Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Dissertação de Mestrado

Modelo de Regressão Log-Logístico discreto com fração de cura para dados de sobrevivência

por

Damião Flávio dos Santos

Orientadora: Profª. Drª. Cira Etheowalda Guevara Otiniano
Coorientadora: Profª. Drª. Juliana Betini Fachini Gomes

Brasília
2017

Damião Flávio dos Santos

Modelo de Regressão Log-Logístico discreto com fração de cura para dados de sobrevivência

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como parte dos requisitos necessários à obtenção do título de Mestre em Estatística.

Brasília
2017

Agradecimentos

Agradeço primeiramente a Deus, pois a realização deste trabalho tornou-se possível graças ao Seu infinito amor e toda Sua graça sobre minha vida.

Aos meus pais José Araújo dos Santos e Maria Lopes dos Santos, que sempre me apoiam em todas as decisões e rezam pelo meu sucesso. A eles que me mostram que a fé, a confiança, o respeito e o amor, são as ferramentas que nos torna pessoas melhores. A eles que ficam felizes com minha felicidade, que choram comigo em momentos difíceis. Aos meus pais, dedico todo o meu sucesso!

Aos meus irmãos Francisco e Francisca, aos meus cunhados Adriana e Claúdio e meus sobrinhos Gustavo e Kauê, que apesar da distância, estão sempre presentes em minha vida.

A minha irmã Fabiana, cunhado Salomão e sobrinha Sthefani que me acolheram em Brasília e que sempre estiveram ao meu lado. Com certeza sem esse apoio, essa caminhada seria mais difícil.

Aos meus primos Dirceu, Daiana, Diego, Daila e Jussara, principalmente por acreditarem em mim e por todo apoio familiar.

A toda minha família, que sempre me deram força e coragem para ir em busca dos meus sonhos.

A minha namorada Anna Hemylia Antunes Pamplona, por ficar ao meu lado nos momentos difíceis e de glória, pela confiança, pelo amor e companheirismo que nos mantém unidos mesmo à 2.120 km de distância.

A família Antunes e Pamplona pelo enorme carinho e respeito.

Aos meus sogros Hélio e Auxiliadora, cunhados Mariane e Mário, por todo apoio, carinho e respeito.

A Dona Sebastiana (in memoriam), por me acolher como neto, por reforçar que o conhecimento é a grande base da vida e que nunca é tarde para aprender.

Aos meus amigos-irmãos Ewerton, Heric, Alex, Aline, Sônia, Deise, Arnete, Pablo, Aninha, André, Izabel, Sueuda, Matheus, Jean, Polyanna, Natasha por acreditarem em mim, pelas orações e por nunca perder o contato.

Aos professores do programa de pós-graduação em Estatística da UnB, pelo excelente trabalho, por toda dedicação, respeito e humildade com os alunos.

As professoras Cira Etheowalda Guevara Otiniano e Juliana Betini Fachini Gomes, pela orientação, por toda a atenção e pelo conhecimento adquirido durante o mestrado. Com certeza as terei para sempre como exemplo de profissionais e de pessoas.

Agradeço ao professor Eduardo Yoshio Nakano pela contribuição na realização deste trabalho e pela atenção por todas às vezes que precisei.

Ao professor Thiago Almeida de Oliveira, por solicitar os dados a Reitoria da Universidade Estadual da Paraíba - Campus I. Sem sua contribuição com certeza esse trabalho teria seguido outros objetivos.

Aos professores Giovana Oliveira Silva e Antônio Eduardo Gomes por aceitarem o convite para participar da banca e por suas contribuições.

A Elisângela e Leandro que foram as pessoas que mais somaram para o meu sucesso durante esses 2 anos de luta. Agradeço imensamente por tudo.

A todos da minha turma do mestrado, que contribuíram para meu aprendizado durante as disciplinas e também pelos bons momentos de confraternização que tivemos.

A todos os funcionários do Departamento de Estatística da UnB, por toda a atenção, respeito e amizade para comigo.

Por fim, a todos que de forma direta ou indireta contribuíram para a realização deste trabalho. Muito obrigado!

Resumo

As técnicas de análise de sobrevivência têm como princípio analisar o tempo até a ocorrência de um determinado evento de interesse. Esse tempo pode ser caracterizado como contínuo ou discreto, e diante disso, as análises serão distintas. Neste trabalho é apresentada uma formulação do modelo de regressão Log-Logístico discreto com fração de cura e sem fração de cura. O comportamento dos estimadores dos parâmetros dos modelos Log-Logístico discreto com fração de cura e sem fração de cura foi avaliado via simulação Monte Carlo com 2.000 réplicas de diferentes tamanhos de amostras. Os modelos propostos foram ilustrados em dois conjuntos de dados sobre a evasão de alunos no ensino superior, nas quais examina-se a influência das covariáveis observadas na variável resposta. As interpretações das estimativas dos parâmetros foram coerentes com as análises descritivas, confirmando um bom ajuste dos modelos propostos. As simulações e análises foram feitas utilizando o *software* livre R.

Palavras-chave: Distribuição Log-Logística Discreta, Fração de Cura, Modelo de Regressão.

Abstract

Survival analysis techniques have as a principle to analyze the time until the occurrence of a particular event of interest, so that this time can be characterized as continuous or discret, in the face of this, the analyses will be distinct. In this work is present a formulation of the discrete time Log-Logistic regression model with healing fraction and no healing fraction. The behavior of the estimators of the parameters of the discrete Log-Logistic models with healing fraction and no healing fraction was evaluated via Monte Carlo simulation with 2.000 replicas of different sample sizes. The proposed models were illustrated in two real data sets on the evasion of students in higher education, in which the influence of the covariates observed in the response variable is examined. The interpretations of the estimates of the parameters were consistent with descriptive analysis, confirming a good adjustment of the proposed models. Simulations and analysis were made using the free *software* R.

Keywords: Distribution Log-Logistic Discrete, Healing Fraction, Regression Model.

Lista de Abreviaturas

S	Suscetíveis
I_S	Indivíduos suscetíveis
N	Não suscetíveis
I_{NS}	Indivíduos não suscetíveis
T_S	Tempo dos indivíduos suscetíveis
T_{NS}	Tempo dos indivíduos não suscetíveis
EMV	Estimadores de máxima verossimilhança
AIC	Critério de informação de Akaike
AICc	Critério de informação de Akaike corrigido
BIC	Critério de informação Bayesiano
EQM	Erro quadrático médio
p.cen	Percentual de censura
cens.t	Percentual de censura total
K-M	Kaplan-Meier
LLD	Log-Logística discreta
LLDFC	Log-Logística discreta com fração de cura
WD	Weibull discreta
WDFC	Weibull discreta com fração de cura
MRLLD	Modelo de regressão Log-Logístico discreto
MRLLDFC	Modelo de regressão Log-Logístico discreto com fração de cura
MRLLDFC1	Primeiro modelo de regressão Log-Logístico discreto com fração de cura
MRLLDFC2	Segundo modelo de regressão Log-Logístico discreto com fração de cura
MRLLDFC3	Terceiro modelo de regressão Log-Logístico discreto com fração de cura

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A análise de sobrevivência é uma das técnicas da estatística que cresceu significativamente nas últimas décadas, devido à possibilidade de aplicação em diversas áreas, com o desenvolvimento de novas técnicas e novas distribuições, combinado com o avanço dos *softwares* estatísticos.

Em análise de sobrevivência, a variável resposta é, geralmente, o tempo até a ocorrência de um evento de interesse. Esse tempo pode ser discreto ou contínuo. Em muitas aplicações com dados discretos, considera-se que os dados "poderiam"ser contínuos e realiza-se a análise utilizando um modelo contínuo. Nakano e Carrasco (2006) estudaram as consequências do uso de um modelo contínuo em dados discretos e mostraram que nem sempre é aceitável esse método, pois em alguns casos pode-se observar um resultado pouco satisfatório.

Sendo a variável resposta o tempo até a ocorrência de um evento de interesse, em algumas situações, é possível ter presença de informações incompletas, denominadas censuradas. Quando a censura ocorre, o indivíduo não chegará a experimentar o evento de interesse por ter saído do estudo ou por ter falhado por razões diferentes das estudadas. Mesmo sendo informações parciais dos indivíduos, as censuras fornecem informações preciosas sobre o tempo decorrido, e desta forma, são importantes para as estimativas. Outra situação em que o indivíduo não experimentará o evento de interesse é o fenômeno denominado fração de cura, ou seja, uma parcela ou fração da população estudada não irá falhar mesmo que o tempo tenda ao infinito.

Uma das distribuições mais utilizadas em estudos de sobrevivência é a Weibull, devido sua flexibilidade. No entanto, segundo Colosimo e Giolo (2006), uma alternativa em muitas situações práticas é a utilização da distribuição Log-Logística.

Além de observar o tempo até a ocorrência de um evento de interesse, é comum observar outras informações sobre os indivíduos em estudo e, a partir disso, torna-se possível identificar quais dessas informações influenciam a variável resposta.

Este trabalho tem como objetivo inicial formular um modelo Log-Logístico discreto (LLD) e um modelo Log-Logístico discreto com fração de cura (LLDFC) para modelar o tempo de sobrevivência de alunos do ensino superior. A acurácia dos estimadores de máxima verossimilhança dos modelos foi testada via simulação Monte Carlo com 2.000 réplicas para diferentes tamanhos de amostras e percentuais de censura. Outro objetivo deste trabalho é examinar a influência das covariáveis na variável resposta. Para realizar esse objetivo, este trabalho propõe um modelo de regressão Log-Logístico com fração de cura e sem fração de cura. Além disso, duas aplicações para o tempo de evasão de alunos de dois cursos do ensino superior foi explorada.

O trabalho está organizado em cinco capítulos. No Capítulo 2, foi feita uma revisão dos conceitos básicos e necessários para desenvolver os capítulos seguintes. No Capítulo 3

desenvolveu-se o modelo de regressão Log-Logístico discreto com fração de cura (MRLLDFC). Esse modelo foi construído gradativamente passando por dois modelos. Ainda neste capítulo, são obtidos os momentos e a função quantil de uma variável aleatória Log-Logística discreta. No Capítulo 4, são apresentadas as simulações de dados com distribuição LLD e LLDFC para testar a acurácia dos estimadores de máxima verossimilhança dos parâmetros dos modelos. Variações de valores de parâmetros, tamanho de amostra e censura foram considerados para todos os modelos. No Capítulo 5, são mostradas duas aplicações dos modelos LLD e LLDFC. Para as simulações e análises estatísticas, utilizou-se o *software* R (TEAM, 2015).

Capítulo 2

Revisão de literatura

Neste capítulo será feita uma revisão de literatura de algumas técnicas de análise de sobrevivência que serão utilizadas para o desenvolvimento deste trabalho. Na Seção 2.1 são apresentados os principais conceitos de análise de sobrevivência, as funções utilizadas e o estimador de Kaplan-Meier. Em seguida, na Seção 2.2, são apresentadas as principais características da distribuição Log-Logística. A forma de discretização de uma variável aleatória contínua é apresentada na Seção 2.3, assim como o desenvolvimento do modelo com fração de cura que é apresentado na Seção 2.4. Na Seção 2.5, são revistos alguns conceitos de inferência clássica, de modo a auxiliar na estimação dos parâmetros do modelo.

2.1 Análise de Sobrevivência

A análise de sobrevivência consiste em uma classe de técnicas e procedimentos estatísticos para estudos com dados em que a variável resposta é, geralmente, o tempo até a ocorrência de um determinado fenômeno de interesse, denominado tempo de falha. Além disso, uma das características principais em análise de sobrevivência é a presença de informações incompletas ou parciais, intituladas como censuras. Mesmo que o indivíduo não tenha experimentado o evento de interesse, não se deve retirá-lo da análise, pois o mesmo contém informações sobre o tempo decorrido até o momento da censura e sua omissão será prejudicial para as estimativas, podendo torná-las viciadas.

Sendo conhecida como análise de sobrevida na área da saúde, ou como análise de confiabilidade na área da engenharia, essas técnicas vêm sendo utilizadas em várias áreas da ciência. Demonstrando, assim, sua flexibilidade no que se refere às áreas de atuação.

É comum esse tipo de análise no âmbito da saúde, em que a variável resposta pode ser o tempo até a morte do paciente, até a cura ou até mesmo a reincidência de uma doença. Em testes de engenharia podem surgir outras escalas de medida, como o número de ciclos, a quilometragem de um carro ou qualquer outra medida de carga (COLOSIMO; GIOLO, 2006).

Ressalta-se a importância de um planejamento e clareza sobre a forma que está sendo conduzido o estudo. Desde a definição das variáveis que serão coletadas até qual método de análise que será utilizado após a coleta. Pois, diante de dados em que não existe a presença de censuras, é possível a utilização de técnicas usuais. Entretanto, na presença de informações incompletas, se faz necessária a inclusão de uma variável indicadora de censura ou falha, em que a mesma é expressa por δ e será igual a 1 quando o indivíduo experimentou o evento de interesse, e igual a 0 caso contrário, como observa-se a seguir:

$$\delta_i = \begin{cases} 0, \text{se o i-ésimo indivíduo foi censurado} \\ 1, \text{se o i-ésimo indivíduo falhou} \end{cases}$$

A censura pode ocorrer de três formas: censura à direita, à esquerda e intervalar. Na forma de censura à direita, o tempo de interesse está à direita do tempo registrado, já na censura à esquerda o evento de interesse está à esquerda do tempo registrado. A censura intervalar é caracterizada pelo fato do acompanhamento dos indivíduos ser em intervalos de tempo e, desta forma, sabe-se apenas que o evento de interesse ocorreu no intervalo entre a visita atual e a anterior, mas não se conhece o tempo exato de falha.

A censura à direita ainda é dividida em três subgrupos: Censura do Tipo I, Censura do Tipo II e Censura Aleatória. Na Figura 2.1 tem-se, em (a), que todos os indivíduos experimentaram o evento de interesse antes do final do estudo, neste caso não há observações censuradas. Em (b), tem-se a representação de dados com a presença de censura do **tipo I**, ou seja, o estudo será terminado após um período de tempo previamente determinado antes do início do estudo, e ao final desse tempo, uma ou mais observações em estudo não falharam. A censura do **tipo II** representada em (c), refere-se ao estudo que termina após ter ocorrido o evento de interesse em um número de indivíduos previamente estabelecido antes do início do estudo. Por fim, em (d), tem-se a representação de dados com censura **aleatória**, que ocorre quando o indivíduo é retirado do estudo sem ter a ocorrência de falha, por motivos não controláveis, ou se o indivíduo falhar por causas diferentes da estudada.

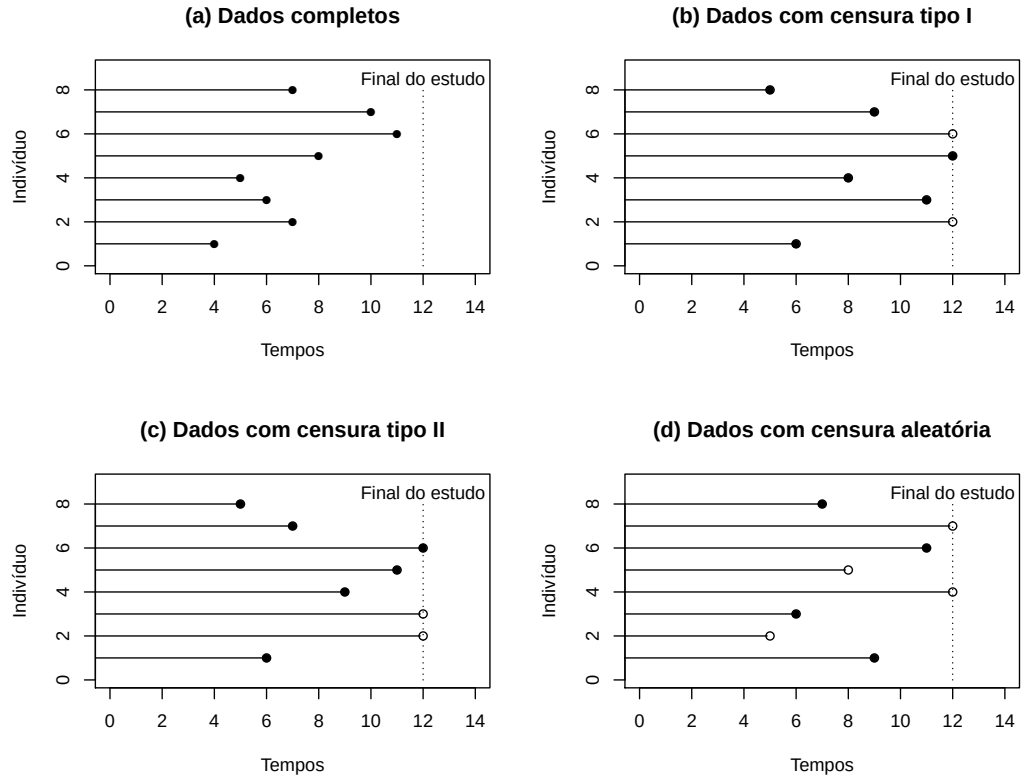


Figura 2.1: Ilustração de alguns mecanismos de censura, em que " • " representa a falha e " ∘ " a censura. Fonte: Adaptado de Colosimo e Giolo (2006)

Além do tempo de sobrevivência e da variável indicadora de censura, também pode-se observar variáveis que representam tanto a heterogeneidade existente na população, quanto possíveis tratamentos aos quais os indivíduos são submetidos. Estas variáveis são conhecidas

como variáveis explicativas ou covariáveis (LOUZADA-NETO; PEREIRA, 2000).

Desta forma, em estudos de sobrevivência é comum o interesse em verificar possíveis correlações entre essas covariáveis e a variável resposta. Em estudos clínicos, é possível identificar possíveis fatores de risco ou prognósticos para uma doença. A Figura 2.2 ilustra o tempo de sobrevivência que é influenciado por três variáveis explicativas, além da interação entre essas. A interação retratada aqui refere-se ao efeito causado entre a combinação de duas ou mais covariáveis no tempo de sobrevivência, sendo que esse efeito pode ser positivo ou negativo.

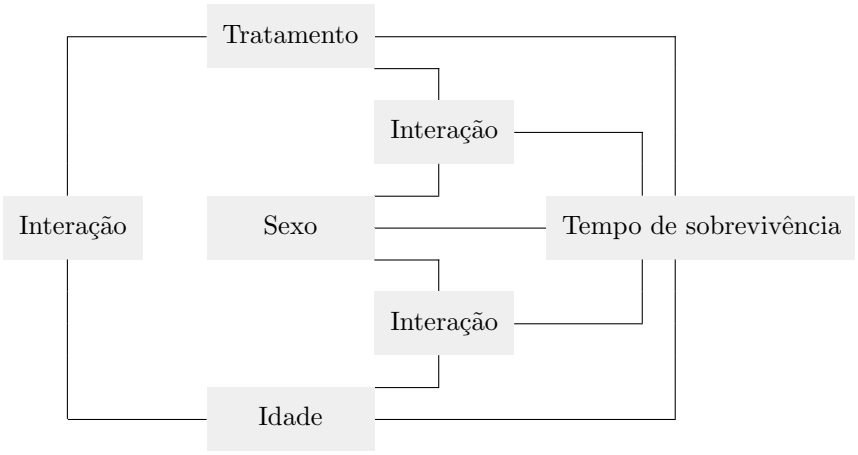


Figura 2.2: Relacionamento entre o tempo de sobrevivência e algumas variáveis explicativas. Fonte: Adaptado de Louzada-Neto e Pereira (2000)

2.1.1 Função de Probabilidade

Seja T uma variável aleatória contínua, sua função densidade de probabilidade é denotada por $f(t)$, de forma que pode ser interpretada como a probabilidade de um indivíduo experimentar o evento de interesse em um intervalo de tempo $[t, t + \Delta t]$ por unidade de Δt, sendo esse o comprimento do intervalo, ou simplesmente por unidade de tempo. A função densidade de probabilidade $f(t)$ é expressa por:

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

em que $f(t) \geq 0$ para todo t e a área abaixo da curva de $f(t)$ é igual a 1. Para o caso discreto, a função de probabilidade é definida como $p(t) = P(T = t)$.

2.1.2 Função de Sobrevivência

A função de sobrevivência $S(t)$ refere-se à probabilidade de um indivíduo sobreviver até o tempo t, que é dada pela seguinte expressão:

$$S(t) = P(T > t). \tag{2.2}$$

É uma das principais funções probabilísticas utilizadas em análise de sobrevivência e tem relação com a função de distribuição acumulada $F(t) = P(T \leq t)$, isto é,

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t).$$

Quando a variável T é contínua e não negativa, no tempo zero tem-se $S(0) = 1$, o que não acontece quando a variável T é discreta, podendo ter ocorrido a falha no tempo zero e $S(0) < 1$, sendo uma das principais características de dados discretos.

Além disso, considerando que todos os indivíduos irão falhar durante o estudo, a probabilidade de sobrevivência por um período de tempo muito grande é 0, ou seja, $\lim_{t\to\infty} S(t) = 0$

Quando a variável T é discreta, ou seja, $t = 0, 1, 2, \ldots$, a função de sobrevivência discreta é dada por:

$$S(t) = P(T > t) = \sum_{k=t+1}^{\infty} P(T = k). \tag{2.3}$$

Na Figura 2.3 é ilustrada a função de sobrevivência no caso contínuo, com um corte em $S(15) = 0,5$, ou seja, em $t = 15$ metade dos indivíduos experimentaram o evento de interesse.

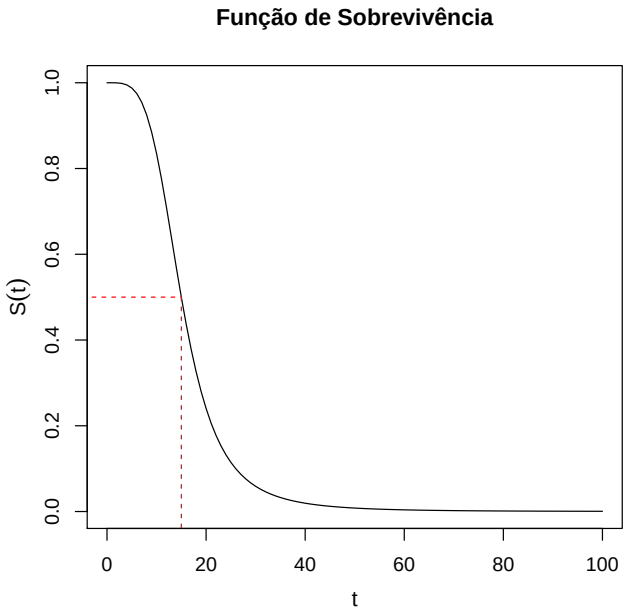


Figura 2.3: Ilustração da função de sobrevivência com corte em $S(15) = 0,5$

2.1.3 Função de Risco

A função de risco se caracteriza por permitir que seja analisado o risco de um indivíduo experimentar o evento de interesse em um determinado tempo t, dado que ele já sobreviveu até aquele momento. Denotada por $h(t)$, é também chamada de taxa de falha. Além disso, segundo Carvalho et al. (2011), é importante ressaltar que apesar de se utilizar o nome risco, $h(t)$ é uma taxa, e não uma probabilidade. Se T é uma variável aleatória contínua, a função de risco é expressa por:

$$h(t) = \lim_{\Delta t \to 0} \frac{P[(t \leq T < t + \Delta t) | T \geq t]}{\Delta t}. \tag{2.4}$$

Além disso, $h(t)$ está relacionada com a função $f(t)$ e $S(t)$ da seguinte forma:

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

No caso discreto, de acordo com Fernandes (2013) a função de risco é igual a zero, exceto nos pontos em que pode ocorrer uma falha. Além disso, a função de risco é definida no intervalo $0 \leq h(t) \leq 1$ e pode ser expressa por:

$$h(t) = P(T = t|T \geq t) = \frac{P(T = t)}{P(T \geq t)} = \frac{P(T = t)}{P(T > t) + P(T = t)} = \frac{p(t)}{S(t) + p(t)}. \tag{2.5}$$

Ressalta-se que a função de risco pode assumir diferentes formas: constante, crescente, decrescente, em forma de banheira e unimodal (inversa banheira) .

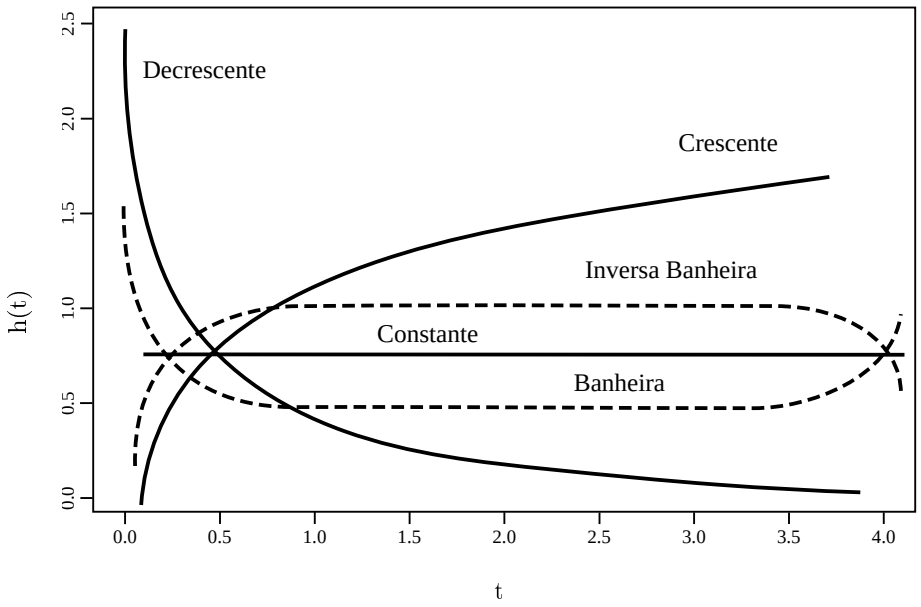


Figura 2.4: Ilustração das diferentes formas da função de risco no caso contínuo. Fonte: Adaptado de Ramos (2014)

Para fazer suposições sobre o modelo que melhor representa os dados em estudo, $h(t)$ em alguns casos é mais informativa do que a $S(t)$. De forma que, diferentes funções de sobrevivência assumem formas semelhantes, enquanto que as funções de risco podem diferir drasticamente.

Outra função utilizada para representar o tempo de sobrevivência é a função de taxa de falha acumulada que é obtida por meio da função de risco, $h(t)$, e é representada por:

$$H(t) = \int_0^t h(u)du. \tag{2.6}$$

Essa função fornece a taxa de falha acumulada do indivíduo e pode ser usada para obter $h(t)$ na estimação não-paramétrica. Outra forma de se obter a função de taxa de falha acumulada é pela relação com a função de sobrevivência, da seguinte forma:

$$H(t) = -[\log(S(t))]. \tag{2.7}$$

Como existem várias formas da função de risco da variável T pode assumir, é importante utilizar uma metodologia gráfica para identificar o modelo mais apropriado para esta variável.

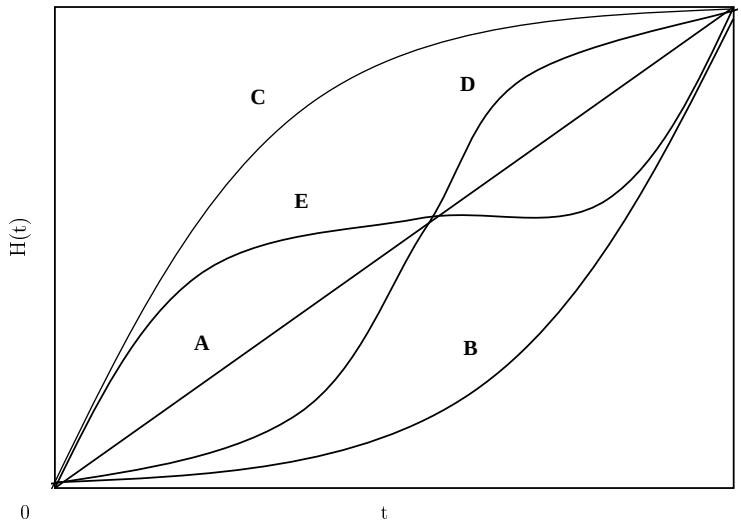


Figura 2.5: Ilustração de algumas formas da função de risco.

Através da Figura 2.5 é possível identificar os seguintes comportamentos da função de risco:

- Reta Diagonal (A): Função de risco constante.
- Curva convexa (B) ou côncava (C): Função de risco monotonicamente crescente ou decrescente, respectivamente.
- Curva convexa e depois côncava (D): Função de risco unimodal.
- Curva côncava e depois convexa (E): Função de risco em forma de banheira ou **U**.

2.1.4 Estimador de Kaplan-Meier

Uma das técnicas mais utilizadas para estimar a função de sobrevivência é o estimador de Kaplan-Meier, sendo conhecido também por limite-produto. Kaplan e Meier (1958) mostraram que $\hat{S}(t)$ é um estimador de máxima verossimilhança não-paramétrico de $S(t)$ e desde então, este estimador vem sendo amplamente utilizado em estudos clínicos e de confiabilidade.

Na ausência de censuras, o estimador de Kaplan-Meier é definido como:

$$\hat{S}(t) = \frac{\text{número de observações que não falharam até o tempo } t}{\text{número total de observações no estudo}}. \tag{2.8}$$

$\hat{S}(t)$ é uma função escada com degraus nos tempos observados de falha de tamanho $1/n$, em que n é o tamanho da amostra. Se existirem empates em um certo tempo t, o tamanho do degrau fica multiplicado pelo número de empates (COLOSIMO; GIOLO, 2006).

Suponha agora que existam n indivíduos no estudo e tem-se $k(\leq n)$ falhas distintas nos tempos $t_1 < t_2 < \ldots < t_k$. O estimador de Kaplan-Meier é expresso por:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j}\right) \tag{2.9}$$

em que, n_j é o número de indivíduos sob risco em t_j, ou seja, aqueles que não falharam e nem foram censurados até o instante imediatamente anterior a t_j, e d_j é o número de falhas em t_j, $j = 1, \ldots, k$.

Ressalta-se que o estimador de Kaplan-Meier é não viciado, é fracamente consistente e além disso, é o estimador de máxima verossimilhança de $S(t)$.

2.2 Distribuição Log-Logística

A distribuição Log-Logística é um modelo de probabilidade utilizada em diversas áreas. Em análise de sobrevivência, devido ao comportamento de sua função de sobrevivência e em finanças e atuária, devido à sua densidade de probabilidade apresentar caudas pesadas. Seja T uma variável aleatória com distribuição Log-Logística. Então, sua função de densidade de probabilidade é dada por:

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1}[1 + (t/\alpha)^\gamma]^{-2}, \tag{2.10}$$

em que $t > 0$, sendo $\alpha > 0$ o parâmetro de escala e $\gamma > 0$ o parâmetro de forma. Uma vantagem do modelo Log-Logístico é apresentar uma expressão simples para as funções de sobrevivência e de risco. A função de sobrevivência é expressa por:

$$S(t) = \frac{1}{1 + (t/\alpha)^\gamma}, \quad t > 0. \tag{2.11}$$

E a função de risco ou taxa de falha é dada por:

$$h(t) = \frac{\gamma(t/\alpha)^{\gamma-1}}{\alpha\,[1 + (t/\alpha)^\gamma]}, \quad t > 0. \tag{2.12}$$

A Figura 2.6 apresenta as formas das funções de densidade de probabilidade, de sobrevivência e de risco para alguns valores de α e γ. Percebe-se que as funções com $\gamma = 1$ tem um comportamento apenas decrescente e com $\gamma > 1$ a densidade de probabilidade e de risco tem formas crescente até um determinado ponto de pico e depois decresce, ou seja, uma função unimodal.

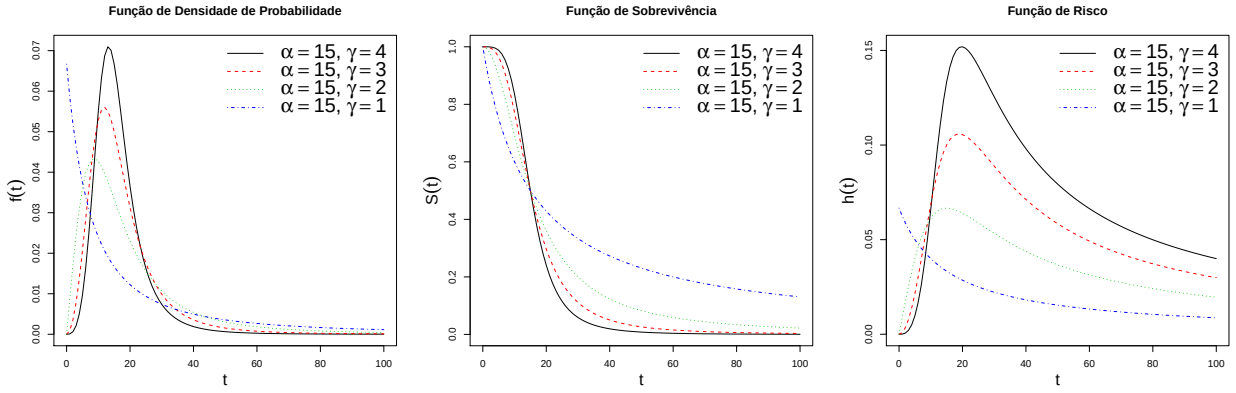


Figura 2.6: Ilustração da forma da densidade de probabilidade, função de sobrevivência e de risco, da distribuição Log-Logística utilizando alguns valores para α e γ.

