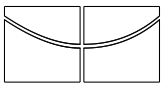

Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Dissertação de Mestrado

Modelo de regressão Weibull discreto com fração de cura em dados de sobrevivência

por

Carolina Andrade Silva

Orientadora: Prof.ª Dr.ª Cira Etheowalda Guevara Otiniano

Coorientador: Prof. Dr. Eduardo Yoshio Nakano

Brasília, 2015

Carolina Andrade Silva

Modelo de regressão Weibull discreto com fração de cura em dados de sobrevivência

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

Orientadora: Prof.ª Dr.ª Cira Etheowalda Guevara Otiniano

Coorientador: Prof. Dr. Eduardo Yoshio Nakano

Universidade de Brasília

Brasília, Dezembro de 2015

Agradecimentos

Agradeço primeiramente a Deus por ter me permitido enfrentar todos os obstáculos para alcançar mais essa vitória.

A minha mãe por sempre acreditar em mim, por estar ao meu lado em todos os momentos me dando apoio incondicional, por me motivar sempre que preciso, por ter feito sempre o seu melhor para me proporcionar uma boa educação, pelo seu amor e dedicação e por ser um exemplo de força e determinação para mim.

Ao Brunno que desde o início desta caminhada esteve ao meu lado me apoiando em todos os momentos. Obrigada por ser meu grande companheiro, por me trazer a calma nos momentos de difíceis e por todo amor, carinho e compreensão quando eu mais precisei.

Aos professores Cira Etheowalda Guevara Otiniano e Eduardo Yoshio Nakano pela orientação, pela confiança e por somarem e compartilharem seus conhecimentos para a realização deste trabalho.

A minha família, em especial ao meu primo Carlos Andrade por me estender a mão na minha mudança para Brasília, fazendo com que eu alcançasse mais um degrau dentre os que faltavam até o mestrado e por acreditar que eu chegaria até aqui.

As minhas amigas da Universidade Federal da Bahia por todos os momentos especiais que passamos juntas na graduação. Aos amigos que fiz em Brasília por alegrarem os meus dias. Aos meus colegas de trabalho pelo incentivo. Aos meus colegas e amigos do mestrado, em especial a Raquel e Thuany, pois estivemos sempre juntas dando apoio uma para a outra, nos ajudando sempre, dividindo as angústias nos momentos difíceis e compartilhando os momentos felizes e de descontração.

Por fim, a todos que direta ou indiretamente me acompanharam e fizeram com que esse objetivo fosse alcançado.

Resumo

Este trabalho apresenta uma formulação de um modelo de regressão para dados de tempo discretos com fração de cura. Para tanto, foi considerado o modelo de mistura, no qual os tempos são modelados através da distribuição Weibull discreta e a probabilidade de cura é modelada a partir de covariáveis utilizando a função de ligação logito. Os parâmetros do modelo foram estimados pelo algoritmo EM usando o software `R`. A avaliação do método proposto foi feita com simulações Monte Carlo utilizando amostras de tamanho $n = 250$, $n = 500$ e $n = 1.000$. O modelo proposto foi ilustrado em dois conjuntos de dados reais com tempos de sobrevivência discretos e presença de uma covariável. Os gráficos das curvas de sobrevivência mostraram que as curvas de sobrevivências estimadas a partir do modelo de regressão Weibull discreto com fração de cura foram bem próximas àquelas obtidas via estimador de Kaplan-Meier, confirmando um bom ajuste do modelo.

Palavras-chave: Distribuição Weibull discreta; Fração de cura; Covariáveis; Algoritmo EM; Modelo de mistura; Modelo de regressão.

Abstract

This work presents a formulation of a regression model for discrete time survival data with cure rate. Therefore, it was considered a mixture model, where the times were modeled though a Discrete Weibull distribution and the probability of cure is modeled by covariates using the logit link function. The parameters of the model were estimated by the EM algorithm using R software. To evaluate the proposed model, Monte Carlo simulations were performed using three different sample sizes: $n = 250$, $n = 500$ and $n = 1.000$. The proposed model was illustrated with two real data sets of discrete survival data with one covariate. The graphs of the survival curves showed that the survival estimated from the discrete Weibull regression model with cure fraction were very close to those obtained from Kaplan-Meier estimator, confirming a good fit of the model.

Keywords: Discrete Weibull distribution; Cure rate; Covarite; EM algorithm; Mixed model; Regression model.

Lista de Tabelas

Lista de Figuras

Lista de abreviaturas e siglas

E - Distribuição exponencial

EM - *Expectation Maximization*

EMV - Estimador de máxima verossimilhança

FC - Fração de cura

FCE - Modelo exponencial com fração de cura

FCG - Modelo geométrico com fração de cura

FCW - Modelo Weibull com fração de cura

FCWD - Modelo Weibull discreto com fração de cura

EKM - Estimador de Kaplan-Meier

KM - Kaplan-Meier

SR - Sob risco

SRE - Modelo exponencial para indivíduos sob risco

SRG - Modelo geométrico para indivíduos sob risco

SRW - Modelo Weibull para indivíduos sob risco

SRWD - Modelo Weibull discreto para indivíduos sob risco

W - Distribuição Weibull

WD - Distribuição Weibull discreta

Sumário

Capítulo 1

Introdução

A análise de sobrevivência pode ser definida como um conjunto de métodos utilizados para analisar dados cuja variável resposta representa o tempo até a ocorrência de um evento de interesse, também chamado de tempo de falha. A principal característica deste tipo de dado é a presença de censura, pois dados que não apresentam censura e são completamente observados podem ser analisados através de técnicas estatísticas clássicas. Entretanto, nem sempre é possível acompanhar todos os indivíduos até a sua falha e os dados censurados, apesar de incompletos, são muito importantes por fornecerem uma informação parcial da resposta, sendo então necessário o uso da análise de sobrevivência.

Por ser uma variável aleatória, o tempo de falha pode ter seu comportamento descrito por uma distribuição de probabilidade, sendo a distribuição Weibull amplamente utilizada para modelar este tipo de dado devido à sua grande flexibilidade e simplicidade. A distribuição Weibull é para ser empregada na modelagem de variáveis contínuas, mas na prática é muito comum a existência de variáveis de tempos discretos, que são obtidas quando o tempo é medido em ciclos, número de impactos que um equipamento eletrônico suporta, dias ou meses completos, por exemplo. Nakano & Carrasco (2006) mostraram que, a depender das características dos dados (quando há muitos tempos empatados, por exemplo), utilizar um modelo contínuo para analisar dados discretos de sobrevivência pode levar a resultados pouco satisfatórios. Neste caso, uma alternativa plausível é utilizar uma distribuição própria para dados discretos. Brunello & Nakano (2015) aplicaram em dados discretos com presença de censura a distribuição Weibull discreta proposta por Nakagawa & Osaki (1975), que é a correspondente discreta da distribuição Weibull contínua e tem como caso especial a distribuição geométrica (correspondente discreta da distribuição exponencial) quando seu parâmetro de forma é igual a 1.

