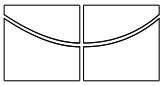

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Dissertação de Mestrado

Modelo de Regressão Log-Beta Burr III para dados Grupados

por

Vanessa Silva Resende

Orientador: Antonio Eduardo Gomes

Co-orientador: Juliana Betini Fachini Gomes

Brasília
2017

Vanessa Silva Resende

Modelo de Regressão Log-Beta Burr III para dados Grupados

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Brasília
2017

Termo de Aprovação

Vanessa Silva Resende

Modelo de Regressão Log-Beta Burr III para dados Grupados

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Data da defesa: 29 de Maio de 2017

Orientador:

Prof. Dr. Antonio Eduardo Gomes - Orientador
Departamento de Estatística, UnB

Comissão Examinadora:

Prof. Dr. Eduardo Yoshio Nakano - Membro da Banca
Departamento de Estatística, UnB

Prof. Dra. Elizabeth Mie Hashimoto - Membro da Banca
Departamento Acadêmico de Matemática, UTFPR

Brasília, Maio de 2017

Ficha Catalográfica

RESENDE, VANESSA SILVA
Modelo de Regressão Log-Beta Burr III para dados grupados, (UnB - IE, Mestre em Estatística, 2017).

Dissertação de Mestrado - Universidade de Brasília. Departamento de Estatística - Instituto de Ciências Exatas.

1. Análise de Sobrevivência 2. Censura intervalar 3. Dados de sobrevivência grupados 4. Distribuição Log-Beta Burr III 5. Modelo de regressão

Aos meus pais,
Everaldo e Francisca pelo apoio e por tudo que me ensinaram.

Aos meus irmãos,
Felipe e Monique pelo carinho e apoio.

Agradecimentos

- Primeiramente agradeço a Deus por tudo, pela minha família, minha saúde, e por ter me dado força para realização deste trabalho.
- Aos professores Dra. Juliana Betini Fachini Gomes e Dr. Antonio Eduardo Gomes, pela orientação, sugestões e ensinamentos neste trabalho.
- À Universidade de Brasília, em especial ao Departamento de Estatística, pela oportunidade de concretização deste trabalho.
- A todos os professores do Departamento de Ciências Exatas, pelos ensinamentos passados.
- A toda minha família, em especial, aos meus pais, ao meu irmão Felipe e minha irmã Monique, pela paciência, incentivo e apoio incondicional.
- Ao Felipe Pereira, pelo companheirismo, amor, paciência, incentivo.
- A todos que de alguma forma contribuíram para a realização deste trabalho.

Sumário

Lista de Abreviaturas

f(t) - Função densidade de probabilidade da variável aleatória tempo (T).
F(t) - Função distribuição de probabilidade da variável aleatória tempo (T).
S(t) - Função de sobrevivência da variável aleatória tempo (T).
h(t) - Função taxa de falha ou função de risco da variável aleatória tempo (T).
H(t) - Função taxa de falha acumulada da variável aleatória tempo (T).
fda - Função de distribuição acumulada.
fdp - Função de densidade de probabilidade.
BIII - Distribuição Burr III
LBIII - Distribuição Log-Burr III
BBIII - Distribuição Beta Burr III
LBBIII - Distribuição Log-Beta Burr III

Lista de Figuras

Lista de Tabelas

Resumo

A censura intervalar ocorre quando não se conhece o tempo de sobrevivência exato, sabe-se somente que ocorreu em um intervalo de tempo. De outra forma, quando todos os indivíduos são avaliados nos mesmos intervalos de tempo, ocasionando muitos empates, tem-se dados grupados. Portanto, os dados de sobrevivência grupados são casos particulares da censura intervalar. Neste trabalho, foi proposto um modelo de regressão para dados de sobrevivência grupados utilizando a distribuição log-beta Burr III, cuja principal característica é a possibilidade da função taxa de falha assumir diferentes formas (crescente, decrescente, unimodal). Posteriormente, foi proposto um modelo de regressão para dados grupados utilizando a distribuição log-Burr III, que um caso particular da distribuição anteriormente citada. Os parâmetros dos modelos foram estimados utilizando os métodos de máxima verossimilhança e bootstrap. Por fim, utilizou-se um conjunto de dados reais para exemplificar a aplicação dos modelos propostos.

Palavras Chave: *Análise de sobrevivência; Censura intervalar; Dados de sobrevivência grupados; Distribuição log-beta Burr III; Modelo de regressão.*

Abstract

Interval-censored data occur when the exact survival times are known only to be in a time interval. When all individuals are evaluated in the same time intervals, causing many ties, we have grouped data. Thus, grouped survival data are particular cases of interval censored data. In this work, a regression model for grouped survival data was developed using log-beta Burr III distribution, which main characteristic is the possibility of the hazard function assuming different forms (increasing, decreasing, unimodal). Also, a regression model for grouped data was developed using log-Burr III distribution, which is a particular case of the aforementioned distribution. The parameters of the models were estimated by maximum likelihood method and bootstrap. Finally, a real data set is used to exemplify the application of the proposed models.

key words: *Survival analysis; Interval censoring; Grouped survival data; log-beta Burr III distribution; Regression model.*

Capítulo 1

Introdução

A análise de sobrevivência é uma área muito importante da estatística, e está presente em vários campos de conhecimento, como na biologia, engenharia, economia, epidemiologia e principalmente, na medicina. Nesse tipo de estudo, a variável resposta é, em geral, o tempo transcorrido até a ocorrência do evento de interesse. Por exemplo, em estudos da medicina, esse tempo de falha pode ser o tempo até a morte de um paciente; em estudos na engenharia, o tempo até a quebra de um equipamento, entre outros.

Além disso, os dados de sobrevivência caracterizam-se por apresentar censuras, que é variável resposta observada parcialmente. Essas observações podem ocorrer por diversos motivos, por exemplo, quando o experimento termina e ainda não ocorreu o evento de interesse, quando o acompanhamento do elemento é interrompido por alguma razão ou, ainda, quando não se sabe o tempo exato de ocorrência do evento de interesse em um determinado intervalo de tempo. Os mecanismos de censura são conhecidos como: censura à direita, à esquerda e a censura intervalar.

Quando não se sabe o tempo exato de ocorrência do evento de interesse e sabe-se apenas que a falha ocorreu em um intervalo de tempo, tem-se o caso da censura intervalar. Essa ocorre, por exemplo, quando o acompanhamento de um equipamento é dado por visitas periódicas, sabe-se que a falha ocorreu em um intervalo de tempo, mas não seu tempo exato. Por outro lado, os dados grupados ocorrem quando todas as unidades amostrais são avaliadas nos mesmos instantes. Vários autores já propuseram diversas abordagens para modelar esses tipos de dados. Prentice e Gloeckler (1978) desenvolveram um modelo de riscos proporcionais modificado para acomodar dados de sobrevivência grupados. Allison (1982) desenvolveu métodos para análise de dados de sobrevivência grupados. Hashimoto et al. (2012) realizaram um estudo sobre o modelo de regressão Log-Burr XII para dados grupados. Diante disso, neste trabalho

são estudados dados grupados, que é um caso particular de censura intervalar.

Ainda, em muitos estudos de sobrevivência a presença de covariáveis podem estar relacionadas com o tempo de sobrevivência, por esse motivo deve-se incorporá-las na análise estatística. A fim de acomodar o efeito dessas covariáveis é utilizado um modelo de regressão para dados censurados.

Assim, o principal objetivo deste estudo é propor um modelo de regressão para dados grupados considerando que o tempo de sobrevivência tem distribuição log-beta Burr III, baseando-se no trabalho de Gomes et al. (2013) e utilizando a metodologia de regressão para dados grupados similar a apresentada por Hashimoto et al. (2012). Além disso, os procedimentos para obter as estimativas dos parâmetros do modelo também são vistas neste trabalho. A grande vantagem desse modelo é a flexibilidade de sua função de risco. Os demais objetivos são: realizar uma revisão teórica das distribuições Burr III, log-Burr III, beta Burr III e log-beta Burr III e aplicar a um banco de dados o modelo proposto.

Mediante ao exposto, o presente trabalho está organizado da seguinte forma: no Capítulo 2 é feita uma revisão bibliográfica de conceitos importantes na análise de sobrevivência e sobre as distribuições de probabilidade: Burr III, log-Burr III, beta Burr III e log-beta Burr III. No Capítulo 3 é proposto o modelo de regressão log-beta Burr III para dados grupados e é apresentada sua função de máxima verossimilhança. No Capítulo 4 é feita uma análise do banco de dados de suplementação de vitamina A (BARRETO et al., 1994), utilizando o modelo log-beta Burr III, e o modelo log-Burr III para dados grupados. Por fim, no Capítulo 5 são apresentadas as principais conclusões e sugestões para trabalhos futuros.

Capítulo 2

Revisão de literatura

Neste capítulo é apresentada uma revisão sobre a análise de sobrevivência e algumas distribuições de probabilidade, formando dessa forma um embasamento teórico para a compreensão do trabalho.

2.1 Notação e conceitos básicos

Em análise de sobrevivência, a variável resposta é considerada o tempo até a ocorrência de um evento de interesse ou censura (COLOSIMO E GIOLO, 2006). O tempo até a ocorrência de um evento de interesse é chamado tempo de falha. Uma das maiores áreas de atuação da análise sobrevivência refere-se a estudos médicos, assim, o tempo de falha pode ser, por exemplo, o tempo até a morte do paciente ou até a cura do indivíduo. Em outras áreas de estudo, como na engenharia, o tempo de falha pode ser considerado o tempo até a ocorrência da falha de determinado equipamento.

A presença da censura, ou seja, de uma observação parcial da resposta é a principal característica dos dados de análise de sobrevivência. As censuras referem-se a situações em que, por algum motivo, o acompanhamento do paciente é interrompido. Por exemplo, morte do indivíduo por uma causa diferente da estudada ou fim do estudo antes do paciente falhar.

Além do mais, dependendo da situação, há alguns mecanismos de censura, por exemplo, a censura à direita, censura à esquerda, ou intervalar. Na censura à direita o tempo de ocorrência do evento de interesse está à direita do tempo de registro. À esquerda ocorre quando o tempo registrado é maior que o tempo de falha e a intervalar ocorre quando não se sabe o tempo exato da falha, sabe-se somente que o evento de interesse ocorreu em um intervalo de tempo, isto é, $T \in (U, V]$. Dessa forma, neste trabalho é adotado o mecanismo de dados grupados, que é um caso particular da

censura intervalar quando as unidades amostrais são avaliadas nos mesmos intervalos de tempo. Por diversas vezes esse tipo de dado é identificado por um número excessivo de empates (COLOSIMO E GIOLO, 2006).

Como visto, o enfoque deste trabalho é o estudo de modelos de regressão para dados de sobrevivência grupados. Na literatura pode-se encontrar trabalhos que propõem esses modelos, como em Thompson (1977), em Colosimo et al. (2000) e Hashimoto et al. (2012).

A censura à direita pode, ainda, ser classificada como do tipo I, do tipo II, e do tipo aleatório. O primeiro mecanismo ocorre quando o estudo termina após um tempo pré-estabelecido e registra indivíduos que não apresentaram o evento de interesse. A censura tipo II ocorre quando o estudo é finalizado após a ocorrência do evento de interesse em um número pré-estabelecido de indivíduos. E a censura do tipo aleatória, que pode ocorrer quando um indivíduo é retirado do estudo sem a ocorrência da falha, ou, se por exemplo, o individuo morrer por uma causa diferente da estudada.

Colosimo e Giolo (2006) apresentam uma representação do mecanismo de censura do tipo aleatória, considerando duas variáveis aleatórias. Seja T uma variável aleatória não-negativa representando o tempo de falha do indivíduo e C, uma variável aleatória independente de T, que representa o tempo de censura. Tem-se, então:

$$t = \min\{T, C\}$$

e

$$\delta = \begin{cases} 1 & se\ T \leq C, \\ 0 & se\ T > C. \end{cases}$$

Em que $\delta = 1$ indica falha e $\delta = 0$ indica censura.

2.1.1 Censura intervalar e dados grupados

Dados com censura intervalar estão naturalmente presentes em estudos médicos e epidemiológicos. Visto que, em muitos estudos não se sabe o tempo exato de falha, só se sabe que ocorreu em um intervalo de tempo. Por exemplo, em um estudo clínico o tempo T de falha é definido como sendo o tempo até o aparecimento de um tumor. Esse tempo T não é exatamente conhecido, dado que o aparecimento do tumor só é identificado em consultas feitas periodicamente. Neste caso, o tempo de falha T ocorreu em um intervalo $(U, V]$, onde U é a última avaliação com resultado negativo e V a primeira avaliação com resultado positivo para o tumor.

Dessa forma, as censuras à direita, à esquerda e os tempos exatos de falha são casos particulares da censura intervalar (LINDSEY E RYAN, 1998), ou seja, quando

$V = \infty$ tem-se o caso da censura à direita, quando $U = 0$ tem-se a censura à esquerda e quando $U = V$ tem-se o tempo exato de falha.

Visto que quando no estudo todas as unidades amostrais são avaliadas nos mesmos instantes tem-se o caso de dados grupados, que são um caso particular de dados de sobrevivência intervalar. Outra característica desses dados é um número excessivo de empates, ou seja, os tempos de vida aparecem repetidas vezes. Frequentemente, encontram-se dados grupados em estudos de medidas repetidas longitudinais, como no estudo envolvendo mangueiras (CHALITA, 1997). Nesse estudo, todas as mangueiras foram avaliadas nas 12 visitas durante o experimento. Houve, então, muitos tempos de falha no mesmo intervalo de tempo, visto que havia poucos tempos de observação, caracterizando-se, assim, dados grupados.

Se ao considerar um conjunto de dados, este apresentar censuras e um número excessivo de empates, os tempos observados $t = \min\{T, C\}$ são grupados em um determinado número de intervalos, a fim de, por exemplo, eliminar os empates. O número de intervalos no qual os tempos observados são grupados é impreciso. Bolfarine et al. (1991) afirmam que a metodologia de dados grupados tem como base métodos de tabela de vida, e portanto, o número de intervalos é arbitrário.

Segundo Hashimoto (2008), os intervalos são construídos de tal forma que o eixo do tempo é dividido em k intervalos definidos por pontos de corte $a_1, ..., a_k$, dessa forma o j-ésimo intervalo é denotado pela expressão $I_j = [a_j, a_{j+1})$, para $j = 1, ...k$.

Sendo assim, o enfoque deste trabalho é propor um modelo de regressão log-beta Burr III para dados grupados presente na Seção 3.2.

2.1.2 Distribuições do tempo de sobrevivência

Seja T uma variável aleatória não-negativa, usualmente contínua, representando o tempo de falha ou de sobrevivência de um indivíduo. Esse tempo T pode ser descrito por sua função de densidade, função de distribuição, função de sobrevivência e função de risco.

