2086 ] ( 792, 965 ] (1037,1725 ] ( 820, 983 ]
 [41] ( 884,4062+] (2262,2291 ] (1121,1171+] (1131,2018 ] ( 878,1394 ]
 [46] (1316,1961 ] (1107,1417 ] (1190,1394 ] ( 393,1737+] (1274,2535 ]
 [51] (1172,1457 ] (2360,2443 ] (2074,2224 ] (1019,2326+] ( 605,1681+]
 [56] (1915,3141 ] (3948,4813+] (1314,2125 ] (1502,4400 ] (1347,1427 ]
 [61] (1379,2346 ] (2352,2970 ] (2625,2860 ] (2586,4822+] (1406,1558 ]
 [66] (1466,2358 ] (3314,3395+] (3413,4498 ] (3712,4785+] (3207,4822+]
 [71] (1592,1627 ] (1537,1827 ] (3018,4798+] (1555,4783+] (1541,1593 ]
 [76] (1589,2322 ] (1609,4822+] (1682,3665 ] (2465,4769+] (1243,1815 ]
 [81] (1667,1688 ] (1605,2877+] (3157,4803+] (2066,2370 ] (1929,2347 ]
 [86] (2216,3070 ] (1809,4782+] (1670,1710 ] (1983,2833 ] (2883,4022 ]
 [91] (1766,2089 ] (3313,4820+] (1977,4694+] (3087,4822+] (2286,2674+]
 [96] (1877,2022 ] (1852,2757 ] (1549,2476 ] (3795,4822+] (2475,3106 ]
[101] (2310,4805+] (2870,4201+] (1935,2558 ] (2199,4767+] (2800,4813+]
[106] (2990,3711 ] (2857,4809+] (3586,3983 ] (2143,2397 ] (3124,4754+]
[111] (3276,4799+] (2208,2354+] (2209,2317 ] (2976,4811+] (2626,3125 ]
[116] (3838,4171 ] (2314,2516+] (2311,4748+] (2280,3295 ] (2684,4822+]
```

```
[121] (2454,2476 ] (2713,4803+] (2311,2490 ] (2370,4809+] (3756,4819+]
[126] (2565,2650+] (2599,2942+] (2553,4768+] (2601,2860 ] (2553,4811+]
[131] (2726,4097 ] (2739,2778 ] (2447,4818+] (3830,4805+] (2429,3381 ]
[136] (2311,4803+] (3299,4777+] (4510,4805+] (2384,3376 ] (3749,4760+]
[141] (2676,3320 ] (2985,3411 ] (4192,4729+] (3159,4613+] (2921,4790+]
[146] (4078,4792+] (2934,4244+] (2645,4729+] (2857,4775+] (3173,4822+]
[151] (4509,4799+] (3082,4767+] (3465,4813+] (3188,3840+] (3271,4655+]
[156] (3276,4747+] (3287,4799+] (4439,4817+] (3446,4798+] (3305,3724 ]
[161] (3391,4817+] (3307,4795+] (3425,4740+] (4117,4760+] (3612,4809+]
[166] (3479,4822+] (3572,4748+] (3796,4740 ] (3527,3867 ] (3921,4802+]
[171] (3798,4713+] (3808,4756+] (3772,4757+] (3557,4799+] (3867,4822+]
[176] (3594,4581+] (3923,4822+] (3733,4789+] (4019,4794 ] (4033,4818+]
[181] (4040,4771+] (4053,4069 ] (4137,4817+] (4208,4229+] (4362,4806+]
[186] (4279,4803+] (4593,4810+] (4320,4760+] (4343,4813+] (4419,4809+]
[191] (4406,4750+] (4199,4777+] (4301,4805+]
```

Exercício 2.7: Para ir se familiarizando com o R e com o banco de dados do Ipec, que será utilizado nos próximos capítulos, faça uma análise exploratória das variáveis de interesse. A seguir estão alguns comandos necessários para arrumar o banco de dados e em seguida explorá-lo.

Lendo o banco de dados:

```
> ipec <- read.table("ipec.csv", header = T, sep = ";")
```

Listando os primeiros cinco registros:

```
> ipec[1:5, ]
```

```
  id  ini  fim tempo status sexo escola idade risco acompan obito anotrat
1  1 1243 2095   852      1    M      3    34     0       1     S    1991
2  2 2800 2923   123      1    M      2    38     6       1     S       9
3  3 1250 2395  1145      1    M     NA    32     0       1     S    1992
4  4 1915 4670  2755      0    M     NA    43     6       0     N    1992
5  5 2653 4770  2117      0    M     NA    40     0       1     N    1992
  tratam doenca propcp
1      1      4      3
2      0      7      4
3      1      3      4
4      1     10      4
5      1      5      4
```

Tamanho do banco de dados (número de registros e número de variáveis, respectivamente):

```
> dim(ipec)

[1] 193  15
```

Listando o nome das variáveis contidas no banco:

```
> names(ipec)

 [1] "id"      "ini"     "fim"     "tempo"   "status"  "sexo"    "escola"
 [8] "idade"   "risco"   "acompan" "obito"   "anotrat" "tratam"  "doenca"
[15] "propcp"
```

No R, o código para informação ignorada é *NA*, logo precisamos substituir os códigos 9, 99 ou I por *NA*

```
> ipec$anotrat[ipec$anotrat == 9] <- NA
> ipec$doenca[ipec$doenca == 99] <- NA
> ipec$obito <- factor(ipec$obito, levels = c("N", "S"))
```

É preciso indicar ao R quais são as variáveis cujos valores numéricos representam categorias. Se você quiser indicar um *label* para estes fatores utilize o comando *factor* com a descrição dos *label*s, acompanhando a ordem numérica do registro. A descrição está na Tabela C.1.

```
> ipec$escola <- factor(ipec$escola, labels = c("sem", "fundam",
+     "medio", "sup"))
> ipec$risco <- factor(ipec$risco)
> ipec$acompan <- factor(ipec$acompan)
> ipec$anotrat <- factor(ipec$anotrat)
> ipec$tratam <- factor(ipec$tratam)
> ipec$doenca <- factor(ipec$doenca)
> ipec$propcp <- factor(ipec$propcp)
```

ANÁLISE EXPLORATÓRIA

No R o nome das variáveis deve ser precedido do nome do objeto seguido do símbolo $, como vínhamos usando anteriormente. Podemos encurtar essa sintaxe utilizando-se a função *attach()* permitindo, desta forma, que os nomes das variáveis do objeto *ipec* sejam usados diretamente.

```
> attach(ipec)
```

Cálculo das medidas resumo:

```
> summary(ipec)
```

```
      id            ini             fim            tempo             status
 Min.   :  1   Min.   :   1   Min.   :  96   Min.   :  16.0   Min.   :0.0000
 1st Qu.: 49   1st Qu.:1406   1st Qu.:2095   1st Qu.: 290.0   1st Qu.:0.0000
 Median : 97   Median :2454   Median :3711   Median : 852.0   Median :0.0000
 Mean   : 97   Mean   :2397   Mean   :3335   Mean   : 938.2   Mean   :0.4663
 3rd Qu.:145   3rd Qu.:3314   3rd Qu.:4790   3rd Qu.:1348.0   3rd Qu.:1.0000
 Max.   :193   Max.   :4593   Max.   :4822   Max.   :3228.0   Max.   :1.0000

 sexo       escola       idade         risco    acompan  obito       anotrat
 F: 49   sem   :59   Min.   :20.00   0   :87   0:57    N   :80   1992   :22
 M:144   fundam:44   1st Qu.:30.00   1   : 9   1:99    S   :92   1995   :20
         medio :55   Median :35.00   2   : 7   2:37    NA's:21   1993   :19
         sup   :24   Mean   :36.55   3   :30                     1996   :18
         NA's  :11   3rd Qu.:43.00   5   :16                     1994   :16
                     Max.   :68.00   6   : 7                     (Other):54
                                     NA's:37                     NA's   :44
 tratam      doenca   propcp
 0: 44   3      :31   0: 38
 1:100   10     :25   2: 24
 2: 35   7      :17   3:  3
 3: 14   1      :12   4:128
         8      :12
         (Other):29
         NA's   :67
```

Para as variáveis categóricas podemos também usar a função *table()* para calcular as freqüências de cada categoria:

```
> table(escola)
```

```
escola
   sem fundam  medio    sup
    59     44     55     24
```

```
> table(tratam)
```

```
tratam
  0   1   2   3
 44 100  35  14
```

Distribuição das idades dos pacientes por gênero:

```
> boxplot(idade ~ sexo, main = "Idade por Gênero", ylab = "Idade",
+     xlab = "Sexo")
```

Número de pacientes por grupo de categoria provável de transmissão e por sexo:

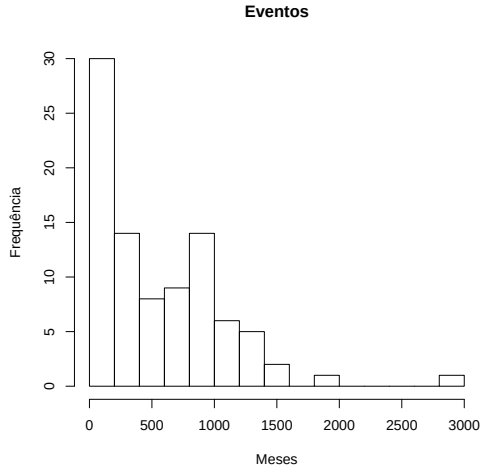
```
> table(risco, sexo)

     sexo
risco  F  M
    0  0 87
    1  2  7
    2  2  5
    3 25  5
    5  2 14
    6  1  6
```

Número de eventos e censuras observadas:

```
> table(status)

status
  0   1
103  90
```

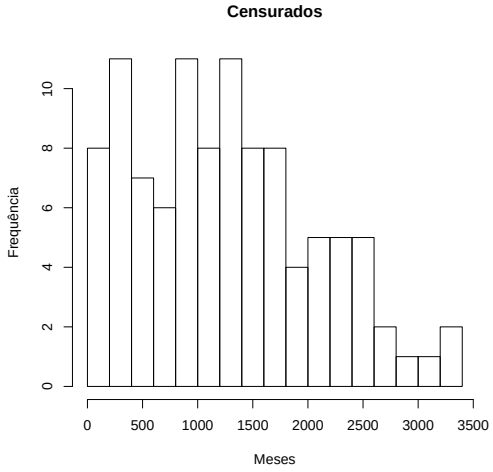
Distribuição dos tempos de sobrevida:

```
> hist(tempo[status == 1], breaks = 12, main = "Eventos", ylab = "Frequência",
+     xlab = "Meses")
```



Distribuição dos tempos de censura:

```
> hist(tempo[status == 0], breaks = 12, main = "Censurados", ylab = "Frequência",
+     xlab = "Meses")
```

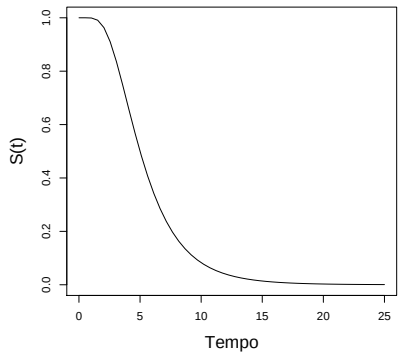


3

Funções básicas de sobrevida

Exercícios

Exercício 3.1: A figura abaixo mostra uma curva de sobrevida. Com base nesta curva, identifique:



1. a probabilidade de sobreviver por mais de 10 dias;

   Resposta: Faça uma reta do tempo igual a 10 dias até encontrar a curva. Procure agora no eixo das ordenadas (vertical) a qual valor de sobrevida corresponde este tempo. Aproximadamente 0,083.

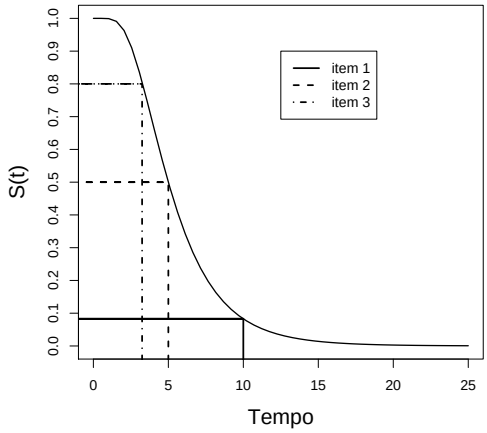
2. a sobrevida mediana e o tempo mediano de sobrevida;

   Resposta: Sobrevida mediana que é também denominada tempo mediano de sobrevida é igual a 5 dias.

3. o tempo em que 80% dos pacientes ainda estavam vivos.

   Resposta: Aproximadamente 4 dias.

Vejam as respostas identificadas no gráfico a seguir:



Exercício 3.2: Sabendo-se que a probabilidade de sobreviver mais que 100 dias após o transplante de coração é igual a 0,7, calcule:

1. a probabilidade de sobreviver até 100 dias (inclusive);

   Resposta: Temos que a probabilidade de viver mais que 100 dias é igual a 0,7, logo $S(100) = Pr(T > 100) = 0,7$. Através da relação $S(t) = 1 - F(t)$, temos $F(t) = 1 - S(t) = 0,3$.

2. o risco acumulado de óbito até 100 dias.

   Resposta: Para calcular o risco acumulado podemos usar a seguinte relação: $\Lambda(t) = -\ln(S(t)) = -\ln(0,7) = 0.37$.

Exercício 3.3: A tabela a seguir mostra o tempo até o óbito de alguns pacientes de uma coorte de 32 indivíduos vivendo com aids. Todos os 32 pacientes morreram antes do fim do estudo, não havendo tempos censurados. Complete as lacunas em branco usando as definições e as relações entre as funções básicas de sobrevida. T é o tempo do óbito, $R(t)$ é o número de pessoas sob risco no início do intervalo de tempo, $N(t)$ é o número de óbitos ocorridos no intervalo de tempo.

Intervalo	$R(t)$	$N(t)$	$\hat{f}(t)$	$\hat{F}(t)$	$\hat{S}(t)$	$\hat{\lambda}(t)$	$\hat{\Lambda}(t)$
(0,3]	32	1	0,0104	0,000	1,000	0,0104	0
(3,18]	31	1	0,0020	0,031	0,968	0,0021	0,031
(18,29]	30	1	0,0028	0,062	0,937	0,0030	0,063
$\vdots$							
(145,151]	20	2	0,0104	0,375	0,625	0,0166	0,460
(151,158]	18	1	0,0044	0,437	0,562	0,0079	0,560
(158,173]	17	1	0,0020	0,468	0,531	0,0039	0,616
(173,194]	16	1	0,00149	0,5000	0,5000	0,00298	0,6751
(194,214]	15	1	0,00156	0,5313	0,4688	0,00333	0,7376
(214,329]	14	1	0,00027	0,5625	0,4375	0,00062	0,8043
(329,331]	13	1	0,01563	0,5938	0,4063	0,03846	0,8757
(331,371]	12	1	0,00078	0,6250	0,3750	0,00208	0,9527

Exemplo do cálculo das funcões básicas para o intervalo (173,194]:

$$\begin{aligned}
\hat{f}_x(t) &= \frac{N_x(t)}{(\text{n}^o \text{ total de ocorrências}) \times \Delta_x} = \frac{1}{32 \times (194-173)} = 0,00149 \\
\hat{S}_x(t) &= \frac{R_x(t)}{\text{n}^o \text{ total de pacientes}} = \frac{18}{32} = 0,5 \\
\hat{F}_x(t) &= 1 - \hat{S}_x(t) = 1 - 0,5 = 0,5 \\
\hat{\lambda}_x(t) &= \frac{N_x(t)}{R_x(t) \times \Delta_x} = \frac{1}{16 \times (194-173)} = 0,00298 \\
\hat{\Lambda}_x(t) &= \sum_{k=1}^{x-1} \hat{\lambda}_x(t) \times \Delta_x = 0,616 + 0,00298 \times (194-173)
\end{aligned}$$

Exercício 3.4: Interprete os valores $S(214)$, $F(214)$, $\lambda(214)$ e $\Lambda(214)$ do exercício anterior, lembrando que o evento é óbito em pacientes vivendo com aids. Olhando a tabela responda qual o tempo mediano de sobrevida desta coorte.

$S(214) = 0,4375$ - A probabilidade de sobreviver mais que 214 dias com Aids é igual a 0,4375.

$F(214) = 0,5625$ - A probabilidade de sobreviver com Aids até 214 dias é igual a 0,5625.

$\lambda(214) = 0,00062$ - o risco de morrer por Aids aos 214 dias é igual a 0,00062.

$\Lambda(214) = 0,8043$ - o risco acumulado de morrer por Aids até 214 dias é 0,8043.

O tempo mediano de sobrevida, isto é quando a $S(t) = 0,50$, é igual a 173 dias.

4

Estimação não-paramétrica

Exercícios

Exercício 4.1: Retorne aos dados de aleitamento do exercício 2.1. Construa, à mão, a tabela Kaplan-Meier e a curva Kaplan-Meier correspondente aos dados:

t_i	$R(t)$	$\Delta N(t)$	$\hat{S}_{KM}(t) = \prod_{t_i \leq t} \frac{R(t_i) - \Delta N(t_i)}{R(t_i)}$
1	15	3	$\frac{(15-3)}{15} = 0,8000$
2	12	2	$0,8000 \times \frac{(12-2)}{12} = 0,6667$
3	10	1	$0,6667 \times \frac{(10-1)}{10} = 0,6000$
5	9	3	$0,6000 \times \frac{(9-3)}{9} = 0,4000$
6	6	2	$0,4000 \times \frac{(6-2)}{6} = 0,2667$
8	4	2	$0,2667 \times \frac{(4-2)}{4} = 0,1333$
10	2	1	$0,1333 \times \frac{(2-1)}{2} = 0,0667$
12	1	1	$0,0667 \times \frac{(1-1)}{1} = 0,0000$

O gráfico do exercício a seguir corresponde a esta tabela.

1. Com base na tabela que você criou, responda: qual a probabilidade de uma criança ser amamentada pelo menos até o sexto mês de vida? Qual a probabilidade de ser amamentada por mais de 3 meses? Qual é a probabilidade de ser

amamentada por mais de 10 meses? Qual foi o tempo mediano de aleitamento?
Resposta:

- a probabilidade de uma criança ser amamentada pelo menos até o sexto mês de vida é $S(6) = 0,2667$
- a probabilidade de ser amamentada por mais de 3 meses é $S(3) = 0,6$
- a probabilidade de ser amamentada por mais de 10 meses é $S(10) = 0,0667$
- o tempo mediano de aleitamento está entre 3 e 5 meses.

2. Construa a tabela Kaplan-Meier e a curva de sobrevida utilizando o pacote estatístico R. Os dados de aleitamento estão no arquivo *leite.txt*.

   Lendo o banco de dados

```
> leite <- read.table("leite.txt", header = T, sep = "")
```

   Criando o objeto que contém os dados de sobrevida

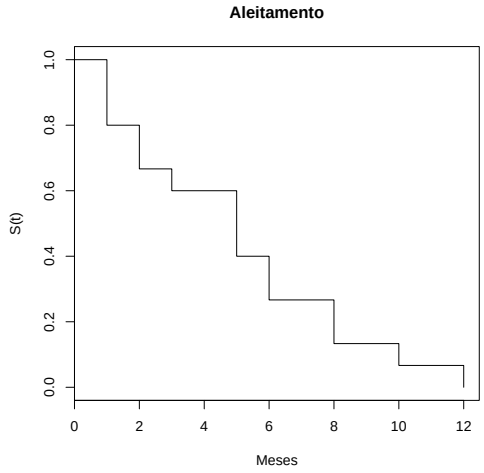
```
> y <- Surv(leite$tempo, leite$status)
```

   Calculando a sobrevida por KM

```
> km <- survfit(y ~ 1, data = leite)
> km

Call: survfit(formula = y ~ 1, data = leite)

      n  events  median 0.95LCL 0.95UCL
     15      15       5       2       8

> summary(km)

Call: survfit(formula = y ~ 1, data = leite)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     15       3   0.8000  0.1033       0.6212        1.000
    2     12       2   0.6667  0.1217       0.4661        0.953
    3     10       1   0.6000  0.1265       0.3969        0.907
    5      9       3   0.4000  0.1265       0.2152        0.743
    6      6       2   0.2667  0.1142       0.1152        0.617
    8      4       2   0.1333  0.0878       0.0367        0.484
   10      2       1   0.0667  0.0644       0.0100        0.443
   12      1       1   0.0000      NA           NA           NA
```

Fazendo o gráfico da função de sobrevida

```
> plot(km, conf.int = F, xlab = "Meses", ylab = "S(t)", main = "Aleitamento")
```



Exercício 4.2: Suponha que os tempos de aleitamento de 60 bebês estejam agrupados em quatro comunidades diferentes:

Comunidade 1: 6 12 10 3 5 1 6 8 1 5 2 2 5 8 1

Comunidade 2: 5 12 10 4 4 3 6 9 2 6 4 1 7 10 1

Comunidade 3: 13 14 20 3 5 1 8 15 2 5 3 2 6 15 1

Comunidade 4: 1 16 20 1 1 1 1 2 2 13 3 1 1 14 2

O arquivo *leite2.txt* contém esses dados. A variável *grupo* indica a comunidade à qual cada criança pertence.

Lendo o bando de dados

```
> leite2 <- read.table("leite2.txt", header = T, sep = "")
> leite2
```

```
   crianca tempo status grupo
1        1     6      1     1
2        2    12      1     1
3        3    10      1     1
4        4     3      1     1
```

5	5	5	1	1
6	6	1	1	1
7	7	6	1	1
8	8	8	1	1
9	9	1	1	1
10	10	5	1	1
11	11	2	1	1
12	12	2	1	1
13	13	5	1	1
14	14	8	1	1
15	15	1	1	1
16	16	5	1	2
17	17	12	1	2
18	18	10	1	2
19	19	4	1	2
20	20	4	1	2
21	21	3	1	2
22	22	6	1	2
23	23	9	1	2
24	24	2	1	2
25	25	6	1	2
26	26	4	1	2
27	27	1	1	2
28	28	7	1	2
29	29	10	1	2
30	30	1	1	2
31	31	13	1	3
32	32	14	1	3
33	33	20	1	3
34	34	3	1	3
35	35	5	1	3
36	36	1	1	3
37	37	8	1	3
38	38	15	1	3
39	39	2	1	3
40	40	5	1	3
41	41	3	1	3
42	42	2	1	3
43	43	6	1	3
44	44	15	1	3
45	45	1	1	3
46	46	1	1	4
47	47	16	1	4
48	48	20	1	4
49	49	1	1	4
50	50	1	1	4
51	51	1	1	4

```
52      52     1      1     4
53      53     2      1     4
54      54     2      1     4
55      55    13      1     4
56      56     3      1     4
57      57     1      1     4
58      58     1      1     4
59      59    14      1     4
60      60     2      1     4
```

1. Ajuste um modelo Kaplan-Meier estratificado por comunidade e compare o tempo mediano de aleitamento em cada comunidade.

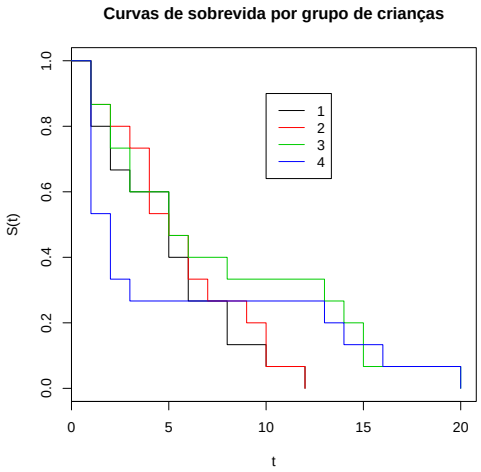
```
> y <- Surv(leite2$tempo, leite2$status)
> km <- survfit(y ~ grupo, data = leite2)
> km

Call: survfit(formula = y ~ grupo, data = leite2)

         n events median 0.95LCL 0.95UCL
grupo=1 15     15      5       2       8
grupo=2 15     15      5       4      10
grupo=3 15     15      5       3      15
grupo=4 15     15      2       1      14

> summary(km)

Call: survfit(formula = y ~ grupo, data = leite2)

                grupo=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     15       3   0.8000  0.1033       0.6212        1.000
    2     12       2   0.6667  0.1217       0.4661        0.953
    3     10       1   0.6000  0.1265       0.3969        0.907
    5      9       3   0.4000  0.1265       0.2152        0.743
    6      6       2   0.2667  0.1142       0.1152        0.617
    8      4       2   0.1333  0.0878       0.0367        0.484
   10      2       1   0.0667  0.0644       0.0100        0.443
   12      1       1   0.0000      NA           NA           NA

                grupo=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     15       2   0.8667  0.0878       0.7106        1.000
    2     13       1   0.8000  0.1033       0.6212        1.000
    3     12       1   0.7333  0.1142       0.5405        0.995
    4     11       3   0.5333  0.1288       0.3322        0.856
```

```
   5      8       1   0.4667  0.1288       0.2717        0.802
   6      7       2   0.3333  0.1217       0.1630        0.682
   7      5       1   0.2667  0.1142       0.1152        0.617
   9      4       1   0.2000  0.1033       0.0727        0.550
  10      3       2   0.0667  0.0644       0.0100        0.443
  12      1       1   0.0000      NA           NA           NA

               grupo=3
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   1     15       2   0.8667  0.0878       0.7106        1.000
   2     13       2   0.7333  0.1142       0.5405        0.995
   3     11       2   0.6000  0.1265       0.3969        0.907
   5      9       2   0.4667  0.1288       0.2717        0.802
   6      7       1   0.4000  0.1265       0.2152        0.743
   8      6       1   0.3333  0.1217       0.1630        0.682
  13      5       1   0.2667  0.1142       0.1152        0.617
  14      4       1   0.2000  0.1033       0.0727        0.550
  15      3       2   0.0667  0.0644       0.0100        0.443
  20      1       1   0.0000      NA           NA           NA

               grupo=4
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   1     15       7   0.5333  0.1288       0.3322        0.856
   2      8       3   0.3333  0.1217       0.1630        0.682
   3      5       1   0.2667  0.1142       0.1152        0.617
  13      4       1   0.2000  0.1033       0.0727        0.550
  14      3       1   0.1333  0.0878       0.0367        0.484
  16      2       1   0.0667  0.0644       0.0100        0.443
  20      1       1   0.0000      NA           NA           NA
```

O tempo mediano de amamentação é menor para a comunidade 4 (igual a 2 meses) e é igual a 5 para as outras comunidades.

2. Desenhe, no mesmo gráfico, as curvas de Kaplan-Meier, estratificadas por grupo. Como se comporta a curva de sobrevivência das outras comunidades quando comparadas à comunidade 1?

```
> plot(km, col = c(1:4), xlab = "t", ylab = "S(t)", conf.int = F)
> legend(10, 0.9, c("1", "2", "3", "4"), lty = 1, col = c(1:4))
> title("Curvas de sobrevida por grupo de crianças")
```



Resposta: No gráfico de sobrevida pode-se ver que a comunidade 4 possui uma probabilidade de amamentar menor que a comunidade 1 até os 8 meses (aproximadamente). O comportamento da comunidade 2 é semelhante ao da comunidade 1 e a comunidade 3 possui um probabilidade maior de amamentar ao longo de todo o tempo de estudo.

3. Calcule o teste log-rank e de Peto para as quatro curvas.

```
> survdiff(y ~ grupo, data = leite2)

Call:
survdiff(formula = y ~ grupo, data = leite2)

         N Observed Expected (O-E)^2/E (O-E)^2/V
grupo=1 15       15     12.4    0.5489    0.8735
grupo=2 15       15     13.4    0.1862    0.3028
grupo=3 15       15     19.7    1.1220    2.1534
grupo=4 15       15     14.5    0.0182    0.0323

 Chisq= 2.5  on 3 degrees of freedom, p= 0.478

> survdiff(y ~ grupo, data = leite2, rho = 1)

Call:
survdiff(formula = y ~ grupo, data = leite2, rho = 1)

         N Observed Expected (O-E)^2/E (O-E)^2/V
grupo=1 15     8.48     8.22   0.00865    0.0187
grupo=2 15     7.57     8.70   0.14764    0.3287
```

```
grupo=3 15     6.97     9.43   0.64499    1.5446
grupo=4 15    10.28     6.95   1.59872    3.3897

 Chisq= 4.1  on 3 degrees of freedom, p= 0.254
```

A que conclusões você chega com esta análise? Existe diferença entre as comunidades quanto ao tempo de aleitamento?

Resposta: Apesar da diferença visual, o teste do Log-rank e o Peto não rejeitam a hipótese nula de igual distribuição dos tempos de amamentação entre as comunidades.

Exercício 4.3: O banco de dados *ipec.csv* contém os dados de uma coorte de 193 pacientes com Aids, da qual fazem parte os dados apresentados na Tabela 4.1. Para este estudo foi definido, como tempo de sobrevida, o tempo entre o diagnóstico de Aids (critério CDC-1993) e o óbito.

1. Abra o banco de dados e observe as variáveis presentes. No Apêndice C1, você encontra um breve histórico da coorte e a descrição das variáveis.

```
> ipec <- read.table("ipec.csv", header = T, sep = ";")
> names(ipec)

 [1] "id"      "ini"     "fim"     "tempo"   "status"  "sexo"    "escola"
 [8] "idade"   "risco"   "acompan" "obito"   "anotrat" "tratam"  "doenca"
[15] "propcp"

> attach(ipec)

        The following object(s) are masked from ipec ( position 3 ) :

         acompan anotrat doenca escola fim id idade ini obito propcp risco sexo status tempo
```

2. Faça uma análise exploratória dos dados no R. Como é esta coorte? Qual é a idade média? Qual é a razão de homens:mulheres? Quantos receberam tratamento? Quantos foram a óbito e quantos foram censurados? Faça um histograma dos tempos censurados: eles ocorreram no fim do estudo? Ou ocorreram durante todo o estudo? (Aproveite para praticar comandos no R para análise exploratória de dados).

   Calculando a idade média:

```
> mean(idade)

[1] 36.55440
```

Número de homens e mulheres:

```
> table(sexo)

sexo
  F   M
 49 144
```

A relação homem:mulher neste grupo de pacientes é igual a $144/49 = 2.93$, aproximadamente 3 homens para cada mulher.

Número de pacientes por grupo de tratamento:

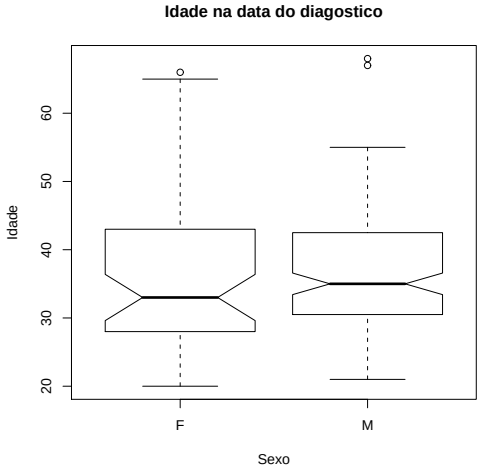
```
> table(tratam)

tratam
  0   1   2   3
 44 100  35  14
```

Número de óbitos e sobreviventes:

```
> table(status)

status
  0   1
103  90
```

Gráfico da distribuição da idade por sexo:

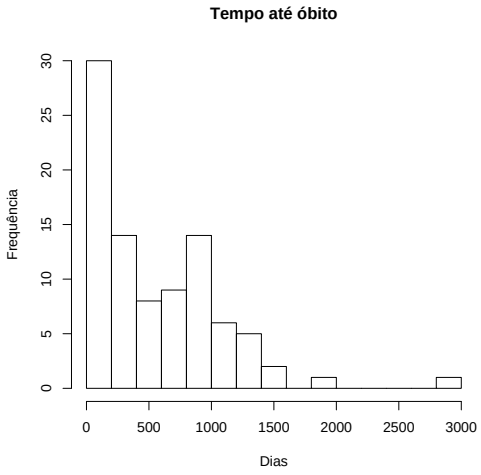
```
> boxplot(idade ~ sexo, main = "Idade na data do diagostico", notch = T,
+     ylab = "Idade", xlab = "Sexo")
```



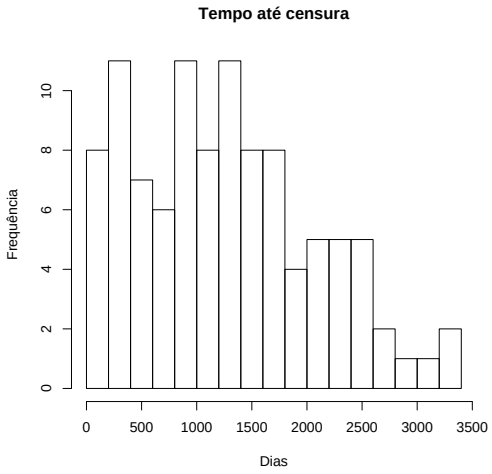
Não há diferença na idade por gênero.

Distribuição do tempo de sobrevida dos pacientes que morreram e dos que sobreviveram até o fim do estudo

```
> hist(tempo[status == 1], breaks = 12, main = "Tempo até óbito",
+     ylab = "Frequência", xlab = "Dias")
```




```
> hist(tempo[status == 0], breaks = 12, ylab = "Frequência", xlab = "Dias",
+     main = "Tempo até censura")
```



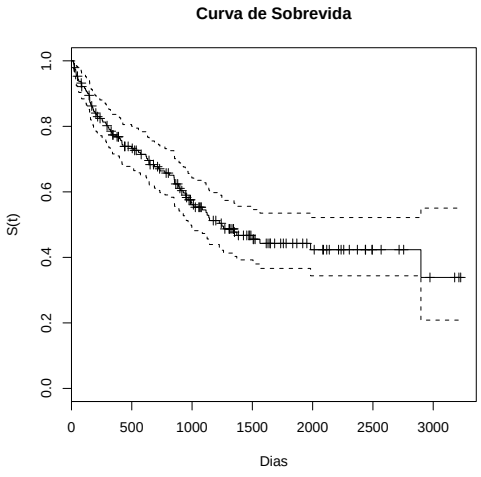
3. Refaça no R a análise não paramétrica dos dados do Ipec apresentada no texto, estratificando o tempo de sobrevida pela variável sexo. Faça os gráficos das curvas de sobrevida e risco e calcule os testes log-rank e Peto (os comandos estão todos ao longo do texto). Existe diferença entre homens e mulheres quanto ao tempo de sobrevida pós-diagnóstico de Aids?

   Lendo o banco de dados e calculando a sobrevida:

```
> require(survival)
> ipec <- read.table("ipec.csv", header = T, sep = ";")
> KM <- survfit(Surv(tempo, status) ~ 1, data = ipec)
> KM

Call: survfit(formula = Surv(tempo, status) ~ 1, data = ipec)

      n  events  median 0.95LCL 0.95UCL
    193      90    1247     992     Inf
```

   Gráfico da sobrevida de Aids dos pacientes do IPEC com intervalo de confiança. Os tempos censurados estão assinalados no gráfico por um traço vertical sobre a curva.

