Universidade de Brasília
IE - Instituto de Ciências Exatas
Departamento de Estatística

# MODELO DE RISCOS MÚLTIPLOS COM FRAÇÃO DE CURA

**NICOLLAS STEFAN SOARES DA COSTA**

Brasília
Julho 2013

**NICOLLAS STEFAN SOARES DA COSTA**

# MODELO DE RISCOS MÚLTIPLOS COM FRAÇÃO DE CURA

Monografia apresentada junto ao Curso de Estatística da Universidade de Brasília, na área de concentração, Análise de Sobrevivência, como requisito parcial à obtenção do título de Bacharel.

Orientadora: Profa: Dra. Juliana Betini Fachini

Brasília
Julho 2013

## DEDICATÓRIA

Aos meus pais,
**Efigênia Maria Soares da Costa e Murilo Vieira da Costa (in memoriam**), que tiveram paciência, dedicação e carinho na minha criação.

“Enquanto você não escolhe, tudo permanece possível.”
(Jaco Van Dormael)

# AGRADECIMENTOS

À professora **Dra. Juliana Betini Fachini**, pela orientação, paciência, conselhos e principalmente por acreditar no meu potencial para que pudesse concluir meu trabalho.

A todos os docentes e funcionários do Departamento de Estatística da UnB, principalmente, **Dr. Lúcio José Vivaldi, Dr. Geraldo da Silva Souza, Ms. Luís Gustavo do Amaral Vinha** e **Dra. Cibele Queiroz da Silva** que transmitiram experiências acadêmicas inesquecíveis e gratificantes que incorporei para o mercado de trabalho, bem como para a vida.

Aos meus amigos que sempre me motivaram e incentivaram para a finalização deste trabalho, e aos dias de boêmia que fizeram a minha cabeça não explodir com tantas fórmulas, números e letras. Em especial para, **Goiano, Bigas, Ed Conchinha, Jéjé, V2, Theteus, Sopão, Amandinha, Allen, Stanley, Sujeira, Teixeira, Juanito, Frango, Padilha, Tutu, Ilka Torres, Malu, Titi, Brubs, Capeta, Carolzinha, Juju, Lauane, Mau Mau, Apache, Ratão, Mayara (saquinho), Gui, Érica (tabaquinho), Raul (pixé), Diogo, Su, Márcia, Juninho** e todos os loucos que passaram pela minha vida.

À **Lara Gabriela**, que em pouco tempo me deu forças e puxões de orelha e principalmente, chamego, carinho e amor.

A todos que de alguma forma influenciaram ou ajudaram para a realização deste trabalho.

# SUMÁRIO

**LISTA DE FIGURAS** ..... **8**

**LISTA DE TABELAS** ..... **9**

**RESUMO** ..... **10**

**1 INTRODUÇÃO** ..... **11**

**2 REVISÃO BIBLIOGRÁFICA** ..... **13**

2.1 Características de dados de sobrevivência ..... 13
2.1.1 Tempo de falha ..... 13
2.1.2 Censura ..... 13
2.1.3 Apresentação dos dados de sobrevivência ..... 15
2.1.3.1 Função densidade de probabilidade ..... 15
2.1.3.2 Função de sobrevivência ..... 16
2.1.3.3 Função de risco ..... 16
2.1.3.4 Estimador Kaplan-Meier ..... 17
2.1.4 Relações entre as funções ..... 19
2.1.5 Modelos paramétricos ..... 19
2.1.5.1 Distribuição Log-Logística ..... 20
2.2 Formas da função risco ..... 22
2.3 Modelos de riscos múltiplos ..... 25
2.3.1 Modelo Log-Logístico múltiplo ..... 26
2.4 Fração de cura ..... 28
2.5 Inferência ..... 30
2.5.1 Teste da razão de verossimilhança ..... 32

**3 APLICAÇÕES** ..... **33**

3.1 Material ..... 33
3.2 Métodos ..... 33
3.2.1 Modelos de riscos múltiplos com fração de cura ..... 33
3.2.2 Estimação ..... 34
3.2.3 Teste da razão de verossimilhança modificada ..... 35

**4 RESULTADOS** ..... **36**

4.1 Análise descritiva ..... 36
4.2 Kaplan-Meier ..... 40
4.3 Curva tempo total em teste (TTT *plot*) ..... 41
4.4 Análise do modelo ..... 41
**5 CONSIDERAÇÕES FINAIS** ..... **45**
**6 REFERÊNCIAS** ..... **46**
**7 ANEXOS** ..... **48**

## LISTA DE FIGURAS

**Figura 1 –** Tipos de mecanismos de censura ..... 15

**Figura 2 –** Gráfico da função de sobrevivência estimada de pacientes com câncer de pulmão ..... 18

**Figura 3 –** Ilustrações de algumas curvas TTT ..... 23

**Figura 4 –** TTT *plot* Reinternações da região metropolitana de Belo Horizonte (SIH-SUS) ..... 24

**Figura 5 –** Gráfico TTT *plot* para pacientes com HIV-positivo (IPEC/Fiocruz) ..... 24

**Figura 6 –** Gráfico da curva de sobrevivência estimada com presença de fração de curados ..... 29

**Figura 7 –** Histograma da variável tempo de sobrevivência dos peixes ..... 36

**Figura 8 –** Diagrama de dispersão da covariável profundidade do rio ..... 37

**Figura 9 –** Diagrama de dispersão comprimento do peixe ..... 38

**Figura 10 –** Diagrama de dispersão transparência da água ..... 39

**Figura 11-** Curva de sobrevivência estimada por Kaplan-Meier para os dados de peixes 40

**Figura 12 –** TTT *plot* para os dados de peixes ..... 41

# LISTA DE TABELAS

**Tabela 1 –** Estatísticas básicas para tempo de sobrevivência dos peixes ..... 37

**Tabela 2 –** Estatísticas básicas para a covariável profundidade do rio ..... 38

**Tabela 3 –** Estatísticas básicas para a covariável comprimento do peixe ..... 39

**Tabela 4 –** Estatísticas básicas para a covariável transparência da água ..... 40

**Tabela 5 –** Estimativas dos parâmetros dos modelos de riscos proporcionais sem fração de cura e modelo de riscos proporcionais com fração de cura. ..... 42

**Tabela 6 –** Estimativas dos parâmetros e erro padrão do modelo de riscos múltiplos para os dados de peixes. ..... 44

# RESUMO

## Modelo de riscos múltiplos com fração de cura

No presente trabalho, será abordado o modelo bi-Log-Logístico com fração de cura. Este modelo possui a vantagem de ser mais flexível em relação ao Log-Logístico por possuir características de acomodar não somente funções de risco monótonas, como também acomodar uma grande variedade de funções não monótonas, como por exemplo, multimodais e em forma de banheira (“U”). E a partir da teoria de fração de cura é construído o modelo que será aplicado a um conjunto de dados reais para ilustrar a teoria apresentada, bem como a seleção do modelo que melhor se ajusta aos dados.

*Palavras-chaves:* Modelos de risco múltiplo, fração de cura, distribuição Log-Logística múltipla com fração de cura, dados censurados.

# 1 INTRODUÇÃO

Em certos estudos em que a variável observada é o tempo até a ocorrência de um evento de interesse, a análise de sobrevivência torna-se uma ferramenta indispensável para a análise dos dados. Usualmente, este tempo é denotado por tempo de falha ou tempo de sobrevivência.

Geralmente, estudos clínicos e médicos utilizam esta ferramenta na análise dos dados. Em pesquisas industriais, o evento pode estar relacionado ao tempo até a falha de um produto, ou o uso da garantia pelo consumidor. Desta forma, podemos generalizar o uso da análise de sobrevivência em variadas áreas, como: economia, seguros, bancos, biologia entre outros.

No entanto, tal técnica possui algumas peculiaridades inerentes ao conjunto de dados. A primeira característica é a restrição da variável contínua T (tempo) possuir o domínio definido nos reais positivos ($R^+$). Assim algumas distribuições usuais perdem importância (distribuição Normal). A segunda particularidade, e mais comum em dados de sobrevivência, é a chamada censura, ou seja, alguns indivíduos ou objetos do estudo têm a resposta observada de forma parcial ou incompleta. Alguns motivos que colaboram para isso são: morte por outros motivos, desistência do estudo, mudança de localidade, término do estudo.

Além dessas características, existe a presença em certos estudos de variáveis auxiliares (covariáveis). Outro aspecto relevante, e comumente encontrado, é a presença de múltiplas causas para o evento de interesse considerado na pesquisa. Neste caso, os estudos são conhecidos na literatura como modelos de riscos múltiplos ou competitivos, em que há mais de uma causa apresentada para definir a ocorrência da falha. Assim, o evento de interesse é analisado levando-se em conta o primeiro fato que acarretou a falha do objeto ou indivíduo de estudo.

Quando tomamos como exemplo o acompanhamento de pacientes com uma determinada doença, um dos focos principais é a quantidade de pacientes que responderam bem ao tratamento, ou seja, a possibilidade de tornarem-se imunes. Assim, considerados curados, não suscetíveis ao evento de interesse.

Modelos que assumem uma proporção de curados são chamados de modelos de fração de cura, e reescrevemos a função de sobrevivência considerando os

indivíduos sujeitos a falha e os indivíduos curados. Essa função é conhecida como função de sobrevivência populacional.

Levando-se em conta a base teórica da análise de sobrevivência enunciada, o presente trabalho abordará a modelagem de dados utilizando os conceitos de riscos múltiplos e fração de cura para analisar os dados de peixes da espécie “Notropis Dourado, crysoleucas de Notemigonus”.

# 2 REVISÃO BIBLIOGRÁFICA

## 2.1 Características de dados de sobrevivência

### 2.1.1 Tempo de falha

Análise de sobrevivência consiste em um aglomerado de procedimentos estatísticos para análise de dados relacionados ao tempo até a ocorrência de um evento de interesse. Geralmente, esse termo é denotado por tempo de falha ou tempo de sobrevivência.

Por ser de suma importância para os dados de sobrevivência, o tempo de falha deve ser bem definido para evitar qualquer tipo de ambiguidade. Assim, devem-se estabelecer três elementos para definir corretamente o tempo de falha: fixar o tempo de início do estudo, a escala de medida a ser utilizada, e a formulação do evento de interesse, comumente considerado como indesejável. Eventualmente, a falha é considerada como a morte do indivíduo, ou até mesmo em certos estudos a recidiva de uma doença.

A falha pode acontecer devido a uma única causa ou por várias causas, e pode ser completamente ou parcialmente conhecida. O caso em que há potencialmente vários motivos determinando a falha denota-se na literatura como riscos competitivos, e geralmente por riscos múltiplos (Prentice et al., 1978).

### 2.1.2 Censura

Os estudos de sobrevivência envolvem resposta temporal, e alguns indivíduos não chegam a experimentar o evento de interesse: a falha. Segundo Colosimo e Giolo (2006), estas observações, denominadas censuras, podem ocorrer por uma variedade de razões, dentre elas, perda de contato com o paciente, efeitos adversos ao tratamento, término do estudo, entre outros motivos. Mesmo sendo observações parciais, os dados censurados não devem ser excluídos das análises, pois podem acarretar em conclusões viciadas.

Desta forma, a introdução de uma nova variável na análise que indica se o valor do tempo para o indivíduo foi ou não observado completamente se faz necessária. Assim, a variável indicadora de censura, ou somente censura, é definida como:

$$\delta_j = \begin{cases} 1, se\ t_j \text{ é } tempo\ de\ falha \\ 0, se\ t_j \text{ é } tempo\ de\ censura \end{cases} \qquad j = 1,2,3,\dots,n.$$

As censuras são definidas em três mecanismos distintos: Tipo I, Tipo II e Aleatório.

(i) Censura tipo I

Ao início do experimento, o pesquisador pré-estabelece um período de tempo $\tau$ em que o experimento irá terminar. Portanto, ao final do estudo, todos os indivíduos que não experimentarem a falha são considerados censurados, ou seja, a informação sobre o tempo foi parcialmente observado.

(ii) Censura tipo II

O estudo é conduzido até que um número de falhas $(k \leq n)$ especificado no começo do experimento se realize. Ao ocorrer o número de falhas desejado, o estudo é encerrado e todos os indivíduos ou objetos que não falharam no período são considerados censurados. Esse tipo de censura é mais comumente encontrado em estudos industriais, onde a técnica é chamada de análise de confiabilidade.

(iii) Censura aleatória

A censura aleatória ocorre, por exemplo, quando um indivíduo experimenta a falha por motivos distintos do estudo e até mesmo por razões como falta de acompanhamento, efeito adverso ou por algum motivo de mudança de localidade. Tal censura é a mais comum em estudos médicos e clínicos.

Os tipos de censura definidos acima são denominados de censura à direita, porém também existem outras classes como, censura à esquerda e censura intervalar, que não serão abordadas no presente trabalho. A seguir, na Figura 1, estão ilustrados os mecanismos de censura, além de um exemplo com dados completos. Os símbolos "•" e "◦" representam as observações de falha e censura, respectivamente.