Em estudos de sobrevivência é comum supor que todas as observações em acompanhamento estão sujeitas à ocorrência do evento de interesse. Entretanto existem situações nas quais parte das observações de uma população não sofre este evento, mesmo quando o período de acompanhamento do estudo é muito longo. Neste contexto, estas observações são consideradas curadas, ou seja, não estão suscetíveis a falhar. Os modelos tradicionais de análise de sobrevivência não são indicados nesse tipo de situação, visto que não assumem a existência da fração de curados na população, considerando apenas que todos estão sob risco. Sendo assim, o modelo de mistura proposto por Berkson & Gage (1952) tem sido bastante utilizado para analisar dados com fração de cura, pois propõe a construção de uma função de sobrevivência imprópria na forma de mistura de duas distribuições: uma representando a distribuição dos tempos dos indivíduos sob risco e outra representando a distribuição dos tempos dos indivíduos que estão curados.

É grande o número de aplicações que os modelos de fração de cura têm tido em diversas áreas, sejam elas de saúde, industrial, financeira, etc. Granzotto *et al* (2010) aplicaram as distribuições Weibull e log-logística, ambas para dados contínuos, com modelo de mistura na análise de dados financeiros com o intuito de estudar o tempo até o cliente deixar de ter relacionamento com uma instituição financeira. Oliveira *et al* (2010) apresentaram o uso do modelo Weibull modificado de longa duração e dos seus casos particulares (Weibull de longa duração, Exponencial de longa duração), bem como os modelos Weibull e exponencial (todos para dados contínuos) para analisar um problema de câncer de mama com fração de curados. Fernandes (2013) apresentou o uso do modelo Weibull discreto com fração de cura em dados de tempos discretos e o aplicou na área de saúde em estudos de esquizofrenia e AIDS.

Quando trabalhamos com fração de curados é importante considerar que a probabilidade de cura pode variar de indivíduo para indivíduo a depender das características intrínsecas a cada um. Diante disto, surge a importância de analisar os efeitos das covariáveis sobre a fração de curados e também sobre o tempo de sobrevivência. Este tipo de situação é apresentado por Aljawadi *et al* (2011), onde os tempos são descritos pela distribuição exponencial, com parâmetro modelado por covariáveis, e é proposta uma abordagem analítica para a estimação paramétrica da fração de cura baseada no modelo BCH (*bounded cumulative hazard*). Paes (2007) apresentou a formulação dos modelos Weibull e de Cox com fração de cura. No modelo Weibull houve a inclusão de covariáveis

para modelar o parâmetro de escala e a fração de curados, enquanto que no modelo de Cox foi considerado o efeito das covariáveis sobre a fração de curados. Kannan *et al* (2010) formularam o modelo de fração de cura baseado na distribuição Exponencial Generalizada incorporando os efeitos das covariáveis na probabilidade de cura. Estes três trabalhos utilizaram o algoritmo EM para obter estimativas de máxima verossimilhança dos parâmetros e, como pode-se observar, utilizaram abordagens contínuas para modelar os tempos de sobrevivência.

Diante do que foi exposto, este trabalho tem como objetivo estudar e formular dentro do contexto de análise de sobrevivência o modelo Weibull discreto com fração de curados incorporando a presença de covariáveis. Como o modelo de fração de cura é uma mistura de distribuições, utilizamos o algoritmo EM para obter as estimativas dos parâmetros do modelo. Foram realizadas simulações Monte Carlo e o modelo foi aplicado em dois conjuntos de dados reais com variável resposta discreta.

A estrutura do trabalho está dividida em mais cinco capítulos. No Capítulo 2 descrevemos os conceitos básicos de análise de sobrevivência: tipos de censuras, as funções que descrevem os tempos, o estimador Kaplan-Meier, as distribuições e modelos Weibull e exponencial contínuos, as distribuições Weibull discreta e geométrica, bem como suas formulações incorporando a fração de cura. O Capítulo 3 apresenta o desenvolvimento da função de verossimilhança do modelo de regressão Weibull discreto com fração de cura e o uso do algoritmo EM para a estimação dos parâmetros. As simulações Monte Carlo e seus resultados são apresentados no Capítulo 4, enquanto que o Capítulo 5 ilustra duas aplicações do modelo proposto em conjuntos de dados reais. Por fim, no Capítulo 6 são apresentadas as considerações finais do trabalho.

Capítulo 2

Conceitos Básicos de Análise de Sobrevivência

A análise de sobrevivência pode ser definida como um conjunto de métodos estatísticos que tem como finalidade analisar dados cuja variável resposta é o tempo até a ocorrência de determinado evento de interesse, a partir de um determinado tempo inicial. Este tempo é denominado tempo de falha e a depender da área de aplicação e do objetivo do estudo pode ser definido como, por exemplo: o tempo desde o diagnóstico de uma doença em um paciente até a sua cura, o tempo até um cliente de um banco se tornar inadimplente desde que lhe foi concedido um crédito, o tempo entre a compra de um automóvel e o seu primeiro defeito mecânico, o tempo entre o ingresso de um aluno num curso de nível superior até a sua diplomação, o tempo desde a contração de um seguro de automóvel até a ocorrência do primeiro sinistro, ou ainda, o tempo que um preso leva desde a sua soltura até cometer um novo delito (CARVALHO, 2011; LIMA JUNIOR *et al*, 2012; BASTOS & ROCHA, 2006). A escala de medida da variável resposta pode ser em segundos, minutos, dias, meses, número de ciclos (na engenharia), etc.

Os estudos que envolvem a análise de sobrevivência são longitudinais, visto que as observações (podem ser indivíduos ou não) que fazem parte deles são acompanhadas ao longo do tempo a fim de se verificar suas características e a ocorrência do evento de interesse. De acordo com Colosimo e Giolo (2006), é de grande importância que duas definições sejam bem determinadas antes de se iniciar um estudo desse tipo. Uma definição é a do tempo de início, pois todas as observações devem começar a ser acompanhadas partindo de um mesmo critério, por exemplo: pacientes que entram num estudo após o diagnóstico confirmado da doença, clientes do banco que passaram a ser

acompanhados a partir do dia em que tomam um empréstimo, etc. Não significa que todos entrarão no estudo ao mesmo tempo. O acompanhamento de cada indivíduo poderá ser iniciado em tempos distintos e seus tempos serão contados a partir do momento em que começarem a ser observados. Assim, dentro de um determinado estudo, todas as observações devem ser comparáveis na sua origem. Outra definição a ser feita é do evento de interesse, também conhecido como falha ou desfecho. Por exemplo: no caso do automóvel que está sendo acompanhado até o seu primeiro defeito mecânico, é necessário deixar claro se será considerado qualquer defeito independente da sua gravidade ou não; no caso do banco é importante deixar claro o conceito utilizado para definir inadimplência.

2.1 Censura

A principal característica presente em dados de sobrevivência é a censura, que é a informação incompleta da resposta e pode ocorrer por diversos motivos. Geralmente não é possível encerrar um estudo somente após todas as observações sofrerem o evento de interesse, ou ainda, no decorrer do estudo pode ocorrer a perda de acompanhamento de algumas observações. Por exemplo, na área de saúde o paciente pode deixar de ser acompanhado devido à sua mudança de cidade ou morte causada por razão diferente da estudada, ou ainda, no acompanhamento de alunos de ensino superior há aqueles que trancam o curso e não retornam mais, ou um seguro de carros é cancelado sem que ocorresse nenhum sinistro. Deste modo, o que se sabe é que até o último instante observado ainda não havia ocorrido a falha, ou seja, o tempo de ocorrência do evento de interesse é maior do que o último tempo observado, o tempo de censura (COLOSIMO & GIOLO, 2006). Esta característica é de extrema importância nos dados de sobrevivência, pois se não houvesse a presença de censura, as análises poderiam ser realizadas através de técnicas estatísticas clássicas. Acontece que essas técnicas precisam de todos os dados completos para serem utilizadas, e os dados censurados apesar de incompletos são muito importantes por fornecerem informações sobre o tempo observado e não devem ser descartados da análise, pois a sua omissão pode levar a resultados incorretos. Portanto, os métodos de sobrevivência são os mais indicados nesse tipo de situação por permitirem incorporar na análise estatística os tempos censurados.