```
> plot(KM, ylab = "S(t)", xlab = "Dias", main = "Curva de Sobrevida")
```



   Calculando a sobrevida estratificada por sexo

```
> KMsexo <- survfit(Surv(tempo, status) ~ sexo, data = ipec)
> KMsexo
```

```
Call: survfit(formula = Surv(tempo, status) ~ sexo, data = ipec)

         n events median 0.95LCL 0.95UCL
sexo=F  49     16    Inf    1371     Inf
sexo=M 144     74   1116     887    1563
```

Não foi possível calcular o tempo mediano para as mulheres porque menos de 50% delas haviam morridos ao fim do estudo.

Gráfico da sobrevida por sexo sem intervalo de confiança

```
> plot(KMsexo, lty = 1:2, col = 1:2, ylab = "S(t)", xlab = "Dias",
+     conf.int = F)
> legend(0, 0.4, c("Fem", "Masc"), lty = 1:2, col = 1:2)
> title("Curvas de sobrevida segundo sexo")
```

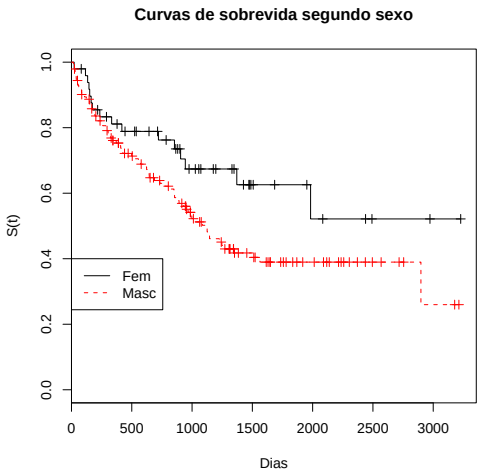
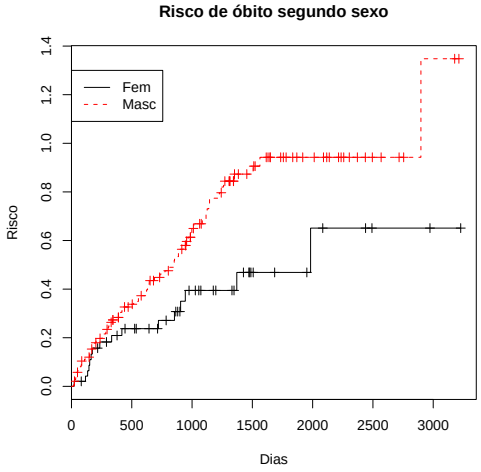


Gráfico do risco de óbito por sexo sem intervalo de confiança

```
> plot(KMsexo, lty = 1:2, fun = "cumhaz", col = 1:2, ylab = "Risco",
+     xlab = "Dias", conf.int = F)
> legend(0, 1.3, c("Fem", "Masc"), lty = 1:2, col = 1:2)
> title("Risco de óbito segundo sexo")
```



Teste log-rank e peto

```
> logrank <- survdiff(Surv(tempo, status) ~ sexo, data = ipec)
> peto <- survdiff(Surv(tempo, status) ~ sexo, data = ipec, rho = 1)
> logrank

Call:
survdiff(formula = Surv(tempo, status) ~ sexo, data = ipec)

         N Observed Expected (O-E)^2/E (O-E)^2/V
sexo=F  49       16     24.5      2.93      4.03
sexo=M 144       74     65.5      1.09      4.03

 Chisq= 4  on 1 degrees of freedom, p= 0.0447

> peto

Call:
survdiff(formula = Surv(tempo, status) ~ sexo, data = ipec, rho = 1)

         N Observed Expected (O-E)^2/E (O-E)^2/V
sexo=F  49     12.1     18.2     2.011      3.54
sexo=M 144     55.1     49.0     0.746      3.54

 Chisq= 3.5  on 1 degrees of freedom, p= 0.0598
```

Resposta: Os resultados do teste log-rank, apesar de limítrofes (p=0.047), mostram uma diferença na sobrevida pós diagnóstico de Aids entre homens e mulheres. Já o teste Peto, que dá maior peso as informações iniciais do estudo,

não rejeita a hipótese nula de igualdade entre os gêneros quanto ao tempo de sobrevida.

4. Refaça a análise acima, agora estratificando-a por tipo de tratamento. Existe diferença na sobrevivência dos pacientes submetidos aos diferentes tipos de tratamento (variável ***tratam***)?

   Calculando a sobrevida estratificada por tratamento:

```
> KMtrat <- survfit(Surv(tempo, status) ~ tratam, data = ipec)
```

   Gráfico da sobrevida por tratamento sem intervalo de confiança:

```
> plot(KMtrat, lty = 1:4, col = 2:5, ylab = "S(t)", xlab = "Dias",
+     conf.int = F)
> legend(1700, 0.3, c("sem tratamento", "monoterapia", "terapia combinada",
+     "potente"), lty = 1:4, col = 2:5, bty = "n")
> title("Curvas de sobrevida segundo tipo de tratamento")
```

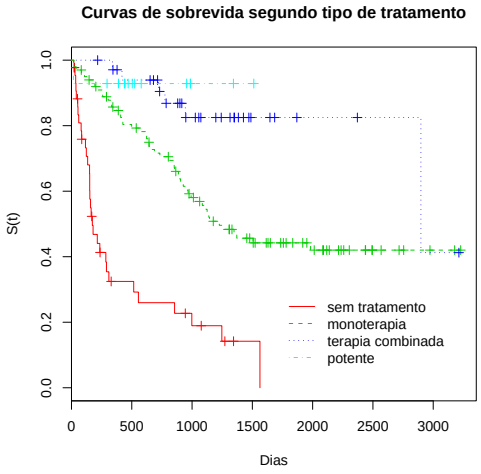
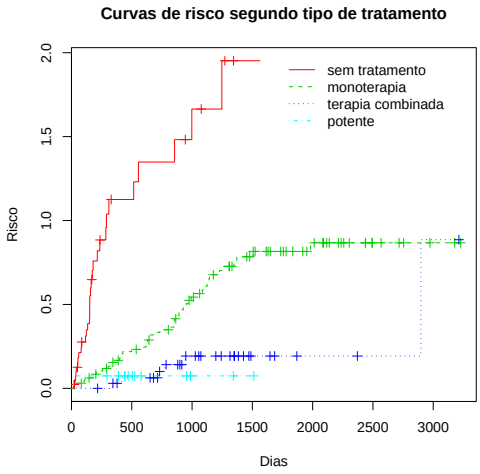


   Gráfico do risco de óbito por tratamento sem intervalo de confiança:

```
> plot(KMtrat, lty = 1:4, fun = "cumhaz", col = 2:5, ylab = "Risco",
+     xlab = "Dias", conf.int = F)
> legend(1700, 2, c("sem tratamento", "monoterapia", "terapia combinada",
+     "potente"), lty = 1:4, col = 2:5, bty = "n")
> title("Curvas de risco segundo tipo de tratamento")
```



Testes log-rank e peto:

```
> survdiff(Surv(tempo, status) ~ tratam, data = ipec)

Call:
survdiff(formula = Surv(tempo, status) ~ tratam, data = ipec)

           N Observed Expected (O-E)^2/E (O-E)^2/V
tratam=0  44       32     9.51    53.229    61.560
tratam=1 100       51    54.73     0.254     0.659
tratam=2  35        6    20.27    10.050    13.049
tratam=3  14        1     5.49     3.674     3.950

 Chisq= 69  on 3 degrees of freedom, p= 6.88e-15

> survdiff(Surv(tempo, status) ~ tratam, data = ipec, rho = 1)

Call:
survdiff(formula = Surv(tempo, status) ~ tratam, data = ipec,
    rho = 1)

           N Observed Expected (O-E)^2/E (O-E)^2/V
tratam=0  44    27.03     7.81    47.230     65.78
tratam=1 100    35.30    40.06     0.565      1.80
tratam=2  35     3.89    14.98     8.206     13.70
tratam=3  14     1.00     4.37     2.597      3.43

 Chisq= 72.2  on 3 degrees of freedom, p= 1.44e-15
```

Resposta: Ambos os testes rejeitaram a hipótese nula de igualdade no tempo de sobrevida entre os diferentes tratamentos.

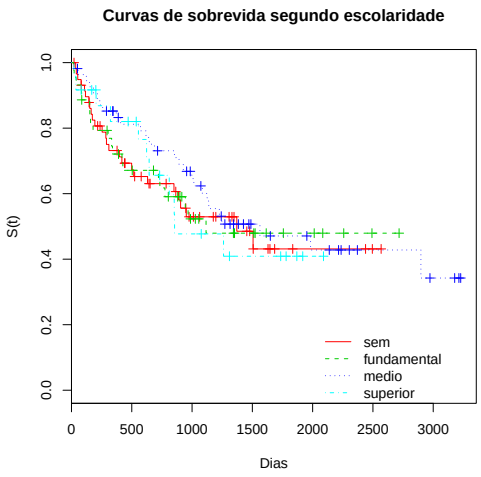
5. Estratifique os dados por nível de escolaridade. Existe diferença na sobrevivência de pacientes com diferentes graus de escolaridade?

   Calculando a sobrevida estratificada por escolaridade:

```
> KMesc <- survfit(Surv(tempo, status) ~ escola, data = ipec)
```

   Gráfico da sobrevida por escolaridade sem intervalo de confiança:

```
> plot(KMesc, lty = 1:4, col = 2:5, ylab = "S(t)", xlab = "Dias",
+     conf.int = F)
> legend(2000, 0.2, c("sem", "fundamental", "medio", "superior"),
+     lty = 1:4, col = 2:5, bty = "n")
> title("Curvas de sobrevida segundo escolaridade")
```



   Testes log-rank e peto:

```
> survdiff(Surv(tempo, status) ~ escola, data = ipec)

Call:
survdiff(formula = Surv(tempo, status) ~ escola, data = ipec)

n=182, 11 observations deleted due to missingness.

           N Observed Expected (O-E)^2/E (O-E)^2/V
escola=0 59       26     24.2    0.1348    0.1910
```

```
escola=1 44       20     19.2    0.0376    0.0490
escola=2 55       27     30.4    0.3785    0.6111
escola=3 24       11     10.3    0.0529    0.0605

 Chisq= 0.6  on 3 degrees of freedom, p= 0.891
```

```
> survdiff(Surv(tempo, status) ~ escola, data = ipec, rho = 1)

Call:
survdiff(formula = Surv(tempo, status) ~ escola, data = ipec,
    rho = 1)

n=182, 11 observations deleted due to missingness.

          N Observed Expected (O-E)^2/E (O-E)^2/V
escola=0 59    20.50    18.55    0.2048    0.3683
escola=1 44    15.96    14.56    0.1348    0.2228
escola=2 55    18.42    22.10    0.6112    1.2316
escola=3 24     8.16     7.84    0.0135    0.0195

 Chisq= 1.3  on 3 degrees of freedom, p= 0.738
```

Resposta: Não existe diferença na sobrevida de Aids para os diferentes níveis de escolaridade.

5

Estimação paramétrica

Exercícios

Exercício 5.1: Em um estudo, ajustou-se um modelo exponencial aos tempos de sobrevida observados nos grupos controle e tratamento. Os modelos encontrados foram:

$$S_c(t) = \exp(-0,07t) \quad \text{para o grupo controle}$$

$$S_{tr}(t) = \exp(-0,04t) \quad \text{para o grupo tratamento}$$

Com base nesses modelos, responda:

1. Qual foi o risco instantâneo estimado para o grupo controle? E para o grupo recebendo tratamento?

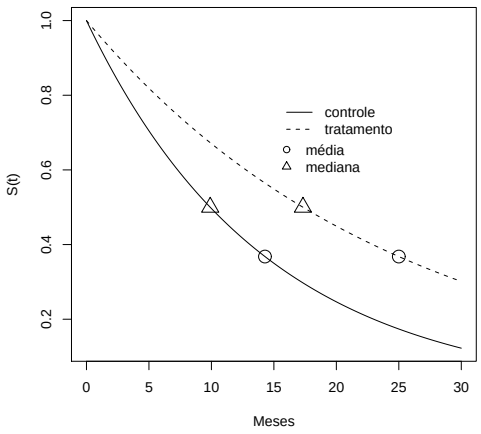
   Resposta: Segundo os resultados acima a estimativa do parâmetro da distribuição exponencial para o grupo controle, digamos $\hat{\lambda}_c$, foi de $\hat{\lambda}_c = 0,07$, e para o grupo tratamento foi de $\hat{\lambda}_{tr} = 0,04$. Assim, o risco estimado em qualquer tempo sob a distribuição exponencial para o grupo controle é 0,07 enquanto que para o grupo tratamento é 0,04.

2. Qual foi a sobrevida média e mediana no grupo controle? E no grupo recebendo tratamento?

   Resposta: O tempo médio de sobrevida é dado por $\bar{T} = \frac{1}{\alpha}$. Logo, para o grupo controle temos $\bar{T} = \frac{1}{0,07} = 14,28$ e para o grupo placebo temos $\bar{T} = \frac{1}{0,04} = 25$. Já o tempo mediano de sobrevida é dado por $T_{mediano} =$

$\frac{\ln(2)}{\alpha}$. Sendo assim, temos que o tempo mediano para o grupo controle igual a $T_{mediano} = \frac{\ln(2)}{0,07} = 9,90$ e $T_{mediano} = \frac{\ln(2)}{0,04} = 17,32$

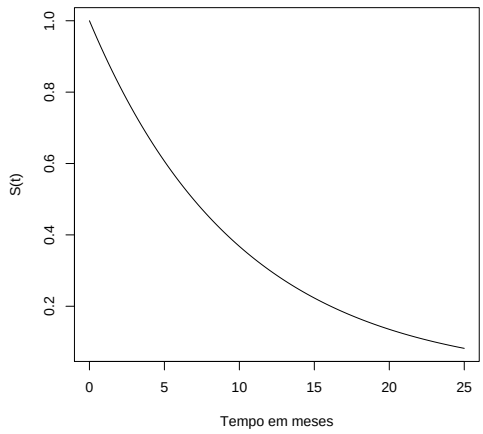
3. As duas curvas estimadas de sobrevida são apresentadas na figura que segue. Localize, nesta, o tempo mediano e médio que você calculou. Com base neste gráfico, você acha que o tratamento teve efeito na sobrevida desses pacientes?



Resposta: O gráfico sugere que o tratamento teve um efeito significativo no aumento do tempo de sobrevida dos pacientes. No entanto para que possamos tirar conclusões estatísticamente conclusivas é importante considerar tanto a variabilidade nas curvas estimadas quanto nos tempos médios e medianos de sobrevida estimados.

Exercício 5.2: No R, faça gráficos da função de sobrevida de acordo com um modelo exponencial utilizando $\alpha = 0,1$. Calcule o tempo mediano de sobrevida de acordo com este modelo. Calcule também o percentil 90 (P90) e o percentil 10 (P10), isto é, o tempo em que 90% e 10% dos pacientes, respectivamente, ainda não tinham sofrido o evento.

```
> alfa <- 0.1
> curve(exp(-alfa * x), from = 0, to = 25, ylab = "S(t)", xlab = "Tempo em meses")
```



Percentil 90

```
> p90 <- log(1/0.9)/alfa
> p90

[1] 1.053605
```

Percentil 10

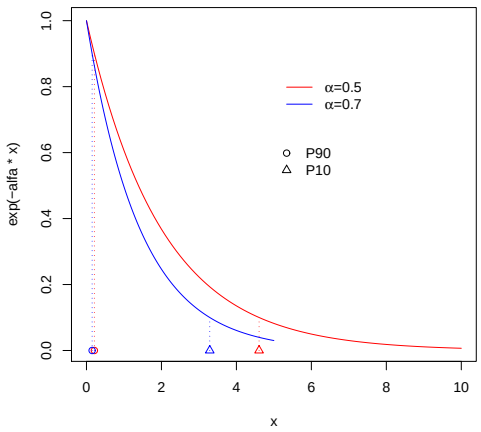
```
> p10 <- log(1/0.1)/alfa
> p10

[1] 23.02585
```

Com base nesses comando do R:

1. Troque o valor do parâmetro para $\alpha = 0,5$ e $\alpha = 0,7$.

```
> alfa <- 0.5
> curve(exp(-alfa * x), from = 0, to = 10, lty = 1, col = "red")
> p90 <- log(1/0.9)/alfa
> p90

[1] 0.2107210

> p10 <- log(1/0.1)/alfa
> p10
```

```
[1] 4.60517

> segments(p90, 0, p90, exp(-alfa * p90), lty = 3, col = "red")
> points(p90, 0, pch = 1, col = "red")
> segments(p10, 0, p10, exp(-alfa * p10), lty = 3, col = "red")
> points(p10, 0, pch = 2, col = "red")
> alfa <- 0.7
> curve(exp(-alfa * x), from = 0, to = 5, add = T, lty = 1, col = "blue")
> p90 <- log(1/0.9)/alfa
> p90

[1] 0.1505150

> p10 <- log(1/0.1)/alfa
> p10

[1] 3.289407

> segments(p90, 0, p90, exp(-alfa * p90), lty = 3, col = "blue")
> points(p90, 0, pch = 1, col = "blue")
> segments(p10, 0, p10, exp(-alfa * p10), lty = 3, col = "blue")
> points(p10, 0, pch = 2, col = "blue")
> legend(5, 0.85, c(expression(paste(alpha, "=0.5")), expression(paste(alpha,
+     "=0.7"))), lty = 1, col = c("red", "blue"), bty = "n")
> legend(5, 0.65, c("P90", "P10"), pch = 1:2, bty = "n")
```



2. Observe o comportamento da função.

Resposta: Note que a função de sobrevivência cai mais rapidamente a medida que aumenta o valor do parâmetro α, e portanto também decrescem os percentis 10% e 90%. Este comportamento é esperado já que o risco instantâneo de falha em qualquer tempo, sob o modelo exponencial, aumenta com o aumento de α.

Exercício 5.3: Com relação ao modelo paramétrico Weibull, responda:

1. Por que o modelo Weibull é considerado mais flexível do que o modelo exponencial?

   Resposta: Porque possui um parâmetro adicional que permite ajustar diferentes formas para a função risco, daí o nome parâmetro de forma.

2. Em que situação particular o modelo Weibull é equivalente ao exponencial?

   Resposta: Na situação em que o parâmetro de forma $\gamma = 1$.

3. Qual a relação entre o parâmetro γ e o comportamento da função de risco?

   Resposta: Quando $\gamma = 1$ a função de risco é constante, ou seja, o risco intantâneo de ocorrência do evento não varia com o passar do tempo; quando $\gamma > 1$ o risco cresce no tempo; e $\gamma < 1$ o risco decresce no tempo.

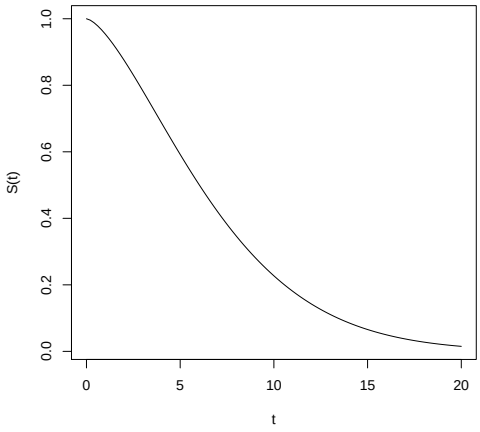
4. Quais das curvas de risco apresentadas na Figura 3.3 não poderiam ser modeladas pela função Weibull, nem mesmo aproximadamente?

   Resposta: As curvas D, E e F não poderiam ser modeladas pela função Weibull pois o comportamento da função risco ao longo de tempo deve ser monotônico: somente crescente ou somente descrescente. O que se vê nos quadros D, E e F são misturas destes comportamentos.

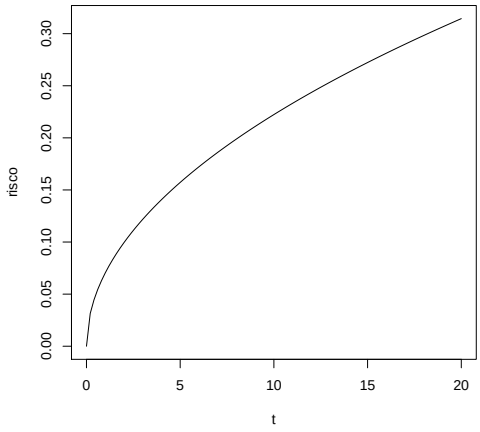
Exercício 5.4: Seja T o tempo de sobrevida até a ocorrência de um evento, que segue uma distribuição Weibull com parâmetros $\gamma = 1,5$ e $\alpha = 0,13$.

1. Escreva as funções $S(t)$, $\lambda(t)$ e $\Lambda(t)$ e use o R para fazer o respectivos gráficos.

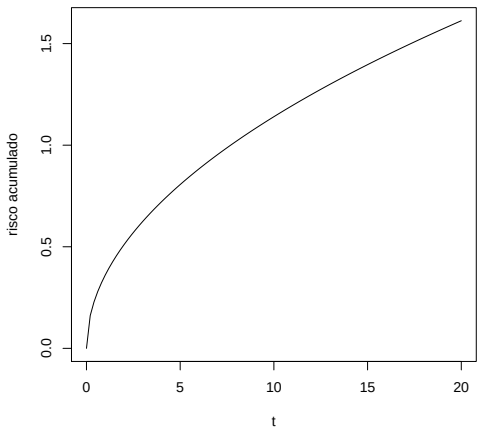
```
> alfa <- 0.13
> gama <- 1.5
> curve(exp(-(alfa * x)^gama), from = 0, to = 20, ylab = "S(t)",
+     xlab = "t")
```




```
> curve(alfa * gama * (alfa * x)^(gama - 1), from = 0, to = 20,
+     ylab = "risco", xlab = "t")
```




```
> curve((alfa * x)^(gama - 1), from = 0, to = 20, ylab = "risco acumulado",
+     xlab = "t")
```



2. Calcule o tempo mediano de sobrevida. Calcule o percentil 80 e o percentil 10 dessa distribuição.

   Tempo mediano ($S(t) = 0.5$)

```
> tmediano <- log(1/0.5)^(1/gama)/alfa
> tmediano

[1] 6.024767
```

   Percentil 80 ($S(t) = 0.80$)

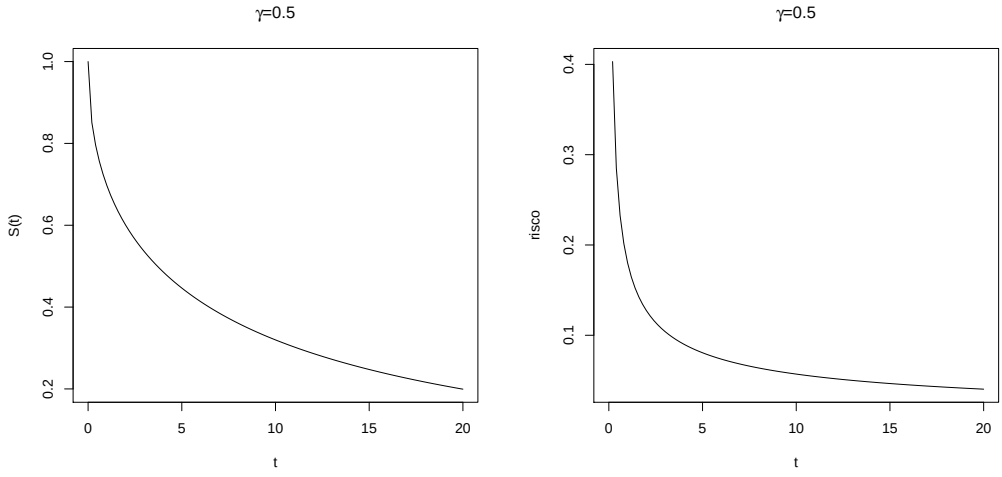
```
> p80 <- log(1/0.8)^(1/gama)/alfa
> p80

[1] 2.829955
```

   Percentil 10 ($S(t) = 0.10$)

```
> p10 <- log(1/0.1)^(1/gama)/alfa
> p10

[1] 13.41324
```

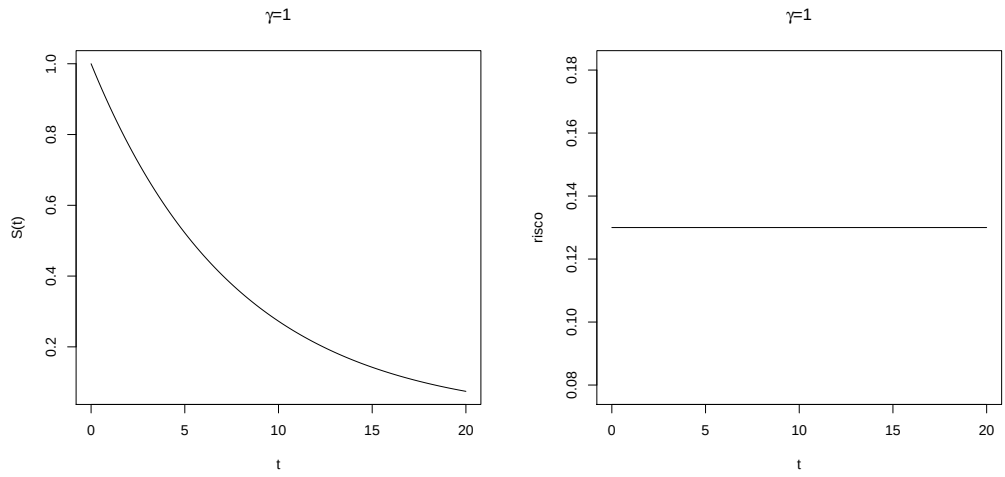
3. Fixe o valor do parâmetro $\alpha = 0,13$ e faça gráficos da função de risco e da função de sobrevida para diversos valores de γ: $0 < \gamma < 1$, $\gamma = 1$ e $\gamma > 1$. Visualize como o parâmetro γ afeta o comportamento do risco e da sobrevida.

```
> par(mfrow = c(3, 2))
> alfa <- 0.13
> gama <- 0.5
> curve(exp(-(alfa * x)^gama), from = 0, to = 20, ylab = "S(t)",
+     xlab = "t", main = expression(paste(gamma, "=0.5")))
> curve(alfa * gama * (alfa * x)^(gama - 1), from = 0, to = 20,
+     ylab = "risco", xlab = "t", main = expression(paste(gamma,
+         "=0.5")))
```



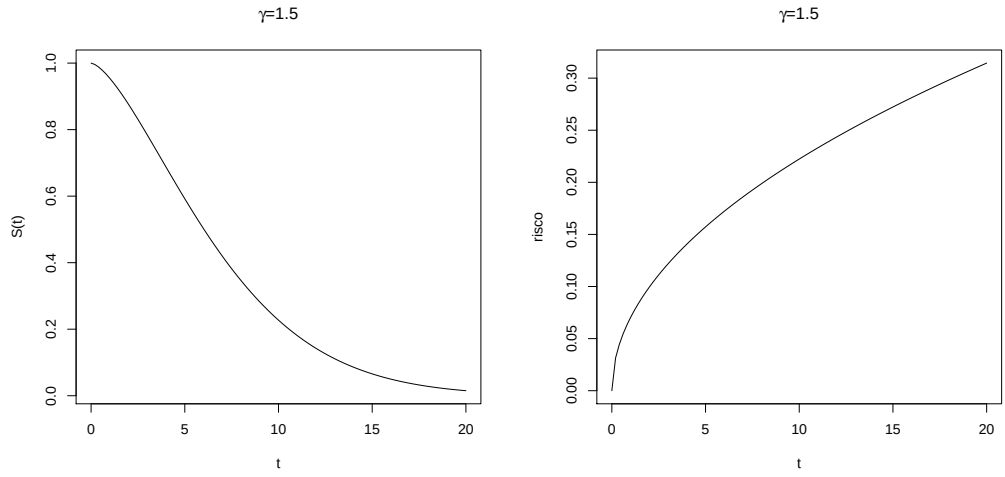
Sobrevida cai rapidamente no início do acompanhamento, e risco é decrescente.

```
> gama <- 1
> curve(exp(-(alfa * x)^gama), from = 0, to = 20, ylab = "S(t)",
+     xlab = "t", main = expression(paste(gamma, "=1")))
> curve(alfa * gama * (alfa * x)^(gama - 1), from = 0, to = 20,
+     ylab = "risco", xlab = "t", main = expression(paste(gamma,
+         "=1")))
```



Sobrevida cai mais suavemente, risco constante.

```
> gama <- 1.5
> curve(exp(-(alfa * x)^gama), from = 0, to = 20, ylab = "S(t)",
+     xlab = "t", main = expression(paste(gamma, "=1.5")))
> curve(alfa * gama * (alfa * x)^(gama - 1), from = 0, to = 20,
+     ylab = "risco", xlab = "t", main = expression(paste(gamma,
+         "=1.5")))
```



Sobrevida cai suavemente no início do período, risco aumenta com o tempo decorrido.

Exercício 5.5: Em um estudo sobre o tempo de incubação de uma infecção, verificou-se que T é adequadamente descrito por uma função Weibull com parâmetros $\gamma = 1,2$ e $\alpha = 0,07$.

1. Calcule o tempo mediano de incubação desta infecção.

```
> alfa <- 0.07
> gama <- 1.2
> tmediano <- log(1/0.5)^(1/gama)/alfa
> tmediano

[1] 10.52583
```

   Resposta: O tempo mediano de incubação é de aproximadamente 10 horas e meia, em outras palavras, segundo este modelo espera-se que 50% das pessoas infectadas comecem a apresentar sintomas depois de 10 horas e meia do contato com o agente infeccioso.

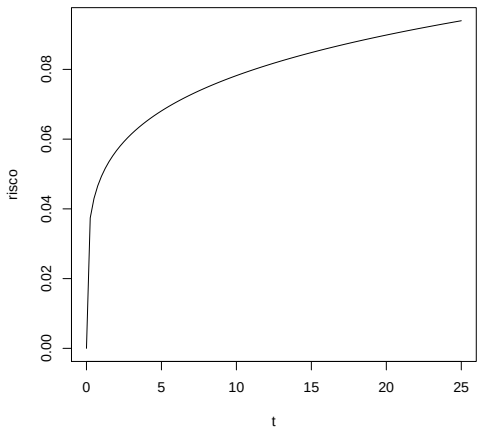
2. É correto dizer que em 10 horas do momento da infecção, espera-se que 80% das pessoas já tenham desenvolvido sintomas?

```
> t <- 10
> S10 <- exp(-(alfa * t)^gama)
> S10

[1] 0.5211044
```

   Resposta: Não, em 10 horas, espera-se que aproximadamente 52% das pessoas não tenham desenvolvido sintomas, ou alternativamente, espera-se que 48% tenham desenvolvido sintomas.

3. O risco de surgimento de sintomas é crescente ou decrescente ao longo do tempo?

```
> par(mfrow = c(1, 1))
> curve(gama * alfa^gama * x^(gama - 1), from = 0, to = 25, ylab = "risco",
+     xlab = "t")
```



Resposta: Na verdade não é nem preciso traçar o gráfico da função risco; basta observar que o parâmetro de forma α é, neste caso, maior do que 1, ou seja, risco crescente.

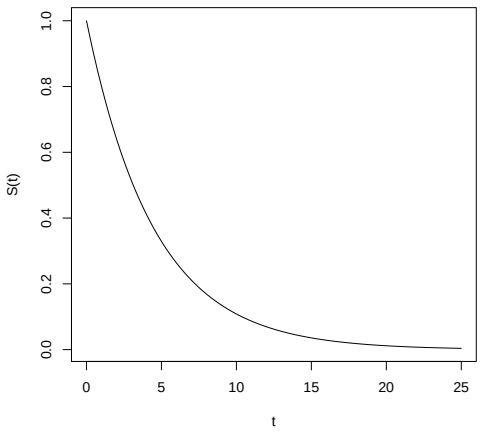
Exercício 5.6: Mil crianças não vacinadas são acompanhadas, a partir do nascimento, em um estudo cujo objetivo é identificar a idade em que adquirem hepatite A. Os resultados do estudo indicam que a idade média de soroconversão das crianças foi de 4,5 anos e que o risco de contrair hepatite A foi constante e independente da idade.

1. Proponha um modelo paramétrico para o tempo até a aquisição de hepatite A.

   Resposta: Neste caso, como o risco de contrair hepatite A é constante no tempo, um modelo simples (i.e, parcimonioso) e adequado (pois possui função risco constante) seria o modelo paramétrico exponencial.

2. Faça no R o gráfico da função de sobrevida, de acordo com esse modelo.

```
> tm <- 4.5
> alfa <- 1/tm
> curve(exp(-alfa * x), from = 0, to = 25, ylab = "S(t)", xlab = "t")
```



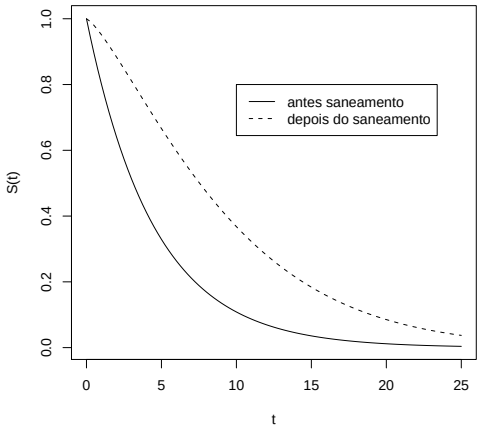
3. Com base nesse modelo, em que idade espera-se ter 90% das crianças soropositivas?

```
> p10 <- log(1/0.1)/alfa
> p10

[1] 10.36163
```

Resposta: Segundo este modelo espera-se que aos 10 anos e 4 meses 90% das crianças sejam soropositivas, ou alternativamente que nesta idade apenas 10% ainda não sejam soropositivas.

4. Após este estudo, um projeto de saneamento é implementado nesta comunidade. Para avaliar o efeito do saneamento na transmissão de hepatite A, uma nova coorte é montada, semelhante à anterior. Ao analisar os dados dessa nova coorte, encontramos que um modelo Weibull com parâmetros $\gamma = 1,3$ e $\alpha = 0,1$ descreve bem a curva de sobrevida. Com base nessa informação, avalie qual foi o efeito do saneamento no risco de contrair hepatite A nessa comunidade. Sugestão: compare os gráficos das funções de sobrevida.

```
> tm <- 4.5
> alfa <- 1/tm
> curve(exp(-alfa * x), from = 0, to = 25, ylab = "S(t)", xlab = "t")
> alfa <- 0.1
> gama <- 1.3
> curve(exp(-(alfa * x)^gama), from = 0, to = 25, add = T, lty = 2)
> legend(10, 0.8, c("antes saneamento", "depois do saneamento"),
+     lty = 1:2)
```



Percentil 10

```
> p10 <- log(1/0.1)^(1/gama)/alfa
> p10

[1] 18.99448
```

Note que a sobrevida aumentou consideravelmente após implantação do projeto de saneamento. Por exemplo, segundo o modelo pós-saneamento espera-se que somente aos 19 anos 10% não sejam soropositivas.

Exercício 5.7: Retorne ao exemplo do Exercício 4.1, sobre tempo de aleitamento de crianças (arquivo: *leite.txt*).

1. Ajuste uma distribuição Weibull ao tempo de aleitamento. Existe evidência de que o modelo Weibull seja mais adequado do que o exponencial?

```
> require(survival)
> leite <- read.table("leite.txt", header = T, sep = "")
> modeloweib <- survreg(Surv(tempo, status) ~ 1, data = leite,
+     dist = "weib")
> summary(modeloweib)

Call:
survreg(formula = Surv(tempo, status) ~ 1, data = leite, dist = "weib")
             Value Std. Error     z        p
(Intercept)  1.713      0.180  9.54 1.38e-21
Log(scale)  -0.415      0.209 -1.99 4.70e-02
```

```
Scale= 0.66

Weibull distribution
Loglik(model)= -37.5   Loglik(intercept only)= -37.5
Number of Newton-Raphson Iterations: 6
n= 15
```

Considerando o parâmetro de escala (`Scale = 0.66`), e que $\gamma = 1/Scale$, então $\gamma = 1.515$, ligeiramente maior do que um, ou seja, risco crescente. O parâmetro de escala é marginalmente significativo: $p = 0,047$. Usando o modelo exponencial assumiríamos que o risco é constante. Ajustando então o modelo exponencial:

```
> modeloexp <- survreg(Surv(tempo, status) ~ 1, data = leite, dist = "exp")
> summary(modeloexp)

Call:
survreg(formula = Surv(tempo, status) ~ 1, data = leite, dist = "exp")
            Value Std. Error    z        p
(Intercept)  1.61      0.258 6.23 4.57e-10

Scale fixed at 1

Exponential distribution
Loglik(model)= -39.1   Loglik(intercept only)= -39.1
Number of Newton-Raphson Iterations: 4
n= 15
```

Existem técnicas para comparar os dois modelos. Baseiam-se na razão de verossimilhança dos dois modelos, que diferem por um grau de liberdade. O modelo Weibull é de fato melhor.

2. Qual o tempo mediano de amamentação, estimado por esse modelo? (Dica: não se esqueça de que a parametrização das distribuições no R difere da vista no texto). Os parâmetros da distribuição Weibul são $alpha = \exp(-intercept)$ e $\gamma = 1/Scale$.

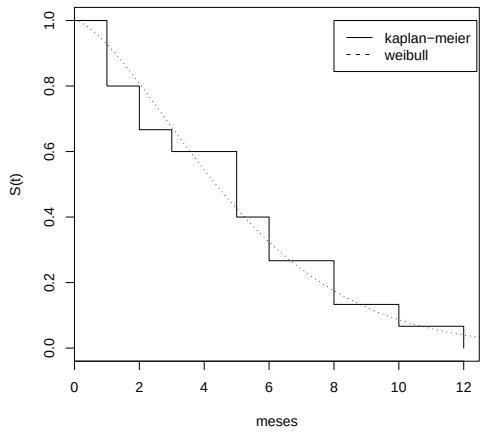
```
> alfa <- as.vector(exp(-modeloweib$coef[1]))
> alfa

[1] 0.1802502

> gama <- 1/modeloweib$scale
> gama
```

```
[1] 1.514787

> tmediano <- log(1/0.5)^(1/gama)/alfa
> tmediano

[1] 4.355558
```

Resposta: O tempo mediano de amamentação estimado por este modelo é de 4.36 meses.

3. Faça um gráfico da curva de sobrevida ajustada pelo modelo Weibull, junto com o gráfico de Kaplan-Meier. O modelo paramétrico representa bem os dados?

