Universidade de Brasília
Departamento de Estatística

Estudo da fração de ejeção via técnicas de análise de sobrevivência

Artur Macedo Rocha

Relatório Final de Monografia apresentado para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Brasília
2018

Artur Macedo Rocha

Estudo da fração de ejeção associada à insuficiência cardíaca sob técnicas de análise de sobrevivência.

Orientador:
Prof. Dr. **Antônio Eduardo Gomes**

Relatório Final de Monografia apresentado para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Brasília
2018

Agradecimentos

Quero começar agradecendo a mim mesmo por sempre acreditar em mim, por realizar todo esse trabalho duro, por não desistir e por sempre ser eu mesmo, sempre.

Agradeço também a minha família, principalmente meu irmão, que está presente em todas situações sempre me apoiando e desejando o melhor para mim e meu futuro, do seu jeito singular.

Não menos importantes, agradeço meus amigos, por me incentivarem a ser a melhor versão de mim mesmo por sempre estarem agregando aos meu valores morais e éticos com suas ações indubitáveis, por me apoiarem em minhas decisões.

Conteúdo

Resumo

A monografia confeccionada tem o intuito de analisar os dados de pacientes de um hospital não identificado, por razões de sigilo das informações referentes, com vista análise de fatores que aumentam o risco de óbito sob a ótica de análise de sobrevivência.

Para modelar a variável de interesse/resposta foram considerados 3 modelos, Log-Logístico, Weibull e o modelo de regressão de Cox, os modelos também foram atribuídos para a variável resposta categorizada de 2 formas, presentes na monografia. Todos os modelos possuem um bom ajuste, mas o modelo escolhido como principal foi o modelo de Cox, para as categorias »40"e «40", considerando os respectivos critérios para tal.

Os resultados obtidos nesta monografia atendem as expectativas da área pois, como esperado, a fração de ejeção reduzida aumenta os riscos que o paciente corre ao decorrer de um tratamento ou acompanhamento médico. Pela modelagem é possível analisar previamente os fatores que impactam no óbito tendo mais tempo para solucionar ou remediar tal fator.

Palavras-chave: Análise de sobrevivência; Fracção de ejeção; Log-Logístico; Weibull; Cox;

1 Introdução

Análise de sobrevivência, conhecido também como análise de confiabilidade, é uma das áreas da estatística que visa a análise de dados com o objetivo de estudar a variável de interesse, o tempo que decorre do início do estudo até o momento em que certo evento de interesse ocorra, e a variável de interesse (tempo decorrido) é chamada de tempo de falha. O estudo a ser feito não é completamente insensível a fatores internos ou externos, o que implica em unidades observacionais deixando o estudo antes da ocorrência do evento de interesse, ou seja, têm-se observações parciais. Quando ocorre o abandono prematuro do estudo, indiferentemente da razão, não é possível inferir seu tempo de falha, é denominado este como um dado censurado e é de igual importância para análise de tempos de sobrevivência pois mesmo incompletas ainda possuem informação.

Esta área de análise de sobrevivência pode ser estendida para pesquisas médicas, entre diversas outras, que auxilia na maior compreensão do comportamento adotado por seus pacientes ou cobaias a partir dos respectivos métodos aplicados e por meio da estimação das curvas de sobrevivência, modelagem da variável de interesse e comparações entre as especificidades de cada conjunto de pacientes, é possível antecipar os fatores impactantes e adotar medidas preventivas com maior antecedência, aumentando as chances do paciente.

Nesta monografia serão utilizados tais métodos de analise de sobrevivência como ferramenta para modelagem do tempo de falha, no referente estudo será o óbito do paciente, de acordo com sua Fração de Ejeção (FE) que pode ser definida brevemente como porcentagem de sangue que é bombeada do coração em sua fase sistólica.

Um estudo será realizado sobre o tempo até o óbito dos paciente de acordo com sua FE, mais precisamente sobre a fração de ejeção do ventrículo esquerdo (FEVE), avaliando a relação da fração de ejeção resultando no óbito, podendo apresentar evidências do desenvolvimento de uma síndrome de insuficiência cardíaca, evidências que auxiliaram no maior conhecimento da situação, aumentando o embasamento e agregando conhecimento adicional ao tópico, possibilitando que medidas de prevenção sejam tomadas antecipadamente, reduzindo a quantidade de pessoas que atingem graus mais elevados de risco de óbito e evitando o aumento em novos detentores de tais evidências.

2 Revisão de Literatura

2.1 Conceitos de análise de sobrevivência

Alguns conceitos serão definidos com fim de padronizar a linguagem utilizada, algumas formas de tratamentos das funções e alguns conhecimento necessários para o entendimento integral do estudo.

2.1.1 Tempo de falha

O tempo de falha, em análise de sobrevivência, é definido como o tempo decorrido até a ocorrência do evento de interesse, ou seja, a falha do experimento. Nesta monografia o evento de interesse é o óbito do paciente em acompanhamento.

2.1.2 Censura

Segundo Colosimo e Giolo, o tempo de censura é definido como a presença de observações parciais ou incompletas, ou seja, o evento de interesse não ocorreu por algum motivo de irrelevante conhecimento, dando assim origem aos tipos de censura.

2.2 Insuficiência cardíaca e fração de ejeção

'A insuficiência cardíaca (IC) é uma complexa síndrome cardiovascular com elevada prevalência, sendo que seu quadro clínico frequentemente é associado à dilatação do ventrículo, à diminuição da contratilidade e à reduzida fração de ejeção do ventrículo esquerdo (FE)' (MESQUITA; JORGE, 2009). Fração de ejeção diz respeito ao percentual de sangue que o ventrículo ejeta para a aorta na sístole, em relação a capacidade máxima, sendo assim pode-se classificar IC com base na porcentagem da fração de ejeção.

2.3 Função de Sobrevivência

Uma das principais funções probabilísticas usadas para descrever estudos de análise de sobrevivência é a função de sobrevivência, que é definida como a probabilidade de uma observação não falhar até um certo tempo t, ou seja, a probabilidade de uma observação sobreviver ao tempo t. Em termos probabilísticos, isto é escrito como:

$$S(t) = P(T \geq t) = \int_t^{\infty} f(x)dx$$

A função de sobrevivência S(t) é uma função monótona não crescente com as seguintes características: (COX, 2018)

$$\lim_{t \to 0} S(t) = 1 \qquad \lim_{t \to \infty} S(t) = 0$$

Em consequência, a função de distribuição acumulada é definida como a probabilidade de uma observação não sobreviver ao tempo t, isto é,

$$S(t) = 1 - F(t).$$

Ou seja, em um estudo médico onde o evento de interesse é a morte, a função de sobrevivência fornece a probabilidade de um indivíduo sobreviver além de um tempo t.

A função de sobrevivência é uma função não crescente no tempo com as propriedades de que a probabilidade de sobreviver pelo menos ao tempo zero é 1 e a probabilidade de sobreviver no tempo infinito é 0.

Para descrever a função de sobrevivência é geralmente utilizada uma representação gráfica de S(t), ou seja, o gráfico de S(t) versus t que é chamado de curva de sobrevivência. Uma curva íngreme representa razão de sobrevivência baixo ou curto tempo de sobrevivência e uma curva de sobrevivência gradual ou plana representam taxa de sobrevivência alta ou sobrevivência longa.

2.4 Função densidade de probabilidade

Muitas áreas da estatística trabalham com função densidade de probabilidade, mas na análise de sobrevivência, exclusivamente, essa função é definida como o limite da probabilidade de um individuo falhar no intervalo com Δt tendendo a 0, e pode ser definida por: (KLEIN; MOESCHBERGER, 2006)

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$

sendo f(t) positiva para todo t≥0 e $\int_0^\infty f(t) = 1$

2.5 Função de risco

A função de risco h(t) é uma parte importante na análise de sobrevivência, ela indica a forma em que a falhas ocorrem ao decorrer do tempo, por isso também conhecida como função taxa de falha e implica em muitas aplicações. A função de risco pode ser definida como:(LAWLESS, 2011)

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

que indica a probabilidade de um individuo falhar no intervalo $[t, t + \Delta t)$ dado que ele não experimentou a falha até o referido tempo t. Algumas relações envolvendo a definições anteriores estão dispostas a seguir:

- $h(t) = \frac{f(t)}{S(t)}$
- $S(t) = \exp\{-H(t)\}$
- $\log S(t)|_0^t = -\int_0^t h(u)du$
- S(t) = 1-F(t)
- $H(t) = -\int_0^t h(u)du$
- $f(t) = \frac{\partial F(t)}{\partial t}$

3 Metodologia

3.1 Estimador Kaplan-Meier

O estimador Kaplan-Meier, atualmente conhecido também como estimador de limite de produto, é o estimador utilizado por grande parte dos estudos sobre análise de sobrevivência, pois incorpora informação de todas as observações disponíveis, tanto censuradas quanto não-censuradas, considerando tempos de sobrevivência e censurados, também é uma estatística não-paramétrica, não requer conhecimento prévio de qual distribuição de probabilidade os dados provem para o uso do estimador.

O estimador de sobrevivência de Kaplan-Meier no tempo t é mostrado na equação (3.1.1). Aqui t_j ,j = 1,2,...,n é o total conjunto de tempos de falhas registradas (com t^+ o tempo máximo de falha), d_j é o número de falhas no tempo t_j , e n_j quantidade total de unidades observacionais sob risco de falha no momento t. O estimador de Kaplan-Meier é definido por:

$$\hat{S}(t) = \prod_{j:t_j \leq t} \frac{(n_j - d_j)}{n_j} = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right), \quad para \quad 0 \leq t \leq t^+ \tag{3.1.1}$$

Seja, $p_j = P(T \geq t_j | T \geq t_{j-1})$ a probabilidade de que a unidade observacional não falhe até t_j dado que não falhou até o instante t_{j-1} e $q_j = P(T \leq t_j | T \geq t_{j-1})$ probabilidade de falhar em t_j dado que sobreviveu até o instante imediatamente anterior à t_{j-1}.

Considerando n_j fixo, Kaplan e Paul Meier afirmaram que d_j ~Binomial(n_j,q_j). Sendo assim pode se mostrar a origem do estimador Kaplan-Meier calculando o respectivo estimador de máxima verosimilhança de $\widehat{S}(t)$, a função de sobrevivência.

$$P(D = d_j) = \binom{n_j}{d_j} q_j^{d_j} (1 - q_j)^{n_j - d_j} = L(p)$$

Aplicando o produtório.

$$\prod_{i=1}^{n} \left[\binom{n_j}{d_j} q_j^{d_j} (1 - q_j)^{n_j - d_j} \right] = \left[\prod_{i=1}^{n} \binom{n_j}{d_j} \right] q_j^{d_j} (1 - q_j)^{(n_j - d_j)}$$

Onde $q_j = 1 - pj$.

$$\begin{aligned} l(p) &= \log \left(\left[\prod_{i=1}^{n} \binom{n_j}{d_j} \right] q_j^{d_j} (1 - q_j)^{(n_j - d_j)} \right) \\ &= \log \left(\prod_{i=1}^{n} \binom{n_j}{d_j} \right) + d_j \log(1 - p_j) + (n_j - d_j) \log\left(1 - (1 - p_j)\right) \end{aligned}$$

Derivando a função log-verossimilhança.

$$\frac{\partial l}{\partial p_j}(p) = d_j \left(\frac{-1}{1 - p_j} \right) + (n_j - d_j) \frac{1}{p_j}$$

Igualando a derivada parcial em relação a p_j a zero.

$$\begin{aligned} 0 &= d_j \left(\frac{-1}{1-p_j} \right) + (n_j - d_j) \frac{1}{p_j} \\ &= \frac{-p_j dj + (1-p_j)(n_j - d_j)}{p_j(1-p_j)} \\ &\Rightarrow \widehat{p_j} = 1 - \frac{d_j}{n_j} \end{aligned}$$

Verificando se p_j é ponto de máximo pelo teste da segunda derivada.

$$\begin{aligned} \frac{\partial^2 l}{\partial^2 p_j}(p) &= \frac{d_j}{(1-p_j)^2} - \frac{(n_j - d_j)}{p_j^2} \\ &= \frac{d_j}{1 - 1 + \frac{d_j}{n_j}} - (n_j - d_j) \frac{n_j^2}{(n_j - d_j)^2} \\ &= n_j - \frac{n_j^2}{n_j - d_j} \\ &= \frac{n_j^2 - n_j d_j - n_j^2}{n_j - d_j} \\ &= \frac{-n_j d_j}{n_j - d_j} < 0 \end{aligned}$$

Como $S(t) = \prod_{j:t_j<T} (1 - \frac{d_j}{n_j}) = \prod_{j:t_j<T} p_j$ e $\widehat{p}_j = 1 - \frac{d_j}{n_j}$ é EMV, então pele principio da invariância.

$$\widehat{S}(t) = \prod_{j:t_j<T} (1 - \frac{d_j}{n_j}) = \prod_{j:t_j<T} \widehat{p}_j$$

3.2 Comparação de curvas

3.2.1 Teste logRank

A comparação entre duas curvas de sobrevivência é importante a fim de determinar se existe diferenças significativas entre elas, concluindo se métodos ou processos possuem influência sob a unidade de estudo. Para construção do teste logRank, serão considerados tempos de 'falha' separadamente por grupos de interesse.

Considerando hipótese nula como ausência de diferença entre os grupos, pode ser feito um teste onde se leva em consideração a diferença na quantidade de 'falhas' observadas nos tempos de falha, em seus respectivos grupos e o numero esperado de falhas.

Em uma tabela de dupla entrada as células da tabela podem ser definidas por apenas d_{1J}, o número de falhas em $t_{[}j]$ no grupo 1 e que será tratada como uma variável aleatória com distribuição hipergeométrica, e que a probabilidade do numero de falhas assuma o valor d_{1J} é definido por:

$$\frac{\binom{d_j}{d_{1j}} \binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}}$$

O valor esperado para variável que segue a distribuição hipergeométrica é dada por $e_{1j} = n_{1j} d_j / n_j$, sendo assim o valor esperado das falhas em $t_{(j)}$ no grupo 1. Sob hipótese

nula não há divergência entre as probabilidades de falha sendo assim em $t_{(j)}$ a probabilidade de falha seria d_j/n_j e multiplicando essa probabilidade pelo numero de indivíduos em risco do grupo 1, n_{1j}, tem-se exatamente o valor esperado. Com o objetivo de obter uma medida geral do desvio do valor esperado e do observado, a forma mais direta é calculando o somatório desses desvios em cada tempo de falha:

$$U_L = \sum_{j=1}^{r} d_{1j} - e_{1j}$$

e a variância de d_{1j} é definida por:

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

logo, a variância de U_L, a diferença entre observado e esperado geral, é definida:

$$var(U_L) = \sum_{j=1}^{r} v_{ij} = V_L$$

E assim podemos definir a estatística de teste, com uma amostra de tamanho satisfatório pode se inferir que a estatística teste segue uma distribuição Normal(0,1) e usando conhecimentos prévios sabe-se que o quadrado de uma variável aleatória normal padrão segue uma distribuição qui-quadrado com 1 grau de liberdade, respectivamente:

$$\frac{U_L}{\sqrt{V_L}} \sim N(0,1) \qquad \frac{U_L^2}{V_L} \sim X_1^2$$

O Teste de logRank pode ser estendido para casos em que será comparado três ou mais curvas/grupos, basta ser calculado os análogos das estatísticas U_L e V_L para o caso de comparação de duas curvas, onde k=1,...,g-1, e g o numero de grupos a ser comparado, as notações para as novas estatísticas de teste são respectivamente:

$$U_{Lk} = \sum_{j=1}^{r} \left(d_{kj} - \frac{n_{kj} d_j}{n_j} \right) \qquad V_{Lkk'} = \sum_{j=1}^{r} \frac{n_{kj} d_j (n_j - d_j)}{n_j (n_{j-1})} \left(\delta_{kk'} - \frac{n_{k'j}}{n_j} \right)$$

onde,

U_{LK}=Valor observado- valor esperado da LK-ésima célula.

