Milene Regina dos Santos

Métodos de inferência Bayesiana e frequentista assumindo a distribuição defectiva de Gompertz na presença de fração de curas: aplicações a dados médicos

Ribeirão Preto

Agosto de 2019

Milene Regina dos Santos

Métodos de inferência Bayesiana e frequentista assumindo a distribuição defectiva de Gompertz na presença de fração de curas: aplicações a dados médicos

Monografia para exame de defesa de Mestrado apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo do Programa de Pós Graduação em Saúde Pública. Orientador: Jorge Alberto Achcar. Coorientador: Edson Zangiacomi Martinez. Linha de Pesquisa: Métodos Quantitativos em Saúde. Versão corrigida da Dissertação de Mestrado apresentado ao Programa de Pós-Graduação em Saúde Pública em 06/09/2019.

Universidade de São Paulo

Saúde Pública

Ribeirão Preto

Agosto de 2019

FICHA CATALOGRÁFICA

Santos, Milene Regina

Métodos de inferência Bayesiana e frequentista assumindo a distribuição defectiva de Gompertz na presença de fração de curas: aplicações a dados médicos. Ribeirão Preto, 2019.
53 p.: il, 30 cm

Dissertação de Mestrado, apresentada à Faculdade de Medicina de Ribeirão Preto/USP. Área de concentração: Saúde Pública.

Orientador: Achcar, Jorge Alberto.

1. Estimadores de Máxima Verossimilhança. 2. Inferência Bayesiana. 3. Distribuição defectiva. 4. Análise de Sobrevida.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001

Agradecimentos

Agradeço primeiramente a Deus, que iluminou as minhas escolhas, colocando pessoas no meu caminho que fizeram todo esse sonho ser concretizado.

Agradeço ao Prof. Dr. Edson Zangiacomi Martinez que depositou sua confiança em mim desde o nosso primeiro e-mail trocado.

Ao Prof. Dr. Jorge Alberto Achcar por fazer grande parte da minha formação em Análise de Sobrevivência, que teve paciência em fazer todas as correções pertinentes e toda a burocracia necessária para o dia da defesa.

Aos meus pais Mauro José dos Santos e Maria Izabel Gonçalves dos Santos por estarem ali sempre eu que precisei, por se preocuparem com a minha felicidade, por me amarem e serem minha fonte de inspiração.

Agradeço a minha irmã Aline Cristina dos Santos e sua filhinha Alice Cristina dos Santos Freitas, por fazerem meus fins semana mais agitados e cheios de fofura.

A Marina Bernardes, doutoranda da Bioengenharia do CAASO, que me falou sobre o Prof. Edson.

Ao Prof. Dr. Rafael Rosales, que me entusiasmou na graduação a seguir esse caminho.

A toda a minha família, amigos, e funcionários do HC USP RP, por fazerem de uma certa forma, parte do meu dia a dia, em especial, aos secretários Sérgio e Paula, aos professores e aos amigos Louraine Marie, José André Queiroz, Maurício Perez, Guilherme Zubatch, Marcos Peres, Vanessa Motozo, Danielle Peralta, Renata Serra, Samara dos Santos, Marília, Lucas, Ricardo Pantaleone, Luiz Ricardo, Aldo, Verena, Juliana e Priscila Queiroz.

Agradeço principalmente ao Ricardo Puziol de Oliveira, que me ajudou muito na condução deste mestrado, que gastou comigo o seu tempo me auxiliando, que me motivou em cada momento de frustração, que tomou iniciativas que eu não tomaria por medo, que me incentivou a apresentar trabalhos fora do estado e por ter me ensinado as linguagem R, do Latex e Análise de Sobrevivência.

Santos, M.R. Métodos de inferência Bayesiana e frequentista assumindo a distribuição defectiva de Gompertz na presença de fração de curas: aplicações a dados médicos. Dissertação de Mestrado, apresentada à Faculdade de Medicina de Ribeirão Preto, 2019. 53 p.

Resumo

A análise de sobrevivência, também chamada de análise de confiabilidade nas aplicações em engenharia, consiste em uma classe de métodos estatísticos usados para estudar o tempo até a ocorrência de um evento de interesse, como o tempo até o óbito de um paciente, o tempo de recuperação após um tratamento médico, o tempo de hospitalização devido a uma doença, entre vários outros eventos de interesse médico. Dados de sobrevivência geralmente apresentam censuras ocasionadas por limitação de tempo no seguimento dos pacientes, perdas de seguimentos por outras causas ou censuras ocasionadas pelo próprio planejamento do experimento. Outra característica comum em análise de sobrevivência é a presença de indivíduos imunes ou uma fração de curas, em que parte dos pacientes não são sujeitos ao evento de interesse. Nesse caso, a literatura estatística apresenta muitos modelos, com destaque especial aos modelos de mistura e não-misturas. Alternativamente, modelos baseados em distribuições impróprias, denominadas como distribuições defectivas podem ser usados para analisar dados com essas características. Nesta dissertação serão apresentadas inferências Bayesianas e de Máxima Verossimilhança para os parâmetros do modelo de fração de cura assumindo a distribuição defectiva de Gompertz na presença de dados censurados e covariáveis. Para ilustrar a metodologia proposta, são consideradas aplicações com dados relacionados aos tempos de sobrevida de pacientes com câncer cervical e pacientes portadores do vírus HIV. Na análise Bayesiana, sumários a posteriori de interesse são obtidos usando métodos de simulação MCMC (Monte Carlo em Cadeias de Markov) para gerar amostras da distribuição a posteriori conjunta de interesse.

Palavras-chaves: Estimadores de Máxima Verossimilhança. Inferência Bayesiana. Distribuições defectivas. Análise de Sobrevida. Distribuição Gompertz Modificada. Métodos MCMC.

Santos, M.R. Bayesian and frequentist inference methods assuming a Gompertz defective distribution in the presence of fraction of cures: applications to medical data. Master's Dissertation, responsible for the Faculty of Medicine of Ribeirão Preto, 2019. 53p.

Abstract

Survival analysis, also called reliability analysis in engineering applications, consists of a class of statistical methods used to study the time to the occurrence of an event of interest, such as time to death of a patient, recovery after a medical treatment, length of hospital stay due to illness, among several other events of medical interest. Survival data usually present censors caused by time limitation in the follow-up of patients, loss of follow-ups due to other causes or censors caused by the planning of the experiment itself. Another common feature in survival analysis is the presence of immune individuals or a fraction of cures, in which part of the patients are not subject to the event of interest. In this case, the statistical literature presents many models, with particular emphasis on mixing models and non-mixtures (see for example, MALLER and ZHOU, 1996). Alternatively, models based on improper distributions, referred to as defective distributions, can be used to analyze data with these characteristics. In this dissertation will be presented Bayesian and Maximum likelihood inferences for the parameters of the curing fraction model assuming the Gompertz defective distribution in the presence of censored and covariable data. To illustrate the proposed methodology, applications with data related to the survival times of patients with cervical cancer and patients with HIV are considered. In Bayesian analysis, a posteriori interest summaries are obtained using Monte Carlo Markov Chain Simulation (MCMC) simulation methods to generate samples of the joint posterior distribution of interest.

Keywords: Maximum Likelihood Estimators. Bayesian Inference. Defective distributions. Survival Analysis. Modified Gompertz distribution. MCMC methods.

Sumário

1 Revisão de literatura

A análise de Sobrevivência é utilizada em estudos que investigam o tempo até a ocorrência de um evento de interesse. Este evento pode ser, por exemplo, o tempo até a recidiva de uma doença, óbito, alta hospitalar, cura de uma morbidade, entre outros. A principal característica desses estudos é a presença de censuras, que é a informação parcial dos dados. As ferramentas da análise de sobrevivência são muito utilizadas na pesquisa médica. Entretanto, é comum assumir que todos os indivíduos estudados sejam sujeitos ao evento sob investigação dentro de um período suficientemente grande de tempo e isto pode não ser real em alguns casos. Por exemplo, em uma pesquisa clínica, uma proporção de pacientes pode responder positivamente ao tratamento aplicado e, consequentemente, não morrer devido à doença considerada. Isto motivou o desenvolvimento de vários modelos estatísticos que incorporam a presença de fração de curas, também chamados de modelos de longa duração (FAREWELL; V.T, 1982; MALLER; ZHOU, 1996).

Modelos de análise de sobrevivência com fração de cura foram introduzidos por vários autores, como Boag (1949) e Berkson e Gage (1952), baseados em uma função de sobrevivência representada por uma mistura de distribuições, descrevendo os indivíduos suscetíveis ao evento de interesse com uma dada probabilidade $(1 - p)$ e os indivíduos não suscetíveis ao evento de interesse com uma dada probabilidade p definida no intervalo $(0, 1)$.

Alternativamente aos modelos de misturas, a literatura apresenta os modelos baseados em distribuições defectivas ou impróprias que são caracterizados por distribuições de probabilidade em que o resultado da integração da sua função densidade de probabilidade (área sob a curva) em todo o seu domínio é diferente de 1. Nesta situação, duas distribuições impróprias introduzidas na literatura e utilizadas em análise de dados de sobrevivência são as distribuições Gompertz e a Gaussiana inversa (BALKA; DESMOND; MCNICHOLAS, 2011). Essas distribuições são úteis no ajuste dos modelos com fração de cura, pois não assumem o pressuposto de que todos os indivíduos sejam suscetíveis ao evento (ROCHA et al., 2016).

Recentemente, outras distribuições de probabilidade defectivas também foram introduzidas na literatura como as distribuições defectivas baseadas na família Marshall-Olkin (ROCHA et al., 2016) e na família Kumaraswamy (ROCHA et al., 2015).

1.1 Exemplos com dados reais

Como ilustração e motivação para o presente estudo, são considerados dois bancos de dados médicos relacionados a dados de sobrevida onde o primeiro banco de dados é referente a mulheres portadoras do carcinoma cervical que receberam o tratamento padrão pela Federação Internacional de Ginecologia e Obstetria e evento de interesse é a reincidência da doença após a cirurgia e o segundo banco e dados é de pacientes portadores do vírus HIV, no qual o evento de interesse é o óbito do paciente após diagnosticado com AIDS. São duas situações diferentes que mostram a necessidade de ajuste de modelos de sobrevivência mais apropriados do que o uso de modelos tradicionais existentes.

1.1.1 Dados de pacientes com câncer cervical

Consideremos os dados de um estudo que incluiu um total de 148 mulheres diagnosticadas e tratadas por carcinoma cervical invasivo entre 1992 e 2002 (BRENNA et al., 2004). Para os propósitos do presente estudo, foi considerada uma subamostra de 118 mulheres que receberam o tratamento padrão recomendado pela Federação Internacional de Ginecologia e Obstetrícia (FIGO). Definimos a sobrevivência livre de doença como o tempo em meses completos a partir da data da cirurgia para o primeiro evento de recorrência da doença. Quase 48% dos dados são observações censuradas. A Figura 1 apresenta a função de sobrevivência para estes dados, estimada pelo método de Kaplan-Meier. Observa-se que após os 60 meses a curva tende a se estabilizar, e os tempos de censuras tornam-se maiores do que os tempos de sobrevivência completos observados.

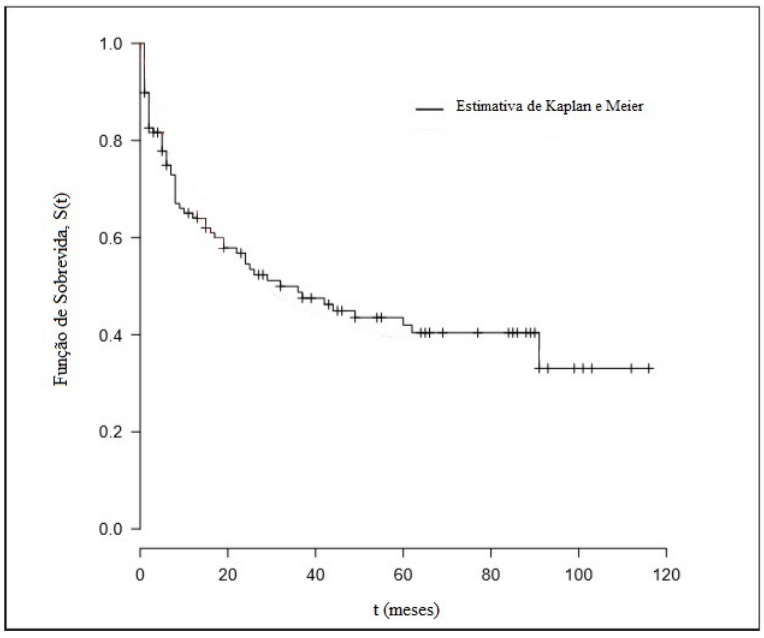

Figura 1 – Gráfico de Kaplan - Meier da função de sobrevida livre da doença considerando os dados de câncer cervical.

1.1.2 Dados de portadores de HIV

A síndrome da imunodeficiência adquirida (AIDS) foi reportada inicialmente na literatura em torno de 1981 para pacientes infectados com o vírus HIV com o relato de alguns casos (BRITO et al., 2001). Essa doença tornou-se mundialmente conhecida na época devido sua alta letalidade e ausência de cura. Apesar de ainda ser uma doença incurável, o avanço de terapias mais eficazes tem retardado cada vez mais o aparecimento da AIDS e consequentemente aumentando sobrevida dos pacientes portadores de HIV, sendo que atualmente é comum encontrar pacientes portadores do vírus HIV que sobrevivem 20 anos ou mais após o diagnóstico. Neste caso, muitos pacientes podem ir a óbito por outras causas diferentes da AIDS, o que leva à necessidade de modelos estatísticos na presença de fração de cura para a análise estatística dos dados.

O banco de dados usado neste estudo está disponível no site da FIOCRUZ <http://sobrevida.fiocruz.br/aidsclassico.html>. Os dados resultam de um estudo de coorte feito no Instituto de Pesquisa Clínica Evandro Chagas (Ipec/Fiocruz) entre os anos de 1986 e 2000, com duração de 4822 dias, incluindo 193 indivíduos portadores do vírus HIV (CARVALHO et al., 2011). De acordo com o livro Carvalho et al. (2011), ao término do estudo ainda haviam 101 indivíduos vivos, o que sugere uma estabilidade na sobrevida. As Tabelas 1 e 2 apresentam características descritivas do banco de dados em relação a variável sexo e o quadro 1.1 exibe todas as variáveis do estudo.

Tabela 1 – Quantidade de dados censurados e não censurados referente ais indivíduos portadores de HIV de acordo com o sexo.

	Mulheres	Homens	Total
Dados censurados	33	70	103
Dados não censurados	16	74	90
Total de indivíduos	49	144	193

Tabela 2 – Variável tratamento do banco de dados de indivíduos portadores de HIV e relação ao gênero.

	Tratamento 0	Tratamento 1	Tratamento 2	Tratamento 3	Total
Mulheres	4	24	17	2	47
Homens	40	76	18	12	146
Total	44	100	35	14	193

A média de sobrevida geral para estes individuos foi de 939 dias, sendo o menor tempo registrado igual a 16 dias e o maior tempo igual a 3228 dias de sobrevida. Na Tabela 3 é apresentada uma parte do banco de dados deste estudo. Por fim, a Figura 2 apresenta o gráfico de Kaplan - Meier (KAPLAN; MEIER, 1958; LAWLESS, 1982; JR; LEMESHOW, 1999; KLEIN; MOESCHBERGER, 1997) para a curva de sobrevivência que permite a inclusão dos dados censurados. Já a Figura 3 apresenta as funções

Variável	Descrição
id	identificação do paciente
ini	data do diagnóstico da Aids (em dias)
fim	data do óbito (ou perda do paciente)
temp	dias de sobrevivência do diagnóstico até o óbito
stat	0 = censura, 1 = óbito
sex	F = feminino, M = masculino
escol	0 = sem escolaridade, 1 = ensino fundamental, 2 = ensino médio, 3 = ensino superior
ida	idade na data do diagnóstico de Aids (20 a 68 anos)
risc	0 = homossexual masculino, 1 = usuário de drogas injetáveis, 2 = transfusão, 3 = contato sexual com HIV+, 5 = hétero c/múltiplos parceiros, 6 = dois fatores de risco
acom	acompanhamento: 0 = ambulatorial/hospital-dia, 1 = internação posterior, 2 = internação imediata
obit	S = óbito, N = não óbito, I = ignorado
anotrat	ano do início do tratamento (1990 a 2000), 9 = sem tratamento
trat	terapia antirretroviral: 0 = nenhum, 1 = mono, 2 = combinada, 3 = potente
doen	de apresentação: 1 = pcp, 2 = pcp pulmonar, 3 = pcp disseminada, 4 = toxoplasmose, 5 = sarcoma, 7 = outra doença, 8 = candidíase, 9 = duas doenças, 10 = herpes, 99 = definido por cd4
prpcp	profilaxia para pneumocistis: 0 = sem profilaxia, 2 = primária, 3 = secundária, 4 = ambas

Quadro 1.1: Descrição das variáveis nas colunas dadas na Tabela 1 referente aos pacientes portadores do vírus HIV.

de sobrevidas estimadas pelo método não-paramétrico de Kaplan-Meier para os dados de sobrevida estratificados pelos quatro tratamentos, em que é possível observar que os tratamentos 2 e 3 apresentam probabilidades de sobrevida maiores em tempos fixados quando comparados aos tratamentos 0 e 1. A estimativa de Kaplan Meier será melhor explicada na seção 1.2.4.

Tabela 3 – Dados de sobrevida dos pacientes portadores do vírus HIV.

id	ini	fim	temp	stat	sex	escol	ida	risc	acom	obit	anotrat	trat	doen	prpcp
1	1243	2095	852	1	M	3	34	0	1	S	1991	1	4	3
2	2800	2923	123	1	M	2	38	6	1	S	9	0	7	4
3	1250	2395	1145	1	M	NA	32	0	1	S	1992	1	3	4
4	1915	4670	2755	0	M	NA	43	6	0	N	1992	1	10	4
5	2653	4770	2117	0	M	NA	40	0	1	N	1992	1	5	4
6	3	332	329	0	M	NA	34	0	1	I	9	0	7	0
7	36	96	60	1	M	NA	27	0	2	S	9	0	7	0
8	1	152	151	1	M	0	22	6	2	S	9	0	3	0
9	544	2107	1563	1	M	2	44	NA	0	S	9	0	10	0
10	71	1318	1247	1	M	2	23	0	2	S	9	0	3	4
11	946	1030	84	1	M	1	40	0	1	S	9	0	3	0
12	802	1016	214	1	M	2	33	0	1	S	9	0	3	4
13	266	291	25	0	M	NA	41	NA	1	I	9	0	4	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
193	4301	4805	504	0	M	1	48	0	0	N	1999	3	99	2

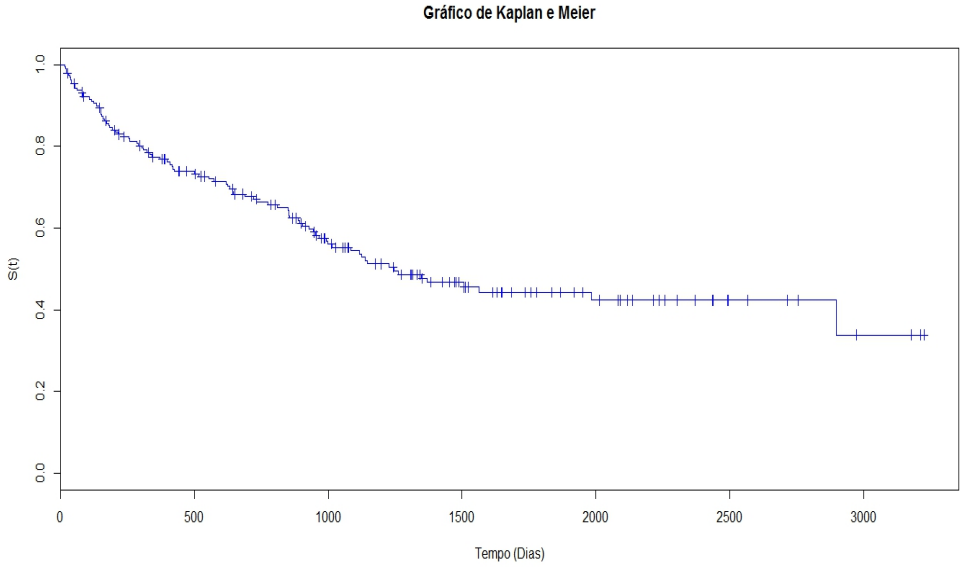

Figura 2 – Gráfico de Kaplan - Meier da função de sobrevida dos pacientes portadores de HIV.