Para encontrar o tempo mediano de sobrevivência, assim como realizar simulações de dados com distribuição Log-Logística, é necessário a utilização da função quantil da distribuição Log-Logística que é dada por:

$$q_m(m;\alpha,\gamma) = \alpha\left[\frac{m}{(1-m)}\right]^{\frac{1}{\gamma}}, \tag{2.13}$$

sendo $0 \leq m \leq 1$.

2.3 Discretização de Distribuições contínuas

Uma variável aleatória X em um espaço de probabilidade $(\Omega, \mathscr{A}, P)$ é uma função real definida no espaço Ω tal que $[X \leq x]$ é evento aleatório para todo $x \in \mathbb{R}$, i.é, $X : \Omega \to \mathbb{R}$ é variável aleatória $[X \leq x] \in \mathscr{A}, \forall x \in \mathbb{R}$ (JAMES, 1981).

Desta forma, sendo X uma variável aleatória contínua, a variável discreta é obtida por $T = [X]$, sendo $[X]$ a parte inteira de X. Assim, a distribuição de probabilidade de T é definida por:

$$\begin{aligned} p(t) &= P(T = t) \\ &= P(t \leq X < t+1) \\ &= P(X < t+1) - P(X \leq t) \\ &= F_X(t+1) - F_X(t) \\ &= [1 - S_X(t+1)] - [1 - S_X(t)] \\ &= S_X(t) - S_X(t+1), \qquad t = 0, 1, 2, \ldots \end{aligned}$$

A função de sobrevivência é definida da seguinte forma:

$$\begin{aligned} S(t) &= P(T > t) \\ &= \sum_{k=t+1}^{\infty} p(k) \\ &= \sum_{k=t+1}^{\infty} S_X(k) - S_X(k+1) \\ &= S_X(t+1), \qquad t = 0, 1, 2, \ldots, \end{aligned}$$

e a função de risco é dada por:

$$\begin{aligned} h(t) &= \frac{p(t)}{S(t) + p(t)} \\ &= \frac{S_X(t) - S_X(t+1)}{S_X(t+1) + S_X(t) - S_X(t+1)} \\ &= \frac{S_X(t) - S_X(t+1)}{S_X(t)} \\ &= 1 - \frac{S_X(t+1)}{S_X(t)} \qquad t = 0, 1, 2, \ldots \end{aligned}$$

Sendo assim, a partir das distribuições de probabilidade contínuas, é possível definir sua forma discretizada e assim, utilizá-las em análises que o tempo observado é registrado em uma escala discreta. No estudo realizado por Nakano e Carrasco (2006), foi avaliado o uso

de modelos contínuos na análise de dados discretos de sobrevivência, e verificou-se que esse uso pode ser adequado quando a variabilidade dos dados é alta. Além disso, o uso do modelo discreto se mostra mais adequado em conjunto de dados com baixa proporção de censuras.

2.4 Fração de Cura

A análise de sobrevivência tem como pressuposto que em algum momento o indivíduo em estudo irá experimentar o evento de interesse. No entanto, há situações em que, para uma parcela de indivíduos, o evento de interesse não ocorrerá, ainda que o estudo seja observado por longos períodos.

Definido como fração de cura, esse fenômeno ocorre quando, mesmo após um longo período de acompanhamento, o indivíduo não falhará. Considere, por exemplo, que o evento de interesse em um estudo clínico seja a morte de pacientes diagnosticados com câncer e a variável resposta seja o tempo de sobrevivência após o tratamento do indivíduo. Nesses estudos é comum observar indivíduos que se curam da doença após o tratamento e, sendo assim, não experimentará o evento de interesse pelas razões estudadas.

Ao observar o gráfico da função de sobrevivência empírica obtida pelo estimador de Kaplan-Meier é possível ter indicação de fração de cura quando o tempo tende ao infinito e a função de sobrevivência não tende a zero.

Um modelo com fração de cura foi proposto por Berkson e Gage (1952). Esse modelo foi definido como um modelo de mistura em que há uma proporção de indivíduos curados ou imunes na população e uma proporção de indivíduos suscetíveis. Para tanto, a população em estudo é dividida em duas subpopulações, de tal forma que uma seja formada pelos indivíduos que estão suscetíveis à falha (S) e a outra, pelos indivíduos não suscetíveis à falha (NS) ou curados. Dessa forma, é considerada a variável aleatória C_i com distribuição Bernoulli associada a cada indivíduo i para indicar se o i-ésimo indivíduo é suscetível ou não suscetível, isto é,

$$C_i = \begin{cases} 1, & \text{se o i-ésimo indivíduo é suscetível} \\ 0, & \text{se o i-ésimo indivíduo é não suscetível.} \end{cases}$$

Considere agora que $P(C_i = 0) = 1 - \phi$ é a probabilidade da i-ésima observação não ser suscetível ou ser curado, e $P(C_i = 1) = \phi$, a probabilidade da i-ésima observação ser suscetível. Ao considerar que $P(\text{NS}) = 1 - \phi$ com função de sobrevivência $S_{NS}(t)$ e $P(\text{S}) = \phi$ com função de sobrevivência $S_S(t)$. Assim, a função de sobrevivência de forma mista é definida da seguinte forma:

$$\begin{aligned} S_{FC}(t) &= P(NS)P(T > t|NS) + P(S)P(T > t|S) \\ &= (1 - \phi)S_{NS}(t) + \phi S_S(t) \\ &= (1 - \phi) + \phi S_S(t) \end{aligned} \quad .$$

Desta maneira, a função de probabilidade e a função de risco para o caso discreto são definidas, respectivamente por:

$$p_{FC}(t) = \phi p_S(t)$$

e

$$h_{FC}(t) = \frac{\phi p_S(t)}{1 - \phi + \phi S_S(t) + \phi p_S(t)}.$$

Destaca-se que sendo $P(NS) = 1 - \phi = 0$ ou $\phi = 1$, tem-se que $S_{FC}(t) = S_S(t)$ e

$\lim_{t\to\infty} S_{FC}(t) = 1 - \phi$, que é a proporção de indivíduos não suscetíveis ao evento de interesse e $\phi \in [0, 1]$.

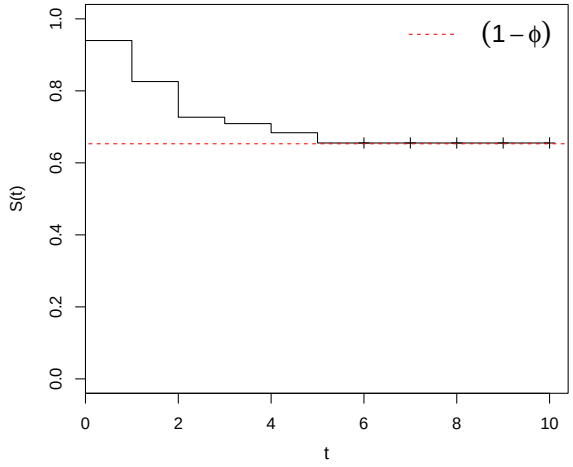


Figura 2.7: Ilustração da forma de uma função de sobrevivência populacional com fração de cura

Ao analisar a Figura 2.7, percebe-se que a função de sobrevivência não converge para zero quando o tempo cresce. Esse fato é um indicativo da presença de indivíduos não suscetíveis ao evento de interesse. Uma estimativa da fração de cura é aproximadamente 0,645, como mostra a linha tracejada.

Destaca-se, ainda, que C_i é uma variável aleatória latente, ou seja, não é observacional. Os valores associados a C_i estão conectados aos valores da variável indicadora δ_i, que é observada no estudo, como mostra a Figura 2.8.

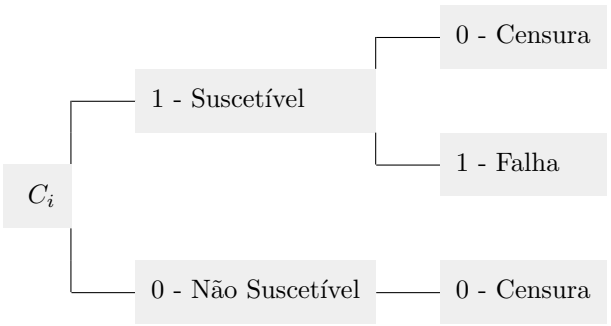


Figura 2.8: Ilustração da variável latente C_i associada a δ_i

2.5 Inferência Estatística

2.5.1 Método de Máxima Verossimilhança

Baseado em uma amostra aleatória observada $t_1, t_2, \ldots, t_n$ de uma variável aleatória discreta T, o método de máxima verossimilhança permite escolher a estimativa dos parâmetros

da distribuição em estudo com maior possibilidade de ter gerado tal amostra. Ao considerar um vetor de parâmetros como sendo $\boldsymbol{\theta}$, a função de verossimilhança para $\boldsymbol{\theta}$ é expressa por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(t_i; \boldsymbol{\theta}). \tag{2.14}$$

A partir dessa função é possível encontrar o valor de $\boldsymbol{\theta}$ que maximiza a probabilidade da amostra observada ocorrer, ou seja, o valor de $\boldsymbol{\theta}$ que maximiza a função $L(\boldsymbol{\theta})$. A expressão (2.14) é utilizada quando todos os indivíduos experimentaram o evento de interesse.

No entanto, em casos em que alguns indivíduos não venham a experimentar o evento de interesse, os dados deverão ser separados de modo que as observações serão divididas em dois conjuntos. Um desses conjuntos com as r observações não censuradas $(1, 2, \ldots, r)$ e, o outro, com as $n - r$ observações censuradas $(r + 1, r + 2, \ldots, n)$. Sendo assim, a função de verossimilhança para dados com a presença de censura, mostrará que a contribuição de cada observação não censurada é $p(t_i; \boldsymbol{\theta})$ e, para cada observação censurada, a contribuição para $L(\boldsymbol{\theta})$ é a sua função de sobrevivência $S(t_i; \boldsymbol{\theta})$. Desta forma, a função de verossimilhança é expressa por:

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{r} p(t_i; \boldsymbol{\theta}) \prod_{i=r+1}^{n} S(t_i; \boldsymbol{\theta}), \tag{2.15}$$

ou ainda, equivalentemente, por:

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} [p(t_i; \boldsymbol{\theta})]^{\delta_i} [S(t_i; \boldsymbol{\theta})]^{(1-\delta_i)}, \tag{2.16}$$

sendo que δ_i é a variável indicadora de falha apresentada na Seção 2.1. Por razões de otimização dos cálculos, é comum trabalhar com o logaritmo da função de verossimilhança. Sendo assim, aplicando o logaritmo em (2.16), tem-se:

$$\log(L(\boldsymbol{\theta})) = \sum_{i=1}^{n} \{\delta_i \log [p\,(t_i; \boldsymbol{\theta})] + (1 - \delta_i) \log [S\,(t_i; \boldsymbol{\theta})]\} + C. \tag{2.17}$$

em que C é uma constante que não depende de $\boldsymbol{\theta}$.

Os estimadores de máxima verossimilhança (EMV) são os valores de $\boldsymbol{\theta}$ que maximizam $L(\boldsymbol{\theta})$ ou, equivalentemente o logaritmo de $L(\boldsymbol{\theta})$. Eles são encontrados resolvendo-se o sistema de equações:

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

A solução deste sistema de equações para um conjunto de dados particular deve ser obtida por meio de um método numérico, o que, usualmente, utiliza-se o método de Newton-Raphson e a utilização de um pacote estatístico. Neste trabalho, o *software* R (TEAM, 2015) foi utilizado para obter $\hat{\boldsymbol{\theta}}$ por meio da função *optim*.

2.5.2 Intervalo de confiança para os parâmetros

Após a estimação dos parâmetros, é importante a construção do intervalo de confiança para os mesmos. O intervalo de confiança é construído a partir da matriz de informação de Fisher definida por:

$$I_f(\boldsymbol{\theta}) = E\left[\left(\frac{\partial \ell(\boldsymbol{t}, \boldsymbol{\delta}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^2\right] = -E\left[\frac{\partial^2 \ell(\boldsymbol{t}, \boldsymbol{\delta}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right],$$

em que $\ell(\boldsymbol{t}, \boldsymbol{\delta}|\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$.

Utiliza-se, então, a distribuição assintótica do estimador de máxima verossimilhança. Para grandes amostras, sob condições de regularidade, essa propriedade estabelece que a distribuição de $\hat{\boldsymbol{\theta}}$ converge assintóticamente para uma distribuição normal multivariada de média $\boldsymbol{\theta}$ e matriz de variância e covariância $Var(\hat{\boldsymbol{\theta}})$, ou seja,

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} N_k\left(\boldsymbol{\theta}, Var(\hat{\boldsymbol{\theta}})\right),$$

em que k é a dimensão de $\boldsymbol{\theta}$. Além disso, tem-se que:

$$Var(\hat{\boldsymbol{\theta}}) \approx \left[I_f(\hat{\boldsymbol{\theta}})\right]^{-1},$$

em que $I_f(\hat{\boldsymbol{\theta}})$ é a informação de Fisher observada da amostra.

Desta forma, um intervalo aproximado de $(1-\alpha)100\%$ de confiança para $\boldsymbol{\theta}$ é dado por:

$$\hat{\boldsymbol{\theta}} \pm Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{\boldsymbol{\theta}})}. \tag{2.18}$$

em que $Z_{1-\alpha/2}$ é o quantil $(1-\alpha/2)$ de uma distribuição normal padrão.

Em alguns casos, torna-se necessário estimar uma função dos parâmetros e uma das propriedades dos estimadores de máxima verossimilhança é a propriedade de invariância, ou seja, se $\hat{\boldsymbol{\theta}}$ é um EMV de $\boldsymbol{\theta}$ e $g(\hat{\boldsymbol{\theta}})$ é uma função bijetora, então $g(\hat{\boldsymbol{\theta}})$ é um EMV de $g(\boldsymbol{\theta})$.

Para a construção de intervalo de confiança para $\boldsymbol{\tau} = g(\boldsymbol{\theta})$, é necessário obter uma estimativa para o erro-padrão de $\hat{\boldsymbol{\tau}} = g(\hat{\boldsymbol{\theta}})$ e isso é feito utilizando o método delta. Além disso, para grandes amostras, tem-se que:

$$g(\hat{\boldsymbol{\theta}}) \overset{a}{\sim} N_k\left(g(\boldsymbol{\theta}), Var(\hat{\boldsymbol{\theta}})[g'(\boldsymbol{\theta})]^2\right), \tag{2.19}$$

ou seja, $\hat{\boldsymbol{\tau}} = g(\hat{\boldsymbol{\theta}})$ converge assintoticamente para uma distribuição normal multivariada com média $\boldsymbol{\tau} = g(\boldsymbol{\theta})$ e matriz de variância e covariância $Var(\hat{\boldsymbol{\theta}})[g'(\boldsymbol{\theta})]^2$, em que k é a dimensão de $g(\hat{\boldsymbol{\theta}})$ e $g'(\boldsymbol{\theta})$ é derivada de primeira ordem de $g(\boldsymbol{\theta})$.

2.5.3 Critérios de Informação

Nesta subseção serão apresentados métodos de seleção baseados na teoria da informação. Os critérios de informação utilizam o valor do logaritmo da verossimilhança do modelo. Os critérios mais utilizados são: critério de Akaike (AIC), Akaike corrigido (AICc) e Informação Bayesiano (BIC).

Critério de informação de Akaike - AIC

O método proposto por Akaike (1974) é conhecido como critério de informação de Akaike (AIC). Sua ideia básica é selecionar um modelo que seja parcimonioso, ou seja, que esteja bem ajustado e tenha um número reduzido de parâmetros. Como o logaritmo da função de verossimilhança cresce com o aumento do número de parâmetros do modelo, uma proposta razoável seria encontrar o modelo com menor valor para a função:

$$AIC = -2\log L(\hat{\boldsymbol{\theta}}) + 2p, \tag{2.20}$$

em que p é o número de parâmetros do modelo.

Critério de informação de Akaike corrigido - AICc

Sugiura (1978) propôs uma correção do critério AIC, pois segundo o mesmo, o AIC pode ter um desempenho ruim se existem muitos parâmetros, em comparação com o tamanho da amostra. Desta forma, o AICc é apenas uma correção, de segunda ordem, do viés de AIC, dado pela seguinte expressão:

$$AICc = -2\log L(\hat{\boldsymbol{\theta}}) + 2p + 2\frac{p(p+1)}{n-p-1}, \tag{2.21}$$

em que p é o número de parâmetros a serem estimados e n é o número de observações da amostra.

Critério de informação Bayesiano - BIC

Proposto por Schwarz (1978) o critério de informação Bayesiano (BIC), que é dado por:

$$BIC = -2\log L(\hat{\boldsymbol{\theta}}) + p\log n, \tag{2.22}$$

em que p é o número de parâmetros a serem estimados e n é o número de observações da amostra. Assim como o AIC e AICc, a proposta do BIC é encontrar o modelo com o menor valor para a função descrita em (2.22).

Capítulo 3

Modelo de regressão Log-Logístico discreto com fração de cura

Neste capítulo, será apresentada a construção do modelo de regressão Log-Logístico discreto com fração de cura e sem fração de cura. Na Seção 3.1, é apresentada a distribuição Log-Logística discreta e uma ilustração do comportamento das funções de probabilidade, de sobrevivência e de risco. Na mesma seção, é apresentada a função quantil, a expressão do r-ésimo momento de T, assim como $E(T)$ e $Var(T)$ e a simulação dos dados para uma distribuição Log-Logística discreta com diversos parâmetros α e γ. Na Seção 3.3, tem-se a formulação do modelo de regressão Log-Logístico discreto, e na Seção 3.4, o modelo de regressão Log-Logístico discreto com fração de cura e a forma de interpretação dos coeficientes de regressão dos modelos.

3.1 Distribuição Log-Logística discreta

Ao utilizar o processo de discretização de distribuições contínuas apresentado na Seção 2.3 e ao considerar a variável aleatória com densidade de probabilidade Log-Logística definida na equação (2.10), a função de probabilidade de uma variável aleatória Log-Logística discreta (LLD), bem como, a função de sobrevivência e função de risco são definidas, respectivamente por:

$$p(t;\alpha,\gamma) = \frac{1}{1+(t/\alpha)^{\gamma}} - \frac{1}{1+[(t+1)/\alpha)]^{\gamma}}, \quad t = 0,1,2,\ldots \tag{3.1}$$

$$S(t;\alpha,\gamma) = \frac{1}{1+[(t+1)/\alpha]^{\gamma}}, \quad t = 0,1,2,\ldots \tag{3.2}$$

e

$$h(t;\alpha,\gamma) = 1 - \frac{1+(t/\alpha)^{\gamma}}{1+[(t+1)/\alpha]^{\gamma}}, \quad t = 0,1,2,\ldots \tag{3.3}$$

sendo $\alpha > 0$ o parâmetro de escala e $\gamma > 0$ o parâmetro de forma.

Na Figura 3.1, mostra-se gráficos das funções (3.1), (3.2) e (3.3), para algumas variações dos parâmetros.

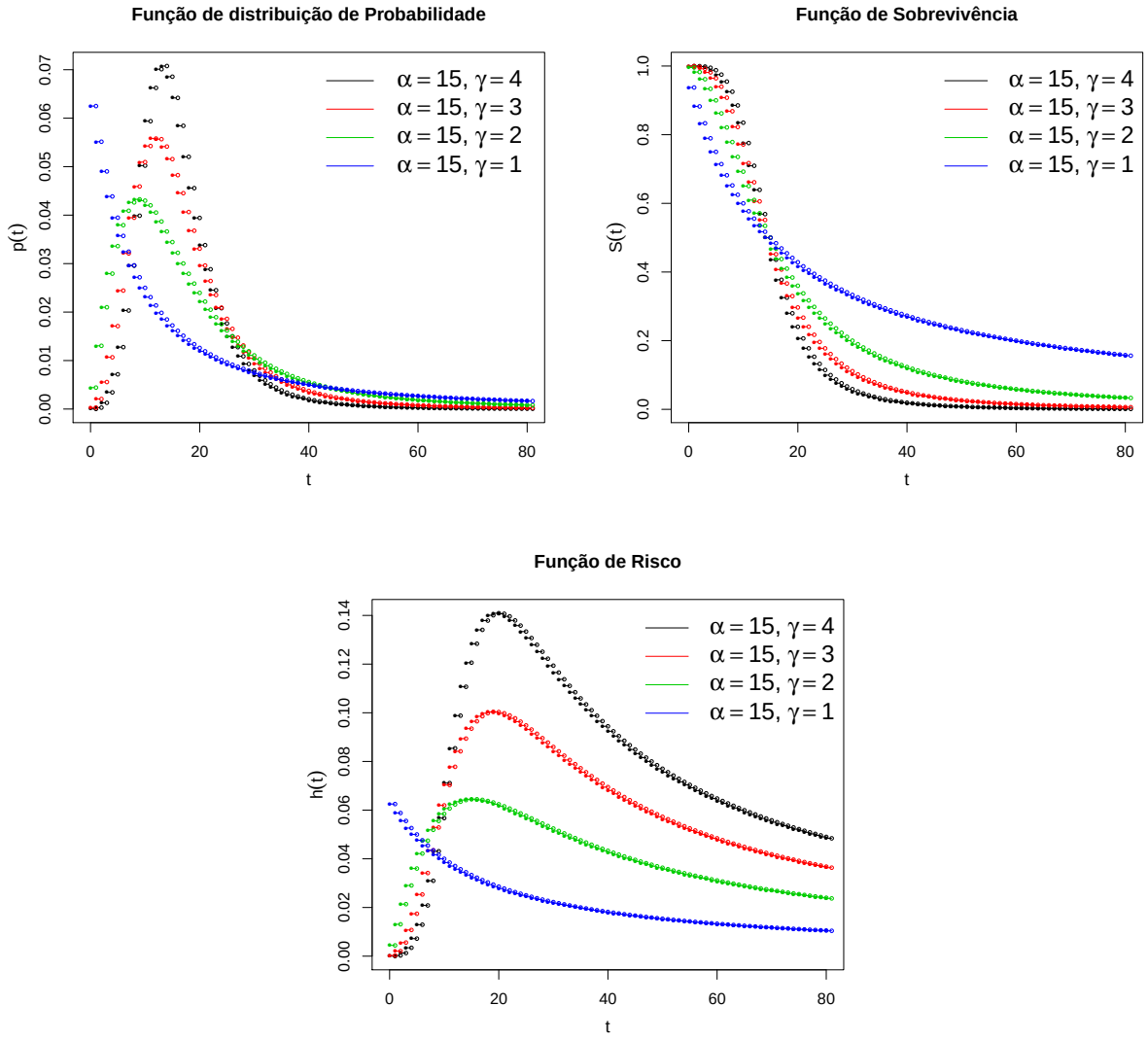


Figura 3.1: Ilustração da forma da função de distribuição de probabilidade, de sobrevivência e de risco da distribuição Log-Logística discreta utilizando alguns valores para α e γ.

Além dessas funções, a função acumulada denotada por $F(t)$ é bastante usada, pois a partir dela é possível encontrar a função quantil. Sendo $S(t)$ uma função de sobrevivência da distribuição Log-Logística discreta, então a $F(t)$ é encontrada usando uma relação com $S(t)$, da seguinte maneira:

$$F(t) = P(T \leq t) = 1 - S(t) = 1 - \frac{1}{1 + [(t+1)/\alpha]^{\gamma}}, \quad t = 0, 1, 2, \ldots. \tag{3.4}$$

Desta forma, a função acumulada da distribuição Log-Logística discreta pode ser reescrita da seguinte maneira:

$$F(t;\alpha,\gamma) = \begin{cases} 0, & t < 0 \\ 1 - \frac{1}{1+\left(\frac{1}{\alpha}\right)^{\gamma}}, & 0 \leq t < 1 \\ 1 - \frac{1}{1+\left(\frac{2}{\alpha}\right)^{\gamma}}, & 1 \leq t < 2 \\ \vdots & \\ 1 - \frac{1}{1+\left(\frac{j+1}{\alpha}\right)^{\gamma}}, & j \leq t < j+1 \end{cases}$$

Na Figura 3.2, tem-se a representação da função acumulada $F(t;\alpha,\gamma)$ com diferentes valores dos parâmetros.

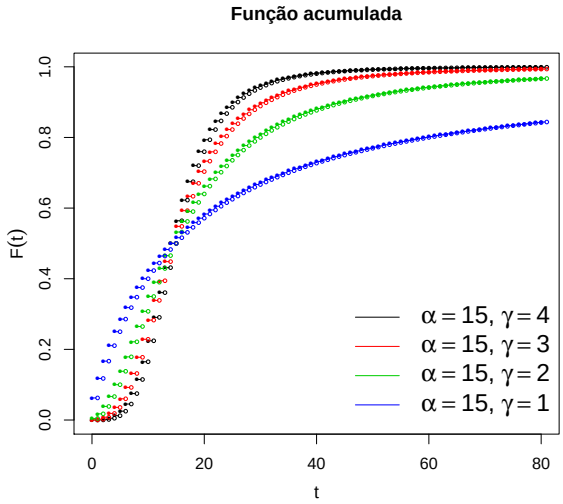


Figura 3.2: Ilustração da forma da função acumulada Log-Logística discreta utilizando alguns valores para α e γ.

Para obter a função quantil de $T \sim LLD(\alpha,\gamma)$, é necessário utilizar a função inversa generalizada da distribuição $F(t)$, definida por:

$$F^{[-1]}(m) = \inf\{t : F(t) \geq m\}, 0 < m \leq 1. \tag{3.5}$$

Ao utilizar a equação (3.5), tem-se que a função quantil da distribuição Log-Logística discreta é dada por:

$$\begin{aligned} q_m(\alpha,\gamma) &= \inf\{t : F(t) \geq m\} = \inf\{t : 1 - S(t) \geq m\} \\ &= \inf\{t : S(t) \leq 1 - m\} \\ &= \inf\left\{t : \frac{1}{1+[(t+1)/\alpha]^{\gamma}} \leq 1 - m\right\} \\ &= \inf\left\{t : \frac{1}{1-m} \leq 1 + [(t+1)/\alpha]^{\gamma}\right\} \\ &= \inf\left\{t : \left[\frac{1}{1-m} - 1\right]^{1/\gamma} \leq \frac{t+1}{\alpha}\right\} \\ &= \inf\left\{t : \alpha\left[\frac{m}{1-m}\right]^{1/\gamma} - 1 \leq t\right\}, \qquad \alpha \geq 1 \end{aligned}$$

Sendo assim, para $m = 0,5$, tem-se a mediana que é dada por:

$$q_{0,5}(\alpha,\gamma) = \begin{cases} \inf\{t : \alpha - 1 \leq t\} & ,\alpha \geq 1 \\ 0 & ,\alpha < 1 \end{cases} \tag{3.6}$$

Seja T uma variável aleatória e $T \sim LLD(\alpha,\gamma)$, o r-ésimo momento de T, dado por:

$$E\left(T^r\right) = \alpha^\gamma \sum_{k=0}^{\infty}\left(\frac{k^r}{\alpha^\gamma + k^\gamma}\right) - \alpha^\gamma \sum_{k=0}^{\infty}\left(\frac{k^r}{\alpha^\gamma + (k+1)^\gamma}\right). \tag{3.7}$$

Para verificar se há alguma restrição para o momento r das séries, utiliza-se o teste da comparação de séries e a partir do primeiro argumento da série definida em (3.7), tem-se que:

$$\frac{k^r}{\alpha^\gamma + k^\gamma} < \frac{k^r}{k^\gamma} = k^{r-\gamma}.$$

Como, $\sum\limits_{k=0}^{\infty} k^{r-\gamma}$ é uma série p, com $p = \gamma - r$, então ela é convergente para $p > 1$. Assim, as séries definidas em (3.7), são convergentes para $r \leq \gamma - 1$. Portanto, o r-ésimo momento dado por (3.7) é finito para um número real r, em que $r \leq \gamma - 1$. Para $r > \gamma - 1$, o r-ésimo momento não existe, o que comprova que a densidade de T tem cauda pesada.

Em particular, quando $r = 1$, a expressão da esperança de T é dada por:

$$E\left(T\right) = \alpha^\gamma \sum_{k=0}^{\infty}\left(\frac{k}{\alpha^\gamma + k^\gamma}\right) - \alpha^\gamma \sum_{k=0}^{\infty}\left(\frac{k}{\alpha^\gamma + (k+1)^\gamma}\right). \tag{3.8}$$

O segundo momento de T é obtido quando $r = 2$ e com isso, encontra-se a variância de T dada por:

$$Var(T) = \alpha^\gamma \left[\sum_{k=0}^{\infty}\left(\frac{k^2}{\alpha^\gamma + k^\gamma}\right) - \sum_{k=0}^{\infty}\left(\frac{k^2}{\alpha^\gamma + (k+1)^\gamma}\right)\right] - E[(T)]^2 \tag{3.9}$$

No intuito de verificar se as séries definidas em (3.8) e (3.9) são expressões calculáveis na prática, realizou-se uma simulação de dados que seguem uma distribuição Log-Logística discreta. Para a simulação dos dados com distribuição Log-Logística discreta será utilizado o método da inversão utilizando a inversa generalizada, a partir do seguinte algoritmo:

1. Gerar $U \sim U(n, 0, 1)$
2. Retorne $T = F^{[-1]}(U)$

Tabela 3.1: Estatísticas para simulação dos dados que seguem uma distribução Log-Logística discreta, com n=100000

$(\alpha;\gamma)$	$E(T)$	$\bar{t}$	$Var(T)$	S^2	$Md(T)$
(35;4)	38,37523	38,34735	413,0257	412,9074	34,00010
(25;4)	27,26802	27,24884	210,7682	210,6832	24,00005
(15;5)	15,53439	15,52563	40,27561	40,19533	14,00007
(7;3)	7,96437	7,95503	46,93826	47,50118	6,00008
(3;2)	4,21237	4,20489	**197,7791**	**146,6396**	2,00011
(3;1)	**34,67820**	**55,77207**	**2.998.546**	**66.831.513**	2,00021

Os resultados da simulação dos dados que seguem uma distribuição Log-Logística discreta, com alguns parâmetros α e γ apresentados na Tabela 3.1, mostram que, tanto os resultados da média quanto da variância são satisfatórios para valores de $\gamma = 4$ e $\gamma = 5$. Isto se deve ao fato de que a condição de convergência é satisfeita, pois $\gamma - 1 > 2$. No entanto, quando a condição não é satisfeita, as equações (3.6) e (3.7) não calculam a média e variância, respectivamente. Além disso, calculou-se a estimativa da mediana de acordo com a equação (3.1) e nota-se que o valor da mediana está de acordo com o valor do parâmetro α menos uma unidade.

As estimativas dos parâmetros do modelo Log-Logístico discreto serão obtidas através do método da máxima verossimilhança, apresentado na Seção 2.5. Deste modo, o logaritmo da função de verossimilhança do modelo Log-Logístico discreto é dado por:

$$\log(L(\boldsymbol{\theta})) = \sum_{i=1}^{n} \left\{ \delta_i \log \left[\frac{1}{1+[t_i/\alpha]^{\gamma}} - \frac{1}{1+[(t_i+1)/\alpha]^{\gamma}} \right] + (1-\delta_i) \log \left[\frac{1}{1+[(t_i+1)/\alpha]^{\gamma}} \right] \right\} + C, \tag{3.10}$$

sendo, $\boldsymbol{\theta} = (\alpha, \gamma)$ e C é uma constante que não depende de $\boldsymbol{\theta}$. Obtém-se as estimativas dos parâmetros derivando a expressão (3.10) e igualando a zero. Com isso, um sistema de equações será obtido e os valores que satisfazem essas equações são os estimadores de máxima verossimilhança para o modelo Log-Logístico discreto.

3.2 Distribuição Log-Logística discreta com fração de cura

A partir do modelo Log-Logístico discreto definido na Seção 3.1 e utilizando o modelo definido na Seção 2.4, é possível obter o modelo Log-Logístico discreto com fração de cura (LLDFC), representado respectivamente, por sua função de probabilidade, função de sobrevivência e função de risco,

$$p(t;\phi,\alpha,\gamma) = \phi \left[\frac{1}{1+(t/\alpha)^{\gamma}} - \frac{1}{1+[(t+1)/\alpha]^{\gamma}} \right], \quad t = 0,1,2,\ldots \tag{3.11}$$

$$S(t;\phi,\alpha,\gamma) = 1 - \phi + \phi \left[\frac{1}{1+[(t+1)/\alpha]^{\gamma}} \right], \quad t = 0,1,2,\ldots \tag{3.12}$$

e

$$h(t;\phi,\alpha,\gamma) = \frac{\phi \left[\frac{1}{1+(t/\alpha)^{\gamma}} - \frac{1}{1+[(t+1)/\alpha]^{\gamma}} \right]}{1 - \phi + \phi \left[\frac{1}{1+(t/\alpha)^{\gamma}} \right]}, \quad t = 0,1,2,\ldots \tag{3.13}$$

em que $0 < \phi < 1$ é o parâmetro de indivíduos suscetíveis, $\alpha > 0$ é o parâmetro de escala e $\gamma > 0$ o parâmetro de forma.

Na Figura 3.3, mostra-se gráficos das funções (3.11), (3.12) e (3.13), para algumas variações dos parâmetros.

