Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Modelagem de fração cura com a distribuição Gompertz

André Gabriel Paulino de Oliveira

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Modelagem de fração cura com a distribuição Gompertz

André Gabriel Paulino de Oliveira
Orientadora: Teresa Cristina Martins Dias

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Setembro 2022

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

Cure rate modelling using the Gompertz distribution

André Gabriel Paulino de Oliveira
Advisor: Teresa Cristina Martins Dias

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
September 2022

André Gabriel Paulino de Oliveira

Modelagem de fração cura com a distribuição Gompertz

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por André Gabriel Paulino de Oliveira e aprovado pela banca examinadora.

Aprovado em 06 de setembro de 2022

Banca Examinadora:

- Teresa Cristina Martins Dias
- Estela Maris Pereira Bereta
- Vera Lucia Damasceno Tomazella

Para meu avô, Benedito.

Agradecimentos

Agradeço aos meus pais, André e Sirlene pela vida e por todo apoio durante todos os momentos. Agradeço aos professores do departamento de Estatística, em especial a minha orientadora Cris pela maravilhosa orientação e, aos membros da banca, Vera e Estela.

Também agradeço aos meus amigos tanto de São José do Rio Pardo por todos os bons momentos, apoio e afins: Léo Baptistella, Luís, Lucas, Matarazzo, Tibério, William entre outros, como os que conheci na graduação, Luben, Luiz Piccin, *Professor*, Vinícius Hideki, Vitor Ramos e Vitor Schiavone.

"From my rotting body, flowers shall grow and I am in them, and that is eternity"

(Edvard Munch)

Resumo

A Análise de Sobrevivência é um campo de estudo bastante amplo e importante da Estatística. A partir dos modelos de sobrevivência é possível estimar a função de sobrevivência ou confiabilidade, descrever as características dos dados, entre outros objetivos. Na área médica, o objeto de estudo é o tempo até a ocorrência de eventos, por exemplo, a cura ou falecimento do paciente. Porém podem existir indivíduos que nunca sofrerão o evento de interesse, que nesse caso pode ser falecimento ou cura. Nesse estudo, descrevemos e exploramos dois dos principais tipos de modelos que consideram fração de "curados", denominados de modelo de fração de cura com mistura e modelo defeituoso. Assumimos uma distribuição Gompertz para o tempo até o evento de interesse. A estimação dos parâmetros envolvidos é realizada sob as abordagens frequentista e Bayesiana. Aplicações em conjunto de dados reais e simulados são consideradas para ilustrar a teoria estudada e comparar as abordagens.

Palavras-chave: *Análise de Sobrevivência, Distribuição Gompertz, Modelos com fração de cura.*

Abstract

Survival Analysis is a broad and crucial field of Statistical Science. Survival models are useful for estimating the reliability function, describe data properties and more. Considering medical studies, the aim is studying the time until an event occurs, like the death of a patient. However, there can be situations in which the subject will not experience the event. In this study, we explore two parametric models for modelling the cure rate, the mixture model and the defective model. The times to event are assumed to follow the Gompertz distribution. Both frequentist and Bayesian approaches are considered for the estimation problem. To exemplify the discussed theory and compare the approaches, applications in real data are discussed.

Keywords: *Survival Analysis, Gompertz Distribution, Cure Models.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Em Estatística, a área de Análise de Sobrevivência e Confiabilidade tem sido bastante aplicada nos mais diversos campos de estudo. Esta parte da Estatística estuda tempos até a ocorrência de um ou mais eventos de interesse nas unidade amostrais observadas. Por exemplo, se são observados pacientes com uma doença, o tempo de cura ou de falecimento é registrado; se a unidade amostral é um equipamento eletrônico, o tempo até a quebra é anotado; se estamos na área financeira, o tempo até o cliente se tornar inadimplente é registrado. Após definir qual é o evento de interesse e consequentemente, como se dará o registro dos tempos observados, um dos objetivos é modelar o comportamento de tais tempos e estimar a função de sobrevivência ou confiabilidade, entre outras funções.

O termo Análise de Sobrevivência refere-se principalmente a estudos da área médica. Porém, em diversas outras áreas, há preocupação em analisar eventos relacionados a tempos de falha, tais como a confiabilidade de um produto, sob certas condições, durar mais que um ano. Na área industrial, o estudo envolvendo modelagem da função de sobrevivência é chamada de Análise de Confiabilidade (Rausand e Hoyland, 2003). Em outras áreas, como criminologia, o foco é o tempo de reincidência ao crime de ex-detentos (Cloyes *et al.*, 2010). Já nas áreas de estudos populacionais e econômicos o foco é a obtenção de estimativas para a probabilidade de uma empresa ainda estar em operação após certo período (Nunes e de Morais Sarmento, 2012), entre outras aplicações.

Nesta área também podem ser consideradas outras características, tais como a ocorrência de eventos múltiplos ou recorrentes, covariáveis, além da presença de informação parcial, ou seja o evento não ocorreu no período de estudo. Esta informação é chamada de censura.

Esta área tem um metodologia própria para ajuste de modelos, pois trata com a variá-

vel aleatória positiva o tempo, geralmente com comportamento assimétrico, sendo possível considerar na análise as características supramencionadas. Existem diversos modelos que podem ser ajustados para descrever os tempos e, considerando os enfoques frequentista e Bayesiano, estimamos as funções de interesse. No enfoque não paramétrico existem algumas técnicas, usadas para estimar diretamente as funções de risco e de sobrevivência.

Desse modo, existe interesse por pesquisadores no desenvolvimento de modelos adequados para descrever o comportamento do tempo até falha, morte ou outro evento. Entre as técnicas mais conhecidas para descrever o comportamento dos tempos observados até o evento, destacamos o estimador não paramétrico produto-limite, mais conhecido como estimador de Kaplan e Meier (Kaplan e Meier, 1958) e o modelo de riscos proporcionais de Cox (Cox, 1972).

Do ponto de vista paramétrico, uma classe de modelos que tem sido bastante utilizados nesta área é o que assume que uma parte da população nunca sofrerá o evento de interesse (morte, cura, falha, etc.). Modelos com esta característica são chamados de modelos de fração de cura (Chernick e Friis, 2003). Dos vários tipos de modelos de fração de cura consideramos o modelo de fração de cura com mistura e o modelo defeituoso. Os modelos com mistura são geralmente compostos por dois submodelos, sendo que um trata da estimação da proporção de indivíduos não suscetíveis ao evento de interesse e o outro que modela a função de sobrevivência para os indivíduos suscetíveis. Já os modelos defeituosos não fazem essa distinção, possuindo uma abordagem alternativa, sem a utilização de submodelos (Balka *et al.*, 2009). Nesse trabalho, consideramos ambos modelos e empregamos a distribuição Gompertz para sua construção.

Temos como objetivo estudar a distribuição Gompertz e sua abordagem em situações nas quais existe uma fração de curados no conjunto de dados. Mais especificamente, ajustar modelos no contexto defeituoso e modelo de mistura padrão.

O trabalho é organizado como segue: no Capítulo 2 tratamos da revisão bibliográfica dos conceitos básicos de Análise de Sobrevivência, apresentamos o estimador não paramétrico de Kaplan-Meier, a distribuição Gompertz e suas propriedades e, como se dá a estimação frequentista dos parâmetros; no Capítulo 3 descrevemos os modelos de fração de cura com mistura e defeituoso; o método Bayesiano de estimação dos parâmetros é explicado em detalhes no Capítulo 4; no Capítulo 5 aplicamos os modelos discutidos em cinco conjuntos de dados, comparando-os em termos de desempenho e, por fim, no Capítulo 6 apresentamos as discussões referentes às aplicações.

Capítulo 2

Análise de Sobrevivência

Neste capítulo apresentamos uma breve revisão dos conceitos básicos da Análise de Sobrevivência. Na Seção 2.1 tratamos das definições de tempo observado e censura, das funções de interesse em um estudo da Área de Sobrevivência e de estimadores para as mesmas, como o estimador de Kaplan-Meier e de Nelson-Aalen (Seção 2.2). Apresentamos a função de verossimilhança, no caso com censuras na Seção 2.3 e a distribuição Gompertz, assim como suas propriedades teóricas e como se dá a estimação de seus parâmetros é apresentada na Seção 2.4.

2.1 Conceitos Básicos

Tempo e Censura

Como dito na Introdução (Capítulo 1), a Análise de Sobrevivência é uma parte da estatística cujo objeto de estudo são os tempos até a ocorrência de determinado evento de interesse. Desse modo, os registros - que são os tempos, são não negativos, cuja distribuição é comumente assimétrica à direita. Podemos definir T como uma variável aleatória (mais informações sobre variáveis aleatórias podem ser encontradas em Casella e Berger (2021)), sujeita à condição de que $P(T < 0) = 0$.

Outra característica dos dados neste contexto, que justifica a utilização de métodos específicos da área, é a presença de informação parcial nos tempos observados (Collett, 2015). Esta informação parcial representa o fato do indivíduo não ter sofrido o evento de interesse durante o tempo de acompanhamento. Se o evento ocorreu durante o período de estudo e, temos conhecimento do tempo até a ocorrência, então essa informação é

registrada. Porém se ao ocorrer o evento, o tempo não for conhecido ou se o evento não ocorrer durante o período observacional, então esse tempo é dito censurado. Isso pode ocorrer por diversas razões, por exemplo, se estamos falando de um estudo médico cujo evento de interesse é o falecimento do paciente, o mesmo pode vir a abandonar o estudo antes do óbito ou procurar tratamento em outra região geográfica, etc. A falta de informação sobre o tempo exato de ocorrência do evento não pode ser ignorada, devendo ser tratada de maneira especial (Collett, 2015). Existem diversos tipos de censura, dentre estas destacamos:

- Censura à direita: é o tipo mais comum e ocorre quando o indivíduo é acompanhado desde um tempo t_0 até um tempo posterior t_1. Nesse caso, a informação extraída é somente que o evento de interesse não ocorreu entre t_0 e t_1.

- Censura à esquerda: se dá quando sabemos que o indivíduo sofreu o evento de interesse em um tempo anterior ao começo do estudo t_0, mas não sabemos exatamente quando. Por exemplo, se o interesse é o aprendizado de `R` para formados em cursos da área de ciência biológicas, sendo que estes declararam não ter conhecimento prévio de `R`. Podem haver pessoas que já possuem algum conhecimento em `R` no começo do estudo, o que configuraria censura à esquerda. Quando o evento de interesse é o falecimento, censura à esquerda não é uma opção. Esse tipo de censura não é tão comum e nos informa que o tempo real é menor que o registrado.

- Censura intervalar: é quando o evento aconteceu em um intervalo de tempo, mas não sabemos o instante exato. Por exemplo, se o evento de interesse é a recidiva de uma doença, o paciente pode ter feito um exame em janeiro, no qual não foi detectada a enfermidade e outro em junho que acuse a presença da doença. O evento ocorreu, possivelmente entre janeiro e junho, mas há falta de conhecimento sobre o instante exato do tempo.

Nesse trabalho consideramos situações da área médica, nas quais:

- Os indivíduos (denotados por i, $i = 1, \ldots, n$) são as unidades amostrais pertencentes a um grupo de estudo;

- Os registros são os tempos no quais ocorreram o evento de interesse;

- Se o evento não aconteceu durante o período de acompanhamento, então o registro é dito censurado (censura à direita do tipo I).

Por fim, a partir dos conceitos de tempo e censura, os registros para um indivíduo i qualquer, são representados por um par de informações (t_i, δ_i), em um grupo de n indivíduos $i = 1, 2, \ldots, n$. Denotamos como δ_i a variável indicadora de ocorrência do evento, definida por

$$\delta_i = \begin{cases} 1, \text{ se } t_i \text{ é tempo em que o evento de interesse aconteceu,} \\ 0, \text{ se } t_i \text{ é tempo referente à censura.} \end{cases} \tag{2.1}$$

Funções de Sobrevivência e de Risco

Como o tempo T até a ocorrência de um evento é uma variável aleatória contínua e não negativa, a função de distribuição de T é dada por

$$F(a) = \int_0^a f(t)dt, a > 0, \tag{2.2}$$

em que $f(t)$ é a função densidade de probabilidade de T, cujo suporte é o conjunto $\mathbb{R}^+$. A expressão em (2.2) representa a probabilidade de que o tempo de vida seja menor do que a. A partir de $F(t)$, é possível definir a função de sobrevivência $S(t)$,

$$S(t) = P(T > t) = 1 - F(t), \tag{2.3}$$

que corresponde à probabilidade do indivíduo sobreviver além do tempo t. Outra função importante é a função de risco instantâneo, definida como

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

que corresponde ao risco de ocorrência em certo tempo, dado que esse evento ainda não ocorreu (para mais detalhes sobre essa expressão ver A.1). Por fim, temos a função de risco acumulado $H(t)$, que corresponde ao total de risco sofrido pelo indivíduo até aquele momento,

$$H(t) = \int_0^t h(t)dt = \int_0^t \frac{f(t)}{S(t)}dt = -\ln(S(t)). \tag{2.4}$$

Podemos notar que quanto menor o valor da função de sobrevivência (2.3) no ponto t, maior o risco acumulado $H(t)$.

2.2 Estimadores Não Paramétricos

Dado o interesse em estimar a função de sobrevivência e, considerando censura à direita, apresentamos o estimador mais simples para tal quantidade, o chamado estimador de Kaplan-Meier (Kaplan e Meier, 1958). Esse estimador é não paramétrico e tem seguinte forma

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right), \ j = 1, 2, \ldots, n, \tag{2.5}$$

em que $t_1, t_2, \ldots, t_n$ são os tempos ordenados de ocorrência do evento, n_j é o número de indivíduos que não foram censurados ou ainda não sofreram o evento de interesse até o tempo imediatamente anterior a t_j e d_j o número de mortes no instante t_j. O estimador em (2.5) tem propriedades bastante interessantes, dentre estas podemos elencar que,

1. é não viesado para grandes amostras;
2. é fracamente consistente;
3. é estimador de máxima verossimilhança para $S(t)$ (Kalbfleisch e Prentice, 2011);
4. converge assintoticamente para um processo Gaussiano.

A partir de uma estimativa pontual, é natural construir um intervalo de confiança para a função de sobrevivência por meio do estimador (2.5) e para tal, é necessário uma expressão para sua variância. Podemos calcular a variância assintótica desse estimador pela fórmula de Greenwood (Kalbfleisch e Prentice, 2011), que é dada por

$$Var(\hat{S}(t)) = \hat{S}(t)^2 \sum_{j:t_j<t} \frac{d_j}{n_j(n_j - d_j)}. \tag{2.6}$$

Uma maneira de calcular esse intervalo é, sob tamanhos amostrais razoáveis, assumir que a distribuição de $\hat{S}(t)$ é normal, com média $\hat{S}(t)$ e variância dada por (2.6). Para um nível de significância $0 < \alpha < 1$ o intervalo com $100(1-\alpha)\%$ de confiança (IC) é dado por

$$IC(S(t), 100(1-\alpha)\%) = \hat{S}(t) \pm z_{1-\alpha/2}\sqrt{Var(\hat{S}(t))},$$

em que $z_{1-\alpha/2}$ é o quantil da distribuição normal padrão. Essas propriedades contribuem para a longevidade e popularidade desse estimador, porém o mesmo é restrito somente a problemas em que há censura à direita. Além disso, não é possível incorporar de maneira natural covariáveis na estimativa de $S(t)$, o que abre espaço para outros tipos de modelos.