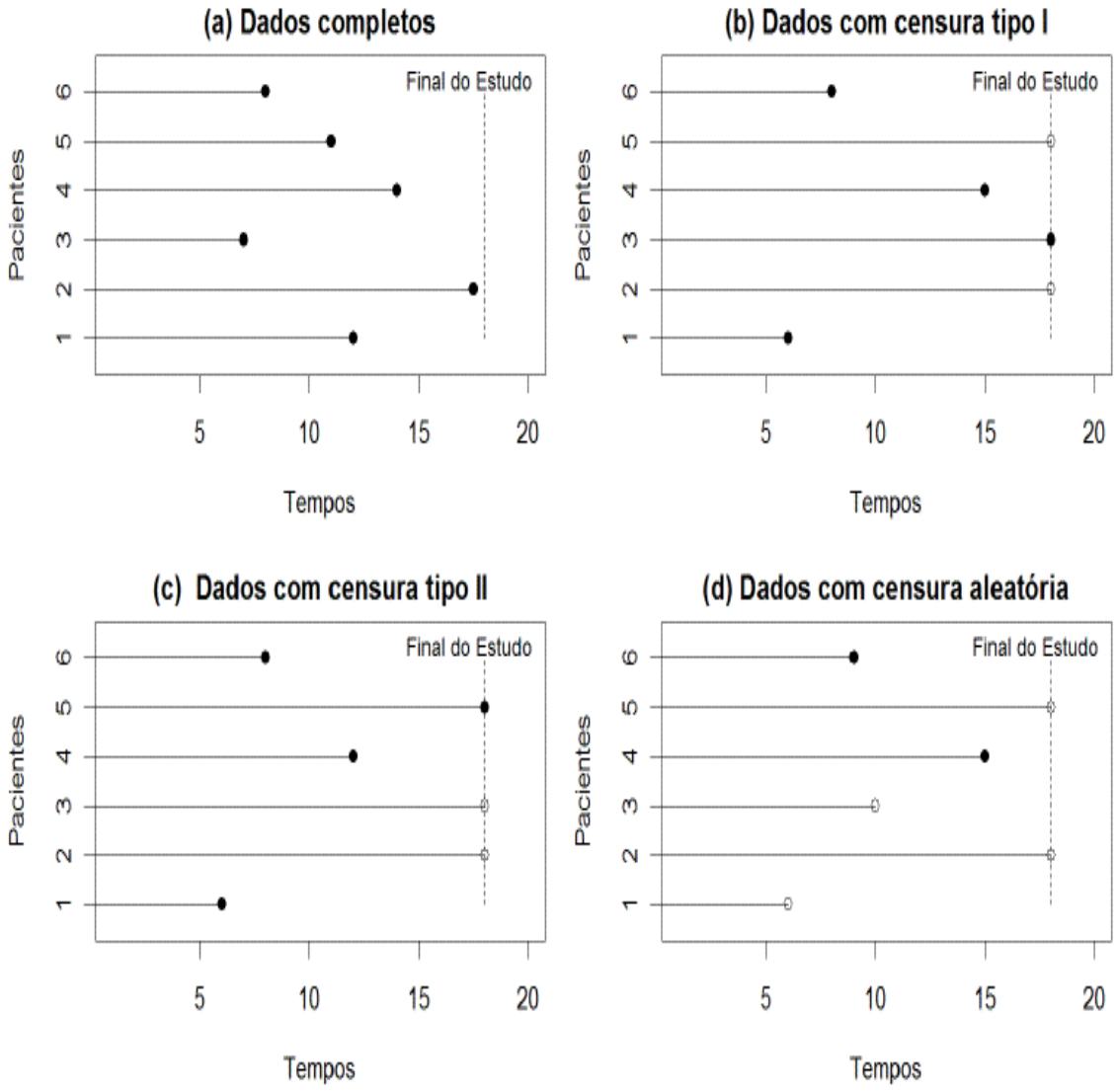


Figura 1: Tipos de mecanismos de censura. (Colosimo;Giolo, 2006).

### 2.1.3 Apresentação dos dados de sobrevivência

Os dados de sobrevivência são usualmente representados pela variável contínua T, que possui a restrição no seu domínio dos reais positivos e pode ser expressa através de diversas funções matemáticas. Dentre estas, temos: a função de densidade de probabilidade $f(t)$, a função de sobrevivência $S(t)$ e a função de risco ou taxa de falha $h(t)$. A seguir, será descrita com mais detalhes cada função, bem como a relação matemática existente entre elas.

#### 2.1.3.1 Função densidade de probabilidade

A função densidade de probabilidade é definida como o limite da probabilidade de um indivíduo falhar no intervalo de tempo por unidade de tempo com $\Delta t \longrightarrow 0$, e é expressa por (KLEIN;MOESCHBERGER,1997):

$$f(t) = \lim_{\Delta t \longrightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}, \quad (1)$$

onde a função $f(t)$ é sempre positiva para todo $t$ e $\int_0^\infty f(t) = 1$.

### 2.1.3.2 Função de sobrevivência

A função de sobrevivência $S(t)$ é a forma mais natural de se apresentar os dados de sobrevivência. Deste modo, é apresentada como a probabilidade de um indivíduo sobreviver além de um tempo $t$, ou equivalentemente, como a probabilidade de um individuo não falhar até certo tempo $t$. Assim, a função $S(t)$ é determinada como:

$$S(t) = P(T > t) = \int_t^\infty f(x)dx. \quad (2)$$

Ainda temos que $S(t)$ é uma função monótona não crescente com as seguintes características: (COX;OAKES,1984)

$$\lim_{t \longrightarrow 0} S(t) = 1 \quad \text{e} \quad \lim_{t \longrightarrow \infty} S(t) = 0,$$

bem como uma importante relação com a função de distribuição acumulada expressa por:

$$S(t) = 1 - F(t). \quad (3)$$

### 2.1.3.3 Função de risco

A função de risco $h(t)$, também conhecida como função taxa de falha, é representada pelo limite da razão da probabilidade de um indivíduo experimentar a falha em um intervalo de tempo $[t, t + \Delta t)$, admitindo-se que este não falhou até o tempo $t$, dividido pelo intervalo de tempo $\Delta t$. Pode-se expressar $h(t)$ por (LAWLESS, 2003):

$$h(t) = \lim_{\Delta t \longrightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \quad (4)$$

A mesma função pode ser enunciada de forma mais simplificada, e incluir a função densidade de probabilidade e a função de sobrevivência:

$$h(t) = \frac{f(t)}{S(t)}. \tag{5}$$

A função taxa de falha também é interpretada como o risco instantâneo do indivíduo experimentar o evento de interesse, ou seja, é um indicador adequado da inclinação a falha após decorrer uma unidade de tempo. Observe que a função de risco assume todos os valores reais positivos, e pode possuir valores acima de um.

No contexto da análise de sobrevivência é comum o uso da relação entre a função de sobrevivência e a função de risco acumulada, que pode ser definida como:

$$S(t) = \exp\{-H(t)\}, \tag{6}$$

onde $H(t) = -\int_0^t h(u)du$.

Como para diversas distribuições de probabilidade a função de sobrevivência pode assumir formas semelhantes, a modelagem da função de risco torna-se uma ferramenta essencial para a análise, pois pode ter forma crescente, decrescente, constante ou não monótona e pode ter uma gama de diferenças entre o conjunto de funções.

#### 2.1.3.4 Estimador Kaplan-Meier

O estimador de Kaplan-Meier é uma técnica não paramétrica bastante conhecida e utilizada na análise de sobrevivência, também chamado de estimador limite-produto por suas características. Tal estimador é uma adaptação da função de sobrevivência empírica.

O estimador é utilizado de forma mais descritiva e auxilia na escolha do modelo paramétrico mais adequado aos dados. Segundo Colosimo e Giolo (2006), é preferível a utilização desse estimador aos estimadores de Nelson-Aalen e Tábua de Vida, pois o estimador de Kaplan-Meier é um estimador de máxima verossimilhança.

Como os dados de sobrevivência possuem observações censuradas, Kaplan e Meier (1958) propuseram um estimador que incorporasse no denominador o número de

indivíduos sob-risco. Desta forma, apenas são considerados os indivíduos sob-risco no instante $t$ que inclui os indivíduos censurados.

Assim, a função de sobrevivência do estimador Kaplan-Meier é definida como:

$$\hat{S}(t) = \prod_{j:t_{j<t}} \left(\frac{n_j - d_j}{n_j}\right) = \prod_{j:t_{j<t}} \left(1 - \frac{d_j}{n_j}\right),$$

onde $n_j$ é o número de indivíduos sob-risco em $t_j$ e $d_j$ o número de falhas em $t_j$, bem como $t_1 < t_2 < \cdots < t_k$ os $j$ - ésimos tempos distintos e ordenados de falha.

A seguir, na Figura 2, um exemplo da curva de sobrevivência estimada pelo método de Kaplan-Meier para um conjunto de dados.

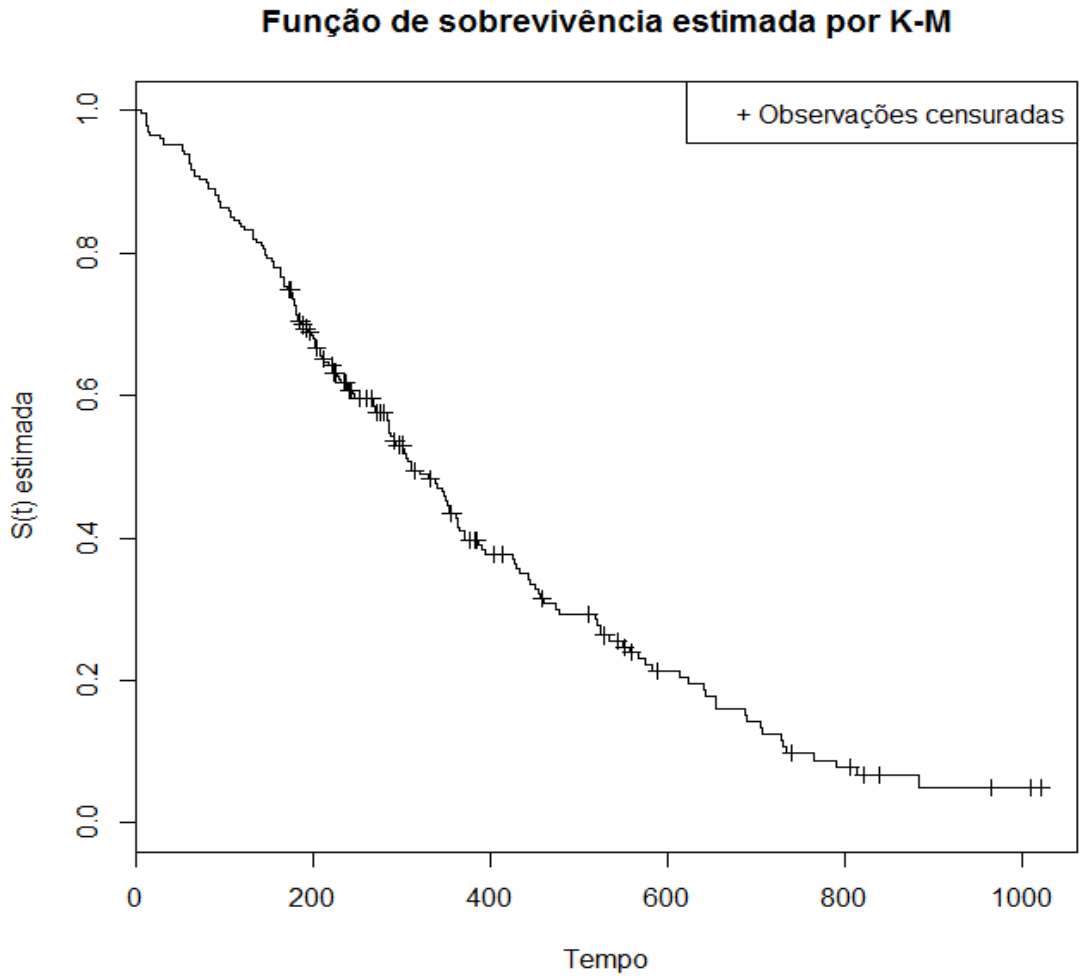


Figura 2 – Gráfico da função de sobrevivência estimada de pacientes com câncer de pulmão – banco de dados do *software* R.

### 2.1.4 Relações entre as funções

Entre as funções caracterizadas anteriormente, estão definidas algumas relações matemáticas. Entre estas, algumas de maior relevância para o trabalho são expressas a seguir:

$$f(t) = \frac{dF(t)}{dt}.$$

E pela expressão em (3), temos:

$$f(t) = \frac{d[1 - S(t)]}{dt} = -S'(t).$$

Assim, considerando a equação em (5) substituindo $f(t) = -S'(t)$ , reescrevemos a equação $h(t)$ como:

$$h(t) = \frac{-S'(t)}{S(t)} = \frac{\partial[-logS(t)]}{\partial t}. \quad (7)$$

E a função densidade de probabilidade pode ser definida, após pequena manipulação, por:

$$f(t) = h(t)S(t). \quad (8)$$

### 2.1.5 Modelos Paramétricos

Geralmente, o tempo de sobrevivência é associado a várias causas do cotidiano, sendo assim de difícil representação matemática. Para contornar tal problema são utilizados modelos paramétricos para modelar de forma mais fidedigno o tempo de sobrevivência até a ocorrência do evento de interesse.

