MAE0514 - Prova 1 - Parte 2

Rubens Santos Andrade Filho[1]

Junho de 2021

Sumário

[1]Número USP: 10370336

Questão 2

Foram considerados os dados provenientes do EBMT (*European Resgistry for Blood and Marrow Transplatation*), discutido em Putter, Fiocco Geskus (2007). Os dados são referentes a 2204 pacientes que reccberam transplante de medula óssea entre 1995 e 1998 reportados ao EBMT. Os dados estão disponíveis no arquivo **bmt3.csv** e a descrição em **bmt3.des**. As variáveis nos dados são:

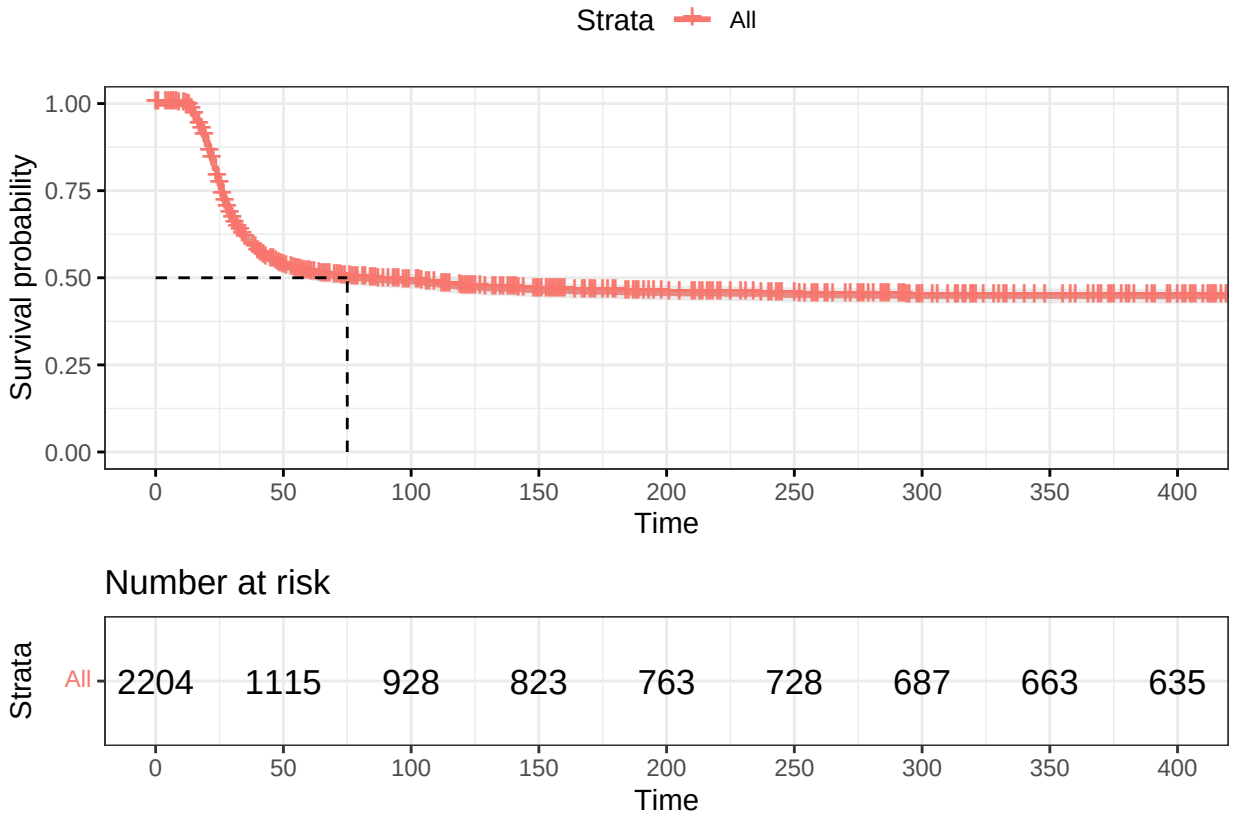
- Tempo, em dias, após o transplante até recuperaço das plaquetas ou perda de acompanhamento
- Indicador de recuperação das plaquetas: 1, se recuperado; 0, se perda de acompanhamento
- Tempo livre de doença: tempo, em dias, após o transplante até óbito ou reincidência do câncer;
- Indicador de evento: 1, se óbito ou reincidência; 0 se censura
- Classificaçao da doença: leucemia linfoblástica aguda (ALL), leucemia mielóide aguda (AML) c leucemia mielóide crônica (CML)
- Idade, em categorias
- Correspondência de gênero doador-receptor
- Esgotamento de células T

Os itens a seguir foram respondidos utilizando esses dados.

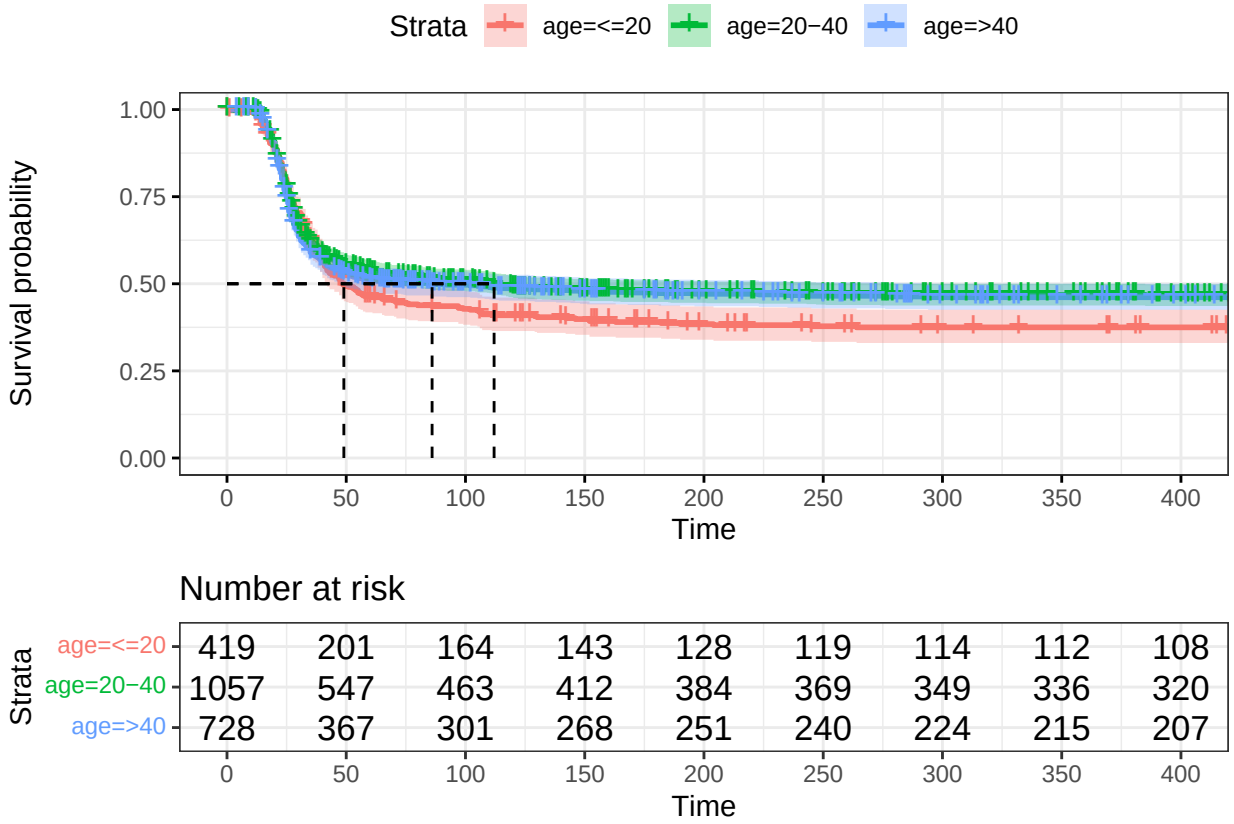
a) Análise descritiva, tempo de recuperação das plaquetas como desfecho

Nesse item, fazemos uma análise descritiva dos dados considerando como desfecho ou variável resposta o tempo até recuperação das plaquetas, `prtime`. Começamos fazendo o gráfico com a estimativa de Kaplan-Meier para a curva geral de sobrevivência do tempo até recuperação das plaquetas.