Alguns diferentes tipos de censuras podem ocorrer, a saber: censura à esquerda, cen-

sura intervalar e censura à direita. Dizemos que uma observação é censurada à esquerda quando o evento de interesse já havia ocorrido quando o experimento começou, isto é, o tempo registrado é maior que o tempo de falha. A censura intervalar é mais comum em situações nas quais as verificações das observações são periódicas. Esta ocorre quando não conhecemos o momento da ocorrência do desfecho, mas sabemos que ele ocorreu entre a última verificação e a mais recente (quando a ocorrência do evento de interesse foi detectada). Neste caso o tempo de falha não é conhecido, mas sabe-se que está contido em um intervalo. Já a censura à direita, a mais comum, é registrada quando o evento de interesse não ocorre até o último instante em que o indivíduo foi observado. Sendo assim, sabe-se que o tempo entre o início da observação e o evento é maior do que o tempo de fato observado (HOSMER *et al*, 2008; CARVALHO *et al*, 2011). Este último tipo de censura, que será o considerado neste trabalho, é caracterizado por outros três tipos: censura do tipo I, censura do tipo II e censura do tipo aleatória.

A censura do tipo I acontece em estudos que tem um período pré-determinado para o seu fim e, assim, ao ser finalizado registram observações que ainda não vieram a falhar. Nestes casos, o tempo de falha é superior ao tempo do fim do estudo.

A censura do tipo II ocorre quando o fim do estudo é condicionado à ocorrência de um determinado número de falhas. Assim, após k observações apresentarem o evento de interesse o estudo é finalizado e aqueles tempos que não são de falha são denominados censuras do tipo II.

Já a censura aleatória é registrada quando se perde o acompanhamento das observações ao longo do estudo antes de ocorrer a falha por motivos que não sejam a ocorrência do evento de interesse. O que se sabe nestes casos é que o tempo de falha destas observações é superior ao último tempo registrado (BASTOS & ROCHA, 2006; COLOSIMO & GIOLO, 2006).

Diante da presença das censuras, os dados de sobrevivência devem ter duas variáveis como forma de resposta para cada indivíduo i, geralmente representadas pelo par (t_i, δ_i), onde t_i representa o tempo de falha ou de censura e δ_i é uma variável dicotômica que indica se aquele determinado tempo é referente à falha ou não, ou seja,

$$\delta_i = \begin{cases} 1, & \text{se} \quad t_i \quad \text{é um tempo de falha,} \\ 0, & \text{se} \quad t_i \quad \text{é um tempo censurado.} \end{cases} \tag{2.1}$$

Nos casos em que, além da resposta, houver ainda a presença de covariáveis, os dados são representados por $(t_i, \delta_i, \mathbf{z}_i)$, na qual $\mathbf{z}_i$ representa o vetor de covariáveis do i-ésimo indivíduo (COLOSIMO & GIOLO, 2006).

Seja T a variável que representa o tempo até a ocorrência do evento de interesse, ou seja, o tempo de falha ou de sobrevivência. T é uma variável aleatória não-negativa, que pode ser contínua ou não, e é geralmente especificada pela sua função densidade de probabilidade (no caso contínuo), distribuição de probabilidade (no caso discreto), pela função de sobrevivência ou pela função de taxa de falha que serão apresentadas a seguir.

2.2 Funções que descrevem o tempo de sobrevivência

2.2.1 Função densidade

Nos casos em que a variável aleatória T é contínua a função densidade de probabilidade de T é denotada por $f(t)$ e pode ser interpretada como a probabilidade de uma observação falhar em um intervalo instantâneo de tempo $[t, t+\epsilon]$ sobre o intervalo ϵ e é definida por:

$$f(t) = \lim_{\epsilon \to 0^+} \frac{P(t \leq T < t+\epsilon)}{\epsilon}$$

na qual ϵ é o incremento de tempo infinitamente pequeno (CARVALHO *et al*, 2011) e $f(t) \geq 0$ para todo $t \geq 0$, com área abaixo da curva igual a 1.

Quando T é uma variável aleatória discreta, $t = 0, 1, 2, \ldots$, a sua distribuição de probabilidade é determinada por:

$$p(t) \quad = \quad P(T = t) = F(t) - F(t-1),$$

na qual $F(t) = P(T \leq t)$ é a distribuição acumulada de T e representa a probabilidade da ocorrência do evento até o instante t.

2.2.2 Função de sobrevivência

A função de sobrevivência, denotada por $S(t)$, é uma das funções mais utilizadas em estudos de sobrevivência. Trata-se de uma função decrescente definida como a probabilidade da falha não ocorrer antes do tempo t, ou seja, é a probabilidade de uma observação vir a "sobreviver"por mais que um tempo t. Quando T é uma variável aleatória contínua, possui como características começar com $S(0) = 1$, ou seja, a probabilidade de uma observação sobreviver por um tempo maior que zero é 1. Partindo do princípio de que todos as observações irão falhar durante o estudo, a probabilidade de sobreviver por um período muito grande é 0, isto é, $\lim_{t\to\infty} S(t) = 0$ (LIMA JUNIOR *et al*, 2012).

$$S(t) = P(T > t)$$

Nos casos em que a variável aleatória T é contínua t assume valores pertencentes a $[0, \infty)$ e a função de sobrevivência, que também é contínua, é definida por:

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = \int_t^\infty f(u)du.$$

Já nos casos em que a variável aleatória T é discreta, esta assume os valores $t = 0, 1, 2, ...$ e a função de sobrevivência discreta é dada por (FERNANDES, 2013):

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

2.2.3 Função de risco

A função de risco, também conhecida como função de taxa de falha condicional, representa o risco (instantâneo) de uma observação sofrer o evento de interesse entre os tempos t e $t + \epsilon$, com $\epsilon \to 0$, dado que até o tempo t o evento não havia ocorrido. Denotada por $\lambda(t)$, é uma função não-negativa e é definida como a probabilidade de uma observação vir a falhar em um intervalo de tempo muito pequeno, ϵ, dado que sobreviveu até o tempo t, dividido pelo comprimento ϵ do intervalo.

Segundo Louzada Neto *et al* (2002), a função de risco tem se destacado entre muitos autores por descrever como a probabilidade instantânea de falha muda com o passar

do tempo, pelo fato de poder assumir formas crescente, decrescente, constante ou não monótona, e por permitir que certos grupos sejam caracterizados de acordo com o comportamento da função de risco ao longo do tempo. Para Colosimo e Giolo (2006) "a função taxa de falha é mais informativa do que a função de sobrevivência. Diferentes funções de sobrevivência podem ter formas semelhantes, enquanto as respectivas funções de taxa de falha podem diferir drasticamente".

