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

Modelagem de fração de cura, aplicado na classificação de clientes com dados segmentados

Douglas de Paula Nestlehner

Trabalho de Conclusão de Curso

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

Modelagem de fração de cura, aplicado na classificação de clientes com dados segmentados

Douglas de Paula Nestlehner
Orientador: José Carlos Fogo

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

São Carlos
Novembro de 2023

Douglas de Paula Nestlehner

Modelagem de fração de cura, aplicado na classificação de clientes com dados segmentados

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Douglas de Paula Nestlehner e aprovado pela banca examinadora.

Aprovado em 20 de Março de 2023

Banca Examinadora:

- Prof. Dr. José Carlos Fogo
- Profa. Dra. Vera Lucia Damasceno Tomazella
- Prof. Dr. Luis Ernesto Bueno Salasar

Resumo

Modelos com fração de cura permitem considerar indivíduos que não foram sujeito ao evento de interesse, podendo assim obter mais resultados em relação aos usuais modelos de sobrevivência. Apresentamos nesse trabalho, um estudo da aplicação de modelos de sobrevivência com fração de cura para dados segmentados. Para isso, utilizamos uma das principais modelagens de fração de cura, o modelo de mistura padrão, realizando ajustes considerando diferentes distribuições, no intuito de se obter o melhor modelo.

Algumas das áreas mais fomentadas do mercado estão relacionadas a vendas de serviços por assinatura, que buscam por meio de estudos de fidelização de clientes, ter conhecimento sobre as ocorrências de churns, ou seja, ter conhecimento sobre o motivo de seus clientes deixarem de adquirirem seus produtos.

A grande maioria desses estudos captam dados de clientes ativos, consequentemente existirá um alto número de observações que não foram sujeitas ao evento de interesse, o churn. Assim sendo, a aplicação de modelos de sobrevivência com fração de cura em estudos de ocorrências de churns, principalmente relacionados a serviços por assinaturas, são adequados e podem trazer resultados relevantes para as tomadas de decisões das empresas.

Desse modo, realizamos a aplicação dos resultados obtidos neste estudo, em uma base de dados de um experimento de ocorrência de *churn* de uma determinada empresa, a qual tem um numero elevado de não ocorrência do evento de interesse, e existe a segmentação dos dados por diversas características importantes.

Palavras-chave: *Fração de Cura, Modelo de Mistura Padrão, Análise de Sobrevivência, Churn.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

O uso de técnicas de análise de sobrevivência trazem resultados de impacto nas mais diversas áreas do conhecimento, auxiliando em tomadas de decisões relacionadas ao tempo de vida do evento em estudo. Por esse motivo, a análise de sobrevivência é uma das áreas da estatística que mais crescem nos últimos anos, principalmente em estudos relacionados a área da saúde e venda de produtos.

Os procedimentos e técnicas de análise de sobrevivência, visam investigar o tempo até a ocorrência de um evento de interesse. Na maioria dos casos, observamos estudos em que o evento de interesse é a cura de uma doença, diagnóstico de uma doença, morte, falha de algum sistema, entre outros. Mas, de modo geral, o evento de interesse pode ser qualquer ocorrência que caracterize uma falha ou sucesso, sendo sempre definida no começo do experimento.

Uma das características que frequentemente observamos em dados coletados no intuito de se realizar algum estudo de sobrevivência, é a presença de censuras, caracterizadas pela observação sem que haja a ocorrência do evento de interesse. Em estudos de fidelização de clientes de serviços por assinatura, teremos dados com essa característica, pois em serviços por assinaturas os clientes tendem a consumir o produto por um longo período de tempo, não apresentando o evento de interesse.

Diversos fatores vêm fazendo com que clientes de serviços por assinatura (streaming, tv a cabo, internet, etc.) cancelem seus planos e/ou troquem por outros, sendo os principais:

- Os efeitos econômicos da pandemia, que causam a necessidade dos clientes procurarem por serviços mais baratos e alternativos;
- O aumento da concorrência, com cada vez mais novas empresas surgindo no mercado,

oferecendo produtos similares e com preços menores;

- A facilidade adquirida pelas tecnologias atuais (internet, telefone, serviços de comunicação etc.) onde conseguimos ter conhecimento dos melhores produtos disponíveis, permitindo a troca de produto quando não estamos satisfeitos, entre outros fatores.

Portanto, a fidelização de clientes é cada vez mais valorizada, pois é de extrema importância para as empresas de assinaturas ter clientes que continuam adquirindo e utilizando os seus produtos, além de que, custa cerca de 5 a 10 vezes mais recrutar um novo cliente do que fidelizar um já existente (Lu, 2003).

Desse modo, a maior diferença em que podemos destacar entre um serviço de assinatura bem-sucedido, é o quanto a empresa consegue reter seus clientes, para isso são criados os estudos relacionados a fidelização, no intuito de ter conhecimento do que está ocorrendo na empresa, e auxiliar nas tomadas de decisões relacionadas a estratégias para manter seus clientes.

Os principais estudos de fidelização de cliente tem como objetivo predizer a probabilidade de um cliente abandonar o serviço (denotamos esse evento como *churn*). Entretanto, como foi dito anteriormente, nesses estudos são coletadas amostras com um grande número de indivíduos que não abandonaram o serviço, portanto também podemos ter como interesse predizer a fração de clientes que permanecem fiéis ao produto em questão. Para isso, podemos fazer o uso de modelos de sobrevivência com fração de cura.

Desse modo, nesse trabalho iremos estudar modelos de análise de sobrevivência com fração de cura, aplicando a um caso de serviço de assinatura, em que o termo "fração de cura" se refere à fração de indivíduos que não experimentaram o evento de interesse, comumente chamados de indivíduos imunes ou curados. Traduzindo para o problema de fidelização, fração de cura trata-se dos indivíduos que não cancelaram o serviço de assinatura até o fim do estudo.

1.1 Objetivos

Nesse trabalho, realizamos um estudo de fidelização de clientes de serviços por assinatura, utilizando a modelagem de fração de cura para estimar a probabilidade de um cliente permanecer fiel por um longo período.

No relacionamento com clientes, muitas empresas têm interesse em classificar seus

clientes, muitas vezes segmentados em grupos com características particulares, possibilitando a implementação de estratégias que visam a retenção desses clientes.

Nessas situações, entretanto, vários clientes acabam abandonando o serviço, num processo que recebe o nome de *churn*, porém, a grande maioria permanece fiel, gerando uma fração de indivíduos que não obtiveram *churn*, em análise de sobrevivência chamamos essa fração de indivíduos como fração de cura.

Assim sendo, o presente trabalho teve como objetivo utilizar modelos de regressão com fração de cura no intuito de predizer a probabilidade de um indivíduo permanecer fiel por um longo período, considerando suas características.

Capítulo 2

Metodologia

Nessa seção apresentamos as metodologias estatísticas que utilizamos para atingir o objetivo proposto.

2.1 Análise de Sobrevivência

Análise de sobrevivência é o termo usado para descrever a análise de dados na forma de tempos desde uma origem temporal bem definida até a ocorrência de algum determinado evento ou ponto final, (Collet, 2015).

Esse tipo de análise tem como principal objetivo investigar o tempo de vida dos indivíduos expostos a um experimento. Ela também pode ser aplicada na duração de componentes até o tempo de falha ou, de maneira geral, em estudos nos quais o tempo de vida é observado até a ocorrência de um evento de interesse, (Fogo, 2007). Para esse estudo, temos como o evento de interesse o tempo até o cliente cancelar o contrato por assinatura (até ocorrer o *churn*).

Nos casos em que o evento de interesse não ocorre temos uma característica muito importante nos dados de análise de sobrevivência, as chamadas censuras, as quais terão mais destaques na Seção 2.1.1.

Denotamos então, como variável resposta T, o tempo até a ocorrência do evento de interesse, sendo T uma variável aleatória contínua e não-negativa. Assim podemos observar as principais medidas de interesse em análise de sobrevivência: função de sobrevivência $S(t)$ e função de risco $h(t)$ que serão descritas com mais detalhes nas Seções 2.2.1 e 2.2.2.

2.1.1 Censura

As censuras ocorrem quando não conseguimos observar o evento de interesse para alguns indivíduos do estudo. Ela é atribuída em geral aos indivíduos que por algum motivo não foram acompanhados até o fim do experimento, ou seja, que possuem tempo de falha superior aquele observado (Lawless, 2011).

Vários motivos podem ocasionar a presença de censura na base de dados, sendo que podemos classificar a censura em três diferentes categorias:

- **Censura do Tipo I:** Ocorre quando o estudo tem um limite final de tempo de execução (denotado por L), pré-definido, e quando o tempo L é atingido, todos os indivíduos que não apresentaram o evento de interesse são considerados como censura do tipo I. Na Figura 2.1 reproduzimos um exemplo de censura do tipo I.

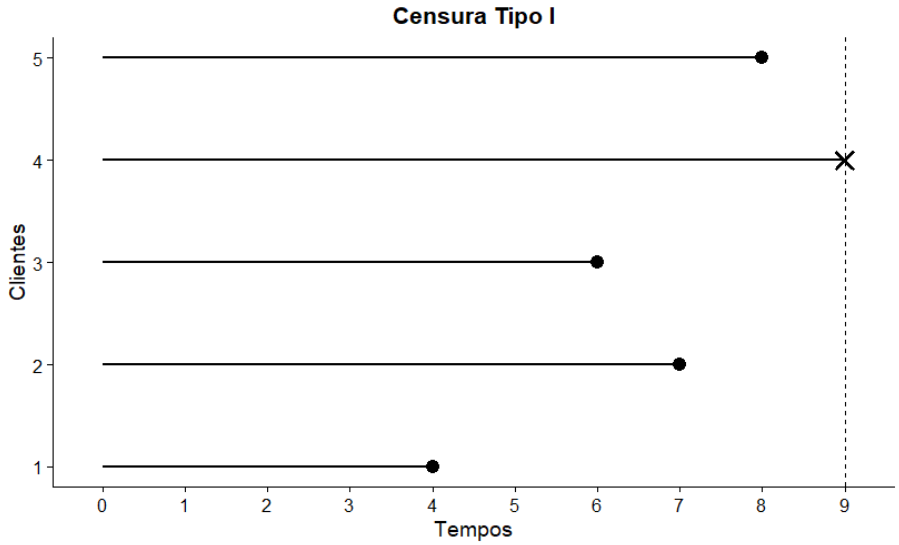

Figura 2.1: Exemplo de um estudo em que o tempo limite é igual a 9, e 5 indivíduos são observados (• indica ocorrência do evento de interesse, e × indica a ocorrência de censura do tipo I).

- **Censura do Tipo II:** Ocorre quando o estudo tem um número pre-definido de eventos de interesse, ou seja, em uma amostra com n indivíduos é observada até a ocorrência de r eventos de interesse ($r < n$), assim os $n - r$ indivíduos restantes são considerados como censura. Na Figura 2.2 reproduzimos um exemplo de censura do tipo II, em que o número de eventos pré-definidos é igual a $r = 3$, ou seja, após a $3^{\underline{a}}$ ocorrência do evento de interesse, todos os demais indivíduos são considerados censurados.

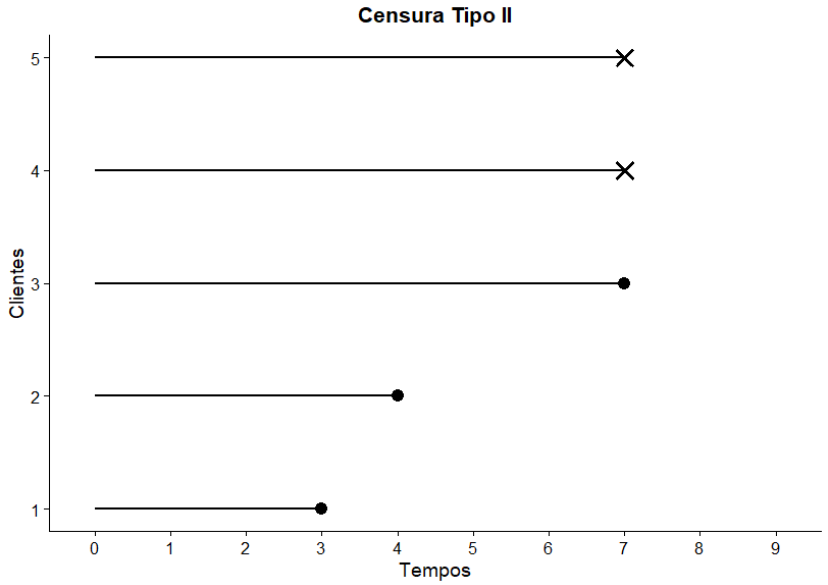

Figura 2.2: Exemplo de um estudo em que o tempo limite é igual a 9, e 5 indivíduos são observados (• indica ocorrência do evento de interesse, e × indica a ocorrência de censura do tipo II).