V_{LK}= Expressão para variância e covariância entre os pares de valores.

$$\delta_{kk'} = \begin{cases} 1 & \text{se } k = k' \\ 0 & \text{caso contrário} \end{cases}$$

Assim será composta a matriz de variância e covariância, onde estarão localizada as variâncias de U_{Lk} na diagonal principal e as covariâncias respectivas fora da diagonal.

$$V_L = \begin{pmatrix} V_{L11} & V_{L12} \\ V_{L21} & V_{L22} \end{pmatrix}$$

e para testar a hipótese nula de que não há diferença entre os grupos, usaremos a estatística

do seguinte resultado $U_L^{'} V_L^{-1} U_L$ que segue uma distribuição qui-quadrado.

3.2.2 Teste de Wilcoxon

O teste de Wilcoxon é usado para comparação de curvas assim como logRank porém não possui os mesmos pré-supostos e também é conhecido como teste de Breslow. De forma similar ao logRank o Wilcoxon pode ser usado para comparar duas curvas de sobrevivência e sua estatísticas consistem em,

$$U_W = \sum_{j=1}^{r} n_j(d_{1j} - e_{1j}) \qquad V_W = \sum_{j=1}^{r} n_j^2 v_{1j}$$

e consequentemente sua estatística teste será expressa por:

$$W_W = \frac{U_W^2}{V_W} \sim X_1^2$$

Assim como o logRank o teste de Wilcoxon pode ser estendido para casos em que é almejada comparação de três ou mais curvas e para isso, de forma análoga, serão calculadas as estatísticas, U_{Wk} e $V_{Wkk'}$ referentes ao teste,

$$U_{Wk} = \sum_{j=1}^{r} n_j \left(d_{kj} - \frac{n_{kj} d_j}{n_j} \right) \qquad V_{Wkk'} = \sum_{j=1}^{r} n_j^2 \left(\frac{n_{kj} d_j (n_j - d_j)}{n_j (n_{j-1})} \right) \left(\delta_{kk'} - \frac{n_{k'j}}{n_j} \right)$$

$V_{Wkk'}$= Expressão para variância e covariância,para o teste de Wilcoxon, entre os pares de valores.

A matriz de variância e covariância para o teste será composta pelas variâncias em sua diagonal principal e suas respectivas covariâncias fora da diagonal principal.

$$V_W = \begin{pmatrix} V_{W11} & V_{W12} \\ V_{W21} & V_{W22} \end{pmatrix}$$

e para testar a hipótese nula de que não há diferença entre os grupos, usaremos a estatística do seguinte resultado $U_W^{'} V_W^{-1} U_W$ que segue uma distribuição qui-quadrado.

3.2.3 Comparação dos testes

Pode se notar que os teste são relativamente parecidos, porém cada um possui suas particularidades. Foi apresentado alguns testes de comparação de curvas mas existem muitos outros que têm estrutura parecida, e derivam de um mesmo 'radical', sendo esse:

$$U_A = \sum_{j=1}^{r} a_j \left(d_{1j} - e_{1j} \right) \qquad var(U_A) = \sum_{j=1}^{r} a_j^2 (v_{1j})$$

e a razão dessas estatísticas resultam em uma distribuição qui-quadrado com 1 grau de liberdade, expressa por:

$$S = \frac{U_A^2}{V_A} = \frac{\sum_{j=1}^{r} a_j \left(d_{1j} - e_{1j} \right)}{\sum_{j=1}^{r} a_j^2 (v_{1j})} \sim X_1^2$$

e cada teste é definido por um a_j diferente,

- se a_j=1 então o teste referido será o logRank, que tem como pressuposto risco proporcional e atribui peso igual para todos o tempos.
- se a_j=n_j será o teste de Wilcoxon, os pesos estão concentrados na proporção inicial do eixo de tempo.
- se a_j=$\sqrt{n_j}$ é o teste de Tarone e Ware, o peso atribuído por este teste é um meio termo do logRank e Wilcoxon.

3.3 Estimação probabilística da curva de sobrevivência

3.3.1 Distribuição Log-logística

É dito que uma variável aleatória T segue uma distribuição Log-logística com parâmetros $\alpha, \gamma > 0$, se ela assumir valores no intervalo [0,∞) e sua função de densidade for dada por:

$$f(t) = \frac{\gamma}{\alpha} t^{\gamma-1} \left[1 + \left(\frac{t}{\alpha}\right)^{\gamma}\right]^{-2}$$

em que α é o parâmetro de escala e $\gamma > 0$ o parâmetro de forma.

Dado isso, obtém-se as funções de sobrevivência e taxa de falha, respectivamente, por:

(i) $S(t) = \frac{1}{1+(t/\alpha)^{\gamma}}$

(ii) $h(t) = \frac{\gamma(t/\alpha)^{\gamma-1}}{\alpha[1+(t/\alpha)^{\gamma}]}$

A função de risco da log-logística também apresenta formas unimodais ($\gamma > 1$) e decrescente (γ 1).Se T tem distribuição Log-logística com parâmetros α e γ, então a variável Y = log(T) tem distribuição logística com parâmetros $-\infty < \mu < \infty$ e $\sigma > 0$, em que $\gamma = 1/\sigma$ e $\alpha = \exp(\mu)$.

3.3.2 Distribuição Weibull

A distribuição Weibull é robusta quando se trata do ajuste de algumas variáveis pois é uma distribuição que devido ao seu parâmetro de forma, torna-se bem volátil, ou seja, se ajusta à grande variedade de funções de risco acumulado ou da curva TTT, em outras palavras, a distribuição Weibull é uma frequente candidata a se considerar para construção de um modelo probabilístico devido a seu parâmetro de forma γ que se:

- $\gamma < 1$ a função de risco é decrescente.
- $\gamma > 1$ a função de risco é crescente.
- $\gamma = 1$ a função de risco é constante.

Caso $\gamma = 1$ a distribuição em questão é uma exponencial que é uma caso particular da Weibull. A distribuição Weibull possui a seguinte função de densidade, sendo γ seu parâmetro de forma e α seu parâmetro de escala.

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\} \tag{3.3.1}$$

Dada a função de densidade é possível determinar as funções de sobrevivência e risco respectivamente representadas a seguir:

(i) $S(t) = exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}$

(ii) $h(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1}$

A distribuição Weibull também possui esperança e variância definidas que auxiliam e agregam clareza ao analisar a variável, que podem ser representadas por:

(i) $\mathrm{E(t)}=\alpha\Gamma\left[1 + (1/\gamma)\right]$

(ii) $\mathrm{Var(t)}=\alpha^2\left[\Gamma\left[1 + (2/\gamma)\right] - \Gamma\left[1 + (1/\gamma)\right]^2\right]$

onde $\Gamma(k) = \int_0^{\infty} x^{k-1} exp\left\{-x\right\} dx$.

3.3.3 Análise de diagnóstico

Uma etapa importante na análise de um ajuste de regressão é a verificação de possíveis afastamentos das suposições feitas para o modelo, especialmente para o componente aleatório e para a parte sistemática do modelo, bem como a existência de observações discrepantes com alguma interferência desproporcional ou inferencial nos resultados do ajuste. Tal etapa, conhecida como análise de diagnóstico, tem longa data, e começou com a análise de resíduos para detectar a presença de pontos aberrantes e avaliar a adequação da distribuição proposta para a variável resposta. (PAULA, 2004)

Na perspectiva de análise de sobrevivência, os resíduos mais utilizados são os de Cox-Snell, martingal e deviance.

Resíduos de Cox-Snell

Os resíduos de Cox-Snell auxiliam a examinar o ajuste global do modelo. Esses resíduos são definidos por:

$$\hat{e_i} = \hat{H}(t_i|x_i)$$

em que $\hat{H}(.)$ é a função de risco acumulado obtida do modelo ajustado.

- Segundo Lawless (1982), os resíduos $\hat{e_i}$ vem de uma população homogênea e devem seguir uma distribuição exponencial padrão se o modelo for adequado.
- O gráfico de $\hat{e_i}$ versus $\hat{H}(\hat{e_i})$ deve ser aproximadamente uma reta para que o modelo exponencial seja adequado.
- Dado que $\hat{H}(\hat{e_i}) = -\log(S(\hat{e_i}))$, o gráfico das curvas de sobrevivência desses resíduos, obtidas por Kaplan-Meier e pelo modelo exponencial padrão, também auxiliam na verificação da qualidade do modelo ajustado. Ou seja, $\exp\{-\hat{e_i}\}$ versus $\hat{S_{KM}}(\hat{e_i})$.

Resíduos Martingal

Os resíduos de martingal são utilizado para identificar as discrepâncias entre o modelo ajustado e o conjunto de dados. Os resíduos são definidos por:

$$\hat{r}_{M_i} = \delta_i - \hat{H}(t_i|x_i) \tag{3.3.2}$$

Onde δ_i é a variável indicadora de falha e $\hat{H}(t_i|x_i)$ é o resíduo de Cox-Snell, definido anteriormente .

- O resíduo de martingal é usado pra avaliar a melhor forma funcional da variável (COLOSIMO; GIOLO, 2014).

Deviance
Para os modelos de regressão paramétricos, os resíduos deviance são definidos por:

$$\hat{r}_{D_i} = sinal(\hat{r}_{M_i}) \left[-2 \left(\hat{r}_{M_i} + \delta_i \log(\delta_i - \hat{r}_{M_i})\right)\right]^{1/2} \tag{3.3.3}$$

- Se o modelo for apropriado, esses resíduos devem apresentar um comportamento aleatório em torno de zero.

- Esses resíduos são uma tentativa de tornar os resíduos martingal mais simétricos em torno de zero, facilitam, em geral, a detecção de pontos atípicos

3.3.4 Método de estimação de máxima verossimilhança

Na estimação paramétrica da curva é de grande importância o conhecimento de seus respectivos parâmetros do modelo probabilístico escolhido e para este objetivo ser alcançado o método utilizado será o de estimação por máxima verossimilhança. Sendo t_1,t_2,...,t_n uma amostra aleatória de tamanho n da variável tempo T e com função de densidade $f(t|\theta)$ em que θ é o vetor de parâmetros, a máxima verossimilhança correspondente é definida por:

$$L(\theta) = \prod_{i=1}^{n} f(t_i|\theta)$$

Este método também suporta dados que contém censuras para a estimação dos parâmetros, levando em consideração a função de sobrevivência, a função de distribuição e uma função indicadora para censura, definida por:

$$\delta_i = \begin{cases} 1, & \text{se houver falha no tempo } t_i \\ 0, & \text{caso contrário} \end{cases}$$

sendo assim a função de máxima verossimilhança que incorpora censuras é representada da forma alternativa, representada por:

$$L(\theta) = \prod_{i=1}^{n} f(t_i;\theta)^{\delta_i} S(t_i;\theta)^{1-\delta_i} \tag{3.3.4}$$

$$= \prod_{i=1}^{n} h(t_i;\theta)^{\delta_i} S(t_i;\theta)$$

Desta forma as observações censuradas contribuem para a função de máxima verossimilhança (3.3.1) com sua respectiva função de sobrevivência e cada observação onde se verificou falha contribui com sua função de densidade.

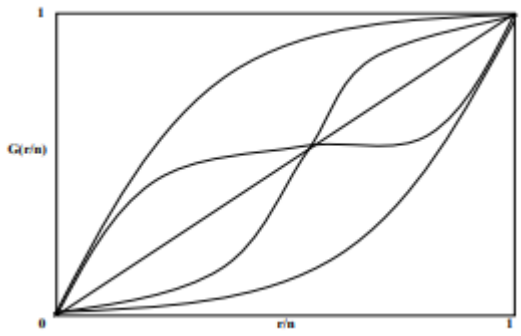
Para determinar qual a função de densidade a variável tempo segue, existem métodos com tal objetivo, um deles é um método visual que baseia-se no gráfico da curva Tempo

Total em Teste descrita pelo gráfico:

$$G(r/n) = \frac{[(\sum_{i=1}^{r} T_{i:n}) + (n-r)\, T_{r:n}]}{(\sum_{i=1}^{n} T_i)}$$

por r/n, onde r e $T_{i:n}$ =1,...,n, são estatísticas de ordem da respectiva amostra.

De acordo com o formato que a curva assumir pode se associa-la a alguma função de densidade cujo comportamento seja semelhante.



Cada curva acima representa a característica que a função de densidade pode assumir, seja ela, monótona crescente, monótona decrescente, unimodal ,*etc...*, a interpretação pode ser feita a partir da função de risco acumulada porém de forma contraria, ambas convergindo ao mesmo resultado da função de densidade.

3.4 Modelo de Cox

Considerando o tempo a variável de interesse, que consiste no tempo até que um certo evento se verifique, podendo ser influenciado por covariáveis, um modelo que candidato para esta situação um modelo adequado é o modelo de Cox, também denominado de modelo de Riscos Proporcionais, o modelo tem como pressuposto que a razão das taxas de risco dos dois indivíduos diferentes é constante no tempo, ou seja, a razão é independente do tempo.

Determinado pela função de risco, o modelo de Cox é composto por um componente paramétrico e outro não paramétrico. Este último componente atribui grande versatilidade ao modelo, trazendo assim uma ampla utilização deste. Por não assumir distribuição de probabilidade para o tempo de sobrevivência este modelo é amplamente utilizado e popular na análise de sobrevivência e por consequência o objetivo neste modelo é estimar o efeito das covariáveis e não estimar os parâmetros da distribuição de tempo de sobrevivência.

A definição do modelo exige a determinação de uma função de risco base, que define o componente não paramétrico, $h_0(t)$, definido por:

$$\hat{H}_0(t) = \sum_{j:t_j<t} \frac{d_j}{\sum_{I\in R_j} exp(x_I^t \hat{\beta})} \tag{3.4.1}$$

e seu componente paramétrico, usado frequentemente na forma multiplicativa.

$$\begin{aligned} G(x^t\beta) &= \exp(x^t\beta) \\ &= \exp(x_1\beta_1 + \cdots + x_p\beta_p) \end{aligned}$$

desta forma o modelo de riscos proporcionais de Cox usual é dado por:

$$h(t|x) = h_0(t)exp(x^t\beta)$$

em que $h_0(t)$ é não negativa no tempo e não especificada, sendo chamada também de função basal, pois, em grande parte dos modelos,quando x=0, $h_0(t) = h(t)$. Essa definição comprova a suposição de riscos proporcionais, ja que:

$$\frac{h_i(t|x_i)}{h_j(t|x_j)} = \frac{\exp(x_i\beta)}{\exp(x_j\beta)} = K$$

em que K é constante no tempo.

3.4.1 Interpretação dos parâmetros

O modelo de Cox também podem ser utilizados para se obter risco relativo entre as categorias delimitadas por certa característica, definindo seus coeficientes e utilizando-o como expoente, exp^{β} sendo β o coeficientes do modelo e sua interpretação é similar a de risco relativo usual.

Além de evidenciar a proporcionalidade de riscos os resíduos de Cox-Snell possuem também a função de avaliar a qualidade do ajuste do modelo de Cox, o resíduo de Cox-Snell é representado por:

$$\hat{e_i} = \hat{H}_0(t_i)\exp\left\{\sum_{k=1}^{p} x_{ip}\hat{\beta}_k\right\}$$

com $\hat{H}_0(t_i)$ estimado por (3.4.1). O ajuste do modelo será evidenciado graficamente, caso o modelo esteja bem ajustado, pelo gráfico de $\hat{H}(\hat{e_i})$ versus $\hat{e_i}$ que deve ser semelhante a uma reta.

Portanto temos que, quando se se trata de variável qualitativa o risco de falha dos indivíduos do grupo i é $\exp(\beta)$ vezes o risco de falha de indivíduos do grupo j, mantendo-se fixa as demais covariáveis. Quando a variável tiver 3 categorias, deve-se escolher uma como referencia e usar duas variáveis indicadoras para representar a terceira;

Quando a variável for quantitativa, com o aumento em uma unidade na covariável x_p, indivíduos com uma unidade a mais tem risco $\exp(\beta_p)$ vezes maior de falhar, mantendo-se fixa as demais covariáveis.