```
> km <- survfit(Surv(tempo, status) ~ 1, data = leite)
> plot(km, ylab = "S(t)", xlab = "meses", conf.int = F)
> alfa <- exp(-1.713)
> gama <- 1/0.66
> curve(exp(-(alfa * x)^gama), from = 0, to = 15, lty = 3, add = T)
> legend(8, 1, c("kaplan-meier", "weibull"), lty = c(1:2))
```



Resposta: O modelo parece se ajustar muito bem aos dados.

6

Modelos de regressão paramétricos

Exercícios

Exercício 6.1: O banco de dados *leite2.txt* contém dados de tempo de aleitamento de crianças de 4 comunidades. No ajuste não-paramétrico a esses dados, observamos que pertencer a uma comunidade não teve efeito no período de aleitamento. Confirme este achado, ajustando um modelo paramétrico a esses dados. Tente o modelo exponencial e o Weibull.

```
> require(survival)
```

```
[1] TRUE
```

```
> leite2 <- read.table("leite2.txt", header = T, sep = "")
> y <- Surv(leite2$tempo, leite2$status)
> modeloE1 <- survreg(y ~ factor(grupo), data = leite2, dist = "exponential")
> modeloW1 <- survreg(y ~ factor(grupo), data = leite2, dist = "weib")
> summary(modeloE1)
```

```
Call:
survreg(formula = y ~ factor(grupo), data = leite2, dist = "exponential")
               Value Std. Error     z        p
(Intercept)    1.609      0.258 6.233 4.57e-10
factor(grupo)2 0.113      0.365 0.310 7.56e-01
factor(grupo)3 0.410      0.365 1.123 2.62e-01
factor(grupo)4 0.052      0.365 0.142 8.87e-01

Scale fixed at 1
```

```
Exponential distribution
Loglik(model)= -165.2   Loglik(intercept only)= -166
        Chisq= 1.58 on 3 degrees of freedom, p= 0.66
Number of Newton-Raphson Iterations: 5
n= 60


> summary(modeloW1)


Call:
survreg(formula = y ~ factor(grupo), data = leite2, dist = "weib")
                Value Std. Error      z        p
(Intercept)     1.651      0.218  7.568 3.78e-14
factor(grupo)2  0.105      0.306  0.344 7.31e-01
factor(grupo)3  0.426      0.307  1.391 1.64e-01
factor(grupo)4  0.126      0.310  0.408 6.83e-01
Log(scale)     -0.175      0.103 -1.703 8.86e-02

Scale= 0.84

Weibull distribution
Loglik(model)= -163.8   Loglik(intercept only)= -164.9
        Chisq= 2.19 on 3 degrees of freedom, p= 0.53
Number of Newton-Raphson Iterations: 7
n= 60
```

Resposta: Note que realmente tanto sob o modelo exponencial quanto o Weibull a variável comunidade (aqui chamada 'grupo') não foi significativa. A estimativa do parâmetro de escala é marginalmente significativo (p-valor=0,089), ou seja, existe alguma indicação de que entre o modelo exponencial e o modelo Weibull o segundo pode ser mais adequado para estes dados.

Exercício 6.2: Um estudo foi realizado para estimar o efeito do transplante de medula óssea na sobrevida de pacientes com leucemia. As covariáveis analisadas foram: idade, fase da doença, ter ou não desenvolvido doença do enxerto crônica e ter ou não desenvolvido doença do enxerto aguda (para mais detalhes acerca desse estudo, refira-se ao Apêndice C.5). Ao se ajustar um modelo exponencial aos dados, obteve-se a seguinte saída do R:

```
                Value Std. Error      z        p
(Intercept)    7.13536     0.4992 14.293 2.44e-46
idade         -0.00179     0.0146 -0.122 9.03e-01
fase interm   -0.79363     0.3651 -2.174 2.97e-02
fase avançada -1.29759     0.4995 -2.598 9.39e-03
```

```
doençacronica  0.92521     0.3335  2.775 5.53e-03
doençaaguda   -1.43654     0.3158 -4.549 5.40e-06

Scale fixed at 1

Exponential distribution
Loglik(model)= -348.3   Loglik(intercept only)= -374.2
        Chisq= 51.96 on 5 degrees of freedom, p= 5.5e-10
Number of Newton-Raphson Iterations: 5
```

Observe a saída do R e responda:

1. O modelo com covariáveis é melhor do que o modelo nulo (sem covariáveis)?

   Resposta: Não. Note que a log-verossimilhança do modelo com covariáveis é muito maior do que a do modelo nulo e o teste da Deviance entre o modelo com covariáveis e o modelo nulo resultou um p-valor praticamente nulo (p=5.5e-10). Portanto temos evidências altamente significativas contra o modelo nulo.

2. Que covariáveis estão associadas com a melhoria da sobrevida? Quais estão associadas com redução da sobrevida?

   Resposta: As covariáveis associadas com a redução da sobrevida são aquelas em que o efeito estimado é negativo: fase intermediária, fase avançada e doença aguda. A covariável idade tem efeito estimado negativo, porém este efeito é não significativo. A única covariável associada a um prognóstico favorável da sobrevida é doença crônica. Cabe observar que como o próprio nome indica, doença crônica necessariamente tem que evoluir durante um tempo razoável, ou seja, o paciente tem que sobreviver por um tempo razoável para apresentá-la.

3. Escreva a função de risco, $\lambda(t)$, estimada para esta coorte.

   Resposta:

$$\begin{aligned}\lambda(t) &= \exp(-(7.13536 - 0.00179 \times \text{idade} - 0.79363 \times \text{fase interm} \\ &- 1.29759 \times \text{fase avançada} + 0.92521 \times \text{doençacronica} \\ &- 1.43654 \times \text{doençaaguda}))\end{aligned}$$

4. Qual seria o risco de óbito de um paciente de 30 anos, em fase intermediária, com doença crônica?

   Resposta:

```
> lambdac <- exp(-(7.13536 - 0.00179 * 30 - 0.79363 + 0.92521))
> lambdac

[1] 0.0007367662
```

5. Qual seria o risco de óbito de um paciente de 30 anos, em fase intermediária, com doença aguda?

   Resposta:

```
> lambdaa <- exp(-(7.13536 - 0.00179 * 30 - 0.79363 - 1.43654))
> lambdaa

[1] 0.007816722
```

   O risco de óbito é 10,6 vezes maior para o paciente com doença aguda.

6. Um segundo modelo, mais simples, foi ajustado aos dados, contendo apenas a covariável fase. O logaritmo da função de verossimilhança deste modelo simples foi de -363.6. Compare este modelo com o mais completo acima e indique se o completo resultou em melhor ajuste.

   Resposta: *l1* e *l2* são as logverossimilhanças dos modelos completo e reduzido, respectivamente.

```
> l1 <- -348.3
> l2 <- -363.6
```

   Calculando a deviance *dev* e os graus de liberdade que é a diferença entre o número de parâmetros do modelo

```
> dev <- 2 * (l1 - l2)
> dev

[1] 30.6

> gl <- 6 - 4
```

   E, por último, calcula-se o p-valor da distribuição χ^2 sob a hipótese nula de que o modelo reduzido é melhor

```
> pvalor <- 1 - pchisq(dev, gl)
> pvalor

[1] 2.26618e-07
```

Rejeitamos o modelo mais simples com p-valor=0,00000023, ou seja, a redução no valor da verossimilhança dada pelo modelo mais completo foi significativa. Em outras palavras, doença crônica ou aguda é um fator prognóstico importante para o tempo de sobrevida.

Exercício 6.3: Em Aids, a terapia anti-retroviral evoluiu da monoterapia para a terapia combinada (2 componentes) e, por fim, para a terapia de alta potência (3 componentes). Espera-se que quanto mais componentes tiver mais efetiva seja a terapia em aumentar a sobrevida. Teste esta hipótese, ajustando um modelo exponencial aos dados da coorte de Aids (*ipec.csv*).

1. Ajuste um modelo com a variável tratamento apenas. O modelo com a variável tratamento é melhor do que o modelo sem covariáveis? Interprete o efeito dos tratamentos na sobrevida. A variável tratamento deve ser modelada como um fator, e não como numérica, lembrando que os efeitos dos tratamentos estão sendo estimados em relação à ausência de tratamento.

```
> ipec <- read.table("ipec.csv", header = T, sep = ";")
> mod.ipec <- survreg(Surv(tempo, status) ~ factor(tratam), data = ipec,
+     dist = "exp")
> summary(mod.ipec)

Call:
survreg(formula = Surv(tempo, status) ~ factor(tratam), data = ipec,
    dist = "exp")
                Value Std. Error     z         p
(Intercept)      6.14      0.177 34.73 2.91e-264
factor(tratam)1  1.59      0.226  7.07  1.58e-12
factor(tratam)2  2.68      0.445  6.01  1.80e-09
factor(tratam)3  3.01      1.016  2.97  3.00e-03

Scale fixed at 1

Exponential distribution
Loglik(model)= -742.9   Loglik(intercept only)= -774.6
        Chisq= 63.49 on 3 degrees of freedom, p= 1.1e-13
Number of Newton-Raphson Iterations: 6
n= 193
```

Resposta: Rejeitamos a hipótese nula de que o modelo nulo é melhor através da estatística de deviance igual a 63,49 que segue uma distribuição χ^2 com 3 graus de liberdade e p-valor menor que 0,001 (*p= 1.1e-13*). Conclusão, o modelo com a covariável tratamento é melhor.

Para calcular o risco temos que substituir os valores das variáveis *dummies* na expressão do risco do modelo exponencial que é $\lambda(t|\boldsymbol{x}) = \alpha(\boldsymbol{x}) = \exp(\boldsymbol{x}\boldsymbol{\beta})$. Como o R parametriza as distribuições de forma diferente, temos que **trocar o sinal dos coeficientes** para interpretarmos de acordo com o texto do capítulo.

Calculando o risco de um paciente sem nenhum tratamento:

```
> trat1 <- 0
> trat2 <- 0
> trat3 <- 0
> lambda0 <- exp(-(mod.ipec$coef[1] + mod.ipec$coef[2] * trat1 +
+     mod.ipec$coef[3] * trat2 + mod.ipec$coef[4] * trat3))
> lambda0

(Intercept)
0.002156625
```

Calculando o risco de um paciente com monoterapia (`tratam = 1`):

```
> trat1 <- 1
> trat2 <- 0
> trat3 <- 0
> lambda1 <- exp(-(mod.ipec$coef[1] + mod.ipec$coef[2] * trat1 +
+     mod.ipec$coef[3] * trat2 + mod.ipec$coef[4] * trat3))
> lambda1

 (Intercept)
0.0004381632
```

Calculando o risco de um paciente com terapia combinada (`tratam = 2`):

```
> trat1 <- 0
> trat2 <- 1
> trat3 <- 0
> lambda2 <- exp(-(mod.ipec$coef[1] + mod.ipec$coef[2] * trat1 +
+     mod.ipec$coef[3] * trat2 + mod.ipec$coef[4] * trat3))
> lambda2

 (Intercept)
0.0001485001
```

Calculando o risco de um paciente com terapia potente (`tratam = 3`):

```
> trat1 <- 0
> trat2 <- 0
> trat3 <- 1
> lambda3 <- exp(-(mod.ipec$coef[1] + mod.ipec$coef[2] * trat1 +
```

```
+     mod.ipec$coef[3] * trat2 + mod.ipec$coef[4] * trat3))
> lambda3

 (Intercept)
0.0001058985
```

Calculando os riscos relativos em relação ao paciente sem nenhum tratamento:

```
> lambda0/lambda1

(Intercept)
   4.921968

> lambda0/lambda2

(Intercept)
   14.52271

> lambda0/lambda3

(Intercept)
   20.36501
```

Todos os tratamentos são altamente significativos no aumento da sobrevida, mas a terapia potente aumenta mais a sobrevida do que a terapia combinada e esta tem um melhor efeito do que a monoterapia. Em termos do risco de óbito, a razão dos riscos de pacientes sem tratamento e com a terapia potente é aproximadamente 20,4, um valor extremamente alto.

2. Faça uma análise gráfica do ajuste do modelo, comparando-o com a curva de Kaplan-Meier estratificada por tratamento. O que você tem a dizer sobre a adequação do modelo exponencial?

```
> km <- survfit(Surv(tempo, status) ~ factor(tratam), data = ipec)
> plot(km, ylab = "S(t)", xlab = "dias", conf.int = F, col = 1:4,
+     mark.time = F)
> title("Tratamento em Aids")
```

Basta adicionar as curvas de sobrevida de acordo com o modelo exponencial.

a. Pacientes sem tratamento

```
> alpha0 <- exp(-6.14)
> sobre0 <- function(x) {
+     exp(-alpha0 * x)
+ }
> curve(sobre0, from = 0, to = 3500, lty = 2, add = T, col = 1)
```

b. Paciente em monoterapia

```
> alpha1 <- exp(-6.14 - 1.59)
> sobre1 <- function(x) {
+     exp(-alpha1 * x)
+ }
> curve(sobre1, from = 0, to = 3500, lty = 2, add = T, col = 2)
```

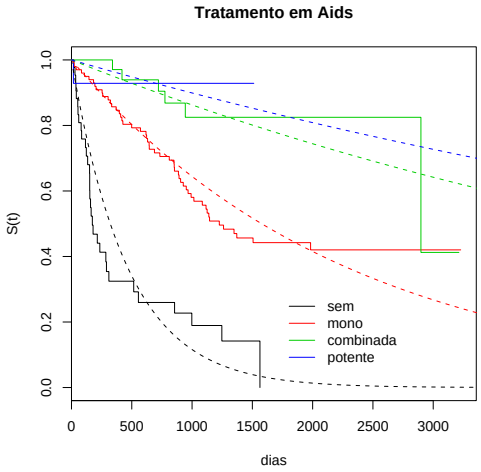
c. Paciente em terapia combinada

```
> alpha2 <- exp(-6.14 - 2.68)
> sobre2 <- function(x) {
+     exp(-alpha2 * x)
+ }
> curve(sobre2, from = 0, to = 3500, lty = 2, add = T, col = 3)
```

d. Paciente em terapia potente

```
> alpha3 <- exp(-6.14 - 3.01)
> sobre3 <- function(x) {
+     exp(-alpha3 * x)
+ }
> curve(sobre3, from = 0, to = 3500, lty = 2, add = T, col = 4)
> legend(1700, 0.3, c("sem", "mono", "combinada", "potente"), bty = "n",
+     col = 1:4, lty = 1)


> km <- survfit(Surv(tempo, status) ~ factor(tratam), data = ipec)
> plot(km, ylab = "S(t)", xlab = "dias", conf.int = F, col = 1:4,
+     mark.time = F)
> title("Tratamento em Aids")
> alpha0 <- exp(-6.14)
> sobre0 <- function(x) {
+     exp(-alpha0 * x)
+ }
> curve(sobre0, from = 0, to = 3500, lty = 2, add = T, col = 1)
> alpha1 <- exp(-6.14 - 1.59)
> sobre1 <- function(x) {
+     exp(-alpha1 * x)
+ }
> curve(sobre1, from = 0, to = 3500, lty = 2, add = T, col = 2)
> alpha2 <- exp(-6.14 - 2.68)
> sobre2 <- function(x) {
+     exp(-alpha2 * x)
+ }
> curve(sobre2, from = 0, to = 3500, lty = 2, add = T, col = 3)
> alpha3 <- exp(-6.14 - 3.01)
> sobre3 <- function(x) {
```

```
+     exp(-alpha3 * x)
+ }
> curve(sobre3, from = 0, to = 3500, lty = 2, add = T, col = 4)
> legend(1700, 0.3, c("sem", "mono", "combinada", "potente"), bty = "n",
+     col = 1:4, lty = 1)
```



O modelo exponencial se ajusta razoavelmente bem para os grupos em que foram observados mais óbitos.

3. Ajuste um outro modelo exponencial, adicionando variáveis de controle (sexo, idade e tipo de atendimento). Quais variáveis tiveram efeito significativo? Quais tiveram efeito protetor?

```
> mod2.ipec <- survreg(Surv(tempo, status) ~ factor(tratam) + sexo +
+     idade + factor(acompan), data = ipec, dist = "exp")
> summary(mod2.ipec)

Call:
survreg(formula = Surv(tempo, status) ~ factor(tratam) + sexo +
    idade + factor(acompan), data = ipec, dist = "exp")
                   Value Std. Error      z        p
(Intercept)       7.95467     0.6554 12.137 6.69e-34
factor(tratam)1   1.38695     0.2972  4.667 3.06e-06
factor(tratam)2   2.21397     0.4656  4.755 1.99e-06
factor(tratam)3   2.98559     1.0165  2.937 3.31e-03
sexoM            -0.07670     0.2833 -0.271 7.87e-01
idade            -0.00292     0.0120 -0.242 8.09e-01
factor(acompan)1 -1.70869     0.4064 -4.205 2.61e-05
factor(acompan)2 -2.23186     0.4664 -4.785 1.71e-06
```

```
Scale fixed at 1

Exponential distribution
Loglik(model)= -723.5   Loglik(intercept only)= -774.6
        Chisq= 102.25 on 7 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 6
n= 193
```

Resposta: Os fatores sexo e idade são não significativos. Dentre os fatores significativos somente tratamento teve um efeito protetor. O tipo de atendimento – internação hospitalar posterior ou imediata, comparadas a tratamento apenas ambulatorial – é significativo, e indica maior risco para pacientes que necessitam internação.

Exercício 6.4: Qual é o efeito da doença de base na sobrevida de pacientes em diálise, quando controlamos por idade? Usando a distribuição Weibull, ajuste o modelo:

```
> dialise <- read.table("dialise.csv", header = T, sep = ",")
> modelo1 <- survreg(Surv(tempo, status) ~ idade + cdiab + crim +
+     congenita, data = dialise)
> summary(modelo1)


Call:
survreg(formula = Surv(tempo, status) ~ idade + cdiab + crim +
    congenita, data = dialise)
              Value Std. Error       z        p
(Intercept)  6.7737    0.14999  45.161 0.00e+00
idade       -0.0428    0.00225 -19.017 1.24e-80
cdiab       -0.3605    0.07353  -4.903 9.44e-07
crim        -0.0384    0.08139  -0.472 6.37e-01
congenita    0.8855    0.27529   3.217 1.30e-03
Log(scale)   0.1951    0.02082   9.373 7.04e-21

Scale= 1.22

Weibull distribution
Loglik(model)= -7857.3   Loglik(intercept only)= -8104.2
        Chisq= 493.87 on 4 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 7
n= 6805
```

Resposta: A diabetes aumenta o risco e as doenças congênitas têm efeito protetor.

Existe evidência a favor da utilização de um modelo mais simples (exponencial)? Ou um modelo com menos variáveis? Remova as variáveis com p-valor menor do que 0,1 e compare o novo modelo com o modelo acima (uma dica: calcule a razão de verossimilhança entre os dois modelos).

Resposta:

Ajustando um modelo exponencial

```
> modeloE <- survreg(Surv(tempo, status) ~ idade + cdiab + crim +
+     congenita, dist = "exp", data = dialise)
> summary(modeloE)


Call:
survreg(formula = Surv(tempo, status) ~ idade + cdiab + crim +
    congenita, data = dialise, dist = "exp")
              Value Std. Error       z        p
(Intercept)  6.1643    0.10897  56.568 0.00e+00
idade       -0.0365    0.00176 -20.676 5.66e-95
cdiab       -0.3092    0.06029  -5.127 2.94e-07
crim        -0.0313    0.06696  -0.467 6.41e-01
congenita    0.7550    0.22616   3.338 8.43e-04

Scale fixed at 1

Exponential distribution
Loglik(model)= -7905.3   Loglik(intercept only)= -8169
        Chisq= 527.4 on 4 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 6
n= 6805
```

Comparando os modelos weibull e exponencial, isto é, testando a hipótese que o o parâmetro de forma γ é igual a 1, através da estatística de deviance

```
> dev <- 2 * (modelo1$loglik[2] - modeloE$loglik[2])
> dev


[1] 95.95186


> gl <- 1
> pvalor <- 1 - pchisq(dev, gl)
> pvalor


[1] 0
```

Alternativamente pode-se testar a redução do modelo usando o comando *anova()*. Observe que na coluna *Deviance* temos o mesmo valor que calculamos anteriormente (*dev = 95.95*)

```
> anova(modeloE, modelo1)

                             Terms Resid. Df    -2*LL Test Df Deviance
1 idade + cdiab + crim + congenita      6800 15810.56      NA       NA
2 idade + cdiab + crim + congenita      6799 15714.61    =  1 95.95186
     P(>|Chi|)
1           NA
2 1.177112e-22
```

Ajustando um modelo sem a covariável causas renais (*crim*)

```
> modelo2 <- survreg(Surv(tempo, status) ~ idade + cdiab + congenita,
+     data = dialise)
> summary(modelo2)

Call:
survreg(formula = Surv(tempo, status) ~ idade + cdiab + congenita,
    data = dialise)
              Value Std. Error      z        p
(Intercept)  6.7623    0.14798  45.70 0.00e+00
idade       -0.0428    0.00225 -19.00 1.70e-80
cdiab       -0.3510    0.07061  -4.97 6.69e-07
congenita    0.8951    0.27454   3.26 1.11e-03
Log(scale)   0.1951    0.02082   9.37 7.05e-21

Scale= 1.22

Weibull distribution
Loglik(model)= -7857.4   Loglik(intercept only)= -8104.2
        Chisq= 493.64 on 3 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 7
n= 6805
```

Testando a hipótese nula de que o modelo reduzido é melhor, ou em outras palavras, que o coeficiente da covariável *crim* não é significativo.

```
> dev <- 2 * (modelo1$loglik[2] - modelo2$loglik[2])
> dev

[1] 0.2218873
```

```
> gl <- modelo1$df - modelo2$df
> gl

[1] 1

> pvalor <- 1 - pchisq(dev, gl)
> pvalor

[1] 0.6376056
```

Alternativamente pode-se testar a redução do modelo usando o comando anova

```
> anova(modelo2, modelo1)

                            Terms Resid. Df    -2*LL  Test Df  Deviance
1        idade + cdiab + congenita      6800 15714.83       NA        NA
2 idade + cdiab + crim + congenita      6799 15714.61 +crim  1 0.2218873
  P(>|Chi|)
1        NA
2 0.6376056
```

Segundo o teste apresentado no sumário do modelo1 (modelo Weibull) e no teste de razão de verossimilhança, a estimativa do parâmetro de escala é significativamente diferente de 1 e portanto a redução para um modelo com menos parâmetros (exponencial) não seria adequada.

Adotando-se o modelo Weibull, testou-se a redução do modelo por exclusão da variável ***crim***, e o modelo reduzido não pode ser rejeitado sendo possível a retirada de tal variável do modelo. Isto significa que, segundo o modelo Weibull, a variável ***crim*** não é um fator prognóstico importante para a sobrevida.

7

Modelos de regressão semiparamétricos

Exercícios

Exercício 7.1: Os dados da coorte de transplante de medula óssea estão no arquivo *tmoclas.dat*. Abra o arquivo no R:

```
> tmo <- read.table("tmoclas.dat", header = T, sep = ",")
```

Nome das variáveis

```
> names(tmo)
```

```
 [1] "id"       "sexo"     "idade"    "status"   "os"       "plaq"
 [7] "tempplaq" "deag"     "tempdeag" "decr"     "tempdecr" "fase"
```

Transformando as variáveis categóricas em factor

```
> tmo$sexo <- factor(tmo$sexo)
> tmo$decr <- factor(tmo$decr)
> tmo$deag <- factor(tmo$deag)
> tmo$fase <- factor(tmo$fase)
```

1. Refaça os gráficos de Kaplan-Meier para as variáveis *sexo, deag, decr* e *fase* e verifique o pressuposto de proporcionalidade.

   Este item tem o objetivo de familiarizar o leitor com os comandos do R. A interpretação dos resultados está discutida ao longo do texto. Ver seção 7.3.2.

   Abrindo a biblioteca survival e criando o objeto sobrevida

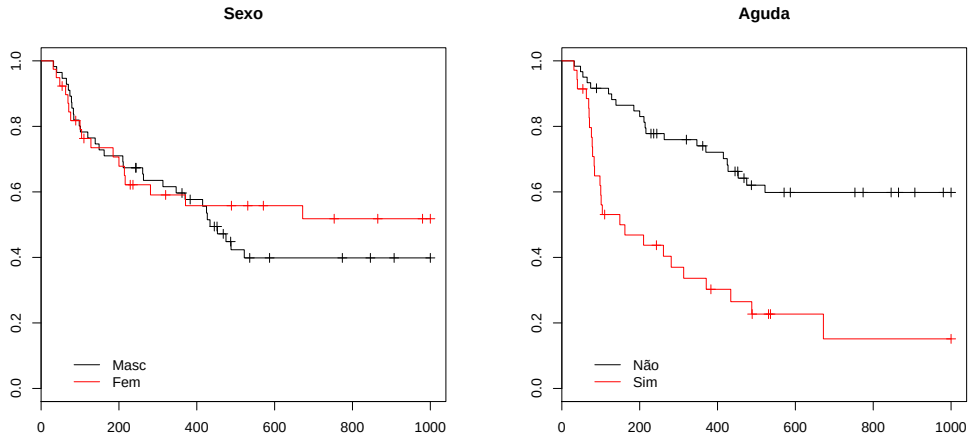
```
> require(survival)
> y <- Surv(tmo$os, tmo$status)
```

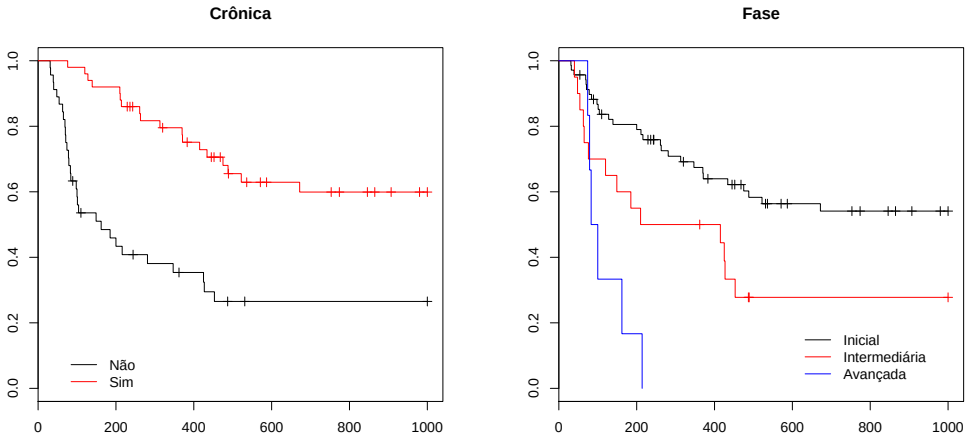
Estimando as curvas de sobrevida pelo Kaplan-Meier por diversas variáveis

```
> KMsexo <- survfit(y ~ sexo, data = tmo)
> KMdeag <- survfit(y ~ deag, data = tmo)
> KMdecr <- survfit(y ~ decr, data = tmo)
> KMfase <- survfit(y ~ fase, data = tmo)
```

Gráfico das curvas de Kaplan-Meier

```
> par(mfrow = c(2, 2))
> plot(KMsexo, main = "Sexo", col = 1:2, legend.tex = c("Masc",
+     "Fem"))
> plot(KMdeag, main = "Aguda", col = 1:2, legend.text = c("Não",
+     "Sim"))
> plot(KMdecr, main = "Crônica", col = 1:2, legend.text = c("Não",
+     "Sim"))
> plot(KMfase, main = "Fase", col = c(1, 2, 4), legend.pos = c(600,
+     0.2), legend.text = c("Inicial", "Intermediária", "Avançada"))
```





2. Refaça os quatro modelos de Cox apresentados no texto e compare os modelos usando a análise de *deviance*. Veja os comentários da seção 7.4

   Estimando os modelos de Cox

```
> mod1 <- coxph(y ~ idade + sexo, data = tmo)
> summary(mod1)

Call:
coxph(formula = y ~ idade + sexo, data = tmo)

  n= 96
         coef exp(coef) se(coef)     z    p
idade -0.0186     0.982   0.0141 -1.32 0.19
sexo2 -0.3299     0.719   0.3219 -1.02 0.31

      exp(coef) exp(-coef) lower .95 upper .95
idade     0.982       1.02     0.955      1.01
sexo2     0.719       1.39     0.383      1.35

Rsquare= 0.022   (max possible= 0.984 )
Likelihood ratio test= 2.16  on 2 df,   p=0.34
Wald test            = 2.11  on 2 df,   p=0.348
Score (logrank) test = 2.11  on 2 df,   p=0.348

> mod2 <- coxph(y ~ idade + sexo + fase, data = tmo)
> summary(mod2)

Call:
coxph(formula = y ~ idade + sexo + fase, data = tmo)
```

```
  n= 96
        coef exp(coef) se(coef)      z       p
idade -0.0255     0.975   0.0146 -1.752 8.0e-02
sexo2 -0.1641     0.849   0.3254 -0.504 6.1e-01
fase2  0.8932     2.443   0.3517  2.540 1.1e-02
fase3  1.9408     6.964   0.4884  3.974 7.1e-05

      exp(coef) exp(-coef) lower .95 upper .95
idade     0.975      1.026     0.947      1.00
sexo2     0.849      1.178     0.448      1.61
fase2     2.443      0.409     1.226      4.87
fase3     6.964      0.144     2.674     18.14

Rsquare= 0.165   (max possible= 0.984 )
Likelihood ratio test= 17.3  on 4 df,   p=0.00169
Wald test            = 19.7  on 4 df,   p=0.000577
Score (logrank) test = 23.5  on 4 df,   p=0.000101

> mod3 <- coxph(y ~ idade + sexo + fase + deag, data = tmo)
> summary(mod3)

Call:
coxph(formula = y ~ idade + sexo + fase + deag, data = tmo)

  n= 96
        coef exp(coef) se(coef)      z       p
idade -0.0163     0.984   0.0144 -1.134 0.26000
sexo2 -0.2053     0.814   0.3185 -0.645 0.52000
fase2  0.9815     2.668   0.3430  2.862 0.00420
fase3  1.5683     4.798   0.4978  3.150 0.00160
deag1  1.1848     3.270   0.3120  3.797 0.00015

      exp(coef) exp(-coef) lower .95 upper .95
idade     0.984      1.016     0.956      1.01
sexo2     0.814      1.228     0.436      1.52
fase2     2.668      0.375     1.362      5.23
fase3     4.798      0.208     1.809     12.73
deag1     3.270      0.306     1.774      6.03

Rsquare= 0.279   (max possible= 0.984 )
Likelihood ratio test= 31.4  on 5 df,   p=7.77e-06
Wald test            = 32.6  on 5 df,   p=4.43e-06
Score (logrank) test = 39.1  on 5 df,   p=2.24e-07

> mod4 <- coxph(y ~ idade + sexo + fase + deag + decr, data = tmo)
> summary(mod4)
```

```
Call:
coxph(formula = y ~ idade + sexo + fase + deag + decr, data = tmo)

  n= 96
          coef exp(coef) se(coef)      z       p
idade -0.00441     0.996   0.0149 -0.296 0.77000
sexo2 -0.22608     0.798   0.3329 -0.679 0.50000
fase2  0.64136     1.899   0.3765  1.703 0.08900
fase3  1.02796     2.795   0.5264  1.953 0.05100
deag1  1.25304     3.501   0.3307  3.789 0.00015
decr1 -0.97759     0.376   0.3404 -2.872 0.00410

      exp(coef) exp(-coef) lower .95 upper .95
idade     0.996      1.004     0.967     1.025
sexo2     0.798      1.254     0.415     1.532
fase2     1.899      0.527     0.908     3.972
fase3     2.795      0.358     0.996     7.844
deag1     3.501      0.286     1.831     6.693
decr1     0.376      2.658     0.193     0.733

Rsquare= 0.34   (max possible= 0.984 )
Likelihood ratio test= 39.9  on 6 df,   p=4.8e-07
Wald test            = 37.8  on 6 df,   p=1.21e-06
Score (logrank) test = 47.1  on 6 df,   p=1.79e-08
```

Análise de Deviance

```
> anova(mod1, mod2, mod3, mod4, test = "Chisq")

Analysis of Deviance Table

Model 1: y ~ idade + sexo
Model 2: y ~ idade + sexo + fase
Model 3: y ~ idade + sexo + fase + deag
Model 4: y ~ idade + sexo + fase + deag + decr
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        94     395.93
2        92     380.78  2    15.14 0.0005146
3        91     366.67  1    14.11 0.0001726
4        90     358.20  1     8.47 0.0036015
```

3. Calcule o índice de prognóstico do modelo 4 para um indivíduo com 35 anos, do sexo feminino (*sexo2 = 1*), na fase intermediária (*fase2 = 1, fase3 = 0*) da doença e que tenha tido doença do enxerto crônica, mas não aguda (*decr = 1, deag = 0*).

Colocando em um vetor as caracteristicas do individuo na ordem em que aparecem no modelo (*idade, sexo, fase, deag, decr*)

```
> individuo <- c(35, 1, 1, 0, 0, 1)
> individuo

[1] 35  1  1  0  0  1
```

Colocando em outro vetor os coeficientes do modelo

```
> coeficientes <- mod4$coef
> coeficientes

       idade        sexo2        fase2        fase3        deag1        decr1
-0.004413461 -0.226083177  0.641355011  1.027956377  1.253039591 -0.977592832
```

Para calcular o índice de prognóstico basta multiplicar os valores das variáveis do paciente (vetor *individuo*) pelos coeficientes do modelo (vetor *coeficientes*) e somar

```
> ip <- sum(coeficientes * individuo)
> ip

[1] -0.7167921
```

Temos então que o índice de prognóstico do indivíduo com as características citadas anteriormente é igual a -0.717.

Exercício 7.2: Três modelos causais aninhados são propostos para explicar a sobrevida de pacientes em diálise. O primeiro considera apenas a variável idade, o segundo inclui as doenças de base e o terceiro inclui uma variável ambiental (tamanho da unidade de tratamento).

Modelo I: sobrevida = idade

Modelo II: sobrevida = idade + cdiab + congenita + crim

Modelo III: sobrevida = idade + cdiab + congenita + crim + grande

1. Utilizando o banco de dados *dialise.csv*, faça gráficos de Kaplan-Meier estratificados por cada covariável categórica e verifique (visualmente) se há fortes indícios de não proporcionalidade (o que inviabilizaria o ajuste do modelo de Cox tradicional).

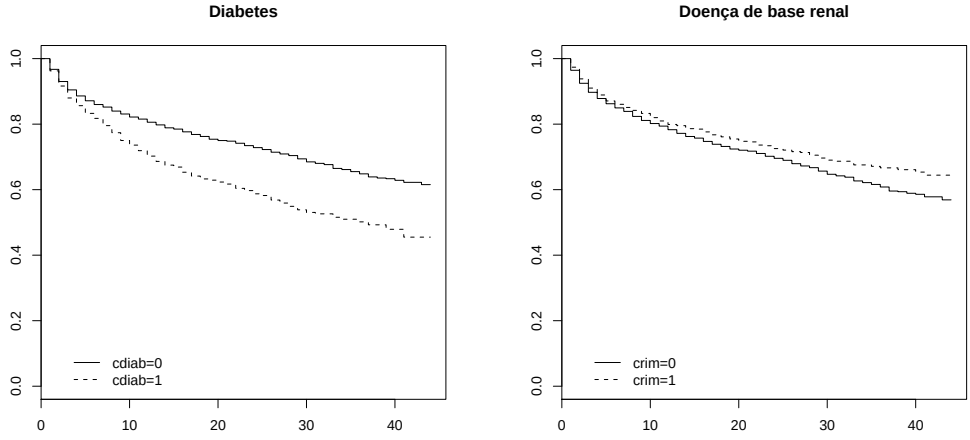
```
> dialise <- read.table("dialise.csv", header = T, sep = ",")
> dialise$grande <- factor(dialise$grande)
> dialise$cdiab <- factor(dialise$cdiab)
> dialise$crim <- factor(dialise$crim)
> dialise$congenita <- factor(dialise$congenita)
> names(dialise)

 [1] "unidade"   "idade"     "inicio"    "fim"       "status"    "tempo"
 [7] "grande"    "causa"     "cdiab"     "crim"      "congenita"

> y <- Surv(dialise$tempo, dialise$status)
```

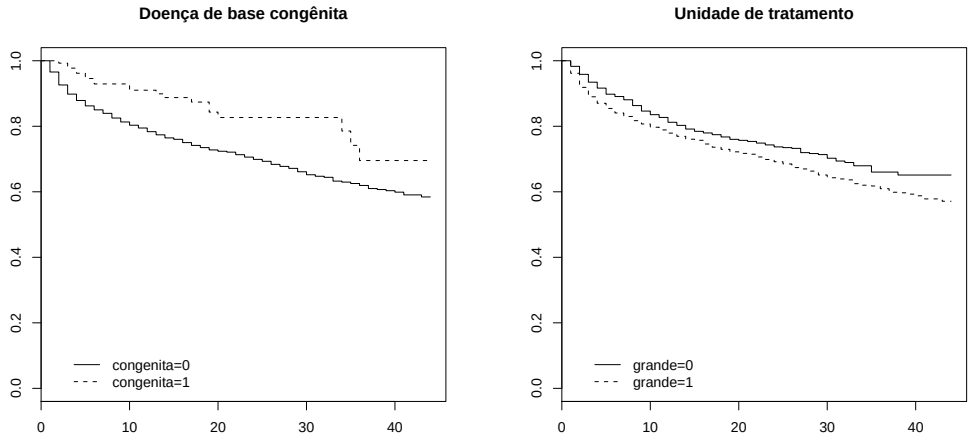
Estimando as curvas de Kaplan-Meier

```
> KMdiab <- survfit(y ~ cdiab, data = dialise)
> KMrim <- survfit(y ~ crim, data = dialise)
> KMcong <- survfit(y ~ congenita, data = dialise)
> KMgrande <- survfit(y ~ grande, data = dialise)
```

Gráficos das curvas KM

```
> par(mfrow = c(2, 2))
> plot(KMdiab, lty = c(1, 2), main = "Diabetes", mark.time = F,
+     legend.text = names(KMdiab$strata))
> plot(KMrim, lty = c(1, 2), main = "Doença de base renal", mark.time = F,
+     legend.text = names(KMrim$strata))
> plot(KMcong, lty = c(1, 2), main = "Doença de base congênita",
+     mark.time = F, legend.text = names(KMcong$strata))
> plot(KMgrande, lty = c(1, 2), main = "Unidade de tratamento",
+     mark.time = F, legend.text = names(KMgrande$strata))
```





Resposta: Podemos dizer que não há evidência de forte desvio do pressuposto de proporcionalidade. Apenas a variável congênita apresenta um padrão mais complexo, mas note que a tendência de queda é a mesma, em média, e os degraus observados na curva superior se deve, em grande parte, ao pequeno número de amostras. Em uma situação como esta, é interessante modelar com e sem esta variável, para verificar o efeito de sua inclusão/exclusão na estimativa das outras.

2. Ajuste cada modelo causal acima utilizando o modelo de riscos proporcionais de Cox, tomando o cuidado de interpretar os parâmetros a cada saída. (Lembre de acrescentar o argumento $x = T$, na especificação da função *coxph()*, para que depois se possam obter os índices de prognóstico.)