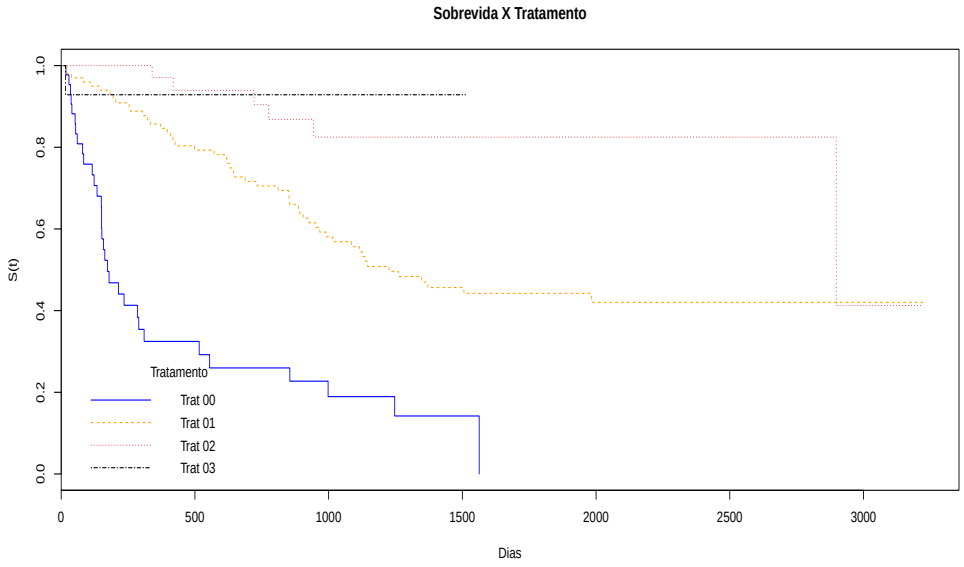

Figura 3 – Gráfico de Kaplan - Meier da função de sobrevida dos pacientes portadores de HIV, para dados classificados de acordo com os tratamentos.

1.2 Uma introdução à Análise de Sobrevivência

A análise de sobrevivência possui como característica básica atuar com dados incompletos e completos conjuntamente. Os dados incompletos representam informações censuradas, as quais nos trazem as informações parciais do tempo real até a ocorrência do evento de interesse. Esta informação parcial significa que de alguma forma perdeu-se o

contato com o paciente antes do término do estudo ou que o paciente não sofreu o evento de interesse (COLOSIMO; GIOLO, 2006). Assim, as ferramentas estatísticas da análise de sobrevivência nos permitem estimar taxas de sobrevida sem perder nenhuma informação fornecida pelo banco de dados.

1.2.1 Dados censurados

Sem a presença de censuras, as técnicas estatísticas clássicas, como a análise de regressão ou análise de variância, poderiam ser utilizadas na análise desses tipos de dados Colosimo e Giolo (2006). Os dados censurados, resultados provenientes de um estudo de sobrevivência não devem ser desprezados na análise, pois fornecem informações importantes sobre o tempo de sobrevida de pacientes e a sua omissão no cálculo das estatísticas de interesse pode acarretar em conclusões enviesadas. Existem várias formas de censuras, sendo a mais usual a censura à direita, que ocorre quando o evento de interesse não é observado até o término do estudo ou até o ultimo instante em que o individuo é acompanhado. Censuras aleatórias são frequentes na área médica; elas acontecem quando um paciente é retirado no decorrer do estudo sem ter ocorrido o evento de interesse ou podem ocorrer caso o paciente seja perdido de seguimento. A Análise de Sobrevivência apresenta diferentes tipos de censuras, são elas:

- **Censura à direita:** os dados apresentam censuras à direita quando o estudo determina uma data para finalizar a pesquisa e mesmo após esta data ser atingida esses dados ainda não passaram pelo evento de interesse.
- **Censura à esquerda:** ocorre quando o indivíduo entra no estudo mesmo já tendo sofrido o evento de interesse. Por exemplo, quando o objetivo é determinar a idade em que crianças aprendem a ler em uma comunidade.
- **Censura intervalar:** ocorre, por exemplo, quando os pacientes não lembram exatamente quando o evento de interesse ocorreu, mas sabem que foi dentro de um certo intervalo de tempo (COLOSIMO; GIOLO, 2006).

Em geral, para indicar se um dado é censurado ou não usualmente é definida uma variável indicadora de censuras para $i = 1, 2, \ldots, n$, onde n igual ao tamanho amostral dada por,

$$\delta_i = \begin{cases} 0, & \text{se o tempo é censurado} \\ 1, & \text{caso contrário.} \end{cases}$$

1.2.2 Função de Sobrevivência

Em Análise de Sobrevivência a probabilidade do evento não ocorrer até o tempo t é dada por $S(t)$, a função densidade de probabilidade é denotada por $f(t)$ e $F(t)$ é a função de distribuição acumulada. As funções $S(t)$ e $F(t)$ são dadas por respectivamente por,

$$S(t) = P(T \geq t) = 1 - F(t), \tag{1.1}$$

e,

$$F(t) = P(T < t), t \in R^{+}. \tag{1.2}$$

De acordo com a função de sobrevida $S(t)$, todos os indivíduos são suscetíveis ao evento de interesse, dado que $\lim_{t\to\infty} S(t) = 0$, o que não é adequado para algumas situações práticas, por exemplo, quando em um conjunto de dados o evento de interesse é o óbito devido uma doença específica e para alguns indivíduos este evento nunca ocorrerá, pois são considerados "curados" ou imunes. Neste contexto, Boag (1949) propôs um modelo de mistura padrão com fração de cura, cuja função de sobrevivência é dada por:

$$S(t) = p + (1 - p)S_0(t), \tag{1.3}$$

em que $S_0(t)$ é uma função basal de sobrevivência descrita por uma distribuição de probabilidade conhecida. O parâmetro p é a probabilidade de indivíduos não-suscetíveis (fração de cura), isto é, que nunca sofrerão o evento de interesse e $1 - p$ é a probabilidade de indivíduos suscetíveis. Neste caso, $S(t)$ não tenderá a zero quando t assumir um número de grande magnitude, ou seja:

$$\lim_{t\to\infty} S(t) = \lim_{t\to\infty}(p + (1 - p)S_0(t)) = p + \lim_{t\to\infty}(1 - p)S_0(t) = p \tag{1.4}$$

pois,

$$S_0(t) = 0, \text{quando t tende ao infinito.} \tag{1.5}$$

Portanto, p é a fração de cura, que descreve a proporção de indivíduos não sujeitos ao evento de interesse, sendo $0 < p < 1$.

1.2.3 Função de Risco

A Função de Risco, denotada por $h(t)$ quantifica numericamente a taxa de um indivíduo sofrer o evento de interesse no instante t, dado que isto ainda não ocorreu (CARVALHO et al., 2011). A função de risco é deduzida a partir da função de densidade de probabilidade e da função de sobrevivência, a partir da relação,

$$h(t) = \frac{f(t)}{S(t),}, sendo \in [t, t + \Delta t] \tag{1.6}$$

Relacionado a isto, a taxa é dada por: $h(t)\Delta t$, em um intervalo $[t, t\Delta]$, sendo Δt uma variação instantânea do tempo (AL-MALKI, 2014). A função de taxa de risco traz muitas informações sobre o grupo. Enquanto diferentes funções de sobrevida podem ser parecidas, suas funções de taxa de risco podem diferenciar-se (COLOSIMO; GIOLO, 2006).

1.2.4 Estimador de Kaplan-Meier para a função de sobrevivência

Um estimador não-paramétrico muito popular e extensivamente usado na literatura médica é dado pelo estimador produto-limite (EPL) de Kaplan - Meier (1958) ilustrado na Figura 4, considerando a situação na presença de fração de curas e a situação sem a presença de fração de curas.

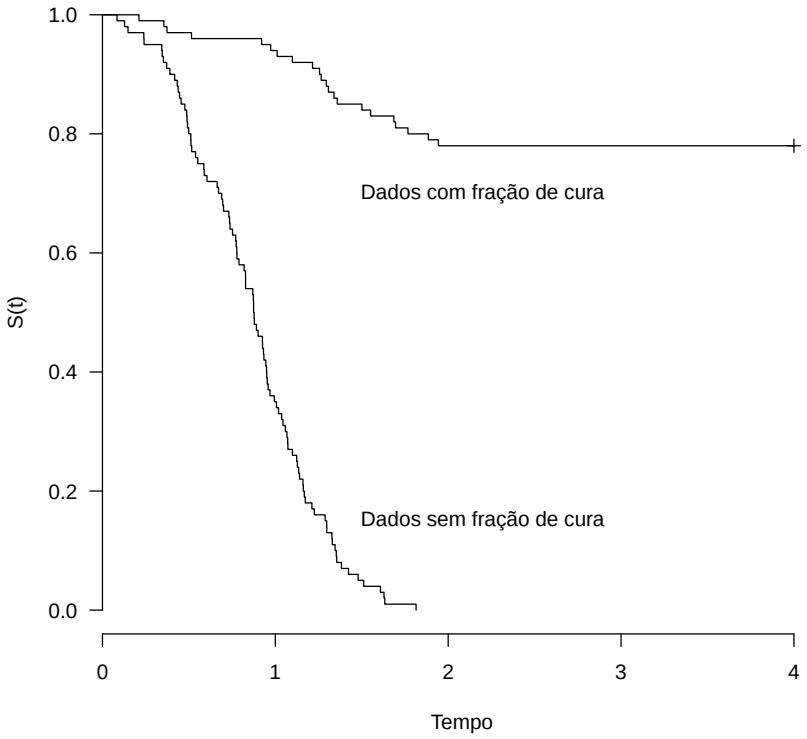

Figura 4 – Estimativa de Kaplan-Meier da função de sobrevivência para dados hipotéticos com e sem fração de cura.

Observar que o EPL de Kaplan-Meier $\hat{S}(t)$ é uma função escada em que os degraus representam os tempos observados em que ocorreu o evento de interesse. Na Figura 4, observa-se um gráfico dos dados com fração de cura convergindo a uma proporção de indivíduos "imunes", enquanto a outra curva, quando o tempo aumenta, a função de sobrevivência se aproxima do valor zero.

O estimador de Kaplan - Meier é um não-viciado e de máxima verossimilhança não-paramétrica para a função de sobrevivência (ver Kaplan e Meier, 1958). Esta técnica nos permite estimar a taxa de sobrevida em um determinado ponto da curva, sendo adequada para amostras com a presença de dados censurados. Sem a presença de dados censurados o estimador de Kaplan e Meier se reduz ao estimador empírico da função de sobrevivência.

Quando os tempos de sobrevida não são sujeitos a censuras, temos o estimador empírico da função de sobrevivência dado por,

$$\hat{S}(t) = \frac{\text{nº de observações que não falharam até o tempo t}}{\text{nº total de observações de estudo}} \tag{1.7}$$

Na presença de censuras assumir que d_i é o número de eventos de interesse observados no tempo t_i e n_i é o número de observações sob risco até a ocorrência do evento no tempo t_i. O estimador Kaplan - Meier e sua variância assintótica são dados respectivamente por:

$$\hat{S}(t) = \prod_{t_i<t}^{n} \frac{(n_i - d_i)}{n_i} \tag{1.8}$$

e,

$$\hat{Var}(\hat{S}(t)) = \sum_{i=0}^{n} \frac{d_i}{n_i(n_i - d_i)} \tag{1.9}$$

1.2.5 Uso de distribuições paramétricas para os dados de sobrevivência

As técnicas paramétricas para a análise de dados de sobrevivência, diferentemente das técnicas não-paramétricas muito populares em estudos médicos, assumem uma distribuição de probabilidade que depende de um ou vários parâmetros que devem ser estimados a partir dos dados.

1.2.5.1 Distribuição Exponencial

A Distribuição Exponencial possui um único parâmetro e caracteriza-se por apresentar uma função risco constante. As expressões para $f(t)$, $S(t)$ e $h(t)$ são dadas respectivamente por

$$f(t) = \frac{1}{\alpha} exp\left\{-\frac{t}{\alpha}\right\}, \tag{1.10}$$

$$S(t) = exp\left\{-\frac{t}{\alpha}\right\}, \tag{1.11}$$

e,

$$h(t) = \frac{1}{\alpha}. \tag{1.12}$$

onde $t > 0$ e o parâmetro desconhecido α (média de T) também maior que zero.

A distribuição exponencial é também muito usada na área de confiabilidade, onde é modelado o tempo de vida de um componente ou de um sistema. Já na área da saúde, dificilmente uma doença não altera a taxa de risco no tempo, porém em certas situações, onde o estudo tem um acompanhamento muito curto, podemos considerar que o risco do evento de interesse ocorrer não se altera.

1.2.5.2 Distribuição Weibull

A distribuição de Weibull foi proposta originalmente por Waloddi Weibull (1951). Sua popularidade em aplicações práticas se deve ao fato dela apresentar uma grande variedade de formas, todas com uma propriedade básica: a sua função de riscos pode ser monótona crescente, decrescente e constante. A função densidade de probabilidade (f.d.p) é dada por,

$$f(t_i) = \frac{\alpha t_i^{\alpha-1} exp\left\{-\left(\frac{t_i}{\lambda}\right)^{\alpha}\right\}}{\lambda^{\alpha}} \tag{1.13}$$

em que, $t_i > 0$ denota os tempos de sobrevida. Os parâmetros $\lambda > 0$ e $\alpha > 0$ denotam respectivamente, os parâmetros de escala e de forma para a distribuição. Diferentes valores de α levam a diferentes formas para a distribuição o que a torna muito flexível na análise de dados para tempos de sobrevida. Na análise de sobrevivência o grande interesse usualmente é na função de sobrevivência $S(t^*) = P(T > t^*)$, em que t^* é um tempo qualquer fixado. Assumindo a distribuição de Weibull com f.d.p. (2.14), a função de sobrevivência é dada por,

$$S(t^*) = exp\left\{-\left(\frac{t^*}{\lambda}\right)^{\alpha}\right\}. \tag{1.14}$$

A função de risco $h(t)$ ou taxa instantânea de falha da distribuição de Weibull (ver, por exemplo, Lawless, 1982) é dada, de $h(t) = f(t)/S(t)$, tal que

$$h(t) = \alpha \frac{t^{\alpha-1}}{\lambda^{\alpha}}. \tag{1.15}$$

Observar que se $\alpha = 1$, temos a distribuição exponencial sendo um caso especial da distribuição de Weibull. A função de risco $h(t)$ dada por (2.16) é estritamente crescente para $\alpha > 1$, estritamente decrescente para $\alpha < 1$ e constante para $\alpha = 1$. Assim, observa-se uma grande flexibilidade de ajuste aos dados. A média e a variância da distribuição de Weibull com densidade dada por (2.14) são dadas respectivamente por:

$$\mu = E(T) = \lambda \Gamma\left(1 + \frac{1}{\alpha}\right) \tag{1.16}$$

e

$$\sigma^2 = Var(t) = \lambda^2 \left\{ \Gamma\left(1 + \frac{2}{\alpha}\right) - \Gamma\left[1 + \frac{1}{\alpha}\right]^2 \right\} \tag{1.17}$$

em que $\Gamma(.)$ denota uma função gama, $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt$.

1.2.5.3 Distribuição Log-Normal

A função densidade de probabilidade de uma variável aleatória T com distribuição log-normal é dada por,

$$f(t; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2}\left[\frac{ln(t) - \mu}{\sigma}\right]^2 \right), t > 0 \tag{1.18}$$

em que $\mu > 0$ e $\sigma > 0$ são respectivamente a média e o desvio-padrão para os logaritmos dos tempos de sobrevida. As funções de sobrevivência e função de risco neste caso, não apresentam uma forma analítica explícita, sendo expressas por,

$$S(t) = \Phi\left(\frac{-ln(t) + \mu}{\sigma}\right) \text{ e } h(t) = \frac{f(t)}{S(t)} \tag{1.19}$$

em que $\Phi(.)$ é a função distribuição acumulada de uma distribuição normal padrão (distribuição normal com média zero e variância igual a um). A função de risco não é monótona como a da distribuição Weibull, ou seja, ela cresce, atinge um valor máximo e depois decresce.

1.2.5.4 Distribuição Log-logística

Se T é uma variável aleatória, tal que $ln(T)$ tem distribuição logística, então T segue uma distribuição Log-logística, com função de densidade de probabilidade dada por,

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \left(1 + \left(\frac{t}{\alpha}\right)^\gamma\right)^{-2}, t > 0 \tag{1.20}$$

em que $\alpha > 0$ é o parâmetro de forma e $\gamma > 0$ o parâmetro de escala. As funções de sobrevivência e de risco são dadas, respectivamente por,

$$S(t) = \frac{1}{1 + \left(\frac{t}{\alpha}\right)^\gamma} \text{ e } h(t) = \frac{\gamma \left(\frac{t}{\alpha}\right)^{\gamma-1}}{\alpha\left[1 + \frac{t}{\alpha}^\gamma\right]} \tag{1.21}$$

em que, para $\gamma > 1$, observa-se comportamento similar ao da distribuição log-normal, isto é, o risco é crescente alcançando um pico e a partir daí começa a declinar; para $\gamma < 1$, o risco é decrescente, similar a função de risco da distribuição Weibull.

1.2.6 Estimação dos parâmetros dos modelos

Existem alguns métodos de estimação conhecidos na literatura (ver por exemplo, Colosimo e Giolo, 2006) sendo que uma abordagem mais usual para dados com censuras é dada pelo método de máxima verossimilhança. O método de estimação inclui os dados censurados, é relativamente simples em termos de interpretação e possui propriedades ótimas para grandes amostras.