3.5 Materiais

O banco de dados possui as seguintes variaveis:

Variável	Descrição
FEVE	Fração de ejeção do ventrículo esquerdo
Óbito	O paciente veio a óbito
Data inicial	Data de inicio do segmento
Data final	Data do final do segmento

Com as variáveis data inicial e data final, foi calculado o tempo de segmento de cada paciente sendo a diferença das datas, em dias. A nova variável (tempo) possui valor observado mínimo de 35 dias e máximo de 2978 dias e media de 1788.

4 Resultados e Discussões

4.1 Análise Descritiva

Para compreender melhor a natureza dos dados, de forma que a modelagem proposta seja de fato viável, foi feita uma análise exploratória dos dados, obtendo-se os seguintes resultados:

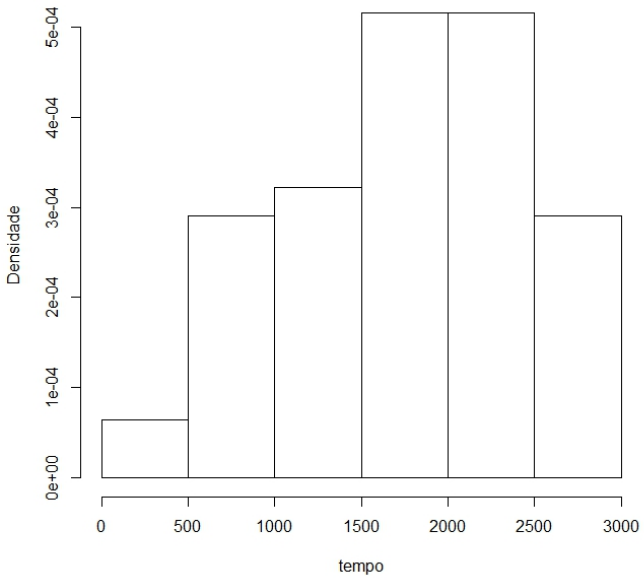


Figura 1: Histograma do tempo de acompanhamento de 62 pacientes.

A Figura 1 mostra que, de início, a frequência dos tempos de sobrevivência não é muito alta , mas que no decorrer do estudo é possível observar um aumento na frequência. Como não possuiu uma concentração de dados observados ao final do histograma, possivelmente não houveram muitos dados que foram censurados ao final do estudo.

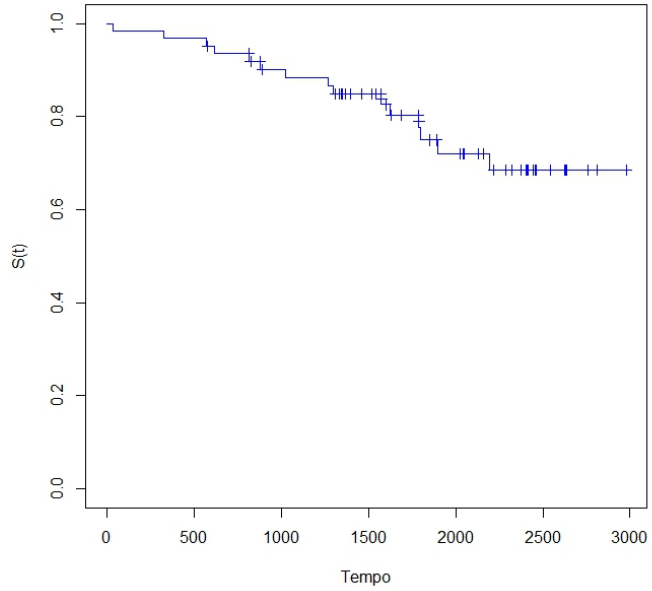


Figura 2: Gráfico da função de sobrevivência estimada através do método de Kaplan-Meier

Analisando visualmente a Figura 2, pode se verificar que a curva de sobrevivência se estabiliza ao decorrer do tempo, o que indica a possível existência de uma fração de cura.

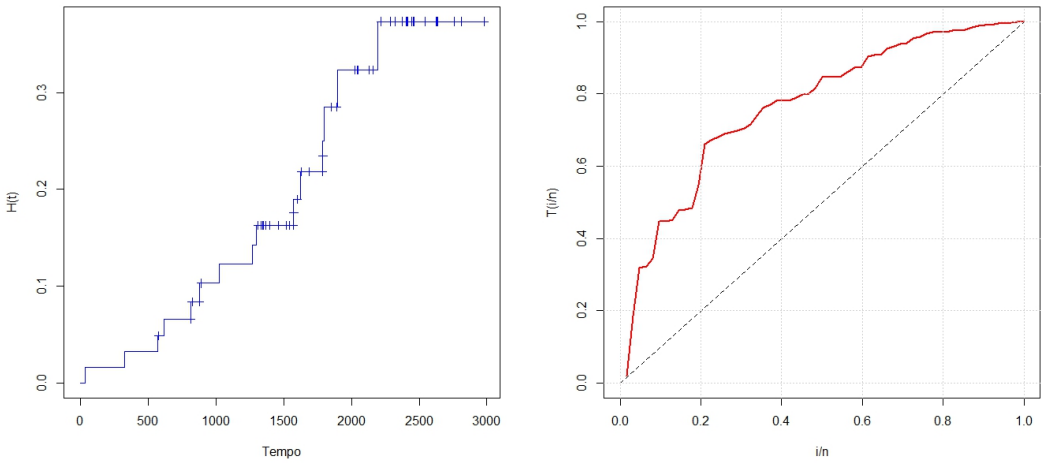


Figura 3: Função de risco acumulado e curva TTT respectivamente

O comportamento da função de risco acumulado e principalmente a curva TTT evidenciam que a função de risco estudada possui um padrão monótono crescente, uma vez que a curva TTT apresenta uma forma nitidamente côncava e o risco acumulado um leve formato convexo.

Devido ao comportamento descrito da curva TTT e função de risco acumulado, as distribuições candidatas para ajuste dos tempos de falha (óbito do paciente), devem apresentar uma função de risco monotonicamente crescente e sendo assim aparecem algumas distribuições possíveis para o ajuste como a Weibull e a Log-logística pois apresentam

as particularidades requiridas de função de risco. Posteriormente serão utilizadas para modelagem dos dados e comparação entre os respectivos modelos.

Para a categoria delimitada pelo valor de referencial 50 (Fração de ejeção de 50 por cento).

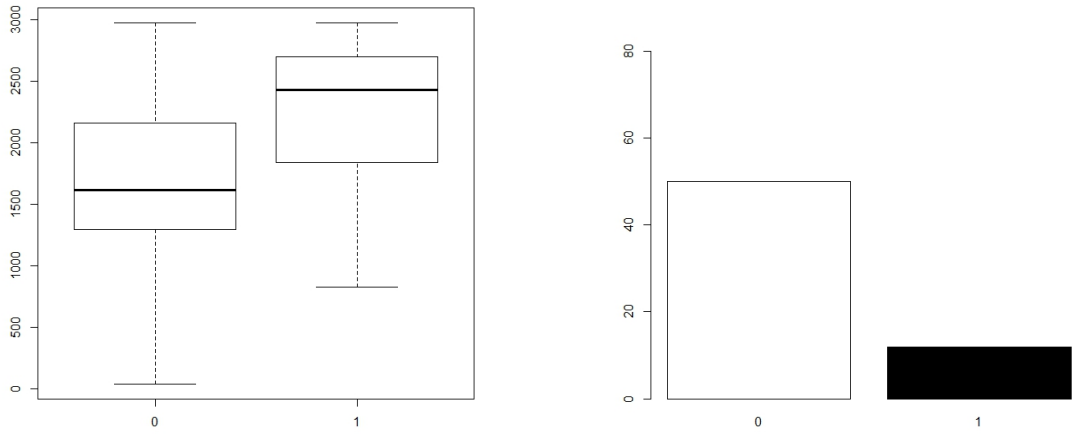


Figura 4: Boxplot dos tempos de acompanhamento e frequência dos pacientes em cada classe sendo 0 '<50' e 1 um pacientes com FEVE '>50'

Para análise inicial,de forma visual, é possível notar que poucos indivíduos estão na categoria de FEVE acima de 50% de acordo com o histograma e ter uma percepção aproximada dos tempos em que o indivíduos de suas respectivas categorias foram acompanhados, até a ocorrência do evento de interesse ou não.

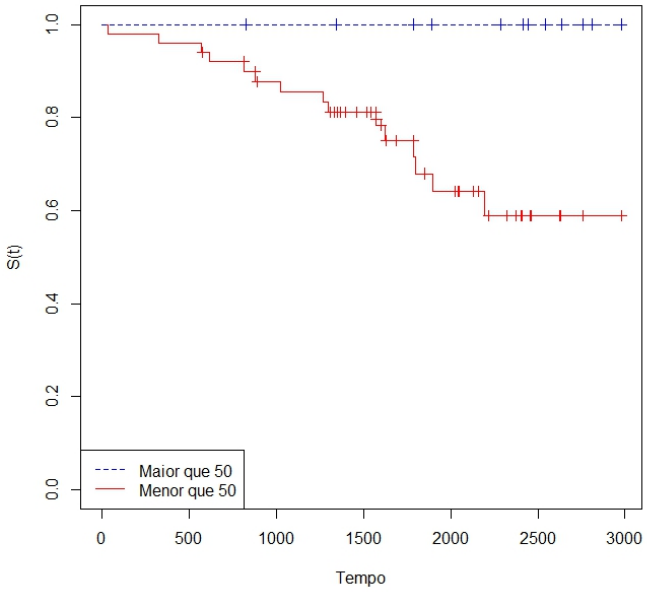


Figura 5: Gráfico da função de sobrevivência, para a variável categorizada ">50 ou <50", estimada através do método de Kaplan-Meier.

Avaliando a Figura 5, as curvas de sobrevivência estimadas por Kaplan-Meier, no início do estudo, são próximas como esperado mas rapidamente se distanciam evidenciando o pressuposto de riscos proporcionais, sendo assim pelo teste log-rank foi obtido um p-valor=0.02, o que significa que as curvas possuem diferença significativa, a categoria limitada por ser maior que 50 é uma reta devido a não observação do evento de interesse nenhum indivíduo pertencente a respectiva. O que já era esperado pois quanto menor a FEVE, maior o risco de falha.

Outra categorização feita foi utilizando o seguinte critério, primeira classe está definida com FEVE entre 49 a 40, segunda de 39 a 30 e a terceira categoria é menor que 30. Dada esta categorização é possível avaliar que:

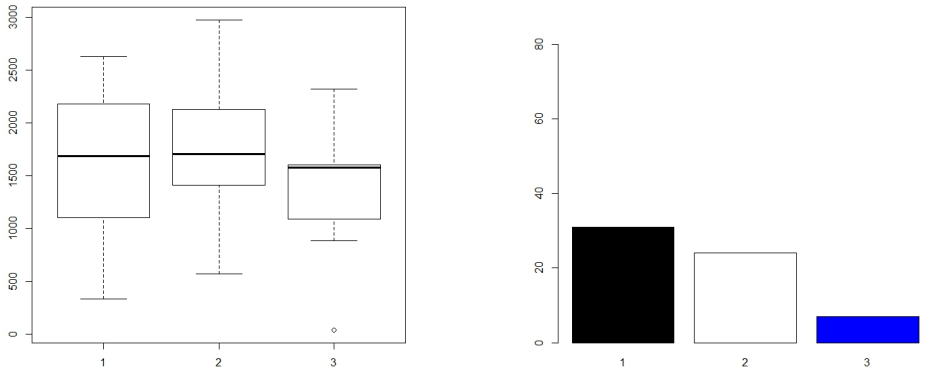


Figura 6: Boxplot dos tempos de acompanhamento e frequência dos pacientes em cada classe sendo 1 '<40', 2 '39-30' e 3 pacientes com FEVE '<30'

O tempo de acompanhamento dos pacientes foi relativamente similar entre as categorias, a mediana das categorias estão do mesmo modo, ou seja, as categorias possuem em sua maioria características semelhantes em relação ao tempo de observação de cada indivíduo.

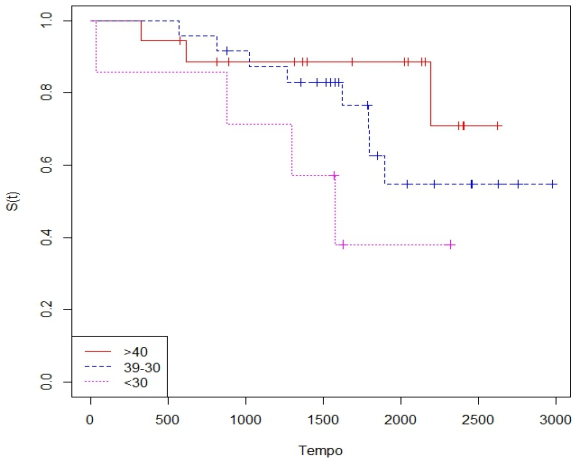


Figura 7: Gráfico da função de sobrevivência, para a variável categorizada ' > 40, 39 a 30 e <30 ', estimada através do método de Kaplan-Meier

Diferentemente das outras categorizações, a análise visual das curvas estimadas por Kaplan-Meier é mais complexa uma vez que estão presentes 3 categorias e consequentemente 3 curvas a serem comparadas, visualmente é possível perceber que no inicio do estudos as categorias possuem curvas muito próximas, como esperado, mas ao decorrer do estudo elas se distanciam de maneira cada vez mais perceptível e para se confirmar o que apontam as evidências foi feito um teste log-rank todavia devido a quantidade de categorias e para comparar duas a duas também foi feito uma correção de Bonferroni obtendo-se os seguintes p-valores:

Tabela 1: Tabela de p-valores das comparações 2 a 2.

	1	2
2	0.082990318	-
3	0.006088399	0.3865899

As curvas ">40","39-30"e "<30"estão representadas respectivamente pelos valor 1,2 e 3 na Tabela 1 e com base nos p-valores obtidos é possível concluir que as únicas curvas que não possuem diferença significativa entre si são as curvas "2"e "3"e a respeito das demais com o respectivo nível de significância, =0.1, o p-valor evidência que a curvas possuem diferença significativa.

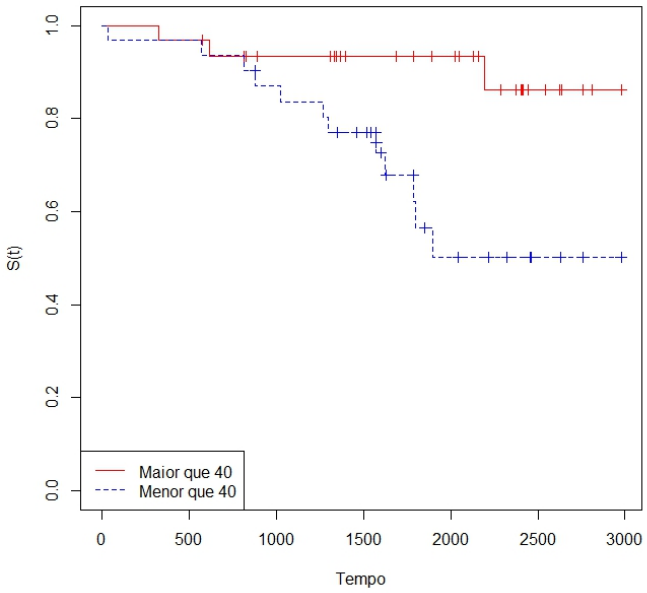


Figura 8: Gráfico da função de sobrevivência, para FEVE categorizada em '>40 ou <40', estimada através do método de Kaplan-Meier

A partir da estimação das funções de sobrevivência pelo método Kaplan-Meier e a análise visual da Figura 8, pode se compreender que as curvas, no inicio do estudo estão relativamente semelhantes e a partir de um certo momento se distanciam evidenciando o pressuposto de riscos proporcionais, possibilitando assim que o teste log-rank fosse executado e assim sendo gerado o p-valor= 0.008, evidencia suficiente para concluir que as curvas se distanciam ao decorrer do estudo de forma significativa, ou seja, são diferentes.