```
> modeloI <- coxph(y ~ idade, data = dialise, x = T)
> summary(modeloI)

Call:
coxph(formula = y ~ idade, data = dialise, x = T)

  n= 6805
       coef exp(coef) se(coef)    z p
idade 0.0349      1.04  0.00173 20.2 0

      exp(coef) exp(-coef) lower .95 upper .95
idade      1.04      0.966      1.03      1.04

Rsquare= 0.062   (max possible= 0.98 )
Likelihood ratio test= 435  on 1 df,   p=0
```

```
Wald test            = 409  on 1 df,   p=0
Score (logrank) test = 415  on 1 df,   p=0

> modeloII <- coxph(y ~ idade + cdiab + crim + congenita, data = dialise,
+     x = T)
> summary(modeloII)

Call:
coxph(formula = y ~ idade + cdiab + crim + congenita, data = dialise,
    x = T)

  n= 6805
              coef exp(coef) se(coef)      z       p
idade       0.0344     1.035  0.00175 19.677 0.0e+00
cdiab1      0.2856     1.331  0.06031  4.736 2.2e-06
crim1       0.0303     1.031  0.06698  0.453 6.5e-01
congenita1 -0.7152     0.489  0.22618 -3.162 1.6e-03

           exp(coef) exp(-coef) lower .95 upper .95
idade          1.035      0.966     1.032     1.039
cdiab1         1.331      0.752     1.182     1.498
crim1          1.031      0.970     0.904     1.175
congenita1     0.489      2.045     0.314     0.762

Rsquare= 0.067   (max possible= 0.98 )
Likelihood ratio test= 473  on 4 df,   p=0
Wald test            = 438  on 4 df,   p=0
Score (logrank) test = 451  on 4 df,   p=0

> modeloIII <- coxph(y ~ idade + cdiab + crim + congenita + grande,
+     data = dialise, x = T)
> summary(modeloIII)

Call:
coxph(formula = y ~ idade + cdiab + crim + congenita + grande,
    data = dialise, x = T)

  n= 6805
              coef exp(coef) se(coef)      z       p
idade       0.0340     1.035  0.00175 19.386 0.0e+00
cdiab1      0.2955     1.344  0.06040  4.892 1.0e-06
crim1       0.0191     1.019  0.06708  0.285 7.8e-01
congenita1 -0.7274     0.483  0.22622 -3.216 1.3e-03
grande1     0.1770     1.194  0.06339  2.792 5.2e-03

           exp(coef) exp(-coef) lower .95 upper .95
idade          1.035      0.967     1.031     1.038
```

```
cdiab1          1.344      0.744     1.194     1.513
crim1           1.019      0.981     0.894     1.163
congenita1      0.483      2.070     0.310     0.753
grande1         1.194      0.838     1.054     1.352

Rsquare= 0.068   (max possible= 0.98 )
Likelihood ratio test= 481  on 5 df,   p=0
Wald test            = 447  on 5 df,   p=0
Score (logrank) test = 461  on 5 df,   p=0
```

Resposta: Ajustando o Modelo I, encontramos que a idade apresenta-se como um fator de risco de 1.04. Isto é, cada ano de idade a mais na data de início da diálise implica em um risco 4% maior de óbito. No modelo II, a idade continua significativa, embora tenha perdido um pouco do efeito. A doença de base diabetes se mostrou um importante fator de risco, com um sobrerisco de 33%. A causa renal também se mostrou um fator de risco, porém seu efeito não foi significativo. A doença de base congênita se mostrou um fator protetor: pessoas sem doença congênita têm risco duas vezes maior de ir a óbito do que as com causa congênita. Este efeito protetor pode ser interpretado como um efeito indireto. Isto é, não é a causa congênita que protege, mas sim a ausência da diabetes ou da causa renal. Outra questão que deve ser levada em consideração é a prevalência baixa de pessoas com doená congênitas neste banco de dados que são somente 142 do total de 6805 pacientes. No modelo III, a inclusão da variável grande não alterou significativamente o efeito das outras variáveis. O tamanho da unidade se mostrou importante, com pacientes atendidos em unidades grandes tendo risco 19% maior de ir a óbito do que aqueles atendidos em unidades menores, com menos de 5 salas.

3. Compare os modelos usando a análise de *deviance* e escolha o modelo com melhor ajuste.

```
> anova(modeloI, modeloII, modeloIII, test = "Chisq")

Analysis of Deviance Table

Model 1: y ~ idade
Model 2: y ~ idade + cdiab + crim + congenita
Model 3: y ~ idade + cdiab + crim + congenita + grande
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      6804    26073.1
2      6801    26034.9    3     38.2 2.507e-08
3      6800    26026.8    1      8.1 4.512e-03
```

Resposta: A inclusão das variáveis de doença de base (Modelo II) melhorou significativamente o ajuste do modelo, quando comparado com o modelo contendo apenas a idade (Deviance = 38.2, p < 0.0001). A inclusão da variável tamanho da unidade (*grande*) melhorou ainda mais o ajuste (comparando os modelos II e III, temos Deviance = 8.2, p < 0.001). Concluimos, então, que o Modelo III é o melhor modelo.

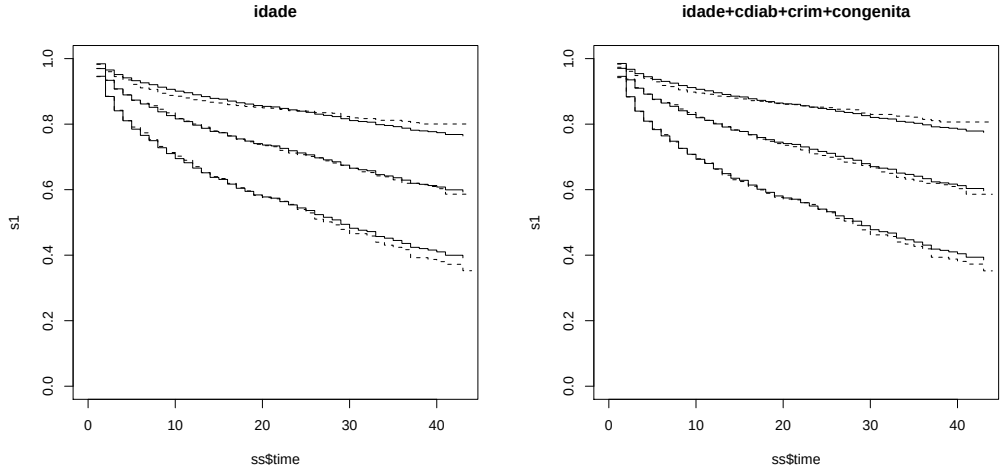
4. Qual o poder explicativo do modelo escolhido? Calcule a razão entre o R^2 do modelo escolhido e o R^2 máximo (ambos estão presentes na saída do comando *summary()*).

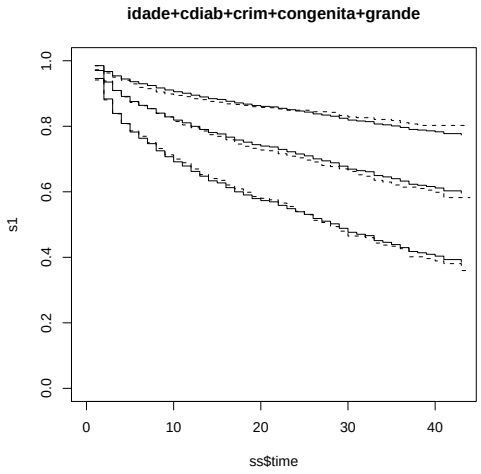
   Resposta: O modelo III explicou $0.068/0.98 \times 100 = 6.9\%$ da variância dos dados.

5. Faça o gráfico dos índices de prognóstico do modelo escolhido, utilizando a função *plot.pi(modelo escolhido)*. Avalie visualmente o ajuste do modelo à curva de Kaplan-Meier. A função *plot.pi()* não existe na biblioteca *survival* do R. Antes de utilizar a função *plot.pi()* precisamos executar o comando *source("Rfun.r")* para criá-la.

```
> source("Rfun.r")

> par(mfrow = c(2, 2))
> plot.pi(modeloI, main = "idade")
> plot.pi(modeloII, main = "idade+cdiab+crim+congenita")
> plot.pi(modeloIII, main = "idade+cdiab+crim+congenita+grande")
```

   A linha sólida é o modelo ajustado e a linha pontinhada é o Kaplan-Meier

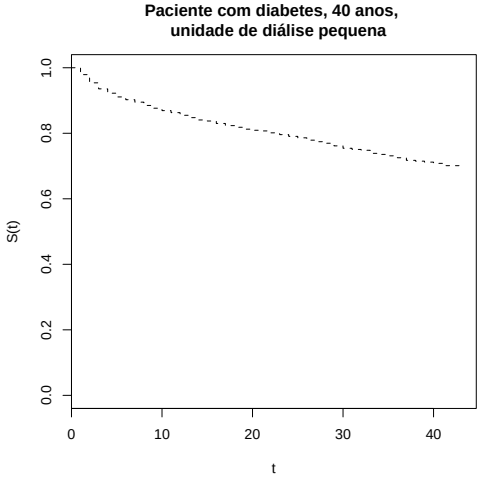




Resposta: O gráfico dos grupos com alto, médio e baixo índice de prognóstico indica que o modelo de Cox é capaz de distinguir entre os três grupos. As curvas estimadas acompanham de perto as curvas de Kaplan-Meier.

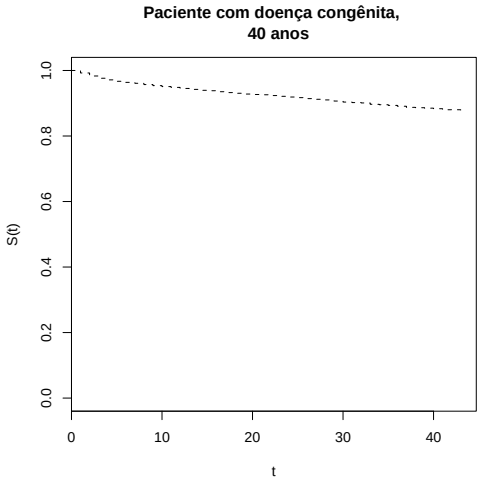
6. Podemos utilizar o modelo ajustado de Cox para obter estimativas de sobrevida para um paciente com determinado perfil. Por exemplo, o gráfico da curva de sobrevida (de acordo com o modelo III), para um paciente com causa de base diabetes, 40 anos de idade, e tratado em uma unidade pequena seria obtido com os comandos:

```
> paciente1 <- survfit(modeloIII, newdata = list(idade = 40, cdiab = factor(1,
+     levels = 0:1), crim = factor(0, levels = 0:1), congenita = factor(0,
+     levels = 0:1), grande = factor(0, levels = 0:1)))
> par(mfrow = c(1, 1))
> plot(paciente1, conf = F, lty = 2, ylab = "S(t)", xlab = "t",
+     main = "Paciente com diabetes, 40 anos, \n  unidade de diálise pequena")
```



7. E um paciente com mesma idade e local de tratamento, mas com doença congênita?

```
> paciente2 <- survfit(modeloIII, newdata = list(idade = 40, cdiab = factor(0,
+     levels = 0:1), crim = factor(0, levels = 0:1), congenita = factor(1,
+     levels = 0:1), grande = factor(0, levels = 0:1)))
> plot(paciente2, conf = F, lty = 2, ylab = "S(t)", xlab = "t",
+     main = "Paciente com doença congênita, \n  40 anos")
```



Exercício 7.3: No exercício 6.3 ajustamos um modelo explicativo para sobrevida em Aids contendo as variáveis sexo, idade, tratamento e tipo de acompanhamento,

utilizando regressão paramétrica (banco de dados *ipec.csv*). Faça agora uma análise desses dados utilizando o modelo de Cox, considerando três modelos explicativos aninhados.

Modelo I: sobrevida = idade + sexo

Modelo II: sobrevida = idade + sexo + acompan

Modelo III: sobrevida = idade + sexo + acompan + tratam

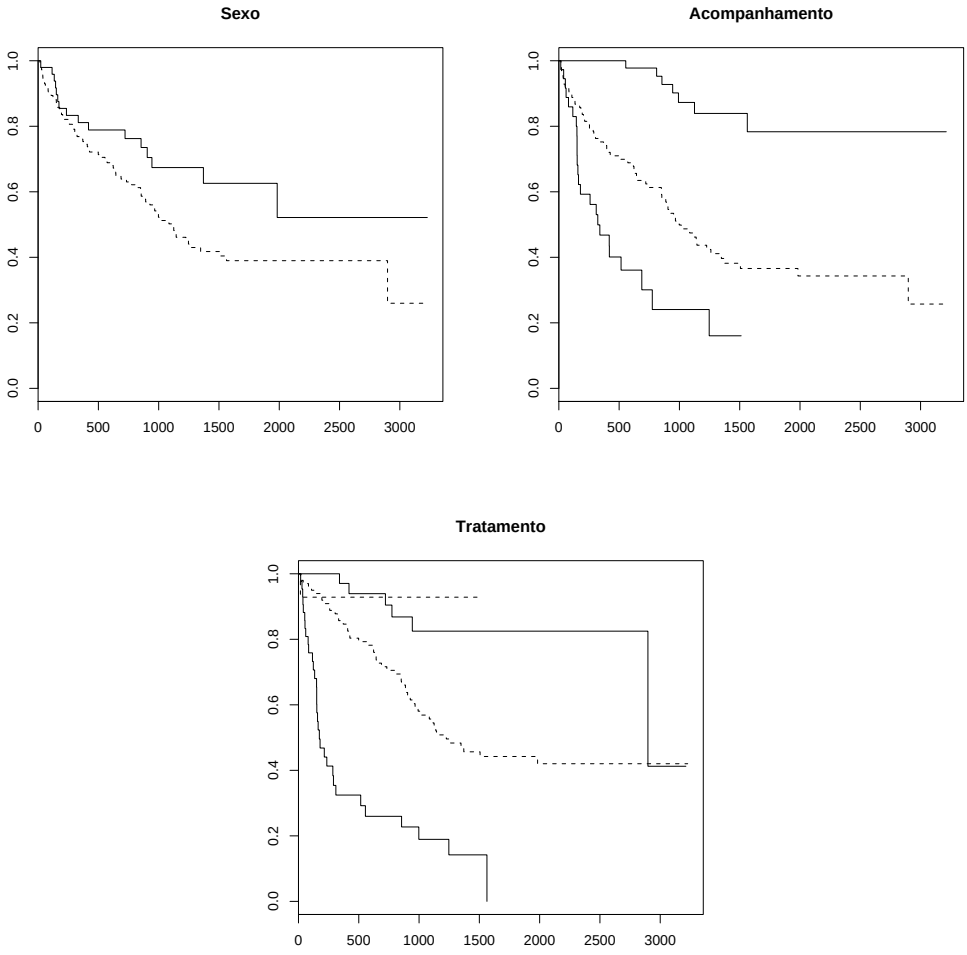
1. Faça gráficos de Kaplan-Meier estratificados por cada covariável categórica para verificar o pressuposto dos modelos de Cox.

   Lendo os dados

```
> ipec <- read.table("ipec.csv", header = T, sep = ";")
> names(ipec)

 [1] "id"      "ini"     "fim"     "tempo"   "status"  "sexo"    "escola"
 [8] "idade"   "risco"   "acompan" "obito"   "anotrat" "tratam"  "doenca"
[15] "propcp"

> ipec$acompan <- factor(ipec$acompan)
> ipec$tratam <- factor(ipec$tratam)
```

   Estimando os Kaplan-Meier

```
> KMsexo <- survfit(Surv(tempo, status) ~ sexo, data = ipec)
> KMacompan <- survfit(Surv(tempo, status) ~ acompan, data = ipec)
> KMtrat <- survfit(Surv(tempo, status) ~ tratam, data = ipec)

> par(mfrow = c(2, 2))
> plot(KMsexo, lty = c(1, 2), main = "Sexo", mark.time = F)
> plot(KMacompan, lty = c(1, 2), main = "Acompanhamento", mark.time = F)
> plot(KMtrat, lty = c(1, 2), main = "Tratamento", mark.time = F)
```



Resposta: Nas duas primeiras variáveis, não há mudança no tempo entre as categorias (não se cruzam e caem relativamente paralelas). A terceira variável tem muitas categorias e fica mais difícil de avaliar. Não há violação óbvia (linhas cruzando por exemplo), mas também não há um padrão claro de proporcionalidade. Vamos prosseguir com a análise, mas com cuidado.

2. Ajuste cada modelo causal acima utilizando o modelo de riscos proporcionais de Cox, tomando o cuidado de interpretar os parâmetros a cada saída. (Lembre de acrescentar o argumento $x = T$, na especificação do coxph, para que se possa utilizá-lo depois para obter os índices de prognóstico.)

```
> modeloI <- coxph(Surv(tempo, status) ~ idade + sexo, data = ipec,
```

```
+     x = T)
> summary(modeloI)

Call:
coxph(formula = Surv(tempo, status) ~ idade + sexo, data = ipec,
    x = T)

  n= 193
        coef exp(coef) se(coef)     z     p
idade -0.0127     0.987   0.0116 -1.10 0.270
sexoM  0.5562     1.744   0.2761  2.01 0.044

      exp(coef) exp(-coef) lower .95 upper .95
idade     0.987      1.013     0.965      1.01
sexoM     1.744      0.573     1.015      3.00

Rsquare= 0.029   (max possible= 0.988 )
Likelihood ratio test= 5.64  on 2 df,   p=0.0597
Wald test            = 5.09  on 2 df,   p=0.0783
Score (logrank) test = 5.18  on 2 df,   p=0.0748

> modeloII <- coxph(Surv(tempo, status) ~ idade + sexo + acompan,
+     data = ipec, x = T)
> summary(modeloII)

Call:
coxph(formula = Surv(tempo, status) ~ idade + sexo + acompan,
    data = ipec, x = T)

  n= 193
             coef exp(coef) se(coef)      z       p
idade    -0.00166     0.998   0.0120 -0.138 8.9e-01
sexoM     0.27218     1.313   0.2818  0.966 3.3e-01
acompan1  1.70732     5.514   0.4010  4.257 2.1e-05
acompan2  2.51763    12.399   0.4452  5.654 1.6e-08

         exp(coef) exp(-coef) lower .95 upper .95
idade        0.998     1.0017     0.975      1.02
sexoM        1.313     0.7617     0.756      2.28
acompan1     5.514     0.1814     2.513     12.10
acompan2    12.399     0.0807     5.181     29.67

Rsquare= 0.231   (max possible= 0.988 )
Likelihood ratio test= 50.7  on 4 df,   p=2.54e-10
Wald test            = 37  on 4 df,   p=1.79e-07
Score (logrank) test = 49  on 4 df,   p=5.84e-10
```

```
> modeloIII <- coxph(Surv(tempo, status) ~ idade + sexo + acompan +
+     tratam, data = ipec, x = T)
> summary(modeloIII)

Call:
coxph(formula = Surv(tempo, status) ~ idade + sexo + acompan +
    tratam, data = ipec, x = T)

  n= 193
             coef exp(coef) se(coef)      z       p
idade     0.00143    1.0014   0.0121  0.118 9.1e-01
sexoM     0.07424    1.0771   0.2858  0.260 8.0e-01
acompan1  1.67618    5.3451   0.4084  4.105 4.0e-05
acompan2  2.15300    8.6107   0.4672  4.608 4.1e-06
tratam1  -1.24192    0.2888   0.3011 -4.124 3.7e-05
tratam2  -2.09674    0.1229   0.4705 -4.456 8.3e-06
tratam3  -2.94502    0.0526   1.0188 -2.891 3.8e-03

         exp(coef) exp(-coef) lower .95 upper .95
idade       1.0014      0.999   0.97792     1.026
sexoM       1.0771      0.928   0.61518     1.886
acompan1    5.3451      0.187   2.40080    11.900
acompan2    8.6107      0.116   3.44629    21.514
tratam1     0.2888      3.462   0.16007     0.521
tratam2     0.1229      8.140   0.04885     0.309
tratam3     0.0526     19.011   0.00714     0.387

Rsquare= 0.372   (max possible= 0.988 )
Likelihood ratio test= 89.8  on 7 df,   p=1.11e-16
Wald test            = 76.9  on 7 df,   p=5.97e-14
Score (logrank) test = 100  on 7 df,   p=0
```

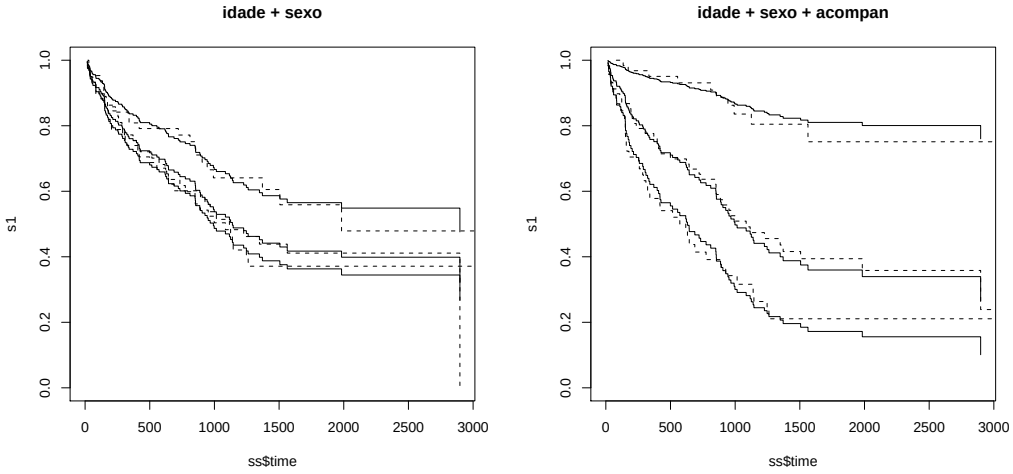
Resposta: O modelo I indica que sexo masculino é um importante fator de risco, com razão de riscos de 1.744. Este modelo, porém, tem um poder explicativo muito baixo e não é significativamente melhor do que o modelo nulo (teste Wald = 5.09, p = 0.0783). A inclusão da variável acompanhamento melhora muito o ajuste ($R^2 = 0.231/0.988 \times 100 = 23.3\%$). Neste modelo, as variáveis demográficas se tornam não significativas. A variável ***acompan*** é um importante fator de risco, o que é razoável, pois pacientes internados têm maior risco de ir a óbito do que pacientes atendidos no ambulatório apenas (uma vez que o tipo de atendimento está associado com a gravidade do quadro clínico). O modelo III mostra que a variável tratamento tem forte efeito protetor, mesmo controlando por tipo de acompanhamento. Um paciente sem tratamento tem 3.5 vezes mais chance de ir a óbito, por unidade de tempo, do

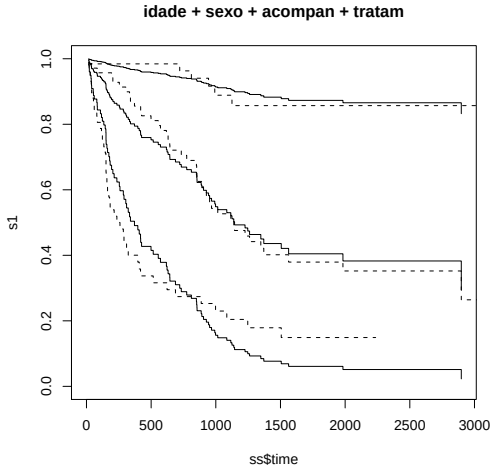
que o paciente com monoterapia, 8.1 vezes mais chance do que um paciente com terapia combinada e 19 vezes mais chance do que o paciente recebendo terapia potente.

3. Compare os modelos usando a análise de deviance e o gráfico dos índices de prognóstico.

```
> anova(modeloI, modeloII, modeloIII, test = "Chisq")

Analysis of Deviance Table

Model 1: Surv(tempo, status) ~ idade + sexo
Model 2: Surv(tempo, status) ~ idade + sexo + acompan
Model 3: Surv(tempo, status) ~ idade + sexo + acompan + tratam
  Resid. Df Resid. Dev  Df Deviance P(>|Chi|)
1       191     844.60
2       189     799.50   2    45.10 1.612e-10
3       186     760.39   3    39.11 1.648e-08

> par(mfrow = c(2, 2))
> plot.pi(modeloI, main = "idade + sexo")
> plot.pi(modeloII, main = "idade + sexo + acompan")
> plot.pi(modeloIII, main = "idade + sexo + acompan + tratam")
```

A linha sólida é o modelo ajustado e a linha pontinhada é o Kaplan-Meier





Resposta: A análise de deviance claramente apoia o modelo III, porém o gráfico de índices de prognóstico mostra que o modelo III está mais descolado da curva Kaplan-Meier do que o modelo II. Isto sugere que a variável tratamento, embora ajude na explicação da variância, provavelmente não teve seu efeito capturado pelo modelo de riscos proporcionais de Cox. Isto confirma a impressão de não proporcionalidade sugerida pela análise dos gráficos de Kaplan-Meier inicial. Concluindo, esta análise sugere que o tratamento tem efeito significativo na sobrevida, mas o valor quantitativo estimado do efeito não é confiável, devido ao desvio do pressuposto de proporcionalidade. Outros modelos, que relaxam o pressuposto de proporcionalidade, são necessários para uma melhor estimativa do efeito do tratamento na sobrevida em aids.

4. Qual o poder explicativo do modelo escolhido? Calcule a razão entre o R^2 do modelo escolhido e o R^2 máximo (ambos estão presentes na saída do comando `summary()`).

   Resposta: Escolhendo o modelo III, temos que seu poder explicativo foi de $0.372/0.988 \times 100 = 37.65\%$.

8

Análise de resíduos para modelos de Cox

Exercícios

Neste capítulo, serão feitas as análises dos resíduos dos modelos ajustados no Capítulo 7. Por isso, é necessário primeiro realizar os exercícios do capítulo anterior.

Exercício 8.1: Encontramos, ao ajustar o modelo de Cox aos dados de transplante de medula óssea (banco de dados *tmoclas.dat*), que o melhor modelo explicativo da sobrevida incluía as covariáveis *idade*, *sexo*, *fase*, *deag* e *decr*. Esse é o modelo que serve de exemplo neste capítulo. Refaça, no R, a análise de resíduos apresentada ao longo do texto.

```
> tmo <- read.table("tmoclas.dat", header = T, sep = ",")
> names(tmo)
```

```
 [1] "id"       "sexo"     "idade"    "status"   "os"       "plaq"
 [7] "tempplaq" "deag"     "tempdeag" "decr"     "tempdecr" "fase"
```

```
> tmo$sexo <- factor(tmo$sexo)
> tmo$decr <- factor(tmo$decr)
> tmo$deag <- factor(tmo$deag)
> tmo$fase <- factor(tmo$fase)
> mod4 <- coxph(Surv(os, status) ~ idade + sexo + fase + deag +
+     decr, data = tmo)
> mod4
```

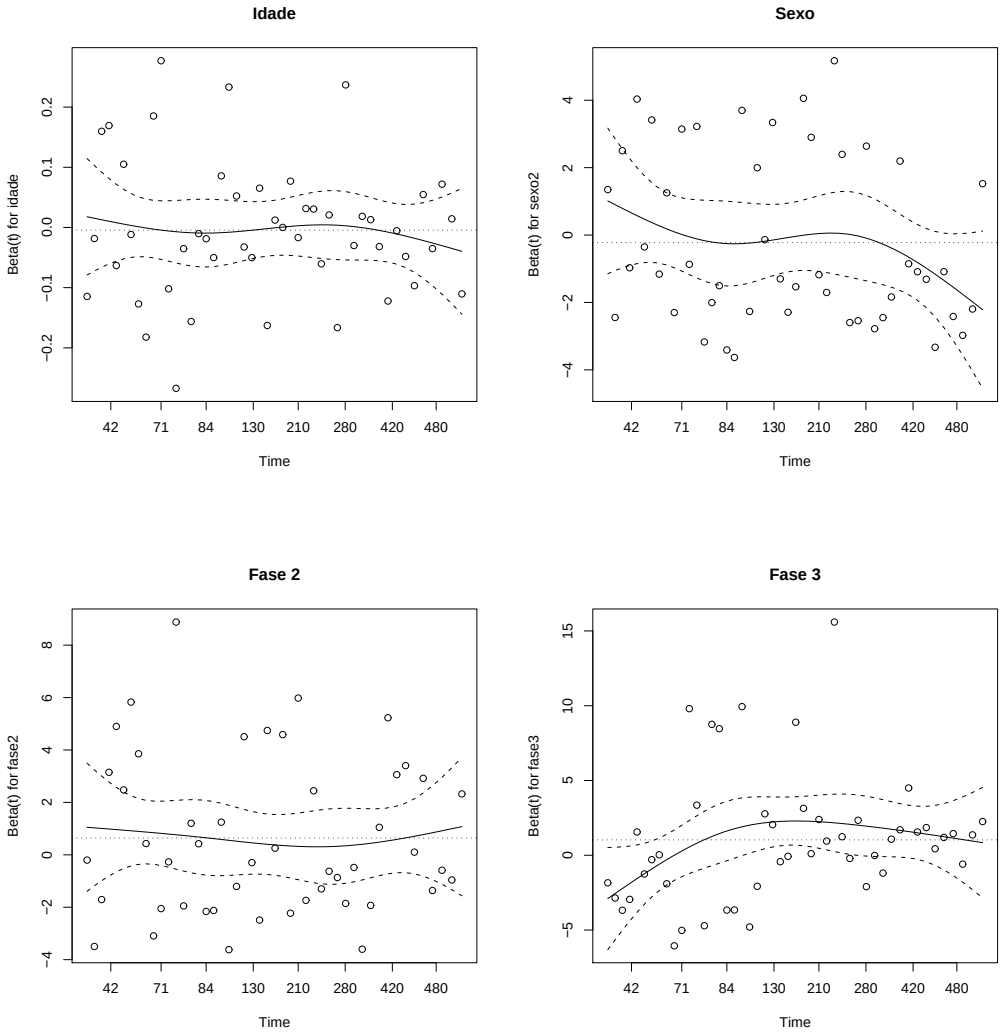
```
Call:
coxph(formula = Surv(os, status) ~ idade + sexo + fase + deag +
    decr, data = tmo)


          coef exp(coef) se(coef)      z       p
idade -0.00441     0.996   0.0149 -0.296 0.77000
sexo2 -0.22608     0.798   0.3329 -0.679 0.50000
fase2  0.64136     1.899   0.3765  1.703 0.08900
fase3  1.02796     2.795   0.5264  1.953 0.05100
deag1  1.25304     3.501   0.3307  3.789 0.00015
decr1 -0.97759     0.376   0.3404 -2.872 0.00410

Likelihood ratio test=39.9  on 6 df, p=4.8e-07  n= 96
```

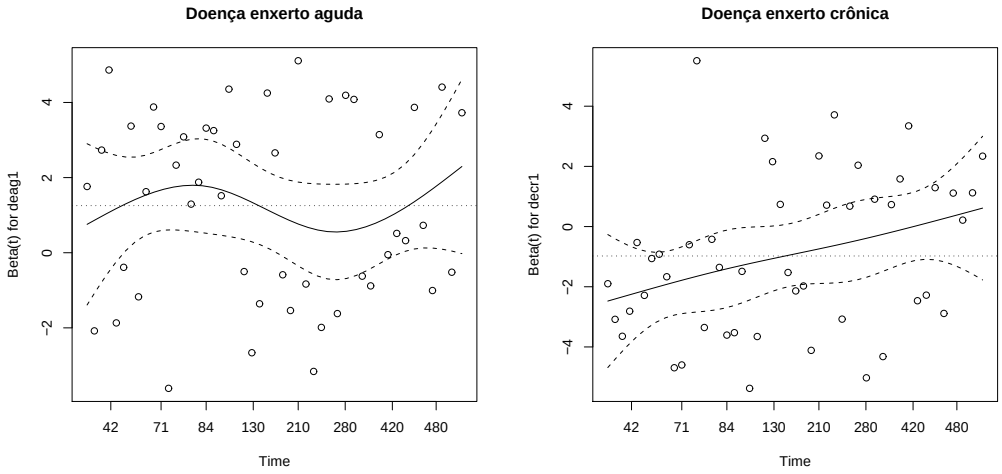
Calculando os resíduos de Schoenfeld

```
> zph <- cox.zph(mod4)
> zph


           rho   chisq      p
idade  -0.0674  0.2547 0.6138
sexo2  -0.2260  2.8393 0.0920
fase2  -0.0317  0.0617 0.8039
fase3   0.2063  2.8416 0.0919
deag1  -0.0147  0.0117 0.9137
decr1   0.3341  6.4078 0.0114
GLOBAL      NA 13.1118 0.0413
```

Gráfico dos resíduos Schoenfeld

```
> par(mfrow = c(2, 3))
> plot(zph[1], main = "Idade")
> abline(h = mod4$coef[1], lty = 3)
> plot(zph[2], main = "Sexo")
> abline(h = mod4$coef[2], lty = 3)
> plot(zph[3], main = "fase 2")
> abline(h = mod4$coef[3], lty = 3)
> plot(zph[4], , main = "Fase 3")
> abline(h = mod4$coef[4], lty = 3)
> plot(zph[5], main = "Doença enxerto aguda")
> abline(h = mod4$coef[5], lty = 3)
> plot(zph[6], main = "Doença enxerto crônica")
> abline(h = mod4$coef[6], lty = 3)
```


Idade
Beta(t) for idade
−0.2
−0.1
0.0
0.1
0.2
42
71
84
130
210
280
420
480
Time
Sexo
Beta(t) for sexo2
−4
−2
0
2
4
42
71
84
130
210
280
420
480
Time
Fase 2
Beta(t) for fase2
−4
−2
0
2
4
6
8
42
71
84
130
210
280
420
480
Time
Fase 3
Beta(t) for fase3
−5
0
5
10
15
42
71
84
130
210
280
420
480
Time

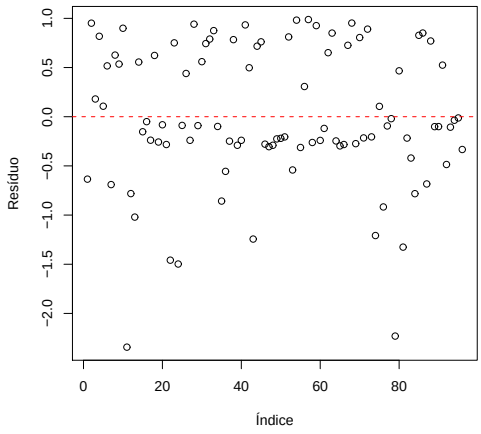


Resíduos de Martingale

