Lista 3 - MAE0514

Guilherme NºUSP: 8943160 e Leonardo NºUSP: 9793436

Exercício 1

Em um estudo com pacientes com mieloma múltiplo, pacientes foram tratados com tratamento padrão e os tempos de sobrevivência dos 25 pacientes, contados a partir do início do tratamento, estão disponíveis na tabela a seguir (tempos censurados à direita são denotados por um sinal "+"). Suponha que deseja-se fazer um estudo para comparar o tratamento padrão com um novo tratamento. Esse novo estudo terá um período de acompanhamento total igual a 4 anos e meio e espera-se que a mortalidade inicial seja reduzida e, após um ano e meio de tratamento, ainda se tenha em torno de 65% de pacientes vivos.

Tempo de sobrevivência, em anos												
0.3	5.9	20.8	28.0^+	1.7	73.6^+	7.2	2.1	6.4	2.5	2.3	0.3	0.4
65.4^+	64.9^+	0.6	23.0	42.6^+	48.0^+	6.9	2.1	43.6^+	42.6	12.0^+	0.8	

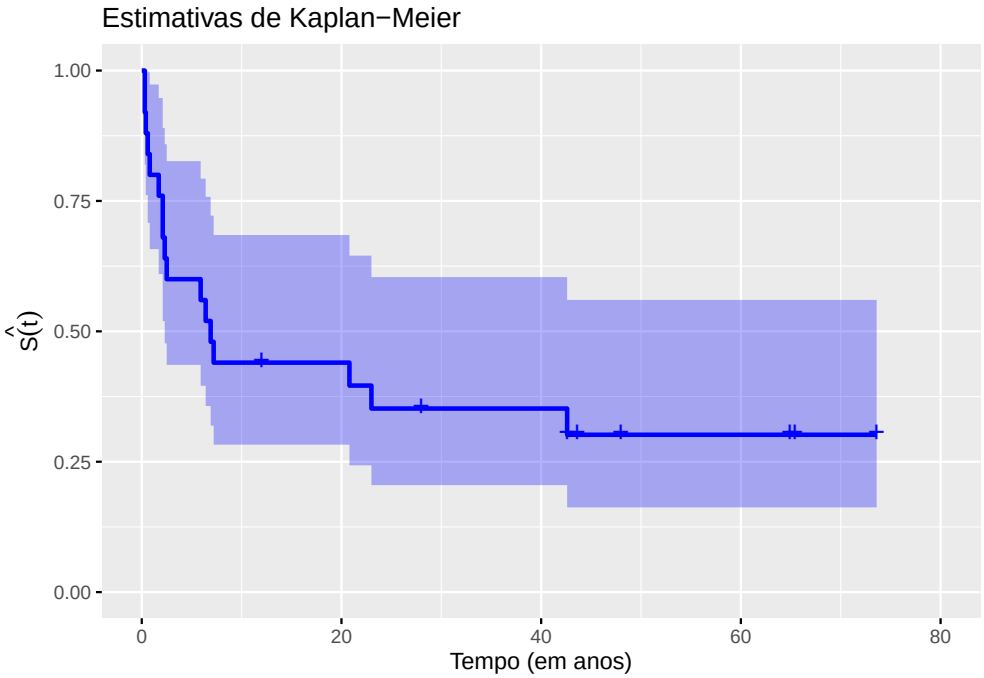
(a) Obtenha o tamanho da amostra necessário se o recrutamento for feito por um período de 2 anos e, nos próximos dois anos e meio, for feito apenas o acompanhamento desses pacientes. Considere 5% e 8% de nível de significância do teste e poder igual a 80%, 85% e 90%.

Resolução

Para calcular os tamanhos amostrais, primeiro vamos fazer a estimativa da função de sobrevivência via Kaplan-Meier.

Tempo	nº em risco	nº de eventos	censura	sobreviv.	desv.pad sobrev.	IC(95%) sup.	IC(95%) inf.
0.3	25	2	0	0.9200000	0.0589768	1.0000000	0.8195712
0.4	23	1	0	0.8800000	0.0738549	1.0000000	0.7614077
0.6	22	1	0	0.8400000	0.0872872	0.9967316	0.7079137
0.8	21	1	0	0.8000000	0.1000000	0.9732180	0.6576122
1.7	20	1	0	0.7600000	0.1123903	0.9472844	0.6097430
2.1	19	2	0	0.6800000	0.1371989	0.8898008	0.5196669
2.3	17	1	0	0.6400000	0.1500000	0.8587371	0.4769795
2.5	16	1	0	0.6000000	0.1632993	0.8263269	0.4356629
5.9	15	1	0	0.5600000	0.1772811	0.7926655	0.3956272
6.4	14	1	0	0.5200000	0.1921538	0.7578180	0.3568139
6.9	13	1	0	0.4800000	0.2081666	0.7218267	0.3191902
7.2	12	1	0	0.4400000	0.2256304	0.6847147	0.2827455
12.0	11	0	1	0.4400000	0.2256304	0.6847147	0.2827455
20.8	10	1	0	0.3960000	0.2490386	0.6451745	0.2430598
23.0	9	1	0	0.3520000	0.2755160	0.6040353	0.2051271
28.0	8	0	1	0.3520000	0.2755160	0.6040353	0.2051271
42.6	7	1	1	0.3017143	0.3157825	0.5602611	0.1624805
43.6	5	0	1	0.3017143	0.3157825	0.5602611	0.1624805
48.0	4	0	1	0.3017143	0.3157825	0.5602611	0.1624805
64.9	3	0	1	0.3017143	0.3157825	0.5602611	0.1624805
65.4	2	0	1	0.3017143	0.3157825	0.5602611	0.1624805
73.6	1	0	1	0.3017143	0.3157825	0.5602611	0.1624805

O estimador Kaplan-Meier para o tempo de sobrevivência dos pacientes com mieloma múltiplo com intervalo de confiança de 95% segue a curva abaixo:



Por fim é obtido o tamanho amostral, com 5% e 8% de nível de significância do teste e poder igual a 80%, 85% e 90%. Os códigos estão anexados nesse trabalho.

	5%	8%
80%	113.8404	97.41371
85%	130.2218	112.60848
90%	152.4063	133.29933

(b) Repita o item (a) assumindo que o recrutamento poderá ser feito ao longo dos 4 anos e meio. Compare com os resultados do item (a).

Resolução

O mesmo é feito alterando o tempo de recrutamento para 4 anos e meio:

	5%	8%
80%	158.1485	135.3284
85%	180.9058	156.4371
90%	211.7248	185.1811

Comparando com o item a), o tamanho amostral aumenta se o tempo de recrutamento aumenta.

Exercício 2

Considere $a_1 < ... < a_J$ uma partição do eixo do tempo. A função de sobrevivência da distribuição Exponencial Segmentada (Piecewise Exponential) é dada por:

$$S(t) = exp\Big\{ -\sum_{j=1}^{J} \lambda_j \nabla_j(t) \Big\},\ t > 0,$$

em que

$$\nabla_j(t) = \begin{cases} 0, & se\ t < a_{j-1} \\ t - a_{j-1}, & se\ a_{j-1} \leq t < a_j \\ a_j - a_{j-1}, & se\ t \geq a_j \end{cases}$$

para $j = 1, ..., J$. Defina $a_0 = 0$ e $a_{J+1} = \infty$.

(a) Mostre que a função de taxa de falha desta distribuição é constante no intervalo (a_j, a_{j+1}), $\forall\ j = 0, 1, ..., J$.