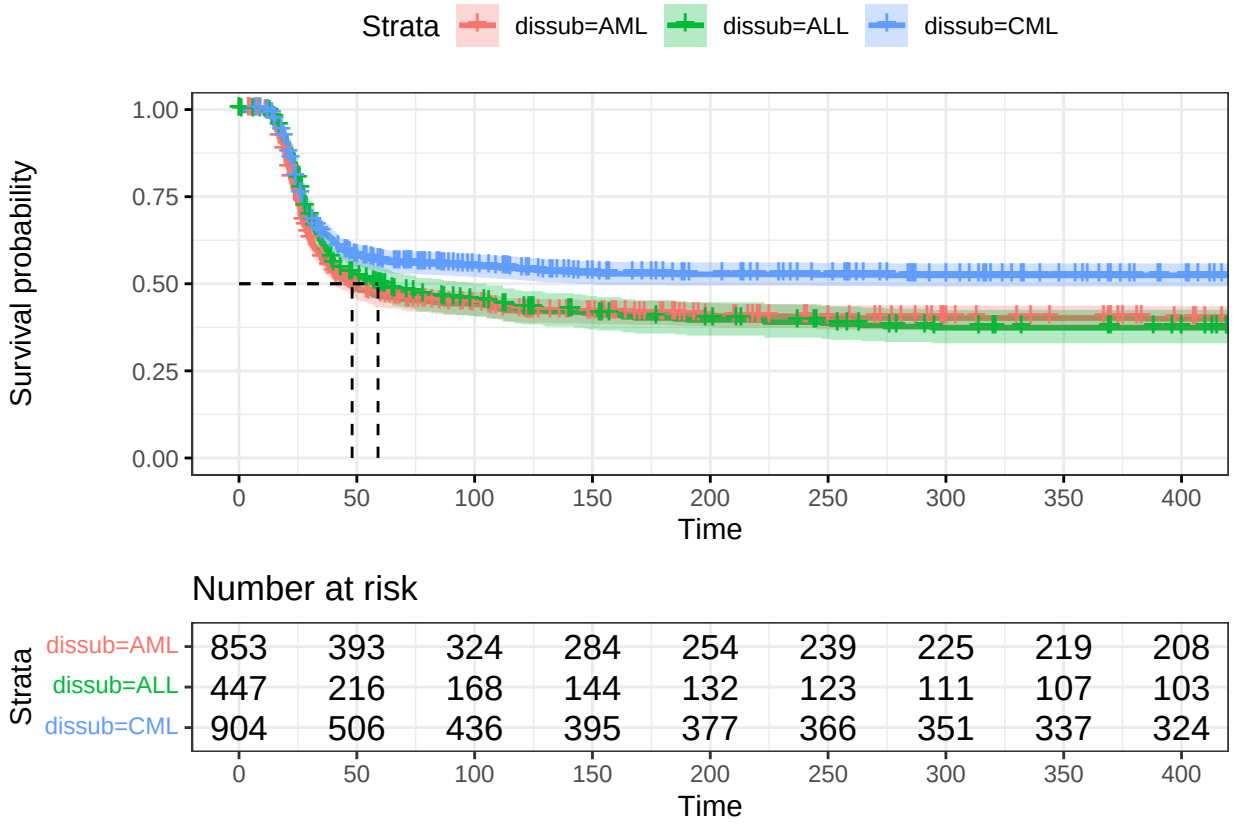
```
## Call: survfit(formula = surv ~ 1, data = dados)
##
##       n  events  median 0.95LCL 0.95UCL
##    2204    1169      75      56     112
```



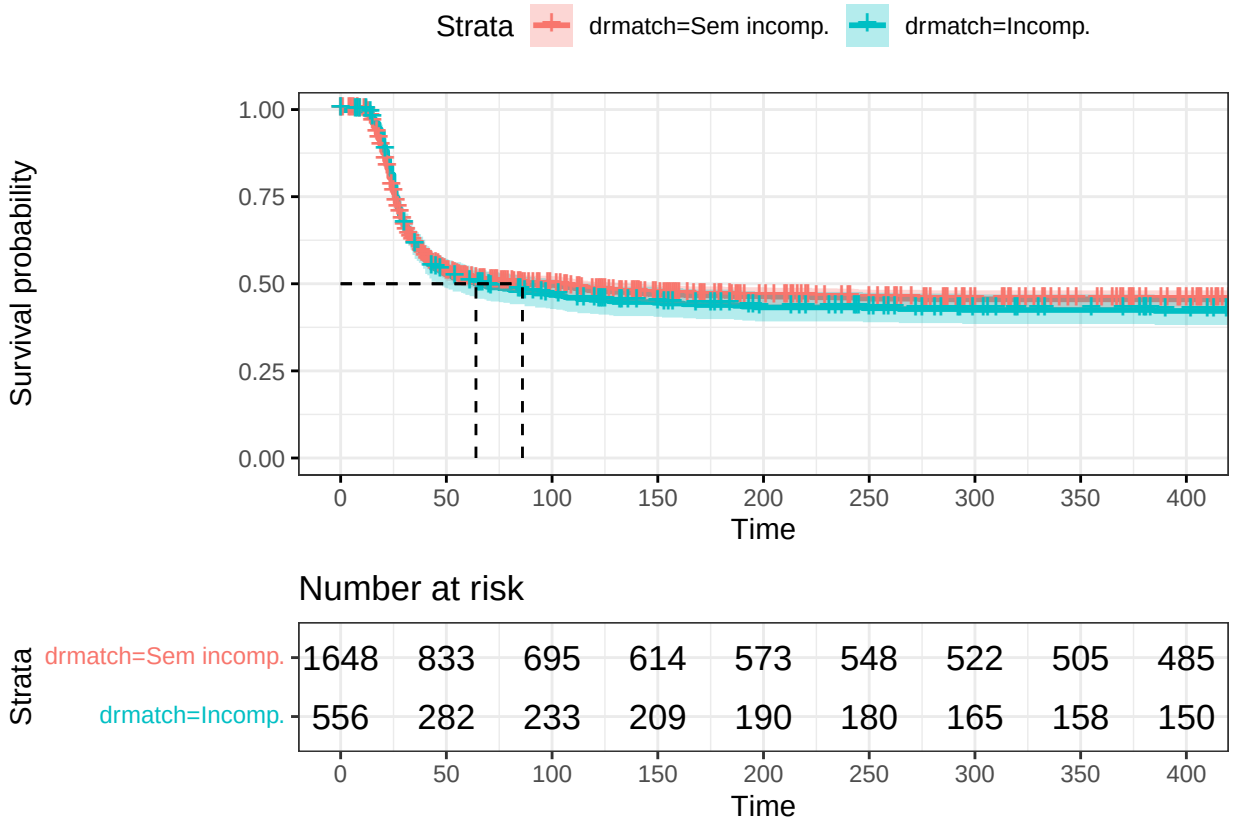

```
## Call: survfit(formula = surv ~ age, data = dados)
##
##               n events median 0.95LCL 0.95UCL
## age=<=20    419    253     49      42      71
## age=20-40  1057    540    112      64     274
## age=>40     728    376     86      48      NA
```