1.2.7 Método de máxima verossimilhança em modelos de sobrevivência

Supor uma amostra de observações não censuradas $t_1, ..., t_n$ de uma população, onde os tempos de sobrevida tenham uma densidade $f(t;\theta)$, onde θ é um parâmetro desconhecido. A função de verossimilhança para o parâmetro θ assumindo dados completos é dada por,

$$L(\theta) = \prod_{i=1}^{r} f(t_i, \theta) \tag{1.22}$$

Para definir a verossimilhança para dados censurados à direita, considere T uma variável aleatória representando o tempo de falha de um paciente e C uma variável aleatória, independente de T, representando o tempo de censura. Para um dado paciente temos como dado observado, $t = min(T, C)$ e defina:

$$\delta = \begin{cases} 1 \text{ se } T \leq C \\ 0 \text{ se } T > C \end{cases} \tag{1.23}$$

sendo que δ é uma variável indicadora de censura. Agora, suponha que os pares (T_i, C_i), para $i = 1, ..., n$ formam uma amostra aleatória de tamanho n. As observações podem ser divididas em duas partes: as r primeiras observações ordenadas são as observações não censuradas $(1, 2, ..., r)$ e as $n - r$ seguintes são observações censuradas $(r + 1, r + 2, ..., n)$. Neste caso, a expressão para a função de verossimilhança é definida como,

$$L(\theta) = \prod_{i=1}^{r} f(t_i, \theta) \prod_{i=r+1}^{n} S(t_i, \theta) \tag{1.24}$$

ou equivalentemente,

$$L(\theta) = \prod_{i=1}^{n} [f(t_i, \theta)]^{\delta_i} [S(t_i, \theta)]^{1-\delta_i} = \prod_{i=1}^{n} [h(t_i, \theta)]^{\delta_i} [S(t_i, \theta)] \tag{1.25}$$

Na prática é sempre conveniente considerar o logaritmo da função de verossimilhança. Os valores que maximizam $L(\theta)$ ou equivalentemente $l(\theta) = logL(\theta)$ são os estimadores de máxima verossimilhança. Eles são encontrados resolvendo-se o sistema de equações dado por:

$$U(\theta) = \frac{\partial logL(\theta)}{\partial \theta} = 0 \tag{1.26}$$

1.3 Uso de métodos Bayesianos em análise de sobrevivência

Os métodos bayesianos têm se mostrado muito eficazes e úteis na análise de dados, inclusive na área da saúde.

O método Bayesiano combina informações objetivas contidas em distribuições de probabilidades a priori com a informação amostral (verossimilhança). A inferência bayesiana se caracteriza pela determinação de uma distribuição a posteriori de um parâmetro ou vetor de parâmetros definidos no espaço paramétrico. Na inferência bayesiana, a incerteza sobre os parâmetros desconhecidos associa-se a uma distribuição de probabilidade Box e Tiao (1992), enquanto que, na inferência frequentista, os parâmetros são valores fixos ou constantes, aos quais não se associam a qualquer distribuição de probabilidade. Sob o enfoque bayesiano, condicionalmente aos dados y observados, descreve-se a incerteza sobre o valor de algum parâmetro θ não observado, em termos de probabilidades ou densidades de probabilidades (BOX e TIAO, 1992). O parâmetro θ pode ser um escalar ou um vetor de parâmetros.

A informação prévia sobre um parâmetro θ, dada por distribuição a priori, é incorporada ao estudo usando a fórmula de Bayes, que combina a informação contida nos dados, resultando na distribuição a posteriori. Dessa forma é possível incluir na análise de dados o conhecimento de um pesquisador ou especialista, quando disponível.

1.3.1 Fórmula de Bayes

Sejam os eventos $A_1, A_2, \ldots, A_k$ formando uma sequência de eventos mutualmente exclusivos e exaustivos formando uma partição do espaço amostral Ω, isto é, $U_{j=1}^{k} A_j = \Omega$ e $A_i \cap A_j = \emptyset$ (conjunto vazio) para $i \neq j$ tal que $P(U_{j=1}^{k} A_j) = \sum_{j=1}^{k} A_j = 1$. Então para qualquer outro evento $B(B \subset \Omega)$, temos

$$P(A_i \mid B) = \frac{p(B \mid A_i)p(A_i)}{\sum_{j=1}^{k} p(B \mid A_j)p(A_j)}, \tag{1.27}$$

para todo i variando de 1 até k.

Considere agora θ um vetor de parâmetros a serem estimados. Logo, pela fórmula de Bayes, tem-se a seguinte distribuição de probabilidade a posteriori para θ,

$$\pi(\theta \mid y) = \frac{\pi(\theta)f(y \mid \theta)}{\int \pi(\theta)f(y \mid \theta)d\theta} \tag{1.28}$$

assumindo que θ seja contínuo, $\pi(\theta)$ é a distribuição a priori conjunta para θ e $f(y \mid \theta) = L(\theta) = \prod_{i=1}^{n} f(y_i \mid \theta)$ é a função de verossimilhança de θ.

Assim, a partir da fórmula de Bayes, temos,

$$\pi(\theta \mid y) \propto L(\theta \mid Y)\pi(\theta) \tag{1.29}$$

Assim temos que a distribuição a posteriori é proporcional à verossimilhança multiplicada pela distribuição a priori. A função de probabilidade a priori representa o conhecimento prévio a respeito dos elementos de θ antes da observação dos dados, refletindo a incerteza em relação aos possíveis valores de θ antes do vetor de dados y ser selecionado. A função a posteriori incorpora o estado de incerteza do conhecimento prévio a respeito do parâmetro θ após a observação dos dados em y e a função de verossimilhança representa a contribuição de y para o conhecimento sobre θ.

1.3.2 Distribuições a priori

Uma distribuição a priori para um parâmetro pode ser elicitada de várias formas. As quais podemos assumir:

(a) Podemos assumir distribuições a priori definidas no domínio de variação do parâmetro de interesse. Como caso particular, poderíamos considerar uma distribuição a priori Beta que é definida no intervalo (0, 1) para proporções que também são definidas no intervalo (0, 1) ou considerar uma priori normal para parâmetros definidos em toda reta;

(b) Podemos assumir uma distribuição a priori baseada em informações de um ou mais especialistas;

(c) Podemos considerar métodos estruturais de elicitação de distribuições a priori (PAULINO et al., 2003);

(d) Podemos considerar distribuições a priori não informativas quando temos total ignorância sobre os parâmetros de interesse;

(e) Podemos usar métodos bayesianos empíricos em dados ou experimentos prévios para construir a priori de interesse.

1.3.3 Métodos de simulação para amostras da distribuição a posteriori

Na obtenção de sumários a posteriori, é necessário resolver integrais múltiplas, muitas vezes, complicadas, o que exige o uso de métodos numéricos ou de aproximações de integrais, especialmente quando a dimensão do vetor de parâmetros é grande. Daí, surge a necessidade do uso de métodos computacionais poderosos, como os métodos de Monte Carlo em cadeias de Markov (MCMC) que incluem alguns algoritmos de simulação

de amostras da distribuição a posteriori conjunta de interesse, como os algoritmos de Metropolis-Hastings e o amostrador de Gibbs. É importante salientar que os métodos com base em simulação de amostras da distribuição a posteriori conjunta de interesse, como, por exemplo, o método de Monte Carlo em cadeias de Markov (MCMC), passaram a ser muito utilizados com o avanço dos recursos computacionais em termos de hardware e software. Esses métodos consistem na simulação de uma variável aleatória através de uma cadeia de Markov, no qual a sua distribuição assintóticamente se aproxima da distribuição a posteriori de interesse (BERNARDO; SMITH, 1994).

A cadeia de Markov é um processo estocástico no qual o próximo estado da cadeia depende somente do estado atual e dos dados. No entanto, como existe certa dependência com os valores iniciais fixados no processo de simulação, na prática uma amostra simulada inicial é descartada após um período de aquecimento, chamada *Burn-in- sample*.

As formas mais usuais de simulação dos métodos MCMC são dadas pelo amostrador de Gibbs e o algoritmo de Metropolis-Hastings. Essas duas formas simulam amostras da distribuição a posteriori conjunta a partir das distribuições condicionais (GELFAND; SMITH, 1990).

O amostrador de Gibbs nos permite gerar amostras da distribuição a posteriori conjunta desde que as distribuições condicionais completas possuam formas fechadas ou conhecidas. Por outro lado, o algoritmo de Metropolis-Hastings permite gerar amostras da distribuição a posteriori conjunta com distribuições condicionais completas possuindo ou não uma forma conhecida ou fechada.

1.3.4 O amostrador de Gibbs

Suponha que $\theta = (\theta_1, ..., \theta_k)$ é um vetor de parâmetros aleatórios e y é o vetor dos dados observados; tem-se como objetivo, obter inferências sobre a distribuição a posteriori conjunta $\pi(\theta \mid y) = \pi(\theta_1, ..., \theta_{k|Y})$ (BERNARDO e SMITH, 1994). Dado um vetor arbitrário de valores iniciais $\theta = (\theta_1^{(\theta)}, ..., \theta_k^{(\theta)})$ para as quantidades desconhecidas, implementa-se o seguinte procedimento iterativo:

Obtém-se $\theta_1^{(1)}$ de $\pi(\theta_1 \mid y, \theta_2^{(0)}, ..., \theta_k^{(0)})$

Obtém-se $\theta_2^{(1)}$ de $\pi(\theta_2 \mid y, \theta_1^{(1)}, \theta_3^{(0)}, ..., \theta_k^{(0)})$

$\vdots$

Obtém-se $\theta_k^{(1)}$ de $\pi(\theta_k \mid y, \theta_1^{(1)}, ..., \theta_{k-1}^{(1)})$

Obtém-se $\theta_1^{(2)}$ de $\pi(\theta_1 \mid y, \theta_2^{(1)}, ..., \theta_k^{(1)})$

$\vdots$

Agora, suponha que este processo é continuado através de i iterações e é independentemente replicado m vezes para que ao final se tenha m replicações do vetor amostrado $\theta^t = (\theta_1^{(t)}, ..., \theta_k^{(t)})$, onde θ^t é uma realização de uma cadeia de Markov com probabilidade de transição dada por,

$$p(\theta^t, \theta^{t+1}) = \prod_{t=1}^{k} \pi(\theta_{kl}^{t+1} \mid y, \theta_1^{t+1}, ..., \theta_{l-1}^{t+1}, \theta_{l+1}^{t}, ..., \theta_k^t) \tag{1.30}$$

Com t tendendo para o infinito, $(\theta_1^{(t)}, ..., \theta_k^{(t)})$ tende para uma distribuição a um vetor aleatório cuja densidade conjunta é $\pi(\theta \mid y)$, ou seja, a distribuição a posteriori de interesse. Em particular, θ_i^t tende em distribuição a uma quantidade aleatória cuja densidade é $\pi(\theta_i \mid y)$, também chamada de densidade marginal a posteriori de θ_i. Desta maneira, para t grande, as replicações $(\theta_{i1}^{(t)}, ..., \theta_{im}^{(t)})$ são aproximadamente uma amostra aleatória de $\pi(\theta_i \mid y)$.

Após a geração de amostras da distribuição a posteriori de interesse, utilizamos essas amostras para obter estimadores de Monte Carlo para sumários a posteriori de interesse como a média a posteriori, o desvio-padrão a posteriori e intervalos de credibilidade de interesse.

1.3.5 O algoritmo de Metropolis-Hastings

Suponha que se deseja simular uma densidade a posteriori $\pi(\theta \mid y)$. Um algoritmo de Metropolis-Hastings se inicia com um valor inicial θ^0 e especifica uma regra para a simulação do t-ésimo valor da sequência θ^t dado o $(t-1)$-ésimo valor da sequência $\theta^{(t-1)}$. Esta regra consiste em uma densidade proposta (ou densidade geradora) a qual simula um valor candidato θ^* e o cálculo da uma probabilidade de aceitação P, que indica a probabilidade do valor candidato ser aceito para ser o próximo valor na sequência. Especificamente, esse algoritmo pode ser descrito da seguinte forma (ALBERT, 2007):

1. Simular um valor candidato θ^* de uma densidade proposta $p(\theta^* \mid \theta^{t-1})$.
2. Calcular a razão:
$$R = \frac{\pi(\theta^* \mid y)p(\theta^{t-1} \mid \theta^*)}{(\pi(\theta^{t-1} \mid y)p(\theta^* \mid \theta^{t-1}))} \tag{1.31}$$
3. Calcular a probabilidade de aceitação $P = min(R, 1)$
4. Amostrar um valor θ^t tal que $\theta^t = \theta^*$ com probabilidade P, caso contrário $\theta^t = \theta^{t-1}$.

Sob certas condições de regularidade facilmente satisfeitas na densidade proposta $p(\theta^* \mid \theta^{t-1})$, a sequência simulada $\theta^1, \theta^2, ...$ convergirá a uma variável aleatória que é distribuída de acordo com a distribuição a posteriori $\pi(\theta \mid y)$ (CHIB; GREENBERG, 1995).

2 Objetivos

2.1 Objetivo Geral

O objetivo deste trabalho é descrever a utilização do modelo de Gompertz baseado em distribuições defectivas sob as abordagens Bayesiana e Frequentista, considerando dados de sobrevivência com fração de cura. E, também mostrar que a distribuição defectiva pode ser a melhor alternativa em dados que apresentam fração de cura, pois não é necessário a inserção do parâmetro taxa de cura. Para ilustrar a metodologia, serão utilizados conjuntos de dados reais, como os dados do artigo de Brenna et al. (2004), que investigou a sobrevivência de mulheres pós tratamento para câncer de colo de útero e também dados coletados pela fundação FIOCRUZ de indivíduos infectados pelo vírus HIV.

2.2 Objetivos Específicos

- Análise Frequentista e Bayesiana para para Gompertz defectivo na presença censura a direita;
- Análise de dados médicos de sobrevivência, em particular, considerar os dados referentes a câncer cervical e pacientes portadores de HIV;
- Comparar os resultados clássicos e bayesianos;
- Utilizar o software R e OpenBugs para encontrar o sumário da análise clássica e da posteriori de interesse;

3 Distribuição de Gompertz Defectiva

Seja f(t) uma função de distribuição qualquer, $T > 0$ uma variável aleatória contínua desta função, a função de sobrevivência é definida como a probabilidade da observação não falhar até o tempo t, isto é,

$$S(t) = P(T \geq t) = S(t) = 1 - F(t), \tag{3.1}$$

pois,

$$F(t) = P(T \leq t), \tag{3.2}$$

sendo $F(x)$ a função de distribuição acumulada. A principal propriedade da função distribuição acumulada é dada por:

$$0 \leq F(x) \leq 1$$

ou seja, o limite superior é igual a 1. Porém quando temos dados com fração de cura, a função de distribuição acumulada é igual a $1 - \eta$. Assim, a presença de fração de cura impede o uso dos métodos tradicionais, o que motivou os modelos de mistura ou utilizar distribuições defectivas, onde a função de distribuição acumulada é diferente de 1.

3.1 A distribuição de Gompertz

A distribuição de Gompertz é muito utilizada em estudos nas áreas de atuária, demografia e em outros estudos da área de sobrevida (GIESER et al., 1998; BECKER, 2015; JAFARI; TAHMASEBI; ALIZADEH, 2014). Essa distribuição possui dois parâmetros, α e β, ambos positivos, sendo β o parâmetro de escala e α o parâmetro de forma.

A função de densidade de probabilidade para a variável aleatória tempo até o evento de interesse T com distribuição de Gompertz, a função de sobrevivência e a função de risco, com α>0 e β>0, são dadas respectivamente por:

$$f(t) = \alpha e^{(\beta t)} e^{\{\frac{\alpha}{\beta}[1-e^{(\beta t)}]\}} \tag{3.3}$$

$$S(t) = e^{\{\frac{\alpha}{\beta}[1-e^{(\beta t)}]\}}, \tag{3.4}$$

e

$$h(t) = \frac{f(t)}{S(t)} = \alpha e^{\beta t} \tag{3.5}$$

sendo

$$\lim_{t\to\infty} S(t) = \lim_{t\to\infty} \left\{ \frac{\alpha}{\beta}[1 - e^{(\beta t)}] \right\} = 0 \tag{3.6}$$

A distribuição apresentada desta forma é própria e portanto não é adequada aos dados com fração de cura, como mostrado na Figura 5. A Figura 5 apresenta o comportamento da função de densidade e probabilidade, da função de sobrevida e da função de risco da Gompertz Defectiva quando o α é fixado e o β varia.

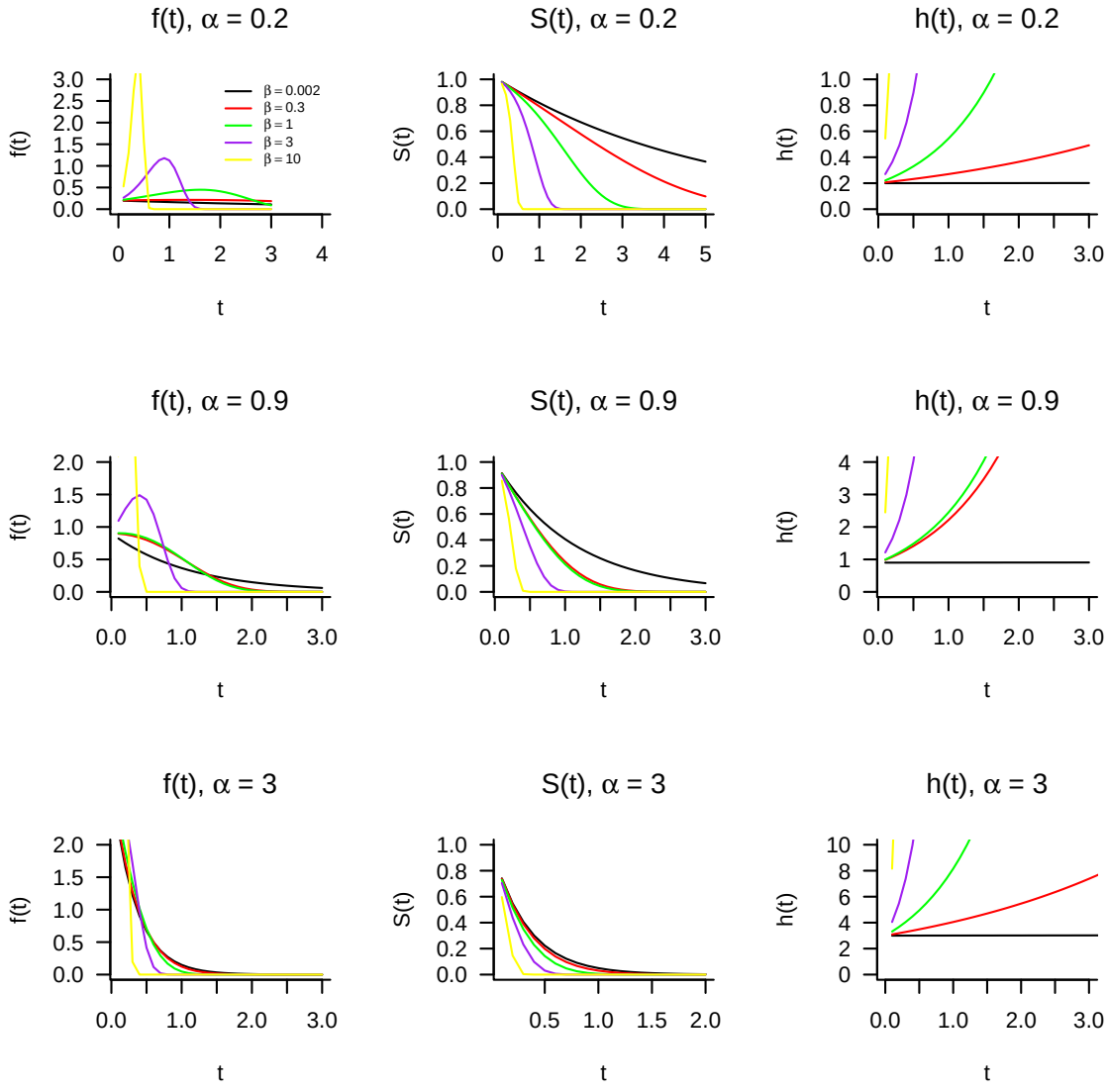

Figura 5 – Gráficos da função de densidade de probabilidade $f(t)$, da função de sobrevivência $S(t)$, e da função risco $h(t)$, para a distribuição de Gompertz, com α=0.2, α=0.9, α=3 e β=0.002, β=0.3, β=1, β=3, β=10.