Notamos também, que como a função de sobrevivência se relaciona à função de risco acumulado por (2.4), podemos também considerar um estimador para essa última. Mais comumente, (2.4) é estimada pelo método, também não paramétrico, de Nelson-Aalen (Colosimo e Giolo, 2006), cuja expressão é dada por

$$\hat{H}(t) = \sum_{t_j \leq t} \frac{d_j}{n_j} = \sum_{t_j \leq t} \hat{h}(t), \tag{2.7}$$

em que n_j é a quantidade de indivíduos com tempo registrado igual a t_j e d_j o total de eventos que ocorreram nesse instante. Podemos por fim, estimar a função de sobrevivência a partir de (2.7),

$$\tilde{S}(t) = e^{-\hat{H}(t)}.$$

É possível demonstrar que as estimativas para $S(t)$ obtidas via Nelson-Aalen são sempre maiores ou iguais às estimativas obtidas por Kaplan-Meier. Borgan (2005) mostra a variância, cálculos dos intervalos de confiança assintóticos e outras propriedades do estimador de Nelson-Aalen. Vale notar que existem outros estimadores para a função de sobrevivência, alguns dos quais consideram censura intervalar (por exemplo, ver Zhang e Sun (2010)).

2.3 Função de Verossimilhança

Assumindo presença de censura à direita do tipo I e uma distribuição conhecida para $f(t)$, é natural estimar os parâmetros indexados na distribuição e, a partir destes obtemos o comportamento da função sobrevivência estimada.

Sejam T uma variável aleatória e $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)$ representando os tempos registrados dos n indivíduos no estudo e, $\boldsymbol{\delta}$ uma variável aleatória que indica censura $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)$, como definido em (2.1). Seja $f(t\ |\boldsymbol{\theta})$ a função densidade de probabilidade que descreve os tempos registrados e $\boldsymbol{\theta}$ o vetor de parâmetros de $f(t|\boldsymbol{\theta})$.

A estimação dos parâmetros é feita utilizando, por exemplo, o método da máxima verossimilhança que encontra o vetor $\hat{\boldsymbol{\theta}}$, que contém estas estimativas. Esses estimadores maximizam a função de verossimilhança $L(.)$ ou de log-verossimilhança $l(.)$. Definimos a contribuição de cada indivíduo i para a função de verossimilhança da seguinte forma

$$L_i(\boldsymbol{\theta}|\delta_i, t_i) = [f(t_i|\boldsymbol{\theta})^{\delta_i}][S(t_i|\boldsymbol{\theta})^{1-\delta_i}],\ i = 1, \ldots, n, \tag{2.8}$$

em que os tempos para cada indivíduo são independentes e identicamente distribuídos. Ressaltamos que, a informação da censura é utilizada para descrever se a contribuição para a função de verossimilhança vem da função densidade ou da função de sobrevivência no instante $t_i,\ i = 1, 2, \ldots, n$. Escrevendo a função de verossimilhança completa, temos

$$L(\boldsymbol{\theta}|\boldsymbol{\delta}, \boldsymbol{t}) = \prod_{i=1}^{n} \left\{ [f(t_i|\boldsymbol{\theta})^{\delta_i}][S(t_i|\boldsymbol{\theta})^{1-\delta_i}] \right\}. \tag{2.9}$$

Aplicando o logaritmo natural (ln) na função (2.9), a função de log-verossimilhança é dada por

$$l(\boldsymbol{\theta}|\boldsymbol{\delta}, \boldsymbol{t}) = \sum_{i=1}^{n} \left\{ \delta_i \ln[f(t_i|\boldsymbol{\theta})] + (1 - \delta_i) \ln[S(t_i|\boldsymbol{\theta})] \right\}.$$

A partir de (2.9), encontramos o EMV. Ressaltamos que em muitos casos é necessário a utilização de métodos numéricos para obter as estimativas, como o algoritmo de Newton-Raphson (Akram e Ann, 2015).

2.4 Distribuição Gompertz

A distribuição Gompertz é uma distribuição de probabilidade contínua, cujo nome é referente a Benjamin Gompertz (1779 - 1865). A distribuição é essencialmente uma extensão da lei de mortalidade proposta pelo mesmo no século XIX, para modelar tempos de vida (ou de morte, dependendo da perspectiva) de pessoas. Essa lei tem forma

$$\mu_x = ae^{bx},\ a, b, x > 0. \tag{2.10}$$

em que μ_x é taxa de mortalidade para pessoas com idade x, a é mortalidade quando $x = 0$ e b é taxa de envelhecimento (Kirkwood, 2015). Como o exponente em (2.10) é sempre positivo, a taxa de mortalidade cresce conforme a idade, logo (2.10) é adequada para descrever o comportamento das taxas de falecimento em algumas populações (Pollard e Valkovics, 1992).

A partir do raciocínio detalhado em Pollard e Valkovics (1992), é possível obter a função densidade de probabilidade, que costuma ser bastante utilizada na modelagem de tempos de vida de adultos na área de Estudos Populacionais. Dentre as diferentes parametrizações existentes na literatura, escolhemos a usada pelo pacote `flexsurv` (Jackson, 2016) do `R`; nesse caso, a função densidade de probabilidade para os tempos T é escrita

como

$$f(t|a,b) = be^{at}\exp\left(-\frac{b}{a}(e^{at}-1)\right), \tag{2.11}$$

em que $t > 0$, esta distribuição é denotada por $Gompertz(a,b)$. Uma função densidade de probabilidade é imprópria quando a mesma não integra 1. Neste caso, quando o parâmetro a é negativo configura-se a distribuição Gompertz imprópria, ou seja

$$\int_0^\infty be^{at}\exp\left[-\frac{b}{a}(e^{at}-1)\right]dt < 1.$$

Portanto, a função de sobrevivência $S(t|.)$ não atinge o valor zero e seu limite inferior é

$$\lim_{t\to\infty}\exp\left[-\frac{b}{a}(e^{at}-1)\right] = e^{\frac{b}{a}} = \pi \in (0,1),$$

pois $\lim_{t\to\infty} e^{at} - 1 = -1$, dado que $a < 0$.

Obtemos a função de distribuição para a Gompertz

$$F(t|a,b) = 1 - \exp(-\frac{b}{a}(e^{at}-1)). \tag{2.12}$$

Usando a relação (2.3), a função de sobrevivência é expressa como

$$S(t|a,b) = 1 - F(t|a,b) = \exp(-\frac{b}{a}(e^{at}-1)) \tag{2.13}$$

e função de risco $h(t)$ é

$$h(t|a,b) = be^{at}. \tag{2.14}$$

A partir de (2.14), observamos que a taxa de risco dessa distribuição é crescente quando $a > 0$, decrescente quando $a < 0$ e constante se $a = 0$. Por fim, a função de risco acumulado é dada por

$$H(t) = -\ln(S(t)) = \frac{b}{a}(e^{at}-1).$$

Na Figura 2.1 mostramos o comportamento da função de densidade (2.11), a Figura 2.2 corresponde à função de sobrevivência (2.13) e na Figura 2.3 exibimos a função de risco. Nas três figuras consideramos diversas situações, com diferentes valores para os parâmetros e estas foram criadas no *software* `R` (R Core Team, 2021). Quando a é negativo, as curvas expostas tem comportamento estritamente decrescente, quando a é positivo, observamos um comportamento unimodal na distribuição.

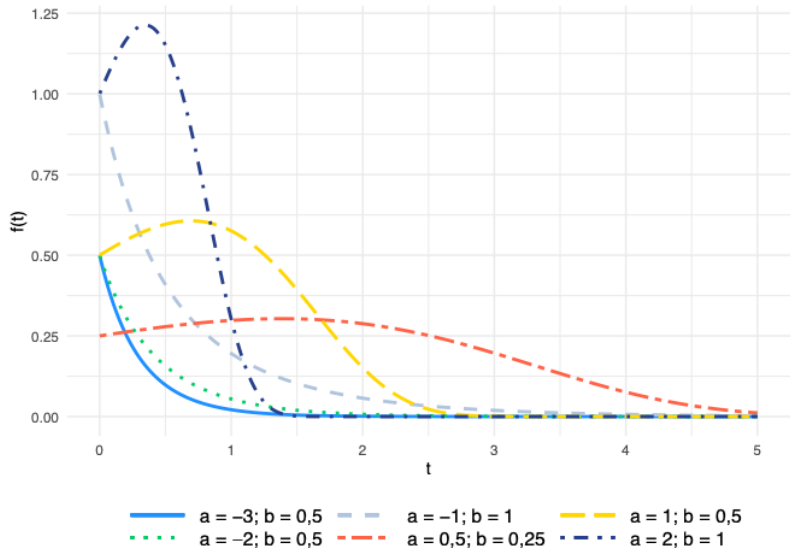


Figura 2.1: Função densidade da distribuição Gompertz.

Na Figura 2.2 mostramos a função de sobrevivência para diversas combinações de a e b. As curvas que cujo parâmetro a é negativo não decaem para 0, existindo naturalmente um patamar ou assíntota. Esta característica torna a distribuição adequada para modelar situações nais quais a função de sobrevivência não atinge 0, como mostraremos no decorrer do trabalho.

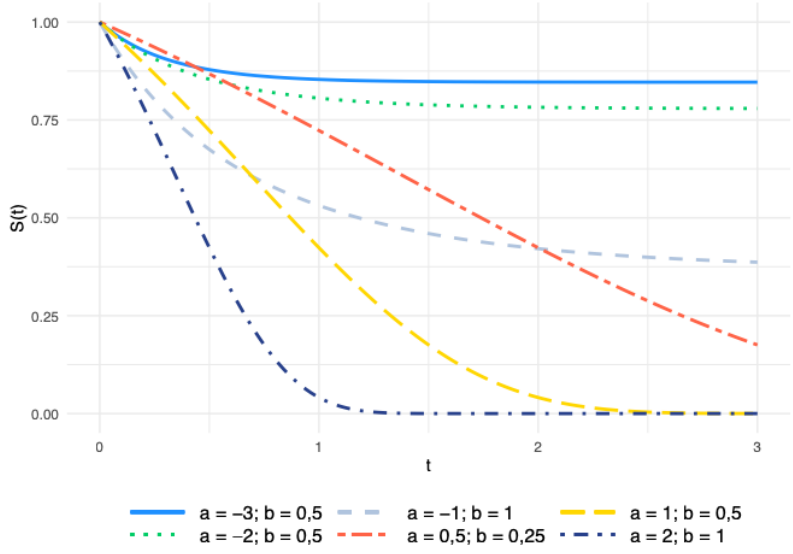


Figura 2.2: Função sobrevivência da distribuição Gompertz.

Em complemento à Figura 2.2, trazemos o valor do patamar na Tabela 2.4.1. Notamos que quanto mais negativo a for para um b fixo, maior é este valor.

Por fim, na Figura 2.3 apresentamos as funções de risco instantâneo. Nesse ponto,

Tabela 2.4.1: Limite inferior de $S(t)$ para cada conjunto de parâmetros.

a	b	π
-3	0,50	0,8465
-2	0,50	0,7788
-1	1,00	0,3679
0,5	0,25	-
1	0,50	-
2	1,00	-

vemos a relação entre a taxa de risco e a sobrevivência. As curvas amarela, azul escura e vermelha (de parâmetros $a = 1; b = 1$, $a = 2; b = 2$ e $a = 0,5; b = 0,25$, respectivamente), cuja $S(t)$ decai para 0, possuem riscos estritamente crescentes. As curvas de risco nos casos $a = -3; b = 0,5$, $a = -2; b = 0,5$ e $a = -1; b = 1$ são decrescentes e suas respectivas curvas de sobrevida não atingem o valor de 0. Notamos que o suporte t nessa figura é menor do que das outras, pois o risco $h(t)$ pode crescer exponencialmente em t.

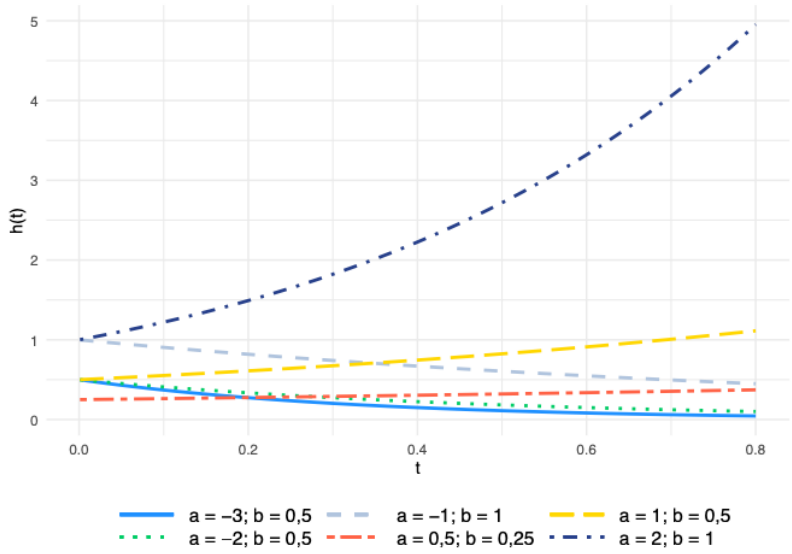


Figura 2.3: Taxa de risco da distribuição Gompertz.