- **Censura Aleatória:** Ocorre quando o indivíduo por algum motivo aleatório deixa o estudo. Observada com frequência na área médica, quando um paciente deixa o estudo por razões distintas do evento de interesse, como exemplo: morte do paciente por outras razões não relacionadas ao estudo; paciente deixa de comparecer ao estudo; etc. Na Figura 2.1.1 reproduzimos um exemplo de censura do tipo Aleatória, em que os indivíduos 2 e 4 deixaram o estudo por algum motivo aleatório.

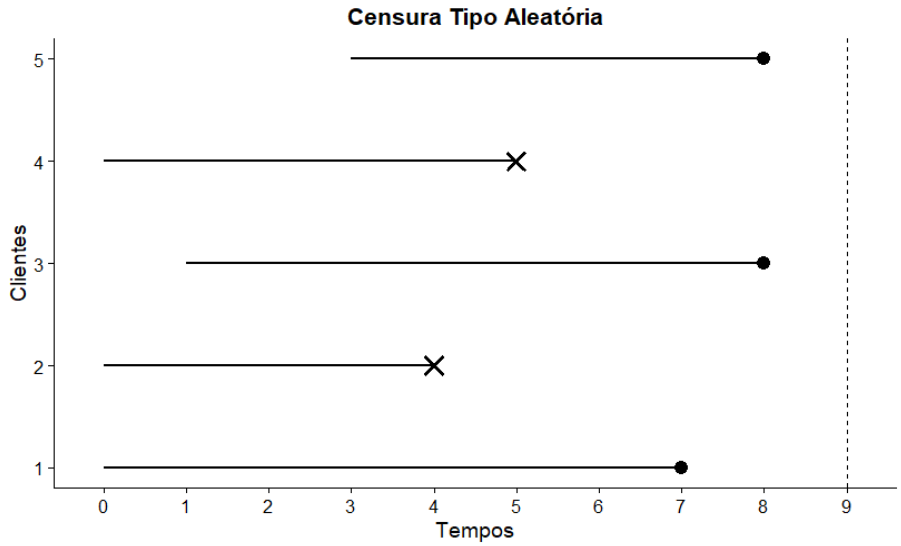

Figura 2.3: Exemplo de um estudo em que o tempo limite é igual a 9, e 5 indivíduos são observados (• indica ocorrência do evento de interesse, e × indica a ocorrência de censura do tipo aleatória).

2.2 Funções de Interesse

Apresentamos a seguir as definições das principais funções utilizadas em análise de sobrevivência.

2.2.1 Função de Sobrevivência

Os dados de sobrevivência para o indivíduo i $(i = 1, ..., n)$ sob estudo, são representados, em geral, por t_i, o tempo de falha ou de censura, e δ_i , a variável indicadora de falha ou censura (Colosimo, 2006), ou seja:

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

A função de sobrevivência tem como intuito fornecer a probabilidade do indivíduo sobreviver ao tempo t, sendo uma função monótona e decrescente.

Temos como propriedades da função de sobrevivência:

1. $S(t) = 1$, se $t = 0$;
2. $S(t) = 0$, se $t \longrightarrow \infty$;
3. $S(t)$ é não crescente.

A função de Sobrevivência $S(t)$ é definida por:

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t). \tag{2.1}$$

Em que $F(t)$ é a função de distribuição acumulada de T, a qual representa a probabilidade do indivíduo apresentar o evento de interesse.

Da teoria da probabilidade, a função densidade de probabilidade é definida por (Cassela, 2002),

$$f(t) = \frac{d}{dt}F(t). \tag{2.2}$$

Na Figura 2.4 temos representado um exemplo de comportamento da função de sobrevivência ao longo do tempo, em que podemos notar as duas propriedade:$S(0) = 1$ e quando $t \longrightarrow \infty$ temos $S(t) = 0$.

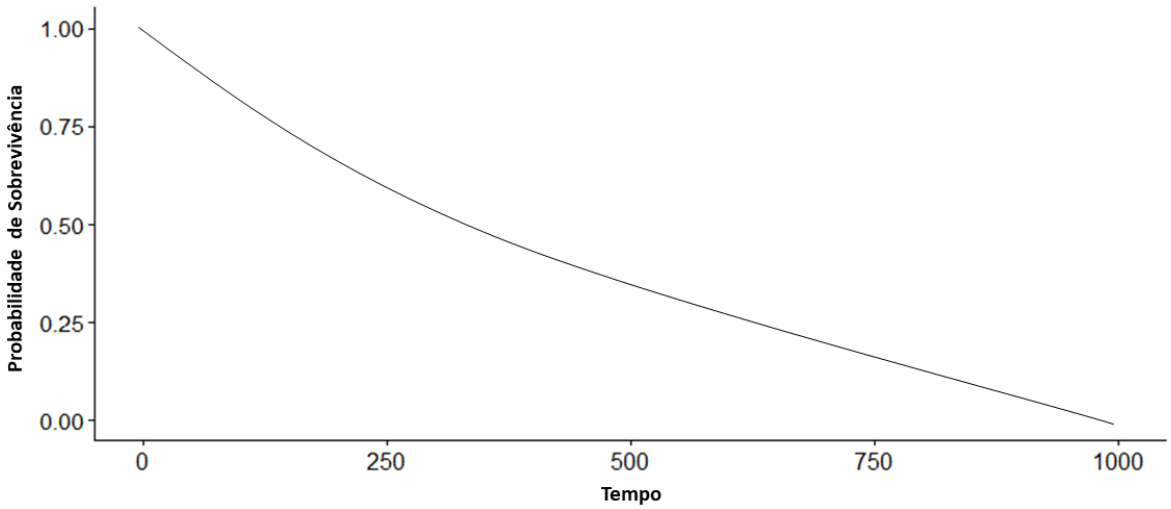

Figura 2.4: Exemplo curva sobrevivência.

Existem diversos métodos para poder estimar a função de sobrevivência $S(t)$, sendo um dos mais utilizados o método de estimação desenvolvido por Kaplan e Meier (1958), o qual é um estimador não paramétrico, ou seja, não é necessário assumir um modelo paramétrico para a variável resposta T.

2.2.2 Função de Risco

Outra medida bastante utilizada em análise de sobrevivência é a função de risco $h(t)$, também chamada de função taxa de risco, fornece o risco do evento ocorrer em um intervalo de tempo muito pequeno $[t, t + dt)$, dada a sobrevivência no início do intervalo, ou seja, no tempo t.

Sendo definida por:

$$h(t) = \lim_{dt \to 0} \frac{P(t < T \leq t + dt | T > t)}{dt},$$

equivalente a:

$$\frac{1}{P(T \geq t)} \lim_{dt \to 0} \frac{P(t < T \leq t + dt)}{dt} = \frac{f(t)}{S(t)}, \tag{2.3}$$

e de (2.3) temos a seguinte relação::

$$h(t) = -\frac{d}{dt} log(S(t)). \tag{2.4}$$

Também definimos a função de risco acumulada $H(t)$:

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

Portanto, por meio de (2.4) e (2.5), temos a seguinte relação para a função de risco acumulada $H(t)$:

$$H(t) = -log(S(t)),$$

de onde escrevemos a função de sobrevivência $S(t)$ em relação apenas da função de risco:

$$S(t) = e^{-H(t)}. \tag{2.6}$$

2.2.3 Função de Verossimilhança

A função de verossimilhança é a principal função utilizada na estimação dos dados pelo método de verossimilhança, portanto a sua definição é a base para o início da modelagem.

Considerando a seguinte situação de estudo:

- Estudo onde n indivíduos de k diferentes grupos foram observados. Em que, n_k representa o número de indivíduos observado no k-ésmio grupo;
- Os tempos são representados por t_{ki} sendo $i = 1, 2, ..., n_k$ e $k = 1, 2, ..., g$, em que os tempos t_{ki} são independentes e identicamente distribuídos com função densidade de probabilidade $f(t_{ki}|\boldsymbol{\theta_k})$, sendo $\boldsymbol{\theta_k}$ o vetor de parâmetros do k-ésimo grupo.
- O indicador de falhas é representado por δ_{ki}, com $i = 1, 2, ..., n_k$ e $k = 1, 2, ..., g$

Desse modo, $D = (t_{ki}, \delta_{ki})$ é o conjunto de dados em estudo, e a função de verossimilhança do k-ésimo grupo é dada por:

$$L_k(\boldsymbol{\theta}_k|D) = \prod_{i=1}^{n_k} \left[f_k(t_{ki})\right]^{\delta_{ki}} \left[S_k(t_{ki})\right]^{1-\delta_{ki}}, \quad i = 1, 2, ..., n_k \quad \text{e} \quad k = 1, 2, ..., g.$$

Podemos ter como interesse expressar a função de verossimilhança em relação a função

de risco, utilizando a relação expressa em (2.6) temos:

$$\begin{aligned} L_k(\boldsymbol{\theta}_k|D) &= \prod_{i=1}^{n_k} \left[h_k(t_{ki})S_k(t_{ki})\right]^{\delta_{ki}} \left[S_k(t_{ki})\right]^{1-\delta_{ki}} \\ &= \prod_{i=1}^{n_k} \left[h_k(t_{ki})\right]^{\delta_{ki}} \left[S_k(t_{ki})\right] \\ &= \prod_{i=1}^{n_k} \left[h_k(t_{ki})\right]^{\delta_{ki}} \exp(-H_k(t_{ki})) \end{aligned}$$

Ressaltando que L_k representa a verossimilhança do k-ésimo grupo, e que os componentes da função: $S_k(t_{ki})$ e $f_k(t_{ki})$ ou $h_k(t_{ki})$ e $H_k(t_{ki})$, são referentes ao k-ésimo grupo.

2.3 Modelo de Mistura Padrão

Comumente os dados de sobrevivência apresentam um certo percentual de não ocorrência do evento de interesse, e uma alternativa para modelar esse tipo de dados é utilizando modelos com fração de cura.

De acordo com (Maller, 1996), para que esse tipo de modelo seja uma opção viável é necessário que o tempo de seguimento dos indivíduos seja suficientemente longo, e a curva de sobrevivência apresente uma estabilização a partir de um determinado tempo, indicando a presença de uma fração razoável de indivíduos que não irão experimentar o evento de interesse, mesmo se forem acompanhados por um longo período de tempo.

Na Figura 2.5 temos representado um exemplo de uma função de sobrevivência com característica de fração de cura, em que p é a probabilidade do indivíduo não apresentar o evento de interesse ao final do experimento.

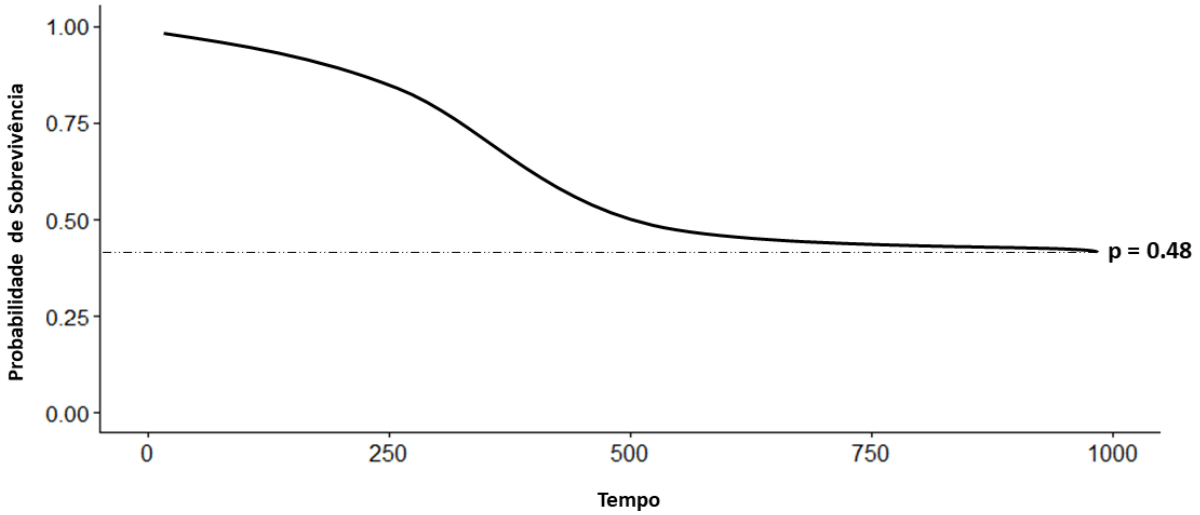

Figura 2.5: Exemplo da curva de sobrevivência com fração de cura.