Quando T é uma variável aleatória contínua a função não tem limite superior, $\lambda(t) \geq 0$, e sua definição é matematicamente expressa por:

$$\lambda(t) = \lim_{\epsilon \to 0^+} \frac{P(t \leq T < t + \epsilon | T \geq t)}{\epsilon}$$

Quando T é uma variável aleatória discreta a função de risco é limitada no intervalo $0 \leq \lambda(t) \leq 1$ e assume valores maiores que zero apenas nos pontos onde ocorre falha (FERNANDES, 2013). É expressa por:

$$\lambda(t) = P(T = t | T \geq t) = \frac{P(T = t)}{P(T \geq t)} = \frac{P(T = t)}{P(T > t) + P(T = t)} = \frac{p(t)}{S(t) + p(t)}$$

2.3 Relações entre funções

As funções básicas de análise de sobrevivência apresentadas são relacionadas entre si e são matematicamente equivalentes. Desta forma, através das suas relações, é possível obter as demais funções a partir do conhecimento de uma delas.

Para o caso em que T é uma variável aleatória contínua e não-negativa, as relações são definidas da seguinte forma:

$$S(t) = \int_t^\infty f(u)du = 1 - F(t) \Leftrightarrow \frac{\partial}{\partial t}\left[1 - S(t)\right] = f(t) \Leftrightarrow -S'(t) = f(t)$$

Tem-se que:

$$
\begin{aligned}
\frac{f(t)}{\lambda(t)} &= \lim_{\epsilon \to 0^+} \frac{P(t \leq T < t+\epsilon)}{\epsilon} \frac{\epsilon}{P(t \leq T < t+\epsilon | T \geq t)} \\
&= \lim_{\epsilon \to 0^+} P(t \leq T < t+\epsilon) \frac{P(T \geq t)}{P(t \leq T < t+\epsilon) \cap P(T \geq t)} \\
&= \lim_{\epsilon \to 0^+} \frac{P(t \leq T < t+\epsilon)}{P(t \leq T < t+\epsilon)} P(T \geq t) \\
&= S(t) \\
&= 1 - F(t).
\end{aligned}
$$

Assim, relacionando estes resultados, temos ainda que:

$$
\frac{-S'(t)}{S(t)} = \lambda(t) \Leftrightarrow -\frac{\partial}{\partial t} \log S(t) = \lambda(t)
$$

$$
\Leftrightarrow -\log S(t) = \Lambda(t) \Leftrightarrow \log S(t) = -\int_0^t \lambda(u) du
$$

$$
\Leftrightarrow S(t) = \exp\left[-\int_0^t \lambda(u) du\right] = \exp\left[-\Lambda(t)\right],
$$

na qual $\Lambda(t) = \int_0^t \lambda(u) du$ é denominada função de taxa de falha acumulada, ou função de risco acumulado, e mede o risco de ocorrência do evento até um determinado tempo t (CARVALHO *et al*, 2011).

Quando T é uma variável aleatória discreta, as relações podem ser descritas como a seguir:

$$
p(t) = F(t) - F(t-1) = [1 - S(t)] - [1 - S(t-1)] = S(t-1) - S(t)
$$

$$
\lambda(t) = \frac{P(T = t \cap T \geq t)}{P(T \geq t)} = \frac{P(T = t)}{P(T > t-1)} = \frac{p(t)}{S(t-1)}
$$

Sabemos que $S(t) + p(t) = S(t-1)$, então:

$$\frac{1}{S(t)+p(t)} = \frac{1}{S(t-1)} \Leftrightarrow \frac{p(t)}{S(t)+p(t)} = \frac{p(t)}{S(t-1)}$$

$$\Leftrightarrow \lambda(t) = \frac{p(t)}{S(t-1)} = \frac{S(t-1)-S(t)}{S(t-1)} = 1 - \frac{S(t)}{S(t-1)}.$$

2.4 Estimador de Kaplan-Meier

O estimador de Kaplan-Meier (EKM), também chamado de estimador produto-limite, é um estimador de máxima verossimilhança não-paramétrico para a função de sobrevivência $S(t)$. É uma adaptação da função de sobrevivência empírica, que na ausência de censura dos dados será estimada no tempo t levando em consideração a proporção de observações que não sofreram o evento de interesse até este momento. Assim, considerando uma amostra de tamanho n, é definida como:

$$\hat{S}(t) = \frac{\text{NS}}{n}, t \geq 0,$$

na qual NS é o número de observações que não sofreram o evento de interesse até o tempo t.

Trata-se de uma função escada, cujos degraus são formados nos instantes nos quais há a ocorrência de falha. Cada degrau tem tamanho $1/n$ quando os tempos observados são distintos, e tamanho a/n quando há empates em um determinado instante t, sendo a o número de empates (COLOSIMO & GIOLO, 2006).

A ideia do EKM é a de que para um indivíduo sobreviver por J intervalos de tempo, ele vai precisar sobreviver a cada intervalo até o J-ésimo, considerando que ele sobreviveu aos anteriores. Assim, tem-se um estimador construído a partir do produto das probabilidades condicionais de sobreviver a cada intervalo de tempo. Este estimador considera tantos intervalos quantos forem o número de falhas distintas, sendo os limites destes intervalos definidos como os tempos de falha da amostra.

Considere uma amostra composta por n observações na qual existam k ($k \leq n$) tempos distintos de falha não censurados e estes tempos estejam dispostos em ordem crescente $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$. Sejam d_j o número de eventos que ocorrem em $t_{(j)}$, $j = 1, \ldots, k$, e n_j o número de indivíduos sob risco no tempo $t_{(j)}$. Assim, o EKM é definido por:

$$\hat{S}_{KM}(t) = \frac{n_1 - d_1}{n_1} \times \frac{n_2 - d_2}{n_2} \times \cdots \times \frac{n_k - d_k}{n_k} = \prod_{j:t_{(j)} \leq t} \left[\frac{n_j - d_j}{n_j} \right].$$

E pode ser escrito também de forma recursiva:

$$\hat{S}_{KM}(t) = \hat{S}_{KM}(t-1) \times \frac{n_j - d_j}{n_j}.$$

Quando $0 \leq t \leq t_{(1)}$, $\hat{S}_{KM}(t) = 1$. Se $t_{(k)}$ é a maior observação registrada, então $\hat{S}_{KM}(t) = 0$ para $t \geq t_{(k)}$, entretanto se a maior observação registrada for um tempo censurado t^*, então $\hat{S}_{KM}(t)$ nunca assume o valor zero, permanecendo constante para $t > t^*$ (BASTOS & ROCHA, 2007). O EKM tem como propriedades a ausência de viés quando estimado em grandes amostras, consistência fraca e convergência assintótica para um processo gaussiano.

2.5 Principais distribuições

O tempo até a ocorrência de um evento de interesse, T, é uma variável aleatória, logo o seu comportamento pode ser descrito por uma distribuição de probabilidade. Duas características que orientam a escolha da distribuição a ser adotada para descrever os tempos são: T assume valores não-negativos e, frequentemente, a distribuição de T apresenta forte assimetria com uma grande cauda à direita.