Então a função densidade de probabilidade (fdp), $f(t)$, é definida como o limite da probabilidade de um indivíduo vir a falhar em um intervalo de tempo $[t, t + \Delta t)$ por unidade de tempo ou de Δ (comprimento do intervalo) e é dada por:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}. \tag{2.1}$$

Dessa forma, a probabilidade do indivíduo experimentar o evento de interesse até o tempo t é obtida pela função de distribuição, $F(t)$ e é expressa por:

$$F(t) = P(T \leq t) = \int_0^t f(u)du.$$

A função de sobrevivência é uma das principais funções probabilísticas utilizadas em estudos com dados de sobrevivência. Essa é definida como sendo a probabilidade de um indivíduo sobreviver, ou seja, não falhar, até um certo tempo t. Ao se obter a função distribuição pode-se utilizar a seguinte relação a fim de obter a função de sobrevivência:

$$F(t) = 1 - S(t). \tag{2.2}$$

Assim, a função de sobrevivência, $S(t)$ é representada por:

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

Segundo Lawless (2003), $S(t)$ é uma função monótona decrescente e contínua e tem as seguintes propriedades: $\lim_{t \to\ 0} S(t) = 1$ e $\lim_{t \to\ \infty} S(t) = 0$.

Além disso, em sobrevivência, outra função muito importante é a função de risco ou taxa de falha $h(t)$. Ela é definida como o limite da probabilidade de um indivíduo experimentar o evento de interesse no intervalo $[t, \Delta t)$, assumindo que esse mesmo indivíduo sobreviveu até o tempo t, é expressa por:

$$h(t) = \lim_{\Delta t \to\ 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{2.3}$$

Logo, a taxa de falha também pode ser expressa em função das funções de densidade e de sobrevivência:

$$h(t) = \frac{f(t)}{S(t)}. \tag{2.4}$$

Ainda, ao considerar a função taxa de falha definida na equação (2.3), pode-se definir uma outra função útil em análise de sobrevivência: a função taxa de falha acumulada. Ela é definida como:

$$H(t) = \int_0^t h(u)du. \tag{2.5}$$

Portanto, utilizando-se a função taxa de falha acumulada definida na equação (2.5), pode-se obter a função de sobrevivência :

$$S(t) = \exp(-H(t)).$$

Então, para se investigar o possível comportamento da função de risco de determinada variável resposta ou banco de dados, pode-se construir o gráfico de tempo total em teste (curva TTT), proposto por Aarset (1987). Essa curva TTT é obtida construindo-se um gráfico de:

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

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

Diante disso, a curva TTT pode assumir as diferentes formas:

- Reta diagonal (A) $\Longrightarrow$ Função de risco é constante.
- Curva convexa (B) ou côncava (C) $\Longrightarrow$ Função de risco é monotonicamente decrescente ou crescente, respectivamente.
- Curva convexa e depois côncava (D) $\Longrightarrow$ Função de risco tem forma de banheira ou **U**.
- Curva côncava e depois convexa (E) $\Longrightarrow$ Função de risco é unimodal.

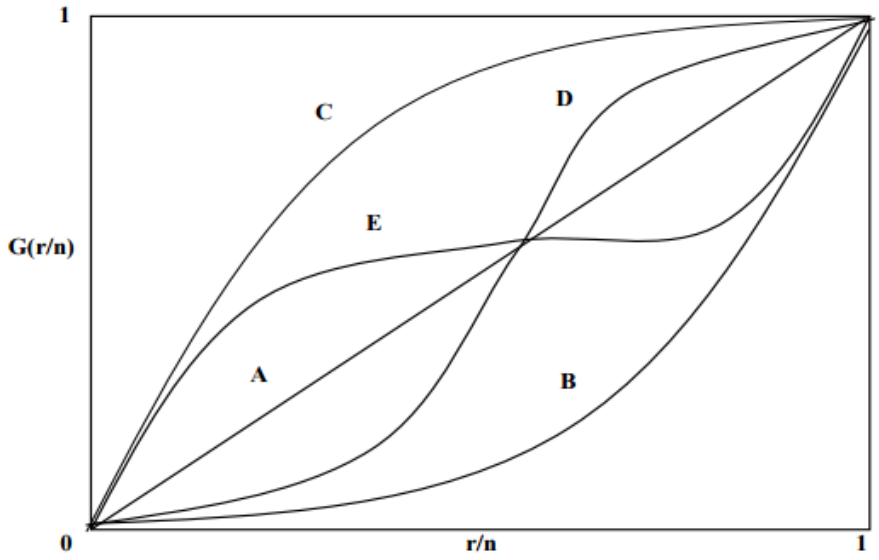

Figura 2.1 – Formas que a curva TTT pode assumir

Segundo Bergman e Klefsjõ (1998), a curva TTT apresentada anteriormente pode ser interpretada em função da função de distribuição empírica, F_n. Essa curva considera a amostra completa, ou seja, os tempos de censura e de falha. Entretanto, se há uma amostra censurada é natural mudar a função distribuição empírica para outro estimador, F_n^c, da função distribuição verdadeira. Assim é construído o gráfico $F_n^c(t(i))$ por $G(r/n)^c$ para os tempos de falha não censurados, $t(i)$, e conecta-se esses pontos. Há duas possibilidades de se fazer isso, a primeira é utilizar o

PiecewiseExponentialEstimator (PEXE) desenvolvido por Kitchin (1980) e discutido por Kim e Proschan (1991). A segunda possibilidade é utilizar o estimador de Kaplan-Meier, esta ideia é indicada no trabalho de Bergman e Klefsjõ (1984).

2.1.3 Estimador de Kaplan-Meier

O estimador de Kaplan-Meier foi proposto por Kaplan e Meier (1958), também chamado de estimador limite-produto é um estimador não-paramétrico desenvolvido para estimar a função de sobrevivência. Na sua construção, esse estimador considera a quantidade de intervalos de tempo quantos forem o número de falhas distintas. Os tempos de falha são considerados os limites dos intervalos de tempo. Considerando-se que:

- $t_1 < t_2 < \cdots < t_k$, os k tempos distintos e ordenados de falha,
- d_j o número de falhas em t_j, $j = 1, ..., k$, e
- n_j o número de indivíduos sob risco em t_j, ou seja, que não falharam e não foram censurados até o tempo imediatamente anterior a t_j.

Assim, tem-se que o estimador de Kaplan-Meier é definido como:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right). \tag{2.6}$$

2.1.4 Função de verossimilhança em análise de sobrevivência

Em análise de sobrevivência, considera-se as censuras na análise estatística. Dessa forma, deve-se incorporar as censuras na função de verossimilhança. Seja uma amostra aleatória observada $(t_1, \delta_1), (t_2, \delta_2), ..., (t_n, \delta_n)$, em que t_i é o tempo de sobrevivência ou censura e δ_i é o indicador de censura a seguir:

$$\delta_i = \begin{cases} 1 & \text{se } t_i \text{ é tempo de falha,} \\ 0 & \text{se } t_i \text{ é tempo de censura.} \end{cases}$$

A função de verossimilhança considerando todos os tipos de censura à direita e considerando que as mesmas são não informativas (não carregam informação sobre os parâmetros) pode ser expressa em termos da função de sobrevivência $S(t)$ e da densidade $f(t)$ ou da função de risco $h(t)$ e da função de sobrevivência $S(t)$. Essas

funções serão definidas na Seção 2.2. Dessa forma, a função de verossimilhança para censura à direita é dada por:

$$L(\theta) \propto \prod_{i=1}^{n} \left[f(t_i,\theta)\right]^{\delta_i} \left[S(t_i,\theta)\right]^{1-\delta_i} = \prod_{i=1}^{n} \left[h(t_i,\theta)\right]^{\delta_i} \left[S(t_i,\theta)\right], \tag{2.7}$$

em que $\boldsymbol{\theta}$ é o vetor de parâmetros desconhecidos, $S(t_i)$, $f(t_i)$ e $h(t_i)$ são as funções de sobrevivência, de densidade de probabilidade e função de risco, respectivamente, para cada variável aleatória T_i. Assim, tem-se que a contribuição de cada observação censurada é sua função de sobrevivência e de cada observação não-censurada é a função densidade.

2.1.5 Método bootstrap

Efron (1979) propôs o método de reamostragem bootstrap não-paramétrico. Esse método é importante por não só avaliar as estimativas dos parâmetros, como também, obter boas estimativas dos erros padrões da distribuição gerada pelas estimativas do parâmetros nas iterações de reamostragem (LEPAGE E BILLARD, 1992). O método de bootstrap trata uma amostra observada como se essa representasse a população. Assim, da informação obtida de tal amostra observada, B amostras bootstrap de tamanho similar ao da amostra observada são geradas, da qual é possível obter a estimativa de diversas características da população, como por exemplo, média e variância.

De acordo com o método bootstrap, a função distribuição F pode ser estimada pela distribuição empírica $\hat{F}$. Seja $\mathbf{T} = (T_1, ..., T_n)$ uma amostra aleatória observada e $\hat{F}$ a distribuição empírica de $\mathbf{T}$. Portanto, uma amostra bootstrap $\mathbf{T}^*$ é construída por amostragem com reposição de n elementos da amostra $\mathbf{T}$. Para as B amostras bootstrap geradas, $\mathbf{T_1^*}, ..., \mathbf{T_B^*}$, a replicação bootstrap do parâmetro de interesse para a b-ésima amostra é dada por $\hat{\theta}_\mathbf{b}^* = s(T_b^*)$, esse é o valor de $\hat{\theta}$ para a amostra T_b^*, $b = 1, ..., B$.

O estimador bootstrap do erro-padrão é o desvio-padrão dessas amostras bootstrap (EFRON E TIBSHIRANI, 1993), esse é denotado por $\hat{EP}_B$ e é expresso pela equação:

$$\hat{EP}_B = \left[\frac{1}{(B-1)} \sum_{b=1}^{B} (\hat{\theta}_\mathbf{b}^* - \hat{\theta}_\mathbf{B})^2\right]^{1/2}, \tag{2.8}$$

em que $\hat{\theta}_\mathbf{B} = \dfrac{1}{B} \displaystyle\sum_{b=1}^{B} \hat{\theta}_\mathbf{b}^*$.

Baseando-se no trabalho de Efron e Tibshirani (1993), é apresentado os principais passos para o precedimento bootstrap:

Passo 1: Construa uma distribuição de probabilidade empírica, F_n, da amostra colocando uma massa de probabilidade de $1/n$ para cada ponto $x_1, x_2, ..., x_n$ da amostra. Essa é a função de distribuição empírica da amostra, que é a estimativa de máxima verossimilhança não-paramétrica da distribuição da população, F.

Passo 2: Da função distribuição empírica, F_n, "obtenha" uma amostra aleatória de tamanho n com reposição. Essa é a reamostra.

Passo 3: Calcule a estatística de interesse, T_n, para essa reamostra, produzindo T_n^*.

Passo 4: Repita os Passos 2 e 3 B vezes, em que B é um número grande, a fim de criar B reamostras. O tamanho ideal de B depende dos testes a serem realizados nos dados. Tipicamente, B é pelo menos igual a 1000 quando se deseja a estimativa do intervalo de confiança para T_n, e para obter uma boa estimativa do erro-padrão são necessárias entre 25 e 200 reamostras.

Passo 5: Construa o histograma da frequência relativa dos B números de T_n^*'s atribuindo probabilidade $1/B$ em cada ponto, T_n^{*1}, T_n^{*2},...,T_n^{*B}. A distribuição obtida é a estimativa bootstrap da distribuição amostral de T_n. Essa distribuição pode agora ser usada para fazer inferências sobre o parâmetro θ, que deve ser estimado por T_n.

Na Figura 2.2 encontra-se uma ilustração explicando o método bootstrap, a fim de facilitar o entendimento do método.

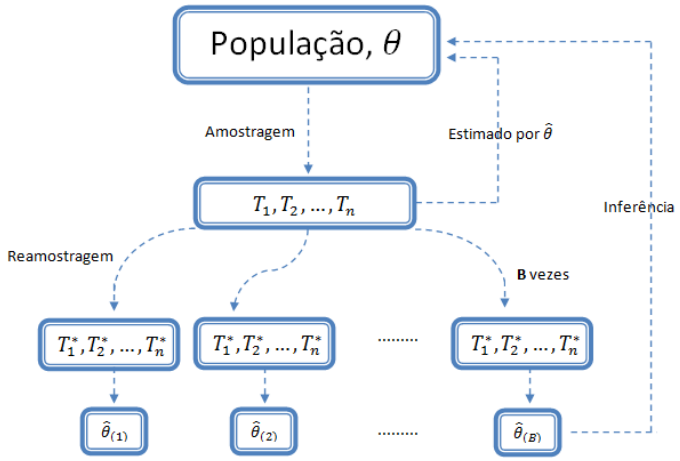

Figura 2.2 – Ilustração do método bootstrap

2.2 Funções de distribuição

2.2.1 Distribuição Burr III

Burr (1942) propôs um sistema de distribuições flexíveis, que inclui, por exemplo a distribuição Burr XII. Seja T uma variável aleatória com distribuição Burr XII, então, T^{-1} tem distribuição Burr III. Essa distribuição tem sido utilizada em vários campos da ciência, bem como em finanças, silvicultura, análise de sobrevivência e em teoria da confiabilidade (ver em Sherrick et al.,1996; Lindsay et al., 1996; Al-Dayian, 1999; Shao, 2000; Hose, 2005; Mokhlis, 2005; Gove et al., 2008). Assim, a distribuição Burr III (α, β, s) tem a seguinte função distribuição de probabilidade:

$$F(t;\alpha,\beta,s) = \left[1+\left(\frac{t}{s}\right)^{-\alpha}\right]^{-\beta} = \left[\frac{(t/s)^{\alpha}}{1+(t/s)^{\alpha}}\right]^{\beta}, t>0, \tag{2.9}$$

em que $\alpha > 0$ e $\beta > 0$ são parâmetros de forma e $s > 0$ parâmetro de escala.