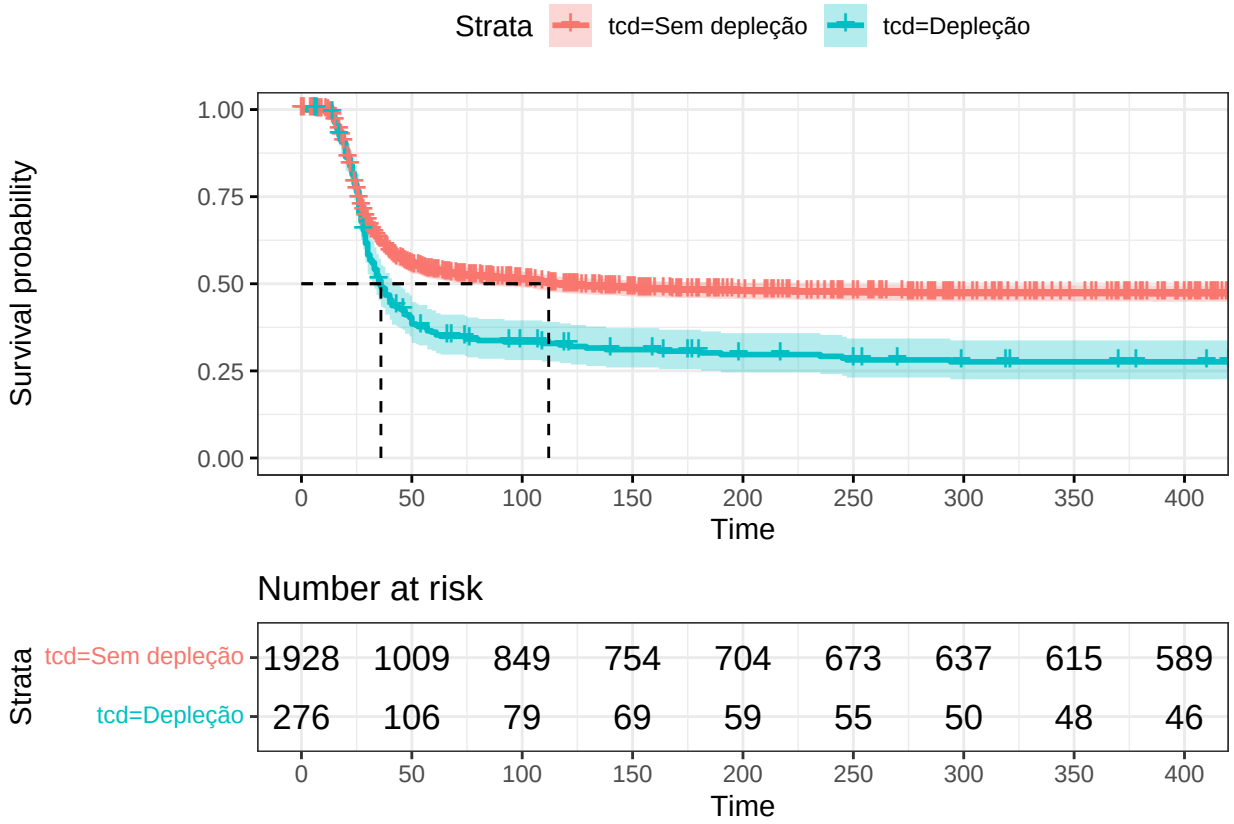

```
## Call: survfit(formula = surv ~ dissub, data = dados)
##
##               n events median 0.95LCL 0.95UCL
## dissub=AML 853    490     48      40      61
## dissub=ALL 447    260     59      42     105
## dissub=CML 904    419     NA     144      NA
```




```
## Call: survfit(formula = surv ~ drmatch, data = dados)
##
##                         n events median 0.95LCL 0.95UCL
## drmatch=Sem incomp. 1648    860     86      55     142
## drmatch=Incomp.      556    309     64      47     118
```




```
## Call: survfit(formula = surv ~ tcd, data = dados)
##
##                      n events median 0.95LCL 0.95UCL
## tcd=Sem depleção 1928    978    112      76     223
## tcd=Depleção      276    191     36      33      43
```



Vemos no primeiro gráfico que a estivativa da curva de sobrevivência decai rapídamente nos primeiros 300 dias para pouco menos de 50%. De fato, observando os dados, vimos que o maior tempo observado de recuperação das plaquetas ocorre no dia 385, com isso estimaviva da curva de sobrevivência não decai mais a partir desse dia. Para observar melhor o que acontece no início da curva, fizemos gráfico, porém, limitando o eixo do tempo até o dia 400. Com isso, podemos ver claramente, seguindo a linha pontilhada, que o tempo de sobrevivência mediano é de 75 dias. Isto é, 75 dias é a estimativa do tempo decorrido após o transplante em que a probabilidade de recuperação das plaquetas é 50%. E uma intervalar de 95% de confiança para o tempo mediano até a recuperação das plaquetas é $[56, 112]$ dias. Além disso, convém notar que o número total de observações é igual a 2204 e o número de recuperações observadas é 1169.

Para a covariável idade, `age`, trata-se da segunda saída do R e gráfico. Construimos o gráfico com as estimativas de Kaplan-Meier para a curva de sobrevivência para cada categoria ("<=20", "20-40" e ">40"). Novamente, para melhor visualizar as curvas, iremos mostrar até o dia 400, uma vez que as estimaivas das curvas são constantes depois disso. O gráfico mostra que os pacientes na faixa etária menor ou igual a 20 anos tem uma aparente menor estimativa da curva de sobrevivencia em relação às outras faixas etárias e em especial após o seu tempo mediano. Antes do tempo mediano, é dificil ver indícios de diferenças entre as curvas. Além disso, as estimativas intervalares (com confiança de 95%) para o tempo mediano se sobrepõem, indicando não haver diferença estatística entre os tempos medianos.

Com relação à Classificaçao da doença, `dissub`, Não parece haver indícios de diferença entre as curvas de sobrevivência para as classificações de leucemia linfoblástica aguda (ALL) e leucemia mielóide aguda (AML). Entretanto, a curva para os pacientes com leucemia mielóide crônica (CML) apresenta uma aparente maior probabilidade de sobrevivência ao longo do tempo. Inclusive nem foi possível estimar o tempo mediano de sobrevivência para esse grupo. Já as estimativas do tempo mediano de sobrevivência, isto é, o tempo no qual metade dos pacientes já recuperaram as plaquetas, são bem próximas para os pacientes com leucemia linfoblástica aguda (ALL) e leucemia mielóide aguda (AML), 48 e 59 respecitamente. A terceira

saída do R também mostra as estimativas intervalares com 95% de confiança.

Já para a covariável de Correspondência de gênero, `drmatch`, não parece haver diferenças entre as curvas de sobrevivência estimadas por Kaplan-Meier. As estimativas pontuais do tempo de sobrevivência mediano também ficaram bem próximas, 86 e 64 dias para pacientes sem e com incompatibilidade de gênero, respectivamente. E com as estimativas intervalares se sobrepondo.

Por fim, quanto a Depleção de células T, `tcd`, observamos que existem indícios de diferença entre as curvas de sobrevivência estimadas. O pacietes com depleção de células T apresentam, ao longo do tempo, uma menor sobrevivência, iso é, um menor tempo de recuperação das plaquetas. Isso também é evidenciado pelo tempo mediano de sobrevivência: a estimativa do tempo até metade dos paciente sem depleção de células T recuperarem as plaquetas é de 112 dias, enquanto que nos pacientes com depleção, essa estimativa é de 36 dias.

b) Testes de log-rank, tempo de recuperação das plaquetas como desfecho

Fizemos os testes de log-rank comparando as categorias das covariáveis no estudo. No teste de log-rank, a hipótese nula é que as funções de sobrevivência em cada nível de um fator são iguais em todos os tempos. A hipótese alternativa é de que pelo menos uma função é diferente para algum tempo dentro de um intervalo de zero a um tempo razoável estabelecido.

```
## Call:
## survdiff(formula = surv ~ age, data = dados)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## age=<=20   419      253      221    4.5047    5.6627
## age=20-40 1057      540      569    1.4323    2.8400
## age=>40    728      376      379    0.0245    0.0369
##
##  Chisq= 6.1  on 2 degrees of freedom, p= 0.05
```

```
## valor p exato
##    0.04797336
```

```
## Call:
## survdiff(formula = surv ~ dissub, data = dados)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## dissub=AML 853      490      428      9.10     14.62
## dissub=ALL 447      260      235      2.64      3.37
## dissub=CML 904      419      506     15.05     27.08
##
##  Chisq= 27.3  on 2 degrees of freedom, p= 1e-06
```

```
## Call:
## survdiff(formula = surv ~ drmatch, data = dados)
##
##                         N Observed Expected (O-E)^2/E (O-E)^2/V
## drmatch=Sem incomp. 1648      860      870     0.120      0.48
## drmatch=Incomp.      556      309      299     0.351      0.48
##
```

```
##  Chisq= 0.5  on 1 degrees of freedom, p= 0.5
```

```
## Call:
## survdiff(formula = surv ~ tcd, data = dados)
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## tcd=Sem depleção 1928      978     1037      3.32      29.9
## tcd=Depleção      276      191      132     25.97      29.9
##
##  Chisq= 29.9  on 1 degrees of freedom, p= 5e-08
```

Para a comparação entre as diferentes faixas etárias da covariável Idade, rejeitamos a hipótese de nula, a um nível de significância de 5% (valor p = 0.048). Observamos o valor p ficou bem próximo do nível de significância e também pelo gráfico, vemos que essa diferença é discutível se notarmos que as curvas para as diferentes faixas etárias aparentam se cruzarem, diminuindo o poder do teste. Ademais, como podemos ver no grafico da curva de sobrevivência para as diferentes faixas etárias e também no resumo do teste log rank, podemos dizer que a curva de sobrevivência dos pacientes com até 20 anos é diferente das demais faixas, que não aparentam diferença.

Quanto à classificação da doença, a um nível de significância de 5%, rejeitamos a hipótese de nula (valor p = 1e-06), isto é, podemos dizer que existe pelo menos alguma diferença entre os tempos até a recuperação das plaquetas entre os níveis da covariável. Em particular, vemos pelo gráfico das curvas de sobrevivência e pela saída do teste que os pacientes com CML, leucemia mielóide crônica, indicam ter um diferente tempo até a recuperação das plaquetas em relação aos outros dois grupos, que não apresentam aparente diferença.