Diversas distribuições de probabilidade podem ser adotadas para modelar o tempo de sobrevivência. Serão descritas aqui as distribuições exponencial e Weibull, por serem amplamente utilizadas e se adaptarem bem a várias situações práticas.

2.5.1 Distribuição exponencial

A distribuição exponencial é considerada uma das mais simples e mais utilizadas dentre as distribuições para modelar dados de sobrevivência. De acordo com Colosimo e Giolo (2006) e Carvalho *et al* (2011), esta distribuição tem sido bastante aplicada em estudos de confiabilidade de sistemas eletrônicos, de produtos e materiais e, na área de saúde, para descrever tempos de vida e de remissão de doenças crônicas e infecciosas.

Trata-se de uma distribuição que apresenta apenas um parâmetro, assume independência do risco ao longo do tempo e possui taxa de falha constante (NAKANO & CARRASCO, 2006; COLOSIMO & GIOLO, 2006). Assim, assumindo que T é uma variável aleatória contínua que segue distribuição Exponencial(α), sua função de densidade de probabilidade pode ser escrita como:

$$f_E(t;\alpha) = \frac{1}{\alpha}\exp\left\{-\left(\frac{t}{\alpha}\right)\right\}, \tag{2.2}$$

na qual $t \geq 0$ e $\alpha > 0$, sendo α o tempo médio de vida e possui a mesma unidade do tempo t.

As funções de sobrevivência $S_E(t;\alpha)$ e de risco $\lambda_E(t;\alpha)$ são dadas, respectivamente, por:

$$S_E(t;\alpha) = \exp\left\{-\left(\frac{t}{\alpha}\right)\right\} \qquad \text{e}$$

$$\lambda_E(t;\alpha) = \frac{f(t)}{S(t)} = \frac{\left(\frac{1}{\alpha}\right)\exp\left\{-\left(\frac{t}{\alpha}\right)\right\}}{\exp\left\{-\left(\frac{t}{\alpha}\right)\right\}} = \frac{1}{\alpha},$$

para $t \geq 0$ e $\alpha > 0$.

Considerar uma distribuição com risco constante significa que, independente do tempo, o risco de ocorrência do evento de interesse é o mesmo. Trata-se da propriedade chamada de falta de memória da distribuição exponencial. Na prática é como se a idade de uma pessoa não interferisse no risco de sua morte, ou ainda, que um equipamento que está com horas de funcionamento tivesse o mesmo risco de falha de um equipamento sem uso, por exemplo. Acontece que, na prática, existem situações nas quais esta suposição é pouco plausível e é necessário trabalhar com alguma distribuição que permita a variação do

risco no tempo, como, por exemplo, a distribuição Weibull, que é uma generalização da distribuição exponencial.

2.5.2 Distribuição Weibull

A distribuição Weibull é amplamente utilizada para modelar dados que representam tempo até a ocorrência de um evento de interesse, principalmente na área biomédica e industrial. Seu frequente uso pode ser explicado em parte pela sua grande flexibilidade e simplicidade. Tem como propriedade possuir função de risco monótona, assim o risco pode variar no tempo, sendo uma função crescente, decrescente ou constante (CARVALHO *et al*, 2011; COLOSIMO & GIOLO, 2006).

Sendo T uma variável aleatória contínua que segue distribuição Weibull(α, β), a sua função densidade de probabilidade é dada por:

$$f_W(t;\alpha,\beta) = \frac{\beta}{\alpha^\beta} t^{\beta-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^\beta\right\}, \tag{2.3}$$

sendo $t \geq 0$, $\beta > 0$ o parâmetro de forma e $\alpha > 0$ o parâmetro de escala. O parâmetro β não tem unidade de medida e α tem a mesma unidade de medida de t. Quando $f_W(t;\alpha,\beta)$ é escrita na forma (2.3), α é aproximadamente o percentil 63% da distribuição da variável aleatória T (LOUZADA NETO *et al*, 2002).

As funções de sobrevivência $S_W(t;\alpha,\beta)$ e de risco $\lambda_W(t;\alpha,\beta)$ são dadas, respectivamente, por:

$$S_W(t;\alpha,\beta) = \exp\left\{-\left(\frac{t}{\alpha}\right)^\beta\right\} \qquad \text{e}$$

$$\lambda_W(t;\alpha,\beta) = \frac{f(t)}{S(t)} = \frac{\frac{\beta}{\alpha^\beta} t^{\beta-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^\beta\right\}}{\exp\left\{-\left(\frac{t}{\alpha}\right)^\beta\right\}} = \frac{\beta}{\alpha^\beta} t^{\beta-1},$$

para $t \geq 0$ e $\alpha, \beta > 0$.

O parâmetro β determina a forma da função de risco. Sendo assim, para $\beta > 1$, a função de risco é estritamente crescente; para $\beta < 1$, a função de risco é estritamente decrescente; e para $\beta = 1$, tem-se a distribuição exponencial, que é um caso particular da Weibull, e possui função de risco constante.

2.5.3 Distribuição Weibull Discreta

Apesar da ampla utilização das distribuições contínuas no contexto de análise de sobrevivência, nem sempre elas são indicadas, pois na prática é muito comum a existência de dados de tempos discretos, que surgem quando o tempo é medido em ciclos, dias ou meses completos, por exemplo. Nakano e Carrasco (2006) estudaram o uso de modelos contínuos em dados discretos e mostraram que dependendo das características dos dados, o uso de um modelo contínuo para analisar tempos discretos pode levar a resultados poucos satisfatórios. Face a esta situação, surge a necessidade de explorar o uso de modelos discretos em estudos que envolvem análise de sobrevivência.

A distribuição Weibull discreta foi proposta por Nakagawa e Osaki (1975) e é a distribuição discreta equivalente à Weibull contínua (FERNANDES, 2013; BRUNELLO & NAKANO, 2015). A obtenção das distribuições discretas pode ser feita a partir das distribuições contínuas através do agrupamento dos tempos em intervalos unitários. Seja Y uma variável aleatória contínua com distribuição Weibull(α, β). A variável discreta é obtida por $T = [Y]$, sendo $[Y]$ a parte inteira de Y (NAKANO & CARRASCO, 2006). Assim, a distribuição de probabilidade de T é definida por:

$$\begin{aligned} p_{WD}(t;\theta) &= P(T = t) \\ &= P(t \leq Y < t+1) \\ &= P(t < Y \leq t+1) \\ &= F_Y(t+1) - F_Y(t) \\ &= [1 - S_Y(t+1)] - [1 - S_Y(t)] \\ &= e^{-\left(\frac{t}{\alpha}\right)^{\beta}} - e^{-\left(\frac{t+1}{\alpha}\right)^{\beta}} \\ &= q^{t^{\beta}} - q^{(t+1)^{\beta}}, \end{aligned}$$

sendo $q = \exp\{-\frac{1}{\alpha^{\beta}}\}$, $0 < q < 1$, $\theta = (q, \beta)$ e $t = 0, 1, 2, \ldots$

A função de sobrevivência neste caso pode ser escrita como:

$$
\begin{aligned}
S_{WD}(t;\theta) &= P(T > t) \\
&= \sum_{i=t+1}^{\infty} p(i) \\
&= \sum_{i=t+1}^{\infty} q^{i^\beta} - q^{(i+1)^\beta} \\
&= (q^{(t+1)^\beta} - q^{(t+2)^\beta}) + (q^{(t+2)^\beta} - q^{(t+3)^\beta}) + \ldots \\
&= q^{(t+1)^\beta}.
\end{aligned}
$$

E a função de risco é dada por:

$$
\lambda_{WD}(t;\theta) = \frac{p_{WD}(t;\theta)}{S_{WD}(t-1;\theta))} = \frac{q^{t^\beta} - q^{(t+1)^\beta}}{q^{(t)^\beta}} = 1 - q^{(t+1)^\beta - t^\beta}.
$$

Assim como no caso contínuo o parâmetro β é o que determina a forma da função do risco: crescente, quando $\beta > 1$; decrescente, quando $\beta < 1$; e constante quando $\beta = 1$. Neste último caso a distribuição Weibull discreta se reduz à distribuição geométrica (que é a distribuição discreta correspondente à distribuição exponencial) com parâmetro $(1-q)$.