2.4.1 Estimação dos parâmetros

Para estimar os parâmetros sob a perspectiva clássica, o método mais comumente utilizado é o da maximização da função de verossimilhança (Casella e Berger, 2021). Para isso, em (2.9) usamos as funções densidade de probabilidade (2.11) e de sobrevivência

(2.13) da distribuição Gompertz, desta forma, obtemos

$$\begin{aligned} L(\boldsymbol{\theta}|\boldsymbol{t},\boldsymbol{\delta}) &= \prod_{i=1}^{n} \left\{ [f(t_i|\boldsymbol{\theta})^{\delta_i}][S(t_i|\boldsymbol{\theta})^{1-\delta_i}] \right\} \\ &= \prod_{i=1}^{n} \left\{ \left[be^{at_i} \exp\left(-\frac{b}{a}(e^{at_i}-1) \right) \right]^{\delta_i} \left[\exp\left(-\frac{b}{a}(e^{at_i}-1) \right) \right]^{1-\delta_i} \right\}, \end{aligned} \tag{2.15}$$

em que $\boldsymbol{\theta} = (a, b)$. Para facilitar a manipulação da expressão (2.15), aplicamos o logaritmo (na base neperiana) na mesma,

$$\begin{aligned} l(\boldsymbol{\theta}|\boldsymbol{t},\boldsymbol{\delta}) &= \ln L(\boldsymbol{\theta}|\boldsymbol{t},\boldsymbol{\delta}) \\ &= \ln \prod_{i=1}^{n} \left\{ \left[be^{at_i} \exp\left(-\frac{b}{a}(e^{at_i}-1) \right) \right]^{\delta_i} \left[\exp\left(-\frac{b}{a}(e^{at_i}-1) \right) \right]^{1-\delta_i} \right\} \\ &= \sum_{i=1}^{n} \ln \left\{ \left[be^{at_i} \exp\left(-\frac{b}{a}(e^{at_i}-1) \right) \right]^{\delta_i} + \ln \left[\exp\left(-\frac{b}{a}(e^{at_i}-1) \right) \right]^{1-\delta_i} \right\} \\ &= \sum_{i=1}^{n} \delta_i \left[\ln be^{at_i} + \ln \exp\left(-\frac{b}{a}(e^{at_i}-1) \right) \right] + (1-\delta_i) \left[\ln \exp\left(-\frac{b}{a}(e^{at_i}-1) \right) \right] \\ &= \sum_{i=1}^{n} \delta_i \left[\ln(b) + at_i - \frac{b}{a}(e^{at_i}-1) \right] + (1-\delta_i) \left[-\frac{b}{a}(e^{at_i}-1) \right]. \end{aligned} \tag{2.16}$$

As estimativas de máxima verossimilhança para o vetor $\boldsymbol{\theta}$ são definidas como o par de parâmetros que maximiza (2.16). Em outras palavras

$$\hat{\boldsymbol{\theta}} = (\hat{a}, \hat{b}) = \arg_{a,b} \max l(\boldsymbol{\theta}|\boldsymbol{t},\boldsymbol{\delta}).$$

Dado que a função de máxima verossimilhança é côncava (Casella e Berger, 2021), sabemos que os pontos de inflexão existentes são pontos de máximo. Portanto, para maximizar essa função e encontrar as estimativas, devemos derivar (2.16) em função de a e b e igualar as derivadas a 0. Para a, temos

$$\frac{d\, l(\boldsymbol{\theta}|\boldsymbol{t},\boldsymbol{\delta})}{d\, a} = \sum_{i=1}^{n} \delta_i \left[t_i + \frac{b}{a^2}(e^{at_i}-1) - be^{at_i} \right] - (1-\delta_i) \left[\frac{b}{a^2}(e^{at_i}-1) - be^{at_i} \right] \tag{2.17}$$

e para b

$$\frac{d\, l(\boldsymbol{\theta}|\boldsymbol{t},\boldsymbol{\delta})}{d\, b} = \sum_{i=1}^{n} = \delta_i \left(\frac{1}{b} + \frac{(e^{at_i}-1)}{a} \right) + (1-\delta_i) \frac{(e^{at_i}-1)}{a}. \tag{2.18}$$

A título de motivação, importância e uso de métodos numéricos, igualamos (2.17) a 0, na busca de uma expressão fechada para o estimador de máxima verossimilhança para a,

$$\sum_{i=1}^{n} \delta_i \left[t_i + \frac{b}{a^2}(e^{at_i} - 1) - be^{at_i} \right] - (1 - \delta_i) \left[\frac{b}{a^2}(e^{at_i} - 1) - be^{at_i} \right] = 0 \implies$$

$$\sum_{i=1}^{n} \delta_i \left[t_i + \frac{b}{a^2}(e^{at_i} - 1) - be^{at_i} \right] = \sum_{i=1}^{n} (1 - \delta_i) \left[\frac{b}{a^2}(e^{at_i} - 1) - be^{at_i} \right]$$

$$\sum_{i=1}^{n} \delta_i \left[t_i + \frac{b}{a^2}(e^{at_i} - 1) - be^{at_i} \right] = \sum_{i=1}^{n} \left[\frac{b}{a^2}(e^{at_i} - 1) - be^{at_i} \right] - \delta_i \left[\frac{b}{a^2}(e^{at_i} - 1) - be^{at_i} \right]$$

$$\sum_{i=1}^{n} \delta_i t_i = \sum_{i=1}^{n} \left[\frac{b}{a^2}(e^{at_i} - 1) - be^{at_i} \right]$$

$$\sum_{i=1}^{n} \delta_i t_i = b \sum_{i=1}^{n} \left[\frac{1}{a^2}(e^{at_i} - 1) - e^{at_i} \right]$$

$$b = \frac{\sum_{i=1}^{n} \delta_i t_i}{\sum_{i=1}^{n} \frac{1}{a^2}(e^{at_i} - 1) - e^{at_i}}.$$

Assim, temos

$$\hat{b} = \frac{\sum_{i=1}^{n} \delta_i t_i}{\sum_{i=1}^{n} \frac{1}{\hat{a}^2}(e^{\hat{a}t_i} - 1) - e^{\hat{a}t_i}}.$$

Embora tenhamos conseguido isolar as expressões que dependem de a, ainda precisamos da estimativa de b. Entretanto não é possível obter expressão fechada para os estimadores. Assim, recorremos aos métodos numéricos como o algoritmo de Newton-Raphson (Akram e Ann, 2015).

2.4.2 Simulação

Para verificar a consistência dos métodos de estimação, observamos as estimativas à medida que o tamanho amostral cresce. Para tal, geramos um conjunto de valores, que representam o tempo de ocorrência até o evento, provenientes de uma distribuição Gompertz com parâmetros $a = 0,5$ e $b = 2$. Nesse momento, não consideramos tempos censurados (de modo que a função de verossimilhança só tenha contribuição da função densidade de probabilidade). Os códigos desenvolvidos para essa simulação e para o trabalho como um todo, se encontram no disponíveis e documentados no repositório https://github.com/kyniaz/TG.

Essa simulação contempla cinco tamanhos amostrais, $n = 25, 50, 100, 200$ e 500. Na Tabela 2.4.2 apresentamos os valores estimados para os parâmetros pelo método de maximização da função de verossimilhança. Observamos que conforme n aumenta, as estimati-

Tabela 2.4.2: Estimativas para o caso sem censura.

Tamanho Amostral	$\hat{a}$	$\hat{b}$
25	2,048	1,235
50	1,294	1,106
100	0,097	2,378
200	0,528	1,799
500	0,657	1,821

vas ficam mais próximas dos valores reais dos parâmetros, o que indica que as técnicas de estimação definidas estão consistentes. Na Figura 2.4 apresentamos as funções de sobrevivência $\hat{S}(t)$ estimadas e sobrepostas ao estimador de Kaplan-Meier (para mais detalhes das estimativas sem censura, ver Apêndice B), construído com a amostra de tamanho $n = 500$. As curvas estimadas de maneira paramétrica e não paramétrica estão próximas.

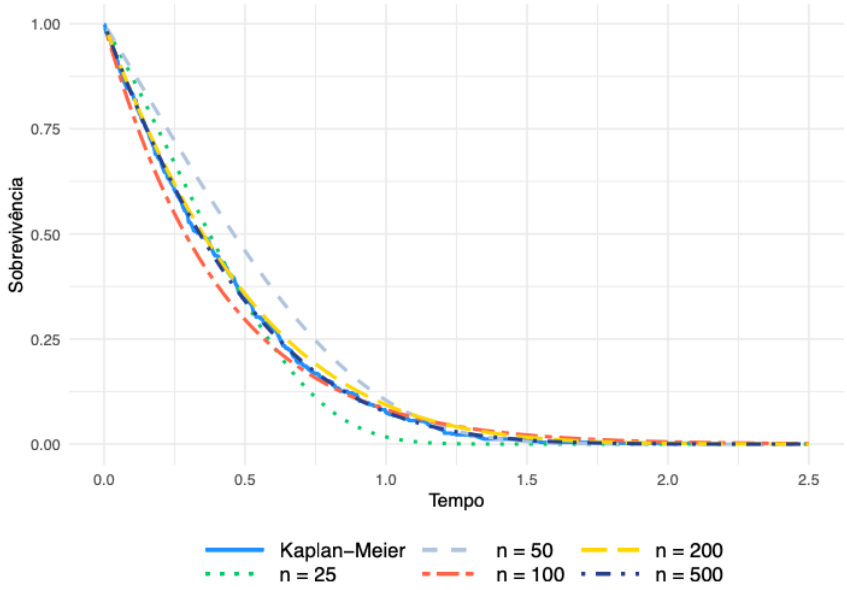


Figura 2.4: Função de sobrevivência estimada via Kaplan-Meier e Gompertz para diversos tamanhos amostrais (n).

Na presença de censura à direita, simulamos tempos da distribuição Gompertz. Para isso, geramos de duas distribuições Gompertz, que denotamos por $X_1 \sim Gompertz(1;5)$ e $X_2 \sim Gompertz(1;1)$, respectivamente. A variável indicadora de censura é definida como

$$\delta_i = \begin{cases} 1, \text{ se } X_{1i} < X_{2i}, \\ 0, \text{ c.c.}, \end{cases} \qquad i = 1, 2, \ldots, n,$$

em que X_{1i} é o i-ésimo tempo registrado pertencente à variável X_1 e X_{2i} é o i-ésimo tempo registrado pertencente à X_2. Desse modo, cada elemento de $\boldsymbol{t}$ representa os tempos observados e é dado por

$$t_i = \begin{cases} X_{1i}, \text{ se } \delta_i = 1 \\ X_{2i}, \text{ c.c.} \end{cases} \qquad i = 1, 2, \ldots, n.$$

Para estimarmos parâmetros a e b da distribuição, usamos o método da maximização da função de verossimilhança (2.15).

Na Tabela 2.4.3 reportamos valores estimados para diferentes quantidades de observações, de acordo com a proporção de tempos censurados. Notamos que os parâmetros estimados se encontram relativamente diferentes dos que foram utilizados para simular os dados. Isso ocorre por conta dos tamanhos amostrais diferentes e especialmente pela quantidade de censura, que altera a relação entre os parâmetros, consequentemente suas estimativas. Entretanto, a partir da Figura 2.5 notamos que as estimativas $\hat{a}$ e $\hat{b}$ representam bem a função de sobrevivência, além de que estão próximas da estimativa de Kaplan-Meier (para mais detalhes das estimativas no caso com censura, ver Apêndice B).

Tabela 2.4.3: Estimativas para o caso com censura.

Tamanho Amostral	$\hat{a}$	$\hat{b}$	Percentual de Censura
25	5,103	2,735	8%
50	1,986	3,676	22%
100	0,620	4,738	15%
200	1,212	5,005	19%
500	1,002	4,669	19,4%

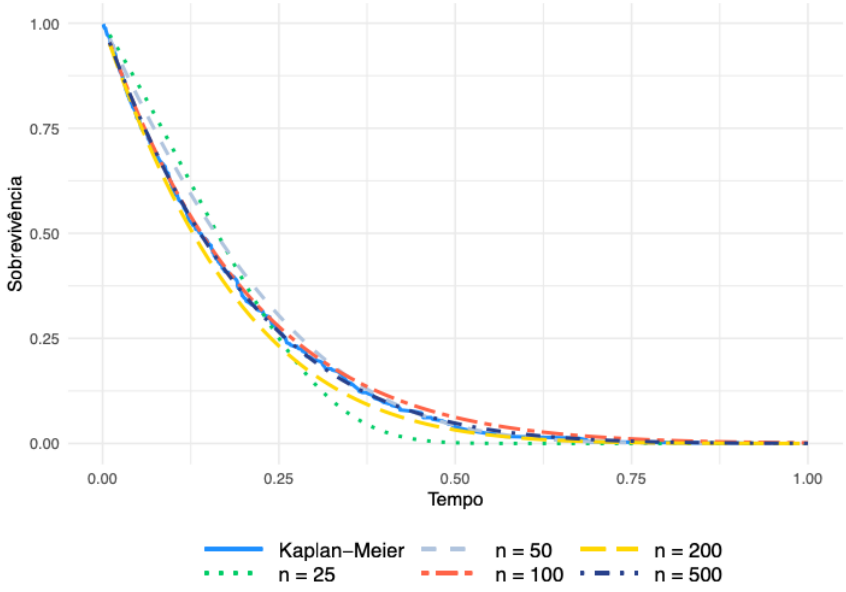


Figura 2.5: Curvas de sobrevivência estimadas para diversos tamanhos amostrais.

Capítulo 3

Modelos Com Fração de Cura

Neste capítulo descrevemos a ideia geral dos modelos de fração de cura. Na Seção 3.1 tratamos da definição de "cura" no contexto de sobrevivência e suas implicações. Nas seções 3.2 e 3.3 definimos os modelos de cura com mistura padrão e "defeituosos" e, como se dá a estimação dos parâmetros em ambos os casos, respectivamente.

3.1 Definição do Conceito de Cura

O conceito de cura em Análise de Sobrevivência não é interpretado literalmente, no sentido de, por exemplo, algum paciente estar doente e ser curado de alguma moléstia. O conceito de "cura" diz respeito a indivíduos no estudo que nunca sofrerão o evento de interesse, denominadas "curadas". A presença de um grupo de indivíduos nesta condição geralmente indica alguma heterogeneidade na população estudada e, os modelos que incorporam esta informação são chamados de modelos com fração de cura (Othus *et al.*, 2012).

Partindo do pressuposto que os tempos analisados contém fração de cura, se faz necessário classificarmos os pacientes existentes em dois grupos:

- os que sofreram o evento, considerados como vulneráveis, não curados, imunes ou suscetíveis;
- os que nunca sofreram o evento (no período de estudo), considerados não vulneráveis, curados, não imunes ou não suscetíveis.

Essa distinção cria conceitos e estende alguns dos modelos mais tradicionais de sobrevivência, o que justifica a utilização de uma abordagem especial. Mais especificamente, se δ_i

$= 1$, sabemos que o indivíduo é vulnerável ao evento, porém se $\delta_i = 0$, há incerteza sobre qual grupo devemos inserir o paciente i $(i = 1, \ldots, n)$ (Amico e Van Keilegom, 2018).

Considerando T como o tempo até o evento acontecer, se o paciente pertence ao grupo de não-curados, ele sofre o evento durante o período observado, isto é, $P(T < \infty) = 1$. Agora, se o mesmo pertence ao grupo de "curados", então nunca observaremos o evento no paciente, o que corresponde a assumir que $T = \infty$. A implicação disso é que a função de sobrevivência passa a ser imprópria. Em outras palavras

$$\lim_{t \to \infty} S(t|.) > 0. \tag{3.1}$$

A partir de (3.1), observamos que conforme o tempo avança, ainda haverá indivíduos que nunca sofrerão o evento de interesse.