4.2 Modelo Log-Logístico

Por evidências da curva TTT foi proposto um modelo com distribuição Log-Logística para analisar e comparar os resultados obtidos entre os modelos de interesse.

Os parâmetros que maximizam a função de densidade da Log-logística para o melhor ajuste estão representados a seguir:

Tabela 2: Tabela dos parâmetros que maximizam a distribuição referida.

α	β
4644.119937	1.350181

É possível representar o ajuste à curva de sobrevivência utilizando os resultados da Tabela 2 e assim obtendo-se o seguinte resultado:

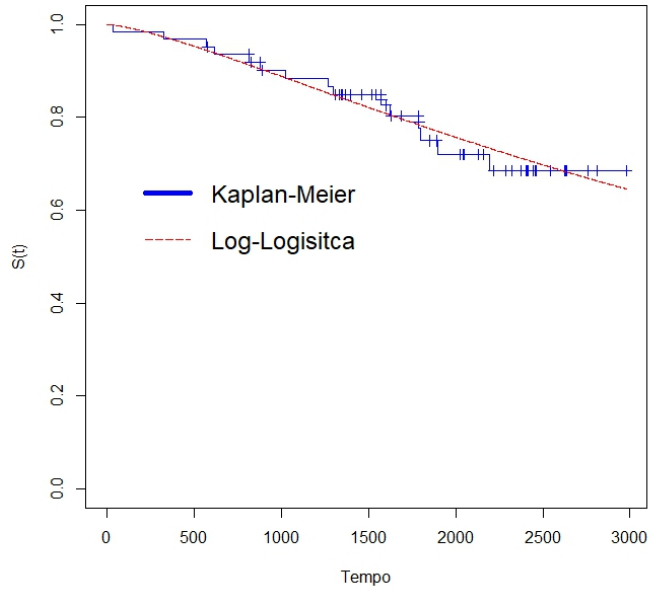


Figura 9: Ajuste do modelo com distribuição Log-Logística e a função de sobrevivência estimada por Kaplan-Meier

O ajuste do modelo proposto com distribuição log-logística sob a curva de sobrevivência estimada por Kaplan-Meier é no mínimo satisfatória, se considerado que as curvas, ao longo do tempo, estão relativamente próximas salvo alguns pontos excepcionais em que é possível notar uma distância um pouco maior que nos demais pontos da curva, todas as conclusões com base na Figura 9 e deduções visuais.

4.2.1 Análise de diagnóstico

Para verificar as suposições do erro do modelo, adequabilidade ou não, uma análise de diagnóstico é realizada e caso durante esta etapa os resíduos do modelo não estiverem de acordo com o esperado o modelo deve ser descartado. Dito isso, obtemos os seguintes resultados:

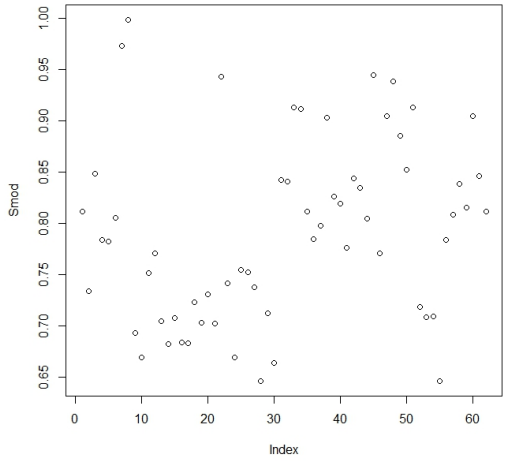


Figura 10: Resíduo de Cox.

O resíduo de Cox deve apresentar um padrão aleatório para que o modelo seja adequado e com auxilio da Figura 10 não pode se notar nenhum padrão definido, ou seja, os resíduos são aleatórios atendendo aos pressupostos para o uso do modelo.

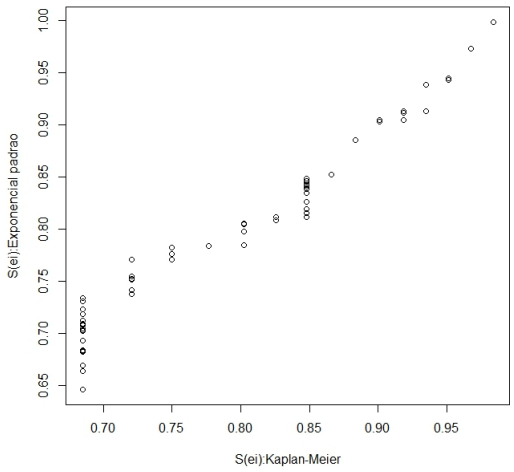
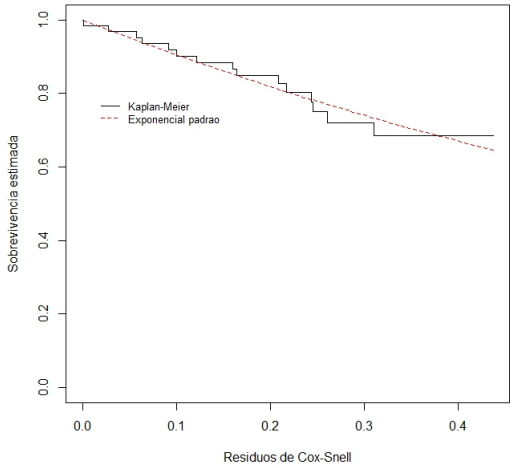


Figura 11: $S(e_i)$ por Kaplan-Meier versus $S(e_i)$ da exponencial padrão.

Pode se notar, pela Figura 11, que o gráfico possui sinais de que o formato representado é uma reta, ou seja, os resíduos do modelo assumem a distribuição de uma exponencial padrão como esperado para afirma que os dados são adequados.



A curva de sobrevivência dos resíduos estimada por Kaplan-Meier, ou seja, $exp\left\{-\hat{e}_i\right\}$ versus $\hat{S}_{KM}(\hat{e}_i)$, também auxilia na verificação de qualidade do modelo ajustado e como observado na Figura 12 pode se concluir que seguem aproximadamente uma distribuição exponencial padrão, formalizando que o ajuste do modelo é razoavelmente adequado.

Figura 12: Curva de sobrevivência do resíduo estimada por kaplan-Meier ajustada por uma exponencial padrão.

O resíduo de martingal também é usado pra examinar a melhor forma funcional paras variáveis do modelo, no caso o modelo log-logístico. Analisando o respectivo resíduo e com auxilio da Figura 13 é possível notar uma evidência para a categorização da variável.

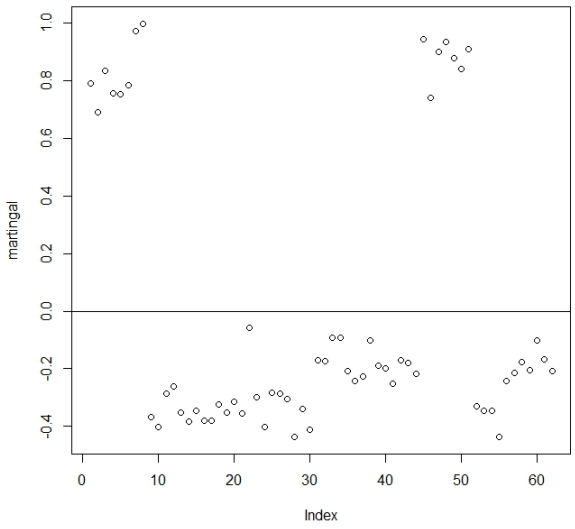


Figura 13: Resíduo de Martingal

Analisando de forma superficial é possível sugerir algumas formas de categorização que podem ser desenvolvidas posteriormente, como a separação em 3 grupo de pacientes ou em 2 grupos de acordo com sua FEVE.

4.2.2 Modelo Log-Logístico para variável categorizada.

Os parâmetros da distribuição log-logística que proporcionam o ajuste para as categorias '>40' e '<40' estão dispostos na tabela a seguir:

Tabela 3: Tabela dos parâmetros para as categorias '>40' e '<40' respectivamente.

	α	γ
Cat >40	9251.41	1.438081
Cat <40	2915.159	1.445982

O modelo Log-logístico proposto para cada variável categorizada possui um ajuste bom ao longo do tempo, salvo alguns pontos onde o modelo superestima a função de sobrevivência mesmo assim é possível concluir que o ajuste está razoável. Visualmente podemos conferir os resultados a seguir:

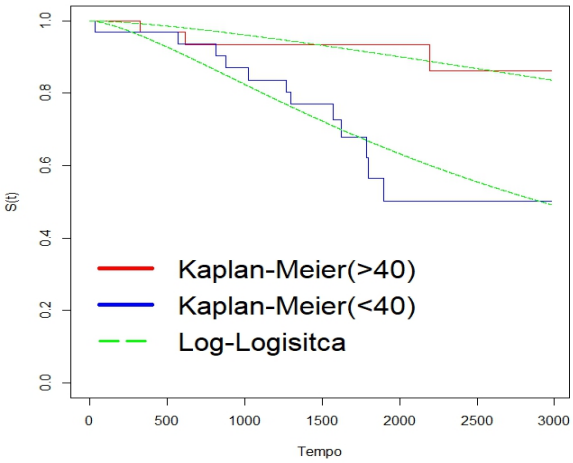


Figura 14: Ajuste para o modelo com as variáveis explicativas

Análise de diagnóstico

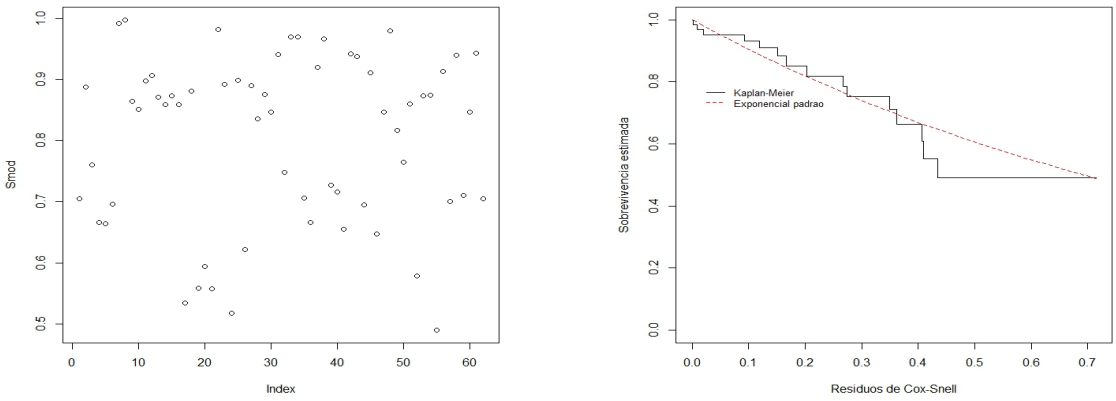


Figura 15: Residuo de Cox-Snell e curva de sobrevivência do resíduo estimada por kaplan-Meier ajustada por uma exponencial padrão.

Ao analisar a Figura 15 não é possível observar nenhum padrão bem definido, concluindo assim que a distribuição de pontos possui princípios de aleatoriedade e também que a curva de sobrevivência do resíduo estimada por kaplan-Meier possui um bom ajuste por uma exponencial padrão, ou seja, os dados provem de uma população homogenia segundo (LAWLESS, 2011).

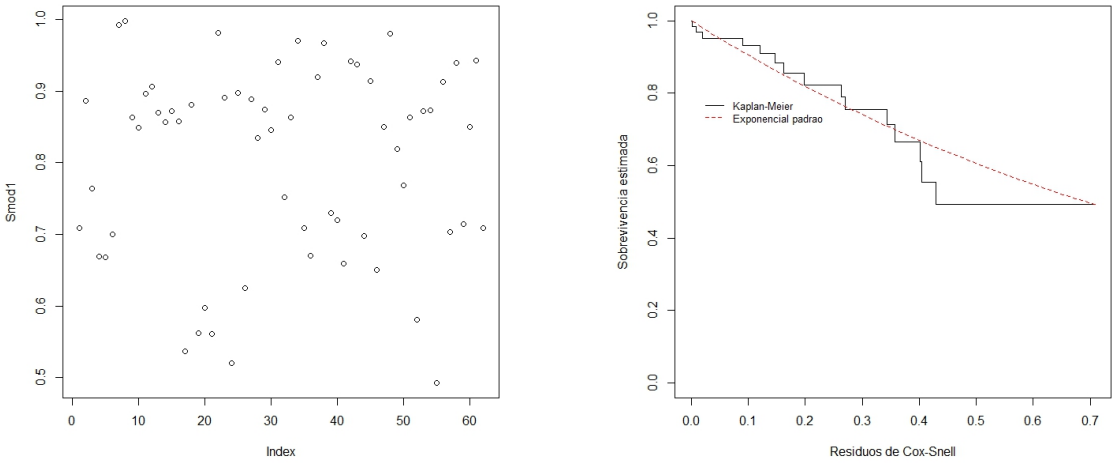


Figura 16: Residuo de Cox-Snell e curva de sobrevivência do resíduo estimada por kaplan-Meier ajustada por uma exponencial padrão.

Os resíduos do modelo para categoria de FEVE'<40' estão bem semelhantes para categoria complementar, levando assim às mesmas conclusões, que os pressupostos foram atendidos.

4.3 Modelo Weibull

Outro modelo candidato para o método paramétrico é o modelo cuja distribuição de probabilidade é a Weibull que analogamente aos outros modelos obtidos será usado para analisar o ajuste e posteriormente contribuirá para a quantidade total de modelos a serem comparados. Os parâmetros que maximizam a função de densidade da Weibull para o melhor ajuste estão representados a seguir:

Tabela 4: Tabela dos parâmetros que maximizam a distribuição referida.

α	β
1.248297	5579.625862

É possível representar graficamente o ajuste à curva de sobrevivência estimada por Kaplan-Meier utilizando os resultados da tabela acima e assim obtendo-se o seguinte resultado:

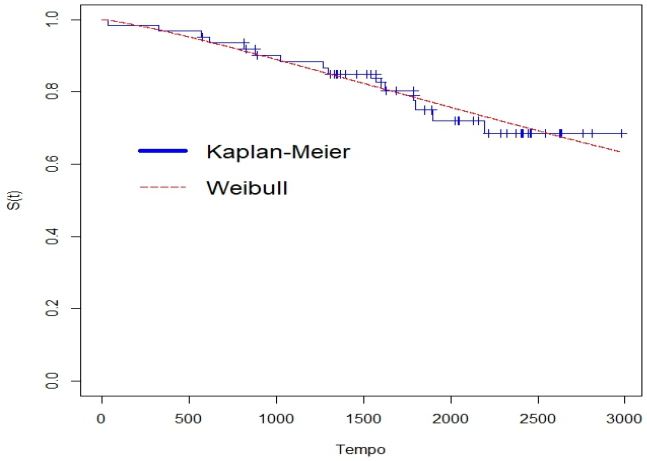


Figura 17: Ajuste do modelo com distribuição Log-Logística e a função de sobrevivência estimada por Kaplan-Meier

Na Figura 17 podemos ver o ajuste graficamente do modelo Weibull sob a curva de sobrevivência estimada por Kaplan-Meier, é possível relatar um ajuste no mínimo razoável, levando em consideração que a curva do modelo Weibull acompanha a curva de sobrevivência estimada por Kaplan-Meier ao longo do tempo e só em certos pontos as curvas possuem uma distancia um pouco mais acentuada.