2.6 Modelos de regressão paramétricos contínuos

Em estudos de análise de sobrevivência geralmente existem características que são observadas em cada indivíduo ou objeto de estudo, que podem estar relacionadas com o tempo de sobrevida. Assim, além do tempo de sobrevivência e da variável indicadora de falha, os dados serão compostos também por estas outras k variáveis, que são chamadas de variáveis explicativas ou covariáveis, e serão representados por $(t_i, \delta_i, \mathbf{z}_i)$, para $i = 1, \ldots, n$, sendo t_i o tempo de sobrevivência, δ_i o indicador de falha e $\mathbf{z}_i' = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{ik})$ o vetor de covariáveis observadas.

Modelos de regressão são utilizados quando o objetivo é avaliar a relação entre os tempos de sobrevivência e as variáveis explicativas. Neste trabalho serão abordados os modelos paramétricos, que serão compostos por dois componentes: um aleatório, que descreve o comportamento do tempo de sobrevivência por meio de uma distribuição de

probabilidade; e um determinístico, que descreve a relação entre os parâmetros da distribuição de probabilidade associada aos tempos e as covariáveis (LOUZADA NETO *et al*, 2002; CARVALHO *et al*, 2011). O relacionamento descrito pela parte determinística pode ser definido por:

$$\alpha(\mathbf{z}) = g(\boldsymbol{\phi}'\mathbf{z}),$$

sendo $g(\cdot)$ uma função positiva e contínua, $\boldsymbol{\phi}' = (\phi_0, \phi_1, \ldots, \phi_k)$ o vetor de k parâmetros que serão estimados e $\mathbf{z}$ o vetor de covariáveis observadas.

2.6.1 Modelo exponencial

O modelo exponencial é estendido a partir de (2.2) e sua função densidade de probabilidade reescrita é expressa por:

$$f_E(t; \boldsymbol{\phi}|\mathbf{z}) = \frac{1}{\alpha(\mathbf{z})} \exp\left\{-\left(\frac{t}{\alpha(\mathbf{z})}\right)\right\},$$

na qual $\alpha(\mathbf{z}) > 0$ representa a taxa de falha constante nos níveis da variável explicativa $\mathbf{z}$.

Neste caso, o parâmetro $\alpha(\mathbf{z})$ se relaciona com o vetor de covariáveis $\mathbf{z}$ através da ligação $g(\cdot) = \exp(\cdot)$:

$$\alpha(\mathbf{z}) = \exp(\boldsymbol{\phi}'\mathbf{z}) = \exp(\phi_0 + \phi_1\mathbf{z}_1 + \ldots + \phi_k\mathbf{z}_k), \tag{2.4}$$

onde $\boldsymbol{\phi}' = (\phi_0, \phi_1, \ldots, \phi_k)$ é o vetor de parâmetros que representam os efeitos das covariáveis, tal que $-\infty < \phi_0, \ldots, \phi_k < \infty$, e $\alpha(\cdot)$ é o parâmetro que define o risco exponencial.

As funções de sobrevivência e de risco são definidas, respectivamente, por:

$$S_E(t; \boldsymbol{\phi}|\mathbf{z}) = \exp\left\{-\left(\frac{t}{\exp(\boldsymbol{\phi}'\mathbf{z})}\right)\right\} \qquad \text{e}$$

$$\lambda_E(t; \boldsymbol{\phi}|\mathbf{z}) = \frac{1}{\exp(\boldsymbol{\phi}'\mathbf{z})}.$$

A estimação dos parâmetros do modelo é feita através do método da máxima verossimilhança. Para dados censurados à direita provenientes de uma amostra aleatória de tamanho n, a função de verossimilhança pode ser escrita como:

$$\begin{aligned} L_E(\boldsymbol{\phi}) &= \prod_{i=1}^{n}[f_E(t_i;\boldsymbol{\phi}|\mathbf{z})]^{\delta_i}[S_E(t_i;\boldsymbol{\phi}|\mathbf{z})]^{1-\delta_i} \\ &= \prod_{i=1}^{n}\left[\frac{1}{\exp(\boldsymbol{\phi}'\mathbf{z})}\right]^{\delta_i}\exp\left\{-\sum_{i=1}^{n}\left(\frac{t_i}{\exp(\boldsymbol{\phi}'\mathbf{z})}\right)\right\}, \end{aligned}$$

na qual t_i é a i-ésima observação com seu respectivo indicador de censura δ_i.

As estimativas de máxima verossimilhança de $\boldsymbol{\phi}$ podem ser obtidas através da maximização da logaritmo da verossimilhança $l_E(\boldsymbol{\phi}) = \log L_E(\boldsymbol{\phi})$.

2.6.2 Modelo Weibull

A inclusão de covariáveis no modelo Weibull para dados contínuos é feita em (2.3). Modelando o parâmetro de escala α pelas covariáveis e considerando a forma funcional (2.4) utilizada na distribuição exponencial, a função densidade de probabilidade estendida é dada por:

$$f_W(t;\boldsymbol{\phi},\beta|\mathbf{z}) = \frac{\beta}{[\exp(\boldsymbol{\phi}'\mathbf{z})]^{\beta}}t^{\beta-1}\exp\left\{-\left(\frac{t}{\exp(\boldsymbol{\phi}'\mathbf{z})}\right)^{\beta}\right\},$$

sendo $t \geq 0$, $\beta > 0$, $\boldsymbol{\phi}' = (\phi_0, \phi_1, \ldots, \phi_k)$ e $\mathbf{z}$ o vetor de covariáveis.

As funções de sobrevivência e de risco são, respectivamente, expressas da seguinte forma:

$$S_W(t;\boldsymbol{\phi},\beta|\mathbf{z}) = \exp\left\{-\left(\frac{t}{\exp(\boldsymbol{\phi}'\mathbf{z})}\right)^{\beta}\right\} \qquad \text{e}$$

$$\lambda_W(t;\boldsymbol{\phi},\beta|\mathbf{z}) = \frac{\beta}{[\exp(\boldsymbol{\phi}'\mathbf{z})]^{\beta}}t^{\beta-1},$$

para $t \geq 0$ e $\beta > 0$.