Para esse tipo de modelagem exitem métodos propostos por diversos autores, entretanto o mais utilizado é o modelo de mistura padrão. O modelo de mistura padrão foi proposto por Berkson (1952) e Boag (1949), e o método consiste na mistura de duas distribuições paramétricas, sendo elas: função de sobrevivência própria de toda a população; e a função de sobrevivência imprópria da fração da população de não curados.

Desta forma, a função de sobrevivência populacional imprópria $S_{pop}(t_i)$, dada pela probabilidade do tempo de vida ser maior que um valor t_i, é definida por:

$$S_{pop}(t_i) = p + (1-p)S(t_i), \tag{2.7}$$

em que $(1-p)$ é a probabilidade do indivíduo estar em risco no instante t_i. Sob esse enfoque, a fração de indivíduos que não apresentam o evento de interesse, também chamada de fração de cura, é dada por p.

Temos como propriedades da função de sobrevivência do modelo de mistura padrão $S_{pop}(t_i)$:

1. $S_{pop}(t_i) = S(t_i)$, se $p = 0$;
2. $S_{pop}(t_i) = 1$, se $t_i = 0$;
3. $S_{pop}(t_i) = p$, se $t_i \longrightarrow \infty$;
4. $S_{pop}(t_i)$ é não crescente.

Assim sendo, a função de densidade imprópria $f_{pop}(t_i)$, e função de risco imprópria $h_{pop}(t_i)$, são dadas por:

$$f_{pop}(t_i) = (1-p)f(t_i), \tag{2.8}$$

$$h_{pop}(t_i) = \frac{(1-p)f(t_i)}{p+(1-p)S(t_i)}. \tag{2.9}$$

Dessa forma, a função de verossimilhança para o modelo com fração de cura, é dada por:

$$\begin{aligned} L(\boldsymbol{\theta}|D) &= \prod_{i=1}^{n} \left[f_{pop}(t_i)\right]^{\delta_i} \left[S_{pop}(t_i)\right]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left[(1-p)f(t_i)\right]^{\delta_i} \left[p+(1-p)S(t_i)\right]^{1-\delta_i} \end{aligned}$$

Considerando k grupos, cada um com uma fração de cura p_k, com $k = 1, 2, ..., g$, a função de verossimilhança para o k-ésimo grupo, é dada por:

$$L_k(\boldsymbol{\theta_k}|D) = \prod_{i=1}^{n_k} \left[(1-p_k)f(t_{ki})\right]^{\delta_{ki}} \left[p_k+(1-p_k)S(t_{ki})\right]^{1-\delta_{ki}} \tag{2.10}$$

2.4 Distribuições do Tempo de Vida

Nessa seção, apresentamos três distribuições de probabilidade, utilizadas ao longo desse trabalho, para caracterizar o tempo até a o coerência do evento de interesse.

2.4.1 Modelo Exponencial

Uma das distribuições mais utilizadas e de fácil aplicação em estudos de sobrevivência é a distribuição exponencial, a fácil aplicação se deve a existência de um único parâmetro λ. A distribuição exponencial é conhecida pela sua propriedade de falta de memória, que pode ser representada pela sua taxa de risco constante (Fogo, 2007).

Na figura 2.6 temos representado o comportamento da função de densidade, função de sobrevivência e função de risco da exponencial, considerando diferentes valores para o parâmetro λ, com $t \in [0, 6]$.

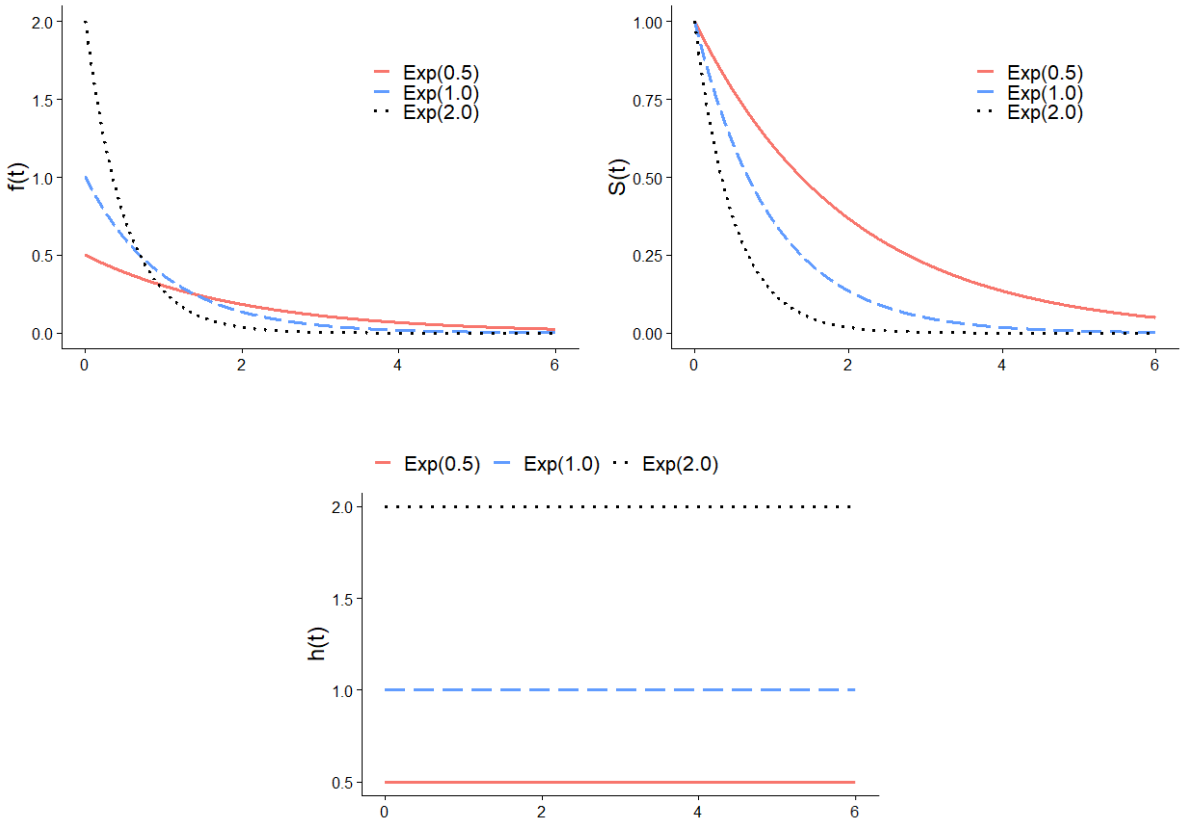

Figura 2.6: Comportamento das funções de densidade, sobrevivência e risco da distribuição exponencial.

Desse modo, considerando a distribuição exponencial para os tempos de vida, a função de densidade $f(t_i)$ é dada por:

$$f(t_i) = \lambda \mathrm{e}^{-\lambda t_i}, \quad t > 0 \text{ e } \lambda > 0,$$

A função de sobrevivência:

$$S(t_i) = \mathrm{e}^{-\lambda t_i},$$

E a função de risco:

$$h(t_i) = \frac{f(t_i)}{S(t_i)} = \lambda$$

Partindo para a modelagem considerando o modelo de mistura padrão, teremos por meio de (2.1) e (2.7) que a sobrevivência imprópria do k-ésimo grupo é dada por:

$$S_{pop_k}(t_{ki}) = p_k + (1 - p_k)\mathrm{e}^{-\lambda_k t_{ki}},$$

e de (2.8) obtemos a densidade imprópria:

$$f_{pop_k}(t_{ki}) = \lambda_k(1 - p_k)\mathrm{e}^{-\lambda_k t_{ki}},$$

Assim, por meio da equação apresentada em (2.10), a função de verossimilhança do k-ésimo grupo, considerando a distribuição exponencial, é dada por:

$$L_k(\boldsymbol{\theta_k}|D) = \prod_{i=1}^{n_k} \left[\lambda_k(1 - p_k)\mathrm{e}^{\lambda_k t_{ki}}\right]^{\delta_{ki}} \left[p_k + (1 - p_k)\mathrm{e}^{-\lambda_k t_{ki}}\right]^{1-\delta_{ki}}$$

2.4.2 Modelo Weibull

Outra distribuição com bastante utilidade para modelar o tempo de vida em estudos de sobrevivência, é a distribuição Weibull. Diferente da exponencial, a distribuição Weibull possui dois parâmetros, um de escala (λ), e outro de forma (α).

Na figura 2.7 temos representado o comportamento das funções de densidade, sobrevivência e risco, considerando diferentes parâmetros λ e α, com $t \in [0, 3]$.

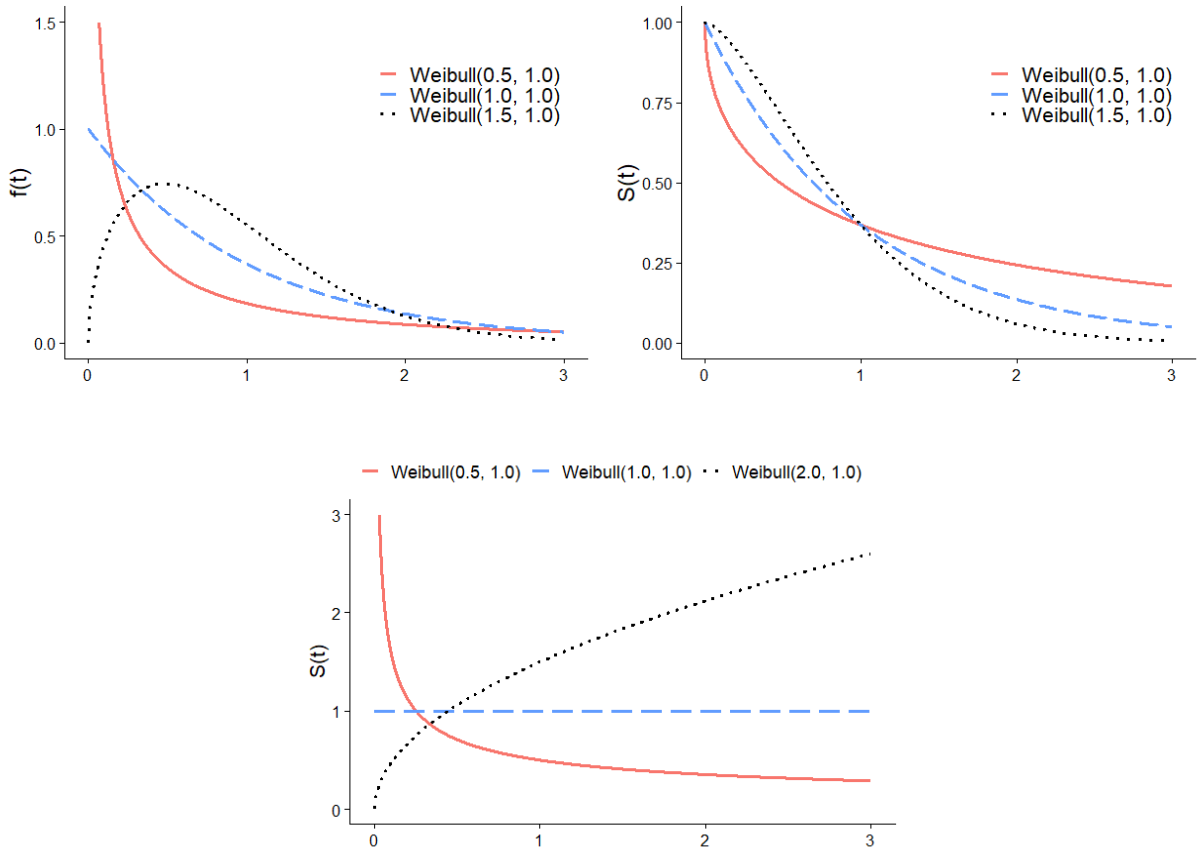

Figura 2.7: Comportamento das funções de densidade, sobrevivência e risco da distribuição Weibull.