Consequentemente, a função densidade de probabilidade, $f(t)$, para a variável aleatória tempo de falha T com distribuição Burr III é expressa por:

$$f(t;\alpha,\beta,s) = \frac{\alpha\beta}{s(t/s)^{\alpha+1}}\left[\frac{(t/s)^{\alpha}}{1+(t/s)^{\alpha}}\right]^{\beta+1}, \tag{2.10}$$

Como descrito na Seção 2.1.2, conhecendo a função distribuição de probabilidade de uma determinada distribuição, pode-se encontrar a função de sobrevivência para essa distribuição. Dessa forma, a função de sobrevivência da distribuição Burr III é definida por:

$$S(t) = 1 - F(t) = 1 - \left[1+\left(\frac{t}{s}\right)^{-\alpha}\right]^{-\beta}. \tag{2.11}$$

Ao considerar as funções definidas nas equações (2.10) e (2.11), define-se a função de risco $h(t)$ da distribuição Burr III:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\dfrac{\alpha\beta}{s(t/s)^{\alpha+1}}\left[\dfrac{(t/s)^{\alpha}}{1+(t/s)^{\alpha}}\right]^{\beta+1}}{1-\left[1+(t/s)^{-\alpha}\right]^{-\beta}} = \frac{\alpha\beta}{t\{[1+(t/s)^{-\alpha}]^{\beta}-1\}}. \tag{2.12}$$

Para entender o comportamento dessa distribuição de probabilidade, alguns gráficos da função de densidade e da função de risco foram obtidos e encontram-se nas Figuras 2.3 e 2.4, respectivamente.

Com base na Figura 2.3 pode-se observar que esse modelo pode comportar dados do tipo assimétrico à direita e com caudas pesadas. Pela Figura 2.4 observa-se que a distribuição Burr III tem grande flexibilidade devido a diversas formas que a sua função de risco pode assumir. Por exemplo, decrescente e unimodal.

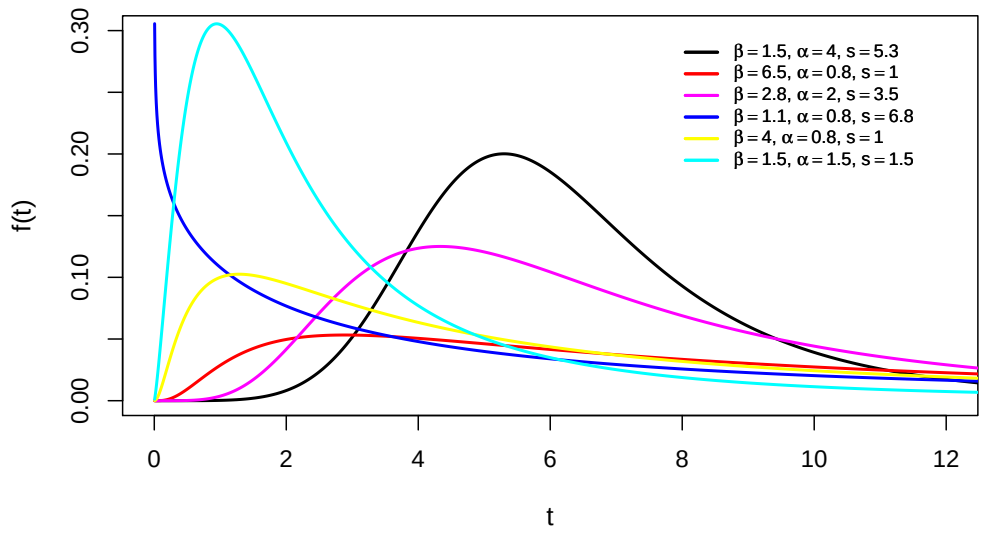

Figura 2.3 – Gráfico da função densidade da distribuição Burr III

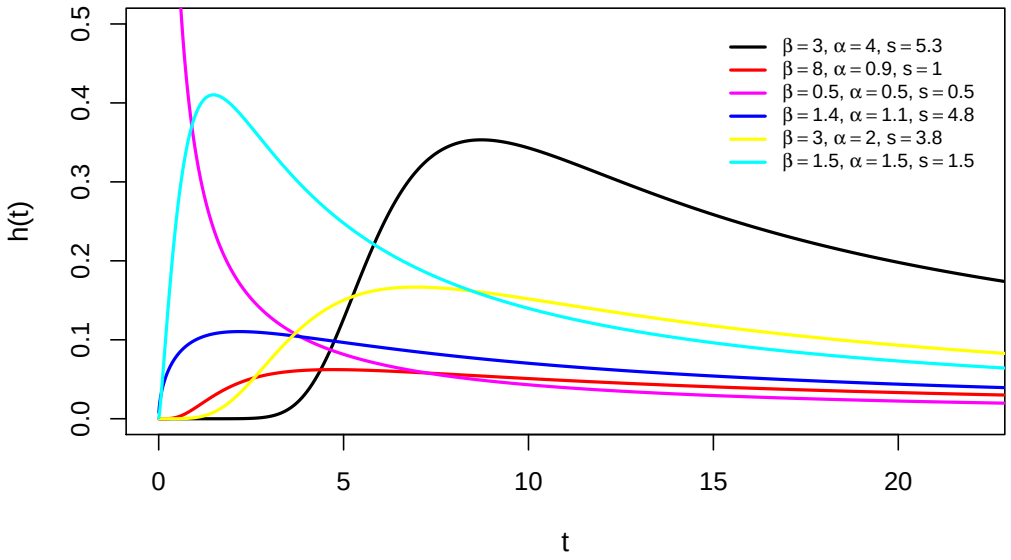

Figura 2.4 – Gráfico da função de risco da distribuição Burr III

2.2.2 Distribuição log-Burr III

Seja T uma variável aleatória com distribuição Burr III como expressa na equação (2.10), então $Y = \log(T)$ tem distribuição log-Burr III. Essa distribuição foi apresentada também no trabalho de Gomes et al. (2013). Considerando-se as seguintes reparametrizações $\alpha = 1/\sigma$ e $s = \exp(\mu)$, pode-se obter a função densidade de probabilidade da distribuição log-Burr III pelo método do jacobiano. A demonstração é dada no Apêndice A. Assim, a fdp de y é expressa por:

$$f(y) = \frac{\beta \left[\exp\left(\frac{y-\mu}{\sigma}\right)\right]^{\beta}}{\sigma \left[1 + \exp\left(\frac{y-\mu}{\sigma}\right)\right]^{\beta+1}} \quad , \tag{2.13}$$

em que $-\infty < y < +\infty$, $-\infty < \mu < \infty$, $\beta > 0$, $\sigma > 0$.

Além disso, ao conhecer função densidade $f(y)$, podemos obter a função distribuição. Dessa forma, a função distribuição da distribuição log-Burr III é dada por:

$$F(y) = \int_{-\infty}^{y} f(u)du$$

$$= \int_{-\infty}^{y} \frac{\beta}{\sigma} \left[\frac{\left[\exp\left(\frac{u-\mu}{\sigma}\right)\right]^{\beta}}{\left[1 + \exp\left(\frac{u-\mu}{\sigma}\right)\right]^{\beta+1}} \right] du = \left[\frac{\exp\left(\frac{u}{\sigma}\right)}{\exp\left(\frac{\mu}{\sigma}\right) + \exp\left(\frac{u}{\sigma}\right)} \right]^{\beta} \Bigg|_{-\infty}^{y}$$

$$= \left[\frac{\exp\left(\frac{u-\mu}{\sigma}\right)}{1 + \exp\left(\frac{u-\mu}{\sigma}\right)} \right]^{\beta} \Bigg|_{-\infty}^{y} = \left[\frac{\exp\left(\frac{y-\mu}{\sigma}\right)}{1 + \exp\left(\frac{y-\mu}{\sigma}\right)} \right]^{\beta} . \tag{2.14}$$

Como visto na Seção 2.2 e baseando-se na equação (2.2), a função de sobrevivência, $S(y)$, para a variável aleatória Y, ou seja, logaritmo de T, é expressa por:

$$S(y) = 1 - \left[\frac{\exp\left(\frac{y-\mu}{\sigma}\right)}{1 + \exp\left(\frac{y-\mu}{\sigma}\right)} \right]^{\beta} . \tag{2.15}$$

Então, sabendo-se as funções densidade e de sobrevivência da distribuição log-Burr III, definidas pela equações (2.13) e (2.15), respectivamente, pode-se definir a função

de risco. Dessa forma, tem-se que a função de risco $h(y)$ da distribuição log-Burr III é dada por:

$$h(y) = \frac{\dfrac{\beta}{\sigma}\left[\dfrac{\left[\exp\left(\dfrac{y-\mu}{\sigma}\right)\right]^{\beta}}{\left[1+\exp\left(\dfrac{y-\mu}{\sigma}\right)\right]^{\beta+1}}\right]}{1-\left[\dfrac{\exp\left(\dfrac{y-\mu}{\sigma}\right)}{1+\exp\left(\dfrac{y-\mu}{\sigma}\right)}\right]^{\beta}} = \frac{\beta}{\sigma\left[\left\{1+\exp\left[-\left(\dfrac{y-\mu}{\sigma}\right)\right]\right\}^{\beta}-1\right]} \quad (2.16)$$

Diante disso, visando verificar o comportamento da distribuição log-Burr III, foram obtidos os gráficos da função densidade de probabilidade e da função de risco para essa distribuição.

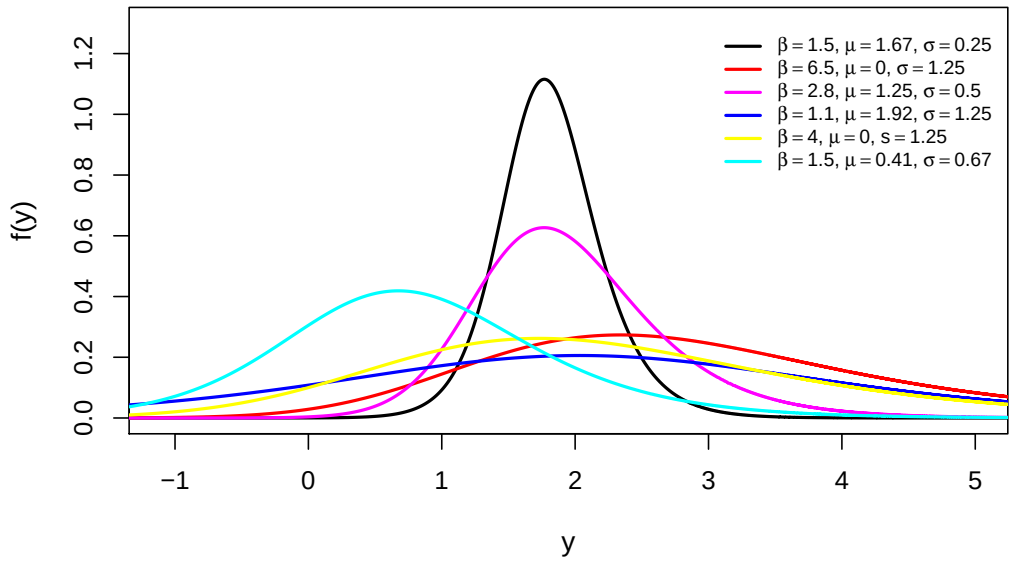

Figura 2.5 – Gráfico da função densidade da distribuição log-Burr III

Pela Figura 2.5 pode-se observar que esse modelo com distribuição log-Burr III pode comportar dados do tipo simétrico. Em relação a função de risco, tem-se pela Figura 2.6 que para esse modelo sua função de risco, à princípio, não assume um comportamento específico, as curvas crescem e depois estabilizam.

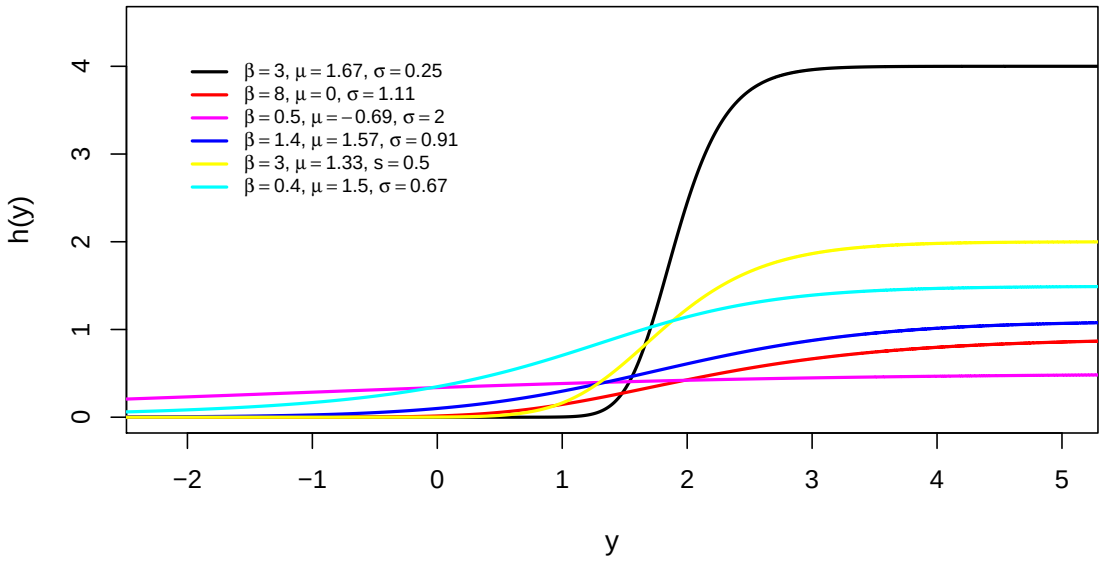

Figura 2.6 – Gráfico da função de risco da distribuição log-Burr III

2.2.3 Distribuição Beta Burr III

Nos últimos anos, muitos estudos têm focado na generalização de classes de funções de distribuição, como a classe generalizada beta (EUGENE et al., 2002). Obtendo, assim, uma distribuição com maior flexibilidade. Se G denota a função distribuição acumulada de uma variável aleatória, a distribuição beta-G é definida por:

$$F(t) = I_{G(t)}(a,b) = \frac{1}{B(a,b)} \int_0^{G(t)} \omega^{a-1}(1-\omega)^{b-1} d\omega, \tag{2.17}$$

em que $a > 0$ e $b > 0$ são parâmetros de forma da distribuição G. Tem-se que $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ é a função beta, $\Gamma(\cdot)$ é a função gama, $B_t(a,b) = \int_0^t w^{a-1}(1-\omega)^{b-1} d\omega$ é a função beta incompleta e a função razão beta incompleta é definida por:

$$I_q(a,b) = B_q(a,b)/B(a,b) \tag{2.18}$$

Gomes et al. (2013) seguindo a mesma ideia desenvolvida por Eugene et al. (2002), propuseram a distribuição beta Burr III (BBIII), a fim de acomodar uma grande variedade de formas. Considerou-se que a função $G(t)$ definida na equação (2.17) é a função distribuição acumulada da distribuição Burr III definida na equação (2.9).