A grande vantagem de se utilizar modelos paramétricos é a possibilidade de extrapolação da curva de sobrevivência para valores de tempo para os quais não se observa falhas. Segundo Latimer (2011), em estudos de custo-efetividade, muitas vezes é necessário extrapolar curvas de sobrevida, visto que, geralmente, os estudos de sobrevivência possuem um tempo de acompanhamento menor do que o esperado pelo pesquisador.

Desta maneira, certas distribuições de probabilidade são bastante utilizadas na análise de sobrevivência. Alguns exemplos são os modelos Exponencial, Weibull, Log-Normal, Log-Logística, além de distribuições mais complexas como Burr XII, Gama Generalizada, Weibull Exponenciada, entre outras. No entanto, neste trabalho será dada mais atenção à distribuição Log-Logística.

#### 2.1.5.1 Distribuição Log-logística

Seja uma variável aleatória não negativa T que segue uma distribuição Log-Logística com parâmetros $\alpha$ e $\gamma$. Então, a função de densidade de probabilidade é descrita como:

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

sendo $\alpha > 0$ e $\gamma > 0$ os parâmetros de escala e forma da distribuição, respectivamente.

Naturalmente, definem-se as funções de sobrevivência, taxa de falha e o p-ésimo percentil, em ordem por:

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

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

$$t_p = \alpha\left[\frac{p}{1-p}\right]^{\gamma}.$$

O primeiro (esperança) e segundo (variância) momento da distribuição Log-Logística são representados por:

$$E(T) = \frac{[\pi\alpha \csc(\pi/\gamma)]}{\gamma}, \qquad para\ \gamma > 0,$$

$$Var(T) = \left[\frac{2\pi\alpha^2 \csc(2\pi/\gamma)}{\gamma}\right] - E(T)^2.$$

Ao se lidar com a distribuição Log-Logística, em certas ocasiões se faz necessário o uso do logaritmo do tempo. Como $T$ é uma variável aleatória que possui

distribuição Log-Logística, por conseguinte $Y = log\ (T)$ que tem distribuição Logística com função densidade de probabilidade denotada por:

$$f(y) = \frac{1}{\delta} exp\left\{\frac{y-\mu}{\sigma}\right\}\left[1 + \exp\left(\frac{y-\mu}{\sigma}\right)\right]^{-2}, \quad (9)$$

sendo os parâmetros de locação e escala, respectivamente, $-\infty < \mu < \infty$ e $\sigma > 0$. Logo, as funções de sobrevivência e risco são expressas por:

$$S(t) = \frac{1}{1 + \exp\left[\frac{y-\mu}{\sigma}\right]},$$

$$h(t) = \frac{1}{\delta} exp\left\{\frac{y-\mu}{\sigma}\right\}\left[1 + \exp\left(\frac{y-\mu}{\sigma}\right)\right]^{-1}.$$

Assim, os parâmetros da distribuição estão relacionados da seguinte forma: $\alpha = exp\ (\mu)$ e $\gamma = 1/\sigma$. Vários pacotes estatísticos trabalham com essa relação que a torna bastante importante.

## 2.2 Formas da função risco

Como a função de sobrevivência possui estruturas semelhantes para distintos modelos, à função de risco passa a desempenhar um papel de grande utilidade na análise dos dados. Por acomodar funções de diferentes tipos, a forma como é representada ganha destaque para a escolha da distribuição que se ajusta aos dados.

Uma metodologia utilizada para selecionar o modelo mais apropriado baseia-se em informações retiradas do gráfico tempo total em teste, ou mais conhecido como curva TTT. Proposto por Aarset (1987) tal gráfico auxilia na escolha do melhor modelo, mesmo antes de qualquer ajuste, para a modelagem dos dados. O gráfico é construído a partir de:

$$G\left(\frac{r}{n}\right) = \frac{\sum_{i=1}^{n} T_{i:n} + (n-r)T_{r:n}}{\sum_{i=1}^{n} T_i} \qquad versus \qquad \frac{r}{n},$$

onde $T_{i:n}$ $i = 1,2,\dots,n$ são as estatísticas de ordem e $r = 1,2,\dots,n$.

Através desse método, tanto informações qualitativas como informações estruturais podem ser determinadas a respeito do estudo em questão. Enquanto a primeira questão é extraída diretamente do gráfico, as informações estruturais cabem ao pesquisador e seu conhecimento prévio sobre o assunto e estudos correlacionados para a análise.

Nesse contexto, na Figura 3, temos a representação de algumas formas para a curva TTT e logo a seguir alguns exemplos com conjunto de dados nas Figuras 4 e 5.

Na Figura 3, caso a curva seja côncava (C) ou convexa (B), a função é crescente ou decrescente monotonicamente. Se a curva possui uma característica diagonal (A), trata-se de uma função de risco constante. Já em casos como (E) onde temos uma curva côncava e em seguida uma curva convexa, a função de risco possui característica unimodal. No caso inverso (D), em que primeiramente temos uma curva convexa e em seguida uma curva côncava, a função de risco toma forma de banheira (“U”).

Na Figura 4, destaca-se a presença inicialmente de uma curva convexa e em seguida uma curva côncava, apontando que a função taxa de falha possui o formato

de banheira (“U”). Deste modo, os modelos Weibull Exponenciada, Weibull Modificada, Burr XII Aditiva, Beta Weibull Generalizada, entre outros, são possíveis indicações de distribuições para a modelagem dos dados.

Já na Figura 5, temos primeiramente uma curva convexa e após a presença de uma parte mais complexa contendo uma reta ou uma leve curva côncava, indicando assim um exemplo de aplicabilidade dos modelos de riscos múltiplos, pois acomodam funções constantes, crescentes, decrescentes, unimodais e banheira.

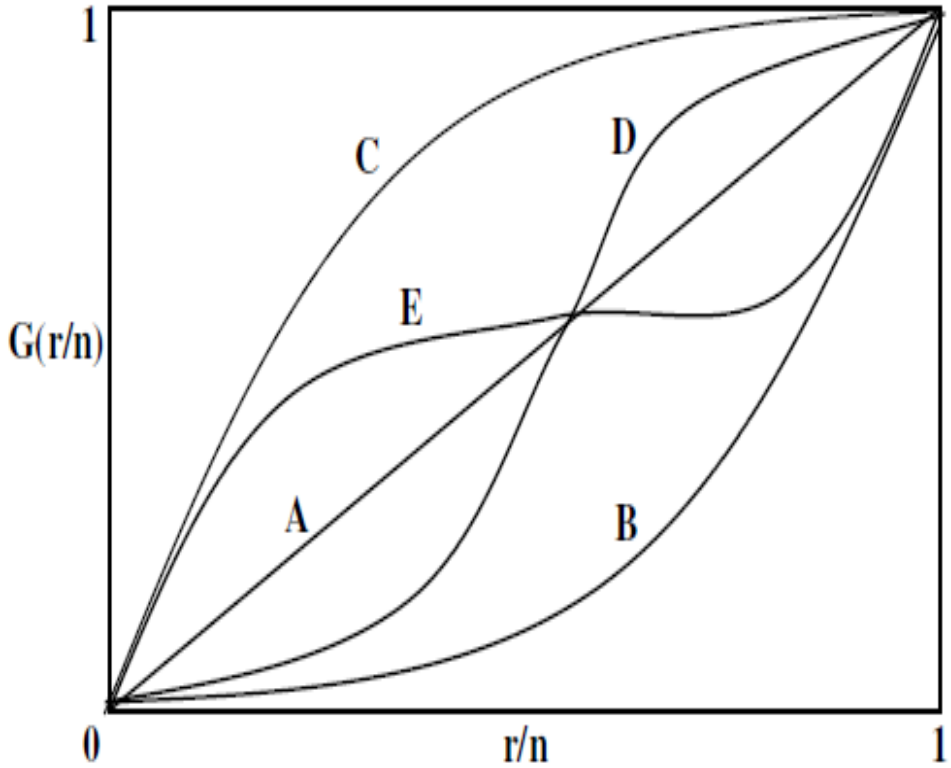


Figura 3 – Ilustrações de algumas curvas TTT.

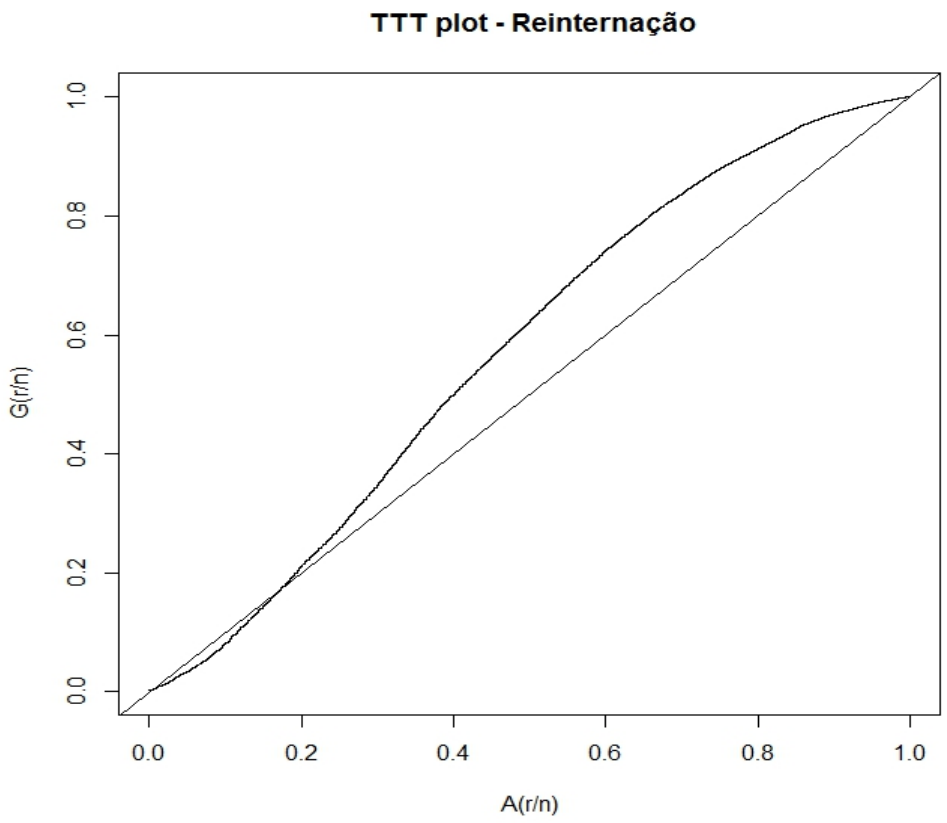


Figura 4 – TTT *plot* - Reinternações da região metropolitana de Belo Horizonte (SIH-SUS).

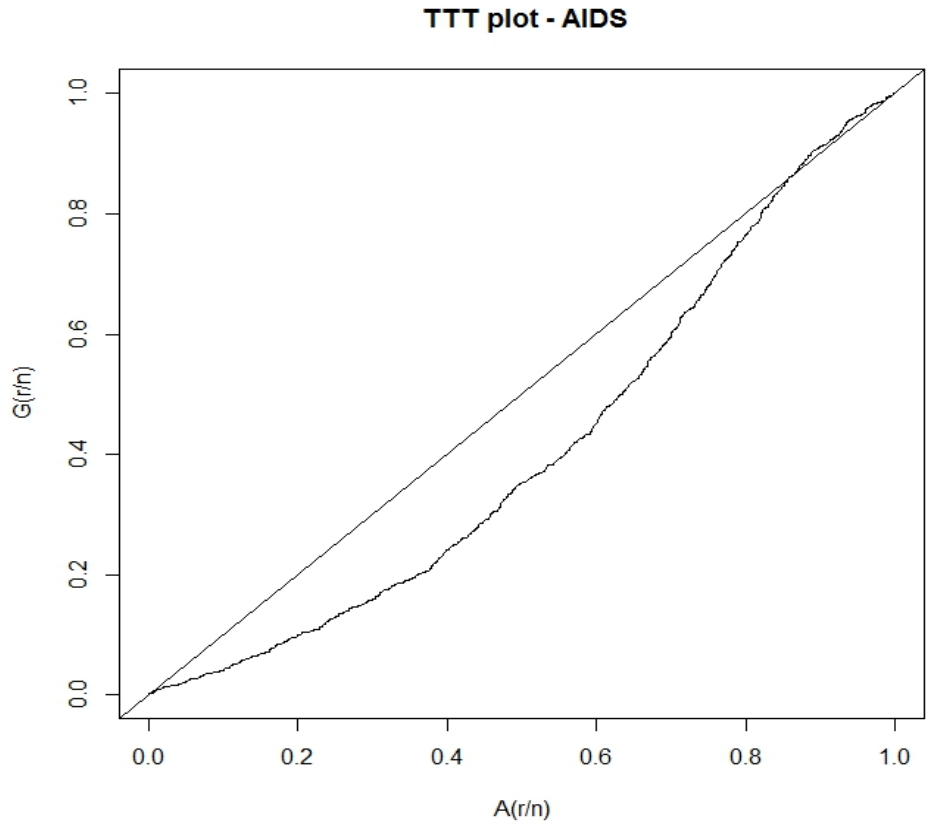


Figura 5 – Gráfico TTT *plot* para pacientes com HIV-positivo (IPEC/Fiocruz).