3.2 O modelo modificado de Gompertz

A distribuição do modelo de Gompertz se torna imprópria ou defectiva quando são considerados valores menores do que zero ao parâmetro β. Essa distribuição de probabilidade definida para valores contínuos positivos foi reportada por diversos autores, como (ROCHA, TOMAZELLA e LOUZADA (2014) e Martinez e Achcar (2017). Sua função

de distribuição de probabilidade é dada por:

$$f(t) = \alpha e^{(-\beta t)} e^{\{\frac{-\alpha}{\beta}\left[1-e^{(-\beta t)}\right]\}} \tag{3.7}$$

para $\alpha > 0$, $\beta > 0$ e $t > 0$, sendo β o parâmetro de escala e α o parâmetro de forma. A função de sobrevivência e função de risco são dadas respectivamente por:

$$S(t) = e^{\{\frac{-\alpha}{\beta}[1-e^{(-\beta t)}]\}}, \tag{3.8}$$

e,

$$h(t) = \frac{f(t)}{S(t)} = \alpha e^{-\beta t}. \tag{3.9}$$

A proporção de imunes ao evento na população é calculada como o limite da função de sobrevivência quando o tempo t tende ao infinito:

$$\lim_{t\to\infty} S(t) = e^{\lim_{t\to\infty}\{\frac{-\alpha}{\beta}[1-e^{(-\beta t)}]\}} = e^{\frac{-\alpha}{\beta}} = p, \tag{3.10}$$

com p pertencente ao intervalo $(0, 1]$. Um comportamento especial para a função de sobrevivência da distribuição de Gompertz é observado quando α se aproxima de β. Isto é, quando β se aproxima do α, observa-se que a fração de cura se aproxima de $e^{\frac{-\alpha}{\beta}} = e^{\frac{-\alpha}{\alpha}} = e^{-1} = 0.36787944117$ (ver a Figura 6).

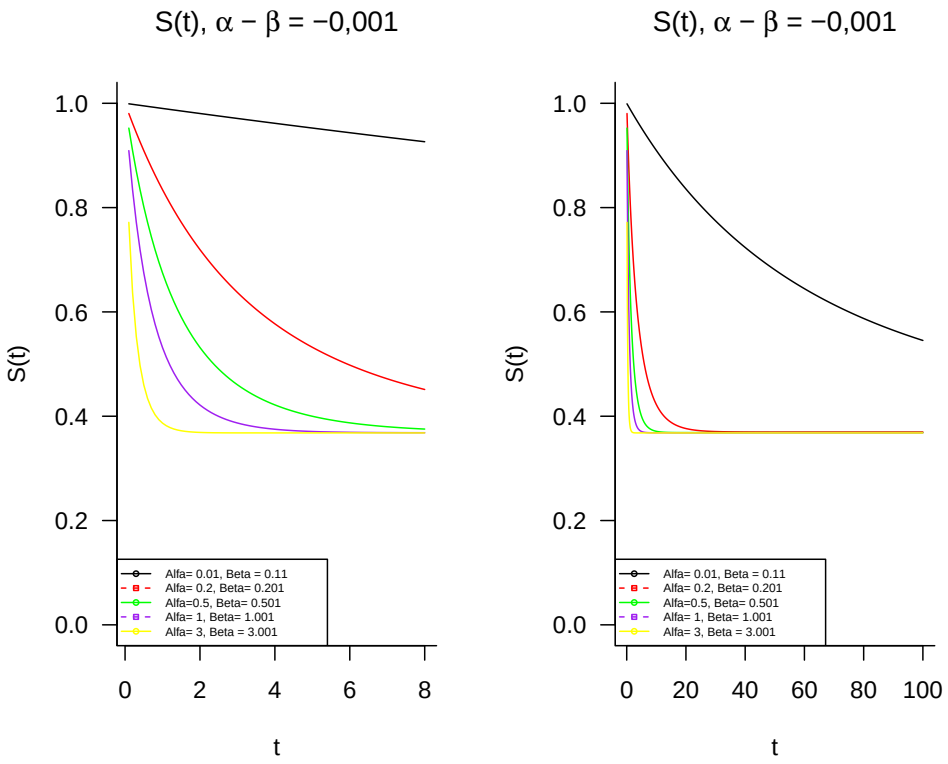

Figura 6 – Função de Sobrevivência com $\alpha - \beta = -0,001$.

Quando é considerado $\alpha > 0$ exatamente igual a $\beta > 0$, temos uma nova distribuição defectiva. Isto porque, quando α é igual β, tem-se uma função de apenas um parâmetro, dada por:

$$f(t) = \alpha e^{(\alpha t)} e \left[1 - e^{(-\alpha t)}\right]. \tag{3.11}$$

A desvantagem de usar esta distribuição é a fração de cura ser uma constante quanto t assume uma valor suficiente alto. A função de sobrevivência é dada por:

$$S(t) = e^{\{1-e^{(-\alpha t)}\}}. \tag{3.12}$$

A Figura 7 apresenta o comportamento das funções de probabilidade, sobrevida e risco quando $\alpha = \beta$ recebe valores positivos e negativos. Quando $\alpha = \beta$ é negativo, tem-se uma função de sobrevida crescente e uma função de risco negativa.

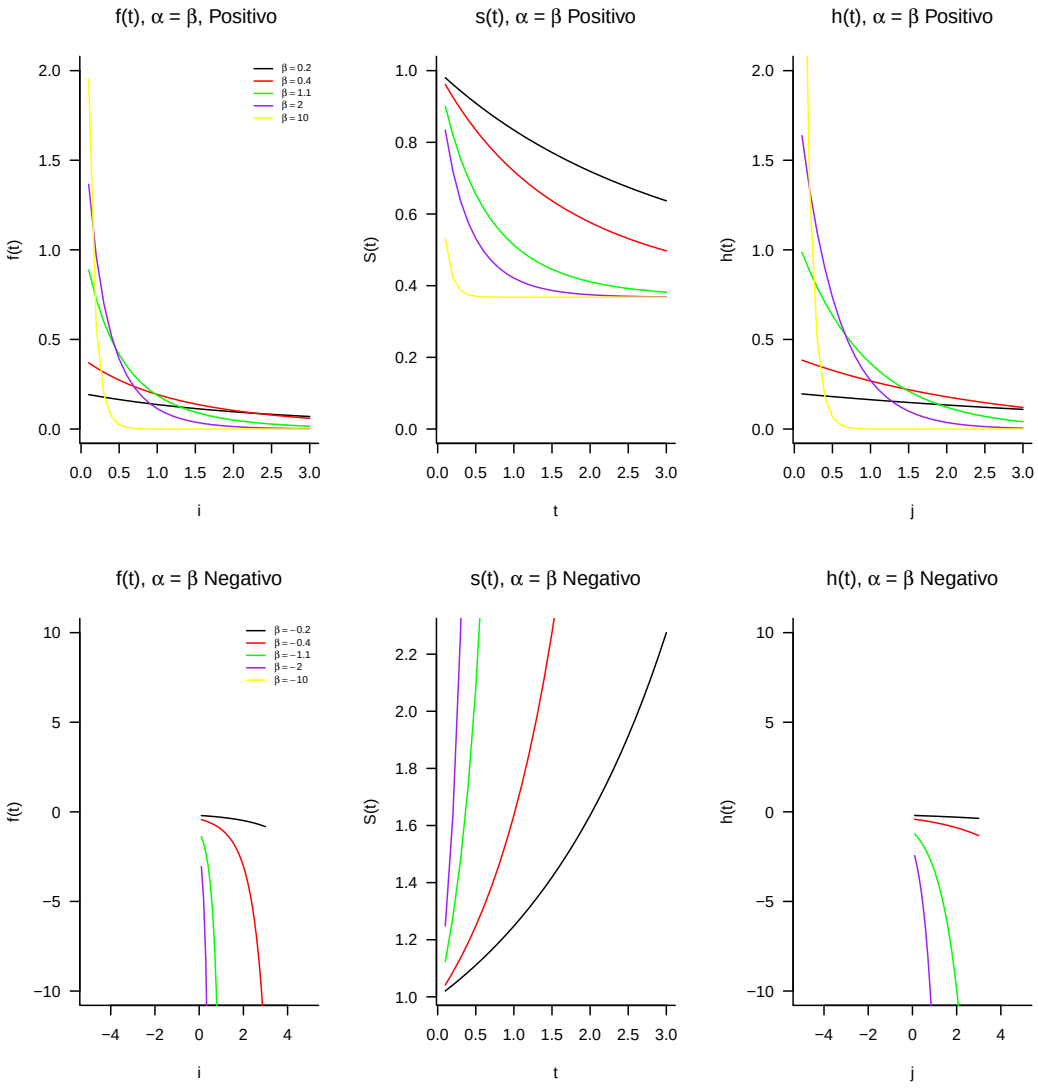

Figura 7 – Função de Sobrevivência com $\alpha = \beta$.

A Figura 8 apresenta alguns casos especiais da distribuição de Gompertz modificada, a qual motivou este trabalho e na qual descorrerão as aplicações. Uma característica desta distribuição é o comportamento da curva de sobrevida quando β é muito maior do que α, tornando a taxa de cura alta, como também mostra a Figura 8. Além disso, nota-se que a função de risco e de sobrevida são inversamente proporcionais, pois quanto maior a sobrevida menor o risco. Intuitivamente já é possível visualizar a aplicação desta distribuição na área médica, por exemplo, em doenças que o tratamento fornece cura para uma parcela dos indivíduos.

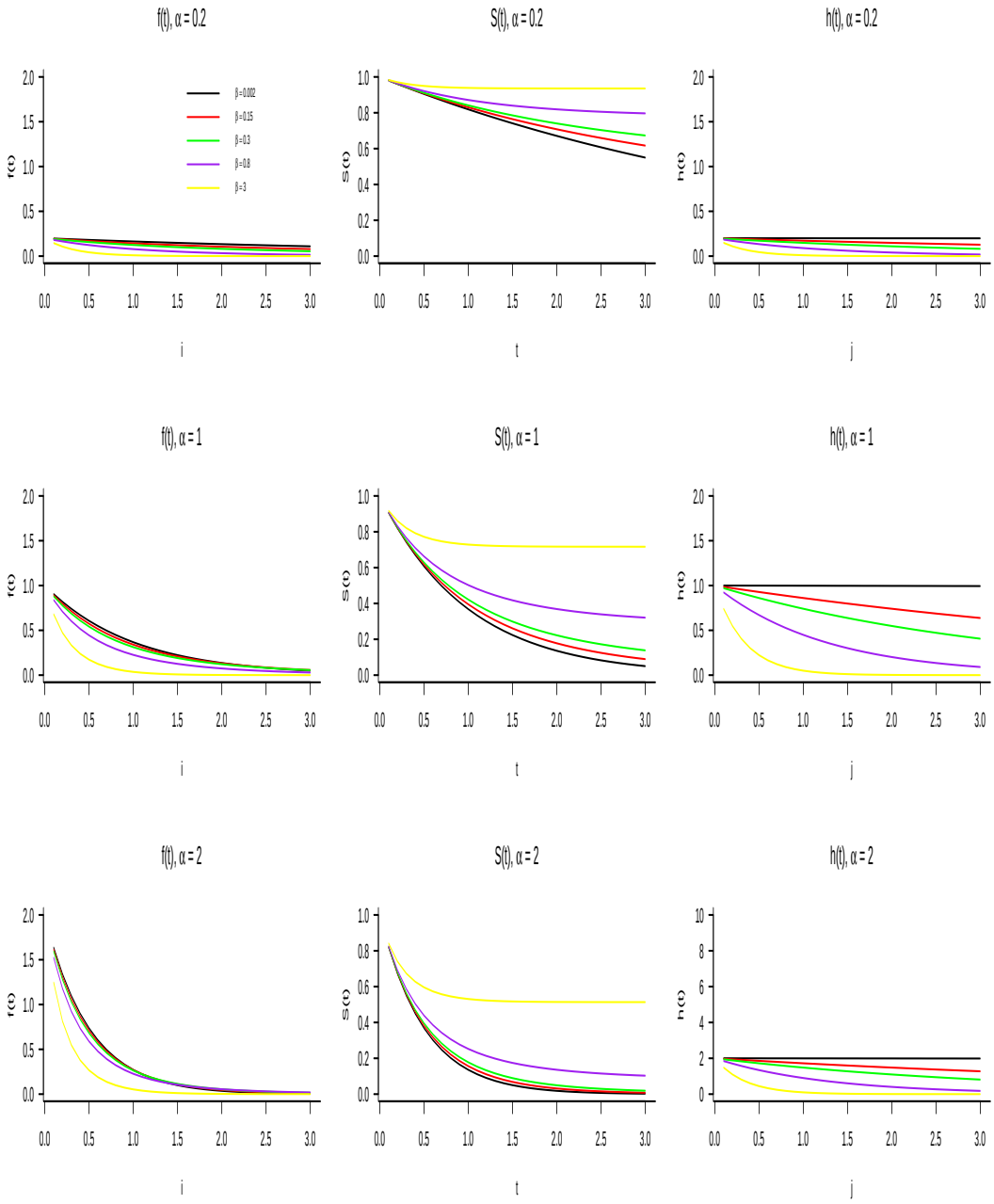

Figura 8 – Gráficos da função de probabilidade, de sobrevivência e risco da Gompertz Defectiva.

3.3 Função de verossimilhança para dados completos

Considerando $t_1, t_2, ..., t_n$ uma amostra aleatória independente e identicamente distribuída (i.i.d), com distribuição de Gompertz Defectiva de tamanho n, $i = 1, 2.., n$, a função de verossimilhança é dada por:

$$L(\alpha, \beta) = \alpha^n e^{-\beta \sum_{i=1}^n t_i} \exp\left\{ \frac{-\alpha}{\beta} \sum_{i=1}^n (1 - e^{-\beta t_i}) \right\}, \tag{3.13}$$

Assim, aplicando o logaritmo, em ambos os lados da equação 3.13, segue,

$$l(\alpha, \beta) = \log L(\alpha, \beta) = n \log \alpha - n\bar{t}\beta - \frac{\alpha}{\beta} \sum_{i=1}^n (1 - e^{-\beta t_i}), \tag{3.14}$$

onde $n\bar{t} = \sum_{i=1}^n t_i$. As primeiras derivadas do logaritmo da função de verossimilhança em relação a α e β são dadas respectivamente por,

$$\frac{\partial l}{\partial \alpha} = \frac{n}{\alpha} - \frac{1}{\beta} \sum_{i=1}^n (1 - e^{-\beta t_i}), \tag{3.15}$$

e

$$\frac{\partial l}{\partial \beta} = -n\bar{t} + \frac{\alpha}{\beta^2} \sum_{i=1}^n (1 - e^{-\beta t_i}) - \frac{\alpha}{\beta} \sum_{i=1}^n t_i e^{-\beta t_i} \tag{3.16}$$

Os estimadores de máxima verossimilhança para α e β são dados pelas soluções das equações: $\frac{\partial l}{\partial \alpha} = 0$ e $\frac{\partial l}{\partial \beta} = 0$. Assim, de $\frac{\partial l}{\partial \alpha} = 0$, temos:

$$\hat{\alpha} = \frac{n\hat{\beta}}{\sum_{i=1}^n (1 - e^{\hat{\beta} t_i})} \tag{3.17}$$

De $\frac{\partial l}{\partial \beta} = 0$, temos:

$$n\bar{t} = \frac{\hat{\alpha}}{\hat{\beta}^2} \sum_{i=1}^n (1 - e^{-} \hat{\beta} t_i) - \frac{\hat{\alpha}}{\hat{\beta}} \sum_{i=1}^n t_i e^{-\hat{\beta} t_i} \tag{3.18}$$

Substituindo 3.17 em 3.18, encontramos:

$$n\bar{t} = \frac{n\hat{\beta}}{\sum_{i=1}^n (1 - e^{-\hat{\beta} t_i})} \frac{1}{\hat{\beta}^2} \sum_{i=1}^n (1 - e^{-\hat{\beta} t_i}) - \frac{n\hat{\beta} \sum_{i=1}^n t_i e^{-\hat{\beta} t_i}}{\hat{\beta} \sum_{i=1}^n (1 - e^{-\hat{\beta} t_i})} \tag{3.19}$$

Assim,

$$n\bar{t} = \frac{n}{\hat{\beta}} - \frac{n \sum_{i=1}^n t_i e^{-\hat{\beta} t_i}}{\sum_{i=1}^n (1 - e^{-\hat{\beta} t_i})} \tag{3.20}$$

Note que, para encontrarmos o EMV para β não temos uma expressão analítica fechada, sendo necessário o uso de métodos iterativos como o método de Newton-Raphson. Observa-se ainda que considerando $f(\hat{\beta}) = 0$, temos que:

$$\sum_{i=1}^n (1 - e^{-\hat{\beta} t_i}) = \frac{\hat{\beta} \sum_{i=1}^n t_i e^{-\hat{\beta} t_i}}{(1 - \hat{\beta}\bar{t})} \tag{3.21}$$

A matriz de informação de Fisher observada, é obtida a partir das segundas derivadas do logaritmo da função de verossimilhança. As segundas derivadas de $l(\alpha, \beta)$ são dadas respectivamente por,

$$\begin{aligned}
\frac{\partial^2 l}{\partial \alpha^2} &= -\frac{n}{\alpha^2} \\
\frac{\partial^2 l}{\partial \alpha \partial \beta} &= \frac{1}{\beta^2}\sum_{i=1}^{n}(1-e^{-\beta t_i}) - \frac{1}{\beta}\sum_{i=1}^{n} t_i e^{-\beta t_i} \\
\frac{\partial^2 l}{\partial \beta^2} &= \alpha\left\{\frac{2}{\beta^3}\sum_{i=1}^{n}(1-e^{-\beta t_i}) + \frac{1}{\beta^2}\sum_{i=1}^{n} t_i e^{-\beta t_i}\right\} - \alpha\left\{\frac{1}{\beta^2}\sum_{i=1}^{n} e^{-\hat{\beta} t_i} + \frac{1}{\beta}\sum_{i=1}^{n}(-t_i^2 e^{-\beta t_i})\right\} \\
&- \frac{2\alpha}{\beta^3}\sum_{i=1}^{n}(1-e^{-\beta t_i} + \frac{\alpha}{\beta^2}\sum_{i=1}^{n} t_i e^{-\beta t_i} + \frac{\alpha}{\beta^2}\sum_{i=1}^{n} t_i e^{-\beta t_i} + \frac{\alpha}{\beta}\sum_{i=1}^{n} t_i^2 e^{-\beta t_i}
\end{aligned} \tag{3.22}$$

Localmente nos EMV encontramos,

$$\begin{aligned}
\frac{\partial^2 l}{\partial \alpha^2}_{\hat{\alpha},\hat{\beta}} &= -\frac{n}{\hat{\alpha}^2} \\
\frac{\partial^2 l}{\partial \alpha \partial \beta}_{\hat{\alpha},\hat{\beta}} &= \frac{\hat{\beta}\sum_{i=1}^{n} t_i e^{-\hat{\beta} t_i}}{\hat{\beta}^2(1-\hat{\beta}\bar{t}} - \frac{1}{\hat{\beta}}\sum_{i=1}^{n} t_i e^{-\hat{\beta} t_i} = \frac{\bar{t}\sum_{i=1}^{n} t_i e^{-\hat{\beta} t_i}}{1-\hat{\beta}\bar{t}} \\
\frac{\partial^2 l}{\partial \beta^2}_{\hat{\alpha},\hat{\beta}} &= -\frac{2\hat{\alpha}\bar{t}}{\hat{\beta}(1-\hat{\beta}\bar{t})}\sum_{i=1}^{n} t_i e^{-\hat{\beta} t_i} + \frac{\hat{\alpha}}{\hat{\beta}}\sum_{i=1}^{n} t_i^2 e^{-\hat{\beta} t_i}
\end{aligned} \tag{3.23}$$