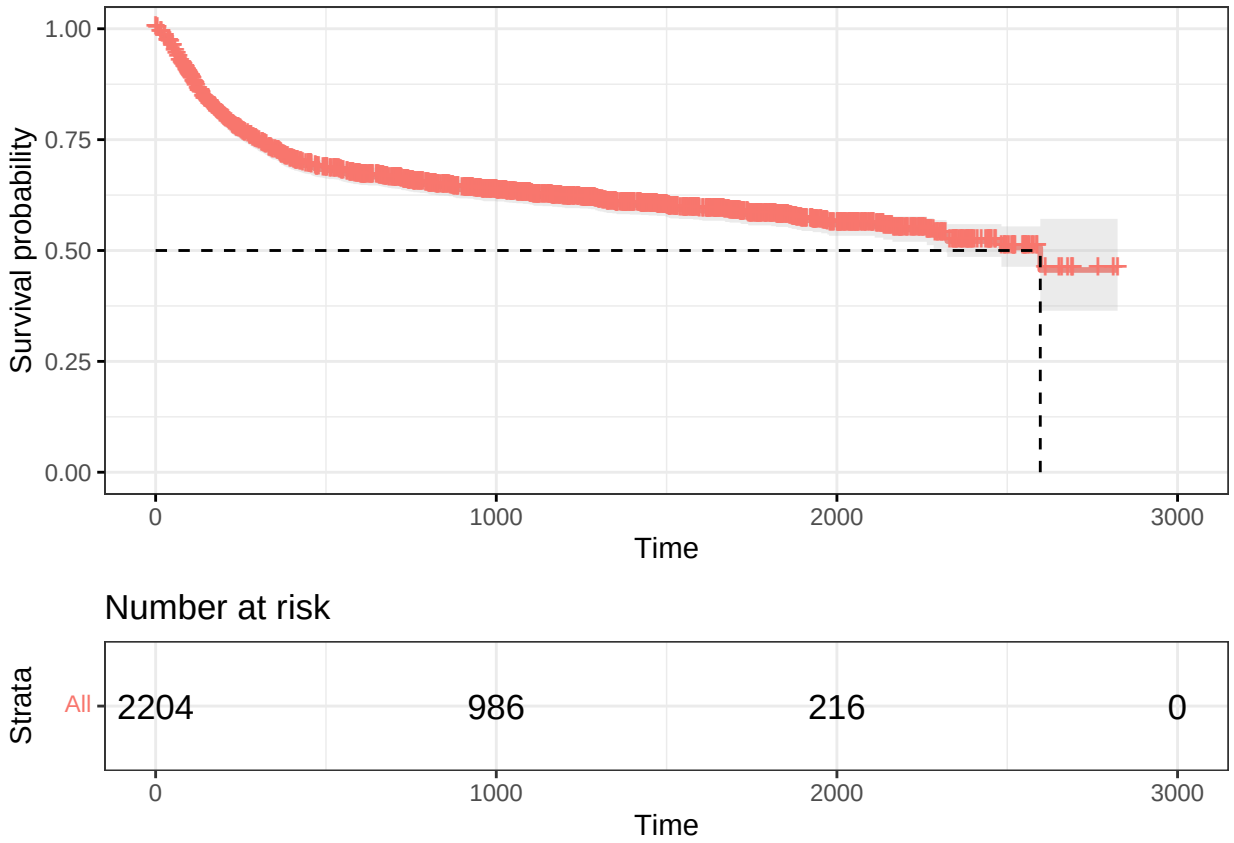
Em vista da depleção ou não de células T, encontramos um valor p igual a 5e-08, rejeitando, dessa forma, a hipótese nula a um nível de significância de 5%. A diferença entre os grupos é bem clara no gráfico e entre os valores observados e esperados. Desse modo, podemos dizer que os pacientes com depleção de céculas T tiveram um menor tempo até a recuperação das plaquetas e a diferença é estatísticamente significativa a 5%.

Por fim, não rejeitamos a hipótese nula quando comparamos os níveis de Correspondência de gênero, que, olhando a saída do teste, apresentam valores observados bem próximos dos esperados. Os testes confirmaram aquilo que pudemos ver nos graficos com as estimativas da curvas de sobreviência para cada nível das covariáveis.

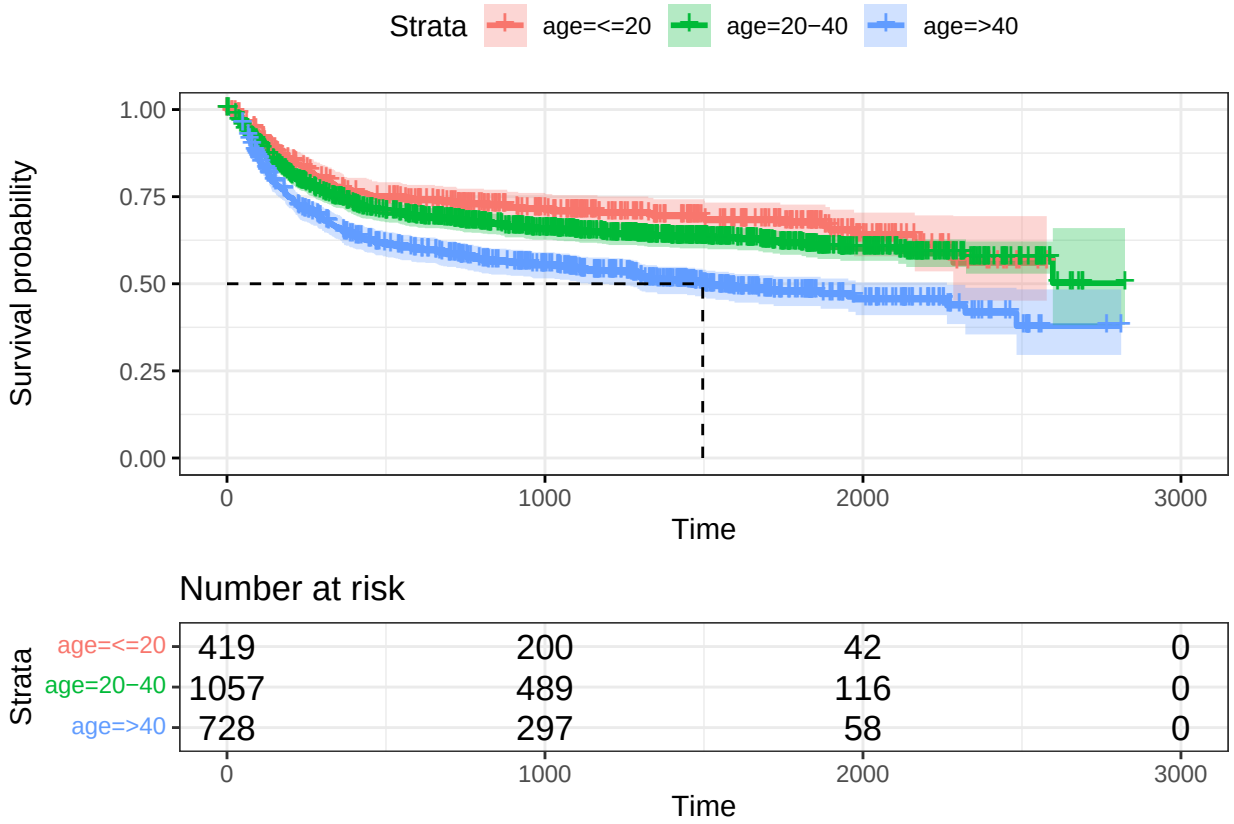
c) Análise descritiva, tempo livre da doença como desfecho

De maneira similar ao item a), fazemos uma análise descritiva dos dados considerando como desfecho ou variável resposta o tempo livre da doença `rfstime`. Começamos fazendo o gráfico com a estimativa de Kaplan-Meier para a curva de sobrevivência do tempo livre da doença. Também mostramos as estimativas pontuais e intervalares do tempo mediano livre da doença, tanto no geral, quanto para os diferentes níveis de cada covariável.