Essas mudanças no comportamento de $S(t|.)$ motivam a aplicação e desenvolvimento de modelos que consideram a existência de uma proporção curada. Comparados a modelos tradicionais da área de sobrevivência, os modelos de fração de cura, podem, por exemplo, auxiliar a identificação de características associadas tanto com os "não imunes", como os "imunes", sendo uma alternativa que possui algumas vantagens. Em especial, quando as curvas da função de sobrevivência populacional não atingem o valor 0, isto é um indicativo de alguma heterocedasticidade na população sob estudo. Segundo Othus *et al.* (2012), descrever e explicitar esse problema é útil e possível na abordagem via fração de cura com mistura.

A identificação da existência da fração de cura pode ser feita através do estimador de Kaplan-Meier (2.5). Se existir um limitante inferior no valor (referente ao último tempo censurado) desta estimativa, é razoável supor a existência de uma proporção de curados. Entre as alternativas de verificação da existência de fração de cura, há um teste de hipótese proposto por Maller e Zhou (1996). Embora a existência de um teste de hipótese seja atrativa, é difícil estender a formulação para um contexto geral de identificação de curados, em especial, este teste só é válido quando o tempo até ocorrência segue algumas distribuições paramétricas específicas, como a distribuição exponencial (Amico e Van Keilegom, 2018).

A título de exemplo, vamos considerar o conjunto de dados provenientes do estudo TGCA (*The Genome Cancer Atlas*). Esse conjunto contém 1050 pacientes de câncer de mama, em que 146 dos mesmos vieram a sofrer o evento de interesse (falecimento).

Os dados se encontram disponíveis em https://portal.gdc.cancer.gov/repository (acesso em 11/01/2022).

Na Figura 3.1 reportamos a estimativa de Kaplan-Meier para $S(t)$ (para mais detalhes, ver Apêndice B). Notamos que, mesmo após 15 anos de acompanhamento, há um grupo de pacientes que não sofreram o evento de interesse, com a curva estimada atingindo um valor mínimo, de $\hat{S}(t) = 0,352$, não apresentando sinais de decrescimento inferior a esse ponto. Portanto, supomos a existência de fração de curados nesse conjunto de dados.

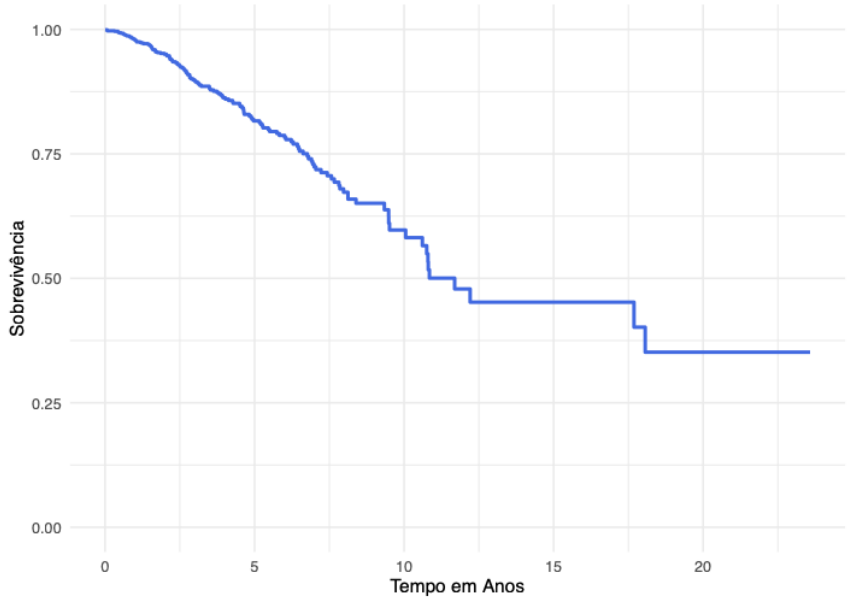


Figura 3.1: Estimador de Kaplan-Meier para pacientes com câncer de mama.

A seguir descrevemos duas classes de modelos que incorporam informação relacionada à presença de imunes, os modelos de fração de cura com mistura na Seção 3.2 e os modelos defeituosos na Seção 3.3.

3.2 Modelos de Fração de Cura com Mistura Padrão

A família mais antiga de modelos de fração cura, que segue a linha de pensamento de Boag (1949) são os modelos com mistura, nos quais se assume que a população de interesse é uma mistura entre a sub-população curada e sub-população não curada (Amico e Van Keilegom, 2018). Denotando como C a sub-população curada, $\bar{C}$ a sub-população não-curada, π a proporção de curados e $1 - \pi$ a proporção de não curados, em termos

formais temos

$$C = \begin{cases} 1, \text{ se o indivíduo é considerado curado,} \\ 0, \text{ c.c.} \end{cases}$$

e $\bar{C}$ podendo ser definida como $1-C$. A partir disso, escrevemos a função de sobrevivência para toda a população

$$\begin{aligned} S_{pop}(t|\boldsymbol{\theta}) &= P(T > t|\boldsymbol{\theta}) \\ &= P(T > t|\boldsymbol{\theta}, C = 1)P(C = 1) + P(T > t|\boldsymbol{\theta}, \bar{C} = 1)P(\bar{C} = 1) \\ &= 1\pi + S^*(t|\boldsymbol{\theta})(1 - \pi), \end{aligned}$$

portanto, neste modelo, a expressão da função de sobrevivência, $S_{pop}(t)$, para a população toda, é dada por

$$S_{pop}(t|\boldsymbol{\theta}) = \pi + (1 - \pi)S^*(t|\boldsymbol{\theta}), \quad t > 0, \ \pi \in (0, 1), \tag{3.2}$$

em que $S^*(t|.)$ representa a função de sobrevivência associada à parte da população que classificamos suscetível ao evento (Ibrahim *et al.*, 2001). Vale ressaltar que a função de sobrevivência descrita em (3.2) é imprópria para qualquer valor de π maior do que 0. Desse modo, se tomarmos o limite de $S_{pop}(t|\boldsymbol{\theta})$, quando $t \to \infty$, temos

$$\lim_{t \to \infty} S_{pop}(t|\boldsymbol{\theta}) = \pi,$$

em que π é a fração de curados.

Partindo da formulação (3.2), obtemos à função densidade para o tempo de vida populacional $f_{pop}(t)$, que é escrita como

$$f_{pop}(t|\boldsymbol{\theta}) = \frac{dS_{pop}(t|\boldsymbol{\theta})}{dt} = (1 - \pi)f^*(t|\boldsymbol{\theta}), \tag{3.3}$$

em que $S_{pop}(t|\boldsymbol{\theta})$ é uma função de sobrevivência imprópria e $f^*(t|\boldsymbol{\theta})$ representa a função densidade (própria) de probabilidade do tempo de vida para os indivíduos suscetíveis.

Ressaltamos que neste trabalho, as funções $f^*(.)$ e $S^*(.)$ são a função densidade de probabilidade e função do sobrevivência da distribuição de Gompertz, dadas em (2.11) e (2.13), respectivamente.

Fazendo uso de (3.3), a contribuição de cada indivíduo $i, i = 1, \ldots, n$ para a função

de verossimilhança no modelo de mistura é,

$$L_i(\boldsymbol{\theta}|.) = (1-\pi)^{\delta_i}[f^*(t|\boldsymbol{\theta})]^{\delta_i}[\pi + (1-\pi)S^*(t_i|\boldsymbol{\theta})]^{1-\delta_i}.$$

Por fim, a função de verossimilhança para esse modelo é dada por,

$$L(\boldsymbol{\theta}|.) = \prod_{i=1}^{n} \left\{(1-\pi)^{\delta_i}[f^*(t_i|\boldsymbol{\theta})]^{\delta_i}[\pi + (1-\pi)S^*(t_i|\boldsymbol{\theta})]^{1-\delta_i}\right\} \tag{3.4}$$

e a função de log-verossimilhança (na qual usamos a função ln), geralmente utilizada para melhorar a estabilidade dos métodos numéricos de maximização, é escrita como

$$l(\boldsymbol{\theta}|.) = \sum_{i=1}^{n} \left\{\delta_i \ln(1-\pi)\delta_i \ln(f^*(t|\boldsymbol{\theta})) + (1-\delta_i)\left[\ln(\pi + (1-\pi)S^*(t_i|\boldsymbol{\theta}))\right]\right\}. \tag{3.5}$$

A partir das funções (3.4) ou (3.5) obtemos as estimativas de máxima verossimilhança. Ressaltamos que a outra classe mais conhecida de modelos de fração de cura (Ibrahim *et al.*, 2001), não abordada neste trabalho, são os modelos sem mistura, que modelam a função de sobrevivência como

$$S_{pop}(t|\boldsymbol{\theta}) = \pi^{F^*(t|\boldsymbol{\theta})} = e^{F^*(t|\boldsymbol{\theta})\ln\pi}, \;\; t > 0, \; \pi \in (0,1),$$

na qual $F^*(t|\boldsymbol{\theta})$ é uma função distribuição própria, utilizada para modelar os indivíduos suscetíveis ao evento de interesse.

3.3 Modelos Defeituosos

Como dito anteriormente, se na modelagem de algum conjunto de dados detectarmos presença de fração de curados, a função de sobrevivência $S(t|.)$ não atinge o valor de zero. A partir da relação entre a função de risco acumulado e a função de sobrevivência, $H(t|.) = -\ln(S(t|.)$, se a sobrevivência é limitada inferiormente, o risco acumulado é limitado superiormente. Por consequência, o risco instantâneo $h(t|.)$ nesse contexto se aproxima de 0 (Gieser *et al.*, 1998).

A ideia dos modelos defeituosos (do inglês - *defective models*) é utilizar de distribuições "naturalmente" impróprias como a Gompertz (Gieser *et al.*, 1998) ou a Gaussiana Inversa (Balka *et al.*, 2009), entre outras, que possibilitem a função de risco instantânea

se aproximar de 0 a partir de algum ponto.

Na Figura 3.2 mostramos o comportamento das funções de sobrevivência, risco e risco acumulado em dados com fração de cura. Observamos que nesse caso, $S(t|.)$ não atinge o valor de zero e que, consequentemente, a função de risco acumulado tem um valor máximo de $0,5$. Já a função de risco instantâneo cai rapidamente para 0.

Essencialmente, o raciocínio por trás dos modelos defeituosos Gompertz, como discutido por Rocha *et al.* (2014), é utilizar a própria densidade mostrada em (2.11). Nesta, as propriedades supramencionadas são validas, com a função de sobrevivência sendo limitada inferiormente em algum valor π - proporção de curados entre 0 e 1, quando o parâmetro a é negativo.

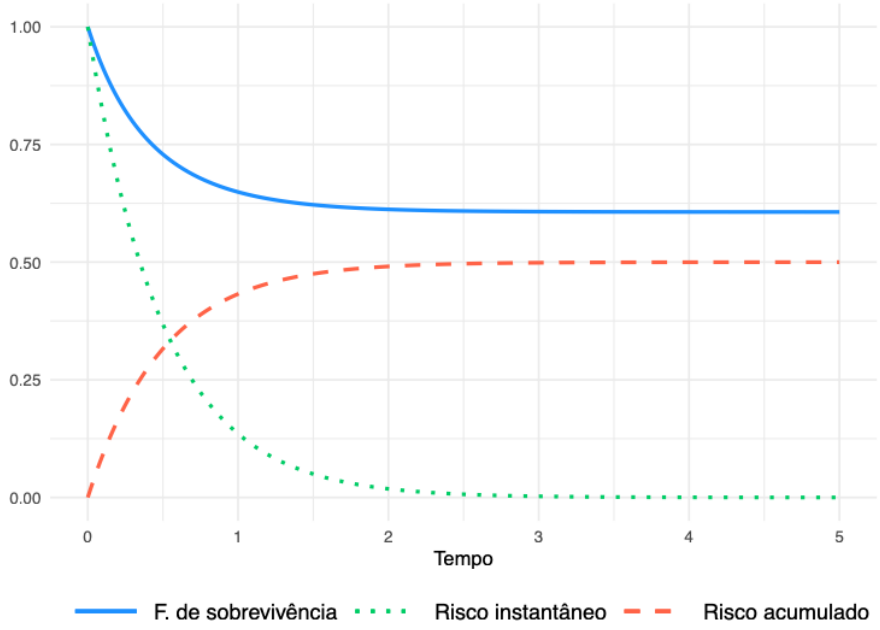


Figura 3.2: Relação entre a função de sobrevivência e o risco em modelos com fração de cura.

Em suma, no caso da distribuição Gompertz, o modelo paramétrico padrão já incorpora naturalmente a existência de uma fração de curados, sem necessidade de ajustes ou suposições adicionais. Podemos simplesmente modelar o problema e se a for negativo, temos evidências de fração de curados.

Ressaltamos que o parâmetro π não é estimado diretamente no modelo defeituoso, sendo calculado de acordo com uma função dos parâmetros

$$\hat{\pi} = \exp(\hat{b}/\hat{a}).$$

Para obtermos a variância deste estimador fizemos o uso do método Delta (Casella e Berger, 2021). Dado que $f(\hat{b}, \hat{a}) = e^{\hat{b}/\hat{a}}$ é uma função derivável, a variância desta função pode ser aproximada por

$$Var(f(\hat{a}, \hat{b})) \approx \frac{1}{n} \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{\delta f}{\delta \hat{a}} \frac{\delta f}{\delta \hat{b}} \sigma_{ij}. \tag{3.6}$$

3.4 Estimação Frequentista dos Parâmetros

Independente do tipo de modelo escolhido, a estimativa do vetor de parâmetros $\boldsymbol{\theta}$ no caso frequentista é realizada pela maximização da função de log-verossimilhança $l(\boldsymbol{\theta}|.)$ em (2.16) ou (3.5). Buscamos $\hat{\boldsymbol{\theta}}$ de tal forma que

$$(\hat{a}, \hat{b}, \hat{\pi}) = \hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|.).$$

No caso do modelo Gompertz,

$$(\hat{a}, \hat{b}) = \hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|.),$$

Para obtermos as estimativas paramétricas é necessário o uso de métodos numéricos implementados em alguma rotina de maximização de funções, como a função `optim` da linguagem `R` (R Core Team, 2021).

Para comparar os modelos supracitados, cuja estimação é descrita sob enfoque frequentista, fazemos uso do critério AIC (do inglês - *Akaike Information Criterion*). Essa métrica utilizada na comparação de modelos é definida, como visto em Paula (2004), da seguinte forma

$$AIC(\boldsymbol{\theta}|.) = -2\ln(L(\boldsymbol{\theta})|.) + 2k, \tag{3.7}$$

em que k é a quantidade de parâmetros e $L(\boldsymbol{\theta}|.)$ é o valor função de verossimilhança apropriada, dada em (2.16) - modelo Gompertz ou em (3.5) - modelo de mistura padrão. Em geral, selecionamos o modelo mais parcimonioso, que é aquele com menor AIC. No próximo capítulo, descrevemos a estimação no contexto Bayesiano.