Das segundas derivadas do logaritmo da função de verossimilhança encontramos a matriz de informação de Fisher observada, dada por:

$$I_0 = \begin{bmatrix} -\frac{\partial^2 l}{\partial \alpha^2}_{(\hat{\alpha},\hat{\beta})} & -\frac{\partial^2 l}{\partial \alpha \partial \beta}_{(\hat{\alpha},\hat{\beta})} \\ -\frac{\partial^2 l}{\partial \alpha \partial \beta}_{(\hat{\alpha},\hat{\beta})} & -\frac{\partial^2 l}{\partial \beta^2}_{(\hat{\alpha},\hat{\beta})} \end{bmatrix},$$

Por convenção, vamos usar as seguintes notações:

$$\begin{aligned}
A &= -\frac{\partial^2 l}{\partial \alpha^2}_{(\hat{\alpha},\hat{\beta})} = \frac{n}{\hat{\alpha}^2} \\
B &= -\frac{\partial^2 l}{\partial \alpha \partial \beta}_{(\hat{\alpha},\hat{\beta})} = \frac{\bar{t}\sum_{i=1}^{n} t_i e^{-\hat{\beta} t_i}}{(1-\hat{\beta}\bar{t})} \\
C &= -\frac{\partial^2 l}{\partial \beta^2}_{(\hat{\alpha},\hat{\beta})} = \frac{2\hat{\alpha}\bar{t}}{\hat{\beta}(1-\hat{\beta}\bar{t})}\sum_{i=1}^{n} t_i e^{-\hat{\beta} t_i} - \frac{\hat{\alpha}}{\hat{\beta}}\sum_{i=1}^{n} t_i^2 e^{-\hat{\beta} t_i}
\end{aligned} \tag{3.24}$$

Assim, a a matriz de informação de Fisher observada, é dada por:

$$I_0 = \begin{bmatrix} A & B \\ B & C \end{bmatrix}.$$

Assumindo tamanhos amostrais suficientemente grandes, os estimadores de máxima verossimilhança para α e β seguem aproximadamente ou assintoticamente uma distribuição normal bivariada dada por,

$$(\hat{\alpha}, \hat{\beta}) \approx N_2\left\{(\alpha, \beta); I_0^{-1}\right\} \tag{3.25}$$

em que N_2 denota uma distribuição normal bivariada e

$$I_0^{-1} = \frac{1}{(AC - B^2)} \begin{bmatrix} C & -B \\ -B & A \end{bmatrix}, \tag{3.26}$$

as variâncias e covariâncias assintóticas de $\hat{\alpha}$ e $\hat{\beta}$ são dadas por:

$$var(\hat{\alpha}) = \frac{C}{(AC - B^2)} \tag{3.27}$$

$$var(\hat{\beta}) = \frac{A}{(AC - B^2)} \tag{3.28}$$

$$var(\hat{\alpha}, \hat{\beta}) = -\frac{B}{(AC - B^2)}. \tag{3.29}$$

Ainda, considerando-se as distribuições assintóticas marginais para $\hat{\alpha}$ e $\hat{\beta}$, temos:

$$\widehat{\alpha} \sim N\left\{\alpha; \frac{C}{AC - B^2}\right\} \tag{3.30}$$

e,

$$\widehat{\beta} \sim N\left\{\beta; \frac{A}{AC - B^2}\right\}. \tag{3.31}$$

Assim, os intervalos de confiança 95% para α e β são dados respectivamente por,

$$\left(\widehat{\alpha} - 1,96\sqrt{\frac{C}{AC - B^2}}; \widehat{\alpha} + 1,96\sqrt{\frac{C}{AC - B^2}}\right) \tag{3.32}$$

e,

$$\left(\widehat{\beta} - 1,96\sqrt{\frac{A}{AC - B^2}}; \widehat{\beta} + 1,96\sqrt{\frac{A}{AC - B^2}}\right). \tag{3.33}$$

3.4 Função de Verossimilhança para dados censurados

Considerando uma amostra aleatória de tamanho n, $i = 1, 2.., n$, a função de verossimilhança com censuras do tipo I é dada por:

$$L(\lambda) = \prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i}, \tag{3.34}$$

em que δ_i é um indicador de censura. Quando $\delta_i = 1$, o dado observado é completo e quando $\delta = 0$ o tempo é censurado. Assumindo o modelo modificado de Gompertz, a função de verossimilhança para $\theta = (\alpha, \beta)$ é dada por:

$$\begin{aligned} L(\alpha, \beta) &= \prod_{i=1}^{n} \left[\alpha e^{(-\beta t_i)} e^{\left\{\frac{-\alpha}{\beta}\left[1 - e^{(-\beta t_i)}\right]\right\}}\right]^{\delta_i} \left[e^{\left\{\frac{-\alpha}{\beta}[1 - e^{(-\beta t_i)}]\right\}}\right]^{1-\delta_i} \\ &= \alpha^{\left[n \sum_{i=1}^{n} \delta_i\right]} e^{\left[-\beta \sum_{i=1}^{n} \delta_i \sum_{i=1}^{n} t_i\right]} e^{\left[\frac{-\alpha}{\beta} \sum_{i=1}^{n} (1 - e^{-\beta t_i}) \sum_{i=1}^{n} \delta_i\right]} e^{\left[\frac{-\alpha}{\beta} \sum_{i=1}^{n} [1 - e^{(-\beta t_i)}] \sum_{i=1}^{n} (1 - \delta_i)\right\}} \end{aligned} \tag{3.35}$$

Assim, aplicando o logaritmo em ambos os lados desta expressão, temos:

$$l(\alpha,\beta) = \log(\alpha)\sum_{i=1}^{n}\delta_i - \frac{\alpha}{\beta}\sum_{i=1}^{n}[1-exp(-\beta t_i)] - \beta\sum_{i=1}^{n}\delta_i t_i \tag{3.36}$$

sendo $l(\alpha,\beta) = \log(L(\alpha,\beta))$ e $n\bar{t} = \sum_{i=1}^{n} t_i$.

Para encontrarmos os estimadores de máxima verossimilhança (EMV), encontramos as derivadas parciais de $l(\alpha,\beta)$ em relação a α β e igualamos a zero. A derivada de $l(\alpha,\beta)$ em relação a β é dada implicitamente, precisando aplicar o método de interação de Newton-Raphson (ver por exemplo, LAWLESS,1982). Posteriormente, calculamos a matriz de Informação de Fisher para a obtenção da distribuição normal assintótica dos EMV e para a determinação das inferências de interesse (testes e hipóteses e intervalo de confiança). A variância assintótica é dado pelo elemento da matriz inversa de Fisher $I(\alpha,\beta)$. As segundas derivadas parciais da função de máxima verossimilhança são dadas por:

$$\frac{\partial^2}{\partial\alpha^2} = l(\alpha,\beta) = -\frac{1}{\alpha^2}\sum_{i=1}^{n}\delta_i, \tag{3.37}$$

$$\frac{\partial^2}{\partial\beta^2}l(\alpha,\beta) = \frac{\alpha}{\beta^2}\left[\sum_{n}^{i=1}\frac{c^{-\beta t_i}(2+\beta t_i-2)}{\beta} + \sum_{i=1}^{n} t_i e^{-\beta t_i}(1+\beta t_i)\right], \tag{3.38}$$

e,

$$\frac{\partial^2}{\partial\alpha\beta}l(\alpha,\beta) = -\sum_{i=1} n\frac{e^{-\beta t_i}(1+\beta t_i-1)}{\beta^2}. \tag{3.39}$$

3.5 Análise Bayesiana para a distribuição de Gompertz modificada

3.5.1 Uma distribuição a priori não informativa para α e β $(\alpha>0,\beta>0)$

A partir da regra de Jeffreys (ver por exemplo, BOX e TIAO, 1973), assumir a seguinte distribuição a priori para α e β,

$$\pi(\alpha,\beta) = \frac{1}{\alpha\beta}, \alpha>0, \beta>0 \tag{3.40}$$

3.5.2 Distribuição a posteriori conjunta para α e β

Combinando-se a distribuição a priori (3.40) com a função de verossimilhança para α e β (dados sem censuras) a distribuição a posteriori conjunta é dada por,

$$\begin{aligned}\pi(\alpha,\beta \mid \mathbf{t}) &= \frac{1}{\alpha\beta}\alpha^n e^{-\beta\sum_{i=1}^{n} t_i}\exp\left\{-\frac{\alpha}{\beta}\sum_{i=1}^{n}\left(1-e^{-\beta t_i}\right)\right\}\\ &= \frac{\alpha^{n-1}}{\beta}e^{-\beta n\bar{t}}\exp\left\{-\frac{\alpha}{\beta}\sum_{i=1}^{n}\left(1-e^{-\beta t_i}\right)\right\}\end{aligned} \tag{3.41}$$

3.5.3 Distribuições a posteriori condicionais para o algoritmo Gibbs Sampling

As distribuições a posteriori condicionais para α e β necessárias para o amostrador de Gibbs são dadas respectivamente por,

$$(i)\pi(a \mid \beta, \mathbf{t}) \sim \alpha^{n-1} \exp\left\{-\frac{\alpha}{\beta}\sum_{i=1}^{n}(1-e)^{-\beta t_i}\right\} \tag{3.42}$$

$$(ii)\pi(\beta \mid \alpha, \mathbf{t}) \sim \beta^{-1} \exp\left\{-\beta n\bar{t}\right\} \exp\left\{-\frac{\alpha}{\beta}\sum_{i=1}^{n}\left(1-e^{-\beta t_i}\right)\right\} \tag{3.43}$$

Notas:

(1) $X \sim$Gama(α, β);então

$$f\left(x \mid \alpha, \beta\right) = \frac{\beta^{\alpha}}{\tau(\alpha)} x^{\alpha-1} \exp\left\{-x\beta\right\}; x > 0 \tag{3.44}$$

A média e a variância para a distribuição gama são dadas respectivamente por,

$$E(x) = \frac{\alpha}{\beta}; var(x) = \frac{\alpha}{\beta^2} \tag{3.45}$$

Assim,

$$f(\alpha \mid \beta, \mathbf{t}) \sim Gama\left(n; \beta - 1\sum_{i=1}^{n}\left(1-e^{-\beta t_i}\right)\right) \tag{3.46}$$

$$E(\alpha \mid \beta, \mathbf{t}) = \frac{n\beta}{\sum\limits_{i=1}^{n}\left(1-e^{-\beta t_i}\right)} \tag{3.47}$$

$$var(\alpha \mid \beta, \mathbf{t}) = \frac{n\beta^2}{\left[\sum\limits_{i=1}^{n}\left(1-e^{-\beta t_i}\right)\right]^2} \tag{3.48}$$

(2) Na simulação de valores β da distribuição a posteriori conjunta para α e β usar o algoritmo de Metropolis-Hastings a partir de $\pi(\beta \mid \alpha, \mathbf{t})$, dado que esta expressão não tem uma forma conhecida.

3.5.4 Uso de distribuições a priori gama informativas para α e β

Outra possibilidade é assumir distribuições a priori gama para os parâmetros α e β dadas por,

$$\alpha \sim Gama(\alpha_1, b_1) \quad \text{e} \quad \beta \sim Gama(a_2, b_2)$$

onde os hiperparâmetros são, $a_1 = b_1 = a_2 = b_2 = 0,001$, isto é, também assumem-se distribuições a priori aproximadamente não-informativas (variâncias grandes). Além disso, usar os métodos MCMC para simular amostras da distribuição a posteriori para α e β.

4 Um estudo de simulação com a distribuição Gompertz modificada

Neste capítulo, iremos abordar um estudo de simulação para a distribuição Gompertz modificada afim de avaliar seus estimadores. Para o nosso estudo, um algoritmo usado para simular uma amostra de tamanho n da distribuição de Gompertz modificada com dados censurados é apresentado como:

1. Valores fixos de α e β.

2. Gerar n amostras aleatórias de.

$$M_i \sim Bernoulli(0, 1-\eta)$$

onde η corresponde a taxa de cura dada por:

$$\eta = exp\left(\frac{-\alpha}{\beta}\right). \tag{4.1}$$

3. Considere:

$$t_i' = \begin{cases} \infty \ if \ M_i = 0 \\ F_Y^{-1}(U_i) \ if \ M_i = 1, \end{cases}$$

onde,

$$F_Y^{-1}(U_i) = \frac{-1}{\beta} ln\left[1 + \frac{\beta}{\alpha} ln(1-U_i)\right] \tag{4.2}$$

e,

$$U_i \sim Uniforme(0, 1-\eta)$$

4. Gerar n amostras aleatórias de,

$$u_i' \sim Uniforme(0, max(t_i')),$$

Nota: Considere apenas o valor finito t_i'.

5. Calcule $t_i = min(t_i', u_i')$.

6. Pares de valores (t_1, δ_1),(t_2, δ_2),...,(t_n, δ_n) são obtidos assim, onde $\delta_i = 1$ se $t_i < u_i'$ e $\delta_i = 0$ se $t_i \geq u_i', i = 1, ..., n$.

Esse algoritmo foi introduzido por Rocha et al. (2017).

Um breve estudo do método de estimação de máxima verossimilhança foi baseado na simulação de amostras de tamanhos $n = 25, 50, 75, 100, 150$ e 200. Cada amostra foi

replicada 5000 vezes. A variância da taxa de cura η foi estimada usando o método Delta. Os resultados do estudo de simulação são apresentadas na Tabela 4, considerando um coeficiente de confiança de 95% e dois conjuntos de valores arbitrários para os parâmetros, dados por $(\alpha, \beta) = (0.4, 0.2)$ e $(\alpha, \beta) = (0.3, 0.7)$.

Tabela 4 – Resultados do estudo de simulação: intervalo de confiança tipo-Wald para os parâmetros α, β e η, viés e o erro quadrático médio, (EQM).

Valores Fixos de α e β	n	Parâmetros	Prob. de Cobertura	Viés	EQM
$\alpha = 0.4$ $\beta = 0.2$	25	α	0.9381	0.0201	0.0223
		β	0.9591	0.0123	0.0255
		η	0.7800	0.0045	0.0134
	50	α	0.9485	0.0090	0.0095
		β	0.9541	0.0022	0.0088
		η	0.8756	-0.0018	0.0064
	75	α	0.9493	0.0046	0.0059
		β	0.9545	-0.0007	0.0046
		η	0.9158	-0.0034	0.0039
	100	α	0.9478	0.0045	0.0043
		β	0.9476	0.0007	0.0031
		η	0.9248	-0.0022	0.0028
	150	α	0.9480	0.0013	0.0027
		β	0.9496	-0.0004	0.0017
		η	0.9344	-0.0016	0.0017
	200	α	0.9482	0.0022	0.0019
		β	0.9538	0.00004	0.0012
		η	0.9426	-0.0014	0.0012
$\alpha = 0.3$ $\beta = 0.7$	25	α	0.9158	0.0928	0.1343
		β	0.9638	0.1241	0.4927
		η	0.7311	-0.1127	0.0978
	50	α	0.9300	0.0245	0.0202
		β	0.9675	0.0886	0.2825
		η	0.9036	-0.0469	0.0338
	75	α	0.9408	0.0192	0.0125
		β	0.9582	0.0585	0.1344
		η	0.9316	-0.0243	0.0160
	100	α	0.9434	0.0095	0.0078
		β	0.9543	0.0284	0.0742
		η	0.9404	-0.0146	0.0087
	150	α	0.9428	0.0075	0.0049
		β	0.9570	0.0225	0.0393
		η	0.9508	-0.0048	0.0034
	200	α	0.9424	0.0069	0.0037
		β	0.9510	0.0185	0.0254
		η	0.9520	-0.0027	0.0021

Os resultados da Tabela 4 mostram que a probabilidade de cobertura do intervalo

de confiança tipo-Wald para os parâmetros α e β está bem próxima do nível de confiança de 95%. Além disso, a probabilidade de cobertura dos intervalos de confiança para η se aproxima de 95% à medida que o tamanho da amostra aumenta. Em todas as simulações, os vieses e os erros quadrados médios (EQM) sempre aproximam-se de zero à medida que o tamanho da amostra aumenta.

4.1 Simulação para $\alpha \cong \beta$

Para fazer o estudo de simulação quando $\alpha \cong \beta$ foram considerados dois cenários, com valores iniciais de $\alpha = 0.2$ e $\beta = 0.201$ e $\alpha = 0.02$ e $\beta = 0.0201$ e tamanhos amostrais de $n = 25, 50, 75, 100, 150$ e 200. Cada amostra foi replicada 5000 vezes. Os resultados do estudo de simulação são apresentados nas Figuras 9, 10, 11, 12 .

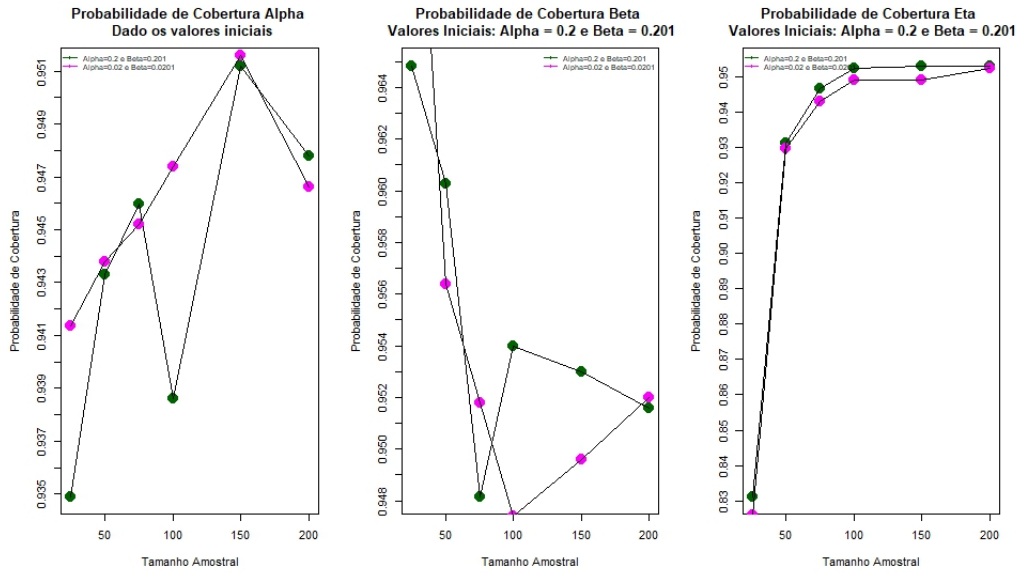

Figura 9 – Probabilidade de Cobertura $\alpha \cong \beta$.