## 2.3 Modelos de riscos múltiplos

A família de modelos de riscos múltiplos é de valiosa importância para análise de certos dados de sobrevivência e confiabilidade, uma vez que são bem flexíveis e acomodam distintas formas de curvas de risco. Vários autores, como Berger e Sun (1993), Louzada-Neto (1999) e Fachini et al. (2008), aplicaram tal teoria em vários exemplos práticos. Tais modelos são geralmente aplicados em situações incluindo riscos competitivos, riscos complementares e sistemas mascarados, ainda que não haja informações completas da origem do motivo da falha.

A grande vantagem do uso dessa modelagem é a possibilidade de não somente ajustar curvas de risco crescente, constante e decrescente, como também acomodar formas não monótonas, como por exemplo, curvas multimodais e em forma de banheira (forma de U). Esse fato pode ser analisado no caso em que a curva TTT apresenta várias regiões côncavas e convexas, ou seja, direcionamentos para riscos multimodais, a utilização dos modelos de riscos múltiplos se torna uma ferramenta efetiva para a melhor adequação do modelo.

Modelos de riscos múltiplos fundamentam que os indivíduos ou objetos estão sujeitos a $k \geq 2$ motivos que, independentes, levam ao evento de interesse. O principal motivo causador da falha pode ser assim, completamente ou parcialmente conhecido no estudo.

Deste modo, se $Z_j,\ j = 1, 2, \ldots, k$, independentes, são os tempos de falha relacionados ao $j$-ésimo motivo de ocorrer o evento de interesse, temos o vetor $v_i^T = (Z_1, Z_2, Z_3, \ldots, Z_k)$, com $i = 1, 2, \ldots, n$ e o vetor $v_i^T$ relacionado aos $k$ motivos de falha dos $i$-ésimos indivíduos ou mecanismos de estudo, sendo que o primeiro motivo que levar o indivíduo a falhar será o tempo considerado para o estudo, ou seja, $T_i = min\ (Z_1, Z_2, \ldots, Z_k)$.

Com a preposição dos $Z_j's$ motivos independentes, e que $f_j(z)$ é a função densidade de probabilidade da variável $Z_j$, conseguinte temos a função de risco para o modelo de risco múltiplo como:

$$h(t) = \sum_{j=1}^{k} h_j\,(t),$$

e proposto por Louzada-Neto (1999), a função de risco é definida por:

$$h(t) = \sum_{j=1}^{k} \frac{\omega_j t^{\omega_{j-1}}}{{\mu_j}^{\omega_j}} q_j(t, \mu_j, \omega_j, \kappa), \tag{10}$$

sendo que $t > 0$ , $\mu_j, \omega_j, \kappa$ são desconhecidos e parâmetros positivos e as funções monótonas de parâmetro de forma $q_j(.)$ iguais a um quando os argumentos restantes são iguais à zero.

### 2.3.1 Modelo Log-Logístico múltiplo

O modelo Log-Logístico múltiplo é especificado como um caso particular da equação (10), e a representação da função de risco é dada por:

$$h(t) = \sum_{j=1}^{k} \frac{\omega_j t^{\omega_j - 1}}{\mu_j^{\omega_j} + t^{\omega_j}}, \tag{11}$$

para

$$q_j(t; \mu_j; \omega_j; \kappa) = \frac{1}{1 + u_j},$$

sendo que, $u_j = \left(\frac{t}{\mu_j}\right)^{\omega_j}$.

No entanto, há situações em que o tempo de falha pode estar associado a um vetor de variáveis explanatórias. Neste caso, o modelo de risco múltiplo pode ser estendido para incluir as variáveis. Reescrevendo a fórmula (11), temos:

$$h(t) = \sum_{j=1}^{k} \frac{\omega_j t^{\omega_j - 1} \exp(x_i^T \beta_j)}{\mu_j + t^{\omega_j} \exp(x_i^T \beta_j)}.$$

A partir da transformação $\beta_{0j} = -log(\alpha_j)$ proposta por Mazucheli; Louzada e Achcar (2001) obtêm-se para o modelo Log-Logístico múltiplo a função taxa de falha expressa por:

$$h(t) = \sum_{j=1}^{k} \frac{\omega_j t^{\omega_j - 1} \exp(x_i^T \beta_j)}{1 + t^{\omega_j} \exp(x_i^T \beta_j)}, \tag{12}$$

sendo que $\beta_j^T = (\beta_{0j}, \beta_{1j}, \dots, \beta_{pj})$, $\beta^T = (\beta_1, \beta_2, \dots, \beta_k)$, $x_i^T = (1, x_{i1}, x_{i2}, \dots, x_{ip})$, $\omega^T = (\omega_1, \omega_2, \dots, \omega_k)$ e $x_i^T \beta_j = \beta_{0j} + \beta_{1j} x_{1i} + \cdots + \beta_{pj} x_{ip}$.

E para melhor exemplificação do modelo, temos a seguinte caracterização dos parâmetros: $\omega_j$, que representa os riscos; $x_i^T$, que representa o vetor de covariáveis do modelo; $\beta_j$ os parâmetros desconhecidos associados ao vetor de covariáveis e $t$, o tempo de sobrevivência dos peixes.

O modelo de riscos múltiplos possui ainda algumas suposições que devem ser verificadas, são essas: as estimativas para os parâmetros de risco devem ser positivas, significativamente diferentes e $\widehat{\omega}_1 < \widehat{\omega}_2$. Caso, estas suposições não sejam válidas, o modelo de riscos múltiplos não é ajustável ao conjunto de dados e retornamos para modelos mais simples ou outros modelos de riscos múltiplos.

## 2.4 Fração de cura

Nos modelos de análise de sobrevivência, supõe-se que no universo dos indivíduos presentes no estudo, todos experimentarão o evento de interesse investigado, e, portanto, durante o acompanhamento acontecerá à falha ou os dados serão considerados censurados (FACHINI, 2011).

Neste caso, a função de sobrevivência estimada pelo estimador Kaplan-Meier possui certas características comuns à maioria dos estudos de sobrevivência. Todavia, em certas situações há a presença no conjunto de dados de uma porção de indivíduos em que não sucede o evento de interesse. Tais indivíduos são frequentemente chamados de curados, e em certas pesquisas, de imunes ou não suscetíveis. Alguns ensaios clínicos, como por exemplo, o estudo da sobrevivência de pacientes que se submeteram a certo tipo de transplante, o foco principal é a não rejeição pelo organismo e conseguinte a sobrevida. Sendo assim, a presença de pacientes não suscetíveis à falha é esperada.

Uma das principais características de dados com a presença de fração de curados é o fato da função de sobrevivência, que naturalmente ao longo do estudo tende a zero, desenvolver uma cauda constante e durante um longo período de tempo em um nível de probabilidade diferente de zero, sendo assim chamada de função de sobrevivência imprópria.

Desta maneira, pela metodologia proposta por Berkson e Gage (1952), a função de sobrevivência é reescrita na forma de mistura e dada por:

$$S_{pop}(t) = (1-\varphi) + \varphi S(t) \qquad (13)$$

sendo que $(1-\varphi)$ é a probabilidade dos indivíduos curados e $\varphi$ a probabilidade relacionada aos indivíduos suscetíveis a falha, onde $\varphi \in [0,1]$ e a função de sobrevivência segue certas características:

$$\lim_{t\longrightarrow 0} S_{pop}(t) = 1 \quad \text{e} \quad \lim_{t\longrightarrow \infty} S_{pop}(t) = (1-\varphi).$$

Quando $\varphi$ assume o valor igual a um, reduzimos a função $(13)$ para a função própria $S(t)$. Por outro lado, quando $\varphi$ assume o valor igual à zero, temos a função (13) reduzida à $S_{pop}(t) = 1$, ou seja, toda a população é de indivíduos curados.

Através do gráfico da função de sobrevivência empírica, é possível a identificação de tal acontecimento. A seguir, na Figura 6, pode-se apreciar um exemplo em um conjunto de dados.

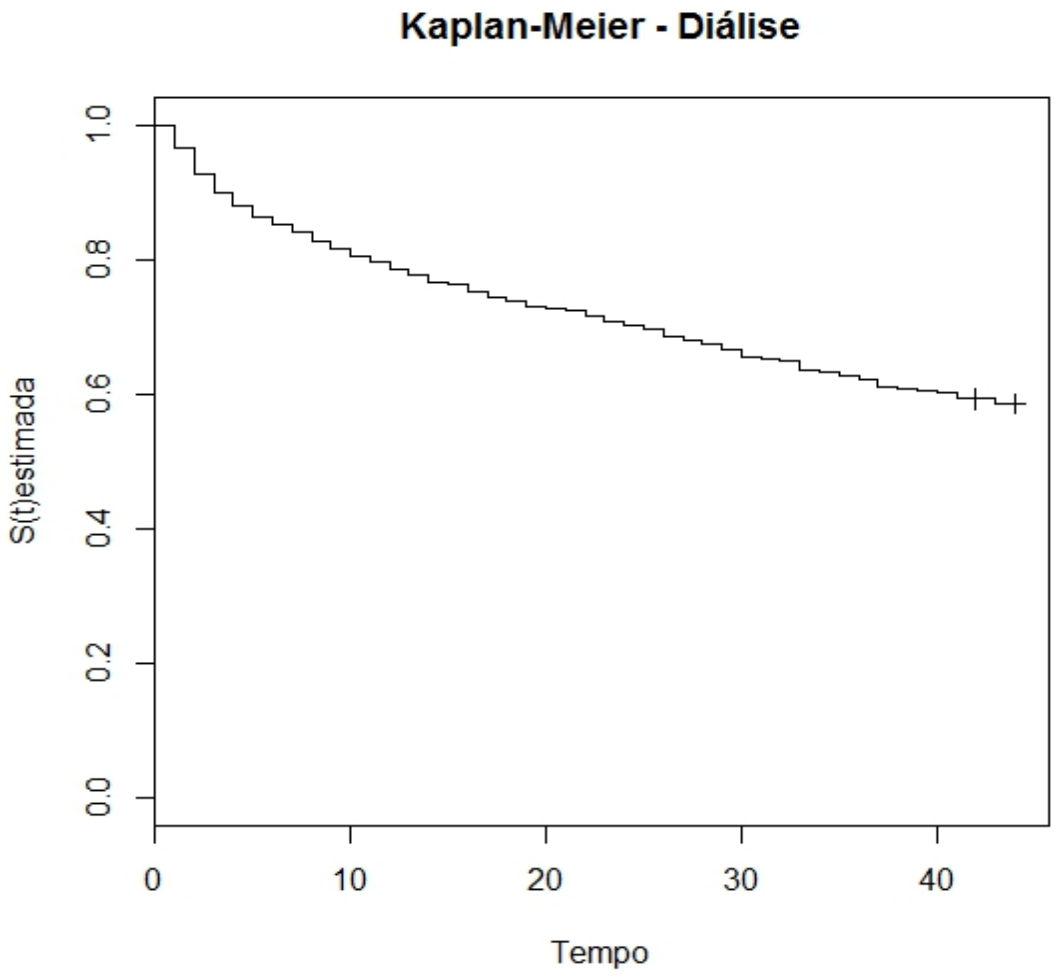


Figura 6 – Gráfico da curva de sobrevivência estimada com presença de fração de curados.

## 2.5 Inferência

A utilização de métodos cujo objetivo é fazer afirmações sobre parâmetros desconhecidos no universo dos dados com base em uma parcela da população (amostra) é chamada Inferência Estatística. Tais amostras são selecionadas de forma aleatória da população e auxiliam nas estimações.

A metodologia mais utilizada, na literatura estatística, é o método dos mínimos quadrados. Contudo, seu uso torna-se inviável nesse tipo de situação pelo fato de não conseguir agregar as observações censuradas para a estimação dos parâmetros. Sendo assim, o método de máxima verossimilhança torna-se adequado para a estimação, pois em grandes amostras possui propriedades desejáveis além dos demais métodos.

Desta maneira, o método de máxima verossimilhança obtém os estimadores através da escolha do valor do parâmetro que maximiza a probabilidade que melhor explica a amostra observada (Fachini, 2006). Supõe-se uma amostra de variáveis aleatórias independentes, $T_1, T_2, \dots, T_n$, tal que $T_i = min\,(Z_1, Z_2, \dots, Z_k)$, $i = 1, 2, \dots, n$, $k \geq 2$ e associada a cada $T_i$, há uma variável indicadora de censura $\delta_i$, sendo que $\delta_i = 1$ se $t_i$ é uma observação de tempo de falha e $\delta_i = 0$, caso seja uma observação censurada.

Temos então, a função de verossimilhança baseada nas $n$-duplas $(t_1, \delta_1)$, $(t_2, \delta_2)$,..., $(t_n, \delta_n)$, que são os tempos de falhas e tempos de censuras e respectivas indicadoras de falha. Desta forma, podem ser dividas em dois grupos tais que, as $r$ primeiras observações são as não censuradas, e as $n - r$ observações restantes são censuradas. E pode-se expressar a função de verossimilhança por:

$$L(t_i; \theta) = \prod_{i=1}^{r} f(t_i; \theta) \prod_{i=r+1}^{n} S(t_i; \theta),$$

que equivale a:

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

utilizando a equação (8) e fazendo algumas modificações temos:

$$L(t_i; \theta) = \prod_{i=1}^{n} [h(t_i; \theta)]^{\delta_i} S(t_i; \theta), \qquad (14)$$

em que $h(t_i)$ é definida em (10) e $\theta$ é o vetor de parâmetros. Assim, como a contribuição de cada observação censurada é sua função de sobrevivência, temos então $S(t_i) = \prod_{j=1}^{k} S_j\,(t_i)$, a função de sobrevivência do modelo de riscos múltiplos.