4.3.1 Análise de diagnóstico

Para verificar as suposições do erro do modelo e adequabilidade a análise de diagnóstico é realizada similarmente aos outros modelos. Dito isso, obtemos os seguintes resultados:

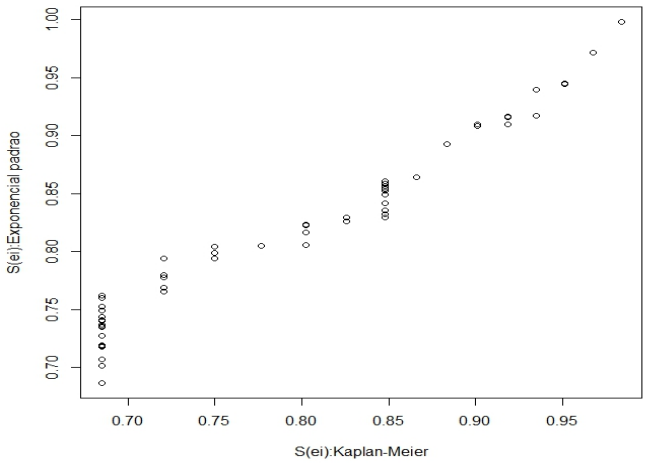


Figura 18: $S(e_i)$ por Kaplan-Meyer versus $S(e_i)$ da exponencial padrão.

Como o gráfico se assemelha a uma reta, consequentemente contatar que os resíduos do modelo assumem a distribuição exponencial padrão assim pode se concluir que o ajuste global do modelo é razoável.

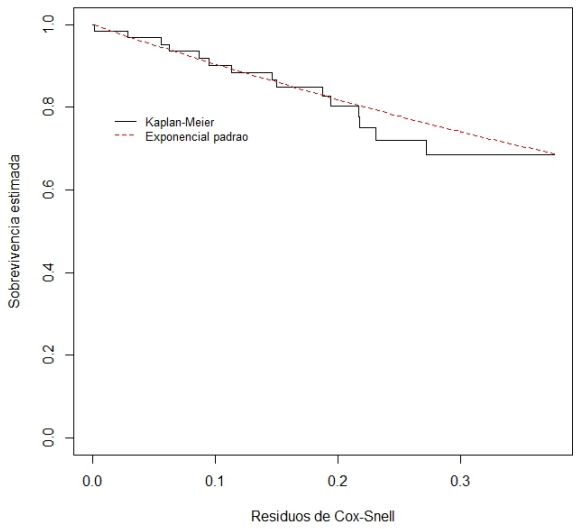


Figura 19: Curva de sobrevivência do resíduo estimada por kaplan-Meier ajustada por uma exponencial padrão.

A curva de sobrevivência dos resíduos estimada por Kaplan-Meier, ou seja, $exp\left\{-\hat{e}_i\right\}$ versus $\hat{S}_{KM}(\hat{e}_i)$, também auxilia na verificação de qualidade do modelo ajustado e como observado na Figura 19 pode se concluir que seguem aproximadamente uma distribuição exponencial padrão, formalizando que o ajuste do modelo é razoavelmente adequado.

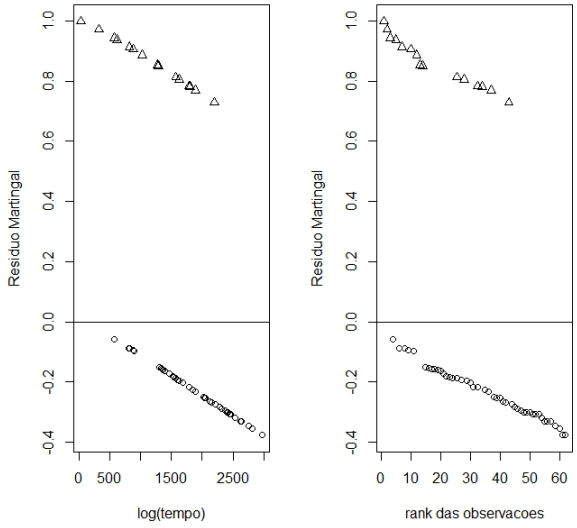


Figura 20: Gráfico do resíduo de martingal versus log(tempo) e rank das observações.

Semelhante aos resíduos de Martingal do modelo log-logístico, os resíduos da Weibull possuem evidências de que umas categorização seria uma candidata a forma funcional da variável também estima o numero de falha em excesso observada nos dados mas não predito pelo modelo. Novamente a presença de evidências visuais é notada para categorização da covariável do respectivo modelo.

4.3.2 Modelo Weibull para variável categorizada.

Os parâmetros da distribuição Weibull que proporcionam o ajuste para as categorias '>40' e '<40' estão dispostos na tabela a seguir:

Tabela 5: Tabela dos parâmetros para as categorias '>40' e '<40' respectivamente.

	α	γ
Cat >40	1.312276	3526.207
Cat <40	1.303491	11400.92

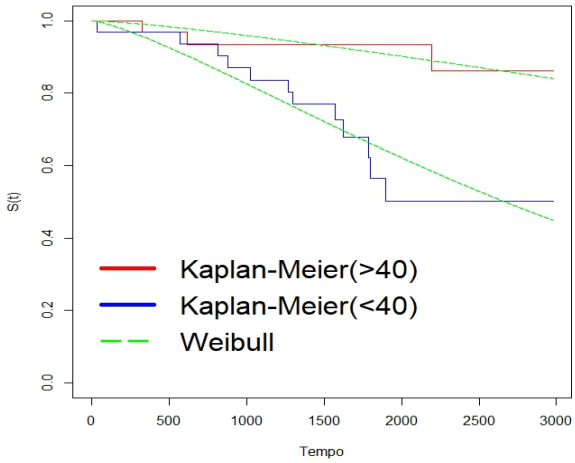


Figura 21: Ajuste para o modelo com as variáveis explicativas

O modelo com dist. Weibull proposto têm um ajuste notavelmente adequado e assim com o modelo Log-logístico ele superestima a função de sobrevivência em alguns pontos, não com tamanha intensidade quanto, e a conclusão que pode se tomar é similar a anterior, o modelo possui um bom ajuste. Com o auxilio da Figura 21 é possível avaliar visualmente a qualidade do ajuste.

Análise de diagnóstico

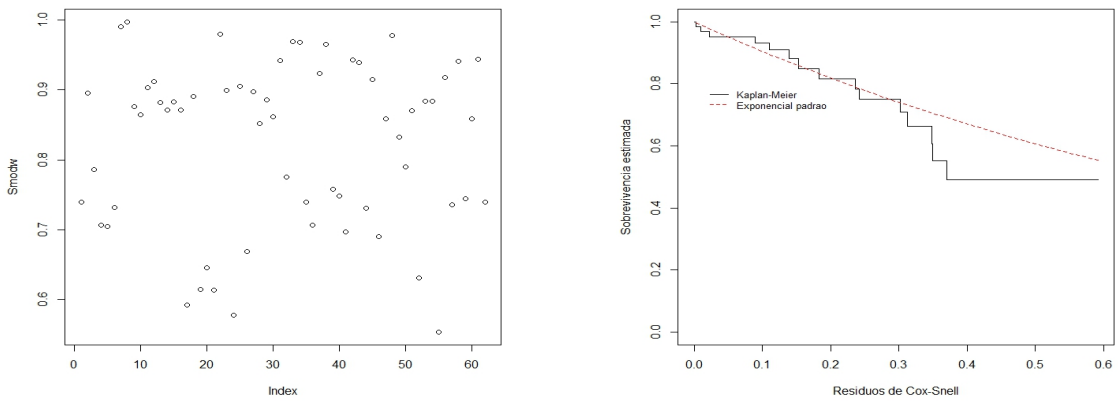


Figura 22: Residuo de Cox-Snell e curva de sobrevivência do resíduo estimada por kaplan-Meier ajustada por uma exponencial padrão.

Ao analisar a Figura 22 não é possível observar nenhum padrão bem definido, concluindo assim que a distribuição de pontos possui princípios de aleatoriedade e também que a curva de sobrevivência do resíduo estimada por kaplan-Meier possui um bom ajuste por uma exponencial padrão, ou seja, os dados provem de uma população homogenia segundo (LAWLESS, 2011).

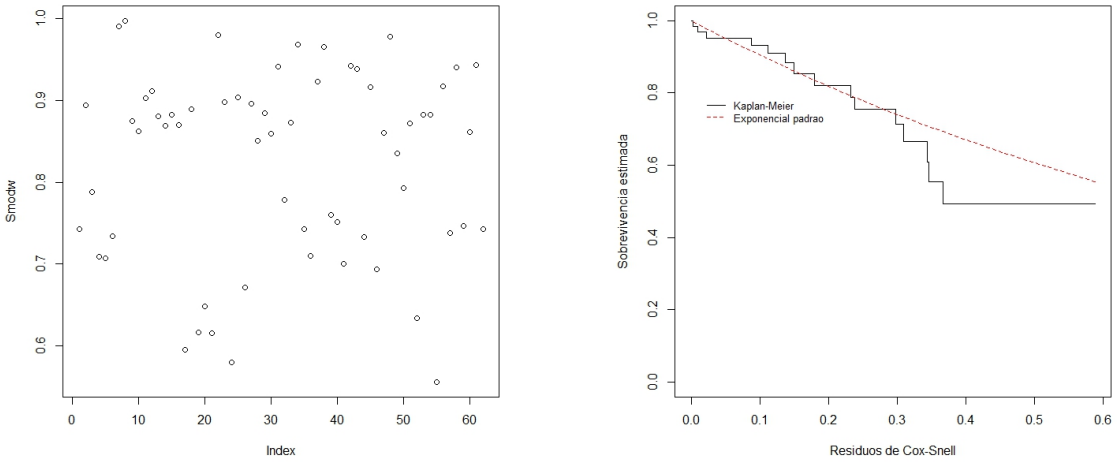


Figura 23: Residuo de Cox-Snell e curva de sobrevivência do resíduo estimada por kaplan-Meier ajustada por uma exponencial padrão.

Os resíduos do modelo para categoria de FEVE'<40' estão bem semelhantes para categoria complementar, levando assim às mesmas conclusões, que os pressupostos foram atendidos.

4.4 Modelo semi-paramétrico de Cox

Pela análise descritiva, podemos observar que a forma com que as categorias foram divididas parecem influenciar nos tempos de sobrevivência dos pacientes e também que há fortes indícios de riscos proporcionais. Dado isso, um método que pode se adequar bem aos dados é propor o modelo de Cox.

O modelo de Cox tem como pressuposto riscos proporcionais e para avaliar se os dados realmente poderiam ser ajustados pelo modelo de Cox, verificou-se a suposição de riscos proporcionais para cada forma de categorização. Para tal, foram ajustados modelos considerando cada covariável para explicar os tempos de sobrevivência. A Tabela abaixo mostra os resultados obtidos considerando que a hipótese a ser testada (H_0) é que os riscos são proporcionais e a alternativa (H_1) de que os riscos não são proporcionais.

Tabela 6: Inferência sobre o Risco Proporcional das Variáveis Categorização 1.

Variável	ρ	p-valor
39-30	0.2506	0.330
<30	0.0886	0.734

A categorização 1 diz respeito as categorias acima de 40 (>40), entre 39 e 30 (39-30), abaixo de 30(<30) e no respectivo modelo de Cox a categoria que é usada como base é FEVE acima de 40. Logo, pode se notar que, estabelecendo 5% como nível de significância, a categorização 1 atende ao pressuposto de risco proporcional, o que são evidências para a proporcionalidade dos riscos nesta forma de categorização, atendendo ao pressuposto do modelo de regressão de Cox para a situação referente.

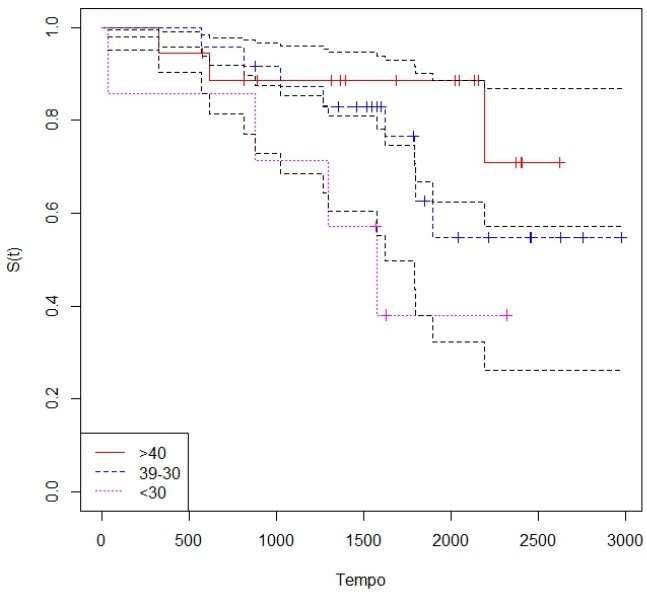


Figura 24: Ajuste do modelo de regressão de Cox e a função de sobrevivência estimada por Kaplan-Meier.

Acima é possível avaliar visualmente o ajuste do modelo de regressão de Cox às curvas

de sobrevivência de cada categoria estimada por Kaplan-Meier, que por sua vez permanecem relativamente próximas entre si ao longo do estudo exceto em alguns raros momentos.

Tabela 7: Coeficientes e seu exponencial do modelo.

Parâmetro	coef	exp(coef)
β_2(39-30)	1.3670	3.924
β_3(<30)	2.2387	9.382

O modelo regressão de Cox correspondente a categorização 1 é dado por:

$$h(t|x) = h_0(t)exp(1.3670 * X_2 + 2.2387 * X_3),$$

em que X_2 é a categoria '39-30' e X_3 é a categoria '<30'. A interpretação dos parâmetros é dada por exp(β_i) onde os β's são os coeficientes do parâmetro, sendo assim exp(β_2)=3.924 significa que a categoria '39-30' possui 3.924 vezes o risco de falha da categoria base, que no caso é a categoria acima de 40 e similarmente o risco de falha para a categoria abaixo de 30 é exp(β_3)=9.382 vezes o risco de falha da categoria base.

4.4.1 Análise de diagnóstico

Na perspectiva de análise de sobrevivência, os resíduos que são comumente utilizados são os de Cox-Snell, Martingal, Deviance e Schoenfeld. Desse modo os seguintes resultados foram obtidos:

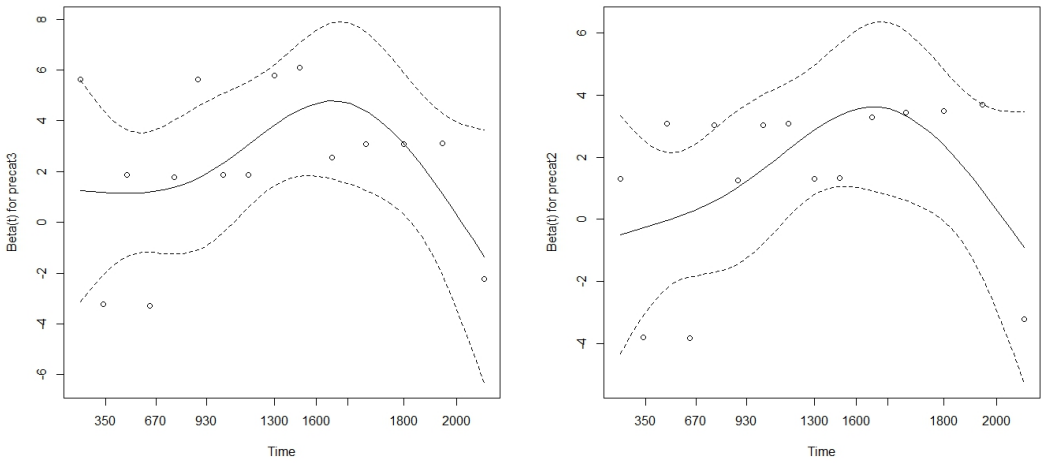


Figura 25: Resíduos de Schoenfeld para o modelo de regressão de Cox com categorização 1.

Com auxílio da Figura 25 é possível notar que o valor do parâmetro estimado em questão esta sempre incluído na banda de confiança e também é possível notar um comportamento particular das observações em questão devido as categorias agrupadas.