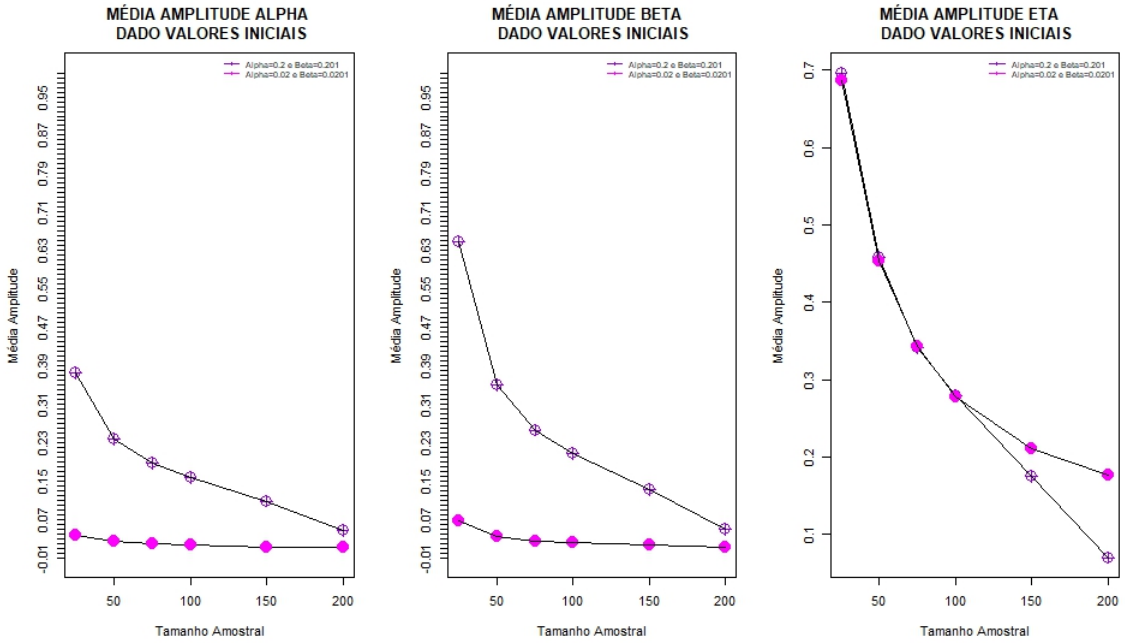

Figura 10 – Média da amplitude para $\alpha \cong \beta$.

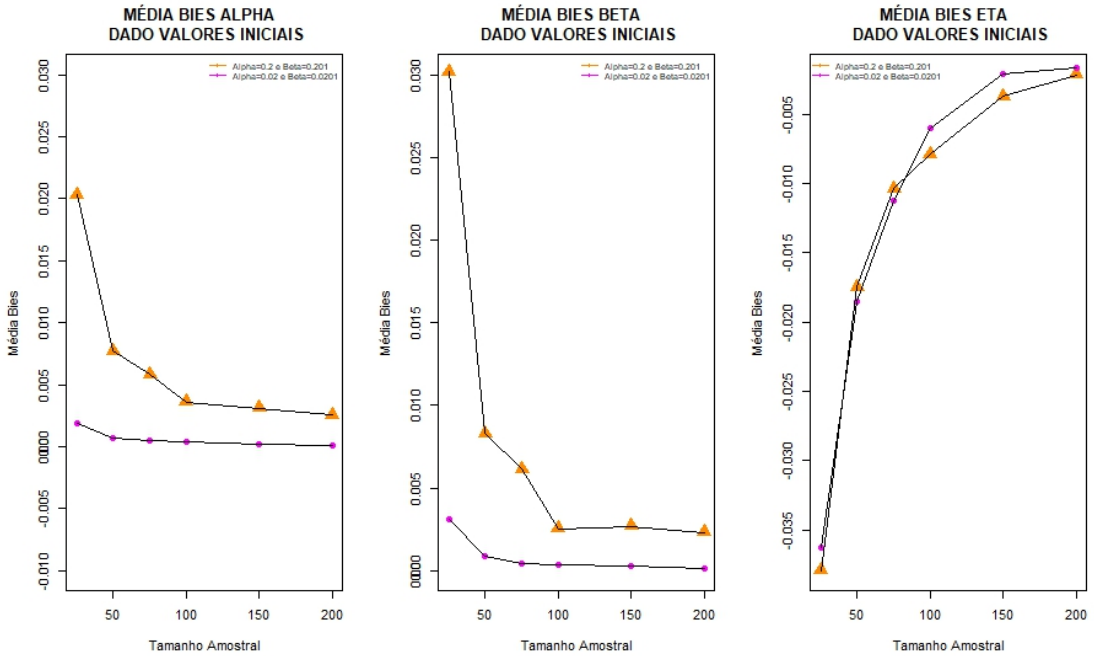

Figura 11 – Média do viés para $\alpha \cong \beta$.

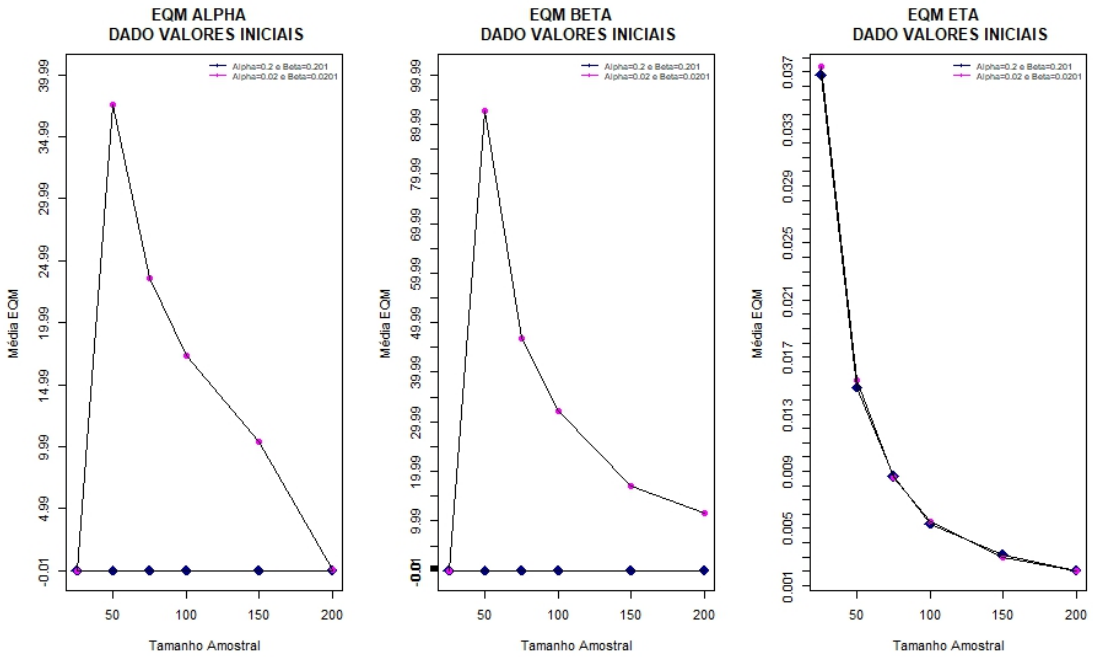

Figura 12 – EQM para $\alpha \cong \beta$.

Das Figuras 9, 11, 12, 10, é possível concluir que quanto maior o tamanho amostral mais a probabilidade de cobertura se aproxima de 95%, menores o viés e o EQM. Outra conclusão importante é extraída da Figura 10, que quanto maior a amostra, menor a média da amplitude do intervalo de credibilidade.

5 Aplicações com dados reais

5.1 Portadoras do Carcinoma Cervical

Os modelos de longa duração são muito utilizados para explicar biologicamente a presença de determinadas covariáveis e como o tempo de sobrevida responde de acordo com elas. Os dados utilizados para ilustrar este trabalho foram obtidos de um estudo desenvolvido por Brenna et al. (2004), como comentado na seção (1.1.2). Esta aplicação utilizará as covariáveis idade, status e estágio da doença.

A Figura 13 é um gráfico de contorno do logaritmo da função de verossimilhança para esses dados, considerando o modelo paramétrico baseado na distribuição modificada de Gompertz. O gráfico indica que o máximo da função log-verossimilhança está em $(\alpha, \beta) = (0,04178, 0,03995)$, como mostrado na Tabela 5.

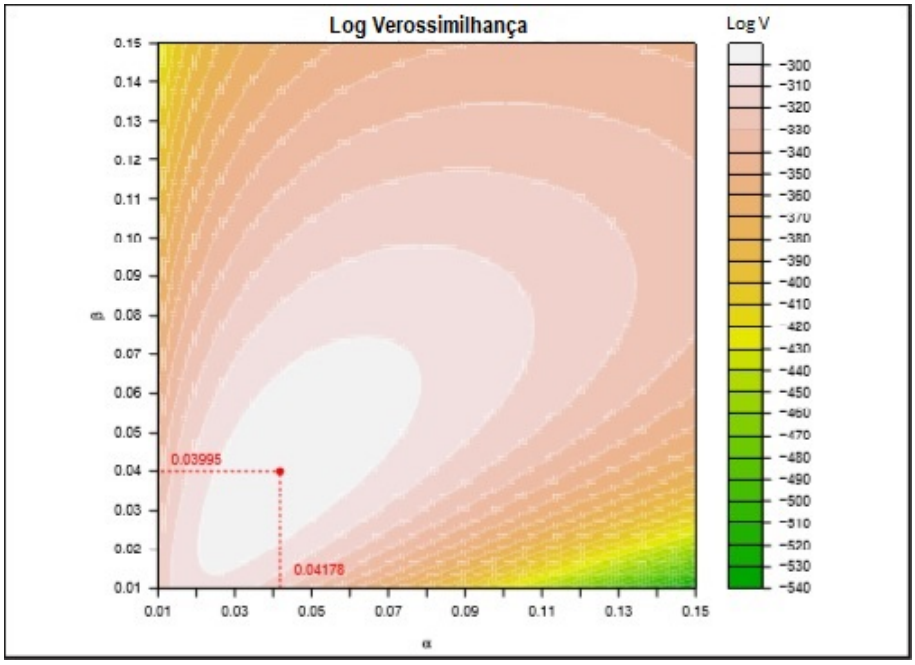

Figura 13 – Gráfico de contorno da função de verossimilhança, considerando os dados do estudo sobre carcinoma cervical. O máximo da função da verossimilhança é dado por $(\alpha, \beta) = (0.04178, 0.03995)$

A Tabela 5 apresenta os estimadores de máxima verossimilhança para α e β e a taxa de cura η, bem como as estimativas de seus erros padrão e intervalos de confiança. A Figura 14 mostra parcelas das estimativas de Kaplan-Meier para a função de sobrevivência e a função de sobrevida ajustada a partir do modelo paramétrico em relação aos tempos (meses).

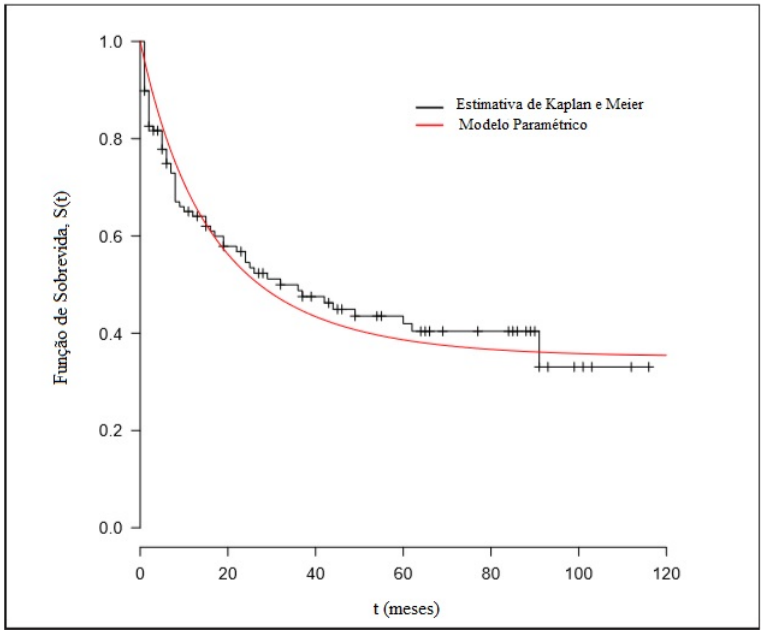

Figura 14 – Gráfico de Kaplan-Meier apresentando fração de cura para as mulheres portadoras do Carcinoma Cervical.

Da Figura 14, é possível notar que os valores preditos obtidos a partir do modelo com base na distribuição Gompertz modificada estão próximos do valores empíricos, sugerindo que este modelo está bem adequado para os dados. A taxa de cura é estimada por:

$$\hat{\eta} = \exp\left(-\frac{\hat{\alpha}}{\hat{\beta}}\right) = 0.35147, \tag{5.1}$$

e o respectivo erro padrão foram obtidos pelo método Delta com uma aproximação de primeira ordem de uma expansão da série de Taylor (Tabela 5).

A Tabela 6 mostra as estimativas bayesianas de α, β e a taxa de cura η, com seus respectivos intervalos de credibilidade com probabilidades iguais à 0,95. Nessa análise foram considerados hiperparâmetros $a1 = b1 = a2 = b2 = 0,001$ na elicitação de distribuições a priori gama invertidas aproximadamente não-informativas, isto é, não-informativos, ou seja, $\alpha \sim IG(0,001, 0,001)$ e $\beta \sim IG(0,001, 0,001)$. Podemos observar que as estimativas bayesianas para os parâmetros são relativamente próximas daquelas obtidas pelo método de máxima verossimilhança (Tabela 5).

Tabela 5 – Estimativa de Máxima Verossimilhança dos dados de carcinoma cervical

Parâmetro	Estimativa	Erro Padrão	Tipo Wald 95%
α	0.04178	0.00747	$(0.0271, 0.0564)$
β	0.03995	0.00795	$(0.0243, 0.0555)$
η	0.35147	0.05440	$(0.2446, 0.4584)$

Tabela 6 – Estimativa Bayesiana dos dados de carcinoma cervical

Parâmetro	Estimativa	Int.Cred. de 95%
α	0.04134	$(0.0282, 0.0574)$
β	0.03928	$(0.0244, 0.0557)$
η	0.3474	$(0.2380, 0.4554)$

5.1.1 Modelo na presença de covariáveis

Para ilustrar o a aplicação do modelo de regressão baseado na distribuição de Gompertz modificado, na presença de covariáveis, assumir as seguintes covariáveis:

- Tratamento dos pacientes que começaram a ser monitorados (x_1), com classificação menor do que 50 anos ($x_1 = 0$) versus maior ou igual a 50 anos ($x_1 = 1$).
- Estágio clínico da doença que quando começou o tratamento, classificado em estágio I, II, ou III. Esta variável representa o modelo de regressão usando duas variáveis dummy (x_2 e x_3), onde $x_2 = 0$ e $x_3 = 0$ se estágio I, $x_2 = 1$ e $x_3 = 0$, se estágio II, e $x_2 = 0$ e $x_3 = 1$ se estágio III.

Assim, a regressão do modelo para os dados considerados é:

$$ln\alpha[\mathbf{x}] = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_{12} x_1 x_2 + \alpha_{13} x_1 x_3, \tag{5.2}$$

e,

$$ln\beta[\mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3, \tag{5.3}$$

onde $\alpha_{12}, \alpha_{13}, \beta_{12}$ e β_{13} são parâmetros associados às interações entre a idade e os estágios da doença. Dessa forma assumimos os seguintes modelos de regressão:

- **Modelo 1:** Modelo 1 considera apenas o efeito da idade (x_1). Neste caso, são considerados $\alpha_2, \alpha_3, \alpha_{12}, \alpha_{13}, \beta_2, \beta_3, \beta_{12}$ e β_{13} iguais a zero.
- **Modelo 2:** Modelo 2 considera apenas o efeito do estágio da doença (variáveis dummys x_2 e x_3). Neste caso, α_1, β_1 e os termos de interação $\alpha_{12}, \alpha_{13}, \beta_{12} e \beta_{13}$, são considerados iguais a zero.
- **Modelo 3:** o modelo 3 considera os efeitos da idade (x_1) e o estágio da doença (x_2) e (x_3), mas não inclui os termos de interação $\alpha_{12}, \alpha_{13}, \beta_{12}$ e β_{13}.
- **Modelo 4:** o Modelo 4 considera os efeitos da idade (x_1) e do estágio da doença (x_2) e (x_3) e inclui a interação dos termos $\alpha_{12}, \alpha_{13}, \beta_{12}$ e β_{13}.

Para todos os coeficientes de regressão, assumimos uma distribuição priori normal com média 0 e variância grande. Os estimadores Bayesianos e os estimadores de

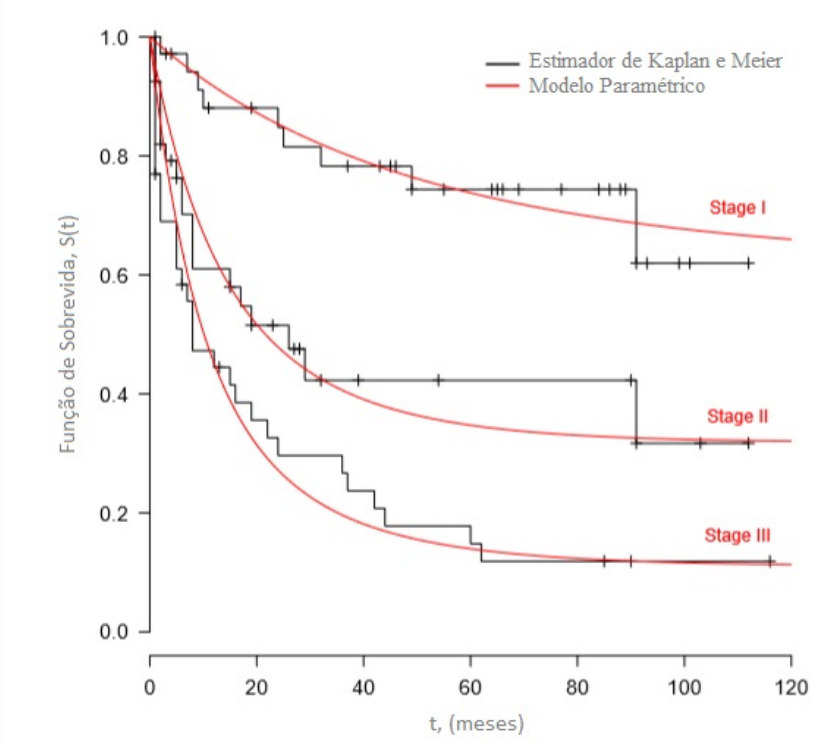

Figura 15 – Estimadores da função de sobrevida de Kaplan - Meier e ajustados pelo modelo paramétrico de acordo com estágio da doença. A curva paramétrica é baseada no modelo 2, com parâmetros estimados pela abordagem da máxima verossimilhança. Observações censuradas são marcadas por traços verticais na curva de Kaplan - Meier.

máxima verossimilhança são apresentados na Tabela 7. O ajuste dos Modelos 1 e 2 também considerou a estimativa da fração de cura (η) para cada nível da respectiva variável independente. Os estimadores e erros-padrão para a fração de curas foram calculados usando o método Delta. Como ilustração, a Figura 15 mostra as curvas de sobrevivência estimadas pelo método Kaplan-Meier e usando o Modelo 2 sob a abordagem de máxima verossimilhança.

A Tabela 7 mostra que a os estimadores de máxima verossimilhança e as estimativas bayesianas são bastante próximas umas das outras. O modelo 1 sugere que a idade não tem um efeito significativo na sobrevivência livre de doença, uma vez que o intervalo de credibilidade para os parâmetros, α_1 e β_1, incluem o valor 0. O modelo 2 sugere um efeito significativo do estágio da doença na sobrevida livre de doença, dado que o intervalo de confiança para os parâmetros de forma α_2 e α_3 não incluem o valor zero.