Resolução

Sabendo do resultado:

$$\alpha(t) = -\frac{d}{dt} ln(S(t))$$

Assim, aplicando o $ln(.)$ em $S(t)$, temos:

$$ln(S(t)) = ln(exp\Big\{ -\sum_{j=1}^{J} \lambda_j \nabla_j(t) \Big\}) = -\sum_{j=1}^{J} \lambda_j \nabla_j(t)$$

Dividindo a derivação em 3 casos, temos:

Para o casa em que $t < a_{j-1} \Rightarrow \nabla_j(t) = 0$, logo:

E derivando em relação a t:

$$-\frac{d}{dt} ln(S(t)) = -\frac{d}{dt} - \sum_{j=1}^{J} \lambda_j * 0 = 0$$

Para o casa em que $a_{j-1} \leq t < a_j \Rightarrow \nabla_j(t) = t - a_{j-1}$, logo:

E derivando em relação a t:

$$-\frac{d}{dt} ln(S(t)) = -\frac{d}{dt} \Big(-\sum_{j=1}^{J} \lambda_j (t - a_{j-1}) \Big) = \sum_{j=1}^{J} \lambda_j$$

Para o casa em que $t \geq a_j \Rightarrow \nabla_j(t) = a_j - a_{j-1}$, logo:

E derivando em relação a t:

$$-\frac{d}{dt} ln(S(t)) = -\frac{d}{dt} \Big(-\sum_{j=1}^{J} \lambda_j (a_j - a_{j-1}) \Big) = 0$$

Com os resultados acima temos que a função de taxa e falha é dada por:

$$\alpha(t) = \begin{cases} 0, & se\ t < a_{j-1} \\ \sum_{j=1}^{J} \lambda_j, & se\ a_{j-1} \leq t < a_j \\ 0, & se\ t \geq a_j \end{cases}$$

E como a função acima não depende de t, logo é uma função constante.

(b) Para uma amostra (censurada) de tamanho n desta distribuição, mostre que o estimador de máxima verossimilhança do vetor $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_J)$ é dado por $\widehat{\boldsymbol{\lambda}} = (\widehat{\lambda}_1, ..., \widehat{\lambda}_J)$, com

$$\widehat{\lambda}_j = \frac{d_j}{\sum_{i=1}^{n} \nabla_j(t_i)},\ j = 1, ..., J,$$

em que d_j representa o número de falhas no j-ésimo intervalo.

Resolução

(c) Para introduzir covariáveis no modelo, considere a matriz de desenho $\boldsymbol{x}$ sem intercepto. Mostre que esta distribuição pertence à classe de modelos de riscos proporcionais em um modelo de regressão com $\lambda_j = exp\{\beta_{0j} + \boldsymbol{x}^T\beta\}$.

Resolução

Para mostrar que esta distribuição pertence à classe de modelos de riscos proporcionais em um modelo de regressão, basta mostrar que:

$$\frac{\alpha(t|X_j)}{\alpha(t|X_i)} = k$$

Ou seja, a razão acima não depende de t, para isso, considerando o item a, como apenas para $a_{j-1} \leq t < a_j$ em que a função $\alpha(t)$ é difirente de zero com $\alpha(t) = \sum_{j=1}^{J} \lambda_j$, assim:

$$\frac{\alpha(t|X_j)}{\alpha(t|X_i)} = \frac{\sum_{j=1}^{J} exp\{\beta_{0j} + \boldsymbol{x}^T\beta\}}{\sum_{i=1}^{J} exp\{\beta_{0i} + \boldsymbol{x}^T\beta\}} = \frac{\sum_{j=1}^{J} exp\{\beta_{0j}\}exp\{\boldsymbol{x}^T\beta\}}{\sum_{i=1}^{J} exp\{\beta_{0i}\}exp\{\boldsymbol{x}^T\beta\}} = \frac{exp\{\boldsymbol{x}^T\beta\}\sum_{j=1}^{J} exp\{\beta_{0j}\}}{exp\{\boldsymbol{x}^T\beta\}\sum_{i=1}^{J} exp\{\beta_{0i}\}} = \frac{\sum_{j=1}^{J} exp\{\beta_{0j}\}}{\sum_{i=1}^{J} exp\{\beta_{0i}\}} \quad (I)$$

Como (I) não depende de t logo esta distribuição pertence à classe de modelos de riscos proporcionais em um modelo de regressão com $\lambda_j = exp\{\beta_{0j} + \boldsymbol{x}^T\beta\}$.

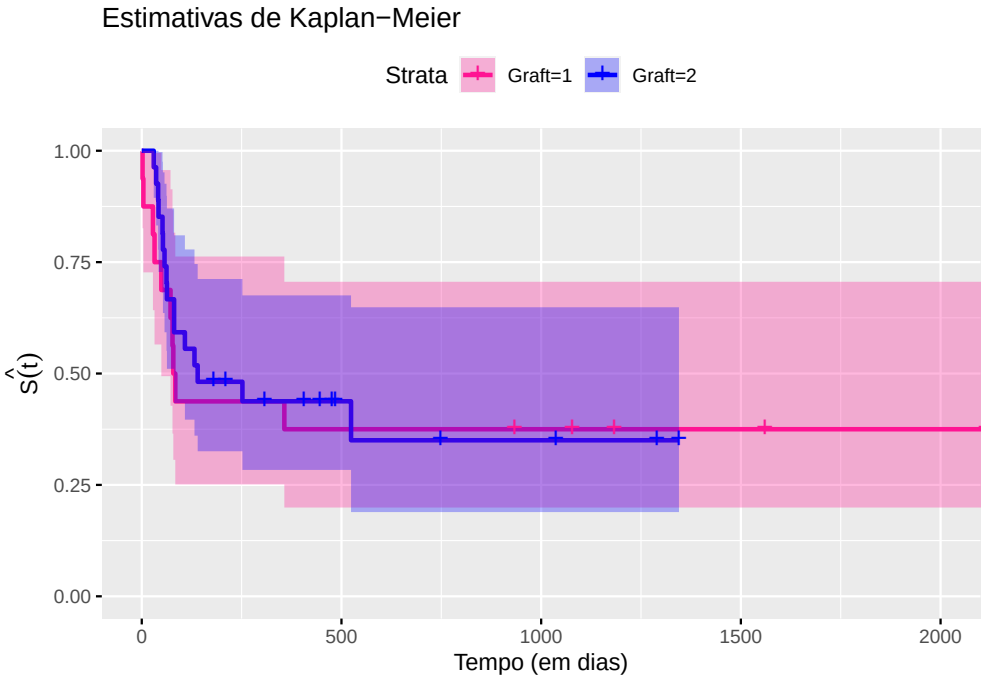
Exercício 3

Os dados no arquivo **HOD NHL.csv** são referentes a 43 pacientes com linfoma de Hodgking ou linfoma não Hodgking, submetidos a transplante de medula óssea. Alguns pacientes receberam transplante de doador aparentado compatível (transplante alogênico) e outros receberam transplante autólogo (ou seja, a própria medula óssea do paciente é coletada e posteriormente infundida). Nos dados também há informação sobre o escore de Karnofsky, que é uma medida de perfomance que classifica os pacientes segundo o bem-estar dos pacientes. O objetivo principal do estudo é a comparação dos tipos de transplante, considerando-se o tempo (dias) livre de doença (i.e., tempo antes de ocorrência da recorrência ou óbito).

(a) Construa curvas de Kaplan-Meier para o tempo de sobrevivência dos pacientes com linfoma para cada um dos grupos de tratamento. Categorize a variável escore de Karnofsky, criando uma variável com duas categorias: escore < 80 ou escore ≥ 80. Construa gráficos de Kaplan-Meier para essa nova variável categorizada. Comente.