Considerando a distribuição Weibull para os tempos de vida, a função de densidade $f(t_i)$ é dada por:

$$f(t_i) = \alpha\lambda(t_i\lambda)^{(\alpha-1)}\mathrm{e}^{-(t_i\lambda)^{\alpha}}, \quad \alpha > 0, \quad \lambda > 0$$

A função de sobrevivência:

$$S(t_i) = \mathrm{e}^{-(t_i\lambda)^{\alpha}}$$

Seguindo a mesma lógica apresentada no modelo exponencial, por meio de (2.1) e (2.7) obtemos a sobrevivência imprópria do k-ésimo grupo:

$$S_{pop_k}(t_{ki}) = p_k + (1 - p_k)\mathrm{e}^{-(t_{ki}\lambda_k)^{\alpha_k}},$$

e de (2.8) a densidade imprópria:

$$f_{pop_k}(t_{ki}) = (1-p_k)\alpha_k\lambda_k(t_{ki}\lambda_k)^{(\alpha_k-1)}\mathrm{e}^{-(t_{ki}\lambda_k)^{\alpha_k}}$$

Assim, a função de verossimilhança do k-ésimo grupo, considerando a distribuição Weibull, é dada por:

$$L_k(\boldsymbol{\theta_k}|D) = \prod_{i=1}^{n_k}\left[(1-p_k)\alpha_k\lambda_k(t_{ki}\lambda_k)^{(\alpha_k-1)}\mathrm{e}^{-(t_{ki}\lambda_k)^{\alpha_k}}\right]^{\delta_{ki}}\left[p_k+(1-p_k)\mathrm{e}^{-(t_{ki}\lambda_k)^{\alpha_k}}\right]^{1-\delta_{ki}}$$

2.4.3 Modelo Gompertz

A distribuição Gompertz é outra distribuição utilizada com uma certa frequência na em estudos de sobrevivência. Proposta por Benjamin Gompertz (1825), essa é uma distribuição amplamente utilizada em áreas de atuária, demografia e outros estudos de sobrevivência (Gieser *et al.*, 1998)

Na figura 2.8 temos representado o comportamento das funções de densidade, sobrevivência e risco, considerando diferentes parâmetros forma a, e de escala b, com $t \in [0, 5]$

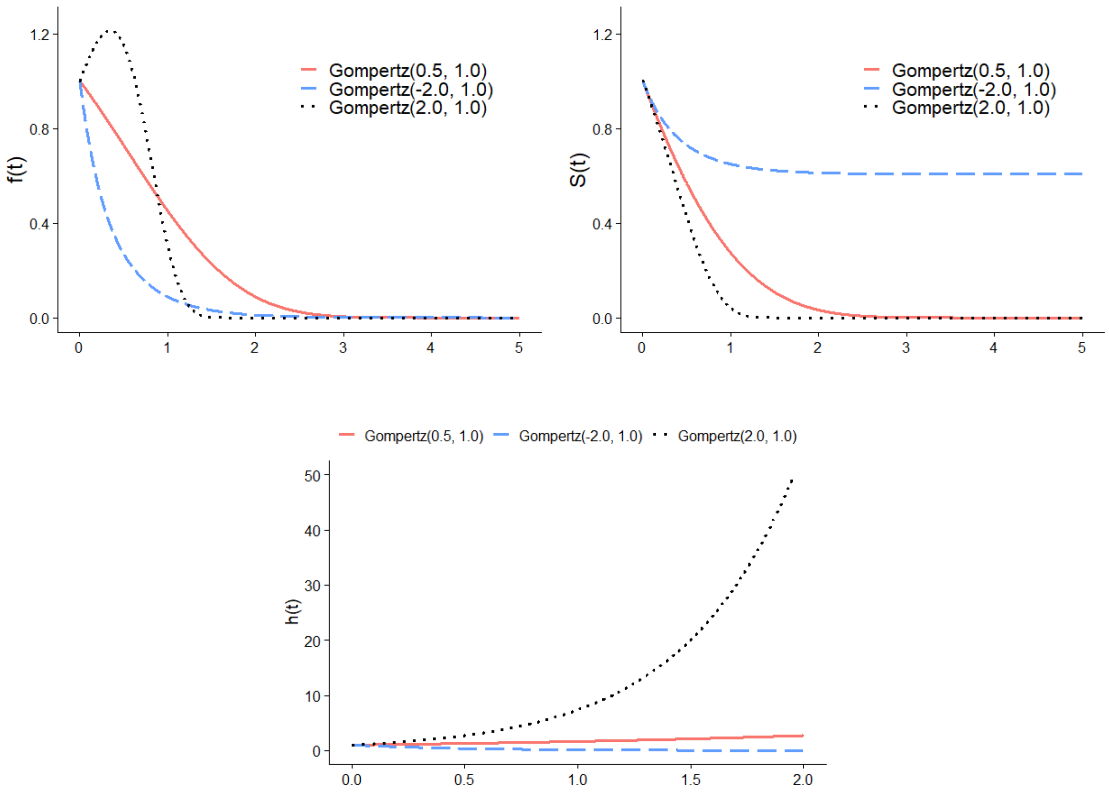

Figura 2.8: Comportamento das funções de densidade, sobrevivência e risco da distribuição Gompertz.

Atribuindo a distribuição Gompertz para os tempos de vida, teremos a função de densidade $f(t_i)$ definida como:

$$f(t_i) = a\mathrm{e}^{bt_i}\mathrm{e}^{\left[-\frac{a}{b}(\mathrm{e}^{bt_i}-1)\right]}, \quad a > 0, \quad b > 0,$$

A função de sobrevivência:

$$S(t_i) = \mathrm{e}^{\frac{a}{b}(\mathrm{e}^{bt_i}-1)}$$

Igualmente ao apresentado nas distribuições exponencial e Weibull, de (2.1) e (2.7) obtemos a sobrevivência imprópria do k-ésimo grupo:

$$S_{pop_k}(t_{ki}) = p_k + (1-p_k)\mathrm{e}^{\frac{a_k}{b_k}(\mathrm{e}^{b_k t_{ki}}-1)}$$

e de (2.8) a densidade imprópria:

$$f_{pop}(t_{ki}) = (1-p_k)a_k\mathrm{e}^{b_k t_{ki}}\mathrm{e}^{\left[-\frac{a_k}{b_k}(\mathrm{e}^{b_k t_{ki}}-1)\right]}$$

Por fim, temos a função de verossimilhança do k-ésimo grupo, considerando a distribuição Gompertz, como sendo:

$$L_k(\boldsymbol{\theta_k}|D) = \prod_{i=1}^{n_k}\left[(1-p_k)a_k\mathrm{e}^{b_k t_{ki}}\mathrm{e}^{\left[-\frac{a_k}{b_k}(\mathrm{e}^{b_k t_{ki}}-1)\right]}\right]^{\delta_{ki}}\left[p_k + (1-p_k)\mathrm{e}^{\frac{a_k}{b_k}(\mathrm{e}^{b_k t_{ki}}-1)}\right]^{1-\delta_{ki}}$$

2.5 Modelos de Regressão de Fração de Cura

Na modelagem tradicional consideramos populações homogêneas, ou seja, não existem características que distinguem os indivíduos em estudo. Entretanto, na grande maioria dos casos, trabalhamos com populações que apresentam uma certa heterogeneidade, contendo covariáveis que expressam as características de cada indivíduo.

Desse modo, os modelos de regressão são capazes de apresentar a influência de determinadas covariáveis na variável resposta. Para o modelo de mistura padrão especificado em 2.3, temos duas maneiras de considerar o efeito das covariáveis (Rodrigues *et al.*, 2009):

- **Efeito na fração de curados e não curados**: ou seja, o modelagem considera que as covariáveis tem efeito na fração de cura e na função de sobrevivência. Nesse caso, um dos modelos mais utilizados é o modelo semiparamétrico de mistura com fração de cura e riscos proporcionais, proposto por Kuk e Chen (1978), que define a sobrevivência populacional imprópria $S_{pop}(t_i|\boldsymbol{Z},\boldsymbol{X})$ como sendo:

$$S_{pop}(t_i|\boldsymbol{Z},\boldsymbol{X}) = p(\boldsymbol{X}) + (1-p(\boldsymbol{X}))S(t_i)^{\exp(\boldsymbol{Zb})},$$

e a função de densidade imprópria $f_{pop}(t_i|\boldsymbol{Z},\boldsymbol{X})$:

$$f_{pop}(t_i|\boldsymbol{Z},\boldsymbol{X}) = (1-p(\boldsymbol{X}))f(t_i)\exp(\boldsymbol{Zb})S(t_i)^{\exp(\boldsymbol{Zb})-1},$$

Desse modo, temos a função de verossimilhança definida como:

$$\begin{aligned} L(\boldsymbol{\theta}|D) &= \prod_{i=1}^{n} \left[f_{pop}(t_i|\boldsymbol{Z})\right]^{\delta_i} \left[S_{pop}(t_i|\boldsymbol{Z},\boldsymbol{x})\right]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left[(1-p(\boldsymbol{X}))f(t_i)\exp(\boldsymbol{Zb})S(t_i)^{\exp(\boldsymbol{Zb})-1}\right]^{\delta_i} \\ &\qquad \left[p(\boldsymbol{X}) + (1-p(\boldsymbol{X}))S(t_i)^{\exp(\boldsymbol{Zb})}\right]^{1-\delta_i} \end{aligned} \tag{2.11}$$

em que:

  - $\boldsymbol{X} = (x_1, x_2, ..., x_p)^t$, p-covariáveis a serem consideradas na fração de cura;
  - $\boldsymbol{Z} = (z_1, z_2, ..., z_p)^t$, p-covariáveis a serem consideradas na na fração de não curados;
  - Podendo $\boldsymbol{X}$ e $\boldsymbol{Z}$ apresentar covariáveis em comum, e situações em que as covariáveis que influenciam a fração de curados e não curados são as mesmas, ou seja, $\boldsymbol{X} = \boldsymbol{Z}$;
  - $p(\boldsymbol{X})$: probabilidade de cura dependendo do efeito da covariável $\boldsymbol{X}$;
  - $S(t|\boldsymbol{Z})$: função de sobrevivência dependendo do efeito da covariável $\boldsymbol{Z}$;
  - $\boldsymbol{b} = (b_0, b_1, ..., b_p)^t$ vetor de parâmetros a ser estimado para cada covariavel de $\boldsymbol{Z}$;
  - Sendo $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^t$ o vetor de parâmetros a ser estimado para cada covariável de $\boldsymbol{X}$.

- **Efeito apenas na fração de cura**: Nesse caso, a modelagem considera o efeito das covariáveis apenas na fração de cura $p(\boldsymbol{X})$. Desse modo, a função de sobrevivência populacional imprópria é dada por:

$$S_{pop}(t_i|\boldsymbol{X}) = p(\boldsymbol{X}) + (1 - p(\boldsymbol{X}))S(t_i)$$

Para modelar os efeitos das covariáveis na fração de cura, podemos utilizar diferentes funções de ligações (Peng e Dear, 2000). A seguir apresentamos as três principais funções de ligação de $p(\boldsymbol{X})$:

$$p(\boldsymbol{X}) = \frac{\exp(\boldsymbol{\beta X})}{1 + \exp(\boldsymbol{\beta X})}, \qquad \text{(Ligação logito)}$$

$$p(\boldsymbol{X}) = \Phi(\boldsymbol{\beta X}), \qquad \text{(Ligação probito)}$$

$$p(\boldsymbol{X}) = \exp(-\exp(\boldsymbol{\beta X})). \qquad \text{(Ligação complemento log-log)}$$

em que $\Phi()$ é a função de distribuição normal padrão.