Seja T uma variável aleatória com distribuição beta Burr III, sua função distribuição de probabilidade é expressa por:

$$F(t) = I_{[1+(t/s)^{-\alpha}]^{-\beta}}(a, b). \tag{2.19}$$

Consequentemente, a função densidade de probabilidade da BBIII (para $t > 0$) é dada por:

$$f(t) = \frac{\alpha\beta}{s(t/s)^{\alpha+1}B(a,b)} \left[\frac{(t/s)^{\alpha}}{1+(t/s)^{\alpha}}\right]^{\beta a+1} \left\{1 - \left[\frac{(t/s)^{\alpha}}{1+(t/s)^{\alpha}}\right]^{\beta}\right\}^{b-1}. \tag{2.20}$$

em que $\alpha > 0, \beta > 0, a > 0, b > 0$ são parâmetros de forma e $s > 0$ parâmetro de escala.

Ao considerar as funções definidas nas equações (2.19) e (2.20), tem-se a função de risco da distribuição BBIII expressa como:

$$h(t) = \frac{\beta\alpha[s(t/s)^{-\alpha-1}]}{B(a,b)I_{1-[1+(t/s)^{-\alpha}]^{-\beta}}(b,a)} \left[\frac{(t/s)^{\alpha}}{1+(t/s)^{\alpha}}\right]^{\beta\alpha+1} \times$$

$$\left\{1 - \left[\frac{(t/s)^{\alpha}}{1+(t/s)^{\alpha}}\right]^{\beta}\right\}^{b-1}. \tag{2.21}$$

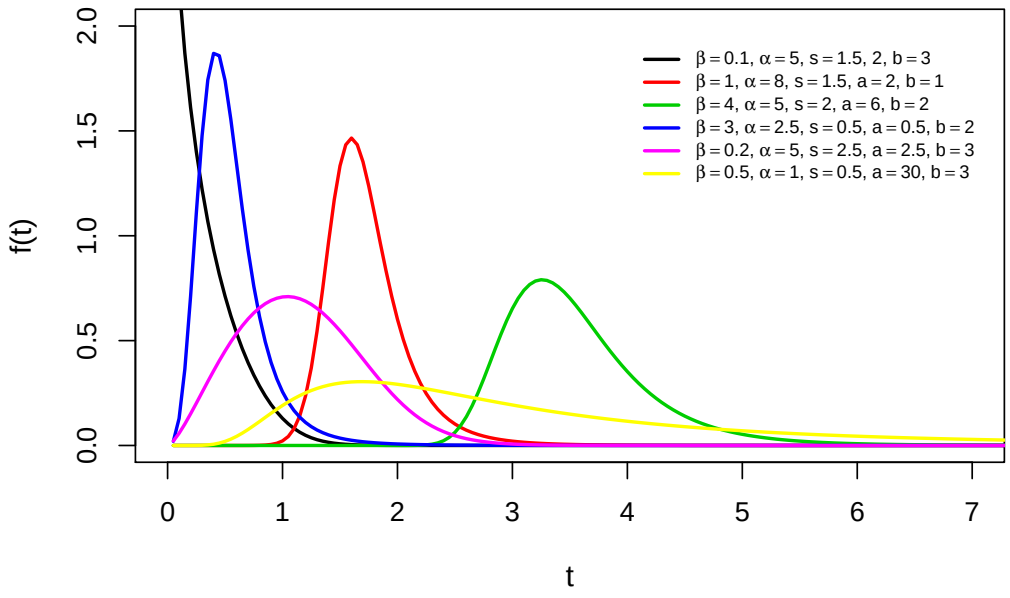

Figura 2.7 – Gráfico da função densidade da distribuição beta Burr III

A função densidade de probabilidade BBIII expressa pela equação (2.20) permite uma maior flexibilidade de suas caudas, incluindo alguns submodelos importantes. Como, por exemplo: distribuição Burr III exponenciada para $b = 1$ e a distribuição Burr III para $a = b = 1$. Essa flexibilidade pode ser observada no gráfico da $f(t)$ da distribuição BBIII presente na Figura 2.7.

Também foi obtido o gráfico da função de risco dessa distribuição a fim de observar seu comportamento. Pela Figura 2.8 pode-se notar que a função de risco pode acomodar, por exemplo, as formas decrescente, unimodal, e possui a forma crescente e depois estabiliza. Caracterizando, assim, a flexibilidade desse modelo.

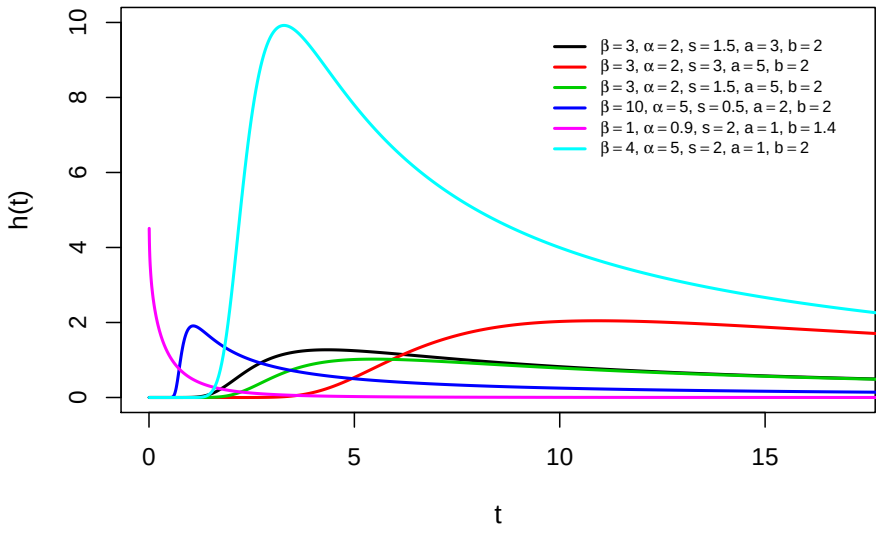

Figura 2.8 – Gráfico da função de risco da distribuição beta Burr III

2.2.4 Distribuição Log-Beta Burr III

Nas Subseções 2.2.1, 2.2.2 e 2.2.3, as distribuições Burr III, log-Burr III e beta Burr III são apresentadas, respectivamente. Nesta Subseção é apresentada a distribuição log-beta Burr III, que é a distribuição considerada no modelo de regressão para dados grupados proposto neste trabalho.

Seja T uma variável aleatória com função distribuição BBIII definida na equação (2.20) da Seção 2.2.3. Considerando-se a transformação $Y = \log(T)$ e as seguintes reparametrizações $\alpha = 1/\ \sigma$ e $s = \exp(\mu)$, tem-se que Y tem distribuição log-beta Burr III (LBBIII) com função densidade de probabilidade dada por:

$$f_y(y) = \frac{\beta}{\sigma \exp\left(\frac{y-\mu}{\sigma}\right) B(a,b)} \left[\frac{\exp\left(\frac{y-\mu}{\sigma}\right)}{1+\exp\left(\frac{y-\mu}{\sigma}\right)}\right]^{\beta a+1} \times$$
$$\left\{1 - \left[\frac{\exp\left(\frac{y-\mu}{\sigma}\right)}{1+\exp\left(\frac{y-\mu}{\sigma}\right)}\right]^{\beta}\right\}^{b-1}. \quad (2.22)$$

Como já foi visto, ao se conhecer a função densidade de probabilidade da LBBIII definida na equação (2.22), pode-se obter sua função distribuição de probabilidade dada por:

$$F(y) = I_{[\exp((y-\mu)/\sigma)/(1+\exp((y-\mu)/\sigma))]^{\beta}}(a,b). \quad (2.23)$$

em que $-\infty < y < \infty$, $-\infty < \mu < \infty$, $\sigma > 0$ e $\beta > 0$. Como visto, essa função distribuição de probabilidade é expressa pela função razão beta incompleta descrita na Equação (2.18).

Então, de acordo com a Seção 2.2, conhecendo a função distribuição de probabilidade expressa na Equação (2.23) e utilizando a relação $S(t) = 1 - F(t)$, pode-se encontrar a função de sobrevivência. Dessa forma, a função de sobrevivência da distribuição LBBIII é dada por:

$$S(y) = 1 - I_{[\exp((y-\mu)/\sigma)/(1+\exp((y-\mu)/\sigma))]^{\beta}}(a,b). \quad (2.24)$$

Ao conhecer a função densidade de probabilidade e a função de sobrevivência da distribuição log-beta Burr III, definidas nas equações (2.22) e (2.24), respectivamente, pode-se definir a função de risco da LBBIII. Assim, a função de risco $h(y)$, que não possui uma forma simplificada, é dada por:

$$h(y) = \frac{1}{[1 - I_{[\exp((y-\mu)/\sigma)/(1+\exp((y-\mu)/\sigma))]^{\beta}}(a,b)]} \left[\frac{\beta}{\sigma \exp\left(\frac{y-\mu}{\sigma}\right) B(a,b)}\right] \times$$
$$\left[\frac{\exp\left(\frac{y-\mu}{\sigma}\right)}{1+\exp\left(\frac{y-\mu}{\sigma}\right)}\right]^{\beta a+1} \left\{1 - \left[\frac{\exp\left(\frac{y-\mu}{\sigma}\right)}{1+\exp\left(\frac{y-\mu}{\sigma}\right)}\right]^{\beta}\right\}^{b-1}. \quad (2.25)$$

Com o objetivo de verificar os possíveis comportamentos da função densidade de probabilidade e as possíveis formas da função de risco, são construídos os gráficos da

função de densidade e da função de risco para a distribuição LBBIII, considerando diversos valores para seus parâmetros. Os gráficos encontram-se nas Figuras 2.9 e 2.10, respectivamente.

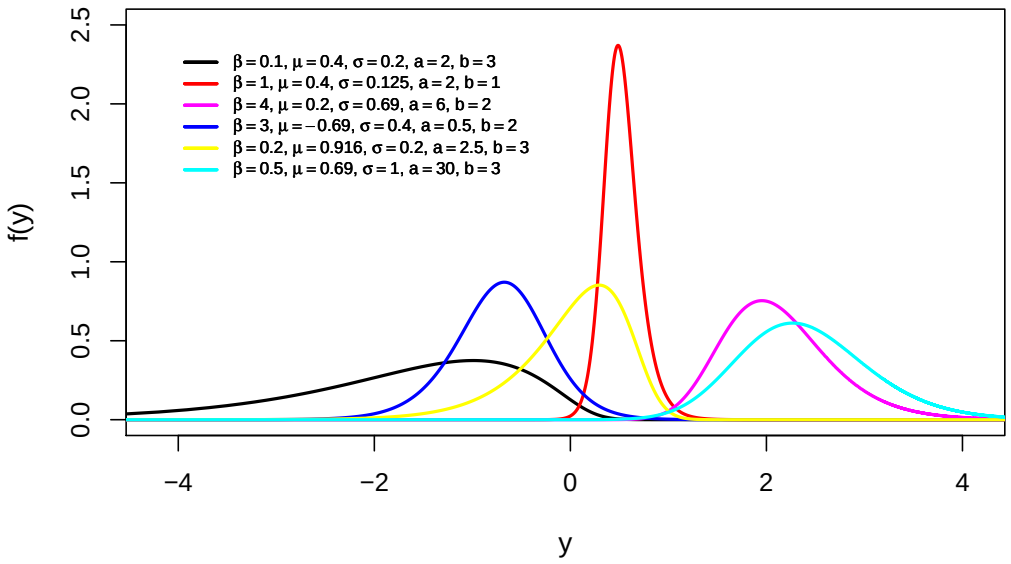

Figura 2.9 – Gráfico da função densidade da distribuição log-beta Burr III

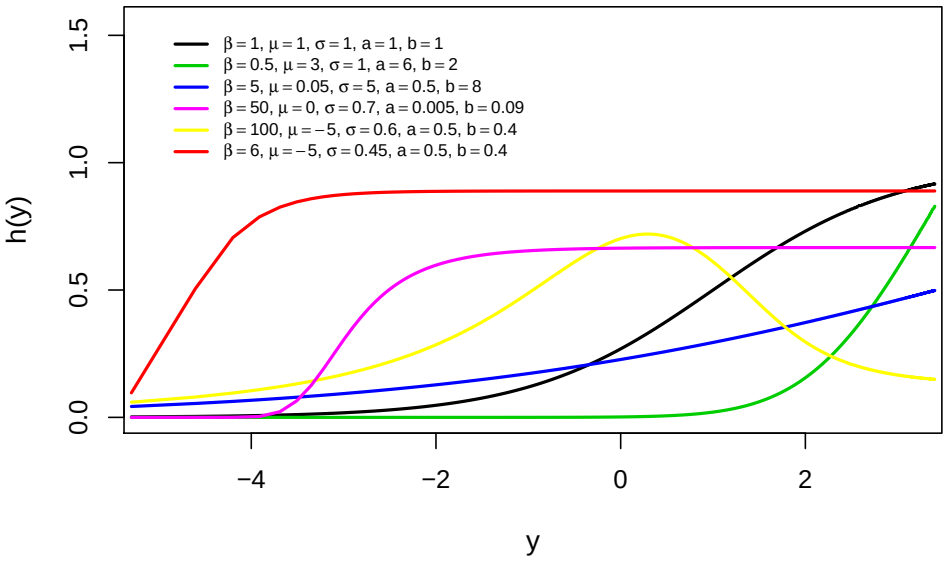

Figura 2.10 – Gráfico da função de risco da distribuição log-beta Burr III

Uma característica da distribuição LBBIII é em relação a sua flexibilidade, que pode ser observada no gráfico de sua função densidade na Figura 2.9. Verifica-se que ela acomodar dados simétricos, assimétricos à esquerda, entre outros. E pela Figura 2.10 se vê que a função de risco pode assumir formas crescente, decrescente e unimodal.

Capítulo 3

Material e métodos

3.1 Material

Visando ilustrar o modelo de regressão proposto é utilizado um banco de dados cedido pelo Instituto de Saúde Coletiva da Universidade Federal da Bahia. Esses dados foram obtidos de um estudo conduzido por Barreto et al. (1994), cujo objetivo foi avaliar o efeito da suplementação de vitamina A na diarreia no nordeste do Brasil entre dezembro de 1990 e dezembro de 1991. O banco de dados é composto de 1207 crianças com idade entre 6 e 48 meses no início do estudo, que receberam placebo ou vitamina A.

O tempo de sobrevivência para o estudo foi definido como sendo o tempo da primeira dose da vitamina A ou placebo até a ocorrência do primeiro episódio de diarreia. Um episódio de diarreia foi considerado como sendo uma sequência de dias com diarreia. Em cada visita, a informação de ocorrência de diarreia foi coletada, essa ocorre no período de 48 a 72 horas. Das 1207 crianças observadas no estudos, 925 apresentaram um episódio de diarreia e 282 apresentaram tempos de censura. Neste estudo foram consideradas algumas variáveis associadas a cada criança, para $i = 1, ..., 1207$, são elas:

- x_{i1}: idade no início do estudo (em meses),
- x_{i2}: tratamento (0=placebo, 1=vitamina A);
- x_{i3}: sexo (0=feminino, 1=masculino)
- t_i é o tempo observado (em horas).