```
> mod4.mar <- resid(mod4, "martingale")
```

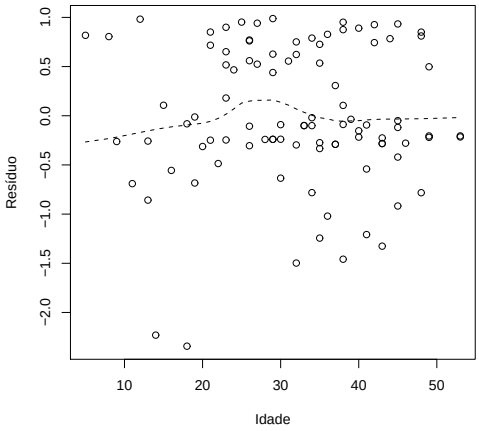
Identificando os pontos mal ajustados através do gráfico dos residuos de Martingale x individuos

```
> plot(mod4.mar, xlab = "Índice", ylab = "Resíduo")
> abline(h = 0, col = 2, lty = 2)
```



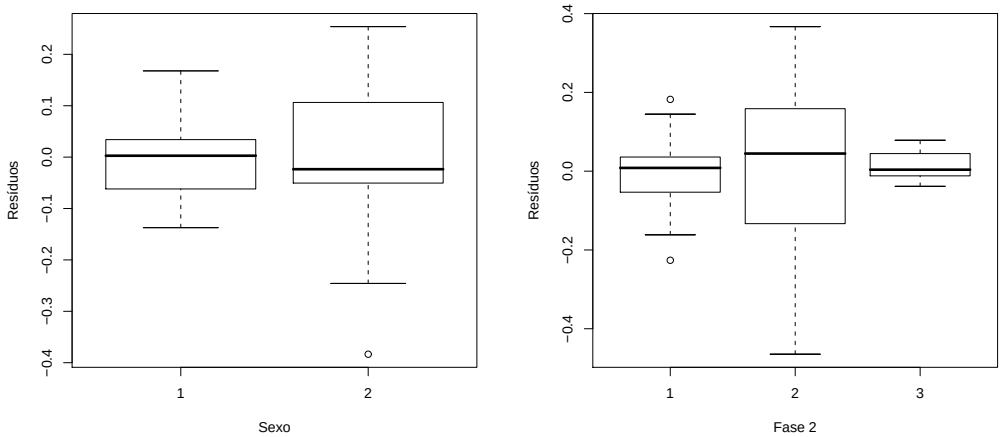
Analisando a forma funcional através do gráfico dos resíduos de Martingale x idade

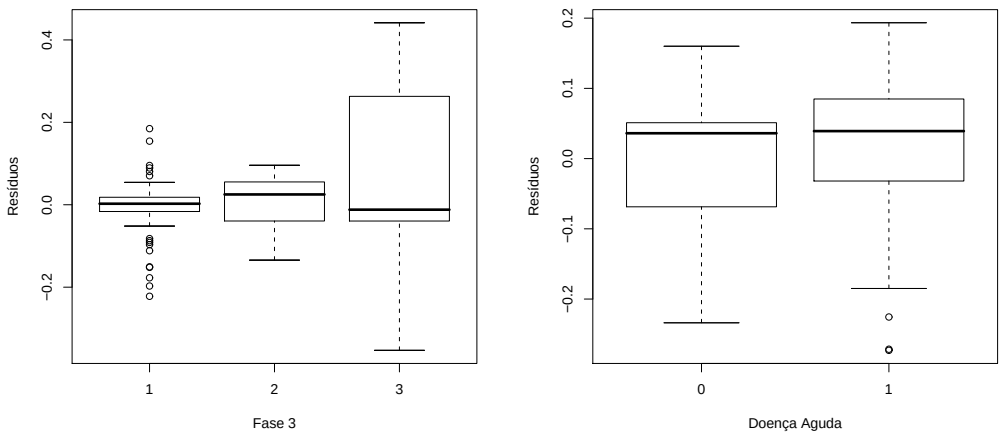
```
> plot(tmo$idade, mod4.mar, xlab = "Idade", ylab = "Resíduo")
> lines(lowess(tmo$idade, mod4.mar, iter = 0), lty = 2)
```

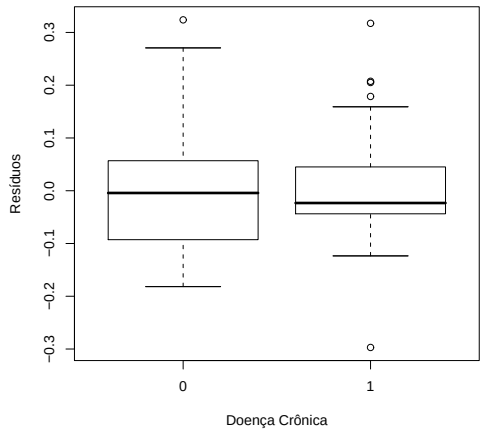


Gráficos dos Residuos scores

```
> mod4.sco2 <- resid(mod4, type = "dfbetas")

> par(mfrow = c(3, 2))
> plot(tmo$sexo, mod4.sco2[, 2], xlab = "Sexo", ylab = "Resíduos")
> plot(tmo$fase, mod4.sco2[, 3], xlab = "Fase 2", ylab = "Resíduos")
> plot(tmo$fase, mod4.sco2[, 4], xlab = "Fase 3", ylab = "Resíduos")
> plot(tmo$deag, mod4.sco2[, 5], xlab = "Doença Aguda", ylab = "Resíduos")
> plot(tmo$decr, mod4.sco2[, 6], xlab = "Doença Crônica", ylab = "Resíduos")
```







Exercício 8.2: No estudo de sobrevida de pacientes em diálise (exercício 7.2), encontramos que o modelo contendo variáveis demográficas, clínicas e ambientais (*idade*, *cdiab*, *crim*, *congenita* e *grande*) foi o que melhor se ajustou aos dados. Pela análise visual do gráfico de Kaplan-Meier realizada no exercício, todas as variáveis pareciam atender ao pressuposto de Cox. A única que parecia levantar dúvidas era a variável *congenita*. Vamos reavaliar essas variáveis, agora utilizando os resíduos de Schoenfeld, fazendo uma análise visual dos resíduos e calculando o teste da correlação.

```
> dialise <- read.table("dialise.csv", header = T, sep = ",")
> y <- Surv(dialise$tempo, dialise$status)
```

```
> modeloIII <- coxph(y ~ idade + cdiab + crim + congenita + grande,
+     data = dialise)
> zph <- cox.zph(modeloIII)
> zph
```

```
               rho   chisq       p
idade      0.05367  5.0743 0.02428
cdiab      0.04419  3.1462 0.07610
crim       0.00773  0.0962 0.75641
congenita  0.05295  4.4936 0.03402
grande    -0.06718  7.2685 0.00702
GLOBAL          NA 19.7751 0.00138
```

Resposta: O teste da correlação linear sugere que idade, doença de base congênita e tamanho da unidade de tratamento não atendem o pressuposto de proporcionalidade. Os gráficos, por outro lado, mostram que a não proporcionalidade ocorre principalmente para os tempos muito longos (> 30 meses). É possível que, censurando estes valores, obtenha-se um modelo com proporcionalidade. A variável idade tem um padrão mais definido, no entanto, com maior variabilidade nos tempos menores.

Faça também o gráfico dos resíduos martingale *versus* o índice dos indivíduos. Há indicação de indivíduos mal ajustados pelo modelo?

```
> mod.mar <- resid(modeloIII, "martingale")
> plot(mod.mar, xlab = "Índice", ylab = "Resíduo")
> abline(h = 0, col = 2, lty = 2)
```

Resposta: O gráfico de resíduo martingale versus indivíduo sugere a ausência de indivíduos mal-ajustados pelo modelo.

Exercício 8.3: No ajuste do modelo de Cox aos dados de sobrevida em Aids (exercício 7.3), vimos que a variável tratamento não parece atender ao pressuposto de Cox. Faça a análise de resíduos do modelo III proposto naquele exercício, e procure confirmar esse achado, calculando os resíduos de Schoenfeld.

Lendo os dados

```
> ipec <- read.table("ipec.csv", header = T, sep = ";")
> names(ipec)
```

```
 [1] "id"      "ini"     "fim"     "tempo"   "status"  "sexo"    "escola"
 [8] "idade"   "risco"   "acompan" "obito"   "anotrat" "tratam"  "doenca"
[15] "propcp"
```

```
> ipec$acompan <- factor(ipec$acompan)
> ipec$tratam <- factor(ipec$tratam)
```

Ajustando o modelo de Cox

```
> modeloIII <- coxph(Surv(tempo, status) ~ idade + sexo + acompan +
+     tratam, data = ipec, x = T)
> summary(modeloIII)


Call:
coxph(formula = Surv(tempo, status) ~ idade + sexo + acompan +
    tratam, data = ipec, x = T)

  n= 193
             coef exp(coef) se(coef)      z       p
idade     0.00143    1.0014   0.0121  0.118 9.1e-01
sexoM     0.07424    1.0771   0.2858  0.260 8.0e-01
acompan1  1.67618    5.3451   0.4084  4.105 4.0e-05
acompan2  2.15300    8.6107   0.4672  4.608 4.1e-06
tratam1  -1.24192    0.2888   0.3011 -4.124 3.7e-05
tratam2  -2.09674    0.1229   0.4705 -4.456 8.3e-06
tratam3  -2.94502    0.0526   1.0188 -2.891 3.8e-03

         exp(coef) exp(-coef) lower .95 upper .95
idade       1.0014      0.999   0.97792     1.026
sexoM       1.0771      0.928   0.61518     1.886
acompan1    5.3451      0.187   2.40080    11.900
acompan2    8.6107      0.116   3.44629    21.514
tratam1     0.2888      3.462   0.16007     0.521
tratam2     0.1229      8.140   0.04885     0.309
tratam3     0.0526     19.011   0.00714     0.387

Rsquare= 0.372   (max possible= 0.988 )
Likelihood ratio test= 89.8  on 7 df,   p=1.11e-16
Wald test            = 76.9  on 7 df,   p=5.97e-14
Score (logrank) test = 100  on 7 df,   p=0
```

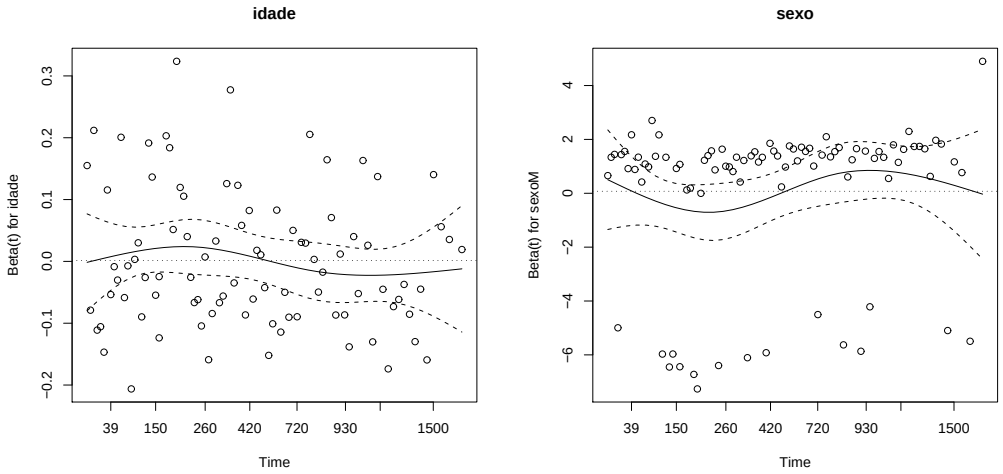
Calculando o resíduo de Schoenfeld e o teste da correlação linear

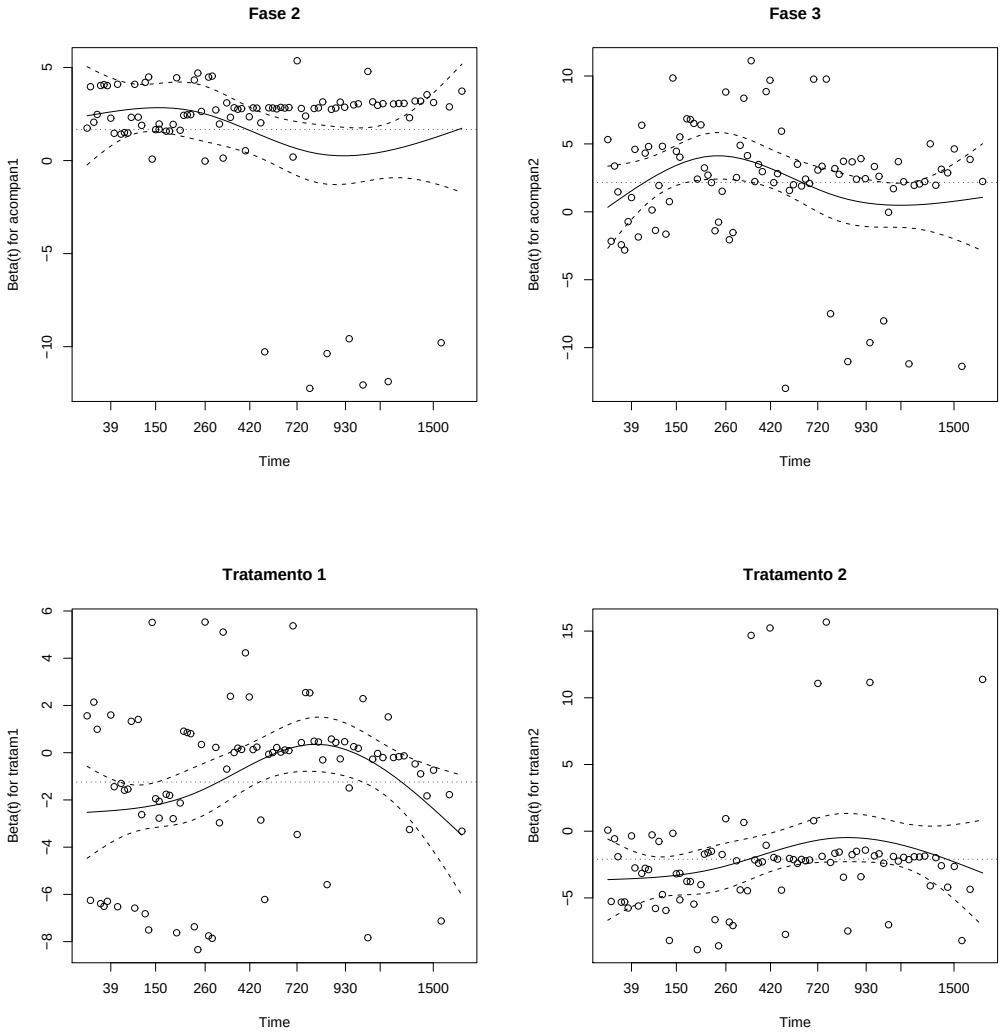
```
> zph <- cox.zph(modeloIII)
> zph


            rho chisq      p
idade    -0.123  1.24 0.2649
sexoM     0.115  1.29 0.2564
```

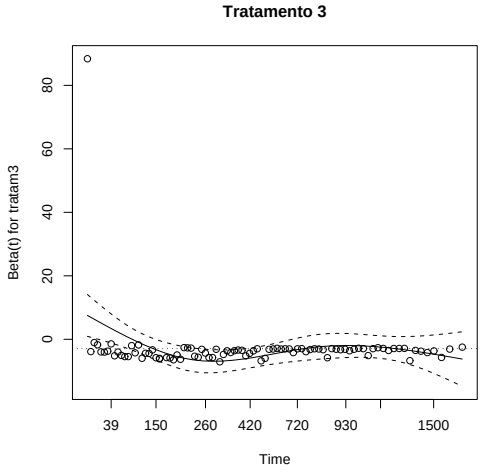
```
acompan1 -0.214  3.97 0.0464
acompan2 -0.153  2.31 0.1288
tratam1   0.153  2.84 0.0920
tratam2   0.177  3.08 0.0794
tratam3  -0.141  1.83 0.1756
GLOBAL       NA 12.17 0.0950
```

Gráficos dos resíduos de Schoenfeld

```
> par(mfrow = c(3, 3))
> plot(zph[1], main = "idade")
> abline(h = modeloIII$coef[1], lty = 3)
> plot(zph[2], main = "sexo")
> abline(h = modeloIII$coef[2], lty = 3)
> plot(zph[3], main = "Fase 2")
> abline(h = modeloIII$coef[3], lty = 3)
> plot(zph[4], main = "Fase 3")
> abline(h = modeloIII$coef[4], lty = 3)
> plot(zph[5], main = "Tratamento 1")
> abline(h = modeloIII$coef[5], lty = 3)
> plot(zph[6], main = "Tratamento 2")
> abline(h = modeloIII$coef[6], lty = 3)
> plot(zph[7], main = "Tratamento 3")
> abline(h = modeloIII$coef[7], lty = 3)
```

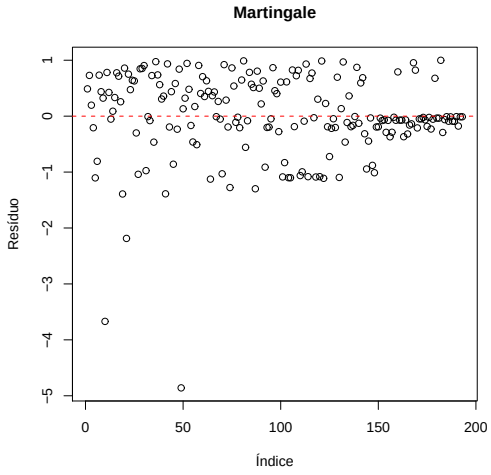



Fase 2
Beta(t) for acompan1
Time
Fase 3
Beta(t) for acompan2
Time
Tratamento 1
Beta(t) for tratam1
Time
Tratamento 2
Beta(t) for tratam2
Time
39
150
260
420
720
930
1500



Resíduos de martingale

```
> mod.mar <- resid(modeloIII, "martingale")
> plot(mod.mar, xlab = "Índice", ylab = "Resíduo", main = "Martingale")
> abline(h = 0, col = 2, lty = 2)
```



Resposta: Observando o teste zph, vê-se que todas as variáveis apresentam correlação baixa isto é, menor do que 0.25. Algumas apresentam $p - valor < 0.05$, mas deve-se tomar cuidado com este indicador devido ao tamanho grande da amostra (n grande pode levar à indicação de valores pequenos da estatística a serem significativamente diferentes de zero. Se este efeito é "biologicamente"diferente de zero, é o que

importa). Quanto ao gráfico dos resíduos de Schoenfeld, nota-se que o tratamento 1 é a variável que mais foge do comportamento linear. As outras, embora contenham variação, estão contidas no intervalo de confiança (linha pontilhada).

9

Covariáveis tempo-dependentes

Exercícios

Recomenda-se que o leitor revisite o conceito de processo de contagem e sua representação gráfica, apresentados no Capítulo 2.

Exercício 9.1: Em um estudo de sobrevida de pacientes infectados pelo HIV, uma covariável importante é o momento em que a contagem de CD4 decresce abaixo de 200. Esse é um critério utilizado para classificar o paciente portador de Aids. A tabela abaixo mostra um trecho do banco de dados que resultaria de um estudo como este, onde `id` é o identificador do paciente e `CD4` é um indicador do nível de CD4 (`CD4=0`, quando a contagem está acima de 200 e, `CD4=1`, quando a contagem está abaixo de 200).

id	inicio	fim	status	CD4
1	25	323	0	0
1	324	768	1	1
2	39	130	0	0
2	131	345	0	1
3	1	56	0	0
3	131	145	1	1

Com base nesses dados, descreva em palavras o que aconteceu com cada paciente ao longo do período de acompanhamento.

Resposta:

- O paciente 1 entrou na coorte no dia 25, no dia 324 seu *CD4* estava abaixo de 200 e no dia 768 ele morreu.
- O paciente 2 entrou na coorte no dia 39, no dia 131 seu *CD4* estava abaixo de 200 e no dia 345 ele deixou de ser acompanhado, ou o estudo terminou.
- O paciente 3 entrou na coorte no dia 1, com *CD4* acima de 200. Entre os dias 56 e 130 ele não foi acompanhado e no dia 131 seu *CD4* estava abaixo de 200. No dia 145 ele morreu.

Exercício 9.2: O tipo de tratamento empregado na terapia de um paciente com Aids pode ser considerado uma covariável tempo-dependente. Considerando 4 tipos possíveis de terapia (0 = sem anti-retrovirais; 1 = monoterapia, 2 = terapia combinada, 3 = terapia potente), poder-se-ia observar o seguinte conjunto de dados:

id	inicio	fim	status	terapia
1	2	124	0	0
1	124	213	0	1
1	213	230	1	3
2	21	210	1	3
3	17	56	0	0
3	56	99	0	1
3	99	154	0	2
3	154	255	1	3

1. Com base nesses dados, descreva em palavras o que aconteceu com cada paciente ao longo do período de acompanhamento.

   Resposta:

   - O paciente 1 entrou na coorte no dia 2, sem nenhum tratamento, no dia 124 ele iniciou a monoterapia, no dia 213 iniciou a terapia potente e no dia 230 veio a falecer.
   - O paciente 2 entrou na coorte no dia 21 já tomando a terapia potente e no dia 210 veio a falecer.
   - O paciente 3 entrou na coorte no dia 17, sem nenhum tratamento, no dia 56 iniciou a monoterapia, no dia 99 iniciou a terapia combinada, no dia 154 iniciou a terapia potente e no dia 255 veio a falecer.

2. Acrescente linhas à tabela acima para incorporar as informações do paciente 4, que entrou na coorte no dia 61, quando passou a ser tratado com a monoterapia.
   No dia 367 do estudo, ele muda para a terapia combinada, no dia 401 muda para a terapia potente, e vem a falecer no dia 460.

   Resposta:

id	inicio	fim	status	terapia
4	61	367	0	1
4	367	401	0	2
4	401	460	1	3

3. Faça o mesmo para o paciente 5, que entrou na coorte no dia 100 e veio a receber seu primeiro tratamento anti-retroviral (monoterapia) no dia 221. Dois meses depois, no dia 281, ele passa para a terapia potente e falece no dia 306.

   Resposta:

id	inicio	fim	status	terapia
5	100	221	0	0
5	221	281	0	1
5	281	306	1	3

4. Por fim, acrescente o paciente 6, que entrou na coorte no dia 47, recebendo monoterapia e passando para terapia combinada no dia 105. Esse paciente saiu da coorte (perda de seguimento) no dia 223.

   Resposta:

id	inicio	fim	status	terapia
6	47	105	0	1
6	105	223	0	2

Exercício 9.3: Abra o arquivo *tmopc.csv* e liste as primeiras 20 linhas. Veja no Apêndice C.5 o significado das variáveis.

```
> tmo <- read.table("tmopc.csv", header = T, sep = ";")
> tmo[1:20, ]
```

	id	sexo	idade	status	inicio	fim	deag	decr	recplaq	fasegr
1	1	2	31	0	0	9	0	0	0	CP1
2	1	2	31	0	9	3527	0	0	1	CP1
3	2	2	38	0	0	28	0	0	0	CP1
4	2	2	38	1	28	39	1	0	0	CP1
5	3	1	23	0	0	27	0	0	0	CP1
6	3	1	23	0	27	36	0	0	1	CP1
7	3	1	23	0	36	268	1	0	1	CP1
8	3	1	23	1	268	434	1	1	1	CP1
9	4	2	5	0	0	24	0	0	0	CP1
10	4	2	5	1	24	69	1	0	0	CP1
11	5	2	15	0	0	22	0	0	0	CP1
12	5	2	15	0	22	83	1	0	0	CP1
13	5	2	15	0	83	446	1	0	1	CP1
14	5	2	15	1	446	672	1	1	1	CP1
15	6	1	23	0	0	22	0	0	0	CP1
16	6	1	23	1	22	98	1	0	0	CP1
17	7	2	11	0	0	32	0	0	0	CP1
18	7	2	11	0	32	4025	0	0	1	CP1
19	8	1	30	0	0	45	0	0	0	Other
20	8	1	30	0	45	278	0	0	1	Other

1. Descreva, em palavras, o que aconteceu com os pacientes 5, 6 e 7.

   Resposta:

   - O paciente 5, entrou no estudo ao ser transplantado, no dia 22 desenvolveu a doença enxerto aguda; no dia 83 teve recuperação de plaquetas; no dia 446 desenvolveu doença enxerto crônica e no dia 672 veio a falecer.
   - O paciente 6, entrou no estudo ao ser transplantado, no dia 22 desenvolveu a doença enxerto aguda e no dia 98 veio a falecer.
   - O paciente 7, entrou no estudo ao ser transplantado, no dia 32 teve recuperação de plaquetas e no dia 4025 quando o estudo terminou ainda estava vivo.

2. Construa as linhas deste banco de dados que correspondem a um paciente que não desenvolveu nenhuma doença-enxerto, recuperou as plaquetas 10 dias após o transplante e saiu da coorte no dia 3.789 (perda de seguimento).

   Resposta:

inicio	fim	status	deag	decr	recplaq
0	10	0	0	0	0
10	3789	0	0	0	1

3. Construa as linhas para um paciente que desenvolveu doença enxerto aguda no dia 20 e faleceu 10 dias depois, sem ter recuperado as plaquetas.

   Resposta:

inicio	fim	status	deag	decr	recplaq
0	20	0	0	0	0
20	30	1	1	0	0

4. Construa também as linhas para o paciente que recuperou as plaquetas no dia 8, mas desenvolveu doença enxerto aguda no dia 31, depois desenvolveu a crônica no dia 210 e faleceu no dia 520.

   Resposta:

inicio	fim	status	deag	decr	recplaq
0	8	0	0	0	0
8	31	0	0	0	1
31	210	0	1	0	1
210	520	1	1	0	1

Exercício 9.4: Ajuste quatro modelos causais para a sobrevida após transplante de medula óssea, usando covariáveis tempo-dependentes (*recplaq*, *deag* e *decr*):

Modelo 1: sobrevida = idade + sexo

Modelo 2: sobrevida = idade + sexo + recplaq

Modelo 3: sobrevida = idade + sexo + recplaq + deag

Modelo 4: sobrevida = idade + sexo + recplaq + deag + decr

```
> tmo <- read.table("tmopc.csv", header = T, sep = ";")
> tmo$sex <- factor(tmo$sexo)
> tmo$recplaq <- factor(tmo$recplaq)
> tmo$deag <- factor(tmo$deag)
> tmo$decr <- factor(tmo$decr)
> tmo.cox1 <- coxph(Surv(inicio, fim, status) ~ idade + sexo, data = tmo)
> summary(tmo.cox1)
```

```
Call:
coxph(formula = Surv(inicio, fim, status) ~ idade + sexo, data = tmo)

  n= 259
        coef exp(coef) se(coef)     z     p
idade -0.0230     0.977   0.0132 -1.74 0.082
sexo  -0.3784     0.685   0.3041 -1.24 0.210

      exp(coef) exp(-coef) lower .95 upper .95
idade     0.977       1.02     0.952      1.00
sexo      0.685       1.46     0.377      1.24

Rsquare= 0.014   (max possible= 0.823 )
Likelihood ratio test= 3.58  on 2 df,   p=0.167
Wald test            = 3.5  on 2 df,   p=0.174
Score (logrank) test = 3.51  on 2 df,   p=0.173

> tmo.cox2 <- coxph(Surv(inicio, fim, status) ~ idade + sexo +
+     recplaq, data = tmo)
> summary(tmo.cox2)

Call:
coxph(formula = Surv(inicio, fim, status) ~ idade + sexo + recplaq,
    data = tmo)

  n= 259
           coef exp(coef) se(coef)      z       p
idade    -0.0171     0.983   0.0135 -1.261 2.1e-01
sexo     -0.2548     0.775   0.3081 -0.827 4.1e-01
recplaq1 -2.2262     0.108   0.4507 -4.940 7.8e-07

         exp(coef) exp(-coef) lower .95 upper .95
idade        0.983       1.02    0.9573     1.010
sexo         0.775       1.29    0.4237     1.418
recplaq1     0.108       9.26    0.0446     0.261

Rsquare= 0.109   (max possible= 0.823 )
Likelihood ratio test= 29.8  on 3 df,   p=1.52e-06
Wald test            = 27.6  on 3 df,   p=4.5e-06
Score (logrank) test = 35.8  on 3 df,   p=8.21e-08

> tmo.cox3 <- coxph(Surv(inicio, fim, status) ~ idade + sexo +
+     recplaq + deag, data = tmo)
> summary(tmo.cox3)

Call:
coxph(formula = Surv(inicio, fim, status) ~ idade + sexo + recplaq +
```

```
    deag, data = tmo)

  n= 259
            coef exp(coef) se(coef)     z       p
idade    -0.0109     0.989   0.0132 -0.82 4.1e-01
sexo     -0.3217     0.725   0.3082 -1.04 3.0e-01
recplaq1 -2.0939     0.123   0.4692 -4.46 8.1e-06
deag1     0.9825     2.671   0.2868  3.43 6.1e-04

         exp(coef) exp(-coef) lower .95 upper .95
idade        0.989      1.011    0.9638     1.015
sexo         0.725      1.379    0.3962     1.326
recplaq1     0.123      8.116    0.0491     0.309
deag1        2.671      0.374    1.5224     4.687

Rsquare= 0.147   (max possible= 0.823 )
Likelihood ratio test= 41.1  on 4 df,   p=2.53e-08
Wald test            = 37.1  on 4 df,   p=1.68e-07
Score (logrank) test = 48.4  on 4 df,   p=7.65e-10

> tmo.cox4 <- coxph(Surv(inicio, fim, status) ~ idade + sexo +
+     recplaq + deag + decr, data = tmo)
> summary(tmo.cox4)

Call:
coxph(formula = Surv(inicio, fim, status) ~ idade + sexo + recplaq +
    deag + decr, data = tmo)

  n= 259
            coef exp(coef) se(coef)      z       p
idade    -0.0122     0.988   0.0134 -0.911 3.6e-01
sexo     -0.3171     0.728   0.3075 -1.031 3.0e-01
recplaq1 -2.0891     0.124   0.4677 -4.467 7.9e-06
deag1     0.9762     2.654   0.2863  3.409 6.5e-04
decr1     0.2213     1.248   0.3903  0.567 5.7e-01

         exp(coef) exp(-coef) lower .95 upper .95
idade        0.988      1.012    0.9622     1.014
sexo         0.728      1.373    0.3986     1.331
recplaq1     0.124      8.078    0.0495     0.310
deag1        2.654      0.377    1.5144     4.653
decr1        1.248      0.801    0.5806     2.681

Rsquare= 0.148   (max possible= 0.823 )
Likelihood ratio test= 41.5  on 5 df,   p=7.6e-08
Wald test            = 37.6  on 5 df,   p=4.51e-07
Score (logrank) test = 48.8  on 5 df,   p=2.44e-09
```

1. Interprete os parâmetros dos modelos.

   Resposta: No primeiro modelo a estimativa do efeito da idade em que realizou o transplante, tem $p < 0,10$ e pode ser considerado significativamente diferente de zero a este nível. Verifica-se que pelo exponencial da estimativa ter valor menor que 1, a sua interpretação é de que a idade tem um efeito protetor e que, por exemplo, uma pessoa que realizou o transplante 10 anos mais velha que a outra tem a cada dia depois do transplante 30% menos chance de morrer.

   No segundo modelo ao incluir a informação a respeito da recuperação de plaquetas, o efeito da idade desaparece. Assim, quando controlado por idade e sexo, a estimativa do efeito da recuperação das plaquetas indica um fortíssimo efeito protetor, diminuindo em 90% o risco do paciente que recuperou as plaquetas de morrer a cada tempo quando comparado a um paciente que não teve esta recuperação.

   No terceiro modelo que inclue a informação do desenvolvimento da doença enxerto aguda, a estimativa do parâmetro referente ao efeito da recuperação da plaqueta continua indicando esta covariável como protetora, mas percebe-se que a estimativa do parâmetro referente ao desenvolvimento da doença enxerto aguda indica que, a cada tempo, este paciente tem quase três vezes mais (2,6) chances de morrer do que os que não desenvolveram esta doença.

   No quarto modelo, a informação a respeito do desenvolvimento da doença enxerto crônica não altera as estimativas dos outros parâmetros de forma substancial, nem o efeito desta covariável é significativamente diferente de um.

2. Avalie a qualidade do ajuste global dos modelos, utilizando a análise de *deviance*

```
> anova(tmo.cox1, tmo.cox2, tmo.cox3, tmo.cox4, test = "Chisq")

Analysis of Deviance Table

Model 1: Surv(inicio, fim, status) ~ idade + sexo
Model 2: Surv(inicio, fim, status) ~ idade + sexo + recplaq
Model 3: Surv(inicio, fim, status) ~ idade + sexo + recplaq + deag
Model 4: Surv(inicio, fim, status) ~ idade + sexo + recplaq + deag + decr
  Resid. Df Resid. Dev  Df Deviance P(>|Chi|)
1       257     444.36
2       256     418.13   1    26.22 3.042e-07
3       255     406.81   1    11.33 7.628e-04
4       254     406.48   1     0.32      0.57
```

Resposta: Ao avaliar a qualidade dos ajustes, a partir da função desvio, verifica-se que a inclusão da variável ***recplaq*** fez decrescer a deviance em 26,22 e este decréscimo é significativo. A inclusão da covariável ***deag*** também melhorou o ajuste de forma significativa, mas a inclusão da variável ***decr*** não alterou a qualidade do ajuste. Assim, o modelo 3 (***tmo.cox3***) seria o escolhido entre estes avaliados.

3. Faça a análise dos resíduos, avaliando os resíduos de Schoenfeld e martingale. O pressuposto de proporcionalidade foi atendido por todas as covariáveis? Há observações mal ajustadas? Nos comandos abaixo, substitua a palavra ***mod*** pelo modelo que você escolheu no item anterior.

   Resposta: Os gráficos dos resíduos foram baseados no modelo escolhido no item anterior que é o modelo 3 (***tmo.cox3***).

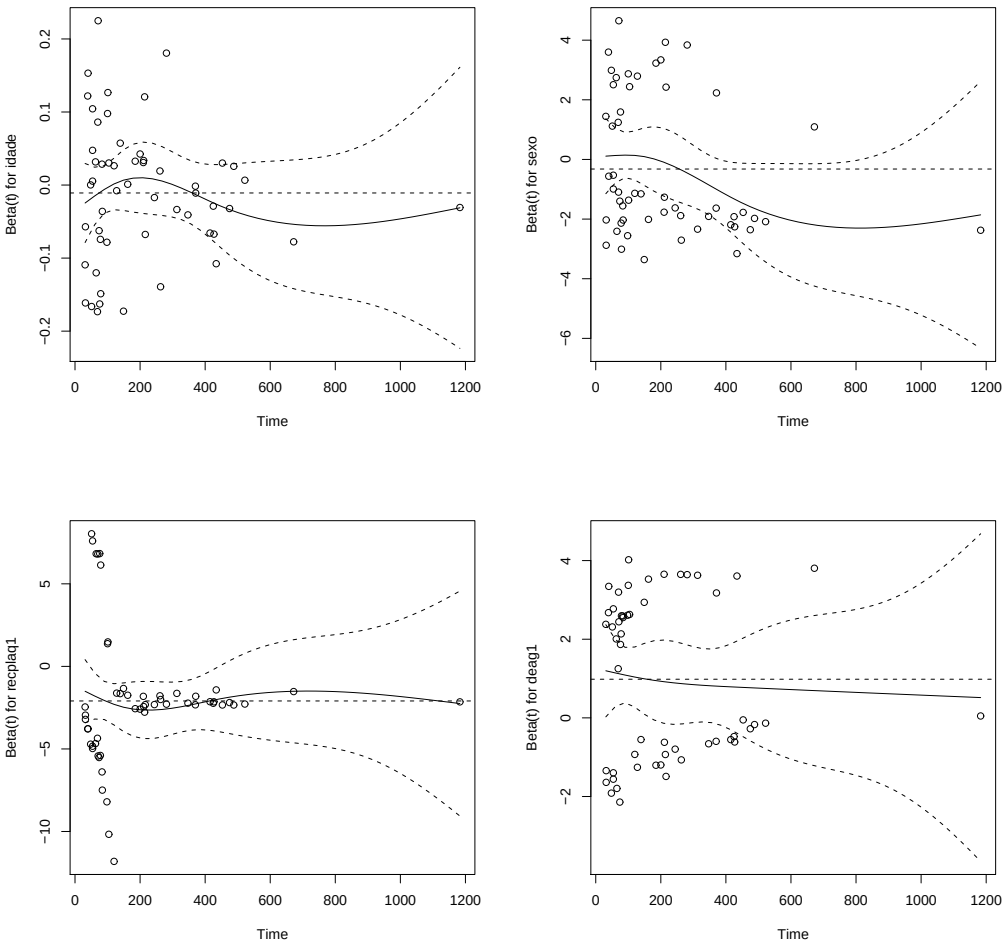
   Resíduos de Schoenfeld

```
> res.sch <- cox.zph(tmo.cox3)
> res.sch

             rho  chisq     p
idade    -0.0655 0.2060 0.650
sexo     -0.2628 3.9107 0.048
recplaq1 -0.0157 0.0178 0.894
deag1    -0.0750 0.2797 0.597
GLOBAL        NA 4.5039 0.342
```

   Os valores do teste indicam bom ajuste global, e somente a variável ***sexo*** seria linearmente correlacionada ao tempo.

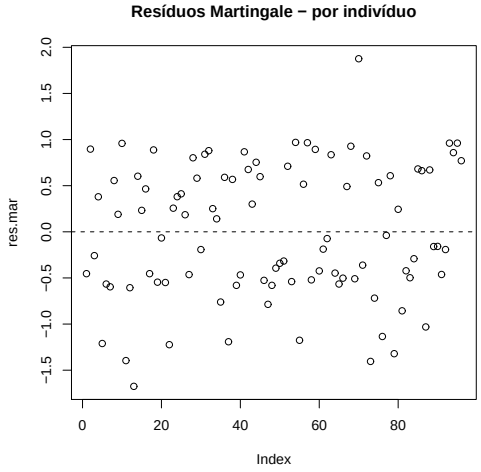
```
> plot(res.sch[1])
> abline(h = tmo.cox3$coef[1], lty = 2)
> plot(res.sch[2])
> abline(h = tmo.cox3$coef[2], lty = 2)
> plot(res.sch[3])
> abline(h = tmo.cox3$coef[3], lty = 2)
> plot(res.sch[4])
> abline(h = tmo.cox3$coef[4], lty = 2)
```

   Graficamente, entretanto, nenhuma das covariáveis apresenta padrão de associação significativo com o tempo (os intervalos de confiança todos incluem o valor do parâmetro). Além disso, a curva visível no gráfico é causada por pouquíssimas observações, e não deve ser valorizada.



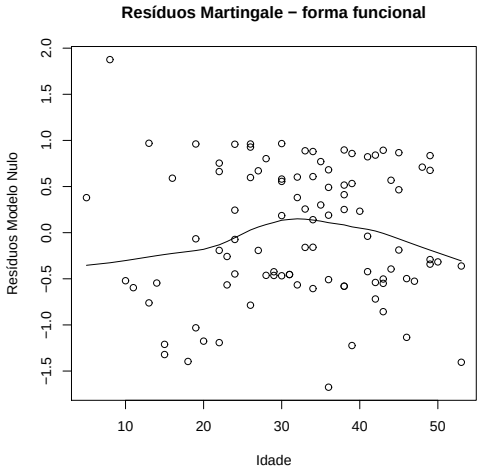
Resíduos de Martingale

```
> res.mar <- resid(tmo.cox3, type = "martingale", collapse = tmo$id)
> plot(res.mar, main = "Resíduos Martingale - por indivíduo")
> abline(h = 0, lty = 2)
```



Pontos homogeneamente distribuidos.