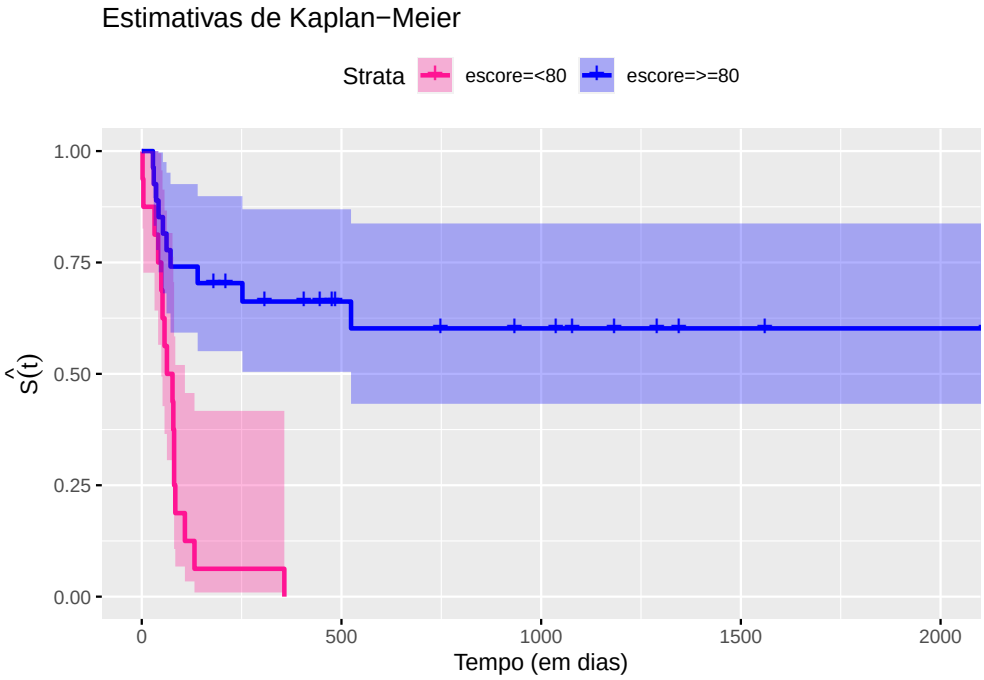
Resolução

O estimador Kaplan-Meier para o tempo de sobrevivência dos pacientes com linfoma para cada um dos grupos de tratamento com intervalo de confiança de 95% segue a curva abaixo:



Em que podemos notar que as curvas de sobrevivência, são bem próximas para os dois tipos de transplante.

O estimador Kaplan-Meier para o tempo de sobrevivência dos pacientes com linfoma para cada um dos escore de Karnofsky com intervalo de confiança de 95% segue a curva abaixo:



Em que podemos notar um maior tempo de sobrevivência para os pacientes com escore de Karnofsky maiores ou iguais a 80, contra os pacientes com escores menores que 80.

(b) Ajuste um modelo de regressão Weibull com a variável tratamento e a variável escore de Karnofsky categorizada conforme o item (a), utilizando algum pacote estatístico (R, Splus, SAS, etc.). Apresente as estimativas de máxima verossimilhança dos parâmetros considerando a representação do modelo Weibull como um modelo de locação-escala e também como um modelo de riscos proporcionais.

Resolução

Utilizando o modelo Weibull com a parametrização locação-escala:

$$Log(T) = \mu + X'\gamma + \sigma w$$

Em que μ é o intercepto, X é a matriz de covariáveis, γ é o vetor de parâmetros associados as covariávies (sem intercepto), σ é o parâmetro de escala e por fim $w \sim Valor\ Extremo$ padrão. Os parâmetros a serem estimados são $\theta = (\mu, \gamma_1, \gamma_2, \sigma)^T$.

Ajustando o modelo obtemos, as estimativas de verossimilhança:

```
##
## Call:
## survreg(formula = Surv(Time, D_R) ~ escore + Graft, data = data,
##     dist = "weibull")
##               Value Std. Error     z       p
## (Intercept)  5.169      0.908  5.70 1.2e-08
## escore>=80   3.325      0.563  5.91 3.5e-09
## Graft       -0.550      0.525 -1.05    0.29
## Log(scale)   0.263      0.171  1.54    0.12
##
## Scale= 1.3
##
## Weibull distribution
## Loglik(model)= -168.4   Loglik(intercept only)= -183.3
##  Chisq= 29.72 on 2 degrees of freedom, p= 3.5e-07
## Number of Newton-Raphson Iterations: 5
## n= 43
```

E para representação do modelo Weibull como um modelo de riscos proporcionais, podemos escrever:

$$\alpha(t|x) = \alpha_0(t)g(x)$$

Sendo $g(x) = e^{-x'\gamma/\sigma}$ e $\alpha_0(t) = \alpha(t|x = 0)$ a função de risco basal, como visto em aula

$$\alpha(t|x) = exp\{-t^{1/\sigma}e^{-\mu/\sigma}e^{-x'\gamma/\sigma}\} \Rightarrow \alpha_0(t) = \alpha(t|x = 0) = exp\{-t^{1/\sigma}e^{-\mu/\sigma}\}$$

Fazendo $\beta = -\gamma/\sigma$, e pela invariância do estimador de máxima verossimilhança temos que as estimativas um modelo weibull de riscos proporcionais é dado por:

$$e^{x'\beta}e^{-\mu/\sigma}$$

Assim, temos:

Variaveis	estimativa
Intercepto	0.0188111
escore>=80	0.0776272
Graft	1.5262664

(c) Encontre uma estimativa pontual para a razão de taxas de falha de pacientes que receberam transplante autólogo e alogênico. Encontre também uma estimativa do fator de aceleração, deixando claro como foi calculado. Faça o mesmo para a outra variável (escore de Karnofsky) incluída no modelo.

Resolução

Para os pacientes que receberam transplante autólogo $Graft = 2$ e alogênico $Graft = 1$, temos:

$$\frac{\alpha(t|Graft = 1)}{\alpha(t|Graft = 2)} = \frac{e^{5.17}}{e^{5.17-0.55}} = \frac{175.91}{101.494} = 1.73$$

Ou seja, risco de indivíduo que recebeu o transplante alogênico morrer é 1.73 vezes maior do que o de morrer com o transplante autólogo.

O modelo de vida acelerado é dado por:

$$S(t|x) = S_0(\psi_x t)$$

Como visto em aula, o fator de aceleração é dado por:

$$\psi_x = e^{-x'\gamma} = e^{-\gamma_1 x_1} = e^{0.55*1} = 1.1733$$

Ou seja, como $\phi_x > 1$, temos que para os pacientes que receberam transplante autólogo se comportam como 'passado' do que os pacientes que receberam transplante alogênico.

Para os pacientes que com escores Karnofsky >=80 e com escores <80, temos:

$$\frac{\alpha(t|escore < 80 = 1)}{\alpha(t|escore < 80 = 0)} = \frac{e^{5.17+3.325}}{e^{5.17}} = \frac{4890.25}{175.91} = 27.8$$

Ou seja, risco de indivíduo que tem o escore de Karnofsky <80 morrer é 27.8 vezes maior do que quem tem o escore de Karnofsky >=80.

Como visto em aula, o fator de aceleração é dado por:

$$\psi_x = e^{-x'\gamma} = e^{-\gamma_1 x_1} = e^{-3.325*1} = 0.036$$

Ou seja, como $\phi_x < 1$, temos que para os pacientes que tem escore de Karnofsky <80 se comportam como 'futuro' do que os pacientes que tem escore de Karnofsky >=80.

(d) Teste a hipótese de igualdade dos tipos de transplante e também das categorias do escore de Karnofsky, utilizando a estatística de Wald, com um nível de significância de 10%. Comente.