Para encontrar os valores de $\theta$ que maximizam o logaritmo de $L(\theta)$, ou seja, o estimador de máxima verossimilhança tem-se que resolver o seguinte sistema de equações:

$$U(\theta) = \frac{\partial log L(\theta)}{\partial \theta} = 0.$$

Dada uma amostra aleatória de variáveis aleatórias independentes $t_1, t_2, t_3, \dots, t_n$, com a função de risco correspondente como (12), de tal maneira que vinculado a $t_i$, exista um vetor de covariáveis $x_i^T$, bem como a variável indicadora de censura $\delta_i$. Desta forma, a representação da função de verossimilhança do modelo Log-Logístico múltiplo é explicitado por:

$$L(t_i; \omega, \beta) = \prod_{i=1}^{n}\left[\sum_{j=1}^{k}\frac{\omega_j t^{\omega_j - 1}\exp\{x_i^T\beta_j\}}{1 + t^{\omega_j}\exp\{x_i^T\beta_j\}}\right]^{\delta_i}\prod_{j=1}^{k}\left[1 + t_i^{\omega_j}\exp\{x_i^T\beta_j\right]^{-1},$$

sendo que $\theta = (\omega_j^T, \beta_j^T)^T$, $S_j(t_i) = \left[1 + t_i^{\omega_j}\exp\{x_i^T\beta_j\}\right]^{-1}$ e $S(t_i) = \prod_{j=1}^{k} S_j(t_i)$ que corresponde à função de sobrevivência do modelo Log-Logístico múltiplo.

Desta maneira, o logaritmo da função de verossimilhança é apresentado como:

$$l(t_i, \omega, \beta) = \sum_{i:\delta_i=1} log\left[\sum_{j=1}^{k}\frac{\omega_j t^{\omega_j - 1}\exp\{x_i^T\beta_j\}}{1 + t^{\omega_j}\exp\{x_i^T\beta_j\}}\right] - \sum_{i=1}^{n}\left[\sum_{j=1}^{k}\log\left(1 + t_i^{\omega_j}\exp\{x_i^T\beta_j\}\right)\right]. \qquad (15)$$

Utilizando o fato de que, assintoticamente, o estimador $\hat{\theta}$ possui distribuição assintótica normal multivariada, e sob certas hipóteses, com média $\theta$ e matriz de variância e covariância $I^{-1}(\theta)$, pode-se construir testes de hipótese e intervalos de confiança, e assim temos:

$$\hat{\theta} \sim N_{[k+k)p+1)]}(\theta, I^{-1}(\theta)),$$

onde $I(\theta) = -E\left[\ddot{L}(\theta)\right]$, em que

$$\ddot{L}(\theta) = \frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^T}.$$

Devido à presença de observações censuradas, não é possível o cálculo da informação de Fisher $I(\theta)$. Portanto, uma alternativa é a utilização da matriz $[\ddot{L}(\theta)]$ avaliada em $\theta = \hat{\theta}$, chamada de matriz de informação observada, que é uma estimativa consistente de $I(\theta)$. Temos, então, a matriz $\ddot{L}(\theta)$ expressa por:

$$\ddot{L}(\theta) = \begin{pmatrix} -\ddot{L}_{\omega\omega} & -\ddot{L}_{\omega\beta} \\ -\ddot{L}_{\beta\omega} & -\ddot{L}_{\beta\beta} \end{pmatrix},$$

e cada submatriz expressa de forma fechada para os modelos de riscos múltiplos.

### 2.5.1 Teste da razão de verossimilhança

Segundo Bozdangan (1987), a escolha do modelo mais adequado, para a estatística, é um tópico de extrema importância para a análise dos dados. A busca por um modelo mais parcimonioso, ou seja, um modelo que envolva um número mínimo possível de parâmetros a serem estimados e que explique o conjunto de dados é de suma importância neste caso.

Desta forma, existem critérios que auxiliam o pesquisador na seleção dos modelos. Alguns exemplos são: Critério de Informação de Akaike (AIC), Critério Bayesiano de Schwarz (BIC) e um dos mais utilizados o Teste da Razão de Verossimilhança (TRV).

O teste da razão de verossimilhança é utilizado nos casos em que os modelos são aninhados, isto é, um é caso particular do outro. O teste usa a estatística $TRV = 2([l_n(\tilde{\theta}_n) - l_n(\tilde{\theta}_{H_\theta})]$, em que é usado, respectivamente, o máximo da função de verossimilhança do modelo ajustado e do modelo saturado. A estatística $TRV$ possui distribuição assintótica $\chi^2_{\delta,v}$, com $\delta$ o parâmetro de não centralidade e $v$ os graus de liberdade da diferença entre o número de parâmetros dos modelos. Caso, $TRV > \chi^2_{\delta,v}$ rejeita-se a hipótese que o modelo saturado é o de melhor ajuste, ou seja, rejeita-se o ajuste pelo modelo com maior número de parâmetros, e assim a escolha do modelo mais simples é de fato o que melhor se ajusta aos dados.

# 3 APLICAÇÕES

## 3.1 Material

O conjunto de dados utilizados no trabalho contém o tempo de sobrevivência de peixes da espécie “Notropis Dourado, crysoleucas de Notemigonus” que foram obtidos através da realização de experimentos no lago Saint Pierre, Quebec, em 2005 (Laplante et al.). Foram feitas medições das seguintes variáveis: $y_i$, tempo de sobrevivência em anos; $\delta_i$, indicador de censura; $x_{i1}$, tamanho do peixe em cm; $x_{i2}$, profundidade do rio em cm e $x_{i3}$, transparência da água, em que $i = 1, 2, \dots, 106$.

O conjunto de dados do IPEC(Fiocruz), utilizado no estágio um, foi substituído pelos dados de peixes citado acima pelo fato de não se ajustar ao modelo sugerido e também ser muito instável nas estimações dos parâmetros. Desta maneira, as análises descritiva e do modelo serão feitas com base nos dados de peixes da espécie “Notropis Dourado, crysoleucas de Notemigonus”.

## 3.2 Métodos

No presente trabalho será utilizada a teoria de modelos de riscos múltiplos citado na seção (2.3) em que será utilizado o modelo log-logístico múltiplo (2.3.1), além da base teórica da seção (2.4) que engloba a parte de fração de cura. Desta forma, o modelo ajustado aos dados será o modelo Log-Logístico múltiplo com fração de cura. Ao modelo será utilizado o método de verossimilhança restrita como estimação para os parâmetros, e verifica-se através do teste da razão de verossimilhança modificado qual o melhor modelo a ser ajustado ao conjunto de dados.

### 3.2.1 Modelo de riscos múltiplos com fração de cura

Através de uma primeira análise preliminar verifica-se através do gráfico Tempo Total em Teste (TTT) e o gráfico da curva de sobrevivência estimada, que um possível modelo aos dados engloba o uso conjunto das teorias de riscos múltiplos e fração de cura. Desta forma, o modelo escolhido para ajustar os dados é expresso pelas funções de sobrevivência populacional e função de risco populacional a seguir:

$$S_{pop}(t) = (1 - \varphi) + \varphi S(t),$$

onde $S(t) = \prod_{j=1}^{k} S_j(t)$, $j = 1, 2, \ldots, k$ citada na seção (2.5). Logo,

$$S_{pop}(t) = (1 - \varphi) + \varphi\left[1 + t_i^{\omega_1}\exp\{x_i^T\beta_1\right]^{-1}\left[1 + t_i^{\omega_2}\exp\{x_i^T\beta_2\right]^{-1}.$$

Da relação da expressão (7), obtemos a função de risco populacional, expressa por:

$$h_{pop}(t) = -\frac{\partial[logS_{pop}(t)]}{\partial t}.$$

E obtemos pelos cálculos,

$$h_{pop}(t) = \frac{\varphi\{(1 + t_i^{\omega_1}\exp(x_i^T\beta_1)^{-2}\omega_1 t^{(\omega_1-1)}(1 + t_i^{\omega_2}\exp(x_i^T\beta_2)^{-1} - (1 + t_i^{\omega_2}\exp(x_i^T\beta_2)^{-2}(1 + t_i^{\omega_1}\exp\left(x_i^T\beta_1\right)^{-1}\omega_2 t^{(\omega_2-1)}\}\}}{(1-\varphi) + \varphi\{(1 + t_i^{\omega_1}\exp(x_i^T\beta_1)^{-1}(1 + t_i^{\omega_2}\exp(x_i^T\beta_2)^{-1}\}\}}.$$

Assim, utilizando a fórmula em (8), temos que a função densidade de probabilidade do modelo utilizado no trabalho é denotada como:

$$f_{pop}(t) = S_{pop}(t)h_{pop}(t).$$

### 3.2.2 Estimação

Para o modelo Log-Logístico múltiplo e Log-Logístico múltiplo com fração de cura, considerando $k = 2$, foi utilizado a função *constrOptim* do *software* R para a estimação dos parâmetros e máxima verossimilhança. O uso da máxima verossimilhança restrita neste caso se fez necessário pelo fato do modelo incorporar os parâmetros $\omega$ e $\varphi$ que possuem restrição no seu espaço paramétrico.

Neste caso, o processo de máxima verossimilhança restrita (Patterson; Thompson, 1971**)** obtém as estimações maximizando a parte da função de verossimilhança que é invariante, ou seja, a parte do modelo que incorpora os betas.

Desta forma, para o modelo de riscos múltiplos com fração de cura, o logaritmo da função de verossimilhança restrita é dado por:

$$l_R(\theta, \vartheta) = l(\theta) + \vartheta\sum_{j=1}^{q}(u_j^T\vartheta - c_j),$$

em que,

$$l(\theta) = \sum_{i:\delta_i=1} log\big[h(t,\theta)_{pop}\big] + \sum_{i=1}^{n}\big[\log\big(S(t,\theta)_{pop}\big)\big].$$

Sendo que $S(t)_{pop}$ e $h(t)_{pop}$ estão definidas na seção (3.2.1) e o vetor de parâmetros é definido como $\boldsymbol{\theta} = (\omega_1, \omega_2, \beta_1^T, \beta_2^T, \varphi)^T$.

Temos também, o parâmetro de ajuste que é uma constante positiva, $\vartheta > 0$ e $u_j^T\vartheta - c_j \geq 0$ é o conjunto de restrições de inequações lineares para $j = 1, 2, \dots, q$.

### 3.2.3 Teste da razão de verossimilhança modificado

Usualmente, quando o pesquisador trabalha com dados estatísticos e se depara com alguns modelos que se ajustam aos dados, há uma dúvida de qual se acomodaria de forma mais satisfatória. Assim, o teste da razão de verossimilhança indicado na seção (2.5.1) dá suporte e teoria para a escolha do modelo mais parcimonioso.

Contudo, segundo (Maller;Zhou 1996), em sobrevivência, há um problema associado com o teste do parâmetro quando esse está na fronteira do espaço paramétrico. E a solução apresentada inclui somente uma pequena modificação na teoria, que toma a forma da seguinte expressão:

$$P(X \leq x) = \frac{1}{2} + \frac{1}{2}P(\chi_1^2 \leq x), \qquad x \geq 0.$$

Desta forma, com o uso desta nova metodologia, rejeita-se a hipótese nula de que o modelo com maior número de parâmetros ajusta-se melhor ao conjunto de dados caso $TRVm > \chi_1^2$. E assim, o uso de modelos com um número menor de parâmetros é utilizado no experimento.

# 4 RESULTADOS

## 4.1 Análise descritiva

É de suma importância em pesquisas estatísticas uma análise descritiva preliminar dos dados para que se possa observar valores discrepantes, bem como se existe algo anormal com os dados. Em pesquisas na área de sobrevivência, muitas vezes o pesquisador divide para cada variável estudada, dois subgrupos, os que experimentaram o evento de interesse (falha) e as observações censuradas.

Desta maneira, pode-se observar se algumas estatísticas sofrem mudanças bruscas para cada grupo e se isso pode interferir na análise posterior. Assim, segue na Figura 7 o histograma da variável tempo de sobrevivência e nas Figuras 8 a 10 os diagramas de dispersão das covariáveis do estudo. E conjuntamente nas Tabelas de 1 a 4 as estatísticas básicas da variável tempo de sobrevivência e as demais variáveis explicativas do estudo.