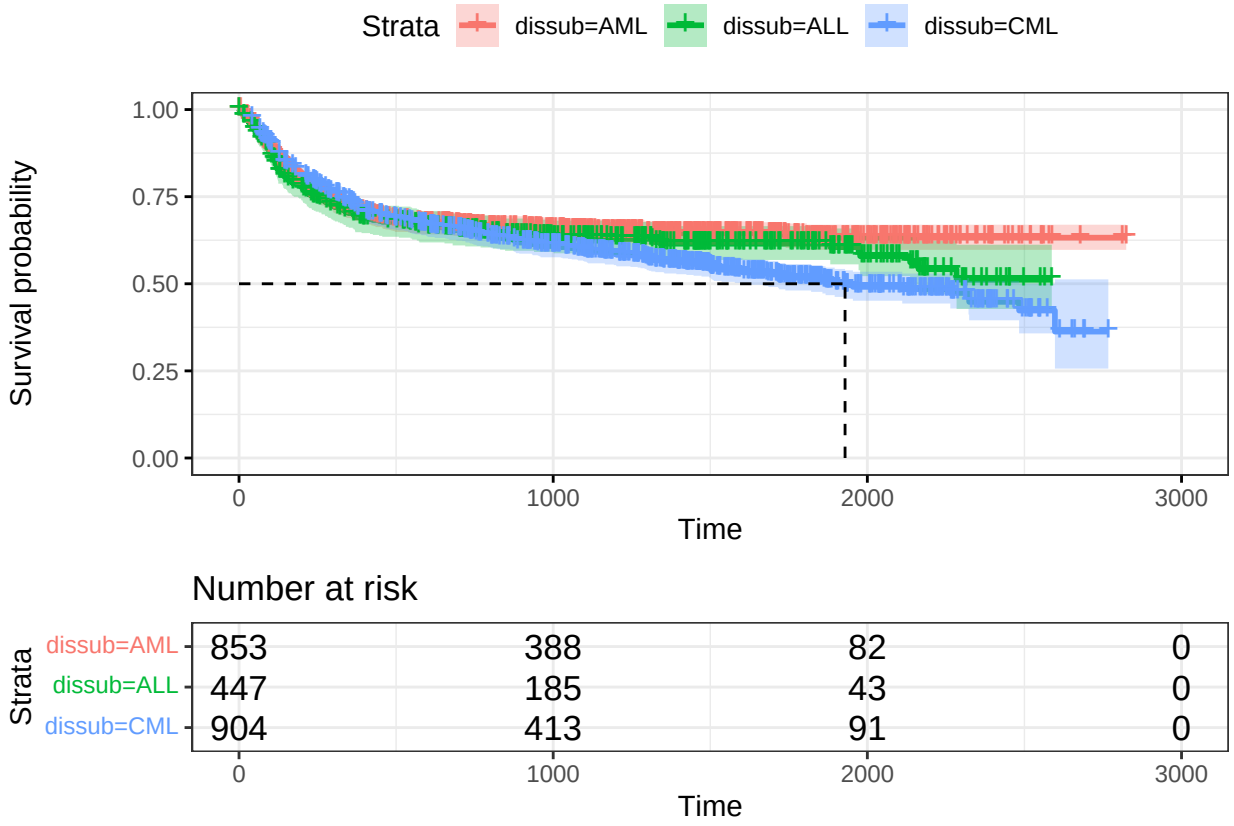
```
## Call: survfit(formula = surv ~ 1, data = dados)
##
##       n  events  median 0.95LCL 0.95UCL
##    2204     841    2597    2322      NA
```



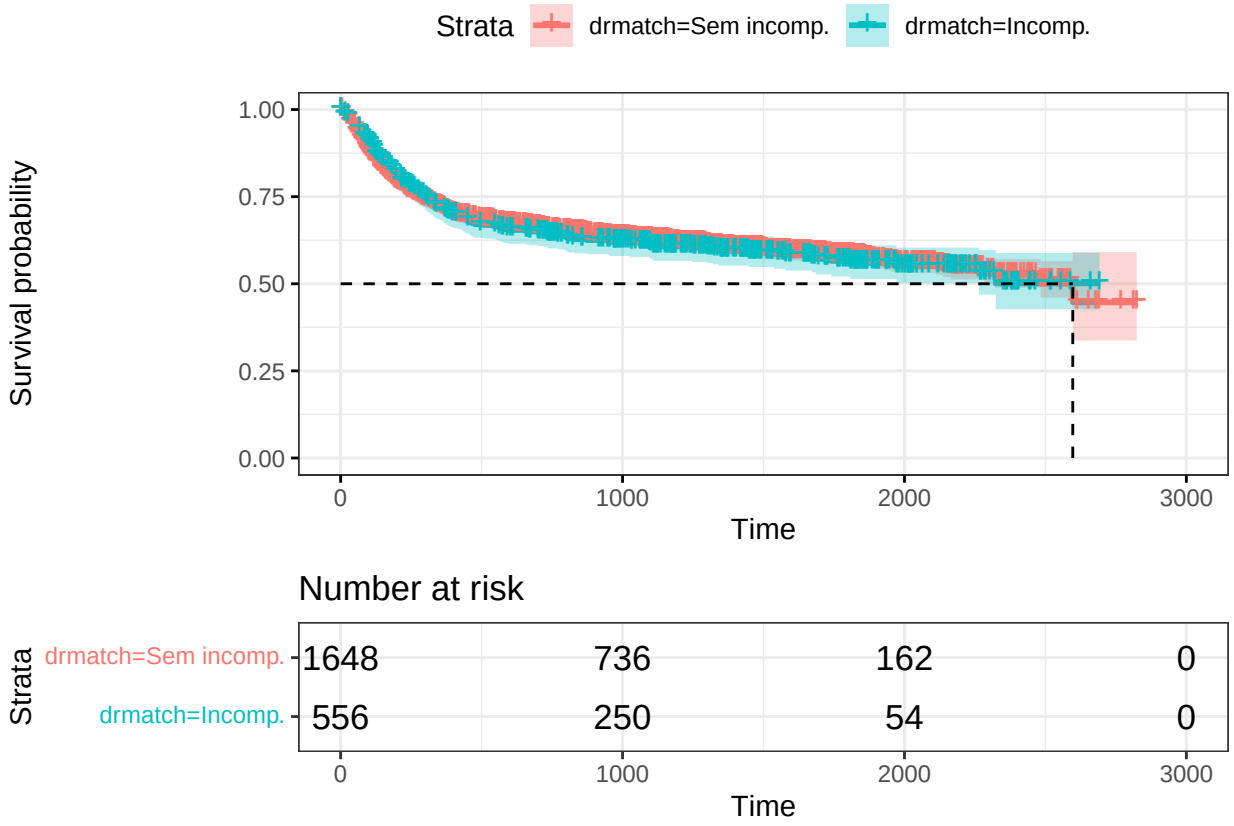

```
## Call: survfit(formula = surv ~ age, data = dados)
##
##               n events median 0.95LCL 0.95UCL
## age=<=20    419    124     NA    2284      NA
## age=20-40  1057    371     NA    2597      NA
## age=>40     728    346   1496    1136    2264
```



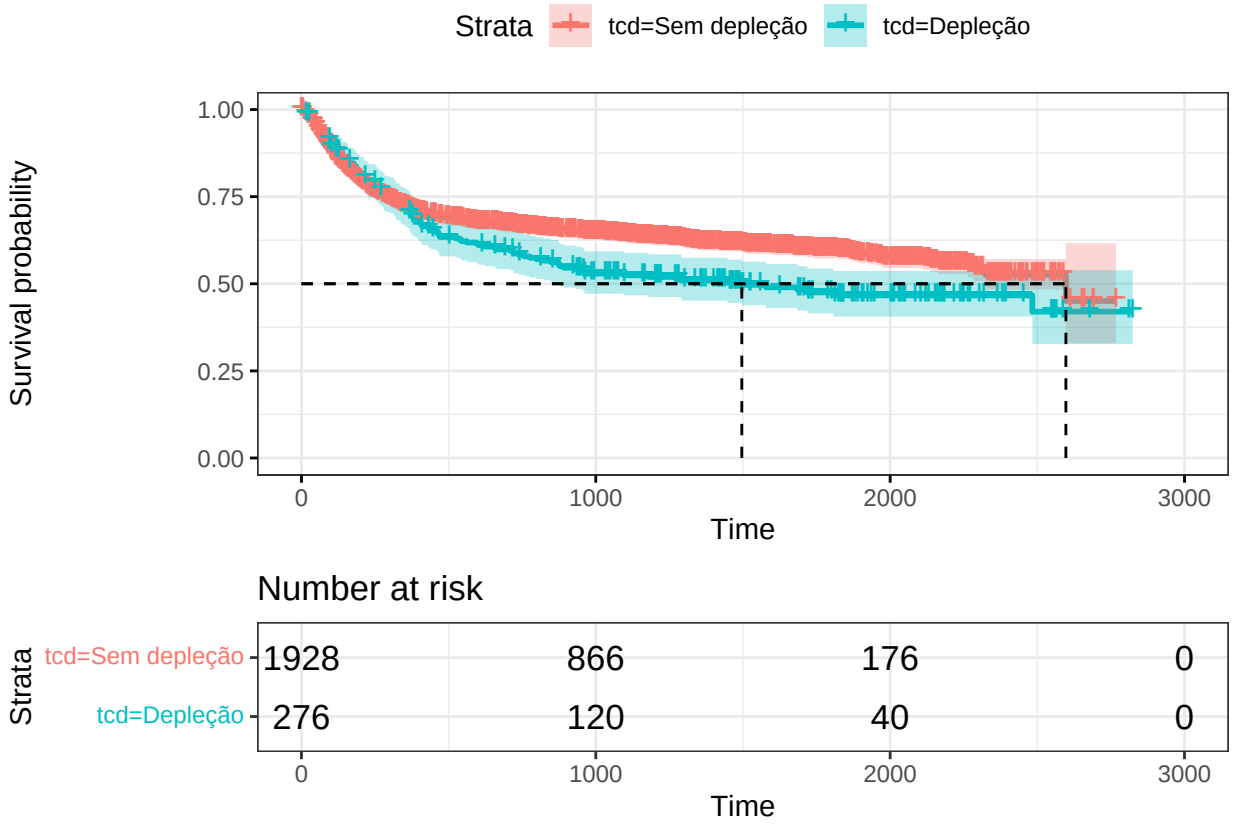

```
## Call: survfit(formula = surv ~ dissub, data = dados)
##
##               n events median 0.95LCL 0.95UCL
## dissub=AML 853    285     NA      NA      NA
## dissub=ALL 447    164     NA    2136      NA
## dissub=CML 904    392   1929    1596      NA
```




```
## Call: survfit(formula = surv ~ drmatch, data = dados)
##
##                         n events median 0.95LCL 0.95UCL
## drmatch=Sem incomp. 1648    625   2597    2322      NA
## drmatch=Incomp.      556    216     NA    2264      NA
```




```
## Call: survfit(formula = surv ~ tcd, data = dados)
##
##                       n events median 0.95LCL 0.95UCL
## tcd=Sem depleção 1928    706   2597    2322      NA
## tcd=Depleção      276    135   1496     873      NA
```