3.2 Métodos

3.2.1 Modelo de regressão log-beta Burr III para dados grupados

Em análise de sobrevivência, muitos estudos envolvem covariáveis que podem estar relacionadas com o tempo de sobrevivência. Por exemplo, o tipo de tratamento, a idade e o sexo de um paciente submetido a um tratamento podem influenciar no tempo de sobrevivência. Dessa forma, essas covariáveis devem estar presentes na análise estatística dos dados. Por isso, nesta seção é apresentada uma extensão da distribuição log-beta Burr III pela inclusão do vetor de covariáveis $\mathbf{x}$.

Neste trabalho é proposto um modelo de regressão de locação e escala para dados grupados baseando-se na função densidade de probabilidade da LBBIII. Esse modelo pode ser escrito como o modelo log-linear:

$$Y = \mu + \sigma Z, \tag{3.1}$$

no qual $Z = (Y - \mu)/\sigma$ é o erro aleatório. Com base na equação (2.22) e utilizando-se o método do jacobiano, tem-se que a função densidade é dada por:

$$f(z) = \frac{\beta \exp(-z)}{B(a,b)} \left[\frac{\exp(z)}{1+\exp(z)}\right]^{\alpha b - 1} \left\{1 - \left[\frac{\exp(z)}{1+\exp(z)}\right]^{\beta}\right\}^{b-1}, \tag{3.2}$$

em que $-\infty < z < \infty$.

Considerando que o parâmetro de escala μ depende do vetor de variáveis regressoras $\mathbf{x} = (\mathbf{x_0}, \mathbf{x_1}, ..., \mathbf{x_p})^T$ por $\mu = \mathbf{x}^T\boldsymbol{\gamma}$, tem-se este modelo $Y|\ \mathbf{x}$ relaciona a variável resposta Y com as covariáveis $\mathbf{x}$ e pode ser representado por:

$$y = \mathbf{x}^T\gamma + \boldsymbol{\sigma} z, \tag{3.3}$$

em que $(\gamma_0, \gamma_1, ..., \gamma_p)^T$, $\sigma > 0$ e $\beta > 0$ são parâmetros desconhecidos e z é o erro aleatório com função densidade expressa na equação (3.2).

Assim, tem-se que a função de sobrevivência é dada por:

$$S(y|\mathbf{x}) = 1 - I_{[\exp((y-\mathbf{x}^T\gamma)/\sigma)/(1+\exp((y-\mathbf{x}^T\gamma)/\sigma))]^{\beta}}(a,b). \tag{3.4}$$

É importante destacar que para o caso especial $a = b = 1$ se tem a distribuição log-burr III padrão.

3.2.2 Especificação do modelo de regressão para dados grupados

Ao utilizar a definição de dados grupados presente na Subseção 2.1.1 especifica-se nesta seção o modelo de regressão para dados grupados, bem como, o método de estimação dos parâmetros para esse modelo.

Os intervalos são construídos de tal modo que os eixos dos tempos são divididos em k intervalos definidos por pontos de corte $a_1, ..., a_k$. Então, o j-ésimo intervalo é denotado por $I_j = [a_j, a_{j+1}]$ para $j = 1, ..., k$ e os logaritmos dos tempos y_i são agrupados em k intervalos. A função de sobrevivência nos pontos $\log(a_{j+1})$ e $\log(a_j)$, dado $\mathbf{x}$ são dados por:

$$S[\log(a_{j+1})|\mathbf{x}] = 1 - I_{[\exp((\log(a_{j+1})-\mathbf{x}^T\boldsymbol{\gamma})/\sigma)/(1+\exp((\log(a_{j+1})-\mathbf{x}^T\boldsymbol{\gamma})/\sigma))]^\beta}(a, b)$$

e

$$S[\log(a_j)|\mathbf{x}] = 1 - I_{[\exp((\log(a_j)-\mathbf{x}^T\boldsymbol{\gamma})/\sigma)/(1+\exp((\log(a_j)-\mathbf{x}^T\boldsymbol{\gamma})/\sigma))]^\beta}(a, b). \quad (3.5)$$

3.2.3 Estimador de máxima verossimilhança

Seja $(y_1, \mathbf{x}_1, ..., y_n, \mathbf{x}_n)$ uma amostra observada de n observações independentes, em que y_i representa o logaritmo do tempo de falha ou o logaritmo do tempo de censura e $\mathbf{x}_i = (\mathbf{x}_{i0}, \mathbf{x}_{i1}, ..., \mathbf{x}_{ip})^T$ é o vetor de variáveis explicativas associado ao i-ésimo indivíduo. Considerando que o logaritmo dos tempos y_i são agrupados em k intervalos denotados por $I_j = [\log(a_j), \log(a_{j+1})]$ para $j = 1, ...k$. A função de verossimilhança pode ser obtida considerando as variáveis explicativas $\mathbf{x}_i$ de modo que a contribuição do i-ésimo indivíduo no j-ésimo intervalo é dado por:

i) Se o i-ésimo indivíduo falhar no j-ésimo intervalo, sua contribuição para a função de verossimilhança é dada por $1 - S[\log(a_{j+1})|\mathbf{x}]/S[\log(a_j)|\mathbf{x}]$.

ii) Se o i-ésimo indivíduo sobreviver (ou seja, estiver sob risco) no j-ésimo intervalo, sua contribuição para a função de verossimilhança é dada por $S[\log(a_{j+1})|\mathbf{x}]/S[\log(a_j)|\mathbf{x}]$.

iii) Se o i-ésimo indivíduo for censurado no tempo c_i no j-ésimo intervalo, sua contribuição para a função de verossimilhança é dada por $S[\log(c_i)|\mathbf{x}]/S[\log(a_j)|\mathbf{x}]$, em que $\log(c_i) \in I_j$.

Assim, tem-se que a função de verossimilhança para o vetor de parâmetros $\boldsymbol{\theta} = (a, b, \sigma, \beta, \gamma^T)^T$ é dado por:

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{k} \left\{ \prod_{i \in F_j} [1 - S[\log(a_{j+1})|\mathbf{x}]/S[\log(a_j)|\mathbf{x}]] \times \right.$$
$$\left. \prod_{i \in R_j} [S[\log(a_{j+1})|\mathbf{x}]/S[\log(a_j)|\mathbf{x}]] \prod_{i \in C_j} [S[\log(c_i)|\mathbf{x}]/S[\log(a_j)|\mathbf{x}]] \right\}, \quad (3.6)$$

em que, F_j denota o conjunto de indivíduos que falharam no j-ésimo intervalo, R_j denota o número de indivíduos sob risco no j-ésimo intervalo, e C_i denota os indivíduos censurados no j-ésimo intervalo. Para Thompson (1977) a equação (3.6) é complicada de ser usada na prática, pois, se os dados forem grupados em intervalos, os tempos de c_i são desconhecidos. Assim, considera-se como contribuição dos indivíduos censurados no j-ésimo intervalo:

$$\frac{S[\log(c_i)]}{S[\log(a_j)]} = \left\{ \frac{S[\log(a_{j+1})]}{S[\log(a_j)]} \right\}^{1/2}. \quad (3.7)$$

Com isso, inserindo-se a equação (3.7) na equação (3.6) e considerando as funções de sobrevivência (3.5), o logaritmo da função de verossimilhança do vetor $\boldsymbol{\theta} = (a, b, \sigma, \beta, \gamma^T)^T$ para o modelo (3.3) tem a seguinte forma:

$$l(\boldsymbol{\theta}) = \sum_{j=1}^{k} \left(\sum_{i \in F_j} \log \left\{ 1 - \left[\frac{1 - I_{[(\exp(z_{ij+1}))/(1+\exp(z_{ij+1}))]^\beta}(a, b)}{1 - I_{[(\exp(z_{ij}))/(1+\exp(z_{ij}))]^\beta} \right] \right\} + \right.$$
$$+ \sum_{i \in R_j} \log \left[\frac{1 - I_{[(\exp(z_{ij+1}))/(1+\exp(z_{ij+1}))]^\beta}(a, b)}{1 - I_{[(\exp(z_{ij}))/(1+\exp(z_{ij}))]^\beta} \right] + \quad (3.8)$$
$$\left. + \frac{1}{2} \sum_{i \in C_j} \log \left[\frac{1 - I_{[(\exp(z_{ij+1}))/(1+\exp(z_{ij+1}))]^\beta}(a, b)}{1 - I_{[(\exp(z_{ij}))/(1+\exp(z_{ij}))]^\beta} \right] \right)$$

em que

$$z_{ij+1} = \frac{\log(a_{j+1}) - \mathbf{x_i}^T \boldsymbol{\gamma}}{\sigma}$$

e

$$z_{ij} = \frac{\log(a_j) - \mathbf{x_i}^T \boldsymbol{\gamma}}{\sigma}$$

.

3.2.4 Modelo de regressão log-Burr III para dados grupados

Como dito na Seção 2.2.3 a distribuição log-beta Burr III tem como caso particular a distribuição log-Burr III (LBIII), quando $a = b = 1$. Desse modo, é interessante também fazer um estudo sobre o modelo de regressão LBIII para dados grupados.

Analogamente um modelo de regressão de locação e escala para dados grupados é escrito como:

$$Y = \mu + \sigma Z, \tag{3.9}$$

no qual $Z = (Y - \mu)/\sigma$ é o erro aleatório. Com base na equação (2.13) e utilizando-se o método do jacobiano, obtem-se a seguinte função densidade:

$$f(z) = \left[\frac{\exp(z)}{1 + \exp(z)}\right]^{\beta}, \tag{3.10}$$

em que $-\infty < z < \infty$.

Sabendo que o parâmetro de escala μ depende do vetor de variáveis regressoras $\mathbf{x} = (\mathbf{x_0}, \mathbf{x_1}, ..., \mathbf{x_p})^T$, sendo $\mu = \mathbf{x}^T\boldsymbol{\gamma}$. Um modelo de regressão baseado na distribuição log Burr III (2.13) relacionando a variável resposta Y ao vetor de variáveis explicativas $\mathbf{x}$, assim o modelo $Y|\mathbf{x}$ pode ser expresso por:

$$y = \mathbf{x}^T\boldsymbol{\gamma} + \sigma z_i, \quad i = 1, ..., n, \tag{3.11}$$

em que $(\gamma_1, ..., \gamma_p)^T$, $\sigma > 0$ e $\beta > 0$ são parâmetros desconhecidos e z_i é o erro aleatório com função densidade expressa na equação 3.10. Desse modo, a função de sobrevivência de $Y|\mathbf{x}$ é dada por:

$$S[(\log(a_j)|\mathbf{x})] = 1 - \left[\frac{\exp\left(\dfrac{y - \mathbf{x}^T\boldsymbol{\gamma}}{\sigma}\right)}{1 + \exp\left(\dfrac{y - \mathbf{x}^T\boldsymbol{\gamma}}{\sigma}\right)}\right]^{\beta}. \tag{3.12}$$

Como visto na Seção anterior, ao considerar dados grupados tem-se que construir intervalos de tal modo que os eixos dos tempos são divididos em k intervalos definidos por pontos de corte $a_1, ..., a_k$. Assim, o j-ésimo intervalo é definido como $I_j = [a_j, a_{j+1}]$ para $j = 1, ..., k$ e os logaritmos dos tempos y_i são agrupados em k intervalos.

A função de sobrevivência nos pontos $\log(a_{j+1})$ e $\log(a_j)$, dado $\mathbf{x}$ para a distribuição LBIII, em que $\mathbf{x_i} = (\mathbf{x_0}, \mathbf{x_1}, ..., \mathbf{x_p})^T$ é o vetor de variáveis explicativas, são dados por:

$$S[(\log(a_{j+1})|\mathbf{x})] = 1 - \left[\frac{\exp\left(\frac{a_{j+1} - \mathbf{x}^T\boldsymbol{\gamma}}{\sigma}\right)}{1 + \exp\left(\frac{a_{j+1} - \mathbf{x}^T\boldsymbol{\gamma}}{\sigma}\right)}\right]^{\beta}.$$

e

$$S[(\log(a_j)|\mathbf{x})] = 1 - \left[\frac{\exp\left(\frac{a_j - \mathbf{x}^T\boldsymbol{\gamma}}{\sigma}\right)}{1 + \exp\left(\frac{a_j - \mathbf{x}^T\boldsymbol{\gamma}}{\sigma}\right)}\right]^{\beta}. \tag{3.13}$$

Ao considerar a função de verossimilhança para dados grupados definida na Seção 3.2.4, a função de verossimilhança do modelo de regressão LBIII para dados grupados é definida por:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{j=1}^{k}\left\{\prod_{i\in F_j}[1 - S[\log(a_{j+1})|\mathbf{x}]/S[\log(a_j)|\mathbf{x}]] \times \right. \\ &\left. \prod_{i\in R_j}[S[\log(a_{j+1})|\mathbf{x}]/S[\log(a_j)|\mathbf{x}]] \prod_{i\in C_j}\left\{\frac{S[\log(a_{j+1})]}{S[\log(a_j)]}\right\}^{1/2}\right\}, \end{aligned} \tag{3.14}$$

Considerando-se as funções de sobrevivência 3.13, o logaritmo da função de verossimilhança do vetor $\boldsymbol{\theta} = (\beta, \sigma, \gamma^T)^T$ é expressa a seguir:

$$\begin{aligned} l(\boldsymbol{\theta}) = \sum_{j=1}^{k}\left(\sum_{i\in F_j}\log\left\{1 - \frac{1 - \left[\frac{\exp(z_{ij+1})}{1+\exp(z_{ij+1})}\right]^{\beta}}{1 - \left[\frac{\exp(z_{ij})}{1+\exp(z_{ij})}\right]^{\beta}}\right\} + \right. \\ \sum_{i\in R_j}\log\left[\frac{1 - \left[\frac{\exp(z_{ij+1})}{1+\exp(z_{ij+1})}\right]^{\beta}}{1 - \left[\frac{\exp(z_{ij})}{1+\exp(z_{ij})}\right]^{\beta}}\right] + \\ \left.\frac{1}{2}\sum_{i\in C_j}\log\left[\frac{1 - \left[\frac{\exp(z_{ij+1})}{1+\exp(z_{ij+1})}\right]^{\beta}}{1 - \left[\frac{\exp(z_{ij})}{1+\exp(z_{ij})}\right]^{\beta}}\right]\right), \end{aligned} \tag{3.15}$$

em que

$$z_{ij+1} = \frac{\log(a_{j+1}) - \mathbf{x}^T\gamma}{\sigma}$$

e

$$z_{ij} = \frac{\log(a_j) - \mathbf{x}^T\gamma}{\sigma}$$

.