**Tempo de sobrevivência**

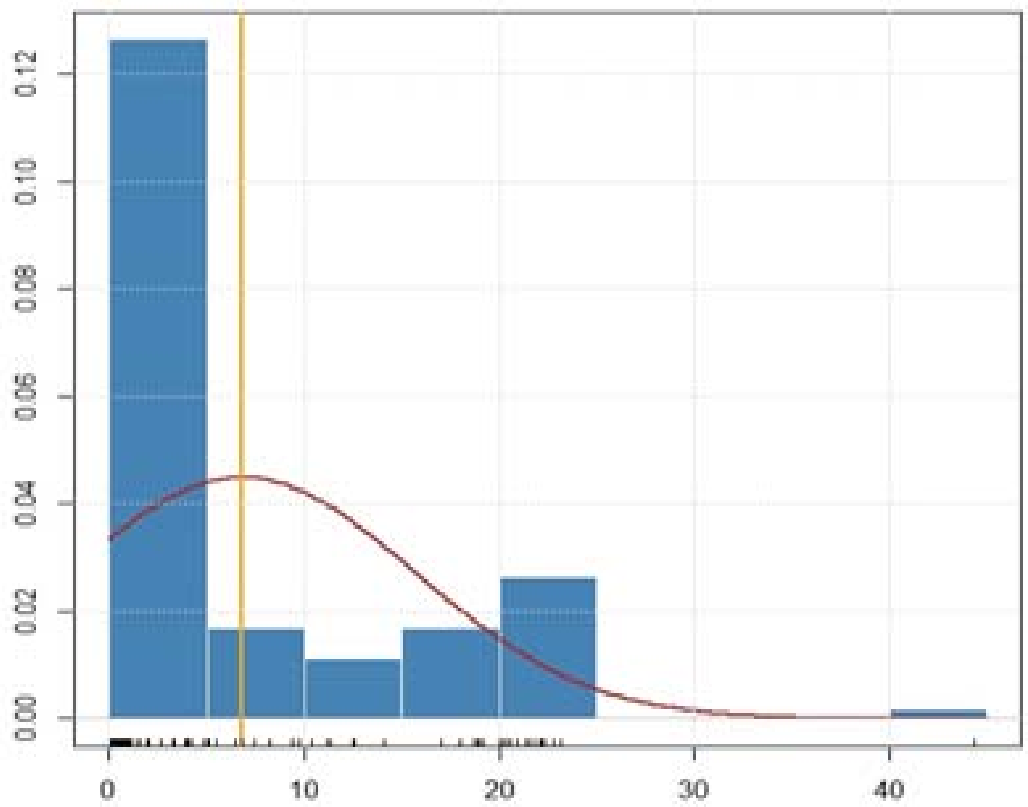


Figura 7 – Histograma da variável tempo de sobrevivência dos peixes.

A Figura 7 representa o histograma do tempo de sobrevivência dos peixes e evidencia claramente um comportamento assimétrico positiva, que condiz com as propriedades de conjunto de dados de sobrevivência.

Tabela 1 – Estatísticas básicas para tempo de sobrevivência dos peixes.

| Estatísticas | Tempo |
|---|---|
| Mínimo | 0,02 |
| Máximo | 44,38 |
| 1º Quartil | 0,385 |
| 3º Quartil | 12,227 |
| Média | 6,808 |
| Mediana | 1,235 |
| Soma | 721,750 |
| LCL Média | 5,100 |
| UCL Média | 8,517 |
| Variância | 78,708 |
| Desvio Padrão | 8,871 |
| Assimetria | 1,350 |

## Profundidade do rio

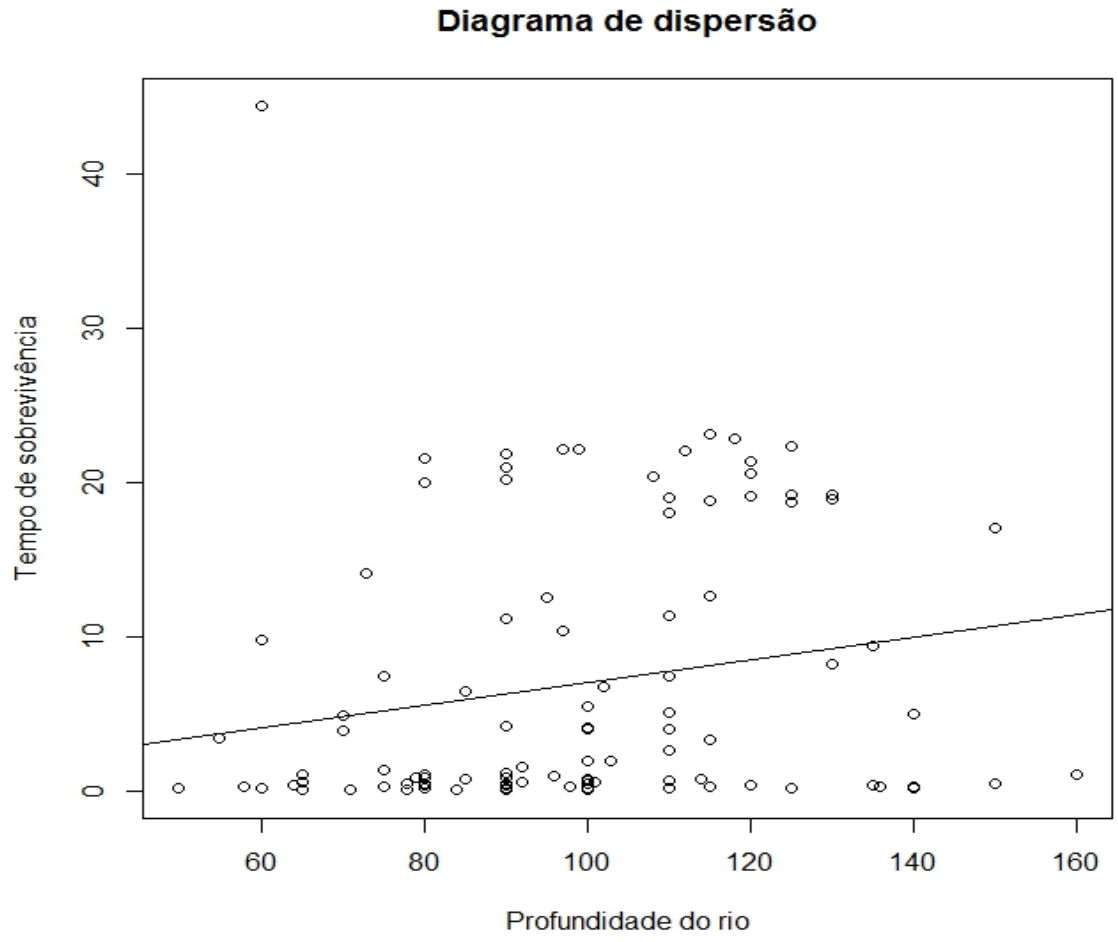


Figura 8 – Diagrama de dispersão da covariável profundidade do rio.

Tabela 2 – Estatísticas básicas para a covariável profundidade do rio.

| Estatísticas | Profundidade do Rio (censura = 0) | Profundidade do Rio (censura = 1) |
|---|---|---|
| Mínimo | 60,0 | 50,0 |
| Máximo | 130,0 | 160,0 |
| 1º Quartil | 98,0 | 80,0 |
| 3º Quartil | 122,50 | 110,0 |
| Média | 108,266 | 96,186 |
| Mediana | 115,0 | 92,0 |
| Soma | 1624,0 | 8753,0 |
| LCL Média | 97,373 | 91,306 |
| UCL Média | 119,159 | 101,067 |
| Variância | 386,923 | 549,153 |
| Desvio Padrão | 19,670 | 23,434 |
| Assimetria | -1,003 | 0,463 |

Através do gráfico de dispersão da Figura 8, e calculado o coeficiente de correlação das variáveis ($r = 0{,}093$), temos que as variáveis tempo de sobrevivência e profundidade do rio possui uma correlação fraca e praticamente nula, sendo assim não interferem uma na outra.

**Comprimento do peixe**

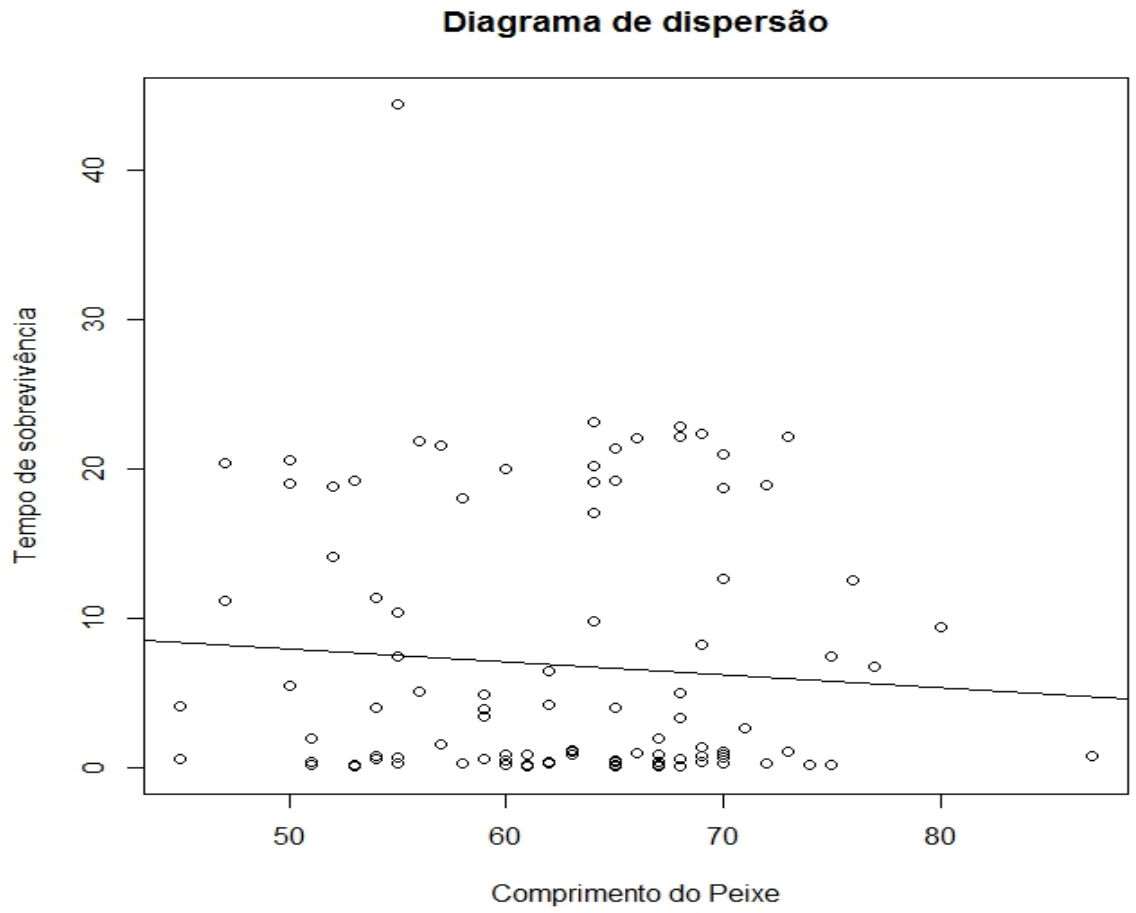


Figura 9 – Diagrama de dispersão da covariável comprimento dos peixes.

Tabela 3 – Estatísticas básicas para a covariável comprimento dos peixes.

| Estatísticas | Tamanho Peixe (censura = 0) | Tamanho do Peixe (censura = 1) |
|---|---|---|
| Mínimo | 47,0 | 45,0 |
| Máximo | 73,0 | 87,0 |
| 1º Quartil | 64,0 | 55,5 |
| 3º Quartil | 68,5 | 68,0 |
| Média | 64,666 | 62,208 |
| Mediana | 65,0 | 62,0 |
| Soma | 970,0 | 5661,0 |
| LCL Média | 60,971 | 60,504 |
| UCL Média | 68,361 | 63,913 |
| Variância | 44,523 | 66,989 |
| Desvio Padrão | 6,672 | 8,184 |
| Assimetria | -1,175 | 0,203 |

Da mesma forma que a covariável anterior, foi feito o gráfico de dispersão e o cálculo da correlação entre as variáveis $(r = -0{,}077)$, e averígua-se que as variáveis possuem uma correlação negativa muito fraca, portanto existe a possibilidade de não serem dependentes uma da outra.