```
> tmo.nulo <- coxph(Surv(inicio, fim, status) ~ 1, data = tmo)
> res.nulo <- resid(tmo.cox3, type = "martingale", collapse = tmo$id)
> idade <- tmo$idade[!duplicated(tmo$id)]
> plot(idade, res.nulo, main = "Resíduos Martingale - forma funcional",
+     xlab = "Idade", ylab = "Resíduos Modelo Nulo")
> lines(lowess(idade, res.nulo))
```



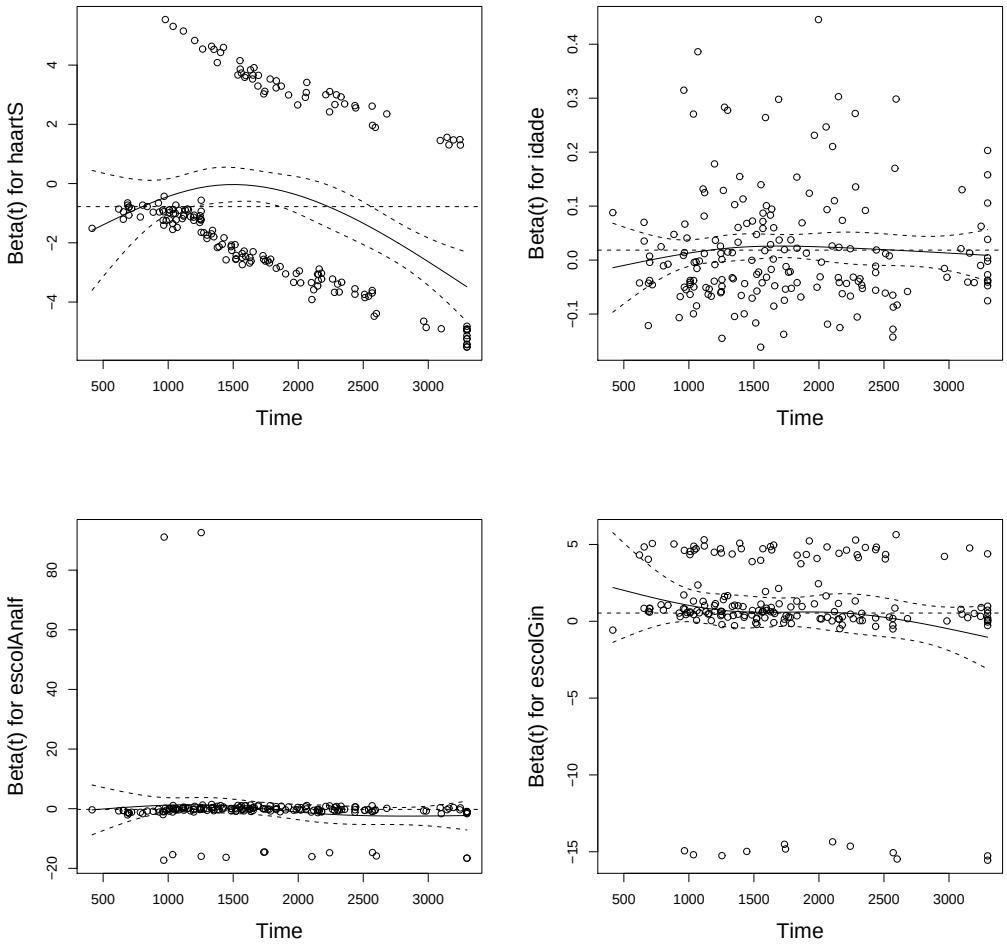
A única variável contínua para a qual pode-se avaliar a forma funcional é a idade. A variação, ainda que presente, é de pouca monta. Se a intensidade fosse maior, pelo formato poderia se investigar uma função quadrática da idade.

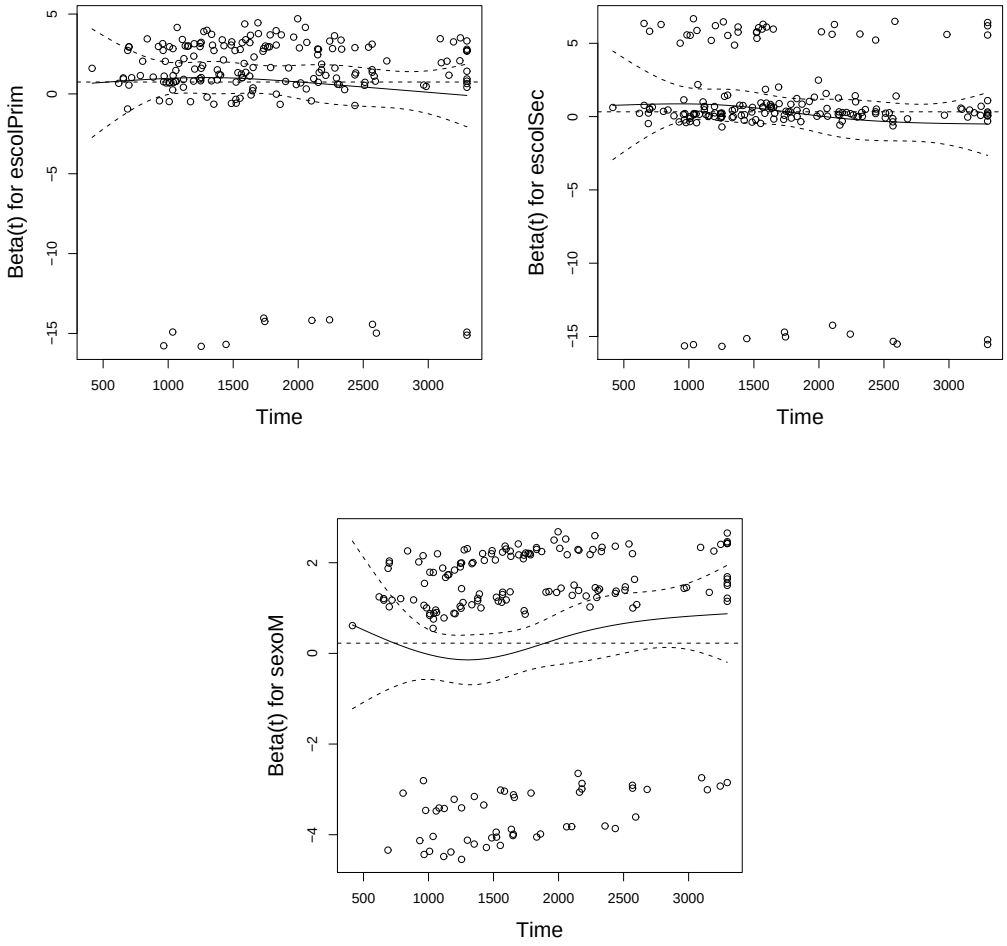
Exercício 9.5: Refaça a análise de sobrevida em Aids apresentada no texto (dados no arquivo *gafcorr.dat*).

Como este exemplo está bem discutido no texto, nos limitaremos aqui a listar comandos e saídas para que o leitor possa comparar com suas saídas.

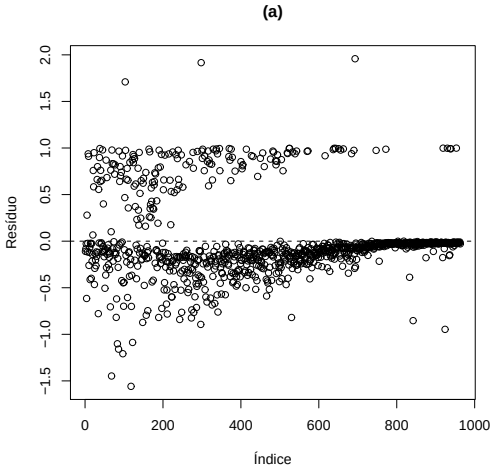
```
> muda <- read.table("gafcorr.dat", header = T)
> names(muda)

[1] "reg"     "haart"   "ini"     "fim"     "sexo"    "escol"   "censura"
[8] "idade"

> muda$escol <- relevel(muda$escol, "Univ")
> muda.cox <- coxph(Surv(ini, fim, censura) ~ haart + idade + escol +
+     sexo, data = muda)
> muda.cox

Call:
coxph(formula = Surv(ini, fim, censura) ~ haart + idade + escol +
    sexo, data = muda)


              coef exp(coef) se(coef)      z       p
haartS     -0.7779     0.459  0.18508 -4.203 2.6e-05
idade       0.0185     1.019  0.00754  2.448 1.4e-02
escolAnalf -0.2342     0.791  0.76547 -0.306 7.6e-01
escolGin    0.5364     1.710  0.32688  1.641 1.0e-01
escolPrim   0.7438     2.104  0.31075  2.394 1.7e-02
escolSec    0.3265     1.386  0.33905  0.963 3.4e-01
sexoM       0.2253     1.253  0.16929  1.331 1.8e-01

Likelihood ratio test=35.1  on 7 df, p=1.08e-05  n= 1377

> muda.sch <- cox.zph(muda.cox)
> muda.sch

                rho    chisq        p
haartS     -0.26583 16.70605 4.36e-05
idade       0.00627  0.00775 9.30e-01
escolAnalf -0.12455  2.86745 9.04e-02
escolGin   -0.12721  3.03844 8.13e-02
escolPrim  -0.07071  0.96321 3.26e-01
escolSec   -0.10421  2.03111 1.54e-01
sexoM       0.12002  2.94786 8.60e-02
GLOBAL           NA 24.41845 9.62e-04
```

```
> plot(muda.sch[1], cex.lab = 1.5)
> abline(h = muda.cox$coefficients[1], lty = 2)
> plot(muda.sch[2], cex.lab = 1.5)
> abline(h = muda.cox$coefficients[2], lty = 2)
> plot(muda.sch[3], cex.lab = 1.5)
> abline(h = muda.cox$coefficients[3], lty = 2)
> plot(muda.sch[4], cex.lab = 1.5)
> abline(h = muda.cox$coefficients[4], lty = 2)
> plot(muda.sch[5], cex.lab = 1.5)
> abline(h = muda.cox$coefficients[5], lty = 2)
> plot(muda.sch[6], cex.lab = 1.5)
> abline(h = muda.cox$coefficients[6], lty = 2)
> plot(muda.sch[7], cex.lab = 1.5)
> abline(h = muda.cox$coefficients[7], lty = 2)
```

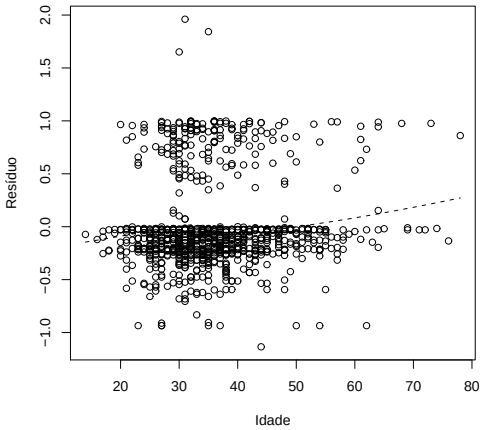





```
> muda.mar <- resid(muda.cox, type = "martingale", collapse = muda$reg)
> plot(muda.mar, xlab = "Índice", ylab = "Resíduo", main = "(a)")
> abline(h = 0, lty = 2)
```

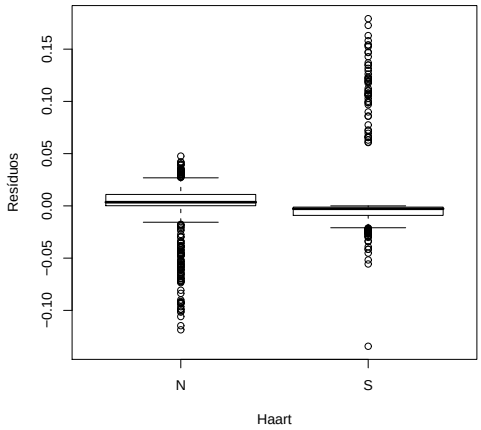


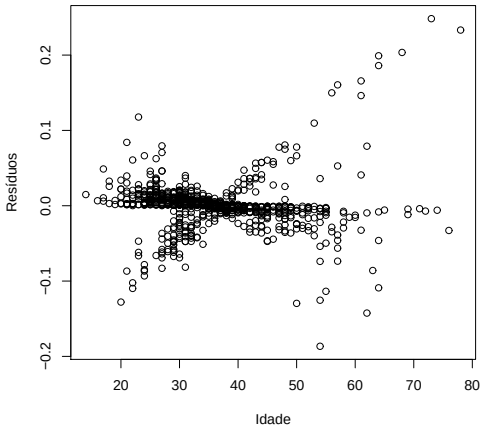

```
> muda.nulo <- coxph(Surv(ini, fim, censura) ~ 1, data = muda)
> res.mar <- resid(muda.nulo, type = "martingale", collapse = muda$reg)
> idade <- muda$idade[!duplicated(muda$reg)]
> plot(idade, res.mar, xlab = "Idade", ylab = "Resíduo")
> lines(lowess(idade, res.mar, iter = 0), lty = 2)
```

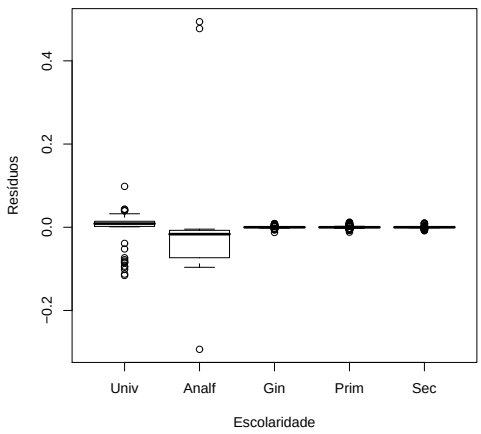


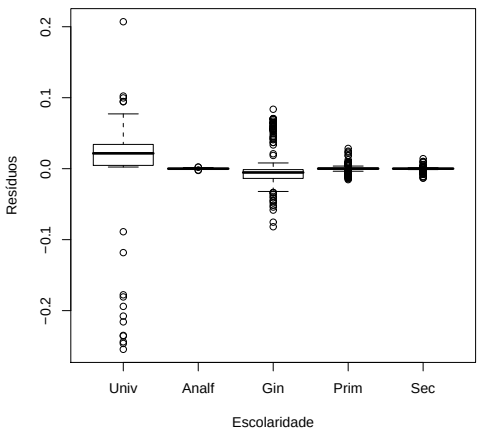

```
> muda.score <- resid(muda.cox, type = "dfbetas", collapse = muda$reg)
> mudaobs.score <- resid(muda.cox, type = "dfbetas")
> indice <- !duplicated(muda$reg)
> plot(muda$haart, mudaobs.score[, 1], xlab = "Haart", ylab = "Resíduos")
> plot(muda$idade[indice], muda.score[, 2], xlab = "Idade", ylab = "Resíduos")
> plot(muda$escol[indice], muda.score[, 3], xlab = "Escolaridade",
```

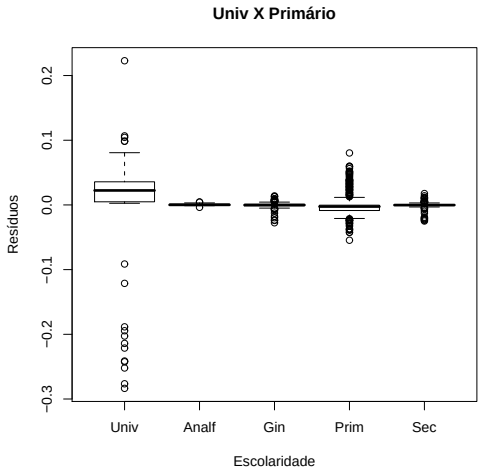
```
+     ylab = "Resíduos", main = "Univ X Analf")
> plot(muda$escol[indice], muda.score[, 4], xlab = "Escolaridade",
+     ylab = "Resíduos", main = "Univ X Ginásio")
> plot(muda$escol[indice], muda.score[, 5], xlab = "Escolaridade",
+     ylab = "Resíduos", main = "Univ X Primário")
> plot(muda$escol[indice], muda.score[, 6], xlab = "Escolaridade",
+     ylab = "Resíduos", main = "Univ X Secundário")
> plot(muda$sexo[indice], muda.score[, 7], xlab = "Sexo", ylab = "Resíduos")
```

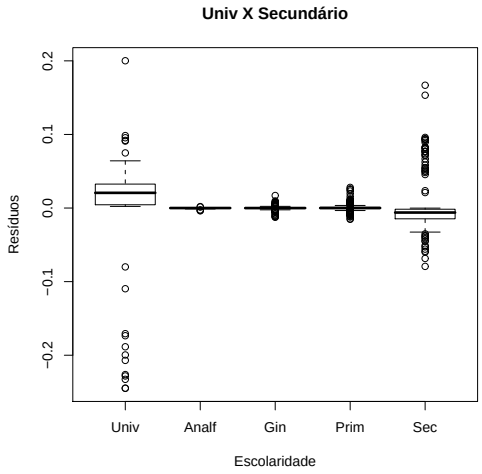


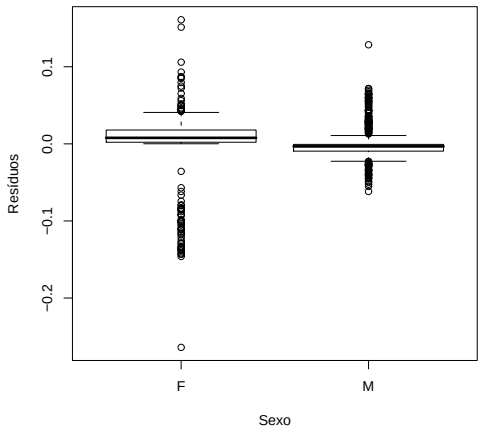







Univ X Primário
Resíduos
0.2
0.1
0.0
−0.1
−0.2
−0.3
Univ
Analf
Gin
Prim
Sec
Escolaridade


Univ X Secundário
Resíduos
0.2
0.1
0.0
−0.1
−0.2
Univ
Analf
Gin
Prim
Sec
Escolaridade


Resíduos
0.1
0.0
−0.1
−0.2
F
M
Sexo

10

Eventos múltiplos

Exercícios

Exercício 10.1: Classifique os eventos múltiplos abaixo (competitivo, paralelo ou ordenado), lembrando que mais de uma interpretação é possível.

1. Estudo de fatores de risco associados à ocorrência de cáries em dentes molares e incisivos.

   Resposta: Evento paralelo – a incidência de cárie acontece para cada dente de forma paralela no tempo

2. Estudo de fatores de risco associados à mortalidade por violência ou Aids em jovens.

   Resposta: Evento competitivo – se morrer de Aids não morre de violência

3. Estudo de fatores de risco associados ao tempo até recaída, para dependentes químicos em tratamento.

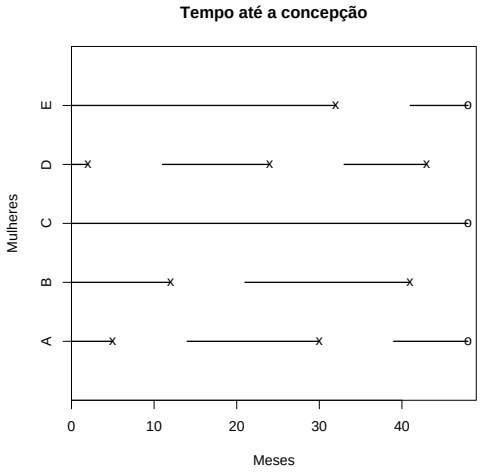
   Resposta: Evento ordenado – qualquer recaída ou recidiva deve necessariamente ser ordenada no tempo

Exercício 10.2: Em um estudo de fertilidade, acompanhou-se uma coorte de mulheres por 48 meses para medir o tempo desde o parto até a próxima concepção, junto com covariáveis potencialmente associadas ao risco de engravidar. Para 5 mulheres, as seguintes datas de concepção foram registradas:

Mulher	Datas	Estado civil	contraceptivo
1	meses 5, 30	casada	Não
2	meses 12 e 41	casada	Não
3	nenhum evento	solteira	Sim
4	meses 2, 24,43	casada	Não
5	mês 32	solteira	Sim

Considerando que cada mulher retorna ao grupo sob risco de gravidez exatamente 9 meses após a concepção:

1. Faça, à mão, um gráfico das trajetórias das cinco mulheres. Use o gráfico referente aos dados de reincidência de diarréia como exemplo.



2. Concepção é um evento múltiplo ordenado. Neste capítulo, foram apresentados três modelos marginais para eventos ordenados. Discuta os pressupostos que você estaria assumindo ao aplicar cada um deles ao estimar o efeito das covariáveis estado civil e uso de contraceptivo no risco de engravidar.

   Resposta: Eventos ordenados independentes – o risco de engravidar seria constante, após cada gravidez se retornaria ao grupo de risco. Este modelo é possível em população de ratas, durante a fase fértil. Em populações humanas, no século XXI e em áreas urbanas, entretanto, o risco de engravidar diminui substancialmente após uma ou duas gestações. Ou seja, a linha de base do risco tende a variar.

Eventos ordenados concomitantes – não é possível no contexto da gravidez. O tempo para cada gravidez não pode ser contado a partir da menarca, por exemplo.

Eventos ordenados com risco condicional – a cada gravidez a linha de base do risco muda. Ou seja, a ordem das gestações é importante na definição da linha de base, e por isso este modelo se chama "condicional". É o mais adequado ao estudo de concepção em populações humanas.

3. Organize uma planilha para esses dados, de forma que se possa ajustar um modelo de incrementos independentes (AG).

Mulher	Início	Fim	Status
A	0	5	1
A	14	30	1
A	39	48	0
B	0	12	1
B	21	41	1
C	0	48	0
D	0	2	1
D	11	24	1
D	33	43	1
E	0	32	1
E	41	48	0

Observe que entre o final de um período de risco e o início de outro decorrem 9 meses da gestação.

4. Organize uma planilha para esses dados, de forma que se possa ajustar um modelo marginal (WLW).

Resposta: Neste modelo devemos assumir uma número máximo de gestações que serão analisadas, para construir o banco de acordo. Suponhamos 3 gestações possíveis no período de 48 meses observado.

Mulher	Início	Fim	Status	Enum
A	0	5	1	1
A	0	30	1	2
A	0	48	0	3
B	0	12	1	1
B	0	41	1	2
B	0	48	0	3
C	0	48	0	1
C	0	48	0	2
C	0	48	0	3
D	0	2	1	1
D	0	24	1	2
D	0	43	1	3
E	0	32	1	1
E	0	48	0	2
E	0	48	0	3

5. Organize uma planilha para esses dados, de forma que se possa ajustar um modelo condicional (PWP).

Mulher	Início	Fim	Status	Enum
A	0	5	1	1
A	14	30	1	2
A	39	48	0	3
B	0	12	1	1
B	21	41	1	2
C	0	48	0	1
D	0	2	1	1
D	11	24	1	2
D	33	43	1	3
E	0	32	1	1
E	41	48	0	2

Exercício 10.3: Abra o arquivo *compete.dat* (detalhes sobre esses dados no Apêndice C.3. Ele contém os dados referentes aos três desfechos competitivos para diálise discutidos no texto: transplante, óbito por causa renal, óbito por outra causa. Liste as primeiras 25 linhas.

```
> compete <- read.table("compete.dat", header = T)
> compete[1:25, ]
```

	id	idade	doenca	status	motivo	tempo	endpoint
1	1	50	outr	1	obrenal	27	1
2	1	50	outr	0	obrenal	27	2
3	1	50	outr	0	obrenal	27	3
4	2	51	hiper	0	censura	28	1
5	2	51	hiper	0	censura	28	2
6	2	51	hiper	0	censura	28	3
7	3	30	diab	0	transpl	18	1
8	3	30	diab	0	transpl	18	2
9	3	30	diab	1	transpl	18	3
10	4	33	diab	0	oboutra	14	1
11	4	33	diab	1	oboutra	14	2
12	4	33	diab	0	oboutra	14	3
13	5	35	diab	1	obrenal	1	1
14	5	35	diab	0	obrenal	1	2
15	5	35	diab	0	obrenal	1	3
16	6	34	hiper	0	censura	43	1
17	6	34	hiper	0	censura	43	2
18	6	34	hiper	0	censura	43	3
19	7	42	outr	0	censura	5	1
20	7	42	outr	0	censura	5	2
21	7	42	outr	0	censura	5	3
22	8	30	outr	0	censura	43	1
23	8	30	outr	0	censura	43	2
24	8	30	outr	0	censura	43	3
25	9	27	rim	0	censura	43	1

1. Descreva o que aconteceu com os pacientes 4, 5, 6 e 8.

   Resposta:

   - O paciente 4, que entrou em insuficiência renal, causada por diabetes aos 33 anos, faleceu no 14^o mês de hemodiálise de causa não relacionada à insufuciência renal.
   - O paciente 5, que entrou em insuficiência renal aos 35 anos, causada por diabetes, faleceu no 1^o mês de hemodiálise de causa relacionada à insufuciência renal.
   - O paciente 6, que entrou em insuficiência renal por causa não especificada aos 34 anos, estava vivo no 43^o mês de acompanhamento.
   - O paciente 8, que entrou em insuficiência renal por causa não especificada aos 30 anos, estava vivo no 43^o mês de acompanhamento.

2. O que caracteriza um banco de dados como sendo elaborado para análise de eventos competitivos?

Resposta:

- Cada paciente terá tantas linhas quantos eventos diferentes analisados.
- Apenas um evento pode ser não censurado (status=1)
- Os demais eventos serão censurados

Exercício 10.4: Podemos dizer que uma criança ao nascer está em risco para dois eventos competitivos, com relação a uma doença imunoprevenível: adquirir imunidade induzida pela vacina ou adquirir imunidade induzida pela doença. Ambos os desfechos, assumindo imunidade duradoura, retiram a criança do grupo sob risco de adquirir imunidade (um terceiro desfecho que tiraria a criança do grupo sob risco seria óbito pela doença ou outra causa). Suponha que 1.000 crianças tenham sido acompanhadas durante o primeiro ano de vida e as datas da vacinação ou da infecção tenham sido registradas. A tabela a seguir mostra os dados das 4 primeiras crianças.

criança	idade(meses)	evento	sexo	prénatal
1	5	infecção	M	Não
2	6	vacina	M	Sim
3	4.5	vacina	F	Sim
4	12	vacina	F	Não

1. Como seriam organizados os dados para ajustar um modelo de riscos competitivos a esses dados?

   Resposta: Cada criança teria duas linhas, uma para cada evento competitivo, suponhamos a primeira linha a vacina e a segunda a infecção. O status seria zero para censura pela causa. O tempo seria a idade da criança.

criança	idade(meses)	evento	sexo	prenatal	status
1	5	infecção	M	Não	0
1	5	infecção	M	Não	1
2	6	vacina	M	Sim	1
2	6	vacina	M	Sim	0
3	4,5	vacina	F	Sim	1
3	4,5	vacina	F	Sim	0
4	12	vacina	F	Não	1
4	12	vacina	F	Não	0

2. Descreva o modelo a ser ajustado no R para estimar o efeito da covariável pré-natal no risco de adquirir imunidade, por ambas as causas.

   Resposta: *coxph(Surv(idade,status) ∼ prenatal+ cluster(criança)+strata(evento),data=dados*

3. Descreva o modelo a ser ajustado no R para estimar o efeito da covariável pré-natal sobre o risco de adquirir imunidade, separadamente para cada desfecho.

   Resposta:

   - Para estimar o efeito do prénatal em adquirir imunidade pela vacina: *coxph(Surv(idade,status) ∼ prénatal+ cluster(criança)+strata(evento), data=dados, subset=(evento==vacina)*
   - Para estimar o efeito do prénatal em adquirir a infecção: *coxph(Surv(idade,status) ∼ prénatal+ cluster(criança)+strata(evento), data=dados, subset=(evento==infecção)*

Exercício 10.5: No texto, vimos o ajuste de dois modelos aos dados de recorrência de diarréia (AG e condicional). Que outros modelos conceituais poderiam ser considerados para esse problema?

Resposta: Não poderia ser eventos paralelos nem competitivos. O modelo WLW poderia ser considerado, supondo, por exemplo, que após a dose de Vitamina A o tempo até cada episódio de diarréia se superpõe. Mas não é muito realista. Um modelo de fragilidade será ajustado no próximo capítulo. Fora do campo da análise de sobrevida, poderia ser utilizado um modelo de regressão de Poisson. Outra abordagem, possivelmente mais adequados à modelagem deste evento, são os modelos para medidas repetidas, marginais ou condicionais a efeitos aleatórios.

Exercício 10.6: Um exemplo de evento ordenado é a reinternação hospitalar. Se o risco de reinternar for independente do fato de o indivíduo ter sofrido uma internação anterior, então o modelo de incrementos independentes é apropriado. Se esse pressuposto não for aceitável, então o modelo condicional é mais apropriado. O arquivo *reint.dat* contém os dados de reinternação de um hospital (recorra ao Apêndice C.8 para detalhes sobre estes dados).

1. Ajuste um modelo de incrementos independentes, com as covariáveis sexo e idade. Que pressuposto está sendo assumido ao se utilizar esse modelo? Compare as estimativas de variância. Como você interpreta esse modelo?

```
> reint <- read.table("reint.dat", header = T, sep = ";")
> names(reint)

 [1] "id"      "spec"    "sexo"    "diaperm" "idade"   "numint"  "obitpos"
 [8] "enum"    "status"  "tempo"   "grupos"  "ini"     "fim"     "ini0"
[15] "fim0"

> reint[1:10, ]

   id spec sexo diaperm idade numint obitpos enum status tempo   grupos  ini
1   1    3    1       7    52      4       0    2      1    41 CAPVIIIC   54
2   1    3    1       5    52      4       0    3      1    26 CAPVIIIC  100
3   1    3    1      51    52      4       0    4      0  1252 CAPVIIIC  177
4   1    3    1      26    52      4       0    1      1    21 CAPVIIIC   26
5   2    1    3       1    29      3       0    3      0  1101  CAPVIIB  328
6   2    3    3       1    29      3       0    1      1   128  CAPVIIB    1
7   3    2    3       1    29      3       0    2      1   197    CAPXI  130
8   4    3    3       9    75      1       0    1      0  1420   CAPIXE    9
9   4    3    3      13    56      4       0    3      1   922  CAPVIIA  397
10  4    3    3       2    56      4       0    4      0   108  CAPVIIA 1321
    fim ini0 fim0
1    95   28   69
2   126   74  100
3  1429  151 1403
4    47    0   21
5  1429  327 1428
6   129    0  128
7   327  129  326
8  1429    0 1420
9  1319  391 1313
10 1429 1315 1423
```

É necessário fazer da variável sexo um fator, pois os valores são 1 e 3, se tratados como numérico estimarão valores errados.

```
> reint$sexo <- factor(reint$sexo, labels = c("masc", "fem"))
> require(survival)

[1] TRUE

> reint.ag <- coxph(Surv(ini, fim, status) ~ idade + sexo + cluster(id),
+     data = reint)
> summary(reint.ag)
```

```
Call:
coxph(formula = Surv(ini, fim, status) ~ idade + sexo + cluster(id),
    data = reint)

  n=21415 (1 observations deleted due to missing)
           coef exp(coef) se(coef) robust se    z       p
idade   0.00817      1.01 0.000727  0.000891 9.17 0.0e+00
sexofem 0.21612      1.24 0.029400  0.038287 5.64 1.7e-08

        exp(coef) exp(-coef) lower .95 upper .95
idade        1.01      0.992      1.01      1.01
sexofem      1.24      0.806      1.15      1.34

Rsquare= 0.008   (max possible= 0.981 )
Likelihood ratio test= 177  on 2 df,   p=0
Wald test            = 132  on 2 df,   p=0
Score (logrank) test = 181  on 2 df,   p=0,   Robust = 125  p=0

  (Note: the likelihood ratio and score tests assume independence of
     observations within a cluster, the Wald and robust score tests do not).
```

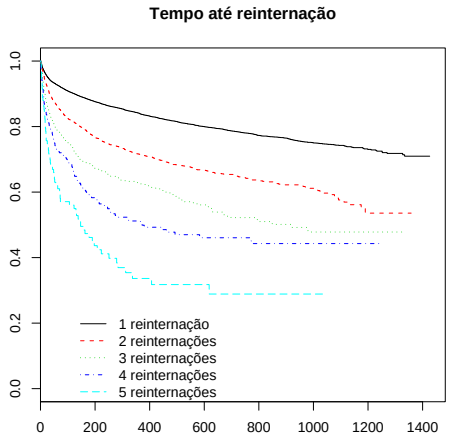
Resposta: Estamos assumindo que as reinternações sucessivas têm o mesmo risco de base. Ao sair de uma internação o paciente está de volta ao estado inicial. As estimativas de variância são:

- para idade: se(coef)=0.000727 e robust se=0.000891, ou seja, a variância robusta é apenas 22% maior.
- para sexo: se(coef)=0.014700 e robust se=0.019143, ou seja, a variância robusta é apenas 30% maior.
- O modelo é significativamente diferente do modelo nulo –`Wald test` e `Score Robust` com p=0. O teste da razão de verossimilhança (`Likelihood ratio test`) e `Score (logrank) test` não deveria ser usado, pois assume independência das internações no mesmo indivíduo.
- O modelo explica apenas 0.008 de um máximo de 0.981, ou seja, 0,82% da variabilidade total.
- A variável idade é significativa (p = 0.000) e a cada ano a mais de idade o risco de internar aumenta em 1%.
- O risco das mulheres se internarem é 24% maior que dos homens.

2. Você sugeriria outro modelo mais apropriado para descrever o risco de nova reinternação?

Resposta: O modelo condicional, onde a linha de base do risco varia conforme as internações. A melhor forma de explorar esta possibilidade é verificando a estimativa não-paramétrica da sobrevida por ordem da internação.