Resolução

Para testar a igualdade dos transplante é equivalente a testar se os parâmetros da regressão são iguais a zero, ou seja:

$$\begin{cases} H_0 : \gamma_i = 0 \ com \ i = 1, 2 \\ H_1 : \gamma_i \neq 0 \end{cases}$$

E para isso a estatística de Wald é dada por:

$$\frac{\hat{\gamma} - \gamma}{\sqrt{I^{-1}(\gamma)}}$$

em que $I^{-1}(\gamma)$ é a variância de γ obtida através da matriz de informação de fisher. Assim, sob H_0, temos:

	Value	Std. Error	z	p
(Intercept)	5.1692031	0.9076140	5.695376	0.0000000
escore>=80	3.3250980	0.5630723	5.905277	0.0000000
Graft	-0.5500871	0.5247505	-1.048283	0.2945082
Log(scale)	0.2631195	0.1706529	1.541840	0.1231125

Em que podemos notar que para variável escore de Karnofsky as categorias são diferentes com um nível de significâcia de 5%, porém para a variável Graft, isso não o ocorre, logo com um nível de significâcia de 5%, os tipos de transplantes são iguais.

Exercício 4

Considere os dados do exercício 3.

(a) Refaça o item (b) utilizando a distribuição log-logística. Especifique claramente qual foi o modelo utilizado e quais foram os parâmetros estimados.

Resolução

Utilizando o modelo log-losgistico com a parametrização locação-escala:

$$Log(T) = \mu + X'\gamma + \sigma w$$

Em que μ é o intercepto, X é a matriz de covariáveis, γ é o vetor de parâmetros associados as covariávies (sem intercepto), σ é o parâmetro de escala e por fim $w \sim Logistica$ padrão.Os parâmetros a serem estimados são $\theta = (\mu, \gamma_1, \gamma_2, \sigma)^T$.

Ajustando o modelo obtemos:

```
##
## Call:
## survreg(formula = Surv(Time, D_R) ~ escore + Graft, data = data,
##     dist = "loglogistic")
##               Value Std. Error     z       p
## (Intercept)  4.5807     1.1040  4.15 3.3e-05
## escore>=80   2.8753     0.6480  4.44 9.1e-06
## Graft       -0.3547     0.6519 -0.54    0.59
## Log(scale)   0.0922     0.1688  0.55    0.58
##
## Scale= 1.1
##
## Log logistic distribution
## Loglik(model)= -170.9   Loglik(intercept only)= -180.2
```

```
##  Chisq= 18.56 on 2 degrees of freedom, p= 9.3e-05
## Number of Newton-Raphson Iterations: 4
## n= 43
```

(b) Encontre uma estimativa pontual para ao fator de aceleração e interprete o resultado.

Resolução

O modelo de vida acelerado é dado por:

$$S(t|x) = S_0(\psi_x t)$$

Como visto em aula, o fator de aceleração é dado por:

$$\psi_x = e^{-x'\gamma} = e^{-\gamma_1 x_1 - \gamma_2 x_2} = e^{-3.325x_1 + 0.55x_2}$$

(c) A chance de sobrevivência após t é definida como

$$\frac{S(t|x)}{1 - S(t|x)}$$

No modelo logístico, mostre que:

$$\frac{S(t|x)}{1 - S(t|x)} = exp[-x^T\beta]\frac{S(t|x=0)}{1 - S(t|x=0)}$$

Resolução

Sabemos que o modelo log-logístico é definido como:

$$Log(T) = \mu + x'\gamma + \sigma w$$

com $w \sim$ *Logistica* padrão. Assim:

$$\frac{S(t|x)}{1 - S(t|x)} = \frac{\mathbb{P}(T > t|x)}{1 - \mathbb{P}(T > t|x)} = \frac{\mathbb{P}(T > t|x)}{\mathbb{P}(T \leq t|x)} = \frac{\mathbb{P}(ln(T) > ln(t|x))}{\mathbb{P}(ln(T) \leq ln(t|x))} = \frac{\mathbb{P}(\mu + x'\gamma + \sigma w > ln(t|x))}{\mathbb{P}(\mu + x'\gamma + \sigma w \leq ln(t|x))}$$

$$\Rightarrow \frac{S(t|x)}{1 - S(t|x)} = \frac{\mathbb{P}(w > (ln(t) - \mu - x'\gamma)/\sigma)}{\mathbb{P}(w \leq (ln(t) - \mu - x'\gamma)/\sigma)} = \frac{\frac{1}{1+e^{(ln(t)-\mu-x'\gamma)/\sigma}}}{\frac{e^{(ln(t)-\mu-x'\gamma)/\sigma}}{1+e^{(ln(t)-\mu-x'\gamma)/\sigma}}} = \frac{1}{e^{(ln(t)-\mu-x'\gamma)/\sigma}} = t^{1/\sigma}e^{-\mu-x'\gamma/\sigma}$$

Definindo $e^{x'\gamma/\sigma}$ como $e^{x'\beta}$

$$\Rightarrow \frac{S(t|x)}{1 - S(t|x)} = t^{1/\sigma}e^{-\mu/\sigma}e^{-x'\beta} \quad (I)$$

Pelo resultado de (I) temos que:

$$\frac{S(t|x=0)}{1 - S(t|x=0)} = t^{1/\sigma}e^{-\mu/\sigma}$$

Logo:

$$\frac{S(t|x)}{1 - S(t|x)} = exp[-x^T\beta]\frac{S(t|x=0)}{1 - S(t|x=0)}$$

(d) Obtenha uma estimativa da razão de chances de sobrevivência após t de pacientes com células anormais e pacientes com células normais. Interprete.

Resolução

(e) Repita o item (d) do exercício 3, utilizando o modelo log-logístico. Compare os resultados e comente.