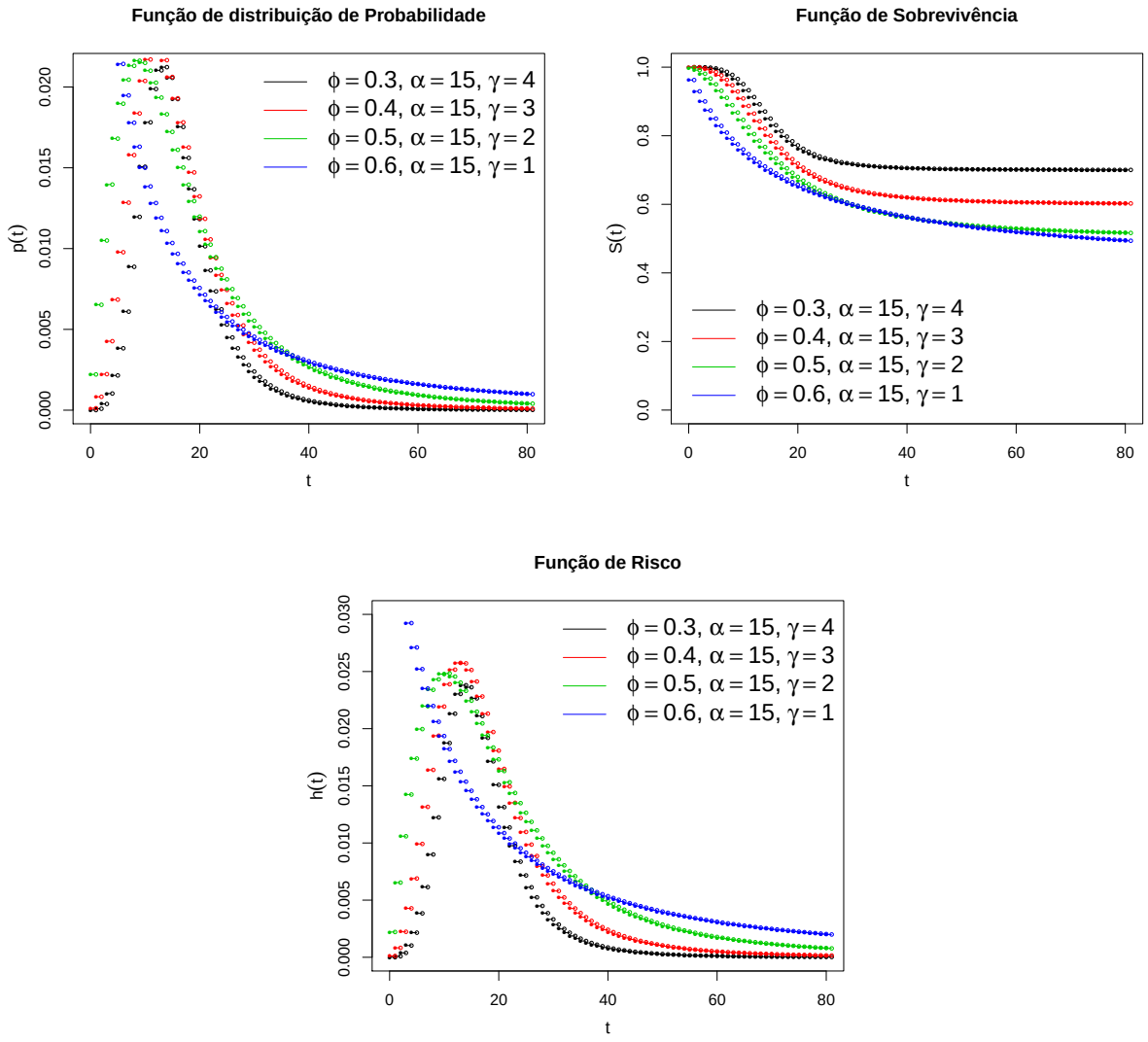


Figura 3.3: Ilustração da forma da função de distribuição de probabilidade, função de sobrevivência e de risco da distribuição Log-Logística discreta com fração de cura, utilizando alguns valores para ϕ, α e γ.

Assim como foi apresentado na Seção 3.1, a função acumulada denotada por $F(t)$ da distribuição LLDFC é definida da seguinte forma:

$$F(t) = P(T \leq t) = 1 - S(t) = \phi - \phi \left[\frac{1}{1 + [(t+1)/\alpha]^{\gamma}} \right], \quad t = 0, 1, 2, \ldots \tag{3.14}$$

Além da expressão (3.14), a $F(t)$ da distribuição LLDFC pode ser reescrita da seguinte maneira:

$$F(t;\phi,\alpha,\gamma) = \begin{cases} 0, & t < 0 \\ \phi - \phi\left[\frac{1}{1+\left(\frac{1}{\alpha}\right)^{\gamma}}\right], & 0 \leq t < 1 \\ \phi - \phi\left[\frac{1}{1+\left(\frac{2}{\alpha}\right)^{\gamma}}\right], & 1 \leq t < 2 \\ \vdots & \\ \phi - \phi\left[\frac{1}{1+\left(\frac{j+1}{\alpha}\right)^{\gamma}}\right], & j \leq t < j+1 \end{cases} \tag{3.15}$$

Na Figura 3.4, tem-se a representação da função acumulada $F(t;\phi,\alpha,\gamma)$ com diferentes valores dos parâmetros.

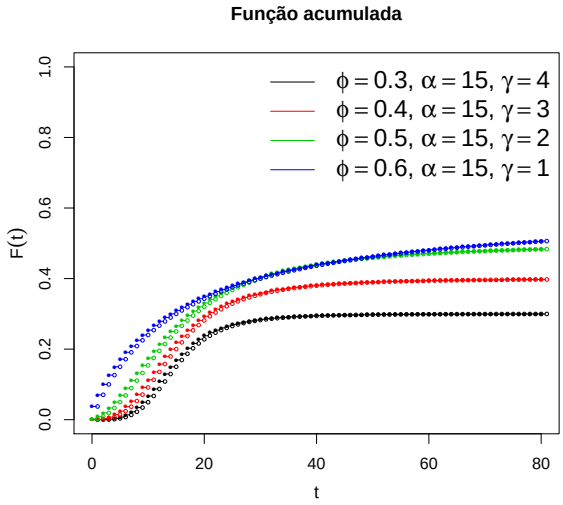


Figura 3.4: Ilustração da forma da função acumulada Log-Logística discreta com fração de cura, utilizando alguns valores para ϕ, α e γ.

Para obter a função quantil de $T \sim LLDFC(\phi,\alpha,\gamma)$, é necessário utilizar a função inversa generalizada definida na equação (3.5). Desta forma, a função quantil da distribuição LLDFC é dada por:

$$q_m(\phi,\alpha,\gamma) = \begin{cases} \inf\left\{t : \alpha\left[\frac{m}{\phi-m}\right]^{1/\gamma} - 1 \leq t\right\}, & \alpha \geq 1 \\ 0, & \alpha < 1, \end{cases} \tag{3.16}$$

para $\phi > m$ e $\gamma > 0$.

Para obter as estimativas dos parâmetros do modelo LLDFC, será utilizado o método da máxima verossimilhança, apresentado na Seção 2.5. Deste modo, o logaritmo da função de verossimilhança do modelo Log-Logístico discreto com fração de cura é dado por:

$$\begin{aligned} \log(L(\boldsymbol{\theta})) = &\sum_{i=1}^{n} \delta_i \log(\phi) + \sum_{i=1}^{n} \delta_i \log\left\{\frac{1}{1+(t_i/\alpha)^{\gamma}} - \frac{1}{1+[(t_i+1)/\alpha]^{\gamma}}\right\} \\ &+ \sum_{i=1}^{n} (1-\delta_i) \log\left\{(1-\phi) + \phi\left[\frac{1}{1+[(t_i+1)/\alpha]^{\gamma}}\right]\right\} + C, \end{aligned} \tag{3.17}$$

sendo, $\boldsymbol{\theta} = (\phi,\alpha,\gamma)$ e C é uma constante que não depende de $\boldsymbol{\theta}$.

Após a estimação, é importante obter um intervalo de confiança para cada um dos

parâmetros. Ao utilizar os parâmetros $\alpha > 0$, $\gamma > 0$ e $0 < \phi < 1$, faz-se necessário realizar uma transformação para construir o intervalo de confiança, pois há restrição no espaço paramétrico. Desta forma, para os parâmetros α e γ foi considerada a transformação logarítmica e para o parâmetro ϕ considerou-se a transformação log-log (Apêndice A).

Assim, os intervalos de confiança para os parâmetros α, γ e ϕ são dados, respectivamente, por:

$$\left[\hat{\alpha}\exp\left(-Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{u})}\right); \hat{\alpha}\exp\left(Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{u})}\right)\right], \quad \hat{u} = \log(\hat{\alpha}), \tag{3.18}$$

$$\left[\hat{\gamma}\exp\left(-Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{v})}\right); \hat{\gamma}\exp\left(Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{v})}\right)\right], \quad \hat{v} = \log(\hat{\gamma}) \tag{3.19}$$

e

$$\left[\hat{\phi}^{\exp\left(Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{w})}\right)}; \hat{\phi}^{\exp\left(-Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{w})}\right)}\right], \quad \hat{w} = \log[-\log(\hat{\phi})]. \tag{3.20}$$

Os valores de $\widehat{Var}(\hat{u})$, $\widehat{Var}(\hat{v})$ e $\widehat{Var}(\hat{w})$, são obtidos numericamente, via *método delta*.

3.3 Modelo de regressão Log-Logístico discreto

Na maioria dos estudos de análise de sobrevivência é possível verificar que algumas covariáveis observadas influenciam o tempo de sobrevivência do indivíduo. O uso dessas covariáveis em um modelo de regressão é uma maneira importante de representar a heterogeneidade em uma população (LAWLESS, 2011).

Sendo $\boldsymbol{x}^T = (1, x_1, \ldots x_p)$ um vetor de covariáveis dos indivíduos em estudo, utiliza-se então uma função de ligação $g(.)$ que conecta a variável resposta às variáveis explicativas. Para um conjunto de p covariáveis, o vetor de parâmetros $\boldsymbol{\theta}$ que será estimado utilizando o vetor $\boldsymbol{x}$, passa a ser definido como:

$$\boldsymbol{\theta} = g(\eta), \tag{3.21}$$

em que $\eta = \boldsymbol{x}^T\boldsymbol{\beta}$ é o preditor linear e $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ é o vetor dos coeficientes de regressão.

Considere T uma variável aleatória com distribuição Log-Logística discreta definida em (3.1) dada por:

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

em que $t = 0, 1, 2, \ldots$, $\alpha > 0$ e $\gamma > 0$. Desta forma, ao utilizar o parâmetro de escala α que é maior que zero, usualmente considera-se $\alpha = g(\eta) = \exp(\boldsymbol{x}^T\boldsymbol{\beta})$. Assim, o modelo de regressão Log-Logístico discreto é definido por:

$$p(t|x) = \frac{1}{1 + [t/\exp(\boldsymbol{x}^T\boldsymbol{\beta})]^\gamma} - \frac{1}{1 + [(t+1)/\exp(\boldsymbol{x}^T\boldsymbol{\beta})]^\gamma}, \quad t = 0, 1, 2, \ldots \tag{3.22}$$

A função de sobrevivência correspondente é dada por:

$$S(t|x) = \frac{1}{1 + [(t+1)/\exp(\boldsymbol{x}^T\boldsymbol{\beta})]^\gamma}, \quad t = 0, 1, 2, \ldots \tag{3.23}$$

e a função de risco é expressa por:

$$h(t|x) = 1 - \frac{1 + \left[t/\exp(\boldsymbol{x}^T\boldsymbol{\beta})\right]^{\gamma}}{1 + [(t+1)/\exp(\boldsymbol{x}^T\boldsymbol{\beta})]^{\gamma}}, \quad t = 0, 1, 2, \ldots \tag{3.24}$$

Para estimar os parâmetros do modelo de regressão Log-Logístico discreto será utilizado o método da máxima verossimilhança. O logaritmo da função de verossimilhança utilizando as funções definidas em (3.22) e (3.23) é expresso por:

$$\begin{aligned}\log(L(\boldsymbol{\theta})) = \sum_{i=1}^{n} \Bigg\{ \delta_i \log \left[\frac{1}{1 + [t_i/\exp(\boldsymbol{x_i}^T\boldsymbol{\beta})]^{\gamma}} - \frac{1}{1 + [(t_i+1)/\exp(\boldsymbol{x_i}^T\boldsymbol{\beta})]^{\gamma}} \right] + \\ (1 - \delta_i) \log \left[\frac{1}{1 + [(t_i+1)/\exp(\boldsymbol{x_i}^T\boldsymbol{\beta})]^{\gamma}} \right] \Bigg\} + C,\end{aligned} \tag{3.25}$$

em que $\boldsymbol{\theta} = (\boldsymbol{\beta}, \gamma)$ e C é uma constante que não depende de $\boldsymbol{\theta}$.

3.4 Modelo de regressão Log-Logístico discreto com fração de cura

Ao utilizar o modelo de fração de cura apresentado na Seção 3.2, neste trabalho se propõe o modelo de regressão Log-Logístico discreto com fração com diferentes formas de inclusão das covariáveis no modelo.

3.4.1 Modelo 1 (MRLLDFC1)

A partir do modelo de regressão Log-Logístico discreto apresentado na Seção 3.3 e o modelo de fração de cura apresentado na Seção 3.2, tem-se o primeiro modelo proposto que inclui as informações das covariáveis no parâmetro α. Assim, as funções de probabilidade, de sobrevivência e de risco associadas ao modelo são definidas, respectivamente, por:

$$p(t|x) = \phi \left\{ \frac{1}{1 + [t/\exp(\boldsymbol{x}^T\boldsymbol{\beta})]^{\gamma}} - \frac{1}{1 + [(t+1)/\exp(\boldsymbol{x}^T\boldsymbol{\beta})]^{\gamma}} \right\}, \quad t = 0, 1, 2, \ldots \tag{3.26}$$

$$S(t|x) = 1 - \phi + \phi \left\{ \frac{1}{1 + [(t+1)/\exp(\boldsymbol{x}^T\boldsymbol{\beta})]^{\gamma}} \right\}, \quad t = 0, 1, 2, \ldots \tag{3.27}$$

e

$$h(t|x) = \frac{\phi \left\{ \frac{1}{1 + [t/\exp(\boldsymbol{x}^T\boldsymbol{\beta})]^{\gamma}} - \frac{1}{1 + [(t+1)/\exp(\boldsymbol{x}^T\boldsymbol{\beta})]^{\gamma}} \right\}}{1 - \phi + \phi \left[\frac{1}{1 + [t/\exp(\boldsymbol{x}^T\boldsymbol{\beta})]^{\gamma}} \right]}, \quad t = 0, 1, 2, \ldots \tag{3.28}$$

em que $0 < \phi < 1$ é o parâmetro dos indivíduos suscetíveis, $\gamma > 0$ é o parâmetro de forma, $\boldsymbol{\beta}_j \in \mathbb{R}, j = 0, 1, \ldots, p$, é o vetor dos coeficientes de regressão e $\boldsymbol{x}^T = (1, x_1, x_2, \ldots, x_p)$ é o vetor de covariáveis associado aos parâmetros desconhecidos $\boldsymbol{\beta}_j$.

Para estimação dos parâmetros do primeiro modelo de regressão Log-Logístico discreto com fração de cura, será utilizado o método da máxima verossimilhança. O logaritmo da

função de verossimilhança do primeiro modelo de regressão Log-Logístico discreto com fração de cura é dado por:

$$\log(L(\boldsymbol{\theta})) = \sum_{i=1}^{n} \delta_i \log(\phi) + \sum_{i=1}^{n} \delta_i \log\left\{\frac{1}{1+[t_i/\exp(\boldsymbol{x_i}^T\boldsymbol{\beta})]^\gamma} - \frac{1}{1+[(t_i+1)/\exp(\boldsymbol{x_i}^T\boldsymbol{\beta})]^\gamma}\right\} + \sum_{i=1}^{n}(1-\delta_i)\log\left\{1-\phi+\phi\left[\frac{1}{1+[(t_i+1)/\exp(\boldsymbol{x_i}^T\boldsymbol{\beta})]^\gamma}\right]\right\} + C, \tag{3.29}$$

sendo, $\boldsymbol{\theta} = (\phi, \boldsymbol{\beta}, \gamma)$ e C é uma constante que não depende de $\boldsymbol{\theta}$.

3.4.2 Modelo 2 (MRLLDFC2)

Ao utilizar o modelo de fração de cura apresentado na Seção 3.2, uma outra forma de obter um modelo de regressão com fração de cura é a partir da inclusão das informações das covariáveis no parâmetro ϕ. Como ϕ assume valores em $[0, 1]$, a relação de ϕ com o vetor de covariáveis $\boldsymbol{x}^T = (1, x_1, x_2, \ldots x_p)$ pode ser feita a partir da função de ligação logito. Desta forma, define-se:

$$\phi(\boldsymbol{\psi}, \boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^T\boldsymbol{\psi})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\psi})}, \tag{3.30}$$

em que $\boldsymbol{\psi} = (\psi_0, \psi_1, \ldots, \psi_p)^T$ é o vetor de p coeficientes de regressão, tal que $-\infty < \boldsymbol{\psi_j} < \infty$, $j = 0, 1, \ldots, p$.

Sendo assim, define-se o modelo de regressão Log-Logístico discreto com fração de cura, utilizando o parâmetro $\phi(\boldsymbol{\psi}, \boldsymbol{x})$ que depende das informações das covariáveis. As funções de probabilidade, de sobrevivência e de risco associadas ao modelo são definidas, respectivamente, por:

$$p(t|x) = \left[\frac{\exp(\boldsymbol{x}^T\boldsymbol{\psi})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\psi})}\right]\left\{\frac{1}{1+(t/\alpha)^\gamma} - \frac{1}{1+[(t+1)/\alpha]^\gamma}\right\}, \quad t = 0, 1, 2, \ldots \tag{3.31}$$

$$S(t|x) = \frac{1}{1+\exp(\boldsymbol{x}^T\boldsymbol{\psi})} + \left[\frac{\exp(\boldsymbol{x}^T\boldsymbol{\psi})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\psi})}\right]\left\{\frac{1}{1+[(t+1)/\alpha]^\gamma}\right\}, \quad t = 0, 1, 2, \ldots \tag{3.32}$$

e

$$h(t|x) = \frac{\left[\frac{\exp(\boldsymbol{x}^T\boldsymbol{\psi})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\psi})}\right]\left\{\frac{1}{1+(t/\alpha)^\gamma} - \frac{1}{1+[(t+1)/\alpha]^\gamma}\right\}}{\frac{1}{1+\exp(\boldsymbol{x}^T\boldsymbol{\psi})} + \left[\frac{\exp(\boldsymbol{x}^T\boldsymbol{\psi})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\psi})}\right]\left[\frac{1}{1+(t/\alpha)^\gamma}\right]}, \quad t = 0, 1, 2, \ldots \tag{3.33}$$

em que $\alpha > 0$ é o parâmetro de escala, $\gamma > 0$ é o parâmetro de forma e $\boldsymbol{\psi} = (\psi_0, \psi_1, \ldots, \phi_p)^T$ é o vetor dos coeficientes de regressão que representam o efeito das covariáveis na proporção de indivíduos não curados.

Para estimação dos parâmetros do segundo modelo de regressão Log-Logístico discreto com fração de cura, será utilizado o método da máxima verossimilhança. Desta forma, o logaritmo da função de verossimilhança do segundo modelo de regressão Log-Logístico discreto com fração de cura é dado por:

$$\log(L(\boldsymbol{\theta})) = \sum_{i=1}^{n} \delta_i \log\left[\frac{\exp(\boldsymbol{x_i}^T\boldsymbol{\psi})}{1+\exp(\boldsymbol{x_i}^T\boldsymbol{\psi})}\right] + \sum_{i=1}^{n} \delta_i \log\left\{\frac{1}{1+(t_i/\alpha)^\gamma} - \frac{1}{1+[(t_i+1)/\alpha]^\gamma}\right\}$$
$$+ \sum_{i=1}^{n} (1-\delta_i) \log\left\{\left[\frac{1}{1+\exp(\boldsymbol{x_i}^T\boldsymbol{\psi})}\right] + \left[\frac{\exp(\boldsymbol{x_i}^T\boldsymbol{\psi})}{1+\exp(\boldsymbol{x_i}^T\boldsymbol{\psi})}\right]\left[\frac{1}{1+[(t_i+1)/\alpha]^\gamma}\right]\right\} + C, \tag{3.34}$$

sendo, $\boldsymbol{\theta} = (\boldsymbol{\psi}, \alpha, \gamma)$ e C é uma constante que não depende de $\boldsymbol{\theta}$.

3.4.3 Modelo 3 (MRLLDFC3)

Ao utilizar o modelo de regressão Log-Logístico discreto com fração de cura apresentado na Subseção 3.4.1 e na Subseção 3.4.2, uma outra maneira de obter um modelo de regressão com fração de cura é a partir da inclusão das informações das covariáveis nos parâmetros α e ϕ, simultaneamente.

Sendo assim, define-se o modelo de regressão Log-Logístico discreto com fração de cura, utilizando o parâmetro $\alpha(\boldsymbol{\beta}, \boldsymbol{x}^T)$ e $\phi(\boldsymbol{\psi}, \boldsymbol{z}^T)$ que dependem, respectivamente, dos vetores de covariáveis $\boldsymbol{x}^T = (1, x_1, x_2, \ldots x_p)$, $\boldsymbol{z}^T = (1, z_1, z_2, \ldots z_p)$ e do vetor de parâmetros desconhecidos $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$, que avalia a influência das covariáveis no tempo de sobrevivência, e o vetor de parâmetros desconhecido $\boldsymbol{\psi} = (\psi_0, \psi_1, \psi_2, \ldots, \psi_p)$, que representa o efeito das covariáveis na proporção de indivíduos não curados. Assim, as funções de probabilidade, de sobrevivência e de risco associadas ao modelo são definidas, respectivamente, por:

$$p(t|x) = \left[\frac{\exp(\boldsymbol{z}^T\boldsymbol{\psi})}{1+\exp\{\boldsymbol{z}^T\boldsymbol{\psi}\}}\right]\left\{\frac{1}{1+[t/\exp(\boldsymbol{x}^T\boldsymbol{\beta})]^\gamma} - \frac{1}{1+[(t+1)/\exp(\boldsymbol{x}^T\boldsymbol{\beta})]^\gamma}\right\}, \quad t = 0, 1, 2, \ldots \tag{3.35}$$

$$S(t|x) = \frac{1}{1+\exp(\boldsymbol{z}^T\boldsymbol{\psi})} + \left[\frac{\exp(\boldsymbol{z}^T\boldsymbol{\psi})}{1+\exp(\boldsymbol{z}^T\boldsymbol{\psi})}\right]\left\{\frac{1}{1+[(t+1)/\exp(\boldsymbol{x}^T\boldsymbol{\beta})]^\gamma}\right\}, \quad t = 0, 1, 2, \ldots \tag{3.36}$$

e

$$h(t|x) = \frac{\left[\frac{\exp(\boldsymbol{z}^T\boldsymbol{\psi})}{1+\exp(\boldsymbol{z}^T\boldsymbol{\psi})}\right]\left\{\frac{1}{1+(t/\exp(\boldsymbol{x}^T\boldsymbol{\beta}))^\gamma} - \frac{1}{1+[(t+1)/\exp(\boldsymbol{x}^T\boldsymbol{\beta})]^\gamma}\right\}}{\frac{1}{1+\exp(\boldsymbol{z}^T\boldsymbol{\psi})} + \left[\frac{\exp(\boldsymbol{z}^T\boldsymbol{\psi})}{1+\exp(\boldsymbol{z}^T\boldsymbol{\psi})}\right]\left\{\frac{1}{1+[t/\exp(\boldsymbol{x}^T\boldsymbol{\beta})]^\gamma}\right\}}, \quad t = 0, 1, 2, \ldots \tag{3.37}$$

em que $\gamma > 0$ é o parâmetro de forma , $-\infty < \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p) < \infty$ é o vetor dos coeficientes de regressão que avalia a influência das covariáveis no tempo de sobrevivência e $-\infty < \boldsymbol{\psi} = (\psi_0, \psi_1, \ldots, \psi_p) < \infty$ representa o efeito causado pelas covariáveis sobre a proporção de indivíduos suscetíveis.

Para estimação dos parâmetros do terceiro modelo de regressão Log-Logístico discreto com fração de cura, será utilizado o método da máxima verossimilhança apresentado na Seção 2.5. O logaritmo da função de verossimilhança do terceiro modelo de regressão Log-Logístico discreto com fração de cura é expresso por:

$$\log(L(\boldsymbol{\theta})) = \sum_{i=1}^{n} \delta_i \log\left[\frac{\exp(\boldsymbol{x}^T\boldsymbol{\psi})}{1+\exp(\boldsymbol{x_i}^T\boldsymbol{\psi})}\right] + \sum_{i=1}^{n} \delta_i \log\left\{\frac{1}{1+[t/\exp(\boldsymbol{x_i}^T\boldsymbol{\beta})]^{\gamma}} - \frac{1}{1+[(t_i+1)/\exp(\boldsymbol{x_i}^T\boldsymbol{\beta})]^{\gamma}}\right\}$$
$$+\sum_{i=1}^{n}(1-\delta_i)\log\left\{\left[\frac{1}{1+\exp(\boldsymbol{x_i}^T\boldsymbol{\psi})}\right] + \left[\frac{\exp(\boldsymbol{x_i}^T\boldsymbol{\psi})}{1+\exp(\boldsymbol{x_i}^T\boldsymbol{\psi})}\right]\left\{\frac{1}{1+[(t_i+1)/\exp(\boldsymbol{x_i}^T\boldsymbol{\beta})]^{\gamma}}\right\}\right\} + C, \tag{3.38}$$

sendo, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\psi}, \gamma)$ e C é uma constante que não depende de $\boldsymbol{\theta}$.

A interpretação dos coeficientes de regressão é definida a partir da função quantil dos modelos Log-Logístico discreto e Log-Logístico discreto com fração de cura. Uma proposta de interpretação é a de se fazer uso da razão de tempos medianos (HOSMER; LEMESHOW, 1999). Ao considerar o modelo LLD, tem-se a seguinte relação:

$$t_{0,5}(\hat{\alpha}, \hat{\gamma}) = \inf\left\{t : \hat{\alpha}\left[\frac{0,5}{(1-0,5)}\right]^{\frac{1}{\hat{\gamma}}} - 1 \leq t\right\} \cong \hat{\alpha} - 1,$$

considerando $\hat{\alpha} = \exp(\hat{\beta}_0 + \hat{\beta}_1 x)$, tem-se que $t_{0,5} + 1 = \exp(\hat{\beta}_0 + \hat{\beta}_1 x)$. Sendo assim, a razão de tempo mediano de um modelo de regressão Log-Logístico discreto composto por uma covariável dicotômica é dada pela seguinte expressão:

$$\frac{1 + t_{0,5}(x=1, \hat{\gamma}, \hat{\boldsymbol{\beta}})}{1 + t_{0,5}(x=0, \hat{\gamma}, \hat{\boldsymbol{\beta}})} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{\exp(\hat{\beta}_0)} = e^{\hat{\beta}_1}$$

Logo, se $\hat{\beta}_1$ for positivo, interpreta-se que o tempo mediano mais uma unidade $(t_{0.5} + 1)$ do indivíduo pertencente ao grupo $x = 1$ é $e^{\hat{\beta}_1}$ vezes o tempo mediano mais uma unidade de um indivíduo do grupo $x = 0$. No entanto, caso $\hat{\beta}_1$ for negativo, conclui-se que o tempo $t_{0,5} + 1$ de um indivíduo do grupo $x = 0$ é $e^{-\hat{\beta}_1}$ ou $\frac{1}{e^{\hat{\beta}_1}}$ vezes o tempo $t_{0,5} + 1$ de um indivíduo do grupo $x = 1$.

O mesmo acontece para o primeiro modelo de regressão Log-Logístico com fração de cura, que apresenta a seguinte relação:

$$t_{0,5}(x, \hat{\phi}, \hat{\gamma}, \hat{\boldsymbol{\beta}}) \cong \exp(\hat{\beta}_0 + \hat{\beta}_1 x)\left[\frac{0,5}{(\hat{\phi} - 0,5)}\right]^{\frac{1}{\hat{\gamma}}} - 1,$$

ao considerar a razão de tempos medianos, tem-se que:

$$\frac{1 + t_{0,5}(x=1, \hat{\phi}, \hat{\gamma}, \hat{\boldsymbol{\beta}})}{1 + t_{0,5}(x=0, \hat{\phi}, \hat{\gamma}, \hat{\boldsymbol{\beta}})} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)\left[\frac{0,5}{(\hat{\phi}-0,5)}\right]^{\frac{1}{\hat{\gamma}}}}{\exp(\hat{\beta}_0)\left[\frac{0,5}{(\hat{\phi}-0,5)}\right]^{\frac{1}{\hat{\gamma}}}} = e^{\hat{\beta}_1}.$$

Para o MRLLDFC2, tem-se ainda os coeficientes $\boldsymbol{\psi} = (\psi_0, \psi_1, \ldots, \psi_p)$ que representam as informações das covariáveis sobre o parâmetro ϕ, que representa os indivíduos não curados. Neste caso, verifica-se o efeito que a covariável x causa no parâmetro ϕ. A interpretação dos coeficientes de regressão $\boldsymbol{\psi}$ será feita utilizando os próprios coeficientes. Sendo assim, se $\hat{\psi}_1$ for positivo, interpreta-se que a fração de indivíduos não curados do grupo $x = 1$ é $\hat{\psi}_1$ vezes maior que a fração de indivíduos não curados do grupo $x = 0$, fixadas as demais covariáveis. De modo análogo, se $\hat{\psi}_1$ for negativo, interpreta-se que a fração de indivíduos não curados do grupo $x = 1$ é $\hat{\psi}_1$ vezes menor que a fração de indivíduos não curados do grupo $x = 0$.

A interpretação dos coeficientes de regressão $\hat{\boldsymbol{\beta}}$ do terceiro modelo de regressão Log-Logístico discreto com fração de cura é feita da mesma forma do MRLLDFC1, assim como

a interpretação dos coeficientes de regressão $\hat{\boldsymbol{\psi}}$, que será interpretado da mesma forma do MRLLDFC2. No entanto, as interpretações serão feitas de forma separada, em que o $\hat{\boldsymbol{\beta}}$ representa o efeito das covariáveis no tempo, mantendo o resto todo constante e $\hat{\boldsymbol{\psi}}$ representa o efeito das covariáveis na proporção de indivíduos não curados, mantendo o resto todo constante.

Capítulo 4

Simulações Computacionais

Neste capítulo, serão apresentadas as simulações computacionais envolvendo os modelos apresentados nas Seções 3.1 e 3.2. As simulações foram obtidas por meio do *software* R (TEAM, 2015). Utilizando-se o método de simulação de Monte Carlo, o objetivo foi gerar dados de sobrevivência com distribuição Log-Logística discreta e Log-Logística discreta com fração de cura, e em seguida, obter as estimativas dos parâmetros dos modelos em estudo.

Na Seção 4.2, considerou-se quatro tamanhos de amostras para as 2.000 réplicas de Monte Carlo, sendo $n = 50, n = 100, n = 200$ e $n = 500$. Para gerar os valores do tempo de sobrevivência, utilizou-se do método da inversão utilizando a inversa generalizada, e para as censuras, considerou-se uma variável indicadora com distribuição Bernoulli. Além disso, destaca-se que o mecanismo de censura utilizado foi o de *censura à direita aleatória* e sua inclusão nas amostras geradas, foi independente do tempo de sobrevivência.

Além da variação no tamanho das amostras, buscou-se pesquisar sobre a influência do percentual de censuras nas estimativas geradas. Desta forma, utilizou-se de quatro percentuais específicos, sendo esses $0\%, 10\%, 20\%$ e 30% de censura.

Devido à distribuição LLD apresentar apenas taxa de falha decrescente e unimodal, buscou-se investigar essas duas situações, utilizando-se de diferentes cenários, ou seja, com diversas combinações dos valores dos parâmetros do modelo simulado.

Para as simulações da distribuição LLDFC, na Seção 4.3 considerou-se três tamanhos de amostras, sendo $n = 50, n = 100$ e $n = 500$. Além disso, utilizou-se três cenários para ser investigada a acurácia dos estimadores, por meio do vício e do erro quadrático médio. Considerou-se o percentual de censura de 10% para os indivíduos suscetíveis em todos os cenários.

4.1 Vício e Erro Quadrático Médio (EQM) dos estimadores

Os principais objetivos da realização da simulação de dados em estudos sobre inferência estatística é verificar o comportamento dos estimadores. O conceito de vício ou viés de um estimador se baseia na diferença média entre o valor do parâmetro θ e o valor estimado de $\hat{\theta}$, ou seja:

$$b(\hat{\theta}) = E[\hat{\theta} - \theta] = E[\hat{\theta}] - \theta, \tag{4.1}$$

e, de acordo com Casella e Berger (2010), o estimador cujo viés é identicamente igual a 0 é chamado de estimador não viesado e satisfaz $E[\hat{\theta}] = \theta$ para todo θ.

Outra medida que incorpora a variabilidade do estimador e também seu viés é o Erro Quadrático Médio (EQM), que mede a diferença quadrática entre o estimador $\hat{\theta}$ e o parâmetro θ, ou seja:

$$EQM(\hat{\theta}) = E[\hat{\theta} - \theta]^2 = Var[\hat{\theta}] + [b(\hat{\theta})]^2. \tag{4.2}$$

Para estimadores não viesados, ou seja, tendo $b[\hat{\theta}] = 0$, a expressão (4.2) se reduz a:

$$EQM(\hat{\theta}) = Var[\hat{\theta}]. \tag{4.3}$$

Sendo assim, um cenário desejável referente aos estimadores é que os mesmos sejam não viciados e que tenham EQM's suficientemente pequenos.