Capítulo 4

Estimação Bayesiana dos Parâmetros

Neste capítulo inicialmente descrevemos os conceitos fundamentais para a estimação Bayesiana dos parâmetros e fazemos uma breve revisão dos métodos de simulação Metropolis-Hastings e amostrador de Gibbs. Por fim, damos maior atenção ao método utilizado, que é o Monte Carlo Hamiltoniano.

4.1 Conceitos Iniciais

Para a estimação dos parâmetros via métodos Bayesianos, precisamos especificar a distribuição conjunta *a priori* do vetor de parâmetros, $p(\boldsymbol{\theta})$, envolvido nos modelos e a distribuição amostral dos dados. Desta forma, para obter a distribuição *a posteriori* $p(\boldsymbol{\theta}|\boldsymbol{t},\boldsymbol{\delta})$, pelo teorema de Bayes (Box e Tiao, 1992), obtendo

$$p(\boldsymbol{\theta}|\boldsymbol{t},\boldsymbol{\delta}) = \frac{p(\boldsymbol{\theta},\boldsymbol{t},\boldsymbol{\delta})}{p(\boldsymbol{t},\boldsymbol{\delta})} = \frac{p(\boldsymbol{\theta})p(\boldsymbol{t},\boldsymbol{\delta}|\boldsymbol{\theta})}{p(\boldsymbol{t},\boldsymbol{\delta})}, \tag{4.1}$$

em que $p(\boldsymbol{t},\boldsymbol{\delta}|\boldsymbol{\theta})$ é a distribuição amostral dos dados. Neste trabalho, estas distribuições estão definidas em (2.11). Considerando que o denominador $p(\boldsymbol{t},\boldsymbol{\delta})$ é constante com relação ao vetor de parâmetros $\boldsymbol{\theta}$, é possível omiti-lo, reescrevendo a expressão (4.1) como

$$p(\boldsymbol{\theta}|\boldsymbol{t},\boldsymbol{\delta}) \propto p(\boldsymbol{\theta})p(\boldsymbol{t},\boldsymbol{\delta}|\boldsymbol{\theta}), \tag{4.2}$$

sendo esta uma forma mais simples de escrever a distribuição *a posteriori* do vetor $\boldsymbol{\theta}$ de parâmetros. Vale ressaltar que ao utilizarmos algoritmos de simulação, usualmente

trabalhamos com o logaritmo natural da densidade *a posteriori* (4.2),

$$\ln p(\boldsymbol{\theta}|\boldsymbol{t},\boldsymbol{\delta}) \propto \ln p(\boldsymbol{\theta}) + \ln p(\boldsymbol{t},\boldsymbol{\delta}|\boldsymbol{\theta}),$$

em que $\ln p(\boldsymbol{\theta})$ e $\ln p(\boldsymbol{t},\boldsymbol{\delta}|\boldsymbol{\theta})$ é o logaritmo natural da densidade conjunta *a priori* dos parâmetros e da função de verossimilhança do modelo em questão, respectivamente. No contexto Bayesiano de estimação, realizar inferência sobre o conjunto de parâmetros $\boldsymbol{\theta}$ equivale a amostrar da distribuição *a posteriori* $p(\boldsymbol{\theta}|\boldsymbol{t},\boldsymbol{\delta})$.

Geralmente, para realizar a inferência do ponto de vista Bayesiano, faz-se necessário a utilização dos métodos de *Markov chain Monte Carlo* (MCMC) para amostrar da distribuição *a posteriori* (Gamerman e Lopes, 2006). O primeiro componente do MCMC é as Cadeia de Markov (*Markov chain*), que é uma sequência de eventos cujo valor sempre depende de uma quantidade fixa de eventos anteriores. Já o segundo é o método de Monte Carlo, utilizado para resolver problemas teóricos complexos a partir de simulações computacionais. Essas simulações são aplicações em diversos problemas como gerar números aleatórios, calcular valores-p em testes de permutação, integrais numéricas etc.

Em suma, o MCMC busca construir cadeias de Markov, que são sequências de valores geradas via Monte Carlo. Essas sequências devem convergir para a distribuição *a posteriori* do vetor de parâmetros $\boldsymbol{\theta}$. Na Seção 4.2, descrevemos alguns desses métodos.

4.2 Métodos de *Markov Chain Monte Carlo*

4.2.1 Metropolis-Hastings

O algoritmo de Metropolis-Hastings é uma generalização do algoritmo de Metropolis e é bastante utilizado para amostrar de distribuições *a posteriori*, especialmente em problemas de baixa dimensão e quando a formulação das distribuições marginais é complexa e, por vezes, sem forma analítica. Esse algoritmo é essencialmente uma adaptação do passeio aleatório com regras de aceitação/rejeição dos valores simulados, de modo a (tentar) obter convergência para a distribuição de interesse (Gelman *et al.*, 2013).

Considerando que queremos amostrar de $p(\boldsymbol{\theta}|\boldsymbol{t},\boldsymbol{\delta})$ como descrito em (4.2), segue uma breve descrição do algoritmo:

1. Iniciar a amostragem de um ponto de partida $\boldsymbol{\theta}^{(0)}$, tal que $p(\boldsymbol{\theta}^{(0)}|\boldsymbol{t},\boldsymbol{\delta}) > 0$.

2. Para as iterações $i = 1, 2, \ldots,$

   (a) especificar uma distribuição de propostas, $J_i(\boldsymbol{\theta}^{(*)}|\boldsymbol{\theta}^{(i-1)})$, e desta amostrar novos valores para os parâmetros, $\boldsymbol{\theta}^{(*)}$,

   (b) calcular a razão
   $$r = \frac{p(\boldsymbol{\theta}^{(*)}|\boldsymbol{t}, \boldsymbol{\delta})/J_i(\boldsymbol{\theta}^{(*)}|\boldsymbol{\theta}^{(i-1)})}{p(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{t}, \boldsymbol{\delta})/J_i(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}^{(*)})},$$

   (c) comparar a razão r com um valor u gerado da distribuição uniforme entre 0 e 1. Atualizar a cadeia gerada conforme a seguinte condição

$$\boldsymbol{\theta}^{(i)} = \begin{cases} \boldsymbol{\theta}^{(*)}, \text{ se } r < u, \\ \boldsymbol{\theta}^{(i-1)}, \text{ c. c.} \end{cases}$$

A estrutura do algoritmo é relativamente simples, porém o custo computacional até a convergência pode ser elevado a depender da complexidade do problema e do número de parâmetros. Além disso, como este algoritmo gera valores aleatórios, pode levar considerável tempo para a cadeia atingir as regiões de maior densidade *a posteriori* (Thomas e Tu, 2021). Mais detalhes sobre esse método podem ser encontrados em Chib e Greenberg (1995).

4.2.2 Amostrador de Gibbs

Um caso especial do algoritmo de Metropolis-Hastings, o amostrador de Gibbs, é uma técnica mais apropriada para simular de distribuições multivariadas, quando a simulação direta é bastante difícil, porém as condicionais completas estão disponíveis. O raciocínio é evitar a computação direta da distribuição, calculando as distribuições marginais que geralmente são mais simples (Casella e George, 1992).

Suponha que $p(\boldsymbol{\theta}|\boldsymbol{t}, \boldsymbol{\delta})$ (como em (4.2) é uma distribuição cuja computação direta é difícil, em que $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$. A distribuição marginal, por exemplo, para θ_1 é definida como

$$p(\theta_1|\theta_2, \theta_3, \ldots, \theta_k, \boldsymbol{t}, \boldsymbol{\delta}) = \int_{\theta_k} \cdots \int_{\theta_3} \int_{\theta_2} p(\theta_1, \theta_2, \theta_3, \ldots, \theta_k, \boldsymbol{t}, \boldsymbol{\delta})\, d\theta_2\, d\theta_3 \ldots d\theta_k,$$

considerando o espaço paramétrico de cada um dos parâmetros. Alternativamente, trabalhar com a função densidade marginal equivale a fixar todos os outros parâmetros de

interesse e trabalhar somente com θ_1 sendo variável aleatória. Essencialmente, estamos transformando um problema multidimensional para $k-$problemas unidimensionais. Segue uma breve descrição do algoritmo:

1. Começar a amostragem de um ponto de partida $\boldsymbol{\theta}^{(0)}$, tal que $p(\boldsymbol{\theta}^{(0)}|\boldsymbol{t}, \boldsymbol{\delta}) > 0$.

2. Para as iterações $i = 1, 2, \ldots,$

   (a) Iniciar com um dos parâmetros, denotado como θ_j de $\theta_1, \theta_2, \ldots, \theta_k$ para atualizar.

   (b) Gerar uma proposta para θ_j^* de $p(\theta_j^{(*)}|\theta_j^{(i-1)})$. A seguir, atualizamos o vetor de parâmetros $\boldsymbol{\theta}$ apenas na coordenada j, com $j = 1, 2, \ldots, k$.

   (c) A iteração só termina quando todos os parâmetros forem atualizados.

Este algoritmo, assim como o de Metropolis-Hastings é relativamente simples, possuindo algumas implementações conhecidas, como os *softwares* da família BUGS (sigla do inglês - *Bayesian inference Using Gibbs Sampling*) como descrito por Lunn *et al.* (2000). Algumas de suas limitações são que o processo até a convergência pode demorar e, em situações cuja dimensão de $\boldsymbol{\theta}$ é grande, pode ser necessário um número impraticável de iterações até convergência (Thomas e Tu, 2021).

4.2.3 Monte Carlo Hamiltoniano

O algoritmo Hamiltoniano surgiu como uma alternativa aos métodos mais conhecidos de simulação - Metropolis-Hastings e amostrador de Gibbs. Esses métodos, embora mais simples tanto a nível teórico quanto de implementação, podem apresentar comportamento de passeio aleatório por muitas iterações, tornando o processo de estimação e inferência demorado (Gelman *et al.*, 2013).

A ideia básica do método de simulação de Monte Carlo Hamiltoniano (MCH) é fazer uso de gradientes e de algumas propriedades da física Hamiltoniana de modo a acelerar a convergência das cadeias simuladas (Thomas e Tu, 2021). Para isso um novo vetor de variáveis $\boldsymbol{\rho}$, com mesma dimensão de $\boldsymbol{\theta}$, denominado como momento é introduzido no espaço paramétrico, de modo que a distribuição *a posteriori* é agora dada por $p(\boldsymbol{\rho}, \boldsymbol{\theta}) = p(\boldsymbol{\rho}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Na mecânica Hamiltoniana o momento é utilizado para movimentar as quantidades no espaço em questão. Uma descrição mais detalhada é dada por Betancourt e Girolami (2015).

Como estamos interessados somente no vetor de parâmetros $\boldsymbol{\theta}$, o vetor de momento $\boldsymbol{\rho}$ é uma quantidade auxiliar, cujos valores mudam de iteração para iteração.

Segundo Gelman *et al.* (2013), o MCH pode ser considerado um método de simulação híbrido, devido à combinação de componentes determinísticos com métodos *Markov chain Monte Carlo*. A adição de parâmetros aumenta a complexidade do processo de simulação e consequentemente o custo computacional é maior por iteração. Entretanto, a convergência mais rápida e robustez do método faz com que o Monte Carlo Hamiltoniano seja uma opção atrativa em aplicações Bayesianas. Existem algumas implementações desse método no *software* `R`, a mais conhecida é o `Rstan`.

Na Figura 4.1 apresentamos uma ilustração simples dessa técnica. Considerando uma curva suave e desprezando atrito, na ilustração (a), aplicamos uma força de intensidade e direção aleatória na partícula (a força é o momento e a partícula representa os parâmetros). Essa força move a partícula subindo a curva para direita. Na ilustração (b), a partícula subiu o máximo de acordo com seu momento e começa o processo contrário, de ir para a região esquerda da curva. Por fim nas ilustrações (c) e (d) o objeto continua o caminho para a região à esquerda da superfície. Como supomos que não existe atrito, não há esgotamento de momento (que é definido aleatoriamente no algoritmo). É essencialmente desse modo que o Monte Carlo Hamiltoniano explora o espaço paramétrico, com cada iteração representando um valor do "caminho" percorrido dentro deste espaço. As equações que descrevem esse movimento são chamadas de equações Hamiltonianas.

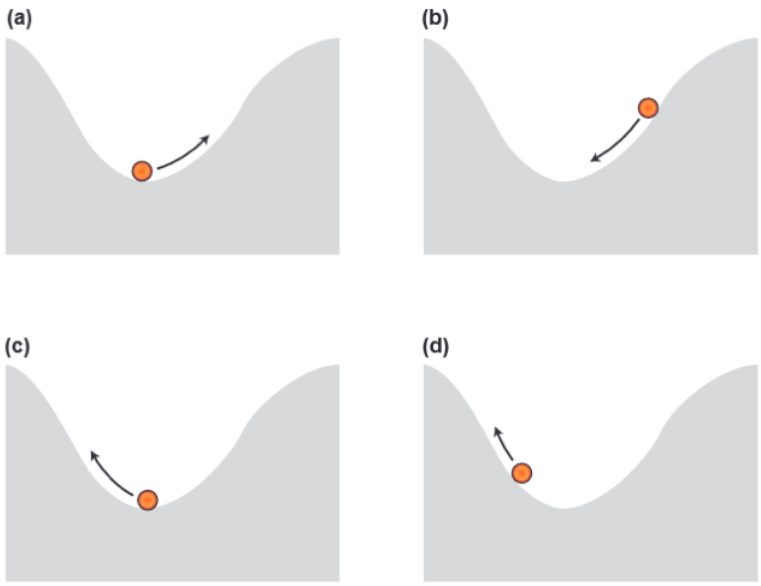


Figura 4.1: Ilustração do Monte Carlo Hamiltoniano. Fonte: Thomas e Tu (2021).

A seguir, descrevemos sucintamente como o método opera, baseado no manual de usuário do `STAN` (Stan Development Team, 2022).

Assim como os algoritmos de Metropolis-Hastings, o MCH também faz uso de um conjunto de parâmetros iniciais $\boldsymbol{\theta}^{(0)}$. Introduzindo a variável auxiliar de momento $\boldsymbol{\rho}$,

$$p(\boldsymbol{\rho}, \boldsymbol{\theta}) = p(\boldsymbol{\rho}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{4.3}$$

geralmente assumimos que $\boldsymbol{\rho}$ tem uma distribuição de probabilidade independente de $\boldsymbol{\theta}$. A distribuição de $\boldsymbol{\rho}$ é comumente considerada pouco importante, sendo comum utilizar a distribuição normal multivariada para tal

$$\boldsymbol{\rho} \sim N_k(0, \boldsymbol{V}),$$

em que $\boldsymbol{V}$ é uma transformação do espaço paramétrico, baseada na covariância estimada no período de *warm-up* (período de "aquecimento", que tem como objetivo levar a cadeia até um estado próximo da convergência, antes da amostragem definitiva) da simulação.