Resolução

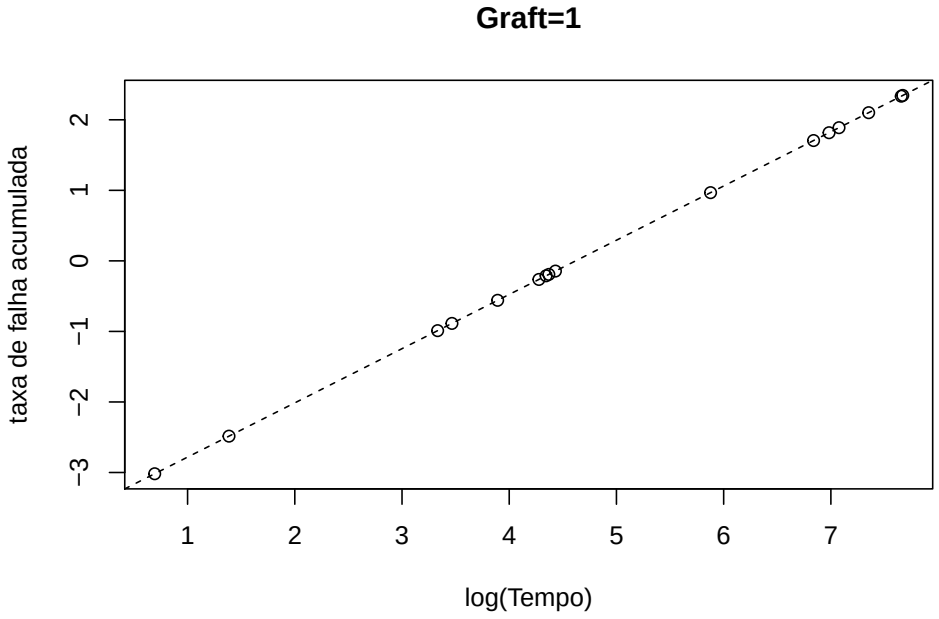
Exercício 5

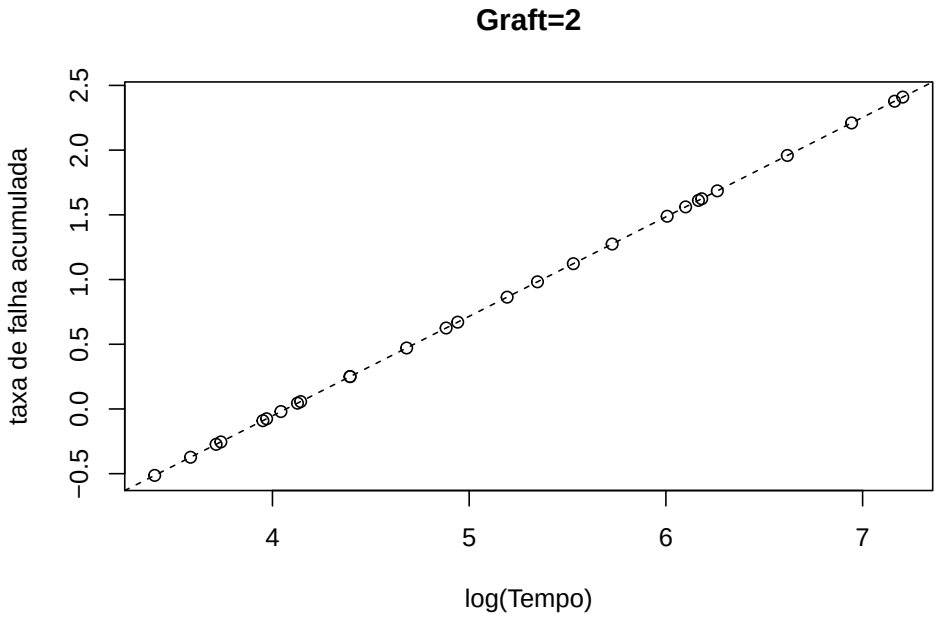
Considerando ainda os dados do exercício 3, faça gráficos apropriados da taxa de falha acumulada para verificar a adequabilidade dos modelos.

Em todos os casos, utilize o estimador de Nelson-Aalen da função de taxa de falha acumulada considerando cada grupo separadamente (ou seja, obtenha estimativas da função de taxa de falha acumulada para cada grupo).

(a) Weibull

Resolução

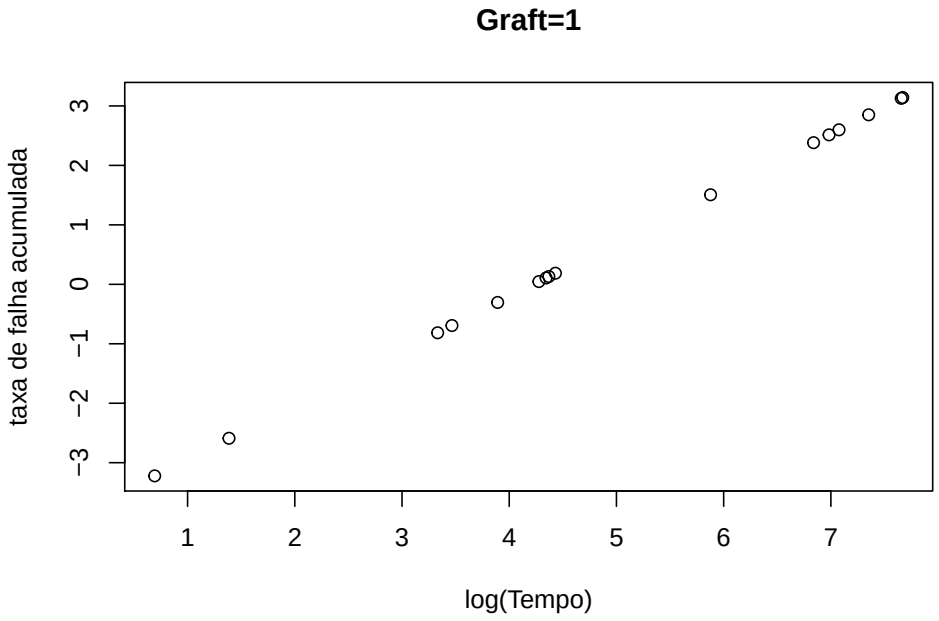


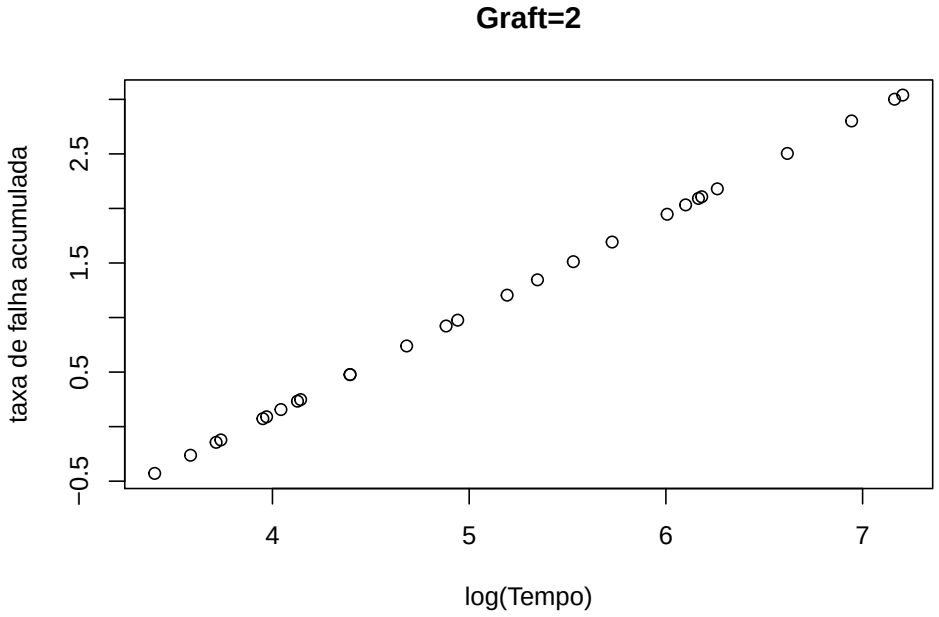


Pode-se observar que os pontos passam por uma reta com uma certa inclinação que cruza a origem.

(b) Log-logístico

Resolução





Era de se esperar que os pontos formassem uma reta. Logo os dois modelos aparentam adequados.

Exercício 6

Considere ainda os dados do exercício 3. Obtenha os resíduos de Cox-Snell e *deviance* para os modelos de regressão Weibull (ajustado no exercício 4) e log-logístico (ajustado no exercício 3). Faça gráficos dos resíduos em função do tempo e comente. Com base em todas as análises feitas, discuta se os modelos (Weibull ou log-logístico) parecem ser adequados para os dados trabalhados.