Tabela 7 – Estimadores de Máxima Verossimilhança e Bayesianos, modelo com covariáveis

Parâmetro	Estimativa	Erro Padrão	Tipo Wald	AIC	Estimativa	Int. Cred. 95%	DIC
Modelo I	**EMV**		**95% IC**	592,9	Bayesianos		593,3
α_0 (intercepto)	-3,041	0,2493	(-3,532 ; -2,551)		-3,091 1	(-3,629 ; 2,585)	
α_1 (idade $\geq$ 50 vs. <50)	-0,258	0,3578	(-0,961 ; 0,445)		-0,285	(-1,032 ; 0,455)	
β_0 (intercepto)	-3,020	0,2637	(-3,539 ; -2,501)		-3,084	(-3711 ; -2,580)	
β_1 (idade $\geq$ 50 vs. <50)	-0,386	0,4036	(-1,178 ; 0,407)		-0,461	(-1,524 ; 0,424)	
η_1 (fração de cura, idade $\geq$ 50)	0,376	0,0756	(0,227 ; 0,524)		0,371	(0,214 ; 0,523)	
η_2 (fração de cura, idade <50)	0,329	0,0794	(0,173 ; 0,485)		0,311	(0,121 ; 0,471)	
Modelo 2				565,2			563,4
α_0 (intercepto)	-4,825	0,5266	(-5,859 ; -3,790)		-4,850	(-5,746 ; -3,972)	
α_2 (estágio II vs. I)	1,814	0,6071	(0,621 ; 3,007)		1,734	(0,636 ; 2,820)	
α_3 (estágio II vs. III)	2,325	0,5789	(1,187 ; 3,462)		2,281	(1,258 ; 3,285)	
β_0 (intercepto)	-4,093	0,8831	(-5,828 ; -2,358)		-4,269	(-6,409 ; -3,070)	
β_2 (estágio II vs. I)	0,943	0,9590	(-0,936 ; 2,832)		0,902	(-0,939 ; 3,153)	
β_3 (estágio III vs. I)	0,803	0,9348	(-1,034 ; 2,639)		0,839	(-0,665 ; 3,085)	
η_1 (fração de cura, estágio I)	0,618	0,1726	(0,278 ; 0,957)		0,560	(0,043 ; 0,814)	
η_2 (fração de cura, estágio II)	0,319	0,1026	(0,117 ; 0,520)		0,291	(0,039 ; 0,506)	
η_3 (fração de cura, estágio II)	0,110	0,0554	(0,001 ; 0,219)		0,106	(0,014 ; 0,231)	
Modelo 3				563,7			560,8
α_0 (intercepto)	-5,490	0,369	(-6,226 ; -4,753)		-5,486	(-6,244 ; -4,828)	
α_1 (idade $\geq$ 50 vs. <50)	0,297	0,340	(-0,396 ; 0,991)		0,260	(-0,433 ; 0,898)	
α_2 (estágio II vs. I)	2,529	0,452	(-1,638 ; 3,421)		2,387	(1,571 ; 3,255)	
α_3 (estágio III vs. I)	2,602	0,415	(1,778 ; 3,424)		2,527	(1,765 ; 3,366)	
β_0 (intercepto)	-12,674	2,443	(-15,654 ; -9,795)		-13,020	(-17,140 ; -9,101)	
β_1 (idade $\geq$ 50 vs. <50)	2,012	0,994	(-0,168 ; 4,192)		2,853	(0,575 ; 6,468)	
β_2 (estágio II vs. I)	8,713	2,680	(5,166 ; 12,359)		8,071	(4,553 ; 11,790	
β_3 (estágio III vs. I)	7,322	2,771	(3,414 ; 11,331)		6,602	(2,946 ; 10,420)	
Modelo 4				562,7			561,9
α_0 (intercepto)	-4,408	0,7133	(-5,806 ; -3,010)		-3,445	(-4,928 ; -2,464)	
α_1 (idade $\geq$ 50 vs. <50)	-0,586	0,8465	(-2,244 ; 1,073)		-1,558	(-2,834 ; 0,037)	
α_2 (estágio II vs. I)	1,472	0,7731	(-0,043 ; 2,987)		0,559	(-0,5678 ; 2,039)	
α_3 (estágio III vs. I)	2,152	0,8377	(0,509 ; 3,793)		1,142	(-0,1288 ; 2,633)	
α_{12} (idade vs. estágio II)	0,477	1,1565	(1,789 ; 2,744)		1,041	(-1,089 ; 2,889)	
α_{13} (idade vs. estágio III)	0,237	1,0077	(-1,737 ; 2,212)		1,123	(-0,5672 ; 2,725)	
β_0 (intercepto)	-3,028	0,6244	(-4,252 ; -1,804)		-3,015	(-9,219 ; -1,836)	
β_1 (idade $\geq$ 50 vs. <50)	-8,734	1,8052	(-12,272 ; -5,196)		-8,243	(-12,19 ; -4,538)	
β_2 (estágio II vs. I)	-1,186	1,1913	(-3,521 ; 1,148)		-1,074	(-3,464 ; 4,538)	
β_3 (estágio III vs. I)	0,008	0,8434	(-1,644 ; 1,661)		0,075	(-2,014 ; 5,786)	
β_{12} (idade vs. estágio II)	10,867	2,1123	(6,727 ; 15,008)		9,907	(5,959 ; 13,85)	
β_{13} (idade vs. estágio III)	8,359	1,9524	(4,532 ; 12,186)		7,568	(3,585 ; 11,62)	

Embora o Modelo 1 não evidencie um efeito significativo da idade na sobrevida livre de doença, o Modelo 4 sugere que a interação entre a idade e os estágios clínicos é importante para entender o papel da idade no tempo até a progressão da doença. Os intervalos de credibilidade para os termos de interação β_{12} e β_{13} não incluem o valor zero, sugerindo um efeito significativo da interação. De fato, as curvas de sobrevivência mostradas na Figura 16 ajuda a entender esse efeito. As curvas na Figura 16 foram obtidas com base no Modelo 4, e mostraram que a idade tem um forte efeito sobre a sobrevivência livre da doença no estágio II, mas a idade não é um importante fator que influencie a sobrevida de pacientes no estágio clínico III, como mostra o terceiro gráfico da Figura 16.

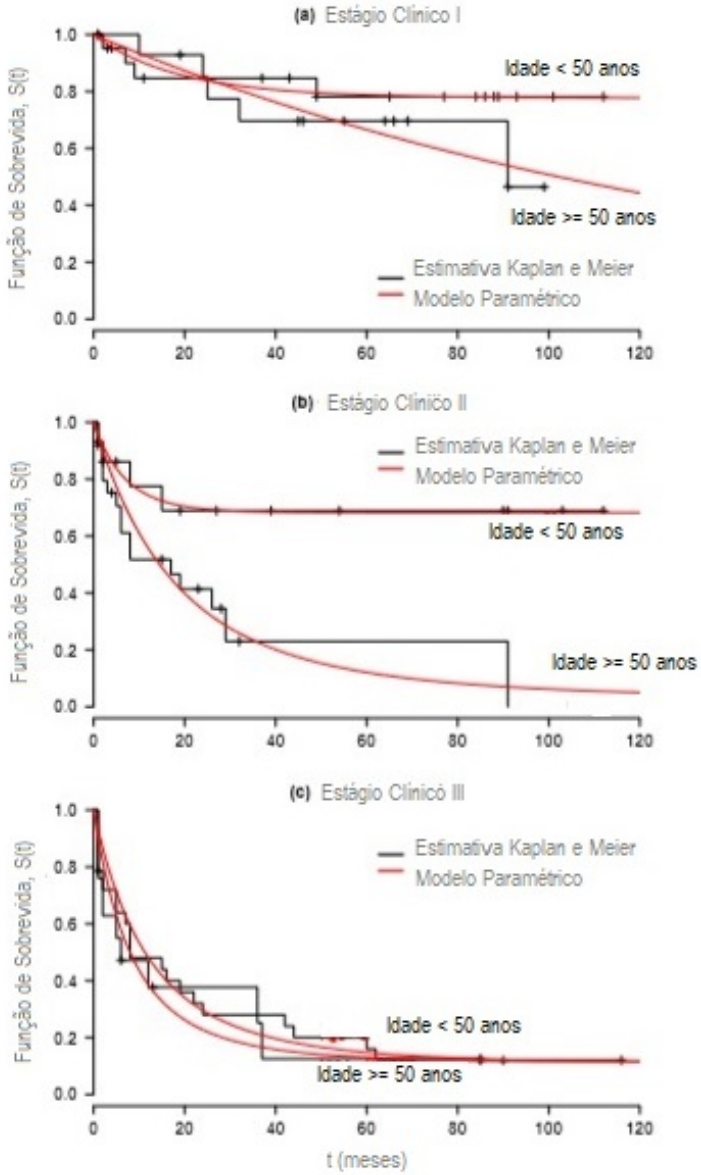

Figura 16 – Kaplan-Meier e curva de sobrevida paramétrica livre da doença de acordo com o estágio da doença no qual começou o tratamento e idade. A curva paramétrica está baseada no Modelo 4, como parâmetros estimados a partir da abordagem de máxima verossimilhança. Observações censuradas são os traços verticais no estimador Kaplan-Meier.

Os modelos 2, 3 e 4 possuem valores similares de AIC e DIC (Tabela 7), e estes valores são inferiores aos respectivos valores calculados para o Modelo 1. No entanto, sob um ponto de vista clínico, o Modelo 4 parece ser o mais apropriado para os dados, dado que a Figura 16 mostra a importância de considerar os termos de interação para a interpretação do padrão de associação entre idade, estágios clínicos e o tempo até a progressão da doença.

5.2 Indivíduos infectados pelo vírus HIV

O segundo banco de dados é do desenho de estudo coorte que durou entre os anos de 1982 e 2000, feito pela fundação FIOCRUZ. Este estudo inclui 193 indivíduos, sendo 144 homens e 49 mulheres, com 104 censuras no total, como também exemplificado na seção 1.1.1. O objetivo é verificar se o modelo da Gompertz Defectivo se ajusta nestes dados, ou seja, se este é um bom modelo para representar a variável resposta sobrevida e com efeito das covariáveis: idade e sexo.

5.2.1 Abordagem frequentista

Em uma primeira análise dos dados dos pacientes portadores de HIV, consideramos o uso do método da máxima verossimilhança. Neste caso, obtemos que a estimativa encontrada usando a teoria clássica sem covariáveis para os parâmetros α e β são respectivamente, $\alpha = 0.0775$, $\beta = 0.0668$. Sendo assim, temos que $\eta = 0.3135351$, pelo cálculo de $\eta = exp(-\alpha/\beta)$, como mostra a equação (5.1). A Figura 17 mostra o gráfico de contorno do log da função de máxima verossimilhança. Este gráfico ilustra o "ponto ótimo" para escolha das estimativas.

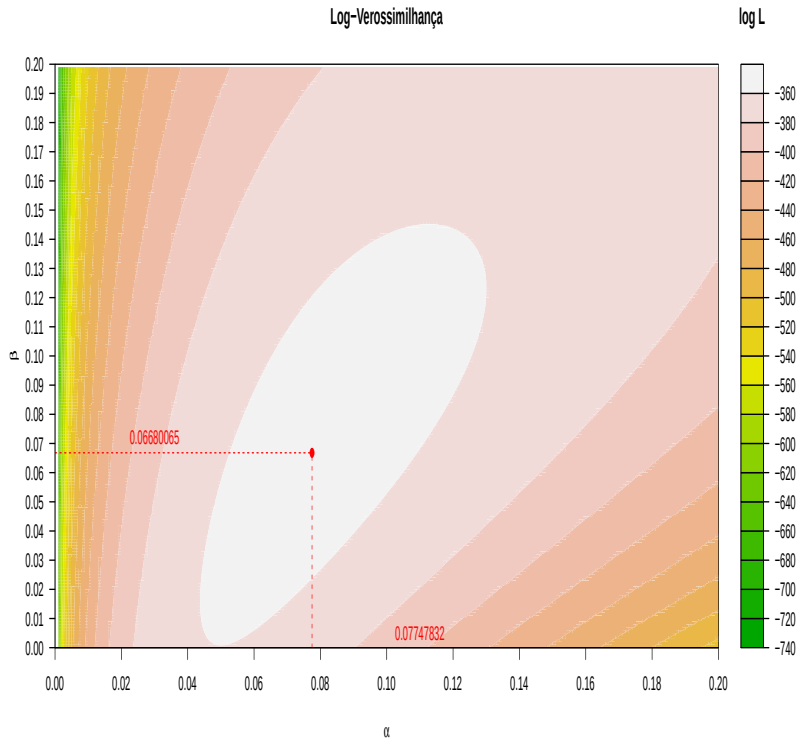

Figura 17 – Gráfico de contorno da função de verossimilhança, considerando os dados do estudo sobre indivíduos infectados pelo vírus HIV. Os estimadores de máxima verossimilhança são dados por $(\alpha, \beta) = (0.0775, 0.0668)$

A Tabela 8 apresenta as estimativas pelo método da máxima verossimilhança dos parâmetros α, β e η assim como os itervalos de confiança obtidos na análise dos dados dos individuos portadores de HIV. Dos resultados da Tabela 8, pode-se notar que a fração de cura estimada pelo modelo Gompertz Defectivo é de aproximadamente 31,35% variando no intervalo de 15,59% a 47,11%. Já a estimativa do estimador de Kaplan - Meier para a fração de cura é de aproximadamente 38%. Uma vez que as estivamtivas são próximos, concluímos que o modelo Gompertz Defectivo é apropriado para a descrição da sobrevida dos pacientes portadores do HIV.

Tabela 8 – Estimadores de máxima verossimilhança para os dados de pacientes portadores de HIV.

Parâmetro	Estimativa	Desvio Padrão	Tipo Wald 95%
α	0.07747829	0.01221153	$(0.05354413; 0.1014124)$
β	0.06680059	0.02081081	$(0.02601214; 0.1075890)$
η	0.31353497	0.08040516	$(0.15594375; 0.4711262)$

A Figura 18 illustra a curva de de Kaplan-Meier para a função de sobrevivência e a função de sobrevida ajustada a partir do modelo paramétrico em relação aos tempo de sobrevida dos pacientes portadores de HIV mostrando a boa qualidade de ajuste do modelo Gompertz Defectivo.

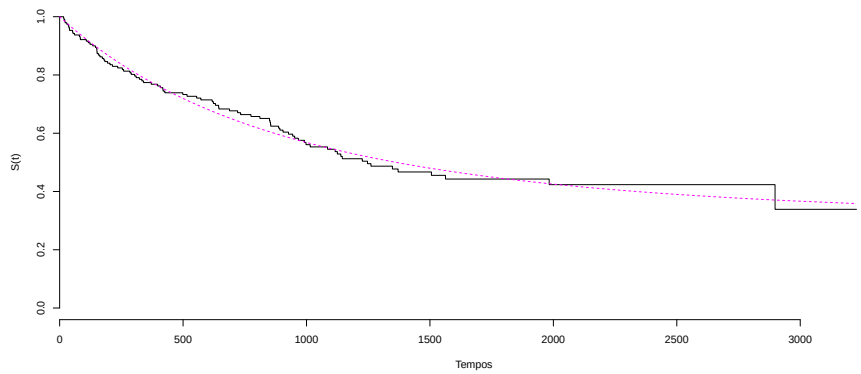

Figura 18 – Estimadores da função de sobrevida de Kaplan - Meier e ajustados pelo modelo paramétrico Gompertz Defectivo assumindo o método da máxima verossimilhança.

5.2.2 Abordagem Bayesiana

Para a análise Bayesiana sem covariáveis foi utilizada uma priori Gamma não informativa. A Tabela 9 apresenta as estimativas Bayesianas dos parâmetros na qual podemos ver uma grande proximidade das estimativas obtidas com as estimativas obtidas pelo método da máxima verossimilhança considerando a priori escolhida.

Tabela 9 – Estimadores Bayesianos para os dados de pacientes portadores de HIV.

Parâmetro	Estimativa	Erro Padrão	IC de 95%
α	0.0745	0.0125	$(0.05198; 0.1008)$
β	0.0599	0.0224	$(0.01423; 0.1040)$
η	0.2730	0.1005	$(0.01989; 0.4336)$

5.2.3 Modelo na presença de covariáveis para os portadores de HIV

O modelo de Gompertz na presença de covariáveis é um modelo de tempo de vida acelerado. Isto significa que a função das covariáveis tem efeito de acelerar ou desacelerar o tempo de vida. Para a aplicação do modelo de regressão baseado na distribuição Gompertz Defectiva consideramos as seguintes covariáveis:

- Idade dos pacientes participantes do estudo (x_1), com classificação ≤ 30 ($x_1 = 0$) versus > 30 anos ($x_1 = 1$).
- Sexo dos pacientes (x_2), com classificação feminino ($x_2 = 0$) versus masculino ($x_2 = 1$).

Assim, considerou-se o seguinte modelo de regressão:

$$ln\alpha(\mathbf{x}) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_1 x_2 \tag{5.4}$$

e,

$$ln\beta(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \tag{5.5}$$

Alguns casos especiais do modelo de regressão são considerados por:

- **Modelo 1:** Modelo 1 considera apenas o efeito da idade (x_1). Neste caso $\alpha_2 = 0$ e $\beta_2 = 0$.
- **Modelo 2:** Modelo 2 considera apenas o efeito do sexo (x_2). Neste caso $\alpha_1 = 0$ e $\beta_1 = 0$.
- **Modelo 3:** Modelo 3 considera o efeito da idade e do sexo.

A tabela 10 traz as informações da estimativa de máxima verossimilhança referente aos modelos dos dados dos portadores de HIV.

Tabela 10 – Estimadores de Máxima conforme banco de dados dos indivíduos portadores de HIV, modelo com covariáveis

Parâmetro	Estimativa	Erro Padrão	Tipo Wald	AIC
Modelo I	**EMV**		**95% IC**	
α_0 (intercepto)	-7.0046	0.2983	(-7.592041, -6.414013)	
α_1 (idade $>$ 30 vs. $\leq$ 30)	-0.2211	0.3525	(-0.9104959, 0.4683569)	
β_0 (intercepto)	-6.8969	0.4327	(-7.740577, -6.053252)	
β_1 (idade $>$ 30 vs. $\leq$ 30)	-0.6237	0.6253	(-1.815263, 0.5679015)	
η_0 (fração de cura, idade $>$ 30)	0.3622	0.4304	(-0.1798186 , 0.904182)	
η_1 (fração de cura, idade $\leq$ 30)	0.7015	0.2765	(0.1595523, 1.243553)	
				1544
Modelo 2				
α_0 (intercepto)	-2.9241	0.3743	(-3.6578, -2.1904)	
α_2 (gênero F vs M)	0.4740	181.458	(-0.3356, 1.2836)	
β_0 (intercepto)	-2.5628	0.6663	(-3.868759,-1.256848)	
β_2 (gênero F vs M)	-0.1794	0.3743	(-1.6597, 1.3008)	
η_0 (fração de cura, gênero F)	0.3195	0.6663	(-355.3335, 355.9725)	
η_2 (fração de cura, gênero M)	14.03652	181.4589	(-341.6165, 369.6895)	
				711.6838
Modelo 3				
α_0 (intercepto)	-2.8230	0.4182	(-3.6427, -2.0033)	
α_1 (idade $>$ 30 vs. $\leq$ 30)	-0.2509	0.2743	(-0.9562, 0.4542)	
α_2 (sexo F vs. M)	0.5694	0.7342	(-0.2372, 0.8379)	
β_0 (intercepto)	-2.3470	0.7342	(-3.7861, -0.9079)	
β_1 (idade $>$ 30 vs. $\leq$ 30)	-0.6375	0.4182	(-1.9080, 0.6330)	
β_2 (Sexo F vs. M)	0.0724	0.2743	(0.1369, 1.2121)	
η_0	0.3003	0.7342	(-0.2372, 0.8379)	
η_1	0.6745	0.2743	(0.1369, 1.2121)	
η_2	0.000385	0.2743	(-0.5372, 0.5380)	
				714.7879

Apesar da Figura 19 aparentemente mostrar que o sexo feminino possui uma sobrevida maior do que a do sexo masculino, na análise de máxima verossimilhança, o intervalo de confiança dos parâmetros do Modelo 2 contém o valor 0. Isto prevalece nos 3 modelos. Portanto de acordo com estes dados, não temos evidências suficientes para afirmar que a idade ou o sexo influenciam no tempo de sobrevida de pacientes infectados pelo vírus HIV.