Portanto, o estimador de máxima verossimilhança $\hat{\boldsymbol{\theta}}$ dos parâmetros do modelo $\boldsymbol{\theta} = ((\beta, \sigma, \gamma^T)^T)$ podem ser obtidos maximizando o logaritmo função de verossimilhança (3.15). Para isso, foi utilizado o $software$ estatístico R.

Capítulo 4

Resultados e Discussão

4.1 Análise descritiva

Nesta Seção é realizada uma análise descritiva dos dados de vitamina A. Uma breve análise de cada covariável referente aos tempos de sobrevivência é feita, bem como um histograma desses tempos a fim de entender o comportamento dos dados. Para auxiliar em toda a análise descritiva dados, bem como nas estimações dos parâmetros dos modelos de regressão propostos é utilizado o *software* R.

Esse histograma dos tempos de sobrevivência encontra-se na Figura 4.1.

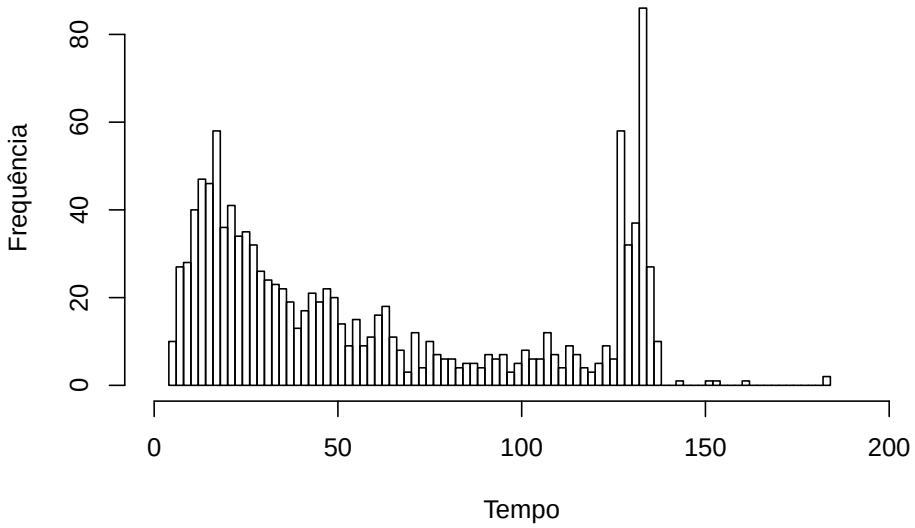

Figura 4.1 – Histograma dos dados de vitamina A

Dessa forma, observa-se pela Figura 4.1 uma frequência maior de observações no início do estudo e entre os tempos 125 a 150, enquanto que nos nos tempos finais há uma baixa frequência.

Na Tabela 4.1 encontra-se a tabela de vida do banco de dados de vitamina A considerando os oito intervalos apresentados. Nota-se pelos resultados dessa tabela que há uma presença considerável de empates. Dessa forma, a metodologia de dados grupados é indicada para modelar os dados de vitamina A.

Tabela 4.1 – Tabela de vida para dados de Vitamina A.

Intervalo	Número de falhas	Número de censuras	Número sob risco	$\hat{S}(t)$
$[4, 21)$	292	0	1207	1,000
$[21, 38)$	243	4	915	0.758
$[38, 55)$	138	6	668	0.556
$[55, 73)$	101	2	524	0.441
$[73, 90)$	46	3	421	0.356
$[90, 108)$	49	6	372	0.317
$[108, 126)$	46	11	317	0.275
$[126, 185)$	10	250	260	0.234

A primeira covariável brevemente analisada é *idade*, a idade média das crianças no início do estudo é de aproximadamente 26 meses. Na Figura 4.2 é apresentado um histograma da cováriavel idades e nota-se que a frequência das observações é parecida para a maior parte das idades.

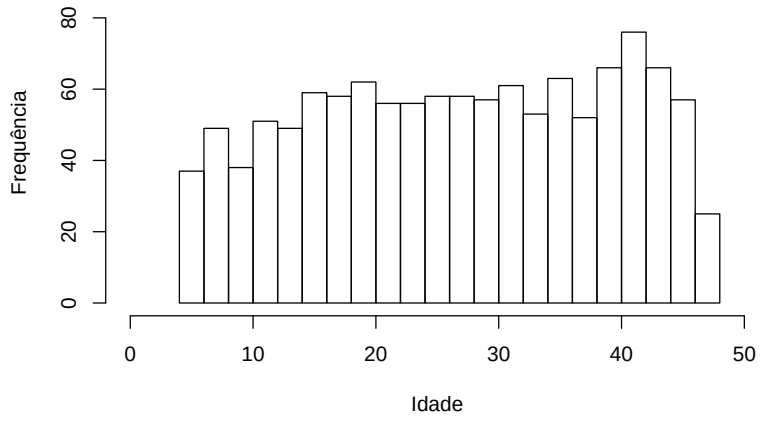

Figura 4.2 – Histograma das idades dos dados de vitamina A

A covariável *tratamento* refere-se ao tipo de tratamento recebido pelos indivíduos: placebo ou vitamina A. Dos 1207 indivíduos, 602 receberam o placebo e 605 vitamina A. O boxplot considerando *tratamento* é apresentado na Figura 4.3.

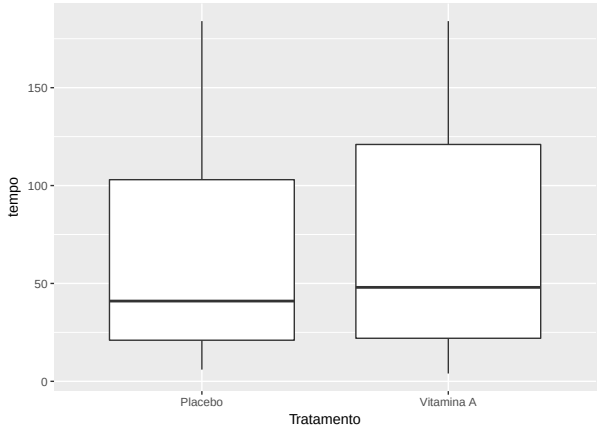

Figura 4.3 – Boxplot do tempo com relação a variável tratamento

Verifica-se que 50% dos indivíduos que receberam vitamina A levaram, em média, aproximadamente 50 dias para a ocorrência do primeiro episódio de diarreia após receberem a primeira dose do tratamento. Enquanto que 50% dos que receberam placebo levaram, em média, aproximadamente 40 para a ocorrência do evento de interesse.

Por fim, é feita uma breve análise da covariável *sexo*. Dos 1207 indivíduos estudados, 575 são do sexo feminino e 632 do masculino. Verifica-se pelo boxplot da variável *sexo* apresentado na Figura 4.4, que tanto para 50% dos indivíduos do sexo feminino quanto masculino, levaram, em média, aproximadamente 46 dias para apresentarem o evento de interesse. Ou seja, em média, 46 dias entre a primeira dose do tratamento e a ocorrência da diarreia.

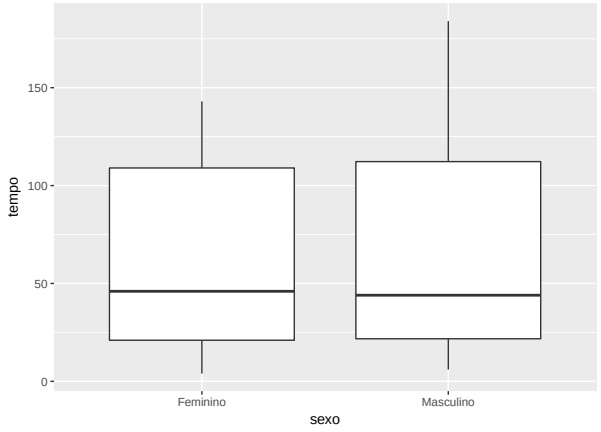

Figura 4.4 – Boxplot do tempo com relação a variável sexo

Como visto na Seção 2.2, a curva TTT é utilizada para identificar o possível comportamento da função de risco dos dados analisados. A curva TTT para os dados de Vitamina A encontra-se na Figura 4.5. Esse gráfico foi construído com base na Curva TTT para dados censurados, e a programação encontra-se no Apêndice B. Observa-se que há um indicativo de que a função de taxa de falha tem forma unimodal, visto que a curva TTT é primeiro côncava e depois convexa. .

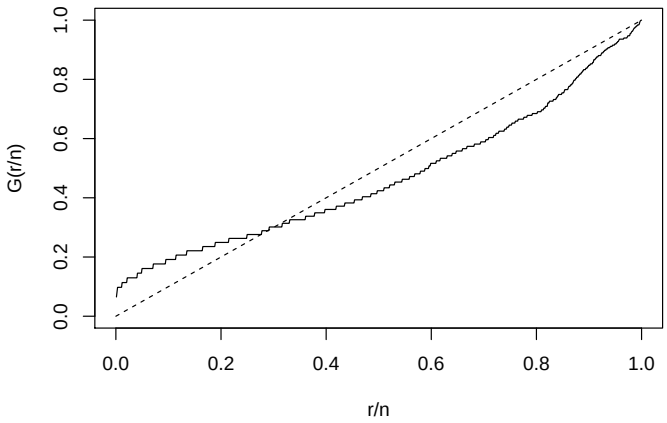

Figura 4.5 – Curva TTT para dados de Vitamina A

Como neste trabalho é aplicado os modelos de regressão LBBIII e LBIII, também foi construído uma curva TTT para o logaritmo do tempos como ilustra a Figura (4.6).

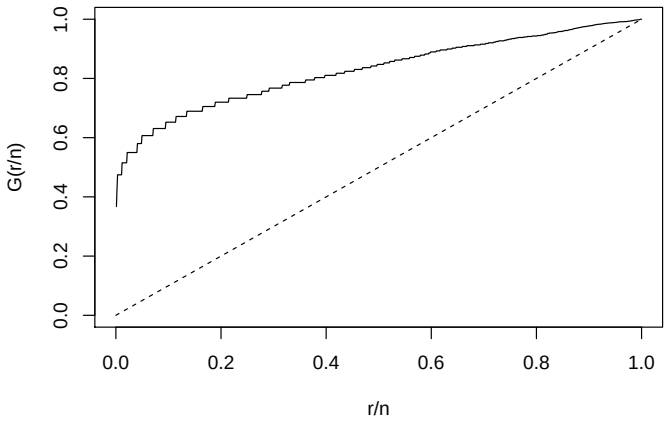

Figura 4.6 – Curva TTT para o logaritmo do tempos de Vitamina A

Dessa forma, observa-se pela Figura 4.6 que a curva TTT considerando o logaritmo

dos tempos assume a forma côncava e, portanto há indicativo de que sua taxa de falha seja monotonicamente crescente.

O estimador não-paramétrico de Kaplan-Meier para a função de sobrevivência considerando o tempo e o logaritmo dos tempos de Vitamina A foi obtido. Na Figura 4.7 encontra-se as estimativas de sobrevivência para os dados em estudo considerando o tempo e na Figura 4.8 considerando o logaritmo dos tempos.

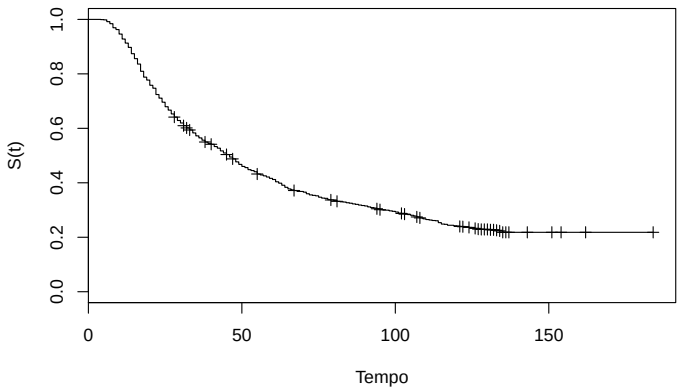

Figura 4.7 – Função de sobrevivência estimada por Kaplan-Meier

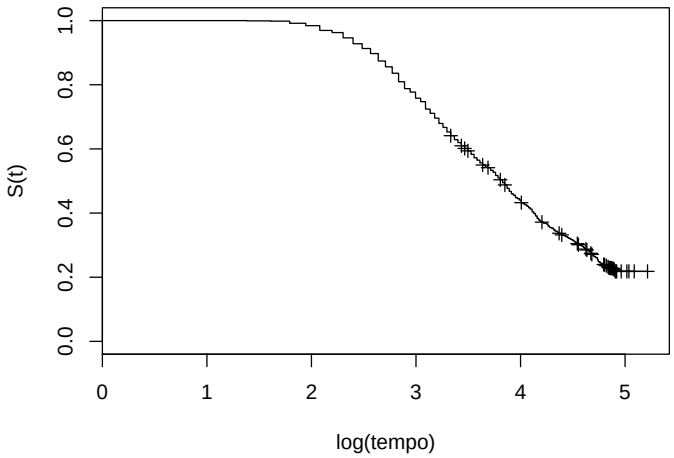

Figura 4.8 – Função de sobrevivência estimada por Kaplan-Meier para logaritmo do tempo

Verifica-se na Figura 4.7 que ao longo do tempo a função de sobrevivência decresce e há uma grande presença de censuras entre os tempos 100 e 150, que pode explicar a

alta frequência de observações nesses tempos. Enquanto que, pela Figura 4.8, nota-se maior número de censuras entre os tempos 3 e 5.

A forma gráfica do estimador de Kaplan-Meier para os tipos de tratamento recebidos pelos indivíduos estudados considerando o logaritmo do tempos, e para o sexo dos indivíduos são apresentadas na Figuras 4.9 e 4.10, respectivamente.