A partir dos dados do arquivo **HOD NHL.csv**, os resíduos de Cox-Snell para o modelo Weibull são obtidos a seguir:

```
##  [1]  0.56758764  0.41207674  0.57176005  0.86525054  2.63119525
##  [6]  8.40377759  9.39068861 10.08613452 12.47585343 15.75857694
## [11] 15.93019088  0.04891378  0.08333317  1.17304200  0.80927421
## [16]  0.82538328  0.06017272  0.07195380  0.04985537  0.05384209
## [21]  0.06531543  0.15181352  0.06531543  0.23852048  0.41870020
## [26]  0.20733041  0.38889446  0.70756696  0.04645991  0.05344918
## [31]  0.03870126  0.04645845  0.08117288  0.08147980  0.09506879
## [36]  0.18416384  0.27760551  0.34413806  0.36991370  0.39390869
## [41]  0.55044465  0.83684300  0.86413500
```

Os resíduos de Cox-Snell para o modelo log-logístico também são obtidos:

```
##  [1] 0.47725849 0.40542613 0.55234918 0.79089565 1.70650452 2.77037407
##  [7] 2.89434298 2.97459303 3.21540094 3.48275921 3.49521670 0.03911720
## [13] 0.07237462 0.89495811 0.74831050 0.76069769 0.06233804 0.07651805
## [19] 0.05967755 0.06519823 0.08132414 0.17632690 0.08132414 0.28485768
## [25] 0.49621705 0.24614999 0.46285485 0.78722442 0.04624097 0.05438128
```

```
## [31] 0.04452922 0.05501419 0.08778073 0.10448254 0.12420825 0.21711935
## [37] 0.33259335 0.41153265 0.44126861 0.46851314 0.63596133 0.90057648
## [43] 0.92335361
```

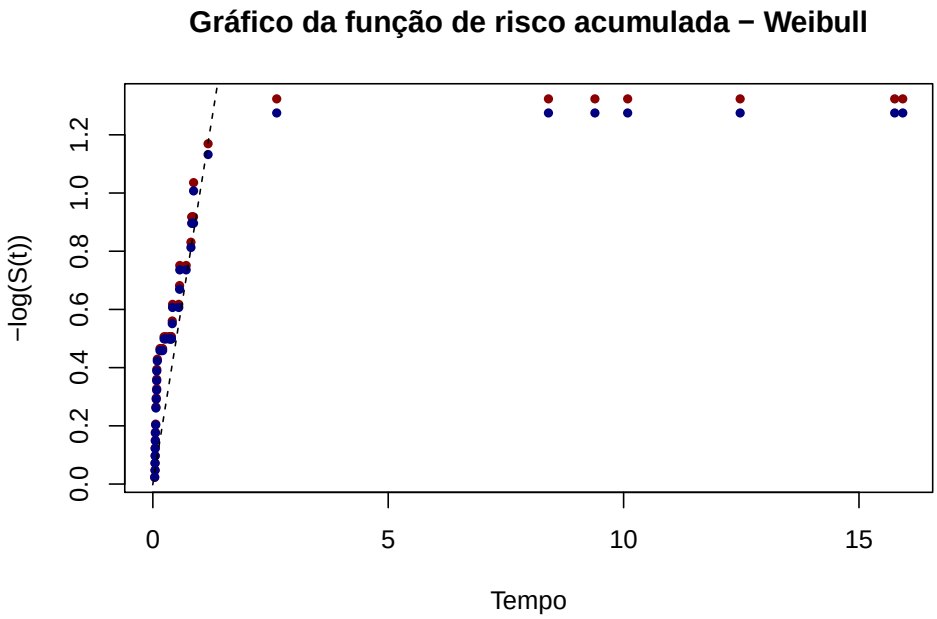
Os resíduos *deviance* também são obtidos para o modelo Weibull:

```
##  [1]  0.5175862  0.7728162  0.5114605  0.1413274 -1.1521780 -4.0997018
##  [7] -4.3337486 -4.4913549 -4.9951684 -5.6140141 -5.6445001  2.0330322
## [13]  1.7710120 -0.1639611  0.2044098  0.1859607  1.9342745  1.8459062
## [19]  2.0240970  1.9877333  1.8941185  1.4400805  1.8941185  1.1591554
## [25]  0.7606580 -0.6439416 -0.8819234 -1.1895940  2.0570005  1.9912173
## [31]  2.1403665  2.0570150  1.7845710  1.7826272  1.7018950 -0.6069001
## [37] -0.7451248 -0.8296241 -0.8601322 -0.8875908 -1.0492327 -1.2937102
## [43] -1.3146368
```

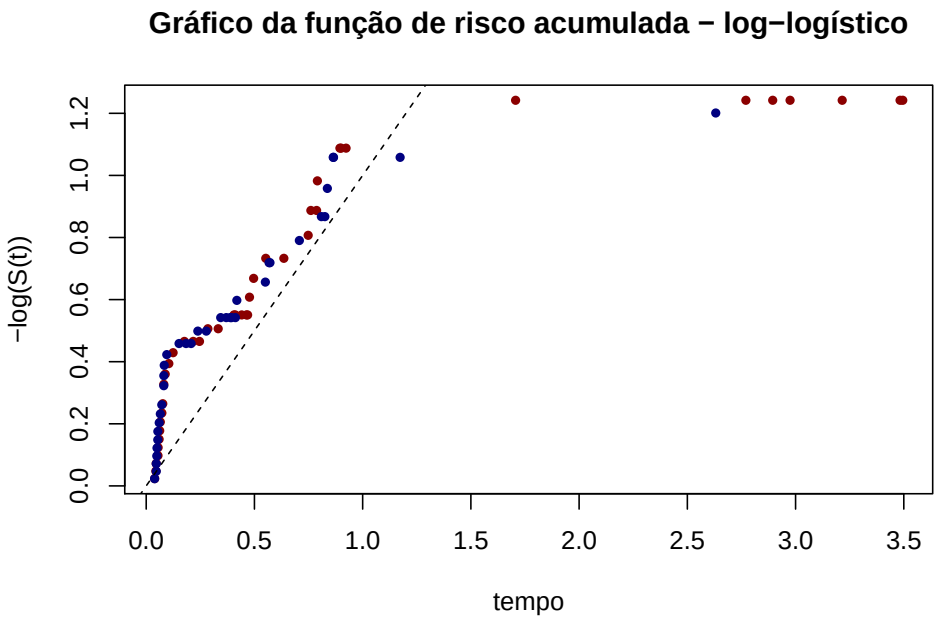
e para o modelo log-logístico:

```
##  [1]  0.6587192  0.7851659  0.5402297  0.2257649 -0.5866130 -2.3538794
##  [7] -2.4059688 -2.4390953 -2.5359026 -2.6392269 -2.6439428  2.1355609
## [13]  1.8429727  0.1089631  0.2765783  0.2615987  1.9170401  1.8147984
## [19]  1.9382864  1.8950046  1.7836123  1.3503648  1.7836123  1.0398300
## [25]  0.6276287 -0.7016409 -0.9621381 -1.2547704  2.0591891  1.9829859
## [31]  2.0766024  1.9774622  1.7439576  1.6511922  1.5556375 -0.6589679
## [37] -0.8155898 -0.9072295 -0.9394345 -0.9680012 -1.1277955 -1.3420704
## [43] -1.3589361
```