O gráfico de Martingal é usados para avaliar a forma funcional da covariável a ser usada no modelo e Deviance uma tentativa de tornar os resíduos martingal mais simétricos em torno de zero e com auxílio da Figura 26 é possível notar a evidência de que a categorização é uma boa escolha para a covariável em questão.

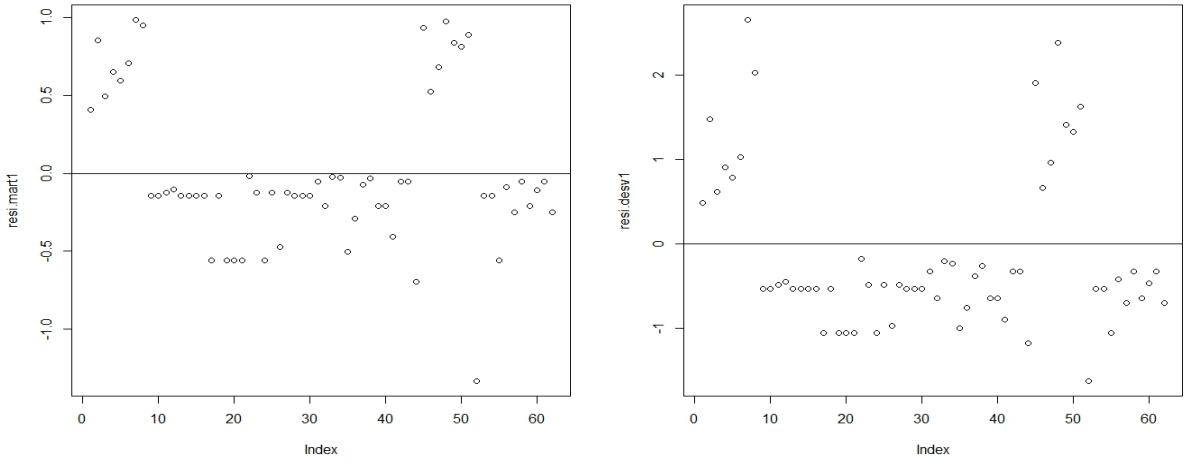


Figura 26: Resíduo de Martingal e Deviance respectivamente para o modelo de regressão de Cox com categorização 1.

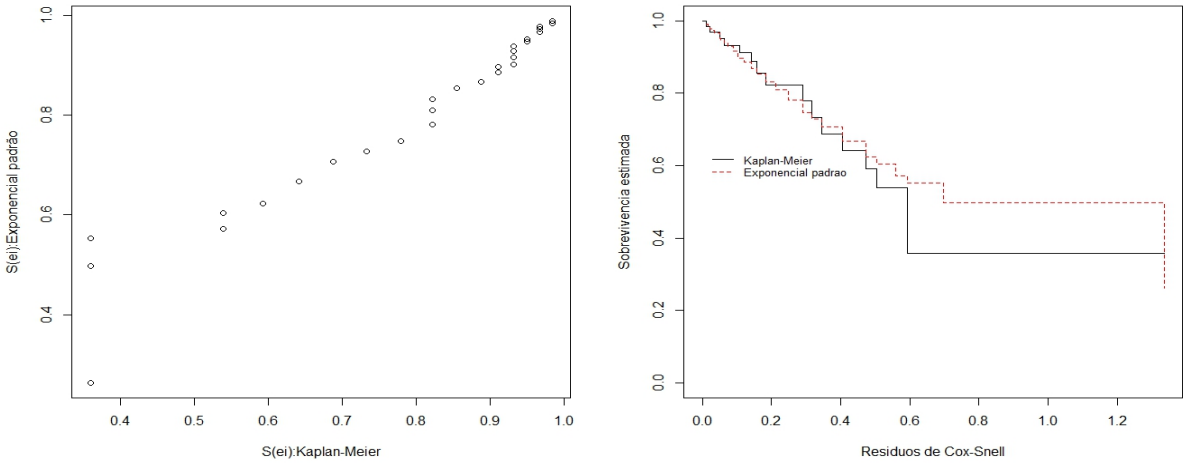


Figura 27: Curva de sobrevivência do resíduo estimada por kaplan-Meier ajustada por uma exponencial padrão e $S(e_i)$ por Kaplan-Meier versus $S(e_i)$ da exponencial padrão respectivamente.

A Figura 27 evidencia que o gráfico de sobrevivência dos resíduos estimada por Kaplan-Meier versus por uma exponencial padrão possui um padrão linear que seria mais nítido caso houvessem mais observações e o gráfico subsequente do bom ajuste da função de sobrevivência do resíduo estimada por Kaplan-Meier versus a função de sobrevivência do resíduo estimada da exponencial padrão, o que confirma as evidencias visuais de que os resíduos deste modelo estão adequados.

Tabela 8: Inferência sobre o Risco Proporcional das Variáveis Categorização 2.

Variável	ρ	p-valor
<40	0.194	0.452

Tabela 9: Inferência sobre o Risco Proporcional das Variáveis Categorização 2

A categorização 2 refere-se as categorias maior que 40(>40) e menor que 40(<40) e o suporte para a comparação deste modelo é a categoria acima de 40. Adotando novamente 5% como nível de significância, a categoria em questão atende ao pressuposto de risco proporcional.

Tabela 10: Coeficientes e seu exponencial do modelo.

Parâmetro	coef	exp(coef)
β_4 (<40)	1.5731	4.821

O modelo de regressão de Cox para essa categorização é dado por:

$$h(t|x) = h_0(t)exp(1.5731 * X_4)$$

Onde X_4 é a categoria '<40' e a interpretação é análoga ao modelo anterior onde a exp(coeficiente) corresponde as chances de falha em comparação a categoria base, neste caso o risco de falha da categoria abaixo de 40 exp(β_4)=4.821 vezes o risco de falha da categoria, neste caso, acima de 40.

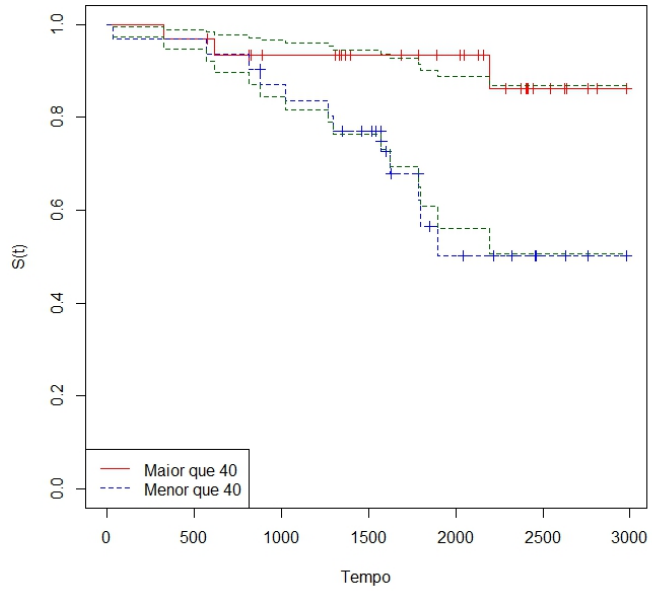


Figura 28: Ajuste do modelo de regressão de Cox e a função de sobrevivência estimada por Kaplan-Meier.

Com auxílio da Figura 28 é possível notar que as curvas do modelo de regressão de Cox e as curvas de sobrevivência de cada categoria estimada pelo método Kaplan-Meier, ao longo do estudo persistem moderadamente próximas e em momentos singulares possuem uma diferença um pouco mais expressiva, caracterizando um ajuste no mínimo razoável.

4.4.2 Análise de diagnóstico

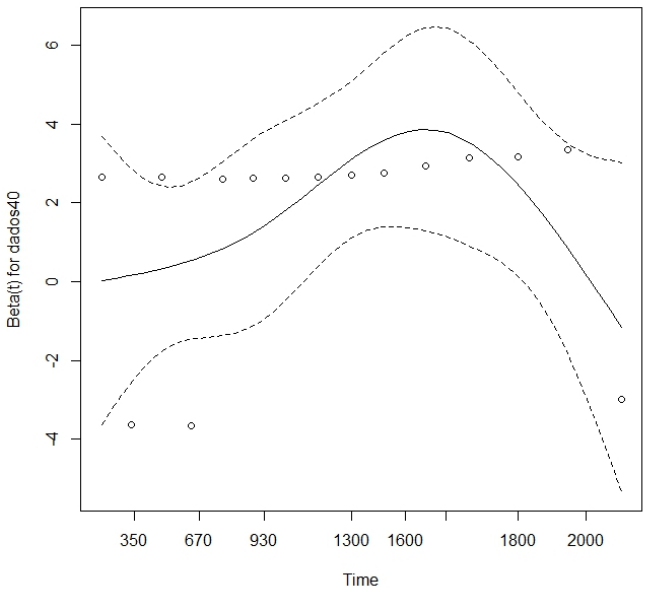


Figura 29: Resíduos de Schoenfeld para o modelo de regressão de Cox com categorização 2.

De acordo com a Figura 29 é possível notar a variação ao passar do tempo e que a estimativa do parâmetro em questão esta sempre entre as bandas de confiança, a distribuição de pontos possui um comportamento singular devido a forma funcional da variável mais nítida nos gráficos de Martigal e Deviance.

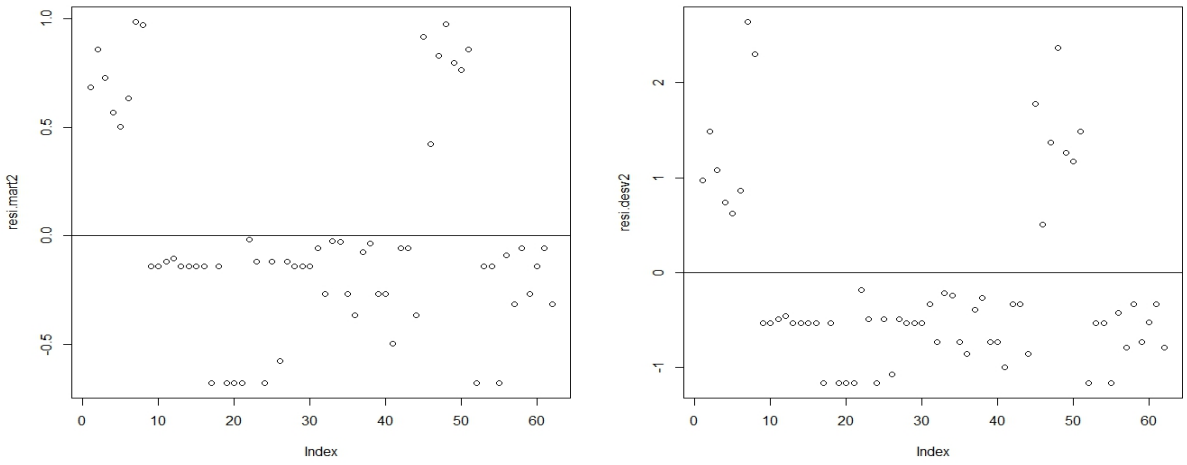


Figura 30: Resíduo de Martingal e Deviance respectivamente para o modelo de regressão de Cox com categorização2.

Com auxilio visual da Figura 30 é possível notar uma tendência a pontos agrupados o que significa que a forma funcional da variável é mais indicada a ser categorizada ainda

que os pontos possuem comportamento aleatório dentro de suas respectivas categorias.

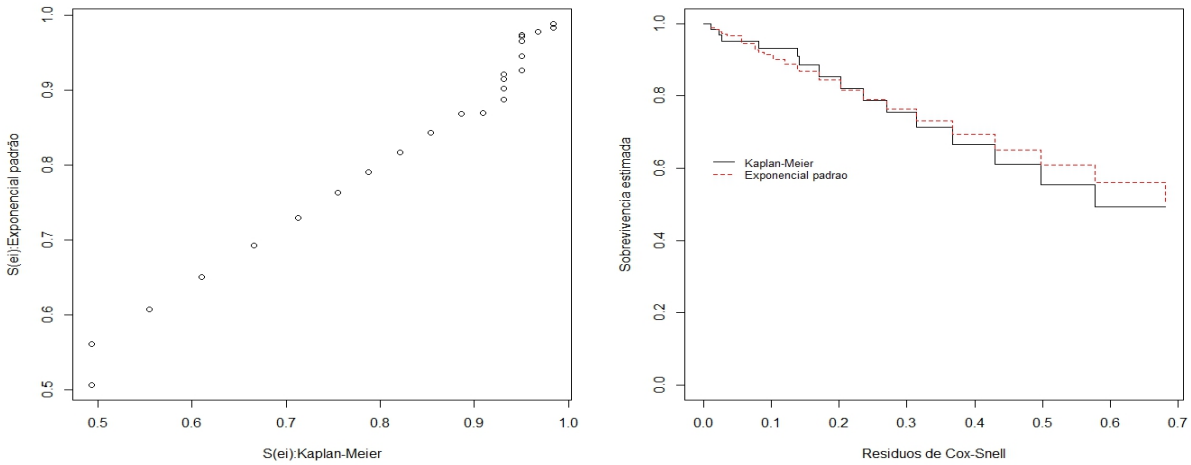


Figura 31: Curva de sobrevivência do resíduo estimada por kaplan-Meier ajustada por uma exponencial padrão e $S(e_i)$ por Kaplan-Meier versus $S(e_i)$ da exponencial padrão respectivamente.

A Figura 31 evidencia que o gráfico de sobrevivência dos resíduos estimada por Kaplan-Meier versus por uma exponencial padrão possui um padrão linear bem nítido e o gráfico subsequente salienta um bom ajuste da função de sobrevivência do resíduo estimada por Kaplan-Meier versus a função de sobrevivência do resíduo estimada da exponencial padrão, o que confirma as evidencias visuais de que os resíduos deste modelo estão adequados e atendem aos pressupostos.

Tabela 11: Inferência sobre o Risco Proporcional das Variáveis Categorização 3

Variável	ρ	p-valor	$\Pr(>\|z\|)$
<50	-0.865	1	0.998

Para a categorização 3 referente as categorias maior que 50($>$50) e menor que 50($<$50) não foram observados nenhum tempo de falha na categoria acima de 50, o que torna os resultados não fidedignos e o modelo de regressão de Cox que seria proposto para tal não terá continuação nesta monografia.

4.5 Comparação dos modelos

Os modelos propostos têm um ajuste notavelmente adequado porém é necessário escolher, dentre todos, um que melhor represente o estudo prezando um bom ajuste, parcimônia e adequação. Sendo assim, foram considerados os critérios de informação de Akaike e Bayesiano, como auxilio para eleição do modelo a ser utilizado, obtendo-se o seguintes resultados:

Tabela 12: Critérios AIC e BIC dos modelos candidatos.

Modelos	log-logistico(cat.2)	Weibull(cat.2)	COX(cat.1)	COX(cat.2)
AIC	295.3815	294.658	107.49	107.27
BIC	304.25	304.10	107.98	108.90

onde, cat.1 (categorização 1) diz respeito à divisão '> 40, 39 a30 e <30', cat.2 (categorização 2) diz respeito à '>40 ou <40' forma de divisão.

Considerando os critérios de informação de Akaike e Bayesiano para selecionar o modelo a ser usado, é notável que os com menores critérios são respectivamente os modelos de Cox, mesmo com uma ínfima diferença o que corresponde afirmar que não tem diferença qual modelo escolher, e dentre eles para manter a parcimônia o modelo final eleito será o modelo de Cox com a categorização 2 (>40 ou <40). Para melhor visualizar a comparação dos ajuste todas as curvas foram representadas em um mesmo gráfico.