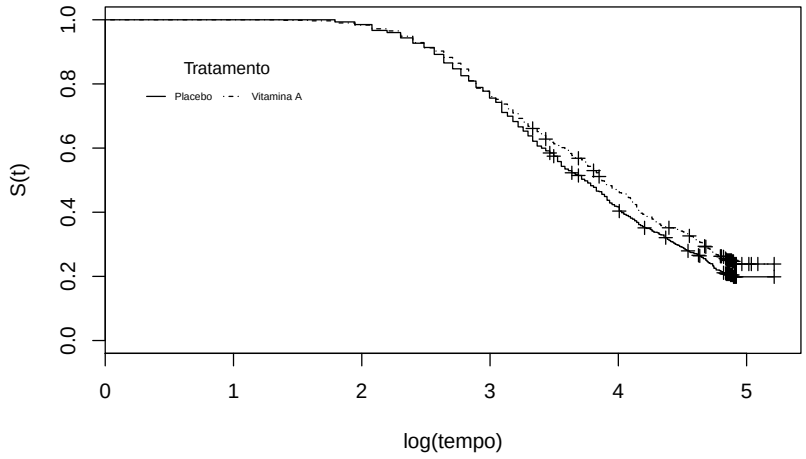

Figura 4.9 – Estimativas de Kaplan-Meier para tratamento placebo e vitamina A dos dados de Vitamina A

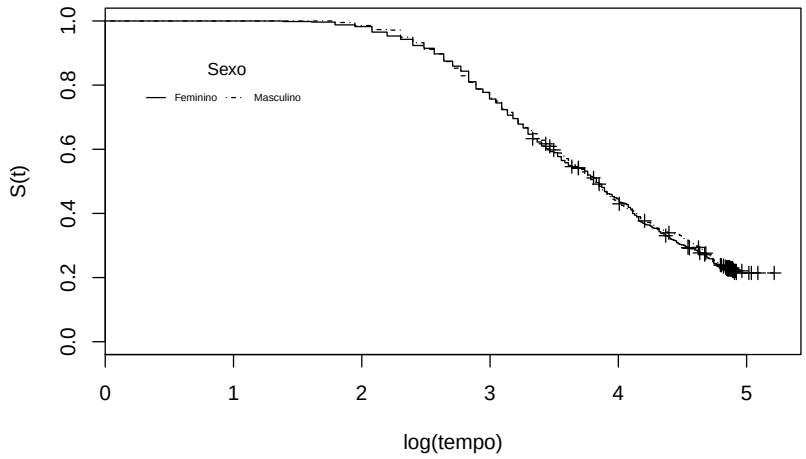

Figura 4.10 – Estimativas de Kaplan-Meier para sexo feminino e masculino dos dados de Vitamina A

Considerando-se apenas a Figura 4.9 não se pode concluir pela existência de di-

ferenças significativas entre o tipos de tratamento placebo e vitamina A. Entretanto, pela Figura 4.10 verifica-se que não há diferença significativa entre as curvas de sobrevivência do sexo feminino e masculino. Visto que não há suposição de riscos proporcionais, ou seja, as curvas se cruzam em determinados períodos de tempo, deve-se utilizar o teste não-paramétrico de $Wilcoxon$ para comparar as curvas de sobrevivência . Esse teste tem as seguintes hipóteses:

$$\begin{cases} H_0 : S_1 = S_2 \\ H_1 : S_1 \neq S_2. \end{cases}$$

Na Tabela são apresentados os resultados dos testes $Wilcoxon$ para comparação das curvas de sobrevivência considerando a covariável *tratamento* e *sexo*.

Tabela 4.2 – Resultados dos testes $Wilcoxon$ para comparação das curvas de sobrevivência.

Covariável	Estatística do teste	Graus de liberdade	$p - valor$
Tratamento	3	1	0.0819
Sexo	0	1	0.8350

Pelos resultados presentes na Tabela , conclui-se que a um nível de significância de 5% não se rejeita a hipótese nula, ou seja, não há evidência de diferença entre as curvas de sobrevivência, tanto considerando a covariável *tratamento* quanto *sexo*.

O próximo passo é verificar se os ajustes dos modelos com a distribuição LBBIII e seu caso particular, a distribuição LBIII são adequados. Na Seção 4.2 é apresentado o ajuste do modelo log-beta Burr III para dados censurados e na Seção 4.3 o ajuste do modelo log-Burr III.

4.2 Modelo log-beta Burr III para dados grupados

Com base no modelo de regressão para dados grupados proposto na Seção 3.2 é realizada uma aplicação ao banco de dados referente a suplementação de Vitamina A. Primeiramente, utilizou-se uma aplicação do modelo de regressão LBBIII para dados grupados, cuja função de verossimilhança é dada pela equação 3.6. Como visto, a função de verossimilhança para dados grupados depende da função de sobrevivência, esta função encontra-se na equação (2.24).

Como foi dito nas seções anteriores, ao utilizar dados grupados deve-se construir intervalos. Não existe um critério para a construção desses intervalos, apenas

a suposição de que se tenha a presença de pelo menos uma falha em cada intervalo. Por conveniência, para o estudo em questão foram construídos 8 intervalos: $\{(4, 21], (21, 38], (38, 55], (55, 73], (73, 90], (90, 108], (108, 126], (126, 185]\}$.

Primeiramente, foi realizada uma análise do modelo de regressão LBBIII para dados grupados sem considerar as covariáveis no modelo. Com auxílio do *software* R obteve-se as estimativas e os erros-padrão do vetor de parâmetros $\boldsymbol{\theta} = (\beta, \sigma, \mu, a, b)$ que encontram-se na Tabela 4.3.

Tabela 4.3 – Estimativas de máxima verossimilhança para os parâmetros do modelo de regressão LBBIII para dados grupados (sem covariável).

Parâmetros	Estimativas	Erro padrão
β	40.52668288	22.715287557
σ	0.53117960	0.107120452
μ	3.90831008	0.339020215
a	0.01514640	0.007846605
b	0.08043491	0.045140597

Observa-se que os erros-padrão do parâmetro β foi relativamente muito alto, entretanto, para os demais parâmetros os erros-padrão são aceitáveis. Com o intuito de verificar a qualidade do ajuste, construiu-se o gráfico da função de sobrevivência empírica (Kaplan-Meier) e da função de sobrevivência estimada pelo ajuste do modelo LBBIII para dados grupados sem covariáveis. Esse gráfico encontra-se na Figura 4.11.

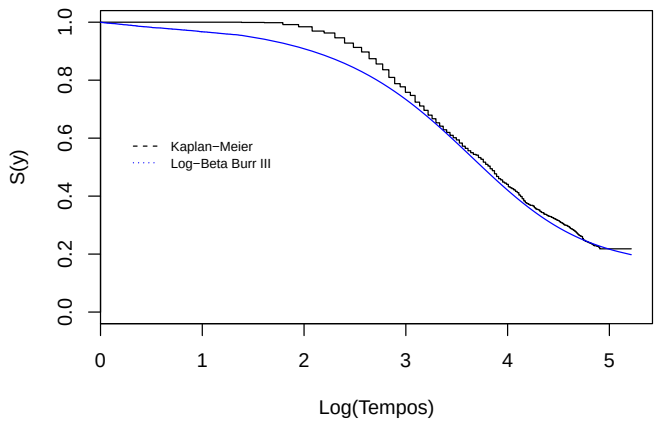

Figura 4.11 – Estimativa da função de sobrevivência para o modelo log-beta Burr III para dados grupados e Kaplan-Meier para os dados de Vitamina A.

Por meio da Figura 4.11 verifica-se um ajuste razoável, pois a curva do modelo ajustado LBBIII acompanha o gráfico da função de sobrevivência estimada por Kaplan-Meier, exceto no intervalo de 0 a 3.

Posteriormente ao ajuste do modelo de regressão LBBIII para dados de sobrevivência grupados sem considerar as covariáveis, foi ajustado um modelo considerando as três covariáveis do estudo: idade (x_{i1}), tratamento (x_{i2}) e sexo (x_{i3}). As estimativas dos parâmetros $\boldsymbol{\theta} = (\beta, \sigma, \gamma_0, \gamma_1, \gamma_2, \gamma_3, a, b)$ estão presentes na Tabela (4.4).

Tabela 4.4 – Estimativas máxima verossimilhança para os parâmetros do modelo de regressão LBBIII para dados grupados.

Parâmetros	Estimativas	Erro padrão	$p - valor$
β	2.00865934	6.541662200	-
σ	0.46968009	0.353522206	-
γ_0	2.88413174	0.890889236	0.0012
γ_1	0.04199152	0.004683613	<0.0001
γ_2	0.18251990	0.091146025	0.0452
γ_3	0.06882961	0.089989206	0.4443
a	0.19706110	0.544553026	-
b	0.48779510	0.449038293	-

Após uma breve análise das covariáveis do modelo, bem como dos resultados da Tabela (4.4), concluiu-se que, a um nível de significância de 5%, a covariável x_{i3}: sexo não é significativa no modelo. Assim, considerou-se um modelo de regressão LBBIII para dados grupados com as covariáveis idade (x_{i1}) e tratamento (x_{i2}).

Para se obter as estimativas e erros-padrão do modelo LBBIII para dados grupados foi utilizado o método de reamostragem bootstrap não-paramétrico apresentado na Seção 3.2.5. Utilizou-se o *software* R para obter as estimativas e os erros-padrão do vetor de parâmetros $\boldsymbol{\theta} = (\beta, \sigma, \gamma_0, \gamma_1, \gamma_2, a, b)$. Foi considerado o logaritmo da função de verossimilhança expresso na equação (3.8) e foram realizadas 300 amostras bootstrap. Com essa quantidade de amostras bootstrap observou-se que as estimativas dos parâmetros estabilizaram. Na Tabela 4.5 encontram-se as estimativas e os erros-padrão do vetor de parâmetros.

Pela Tabela 4.5 vê-se que os erros-padrão considerando o método de máxima verossimilhança não são bons para os parâmetros β, a e b. Ao considerar o método de reamostragem bootstrap os erros-padrão são relativamente bons para todas as covariáveis, porém nota-se que para os parâmetros a e b os erros não são tão bons.

Tabela 4.5 – Estimativas MV e bootstrap para os parâmetros do modelo de regressão LBBIII para dados grupados (sem γ_3).

	Estimativas Máxima Verossimilhança			Estimativas Bootstrap		
Parâmetros	Estimativas	Erro padrão	$p-valor$	Estimativas	Erro padrão	I.C. (95%)
β	1.59084678	6.286506574	-	1.75006721	0.69404194	(0.9374,2.8302)
σ	0.45407987	0.255118445	-	0.49301414	0.27236753	(0.1955,0.9832)
γ_0	2.86634496	1.266948081	0.0237	2.82433059	0.26405884	(2.3643,3.1568)
γ_1	0.04197961	0.004700726	<0.0001	0.04120465	0.00665836	(0.0308,0.0507)
γ_2	0.18449562	0.091070733	0.0428	0.18460312	0.09385609	(0.0221,0.3203)
a	0.23742384	0.976093495	-	0.29711059	0.26507999	(0.0700,0.7337)
b	0.47033184	0.342178240	-	0.59802116	0.56502565	(0.1747,1.4861)

Nota: I.C.=Intervalo de confiança

Observa-se, então, a importância do uso do método bootstrap.

4.3 Modelo log-Burr III para dados grupados

Considerando o caso particular da distribuição LBBIII, ou seja, a distribuição LBIII, primeiramente é ajustado um modelo de regressão para dados grupados sem covariáveis. Para isso utiliza-se as funções de sobrevivência e de máxima verossimilhança expressas nas equações (3.13) e (3.15), respectivamente. Na Tabela 4.6 encontram-se as estimativas dos parâmetros $\boldsymbol{\theta} = (\beta, \sigma, \mu)$ e dos erros-padrão.

Tabela 4.6 – Estimativas de máxima verossimilhança para os parâmetros do modelo de regressão LBIII para dados grupados (sem covariável).

Parâmetros	Estimativas	Erro padrão
β	31.9994828	41.02386328
σ	1.1157723	0.04219508
μ	-0.4301267	1.50984136

Então, construiu-se o gráfico da função de sobrevivência empírica (Kaplan-Meier) e da função de sobrevivência estimada pelo ajuste do modelo LBIII para dados grupados sem covariáveis a fim de verificar se o modelo é adequado. Esse gráfico encontra-se na Figura 4.12.

Observa-se pela Figura 4.12, que apesar dos erros-padrões relativamente ruins, a

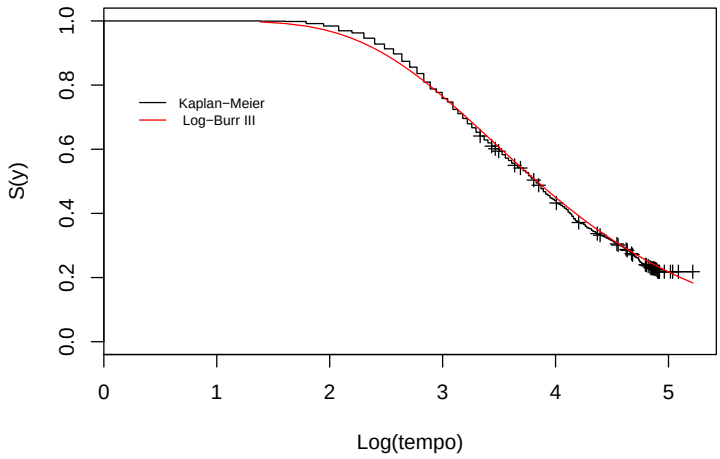

Figura 4.12 – Estimativa da função de sobrevivência para o modelo log-Burr III para dados grupados e Kaplan-Meier para os dados de Vitamina A.

qualidade do ajuste foi boa, visto que a curva do modelo ajustado LBIII acompanha bem o gráfico da função de sobrevivência estimada por Kaplan-Meier.

Então, após o ajuste de um modelo LBIII para dados grupados sem levar em consideração as covariáveis, foi realizado um ajuste considerando as covariáveis idade (x_{i1}) e tratamento (x_{i2}). Como foi dito, a covariável sexo (x_{i3}) não é significativa e por este motivo não foi utilizada no modelo. Utilizou-se o método de máxima verossimilhança para obter as estimativas dos parâmetros $\boldsymbol{\theta} = (\beta, \sigma, \gamma_0, \gamma_1, \gamma_2)$ do modelo log Burr III para dados grupados. Essas estimativas e seus respectivos erros-padrão estão presentes na Tabela 4.7.

Tabela 4.7 – Estimativas de máxima verossimilhança para os parâmetros do modelo de regressão LBIII para dados grupados.

Parâmetros	Estimativas	Erro padrão	$p-valor$
β	0.82614806	0.297368420	-
σ	0.77717560	0.076201949	-
γ_0	2.76068932	0.317056950	<0.0001
γ_1	0.04250712	0.004627486	<0.0001
γ_2	0.18148268	0.090947912	0.0459

Por meio dessa tabela, observa-se que todos os parâmetros têm erros-padrões aceitáveis, indicando que a presença das covariáveis é necessária para o modelo.