Considerando uma amostra aleatória de tamanho n, a verossimilhança para dados censurados à direita é definida por:

$$\begin{aligned} L_W(\boldsymbol{\phi}, \beta) &= \prod_{i=1}^{n} [f_W(t_i; \boldsymbol{\phi}, \beta | \mathbf{z})]^{\delta_i} [S_W(t_i; \boldsymbol{\phi}, \beta | \mathbf{z})]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left[\frac{\beta}{[\exp(\boldsymbol{\phi}'\mathbf{z})]^{\beta}} t_i^{\beta-1} \right]^{\delta_i} \exp\left\{ -\sum_{i=1}^{n} \left(\frac{t}{\exp(\boldsymbol{\phi}'\mathbf{z})} \right)^{\beta} \right\}. \end{aligned}$$

As estimativas de máxima verossimilhança dos parâmetros β e $\boldsymbol{\phi}' = (\phi_0, \phi_1, \ldots, \phi_k)$ são obtidas maximizando $l_W(\boldsymbol{\phi}, \beta) = \log L_W(\boldsymbol{\phi}, \beta)$.

2.7 Fração de cura

Em estudos de sobrevivência pressupõe-se que todas as observações são suscetíveis à ocorrência do evento de interesse e em algum momento irão experimentá-lo. No entanto, às vezes, em parte dessas observações o evento de interesse não ocorre e, mesmo após um longo tempo de acompanhamento, estas não vem a falhar. Neste contexto, são observações chamadas de fração de curadas, não-suscetíveis, imunes ao evento ou ainda observações de longa duração. São, por exemplo, os indivíduos diagnosticados com determinada doença e, após o tratamento, se curam e não vem a óbito, os alunos que entram num curso de nível superior e não se graduam até o fim do acompanhamento, ou ainda, na área financeira, a fração de indivíduos que tomaram um certo tipo de empréstimo e quitaram toda a dívida em dia sem se tornarem inadimplentes, ou as pessoas que contratam um seguro de carro e passam muito tempo sem acioná-lo.

A partir da construção de um gráfico da função de sobrevivência empírica estimada através do estimador produto-limite de Kaplan-Meier, é possível observar indícios da fração de observações curadas nos dados. Caracteriza-se pela presença de grande quantidade de censuras ao final do estudo e pelo fato da cauda da função estar tendendo a um valor constante diferente de zero durante um período longo de tempo (MALLER e ZHOU, 1996). Um exemplo de função de sobrevivência de uma população com fração de curados é ilustrado na Figura 2.1.

Os modelos tradicionais de análise de sobrevivência assumem que a fração de curados é zero ao longo do tempo e, assim, a função de sobrevivência converge para zero quando

o tempo tende ao infinito (função de sobrevivência própria). Deste modo, utilizá-los em dados com parcela de curados pode levar a resultados equivocados, sendo, então, modelos com fração de cura os mais apropriados para modelar este tipo de dado.

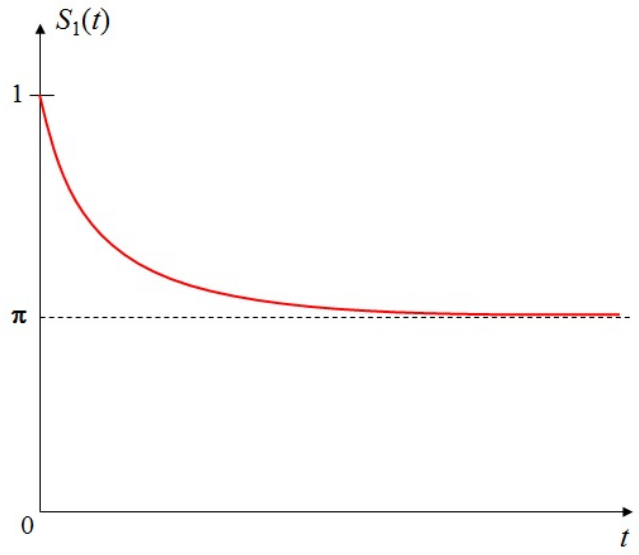

Figura 2.1: Função de sobrevivência com fração de curados.
Fonte: Fernandes (2013)

Proposto por Berkson e Gage (1952), o modelo de mistura propõe a construção de uma função de sobrevivência populacional imprópria ($S_{FC}(t)$) na forma de mistura de duas distribuições paramétricas. Para isto, a população em estudo é dividida em duas subpopulações, sendo uma composta por indivíduos que estão sob risco durante o estudo (SR), e a outra formada por indivíduos curados, ou seja, indivíduos não-suscetíveis à ocorrência do evento de interesse.

A situação de cura ou não da observação em um estudo pode ser indicada a partir da variável aleatória C com distribuição Bernoulli. Assim, considera-se que para a i-ésima observação:

$$\mathrm{c}_i = \begin{cases} 0, & \text{se a observação é curada,} \\ 1, & \text{se a observação é não curada.} \end{cases} \tag{2.5}$$

Desta maneira, $P(\mathrm{c}_i = 0) = \pi$ é a probabilidade da i-ésima observação ser não-suscetível, ou curada, enquanto que a probabilidade de uma observação i ser suscetível é

$P(\mathrm{c}_i = 1) = (1 - \pi)$.

Sejam, então: $\pi \in [0, 1]$ a proporção de observações curadas na população estudada, com $P(C) = \pi$ e função de sobrevivência $S_C(t)$; e $(1 - \pi)$ a proporção de observações que se encontram sob risco, com $P(\mathrm{SR}) = 1 - P(C) = (1 - \pi)$ e função de sobrevivência $S_{SR}(t)$. Assim, a função de sobrevivência populacional em forma de mistura é definida por:

$$\begin{aligned} S_{FC}(t) &= P(C)P(T > t|C) + P(SR)P(T > t|SR) \\ &= \pi S_C(t) + (1 - \pi)S_{SR}(t) \\ &= \pi + (1 - \pi)S_{SR}(t), \end{aligned}$$

e possui como propriedade: $\lim_{t\to\infty} S_{FC}(t) = \pi$. Se $\pi = 0$, então $S_{FC}(t) = S_{SR}(t)$.

2.8 Modelos Weibull, exponencial e geométrico com fração de cura

Considerando que T é uma variável aleatória contínua com distribuição Weibull(α, β), as funções densidade e de sobrevivência considerando a presença de fração de cura são dadas por:

$$\begin{aligned} f_{FCW}(t; \pi, \alpha, \beta) &= (1 - \pi) f_{SRW}(t; \alpha, \beta) \\ &= (1 - \pi)\frac{\beta}{\alpha^\beta} t^{\beta-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^\beta\right\} \qquad \text{e} \end{aligned}$$

$$\begin{aligned} S_{FCW}(t; \pi, \alpha, \beta) &= \pi + (1 - \pi)S_{SRW}(t; \alpha, \beta)) \\ &= \pi + (1 - \pi)\exp\left\{-\left(\frac{t}{\alpha}\right)^\beta\right\}, \end{aligned}$$

em que $\alpha, \beta > 0$, $0 \leq \pi \leq 1$ e $t \geq 0$.