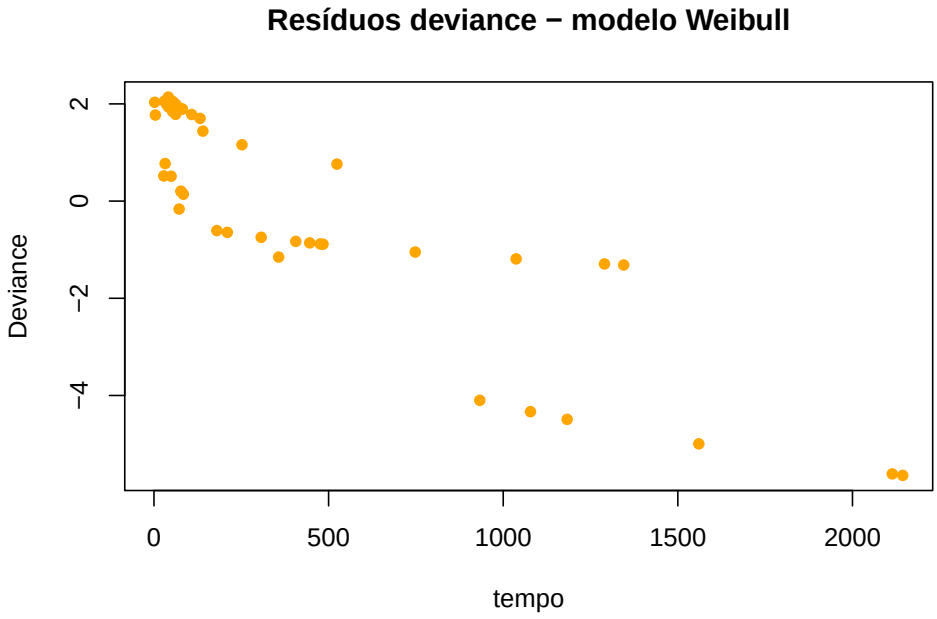
Com esses resultados, é possível elaborar gráficos desses resíduos para a análise da escolha do modelo. Uma opção é realizar um gráfico da função de risco acumulada para os resíduos de Cox-Snell, utilizando os estimadores de Kaplan-Meier (em vermelho) e Nelson_Aalen (em azul), primeiramente para o modelo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



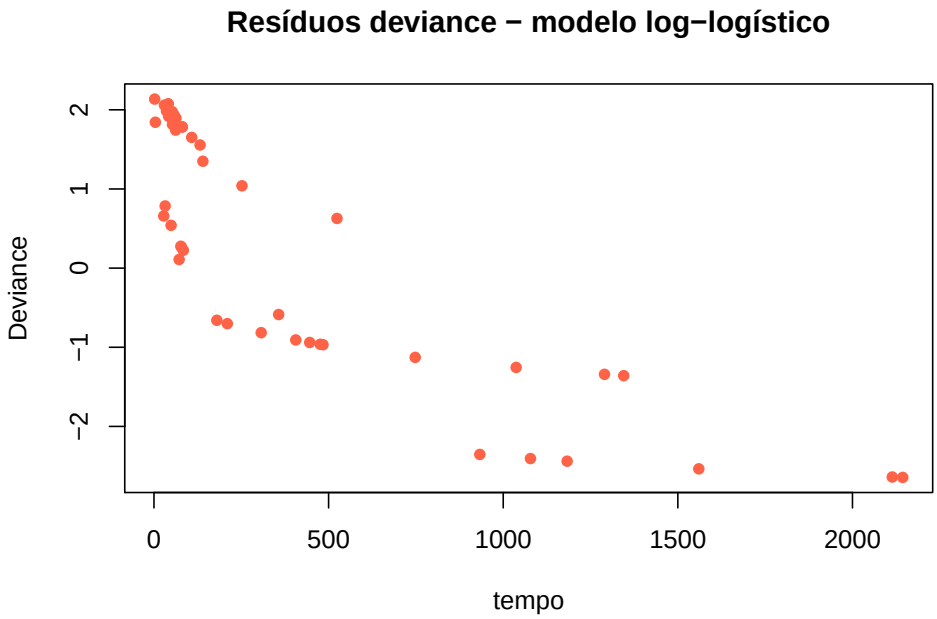
O mesmo é feito para o modelo log-logístico:



O esperado é que os resíduos acompanhem a linha pontilhada. Graficando os resíduos deviance em relação ao tempo para o modelo Weibull:



O mesmo processo para o modelo log-logístico:



Para interpretar esses gráficos é necessário verificar se os pontos estão próximos ou mais distantes entre si, quanto mais próximos melhor o ajuste do modelo. Logo, nota-se que pelos gráficos do resíduos de Cox-Snell, o modelo Weibull possuem os primeiros pontos mais próximos da reta pontilhada do que o modelo log-logístico. Pelos gráficos dos resíduos deviance pode-se dizer que os resíduos estão mais concentrados para o modelo log-logístico, dos quais variam de -3 e 3, do que os resíduos para o modelo Weibull variam de -6 e 3, entretanto os dois possuem pontos bem distântes entre si. Portanto, pode-se dizer que o modelo Weibull aparentemente está melhor ajustado. Pode-se fazer os cretérios AIC e BIC para confirmar a escolha. O AIC para o modelo Weibull é de 344.8589 e para o modelo log-logístico é 349.8397. O BIC para o modelo Weibull é de -174,0712 e para o modelo log-logístico é -176.5617. Observando o menor AIC e o maior BIC, confirma-se que o modelo Weibull está melhor ajustado.