Considerando as diferentes faixas etárias, nota-se que o grupo com idade maior ou igual a 40 anos apresenta uma menor probabilidade de sobrevivencia em comparação aos outros grupos. Nota-se ainda que as curvas para os grupos com até 20 anos e entre 20 e 40 são bem próximas e até se cruzam. Esse cruzamento pode diminuir o poder do teste log-rank pela maneira como é definida a estatística do teste.

Observando agora o gráfico com as estimativas de Kaplan-Meier para as três classificações de doenças, nota-se que as curvas são bem próximas nos primeiros 1000 dias, mas com o passar do tempo aparecem diferenças, em especial entre os grupos com AML e CML. Entretando , nesse gráfico também, podemos notar que as curvas se sobrepõem e se cruzam.

Para a correspondencia de gênero, não parece haver diferenças entre as estimativas das curvas de sobrevivência. E por fim, com relação à depleção de células T, nota-se que as curvas se diferenciam principalmente no centro. É interessante notar também que em vários grupos não foi possível encontrar o tempo mediano livre da doença.

d) Testes de log-rank, tempo livre da doença como desfecho

Fizemos os testes de log-rank comparando as curvas de sobrevivência das categorias para cada covariável no estudo. Consideramos também diferentes ponderações, além do peso igual a 1 que é o caso do teste log-rank.

```
##   variable        method        pval
## 1      age      Log-rank 1.08622e-09
## 2      age Gehan-Breslow 0.00000e+00
```

```
## 3      age     Peto-Peto 0.00000e+00
```

Comparando os grupos etários, obtemos p-valores significativos, a um nível de 5%, em todos os testes. Dessa forma, rejeitamos a hipótese nula, isto, concluimos que existe pelo menos uma diferença nas curvas de sobrevivência.

```
##   variable        method       pval
## 1   dissub      Log-rank 0.01828238
## 2   dissub Gehan-Breslow 0.50190000
## 3   dissub     Peto-Peto 0.14410000
```

Agora, testando as curvas dos níveis da classificação da doença, obtemos valores p bem diferentes entre si. O teste de log-rank rejeita a hipótese nula a um nível de 5%. O teste de Gehan-Breslow é melhor para detectar diferenças no início das curvas, e como podemos ver no gráfico, no inicio não existe tanta diferença, isso explica o valor p de 0.5. Enquanto que no teste modificado Peto-Peto, obtemos um menor valor p, mas ainda não significativo a um nível de 5% e, com isso, não rejeitamos a hipótese nula.

```
##   variable        method      pval
## 1  drmatch      Log-rank 0.8360478
## 2  drmatch Gehan-Breslow 0.8948000
## 3  drmatch     Peto-Peto 0.9500000
```

Comparando a compatibildade de gênero, em todos os testes, rejeitamos não rejeitamos a hipótese nula. Ou seja, não parece haver diferença nas curvas de sobrevivência do tempo livre de doença.

```
##   variable        method        pval
## 1      tcd      Log-rank 0.007470226
## 2      tcd Gehan-Breslow 0.063200000
## 3      tcd     Peto-Peto 0.035100000
```

Por fim, quanto à depleção de células T, obtemos um valor p bem significante no teste de log-rank, rejeitando a hipótese nula a um nível de 5%, assim como no teste de Peto-Peto modificado com um valor p de 0.0351. Mas encontrando um valor p não significativo no teste de Gehan-Breslow, a 5%. Ainda assim, pela pelo fato do teste de Peto-Peto modificado ser mais robusto e pela análise descritiva, concluímos que existe diferença nas curvas de sobrevivencia para o tempo livre de doença no que diz respeito à depleção de células T.

e) Modelo Weibull, tempo livre da doença como desfecho

Considerando o tempo livre da doença, ajustamos um modelo Weibull aos dados. A seguir, apresentamos os resultados do modelo completo, com todas as covariáveis incluídas.

```
##
## Call:
## survreg(formula = surv ~ dissub + age + drmatch + tcd, data = dados %>%
##     mutate(age = as.character(age)), dist = "weibull")
##                  Value Std. Error     z       p
## (Intercept)     9.0513     0.1899 47.67 < 2e-16
```

```
## dissubALL      -0.3592     0.1700 -2.11   0.035
## dissubCML      -0.2645     0.1355 -1.95   0.051
## age>40         -0.9660     0.1925 -5.02 5.2e-07
## age20-40       -0.3089     0.1827 -1.69   0.091
## drmatchIncomp. -0.0746     0.1349 -0.55   0.580
## tcdDepleção    -0.3309     0.1626 -2.04   0.042
## Log(scale)      0.5323     0.0308 17.29 < 2e-16
##
## Scale= 1.7
##
## Weibull distribution
## Loglik(model)= -7183.3   Loglik(intercept only)= -7209.5
##  Chisq= 52.26 on 6 degrees of freedom, p= 1.7e-09
## Number of Newton-Raphson Iterations: 5
## n= 2204
```

Ajustamos o modelo com todas as covariáveis, o modelo "cheio". Em seguida, para cada covariável, ajustamos um modelo sem essa covariável, calculamos a razão de verossimilhança entre esse modelo e o modelo cheio e o valor p do teste.

Supondo que o espaço parâmetrico do modelo cheio é Θ e o espaço parâmetrico do modelo sem uma covariável é $\Theta_0 \in \Theta$, o teste de razão de verossimilhanças testa

$$\begin{aligned} H_0: &\quad \theta \in \Theta_0 \\ H_1: &\quad \theta \notin \Theta_0 \end{aligned}$$

A estatística do teste de razão de verossimilhanças pode ser escrita como

$$\lambda_{\text{RV}} = -2\left[\ell\left(\theta_0\right) - \ell(\hat{\theta})\right] \to \chi^2(p)$$

onde $\ell\left(\theta_0\right)$ é a máxima log-verossimilhança do modelo sem a covariável, $\ell(\hat{\theta})$ é a máxima log-verossimilhança do modelo cheio e p é a diferença de dimensionalidade entre os espaços paramétricos.

Por fim, retiramos as covariável cujo teste não foi signifcativo a 5%.

Tabela 1: Teste de RV entre o modelo cheio e um modelo removendo cada covariavel separadamente.

Removendo	RV	Valor-p
age	36.0364211	0.0000000
tcd	3.9945328	0.0456481
dissub	5.9655426	0.0506523
drmatch	0.3036512	0.5816025

Com isso, as variáveis que ficaram no modelo final foram `age` e `tcd`.

```
##
## Call:
## survreg(formula = surv ~ age + tcd, data = dados %>% mutate(age = as.character(age)),
##     dist = "weibull")
##               Value Std. Error     z       p
```

```
## (Intercept)  8.8371     0.1632 54.14 < 2e-16
## age>40      -0.9553     0.1815 -5.26 1.4e-07
## age20-40    -0.3045     0.1778 -1.71   0.087
## tcdDepleção -0.3190     0.1617 -1.97   0.049
## Log(scale)   0.5323     0.0308 17.29 < 2e-16
##
## Scale= 1.7
##
## Weibull distribution
## Loglik(model)= -7186.5   Loglik(intercept only)= -7209.5
##  Chisq= 45.85 on 3 degrees of freedom, p= 6.1e-10
## Number of Newton-Raphson Iterations: 5
## n= 2204
```

f) Interpretação do modelo final em (e).