Nesse trabalho exploramos o modelo de regressão para o modelo de mistura padrão que considera o efeitos das covariáveis na fração de indivíduos curados e não curados.

2.5.1 Modelo de Regressão de Fração de Cura Weibull

Utilizando a distribuição Weibull para caracterizar o tempo até a ocorrência do evento de interesse, no modelo de regressão de fração de cura apresentado em (2.11). Temos que a função de verossimilhança é dada por:

$$L(\boldsymbol{\theta}|D) = \prod_{i=1}^{n} \left[(1 - p(\boldsymbol{X}))\alpha\lambda(t_i\lambda)^{(\alpha-1)}\mathrm{e}^{-(t_i\lambda)^{\alpha}}\exp(\boldsymbol{Zb})\mathrm{e}^{-(t_i\lambda)^{\alpha}\exp(\boldsymbol{Zb})-1}\right]^{\delta_i} \left[p(\boldsymbol{X}) + (1 - p(\boldsymbol{X}))\mathrm{e}^{-(t_i\lambda)^{\alpha}\exp(\boldsymbol{Zb})}\right]^{1-\delta_i} \tag{2.12}$$

Permitindo estimar os parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{b}$, e fazendo o uso de uma função de ligação, estimamos a fração de cura $p(\boldsymbol{X})$.

2.6 Seleção de Modelos

Quando realizado ajustes de modelos, podemos ter como interesse identificar o quão bom é aquele modelo, ou realizar comparações, no intuito de identificar o melhor modelo possível para o problema em questão. Para isso, exitem diversos critérios para seleção de modelos, dentre eles, os mais utilizados são o teste da razão de verossimilhança (TRV), o critério de informação de Akaike (AIC) e o critério Bayesiano de Schwarz (BIC) (Cintra, 2021).

O critério de informação de Akaike é uma métrica que visa mensurar a qualidade do modelo de maneira simples. Este critério é baseado na Divergência de Kullback-Leibler, que é uma medida da "distância" entre o modelo identificado e um teórico "modelo real". Como o modelo real não é conhecido, Akaike desenvolveu uma forma de estimar esta distância através dos dados utilizados na modelagem, usando a função de verossimilhança e a ordem do modelo (Barreto, 2016).

Desse modo, a medida AIC, proposta por Akaike (1974), é definida como:

$$\text{AIC} = -2\log(L(\boldsymbol{\theta})) + 2p, \tag{2.13}$$

em que p é o número de parâmetros utilizados no ajuste do modelo.

À medida que a verossimilhança aumenta, o termo $-2\log(L(\boldsymbol{\theta}))$ decresce, enquanto o termo $2p$ cresce sempre que a ordem do modelo for maior. Dessa forma, o critério de Akaike pondera entre a adequação aos dados e a complexidade do modelo (Barreto, 2016).

Com a criação da medida proposta por Akaike diversas outras medidas foram criadas com base no AIC. Uma das mais utilizadas atualmente é o critério Bayesiano de Schwarz (BIC), proposta por Schwarz (1978), o qual funciona de maneira semelhante ao AIC, porém com outro termo de penalização.

A medida BIC é dada por:

$$\text{BIC} = -2\log(L(\boldsymbol{\theta})) + p\log(n), \tag{2.14}$$

em que p é o número de parâmetros utilizados no ajuste do modelo, e n o número de observações da amostra.

Capítulo 3

Aplicação

Na introdução e metodologia, discutimos o problema do alto número de censuras que ocorre em estudos de fidelização de clientes de serviço por assinatura, além da segmentação dos dados por diversas características.

Nessa seção, realizamos aplicação dos estudos em uma base de dados de planos de internet, disponível na plataforma Maven Analytics, tendo como objetivo estimar a probabilidade de um cliente permanecer fiel ao plano de internet por um longo período.

Os dados são referente a informações de clientes de serviços por assinatura de uma empresa da Califórnia, os quais foram coletados no segundo trimestre de 2022, no intuito de obter informações sobre *churn* de seus clientes.

Foram coletadas informações dos últimos 72 meses, totalizando 7043 clientes, em que cada registro de cliente contém detalhes sobre seus dados de assinatura: tipo de produto, tempo com o produto, status, dados demográficos, dados pessoais, etc.

3.1 Análise Exploratória

Iniciamos o estudo realizando uma breve análise exploratória, no intuito de trazer as principais características da base, e poder verificar algum tipo de anomalia nos dados, para então podermos realizar os ajustes dos modelos.

Como dito anteriormente, os dados trazem informações dos clientes que adquiriram o produto da empresa nos últimos 72 meses da data de coleta. O produto em questão é o serviço de disponibilização de internet, podendo existir diversas formas e tipos.

A base completa tem um total de 38 covariáveis, entretanto nesse estudo não fizemos o

uso de todas, e sim as que julgamos mais importantes. Na tabela 3.1 temos representado as principais covariáveis presente na base de dados, quais suas características e informações.

Tabela 3.1: Descrição das principais covariáveis da base.

Covariável	**Tipo**	**Descrição**
ID Cliente	Categórica	Covariável indicativa de cliente, de qual cliente são as informações.
Sexo	Categórica	Covariável indicativa do sexo do cliente. Possíveis valores: Feminino ou Masculino.
Estado civil	Categórica	Covariável indicativa de status casamento. Possíveis valores: Sim ou Não.
Idade	Quantitativa	Covariável indicativa da idade do cliente. Possíveis valores: valores numéricos inteiros.
Referencias	Quantitativa	Covariável indicativa do número de referencias/indicações do cliente. Possíveis valores: valores numéricos inteiros.
GB	Quantitativa	Covariável indicativa do valor de gigabytes da internet contratada pelo cliente. Possíveis valores: valores numéricos inteiros.
Salário	Quantitativa	Covariável indicativa do valor total da receita do cliente por mês. Possíveis valores: valores numéricos.
Tipo de Internet	Categórica	Covariável indicativa do tipo de internet contratada pelo cliente. Possíveis valores: “Cable”, “Fiber Optic”, “DSL” ou “Plus”.
Tempo	Quantitativa	Covariável indicativa do número em meses da permanência com o produto. Possíveis valores: valores numéricos inteiros.

Status	Categórica	Covariável indicativa de status do cliente. Possíveis valores: “Stayed” ou “Churned”.

Afim de representar o formato do conjunto de dados, construímos a Tabela 3.2, com apenas as principais informações, já pensando na aplicação em modelos de sobrevivência. Em que “Status” é a covariável indicativa do evento de interesse (“Churned”) e censura (“Stayed”), e “Tempo” a covariável indicativa do tempo até o evento.

Tabela 3.2: Representação da base de dados.

ID	Tipo de Internet	Status	Tempo
0002-ORFBO	Cable	Stayed	58
0003-MKNFE	Fiber Optic	Churned	9
0004-TLHLJ	DSL	Churned	39
0011-IGKFF	Plus	Stayed	13
0013-EXCHZ	Cable	Churned	71
⋮	⋮	⋮	⋮

Na covariável “Status”, ressaltamos que “Stayed” indica que o cliente não apresentou o evento de interesse, portanto consideramos como censura; e “Churned” indica que o cliente cancelou o serviço, ou seja, apresentou o evento de interesse *churn*. Também iremos representar “Churned”$= 1$ e “Stayed” $= 0$.

Com esses dados, tivemos como principal objetivo, estimar a fração de indivíduos que permanecem fiéis a assinatura de internet por um longo período. Para isso, utilizamos os modelos de mistura padrão.

Para que os modelos de mistura padrão seja um método relevante, é necessário que exista um certo número de censuras, permitindo assim trazer inferência sobre a fração de indivíduos censurados.

Na Figura 3.1 temos representado a proporção de *chrun* dos 7043 clientes observados.

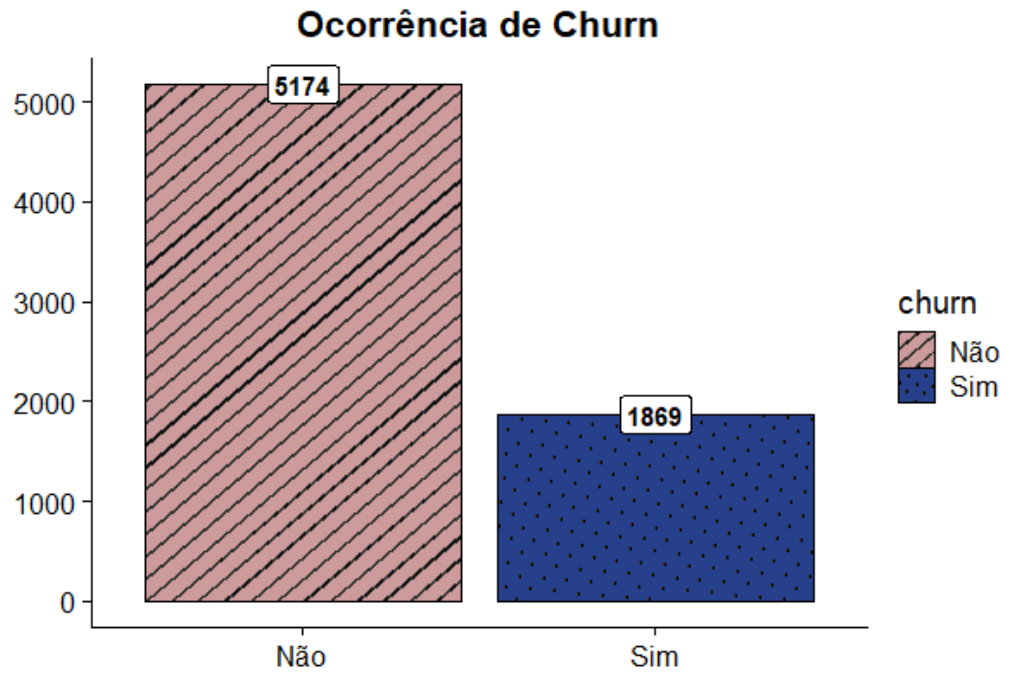

Figura 3.1: Gráfico de barras da quantidade de ocorrências ou não de *churn*.

Notamos uma grande fração de observações "Não", portanto um alto número de não ocorrência do evento de interesse, ou seja, um alto número de censuras, e temos como interesse verificar o desempenho dos modelos de mistura padrão em dados com essa característica.

3.1.1 Segmentação

Uma das principais covariáveis que observamos ao interpretar os dados, é a covariável "Tipo de Internet". Essa covariável traz informações sobre o tipo de internet que o cliente adquiriu, e por se tratar de diferentes tipos, diferentes serviços são oferecidos, portanto a experiencia obtida pelos clientes de diferentes tipos de internet são distintas.

Desse modo, segmentamos o estudo com base na covariável "Tipo de Internet" ("Cable", "Fiber Optic", "DSL" e "Plus"), no intuito de observar e comparar as estimativas obtidas para cada grupo de indivíduos com produtos diferentes.

Na Tabela 3.3, representamos a proporção de *churn* e censura para cada grupo.

Tabela 3.3: Ocorrência de *churn* em cada plano.

	Cable	**Fiber Optic**	**DSL**	**Plus**
Ocorrência *churn*	25.66%	40.25%	18.58%	7.40%
Censura	74.34%	59.75%	81.42%	92.60%

Afim de tornar a visualização mais clara, representamos as proporções de *churn* e censura para cada grupo na Figura 3.2.