Anexo

Códigos

```
 # Pacotes
library(ggplot2)
library(asaur)
library(survival)
library(survminer)
library(sqldf)
library(mice)
library(KMsurv)
library(mice)


# Exercício 1
```

```r
# item a
Tempo <- c(0.3,5.9,20.8,28.0,1.7,73.6,7.2,2.1,6.4,2.5,2.3,0.3,0.4,65.4,64.9,0.6,23.0,42.6,48.0,6.9,2.1,4
Censura <- c(1,1,1,0,1,0,1,1,1,1,1,1,1,0,0,1,1,0,0,1,1,0,1,0,1)

library(survival)
library(survminer)
Ex1 <- data.frame(Tempo,Censura)

KM1 <- survfit(Surv(Ex1$Tempo, Ex1$Censura)~1)


# Tabela com estimativas de Kaplan-Meier
knitr::kable(surv_summary(KM1),col.names = c("Tempo","nº em risco","nº de eventos",
                                             "censura","sobreviv.","desv.pad sobrev.",
                                             "IC(95%) sup.","IC(95%) inf."))

# Grafico Kaplan-Meier
ggsurvplot(KM1, data = Ex1,palette = c('blue'),
           ggtheme=theme_gray(), legend = 'none') +
  labs(x="Tempo (em anos)",
       y=expression(hat(S(t))),
       title = "Estimativas de Kaplan-Meier")

x <- (18-12)/((20.8-12)/(0.396-0.44))+0.44

#log(0.65)/log(0.41)

#log(0.4831582)

# a =5% b= 80%
d <- data.frame()
d[1,1] = 4*(1.96+0.8416)^2/(-0.7274111)^2

# a = 8% b = 80%
d[1,2] = 4*(1.75+0.8416)^2/(-0.7274111)^2

# a =5% b= 85%
d[2,1] = 4*(1.96+1.0364)^2/(-0.7274111)^2

# a = 8% b = 85%
d[2,2] = 4*(1.75+1.0364)^2/(-0.7274111)^2

# a =5% b= 90%
d[3,1] = 4*(1.96+1.2816)^2/(-0.7274111)^2

# a = 8% b = 90%
d[3,2] = 4*(1.75+1.2816)^2/(-0.7274111)^2

a= 2

S30 <- (30-28)/((42.6-28)/(0.3017143-0.3520000))+0.3520000
S42 <- (42-28)/((42.6-28)/(0.3017143-0.3520000))+0.3520000
S54 <- 0.3017143
```

```r
ppad <- 1-(S30+4*S42+ S54)/6

S30N <- S30^-log(ppad)
S42N <- S42^-log(ppad)
S54N <- S54^-log(ppad)

pnovo <- 1-(S30N+4*S42N+S54N)/6

pf <- (ppad+pnovo)/2

n24 <- d/pf
colnames(n24) <- c("5%","8%")
rownames(n24) <- c("80%","85%","90%")
knitr::kable(n24)

# item b
a= 4.5

S0 <- 1
S27<- 0.3520000
S54 <- 0.3017143
ppad <- 1-(S0+S27+4*S54)/6

S0N <- S0^-log(ppad)
S27N <- S27^-log(ppad)
S54N <- S54^-log(ppad)

ppad <- 1-(S0N+S27N+4*S54N)/6

pf <- (ppad+pnovo)/2

n54 <- d/pf
colnames(n54) <- c("5%","8%")
rownames(n54) <- c("80%","85%","90%")
knitr::kable(n54)

# Exercício 3

data <- read.csv("HOD_NHL.csv",header = T,sep=';')

# item a

ekm_ex3 <- survfit(Surv(Time, D_R)~ Graft,data = data)

# Grafico Kaplan-Meier
ggsurvplot(ekm_ex3, data = data, palette = c('deeppink','blue'),conf.int = T,
           ggtheme=theme_gray()) +
  labs(x="Tempo (em dias)",
       y=expression(hat(S(t))),
       title = "Estimativas de Kaplan-Meier")

# categorizando variável Karnofsky
data$escore <- sapply(data$Karnofsky,
```

```
                          function(x){
                            if (x < 80) x = '<80'
                            else x = '>=80'
                          })
data$escore <- as.factor(data$escore)

ekm_ex3_esc <- survfit(Surv(Time, D_R)~ escore,data = data)

# Grafico Kaplan-Meier
ggsurvplot(ekm_ex3_esc, data = data, palette = c('deeppink','blue'),conf.int = T,
           ggtheme=theme_gray()) +
  labs(x="Tempo (em dias)",
       y=expression(hat(S(t))),
       title = "Estimativas de Kaplan-Meier")

# item b

Modelo.wei <- survreg(Surv(Time, D_R)~ escore+Graft, dist='weibull',data = data)

summary(Modelo.wei)

df <- data.frame(Variaveis=c("Intercepto","escore>=80","Graft"), estimativa=c(exp(-Modelo.wei$coefficie

knitr::kable(df,row.names = FALSE)

# item d

knitr::kable(summary(Modelo.wei)$table)


# Exercício 4

# item a

# Modelo log-logistico
Modelo.ll <- survreg(Surv(Time, D_R)~ escore+Graft, dist='loglogistic',data = data)

summary(Modelo.ll)

# item b

exp(Modelo.ll$coefficient[2]+Modelo.ll$coefficient[3])

# Exercício 5

# item a

data <- read.csv("HOD_NHL.csv",header = T,sep=';')
data$Graft <- as.factor(data$Graft)
data$escore <- sapply(data$Karnofsky,
                      function(x){
                        if (x < 80) x = '<80'
                        else x = '>=80'
```

```
                      })
data$escore <- as.factor(data$escore)

Modelo.wei <- survreg(Surv(Time, D_R)~ escore+Graft, dist='weibull',data = data)
Modelo.llog <- survreg(Surv(Time, D_R)~ escore+Graft, dist='loglogistic',data = data)

beta <- as.vector(-Modelo.wei$coef/Modelo.wei$scale)
gama <- 1/Modelo.wei$scale

dataG1 <- data[data$Graft==1,]
dataG2 <- data[data$Graft==2,]

# Weibull
a <- exp(beta[1])*dataG1$Time^(gama)
A <- log(a)
plot(log(dataG1$Time),A,ylab="taxa de falha acumulada",xlab="log(Tempo)",main="Graft=1")
abline(log(exp(beta[1])),gama,lty=2)

b <- exp(beta[1]+beta[3])*dataG2$Time^(gama)
B <- log(b)
plot(log(dataG2$Time),B,ylab="taxa de falha acumulada",xlab="log(Tempo)",main="Graft=2")
abline(log(exp(beta[1]+beta[3])),gama,lty=2)

# item b

#log-logistico

beta <- Modelo.llog$coefficients
gama <- 1/Modelo.llog$scale

a <- log(1+exp(-beta[1]*gama)*dataG1$Time^gama)
A <- log(exp(a)-1)
plot(log(dataG1$Time),A,ylab="taxa de falha acumulada",xlab="log(Tempo)",main="Graft=1")

b <- log(1+exp(-(beta[1]+beta[3])*gama)*dataG2$Time^gama)
B <- log(exp(b)-1)
plot(log(dataG2$Time),B,ylab="taxa de falha acumulada",xlab="log(Tempo)",main="Graft=2")

# Exercício 6

v2 <- ifelse(data$Graft==2,1,0)
v3 <- ifelse(data$escore==">=80",1,0)

xb_wei<- Modelo.wei$coef[1]+Modelo.wei$coef[2]*v2+Modelo.wei$coef[3]*v3
res_wei<- (log(data$Time)-xb_wei)/Modelo.wei$scale

xb_llog<- Modelo.llog$coef[1]+Modelo.llog$coef[2]*v2+Modelo.llog$coef[3]*v3
res_llog<- (log(data$Time)-xb_llog)/Modelo.llog$scale

resid_wei<-exp(res_wei)
resid_llog<-exp(res_llog)

coxsnell_wei<- (data$Time^(1/Modelo.wei$scale))*exp(-xb_wei/Modelo.wei$scale)
```

```
coxsnell_llog<- log(1+(data$Time^(1/Modelo.llog$scale))*exp(-xb_llog/Modelo.llog$scale))

coxsnell_wei

coxsnell_llog

# RESÍDUOS DEVIANCE

m_wei<- data$D_R- coxsnell_wei
m_llog<- data$D_R- coxsnell_llog
deviance_wei <-  sqrt(-2*(m_wei+data$D_R*log(data$D_R-m_wei)))*ifelse(m_wei<0,-1,1)
deviance_llog <-  sqrt(-2*(m_llog+data$D_R*log(data$D_R-m_llog)))*ifelse(m_llog<0,-1,1)

deviance_wei

deviance_llog

#WEIBULL
# Curva de Kaplan-Meier
KM_wei <-  survfit(Surv(coxsnell_wei, data$D_R)~1)
TFAcum_KM_wei <- -log(KM_wei$surv)
# Estimador de Nelson_Aalen
Surv_Aa_wei <- survfit(coxph(Surv(coxsnell_wei, data$D_R)~1))
TFAcum_Aa_wei <- -log(Surv_Aa_wei$surv)
#GrÃ¡fico
plot(KM_wei$time,TFAcum_KM_wei, col="dark red", pch=16, main="Gráfico da função de risco acumulada - Wei
points(Surv_Aa_wei$time,TFAcum_Aa_wei, col="navy blue", pch=16, cex=0.8)
abline(0,1,lty=2)

#LOG-LOGÍSTICA
# Curva de Kaplan-Meier
KM_llog <-  survfit(Surv(coxsnell_llog, data$D_R)~1)
TFAcum_KM_llog <- -log(KM_llog$surv)
# Estimador de Nelson_Aalen
Surv_Aa_llog <- survfit(coxph(Surv(coxsnell_llog, data$D_R)~1))
TFAcum_Aa_llog <- -log(Surv_Aa_llog$surv)
#Gráfico

plot(KM_llog$time,TFAcum_KM_llog, col="dark red", pch=16, main="Gráfico da função de risco acumulada - l
points(Surv_Aa_wei$time,TFAcum_Aa_llog, col="navy blue", pch=16, cex=0.8)
abline(0,1,lty=2)

plot(data$Time, deviance_wei, pch=16, col="orange", main="Resíduos deviance - modelo Weibull",xlab="temp

plot(data$Time, deviance_llog, pch=16, col="tomato1", main="Resíduos deviance - modelo log-logístico",xl

# CRITÉRIOS DE AKAIKE E BIC (Klein e Moeschberger)

AIC_wei<- -2*Modelo.wei$loglik[2]+2*4
AIC_llog<- -2*Modelo.llog$loglik[2]+2*4
n<- length(data$Time)
BIC_wei <- Modelo.wei$loglik[2]-(3/2)*log(n)
BIC_llog <- Modelo.llog$loglik[2]-(3/2)*log(n)
```