Se $T \sim \text{Weibull}(\lambda, \rho)$, a função de taxa de falha é dada por

$$\alpha(t) = \lambda \rho t^{\rho - 1}$$

Devido à parametrização implementada em `survreg`, na saída do `R`, obtemos as estimativas $\hat{\gamma}$ e $\hat{\sigma}$. Estamos interessados em $\hat{\beta} = -\hat{\gamma}/\hat{\sigma}$ e $\hat{\rho} = 1/\hat{\sigma}$. Da saida do modelo, obtemos que $\hat{\rho} = 1/\hat{\sigma} \approx 0.587$ e calculando $\hat{\beta}$ obtemos

```
## (Intercept)      age>40    age20-40 tcdDepleção
##  -5.1895046   0.5610175   0.1788400   0.1873444
```

Dado um vetor de observações de covariáveis $x_i' = (x_{i1}, x_{i2}, ..., x_{ip})'$, temos que $\hat{\lambda_i} = \exp\{x_i'\hat{\beta}\}$ é a estimativa para λ_i e a estimativa da função de taxa de falha é

$$\hat{\alpha}(t \mid x_i) = \hat{\lambda}\hat{\rho}t^{\hat{\rho}-1} = \exp\{x_i'\hat{\beta}\}\hat{\rho}t^{\hat{\rho}-1}$$

Sejam x_j e x_i, tal que $x_{.k}$ seja uma covariável binária com $x_{jk} = 1, x_{ik} = 0$ para um certo k e $x_{jl} = x_{il}$ para $l \neq k$. Para interpretar as estimativas do modelo, consiredemos a razão de riscos, ou razão de taxas de falha, entre x_j e x_i,

$$\frac{\hat{\alpha}(t \mid x_j)}{\hat{\alpha}(t \mid x_i)} = \frac{\exp\{x_j'\hat{\beta}\}\hat{\rho}t^{\hat{\rho}-1}}{\exp\{x_i'\hat{\beta}\}\hat{\rho}t^{\hat{\rho}-1}} = \exp\{(x_j - x_i)'\hat{\beta}\} = \exp\{\hat{\beta}_k\}$$

Isso nos mostra que, se $\hat{\beta}_k > 0$ (< 0), o risco do evento aumenta (diminui) em $\exp\{\hat{\beta}\}$ vezes quando a covariável k ocorre ($x_{.k} = 1$) em relação a quando não ocorre ($x_{.k} = 0$), mantendo constantes as outras covariáveis.

Com isso, calculamos $\exp\{\hat{\beta}\}$,

```
## (Intercept)      age>40    age20-40 tcdDepleção
## 0.005574768 1.752454632 1.195829377 1.206042596
```

Isso significa que (lembrando que o `R` adota o estilo de casela de refência):

- Um paciente com mais de 40 anos tem um risco de óbito ou reincidência da doença 75% maior que um paciente com até 20 anos, mantendo-se as outras covariáveis constantes, no caso. `tcd`.
- Um paciente com idade entre 20 e 40 anos tem um risco de óbito ou reincidência da doença 20% maior que um paciente com até 20 anos, mantendo-se as outras covariáveis constantes, no caso. `tcd`.
- De forma análoga, um paciente com depleção de células T tem um risco de óbito ou reincidência da doença 21% maior que um paciente sem depleção de células T, mantendo-se as outras covariáveis constantes, no caso, a idade, `age`.

g) Modelo log-logístico, tempo livre da doença como desfecho

Considerando o tempo livre da doença, ajustamos um modelo log-logístico aos dados. A seguir, apresentamos os resultados do modelo completo, com todas as covariáveis incluídas.

```
##
## Call:
## survreg(formula = surv ~ dissub + age + drmatch + tcd, data = dados %>%
##     mutate(age = as.character(age)), dist = "loglogistic")
##                   Value Std. Error     z       p
## (Intercept)      8.4139     0.1892 44.47 < 2e-16
## dissubALL       -0.3814     0.1794 -2.13   0.034
## dissubCML       -0.1349     0.1429 -0.94   0.345
## age>40          -1.0871     0.1987 -5.47 4.5e-08
## age20-40        -0.3912     0.1862 -2.10   0.036
## drmatchIncomp. -0.0470     0.1432 -0.33   0.743
## tcdDepleção     -0.2663     0.1777 -1.50   0.134
## Log(scale)       0.3665     0.0298 12.32 < 2e-16
##
## Scale= 1.44
##
## Log logistic distribution
## Loglik(model)= -7153.7   Loglik(intercept only)= -7177.6
##  Chisq= 47.8 on 6 degrees of freedom, p= 1.3e-08
## Number of Newton-Raphson Iterations: 4
## n= 2204
```

A seguir a tabela com os valores p dos teste de razão de verossimilhança.

Tabela 2: Teste de RV entre o modelo cheio e um modelo removendo cada covariavel separadamente.

Removendo	RV	Valor-p
age	38.3826137	0.0000000
dissub	4.5069757	0.1050322
tcd	2.2264103	0.1356687
drmatch	0.1073288	0.7432060

Com isso, a única variável que ficou no modelo final foi `age`.

```
##
## Call:
## survreg(formula = surv ~ age, data = dados %>% mutate(age = as.character(age)),
##     dist = "loglogistic")
##               Value Std. Error     z       p
## (Intercept)  8.2097     0.1608 51.04 < 2e-16
## age>40      -1.0537     0.1857 -5.67 1.4e-08
## age20-40    -0.3736     0.1801 -2.07   0.038
## Log(scale)   0.3671     0.0297 12.35 < 2e-16
##
## Scale= 1.44
##
## Log logistic distribution
## Loglik(model)= -7157.2   Loglik(intercept only)= -7177.6
##  Chisq= 40.89 on 2 degrees of freedom, p= 1.3e-09
## Number of Newton-Raphson Iterations: 4
## n= 2204
```

h) Interpretação do modelo final em (g).

Se $T \sim$ log-logística(λ, ρ), a função de taxa de falha é dada por

$$\alpha(t) = \frac{\lambda\rho t^{\rho-1}}{1+\rho t^{\rho}}$$

Com a parametrização implementada em `survreg`, na saída do `R`, obtemos as estimativas $\hat{\gamma}$ e $\hat{\sigma}$. No caso do modelo log-logístico, também $\hat{\beta} = -\hat{\gamma}/\hat{\sigma}$ e $\hat{\rho} = 1/\hat{\sigma}$. Da saida do modelo, obtemos que $\hat{\rho} = 1/\hat{\sigma} \approx 0.693$ e calculando $\hat{\beta}$ obtemos

```
## (Intercept)      age>40    age20-40
##  -5.6873833   0.7299988   0.2588065
```

Dado um vetor de observações de covariáveis $x_i' = (x_{i1}, x_{i2}, ..., x_{ip})'$, temos que $\hat{\lambda}_i = \exp\{x_i'\hat{\beta}\}$ é a estimativa para λ_i e a estimativa da função de taxa de falha é

$$\hat{\alpha}(t \mid x_i) = \frac{\hat{\lambda}\hat{\rho}t^{\hat{\rho}-1}}{1+\hat{\rho}t^{\hat{\rho}}} = \frac{\exp\{x_i'\hat{\beta}\}\hat{\rho}t^{\hat{\rho}-1}}{1+\hat{\rho}t^{\hat{\rho}}}$$

Sejam x_j e x_i, tal que $x_{.k}$ seja uma covariável binária com $x_{jk} = 1, x_{ik} = 0$ para um certo k e $x_{jl} = x_{il}$ para $l \neq k$. Para interpretar as estimativas do modelo, consiredemos a razão de riscos, ou razão de taxas de falha, entre x_j e x_i,

$$\frac{\hat{\alpha}(t \mid x_j)}{\hat{\alpha}(t \mid x_i)} = \frac{\frac{\exp\{x_j'\hat{\beta}\}\hat{\rho}t^{\hat{\rho}-1}}{1+\hat{\rho}t^{\hat{\rho}}}}{\frac{\exp\{x_i'\hat{\beta}\}\hat{\rho}t^{\hat{\rho}-1}}{1+\hat{\rho}t^{\hat{\rho}}}} = \exp\{(x_j - x_i)'\hat{\beta}\} = \exp\{\hat{\beta}_k\}$$

Isso nos mostra que, se $\hat{\beta}_k > 0$ (< 0), o risco do evento aumenta (diminui) em $\exp\{\hat{\beta}\}$ vezes quando a covariável k ocorre $(x_{.k} = 1)$ em relação a quando não ocorre $(x_{.k} = 0)$, mantendo constantes as outras covariáveis.