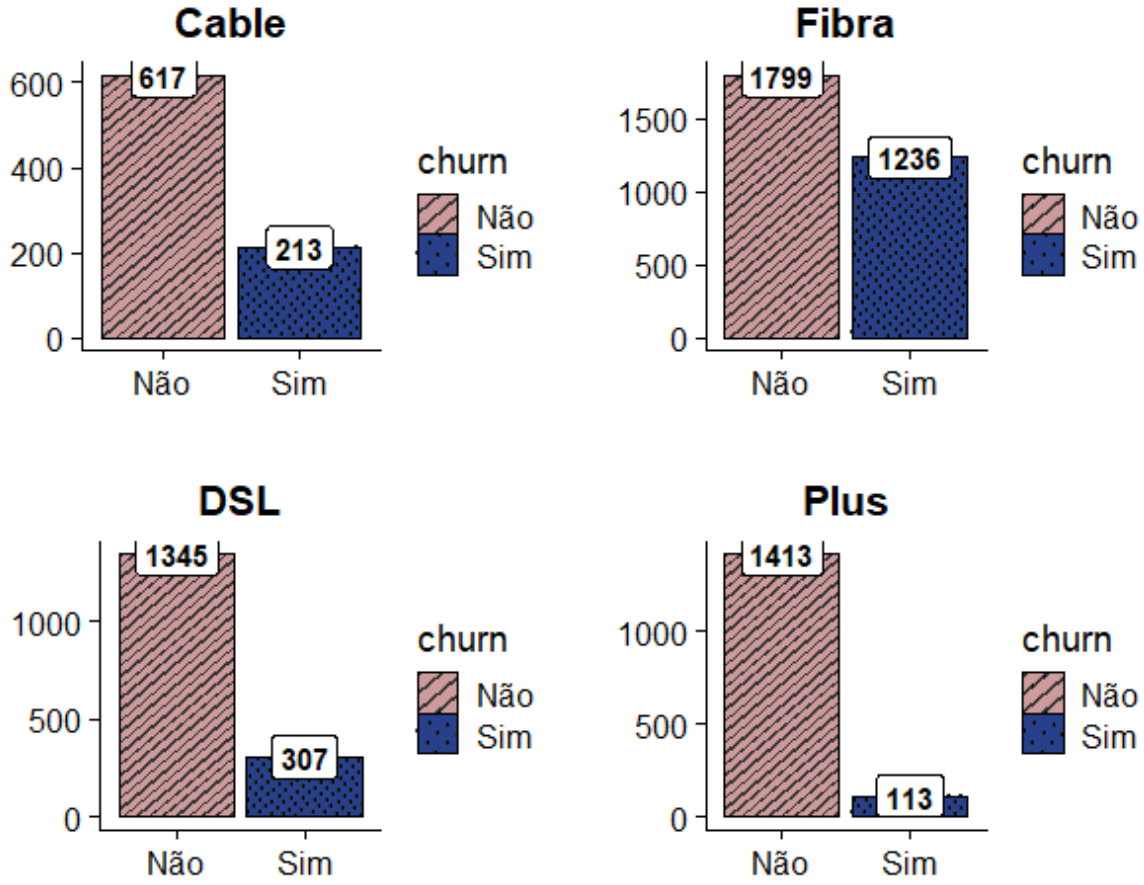

Figura 3.2: Gráfico de barras da quantidade de ocorrências ou não de *churn* segmentado por tipo de plano.

Por se tratar de diferentes produtos, já esperávamos diferentes proporções de ocorrência de *churn*, nos diferentes tipos de internet. Mostrando que a segmentação dos dados faz sentido para o problema em questão.

3.1.2 Kaplan-Meier

No intuito de observar o comportamento da função de sobrevivência para as diferentes segmentações, utilizamos o estimador de Kaplan-Meier. O resultado obtido para cada grupo está representado na Figura 3.3

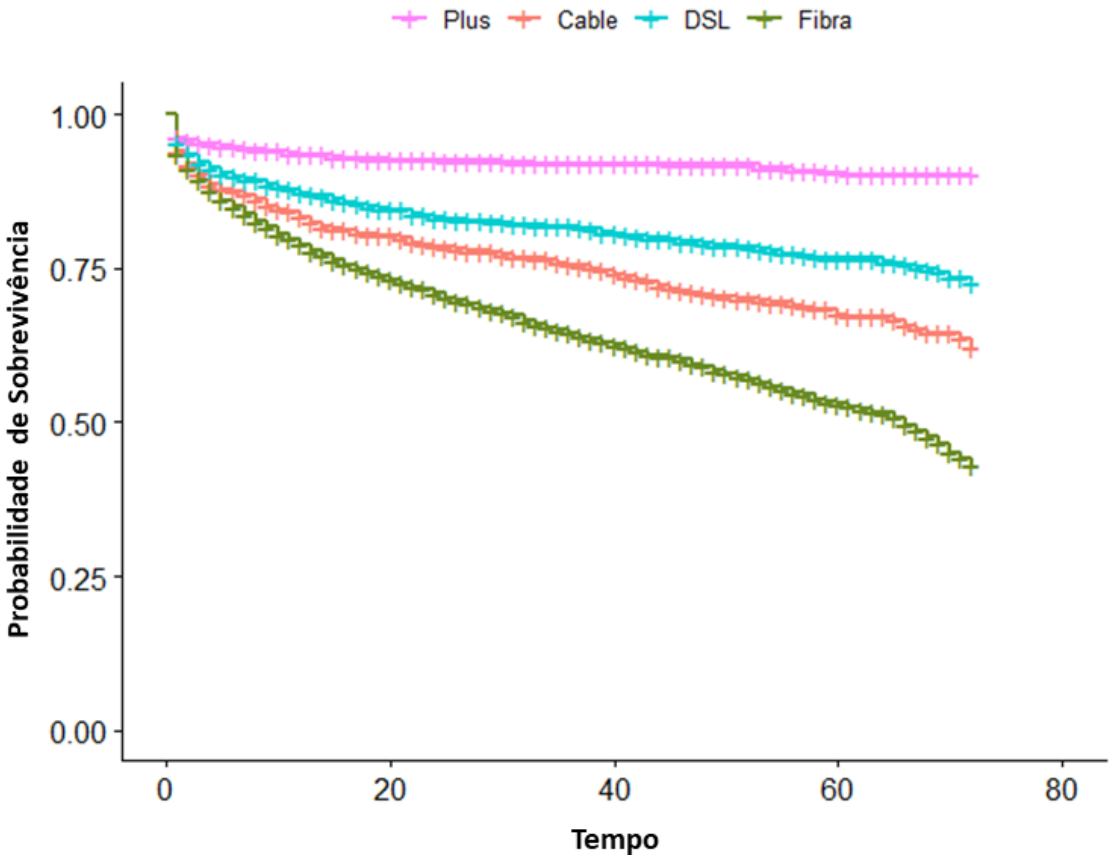

Figura 3.3: Curvas de sobrevivência estimadas pelo método de Kaplan-Meier para cada grupo.

É possível notar que o comportamento das curvas de sobrevivência para os quatro grupos são completamente diferentes, além de que, todas as curvas terminam bem distante de 0, o que é um indicativo de que os dados tem um alto número de observações censuradas, e que o uso do modelo de mistura padrão pode ser adequado e trazer estimativas relevantes para cada grupo.

Contudo, pudemos notar por meio da análise exploratória, e principalmente pelo Kaplan Meier, que o ajuste de modelos de mistura padrão aparenta ser uma opção viável para os dados. Desse modo, realizamos os ajustes com o objetivo de estimar a fração de indivíduos censurados para cada segmentação. De forma mais especifica, utilizamos os modelos de mistura padrão para estimar a probabilidade do cliente continuar fiel por um

longo período, para cada tipo de internet.

3.2 Modelos Ajustados

Nessa seção apresentamos os resultados obtidos nos ajustes dos modelos, e suas respectivas interpretações.

3.2.1 Definições e Resultados

Para obter o modelo que melhor se adequava aos dados, realizamos os ajustes dos seguintes modelos:

- **Modelo de mistura padrão Exponencial**, o qual denotamos como **Modelo 1**;
- **Modelo de mistura padrão Weibull**, o qual denotamos como **Modelo 2**;
- **Modelo de mistura padrão Gompertz**, o qual denotamos como **Modelo 3**.

Para cada modelo definido, realizamos ajustes separados para cada segmentação em estudo, ou seja, cada modelo foi abordado nas quatro segmentações.

Desse modo, realizando os ajustes, observamos inicialmente o comportamento da curva de sobrevivência estimada para cada ajuste, comparando os resultados obtidos por segmentação. Em geral, as estimativas obtidas foram bem próximas ao Kaplan Meier, nas quatro segmentações, indicando bons resultados.

No Apêndice A, temos representado as estimativas da curva de sobrevivência, e dos parâmetros estimados de cada modelo e segmentação.

Antes de realizar a interpretação dos parâmetros estimados de cada segmentação, principalmente o da fração de cura p, escolhemos o melhor modelo ajustado. Portanto, para definir qual dos três modelos ajustados é o que apresenta melhores resultados, calculamos o AIC e o BIC.

Na Tabela 3.4 temos representado o valor estimado do AIC e BIC para cada um dos modelos.

Tabela 3.4: AIC e BIC estimados nos modelos 1, 2 e 3, para cada segmentação.

Segmentação	**Modelo**	**AIC**	**BIC**
Cable	Modelo 1	2427.064	2436.507
Cable	Modelo 2	2393.941	2408.105
Cable	Modelo 3	2415.053	2429.218
Fibra	Modelo 1	13265.460	13277.490
Fibra	Modelo 2	13140.490	13158.550
Fibra	Modelo 3	13233.700	13251.760
DSL	Modelo 1	3666.773	3677.592
DSL	Modelo 2	3618.945	3635.174
DSL	Modelo 3	3642.223	3658.453
Plus	Modelo 1	1459.898	1470.559
Plus	Modelo 2	1430.704	1446.695
Plus	Modelo 3	1432.641	1448.632

Em todas as segmentações observamos que o **Modelo 2**, o modelo de mistura padrão Weibull, foi o que apresentou melhores resultados, seguido pelo modelo Gompertz, e por ultimo o modelo exponencial.

Na figura 3.4 temos representado as estimativas das curvas de sobrevivência $S(t)$, e os respectivos Kaplan-Meier para as quatro segmentações, pelo modelo de mistura padrão Weibull.

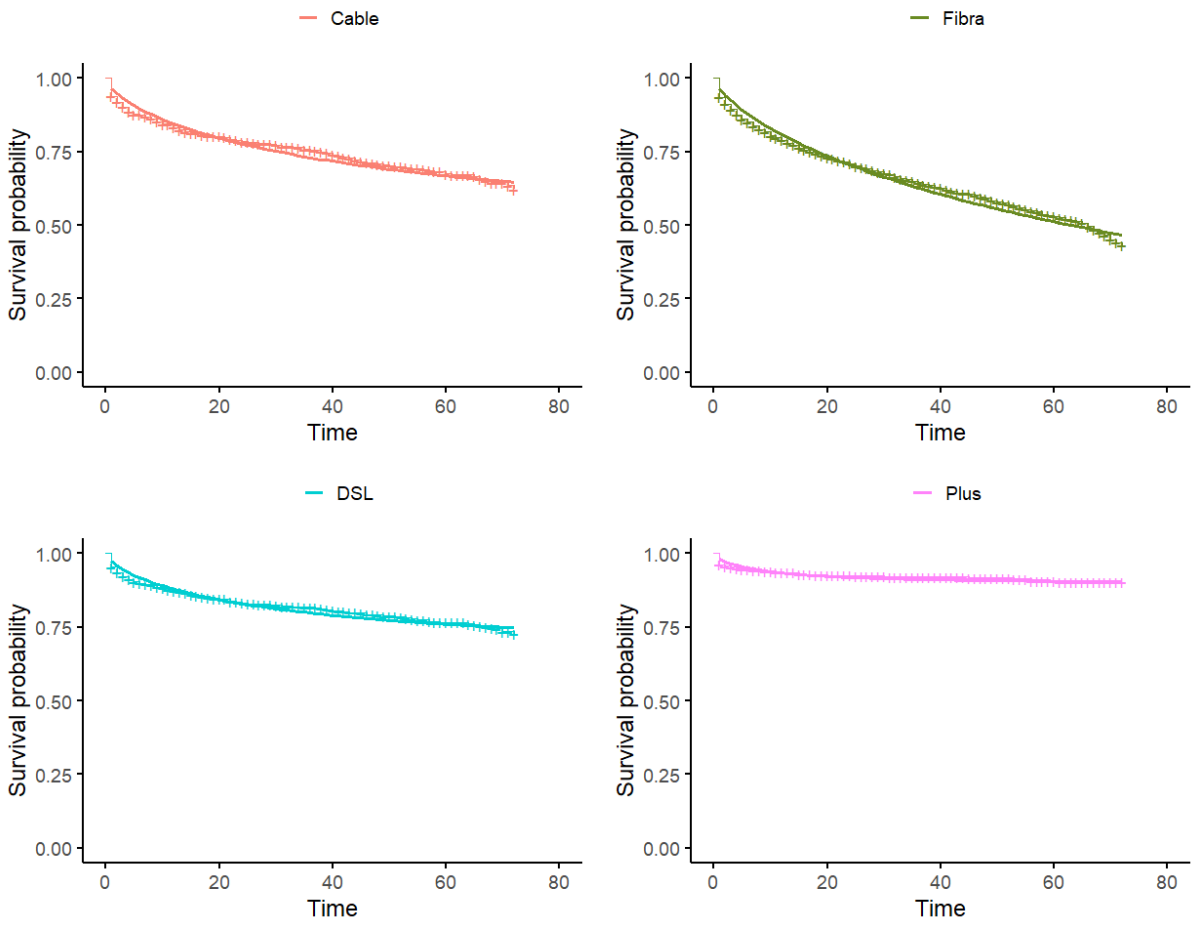

Figura 3.4: Curva de sobrevivência estimada pelo modelo de mistura padrão Weibull.