4.2 Simulação da distribuição LLD

Para a simulação da distribuição LLD foram considerados os cenários descritos na Tabela 4.1, levando em consideração a distância entre os valores dos parâmetros α e γ, assim como o comportamento da função de risco.

Tabela 4.1: Cenários utilizados na simulação da distribuição LLD.

Cenários	α	γ	Função de risco
1	7,4	7	Unimodal
	7,4	5,2	Unimodal
	7,4	3,2	Unimodal
	7,4	1	Decrescente
2	5	7	Unimodal
	5	5,2	Unimodal
	5	3,2	Unimodal
	5	1	Decrescente
3	3	7	Unimodal
	3	5,2	Unimodal
	3	3,2	Unimodal
	3	1	Decrescente
4	0,9	7	Unimodal
	0,9	5,2	Unimodal
	0,9	3,2	Unimodal
	0,9	1	Decrescente

A simulação dos dados com distribuição LLD, será feita a partir do método da inversão utilizando a função inversa generalizada, com o seguinte algoritmo:

1. Gerar $U \sim U(n, 0, 1)$
2. Retorne $T = F^{[-1]}(U)$
3. Gerar $\delta_i \sim Bin(n, p)$

A seguir, serão apresentadas as tabelas e os gráficos das simulações e, na sequência, serão apresentadas as conclusões das simulações para cada cenário.

Tabela 4.2: Estimativas dos parâmetros do cenário 1, do vício e EQM, via simulação de Monte Carlo com 2000 réplicas para diferentes percentuais de censura.

N		$\boldsymbol{\theta_1}(\alpha=7,4;\gamma=7)$							
		30%		20%		10%		0%	
		α	γ	α	γ	α	γ	α	γ
50	$\hat{\boldsymbol{\theta}}$	8,1709	6,1589	7,8660	6,4912	7,6161	6,8344	7,4029	7,1726
	$b(\hat{\boldsymbol{\theta}})$	0,7709	-0,8411	0,4660	-0,5088	0,2161	-0,1656	0,0029	0,1726
	$EQM(\hat{\boldsymbol{\theta}})$	0,7361	1,4841	0,3245	1,0308	0,1336	0,7878	0,0663	0,7975
100	$\hat{\boldsymbol{\theta}}$	8,1615	6,0342	7,8578	6,3754	7,6081	6,7284	7,4043	7,0863
	$b(\hat{\boldsymbol{\theta}})$	0,7615	-0,9658	0,4578	-0,6246	0,2081	-0,2716	0,0043	0,0863
	$EQM(\hat{\boldsymbol{\theta}})$	0,6481	1,3186	0,2603	0,7793	0,0831	0,4598	0,0342	0,3756
200	$\hat{\boldsymbol{\theta}}$	8,1548	5,9633	7,8548	6,3088	7,6070	6,6714	7,4021	7,0467
	$b(\hat{\boldsymbol{\theta}})$	0,7548	-1,0367	0,4548	-0,6912	0,2070	-0,3286	0,0021	0,0467
	$EQM(\hat{\boldsymbol{\theta}})$	0,6037	1,2509	0,2320	0,6541	0,0627	0,2888	0,0167	0,1952
500	$\hat{\boldsymbol{\theta}}$	8,1489	5,9561	7,8500	6,3030	7,6036	6,6585	7,3996	7,0107
	$b(\hat{\boldsymbol{\theta}})$	0,7489	-1,0439	0,4500	-0,6970	0,2036	-0,3415	**-0,0004**	0,0107
	$EQM(\hat{\boldsymbol{\theta}})$	0,5740	1,1588	0,2125	0,5541	0,0495	0,1859	0,0067	0,0740
N		$\boldsymbol{\theta_2}(\alpha=7,4;\gamma=5,2)$							
		30%		20%		10%		0%	
		α	γ	α	γ	α	γ	α	γ
50	$\hat{\boldsymbol{\theta}}$	8,4078	4,6422	8,0069	4,8678	7,6812	5,1002	7,4033	5,3273
	$b(\hat{\boldsymbol{\theta}})$	1,0078	-0,5578	0,6069	-0,3322	0,2812	-0,0998	0,0033	0,1273
	$EQM(\hat{\boldsymbol{\theta}})$	1,2814	0,7491	0,5670	0,5323	0,2374	0,4213	0,1194	0,4276
100	$\hat{\boldsymbol{\theta}}$	8,3926	4,5529	7,9944	4,7882	7,6696	5,0248	7,4064	5,2593
	$b(\hat{\boldsymbol{\theta}})$	0,9926	-0,6471	0,5944	-0,4118	0,2696	-0,1752	0,0064	0,0593
	$EQM(\hat{\boldsymbol{\theta}})$	1,1138	0,6347	0,4481	0,3868	0,1469	0,2442	0,0629	0,2023
200	$\hat{\boldsymbol{\theta}}$	8,3829	4,5035	7,9897	4,7413	7,6678	4,9855	7,4036	5,2314
	$b(\hat{\boldsymbol{\theta}})$	0,9829	-0,6965	0,5897	-0,4587	0,2678	-0,2145	0,0036	0,0314
	$EQM(\hat{\boldsymbol{\theta}})$	1,0287	0,5834	0,3937	0,3083	0,1075	0,1452	0,0306	0,1038
500	$\hat{\boldsymbol{\theta}}$	8,3748	4,4955	7,9832	4,7344	7,6631	4,9757	7,3998	5,2077
	$b(\hat{\boldsymbol{\theta}})$	0,9748	-0,7045	0,5832	-0,4656	0,2631	-0,2243	**-0,0002**	0,0077
	$EQM(\hat{\boldsymbol{\theta}})$	0,9743	0,5357	0,3580	0,2555	0,0837	0,0889	0,0119	0,0395
N		$\boldsymbol{\theta_3}(\alpha=7,4;\gamma=3,2)$							
		30%		20%		10%		0%	
		α	γ	α	γ	α	γ	α	γ
50	$\hat{\boldsymbol{\theta}}$	8,8839	3,1662	8,2859	3,3062	7,8097	3,4470	7,4123	3,5857
	$b(\hat{\boldsymbol{\theta}})$	1,4839	-0,3338	0,8859	-0,1938	0,4097	-0,0530	0,0123	0,0857
	$EQM(\hat{\boldsymbol{\theta}})$	2,8391	0,3132	1,2428	0,2311	0,5218	0,1879	0,2624	0,1931
100	$\hat{\boldsymbol{\theta}}$	8,8579	3,1071	8,2658	3,2533	7,7914	3,3992	7,4117	3,5397
	$b(\hat{\boldsymbol{\theta}})$	1,4579	-0,3929	0,8658	-0,2467	0,3914	-0,1008	0,0117	0,0397
	$EQM(\hat{\boldsymbol{\theta}})$	2,4300	0,2558	0,9657	0,1609	0,3176	0,1073	0,1367	0,0909
200	$\hat{\boldsymbol{\theta}}$	8,8395	3,0724	8,2557	3,2213	7,7866	3,3724	7,4065	3,5223
	$b(\hat{\boldsymbol{\theta}})$	1,4395	-0,4276	0,8557	-0,2787	0,3866	-0,1276	0,0065	0,0223
	$EQM(\hat{\boldsymbol{\theta}})$	2,2216	0,2292	0,8378	0,1232	0,2300	0,0618	0,0660	0,0474
500	$\hat{\boldsymbol{\theta}}$	8,8261	3,0662	8,2453	3,2157	7,7788	3,3641	7,4004	3,5052
	$b(\hat{\boldsymbol{\theta}})$	1,4261	-0,4338	0,8453	-0,2843	0,3788	-0,1359	0,0004	0,0052
	$EQM(\hat{\boldsymbol{\theta}})$	2,0911	0,2063	0,7559	0,0986	0,1757	0,0358	0,0259	0,0179
N		$\boldsymbol{\theta_4}(\alpha=7,4;\gamma=1)$							
		30%		20%		10%		0%	
		α	γ	α	γ	α	γ	α	γ
50	$\hat{\boldsymbol{\theta}}$	14,1334	0,9114	11,1217	0,9488	9,0890	0,9864	7,5943	1,0221
	$b(\hat{\boldsymbol{\theta}})$	6,7334	-0,0886	3,7217	-0,0512	1,6890	-0,0136	0,1943	0,0221
	$EQM(\hat{\boldsymbol{\theta}})$	**66,3408**	0,0270	24,3114	0,0206	8,9046	0,0173	3,5264	0,0170
100	$\hat{\boldsymbol{\theta}}$	13,7312	0,8947	10,8590	0,9346	8,8934	0,9733	7,5206	1,0099
	$b(\hat{\boldsymbol{\theta}})$	6,3312	-0,1053	3,4590	-0,0654	1,4934	-0,0267	0,1206	0,0099
	$EQM(\hat{\boldsymbol{\theta}})$	**49,0883**	0,0208	16,5671	0,0136	4,8981	0,0096	1,7821	0,0083
200	$\hat{\boldsymbol{\theta}}$	13,5200	0,8844	10,7413	0,9249	8,8231	0,9658	7,4611	1,0058
	$b(\hat{\boldsymbol{\theta}})$	6,1200	-0,1156	3,3413	-0,0751	1,4231	-0,0342	0,0611	0,0058
	$EQM(\hat{\boldsymbol{\theta}})$	**41,6915**	0,0179	13,3293	0,0100	3,2710	0,0054	0,8303	0,0043
500	$\hat{\boldsymbol{\theta}}$	13,3813	0,8830	10,6490	0,9238	8,7639	0,9637	7,4165	1,0013
	$b(\hat{\boldsymbol{\theta}})$	5,9813	-0,1170	3,2490	-0,0762	1,3639	-0,0363	0,0165	0,0013
	$EQM(\hat{\boldsymbol{\theta}})$	**37,3586**	0,0155	11,3924	0,0075	2,3565	0,0029	**0,3209**	**0,0016**

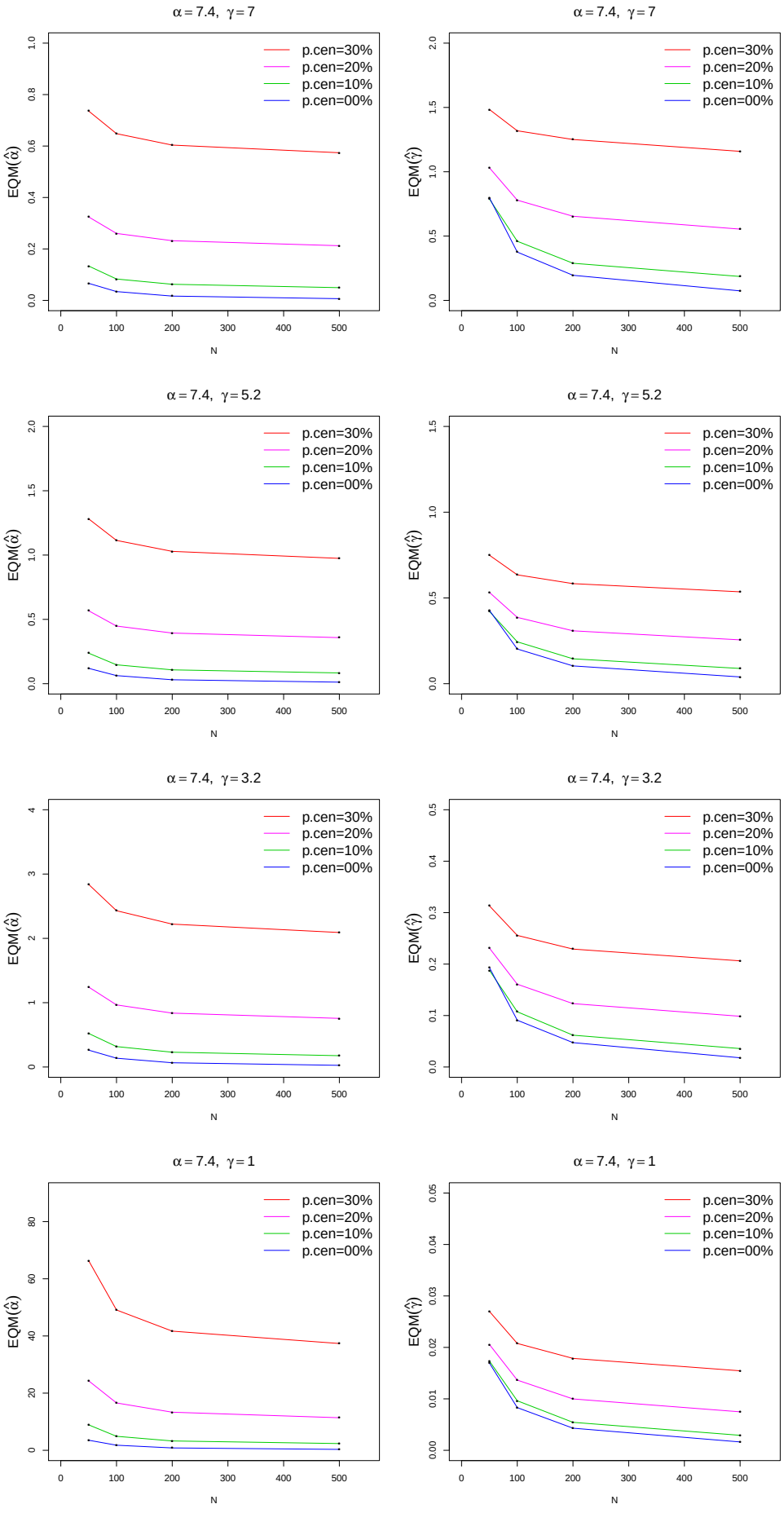


Figura 4.1: Gráfico de linhas dos EQM's dos parâmetros, segundo o tamanho da amostra, para o cenário 1.

Tabela 4.3: Estimativas dos parâmetros do cenário 2, do vício e EQM, via simulação de Monte Carlo com 2000 réplicas para diferentes percentuais de censura.

		$\boldsymbol{\theta_1}(\alpha=5;\gamma=7)$							
N		30%		20%		10%		0%	
		α	γ	α	γ	α	γ	α	γ
50	$\hat{\boldsymbol{\theta}}$	5,5713	5,9850	5,3451	6,3602	5,1599	6,7655	5,0016	7,1775
	$b(\hat{\boldsymbol{\theta}})$	0,5713	-1,0150	0,3451	-0,6398	0,1599	-0,2345	0,0016	0,1775
	$EQM(\hat{\boldsymbol{\theta}})$	0,3968	1,7775	0,1711	1,1731	0,0668	0,8387	0,0312	0,8485
100	$\hat{\boldsymbol{\theta}}$	5,5636	5,8699	5,3387	6,2548	5,1537	6,6625	5,0035	7,0874
	$b(\hat{\boldsymbol{\theta}})$	0,5636	-1,1301	0,3387	-0,7452	0,1537	-0,3375	0,0035	0,0874
	$EQM(\hat{\boldsymbol{\theta}})$	0,3516	1,6553	0,1395	0,9533	0,0430	0,5278	0,0165	0,4087
200	$\hat{\boldsymbol{\theta}}$	5,5593	5,8008	5,3370	6,1848	5,1536	6,6027	5,0021	7,0487
	$b(\hat{\boldsymbol{\theta}})$	0,5593	-1,1992	0,3370	-0,8152	0,1536	-0,3973	0,0021	0,0487
	$EQM(\hat{\boldsymbol{\theta}})$	0,3296	1,6160	0,1257	0,8466	0,0331	0,3486	0,0081	0,2045
500	$\hat{\boldsymbol{\theta}}$	5,5551	5,7902	5,3335	6,1764	5,1511	6,5869	5,0003	7,0149
	$b(\hat{\boldsymbol{\theta}})$	0,5551	-1,2098	0,3335	-0,8236	0,1511	-0,4131	0,0003	0,0149
	$EQM(\hat{\boldsymbol{\theta}})$	0,3146	1,5310	0,1160	0,7464	0,0266	0,2414	0,0032	0,0784
		$\boldsymbol{\theta_2}(\alpha=5;\gamma=5,2)$							
N		30%		20%		10%		0%	
		α	γ	α	γ	α	γ	α	γ
50	$\hat{\boldsymbol{\theta}}$	5,7271	4,5672	5,4372	4,8125	5,2023	5,0722	5,0038	5,3354
	$b(\hat{\boldsymbol{\theta}})$	0,7271	-0,6328	0,4372	-0,3875	0,2023	-0,1278	0,0038	0,1354
	$EQM(\hat{\boldsymbol{\theta}})$	0,6570	0,8320	0,2851	0,5771	0,1150	0,4412	0,0563	0,4685
100	$\hat{\boldsymbol{\theta}}$	5,7171	4,4717	5,4290	4,7289	5,1945	4,9943	5,0048	5,2648
	$b(\hat{\boldsymbol{\theta}})$	0,7171	-0,7283	0,4290	-0,4711	0,1945	-0,2057	0,0048	0,0648
	$EQM(\hat{\boldsymbol{\theta}})$	0,5767	0,7428	0,2294	0,4378	0,0731	0,2579	0,0295	0,2133
200	$\hat{\boldsymbol{\theta}}$	5,7099	4,4201	5,4255	4,6783	5,1931	4,9507	5,0028	5,2327
	$b(\hat{\boldsymbol{\theta}})$	0,7099	-0,7799	0,4255	-0,5217	0,1931	-0,2493	0,0028	0,0327
	$EQM(\hat{\boldsymbol{\theta}})$	0,5342	0,7099	0,2028	0,3731	0,0540	0,1655	0,0143	0,1097
500	$\hat{\boldsymbol{\theta}}$	5,7044	4,4150	5,4211	4,6740	5,1899	4,9417	5,0002	5,2083
	$b(\hat{\boldsymbol{\theta}})$	0,7044	-0,7850	0,4211	-0,5260	0,1899	-0,2583	0,0002	0,0083
	$EQM(\hat{\boldsymbol{\theta}})$	0,5077	0,6544	0,1858	0,3147	0,0428	0,1054	0,0055	0,0425
		$\boldsymbol{\theta_3}(\alpha=5;\gamma=3,2)$							
N		30%		20%		10%		0%	
		α	γ	α	γ	α	γ	α	γ
50	$\hat{\boldsymbol{\theta}}$	6,1471	2,8785	5,6836	3,0120	5,3170	3,1488	5,0104	3,2802
	$b(\hat{\boldsymbol{\theta}})$	1,1471	-0,3215	0,6836	-0,1880	0,3170	-0,0512	0,0104	0,0802
	$EQM(\hat{\boldsymbol{\theta}})$	1,6861	0,2730	0,7283	0,2007	0,2989	0,1632	0,1449	0,1676
100	$\hat{\boldsymbol{\theta}}$	6,1240	2,8198	5,6650	2,9590	5,3001	3,0996	5,0106	3,2397
	$b(\hat{\boldsymbol{\theta}})$	1,1240	-0,3802	0,6650	-0,2410	0,3001	-0,1004	0,0106	0,0397
	$EQM(\hat{\boldsymbol{\theta}})$	1,4401	0,2297	0,5654	0,1426	0,1825	0,0924	0,0760	0,0785
200	$\hat{\boldsymbol{\theta}}$	6,1104	2,7878	5,6580	2,9290	5,2968	3,0731	5,0059	3,2207
	$b(\hat{\boldsymbol{\theta}})$	1,1104	-0,4122	0,6580	-0,2710	0,2968	-0,1269	0,0059	0,0207
	$EQM(\hat{\boldsymbol{\theta}})$	1,3196	0,2092	0,4931	0,1118	0,1331	0,0546	0,0366	0,0404
500	$\hat{\boldsymbol{\theta}}$	6,0984	2,7843	5,6486	2,9260	5,2898	3,0681	5,0009	3,2051
	$b(\hat{\boldsymbol{\theta}})$	1,0984	-0,4157	0,6486	-0,2740	0,2898	-0,1319	0,0009	0,0051
	$EQM(\hat{\boldsymbol{\theta}})$	1,2396	0,1881	0,4442	0,0901	0,1018	0,0322	0,0142	0,0156
		$\boldsymbol{\theta_4}(\alpha=5;\gamma=1)$							
N		30%		20%		10%		0%	
		α	γ	α	γ	α	γ	α	γ
50	$\hat{\boldsymbol{\theta}}$	9,6417	0,9082	7,5578	0,9472	6,1588	0,9864	5,1304	1,0229
	$b(\hat{\boldsymbol{\theta}})$	4,6417	-0,0918	2,5578	-0,0528	1,1588	-0,0136	0,1304	0,0229
	$EQM(\hat{\boldsymbol{\theta}})$	**31,3409**	0,0288	11,3752	0,0219	4,1338	0,0186	1,6188	0,0180
100	$\hat{\boldsymbol{\theta}}$	9,3708	0,8904	7,3830	0,9314	6,0277	0,9719	5,0838	1,0113
	$b(\hat{\boldsymbol{\theta}})$	4,3708	-0,1096	2,3830	-0,0686	1,0277	-0,0281	0,0838	0,0113
	$EQM(\hat{\boldsymbol{\theta}})$	**23,3139**	0,0223	7,8122	0,0146	2,2861	0,0102	0,8169	0,0090
200	$\hat{\boldsymbol{\theta}}$	9,2261	0,8801	7,3022	0,9216	5,9795	0,9639	5,0419	1,0060
	$b(\hat{\boldsymbol{\theta}})$	4,2261	-0,1199	2,3022	-0,0784	0,9795	-0,0361	0,0419	0,0060
	$EQM(\hat{\boldsymbol{\theta}})$	**19,8292**	0,0192	6,3002	0,0108	1,5371	0,0059	0,3815	0,0045
500	$\hat{\boldsymbol{\theta}}$	9,1292	0,8792	7,2401	0,9212	5,9398	0,9624	5,0123	1,0015
	$b(\hat{\boldsymbol{\theta}})$	4,1292	-0,1208	2,2401	-0,0788	0,9398	-0,0376	0,0123	0,0015
	$EQM(\hat{\boldsymbol{\theta}})$	**17,7907**	0,0164	5,4034	0,0080	1,1116	0,0031	**0,1475**	**0,0017**

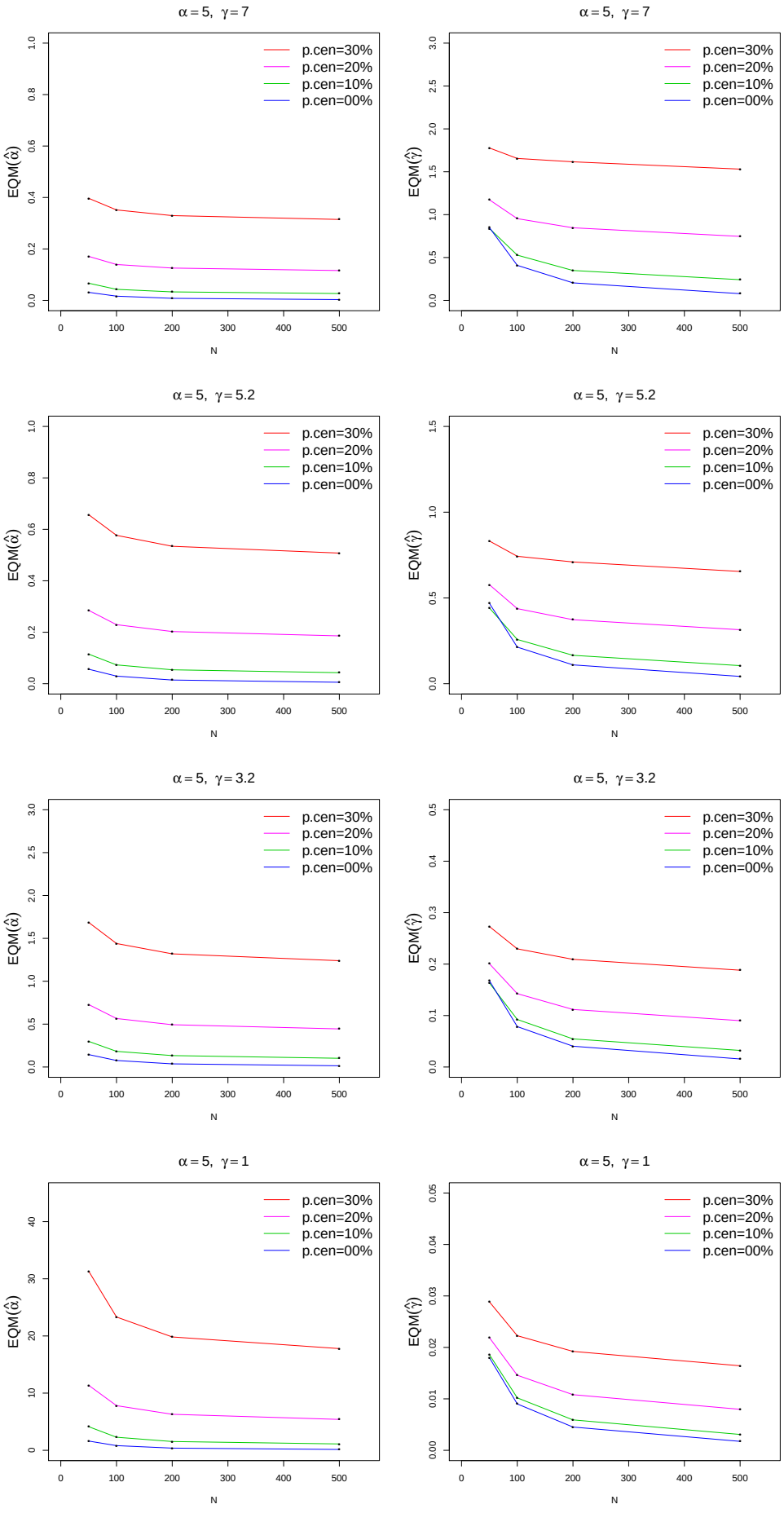


Figura 4.2: Gráfico de linhas dos EQM's dos parâmetros, segundo o tamanho da amostra para o cenário 2.

Tabela 4.4: Estimativas dos parâmetros do cenário 3, do vício e EQM, via simulação de Monte Carlo com 2000 réplicas para diferentes percentuais de censura.

N		$\boldsymbol{\theta_1}(\alpha=3;\gamma=7)$							
		30%		20%		10%		0%	
		α	γ	α	γ	α	γ	α	γ
50	$\hat{\boldsymbol{\theta}}$	3,4069	5,5902	3,2435	6,0607	3,1111	6,6103	3,0019	7,2577
	$b(\hat{\boldsymbol{\theta}})$	0,4069	-1,4098	0,2435	-0,9393	0,1111	-0,3897	0,0019	0,2577
	$EQM(\hat{\boldsymbol{\theta}})$	0,1972	**2,6786**	0,0818	1,6693	0,0294	1,0527	0,0125	1,1682
100	$\hat{\boldsymbol{\theta}}$	3,4036	5,4709	3,2410	5,9400	3,1086	6,4872	3,0018	7,1241
	$b(\hat{\boldsymbol{\theta}})$	0,4036	-1,5291	0,2410	-1,0600	0,1086	-0,5128	0,0018	0,1241
	$EQM(\hat{\boldsymbol{\theta}})$	0,1784	**2,6761**	0,0689	1,5070	0,0198	0,7078	0,0066	0,4965
200	$\hat{\boldsymbol{\theta}}$	3,4013	5,4025	3,2405	5,8643	3,1089	6,4118	3,0015	7,0661
	$b(\hat{\boldsymbol{\theta}})$	0,4013	-1,5975	0,2405	-1,1357	0,1089	-0,5882	0,0015	0,0661
	$EQM(\hat{\boldsymbol{\theta}})$	0,1685	**2,7091**	0,0631	1,4607	0,0158	0,5433	0,0032	0,2518
500	$\hat{\boldsymbol{\theta}}$	3,3976	5,3970	3,2374	5,8615	3,1066	6,3986	3,0004	7,0170
	$b(\hat{\boldsymbol{\theta}})$	0,3976	-1,6030	0,2374	-1,1385	0,1066	-0,6014	0,0004	0,0170
	$EQM(\hat{\boldsymbol{\theta}})$	0,1611	**2,6326**	0,0585	1,3648	0,0130	0,4420	0,0013	0,0953
N		$\boldsymbol{\theta_2}(\alpha=3;\gamma=5,2)$							
		30%		20%		10%		0%	
		α	γ	α	γ	α	γ	α	γ
50	$\hat{\boldsymbol{\theta}}$	3,4979	4,3756	3,2982	4,6691	3,1369	4,9934	3,0024	5,3385
	$b(\hat{\boldsymbol{\theta}})$	0,4979	-0,8244	0,2982	-0,5309	0,1369	-0,2066	0,0024	0,1385
	$EQM(\hat{\boldsymbol{\theta}})$	0,3001	1,1025	0,1263	0,7254	0,0474	0,4995	0,0213	0,5050
100	$\hat{\boldsymbol{\theta}}$	3,4926	4,2905	3,2942	4,5908	3,1332	4,9190	3,0026	5,2683
	$b(\hat{\boldsymbol{\theta}})$	0,4926	-0,9095	0,2942	-0,6092	0,1332	-0,2810	0,0026	0,0683
	$EQM(\hat{\boldsymbol{\theta}})$	0,2682	1,0372	0,1045	0,5931	0,0314	0,3121	0,0113	0,2427
200	$\hat{\boldsymbol{\theta}}$	3,4892	4,2363	3,2932	4,5377	3,1335	4,8735	3,0024	5,2401
	$b(\hat{\boldsymbol{\theta}})$	0,4892	-0,9637	0,2932	-0,6623	0,1335	-0,3265	0,0024	0,0401
	$EQM(\hat{\boldsymbol{\theta}})$	0,2517	1,0270	0,0946	0,5404	0,0243	0,2148	0,0055	0,1245
500	$\hat{\boldsymbol{\theta}}$	3,4845	4,2305	3,2892	4,5327	3,1303	4,8611	3,0007	5,2094
	$b(\hat{\boldsymbol{\theta}})$	0,4845	-0,9695	0,2892	-0,6673	0,1303	-0,3389	0,0007	0,0094
	$EQM(\hat{\boldsymbol{\theta}})$	0,2395	0,9780	0,0871	0,4850	0,0197	0,1574	0,0022	0,0467
N		$\boldsymbol{\theta_3}(\alpha=3;\gamma=3,2)$							
		30%		20%		10%		0%	
		α	γ	α	γ	α	γ	α	γ
50	$\hat{\boldsymbol{\theta}}$	3,7474	2,8171	3,4440	2,9645	3,2046	3,1226	3,0071	3,2810
	$b(\hat{\boldsymbol{\theta}})$	0,7474	-0,3829	0,4440	-0,2355	0,2046	-0,0774	0,0071	0,0810
	$EQM(\hat{\boldsymbol{\theta}})$	0,7015	0,3236	0,2961	0,2284	0,1158	0,1766	0,0538	0,1850
100	$\hat{\boldsymbol{\theta}}$	3,7337	2,7608	3,4333	2,9158	3,1951	3,0741	3,0060	3,2367
	$b(\hat{\boldsymbol{\theta}})$	0,7337	-0,4392	0,4333	-0,2842	0,1951	-0,1259	0,0060	0,0367
	$EQM(\hat{\boldsymbol{\theta}})$	0,6077	0,2788	0,2351	0,1681	0,0731	0,1029	0,0282	0,0854
200	$\hat{\boldsymbol{\theta}}$	3,7242	2,7295	3,4283	2,8855	3,1928	3,0495	3,0042	3,2204
	$b(\hat{\boldsymbol{\theta}})$	0,7242	-0,4705	0,4283	-0,3145	0,1928	-0,1505	0,0042	0,0204
	$EQM(\hat{\boldsymbol{\theta}})$	0,5580	0,2622	0,2063	0,1393	0,0541	0,0643	0,0137	0,0432
500	$\hat{\boldsymbol{\theta}}$	3,7174	2,7245	3,4230	2,8808	3,1889	3,0426	3,0008	3,2057
	$b(\hat{\boldsymbol{\theta}})$	0,7174	-0,4755	0,4230	-0,3192	0,1889	-0,1574	0,0008	0,0057
	$EQM(\hat{\boldsymbol{\theta}})$	0,5275	0,2414	0,1880	0,1170	0,0424	0,0399	0,0054	0,0163
N		$\boldsymbol{\theta_4}(\alpha=3;\gamma=1)$							
		30%		20%		10%		0%	
		α	γ	α	γ	α	γ	α	γ
50	$\hat{\boldsymbol{\theta}}$	5,8940	0,8994	4,5854	0,9401	3,7113	0,9826	3,0749	1,0235
	$b(\hat{\boldsymbol{\theta}})$	2,8940	-0,1006	1,5854	-0,0599	0,7113	-0,0174	0,0749	0,0235
	$EQM(\hat{\boldsymbol{\theta}})$	**12,0556**	0,0325	4,3086	0,0245	1,5309	0,0207	0,5907	0,0200
100	$\hat{\boldsymbol{\theta}}$	5,7288	0,8827	4,4785	0,9257	3,6323	0,9682	3,0470	1,0110
	$b(\hat{\boldsymbol{\theta}})$	2,7288	-0,1173	1,4785	-0,0743	0,6323	-0,0318	0,0470	0,0110
	$EQM(\hat{\boldsymbol{\theta}})$	**9,0322**	0,0250	2,9820	0,0165	0,8566	0,0115	0,3000	0,0101
200	$\hat{\boldsymbol{\theta}}$	5,6393	0,8727	4,4314	0,9163	3,6057	0,9608	3,0223	1,0055
	$b(\hat{\boldsymbol{\theta}})$	2,6393	-0,1273	1,4314	-0,0837	0,6057	-0,0392	0,0223	0,0055
	$EQM(\hat{\boldsymbol{\theta}})$	**7,7051**	0,0216	2,4221	0,0122	0,5798	0,0067	0,1404	0,0050
500	$\hat{\boldsymbol{\theta}}$	5,5843	0,8718	4,3974	0,9159	3,5843	0,9596	3,0059	1,0014
	$b(\hat{\boldsymbol{\theta}})$	2,5843	-0,1282	1,3974	-0,0841	0,5843	-0,0404	0,0059	0,0014
	$EQM(\hat{\boldsymbol{\theta}})$	**6,9548**	0,0185	2,0957	0,0090	0,4251	0,0035	**0,0540**	**0,0019**

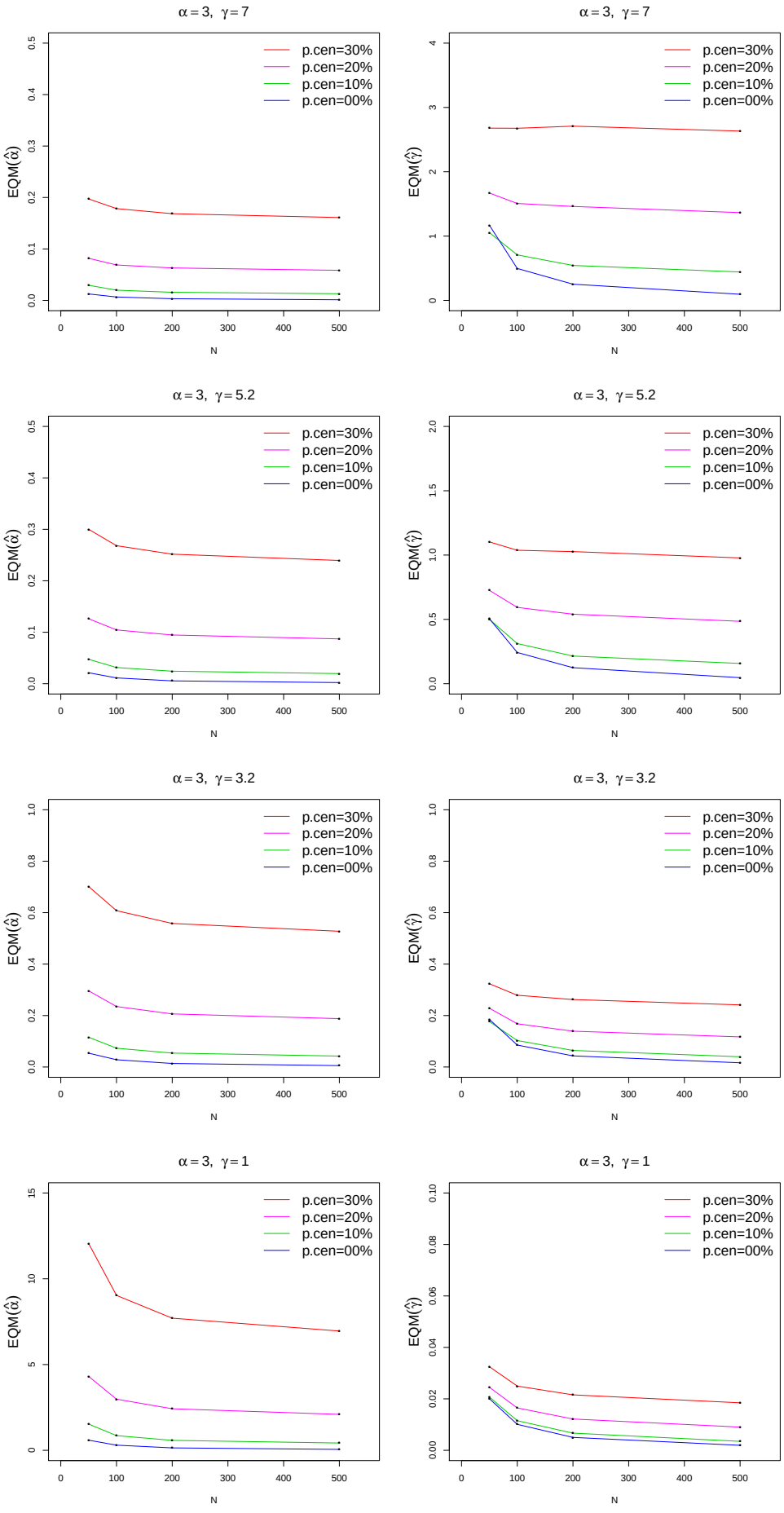


Figura 4.3: Gráfico de linhas dos EQM's dos parâmetros, segundo o tamanho da amostra para o cenário 3.