**Transparência da água**

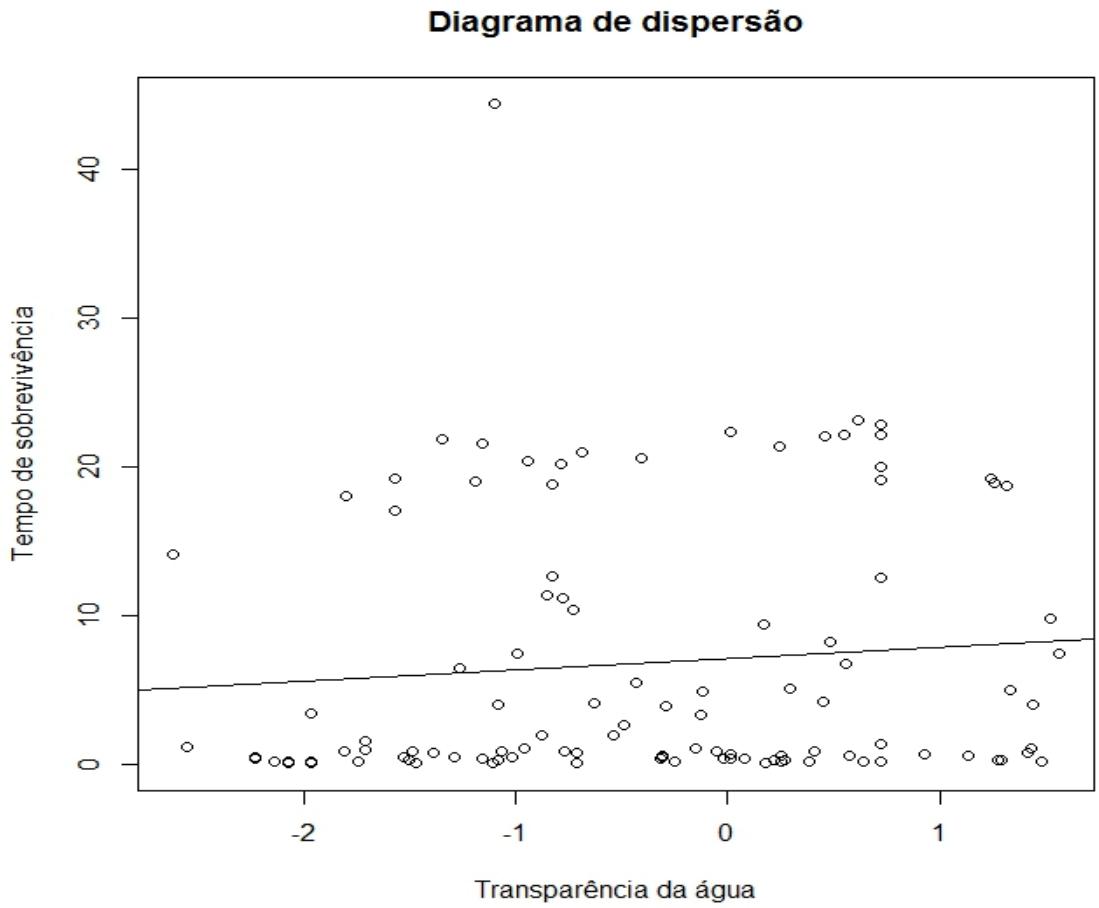


Figura 10 – Diagrama de dispersão da covariável transparência da água.

Tabela 4 – Estatísticas básicas para a covariável transparência da água.

| Estatísticas | Transparência da Água (censura = 0) | Transparência da Água (censura = 1) |
|---|---|---|
| Mínimo | -1,10 | -2,610 |
| Máximo | 1,310 | 1,560 |
| 1º Quartil | 0,125 | -1,320 |
| 3º Quartil | 0,72 | 0,26 |
| Média | 0,382 | -0,508 |
| Mediana | 0,61 | -0,63 |
| Soma | 5730,0 | -45,85 |
| LCL Média | -0,046 | -0,728 |
| UCL Média | 0,810 | -0,279 |
| Variância | 0,599 | 1,163 |
| Desvio Padrão | 0,774 | 1,078 |
| Assimetria | -0,711 | 0,153 |

Dentre as variáveis explicativas, a transparência da água foi a que mais se distanciou para alguns valores como mínimo, média, mediana, desvio padrão, e isso são vistos na Tabela 4. Analisando o gráfico da Figura 10 e o coeficiente de correlação $(r = 0.1926)$, temos que estas variáveis possuem uma correlação mais forte que as demais e isso pode acarretar em uma dependência que influencia nos resultados futuros.

## 4.2 Kaplan Meier

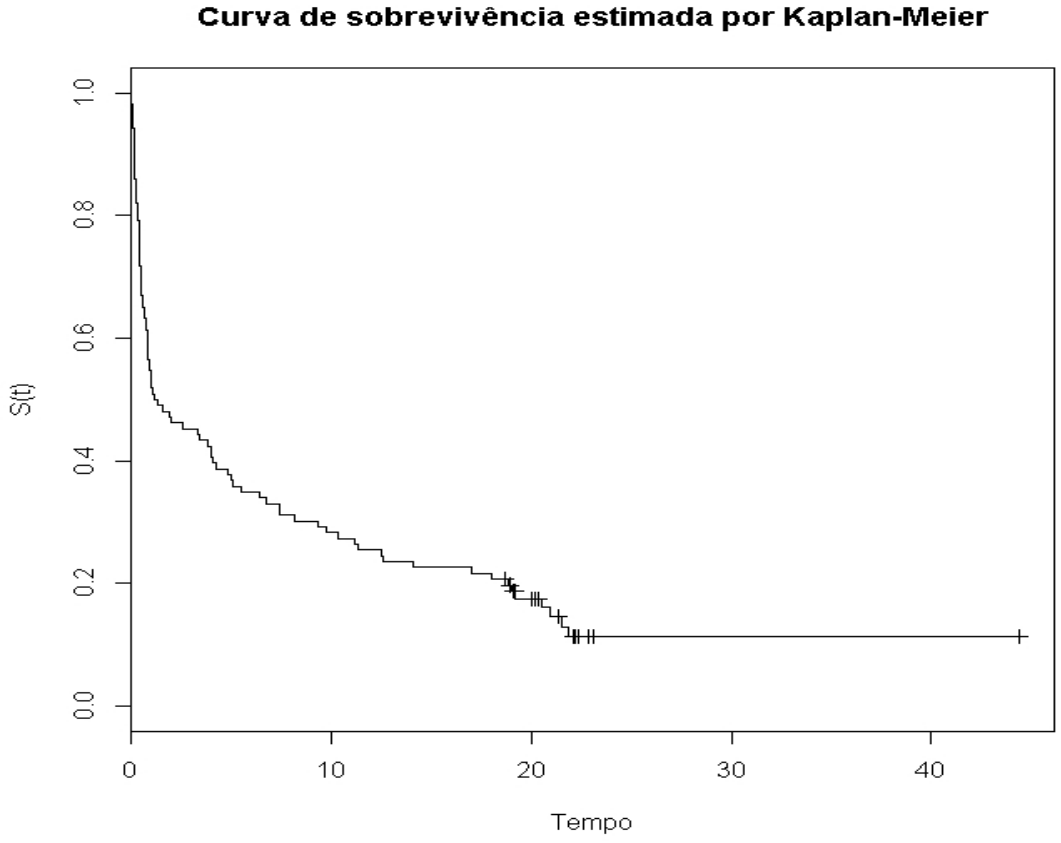


Figura 11 – Curva de sobrevivência estimada por Kaplan-Meier para os dados de peixes.

Na Figura 11, temos a curva de sobrevivência estimada do conjunto de dados de peixes, e pelas características da função de sobrevivência, há um pequeno indicativo de indivíduos curados pelo fato da curva não tender a zero quando o tempo vai para infinito.

### 4.3 Curva Tempo Total em Teste (TTT)

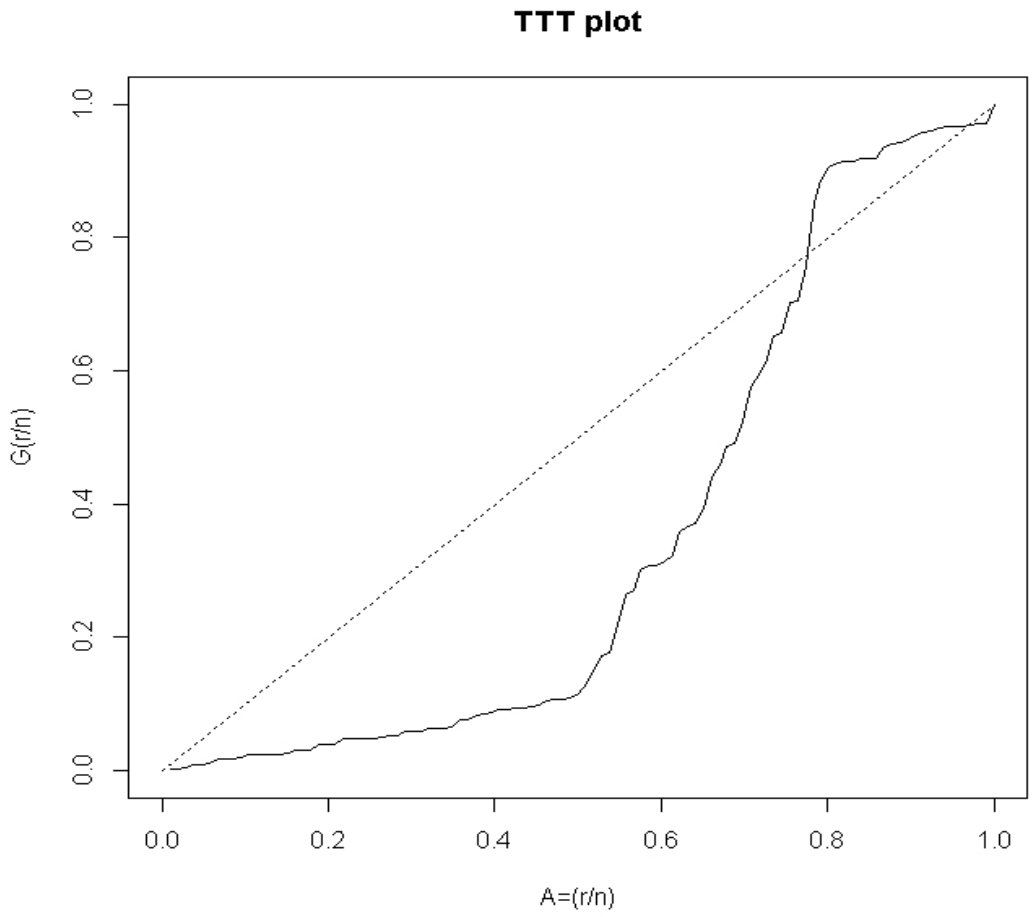


Figura 12 – TTT *plot* para os dados de peixes.

Pela teoria do gráfico TTT *plot,* temos o pressuposto da utilização de um modelo mais complexo para a modelagem do conjunto de dados. Pela análise da Figura 12, temos que um ajuste por um modelo multimodal seria uma ferramenta mais adequada, e com isso o ajuste pelo modelo de riscos múltiplos é válido nesse caso.

### 4.4 Análise do Modelo

Devido às características apresentadas na Figura 11 e 12, conjuntamente

com a análise descritiva do conjunto de dados de peixes, será ajustado o modelo de riscos múltiplo e o modelo de riscos múltiplos com fração de cura para verificar se de fato existem evidências da presença de indivíduos curados, pois há um pequeno indicativo observado na Figura 11. Na Tabela 5, respectivamente, temos as estimativas dos parâmetros dos modelos de riscos proporcionais sem fração de cura e do modelo de riscos proporcionais com fração de cura.

Contudo, o modelo de riscos múltiplos com fração de cura não obteve um ajustamento satisfatório ao conjunto de dados. Uma possível justificativa é a possibilidade de não identificabilidade do modelo somado a uma superfície de verossimilhança que não possui um máximo global, mas sim vários máximos locais. Desta forma, impossibilitou o cálculo da inversa da matriz hessiana, e não foi possível o cálculo do erro padrão deste modelo.

Tabela 5 – Estimativas dos parâmetros dos modelos de riscos proporcionais sem fração de cura e modelo de riscos proporcionais com fração de cura.

| Parâmetro | Estimativa (s/ fração de cura) | Estimativa (c/ fração de cura) |
|---|---|---|
| $\omega_1$ | 0,7475 | 0,4511 |
| $\beta_{01}$ | -0,8133 | -1,5029 |
| $\beta_{11}$ | 0,0357 | 0,0453 |
| $\beta_{21}$ | -0,0216 | -0,0161 |
| $\beta_{31}$ | -0,2601 | 0,0545 |
| $\omega_2$ | 8,8227 | 6,9911 |
| $\beta_{02}$ | -27,9656 | -20,0035 |
| $\beta_{12}$ | -0,2309 | -0,1209 |
| $\beta_{22}$ | 0,0811 | -0,3783 |
| $\beta_{32}$ | -3,6009 | 2,0016 |
| $\varphi$ | - | 0,9999 |

Desta forma, podemos testar a hipótese a um nível de confiança de 5%, de que há total presença de indivíduos suscetíveis a falha, ou seja, testar o parâmetro $\varphi$ do modelo ajustado, e isto é expresso por:

$$H_0\colon\ \varphi_0 = 1.$$

E os respectivos valores de verossimilhança para os modelos de risco múltiplo e risco múltiplo com fração de cura são -221,7404 e -211,9045, desta forma temos:

$$TRVm = 2(\left[l_n(\tilde{\theta}_n) - l_n(\tilde{\theta}_{H_\theta})\right] = 2[((-211{,}9045) - (-221{,}7404)] = \ 19{,}6718.$$

E pelo teste levemente modificado da razão de verossimilhança, temos:

$$\frac{1}{2} + \frac{1}{2}P(\chi_1^2 \leq c_{0{,}95}) = 0{,}95.$$

Como $c_{0.95}$ satisfaz $P(\chi_1^2 \leq c_{0{,}95}) = 0{,}9$ e pela tabela do $\chi_1^2$, obtemos $c_{0.95} = 2{,}71$. Assim, concluímos que $19{,}6718 > 2{,}71$, ou seja, rejeita-se a hipótese nula a um nível de significância de 5% e consideramos que existem fortes evidências que $\varphi_0 \neq 1$.