Pela visualização gráfica, conseguimos observar indícios que os ajustes apresentam bons resultados, para a fração de não curados, pois as estimativas de $S(t)$ estão próximas ao Kaplan Meier. Entretanto nosso objetivos é trazer a estimativa da fração de cura, portanto observamos os parâmetros estimados dos modelos.

As estimativas dos parâmetros do **Modelo 2**, para cada tipo de internet, estão representado na Tabela 3.5.

Tabela 3.5: Parâmetros estimados do Modelo 2.

Parâmetro	Segmentação	Estimativa	IC (95%)
p	Cable	0.6326	(0.5816 - 0.6808)
	Fibra	0.3855	(0.3449 - 0.4277)
	DSL	0.7440	(0.7143 - 0.7715)
	Plus	0.9051	(0.8865 - 0.9210)

λ	Cable	51.801	(19.917 - 134.72)
	Fibra	106.00	(95.400 - 117.00)
	DSL	32.407	(18.956 - 55.401)
	Plus	9.9599	(6.4968 - 15.269)
α	Cable	0.6721	(0.5731 - 0.7882)
	Fibra	0.7100	(0.6750 - 0.7460)
	DSL	0.6905	(0.6109 - 0.7805)
	Plus	0.6800	(0.5802 - 0.7970)

Podemos notar que as estimativas de p para os diferentes tipos de internet estão próximas as proporções de censuras, apresentadas na Tabela 3.3, indicativo que os modelos tiveram um bom desempenho.

Entretanto, ainda não realizamos a interpretação dos parâmetros, pois a seguir exploramos a modelagem do modelo de mistura padrão na presença de covariáveis, com objetivo de verificar se o desempenho do modelo é afetado.

3.2.2 Regressão

Apos obter estimativas gerais com o modelo de mistura padrão Weibull sem a presença de covariáveis, para cada tipo de internet, decidimos explorar a modelagem do modelo de mistura padrão com regressão, no intuito de observar se incorporação de covariáveis causaria um impacto positivo nas estimativas da fração de cura e no resultados do estudo.

Analisando as covariáveis presente na base de dados, iniciamos os ajustes considerando as covariáveis “GB”, “Salário” e “Idade”, pois foram as informações que julgamos mais importantes para o contexto da aplicação. Desse modo, realizamos o ajuste do modelo de mistura padrão Weibull, considerando as covariáveis “GB”, “Salário” e “Idade”, tanto na fração de curados, quanto na fração de não curados.

Os ajustes para todas as segmentações foram realizados pelo software R, com o auxílio da biblioteca ”cuRe”. Tivemos um alto custo computacional para obter os resultados considerando as três covariáveis citadas anteriormente, tornando inviável a inclusão de mais covariáveis.

Realizando o ajuste nas diferentes segmentações, observamos que “GB” e “Salário” não eram significativas para o modelo, desse modo, o ajuste final é dado apenas com a inclusão da covariavel “Idade”.

Na Tabela 3.6 temos representando as estimativas obtidas na segmentação “Cable”, pelo modelo de mistura padrão Weibull na presença da covariável idade na fração de curados e não curados.

Tabela 3.6: Parâmetros estimados do modelo de regressão Weibull na presença da covariável idade, para a segmentação Cable.

Parâmetro	Estimativa	Desvio
λ	-2.1363	0.4116
α	-0.3549	0.0744
β_0	1.3470	0.4290
β_1	-0.0243	0.0087
b_1	-0.0087	0.0074

Sendo β_0 e β_1 os parâmetros estimados da fração de cura, e b_1 o parâmetro estimado da função de sobrevivência.

A função ligação utilizada no ajuste do modelo foi a logito, desse modo, com as estimativas obtidas na Tabela 3.6, e a definição da ligação logito apresentada na seção 2.5 da metodologia, podemos calcular a proporção de indivíduos curados.

$$p_{(\text{cable})}(\boldsymbol{X}) = \frac{\exp(\boldsymbol{\beta X})}{1+\exp(\boldsymbol{\beta X})},$$

$$p_{(\text{cable})}(\boldsymbol{X}) = \frac{\exp(\beta_0 + \beta_1 \cdot \text{Idade})}{1+\exp(\beta_0 + \beta_1 \cdot \text{Idade})}$$

Para representar um valor geral de idade para todo o grupo de pessoas com internet “cable”, utilizamos a idade média, a qual apresentou um total de 44.70 anos. Assim sendo,

temos:

$$\begin{aligned} p_{(\text{cable})}(\boldsymbol{X}) &= \frac{\exp(1.3470 - 0.0243 \cdot 44.70)}{1 + \exp(1.3470 - 0.0243 \cdot 44.70)} \\ &= 0.5644 \end{aligned}$$

Ou seja, a fração de cura estimada pelo modelo de regressão considerando a covariável idade, foi de 56.44%.

Também realizamos os mesmos procedimentos para as segmentações "Fibra", "DSL" e "Plus", as estimativas dos parâmetros dos ajustes estão representadas no Apêndice A.

E a fração de cura estimada pelos modelos, para as segmentações "Fibra", "DSL" e "Plus", são dadas por:

$$\begin{aligned} p_{(\text{fibra})}(\boldsymbol{X}) &= \frac{\exp(0.1056 - 0.0115 \cdot 49.80)}{1 + \exp(0.1056 - 0.0115 \cdot 49.80)} \\ &= 0.3851 \end{aligned}$$

$$\begin{aligned} p_{(\text{DSL})}(\boldsymbol{X}) &= \frac{\exp(2.1242 - 0.0276 \cdot 44.83)}{1 + \exp(2.1242 - 0.0276 \cdot 44.83))} \\ &= 0.7086 \end{aligned}$$

$$\begin{aligned} p_{(\text{plus})}(\boldsymbol{X}) &= \frac{\exp(2.2412 - 0.0002 \cdot 42.77)}{1 + \exp(2.2412 - 0.0002 \cdot 42.77)} \\ &= 0.9032 \end{aligned}$$

Em geral, as estimativas obtidas foram bem próximas ao observado no modelo sem a presença de covariáveis.

Destacamos que a modelagem com a presença de covariáveis permite uma maior aplicação na interpretação dos resultados. Na Tabela 3.7, apresentamos as estimativas de fração de cura considerando idades a partir de 20 anos:

Tabela 3.7: Comparação entre os modelos ajustados.

Idade	$p(X)$
20	0.7026
30	0.6494
40	0.5922
44.70*	0.5644
50	0.5323
60	0.4716
70	0.4116
80	0.3542

Podemos observar que quanto maior a idade da pessoa, menor é a fração de cura, ou seja, quanto maior a idade, menor é a probabilidade do cliente permanecer fiel a internet do tipo cable.

Para tornar a visualização mais clara, construímos o gráfico representado na Figura 3.5, o qual apresenta as estimativas de fração de cura para todas as idades, na segmentação cable.

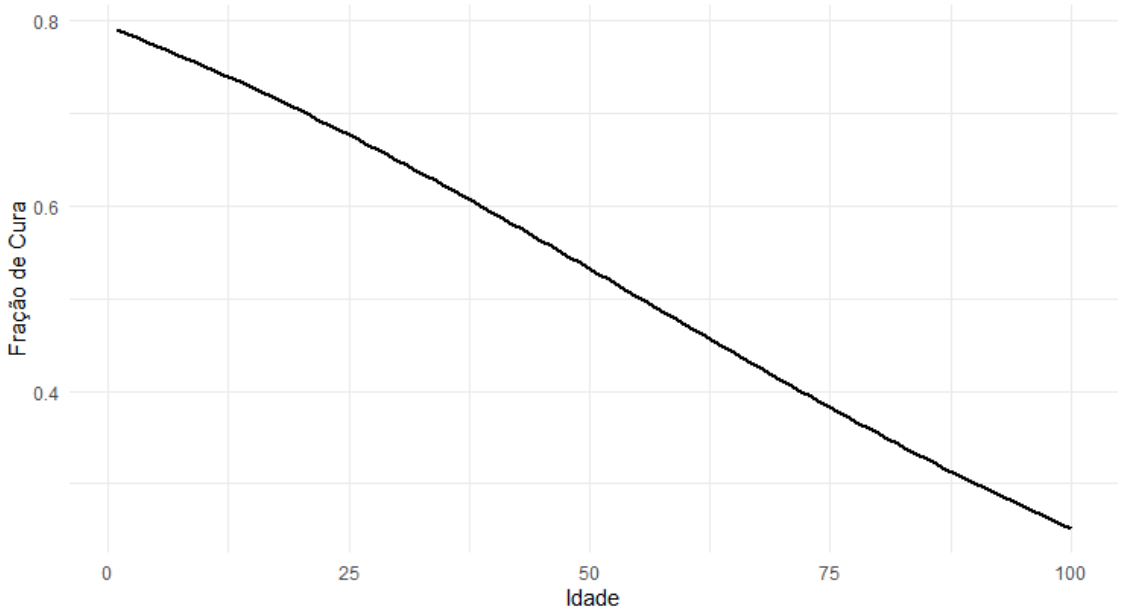

Figura 3.5: Estimativas de $p(x)$ para as idades de 0 a 100 anos.

A Figura 3.5 apresenta o comportamento das estimativas para as idades especificadas, permitindo traçar ações e soluções relacionadas a fidelização, para cada perfil de pessoa que possuem a internet do tipo cable.

Comparações com as estimativas obtidas nas demais segmentações ("Fibra", "DSL" e "Plus") também permite observar características que auxiliem nas mais diversas tomadas de decisões relacionadas a fração de cura, no contexto da aplicação, tomadas de decisões relacionadas a fidelização de clientes.

Contudo, pelo modelo de mistura padrão Weibull sem a presença de covariável, estimamos a fração de cura para cada segmentação de uma maneira geral, permitindo caracterizar e comparar os tipos de internet. Já no modelo modelo de mistura padrão Weibull com regressão, utilizamos a característica de idade na modelagem, permitindo estimar a fração de cura para a idade desejada, ampliando a interpretação e aplicação dos resultados.

Capítulo 4

Conclusão

Durante este trabalho, estudamos a modelagem de fração de cura pelo método de mistura padrão, considerando situações com e sem a presença de covariáveis. Também exploramos as distribuições Exponencial, Weibull e Gompertz para caracterizar o tempo até a ocorrência do evento.

Aplicamos os estudos em um problema de fidelização de clientes de serviços de assinatura, mais especificamente, em um problema de ocorrência de *churn*, para dados segmentados. O alto número de censuras em dados coletados em estudos de fidelização é uma característica comum, característica a qual também é observada na modelagem de fração de cura. Desse modo, o objetivo proposto foi de utilizar da modelagem de fração de cura para estimar a probabilidade de um cliente permanecer fiel a um produto, levando em conta as características tanto do cliente quanto do produto.

No Capitulo 3, apresentamos as informações sobre a aplicação. Utilizamos uma base de uma empresa que fornece serviços de internet, trazendo dados sobre ocorrência de *churn* de seus clientes, e suas características. Observamos que os dados podiam ser segmentados de acordo com os diferentes tipos de internet, já que cada tipo apresentava padrões de informação distintos. Portanto, realizamos toda a análise considerando a segmentação dos dados por tipo de internet.