Tabela 4.5: Estimativas dos parâmetros do cenário 4, do vício e EQM, via simulação de Monte Carlo com 2000 réplicas para diferentes percentuais de censura.

N		$\boldsymbol{\theta_1}(\alpha=0,9;\gamma=7)$							
		30%		20%		10%		0%	
		α	γ	α	γ	α	γ	α	γ
50	$\hat{\boldsymbol{\theta}}$	1,0454	2,7457	0,9485	3,7591	0,9023	6,7457	0,9475	19,0572
	$b(\hat{\boldsymbol{\theta}})$	0,1454	-4,2543	0,0485	-3,2409	0,0023	-0,2543	0,0475	12,0572
	$EQM(\hat{\boldsymbol{\theta}})$	0,0367	20,0401	0,0128	22,5745	0,0072	46,6584	0,0048	**186,7429**
100	$\hat{\boldsymbol{\theta}}$	1,0447	2,5956	0,9467	3,1256	0,8939	4,6216	0,9382	16,8482
	$b(\hat{\boldsymbol{\theta}})$	0,1447	-4,4044	0,0467	-3,8744	-0,0061	-2,3784	0,0382	9,8482
	$EQM(\hat{\boldsymbol{\theta}})$	0,0279	19,6606	0,0067	15,6978	0,0033	17,4951	0,0038	**155,7016**
200	$\hat{\boldsymbol{\theta}}$	1,0425	2,5380	0,9458	3,0214	0,8906	3,9245	0,9256	13,6306
	$b(\hat{\boldsymbol{\theta}})$	0,1425	-4,4620	0,0458	-3,9786	-0,0094	-3,0755	0,0256	6,6306
	$EQM(\hat{\boldsymbol{\theta}})$	0,0237	20,0284	0,0043	15,9981	0,0016	10,0166	0,0025	**104,9254**
500	$\hat{\boldsymbol{\theta}}$	1,0415	2,5373	0,9457	3,0227	0,8906	3,8823	0,9094	9,4137
	$b(\hat{\boldsymbol{\theta}})$	0,1415	-4,4627	0,0457	-3,9773	-0,0094	-3,1177	0,0094	2,4137
	$EQM(\hat{\boldsymbol{\theta}})$	0,0213	19,9636	0,0030	15,8869	**0,0007**	**9,8448**	0,0010	**37,2721**
N		$\boldsymbol{\theta_2}(\alpha=0,9;\gamma=5,2)$							
		30%		20%		10%		0%	
		α	γ	α	γ	α	γ	α	γ
50	$\hat{\boldsymbol{\theta}}$	1,1001	2,6439	0,9930	3,2840	0,9301	4,9439	0,9258	12,3952
	$b(\hat{\boldsymbol{\theta}})$	0,2001	-2,5561	0,0930	-1,9160	0,0301	-0,2561	0,0258	7,1952
	$EQM(\hat{\boldsymbol{\theta}})$	0,0563	7,2490	0,0190	8,1531	0,0076	21,5745	0,0051	**128,8790**
100	$\hat{\boldsymbol{\theta}}$	1,0991	2,5508	0,9919	3,0037	0,9258	3,8482	0,9134	8,6768
	$b(\hat{\boldsymbol{\theta}})$	0,1991	-2,6492	0,0919	-2,1963	0,0258	-1,3518	0,0134	3,4768
	$EQM(\hat{\boldsymbol{\theta}})$	0,0471	7,2245	0,0131	5,3025	0,0037	3,9932	0,0026	**61,5490**
200	$\hat{\boldsymbol{\theta}}$	1,0964	2,5020	0,9912	2,9297	0,9244	3,6476	0,9045	6,0664
	$b(\hat{\boldsymbol{\theta}})$	0,1964	-2,6980	0,0912	-2,2703	0,0244	-1,5524	0,0045	0,8664
	$EQM(\hat{\boldsymbol{\theta}})$	0,0422	7,3742	0,0106	5,2810	0,0021	2,6291	0,0011	**12,5291**
500	$\hat{\boldsymbol{\theta}}$	1,0946	2,5001	0,9902	2,9284	0,9239	3,6256	0,9010	5,3167
	$b(\hat{\boldsymbol{\theta}})$	0,1946	-2,6999	0,0902	-2,2716	0,0239	-1,5744	0,0010	0,1167
	$EQM(\hat{\boldsymbol{\theta}})$	0,0393	7,3286	0,0091	5,2117	0,0012	2,5609	0,0004	0,3323
N		$\boldsymbol{\theta_3}(\alpha=0,9;\gamma=3,2)$							
		30%		20%		10%		0%	
		α	γ	α	γ	α	γ	α	γ
50	$\hat{\boldsymbol{\theta}}$	1,1986	2,2199	1,0614	2,4942	0,9664	2,9075	0,9032	3,8419
	$b(\hat{\boldsymbol{\theta}})$	0,2986	-0,9801	0,1614	-0,7058	0,0664	-0,2925	0,0032	0,6419
	$EQM(\hat{\boldsymbol{\theta}})$	0,1146	1,2339	0,0427	0,8285	0,0158	1,4968	0,0078	**10,1046**
100	$\hat{\boldsymbol{\theta}}$	1,1959	2,1599	1,0603	2,4276	0,9653	2,7908	0,9026	3,3303
	$b(\hat{\boldsymbol{\theta}})$	0,2959	-1,0401	0,1603	-0,7724	0,0653	-0,4092	0,0026	0,1303
	$EQM(\hat{\boldsymbol{\theta}})$	0,0995	1,1985	0,0333	0,7340	0,0095	0,3544	0,0038	0,7584
200	$\hat{\boldsymbol{\theta}}$	1,1920	2,1258	1,0586	2,3842	0,9641	2,7392	0,9016	3,2602
	$b(\hat{\boldsymbol{\theta}})$	0,2920	-1,0742	0,1586	-0,8158	0,0641	-0,4608	0,0016	0,0602
	$EQM(\hat{\boldsymbol{\theta}})$	0,0911	1,2122	0,0288	0,7323	0,0066	0,2998	0,0018	0,1339
500	$\hat{\boldsymbol{\theta}}$	1,1901	2,1281	1,0578	2,3890	0,9638	2,7331	0,9003	3,2203
	$b(\hat{\boldsymbol{\theta}})$	0,2901	-1,0719	0,1578	-0,8110	0,0638	-0,4669	0,0003	0,0203
	$EQM(\hat{\boldsymbol{\theta}})$	0,0863	1,1718	0,0263	0,6841	0,0051	0,2507	**0,0007**	**0,0497**
N		$\boldsymbol{\theta_4}(\alpha=0,9;\gamma=1)$							
		30%		20%		10%		0%	
		α	γ	α	γ	α	γ	α	γ
50	$\hat{\boldsymbol{\theta}}$	1,9625	0,8823	1,4740	0,9310	1,1539	0,9828	0,9271	1,0381
	$b(\hat{\boldsymbol{\theta}})$	1,0625	-0,1177	0,5740	-0,0690	0,2539	-0,0172	0,0271	0,0381
	$EQM(\hat{\boldsymbol{\theta}})$	1,5596	0,0511	0,5305	0,0417	0,1776	0,0366	0,0670	0,0374
100	$\hat{\boldsymbol{\theta}}$	1,9118	0,8619	1,4433	0,9123	1,1316	0,9654	0,9166	1,0187
	$b(\hat{\boldsymbol{\theta}})$	1,0118	-0,1381	0,5433	-0,0877	0,2316	-0,0346	0,0166	0,0187
	$EQM(\hat{\boldsymbol{\theta}})$	1,2108	0,0364	0,3873	0,0249	0,1070	0,0188	0,0342	0,0181
200	$\hat{\boldsymbol{\theta}}$	1,8815	0,8483	1,4272	0,8991	1,1223	0,9535	0,9092	1,0105
	$b(\hat{\boldsymbol{\theta}})$	0,9815	-0,1517	0,5272	-0,1009	0,2223	-0,0465	0,0092	0,0105
	$EQM(\hat{\boldsymbol{\theta}})$	1,0491	0,0316	0,3198	0,0187	0,0734	0,0107	0,0164	0,0087
500	$\hat{\boldsymbol{\theta}}$	1,8629	0,8459	1,4163	0,8974	1,1145	0,9500	0,9024	1,0031
	$b(\hat{\boldsymbol{\theta}})$	0,9629	-0,1541	0,5163	-0,1026	0,2145	-0,0500	0,0024	0,0031
	$EQM(\hat{\boldsymbol{\theta}})$	0,9598	0,0270	0,2833	0,0137	0,0558	0,0056	**0,0064**	**0,0033**

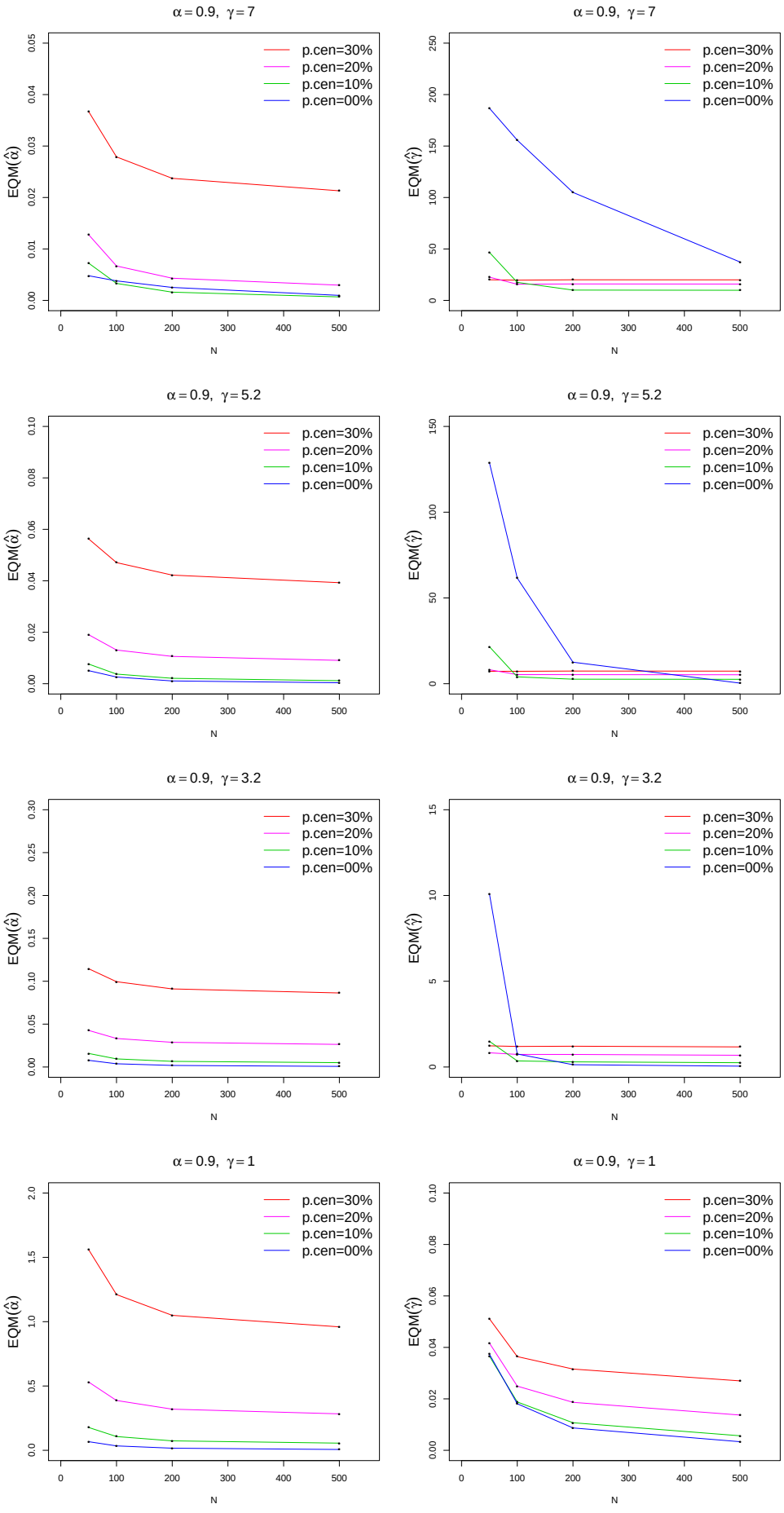


Figura 4.4: Gráfico de linhas dos EQM's dos parâmetros, segundo o tamanho da amostra para o cenário 4.

Cenário 1:

Neste primeiro cenário, pode-se comentar primeiramente sobre a distância entre os valores dos parâmetros α e γ, pois utilizando valores de α próximos de γ como em $\boldsymbol{\theta}_1(\alpha = 7,4; \gamma = 7)$, os EQM's dos estimadores são relativamente pequenos mesmo com amostras pequenas e à medida que aumenta o tamanho da amostra, o EQM decresce. Além disso, as estimativas sofrem influência do percentual de censuras, ou seja, à medida em que diminui o percentual de censura, o EQM também decresce.

Conforme a diferença entre os valores dos parâmetros α e γ aumenta, ocorre um aumento significativo relacionado a $b(\hat{\alpha})$ e, consequentemente, a $EQM(\hat{\alpha})$, e um decréscimo no $EQM(\hat{\gamma})$. Ao utilizar o cenário com função de risco decrescente, percebe-se que, com uma amostra de tamanho 50 e com percentual de censura 30%, tem-se o $EQM(\hat{\alpha}) = 66,3408$. No entanto, ocorre uma queda significativa quando o tamanho da amostra aumenta para $n = 500$ e percentual de censura igual a 0%, quando $EQM(\hat{\alpha}) = 0,3209$.

Observa-se, também que, quando há algum percentual de censura, o vício do estimador $\hat{\alpha}$ em todos os casos é positivo. Em contrapartida, o vício do estimador $\hat{\gamma}$ é negativo. Em primeira instância, o que se pode concluir é que existe uma superestimação por parte de $\hat{\alpha}$ e uma subestimação por parte de $\hat{\gamma}$, e que há uma relação entre os dois parâmetros. No entanto, quando não existe censura nos dados, o estimador $\hat{\alpha}$ em alguns momentos está também subestimando α e o $\hat{\gamma}$ superestimando γ.

Cenário 2:

Para o cenário 2, buscou-se investigar a influência na redução do parâmetro α, juntamente com o mesmo conjunto de valores do parâmetro γ apresentado anteriormente no cenário 1. Observa-se que o comportamento das estimativas do cenário 2 foram similares ao do cenário 1. Como houve a redução em 2,4 unidades do valor do parâmetro α em comparação ao cenário 2, ocorreu um decréscimo significativo no $EQM(\hat{\alpha})$ em todas as combinações dos parâmetros. Houve uma redução de cerca de 53% para o $EQM(\hat{\alpha})$ do cenário com taxa de falha decrescente com amostra de tamanho 50 e percentual de censuras 30%.

Ao contrário do cenário 1, em todos os casos do cenário 2, $b(\hat{\alpha}) > 0$, e sem a presença de censura, $b(\hat{\gamma})$ também é positivo, indicando que $\hat{\alpha}$ está superestimando α em todos os casos, e $\hat{\gamma}$ está superestimando γ apenas quando não há a presença de censura.

Neste cenário, também observa-se a relação do tamanho da amostra com a acurácia das estimativas, ou seja, à medida que o tamanho da amostra aumenta, os erros quadráticos médios diminuem.

Cenário 3:

Ao reduzir em duas unidades o parâmetro α no cenário 3, em comparação com o cenário 2, percebe-se que há uma redução significativa no $EQM(\hat{\alpha})$. Além disso, há um acréscimo no $EQM(\hat{\gamma})$, fato que pode ser explicado pela relação que existe entre os dois parâmetros. Destaca-se, ainda, o comportamento aparentemente constante de $EQM(\hat{\gamma})$ ao utilizar $\alpha = 3$, $\gamma = 7$ e o percentual de censura igual a 30%, pois mesmo aumentando o tamanho da amostra, ainda assim, o $EQM(\hat{\gamma})$ se mantém.

Cenário 4:

Neste último cenário, observou-se a importância que a distância entre os valores dos parâmetros tem sobre as estimativas dos mesmos. Com $\alpha = 0,9$ e $\gamma = 7$ observa-se que, à medida que o percentual de censura aumenta, o $EQM(\hat{\gamma})$ diminui, e com nenhuma censura e

com $n = 50$, tem-se $EQM(\hat{\gamma}) = 186,7429$, ou seja, um erro muito grande para a estimativa. Há ainda uma relação entre o tamanho da amostra e o valor do EQM, ou seja, à medida que o tamanho da amostra aumenta, o EQM dos estimadores diminui. Além disso, observa-se que o menor EQM para $\alpha = 0.9$ e $\gamma = 7$ ocorreu com o tamanho de amostra igual a 500 e com 10% de censura.

Assim como nos cenários anteriores, conforme diminui a distância entre os parâmetros, ocorre uma maior acurácia nos estimadores. Desta forma, utilizando $\alpha = 0,9$ e $\gamma = 1$, percebe-se que o vício e o EQM para ambos são bem pequenos e satisfaz a relação entre o percentual de censura e também o tamanho da amostra sobre as estimativas.

De maneira geral, percebe-se que as estimativas dos parâmetros sofrem influência principalmente no que se refere à distância entre os valores de α e γ. Quanto mais próximos forem os valores de α e γ, menores serão os EQM's. Além disso, valores altos dos parâmetros, mesmo sendo próximos, contribuem para o aumento dos erros quadráticos médios. Na maioria dos casos, há a influência do percentual de censura sobre as estimativas. No entanto, em casos em que há uma diferença significativa entre os valores dos parâmetros, essa relação não satisfaz, como pode ser visto no cenário 4 com $\alpha = 0,9$ e $\gamma = 7$, em que os menores EQM's ocorreram com percentual de censura igual a 10% e com amostra de tamanho 500.

Para todos os casos, o aumento do tamanho da amostra tem influência sobre a acurácia dos estimadores. No entanto, as diferenças entre as estimativas com amostra de tamanho 200 e 500 são mínimas, o que na prática é bastante satisfatório, pois, com uma amostra de tamanho 200, já é possível obter boas estimativas, reduzindo assim, os custos de pesquisa.

4.3 Simulação da distribuição LLDFC

Para a simulação dos dados com distribuição Log-Logística discreta com fração de cura considerou-se os cenários apresentados na Tabela 4.6.

Tabela 4.6: Cenários utilizados na simulação da distribuição LLDFC.

Cenários	ϕ	α	γ	Função de risco
1	0,95	0,9	1	Decrescente
	0,80	0,9	1	Decrescente
	0,60	0,9	1	Decrescente
2	0,95	5,2	3	Unimodal
	0,80	5,2	3	Unimodal
	0,60	5,2	3	Unimodal
3	0,95	3	5,2	Unimodal
	0,80	3	5,2	Unimodal
	0,60	3	5,2	Unimodal

Na a simulação da distribuição LLD foram considerados cenários com quatro percentuais de censuras diferentes (0%, 10%, 20% e 30%). Para a simulação da distribuição LLDFC, será considerado o percentual de censura de 10% para os indivíduos suscetíveis em todos os cenários e a soma desses percentuais com o percentual de censura dos indivíduos não suscetíveis, será o percentual total de censura.

Seja T o tempo até a ocorrência do evento de interesse e $t_1, \ldots, t_N$ uma amostra aleatória simples de T. Seja n o número de indivíduos suscetíveis (I_S), $N-n$ o número de indivíduos não-suscetíveis (I_{NS}) e $n = n_1 + n_2$, sendo n_1 a quantidade de censuras nos indivíduos suscetíveis e n_2 a quantidade de falhas nos indivíduos suscetíveis. Desta forma, tem-se que:

$$P(\text{censura}) = \frac{n_1}{n} \times \frac{n}{N} + \frac{N-n}{N} = \frac{n_1}{N} + \frac{N-n}{N}, \tag{4.4}$$

$$P(\text{falha}) = \frac{n}{N} \times \frac{n_2}{n} = \frac{n_2}{N}, \tag{4.5}$$

$$P(I_S) = \frac{n}{N} = \phi, \tag{4.6}$$

$$P(I_{NS}) = \frac{N-n}{N} = 1 - \phi, \tag{4.7}$$

cujas proporções são mostradas na Figura 4.5.

Desta forma, a simulação dos dados da distribuição LLDFC será realizada utilizando o percentual de 10% de censura para os indivíduos suscetíveis e o percentual de censura total será dado pela média dos percentuais totais das 2000 réplicas de Monte Carlo.

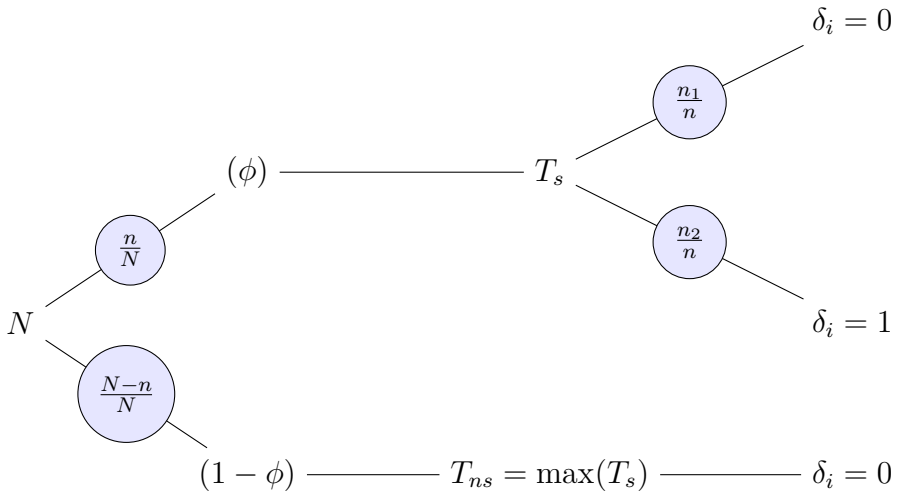


Figura 4.5: Diagrama para simulação de tempos de sobrevivência com fração de cura.

O algoritmo para a simulação dos dados de sobrevivência com distribuição LLDFC é dado por:

1. Gerar a quantidade n de indivíduos suscetíveis (I_S) de acordo com a distribuição binomial com probabilidade ϕ e a quantidade de indivíduos não suscetíveis (I_{NS}) será $N - n$;

2. Gerar o tempo de sobrevivência dos indivíduos suscetíveis com distribuição LLD, (T_S) ;

3. Gerar o vetor de censuras dos indivíduos suscetíveis, segundo a distribuição binomial com probabilidade igual a *p.censura*, em que *p.censura* é o percentual de censura dos suscetíveis;

4. Gerar o tempo dos indivíduos não suscetíveis (T_{NS}) que será um vetor composto pelo tempo máximo de T_S com tamanho igual a $N - n$;

5. Gerar o vetor de censuras dos indivíduos não suscetíveis, que terá todos os indivíduos censurados, com tamanho $N - n$;

6. Gerar o vetor dos tempos de sobrevivência de todos os indivíduos, que será a junção de T_S com T_{NS};

7. Gerar o vetor de censuras de todos os indivíduos, que será dado pela junção do vetor de censura dos indivíduos suscetíveis e não suscetíveis.

Para o cenário 1, com taxa de falha decrescente, observa-se que os EQM's dos estimadores são decrescentes à medida que o tamanho da amostra aumenta. Como há uma relação direta entre o valor de ϕ e o percentual de censuras, à medida que o valor de ϕ diminui, ocorre um aumento na quantidade de censuras e com isso, há um aumento significativo no erro quadrático médio do estimador. Para o caso $\boldsymbol{\theta_3}$ do cenário 1, observa-se que, com $N = 50$, $EQM(\hat{\alpha}) = 390,21$. Já com $N = 100$, esse valor decai para $0,729$, ou seja, o tamanho da amostra e o valor do parâmetro ϕ interferem na acurácia dos estimadores. Na Tabela 4.7, apresenta-se o percentual médio do total de censura (cens.t), de acordo com equação (4.4).

Tabela 4.7: Estimativas dos parâmetros do cenário 1, do vício e EQM, via simulação de Monte Carlo com 2000 réplicas para o modelo LLDFC com 10% de censura dos indivíduos suscetíveis.

Estimativas	$\boldsymbol{\theta}_1$ ($\phi = 0,95; \alpha = 0,9; \gamma = 1$)								
	N=50	cens.t=14,45%		N=100	cens.t=14,52%		N=500	cens.t=14,49%	
	ϕ	α	γ	ϕ	α	γ	ϕ	α	γ
$\hat{\boldsymbol{\theta}}$	0,96212	1,17521	0,90746	0,94742	1,11543	0,91962	0,93283	1,07307	0,96361
$b(\hat{\boldsymbol{\theta}})$	0,01212	0,27521	-0,09254	-0,00258	0,21543	-0,08038	-0,01717	0,17307	-0,03639
$EQM(\hat{\boldsymbol{\theta}})$	0,00187	0,21812	0,05350	0,00110	0,10996	0,02824	0,00047	0,03957	0,00532
Estimativas	$\boldsymbol{\theta}_2$ ($\phi = 0,80; \alpha = 0,9; \gamma = 1$)								
	N=50	cens.t=27,96%		N=100	cens.t=28,03%		N=500	cens.t=28,02%	
	ϕ	α	γ	ϕ	α	γ	ϕ	α	γ
$\hat{\boldsymbol{\theta}}$	0,81447	1,24489	0,86212	0,78006	1,06283	0,91249	0,76234	1,01074	0,97906
$b(\hat{\boldsymbol{\theta}})$	0,01447	0,34489	-0,13788	-0,01994	0,16283	-0,08751	-0,03766	0,11074	-0,02094
$EQM(\hat{\boldsymbol{\theta}})$	0,00819	0,50788	0,09140	0,00329	0,10657	0,03799	0,00180	0,02282	0,00530
Estimativas	$\boldsymbol{\theta}_3$ ($\phi = 0,60; \alpha = 0,9; \gamma = 1$)								
	N=50	cens.t=46,14%		N=100	cens.t=46,07%		N=500	cens.t=46,01%	
	ϕ	α	γ	ϕ	α	γ	ϕ	α	γ
$\hat{\boldsymbol{\theta}}$	0,63347	2,85023	0,84339	0,57937	1,10085	0,90473	0,56004	0,97239	0,97986
$b(\hat{\boldsymbol{\theta}})$	0,03347	1,95023	-0,15661	-0,02063	0,20085	-0,09527	-0,03996	0,07239	-0,02014
$EQM(\hat{\boldsymbol{\theta}})$	0,01952	**390,216**	0,31324	0,00441	0,72940	0,05387	0,00211	0,01964	0,00747

Além disso, existe uma relação entre os EQMs dos três estimadores, ou seja, percebe-se, através da Figura 4.6, que mesmo os valores dos três parâmetros sendo próximos, o valor de $EQM(\hat{\phi})$ é menor do que $EQM(\hat{\gamma})$ e $EQM(\hat{\alpha})$.

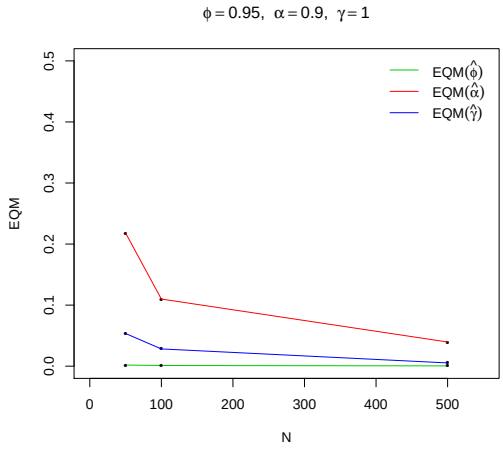


Figura 4.6: Gráfico de linhas dos EQM's dos estimadores, segundo o tamanho da amostra para o cenário 1.1.

A Figura 4.7 apresenta os gráficos de linha para o comportamento dos EQM's dos estimadores de acordo com o tamanho da amostra, apresentado na Tabela 4.6.