Com isso, calculamos $\exp\{\hat{\beta}\}$,

```
## (Intercept)      age>40    age20-40
## 0.003388448 2.075078130 1.295383134
```

De forma análoga ao item e), iterpretamos que:

- Um paciente com mais de 40 anos tem um risco de óbito ou reincidência da doença 2 vezes maior que um paciente com até 20 anos.
- Um paciente com idade entre 20 e 40 anos tem um risco de óbito ou reincidência da doença 30% maior que um paciente com até 20 anos.

Código Completo

```
library(knitr)
library(tidyverse)
library(dplyr)
library(readr)
library(ggplot2)
library(survival)
library(survminer)
library(gtsummary)

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



# helper com padroes predefinidos
ggsurv <- function(
  fit,
  conf.int = T,
  surv.median.line = "hv",
  ggtheme = theme_bw(),
  xlim=NULL,break.time.by = NULL,
  risk.table = T, tables.height = 0.32,
  legend='top', ...) {

  ggsurvplot(
    fit,conf.int = conf.int,
    surv.median.line = surv.median.line,
    ggtheme = ggtheme,
    xlim=xlim,
    risk.table = risk.table,
```

```
    tables.height = tables.height,
    break.time.by = break.time.by,
    legend=legend,
    ...
  )
}

# QUESTAO 2 ----

dados_raw <- readr::read_csv(
  'ebmt3.csv',
  col_types = readr::cols_only(
    id = col_integer(),
    prtime = col_double(),
    prstat = col_integer(),
    rfstime = col_double(),
    rfsstat = col_integer(),
    dissub = col_factor(c("AML", "ALL", "CML")),
    age = col_factor(levels = c("<=20", "20-40", ">40"), ordered = T),
    drmatch = col_factor(c("No gender mismatch", "Gender mismatch")),
    tcd = col_factor(c("No TCD", "TCD"))
  )
)

labels <- list(
  id="Identificação do paciente",
  prtime="Tempo de recuperação das plaquetas",
  prstat="Indicador de recuperação das plaquetas",
  rfstime="Tempo livre de doença",
  rfsstat="Indicador de evento",
  dissub="Subclassificação da doença",
  age="Idade",
  drmatch="Correspondência de gênero",
  tcd="Depleção de células T"
)

labelled::var_label(dados_raw) <- labels

dados_raw %>% str

# traduz fatores
dados <- dados_raw %>%
  mutate(
    #dissub = factor(dissub, c("AML", "ALL", "CML")),
    # age = forcats::fct_recode(
    #   age, "Até 20"="<=20", "Entre 20 e 40"="20-40", "Mais que 40"=">40"),
    drmatch = forcats::fct_recode(
      drmatch, "Sem incomp."="No gender mismatch",
      "Incomp."="Gender mismatch"),
    tcd = forcats::fct_recode(tcd, "Sem depleção"="No TCD", "Depleção"="TCD")
  )
dados
```

```r
# QUESTÃO 2.a) geral ----
surv <- with(dados,Surv(prtime, prstat))

fit <- survfit(surv~1, dados)
print(fit)
ggsurv(fit, xlim=c(0,400), break.time.by = 50)

# QUESTÃO 2.a) age ----
fit <- survfit(surv~age, dados)
print(fit)
ggsurv(fit, xlim=c(0,400), break.time.by = 50)

# QUESTÃO 2.a) dissub ----
fit <- survfit(surv~dissub, dados)
print(fit)
ggsurv(fit, xlim=c(0,400), break.time.by = 50)

# QUESTÃO 2.a) drmatch ----
fit <- survfit(surv~drmatch, dados)
print(fit)
ggsurv(fit, xlim=c(0,400), break.time.by = 50)

# QUESTÃO 2.a) tcd ----
fit <- survfit(surv~tcd, dados)
print(fit)
ggsurv(fit, xlim=c(0,400), break.time.by = 50)
# QUESTÃO 2.b) ----
(sdiff <- survdiff(surv ~ age, dados))
c("valor p exato" = pchisq(sdiff$chisq, df = 2, lower.tail = F))
survdiff(surv ~ dissub, dados)
survdiff(surv ~ drmatch, dados)
survdiff(surv ~ tcd, dados)
# QUESTÃO 2.c) geral ----
surv <- with(dados,Surv(rfstime, rfsstat))

fit <- survfit(surv~1, dados)
print(fit)
ggsurv(fit, legend = 'none')

# QUESTÃO 2.c) age ----
fit <- survfit(surv~age, dados)
print(fit)
ggsurv(fit)

# QUESTÃO 2.c) dissub ----
fit <- survfit(surv~dissub, dados)
print(fit)
ggsurv(fit)

# QUESTÃO 2.c) drmatch ----
fit <- survfit(surv~drmatch, dados)
print(fit)
ggsurv(fit)
```

```r
# QUESTÃO 2.c) tcd ----
fit <- survfit(surv~tcd, dados)
print(fit)
ggsurv(fit)

# QUESTÃO 2.d) ----

library(survminer)

survfit(surv~age, dados) %>%
  {bind_rows(
    surv_pvalue(., method = "log-rank"),
    surv_pvalue(., method = "gehan-breslow"),
    surv_pvalue(., method = "S1")
  )} %>%
  dplyr::select(variable, method, pval)

survfit(surv ~ dissub, dados) %>%
  {bind_rows(
    surv_pvalue(., method = "log-rank"),
    surv_pvalue(., method = "gehan-breslow"),
    surv_pvalue(., method = "S1")
  )} %>%
  dplyr::select(variable, method, pval)
survfit(surv ~ drmatch, dados) %>%
  {bind_rows(
    surv_pvalue(., method = "log-rank"),
    surv_pvalue(., method = "gehan-breslow"),
    surv_pvalue(., method = "S1"),
  )} %>%
  dplyr::select(variable, method, pval)
survfit(surv ~ tcd, dados) %>%
  {bind_rows(
    surv_pvalue(., method = "log-rank"),
    surv_pvalue(., method = "gehan-breslow"),
    surv_pvalue(., method = "S1")
  )} %>%
  dplyr::select(variable, method, pval)

# QUESTÃO 2.e)
surv <- with(dados, Surv(rfstime, rfsstat))
fitfull <- survreg(surv ~ dissub+age+drmatch+tcd,
                   dados %>% mutate(age=as.character(age)), dist = "weibull")

summary(fitfull)

# teste de RV entre o modelo cheio e um modelo sem a covariavel para cada covariavel
bind_rows(
  anova(update(fitfull, ~ . - dissub), fitfull)[2,],
  anova(update(fitfull, ~ . - age), fitfull)[2,],
  anova(update(fitfull, ~ . - drmatch), fitfull)[2,],
  anova(update(fitfull, ~ . - tcd), fitfull)[2,]) %>%
  dplyr::transmute(
```

```
    Removendo = c("dissub", "age", "drmatch", "tcd"),
    RV = Deviance,
    `Valor-p` = `Pr(>Chi)`
  ) %>%
  arrange(`Valor-p`) %>%
  kable(caption=paste0("Teste de RV entre o modelo cheio e um modelo ",
                       "removendo cada covariavel separadamente."))

fit_final <- update(fitfull, ~ age + tcd)
summary(fit_final)

# QUESTÃO 2.f)
beta_hat = - fit_final$coefficients / fit_final$scale
beta_hat
exp(beta_hat)
# QUESTÃO 2.g)
surv <- with(dados, Surv(rfstime, rfsstat))
fitfull <- survreg(surv ~ dissub+age+drmatch+tcd,
                   dados %>% mutate(age=as.character(age)), dist = "loglogistic")

summary(fitfull)

# teste de RV entre o modelo cheio e um modelo sem a covariavel para cada covariavel
bind_rows(
  anova(update(fitfull, ~ . - dissub), fitfull)[2,],
  anova(update(fitfull, ~ . - age), fitfull)[2,],
  anova(update(fitfull, ~ . - drmatch), fitfull)[2,],
  anova(update(fitfull, ~ . - tcd), fitfull)[2,]) %>%
  dplyr::transmute(
    Removendo = c("dissub", "age", "drmatch", "tcd"),
    RV = Deviance,
    `Valor-p` = `Pr(>Chi)`
  ) %>%
  arrange(`Valor-p`) %>%
  kable(caption="Teste de RV entre o modelo cheio e um modelo removendo cada covariavel separadamente.")

fit_final <- update(fitfull, ~ age)
summary(fit_final)

# QUESTÃO 2.h)
beta_hat = - fit_final$coefficients / fit_final$scale
beta_hat

exp(beta_hat)
```