Ajustando os modelos sem a presença de covariáveis, observamos que o modelo de mistura padrão Weibull foi o que apresentou melhores resultados, em todas as segmentações em estudo. Permitindo estimar a probabilidade de cura para cada segmentação, ou seja, estimar a probabilidade de um cliente permanecer fiel ao serviço, para os diferentes tipos de internet.

Em seguida, a fim de verificar se a inclusão de covariáveis no modelo Weibull causaria

um impacto positivo nos resultados, apresentamos o modelo de mistura padrão Weibull com a covariável idade na modelagem. Utilizando a média de idade, estimamos a fração de cura geral de cada plano de internet, permitindo observar a probabilidade de um cliente permanecer fiel ao produto. Entretanto, também apresentamos os resultados das estimativas para cada idade, ou seja, a probabilidade de uma pessoa com idade x, sendo $x \in [0, 100]$, permanecer fiel ao plano de internet analisado, ampliando os resultados e aplicações para todas as segmentações.

Referências Bibliográficas

Akaike, H. (1974). *PA new look at the statistical model identification.*. IEEE Transactions on Automatic Control, 19, 716–723.

Barreto, G. e Sobral, T. E. L. (2016). *Utilização dos critérios de informação na seleção de modelos de regressão linear*. Proceeding Series of the Brazilian Society of Applied and Computational Mathematics, Vol. 4, N. 1.

Berkson, J. & Gage, R. (1952). *Survival curve for cancer patients following treatment*. Journal of the American Statistical Association, 47(259), 501–515.

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.

Cassela, G. Berger, R. L. (2002). *Statistical inference, volume 2*. Duxbury Pacific Grove, CA.

Cintra, A. C. H. (2021). *Análise de sobrevivência em marketing, considerando o modelo exponencial com fragilidade compartilhada, na predição de churn.*. Dissertação de Graduação, Universidade Federal de São Carlos.

Collet, D. (2015). *Modelling survival data in medical research*. CRC press.

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

Fogo, J. C. (2007). *Modelo de regressão para um processo de renovação Weibull com termo de fragilidade*. Tese (Doutorado em agronomia), Escola superior de Agronomia Luiz de Queiroz, Universidade de São Paulo, SP.

Gieser, P. W., Chang, M. N., Rao, P., Shuster, J. e Pullen, J. (1998). *Modelling cure rates*

using the gompertz model with covariate information. Statistics in medicine, 17(8), 831–839.

Kuk, A. e Chen, C. (1978). *A mixture model combining logistic regression with proportional hazards regression*. Biometrika, 79, 531-541.

Lawless, J. F. (2011). *Statistical models and methods for lifetime data, volume 362*. John Wiley & Sons.

Lu, J. e Park, O. (2003). Modeling customer lifetime value using survival analysis—an application in the telecommunications industry. *Data Mining Techniques*, páginas 120–128.

Maller, R. A. e Zhou, X. (1996). Survival analysis with long-term survivors. *New York: Wiley*.

Peng, Y. e Dear, K. (2000). *A nonparametric mixture model for cure rate estimation*. Biometrics, 56, 237-243.

Rodrigues, J., Castro, M., Cancho, V. e Balakrishnan, N. (2009). *Poisson cure rate survival models and an application to a cutaneous melanoma data.*. Journal of Statistical Planning and Inference, v.139, p.3605-3611.

Schwarz, G. (1978). *Estimating the dimension of a model.*. The Annals of Statistics, 6, 461–464.

Apêndice A

Tabelas e gráficos

A seguir, temos representado as curvas de sobrevivência estimada pelo modelo de mistura padrão Exponencial, e modelo de mistura padrão Gompertz.

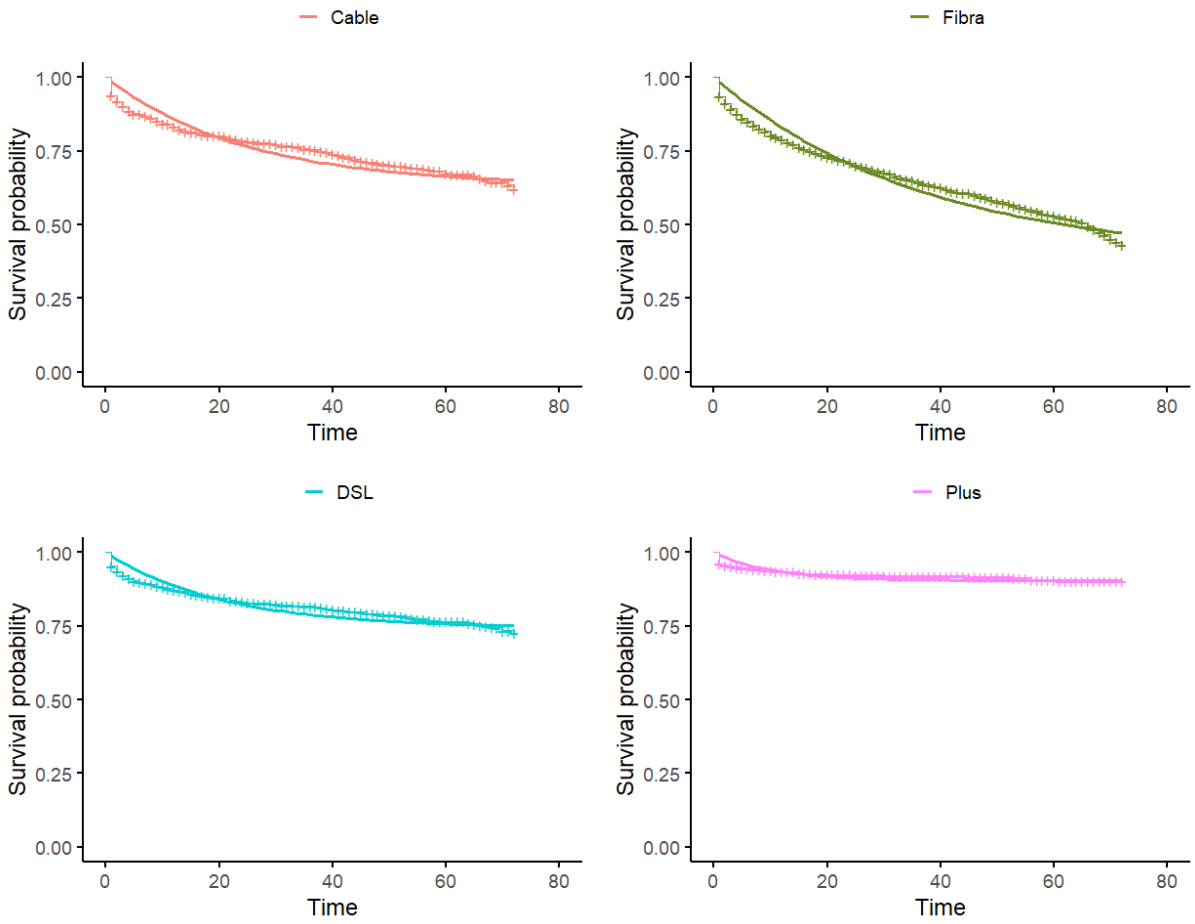

Figura A.1: Estimativas obtida no modelo de mistura padrão Exponencial.

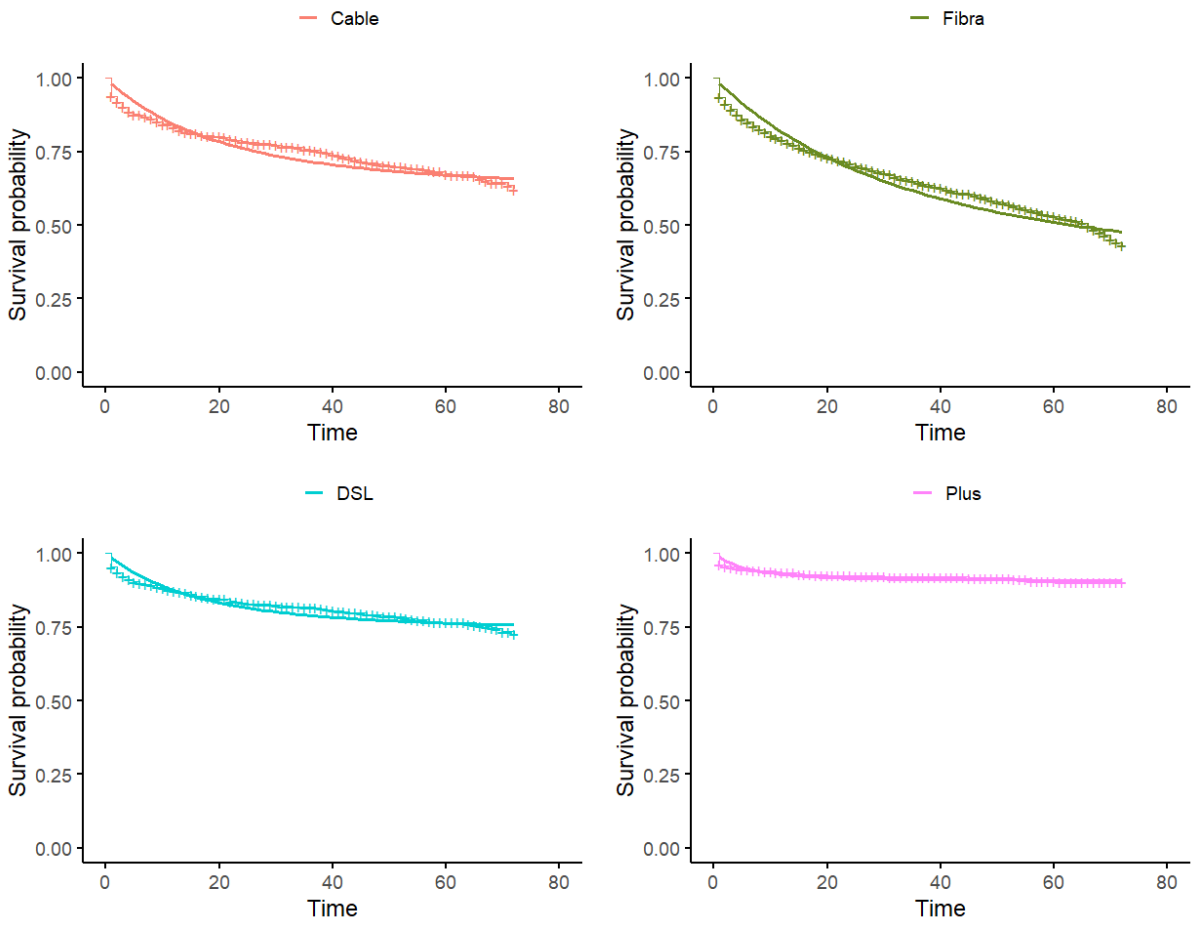

Figura A.2: Estimativas obtida no modelo de mistura padrão Gompertz.

Na Tabela A.1 representamos as estimativas obtidas no ajuste do modelo de mistura padrão Weibull com a presença da covariável "Idade", para a segmentação fibra.

Tabela A.1: Parâmetros estimados do modelo de regressão, segmentação Fibra.

Parâmetro	Estimativa	Desvio
λ	-3.5461	0.1753
α	-0.4470	0.0789
β_0	0.1056	0.2372
β_1	-0.0115	0.0047
b_1	-0.0011	0.0032

Na Tabela A.2 representamos as estimativas obtidas no ajuste do modelo de mistura padrão Weibull com a presença da covariável "Idade", para a segmentação DSL.

Tabela A.2: Parâmetros estimados do modelo de regressão, segmentação DSL.

Parâmetro	Estimativa	Desvio
λ	-1.5180	0.3156
α	-0.3477	0.0589
β_0	2.1243	0.2977
β_1	-0.0276	0.0075
b_1	-0.0183	0.0067

Na Tabela A.3 representamos as estimativas obtidas no ajuste do modelo de mistura padrão Weibull com a presença da covariável "Idade", para a segmentação plus.

Tabela A.3: Parâmetros estimados do modelo de regressão, segmentação Plus.

Parâmetro	Estimativa	Desvio
λ	-1.7957	0.4296
α	-0.3833	0.0810
β_0	2.2412	0.3495
β_1	-0.0002	0.0075
b_1	0.0053	0.0088