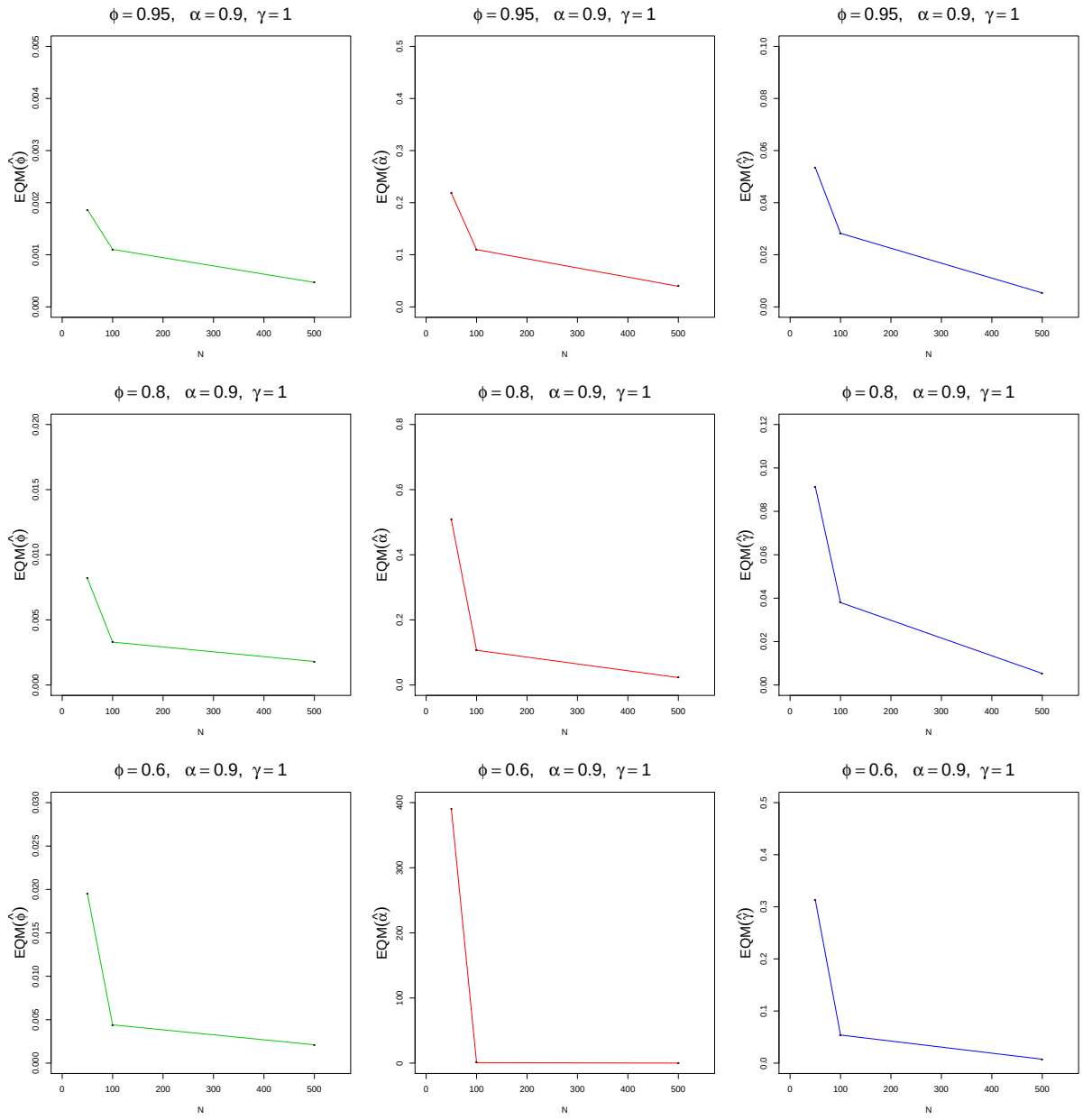


Figura 4.7: Gráficos de linhas dos EQM's dos estimadores, segundo o tamanho da amostra para o cenário 1.

Para o cenário 2, com taxa de falha unimodal, observa-se que também existe a relação da acurácia dos estimadores com o tamanho da amostra. Neste cenário, com uma distância pequena entre os parâmetros α e γ, o que se observa é que, mesmo com o valor do parâmetro ϕ diminuindo, (o que acarreta o aumento de censura total), não há um aumento significativo ao erro quadrático médio dos estimadores.

Tabela 4.8: Estimativas dos parâmetros do cenário 2, do vício e EQM, via simulação de Monte Carlo com 2000 réplicas para o modelo LLDFC com 10% de censura dos indivíduos suscetíveis.

Estimativas	$\boldsymbol{\theta}_1$ ($\phi = 0,95; \alpha = 5,2; \gamma = 3$)								
	N=50	cens.t=14,45%		N=100	cens.t=14,52%		N=500	cens.t=14,49%	
	ϕ	α	γ	ϕ	α	γ	ϕ	α	γ
$\hat{\boldsymbol{\theta}}$	0,95430	5,59727	2,88721	0,94352	5,52199	2,87889	0,93343	5,45026	2,91570
$b(\hat{\boldsymbol{\theta}})$	0,00430	0,39727	-0,11279	-0,00648	0,32199	-0,12111	-0,01657	0,25026	-0,08430
$EQM(\hat{\boldsymbol{\theta}})$	0,00172	0,44284	0,18641	0,00102	0,23725	0,09433	0,00044	0,08586	0,02252
Estimativas	$\boldsymbol{\theta}_2$ ($\phi = 0,80; \alpha = 5,2; \gamma = 3$)								
	N=50	cens.t=27,96%		N=100	cens.t=28,03%		N=500	cens.t=28,02%	
	ϕ	α	γ	ϕ	α	γ	ϕ	α	γ
$\hat{\boldsymbol{\theta}}$	0,79147	5,57510	2,88473	0,77525	5,43968	2,90025	0,76482	5,36109	2,95613
$b(\hat{\boldsymbol{\theta}})$	-0,00853	0,37510	-0,11527	-0,02475	0,23968	-0,09975	-0,03518	0,16109	-0,04387
$EQM(\hat{\boldsymbol{\theta}})$	0,00529	0,50884	0,23118	0,00287	0,20968	0,10296	0,00162	0,05270	0,02121
Estimativas	$\boldsymbol{\theta}_3$ ($\phi = 0,60; \alpha = 5,2; \gamma = 3$)								
	N=50	cens.t=46,14%		N=100	cens.t=46,07%		N=500	cens.t=46,01%	
	ϕ	α	γ	ϕ	α	γ	ϕ	α	γ
$\hat{\boldsymbol{\theta}}$	0,58881	5,62228	2,88154	0,57190	5,42354	2,90276	0,56216	5,31176	2,96417
$b(\hat{\boldsymbol{\theta}})$	-0,01119	0,42228	-0,11846	-0,02810	0,22354	-0,09724	-0,03784	0,11176	-0,03583
$EQM(\hat{\boldsymbol{\theta}})$	0,00714	0,87510	0,32100	0,00339	0,26002	0,14375	0,00194	0,04981	0,02868

Assim como no cenário 1, no cenário 2 a estimativa do parâmetro ϕ é mais acurada em relação às estimativas dos outros parâmetros.

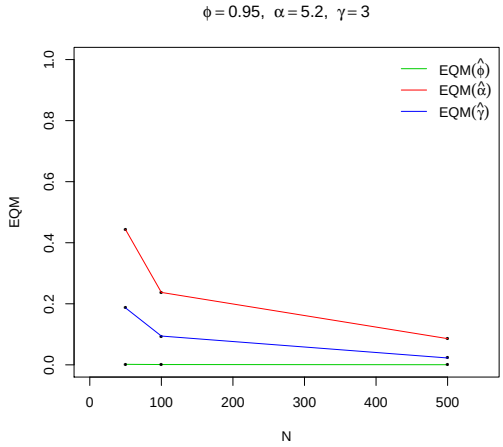


Figura 4.8: Gráfico de linhas dos EQM's dos estimadores, segundo o tamanho da amostra para o cenário 2.1.

A Figura 4.9 mostra o comportamento decrescente dos EQM's dos estimadores à medida que o tamanho da amostra aumenta.

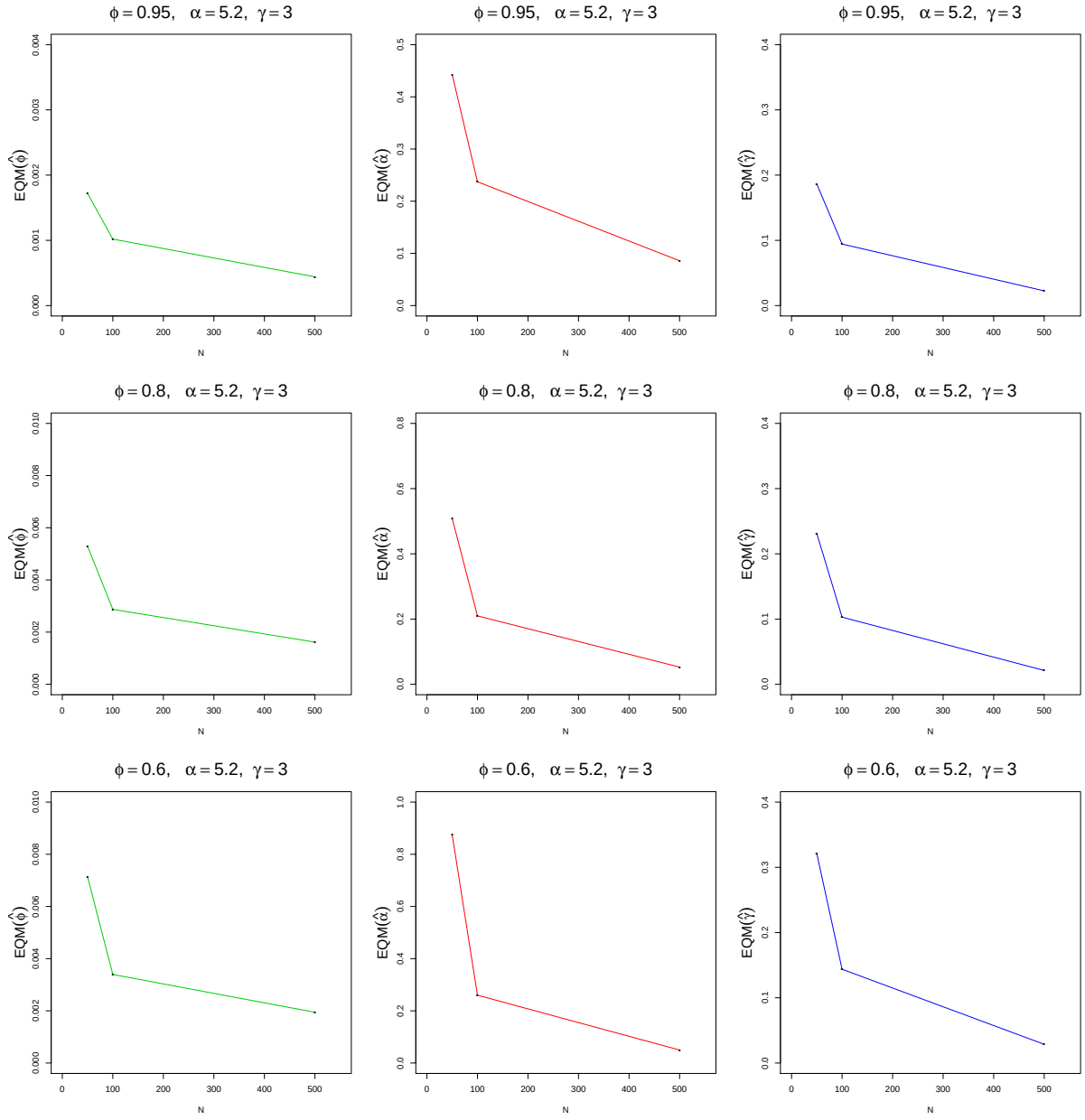


Figura 4.9: Gráficos de linhas dos EQM's dos estimadores, segundo o tamanho da amostra para o cenário 2.

No cenário 3, com taxa de falha também unimodal e com $\gamma > \alpha$, observa-se que, assim como no cenário 2, os erros quadráticos médios dos estimadores são pequenos e, à medida que o tamanho da amostra aumenta, os erros tendem a diminuir, como mostra a Figura 4.11.

Tabela 4.9: Estimativas dos parâmetros do cenário 3, do vício e EQM, via simulação de Monte Carlo com 2000 réplicas para o modelo LLDFC com 10% de censura dos indivíduos suscetíveis.

	$\boldsymbol{\theta}_1$ ($\phi = 0,95; \alpha = 3; \gamma = 5,2$)								
Estimativas	N=50	cens.t=14,45%		N=100	cens.t=14,52%		N=500	cens.t=14,49%	
	ϕ	α	γ	ϕ	α	γ	ϕ	α	γ
$\hat{\boldsymbol{\theta}}$	0,94522	3,13219	4,96728	0,93601	3,11393	4,95539	0,92813	3,09496	4,98695
$b(\hat{\boldsymbol{\theta}})$	-0,00478	0,13219	-0,23272	-0,01399	0,11393	-0,24461	-0,02187	0,09496	-0,21305
$EQM(\hat{\boldsymbol{\theta}})$	0,00193	0,05026	0,63322	0,00122	0,02830	0,34406	0,00066	0,01178	0,09762
	$\boldsymbol{\theta}_2$ ($\phi = 0,80; \alpha = 3; \gamma = 5,2$)								
Estimativas	N=50	cens.t=27,96%		N=100	cens.t=28,03%		N=500	cens.t=28,02%	
	ϕ	α	γ	ϕ	α	γ	ϕ	α	γ
$\hat{\boldsymbol{\theta}}$	0,77805	3,10672	5,03538	0,76706	3,07618	5,02528	0,75923	3,05769	5,09361
$b(\hat{\boldsymbol{\theta}})$	-0,02195	0,10672	-0,16462	-0,03294	0,07618	-0,17472	-0,04077	0,05769	-0,10639
$EQM(\hat{\boldsymbol{\theta}})$	0,00526	0,05081	0,83975	0,00329	0,02325	0,36278	0,00204	0,00653	0,07482
	$\boldsymbol{\theta}_3$ ($\phi = 0,60; \alpha = 3; \gamma = 5,2$)								
Estimativas	N=50	cens.t=46,14%		N=100	cens.t=46,07%		N=500	cens.t=46,01%	
	ϕ	α	γ	ϕ	α	γ	ϕ	α	γ
$\hat{\boldsymbol{\theta}}$	0,57519	3,10449	5,09653	0,56542	3,06395	5,07215	0,55861	3,03844	5,12865
$b(\hat{\boldsymbol{\theta}})$	-0,02481	0,10449	-0,10347	-0,03458	0,06395	-0,12785	-0,04139	0,03844	-0,07135
$EQM(\hat{\boldsymbol{\theta}})$	0,00671	0,07360	1,15337	0,00373	0,02768	0,48782	0,00222	0,00581	0,09720

Como nesse cenário tem-se que $\phi < \alpha < \gamma$, percebe-se que $EQM(\hat{\phi})$ continua menor, mas bem próximo de $EQM(\hat{\alpha})$ e $EQM(\hat{\gamma})$, tendo este último o maior valor, como mostra a Figura 4.10.

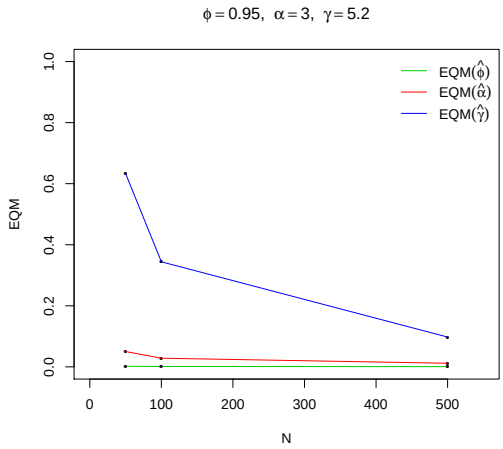


Figura 4.10: Gráfico de linhas dos EQM's dos estimadores, segundo o tamanho da amostra para o cenário 2.1.

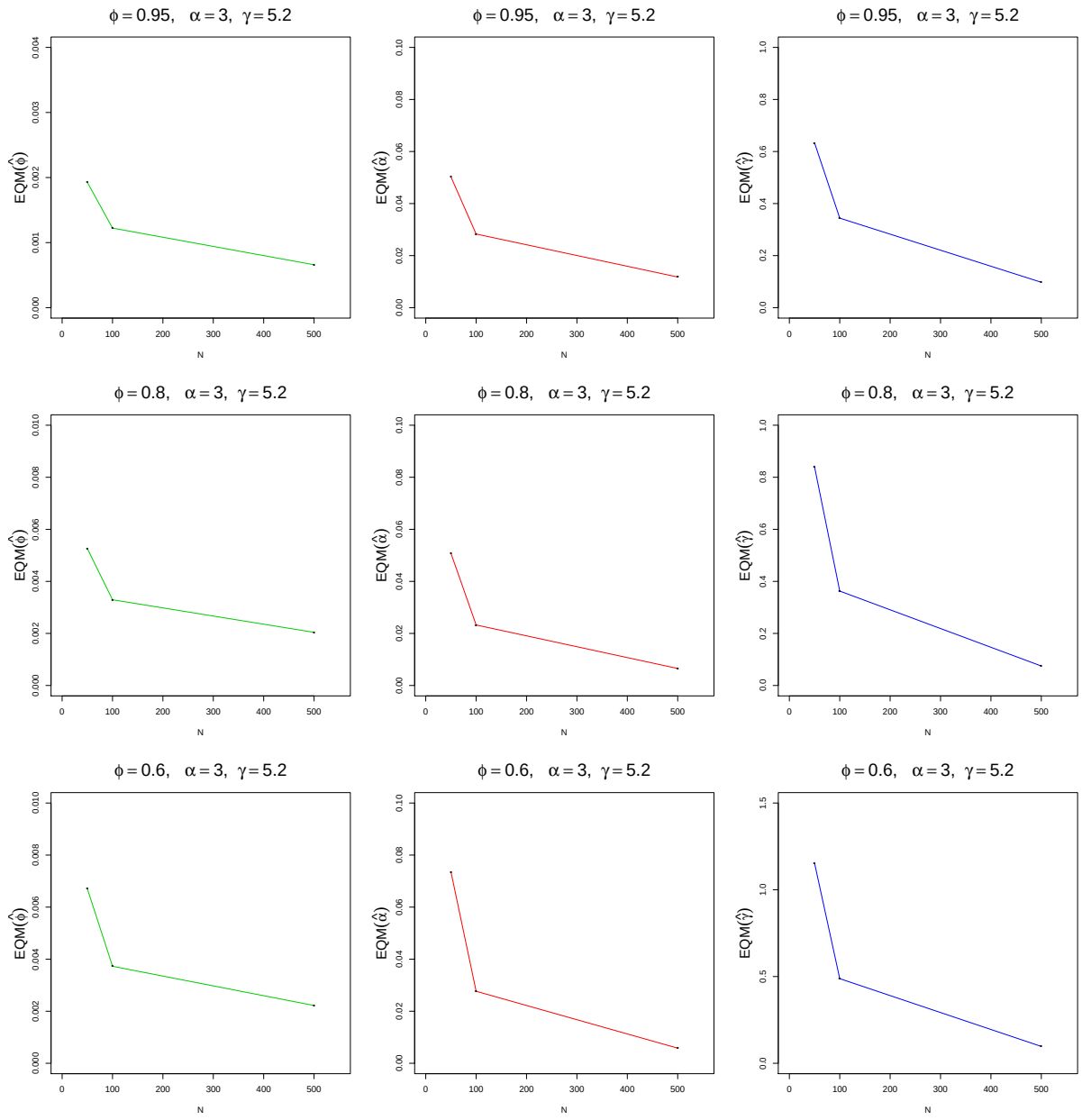


Figura 4.11: Gráficos de linhas dos EQM's dos estimadores, segundo o tamanho da amostra para o cenário 3.

De maneira geral, percebe-se que em todos os cenários prevalece a relação do tamanho da amostra com o valor do EQM dos estimadores. Nesse mesmo contexto, no caso $\boldsymbol{\theta_3}$ do cenário 1, observou-se que, com amostra de tamanho 50, $EQM(\hat{\alpha})$ foi muito elevado. Isso pode ser explicado pelo valor do parâmetro ϕ, que acarreta em um percentual médio de censura total em torno de 46%. Além disso, essa situação também pode ser explicada pelo fato que $\gamma = 1$, tendo então, taxa de falha decrescente, e como foi observado na Seção 4.2, este cenário compromete $\hat{\alpha}$ e favorece $\hat{\gamma}$. Observa-se que, em todos os casos, as estimativas do parâmetro ϕ foram bem próximas do valor real do parâmetro, mostrando assim, que o modelo tem uma boa precisão.

Capítulo 5

Aplicação em dados reais

O objetivo deste capítulo é mostrar uma aplicação do modelo de regressão Log-Logístico discreto e uma aplicação com o modelo Log-Logístico discreto com fração de cura em dados reais. Os dados foram obtidos junto ao Departamento de Estatística da Universidade Estadual da Paraíba (UEPB) - Campus I. O conjunto de dados refere-se às informações dos alunos do curso de Bacharelado em Engenharia Ambiental e de Computação da Universidade Estadual da Paraíba - Campus I .

5.1 Aplicação 1 - LLD

5.1.1 Banco de dados

Para a aplicação 1 será utilizado o banco de dados dos alunos do curso de Computação da UEPB - Campus I, com o total de 709 alunos. O período de acompanhamento é compreendido entre o primeiro semestre de 2010 e o segundo semestre de 2016.

O evento de interesse é a evasão do aluno, logo a variável resposta T é o tempo, em semestres, desde a matrícula até a ocorrência do evento de interesse ou uma censura. Além da variável T, tem-se a presença de outras informações dos alunos, denotadas como covariáveis. Neste banco de dados, a proporção de censuras é de 53,4%. Na Tabela 5.1, tem-se a descrição das covariáveis que será analisada.

Tabela 5.1: Covariáveis dos alunos de Computação

Covariáveis	Categorias	N (%)
Sexo	0 - Masculino	599 (84,5%)
	1 - Feminino	110 (15,5%)
Idade	0 - $<$ 20 anos	301 (42,5%)
	1 - $\geq$ 20 anos	408 (57,5%)
Origem	0 - Outras Cidades	242 (34,1%)
	1 - Campina Grande	467 (65,9%)
Tipo de Escola	0 - Privado	274 (38,6%)
	1 - Público	435 (61,4%)
Forma de Ingresso	0 - Vestibular	300 (42,3%)
	1 - ENEM	409 (57,7%)
Turno	0 - Diurno	344 (48,5%)
	1 - Noturno	365 (51,5%)

5.1.2 Análise descritiva

Inicia-se a análise descritiva dos dados observando o comportamento da curva de sobrevivência obtida pelo estimador de Kaplan e Meier (1958). A partir do gráfico da função de sobrevivência e da função de risco acumulada, pode-se obter informações sobre modelos que podem se ajustar aos dados.

A Figura 5.1 mostra o comportamento empírico da função de sobrevivência e pode-se obter informação sobre o decaimento pouco acentuado da curva de sobrevivência. No tempo zero a probabilidade de sobrevivência não é 1, ou seja, neste tempo ocorreram falhas, sendo uma característica de dados discretos. Observa-se que em todos os tempos ocorreram censuras e que, aparentemente, o tempo mediano está em torno de $t = 5$.

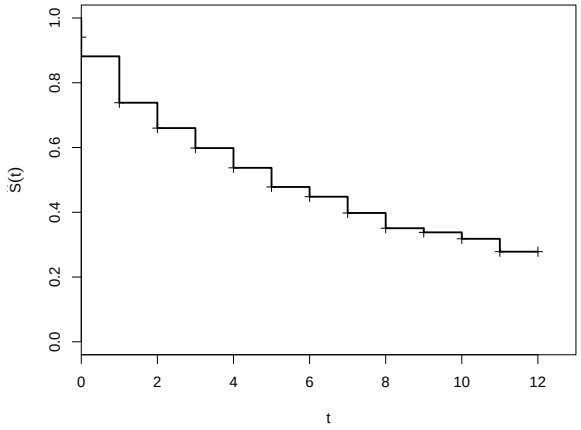


Figura 5.1: Curva de sobrevivência estimada pelo método de Kaplan e Meier (1958) para os dados dos alunos do curso de Computação.

Além disso, pela Figura 5.2, observa-se que a função de risco apresenta forma decrescente de acordo com a metodologia gráfica apresentada na Subseção 2.1.3.

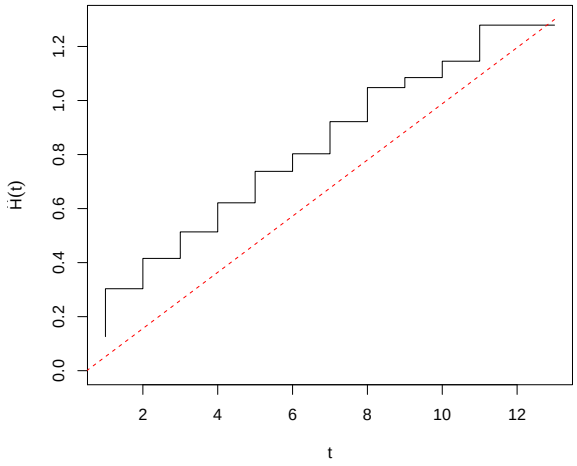


Figura 5.2: Curva do risco acumulado dos alunos do curso de Computação.

Com base nessas análises descritivas, é possível supor o modelo Log-Logístico discreto apresentado na Seção 3.1. Outra distribuição que atende as essas suposições é a Weibull

discreta proposta por Nakagawa e Osaki (1975), que foi estudada por Brunello e Nakano (2015), e que tem a função de distribuição de probabilidade, função de sobrevivência e função de risco, respectivamente, por:

$$p(t) = q^{t^\gamma} - q^{(t+1)^\gamma}, \qquad S(t) = q^{(t+1)^\gamma}, \qquad h(t) = 1 - q^{(t+1)^\gamma - t^\gamma} \qquad t = 0, 1, 2, ... \tag{5.1}$$

em que $q = \exp\left\{-\left(\frac{1}{\alpha}\right)^\gamma\right\}$ é interpretado como uma probabilidade, pois $0 < q < 1$ para quaisquer γ e α maiores que zero.

Ao ajustar os dois modelos, conforme mostra a Figura 5.3, percebe-se que o modelo Log-Logístico discreto apresenta maior proximidade com a curva empírica do que o modelo Weibull discreto (WD).

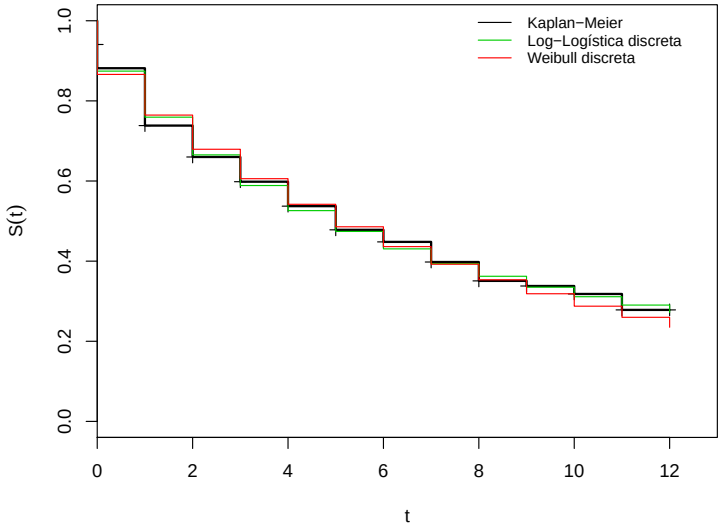


Figura 5.3: Curvas de sobrevivência estimadas pelo método de Kaplan e Meier (1958) e pelos modelos Log-Logístico discreto e Weibull discreto.

Na Tabela 5.2, apresenta-se as estimativas para os modelos LLD e WD, com seus respectivos erros padrão e o intervalo de confiança de 95% de confiança utilizando a transformação logarítmica.

Tabela 5.2: Estimativas dos modelos Log-Logístico discreto e Weibull discreto.

Modelo	Parâmetros	Estimativas	Erro Padrão	Intervalo de confiança (95%)
LLD	α	5,479414	0,37952262	$[4,82663; 6,22048]$
	γ	1,139994	0,05985065	$[1,03549; 1,25505]$
WD	q	0,866043	0,01168612	$[0,85268; 0,87962]$
	γ	0,900524	0,04686015	$[0,84811; 0,95618]$

Segundo Nakano e Carrasco (2006), o erro máximo cometido na estimação pode ser definido como um teste estatístico de ajuste de modelos, utilizando as diferenças entre as estimativas do modelo estudado e as estimativas pelo método de Kaplan e Meier (1958), sendo expresso por:

$$\epsilon = \max |\hat{S}(t) - \hat{S}_{km}(t)|. \tag{5.2}$$

Desta forma, foram calculadas as estimativas da função de sobrevivência dos modelos Log-Logístico discreto, Weibull discreto e as estimativas da função de sobrevivência empírica pelo método de Kaplan e Meier (1958).

Tabela 5.3: Estimativas da função de sobrevivência pelo método de Kaplan e Meier (1958), pelo modelo Log-Logístico discreto e Weibull discreto.

Tempo	K-M	LLD	WD
0	0,881523	0,874256	0,866043
1	0,738356	0,759318	0,764543
2	0,659963	0,665236	0,679230
3	0,598284	0,588740	0,605817
4	0,537235	0,526071	0,541876
5	0,478167	0,474156	0,485757
6	0,448073	0,430650	0,436237
7	0,397821	0,393787	0,392363
8	0,350711	0,362231	0,353365
9	0,337958	0,334967	0,318610
10	0,318078	0,311212	0,287569
11	0,278318	0,290357	0,259791
12	0,278318	0,271921	0,234895
	ϵ	**0,020962**	0,043424

De acordo com a Tabela 5.3, percebe-se que as estimativas dos dois modelos estão próximas das estimativas da função empírica pelo método de Kaplan e Meier (1958). No entanto, percebe-se que nos últimos tempos o modelo LLD se aproxima mais do que o WD. Assim, ao utilizar o valor de $\epsilon = 0,020962$ do modelo LLD, tem-se mais indícios que o modelo mais adequado é o LLD.

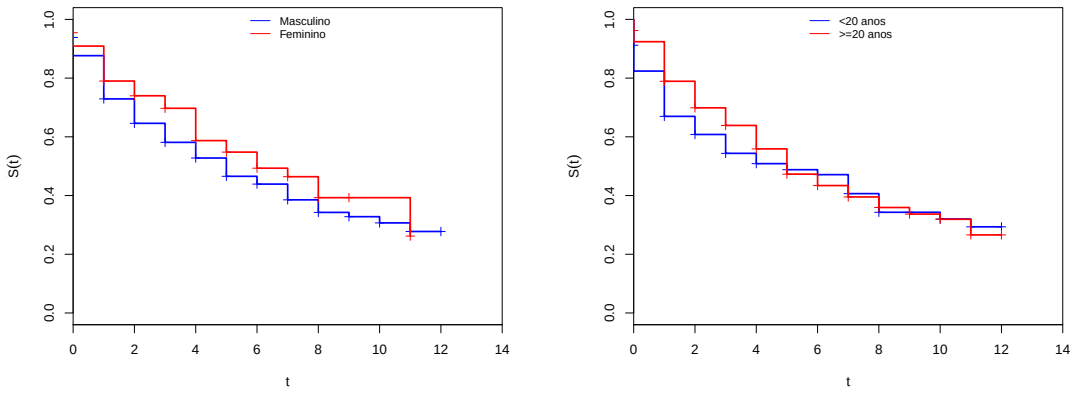
Outros procedimentos para seleção de modelos utilizam critérios de informação. Na Tabela 5.4, apresenta-se os critérios AIC, AICc e BIC para os modelos LLD e WD. Nos três critérios, o modelo LLD apresentou menores valores, e isso significa que a distribuição Log-Logística discreta se ajusta melhor aos dados do que a Weibull discreta.

Tabela 5.4: Critérios de informação AIC, AICc e BIC segundo os modelos LLD e WD.

Modelos	AIC	AICc	BIC
LLD	**2036,226**	**2036,260**	**2043,998**
WD	2041,763	2041,797	2049,535

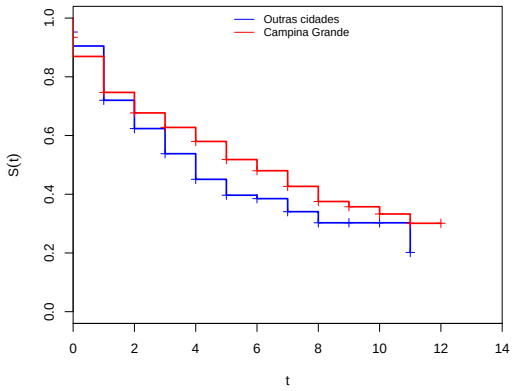
Pelas análises anteriores, conclui-se que um modelo adequado para modelar o tempo de sobrevivência dos alunos do curso de Computação é o modelo LLD. Como neste banco de dados tem-se a presença das covariáveis dos alunos, o próximo passo é a construção de um modelo de regressão, conforme apresentado na Seção 3.3. Para tanto, primeiramente será verificado quais covariáveis estão influenciando no tempo de sobrevivência por meio da análise descritiva.

Uma forma preliminar de verificar a influência da covariável é estimar a função de sobrevivência pelo método de Kaplan e Meier (1958) em função da covariável de interesse. Caso as curvas de sobrevivência empírica apresentem formas diferentes, há indícios que essa covariável está influenciando a resposta e ela deve ser considerada na modelagem paramétrica.