```
> plot(survfit(Surv(tempo, status) ~ enum, data = reint, subset = (enum <
+     6)), col = 1:5, lty = 1:5, legend.text = c("1 reinternação",
+     "2 reinternações", "3 reinternações", "4 reinternações",
+     "5 reinternações", ), legend.pos = c(100, 0.25), mark.time = F,
+     main = "Tempo até reinternação")
```



Observe que no gráfico é visível a diferença no tempo até cada reinternação, diminuindo à medida em que aumenta o número.

3. Ajuste um modelo condicional aos dados e interprete-o.

```
> reint.pwp <- coxph(Surv(ini, fim, status) ~ idade + sexo + cluster(id) +
+     strata(enum), data = reint)
> summary(reint.pwp)

Call:
coxph(formula = Surv(ini, fim, status) ~ idade + sexo + cluster(id) +
    strata(enum), data = reint)

  n=21415 (1 observations deleted due to missing)
           coef exp(coef) se(coef) robust se     z       p
idade   0.0081      1.01 0.000731    0.0008 10.13 0.0e+00
sexofem 0.1488      1.16 0.029736    0.0335  4.44 8.9e-06

        exp(coef) exp(-coef) lower .95 upper .95
```

```
idade        1.01      0.992      1.01      1.01
sexofem      1.16      0.862      1.09      1.24

Rsquare= 0.007   (max possible= 0.969 )
Likelihood ratio test= 141  on 2 df,   p=0
Wald test            = 126  on 2 df,   p=0
Score (logrank) test = 144  on 2 df,   p=0,   Robust = 119  p=0

  (Note: the likelihood ratio and score tests assume independence of
     observations within a cluster, the Wald and robust score tests do not).
```

O modelo explica 0,72% da variabilidade total da sobrevida, sendo melhor que o modelo nulo (p-valor do Wald teste = 0). As duas variáveis são significativas: para cada ano a mais de idade o risco de reinternar aumenta em 1%, e as mulheres internam 16% mais que os homens.

Exercício 10.7: Um pesquisador está estudando fatores de risco associados à ocorrência de algumas doenças oportunistas em pacientes com Aids. Para isso, ele acompanha uma coorte de pacientes e registra a data do diagnóstico da Aids (que define o início do período de acompanhamento) e as datas de ocorrência das três doenças de interesse: alterações hematológicas (hemato), herpes e candidíase.

1. Como você classifica esse problema? É um caso de eventos competitivos, paralelos ou ordenados? Por quê?

   Resposta: São eventos não competitivos, pois o paciente pode ter mais de uma doença. Podem ser tratados como eventos paralelos não ordenados ou ordenados do tipo WLW.

2. Abra, no R, o arquivo *oport.dat* e liste as primeiras 10 linhas. Veja, no Apêndice C.8, a descrição das variáveis e o comando para abrir o arquivo.

```
> oport <- read.table("oport.dat", header = T)
> oport[1:10, ]
```

	reg	sex	esc	idade	udi	sexual	ini	oport	fim	status	tempo
1	1	F	1	45	0	0	3127	Candida	3141	1	14
2	1	F	1	45	0	0	3127	Caquex	3141	1	14
3	2	M	2	38	1	0	2884	Candida	3343	1	459
4	2	M	2	38	1	0	2884	Diarreia	4274	1	1390
5	2	M	2	38	1	0	2884	Herpes	3133	1	249
6	2	M	2	38	1	0	2884	Pneumocist	3133	1	249
7	3	F	4	34	1	0	3330	Hemato	3332	1	2
8	4	M	3	43	0	0	3780	Hemato	3964	1	184
9	5	F	3	32	0	1	2923	Candida	3588	1	665
10	5	F	3	32	0	1	2923	Caquex	3987	1	1064

3. Descreva o que aconteceu com os pacientes 1 a 4. Observe a variável *status*, cujo valor é sempre 1. O que isso quer dizer em termos da composição do grupo sob risco?

   Resposta:

   - Paciente 1 – sexo feminino, escolaridade menor do que 4 anos, com 45 anos, não é usuária de droga injetável (UDI) e o comportamento sexual não foi classificado como de risco, 14 dias após o diagnóstico de Aids apresentou candidíase e caquexia.
   - Paciente 2 – sexo masculino, ensino fundamental, 38 anos, UDI, 249 dias após o diagnóstico de Aids apresentou Herpes e pneumocistose, 459 dias após, candidíase e 1390 dias, diarréia.
   - Paciente 3 – sexo feminino, curso superior, 34 anos, UDI, 2 dias após o diagnóstico de Aids apresentou alterações hematólogicas.
   - Paciente 4 – sexo masculino, ensino médio, 43 anos, comportamento sexual de risco, 184 dias após o diagnóstico de Aids apresentou alterações hematólogicas.

   O grupo de risco são todos os pacientes em observação: ninguém sai do risco porque apresentou qualquer doença oportunista. Só se sai do risco por óbito ou perda de acompanhamento.

4. Ajuste um modelo para cada desfecho (hemato, herpes e candidíase) separadamente (use o comando *subset*, como descrito na seção 10.3.2 e avalie o efeito das covariáveis sexo, idade e comportamento de risco (*udi* e *sexual*) como fatores de risco.

```
> hem.fit <- coxph(Surv(tempo, status) ~ udi + sexual + sex + factor(esc),
+     data = oport, subset = (oport == "Hemato"))
> her.fit <- coxph(Surv(tempo, status) ~ udi + sexual + sex + factor(esc),
+     data = oport, subset = (oport == "Herpes"))
> can.fit <- coxph(Surv(tempo, status) ~ udi + sexual + sex + factor(esc),
+     data = oport, subset = (oport == "Candida"))
> summary(hem.fit)

Call:
coxph(formula = Surv(tempo, status) ~ udi + sexual + sex + factor(esc),
    data = oport, subset = (oport == "Hemato"))

  n=442 (52 observations deleted due to missing)
              coef exp(coef) se(coef)      z      p
udi          -0.220     0.803    0.209 -1.051 0.2900
```

```
sexual       -0.312     0.732    0.174 -1.794 0.0730
sexM          0.312     1.366    0.107  2.921 0.0035
factor(esc)1 -0.314     0.730    0.420 -0.748 0.4500
factor(esc)2 -0.309     0.734    0.426 -0.725 0.4700
factor(esc)3 -0.340     0.712    0.428 -0.796 0.4300
factor(esc)4 -0.674     0.510    0.448 -1.506 0.1300

             exp(coef) exp(-coef) lower .95 upper .95
udi              0.803      1.246     0.533      1.21
sexual           0.732      1.366     0.520      1.03
sexM             1.366      0.732     1.108      1.68
factor(esc)1     0.730      1.369     0.320      1.66
factor(esc)2     0.734      1.362     0.318      1.69
factor(esc)3     0.712      1.405     0.308      1.65
factor(esc)4     0.510      1.962     0.212      1.23

Rsquare= 0.043   (max possible= 1 )
Likelihood ratio test= 19.3  on 7 df,   p=0.00741
Wald test            = 18.3  on 7 df,   p=0.0107
Score (logrank) test = 18.4  on 7 df,   p=0.0102

> summary(her.fit)

Call:
coxph(formula = Surv(tempo, status) ~ udi + sexual + sex + factor(esc),
    data = oport, subset = (oport == "Herpes"))

  n=108 (11 observations deleted due to missing)
               coef exp(coef) se(coef)      z    p
udi          -0.818     0.441    0.482 -1.698 0.09
sexual       -0.139     0.870    0.496 -0.281 0.78
sexM         -0.301     0.740    0.233 -1.292 0.20
factor(esc)1 -0.669     0.512    1.039 -0.643 0.52
factor(esc)2 -0.569     0.566    1.024 -0.555 0.58
factor(esc)3 -0.559     0.572    1.023 -0.547 0.58
factor(esc)4 -0.722     0.486    1.066 -0.677 0.50

             exp(coef) exp(-coef) lower .95 upper .95
udi              0.441       2.27    0.1718      1.13
sexual           0.870       1.15    0.3293      2.30
sexM             0.740       1.35    0.4683      1.17
factor(esc)1     0.512       1.95    0.0668      3.93
factor(esc)2     0.566       1.77    0.0760      4.22
factor(esc)3     0.572       1.75    0.0770      4.25
factor(esc)4     0.486       2.06    0.0602      3.92

Rsquare= 0.059   (max possible= 0.999 )
```

```
Likelihood ratio test= 6.57  on 7 df,   p=0.475
Wald test            = 5.8  on 7 df,   p=0.564
Score (logrank) test = 6.01  on 7 df,   p=0.538

> summary(can.fit)

Call:
coxph(formula = Surv(tempo, status) ~ udi + sexual + sex + factor(esc),
    data = oport, subset = (oport == "Candida"))

  n=352 (39 observations deleted due to missing)
                coef exp(coef) se(coef)      z    p
udi          -0.0635     0.938    0.214 -0.297 0.77
sexual       -0.2622     0.769    0.187 -1.402 0.16
sexM          0.0630     1.065    0.118  0.534 0.59
factor(esc)1 -0.3683     0.692    0.589 -0.626 0.53
factor(esc)2 -0.3735     0.688    0.593 -0.629 0.53
factor(esc)3 -0.5685     0.566    0.595 -0.955 0.34
factor(esc)4 -0.5726     0.564    0.611 -0.937 0.35

             exp(coef) exp(-coef) lower .95 upper .95
udi              0.938      1.066     0.617      1.43
sexual           0.769      1.300     0.533      1.11
sexM             1.065      0.939     0.845      1.34
factor(esc)1     0.692      1.445     0.218      2.19
factor(esc)2     0.688      1.453     0.215      2.20
factor(esc)3     0.566      1.766     0.176      1.82
factor(esc)4     0.564      1.773     0.170      1.87

Rsquare= 0.017   (max possible= 1 )
Likelihood ratio test= 5.88  on 7 df,   p=0.553
Wald test            = 5.73  on 7 df,   p=0.571
Score (logrank) test = 5.77  on 7 df,   p=0.567
```

Para a ocorrência de complicações hematológicas o único fator de risco sigificativo é ser do sexo masculino. O modelo é significativamente diferente do modelo nulo, e explica apenas 4% da variância. Em relação ao Herpes e a candidíase, nenhuma das variáveis é significativa, e o modelo não difere do modelo nulo.

5. Compare o modelo acima com uma abordagem alternativa, onde o efeito das covariáveis é modelado simultaneamente para os três desfechos, porém permitindo um risco basal diferente para cada uma (eventos paralelos).

```
> todas.fit <- coxph(Surv(tempo, status) ~ udi + sexual + sex +
+     factor(esc) + cluster(reg) + strata(oport), data = oport,
```

```
+     subset = (oport == "Hemato" || oport == "Herpes" || oport ==
+         "Candida"))
```

O modelo agora é significativamente diferente do modelo nulo pelo teste de Wald, mas limítrofe segundo o teste score robusto (p=0.0574). Nenhuma das variáveis é signficativa.

6. Uma terceira opção de modelagem seria considerar a ordem em que ocorrem as doenças oportunistas e estudar o efeito de covariáveis no risco de desenvolver a primeira, segunda, terceira etc. Como você organizaria o banco de dados *oport.dat* para aplicar um modelo ordenado do tipo WLW?

   Resposta:

   (a) Cada paciente teria 3 linhas (para analisar as três doenças), ordenadas conforme o aparecimento.

   (b) Criação de uma variável de enumeração (1,2,3) para cada linha do paciente, respeitando a sequência do aparecimento das doenças.

   (c) O paciente que não teve a doença nesta linha teria zero na variável *status*.

Exercício 10.8: O problema de reações adversas ao uso de medicamentos apresentado no item 10.2.3 pode ser analisado de outras formas. Sugira duas outras abordagens possíveis e descreva como ficariam:

1. contagem do tempo: início, fim e tempo total;
2. censura;
3. repetição ou não de linhas;
4. grupo de risco;
5. risco de base;
6. modelo no R.

Resposta: O aparecimento de efeitos colaterais de medicamentos poderia ser analisado como evento paralelo, sem ordenação. O aparecimento de efeitos colaterias de medicamentos não pode ser tratado como eventos recorrentes, pois não poderiam acontecer simultaneamente, uma vez que estes são ordenados pelo início e fim de

cada período entre duas intercorrências. Poderiamos modelar simplesmente apenas o tempo até o primeiro efeito colateral, ou o tempo até cada ocorrência de interesse separadamente.

No caso de modelo para **eventos paralelos**, a contagem do tempo começa sempre em zero, ou seja, basta colocar o tempo entre o uso do medicamento e o aparecimento do efeito colateral. Censurados serão somente os pacientes que não apresentam qualquer evento durante o tempo de observação. Cada paciente terá tantas linhas quantos efeitos colaterais forem registrados, sem repetição de linhas. O grupo de risco é sempre todos os que usaram o medicamento, exceto as perdas. No R fica assim:

```
ajuste <- coxph(Surv(tempo,status) covariaveis+
cluster(id)+strata(efeito),data=dado)
```

Para o modelo simples de Cox, cada paciente tem apenas uma linha, com o primeiro evento. Ou então, analisando separadamente cada efeito colateral, seleciona-se no banco o evento de interesse, através do comando *subset*.

11

Fragilidade

Exercícios

Exercício 11.1: Discuta, para cada uma das situações abaixo, por que utilizar o modelo com fragilidade.

1. Em um estudo de reinternação, em que se procura estimar o efeito de covariáveis associadas ao hospital (tamanho, especialidade clínica) no risco de ocorrer reinternação.

   Resposta: Em se tratando de diversos hospitais, a incorporação de efeitos aleatórios ao modelo visa dar conta da estrutura de dependência gerada pelo risco comum dos pacientes do mesmo serviço, e para permitir a estimação das variáveis (tamanho, especialidade clínica) neste nível. A inclusão dos hospitais como uma variável *dummy*, uma para cada hospital, permite estimar o risco do hospital, mas não permite estimar simultaneamente o efeito do hospital e o efeito das covariáveis relacionadas ao serviço.

2. Em um estudo sobre efeito do tratamento na reincidência de doenças oportunistas em pacientes com Aids, no qual medidas repetidas são obtidas para cada indivíduo.

   Resposta: Neste caso o efeito aleatório deve ser incluído para cada paciente, de forma a permitir a estimação correta dos parâmetros na presença de estrutura de correlação intra-indivíduo. Além disso, a inclusão do efeito aleatório permite estimar o efeito de uma "fragilidade"particular de cada indivíduo gerada por covariáveis não medidas.

Exercício 11.2: Um dos exemplos apresentados no texto é sobrevida em diálise. Refaça a análise dos dados de diálise apresentada no texto, utilizando os comandos do R disponíveis no texto. Estude os comandos e seus usos.

1. Abra o arquivo *dialmenor.dat* e liste as 10 primeiras linhas.

```
> dialmenor <- read.table("dialmenor.dat", header = T)
> dialmenor[1:10, ]

   unidade idade sexo inicio fim status tempo grande causa
1      128    52    1     26  45      0    19      1   out
2      128    76    0     32  33      0     1      1   out
3      128    61    1     22  24      0     2      1   out
4      128    35    0      7  13      0     6      1   out
5      128    42    0      2  13      0    11      1   out
6      128    44    1      6  30      0    24      1   hip
7      128    41    1      1   6      1     5      1   out
8      128    39    1     10  13      0     3      1   out
9      128    57    0      7  45      0    38      1   out
10     128    71    1     16  33      0    17      1   out
```

2. Coloque a hipertensão como referência, na variável causa.

   Para ver qual a categoria que está como referência, é útil fazer uma tabela da variável. A primeira categoria é sempre a categoria de referência:

```
> table(dialmenor$causa)

con dia hip out ren
 49 180 307 133 192

> dialmenor$causa <- relevel(dialmenor$causa, "hip")
> table(dialmenor$causa)

hip con dia out ren
307  49 180 133 192
```

3. Ajuste um modelo de Cox clássico considerando apenas as variáveis sexo e idade, e outro contendo a variável unidade como fator. Este último modelo ajusta um efeito para cada unidade de diálise.

```
> require(survival)

[1] TRUE
```

```
> fit1.cox <- coxph(Surv(inicio, fim, status) ~ idade + sexo, data = dialmenor)
> fit2.cox <- coxph(Surv(inicio, fim, status) ~ idade + sexo +
+     factor(unidade), data = dialmenor)
```

Ao ajustar o modelo com as unidades como fator, uma mensagem de aviso indica que o beta de uma unidade pode ser infinito. Isso em geral significa que não houve óbito nesta unidade. Veja o número de pacientes e de óbitos na tabela:

```
> table(dialmenor$unidade, dialmenor$status)

       0  1
  128  24  5
  217   4  1
  344  47  4
  561  30 16
  562  27 16
  641  53 23
  741   6  5
  1048 39  2
  1051 73  6
  1053 26  7
  1070 92 23
  1071  4  1
  1100 59 19
  1159 40 25
  1654 39  7
  1692 52 16
  2811  5  4
  2844 29 20
  5681  5  0
  5692  6  1
```

Analisando os resultados dos dois modelos ajustados:

```
> summary(fit1.cox)

Call:
coxph(formula = Surv(inicio, fim, status) ~ idade + sexo, data = dialmenor)

  n= 861
        coef exp(coef) se(coef)      z       p
idade  0.0356     1.036   0.0052  6.845 7.6e-12
sexo  -0.1289     0.879   0.1418 -0.909 3.6e-01

      exp(coef) exp(-coef) lower .95 upper .95
```

```
idade     1.036      0.965     1.026      1.05
sexo      0.879      1.138     0.666      1.16

Rsquare= 0.058   (max possible= 0.929 )
Likelihood ratio test= 51.1  on 2 df,   p=7.9e-12
Wald test            = 47.5  on 2 df,   p=4.74e-11
Score (logrank) test = 48.8  on 2 df,   p=2.54e-11
```

```
> summary(fit2.cox)

Call:
coxph(formula = Surv(inicio, fim, status) ~ idade + sexo + factor(unidade),
    data = dialmenor)

  n= 861
                        coef exp(coef) se(coef)       z       p
idade                 0.0342  1.03e+00 5.47e-03  6.2528 4.0e-10
sexo                 -0.1780  8.37e-01 1.46e-01 -1.2180 2.2e-01
factor(unidade)217   -0.1142  8.92e-01 1.10e+00 -0.1041 9.2e-01
factor(unidade)344   -0.6101  5.43e-01 6.74e-01 -0.9046 3.7e-01
factor(unidade)561    0.7304  2.08e+00 5.15e-01  1.4192 1.6e-01
factor(unidade)562    0.9614  2.62e+00 5.16e-01  1.8639 6.2e-02
factor(unidade)641    0.6416  1.90e+00 4.94e-01  1.2986 1.9e-01
factor(unidade)741    1.9697  7.17e+00 6.38e-01  3.0855 2.0e-03
factor(unidade)1048  -1.6302  1.96e-01 8.38e-01 -1.9460 5.2e-02
factor(unidade)1051  -0.8905  4.10e-01 6.11e-01 -1.4580 1.4e-01
factor(unidade)1053   0.3384  1.40e+00 5.88e-01  0.5755 5.6e-01
factor(unidade)1070  -0.0196  9.81e-01 4.94e-01 -0.0396 9.7e-01
factor(unidade)1071  -0.0679  9.34e-01 1.10e+00 -0.0619 9.5e-01
factor(unidade)1100   0.2468  1.28e+00 5.04e-01  0.4894 6.2e-01
factor(unidade)1159   2.5284  1.25e+01 4.94e-01  5.1143 3.1e-07
factor(unidade)1654  -0.1994  8.19e-01 5.88e-01 -0.3390 7.3e-01
factor(unidade)1692   0.4452  1.56e+00 5.21e-01  0.8553 3.9e-01
factor(unidade)2811   0.5290  1.70e+00 6.72e-01  0.7878 4.3e-01
factor(unidade)2844   1.6087  5.00e+00 5.02e-01  3.2041 1.4e-03
factor(unidade)5681 -14.9412  3.24e-07 1.32e+03 -0.0113 9.9e-01
factor(unidade)5692   0.2720  1.31e+00 1.10e+00  0.2475 8.0e-01

                    exp(coef) exp(-coef) lower .95 upper .95
idade                1.03e+00   9.66e-01    1.0238      1.05
sexo                 8.37e-01   1.19e+00    0.6285      1.11
factor(unidade)217   8.92e-01   1.12e+00    0.1038      7.66
factor(unidade)344   5.43e-01   1.84e+00    0.1449      2.04
factor(unidade)561   2.08e+00   4.82e-01    0.7570      5.69
factor(unidade)562   2.62e+00   3.82e-01    0.9516      7.19
factor(unidade)641   1.90e+00   5.26e-01    0.7213      5.00
factor(unidade)741   7.17e+00   1.39e-01    2.0513     25.05
```

```
factor(unidade)1048  1.96e-01   5.11e+00    0.0379      1.01
factor(unidade)1051  4.10e-01   2.44e+00    0.1240      1.36
factor(unidade)1053  1.40e+00   7.13e-01    0.4431      4.44
factor(unidade)1070  9.81e-01   1.02e+00    0.3724      2.58
factor(unidade)1071  9.34e-01   1.07e+00    0.1087      8.03
factor(unidade)1100  1.28e+00   7.81e-01    0.4764      3.44
factor(unidade)1159  1.25e+01   7.98e-02    4.7562     33.03
factor(unidade)1654  8.19e-01   1.22e+00    0.2587      2.59
factor(unidade)1692  1.56e+00   6.41e-01    0.5626      4.33
factor(unidade)2811  1.70e+00   5.89e-01    0.4552      6.33
factor(unidade)2844  5.00e+00   2.00e-01    1.8676     13.37
factor(unidade)5681  3.24e-07   3.08e+06    0.0000       Inf
factor(unidade)5692  1.31e+00   7.62e-01    0.1522     11.32

Rsquare= 0.209   (max possible= 0.929 )
Likelihood ratio test= 202  on 21 df,   p=0
Wald test            = 203  on 21 df,   p=0
Score (logrank) test = 293  on 21 df,   p=0
```

Os efeitos das variáveis do indivíduo não varia muito entre os modelos, o poder explicativo aumenta substancialmente ao incluir as unidade como covariáveis (passa de 0,058 para 0,209 em um máximo possível de 0,929). Entre as unidades, apenas 3 têm efeito significativo.

4. Modele a unidade hospitalar como um termo de fragilidade gama.

```
> fit.gama <- coxph(Surv(inicio, fim, status) ~ idade + sexo +
+     frailty(unidade, sparse = T), data = dialmenor)
> summary(fit.gama)

Call:
coxph(formula = Surv(inicio, fim, status) ~ idade + sexo + frailty(unidade,
    sparse = T), data = dialmenor)

  n= 861
                          coef    se(coef) se2     Chisq  DF   p
idade                      0.0338 0.00549  0.00546  37.90  1.0 7.4e-10
sexo                      -0.1638 0.14578  0.14529   1.26  1.0 2.6e-01
frailty(unidade, sparse =                           136.66 16.8 0.0e+00

      exp(coef) exp(-coef) lower .95 upper .95
idade     1.034      0.967     1.023      1.05
sexo      0.849      1.178     0.638      1.13

Iterations: 10 outer, 26 Newton-Raphson
     Variance of random effect= 0.855   I-likelihood = -1068.2
```

```
Degrees of freedom for terms=  1.0  1.0 16.8
Rsquare= 0.204   (max possible= 0.929 )
Likelihood ratio test= 197  on 18.8 df,   p=0
Wald test            = 38.5  on 18.8 df,   p=0.00464
```

A distribuição *default* para o efeito aleatório é gama. O efeito das variáveis do indivíduo não se alterou, e a variância do efeito aleatório foi de 0,855.

5. Ajuste o mesmo modelo, utilizando fragilidade gaussiana.

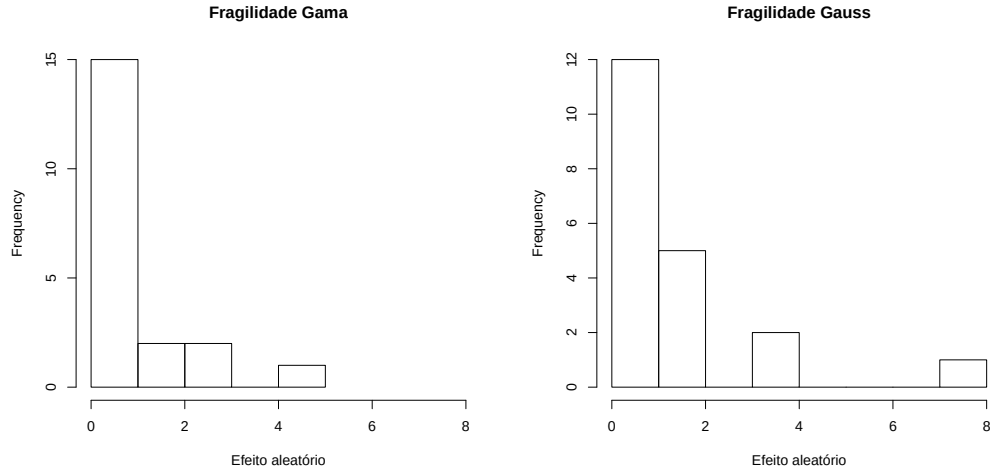
```
> fit.gauss <- coxph(Surv(inicio, fim, status) ~ idade + sexo +
+     frailty(unidade, sparse = T, dist = "gauss"), data = dialmenor)
> summary(fit.gauss)

Call:
coxph(formula = Surv(inicio, fim, status) ~ idade + sexo + frailty(unidade,
    sparse = T, dist = "gauss"), data = dialmenor)

  n= 861
                          coef   se(coef) se2     Chisq  DF   p
idade                      0.034 0.00547  0.00544  38.65  1.0 5.1e-10
sexo                      -0.165 0.14590  0.14547   1.29  1.0 2.6e-01
frailty(unidade, sparse =                          151.29 15.7 0.0e+00

      exp(coef) exp(-coef) lower .95 upper .95
idade     1.035      0.967     1.024      1.05
sexo      0.848      1.180     0.637      1.13

Iterations: 5 outer, 16 Newton-Raphson
     Variance of random effect= 0.792
Degrees of freedom for terms=  1.0  1.0 15.7
Rsquare= 0.204   (max possible= 0.929 )
Likelihood ratio test= 197  on 17.7 df,   p=0
Wald test            = 39.3  on 17.7 df,   p=0.00221
```

Resposta: O resultado é muito parecido com o anterior. Os termos aleatórios nos dois modelos são significativamente diferentes de zero (ver a coluna `p`), com variância respectivamente de 0,855 (gama) e 0,792 (gauss).

Os efeitos aleatórios estimados por cada um dos modelos pode ser visualizado através de um histograma. Podemos usar a exponencial da fragilidade para obtermos a estimativa de risco. Os valores ficam arquivados sob o nome *frail* no objeto resultado do modelo.

```
> hist(exp(fit.gama$frail), xlim = c(0, 8), main = "Fragilidade Gama",
+     xlab = "Efeito aleatório")
```

```
> hist(exp(fit.gauss$frail), xlim = c(0, 8), main = "Fragilidade Gauss",
+     xlab = "Efeito aleatório")
```



Exercício 11.3: Um estudo sobre infarto busca avaliar o efeito de variáveis relacionadas ao paciente (sexo, idade) e variáveis relacionadas ao hospital, na sobrevida de pacientes lá atendidos. Para isso, um banco de dados foi criado, com informações de diversos hospitais. Abra o banco de dados *infarto.dat* no R e identifique as variáveis existentes para os indivíduos e para os hospitais. Note que cada paciente só tem uma linha, e que cada hospital (segundo nível) tem uma linha por paciente atendido, caracterizando uma estrutura aninhada. Para maiores detalhes da descrição desse banco consulte o Apêndice C.9.

```
> infarto <- read.table("infarto.dat", header = T)
> names(infarto)
```

```
 [1] "hospital" "id"       "ini"      "fim"      "status"   "sexo"
 [7] "idade"    "natureza" "volume"   "luti"
```

1. Vamos começar supondo que temos apenas as variáveis no nível de indivíduo (sexo e idade). Ajuste um modelo de riscos proporcionais (clássico) e interprete o resultado.

   Mod1: sobrevida = sexo + idade

   ```
   > mod1 <- coxph(Surv(ini, fim, status) ~ idade + sexo, data = infarto)
   > summary(mod1)
   ```

```
Call:
coxph(formula = Surv(ini, fim, status) ~ idade + sexo, data = infarto)

  n= 3176
        coef exp(coef) se(coef)     z     p
idade  0.0452     1.046  0.00342 13.23 0.000
sexoM -0.2801     0.756  0.08531 -3.28 0.001

      exp(coef) exp(-coef) lower .95 upper .95
idade     1.046      0.956     1.039     1.053
sexoM     0.756      1.323     0.639     0.893

Rsquare= 0.065   (max possible= 0.868 )
Likelihood ratio test= 213  on 2 df,   p=0
Wald test            = 209  on 2 df,   p=0
Score (logrank) test = 214  on 2 df,   p=0
```

Resposta: O modelo ajustado é diferente do modelo nulo pelos testes da razão de verossimilhança, score e Wald, explicando 7,5% da variabilidade total (Rsquare= 0.065 (max possible= 0.868)). O risco de infarto aumenta em 4,6% a cada ano a mais de idade, e o sexo masculino tem risco menor que o feminino, uma vez internados em hospital. O sobrerisco do sexo feminino é de 1,323 – exp(-coef).

2. Como existe uma estrutura de correlação, por termos indivíduos atendidos no mesmo hospital, podemos ajustar um modelo com um termo aleatório para cada hospital para estimar um perfil de risco neste nível (apesar da ausência de covariável neste nível). Registre o nível de significância do termo aleatório.

   Mod1f: sobrevida = sexo + idade +
   frailty(hospital, sparse=F)

```
> mod1f <- coxph(Surv(ini, fim, status) ~ idade + sexo + frailty(hospital,
+     sparse = F), data = infarto)
> summary(mod1f)

Call:
coxph(formula = Surv(ini, fim, status) ~ idade + sexo + frailty(hospital,
    sparse = F), data = infarto)

  n= 3176
                          coef    se(coef) se2     Chisq DF   p
idade                    0.0449 0.00342  0.00342 172.1  1.0 0.0e+00
sexoM                   -0.3069 0.08602  0.08582  12.7  1.0 3.6e-04
frailty(hospital, sparse                          57.1 14.3 4.6e-07
```

```
        exp(coef) exp(-coef) lower .95 upper .95
idade       1.046      0.956     1.039     1.053
sexoM       0.736      1.359     0.622     0.871
gamma:a     0.852      1.174     0.570     1.273
gamma:b     0.925      1.081     0.635     1.348
gamma:c     0.862      1.161     0.522     1.423
gamma:d     1.007      0.993     0.696     1.457
gamma:e     1.221      0.819     0.874     1.705
gamma:f     0.634      1.578     0.417     0.964
gamma:g     0.831      1.203     0.528     1.309
gamma:h     1.033      0.968     0.746     1.431
gamma:i     1.366      0.732     1.068     1.747
gamma:j     0.667      1.500     0.476     0.934
gamma:k     0.667      1.499     0.495     0.899
gamma:l     0.924      1.082     0.542     1.577
gamma:m     0.866      1.154     0.639     1.175
gamma:n     1.163      0.860     0.830     1.630
gamma:o     1.102      0.908     0.647     1.877
gamma:p     0.593      1.686     0.399     0.882
gamma:q     1.051      0.951     0.690     1.602
gamma:r     1.158      0.864     0.666     2.012
gamma:s     1.113      0.898     0.676     1.833
gamma:t     0.786      1.272     0.532     1.161
gamma:u     1.348      0.742     1.002     1.813
gamma:v     1.228      0.814     0.895     1.685
gamma:w     1.144      0.874     0.788     1.662
gamma:x     1.374      0.728     1.043     1.809
gamma:y     1.086      0.921     0.707     1.668

Iterations: 10 outer, 25 Newton-Raphson
     Variance of random effect= 0.0945   I-likelihood = -3095
Degrees of freedom for terms=  1.0  1.0 14.3
Rsquare= 0.086   (max possible= 0.868 )
Likelihood ratio test= 287  on 16.3 df,   p=0
Wald test            = 267  on 16.3 df,   p=0
```

Resposta: O termo aleatório é significativametne diferente de zero.

3. Agora vamos adicionar ao modelo as variáveis de nível hospitalar. Elas devem explicar parte da fragilidade mensurada no modelo acima. Ajuste cada um dos modelos abaixo. Observe o efeito da inclusão das novas covariáveis no termo aleatório.

   Mod2f: sobrevida = sexo + idade + luti +
   frailty(hospital, sparse=F)

Mod3f: sobrevida = sexo + idade + luti + natureza +
frailty(hospital, sparse=F)

Mod4f: sobrevida = sexo + idade + luti + natureza +
volume + frailty(hospital, sparse=F)

Mod2f: sobrevida = sexo + idade + luti + frailty(hospital, sparse=F)

```
> mod2f <- coxph(Surv(ini, fim, status) ~ idade + sexo + luti +
+     frailty(hospital, sparse = F), data = infarto)
> summary(mod2f)

Call:
coxph(formula = Surv(ini, fim, status) ~ idade + sexo + luti +
    frailty(hospital, sparse = F), data = infarto)

  n= 3176
                           coef    se(coef) se2     Chisq  DF   p
idade                       0.0449 0.00342  0.00342 172.24  1.0 0.0e+00
sexoM                      -0.3084 0.08597  0.08580  12.86  1.0 3.3e-04
luti25+                     0.1420 0.17153  0.10690   0.69  1.0 4.1e-01
lutin                       0.2762 0.23887  0.14753   1.34  1.0 2.5e-01
frailty(hospital, sparse                             42.76 12.1 2.7e-05

        exp(coef) exp(-coef) lower .95 upper .95
idade       1.046      0.956     1.039     1.053
sexoM       0.735      1.361     0.621     0.869
luti25+     1.153      0.868     0.823     1.613
lutin       1.318      0.759     0.825     2.105
gamma:a     0.827      1.209     0.548     1.248
gamma:b     0.891      1.123     0.601     1.319
gamma:c     0.854      1.171     0.530     1.378
gamma:d     1.052      0.951     0.734     1.507
gamma:e     1.262      0.793     0.909     1.751
gamma:f     0.687      1.455     0.458     1.030
gamma:g     0.817      1.224     0.523     1.278
gamma:h     1.085      0.922     0.786     1.498
gamma:i     1.281      0.781     0.939     1.746
gamma:j     0.719      1.390     0.514     1.007
gamma:k     0.719      1.390     0.531     0.974
gamma:l     0.886      1.129     0.533     1.473
gamma:m     0.923      1.084     0.679     1.254
gamma:n     1.104      0.906     0.770     1.584
gamma:o     1.096      0.913     0.669     1.795
gamma:p     0.639      1.565     0.436     0.937
gamma:q     0.949      1.054     0.605     1.490
gamma:r     1.133      0.883     0.680     1.885
```

```
gamma:s     1.114      0.898     0.699     1.776
gamma:t     0.838      1.193     0.574     1.224
gamma:u     1.397      0.716     1.039     1.878
gamma:v     1.275      0.784     0.932     1.743
gamma:w     1.178      0.849     0.820     1.691
gamma:x     1.165      0.858     0.773     1.756
gamma:y     1.111      0.900     0.738     1.672

Iterations: 8 outer, 22 Newton-Raphson
     Variance of random effect= 0.0781   I-likelihood = -3094.2
Degrees of freedom for terms=  1.0  1.0  0.8 12.1
Rsquare= 0.086   (max possible= 0.868 )
Likelihood ratio test= 285  on 14.9 df,   p=0
Wald test            = 265  on 14.9 df,   p=0
```

Mod3f: sobrevida = sexo + idade + luti + natureza + frailty(hospital, sparse=F)