4.4 Comparação dos modelos LBBIII e LBIII

Na Seção 3.2 foram propostos os modelos de regressão para dados grupados considerando as distribuições log-beta Burr III e log-Burr III. Esses modelos foram aplicados ao banco de dados de Vitamina A e na Seção 4.2 e 4.3 foram apresentados seus resultados. Nesta seção, o objetivo é comparar os dois modelos propostos neste trabalho. Para isso, utilizou-se três critérios de seleção de modelos: AIC (critério de informação Akaike), AICc (critério de informação Akaike corrigido) e BIC (critério de informação Bayesiano). Akaike (1974), definiu seu critério de informação AIC como:

$$AIC = -2\log L(\hat{\boldsymbol{\theta}}) + 2(p).$$

Enquanto que o AICc foi proposto por Bozdogan (1987) e é definido por:

$$AICc = -2\log L(\hat{\boldsymbol{\theta}}) + 2(p) + 2\frac{p(p+1)}{n-p-1}.$$

O BIC foi proposto por Schwarz (1978) e é definido por:

$$BIC = -2\log L(\hat{\boldsymbol{\theta}}) + p\log(n),$$

em que p é o número de parâmetros a serem estimados no modelo, n é o número de observações da amostra e $L(\hat{\boldsymbol{\theta}})$ é a função de verossimilhança maximizada.

Diante do exposto, na Tabela 4.8 encontra-se um quadro comparativo das estimativas do modelo LBBIII pelo método de reamostragem bootstrap e LBIII pelo método de verossimilhança, bem como os valores dos critérios AIC, AICc e BIC. Pelo fato dos modelos LBBIII e LBIII serem modelos encaixados, pode-se realizar o Teste de Razão de Verossimilhança (TRV). Assim, na Tabela 4.9 encontram-se as estatísticas de Razão de Verossimilhança.

Pelo fato de os valores AIC, AICc e BIC do modelo LBIII serem menor do que o do modelo LBBIII, conclui-se que a distribuição LBIII fornece uma maior evidência favorável a esse modelo para dados de Vitamina A considerando dados grupados.

Portanto, baseando-se em todas as análises e comparações (ver Tabela 4.9) realizadas neste presente trabalho, tem-se que o modelo de regressão log-Burr III para dados grupados é o mais apropriado para o ajuste dos dados de vitamina A. Logo, as estimativas dos parâmetros do modelo final são apresentadas na Tabela 4.7.

Tabela 4.8 – Estimativas dos parâmetros para dados de Vitamina A, respectivos Erros-padrão (em parênteses) e estatística AIC, AICc e BIC.

Parâmetros	LBBIII	LBIII
β	1.75006721	0.82614806
	(0.69404194)	(0.297368420)
σ	0.49301414	0.77717560
	(0.27236753)	(0.076201949)
γ_0	2.82433059	2.76068932
	(0.238106958)	(0.317056950)
γ_1	0.04120465	0.04250712
	(0.00665836)	(0.004627486)
γ_2	0.18460312	0.18148268
	(0.09385609)	(0.090947912)
a	0.29711059	1
	(0.26507999)	-
b	0.59802116	1
	(0.56502565)	-
AIC	4443.972	4440.742
AICc	4444.065	4440.792
BIC	4479.643	4466.221

Tabela 4.9 – Estatística Razão de Verossimilhança para dados de Vitamina A.

Modelo	Hipótese	Estatística w	$p - valor$
LBBIII vs LBIII	H_0 : a=b=1 vs H_1 : H_0 é falso	0.77	0.6804506

Capítulo 5

Conclusões e trabalhos futuros

5.1 Conclusões

Dados de sobrevivência grupados são adequados em situações em que há falhas ou até mesmo censuras nas mesmas unidades de tempo, havendo assim empates.

Neste trabalho foi proposto o modelo de regressão log-beta Burr III para dados grupados, e seu caso particular, o modelo de regressão log-Burr III. Aplicou-se, então, o banco de dados de vitamina A nos modelos apresentados. Após uma breve análise das covariáveis, conclui-se que a covariável *sexo* não é significativa. É importante destacar que considerando os modelos propostos, a covariável $trata-mento$ foi significativa, a um nível de significância de 5%. Assim, tem-se que os indivíduos do estudo que receberam vitamina A possuem um tempo maior de sobrevivência, ou seja, levam mais tempo para apresentarem o evento de interesse.

Os parâmetros dos modelos foram estimados utilizando método de máxima verossimilhança e o método de reamostragem bootstrap. Para as estimações, utilizou-se funções já existentes do $software$ R e outras foram implementadas no mesmo. Verificou-se que o processo de otimização é sensível à escolha dos valores iniciais utilizados nos algoritmos. Com relação ao método de reamostragem bootstrap, os valores inicias utilizados no algoritmo correspondem aos parâmetros estimados pelo método de máxima verossimilhança.

Com os estudos desenvolvidos neste trabalho, observou-se que os modelos log-Burr III e log-beta Burr III para dados grupados foram adequados sem a presença da covariável *sexo*, entretanto, o modelo LBIII mostrou-se mais adequado considerando o conjunto de dados de Vitamina A.

5.2 Trabalhos futuros

Para trabalhos futuros pode-se considerar os seguintes temas:

1. Realizar uma análise de resíduos, bem como de influência global e local para os modelos de regressão LBBIII e LBIII para dados grupados propostos neste trabalho;

2. Considerar outras distribuições de probabilidade originando novos modelos de regressão para dados grupados.

3. Considerar um modelo de regressão para dados grupados com fração de cura.

4. Desenvolver a metodologia de dados de sobrevivência grupados sob o enfoque bayesiano.

Referências Bibliográficas

[1] Aarset, M. V., 1987."How to identify bathtub hazard rate." *IEEE Transactions on Reliability*, 36, 106-108.

[2] Akaike, H., 1974. "A new look at the statistical model identification." *IEEE Transactions on Automatic Control.,Boston,* 19, 6, 716-723.

[3] Al-Dayian, G. R., 1999. "Burr type III distribution: Properties and estimation." *The Egyptian Statistical Journal* 43, 102–116.

[4] Allison, P. D. 1982. "Discrete-time methods for the analysis of event histories." *Sociological Methodology, Washington*, 13, 61-98.

[5] Barreto, M. L.; Santos, L. M. P.; Assis, A. M. O; Araújo, M. P. N.; Farenzena, G. G.; Santos, P. A. B.; Fiaccone, R. L., 1994. "Effect of vitamin A supplementation on diarrhoea and acute lower-respiratory-tract infections in young children in Brazil." *Lancet* 344, 228-231.

[6] Bergman, Bo, e Klefsjõ, B. E. N. G. T., 1984. "Operations Research." 32,596-606.

[7] Bergman, Bo, e Klefsjõ, B. E. N. G. T., 1998. "Recent applications of the TTT-plotting technique." *Frontiers in Reliability. World Scientific New York*, 47-61.

[8] Bolfarine, H.; Rodrigues, J.; Achcar, J.A., 1991. "Análise de sobrevivência." *Associação Brasileira de Estatística, Rio de Janeiro*, 22p.

[9] Bozdogan, H., 1987. "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions." *Psychometrica*, 52, 345-370.

[10] Burr, I. W.,1942. "Cumulative frequency functions." *Annals of Mathematical Statistics, Chicago*, 13, 215-232.

[11] Chalita, L. V., A, S, 1997. "Modelos para dados agrupados e censurados." Tese (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba.

[12] Colosimo, E. A.; Chalita, L. V. A. S.; Demétrio, G. B.,2000. "Tests of proportional hazards and proportional odds models for grouped survival data." *Biometrics, Washington*, 56, 1233-1240.

[13] Colosimo, E. A. e Giolo, S. R., 2006. "Análise de Sobrevivência Aplicada." *Editora Blucher, São Paulo.*

[14] Efron, B., 1979."Bootstrap methods: another look at the jackknife." *The Annuals of Statistics*, 7, 1-26.

[15] Efron, B. e Tibshirani, R. J., 1993. "An introduction to the bootstrap." *Chapman e Hall, New York.*

[16] Eugene, N.; Lee, C. e Famoye, F., 2002. "Beta-normal distribution and its applications." *Communication in Statistics—Theory and Methods* 31, 497–512.

[17] Gomes, A. E.; Da Silva, C. Q.; Cordeiro, G. M. e Ortega, E. M. M., 2013. "The Beta Burr III model of lifetime data." *Brazilian Journal of Probability and Statistics*, 27, 502-543.

[18] Gove, J. H.; Ducey, M. J.; Leak, W. B. e Zhang, L., 2008. "Rotated sigmoid structures in managed uneven-aged northern hardwood stands: A look at the Burr type III distribution." *Forestry* , 81, 161–176.

[19] Hashimoto, E. M.,2008. "Modelo de regressão para dados com censura intervalar e dados de sobrevivência grupados." Dissertação (Mestrado) - Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba.

[20] Hashimoto, E. M.; Ortega, E. M. M; Cordeiro, G. M. e Barreto, M. L.,2012 "The Log-Burr XII Regression Model for Grouped Survival Data." *Journal of Biopharmaceutical Statistics*, 22:1, 141-159.

[21] Hose, G. C., 2005. "Assessing the need for groundwater quality guidlines for pesticides using the species sensitivity distribution approach." *Human and Ecological Risk Assessment*, 11, 951–966.

[22] Kaplan, E. L. e Meier, P., 1958. "Nonparametric estimation from incomplete observations." *Journal of the American Statistical Association*, 53, 457-481.

[23] Kim, J.S. e Proschan, F., 1991. *IEEE Transactions on Realiability*, R-40, 134-139.

[24] Kitchin, J., 1980. "A new method for estimating life distributions from incomplete data." *Dissertação PhD não-publicada - Universidade do Estado da Flórida.*

[25] Lawless, J. F, 2003. "Statistical Models and Methods for Lifetime Data." *John Wiley and Sons, New York, 2nd edition.*

[26] Lepage, R.; Billard, L., 1992. "Exploring the limits of bootstrap." *John Wiley & Sons, New York.*, 632.

[27] Lindsay, S. R.; Wood,G. R. e Woollons R. C., 1996. "Modelling the diameter distribution of forest stands using the Burr distribution." *Journal of Applied Statistics*, 23, 609–620.

[28] Lindsey, J. C.; Ryan, L. M., 1998. "Tutorial in biostatistics methods for interval-censored data." *Statistics in Medicine, Chichester*, 17, 129-238.

[29] Mokhlis, N. A., 2005. "Reliability of a stress-strength model with Burr type III distributions." *Communications in Statistics—Theory and Methods*, 34, 1643–1657.

[30] Prentice, R. L.; Gloeckler, L. A. 1978. "Regression analysis of grouped survival data with application to breast cancer data." *Biometrics, Washington*, 34, 57-67.

[31] Shao, Q., 2000. "Estimation for hazardous concentrations based on NOEC toxicity data: An alternative approach." *Environmetrics,* 11, 583-595.

[32] Shao, Q.; Chen, Y. D. e Zhang, L., 2008. "An extension of three-parameter Burr III distribution for low-flow frequency analysis." /em Computational Statistics and Data Analysis, 52, 1304-1314.

[33] Schwarz, G., 1978. "Estimating the dimensional of a model." *Annals of Statistics, Hayward,* 6, 2, 461-464.

[34] Sherrick, B. J.; Garcia, P. e Tirupattur,V.,1996. "Recovering probabilistic information from option markets: Test of distributional assumptions." *The Journal of Future Markets*, 16, 545–560.

[35] Thompson, W. A, Jr.,1977 "On the treatment of grouped observation in life studies." *Biometrics, Washington*, 33, 463-470.

Apêndice A - Função densidade de probabilidade da log-Burr III

Pelo método do jacobiano a função densidade de probabilidade da log-Burr III, considerando as reparametrizações $\alpha = 1/\sigma$ e $s = \exp(\mu)$, pode ser obtida:

$$f_y(y) = \frac{\alpha\beta}{s(t/s)^{\alpha+1}}\left[\frac{(t/s)^{\alpha}}{1+(t/s)^{\alpha}}\right]^{\beta+1} \times |J| =$$

$$= \frac{\beta/\sigma}{\exp(\mu)[\exp(y)/\exp(\mu)]^{\frac{1}{\sigma}+1}}\left[\frac{[\exp(y)/\exp(\mu)]^{\frac{1}{\sigma}}}{1+[\exp(y)/\exp(\mu)]^{\frac{1}{\sigma}}}\right]^{\beta+1} \times \exp(y) =$$

$$= \frac{\beta}{\sigma\exp\left(\dfrac{y-\mu}{\sigma}\right)}\left[\frac{\exp\left(\dfrac{y-\mu}{\sigma}\right)}{1+\exp\left(\dfrac{y-\mu}{\sigma}\right)}\right]^{\beta+1}$$

$$= \frac{\beta\left[\exp\left(\dfrac{y-\mu}{\sigma}\right)\right]^{\beta}}{\sigma\left[1+\exp\left(\dfrac{y-\mu}{\sigma}\right)\right]^{\beta+1}} \quad ,$$

em que $-\infty < y < +\infty$, $-\infty < \mu < \infty$, $\beta > 0$, $\sigma > 0$ e $|J| = \dfrac{\partial t}{\partial y} = \dfrac{\partial \exp(y)}{\partial y} = \exp(y)$.

A fim de se verificar que $f(y)$ é realmente uma função densidade de probabilidade tem-se que obter:

$$\int_{-\infty}^{\infty} f(u)du = 1.$$

Dessa maneira, utilizando-se a equação 2.13 tem-se que:

$$\int_{-\infty}^{\infty} \frac{\beta \left[\exp\left(\frac{u-\mu}{\sigma}\right)\right]^{\beta}}{\sigma \left[1+\exp\left(\frac{u-\mu}{\sigma}\right)\right]^{\beta+1}} du = 1.$$

Apêndice B - Programa no *software* R para a curva TTT com dados censurados

```
n <- length(time)
    oo <- order(time)
    if(is.null(cens))cens = rep(1, n)
    sorttime <- time[oo]
    sortcens <- cens[oo]
    ttt <- numeric(n)
    ttt[1] <- n * sorttime[1]
    for(i in 2 : n)
    {
    ttt[i] <- ttt[i - 1] + (n - i + 1) * (sorttime[i] - sorttime[i - 1])
    }
    ttt <- data.frame(time = sorttime, cens = sortcens,
    rank = 1 : n, "cumtotaltime" = ttt)
    colnames(ttt) <- c("time","cens","rank","Cum.Total.Time")#,"i/n","TTT")
    TTT <- ttt[ttt$cens == 1, ]
    nfail <- dim(TTT)[1]
    TTT <- cbind(TTT[, -3], TTT$Cum.Total.Time/TTT$Cum.Total.Time[nfail],
    (1 : nfail)/nfail)
    colnames(TTT) <- c("time","cens","Cum.Total.Time","TTT","i/n")
    plot(TTT$"i/n", TTT$TTT, xlab = "r/n", ylab = "G(r/n)", xlim = c(0, 1),
    ylim = c(0, 1), pch = 15, cex = 0.65, type = "l")
    lines(c(0, 1), c(0, 1), lty = 2)
```