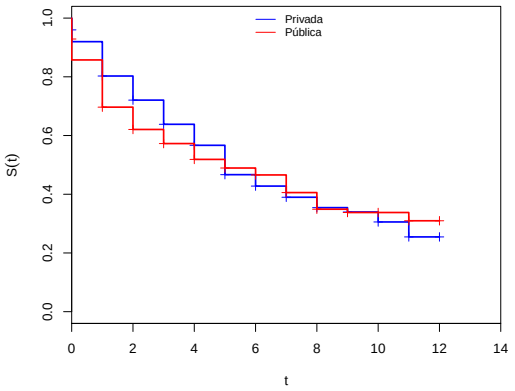


(a) Sexo

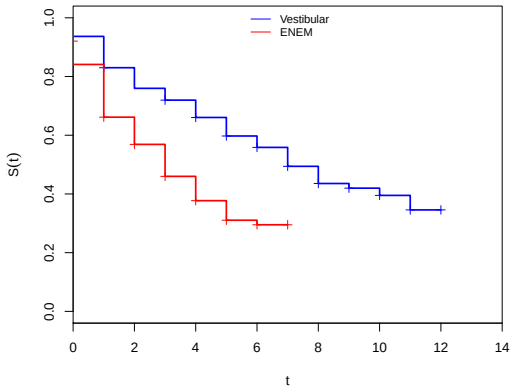
(b) Idade



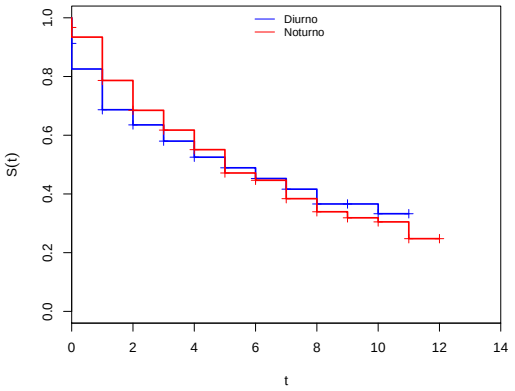
(c) Origem



(d) Tipo de Escola



(e) Forma de Ingresso



(f) Turno

Figura 5.4: Curvas de sobrevivência estimadas pelo método de Kaplan e Meier (1958) para as covariáveis dos alunos do curso de Computação.

Pela Figura 5.4, observa-se o comportamento das covariáveis em estudo. Dentre as seis covariáveis estudadas, a que apresentou uma diferença mais relevante foi a *forma de ingresso*, apresentada na Figura 5.4(e). As demais covariáveis aparentam ter efeito menos significativos. Uma outra metodologia que pode ser utilizada para verificar a influência das covariáveis no tempo de sobrevivência são os testes de hipóteses não paramétricos de Wilcoxon ou de Log-Rank. Esses testes consideram as seguintes hipóteses:

$$\begin{cases} H_0 : \text{As curvas de sobrevivência são iguais,} \\ H_1 : \text{As curvas de sobrevivência são diferentes.} \end{cases}$$

Ao considerar o nível de significância de 5%, como mostra a Tabela 5.5, as covariáveis *idade* e *forma de ingresso* apresentam curvas de sobrevivência diferentes. E ao nível de 10%, as covariáveis *tipo de escola* e *turno* também apresentam curvas de sobrevivência diferentes. Com esse resultado, pode-se construir um modelo de regressão LLD.

Tabela 5.5: Covariáveis dos alunos de Computação

Covariáveis	Estatística do teste	p-valor
Sexo	$S = 2,5$	0,1110
Idade	$S = 6,2$	0,0131
Origem	$S = 2,0$	0,1560
Tipo de Escola	$S = 3,3$	0,0677
Forma de Ingresso	$T = 41,3$	$1,32 \times 10^{-10}$
Turno	$S = 3,5$	0,0605

Para as covariáveis *Sexo*, *Idade*, *Origem*, *tipo de escola* e *turno* foi utilizado o teste similar de Wilcoxon devido ao comportamento das curvas de sobrevivência. Para a covariável *forma de ingresso* foi utilizado o teste Log-Rank, por não apresentar nenhum cruzamento entre as curvas de sobrevivência.

5.1.3 Modelo de regressão LLD

Por meio da análise descritiva realizada na Subseção 5.1.2, foi possível identificar as covariáveis que apresentam diferenças mais relevantes nas curvas de sobrevivência. Essa diferença pode ser indício de que essas covariáveis estão exercendo influência sobre a variável resposta.

Para a construção do modelo de regressão LLD, utilizou-se o processo de seleção de modelo *backward*, que incorpora inicialmente todas as covariáveis e retira-se aquelas com pior desempenho, até obter-se um modelo com todas as covariáveis significativas a um determinado nível de significância.

Dessa forma, a Tabela 5.6 apresenta as estimativas do modelo de regressão Log-Logístico discreto.

Ao analisar os resultados das estimativas dos parâmetros do modelo de regressão Log-Logístico apresentado na Tabela 5.6, observa-se o valor de $\gamma = 2,022$. Sendo assim, o modelo ajustado tem taxa de falha unimodal. Em relação à interpretação dos coeficientes de regressão, tem-se que o tempo mediano mais um semestre até a evasão para alunos com idade maior ou igual a 20 anos é em torno de 1,83 vezes maior ou 83% maior do que para alunos com idade menor que 20 anos.

Ao fazer $e^{-\beta_2}$, tem-se que o tempo mediano mais um semestre até a evasão dos alunos que terminaram o ensino médio em escola privada é cerca de 1,44 vezes maior que para os alunos que terminaram o ensino médio em escola pública.

Tabela 5.6: Estimativas dos parâmetros do modelo de regressão Log-Logístico discreto.

Modelo	Parâmetros	Estimativas	e^{β}	Erro Padrão	p-valor
MRLLD	β_0	1,0380590	2,823731	0,1151007	$0,000000 \times 10^{-0}$
	β_1(Idade $\geq$ 20 anos)	0,6025797	1,826825	0,1102721	$4,642670 \times 10^{-8}$
	β_2(Escola= Pública)	-0,3652644	0,694013	0,1056483	$5,454938 \times 10^{-4}$
	β_3(Ingresso= ENEM)	-0,7235238	0,485040	0,1062902	$9,961365 \times 10^{-12}$
	β_4(Turno= Noturno)	0,3374152	1,401321	0,1059122	$1,443541 \times 10^{-3}$
	γ	2,0220465	-	0,1202715	-

Através da análise descritiva realizada para a covariável *forma de ingresso*, verificou-se que a forma de ingresso do aluno no curso é um fator que exerce bastante influência sobre o tempo até a evasão do aluno. Ao fazer $e^{-\beta_3}$, tem-se que os alunos que ingressaram no curso via vestibular têm o tempo mediano mais um semestre 2,06 vezes maior do que os alunos que ingressaram pelo ENEM.

O último coeficiente de regressão está relacionado com o turno do curso. Como tem-se a estimativa do coeficiente β_4 positiva, significa que os alunos que pertencem ao curso noturno têm o tempo mediano mais um semestre 1,40 vezes maior do que aqueles que pertencem ao curso no turno diurno.

De maneira geral, essas interpretações estão coerentes com a análise descritiva dos dados. Isso demonstra a consistência do modelo ajustado.

5.2 Aplicação 2 - LLDFC

5.2.1 Banco de dados

Os dados foram obtidos junto ao Departamento de Estatística da Universidade Estadual da Paraíba - Campus I, com um total de 360 alunos do curso de Engenharia Ambiental. O período de acompanhamento é compreendido entre o primeiro semestre de 2010 e o segundo semestre de 2016.

O evento de interesse é a evasão do aluno, logo a variável resposta T é o tempo, em semestres, desde a matrícula até a ocorrência do evento de interesse ou uma censura. Além da variável T, tem-se a presença de outras informações dos alunos, denotadas como covariáveis. Neste banco de dados há 65,6% de censuras. A descrição das covariáveis está na Tabela 5.7.

Tabela 5.7: Covariáveis dos alunos de Engenharia Ambiental

Covariáveis	Categorias	N (%)
Sexo	0 - Masculino	201 (55,8%)
	1 - Feminino	159 (44,2%)
Idade	0 - $<$ 20 anos	172 (47,8%)
	1 - $\geq$ 20 anos	188 (52,2%)
Origem	0 - Outras Cidades	118 (32,8%)
	1 - Campina Grande	242 (67,2%)
Tipo de Escola	0 - Privado	174 (48,3%)
	1 - Público	186 (51,7%)
Forma de Ingresso	0 - Vestibular	161 (44,7%)
	1 - ENEM	199 (55,3%)

5.2.2 Análise descritiva

Inicia-se na análise descritiva dos dados observando o comportamento da curva de sobrevivência estimada pelo método de Kaplan e Meier (1958), através da Figura 5.5. Observa-se que a probabilidade de sobrevivência no tempo zero é menor do que 1, pois nesse mesmo tempo ocorreram censuras e também falhas, fato esse bastante comum em dados discretos. Além disso, ocorreram censuras em todos os tempos e, a partir do oitavo período de observação, a probabilidade de sobrevivência se estabilizou, sendo assim, um forte indício de que haja fração de curados dentre os indivíduos em estudo.

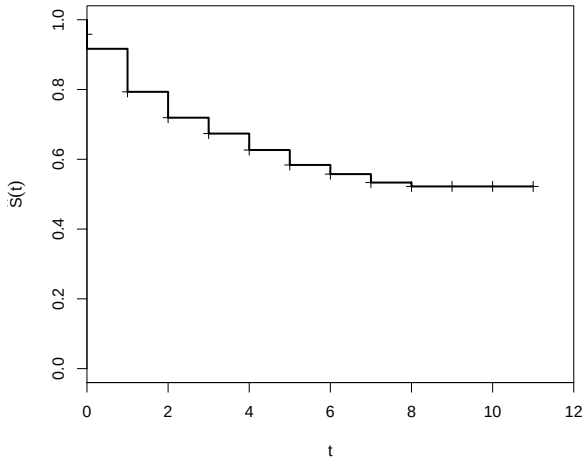


Figura 5.5: Curva de sobrevivência estimada pelo método de Kaplan e Meier (1958) para os dados dos alunos do curso de Engenharia Ambiental.

Outra forma de investigar qual o modelo mais apropriado para o ajuste do tempo de sobrevivência é através do gráfico da função de risco acumulado.

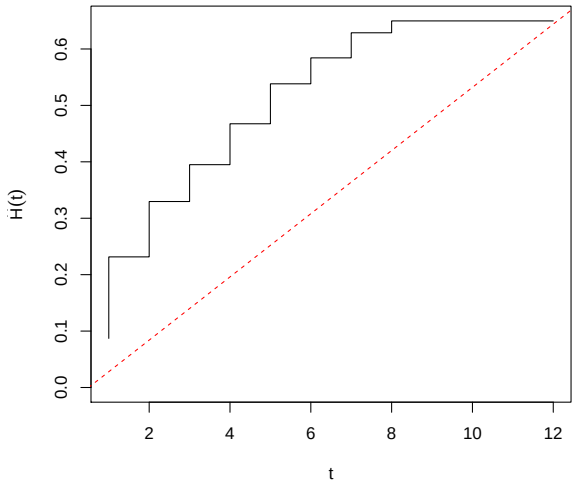


Figura 5.6: Função de risco acumulada do tempo de sobrevivência dos alunos do curso de Engenharia Ambiental.

De acordo com a Figura 5.6, observa-se que o modelo mais adequado para o ajuste dos dados deve ser o que apresenta função de risco decrescente. Desta forma, a distribuição Log-Logística discreta com fração de cura pode ser adequada para modelar esses dados ou outra distribuição que também tenha forma decrescente na função de risco.

O modelo Weibull discreto com fração de cura (WDFC) proposto por Silva (2015) também poderá ser estudado nesse conjunto de dados, pois o mesmo também compreende as suposições descritas sobre a distribuição dos dados. O modelo WDFC tem a função de distribuição de probabilidade, função de sobrevivência e função de risco, respectivamente, por:

$$p(t) = \phi[q^{t^\gamma} - q^{(t+1)^\gamma}], \qquad S(t) = 1 - \phi + \phi[q^{(t+1)^\gamma}], \qquad h(t) = \frac{\phi(q^{t^\gamma} - q^{(t+1)^\gamma})}{1 - \phi + \phi[q^{t^\gamma}]} \tag{5.3}$$

em que $q = \exp\left\{-\left(\frac{1}{\alpha}\right)^\gamma\right\}$ é interpretado como uma probabilidade, pois $0 < q < 1$ para quaisquer $\gamma > 0$, $\alpha > 0$ e $0 < \phi < 1$ é o parâmetro dos indivíduos suscetíveis.

Com base nas análises descritivas realizadas anteriormente, decidiu-se estudar modelos com fração de cura. Sendo assim, será ajustado o modelo Log-Logístico discreto com fração de cura e Weibull discreto com fração de cura.

De acordo com a Figura 5.7, é possível verificar que o modelo Log-Logístico discreto com fração de cura se comporta de forma similar com o modelo Weibull discreto com fração de cura (WDFC). No entanto, nos tempos iniciais o modelo LLDFC está mais próximo da curva empírica estimada pelo método de Kaplan e Meier (1958) do que o modelo WDFC.

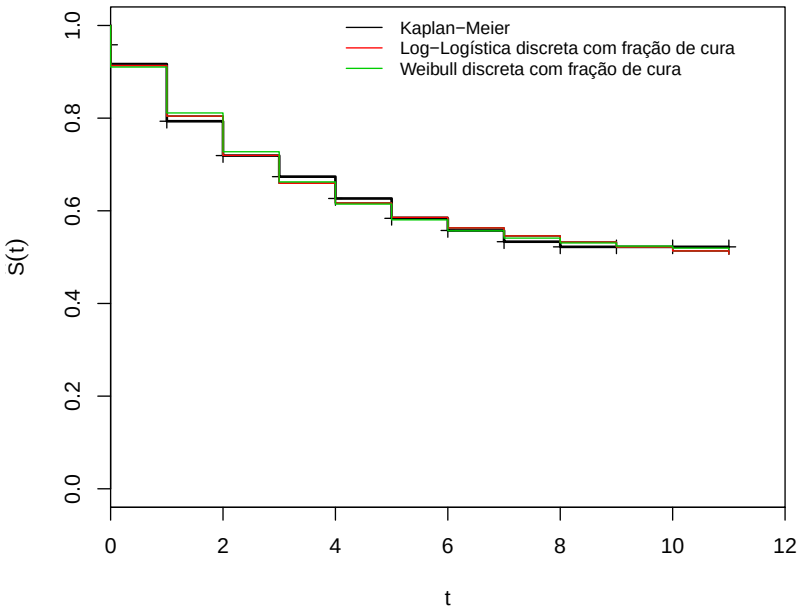


Figura 5.7: Curvas de sobrevivência estimadas pelo método de Kaplan e Meier (1958) e pelos modelos Log-Logística discreta e Log-Logística discreta com Fração de Cura

A Tabela 5.8 mostra as estimativas dos parâmetros dos modelos propostos e os respectivos erros padrão das estimativas. Além disso, construiu-se os intervalos de confiança utilizando a transformação logarítmica para $\hat{\alpha}, \hat{\gamma}$ e a transformação log-log para $\hat{\phi}$ e $\hat{q}$.

Tabela 5.8: Estimativas dos modelos Log-Logístico discreto com fração de cura e Weibull discreto com fração de cura.

Modelo	Parâmetros	Estimativas	Erro Padrão	Intervalo de confiança (95%)
LLDFC	α	2,9058573	0,48140023	$[2,30994; 3,53843]$
	γ	1,5676821	0,20376877	$[1,30994; 1,87614]$
	ϕ	0,5469946	0,05724406	$[0,47053; 0,61705]$
WDFC	q	0,8159071	0,02785543	$[0,64345; 1,03459]$
	γ	1,2676619	0,12301420	$[1,03449; 1,55339]$
	ϕ	0,4874477	0,03903247	$[0,40429; 0,58128]$

De acordo com a Tabela 5.9, nota-se que os modelos LLDFC e WDFC apresentam estimativas próximas das estimativas empíricas (K-M). No entanto, utilizando o erro máximo, há indícios que o modelo LLDFC tem melhor ajuste ao dados por apresentar $\epsilon = 0,01574$, que é menor do que o erro máximo do modelo WDFC.

Tabela 5.9: Estimativas da função de sobrevivência pelo método de Kaplan e Meier (1958), pelo modelo Log-Logístico discreto com fração de cura e Weibull discreto com fração de cura.

Tempo	K-M	LLDFC	WDFC
0	0,91667	0,91351	0,91026
1	0,79333	0,80438	0,81122
2	0,71919	0,71967	0,72745
3	0,67372	0,65939	0,66242
4	0,62661	0,61670	0,61447
5	0,58377	0,58589	0,58040
6	0,55753	0,56311	0,55688
7	0,53329	0,54584	0,54104
8	0,52218	0,53246	0,53058
9	0,52218	0,52189	0,52381
10	0,52218	0,51338	0,51949
11	0,52218	0,50644	0,51678
	ϵ	**0,01574**	0,01788

Segundo Sengupta (1995), alguns procedimentos de diagnósticos utilizando gráficos podem ser utilizados. Baseia-se no princípio da comparação de uma estimativa não paramétrica da função de sobrevivência com a estimativa paramétrica correspondente. Desta forma, um dos gráficos que pode-se utilizar é um gráfico de dispersão de $\hat{F}_0$ vs $\hat{F}$. Desta forma, a Figura 5.8 mostra os gráficos de dispersão da função acumulada $\hat{F}_0$ dos modelos LLDFC e WDFC vs a função acumulada empírica $\hat{F}$. Observa-se que o comportamento da dispersão para os modelos LLDFC e WDFC se assemelha a reta $x = y$, tendo assim, mais indícios que os dois modelos se ajustam bem aos dados.

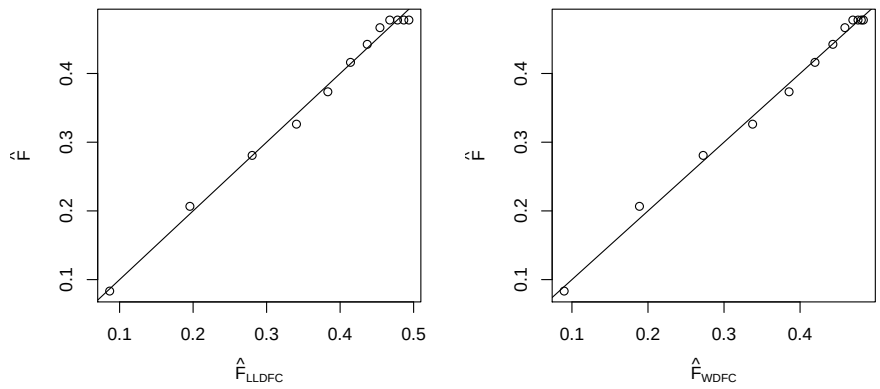


Figura 5.8: Gráfico de dispersão das estimativas da função acumulada dos modelos LLDFC e WDFC vs a função acumulada empírica.

Além disso, outras medidas importantes para a escolha do melhor modelo, são os critérios de informação AIC, AICc e BIC.

Tabela 5.10: Critérios de informação AIC, AICc e BIC segundo os modelos LLD, WD, LLDFC e WDFC.

Modelos	AIC	AICc	BIC
LLDFC	**847,9671**	**848,0345**	**859,6254**
WDFC	848,5053	848,5727	860,1636

De acordo com a Tabela 5.10, conclui-se que o modelo mais adequado para o ajuste sobre o tempo de sobrevivência dos alunos do curso de Engenharia Ambiental é o modelo Log-Logístico discreto com fração de cura, pois os valores calculados dos critérios AIC, AICc e BIC foram inferiores aos valores dos critérios referentes ao modelo WDFC.

Para as covariáveis associadas ao banco de dados, é necessário realizar uma análise descritiva com a necessidade de verificar aquelas que podem estar influenciando na variável resposta, o tempo de sobrevivência. Essa análise é feita, preliminarmente, utilizando o método gráfico e, posteriormente, um teste de hipótese baseado nas seguintes hipóteses:

$$\begin{cases} H_0 : \text{As curvas de sobrevivência são iguais,} \\ H_1 : \text{As curvas de sobrevivência são diferentes,} \end{cases}$$

de tal forma que, as covariáveis que apresentarem diferenças entre as curvas de sobrevivência serão introduzidas no modelo de regressão Log-Logístico discreto com fração de cura.

A primeira covariável a ser analisada é o sexo dos alunos. Como pode-se observar na Figura 5.9, na maioria dos tempos as duas curvas de sobrevivência se comportam de forma semelhante. No entanto, faz-se necessário realizar um teste de hipóteses. Neste caso, como houve alguns cruzamentos durante o tempo, o teste mais indicado é o similar ao de Wilcoxon. Assim sendo, o valor da estatística S do teste de hipótese similar ao de Wilcoxon resultou em $S = 0,3$ e correspondente $p - valor = 0,611$, indicando igualdade entre as curvas de sobrevivência.

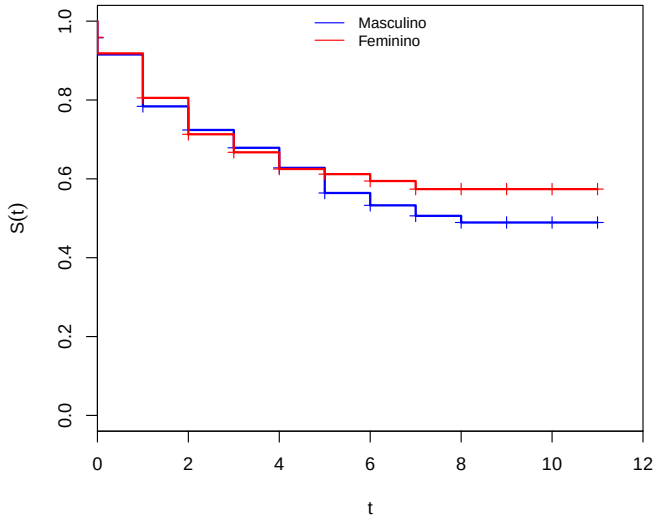


Figura 5.9: Curvas de sobrevivência estimadas pelo método de Kaplan e Meier (1958) para a covariável *Sexo*.

Na covariável *idade*, realizou-se uma categorização de acordo com o seu valor mediano igual a 20 anos. De acordo com a Figura 5.10, aparentemente, há uma diferença nas curvas de sobrevivência nos tempos finais. No entanto, ao realizar o teste similar ao de Wilcoxon, obteve-se a estatística $S = 0,5$ e correspondente $p-valor = 0,476$, indicando que não existe diferença significativa entre as curvas de sobrevivência.

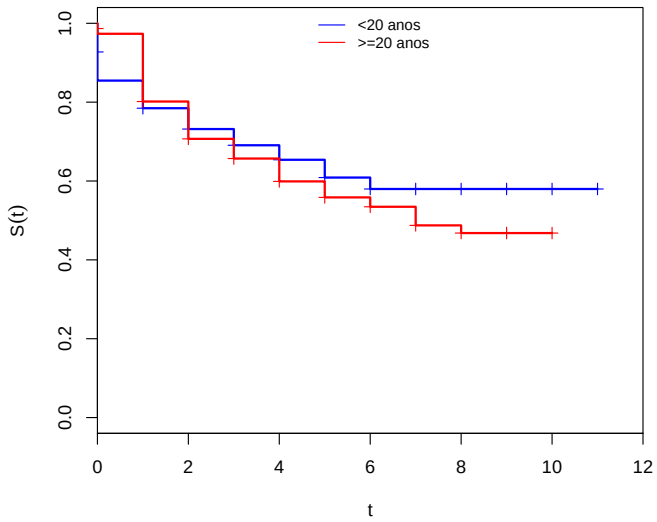


Figura 5.10: Curvas de sobrevivência estimadas pelo método de Kaplan e Meier (1958) para a covariável *Idade*.

A covariável *origem* está baseada na cidade em que o aluno reside e que foi tratada em duas categorias: Outras cidades e Campina Grande. Partiu-se do princípio de que os alunos que residem em Campina Grande têm mais chances de sobrevivência do que os alunos de outras cidades, devido à facilidade de acesso e outros fatores. Desta forma, aplicou-se o teste similar de Wilcoxon e com a estatístia $S = 0$ e correspondente $p - valor = 0,884$, conclui-se que não existe diferença significativa entre as curvas de sobrevivência.

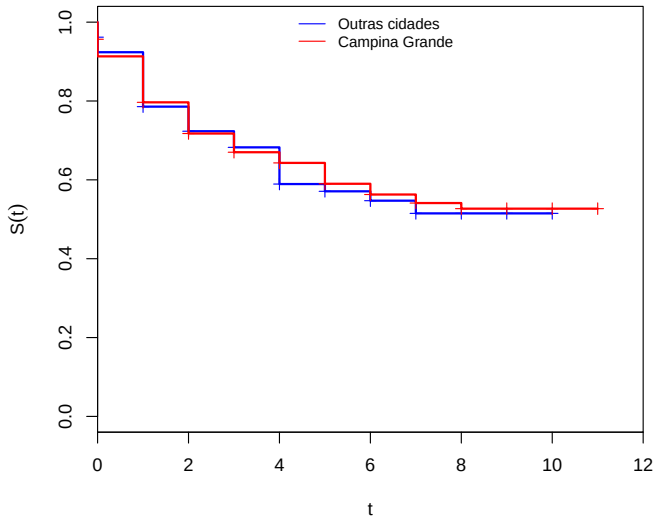


Figura 5.11: Curvas de sobrevivência estimadas pelo método de Kaplan e Meier (1958) para a covariável *Origem*.

O tipo de escola em que o aluno cursou o ensino médio, que contempla a escola privada ou pública, também foi abservado.

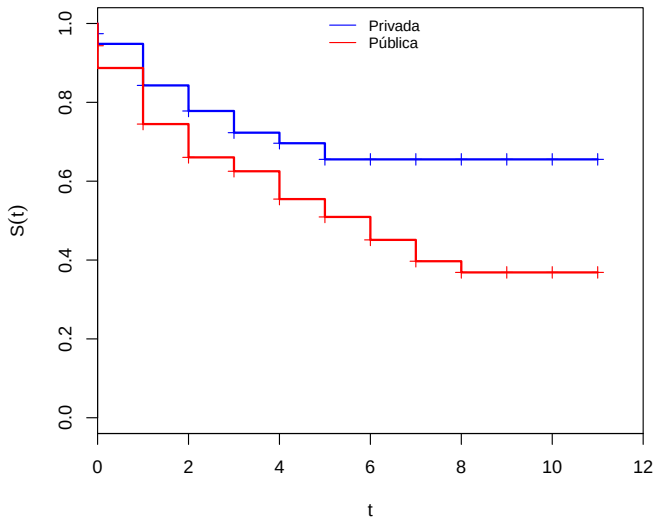


Figura 5.12: Curvas de sobrevivência estimadas pelo método de Kaplan e Meier (1958) para a covariável *Tipo de escola que cursou o ensino médio.*

Percebe-se, através da Figura 5.12, que os alunos que cursaram o ensino médio em escola privada têm maior probabilidade de sobrevivência do que os que cursaram em escola pública. Além disso, nota-se que, a partir do quinto período, há indicativo de fração de cura dentre os alunos que cursaram o ensino médio na rede privada, ou seja, a probabilidade de sobrevivência se mantém ao longo do tempo. Como não houve nenhum cruzamento entre as curvas de sobrevivência ao longo do tempo, realizou-se o teste Log-Rank para testar a existência de igualdade entre as curvas de sobrevivência. Através da estatística $T = 12,2$ e correspondente $p - valor = 0,000483$, conclui-se que existe diferença entre as curvas e que há indícios para explicar a fração de curados.

Por fim, a última covariável a ser estudada é a *forma de ingresso* no curso de Engenharia Ambiental. De acordo com a Figura 5.13, é possível verificar que a forma de ingresso no curso está influenciando no tempo de sobrevivência dos alunos. Aqueles que entraram no curso pelo vestibular têm maiores chances de se manter no curso do que aqueles que ingressaram pelo ENEM. Ressalta-se que o ENEM foi inserido na Universidade Estadual da Paraíba a partir do primeiro semestre de 2012. Por esse motivo, o período de acompanhamento para os alunos que ingressaram pelo vestibular é maior do que para aqueles que ingressaram pelo ENEM.

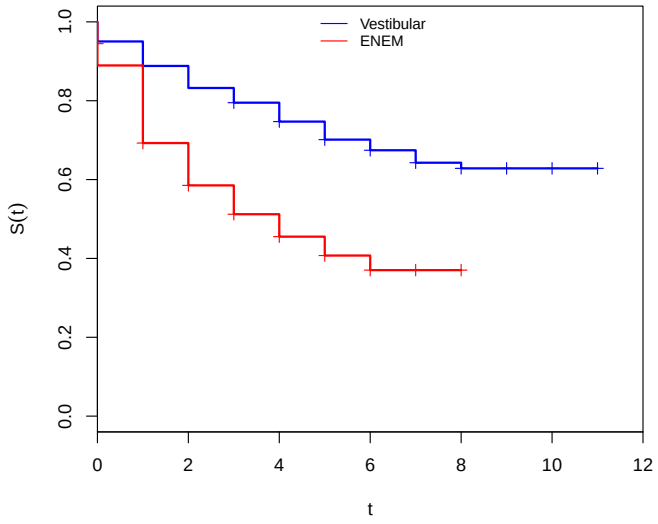


Figura 5.13: Curvas de sobrevivência estimadas pelo método de Kaplan e Meier (1958) para a covariável *Forma de ingresso no curso.*

De acordo com o teste de hipótese de Log-Rank, através da estatística $T = 26,1$ e correspondente $p - valor = 3,32 \times 10^{-7}$, conclui-se que a forma de ingresso no curso está influenciando o tempo de permanência do aluno, ou melhor, que as curvas de sobrevivência diferem significativamente, ao nível de 5% de significância. Com isso, há evidências para introduzir essa covariável no modelo proposto.

5.2.3 Modelos de regressão LLDFC

A análise descritiva realizada na Subseção 5.2.2 revelou que o modelo indicado para modelar o tempo de sobrevivência dos alunos do curso de Engenharia Ambiental é o Log-Logístico discreto com fração de cura. Ao analisar as covariáveis, verificou-se uma diferença mais relevante entre as curvas de sobrevivência da covariável *tipo de escola que cursou o ensino médio* e a *forma de ingresso* no curso. No entanto, para construir o modelos apresentados na Tabela 5.11, foi necessário realizar o processo de seleção de covariáveis utilizando o método *backward*.

Para a estimação dos coeficientes de regressão utilizou-se a função *optim* do *software* R, e para o modelo MRLLDFC1 e MRLLDFC2 não houve dificuldade quanto aos chutes iniciais para as estimativas dos coeficientes de regressão. No entanto, para o modelo MRLLDFC3, ao incluir todas as covariáveis no modelo e definir os chutes iniciais, de modo que houvesse convergência do método, não foi um processo rápido. Porém, após encontrar o modelo com todas as covariáveis, ao utilizar o processo *backward*, o processo para definir os chutes iniciais não apresentou nenhuma dificuldade.

Tabela 5.11: Estimativas dos parâmetros dos modelos de regressão Log-Logístico discreto com fração de cura.

Modelo	Parâmetros	Estimativas	Erro Padrão	p-valor
MRLLDFC1	γ	1,2486397	0,1914902	-
	β_0	2,7550303	0,4980269	$3,167981 \times 10^{-8}$
	β_1(Escola= Pública)	-0,5607788	0,2003517	$5,126559 \times 10^{-3}$
	β_2(Ingresso= ENEM)	-1,0356574	0,1878044	$3,496694 \times 10^{-8}$
	ϕ	0,9390689	0,2493794	-
MRLLDFC2	γ	1,5816794	0,1806450	-
	α	3,2967547	0,5074809	-
	ψ_0	-0,9178724	0,2863194	$1,347046 \times 10^{-3}$
	ψ_1(Escola= Pública)	1,2996302	0,4020341	$1,226534 \times 10^{-3}$
	ψ_2(Ingresso= ENEM)	1,9137643	0,5906139	$1,194053 \times 10^{-3}$
MRLLDFC3	γ	1,6880366	0,1957654	-
	β_0	0,8735784	0,1969317	$9,166851 \times 10^{-6}$
	β_1(Idade $\geq$ 20 anos)	0,3900659	0,1850706	$3,506038 \times 10^{-2}$
	ψ_0	-0,9947017	0,2693839	$2,220549 \times 10^{-04}$
	ψ_1(Escola= Pública)	1,2338352	0,3716076	$8,993259 \times 10^{-4}$
	ψ_2(Ingresso= ENEM)	1,7070823	0,4926850	$5,305169 \times 10^{-4}$

No primeiro modelo de regressão LLDFC, tem-se a estimativa do parâmetro γ igual a 1,25, o que indica que o modelo apresenta taxa de falha unimodal. A estimativa do parâmetro ϕ igual a 0,939 indica que há em torno de 6,1% de indivíduos curados nos dados analisados. É importante ressaltar que os indivíduos curados são os alunos que concluíram o curso e essa informação está disponível no banco de dados. Trata-se de 21 alunos em 360, ou seja, 5,83% de alunos que concluíram o curso. Esse resultado indica que o modelo foi bastante eficiente quanto à precisão da estimativa da fração de curados mesmo sem usar informação das covariáveis para a estimação do parâmetro de fração de cura, ou seja, usando apenas a variável latente apresentada na Seção 2.4. Além disso, esse resultado corrobora com os resultados apresentados nas simulações da Seção 4.3.

Ao fazer $e^{0,5607788} = 1,7520$, tem-se que os alunos que concluíram o ensino médio em escola privada têm o tempo mediano mais um semestre 1,75 vezes maior do que aqueles que

concluíram o ensino médio em escola pública.

A interpretação da estimativa do coeficiente β_2 do MRLLDFC1 é feita utilizando $e^{1,0346574} = 2,816957$. Dessa forma, interpreta-se que os alunos que ingressaram pelo vestibular têm o tempo mediano mais um semestre 2,82 vezes maior do que aqueles que ingressaram pelo ENEM.