Podemos reescrever (4.3) como uma equação diferencial Hamiltoniana (HA),

$$\begin{aligned} HA(\boldsymbol{\rho}, \boldsymbol{\theta}) &= -\ln p(\boldsymbol{\rho}, \boldsymbol{\theta}) \\ &= -\ln p(\boldsymbol{\rho}|\boldsymbol{\theta}) - \ln p(\boldsymbol{\theta}) \\ &= K(\boldsymbol{\rho}, \boldsymbol{\theta}) + V(\boldsymbol{\theta}) \end{aligned}$$

e, ressaltamos o fato que $\ln p(\boldsymbol{\theta})$ é o logaritmo natural da densidade *a posteriori* dos parâmetros, que é a quantidade que temos interesse em estimar.

Transições são geradas pelo algoritmo dentro do espaço de parâmetros de modo que os valores gerados se encontram na região de maior densidade (o máximo da densidade *a posteriori*). Uma breve descrição do algoritmo é:

1. para cada iteração $i = 1, 2, \ldots,$
2. gerar um novo momento da distribuição de $\boldsymbol{\rho}^i$ e construir o sistema de equações diferencias Hamiltoniano,

$$\begin{aligned} \frac{-d\boldsymbol{\theta}}{dk} &= \frac{-dK(\boldsymbol{\rho}, \boldsymbol{\theta})}{\delta \boldsymbol{\rho}} \\ \frac{-d\boldsymbol{\rho}}{dk} &= -\frac{\delta V(\boldsymbol{\theta})}{d\boldsymbol{\theta}}. \end{aligned} \tag{4.4}$$

3. a resolução da equação diferencial (4.4) é feita através de métodos numéricos. O método comumente utilizado no MCH é o *leapfrog integrator*, que foi desenvolvido especialmente para garantir resultados estáveis para equações Hamiltonianas. Mais detalhes sobre essa técnica de integração, podem ser conferidos em Leimkuhler e Reich (2004),

4. por fim, uma nova quantidade para o vetor de parâmetros expandido $p(\boldsymbol{\rho}^*, \boldsymbol{\theta}^*)$ é gerada. Assim como no método de Metropolis-Hastings, há um critério de aceitação da proposta, no qual a probabilidade de aceitação é dada por

$$\min(1, \exp(HA(\boldsymbol{\rho}, \boldsymbol{\theta})) - \exp(HA(\boldsymbol{\rho}^*, \boldsymbol{\theta}^*))).$$

   Se a proposta não for aceita, os parâmetros da iteração anterior são repetidos e utilizados para uma nova tentativa.

Alguns detalhes da mecânica Hamiltoniana do ponto de vista físico e da resolução do sistema diferencial foram omitidos, e podem ser consultados em Gelman *et al.* (2013) e em Betancourt e Girolami (2015).

No contexto Bayesiano, para seleção de modelos, um critério disponível é o DIC (do inglês - *Deviance Information Criteria*), que é definido como

$$DIC(\boldsymbol{\theta}|.) = -2l(\boldsymbol{\theta}|.) - 2p_{DIC},$$

em que

$$2p_{DIC} = 2(l(\boldsymbol{\theta}|.) - E_{posteriori}[l(\boldsymbol{\theta}|.)]).$$

Neste critério, $2p_{DIC}$ é chamado de número efetivo de parâmetros do modelo. Mais informações sobre este critério podem ser encontrados no artigo de Spiegelhalter *et al.* (2002).

Capítulo 5

Aplicações

Neste capítulo ajustamos os modelos discutidos em cinco conjuntos de dados, referentes a câncer de mama, de ovário, de pele melanoma e de cólon e, dados de leucemia. Mais detalhes sobre as análises são dados nas seções específicas para cada conjunto. Para todas a situações problema, estimamos a função de sobrevivência considerando a distribuição Gompertz em dois circunstâncias diferentes, como escrito no Capítulo 3, assumindo

- existência de porcentagem de curados e estimando-a com o modelo de mistura padrão;
- modelagem a partir do modelo Gompertz, cuja fração de cura é estimada naturalmente;

Estimamos sob os enfoques frequentista e Bayesiano os parâmetros de cada modelo, considerando as especificidades em cada caso e, utilizamos os critérios de AIC e DIC para selecionar o melhor modelo. Por fim, nas seções 5.1 a 5.5, apresentamos as aplicações e discutimos os pormenores dos métodos e resultados.

5.1 Aplicação para Câncer de Mama

Retomando o conjunto de dados extraídos do estudo *TGCA - The Cancer Genome Atlas* (Seção 3.1), apresentamos as estimativas frequentistas e Bayesianas. Começando com os estimadores de máxima verossimilhança, na Tabela 5.1.1 reportamos as estimativas para os parâmetros, seus respectivos intervalos de confiança, para cada uma das dois situações e também a métrica AIC.

Tabela 5.1.1: Resultados (frequentistas) para o conjunto de dados TGCA.

Modelo	Parâmetro	Estimativa	2,5%	97,5%	AIC
Gompertz	a	0,0653	0,0274	0,1031	
	b	0,0318	0,0246	0,0390	1223,451
	π	Não há	-	-	
Mistura padrão Gompertz	a	0,2064	0,1060	0,1944	
	b	0,0391	0,0261	0,0522	1215,403
	π	0,3352	0,1941	0,4761	

A partir da Figura 5.1a observamos que o ajuste modelo Gompertz defeituoso (2.11) não captou a fração de curados, enquanto o modelo com mistura padrão sim. Para realização das estimativas sob enfoque Bayesiano, especificamos as distribuições *a priori* dos parâmetros, nos quais supomos independência. Seguem as distribuições *a priori*

- Modelo Gompertz:

$$a \sim Normal(-0,5;1) \text{ e}$$
$$b \sim Gama(1,25;1).$$

- Modelo de Mistura padrão Gompertz:

$$a \sim Normal(0,2;1),$$
$$b \sim Gama(0,05;2) \text{ e}$$
$$\pi \sim Uniforme(0;1).$$

Para os dois modelos, a distribuição $Gama(\alpha, \beta)$ utilizada tem função densidade de probabilidade dada por,

$$f(x|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \text{ para } x,\ \alpha,\ \beta > 0,$$

em que X é a variável aleatória, α é parâmetro de forma e β de taxa.

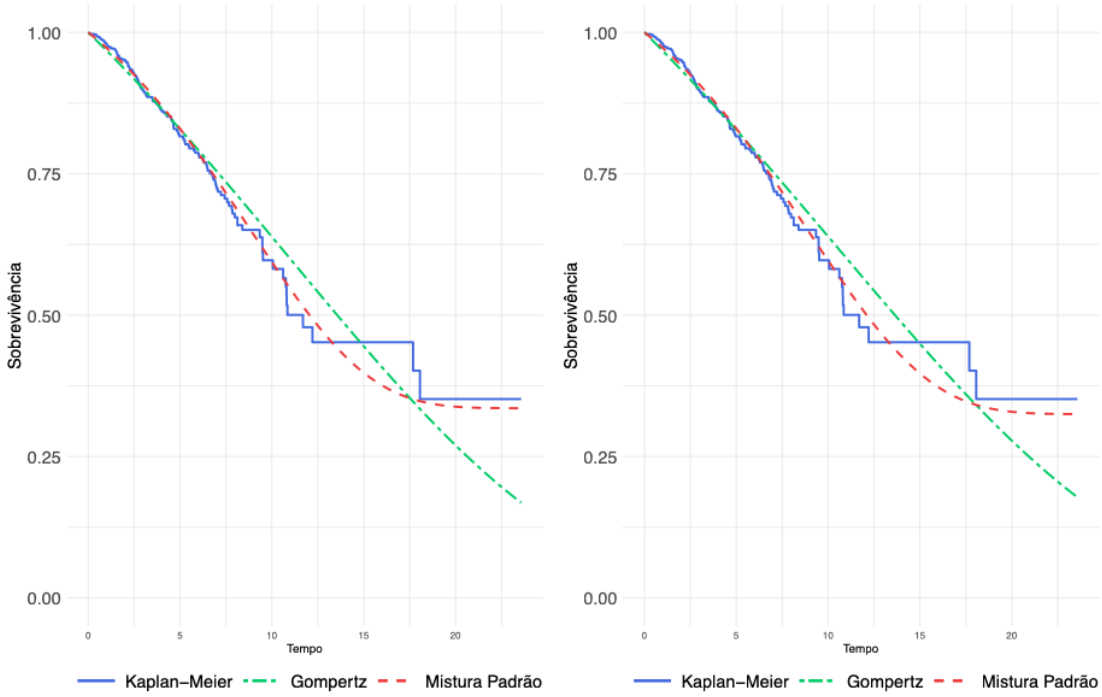
A partir das densidades *a posteriori*, simulamos duas cadeias distintas que são permutadas para diminuir os impactos da autocorrelação de valores gerados. Os valores simulados foram retirados de duas cadeias distintas, cada uma com período de *burn-in* ou aquecimento de 1000 valores e 5000 iterações consideradas para a inferência, totalizando 10000 valores amostrados. Na Tabela 5.1.2 apresentamos as estimativas pontuais (medi-

anas), intervalos de credibilidade para os parâmetros de cada modelo e o valor do critério DIC.

Tabela 5.1.2: Resultados (Bayesianos) para o conjunto de dados TGCA.

Modelo	Parâmetro	Estimativa	2,5%	97,5%	DIC
Gompertz	a	0,0624	0,0240	0,0995	
	b	0,0322	0,0256	0,0399	1223,4147
	π	Não há	-	-	
Mistura padrão Gompertz	a	0,1521	0,0915	0,2011	
	b	0,0387	0,0282	0,0540	1215,4829
	π	0,3345	0,1632	0,4588	

Observamos que novamente o modelo Mistura padrão Gompertz teve performance melhor, ou foi o mais parcimonioso, conseguindo captar a fração de cura existente. A partir da Figura 5.1b, reforçamos esta conclusão, pois o ajuste Mistura padrão Gompertz é o único que atinge o patamar apresentado pelo conjunto de dados. Desta forma, esse foi o modelo que consideramos o mais adequado e, os critérios de AIC e BIC suportam nossa afirmação. Ressaltamos que a proporção de cura estimada pelo modelo com mistura em ambos os casos Bayesianos e frequentista ficou ao redor de 33%.



(a) Estimativas frequentistas.

(b) Estimativas Bayesianas.

Figura 5.1: Estimativas para $S(t)$, segundo os dois modelos paramétricos e o Kaplan-Meier.

5.2 Aplicação para Câncer de Ovário

Nesta seção, analisamos o conjunto `Ovarian` presente na biblioteca `survival` (Therneau, 2022) da linguagem `R`. Neste conjunto as informações disponíveis são relacionadas aos tempos (anos) até falecimento de 26 mulheres com câncer de ovário, sendo que destes registros, 54% são censurados à direita. Este câncer é caracterizado por anomalias na multiplicação de células no ovário, afetando em especial mulheres mais velhas, sendo de diagnóstico difícil, devido à internalidade do problema.

Na cenário frequentista mostramos, na Tabela 5.2.1, os dois modelos ajustados (com mistura e defeituoso), os respectivos intervalos de confiança, além do valor do critério AIC. No modelo defeituoso, a estimativa da fração de curados é estimada como 16,10%, bem distante dos 49,50% estimados no modelo com mistura.

Tabela 5.2.1: Resultados (frequentistas) para o conjunto de dados sobre câncer de ovário.

Modelo	**Parâmetro**	**Estimativa**	**2,5%**	**97,5%**	**AIC**
Gompertz	a	-0,1859	-0,9373	0,5656	
	b	0,3394	0,0277	0,6512	58,2216
	π	0,1610	-0,8292	1,1511	
Mistura padrão Gompertz	a	2,0021	0,6556	3,3486	
	b	0,1979	-0,0830	0,4789	53,8681
	π	0,4951	0,2886	0,7015	

Na Figura 5.2a mostramos a curva de sobrevivência ajustada e, embora o modelo defeituoso tenha estimativa de a como negativa (existência de patamar), a curva estimada por este ficou distante da do ajuste com mistura padrão. Apenas o ajuste com mistura conseguiu acompanhar o comportamento da função de sobrevivência obtida via Kaplan-Meier.

Para obtermos as estimativas Bayesianas utilizamos *prioris* informadas com base nas estimativas de máxima verossimilhança (5.2.1). Novamente, utilizamos tamanho de cadeia simulada igual a 10000, com 2000 valores utilizados para "aquecer" a cadeia. Seguem as distribuições *a priori*

- Modelo Gompertz:

$$a \sim Normal(-0,3;1) \text{ e}$$
$$b \sim Gama(0,5;1).$$

- Modelo de mistura padrão Gompertz:

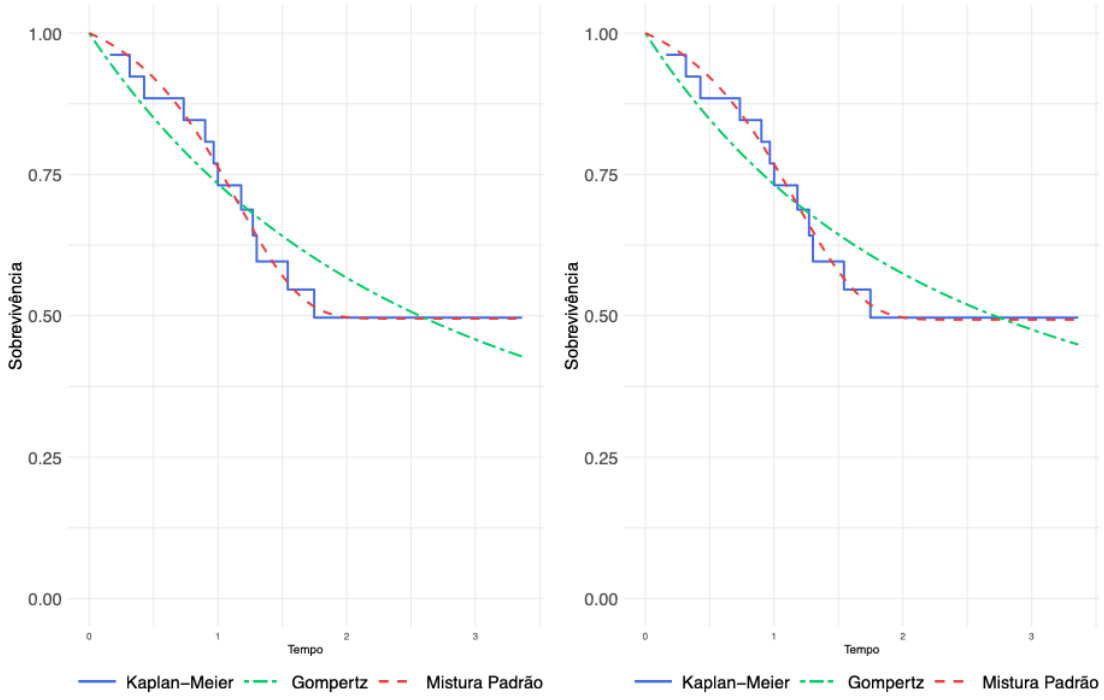
$$a \sim Normal(2; 1),$$
$$b \sim Gama(0,3; 1) \text{ e}$$
$$\pi \sim Uniforme(0; 1).$$

Sob a metodologia Bayesiana, mostramos na Tabela 5.2.2 as estimativas pontuais (mediana), intervalos de credibilidade e valor do DIC para cada modelo. Pela Figura

Tabela 5.2.2: Resultados (Bayesianos) para o conjunto de dados sobre câncer de ovário.