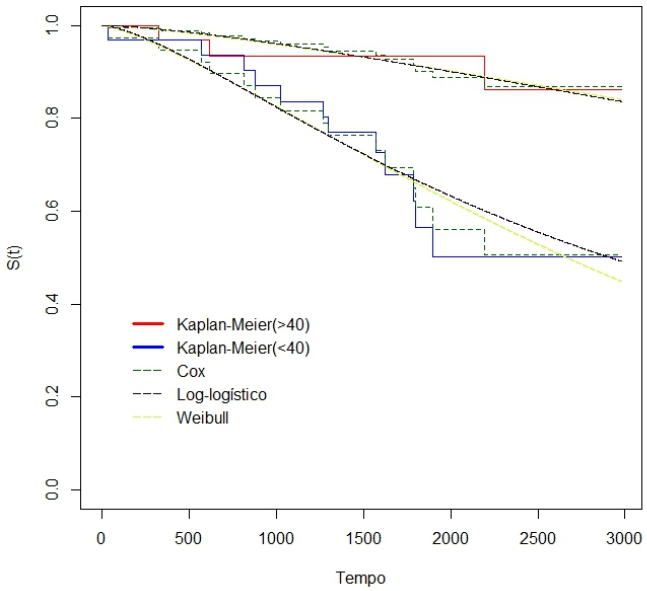


Figura 32: Ajuste das curvas de cada modelo respectivamente.

Com auxílio da Figura 32 é possível comparar visualmente o ajuste das curvas de sobrevivência estimadas por seus respectivos modelos, concluindo assim que em sua maioria a curva que permanece mais próxima por mais tempo é a curva estimada pelo modelo de Cox, que é mais evidente ao se aproximar do final do estudo onde fica mais nítida a diferença das curvas e da superestimação pelos modelos Log-logístico e Weibull porém no intermédio do estudo as curvas estão todas muito próximas não possuindo diferença proveniente do modelo em geral.

4.6 Conclusão

Todos os modelos encontrados possuem um bom ajuste aos dados, respeitando as suposições e o diagnóstico do modelo permite suas respectivas aplicações e analisando o AIC, as comparações mostram que os modelos de Cox possuem um critérios de informação de Akaike e Bayesiano significantemente diferente dos demais, Log-logístico e Weibull respectivamente, tornando-os mais indicados para a situação em questão por possuírem menores valores contudo o modelos Log-logistico e Weibull também podem ser usados sem muita discrepância nos resultados.

Outro ponto que foi considerado para escolha do modelo de Cox é sua fama na área médica por possuir interpretações familiares para profissionais de outras áreas relacionadas e por fornecer as estimativas das razões de risco que avalia o impacto que alguns fatores tem ate o evento de interesse que possibilita uma acessibilidade às interpretações dos resultados, em termos práticos e do modelo escolhido para representar o estudo (modelo semi-paramétrico de Cox com categorização 2) em questão, pacientes que possuem FEVE menor que 40 possuem 4.821 vezes o risco de falha dos que possuem FEVE maior que 40, consequentemente concluindo que a FEVE possui um impacto significativo no tempo de falha (óbito).

5 Referências

Referências

BOHORIS, G. Comparison of the cumulative-hazard and kaplan-meier estimators of the survivor function. *IEEE Transactions on Reliability*, IEEE, v. 43, n. 2, p. 230–232, 1994.

COLLETT, D. *Modelling survival data in medical research*. [S.l.]: Chapman and Hall/CRC, 2015.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de sobrevivência aplicada*. [S.l.]: E. Blucher,, 2014. 18

COX, D. R. *Analysis of survival data*. [S.l.]: Routledge, 2018. 10

KALBFLEISCH, J. D.; PRENTICE, R. L. *The statistical analysis of failure time data*. [S.l.]: John Wiley & Sons, 2011. v. 360.

KLEIN, J. P.; MOESCHBERGER, M. L. *Survival analysis: techniques for censored and truncated data*. [S.l.]: Springer Science & Business Media, 2006. 11

LAWLESS, J. F. *Statistical models and methods for lifetime data*. [S.l.]: John Wiley & Sons, 2011. v. 362. 11, 30, 34

MESQUITA, E. T.; JORGE, A. J. L. Insuficiência cardíaca com fração de ejeção normal-novos critérios diagnósticos e avanços fisiopatológicos. *Arq Bras Cardiol*, SciELO Brasil, v. 93, n. 2, p. 180–7, 2009. 10

PAULA, G. A. *Modelos de regressão: com apoio computacional*. [S.l.]: IME-USP São Paulo, 2004. 17