Para o MRLLDFC2, agora com as informações das covariáveis sendo introduzidas no parâmetro ϕ, observa-se que as mesmas covariáveis do primeiro modelo foram selecionadas para o segundo modelo. Como a estimativa do coeficiente ψ_1 do segundo modelo é positiva, interpreta-se que a fração de não curados dos alunos que concluíram o ensino médio em escola pública é em torno de 1,3 vezes maior do que entre aqueles que concluíram o ensino médio em escola particular.

Ao analisar a estimativa do coeficiente ψ_2 do segundo modelo de regressão, tem-se que a fração de não curados dos alunos que ingressaram no curso de Engenharia Ambiental via ENEM é em torno de 1,9 vezes maior do que entre os alunos que ingressaram pelo vestibular. Esse resultado corrobora com o resultado do primeiro modelo, em que observou-se que a forma de ingresso interfere de maneira significativa no tempo mediano mais um semestre dos alunos.

Além disso, foi calculada a estimativa média de ϕ, sendo esta igual a 0,6506169, ou seja, o segundo modelo de regressão Log-Logístico discreto com fração de cura indicou a presença em torno de 65,1% de indivíduos suscetíveis nos dados em estudo. Esse resultado se aproxima da estimativa do parâmetro ϕ realizada na análise descritiva sem a presença das covariáveis. No entanto, é diferente do resultado do modelo MRLLDFC1, em que se conseguiu estimar o percentual de indivíduos curados bem próxima ao verdadeiro valor.

Ao analisar o MRLLDFC3, tem-se a junção das informações dos dois modelos anteriores, ou seja, observar o que está influenciando conjuntamente no tempo mediano mais um semestre e na fração de não curados.

Ao utilizar a estimativa do coeficiente β_1 e fazendo $e^{0,3900659} = 1,47707813$, interpreta-se que o tempo mediano mais um semestre dos alunos com idade maior ou igual a 20 anos é em torno de 1,48 vezes maior do que para os alunos com idade menor que 20 anos, mantendo as outras covariáveis fixas, inclusive as covariáveis relacionadas ao parâmetro ϕ.

De acordo com a estimativa de $\psi_1 = 1,2338352$, interpreta-se que os alunos que concluíram o ensino médio em escola pública têm fração de indivíduos não curados em torno de 1,23 vezes maior do que os alunos que concluíram o ensino médio em escola privada. Essa interpretação está de acordo com a interpretação da estimativa de ψ_1 do segundo modelo.

Aqueles que ingressaram no curso de Engenharia Ambiental via ENEM, apresentam fração de indivíduos não curados em torno de 1,71 vezes maior do que os alunos que ingressaram por meio do vestibular.

Além disso, foi também calculada a estimativa média de ϕ, sendo esta igual a 0,6148651, ou seja, o terceiro modelo de regressão Log-Logístico discreto com fração de cura indicou a presença de 61,5% de indivíduos suscetíveis nos dados em estudo. Esse resultado corroborou com a discussão realizada sobre a estimativa do parâmetro ϕ do modelo MRLDFC2. Desta forma, o segundo e o terceiro modelo de regressão LLDFC apresentam estimativas próximas para o parâmetro ϕ e coerência com o comportamento dos dados.

Ressalta-se que cada modelo tem sua particularidade e sua forma de interpretação. Nos três modelos apresentados, houve um ganho significativo de informação, comparando-se a modelos sem fração de cura. Diante disso, no intuito de utilizar os modelos para previsão, decidiu-se calcular a probabilidade de sobrevivência utilizando os três modelos de regressão, com base nas categorias das covariáveis utilizadas pelos modelos.

De acordo com a Figura 5.14, para o modelo MRLLDFC1 nota-se que os alunos que

concluíram o ensino médio em escola privada e ingressaram pelo vestibular têm probabilidade de sobrevivência maior do que os outros alunos.

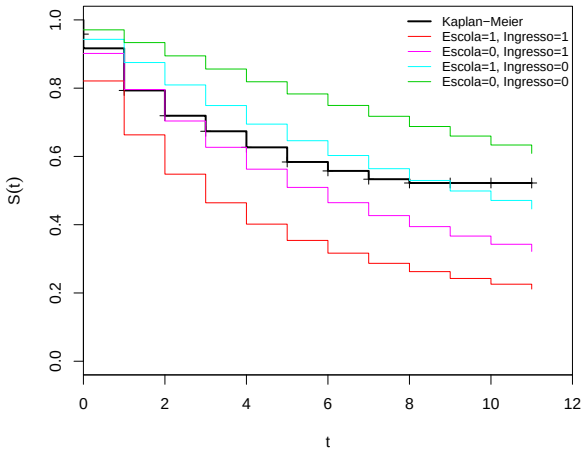


Figura 5.14: Curvas de sobrevivência estimadas pelo modelo MRLLDFC1 de acordo com as combinações das categorias das covariáveis.

Como é observado na Figura 5.15, a probabilidade de sobrevivência dos alunos que concluíram o ensino médio em escola particular e ingressaram no curso de Engenharia Ambiental pelo vestibular é maior do que para os outros alunos. Os resultados desse segundo modelo de regressão estão de acordo com os resultados obtidos pelo MRLLDFC1, com apenas a diferença na informação da fração de cura, pois percebe-se que os indivíduos que estão mais contribuindo para a fração de cura são os alunos que concluíram o ensino médio em escola privada e que ingressaram no curso pelo vestibular.

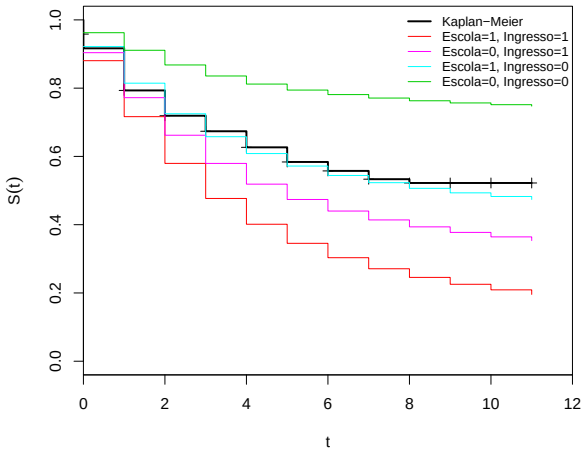


Figura 5.15: Curvas de sobrevivência estimadas pelo modelo MRLLDFC2 de acordo com as combinações das categorias das covariáveis.

Com a inclusão da covariável *idade* no modelo MRLLDFC3, nota-se que a probabilidade de sobrevivência dos alunos do curso de Engenharia Ambiental é maior para aqueles com idade igual ou superior a 20 anos, que concluíram o ensino médio em escola particular e

ingressaram no curso pelo vestibular. Esse resultado está de acordo com a interpretação do modelo MRLLDFC2 e mostra um ganho de informação acerca da probabilidade de sobrevivência dos alunos do curso de Engenharia Ambiental da Universidade Estadual da Paraíba, Campus I.

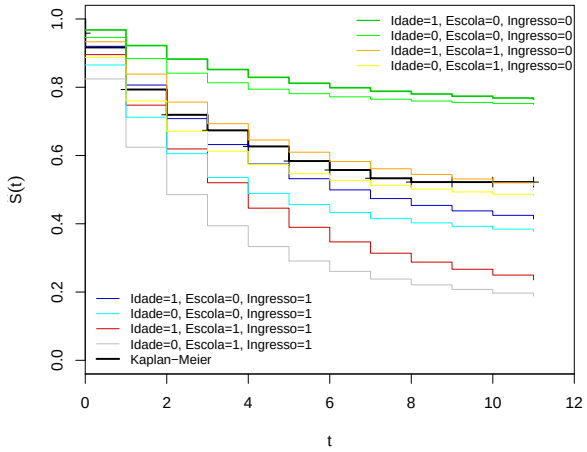


Figura 5.16: Curvas de sobrevivência estimadas pelo modelo MRLLDFC3 de acordo com as combinações das categorias das covariáveis.

Capítulo 6

Considerações finais

Neste trabalho foi apresentado o processo de discretização da distribuição Log-Logística no contexto de análise de sobrevivência e, dessa forma, construiu-se o modelo Log-Logístico discreto. Para populações em que há uma fração de indivíduos curados, foi proposto o modelo Log-Logístico discreto com fração de cura.

As simulações via Monte Carlo realizadas na Seção 4.2 deram suporte para analisar a acurácia dos estimadores dos modelos. Ao utilizar tamanhos de amostras maiores, os EQM's dos estimadores se reduz. No entanto, percebeu-se que em alguns casos, o tamanho da amostra igual a 200 apresenta resultados semelhantes a uma amostra de tamanho 500. Além do tamanho da amostra, outro fator que corrobora para a acurácia dos estimadores são as distâncias entre os valores dos parâmetros dos modelos. Percebeu-se que, à medida que o valor dos parâmetros e suas distâncias diminuem, há um efeito benéfico no que se refere aos erros quadráticos médios, ou seja, há uma redução nos EQMs dos estimadores.

De acordo com as simulações dos dados com distribuição Log-Logística discreta com fração de cura realizada na Seção 4.3, os EQM's dos estimadores $\hat{\phi}, \hat{\alpha}$ e $\hat{\gamma}$ diminuem à medida que o tamanho da amostra aumenta. Foi mostrado que a distância entre os valores dos parâmetros α e γ interfere na acurácia das estimativas dos parâmetros. Destaca-se ainda que as estimativas do parâmetro ϕ foram bem precisas em várias situações nos três cenários.

As aplicações apresentadas no Capítulo 5 permitiram ilustrar o uso do modelo Log-Logístico discreto e o modelo Log-Logístico discreto com fração de cura. Para a primeira aplicação, utilizou-se o banco de dados de alunos do curso de Computação, e através das análises preliminares decidiu-se estudar os modelos Weibull discreto e Log-Logístico discreto. Inicialmente verificou-se que o ajuste do modelo LLD foi melhor em relação ao WD. Entretanto, a conclusão do melhor ajuste deu-se a partir do erro máximo cometido pela estimação, assim como através das medidas AIC, AICc e BIC. Em todos esses procedimentos, verificou-se que o modelo mais adequado para modelar os dados em estudo é o Log-Logístico discreto. Em seguida, verificou-se o comportamento das curvas de sobrevivência das covariáveis em estudo, no intuito de construir o modelo de regressão Log-Logístico discreto.

Na segunda aplicação, utilizou-se o banco de dados dos alunos do curso de Engenharia Ambiental. Na análise preliminar, através do gráfico da função de sobrevivência estimada pelo método de Kaplan e Meier (1958), verificou-se a existência de indícios de fração de indivíduos curados nos dados. A partir disso, buscou-se investigar os modelos com fração de cura como o WDFC e o LLDFC. A partir das análises descritivas, decidiu-se que o modelo mais adequado para a modelagem dos dados é o Log-Logístico discreto com fração de cura. Diante disso, foi realizada a análise descritiva das covariáveis no intuito de identificar as covariáveis mais significativas para seguir com a construção do modelo de regressão LLDFC.

De acordo com as estimativas dos parâmetros, de maneira geral, verificou-se que os

alunos com idade superior ou igual a 20 anos têm maior probabilidade de sobrevivência do que aqueles com menos de 20 anos. Além disso, os alunos que concluíram o ensino médio em escola privada têm probabilidade de sobrevivência maior do que os que concluíram em escola pública e observou-se, também, que os alunos que ingressaram no curso pelo vestibular têm maior probabilidade de sobrevivência do que os que ingressaram no curso via ENEM. Esse resultado mostrou que existe uma coerência entre os dados dos alunos do ensino superior, mesmo que seja de cursos diferentes, pois de maneira geral, a mesma conclusão foi obtida pelo modelo de regressão LLD para a aplicação 1 com os dados dos alunos do curso de Computação.

Em relação ao modelo MRLLDFC1 observou-se que, por não usar as informações das covariáveis e apenas a variável latente para estimar o parâmetro ϕ, o modelo foi capaz de estimar de maneira bem precisa o verdadeiro valor da fração de indivíduos curados. A partir das probabilidades de sobrevivência calculadas para os três modelos de regressão LLDFC, observou-se principalmente que os três modelos são coerentes com a realidade dos dados. Além disso, a parte mais importante dessa última análise é observar as categorias dos indivíduos que estão potencializando a fração de cura nos dados em estudo.

De maneira geral, os modelos de regressão Log-Logístico discreto e Log-Logístico discreto com fração de cura para as duas aplicações revelaram bons ajustes e, principalmente, resultados bem coerentes com os dados analisados.

Para proposta de trabalhos futuros, sugere-se:

1. Fazer a análise de resíduos dos modelos;

2. Utilizar outras funções de ligação para estimar o parâmetro ϕ;

3. Construir os pacotes no R para LLD e LLDFC.

Referências Bibliográficas

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, IEEE, v. 19, n. 6, p. 716–723, 1974. 14

BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, v. 47, n. 259, p. 501–515, 1952. 11

BRUNELLO, G. H. V.; NAKANO, E. Y. Inferência bayesiana no modelo weibull discreto em dados com presença de censura. *TEMA - Tend. Mat. Apl. Comput.*, v. 16, p. 1–14, 2015. 53

CARDIAL, M. R. P. *Modelo Weibull Exponenciado para dados discretos em Análise de Sobrevivência: Uma abordagem clássica e bayesiana*. Dissertação (Mestrado) — Universidade de Brasília, 2016. 75

CARVALHO, M. S. et al. *Análise de Sobrevivência: teoria e aplicações em saúde*. Rio de Janeiro: SciELO-Editora FIOCRUZ, 2011. 6

CASELLA, G.; BERGER, R. L. *Inferência Estatística*. São Paulo: Cengage Learning, 2010. 31

COLOSIMO, E. A.; GIOLO, S. R. *Análise de Sobrevivência Aplicada*. São Paulo: Edgard Blücher, 2006. ix, 1, 3, 4, 8

FERNANDES, L. M. *Inferência Bayesiana em modelos discretos com fração de cura*. Dissertação (Mestrado) — Universidade de Brasília, 2013. 7

HOSMER, D.; LEMESHOW, S. *Applied Survival Analysis*. New York: Wiley, 1999. 28

JAMES, B. R. *Probabilidade: um curso em nível intermediário*. Rio de Janeiro: Instituto de Matemática Pura e Aplicada, CNPq, 1981. v. 12. 10

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, v. 53, n. 282, p. 457–481, 1958. x, xi, xiii, xiv, 8, 52, 53, 54, 55, 58, 59, 60, 62, 63, 64, 69

LAWLESS, J. F. *Statistical Models and Methods for Lifetime Data*. New York: John Wiley & Sons, 2011. v. 362. 24

LOUZADA-NETO, F.; PEREIRA, B. de B. Modelos em análise de sobrevivência. *Cadernos Saúde Coletiva, Rio de Janeiro*, v. 8, n. 1, p. 8–26, 2000. ix, 5

NAKAGAWA, T.; OSAKI, S. The discrete weibull distribution. *IEEE Transactions on Reliability*, IEEE, v. 24, n. 5, p. 300–301, 1975. 53

NAKANO, E. Y.; CARRASCO, C. G. Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência. *Trends in Applied and Computational Mathematics*, v. 7, n. 1, p. 91–100, 2006. 1, 10, 53

RAMOS, P. L. *Aspectos computacionais para Inferência na distribuição Gama generalizada*. Dissertação (Mestrado) — Universidade Estadual Paulista, 2014. ix, 7

SCHWARZ, G. Estimating the dimensional of a model. *Annals of Statistics, Hayward*, v. 6, p. 461–464, 1978. 15

SENGUPTA, D. Graphical tools for censored survival data. *Lecture Notes-Monograph Series*, JSTOR, p. 193–217, 1995. 60

SILVA, C. A. *Modelo de regressão Weibull discreto com fração de cura em dados de sobrevivência*. Dissertação (Mestrado) — Universidade de Brasília, 2015. 59

SUGIURA, N. Further analysis of the data by akaike's information criterion and the finite corrections: Further analysis of the data by akaike's. *Communications in Statistics-Theory and Methods*, v. 7, n. 1, p. 13–26, 1978. 15

TEAM, R. C. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*, 2015. Disponível em: ⟨http://www.R-project.org/⟩. 2, 13, 31

Apêndice A

Intervalo de confiança para os parâmetros

A.1 Parâmetro α

Considerando que $\alpha > 0$ e ao utilizar a transformação:

$$u = \log(\alpha) \Rightarrow \alpha = \exp(u),$$

e que:

$$\hat{u} = \log(\hat{\alpha}) \Rightarrow \hat{\alpha} = \exp(\hat{u}),$$

Sendo assim:

$$P\left(-Z_{1-\alpha/2} < \frac{\hat{u} - u}{\sqrt{\widehat{Var}(\hat{u})}} < Z_{1-\alpha/2}\right) = 1 - \alpha$$

$$-Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{u})} < \hat{u} - u < Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{u})}$$

$$\hat{u} - Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{u})} < u < \hat{u} + Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{u})}.$$

Ao aplicar exp(), na desigualdade, tem-se:

$$\exp\left(\hat{u} - Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{u})}\right) < \exp(u) < \exp\left(\hat{u} + Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{u})}\right).$$

Como tem-se $\alpha = \exp(u)$ e $\hat{\alpha} = \exp(\hat{u})$, então:

$$\exp(\hat{u})\exp\left(-Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{u})}\right) < \exp(u) < \exp(\hat{u})\exp\left(Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{u})}\right)$$

$$\hat{\alpha}\exp\left(-Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{u})}\right) < \alpha < \hat{\alpha}\exp\left(Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{u})}\right).$$

Sendo assim, o intervalo de confiança para o parâmetro α é dado por:

$$\left[\hat{\alpha}\exp\left(-Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{u})}\right); \hat{\alpha}\exp\left(Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{u})}\right)\right]$$

A.2 Parâmetro γ

Considerando que $\gamma > 0$ e ao utilizar a transformação:

$$v = \log(\gamma) \Rightarrow \gamma = \exp(v),$$

e que:

$$\hat{v} = \log(\hat{\gamma}) \Rightarrow \hat{\gamma} = \exp(\hat{v}),$$

Sendo assim:

$$P\left(-Z_{1-\alpha/2} < \frac{\hat{v} - v}{\sqrt{\widehat{Var}(\hat{v})}} < Z_{1-\alpha/2}\right) = 1 - \alpha$$

$$-Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{v})} < \hat{v} - v < Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{v})}$$
$$\hat{v} - Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{v})} < v < \hat{v} + Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{v})}.$$

Ao aplicar exp(), na desigualdade, tem-se:

$$\exp\left(\hat{v} - Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{v})}\right) < \exp(v) < \exp\left(\hat{v} + Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{v})}\right).$$

Como tem-se $\gamma = \exp(v)$ e $\hat{\gamma} = \exp(\hat{v})$, então:

$$\exp(\hat{v})\exp\left(-Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{v})}\right) < \exp(v) < \exp(\hat{v})\exp\left(Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{v})}\right)$$

$$\hat{\gamma}\exp\left(-Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{v})}\right) < \gamma < \hat{\gamma}\exp\left(Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{v})}\right).$$

Sendo assim, o intervalo de confiança para o parâmetro α é dado por:

$$\left[\hat{\gamma}\exp\left(-Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{v})}\right); \hat{\gamma}\exp\left(Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{v})}\right)\right]$$

A.3 Parâmetro ϕ

Considerando que $0 < \phi < 1$ e ao utilizar a transformação:

$$w = \log(-\log\phi) \Rightarrow \phi = \exp[-\exp(w)],$$

e que:

$$\hat{w} = \log(-\log\hat{\phi}) \Rightarrow \hat{\phi} = \exp[-\exp(\hat{w})].$$

Sendo assim:

$$P\left(-Z_{1-\alpha/2} < \frac{\hat{w} - w}{\sqrt{\widehat{Var}(\hat{w})}} < Z_{1-\alpha/2}\right) = 1 - \alpha$$

$$-Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{w})} < \hat{w} - w < Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{w})}$$
$$\hat{w} - Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{w})} < w < \hat{w} + Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{w})}.$$

Ao aplicar exp[− exp()], na desigualdade, tem-se:

$$\exp\left[-\exp\left(\hat{w}-Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{w})}\right)\right]>\exp\left[-\exp\left(w\right)\right]>\exp\left[-\exp\left(\hat{w}+Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{w})}\right)\right]$$

Como tem-se $\phi=\exp[-\exp(w)]$ e $\hat{\phi}=\exp[-\exp(\hat{w})]$, então:

$$\exp\left[-\exp\left(\hat{w}\right)\exp\left(-Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{w})}\right)\right]>\phi>\exp\left[-\exp\left(\hat{w}\right)\exp\left(Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{w})}\right)\right]$$

$$\hat{\phi}^{\exp\left(Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{w})}\right)}<\phi<\hat{\phi}^{\exp\left(-Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{w})}\right)}.$$

Sendo assim, o intervalo de confiança para o parâmetro ϕ é dado por:

$$\left[\hat{\phi}^{\exp\left(Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{w})}\right)};\hat{\phi}^{\exp\left(-Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{w})}\right)}\right].$$

A.4 Parâmetro q

De acordo com Cardial (2016) o intervalo de confiança para o parâmetro q do modelo Weibull discreto é dado por:

$$\left[\hat{q}^{\exp\left(Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{u})}\right)};\hat{q}^{\exp\left(-Z_{1-\alpha/2}\sqrt{\widehat{Var}(\hat{u})}\right)}\right],$$

sendo $\hat{u}=log[-log(\hat{q})]$.

Apêndice B

Script em R

B.1 Script - Simulação

```
################Pacotes utilizados
require(survival)
require(emdbook)

################Simulação LLD

rLLD<-function(n,alpha,gamma){
u<-runif(n)
t<-floor((alpha*((u/(1-u))^(1/gamma))))
return(t)}

################Estimação LLD

dens.LL<-function(t,alpha,gamma){
(1/(1+(t/alpha)^gamma)) - (1/(1+((t+1)/alpha)^gamma)) }
sobr.LL<-function(t,alpha,gamma){
( 1+((t+1)/alpha)^gamma)^-1}

VLL<-function(parametro,tempo,delta){
L1<-log(dens.LL(tempo,parametro[1],parametro[2]))
L2<-log(sobr.LL(tempo,parametro[1],parametro[2]))
-sum(L1*delta + L2*(1-delta) )  }  #VEROSSIMILHANÇA

m0 <- optim(c(chute1,chute2),VLL,hessian = T,delta=censura,tempo=tempo)
hessianam0 <- m0$hessian
invhm0 <- solve(hessianam0)
epm0 <- sqrt(diag(invhm0)) # Erro padrão das estimativas

################Intervalo de confiança das estimativas LLD

var.v1<-deltavar(fun = log(alpha),Sigma=solve(m0$hessian[1,1]),
meanval=c(alpha = m0$par[1]))
LI.alpha<-(m0$par[1])*(exp(-qnorm(0.975)*sqrt(var.v1)))
LS.alpha<-(m0$par[1])*(exp(qnorm(0.975)*sqrt(var.v1)))

var.w1<-deltavar(fun = log(gamma),Sigma=solve(m0$hessian[2,2]),
meanval=c(gamma = m0$par[2]))
LI.gamma<-(m0$par[2])*(exp(-qnorm(0.975)*sqrt(var.w1)))
LS.gamma<-(m0$par[2])*(exp(qnorm(0.975)*sqrt(var.w1)))
```

```
################Estimação do MRLLD
VLLD<-function(par,t,censura){
     gamma<-par[1]
     beta0<-par[2]
     beta1<-par[3]
     beta2<-par[4]
     beta3<-par[5]
     beta4<-par[6]
     alpha<-exp(beta0+beta1*x1+beta2*x2+beta3*x3+beta4*x4)
     if ( (alpha>0) &&(gamma>0))
     return (-1*(
     sum(censura*log( (1/(1+(t/alpha)^gamma)) - (1/(1+((t+1)/alpha)^gamma))
         ))
     +sum((1-censura)* ( 1+((t+1)/alpha)^gamma)^-1)))
     else return (-Inf)
}
m1<- optim(c(chute1,chute2,chute3,chute4,chute5,chute6),VLLD,hessian = T,
    censura=censura1,t=tempo1)

################Simulação via Monte Carlo - LLD

SimulaLLD<-function(M,n,alpha,gamma,p.censura){
set.seed(2017)
alpha.est<-gamma.est<-numeric(M)
for(i in 1:M){
censura<-rbinom(n,1,1 - p.censura)
tempo<-rLLD(n,alpha,gamma)
m01 <- optim(c(alpha,gamma),VLL,delta=censura,tempo=tempo)
alpha.est[i]<-m01$par[1]
gamma.est[i]<-m01$par[2]}
a<-mean(alpha.est[])
g<-mean(gamma.est[])
vicio.alpha<-mean(alpha.est[]-alpha)
EQM.alpha<-mean((alpha.est[]-alpha)^2)
vicio.gamma<-mean(gamma.est[]-gamma)
EQM.gamma<-mean((gamma.est[]-gamma)^2)
return(list(alpha=a,vicio.alpha=vicio.alpha,EQM.alpha=EQM.alpha,gamma=g,
    vicio.gamma=vicio.gamma,EQM.gamma=EQM.gamma))}

################Simulação Via Monte Carlo - LLDFC

SimulaLLDFC<-function(M,n,phi,alpha,gamma,p.censura){
set.seed(2017)
alpha.est<-gamma.est<-phi.est<-p.censura.T<-numeric(M)
for(i in 1:M){
tempo<-numeric(n)
censura<-numeric(n)
suscept<-rbinom(1,n,phi) # gerando a quantidade de Suscetíveis
x<-rLLD(suscept,alpha,gamma)
censura.suscept<-rbinom(suscept,1,1-p.censura)
tempo<-c(x,rep(max(x),n-suscept))
censura<-c(censura.suscept,rep(0,n-suscept))
mfc<-optim(c(phi,alpha,gamma),VLLFC,censura=censura,t=tempo)
phi.est[i]<-mfc$par[1]
alpha.est[i]<-mfc$par[2]
gamma.est[i]<-mfc$par[3]
p.censura.T[i]=((suscept*p.censura)+(n-suscept))/n}
```

```
f<-mean(phi.est[])
a<-mean(alpha.est[])
g<-mean(gamma.est[])
p.cens.T<-mean(p.censura.T[])
vicio.phi<-mean(phi.est[]-phi)
EQM.phi<-mean((phi.est[]-phi)^2)
vicio.alpha<-mean(alpha.est[]-alpha)
EQM.alpha<-mean((alpha.est[]-alpha)^2)
vicio.gamma<-mean(gamma.est[]-gamma)
EQM.gamma<-mean((gamma.est[]-gamma)^2)
return(list(p.Censura.Total=p.cens.T, phi=f, vicio.phi=vicio.phi,EQM.phi=EQM
    .phi, alpha=a, vicio.alpha=vicio.alpha,EQM.alpha=EQM.alpha, gamma=g, vicio.
    gamma=vicio.gamma,EQM.gamma=EQM.gamma))}


################Estimação LLDFC
sobr.LLFC<-function(t,phi,alpha,gamma){(1-phi)+phi*( 1+((t+1)/alpha)^gamma
    )^-1}

VLLFC<-function(par,t,censura){
phi<-par[1]
alpha<-par[2]
gamma<-par[3]
if ((phi>0) && (phi<1) && (alpha>0) && (gamma>0))
return (-1*(
sum(censura*log(phi))
+sum(censura*log( (1/(1+(t/alpha)^gamma)) - (1/(1+((t+1)/alpha)^gamma)) ))
+sum((1-censura)*log((1-phi) + (phi)* ( 1+((t+1)/alpha)^gamma)^-1))))
else return (-Inf)
}

m1 <- optim(c(chute1,chute2,chute3),VLLFC, hessian = T, censura=censura, t=
    tempo)
hessianam1 <- m1$hessian
invhm1 <- solve(hessianam1)
epm1 <- sqrt(diag(invhm1)) # Erro padrão das estimativas

################Intervalo de confiança das estimativas LLDFC

var.u<-deltavar(fun = log(-log(phi)),Sigma=solve(m1$hessian[1,1]),
meanval=c(phi = m1$par[1]))
LI.phi<-(m1$par[1])^(exp(qnorm(0.975)*sqrt(var.u)))
LS.phi<-(m1$par[1])^(exp(-qnorm(0.975)*sqrt(var.u)))

var.v<-deltavar(fun = log(alpha),Sigma=solve(m1$hessian[2,2]),
meanval=c(alpha = m1$par[2]))
LI.alpha<-(m1$par[2])*(exp(-qnorm(0.975)*sqrt(var.v)))
LS.alpha<-(m1$par[2])*(exp(qnorm(0.975)*sqrt(var.v)))

var.w<-deltavar(fun = log(gamma),Sigma=solve(m1$hessian[3,3]),
meanval=c(gamma = m1$par[3]))
LI.gamma<-(m1$par[3])*(exp(-qnorm(0.975)*sqrt(var.w)))
LS.gamma<-(m1$par[3])*(exp(qnorm(0.975)*sqrt(var.w)))
```

```
################Estimação WD
sobr.WD<-function(t,q,gamma){q^((t+1)^gamma)}

VWD<- function(par,tempo,censura,x){
q <- par[1]
gamma <- par[2]
if((q > 0) && (q <1) && (gamma > 0))
return(-1*(sum(censura*log(q^(tempo^gamma) - q^((tempo+1)^gamma)) +
(1 - censura)*log(q^((tempo + 1)^gamma)))))
else return(-Inf)
}
moWD<-optim(c(chute1,chute2), VWD,hessian=T,tempo=tempo,censura=censura)

################Estimação WDFC

sobr.WDFC<-function(t,f,q,b){(1-f) + f* q^((t+1)^b)}

VWDFC<-function(par,tempo,censura,x){
f<-par[1]
q<-par[2]
b<-par[3]
if ( (q>0) && (q<1) && (f>0) && (f<1) && (b>0)  )
return (-1*(
sum( censura*log(f) )
+sum( censura * log( q^(tempo^b) - q^((tempo+1)^b) ) )
+sum(   (1-censura)*log((1-f) + f* q^((tempo+1)^b) ) )
))
else return (-Inf)
}
moWDFC<-optim(c(chute1,chute2,chute3),hessian=T, VWDFC,tempo=tempo,censura
    =censura)
moWDFC

################Estimação do MRLLDFC1
VLLFCM2<-function(par,t,censura){
phi<-par[1]
gamma<-par[2]
beta0<-par[3]
beta1<-par[4]
beta2<-par[5]
alpha<-exp(beta0+beta1*x1+beta2*x2)
if ((phi>0) && (phi<1) && (alpha>0) && (gamma>0))
return (-1*(
sum(censura*log(phi))
+sum(censura*log( (1/(1+(t/alpha)^gamma)) - (1/(1+((t+1)/alpha)^gamma)) ))
+sum((1-censura)*log((1-phi) + (phi)* ( 1+((t+1)/alpha)^gamma)^-1))))
else return (-Inf)
}
m1.1 <- optim(c(chute1,chute2,chute3,chute4,chute5),VLLFCM2,hessian = T,
    censura=censura,t=tempo)
```

```
################Estimação do MRLDFC2
VLLFCM2P<-function(par,t,censura){
alpha<-par[1]
gamma<-par[2]
psi0<-par[3]
psi1<-par[4]
psi2<-par[5]
phi<-exp(psi0+psi1*x1+psi2*x2)/(1+exp(psi0+psi1*x1+psi2*x2))
if ((phi>0) && (phi<1)&& (alpha>0) && (gamma>0))
return (-1*(
sum(censura*log(phi))
+sum(censura*log( (1/(1+(t/alpha)^gamma)) - (1/(1+((t+1)/alpha)^gamma)) ))
+sum((1-censura)*log((1-phi) + (phi)* ( 1+((t+1)/alpha)^gamma)^-1))))
else return (-Inf)
}
m1.1P <- optim(c(chute1,chute2,chute3,chute4,chute5),VLLFCM2P,hessian = T,
    censura=censura,t=tempo)

################Estimação do MRLLDFC3
VLLFCM2PA<-function(par,t,censura){
gamma<-par[1]
psi0<-par[2]
psi1<-par[3]
psi2<-par[4]
beta0<-par[5]
beta1<-par[6]
phi<-exp(psi0+psi1*z1+psi2*z2)/
(1+exp(psi0+psi1*z1+psi2*z2))
alpha<-exp(beta0+beta1*x1)
if ((phi>0) && (phi<1)&&(alpha>0) && (gamma>0))
return (-1*(
sum(censura*log(phi))
+sum(censura*log( (1/(1+(t/alpha)^gamma)) - (1/(1+((t+1)/alpha)^gamma)) ))
+sum((1-censura)*log((1-phi) + (phi)* ( 1+((t+1)/alpha)^gamma)^-1))))
else return (-Inf)
}
m1.1PA <- optim(c(chute1,chute2,chute3,chute4,chute5,chute6),VLLFCM2PA,
    hessian = T,censura=censura,t=tempo)


################Estimativa da função de sobrevivência por K-M

KM<-survfit(Surv(tempo,censura)~1, conf.int=F)
summary(KM)
plot(KM,lwd=2, mark=3, conf.int = FALSE, mark.time=KM$time[KM$n.censor >
    0],xlim=c(0,12),xlab="t",ylab=substitute(hat(S)(t)))
################Estimativa da função de risco acumulado por K-M

Kaplan<-cumprod(1-KM$n.event/KM$n.risk)
Ht<--(log(Kaplan))
plot(Ht,type="S",col=1,xlab="t",ylim=c(0,1),ylab=expression(hat(H)(t)))
points(c(0.5,12),c(0,1.3),type="l",lty=2, col=2)
```