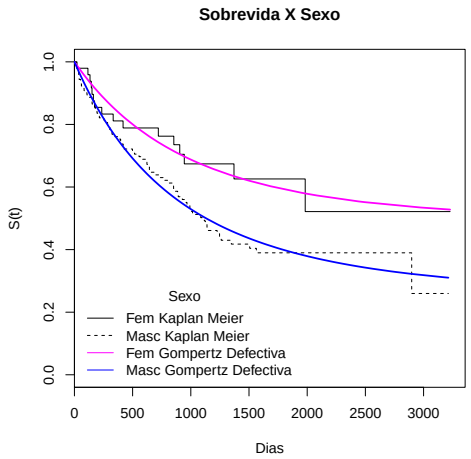

Figura 19 – Gráfico Estratificado por sexo nas estimativas paramétricas e não paramétricas. As curvas paramétricas são baseada no modelo 2, com parâmetros estimados pela abordagem da máxima.

5.3 Conclusão

Os dados do tempo para um evento, incluindo uma proporção de indivíduos que são imunes ao evento de interesse, são comuns, especialmente em estudos médicos. As ferramentas básicas para análise de sobrevivência geralmente consideram que a função de sobrevivência $S(t)$ tende a zero à medida que o tempo t tende até o infinito, e essa suposição não é realista se indivíduos imunes estão presentes. O uso de modelos baseados em distribuições defectivas é uma maneira adequada de analisar dados nesta situação. Desta forma, o modelo paramétrico baseado na distribuição modificada de Gompertz permite a estimativa da fração de cura e permite a inserção de um vetor de covariáveis. Além disso, o modelo pode ser facilmente implementado em programas computacionais como SAS, R e OpenBUGS, conforme mostrado no Apêndice. Para ilustrar o uso do modelo usamos um conjunto de dados reais de um estudo de câncer cervical e portadores de HIV. Podemos notar que o modelo se adequou aos dados.

Referências

AL-MALKI, S. J. Statistical analysis of lifetime data using new modified weibull distributions. The University of Manchester, Manchester, UK, 2014. Citado na página 17.

ALBERT, J. *Bayesian computation with R*. [S.l.]: New York: Springer-Verlag, 2007. Citado na página 25.

BALKA, J.; DESMOND, A. F.; MCNICHOLAS, P. D. Bayesian and likelihood inference for cure rates based on defective inverse gaussian regression models. *Journal of Applied Statistics*, Taylor & Francis, v. 38, n. 1, p. 127–144, 2011. Citado na página 10.

BECKER, J. L. *Estatística básica: transformando dados em informação*. [S.l.]: Bookman Editora, 2015. Citado na página 27.

BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 47, n. 259, p. 501–515, 1952. Citado na página 10.

BERNARDO; SMITH. *Smith. Bayesian Theory*. [S.l.]: John Wiley & Sons, New York, 1994. Citado na página 24.

BOAG, J. W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, v. 11, n. 1, p. 15–53, 1949. Citado 2 vezes nas páginas 10 e 16.

BOX, G. E.; TIAO, G. C. *Bayesian Inference in Statistical Analysis*. [S.l.]: New York: J. Wiley Interscience, 1992. Citado na página 22.

BRENNA, S. et al. Prognostic value of p53 codon 72 polymorphism in invasive cervical cancer in brazil. *Gynecologic oncology*, Elsevier, v. 93, n. 2, p. 374–380, 2004. Citado 3 vezes nas páginas 11, 26 e 42.

BRITO, A. M. d. et al. Aids e infecção pelo HIV no brasil: uma epidemia multifacetada. SciELO Brasil, 2001. Citado na página 12.

CARVALHO, M. S. et al. *Análise de Sobrevivência: teoria e aplicações em saúde*. [S.l.]: SciELO-Editora FIOCRUZ, 2011. Citado 2 vezes nas páginas 12 e 16.

CHIB, S.; GREENBERG, E. Understanding the metropolis-hastings algorithm. *The american statistician*, Taylor & Francis Group, v. 49, n. 4, p. 327–335, 1995. Citado na página 25.

COLOSIMO, E. A.; GIOLO, S. R. Análise de sobrevivência aplicada. In: *ABE-Projeto Fisher*. [S.l.]: Edgard Blücher, 2006. Citado 2 vezes nas páginas 15 e 17.

FAREWELL; V.T. The use of mixture models for the analysis of survival data with long-term survivors. *BIOMETRICS*, v. 38, p. 1041–1046, 1982. Citado na página 10.

GELFAND, A. E.; SMITH, A. F. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, Taylor & Francis Group, v. 85, n. 410, p. 398–409, 1990. Citado na página 24.

GIESER, P. W. et al. Modelling cure rates using the gompertz model with covariate information. *Statistics in medicine*, Wiley Online Library, v. 17, n. 8, p. 831–839, 1998. Citado na página 27.

JAFARI, A. A.; TAHMASEBI, S.; ALIZADEH, M. The beta-gompertz distribution. *Revista Colombiana de Estadística*, v. 37, n. 1, p. 141–158, 2014. Citado na página 27.

JR, D. W. H.; LEMESHOW, S. Applied survival analysis: regression modelling of time to event data (1999). *Eur Orthodontic Soc*, p. 561–2, 1999. Citado na página 12.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958. Citado na página 12.

KLEIN, J.; MOESCHBERGER, M. *Survival analysis: Techniques for censored and truncated regression*. [S.l.]: New York, NY: Springer-Verlag, 1997. Citado na página 12.

LAWLESS, J. *Statistical Models and Methodsfor Lifetime Data. New York: J.* [S.l.]: Wiley and, 1982. Citado na página 12.

MALLER, R. A.; ZHOU, X. *Survival analysis with long-term survivors*. [S.l.]: Wiley New York, 1996. Citado na página 10.

MARTINEZ, E. Z.; ACHCAR, J. A. The defective generalized gompertz distribution and its use in the analysis of lifetime data in presence of cure fraction, censored data and covariates. *Electronic Journal of Applied Statistical Analysis*, v. 10, n. 2, p. 463–484, 2017. Citado na página 28.

PAULINO, C. et al. estatística bayesiana fundação clouste gulbenkian lisboa. 2003. Citado na página 23.

ROCHA, R. et al. New d efective models based on the Kumaraswamy family of distributions with application to cancer data sets. *Statistical methods in medical research*, SAGE Publications, p. 0962280215587976, 2015. Citado na página 10.

ROCHA, R. et al. Two new defective distributions based on the marshall–olkin extension. *Lifetime data analysis*, Springer, v. 22, n. 2, p. 216–240, 2016. Citado na página 10.

ROCHA, R. et al. New defective models based on the kumaraswamy family of distributions with application to cancer data sets. *Statistical methods in medical research*, SAGE Publications Sage UK: London, England, v. 26, n. 4, p. 1737–1755, 2017. Citado na página 37.

ROCHA, R. F. d. et al. Defective models for cure rate modeling. Universidade Federal de São Carlos, 2016. Citado na página 10.

ROCHA, R. F. da; TOMAZELLA, V. L. D.; LOUZADA, F. Inferência classica e bayesiana para o modelo de frac ao de cura gompertz defeituoso. *Rev. Bras. Biom*, v. 32, n. 1, p. 104–114, 2014. Citado na página 28.

A Códigos em R

```
#Figura 7
#FDP GOMPERTZ IMPRÓPRIA
fdpgompertz <- function(i, alpha, beta){
alpha*exp(-beta*t)*exp(alpha/-beta*(1-exp(-beta*t)))
}
t <- seq(0.01, 3, by = 0.1)
pa1<- 2
pa2<- 0.5
pa3<- 2
pa4<- 1.5
plot(t, fdpgompertz(t, pa1, pa2), type = 'l',bty="l",ylim=c(0,2), ylab="f(t)",
main=expression(paste("f(t), ",alpha,"= 0.2")),las=1)
lines(t,fdpgompertz(t, pa3, pa4), type='l', pch=1, lty=1, col="red")
legend("topright", cex = 0.65,col=c("black","red"), lty = 1,
legend= c(expression(beta == 0.5), expression(beta == 1.5)), bty = 'n')
#FUNÇÃO DE SOBREVIDA GOMPERTZ IMPRÓPRIA
sgompertz <- function(t, alpha, beta){
exp(alpha/-beta*(1-exp(-beta*t)))
}
plot(t, sgompertz(t, pa1, pa2), type = 'l',bty="l",ylim=c(0,1), ylab="S(t)",
main=expression(paste("S(t), ",alpha,"= 0.2")),las=1) lines(t,sgompertz(t, pa3, pa4), type='l',
pch=1, lty=1, col="red")
legend("topright", cex = 0.65,col=c("black","red"), lty = 1,
legend= c(expression(beta == 0.5), expression(beta == 1.5)), bty = 'n')
#FUNÇÃO DE RISCO DA GOMPERTZ IMPRÓPRIA
hgompertz <- function(t, alpha, beta){
alpha*exp(-beta*t) }
plot(t, hgompertz(t, pa1, pa2), type = 'l',bty="l",ylim=c(0,2), ylab="h(t)",
main=expression(paste("h(t), ",alpha,"= 0.2")),las=1) lines(t,hgompertz(t, pa3, pa4), type='l',
pch=1, lty=1, col="red")
legend("topright", cex = 0.65,col=c("black","red"), lty = 1,
legend= c(expression(beta == 0.5), expression(beta == 1.5)), bty = 'n')
#Gráfico de Kaplan Meier com os dados dos indivíduos portadores de HIV#
#lembrando que o banco de dados está no link http:
```

```
sobrevida.fiocruz.br/aidsclassico.html #chamando o banco de dados de ipec

        #Gráfico de Kaplan-Meier Geral
require(survival)
attach(ipec)
Surv(tempo,status)
sobrev.km<-survfit( formula = Surv(tempo, status) 1,conf.type="plain")
summary(sobrev.km)
plot(sobrev.km, conf.int = F, xlab = "Tempo (Dias)", ylab = "S(t)",
main="Gráfico de Kaplan e Meier")

        #Estimando os Parâmetros da Gompertz
ekm<-survfit(Surv(tempo,status) 1)
time<-ekm$time
log.f <- function(parms){
alpha <- parms[1]
beta <- parms[2]
if (parms[1]<0) return(-Inf)
if (parms[2]<0) return(-Inf)
St <- exp(-alpha/beta*(1-exp(-beta*tempo)))
ht <- alpha*exp(-beta*tempo)
like <- St * htŝtatus
L <- sum(log(like))
if (is.na(L)==TRUE) {return(-Inf)}
else {return(L)} }
library(maxLik)
mle <- maxLik(logLik=log.f,start=c(.06,.06))
summary(mle)
alpha<-mle$estimate[1]
beta<-mle$estimate[2]
stgompertz <- exp(-alpha/beta*(1-exp(-beta*time)))

        #Gráfico de Kaplan-Meier ajustado com a Gompertz Defectiva
plot(sobrev.km, conf.int = F, xlab = "Tempo (Dias)", ylab = "S(t)",
main="Gráfico de Kaplan e Meier")
lines(c(0,time),c(1,stgompertz),lty=2,col="magenta")

        #Gráfico Kaplan Meier com estrato do sexo
survaids <- survfit(Surv(tempo, status) sexo, data = ipec)
survaids
summary(survaids)
plot(survaids, conf.int = F, main=("Sobrevida X Sexo"),
```

```
xlab = "Dias", ylab = "S(t)", mark.time = F, lty = c(1, 2))
legend( x = "bottomleft", legend = c("Fem", "Masc"),
lty = c(1, 2), title = "Sexo", bty = "n")
```

```
        #Gráfico de Contorno
logLDp<-function(x,d,alpha,beta){
a0<-log(alpha)*sum(d)
b0<-alpha/beta*sum(1-exp(-beta*x))
c0<-beta*sum(d*x)
lofLDp<-a0-b0-c0 }
xth<-seq(0.001,0.15,0.001)
xmu<-seq(0.00001,0.15,0.001)
mat<-matrix(NA,nrow=length(xth), ncol=length(xmu))
for (i in 1:length(xth)){
for (j in 1:length(xmu)){
mat[i,j] <- logLDp(t,d,xth[i],xmu[j]) }}
a<-mle$estimate[1]
b<-mle$estimate[2]
filled.contour(xth,xmu,mat, color = cm.colors,ylim=c(0,0.15),
plot.title=title(main="Log-Verossimilhança", xlab=expression(alpha),
ylab=expression(beta)), plot.axes=axis(1,seq(0,0.15,by=0.01))
axis(2,seq(0,0.15,by=0.01))
thMLE<-a
muMLE<-b
points(thMLE,muMLE,pch=19,col="red")
lines(c(thMLE,thMLE), c(0,muMLE), lty="dashed",col="red")
lines(c(0,thMLE), c(muMLE,muMLE), lty="dashed",col="red")
text(0.07747832,0.01,"0.07747832", col="red")
text(0.06680065,0.01,"0.06680065", col="red")
},
key.title=title(main="log L"),
key.axes=axis(4,seq (-540, max(mat),by=10)))
```

```
        # Gráfico - Curvas de sobrevivencia por idade
idade40<-ifelse(idade>40,1,0)
idade40
table(idade40)
ipec$idade40 = ifelse(idade>40,1,0)
head(ipec)
ipec
attach(ipec)
```

```
survaidade40 <- survfit(Surv(tempo,
status) idade40, data = ipec)
survaidade40
summary(survaidade40)
plot(survaidade40, conf.int = F, main=("Idade>40 ou <= 40 anos"),
xlab = "Dias", ylab = "S(t)", mark.time = F, lty = c(1, 2))
legend( x = "bottomleft", legend = c("Idade > 40 anos", "Idade <= 40 anos"),
lty = c(1,2), title = "Idade", bty = "n")
```

Estudo de Simulação para $\alpha = \beta$

```
rMGompertz<-function(n=50,beta=0.1)
eta<-exp(-beta/beta)
m<-rbinom(n,prob=1-eta,size=1)
u<-runif(n,0,1-eta)
y0<- -(1/beta)*log(1+(beta/beta)*log(1-u))
t0<-ifelse(m,y0,Inf)
maxti<-max(y0*m)
w<-runif(n,0,maxti)
t<-pmin(t0,w)
d<-as.numeric(t0<w)
dados<-data.frame(t,d)
return(dados)
a<-rMGompertz()
a
```

#Estudo de Simulação para $\alpha \approx \beta$

Função para calcular os estimadores de máxima-verossimilhança

```
rm(list=ls())
```

Importando o Banco de Dados

```
library(maxLik)
rMGompertz <- function(n,alpha,beta)
eta <- exp(-alpha/beta)
m <- rbinom(n,prob=1-eta,size=1)
u <- runif(n,0,1-eta)
y0 <- -(1/beta)*log(1+beta/alpha*log(1-u))
t0 <- ifelse(m,y0,Inf)
maxti <- max(y0*m)
w <- runif(n,0,maxti)
t <- pmin(t0,w)
d <- as.numeric(t0<w)
dados <- data.frame(t,d)
```

```
return(dados)
log.f <- function(parms)
alpha <- parms[1]
beta <- parms[2]
if (parms[1]<0) return(-Inf)
if (parms[2]<0) return(-Inf)
St <- exp(-alpha/beta*(1-exp(-beta*t)))
ht <- alpha*exp(-beta*t)
like <- St * ht^d
L <- sum(log(like))
if (is.na(L)==TRUE) return(-Inf)
else return(L)
library(maxLik)
mle <- maxLik(logLik=log.f,start=c(.06,.06))
summary(mle)
maxveros<-function(t,d,alpha,beta)
mle<-maxLik(logLik=log.f,start=c(alpha,beta))
a<-amp<-cob<-liminf<-limsup<-rep(NA,3)
eta<-exp(-alpha/beta)
a[1]<-mle$estimate[1]
a[2]<-mle$estimate[2]
a[3]<-exp(-a[1]/a[2])
S<-matrix(NA,2,2)
sa<-vcov(mle)
S[1,1] <- sa[1,1]
S[1,2] <- sa[1,2]
S[2,1] <- sa[2,1]
S[2,2] <- sa[2,2]
der1 <- -a[3]/a[2]
der2 <- a[3]*a[1]/(a[2]*a[2])
var3 <- der1*(der1*S[1,1]+der2*S[2,1]) + der2*(der1*S[1,2]+der2*S[2,2])
liminf[1]<-a[1] - 1.96*sqrt(S[1,1])
limsup[1]<-a[1] + 1.96*sqrt(S[1,1])
liminf[2]<-a[2] - 1.96*sqrt(S[2,2])
limsup[2]<-a[2] + 1.96*sqrt(S[2,2])
liminf[3]<-a[3] - 1.96*sqrt(var3)
limsup[3]<-a[3] + 1.96*sqrt(var3)
amp[1]<-limsup[1] - liminf[1]
amp[2]<-limsup[2] - liminf[2]
```

```
amp[3]<-limsup[3] - liminf[3]
cob[1] <- ifelse(alpha>liminf[1]  alpha<limsup[1],1,0)
cob[2] <- ifelse(beta>liminf[2]  beta<limsup[2],1,0)
cob[3] <- ifelse(eta>liminf[3]  eta<=limsup[3],1,0)
dados <- data.frame(a,liminf,limsup,amp,cob)
return(dados)
Fixando os valores nominais
alpha<- 0.2
beta <- 0.201
eta <- exp(-alpha/beta)
Tamanho da amostra
n<- 200
Número de simulações
B<-5000
#Criando os vetores para os resultados da simulação#
a.mv.cob<-b.mv.cob<-p.mv.cob<-a.amplt.mv<-b.amplt.mv<-a.mv<-b.mv<-p.mv<-p.amplt.mv<-
rep(NA,B)
for(k in 1:B)
dados<-rMGompertz(n,alpha,beta)
t<-dados[,1]
d<-dados[,2]

Estimadores de máximo-verossimilhança
mv.res<-maxveros(t,d,alpha,beta)
a.mv[k] <- mv.res[1,1]
b.mv[k] <- mv.res[2,1]
p.mv[k] <- mv.res[3,1]
a.mv.cob[k] <- mv.res[1,5]
b.mv.cob[k] <- mv.res[2,5]
p.mv.cob[k] <- mv.res[3,5]
a.amplt.mv[k] <- mv.res[1,4]
b.amplt.mv[k] <- mv.res[2,4]
p.amplt.mv[k] <- mv.res[3,4]

Probabilidade de Cobertura
mean(a.mv.cob,na.rm=T)
mean(b.mv.cob,na.rm=T)
mean(p.mv.cob,na.rm=T)
Amplitude (médias)
```

```
mean(a.amplt.mv,na.rm=T)
mean(b.amplt.mv,na.rm=T)
mean(p.amplt.mv,na.rm=T)
Bias
a.bias <- mean(a.mv-alpha)
b.bias <- mean(b.mv-beta)
p.bias <- mean(p.mv-eta)
Erro quadrático médio
a.MSE <- sum((a.mv-alpha)*(a.mv-alpha))/B
b.MSE <- sum((b.mv-beta)*(b.mv-beta))/B
p.MSE <- sum((p.mv-eta)*(p.mv-eta))/B
a.bias
b.bias
p.bias
a.MSE
b.MSE
p.MSE
```