5.1 Códigos

```
  library(readxl)
dados<- read_excel("d:/Users/User/Desktop/Grupo4_1.xlsx",
                   col_types = c("numeric", "text", "numeric",
                                 "numeric", "numeric", "numeric",
                                 "numeric", "numeric", "numeric",
                                 "numeric", "numeric", "date", "date"))


head(dados)
require(date)
attach(dados)
library(survival)
library(survminer)
library(AdequacyModel)
x11()

dados[43,13] <- mdy.date(2,23,2016)
dados <- dados[-35,]
dados <- dados[,-11]
tempo <-  as.numeric(difftime(dados$‘Data final seguimento‘,dados$‘Data início seguimento‘,
units = "days"))
status <- dados$‘Óbito ou Tx N=0;  S=1‘
dados <- cbind(dados,tempo,status)

dados
################ Análise descritiva dos dados ##################
summary(dados$tempo)


KM <-  survfit(Surv(tempo,status)~1, conf.int=F)
summary(KM)
plot(KM,conf.int=T, xlab="Tempo", ylab="S(t)",mark.time=T,main="",col = "blue")

table(status)
plot(KM, conf.int=F, fun="cumhaz",mark.time=T, xlab="Tempo", ylab="H(t)",col = "blue")

hist(tempo, freq=F,ylab = 'Densidade',main ='' )
plot(density(tempo),main="Desidade dos tempos")

TTT(tempo, col="red", lwd=2, grid=TRUE, lty=2)


################ Análise descritiva das covariáveis ##################


#Maior ou menor que 50
KMs<-survfit(Surv(tempo,status)~dados$‘FEVE > ou = 50 N=0;  S=1‘, conf.int=F)
summary(KMs)
plot(KMs, conf.int=F, xlab="Tempo", ylab="S(t)", lty=c(1,2),col=c(2,4),mark.time = T)
legend("bottomleft",c('Maior que 50', 'Menor que 50'), lty=c(2,1),col=c(4,2))
survdiff(Surv(tempo, status)~dados$‘FEVE > ou = 50 N=0;  S=1‘, rho=0)

boxplot(tempo~dados$‘FEVE > ou = 50 N=0;  S=1‘)
barplot(table(dados$‘FEVE > ou = 50 N=0;  S=1‘), ylim=c(0,90),col=c(0,1))

#Entre 40-49 ou 30-39 ou >30
```

```
dados$precat <- NA
dados$precat[dados$‘FEVE 40-49 N=0;  S=1‘==1] <- 1
dados$precat[dados$‘FEVE 30-39 N=0;  S=1‘==1] <- 2
dados$precat[dados$‘FEVE < 30 N=0;  S=1‘==1] <- 3
dados$precat[is.na(dados$precat)] <- 1
dados$precat <- as.factor(dados$precat)

KMs1<-survfit(Surv(tempo,status)~dados$precat, conf.int=F)
summary(KMs1)
plot(KMs1, conf.int=F, xlab="Tempo", ylab="S(t)", lty=c(1,2,3),col=c(2,4,6),mark.time = T)
legend("bottomleft",c('>40', '39-30','<30'), lty=c(1,2,3),col=c(2,4,6))

precat <- dados$precat
pairwise_survdiff(Surv(tempo,status)~precat,data=dados,rho=0,p.adjust.method='bonferroni')

x11()
boxplot(tempo~dados$precat)
barplot(table(dados$precat), ylim=c(0,90),col=c(1,0,4))

#<40 ou >40
KMs3<-survfit(Surv(tempo,status)~dados$‘FEVE < 40 N=0;  S=1‘, conf.int=F)
summary(KMs3)
plot(KMs3, conf.int=F, xlab="Tempo", ylab="S(t)", lty=c(1,2),col=c(2,4),mark.time = T)
legend("bottomleft",c('Maior que 40', 'Menor que 40'), lty=c(1,2),col=c(2,4))
survdiff(Surv(tempo, status)~dados$‘FEVE < 40 N=0;  S=1‘, rho=0)

boxplot(tempo~dados$‘FEVE < 40 N=0;  S=1‘)
barplot(table(dados$‘FEVE < 40 N=0;  S=1‘), ylim=c(0,90),col=c(2,4))
######################COX MODEL############################

#Modelo de cox
require(survival)
#50<
dados50 <- dados$‘FEVE > ou = 50 N=0;  S=1‘
mod50 <- coxph(Surv(tempo,status)~dados50, x=TRUE)
cox.zph(mod50, transform="identity", global=TRUE)
plot(cox.zph(mod50))
summary(mod50)

resi.mart <- resid(mod50, type = 'martingal')
plot(resi.mart)
abline(h=0)

resi.desv <- resid(mod50, type = 'deviance')
plot(resi.desv)
abline(h=0)


#49-40
modcat <- coxph(Surv(tempo,status)~precat, x=TRUE)
cox.zph(modcat, transform="identity", global=TRUE)
plot(cox.zph(modcat)[1])
summary(modcat)


#grafico de S(t) e H(t) estimado pelo modelo fit1#
fb2<-basehaz(modcat, centered = FALSE)
```

```
temp2<-fb2$time
h02<-fb2$hazard
S02<-exp(-h02)


beta11<-modcat$coef[1]
beta12<-modcat$coef[2]

Sg111<-S02
Sg211<-S02^exp(beta11)
Sg311<-S02^exp(beta12)

plot(KMs1, conf.int=F, xlab="Tempo", ylab="S(t)", lty=c(1,2,3),col=c(2,4,6),mark.time = T)
lines(temp2,Sg111, ylim=c(0,1), type="s",lty=2)
lines(temp2,Sg211,type="s",lty=2)
lines(temp2,Sg311,type="s",lty=2)
legend("bottomleft",c('>40', '39-30','<30'), lty=c(1,2,3),col=c(2,4,6))

#funcao risco acumulada#

Hg111<--log(Sg111)
Hg211<--log(Sg211)
Hg311<--log(Sg311)
plot(KMs1, conf.int=F, fun="cumhaz", xlab="Tempo", ylab="H(t)",col=c(2,4,6))
lines(temp2,Hg111, ylim=c(0,1), type="s",lty=2,xlab="Tempo", ylab="H(t|x)estimada")
lines(temp2,Hg211,type="s",lty=2)
lines(temp2,Hg311,type="s",lty=2)
legend("topleft",c('>40', '39-30','<30'), lty=c(1,1,1),col=c(2,4,6))

#Analise de diagnostico Mod3##
resi.mart1 <- resid(modcat, type = 'martingal')
plot(resi.mart1)
abline(h=0)

resi.desv1 <- resid(modcat, type = 'deviance')
plot(resi.desv1)
abline(h=0)

cox3 <- status-resi.mart1
b <- abs(cox3)
KMe3<-survfit(Surv(b,status)~1,conf.int=F)
te3<-KMe3$time
ste3<-KMe3$surv
sexpw3<-exp(-te3)
plot(ste3,sexpw3, xlab="S(ei):Kaplan-Meier", ylab="S(ei):Exponencial padrão")
plot(KMe3,conf.int=F, xlab="Residuos de Cox-Snell", ylab="Sobrevivencia estimada")
lines(te3,sexpw3,lty=2, col=2, type="s")
legend(0.0,0.65,lty=c(1,2), col=c(1,2),c("Kaplan-Meier", "Exponencial padrao"), cex=0.8, bty="n")



#<40 ou >40
dados40 <- dados$`FEVE < 40 N=0;  S=1`
mod40 <- coxph(Surv(tempo,status)~dados40, x=TRUE)
cox.zph(mod40, transform="identity", global=TRUE)
plot(cox.zph(mod40))#Banda de confiança sempre incluem o valor do parâmetro estimado,
comportamento aleatorio#
summary(mod40)
```

```
# ajuste modelo de cox#

fb1<-basehaz(mod40, centered = FALSE)
temp1<-fb1$time
h01<-fb1$hazard
S01<-exp(-h01)


beta1<-mod40$coef[1]

Sg11<-S01
Sg21<-S01^exp(beta1)
X11()
plot(KMs3, conf.int=F, xlab="Tempo", ylab="S(t)", lty=c(1,2),col=c(2,4),mark.time = T)
lines(temp1,Sg11, ylim=c(0,1), type="s",lty=2,lwd=1,col='darkgreen')
lines(temp1,Sg21,type="s",lty=2,col='darkgreen')
legend("bottomleft",c('Maior que 40', 'Menor que 40'), lty=c(1,2),col=c(2,4))

#funcao risco acumulada 40#

Hg11<--log(Sg11)
Hg21<--log(Sg21)
plot(KMs3, conf.int=F, fun="cumhaz", xlab="Tempo", ylab="H(t)",col=c(2,4))
lines(temp1,Hg11, ylim=c(0,1), type="s",lty=2,xlab="Tempo", ylab="H(t|x)estimada")
lines(temp1,Hg21,type="s",lty=2)
legend("topleft",c('Maior que 40', 'Menor que 40'), lty=c(1,1),col=c(2,4))

#Analise de diagnostico Mod40##

resi.mart2 <- resid(mod40, type = 'martingal')
plot(resi.mart2)
abline(h=0)

resi.desv2 <- resid(mod40, type = 'deviance')
plot(resi.desv2)
abline(h=0)

cox40 <- status-resi.mart2
a <- abs(cox40)
KMe40<-survfit(Surv(a,status)~1,conf.int=F)
te40<-KMe40$time
ste40<-KMe40$surv
sexpw40<-exp(-te40)
plot(ste40,sexpw40, xlab="S(ei):Kaplan-Meier", ylab="S(ei):Exponencial padrão")
plot(KMe40,conf.int=F, xlab="Residuos de Cox-Snell", ylab="Sobrevivencia estimada")
lines(te40,sexpw40,lty=2, col=2, type="s")
legend(0.0,0.65,lty=c(1,2), col=c(1,2),c("Kaplan-Meier", "Exponencial padrao"), cex=0.8, bty="n")



################## LOG LOGISTICA ################

dens<-function(t,mu,beta){                              # sub-funções que realizarão o cálculo
  (beta*(t/mu)^(beta-1)) / ( mu* (1+(t/mu)^beta)^2 ) } # da função densidade

sobrev<-function(t,mu,beta){                            # sub-funções que realizarão o cálculo
```

```
  (1+(t/mu)^beta)^-1}

like.loglogist<-function(parametro,tempo,delta){
  L1<-log(dens(tempo,parametro[1],parametro[2]))
  L2<-log(sobrev(tempo,parametro[1],parametro[2]))
  -sum(L1*delta + L2*(1-delta) ) }

chute.inicial<-c(1,1)
emv<-nlm(like.loglogist,chute.inicial,hessian=TRUE,tempo=tempo,delta=status)
emv
solve(emv$hessian)
#####

alphas <- seq(4000,5000,10)
gamas <- seq(1,2,.01)
z <- matrix(0,nrow = length(alphas),ncol=length(gamas))
for (i in 1:length(alphas)) {
  for (j in 1:length(gamas)) {
    z[i,j] <--1*like.loglogist(parametro = c(alphas[i],gamas[j]),tempo=tempo,delta = status)
  }
}
persp(alphas,gamas,z,zlim = c(min(z)-1,max(z)+1),theta =310, phi = 18,
      expand = .6,ltheta = 10, shade = 0.9, ticktype = "detailed")
points(p%*%p,col="red",pch=16)
#####

mll <- survreg(Surv(tempo,status)~1, dist="loglogistic")
summary(mll)
alphall<-exp(mll$coefficients[1])
gamall<- 1/mll$scale

mll1 <- survreg(Surv(tempo,status)~dados$precat, dist="loglogistic")
summary(mll1)
alphall1<-exp(mll1$coefficients[1])
gamall1<- 1/mll1$scale



plot(KM,conf.int=T, xlab="Tempo", ylab="S(t)",mark.time=T,main="",col = "blue")
plot(KMs3, conf.int=F, xlab="Tempo", ylab="S(t)", lty=c(1,1),col=c(2,4))
m<-max(tempo)*100
t<-(1:m)/100

estimativa<-emv$estimate
s.loglogistica<-sobrev(t,estimativa[1],estimativa[2])
s.loglogistica<-sobrev(t,alphall1,gamall1)


x11()
plot(KMs3, conf.int=F, xlab="Tempo", ylab="S(t)", lty=c(1,1),col=c(2,4),mark.time = T)

points(t,s.loglogistica,type="l",lty=5,col="green")
points(t,s.loglogistica1,type="l",lty=5,col="green")


legend(-100,0.4,lty=1,lwd=5,c("Kaplan-Meier(>40)"), bty = 'n', cex=2,col=2)
legend(-100,0.3,lty=1,lwd=5,c("Kaplan-Meier(<40)"), bty = 'n', cex=2,col=4)
legend(-100,0.2,lty=5,lwd=,col="green",c("Log-Logisitca"),bty="n", cex = 2)
```

```
x11()
plot(KM,conf.int=T, xlab="Tempo", ylab="S(t)",mark.time=T,main="",col = "blue")
legend(100,0.7,lty=1,lwd=5,c("Kaplan-Meier"), bty = 'n', cex=1.5,col=4)
legend(100,0.6,lty=5,lwd=,col="red",c("Log-Logisitca"),bty="n", cex = 1.5)
points(t,s.loglogistica,type="l",lty=5,col="red")

##############LOG varial cat ######################################################

mll2 <- survreg(Surv(tempo,status)~dados$`FEVE < 40 N=0;  S=1`, dist="loglogistic")
summary(mll2)
alphall2<-exp(mll2$coefficients[1])
gamall2<- 1/mll2$scale

mll3 <- survreg(Surv(tempo,status)~dados$`FE VE > 40 N=0;  S=1`, dist="loglogistic")
summary(mll3)
alphall3<-exp(mll3$coefficients[1])
gamall3<- 1/mll3$scale

s.loglogistica<-sobrev(t,alphall2,gamall2)
s.loglogistica1<-sobrev(t,alphall3,gamall3)

plot(KMs3, conf.int=F, xlab="Tempo", ylab="S(t)", lty=c(1,1),col=c(2,4),mark.time = T)
points(t,s.loglogistica,type="l",lty=5,col="green")
points(t,s.loglogistica1,type="l",lty=5,col="green")


###1
#Cox-Snell#
y=tempo
alphall2 <- exp(mll2$linear.predictors) # mi=mip
gamall2 <- 1/mll2$scale
Smod<-(1+(y/alphall2)^gamall2)^-1
plot(Smod)

ei<-(-log(Smod)) # residuo de Cox-Snell

KMew<-survfit(Surv(ei,status)~1,conf.int=F)
te<-KMew$time  # res?duo de Cox-Snell
ste<-KMew$surv
sexp<-exp(-te)

plot(ste,sexp, xlab="S(ei):Kaplan-Meier", ylab="S(ei):Exponencial padrao")
##deve ser aproximadamente reto##

plot(KMew,conf.int=F, xlab="Residuos de Cox-Snell", ylab="Sobrevivencia estimada")
lines(te,sexp,lty=2, col=2)
legend(0.01,0.8,lty=c(1,2), col=c(1,2),c("Kaplan-Meier", "Exponencial padrao"), cex=0.8, bty="n")


###2
#Cox-Snell#
y=tempo
alphall3 <- exp(mll3$linear.predictors) # mi=mip
gamall3 <- 1/mll3$scale
Smod1<-(1+(y/alphall3)^gamall3)^-1
plot(Smod1)
```

```
ei1<-(-log(Smod1)) # residuo de Cox-Snell

KMew1<-survfit(Surv(ei1,status)~1,conf.int=F)
te1<-KMew1$time  # res?duo de Cox-Snell
ste1<-KMew1$surv
sexp1<-exp(-te1)

plot(ste1,sexp1, xlab="S(ei):Kaplan-Meier", ylab="S(ei):Exponencial padrao")
##deve ser aproximadamente reto##

plot(KMew1,conf.int=F, xlab="Residuos de Cox-Snell", ylab="Sobrevivencia estimada")
lines(te1,sexp1,lty=2, col=2)
legend(0.01,0.8,lty=c(1,2), col=c(1,2),c("Kaplan-Meier", "Exponencial padrao"), cex=0.8, bty="n")



###################### Weibull  ###################
like.weibull<-function(parametro,tempo,delta){
  L1<-dweibull(tempo,parametro[1],parametro[2],log=TRUE)
  L2<-pweibull(tempo,parametro[1],parametro[2],log.p=TRUE,lower.tail=FALSE)
  -sum(L1*delta + L2*(1-delta) ) }

chute.inicial<-c(1,1)
emvw<-nlm(like.weibull,chute.inicial,tempo=tempo,delta=status)
emvw
solve(emv$hessian)

mwe<-survreg(Surv(tempo,status)~1, dist="weibull")
mwe
summary(mwe)

alphaw<-1/mwe$scale
gamaw<-exp(mwe$coefficients[1])

####### Ajuste ###############
x11()
plot(KM,conf.int=T, xlab="Tempo", ylab="S(t)",mark.time=T,main="",col = "blue")
m<-max(tempo)*100
t<-(1:m)/100

estimativaw<-emvw$estimate
s.weibull<-pweibull(t,estimativaw[1],estimativaw[2],lower.tail=F)
points(t,s.weibull,type="l",lty=5,col='red')
legend(100,0.7,lty=1,lwd=5,c("Kaplan-Meier"), bty = 'n', cex=1.5,col=4)
legend(100,0.6,lty=5,lwd=,col="red",c("Weibull"),bty="n", cex = 1.5)


legend(110,0.9,lty=1,c("Kaplan-Meier"), bty = 'n', cex=0.8)
legend(110,0.8,lty=5,col="green",c("Log-Logisitca"),bty="n", cex = 0.8)
legend(110,0.85,lty=5,col="blue",c("Weibull"),bty="n", cex = 0.8)

# ajuste pra categopria 2#
mw2 <- survreg(Surv(tempo,status)~dados$`FEVE < 40 N=0;  S=1`, dist="weibull")
summary(mw2)
gamaw2<-exp(mw2$coefficients[1])
alphaw2<- 1/mw2$scale

mw3 <- survreg(Surv(tempo,status)~dados$`FE VE > 40 N=0;  S=1`, dist="weibull")
```

```
summary(mw3)
gamaw3<-exp(mw3$coefficients[1])
alphaw3<- 1/mw3$scale

x11()
plot(KMs3, conf.int=F, xlab="Tempo", ylab="S(t)", lty=c(1,1),col=c(2,4))
m<-max(tempo)*100
t<-(1:m)/100
s.weibuw<-pweibull(t,alphaw2,gamaw2,lower.tail=F)
s.weibuw2<-pweibull(t,alphaw3,gamaw3,lower.tail=F)
points(t,s.weibuw,type="l",lty=5,col="green")
points(t,s.weibuw2,type="l",lty=5,col="green")
legend(-100,0.4,lwd=5,lty=1,c("Kaplan-Meier(>40)"), bty = 'n', cex=2,col=2)
legend(-100,0.3,lwd=5,lty=1,c("Kaplan-Meier(<40)"), bty = 'n', cex=2,col=4)
legend(-100,0.2,lwd=3,lty=5,col="green",c("Weibull"),bty = 'n', cex =2)

########################Analise de Residuo Log-L ####################

#Cox-Snell#
y=tempo
alphalll <- exp(mll$linear.predictors) # mi=mip
gamalll <- 1/mll$scale
Smod<-(1+(y/alphalll)^gamalll)^-1
plot(Smod)

ei<-(-log(Smod)) # residuo de Cox-Snell

KMew<-survfit(Surv(ei,status)~1,conf.int=F)
te<-KMew$time  # res?duo de Cox-Snell
ste<-KMew$surv
sexp<-exp(-te)

plot(ste,sexp, xlab="S(ei):Kaplan-Meier", ylab="S(ei):Exponencial padrao")
##deve ser aproximadamente reto##

plot(KMew,conf.int=F, xlab="Residuos de Cox-Snell", ylab="Sobrevivencia estimada")
lines(te,sexp,lty=2, col=2)
legend(0.01,0.8,lty=c(1,2), col=c(1,2),c("Kaplan-Meier", "Exponencial padrao"), cex=0.8, bty="n")
#Dado que Hb(ebi) = ???log(S(ebi)), o grafico das curvas de sobreviv^encia desses residuos,
obtidas por Kaplan-Meier e pelo#
#modelo exponencial padr~ao, tambem auxiliam na verifica¸c~aoda qualidade do modelo ajustado.
Ou seja, exp{???dei} versusSdKM(ebi).#


#Residuo martingal
x11()
martingal<-status-ei

par(mfrow=c(1,2))
plot(y,martingal,xlab="log(tempo)", ylab="Residuo Martingal",pch=status+1)
plot(rank(y),martingal,xlab="rank das observacoes", ylab="Residuo Martingal",pch=status+1)
par(mfrow=c(1,1))
plot(martingal)
abline(h=0)

#Residuo Deviance

devw<-(martingal/abs(martingal))*(-2*(martingal+status*log(status-martingal)))^(1/2)
```

```
plot(y,devw,xlab="log(tempo)", ylab="Residuo Deviance",pch=status+1)
plot(rank(y),devw,xlab="rank das observacoes", ylab="Residuo Deviance",pch=status+1)
abline(h=0)

#######################Analise de Residuo Weibull ####################
x11()
#Cox-Snell#
y=tempo
alphaww <- exp(mwe$linear.predictors) # mi=mip
gamaww <- 1/mwe$scale
Smodw<-(1+(y/alphaww)^gamaww)^-1
eiw<-(-log(Smodw)) # residuo de Cox-Snell

KMeww<-survfit(Surv(eiw,status)~1,conf.int=F)
tew<-KMeww$time  # res?duo de Cox-Snell
stew<-KMeww$surv
sexpw<-exp(-tew)
plot(Smodw)

plot(stew,sexpw, xlab="S(ei):Kaplan-Meier", ylab="S(ei):Exponencial padrao")
##deve ser aproximadamente reto##

plot(KMeww,conf.int=F, xlab="Residuos de Cox-Snell", ylab="Sobrevivencia estimada")
lines(tew,sexpw,lty=2, col=2)
legend(0.01,0.8,lty=c(1,2), col=c(1,2),c("Kaplan-Meier", "Exponencial padrao"), cex=0.8, bty="n")
#Dado que Hb(ebi) = ???log(S(ebi)), o grafico das curvas de sobreviv^encia desses residuos,
obtidas por Kaplan-Meier e pelo#
#modelo exponencial padr~ao, tambem auxiliam na verifica¸c~aoda qualidade do modelo ajustado.
Ou seja, exp{???dei} versusSdKM(ebi).#


#Residuo martingal
x11()
martingalw<-status-eiw

par(mfrow=c(1,2))
plot(y,martingalw,xlab="log(tempo)", ylab="Residuo Martingal",pch=status+1)
plot(rank(y),martingalw,xlab="rank das observacoes", ylab="Residuo Martingal",pch=status+1)
par(mfrow=c(1,1))
plot(martingalw)
abline(h=0)

#Residuo Deviance

devww<-(martingalw/abs(martingalw))*(-2*(martingalw+status*log(status-martingalw)))^(1/2)
plot(y,devww,xlab="log(tempo)", ylab="Residuo Deviance",pch=status+1)
plot(rank(y),devww,xlab="rank das observacoes", ylab="Residuo Deviance",pch=status+1)
abline(h=0)

############ WEIBULL p var categorizada
mw2 <- survreg(Surv(tempo,status)~dados$‘FEVE < 40 N=0;  S=1‘, dist="weibull")
summary(mw2)
gamaw2<-exp(mw2$coefficients[1])
alphaw2<- 1/mw2$scale

mw3 <- survreg(Surv(tempo,status)~dados$‘FE VE > 40 N=0;  S=1‘, dist="weibull")
summary(mw3)
gamaw3<-exp(mw3$coefficients[1])
```

```
alphaw3<- 1/mw3$scale

y=tempo
alphaww <- exp(mw2$linear.predictors) # mi=mip
gamaww <- 1/mw2$scale
Smodw<-(1+(y/alphaww)^gamaww)^-1
eiw<-(-log(Smodw)) # residuo de Cox-Snell

KMeww<-survfit(Surv(eiw,status)~1,conf.int=F)
tew<-KMeww$time  # res?duo de Cox-Snell
stew<-KMeww$surv
sexpw<-exp(-tew)
plot(Smodw)

plot(stew,sexpw, xlab="S(ei):Kaplan-Meier", ylab="S(ei):Exponencial padrao")
##deve ser aproximadamente reto##

plot(KMeww,conf.int=F, xlab="Residuos de Cox-Snell", ylab="Sobrevivencia estimada")
lines(tew,sexpw,lty=2, col=2)
legend(0.01,0.8,lty=c(1,2), col=c(1,2),c("Kaplan-Meier", "Exponencial padrao"), cex=0.8, bty="n")

y=tempo
alphaww <- exp(mw3$linear.predictors) # mi=mip
gamaww <- 1/mw3$scale
Smodw<-(1+(y/alphaww)^gamaww)^-1
eiw<-(-log(Smodw)) # residuo de Cox-Snell

KMeww<-survfit(Surv(eiw,status)~1,conf.int=F)
tew<-KMeww$time  # res?duo de Cox-Snell
stew<-KMeww$surv
sexpw<-exp(-tew)
plot(Smodw)

plot(stew,sexpw, xlab="S(ei):Kaplan-Meier", ylab="S(ei):Exponencial padrao")
##deve ser aproximadamente reto##

plot(KMeww,conf.int=F, xlab="Residuos de Cox-Snell", ylab="Sobrevivencia estimada")
lines(tew,sexpw,lty=2, col=2)
legend(0.01,0.8,lty=c(1,2), col=c(1,2),c("Kaplan-Meier", "Exponencial padrao"), cex=0.8, bty="n")


######## Comparação####
AIC(mwe)
pws<-2
n <- length(tempo)
AICws<-(-2*mw2$loglik[1])+(2*pws)
AICcws<-AICws + ((2*pws*(pws+1))/(n-pws-1))
BICws<-(-2*mw2$loglik[1]) + pws*log(n)

AICll<-(-2*mll2$loglik[1])+(2*pws)
AICcll<-AICll + ((2*pws*(pws+1))/(n-pws-1))
BIClls<-(-2*mll2$loglik[1]) + pws*log(n)

AICcat <- AIC(modcat)
BICcat <- BIC(modcat)

AIC40 <- AIC(mod40)
BIC40 <- BIC(mod40)
```

```
maic <- cbind(AICcll,AICcws,AICcat,AIC40)
mbic <- cbind(BIClls,BICws,BICcat,BIC40)
require(xtable)
xtable(rbind(maic,mbic))

x11()
plot(KMs3, conf.int=F, xlab="Tempo", ylab="S(t)", lty=c(1,1),col=c(2,4),lwd = 1.5)
lines(temp1,Sg11, ylim=c(0,1), type="s",lty=2,lwd=1,col='darkgreen')
lines(temp1,Sg21,type="s",lty=2,col='darkgreen')
points(t,s.loglogistica,type="l",lty=5,col="#1a001a")
points(t,s.loglogistica1,type="l",lty=5,col="#1a001a")
points(t,s.weibuw,type="l",lty=5,col="#ccff33")
points(t,s.weibuw2,type="l",lty=5,col="#ccff33")

legend(100,0.4,lty=1,lwd=3,c("Kaplan-Meier(>40)"), bty = 'n', cex=1,col=2)
legend(100,0.35,lty=1,lwd=3,c("Kaplan-Meier(<40)"), bty = 'n', cex=1,col=4)
legend(100,0.3,lty=5,lwd=1,col="darkgreen",c("Cox"),bty="n", cex = 1)
legend(100,0.25,lty=5,lwd=1,col="#1a001a",c("Log-logístico"),bty="n", cex = 1)
legend(100,0.2,lty=5,lwd=1,col="#ccff33",c("Weibull"),bty="n", cex = 1)
```