Desta forma, apesar de não rejeitar a hipótese nula, através da análise do conjunto de dados, o modelo sem cura será ajustado, pois pelo valor do parâmetro $\varphi$ estimado a estimativa de indivíduos não suscetíveis a falha é quase nula e como observado na Figura 11 existe pouco indicativo de indivíduos curados nos dados, então usando o princípio da parcimônia será considerado o modelo de riscos múltiplos.

Com base na análise do teste da razão de verossimilhança modificado, análise das estimativas obtidas e gráficos de suporte, como as Figuras 11 e 12, o modelo a ser utilizado e ajustado ao conjunto de dados será o modelo Log-Logístico múltiplo citado na seção (2.3.1), e em geral, aplica-se a esta teoria considerando $k = 2$ causas que podem causar a falha do indivíduo pela facilidade dos cálculos e aplicabilidade do modelo.

Esse modelo é expresso por:

$$h(t) = \sum_{j=1}^{2} \frac{\omega_j t^{\omega_j - 1} \exp\{x_i^T \beta_j\}}{1 + \ t^{\omega_j} \exp\{x_i^T \beta_j\}} \quad \text{e} \quad S(t) = \prod_{j=1}^{2} \left[1 + \ t_i^{\omega_j} \exp\{x_i^T \beta_j\right]^{-1},$$

em que, $t$ indica o tempo de sobrevivência dos peixes e o vetor de parâmetros é representado por $\boldsymbol{\theta} = (\omega_1, \omega_2, \beta_1^T, \beta_2^T)^T$.

São apresentados na Tabela 6 as estimativas dos parâmetros, estimativas de máxima verossimilhança, erro padrão e p-valor. Apenas as estimativas dos parâmetros

de risco, o intercepto para o risco dois e a covariável profundidade do rio relacionada ao risco um, são significativas ao nível de 5%. As estimativas dos parâmetros de risco são positivas, bem como são significativamente diferentes, e $\widehat{\omega}_1 < \widehat{\omega}_2$. Desta maneira, o modelo bi-Log-Logístico não contraria as suposições e assim, ajusta os dados de peixes.

Tabela 6 – Estimativas dos parâmetros e erro padrão do modelo de riscos múltiplos para os dados de peixes.

| Parâmetro | Estimativa | Erro padrão | p-valor |
|---|---|---|---|
| $\omega_1$ | 0,7475 | 0,0692 | **<0,0001*** |
| $\beta_{01}$ | -0,8133 | 1,5838 | 0,6065 |
| $\beta_{11}$ | 0,0357 | 0,0223 | 0,1092 |
| $\beta_{21}$ | -0,0216 | 0,0082 | **0,0083*** |
| $\beta_{31}$ | -0,2601 | 0,1764 | 0,1411 |
| $\omega_2$ | 8,8227 | 3,1057 | **0,0047*** |
| $\beta_{02}$ | -27,9656 | 11,1828 | **0,0103*** |
| $\beta_{12}$ | -0,2309 | 0,2791 | 0,4850 |
| $\beta_{22}$ | 0,0811 | 0,0657 | 0,2560 |
| $\beta_{32}$ | -3,6009 | 2,7134 | 0,2171 |
| *Estatística* | Valor | | |
| AIC | 463,4808 | | |

(*) significativo a 5%

# 5 CONSIDERAÇÕES FINAIS

Incialmente, o trabalho propôs o ajuste do conjunto de dados através do modelo de riscos múltiplos com fração de cura, e de forma resumida foi apresentado à parte inferencial do modelo. A estimação dos parâmetros foi feita através da função *constrOptim* do *software* R.

Desta forma, os resultados obtidos pelo teste da razão da verossimilhança modificada, indicou um modelo menos complexo para o ajuste e pelo princípio da parcimônia o ajuste do conjunto de dados de peixes da espécie “Notropis Dourado, crysoleucas de Notemigonus”, foi adequado com a teoria desenvolvida para o modelo bi-Log-Logístico.

Uma possibilidade para trabalhos futuros é a utilização do banco de dados do IPEC(Fiocruz), que possivelmente pode ser ajustado por outro modelo de riscos múltiplo, ou de fato, o conjunto de dados deve ser abordado com outra metodologia.

# 6 REFERÊNCIAS


AARSET, M. V., **How to identify bathtub hazard rate.** IEEE Transactions Reliability, 36, 1987.106-108.

BERGER, J.O.; SUN, D., Bayesian analysis for the Poly-Weibull distribution. **Journal of the American Statistical Association**, v. 88, p.1412-1418, 1993.

BERKSON, J.; GAGE, R. P., Survival curve for cancer patients following treatment. **Journal of the American Statistical Association**, v.47, p.501-511, 1952.

BOZDOGAN. H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. **Psychometrika**. v.52, n.3, 345-370, Sep. 1987.

COLOSIMO, E. A.; GIOLO S. R., **Análise de sobrevivência aplicada**, São Paulo: Blucher, 2006. 392 p.

COX, D. R.; OAKES, D., **Analysis of survival data**. New York: Chapman e Hall, 1984. 201 p.

FACHINI, J. B., **Análise de influência local nos modelos de risco múltiplos**. 2006. 77p. Tese (Mestrado em Estatística e Experimetação Agronomica) – Escola Superior de Agricultura “Luis de Queiroz”, Universidades de São Paulo, Piracicaba, 2006.

FACHINI, J. B., **Modelos de regressão com e sem fração de cura para dados bivariados em análise de sobrevivência**. 2011. 140p. Tese (Doutorado em Estatística e Experimetação Agronomica) – Escola Superior de Agricultura “Luis de Queiroz”, Universidades de São Paulo, Piracicaba, 2011.

FACHINI, J. B.; ORTEGA, E. M.; LOUZADA-NETO, F., Influence diagnostics for polyhazard models in the presence of covariates. **Statistical Methods and Applications**, New York, v.17, p.413-433, 2008.

KALBFLEISCH, J. D.; PRENTICE, R. L., The Statistical Analysis of Failure Time Data. 2nd ed. **John Wiley and Sons,** New York, 2002, 439p.

KAPLAN, E. L.; MEIER, P., Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, v.53, p.457-481, 1958.

KLEIN, J. P.; MOESCHBERGER, M. L., **Survival analysis:** techniques for censored and truncated data. New York: Springer Verlag, 1997. 536 p.

LATIMER, N., Survival Analysis for Economic Evaluations Alongside Clinical Trials – Extrapolation with Patient-Level Data, **Relatório Técnico do NICE** (disponível online, acessado em 17 de janeiro de 2013, http://www.nicedsu.org.uk/NICE%20DSU%20TSD%20Survival%20analysis_finalv2.pdf).

LAWLESS, J. F., **Statistical models and methods for lifetime data**, 2nd ed., New York: Wiley, 2003. 439 p.

LOUZADA-NETO, F., Polyhazard models for lifetime data. **Biometrics,** Washington, v.55, p.1281-1285, 1999.

MALLER, R. A.; ZHOU, X., **Survival Analysis with Long-Term Survivors**, 1st ed., John Wiley & Sons, 1996. 304 p.

MAZUCHELI, J.; LOUZADA-NETO, F.; ACHCAR, J. A., Bayesian Inference for polyhazards models in the presence of covariates. **Computational Statistics and Data Analysis**, New York, v.38, p.1-14, 2001.

PATTERSON, H.D.; THOMPSON, R., Recovery of interr-block information when blocks sizes are unequal. **Biometrika**, vol.58, p. 545-554, 1971.

PRENTICE, R. L., KALBFLEISCH, J. D.; PETERSON, A. V. Jr; FLOURNOY, N.; FAREWELL, V. T.; BRESLOW, N. E., The Analysis of Failure Times in the Presence of Competing Risks. **Biometrics**, v.34, p.541-554, 1978.

R DEVELOPMENT CORE TEAM (2003)., R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Viena, Austria. ISBN 3-900051-00-3, URL http://www.r-project.org.

# 7 ANEXOS

```
##Leitura dos dados
dados = read.table("peixe.txt")
t = dados$V1
cens = dados$V2
longf = dados$V4
prof1 = dados$V5
tran_edw = dados$V6


##### Modelo de riscos múltiplos com fração de cura ####



##Função de verossimilhança
log.vero = function(para) {
  beta01 = para[1]
  beta11 = para[2]
  beta21 = para[3]
  beta31 = para[4]
  beta02 = para[5]
  beta12 = para[6]
  beta22 = para[7]
  beta32 = para[8]
  risco1 = para[9]
  risco2 = para[10]
     phi = para[11]

  beta1X = beta01 + beta11*longf + beta21*prof1 + beta31*tran_edw
  beta2X = beta02 + beta12*longf + beta22*prof1 + beta32*tran_edw

  s = ((1 - phi) + phi*(1 + ((t^(risco1))*exp(beta1X))^-1)*((1 + (t^(risco2))*exp(beta2X))^-1))
  h1 = (phi*(((risco1*(t^(risco1 -1))*exp(beta1X)))*((1 + ((t^(risco1))*exp(beta1X)))^-2)*(1 +
((t^(risco2))*exp(beta2X)))^-1))
  h2 = (phi*(((risco2*(t^(risco2 -1))*exp(beta2X)))*((1 + ((t^(risco2))*exp(beta2X)))^-2)*(1 +
((t^(risco1))*exp(beta1X)))^-1))

  logL = (cens*(log(h1 + h2)/s) - log(s))
  return(-sum(logL))
}

##Gradiente
grad = function(para) {
beta01 = para[1]
  beta11 = para[2]
  beta21 = para[3]
  beta31 = para[4]
  beta02 = para[5]
  beta12 = para[6]
```

```
  beta22 = para[7]
  beta32 = para[8]
  risco1 = para[9]
  risco2 = para[10]
     phi = para[11]


dbeta01 = D(f, "beta01")
dbeta11 = D(f, "beta11")
dbeta21 = D(f, "beta21")
dbeta31 = D(f, "beta31")
dbeta02 = D(f, "beta02")
dbeta12 = D(f, "beta12")
dbeta22 = D(f, "beta22")
dbeta32 = D(f, "beta32")
drisco1 = D(f, "risco1")
drisco2 = D(f, "risco2")
dphi    = D(f, "phi")


return(c(-dbeta01,-dbeta11,-dbeta21,-dbeta31,-dbeta02,-dbeta12,-dbeta22,-dbeta32,-
drisco1,-drisco2,-dphi))
}
```

**##Estimação dos parâmetros**

```
chute = c(0.1,0.01,0.01,0.2,-20,0.1,0.01,2,0.7,7,0.959)
estima = constrOptim(chute, log.vero, grad, method = "BFGS",
        ui=rbind(c(0,0,0,0,0,0,0,0,1,0,0), c(0,0,0,0,0,0,0,0,0,1,0),
             c(0,0,0,0,0,0,0,0,0,0,1), c(0,0,0,0,0,0,0,0,0,0,-1)) , ci=c(0,0,0,-1),hessian = TRUE)
```

**##Cálculo do erro padrão e p-valor**

```
estima$par
estima$value
veroMe<-estima$value
veroMe
paraMe<-estima$par
paraMe
hessiMe<-estima$hessian
invMe<-solve(hessiMe)
varianciaMe<-diag(invMe)
eppMe<-sqrt(varianciaMe)
eppMe
zMe=paraMe/eppMe
pvalorMe<-2*(1-pnorm(abs(zMe)))
pvalorMe
```

#### **Modelo de riscos múltiplos sem fração de cura** ####

```
##Função de verossimilhança
log.vero = function(para) {
  beta01 = para[1]
  beta11 = para[2]
  beta21 = para[3]
  beta31 = para[4]
  beta02 = para[5]
  beta12 = para[6]
  beta22 = para[7]
  beta32 = para[8]
  risco1 = para[9]
  risco2 = para[10]

  beta1X = beta01 + beta11*longf + beta21*prof1 + beta31*tran_edw
  beta2X = beta02 + beta12*longf + beta22*prof1 + beta32*tran_edw

  a = (risco1*(t^(risco1 -1))*exp(beta1X))/(1 + (t^(risco1))*exp(beta1X))
  b = (risco2*(t^(risco2 -1))*exp(beta2X))/(1 + (t^(risco2))*exp(beta2X))
  c = log(1 + (t^(risco1))*exp(beta1X))
  d = log(1 + (t^(risco2))*exp(beta2X))

  logL = (cens*(log(a + b)) - (c + d))
  return(-sum(logL))
}

##Estimação dos parâmetros
chute = c(-0.1,0.01,0.01,0.2,-20,-0.1,0.01,2,0.7,7)
estima = constrOptim(chute, log.vero, NULL,
       ui=rbind(c(0,0,0,0,0,0,0,0,1,0), c(0,0,0,0,0,0,0,0,0,1)), ci=c(0,0))
estima

est<-optim(c(-0.1,0.01,0.01,0.2,-20,-0.1,0.01,2,0.7,7), log.vero, NULL, method = "BFGS",
hessian = TRUE)

##Cálculo do erro padrão e p-valor
paraMe<-est$par
hessiMe<-est$hessian
invMe<-solve(hessiMe)
varianciaMe<-diag(invMe)
eppMe<-sqrt(varianciaMe)
eppMe
zMe=paraMe/eppMe
pvalorMe<-2*(1-pnorm(abs(zMe)))
pvalorMe
```