Modelo	**Parâmetro**	**Estimativa**	**2,5%**	**97,5%**	**DIC**
Gompertz	a	-0,2509	-1,0120	0,4260	57,9375
	b	0,3524	0,1369	0,7762	
	π	0,2454	-0,5357	1,0265	
Mistura padrão Gompertz	a	1,9379	0,7247	3,0442	53,0727
	b	0,1993	0,0567	0,5851	
	π	0,4931	0,2899	0,6805	

5.2b, observamos que as curvas estimadas pelos modelos defeituosos e com mistura estão próximas. Entretanto há diferenças consideráveis na fração de cura estimada por ambos modelos, sendo de 24,5% no modelo defeituoso, comparado a 48,6% no modelo com mistura. Vale ressaltar que pelo critério do DIC, o modelo com mistura é o mais parcimonioso.



(a) Estimativas frequentistas.

(b) Estimativas Bayesianas.

Figura 5.2: Estimativas para a função de sobrevivência, segundo os dois modelos paramétricos e o Kaplan-Meier.

5.3 Aplicação para Câncer de Pele Melanoma

Esta análise diz respeito a pacientes com câncer de pele do tipo melanoma. Esta condição ocorre quando as células melanócitos (que dão coloração a pele) se tornam cancerígenas. A lesão causada é bastante perigosa, podendo evoluir para metástase (invasão de células cancerígenas para outras regiões do corpo) e com mortalidade elevada em casos mais graves. Estes dados foram extraídos do livro de Ibrahim *et al.* (2001), cujas informações são tempo de falecimento (em anos) de 417 pacientes, apresentando 55% de censura à direita. Na Tabela 5.3.1 explicitamos as estimativas frequentistas, seus respectivos intervalos de confiança e o AIC.

Tabela 5.3.1: Resultados (frequentistas) para o conjunto de dados sobre câncer melanoma.

Modelo	Parâmetro	Estimativa	2,5%	97,5%	AIC
defeituoso	a	-0,1313	-0,2372	-0,0254	
	b	0,1792	0,1367	0,2217	1096,4674
	π	0,2555	0,0342	0,4767	
com mistura	a	0,3373	0,1954	0,4791	
	b	0,2794	0,2079	0,3507	1085,4587
	π	0,5077	0,4504	0,5650	

Pela Figura 5.3a, os modelos Gompertz e com mistura padrão acompanham o patamar da curva de sobrevivência, exibido pelo estimador de Kaplan-Meier. A diferença entre os modelos está no valor de π, o modelo com mistura estima como 50,77% e o modelo defeituoso 25,55%, diferença considerável.

Assim como anteriormente, utilizamos duas cadeias para simulação, com tamanho amostral total de 10000 valores e, 2000 amostras de aquecimento. Seguem as *prioris* adotadas para as estimativas Bayesianas:

- Modelo Gompertz:

$$a \sim Normal(-0,1; 0,5) \text{ e}$$
$$b \sim Gama(0,2; 2).$$

- Modelo de mistura padrão Gompertz:

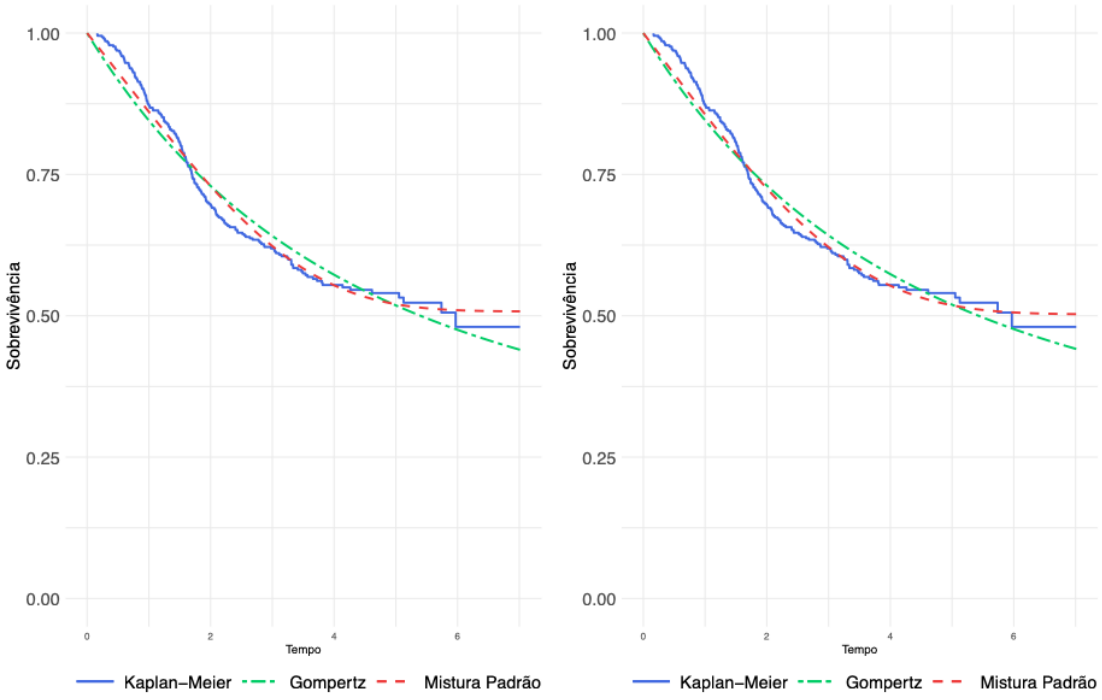
$$a \sim Normal(0,3; 0,5),$$
$$b \sim Gama(0,3; 1) \text{ e}$$
$$\pi \sim Uniforme(0; 1).$$

Na Tabela 5.3.2 mostramos as estimativas pontuais, intervalares e o DIC obtidas por métodos Bayesianos e, notamos que estas ainda continuam próximas das estimativas na Tabela 5.3.1.

Tabela 5.3.2: Resultados (Bayesianos) para o conjunto de dados sobre câncer melanoma.

Modelo	**Parâmetro**	**Estimativa**	**2,5%**	**97,5%**	**DIC**
Gompertz	a	-0,1316	-0,2346	-0,0305	1096,3502
	b	0,1786	0,1411	0,2243	
	π	0,2544	0,1502	0,3644	
Mistura padrão Gompertz	a	0,3050	0,1230	0,4406	1084,8707
	b	0,2918	0,2259	0,3728	
	π	0,5026	0,4306	0,5604	

Com relação ao porcentual de curadas do banco de dados, esta fração foi estimada em 50,01% pelo modelo com mistura e o modelo defeituoso estimou 25,44%.



(a) Estimativas frequentistas.

(b) Estimativas Bayesianas.

Figura 5.3: Estimativas para a função de sobrevivência, segundo os dois modelos paramétricos e o Kaplan-Meier.

5.4 Aplicação para Câncer de Cólon

Nesta aplicação o conjunto de dados é referente a câncer de cólon, um tipo de tumor que se desenvolve no intestino grosso e é bastante comum tanto em homens como mulheres. Este conjunto de dados, `colon`, faz parte da biblioteca `survival` (Therneau, 2022). Os tempos de vida para 1858 pacientes foram registrados, sendo que 50,5% são censurados à direita.

Na Tabela 5.4.1 mostramos as estimativas frequentistas, seus respectivos intervalos de confiança e o AIC. Observamos que ambos os modelos defeituoso e com mistura conseguiram captar a fração de curados. Assim como nos casos anteriores, utilizamos duas cadeias

Tabela 5.4.1: Resultados (frequentistas) para o conjunto de dados sobre câncer de cólon.

Modelo	Parâmetro	Estimativa	2,5%	97,5%	AIC
Gompertz	a	-0,2563	-0,2944	-0,2182	
	b	0,2195	0,1975	0,2415	5585,3965
	π	0,4247	0,3895	0,4598	
Mistura padrão Gompertz	a	0,0603	-0,0130	0,1337	
	b	0,3709	0,3326	0,4092	5576,9578
	π	0,4698	0,4391	0,5006	

com tamanho 6000 (1000 valores de *warm-up* ou aquecimento) e hiperparâmetros dados pelas estimativas frequentistas supramencionadas. Seguem as distribuições *a priori*

- Modelo Gompertz:

$$a \sim Normal(-0,25; 1) \text{ e}$$
$$b \sim Gama(0,2; 1).$$

- Modelo de mistura padrão Gompertz:

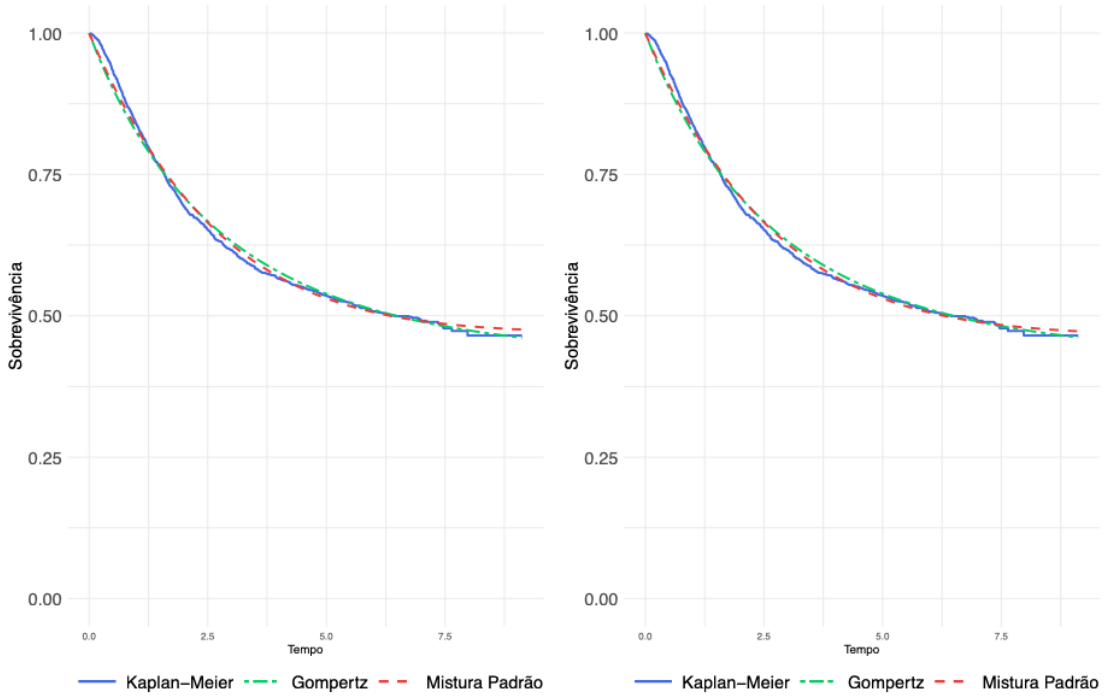
$$a \sim Normal(0,05; 0,5),$$
$$b \sim Gama(1; 2) \text{ e}$$
$$\pi \sim Uniforme(0; 1).$$

Nas estimativas Bayesianas (Tabela 5.4.2) o mesmo padrão das frequentistas essencialmente se repetiu. Ambas as estimativas do modelo com mistura e defeituoso captaram a presença de fração de cura, com o primeiro estimando a mesma em 44,3% e o segundo em 42,52%.

Tabela 5.4.2: Resultados (Bayesianos) para o conjunto de dados sobre câncer de cólon.

Modelo	**Parâmetro**	**Estimativa**	**2,5%**	**97,5%**	**DIC**
Gompertz	a	-0,2565	-0,2946	-0,2181	5585,3820
	b	0,2194	0,1984	0,2424	
	π	0,4252	0,3900	0,4605	
Mistura padrão Gompertz	a	0,0510	-0,0618	0,1218	5577,2664
	b	0,3693	0,3310	0,4099	
	π	0,4426	0,4700	0,4963	

Pelo critério DIC, escolhemos o modelo com mistura como mais parcimonioso e comparando as curvas com base na Figura 5.4b, vemos que os modelos defeituoso e com mistura estão extremamente próximos.



(a) Estimativas frequentistas.

(b) Estimativas Bayesianas.

Figura 5.4: Estimativas para a função de sobrevivência, segundo os dois modelos paramétricos e o Kaplan-Meier.

5.5 Aplicação para dados de Leucemia

O último conjunto de dados analisado diz respeito à leucemia, doença caracterizada por anomalias na produção dos glóbulos brancos no sangue; embora seja uma doença grave, muitas das sub condições da leucemia tem bom prognóstico. Os registros analisados provem do artigo de Kersey *et al.* (1987) e são referentes aos tempos de vida de 44 pacientes, sendo 21% dos registros censurados à direita.

Na Tabela 5.5.1 mostramos as estimativas de máxima verossimilhança para os modelos Gompertz e com mistura padrão. Ambos acompanham a fração de cura exibida pelo estimador de Kaplan-Meier, sendo bastante próximos, com o primeiro possuindo um AIC menor (mais parcimonioso). A proporção de cura de ambos é bastante parecida, sendo 20,42% (mistura padrão) e 20,73% (Gompertz). Por fim, na Figura 5.5a ilustramos os ajustes obtidos.

A partir das estimativas de EMV, definimos *prioris* para os parâmetros sob o enfoque Bayesiano (mesmas condições e tamanhos de cadeias das aplicações anteriores). Seguem as distribuições *a priori*

Tabela 5.5.1: Resultados (frequentistas) para o conjunto de dados sobre leucemia.