```
> mod3f <- coxph(Surv(ini, fim, status) ~ idade + sexo + luti +
+     natureza + frailty(hospital, sparse = F), data = infarto)
> summary(mod3f)

Call:
coxph(formula = Surv(ini, fim, status) ~ idade + sexo + luti +
    natureza + frailty(hospital, sparse = F), data = infarto)

  n= 3176
                          coef    se(coef) se2     Chisq  DF   p
idade                      0.0453 0.00342  0.00342 175.04 1.00 0.00000
sexoM                     -0.3100 0.08585  0.08569  13.04 1.00 0.00031
luti25+                    0.5475 0.16462  0.12292  11.06 1.00 0.00088
lutin                      0.1501 0.20863  0.15813   0.52 1.00 0.47000
naturezaPE                 0.1140 0.33698  0.28888   0.11 1.00 0.74000
naturezaPFU               -0.6889 0.38916  0.33024   3.13 1.00 0.07700
naturezaPM                -0.3447 0.35318  0.30426   0.95 1.00 0.33000
frailty(hospital, sparse                             12.80 5.81 0.04200

           exp(coef) exp(-coef) lower .95 upper .95
idade          1.046      0.956     1.039     1.053
sexoM          0.733      1.363     0.620     0.868
luti25+        1.729      0.578     1.252     2.387
lutin          1.162      0.861     0.772     1.749
naturezaPE     1.121      0.892     0.579     2.169
naturezaPFU    0.502      1.992     0.234     1.077
naturezaPM     0.708      1.412     0.355     1.415
gamma:a        0.949      1.054     0.720     1.251
gamma:b        0.988      1.012     0.755     1.294
```

```
gamma:c         0.959      1.043     0.715     1.287
gamma:d         0.944      1.059     0.725     1.229
gamma:e         1.215      0.823     0.953     1.550
gamma:f         0.882      1.134     0.669     1.163
gamma:g         0.863      1.158     0.644     1.157
gamma:h         1.138      0.879     0.893     1.449
gamma:i         1.067      0.937     0.825     1.380
gamma:j         0.899      1.112     0.699     1.157
gamma:k         0.897      1.115     0.706     1.139
gamma:l         0.987      1.014     0.731     1.332
gamma:m         1.052      0.951     0.830     1.332
gamma:n         1.116      0.896     0.864     1.440
gamma:o         1.027      0.974     0.761     1.386
gamma:p         0.943      1.061     0.710     1.251
gamma:q         0.973      1.028     0.715     1.323
gamma:r         1.046      0.956     0.773     1.414
gamma:s         1.021      0.979     0.762     1.368
gamma:t         0.825      1.212     0.628     1.084
gamma:u         1.110      0.901     0.877     1.406
gamma:v         1.047      0.955     0.820     1.338
gamma:w         1.013      0.987     0.779     1.317
gamma:x         1.041      0.961     0.779     1.390
gamma:y         0.999      1.001     0.757     1.318

Iterations: 8 outer, 25 Newton-Raphson
     Variance of random effect= 0.0249   I-likelihood = -3088.2
Degrees of freedom for terms= 1.0 1.0 1.1 2.0 5.8
Rsquare= 0.084   (max possible= 0.868 )
Likelihood ratio test= 278  on 10.9 df,   p=0
Wald test            = 264  on 10.9 df,   p=0
```

Mod4f: sobrevida = sexo + idade + luti + natureza + volume + frailty(hospital, sparse=F)

```
> mod4f <- coxph(Surv(ini, fim, status) ~ idade + sexo + luti +
+     natureza + volume + frailty(hospital, sparse = F), data = infarto)
> summary(mod4f)

Call:
coxph(formula = Surv(ini, fim, status) ~ idade + sexo + luti +
    natureza + volume + frailty(hospital, sparse = F), data = infarto)

  n= 3176
                          coef    se(coef) se2     Chisq  DF   p
idade                      0.0454 0.00343  0.00343 175.44 1.00 0.00000
sexoM                     -0.3104 0.08586  0.08570  13.07 1.00 0.00030
luti25+                    0.5576 0.16512  0.12320  11.40 1.00 0.00073
```

```
lutin                       0.1353 0.20384  0.15596   0.44 1.00 0.51000
naturezaPE                  0.1532 0.33254  0.28858   0.21 1.00 0.65000
naturezaPFU                -0.6491 0.38329  0.32838   2.87 1.00 0.09000
naturezaPM                 -0.3070 0.34623  0.30202   0.79 1.00 0.38000
volumevp                    0.3402 0.30606  0.28665   1.24 1.00 0.27000
frailty(hospital, sparse                              12.89 5.78 0.03900

            exp(coef) exp(-coef) lower .95 upper .95
idade           1.046      0.956     1.039     1.054
sexoM           0.733      1.364     0.620     0.868
luti25+         1.746      0.573     1.264     2.414
lutin           1.145      0.873     0.768     1.707
naturezaPE      1.166      0.858     0.607     2.237
naturezaPFU     0.522      1.914     0.247     1.107
naturezaPM      0.736      1.359     0.373     1.450
volumevp        1.405      0.712     0.771     2.560
gamma:a         0.949      1.054     0.719     1.252
gamma:b         0.988      1.013     0.754     1.294
gamma:c         0.959      1.043     0.714     1.288
gamma:d         0.947      1.056     0.727     1.233
gamma:e         1.220      0.820     0.956     1.557
gamma:f         0.885      1.130     0.671     1.167
gamma:g         0.863      1.159     0.643     1.157
gamma:h         1.143      0.875     0.897     1.456
gamma:i         1.068      0.936     0.825     1.383
gamma:j         0.904      1.107     0.702     1.164
gamma:k         0.903      1.108     0.710     1.147
gamma:l         0.957      1.044     0.703     1.304
gamma:m         1.058      0.946     0.835     1.340
gamma:n         1.116      0.896     0.864     1.441
gamma:o         1.012      0.988     0.747     1.371
gamma:p         0.945      1.058     0.711     1.255
gamma:q         0.988      1.012     0.727     1.342
gamma:r         1.045      0.957     0.772     1.414
gamma:s         0.986      1.014     0.729     1.334
gamma:t         0.828      1.208     0.630     1.088
gamma:u         1.116      0.896     0.881     1.413
gamma:v         1.052      0.951     0.823     1.344
gamma:w         1.016      0.984     0.781     1.322
gamma:x         1.055      0.948     0.792     1.405
gamma:y         1.001      0.999     0.758     1.322

Iterations: 8 outer, 25 Newton-Raphson
     Variance of random effect= 0.0251   I-likelihood = -3087.7
Degrees of freedom for terms= 1.0 1.0 1.1 2.0 0.9 5.8
Rsquare= 0.084   (max possible= 0.868 )
Likelihood ratio test= 280  on 11.8 df,   p=0
```

```
Wald test            = 265  on 11.8 df,   p=0
```

Resposta: Uma boa forma de comparar vários modelos é resumir em uma tabela como esta:

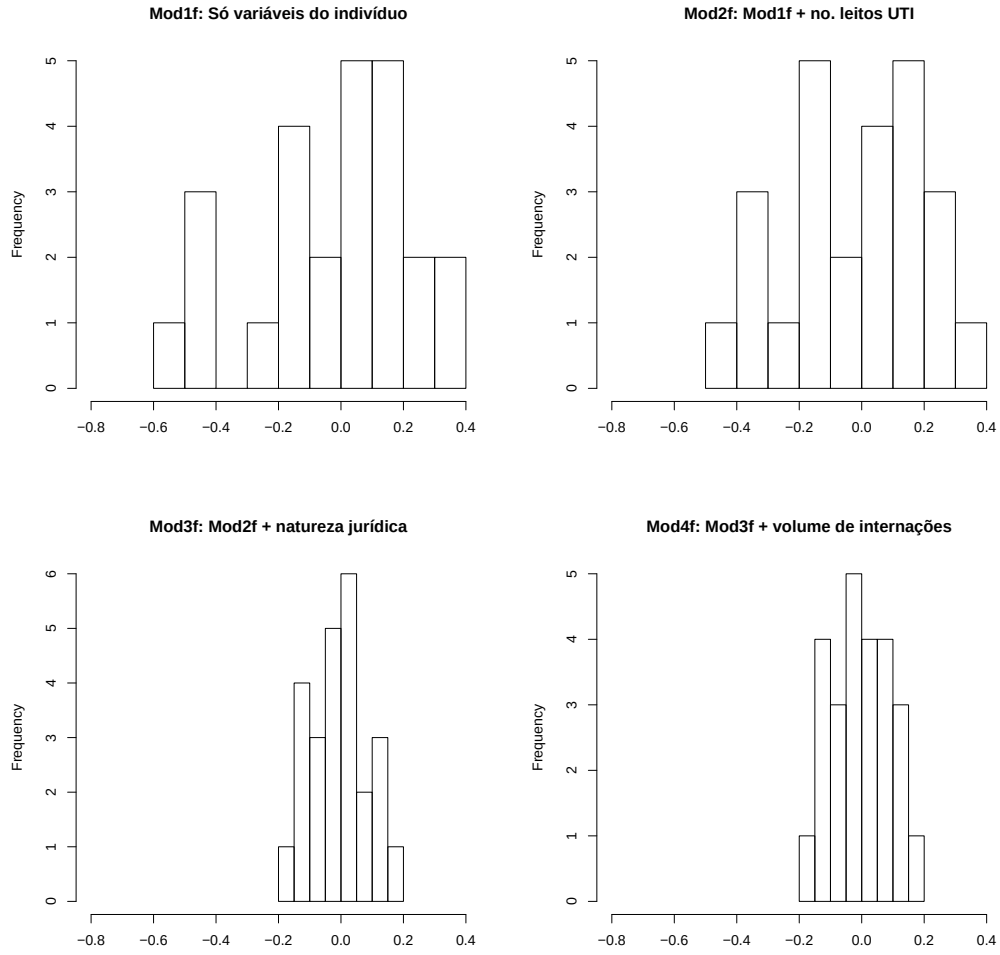
	Modelos			
Variáveis	**1**	**2**	**3**	**4**
Idade	1.046*	1.046*	1.046*	1.046*
Sexo	0.736*	0.735*	0.733*	0.733*
Leitos UTI 25+	-	1.153	1.729*	1.746*
Sem Leitos UTI	-	1.318	1.162	1.145
Natureza Estadual	-	-	1.121	1.166
Natureza Federal/Universitário	-	-	0.502	0.522
Natureza Municipal	-	-	0.708	0.736
Volume <25 internações	-	-	-	1.405
Fragilidade (variância)	0.0945*	0.0781*	0.0249*	0.0251*

*valores significativos para intervalo de confiança de 90%

Observe que a variância do efeito aleatório, único valor de fato estimado, diminui com a inclusão das variáveis no nível do hospital, principalmente o número de leitos de UTI e a natureza jurídica do hospital.

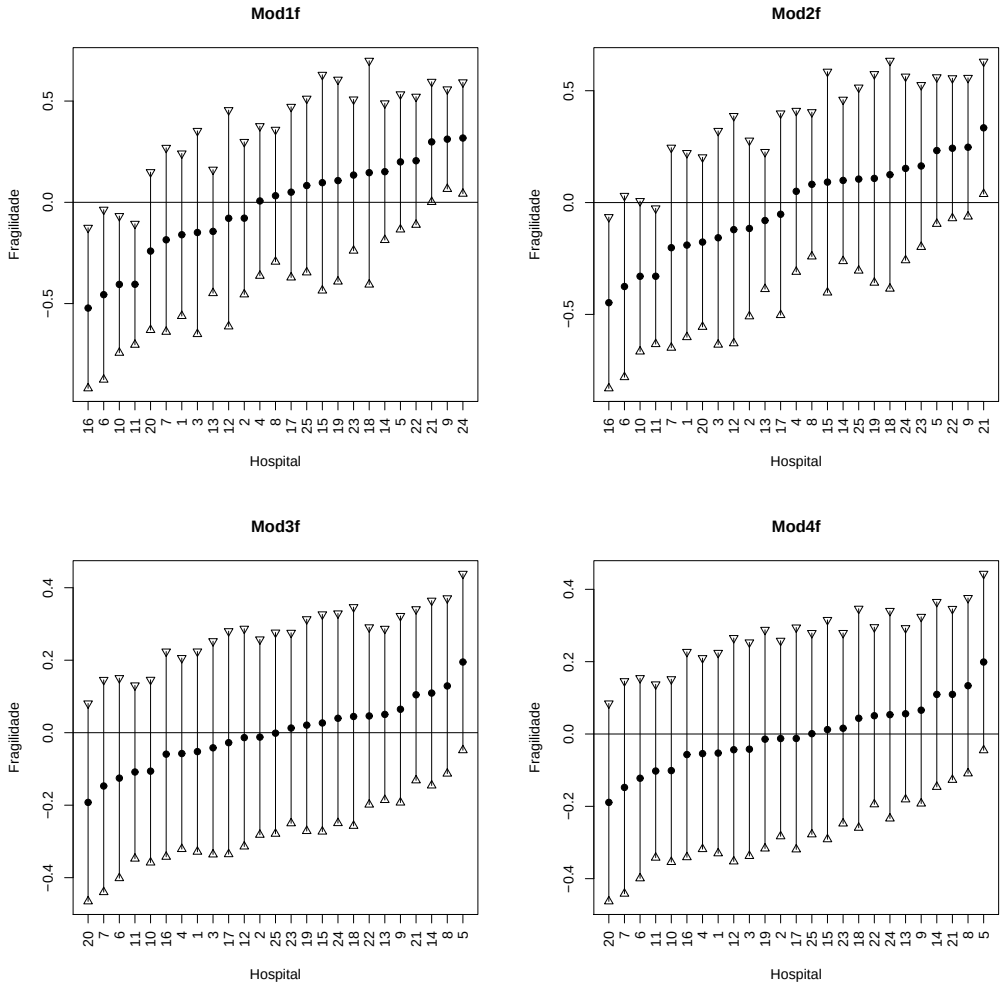
4. Faça um gráfico das fragilidades em cada modelo e observe a dispersão das fragilidades. Por que à medida que se incluem novas variáveis do nível do hospital diminui a variância da fragilidade?

```
> hist(mod1f$coeff[3:27], main = "Mod1f: Só variáveis do indivíduo",
+     xlab = "", xlim = c(-0.8, 0.4))
> hist(mod2f$coeff[5:29], main = "Mod2f: Mod1f + no. leitos UTI",
+     xlab = "", xlim = c(-0.8, 0.4))
> hist(mod3f$coeff[8:32], main = "Mod3f: Mod2f + natureza jurídica",
+     xlab = "", xlim = c(-0.8, 0.4))
> hist(mod4f$coeff[9:33], main = "Mod4f: Mod3f + volume de internações",
+     xlab = "", xlim = c(-0.8, 0.4))
```



Uma forma de olhar as fragilidades é com um gráfico do tipo a seguir. PAra a elaboração desse gráfico é necessário carregar as funções que estão no arquivo *Rfun.r* disponibilizada na página http://dengue.procc.fiocruz.br/∼sobrevida/dados/

```
> source("Rfun.r")
> plot.frail(infarto$hospital, mod1f, ylab = "Fragilidade", main = "Mod1f",
+     xlab = "Hospital")
> plot.frail(infarto$hospital, mod2f, ylab = "Fragilidade", main = "Mod2f",
+     xlab = "Hospital")
> plot.frail(infarto$hospital, mod3f, ylab = "Fragilidade", main = "Mod3f",
+     xlab = "Hospital")
> plot.frail(infarto$hospital, mod4f, ylab = "Fragilidade", main = "Mod4f",
+     xlab = "Hospital")
```



Resposta: Pode-se observar que os gráficos das fragilidades referentes aos modelos 1 (somente variáveis do indivíduo) e 2 (incluindo a variável *nº de leitos UTI*) são muito semelhantes, com as mesmas unidades com menor risco (unidades 16, 6, 10 e 11), embora o intervalo de confiança no segundo modelo inclua o zero. Na outra ponta, o unidade 24, que apresentava o maior risco agora tem perfil de risco médio. A inclusão das demais variáveis do nível do hospital, entretanto, fez com que nenhuma unidade tenha perfil de risco significativamente diferente da média. Ou seja, a inclusão da variável *Natureza Jurídica* do hospital permitiu explicar o excesso de risco estimado através do modelos de sobrevida com efeitos aleatórios. A variável *Volume de internações* não acrescentou muita informação ao modelo: por um lado, seu

efeito não era significativamente diferente de um; por outro, a variabilidade entre os hospitais também não se alterou.

Exercício 11.4: Compare as estimativas do modelo *mod3f* com fragilidade, estimado no exercício anterior, com um modelo semelhante que utilize uma distribuição gaussiana. Houve diferença nas estimativas?

```
> mod3f.gauss <- coxph(Surv(ini, fim, status) ~ idade + sexo +
+     luti + natureza + frailty(hospital, sparse = F, dist = "gauss"),
+     data = infarto)
> summary(mod3f.gauss)


Call:
coxph(formula = Surv(ini, fim, status) ~ idade + sexo + luti +
    natureza + frailty(hospital, sparse = F, dist = "gauss"),
    data = infarto)

  n= 3176
                          coef    se(coef) se2     Chisq  DF   p
idade                      0.0451 0.00342  0.00342 173.48 1.00 0.0000
sexoM                     -0.3112 0.08602  0.08585  13.08 1.00 0.0003
luti25+                    0.5375 0.20180  0.13405   7.09 1.00 0.0077
lutin                      0.1259 0.25463  0.17014   0.24 1.00 0.6200
naturezaPE                 0.0693 0.38121  0.29682   0.03 1.00 0.8600
naturezaPFU               -0.7211 0.44329  0.34173   2.65 1.00 0.1000
naturezaPM                -0.3957 0.39759  0.31383   0.99 1.00 0.3200
frailty(hospital, sparse                             19.51 8.83 0.0190

            exp(coef) exp(-coef) lower .95 upper .95
idade           1.046      0.956     1.039     1.053
sexoM           0.733      1.365     0.619     0.867
luti25+         1.712      0.584     1.153     2.542
lutin           1.134      0.882     0.689     1.868
naturezaPE      1.072      0.933     0.508     2.263
naturezaPFU     0.486      2.057     0.204     1.159
naturezaPM      0.673      1.485     0.309     1.467
gauss:a         0.921      1.086     0.641     1.323
gauss:b         0.983      1.018     0.688     1.403
gauss:c         0.926      1.080     0.620     1.383
gauss:d         0.919      1.088     0.657     1.285
gauss:e         1.391      0.719     0.998     1.939
gauss:f         0.828      1.207     0.585     1.173
gauss:g         0.798      1.253     0.546     1.167
gauss:h         1.228      0.814     0.892     1.691
gauss:i         1.101      0.908     0.773     1.568
```

```
gauss:j         0.869      1.151     0.636     1.186
gauss:k         0.875      1.143     0.653     1.172
gauss:l         0.979      1.022     0.639     1.500
gauss:m         1.081      0.925     0.799     1.461
gauss:n         1.189      0.841     0.839     1.685
gauss:o         1.053      0.949     0.679     1.634
gauss:p         0.909      1.101     0.622     1.328
gauss:q         0.949      1.053     0.612     1.472
gauss:r         1.104      0.906     0.701     1.737
gauss:s         1.043      0.959     0.690     1.577
gauss:t         0.760      1.315     0.544     1.062
gauss:u         1.165      0.859     0.857     1.583
gauss:v         1.072      0.933     0.782     1.469
gauss:w         1.023      0.978     0.725     1.443
gauss:x         1.076      0.929     0.710     1.632
gauss:y         0.999      1.001     0.689     1.449

Iterations: 7 outer, 17 Newton-Raphson
     Variance of random effect= 0.054
Degrees of freedom for terms= 1.0 1.0 0.9 1.5 8.8
Rsquare= 0.099   (max possible= 0.87 )
Likelihood ratio test= 332  on 13.2 df,   p=0
Wald test            = 271  on 13.2 df,   p=0
```

Resposta: Os efeitos das covariáveis fixas (ver tabela a seguir) praticamente não se altera.

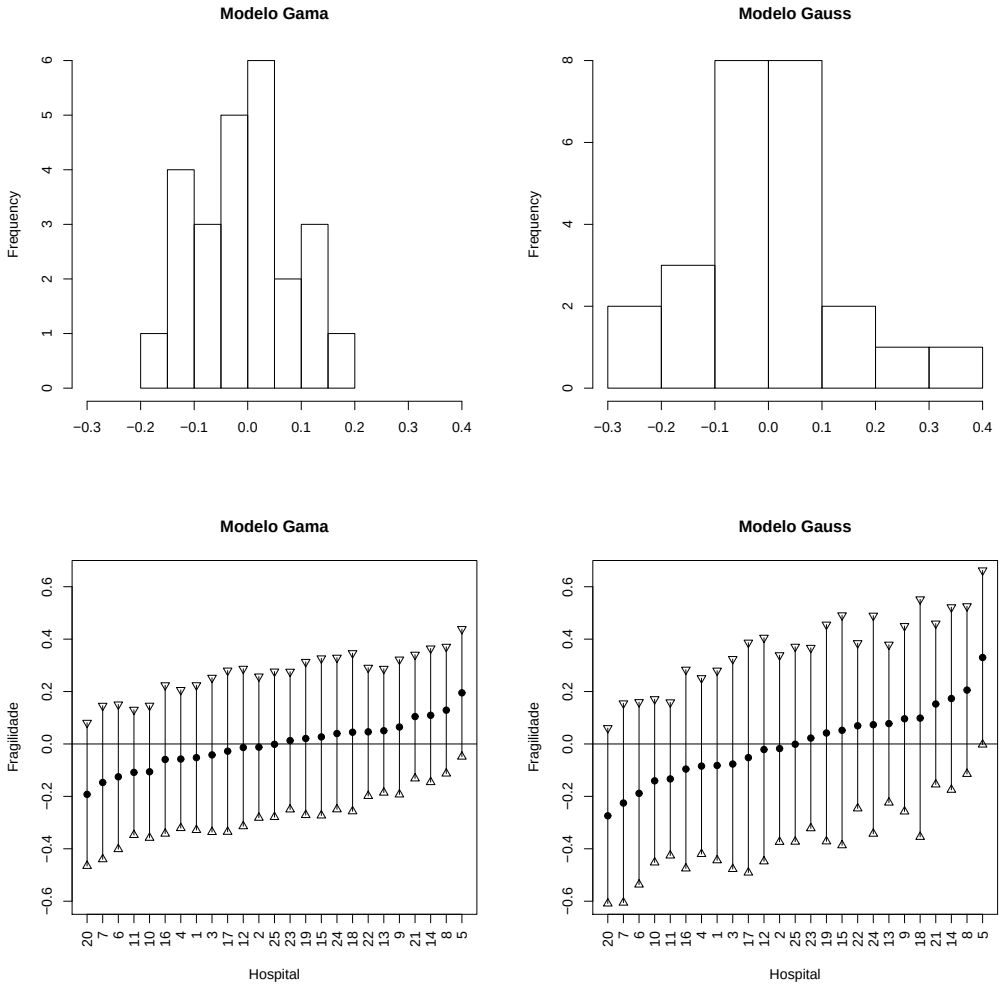
	Modelos	
Variáveis	**Gama**	**Gauss**
Idade	1.046*	1.046*
Sexo	0.733*	0.733*
Leitos UTI 25+	1.729*	1.712*
Sem Leitos UTI	1.162	1.134
Natureza Estadual	1.121	1.072
Natureza Federal/Universitário	0.502	0.486
Natureza Municipal	0.708	0.673
Fragilidade (variância)	0.0249*	0.054*

*valores significativos para intervalo de confiança de 90%

As fragilidades estimadas com distribuição gaussiana apresentam uma maior dispersão, mas no essencial o modelo é muito semelhante.

```
> hist(mod3f$coeff[8:32], xlab = "", xlim = c(-0.3, 0.4), main = "Modelo Gama")
> hist(mod3f.gauss$coeff[8:32], xlab = "", xlim = c(-0.3, 0.4),
```

```
+     main = "Modelo Gauss")
> plot.frail(infarto$hospital, mod3f, xlab = "Hospital", ylim = c(-0.6,
+     0.65), ylab = "Fragilidade", main = "Modelo Gama")
> plot.frail(infarto$hospital, mod3f.gauss, xlab = "Hospital",
+     ylim = c(-0.6, 0.65), ylab = "Fragilidade", main = "Modelo Gauss")
```



Exercício 11.5: No Capítulo 10, sobre eventos múltiplos, ajustou-se modelos marginais para se avaliar o efeito do tratamento com vitamina A no risco de ocorrência de diarréia infantil. Como múltiplos eventos eram observados por criança, uma primeira opção de modelo era o modelo marginal, que corrigia a variância dos efeitos para a presença de correlação entre tempos de sobrevida observados para a mesma criança. Uma outra abordagem para esse problema pode ser feita agora,

utilizando-se fragilidade. Podemos considerar que *criança* é a unidade de segundo nível, que agrega as medidas repetidas de tempo. Neste caso, a adição de um termo aleatório irá estimar o efeito da fragilidade das crianças, não explicada pelos efeitos fixos. Abra o arquivo *multdiarreia.txt* e selecione as primeiras 100 crianças:

```
> diar <- read.table("multdiarreia.txt", header = TRUE)
> diar <- diar[diar$numcri <= 100, ]
```

1. Ajuste um modelo marginal de incrementos independentes e um modelo com fragilidade aos dados.

```
> modelo.inc <- coxph(Surv(ini, fim, status) ~ grupo + idade +
+     cluster(numcri), data = diar)
> summary(modelo.inc)

Call:
coxph(formula = Surv(ini, fim, status) ~ grupo + idade + cluster(numcri),
    data = diar)

  n= 643
            coef exp(coef) se(coef) robust se     z       p
grupovit -0.3513     0.704  0.08694   0.18257 -1.92 5.4e-02
idade    -0.0408     0.960  0.00361   0.00716 -5.70 1.2e-08

         exp(coef) exp(-coef) lower .95 upper .95
grupovit     0.704       1.42     0.492     1.007
idade        0.960       1.04     0.947     0.974

Rsquare= 0.189   (max possible= 0.999 )
Likelihood ratio test= 135  on 2 df,   p=0
Wald test            = 32.8  on 2 df,   p=7.6e-08
Score (logrank) test = 143  on 2 df,   p=0,   Robust = 16.8  p=0.000224

  (Note: the likelihood ratio and score tests assume independence of
     observations within a cluster, the Wald and robust score tests do not).

> modelo.frag <- coxph(Surv(ini, fim, status) ~ grupo + idade +
+     frailty(numcri, sparse = F, dist = "gamma"), data = diar)
> summary(modelo.frag)

Call:
coxph(formula = Surv(ini, fim, status) ~ grupo + idade + frailty(numcri,
    sparse = F, dist = "gamma"), data = diar)

  n= 643
```

```
                         coef    se(coef) se2     Chisq  DF   p
grupovit                -0.2942 0.18680  0.09192   2.48  1.0 1.2e-01
idade                   -0.0424 0.00829  0.00401  26.11  1.0 3.2e-07
frailty(numcri, sparse =                          279.56 73.4 0.0e+00

          exp(coef) exp(-coef) lower .95 upper .95
grupovit      0.745      1.342    0.5167     1.075
idade         0.959      1.043    0.9431     0.974
gamma:1       2.048      0.488    1.1372     3.687
gamma:2       1.099      0.910    0.4585     2.632
gamma:3       1.212      0.825    0.5753     2.553
gamma:4       3.005      0.333    1.6474     5.481
gamma:5       0.949      1.053    0.3998     2.254
gamma:6       1.748      0.572    0.9662     3.164
gamma:7       0.300      3.335    0.0608     1.479
gamma:8       0.344      2.906    0.1184     1.000
gamma:9       0.652      1.533    0.2536     1.678
gamma:10      0.241      4.141    0.0490     1.189
gamma:11      0.729      1.372    0.2823     1.882
gamma:12      0.624      1.603    0.1789     2.175
gamma:13      0.783      1.278    0.3535     1.732
gamma:14      1.094      0.914    0.4625     2.587
gamma:15      0.441      2.269    0.1523     1.276
gamma:16      0.993      1.007    0.4168     2.364
gamma:17      0.515      1.940    0.2136     1.243
gamma:18      2.293      0.436    1.3586     3.869
gamma:19      0.527      1.897    0.1517     1.832
gamma:20      0.327      3.056    0.0662     1.617
gamma:21      0.296      3.375    0.0602     1.457
gamma:22      0.353      2.832    0.0719     1.735
gamma:23      1.242      0.805    0.6370     2.421
gamma:24      0.497      2.010    0.1426     1.736
gamma:25      0.558      1.793    0.1594     1.951
gamma:26      0.520      1.923    0.2024     1.337
gamma:27      0.219      4.558    0.0445     1.083
gamma:28      0.248      4.032    0.0704     0.874
gamma:29      0.972      1.029    0.4938     1.912
gamma:30      3.747      0.267    2.1035     6.673
gamma:31      1.308      0.764    0.6705     2.552
gamma:32      0.614      1.629    0.2823     1.334
gamma:33      0.924      1.083    0.4169     2.046
gamma:34      0.918      1.089    0.3896     2.164
gamma:35      2.093      0.478    1.0265     4.269
gamma:36      1.592      0.628    0.7892     3.212
gamma:37      0.953      1.050    0.4847     1.872
gamma:38      0.770      1.298    0.2977     1.994
gamma:39      0.479      2.089    0.2004     1.143
```

gamma:40	0.723	1.382	0.2809	1.863
gamma:41	1.399	0.715	0.6805	2.876
gamma:42	0.910	1.099	0.4395	1.882
gamma:43	0.389	2.572	0.1117	1.353
gamma:44	1.160	0.862	0.5523	2.435
gamma:45	0.250	4.003	0.0505	1.235
gamma:46	0.407	2.457	0.1158	1.431
gamma:47	0.417	2.396	0.1436	1.213
gamma:48	0.608	1.644	0.2367	1.563
gamma:49	0.329	3.042	0.0943	1.146
gamma:50	0.498	2.008	0.1421	1.746
gamma:51	0.574	1.743	0.1639	2.008
gamma:52	0.656	1.525	0.2242	1.917
gamma:53	0.291	3.436	0.0590	1.435
gamma:54	0.266	3.766	0.0539	1.308
gamma:55	1.007	0.993	0.3906	2.597
gamma:56	0.849	1.177	0.4245	1.699
gamma:57	3.027	0.330	1.8405	4.977
gamma:58	0.541	1.848	0.2559	1.145
gamma:59	0.871	1.148	0.3696	2.051
gamma:60	0.898	1.113	0.4730	1.706
gamma:61	1.216	0.823	0.6544	2.258
gamma:62	1.925	0.519	1.1034	3.359
gamma:63	1.063	0.940	0.4777	2.367
gamma:64	0.587	1.704	0.2721	1.266
gamma:65	0.591	1.693	0.2030	1.719
gamma:66	0.229	4.360	0.0651	0.808
gamma:67	0.522	1.916	0.1500	1.815
gamma:68	1.100	0.909	0.5448	2.219
gamma:69	1.800	0.556	0.9465	3.422
gamma:70	1.711	0.584	0.9660	3.031
gamma:71	0.737	1.356	0.2539	2.141
gamma:72	1.828	0.547	1.0718	3.118
gamma:73	0.829	1.206	0.4062	1.693
gamma:74	1.490	0.671	0.7403	2.998
gamma:75	0.518	1.932	0.1054	2.541
gamma:76	1.610	0.621	0.7591	3.413
gamma:77	0.401	2.491	0.1148	1.404
gamma:78	0.733	1.364	0.3103	1.731
gamma:79	0.794	1.260	0.4045	1.558
gamma:80	1.240	0.807	0.6148	2.500
gamma:81	1.484	0.674	0.7183	3.066
gamma:82	1.245	0.803	0.3624	4.274
gamma:83	1.653	0.605	0.9440	2.893
gamma:84	1.956	0.511	1.1630	3.290
gamma:85	0.831	1.204	0.3235	2.134
gamma:86	1.342	0.745	0.6405	2.812

```
gamma:87      1.025      0.975    0.4875     2.156
gamma:88      0.827      1.209    0.3737     1.830
gamma:89      2.697      0.371    1.6287     4.466
gamma:90      2.316      0.432    1.3914     3.855
gamma:91      0.299      3.342    0.0607     1.475
gamma:92      0.804      1.243    0.3992     1.621
gamma:93      1.202      0.832    0.5382     2.684
gamma:94      0.316      3.162    0.1047     0.955
gamma:95      0.259      3.865    0.0526     1.273
gamma:96      1.136      0.880    0.5140     2.511
gamma:97      0.322      3.105    0.1078     0.963
gamma:98      0.847      1.181    0.2924     2.452
gamma:99      3.393      0.295    2.0468     5.624
gamma:100     0.828      1.207    0.2843     2.413

Iterations: 9 outer, 26 Newton-Raphson
     Variance of random effect= 0.651   I-likelihood = -2287.3
Degrees of freedom for terms=  0.2  0.2 73.4
Rsquare= 0.534   (max possible= 0.999 )
Likelihood ratio test= 492  on 73.8 df,   p=0
Wald test            = 400  on 73.8 df,   p=0
```

2. O termo aleatório foi significativo? Qual a sua variância?

   Resposta: Sim, com variância 0,651.

3. Os valores estimados para os efeitos fixos mudaram de um modelo para o outro? A interpretação deles também. Explique seu significado em cada modelo.

   Resposta:

	Modelos	
Variáveis	AG (IC)	Fragilidade (IC)
grupovit	0.704 (0.492, 1.007)	0.745 (0.5167, 1.075)
idade	0.960 (0.947, 0.974)	0.959 (0.9431, 0.974)

   O efeito da idade não se alterou, mas o efeito protetor da Vitamina A diminuiu no modelo com efeitos aleatórios. O intervalo de confiança no modelo com fragilidade é um pouco maior.
   Modelo AG:
   O efeito da idade é protetor: para cada mes a mais o risco de ter diarréia diminui **em média** em 4%, porém não é significativo. O efeito da vitamina A é de diminui o risco da diarréia em aproximadamente 30%.

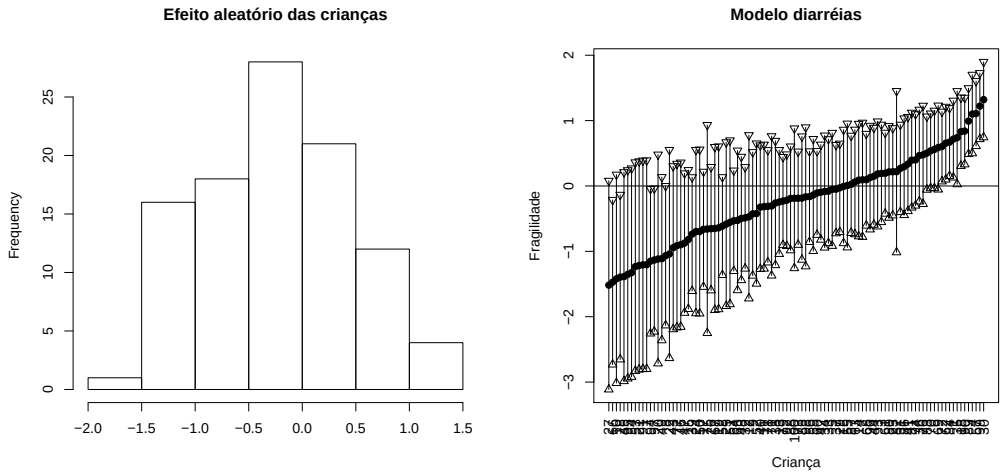
Modelo de Efeitos Aleatórios:
O efeito das duas variáveis é semelhante, mas a formulação correta é: O efeito da vitamina A, **condicionado na fragilidade individual**, não é significativo. Já o efeito da idade é de proteção de cerca de 4%, dada a variabilidade entre os indivíduos.

4. Faça um histograma das fragilidades estimadas. O que ele sugere?

   Resposta: Neste caso é interessante fazer o histograma sobre o risco usando $(exp(fragilidade))$.

   Fazendo os dois gráficos:

```
> hist(modelo.frag$coeff[3:length(modelo.frag$coeff)], xlab = "",
+     main = "Efeito aleatório das crianças")
> plot.frail(diar$numcri, modelo.frag, xlab = "Criança", ylab = "Fragilidade",
+     main = "Modelo diarréias")
```



O histograma sugere que algumas crianças apresentam risco muito maior do que a média (até 4 vezes). Por outro lado, o intervalo de confiança das fragilidades individuais diminui à medida que aumenta o risco, exatamente porque têm mais episódios de diarréia.