Modelo	Parâmetro	Estimativa	2,5%	97,5%	AIC
Gompertz	a	-1,5103	-2,2347	-0,7859	52,5762
	b	2,3767	1,3633	3,3901	
	π	0,2073	0,0791	0,3354	
Mistura padrão Gompertz	a	0,2143	-0,5592	0,9879	
	b	2,4968	1,4139	3,5796	50,7410
	π	0,2042	0,0849	0,3234	

- Modelo Gompertz:

$$a \sim Normal(-1,5; 0,5) \text{ e}$$
$$b \sim Gama(2,1; 1).$$

- Modelo de Mistura padrão Gompertz:

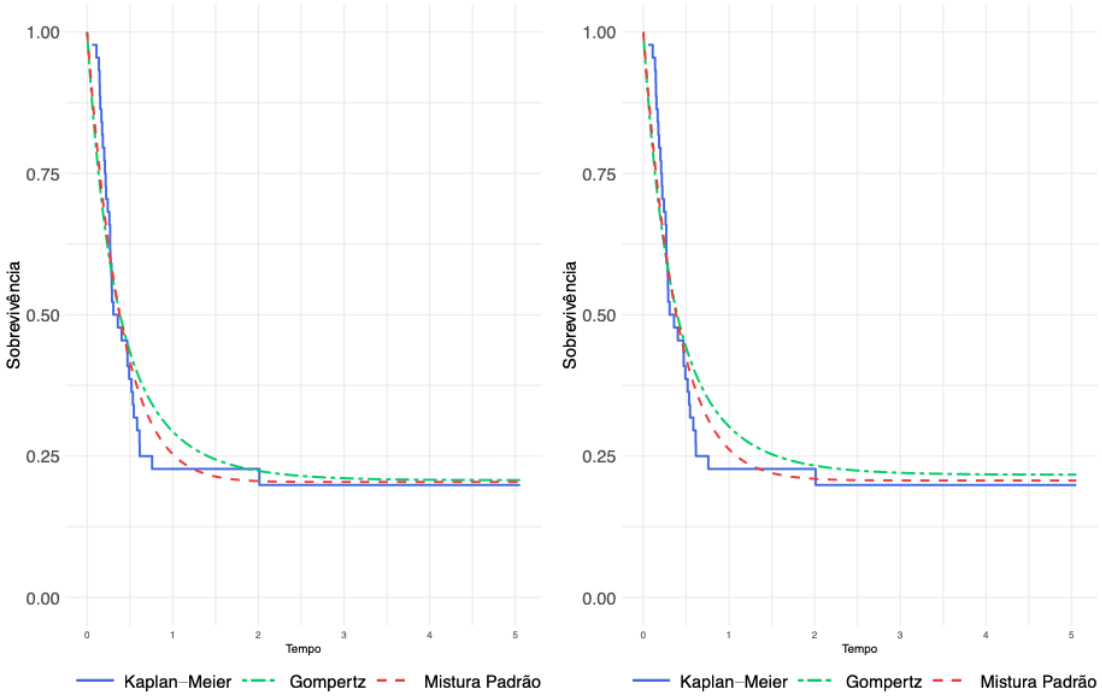
$$a \sim Normal(0,2; 1),$$
$$b \sim Gama(3; 1) \text{ e}$$
$$\pi \sim Uniforme(0; 1).$$

Na Tabela 5.5.2 mostramos as estimativas pontuais e intervalares e o critério DIC obtidas no contexto Bayesiano, na qual observamos que os modelos com mistura e defeituoso tiveram valores bastante próximos em termos de DIC. Há diferença mínima entre os valores estimados para a fração de cura, com o modelo com mistura ficando em 21,46% e o defeituoso, 21,70%.

Tabela 5.5.2: Resultados (Bayesianos) para o conjunto de dados sobre leucemia.

Modelo	Parâmetro	Estimativa	2,5%	97,5%	DIC
Gompertz	a	-1,5268	-2,1180	-0,9812	
	b	2,3330	1,5734	3,3324	51,7008
	π	0,2170	0,0930	0,3409	
Mistura padrão Gompertz	a	0,0904	-1,2000	0,7919	
	b	2,5447	1,6817	3,7829	50,3240
	π	0,2146	0,0844	0,3419	

Na Figura 5.5b mostramos o comportamento das funções de sobrevivência estimadas de acordo com cada modelo, notamos que o modelo defeituoso e com mistura estão próximos em termos de curva ajustada.



(a) Estimativas frequentistas.

(b) Estimativas Bayesianas.

Figura 5.5: Estimativas para a função de sobrevivência, segundo os dois modelos paramétricos e o Kaplan-Meier.

Capítulo 6

Conclusão

Para desenvolver este trabalho, realizamos uma extensiva revisão bibliográfica dos modelos de fração de cura, tanto com mistura como defeituoso. Além disso, trabalhamos consideravelmente os conceitos de programação, em especial com relação ao ajuste de modelos Bayesianos na linguagem `R`.

Ao longo deste projeto, descrevemos duas maneiras distintas de modelar dados de sobrevivência, utilizando modelos paramétricos da família Gompertz. A primeira é por meio de modelos com mistura, que possuem um parâmetro adicional destinado a modelar a proporção de curados. A segunda técnica é por meio do modelo defeituoso Gompertz, que consiste em relaxar a distribuição de modo a configurar uma função densidade imprópria, captando a fração de curados.

No Capítulo 5 comparamos o desempenho dos modelos supramencionados para cinco conjuntos de dados diferentes. Porém, todos os conjuntos possuem covariáveis e particularidades não exploradas ao longo deste estudo. Se considerarmos apenas as métricas de AIC e DIC, o modelo com mistura foi o se ajustou mais adequadamente nos cinco conjuntos de dados estudados. Entretanto, para quatro dos cinco conjuntos de dados estudados (exceção da aplicação na Seção 5.1), o modelo defeituoso também captou a fração de curados. Além disso, as curvas de sobrevivência estimadas por ambos modelos são muito próximas.

Vale ressaltar que quanto mais parâmetros a serem estimados, mais instável e custoso computacionalmente se torna o processo as estimativas. Neste caso, o modelo defeituoso - que tem um parâmetro a menos, pode se tornar um ajuste mais interessante e menos custoso de se obter (em especial no caso com covariáveis e/ou sob enfoque Bayesiano). Como neste trabalho consideramos apenas os tempos de vida e a informação de censura,

uma sugestão natural para outros estudos é explorar o impacto da inclusão de covariáveis nestes ajustes.

Referências Bibliográficas

Akram, S. e Ann, Q. U. (2015). Newton-Raphson method. *International Journal of Scientific & Engineering Research*, **6**(7), 1748–1752.

Amico, M. e Van Keilegom, I. (2018). Cure odels in survival analysis. *Annual Review of Statistics and Its Application*, **5**, 311–342.

Balka, J., Desmond, A. F. e McNicholas, P. D. (2009). Review and implementation of cure models based on first hitting times for Wiener processes. *Lifetime Data Analysis*, **15**(2), 147–176.

Betancourt, M. e Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. *Current Trends In Bayesian Methodology with Applications*, **79**(30), 2–4.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, **11**(1), 15–53.

Borgan, Ø. (2005). Nelson–Aalen Estimator. volume 5. Wiley Online Library.

Box, G. E. e Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis*, volume 40. John Wiley & Sons, 1^a edição.

Casella, G. e Berger, R. (2021). *Statistical Inference*. Cengage Learning, 2^a edição. ISBN 9780357753132.

Casella, G. e George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, **46**(3), 167–174.

Chernick, M. R. e Friis, R. H. (2003). *Introductory biostatistics for the health sciences: modern applications including bootstrap*. John Wiley & Sons, 1^a edição.

Chib, S. e Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, **49**(4), 327–335.

Cloyes, K. G., Wong, B., Latimer, S. e Abarca, J. (2010). Time to prison return for offenders with serious mental illness released from prison: A survival analysis. *Criminal Justice and Behavior*, **37**(2), 175–187.

Collett, D. (2015). *Modelling Survival Data in Medical Research*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 3ª edição. ISBN 9781498731690.

Colosimo, E. A. e Giolo, S. R. (2006). *Análise de Sobrevivência Aplicada*. Editora Blucher.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**(2), 187–202.

Gamerman, D. e Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press, 2ª edição.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. e Rubin, D. B. (2013). *Bayesian data analysis*. CRC Press, 3ª edição.

Gieser, P. W., Chang, M. N., Rao, P., Shuster, J. J. e Pullen, J. (1998). Modelling cure rates using the Gompertz model with covariate information. *Statistics in Medicine*, **17**(8), 831–839.

Ibrahim, J., Chen, M. e Sinha, D. (2001). *Bayesian Survival Analysis*. Springer Series in Statistics. Springer, 1ª edição. ISBN 9783540952770.

Jackson, C. (2016). flexsurv: A platform for parametric survival modeling in R. *Journal of Statistical Software*, **70**(8), 1–33.

Kalbfleisch, J. e Prentice, R. (2011). *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. Wiley, 2ª edição. ISBN 9781118031230.

Kaplan, E. L. e Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**(282), 457–481.

Kersey, J. H., Weisdorf, D., Nesbit, M. E., LeBien, T. W., Woods, W. G., McGlave, P. B., Kim, T., Vallera, D. A., Goldman, A. I., Bostrom, B. *et al.* (1987). Comparison of autologous and allogeneic bone marrow transplantation for treatment of high-risk

refractory acute lymphoblastic leukemia. *New England Journal of Medicine*, **317**(8), 461–467.

Kirkwood, T. B. L. (2015). Deciphering death: a commentary on Gompertz (1825) 'On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **370**(1666), 20140379.

Leimkuhler, B. e Reich, S. (2004). *Simulating hamiltonian dynamics*. Number 14 Em Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press.

Lunn, D. J., Thomas, A., Best, N. e Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, **10**(4), 325–337.

Maller, R. A. e Zhou, X. (1996). *Survival Analysis With Long-Term Survivors*. Wiley New York.

Nunes, A. e de Morais Sarmento, E. (2012). Business demography dynamics in Portugal: a non-parametric survival analysis. Em *The Shift to the Entrepreneurial Society*. Edward Elgar Publishing.

Othus, M., Barlogie, B., LeBlanc, M. L. e Crowley, J. J. (2012). Cure models as a useful statistical tool for analyzing survival. *Clinical Cancer Research*, **18**(14), 3731–3736.

Paula, G. A. (2004). *Modelos de regressão: com apoio computacional*. IME-USP São Paulo.

Pollard, J. H. e Valkovics, E. J. (1992). The Gompertz distribution and its applications. *Genus*, **48**, 15–28.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rausand, M. e Hoyland, A. (2003). *System reliability theory: models, statistical methods, and applications*. John Wiley & Sons, 3^a edição.

Rocha, R. F., Tomazella, V. L. D. e Louzada, F. (2014). Inferência clássica e Bayesiana para o modelo de fração de cura Gompertz defeituoso. *Revista Brasileira de Biometria*, **32**(1), 104–114.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. e Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(4), 583–639.

Stan Development Team (2022). Stan modeling language users guide and reference manual, version 2.29.0.

Therneau, T. M. (2022). *A Package for Survival Analysis in R*. R package version 3.3-1.

Thomas, S. e Tu, W. (2021). Learning Hamiltonian Monte Carlo in R. *The American Statistician*, **75**(4), 403–413.

Zhang, Z. e Sun, J. (2010). Interval censoring. *Statistical Methods in Medical Research*, **19**(1), 53–70.

Apêndice A

Demonstrações Adicionais

A.1 Função de Risco

Supondo que $S(t)$ representa a função de sobrevivência no tempo t, a probabilidade de ocorrência do evento dentro do intervalo definido por $[t_1, t_2)$, é dada por

$$S(t_1) - S(t_2),$$

lembrando que $S(t)$ é decrescente em t. A taxa de ocorrência nesse intervalo é dada pela razão entre probabilidade de ocorrência e a probabilidade do evento não ter acontecido até o inicio do intervalo $[t_1, t_2)$

$$\frac{S(t_1) - S(t_2)}{(t_2 - t_1)S(t_1)}, \tag{A.1}$$

a presença e $S(t_1)$ no denominador é justificada pelo fato de que, assumimos que o evento não ocorreu até t_1. Se definirmos $t_2 = t_1 + \Delta t_1$, tal que $\Delta > 0$, a expressão (A.1) passa a ser escrita como

$$\frac{S(t_1) - S(t_1 + \Delta t_1)}{(t_1 + \Delta t_1 - t_1)S(t_1)} = \frac{S(t_1) - S(t_1 + \Delta t_1)}{\Delta t_1 S(t_1)}.$$

A função de risco no instante t, que intitulamos como $h(t)$, é obtida conforme $\Delta t_1 \to 0$,

$$h(t_1) = \lim_{\Delta t_1 \to 0} \frac{S(t_1) - S(t_1 + \Delta t_1)}{\Delta t_1 S(t_1)}. \tag{A.2}$$

Para resolvermos o limite em (A.2), podemos utilizar a relação dada por Kalbfleisch e Prentice (2011), de que

$$rf(t) \approx S(t) - S(t + r), \tag{A.3}$$

quando r é uma quantidade muito pequena. Substituindo $r = \Delta t_1$ em (A.3), temos $\Delta t_1 f(t) = S(t) - S(t_1 + \Delta t_1)$. Assim, (A.2) pode ser escrita como,

$$
\begin{aligned}
h(t_1) &= \lim_{\Delta t_1 \to 0} \frac{S(t_1) - S(t_1 + \Delta t_1)}{\Delta t_1 S(t_1)} \\
h(t_1) &= \lim_{\Delta t_1 \to 0} \Delta t_1 f(t_1) \Delta t_1 S(t_1) \\
h(t_1) &= \frac{f(t_1)}{S(t_1)}.
\end{aligned}
$$

Desse modo, mostramos que a função de risco $h(t)$ pode ser representada pelo quociente entre a função densidade e a função de sobrevivência nesse instante.

Apêndice B

Estimadores de Kaplan-Meier

Na Tabela B.1 reportamos as estimativas para $S(t)$ obtidas via Kaplan-Meier no caso sem e com censura e, na Tabela B.2 reportamos as estimativas por Kaplan-Meier para o conjunto de dados *TGCA*.

Tabela B.1: Estimador de Kaplan-Meier para os dados simulados.

Sem censura			Com Censura		
Tempo	Sobrevivência Estimada	Evento	Tempo	Sobrevivência Estimada	Evento
0,000	0,998	1	0,001	0,998	1
0,002	0,996	1	0,001	0,996	1
0,003	0,994	1	0,002	0,994	1
0,004	0,992	1	0,002	0,992	1
0,004	0,990	1	0,003	0,990	1
0,004	0,988	1	0,003	0,988	1
0,004	0,986	1	0,004	0,988	0
0,005	0,984	1	0,004	0,986	1
0,006	0,982	1	0,005	0,984	1
0,006	0,980	1	0,005	0,982	1
...	...	...	...	...	...
1,320	0,018	1	0,503	0,038	1
1,328	0,016	1	0,513	0,034	1
1,336	0,014	1	0,525	0,029	1
1,345	0,012	1	0,541	0,025	1
1,481	0,010	1	0,562	0,021	1
1,498	0,008	1	0,570	0,017	1
1,513	0,006	1	0,674	0,013	1
1,547	0,004	1	0,686	0,008	1
1,770	0,002	1	0,687	0,004	1
2,088	0,000	1	0,927	0,000	1

Tabela B.2: Estimador de Kaplan-Meier apresentado na Figura 3.1.

Tempo	Sobrevivência Estimada	Evento
0,003	0,999	1
0,014	0,999	0
0,019	0,999	0
0,022	0,999	0
0,025	0,999	0
0,027	0,999	0
0,030	0,999	0
0,044	0,999	0
0,052	0,999	0
0,058	0,999	0
...	...	...
17,238	0,452	0
17,688	0,402	1
18,063	0,352	1
19,468	0,352	0
19,523	0,352	0
21,307	0,352	0
21,940	0,352	0
22,989	0,352	0
23,441	0,352	0
23,575	0,352	0