Desta maneira, é possível obter a função de risco:

$$\begin{aligned}\lambda_{FCW}(t;\pi,\alpha,\beta) &= \frac{f_{FCW}(t;\pi,\alpha,\beta)}{S_{FCW}(t;\pi,\alpha,\beta)} \\ &= \frac{(1-\pi)\frac{\beta}{\alpha^\beta}t^{\beta-1}\exp\left\{-\left(\frac{t}{\alpha}\right)^\beta\right\}}{\pi+(1-\pi)\exp\left\{-\left(\frac{t}{\alpha}\right)^\beta\right\}}.\end{aligned}$$

A obtenção das funções anteriormente citadas para o caso em que a variável aleatória T assume distribuição Exponencial(α) é feita assumindo $\beta = 1$. Sendo assim, as funções densidade,de sobrevivência e de risco na presença de fração de cura são definidas, respectivamente, por:

$$\begin{aligned}f_{FCE}(t;\pi,\alpha) &= (1-\pi)f_{SRE}(t;\alpha) \\ &= \frac{(1-\pi)}{\alpha}\exp\left\{-\left(\frac{t}{\alpha}\right)\right\},\end{aligned}$$

$$\begin{aligned}S_{FCE}(t;\pi,\alpha) &= \pi+(1-\pi)S_{SRE}(t;\alpha) \\ &= \pi+(1-\pi)\exp\left\{-\left(\frac{t}{\alpha}\right)\right\} \qquad \text{e}\end{aligned}$$

$$\begin{aligned}\lambda_{FCE}(t;\pi,\alpha) &= \frac{f_{FCE}(t;\pi,\alpha)}{S_{FCE}(t;\pi,\alpha)} \\ &= \frac{\frac{(1-\pi)}{\alpha}\exp\left\{-\left(\frac{t}{\alpha}\right)\right\}}{\pi+(1-\pi)\exp\left\{-\left(\frac{t}{\alpha}\right)\right\}},\end{aligned}$$

em que $\alpha > 0, 0 \leq \pi \leq 1$ e $t \geq 0$.

Para os casos em que T é uma variável aleatória discreta que segue distribuição Weibull(α,β) discreta, a função de sobrevivência é definida por:

$$\begin{aligned}S_{FCWD}(t;\pi,\theta) &= \pi+(1-\pi)S_{SRWD}(t;\theta) \\ &= \pi+(1-\pi)(q^{(t+1)^\beta}),\end{aligned} \tag{2.6}$$

sendo $q = \exp\{-\frac{1}{\alpha^\beta}\}$, $0 < q < 1, \theta = (q,\beta)$ e $t = 0, 1, 2, \ldots$

A partir da função de sobrevivência é possível obter a distribuição de probabilidade de T:

$$\begin{aligned} p_{FCWD}(t;\pi,\theta) &= S_{FCWD}(t-1;\pi,\theta) - S_{FCWD}(t;\pi,\theta) \\ &= [\pi + (1-\pi)q^{t^\beta}] - [\pi + (1-\pi)q^{(t+1)^\beta}] \\ &= (1-\pi)[q^{t^\beta} - q^{(t+1)^\beta}] = (1-\pi)p_{WD}(t;\theta), \end{aligned}$$

em que $t = 0, 1, 2, \ldots$.

A função de risco é obtida a partir da relação entre $S_{FCWD}(t;\pi,\theta)$ e $p_{FCWD}(t;\pi,\theta)$:

$$\begin{aligned} \lambda_{FCWD}(t;\pi;\theta) &= \frac{p_{FCWD}(t;\theta)}{S_{FCWD}(t-1;\theta)} \\ &= \frac{(1-\pi)[q^{t^\beta} - q^{(t+1)^\beta}]}{\pi + (1-\pi)(q^{t^\beta})}, \end{aligned}$$

em que $t = 0, 1, 2, \ldots$.

Ao considerar $\beta = 1$ é possível determinar as funções anteriormente descritas para o caso da distribuição geométrica com fração de cura:

$$\begin{aligned} S_{FCG}(t;\pi,q) &= \pi + (1-\pi)f_{SRG}(t;q) \\ &= \pi + (1-\pi)(q^{(t+1)}), \end{aligned}$$

$$\begin{aligned} p_{FCG}(t;\pi,q) &= S_{FCG}(t-1;\pi,q) - S_{FCG}(t;\pi,q) \\ &= (1-\pi)[q^t - q^{(t+1)}] \qquad \text{e} \end{aligned}$$

$$\begin{aligned} \lambda_{FCG}(t;\pi,q) &= \frac{p_{FCG}(t;\pi,q)}{S_{FCG}(t-1;\pi,q)} \\ &= \frac{(1-\pi)[q^t - q^{(t+1)}]}{\pi + (1-\pi)(q^t)}, \end{aligned}$$

em que $t = 0, 1, 2, \ldots$.

Capítulo 3

Modelo de regressão Weibull discreto com fração de cura

Conforme apresentado no capítulo anterior, estudos com fração de cura assumem que a população de estudo é dividida em dois subgrupos, sendo um formado por observações suscetíveis ao evento de interesse (SR), e outro de observações curadas (C). Neste capítulo serão considerados tempos de vida que seguem distribuição Weibull(q, β) discreta com a presença de fração de cura.

A inclusão de covariáveis no modelo Weibull discreto permite acrescentar informações intrínsecas a cada observação em estudo. Isto se justifica pelo fato de que a probabilidade de cura pode ser diferente em pessoas do sexo feminino quando comparados com os de pessoas do sexo masculino, ou em grupos que estão sujeitos a diferentes tipos de medicação, ou ainda quando pessoas têm mais ou menos idade, por exemplo. Diante disto, a proporção de pessoas curadas, π, pode ser modelada a partir de um conjunto de variáveis $\mathbf{z}$. Como π assume valores em $[0, 1]$, a relação de π com as covariáveis $\mathbf{z}$ pode ser feita a partir da função de ligação logito. Sendo assim, define-se:

$$\pi(\boldsymbol{\phi}, \mathbf{z}) \;=\; \frac{e^{\boldsymbol{\phi}'\mathbf{z}}}{1 + e^{\boldsymbol{\phi}'\mathbf{z}}},$$

em que $\boldsymbol{\phi}' = (\phi_0, \phi_1, \ldots, \phi_k)$ é o vetor de parâmetros que representam os efeitos das covariáveis, tal que $-\infty < \phi_0, \ldots, \phi_k < \infty$ e $\mathbf{z}' = (\mathbf{z}_1, \ldots, \mathbf{z}_k)$ é o vetor de covariáveis observadas.

Sabemos que $\pi(\boldsymbol{\phi}, \mathbf{z})$ é a proporção de observações curadas, então temos:

$$\mathrm{c}_i = \begin{cases} 0, & \text{com probabilidade } \pi_i(\boldsymbol{\phi}, \mathbf{z}) = \frac{e^{\boldsymbol{\phi}'\mathbf{z}}}{1+e^{\boldsymbol{\phi}'\mathbf{z}}} \\ 1, & \text{com probabilidade } 1 - \pi_i(\boldsymbol{\phi}, \mathbf{z}) = \frac{1}{1+e^{\boldsymbol{\phi}'\mathbf{z}}} \end{cases}$$

A partir disso é possível, então, definir o modelo de regressão Weibull discreto com fração de cura, modelando o parâmetro $\pi(\boldsymbol{\phi}, \mathbf{z})$ que depende do vetor de covariáveis $\mathbf{z}$. A distribuição de probabilidade para este modelo é definida como:

$$\begin{aligned} p_{FCWD}(t;\Phi|\mathbf{z}) &= [1-\pi(\boldsymbol{\phi},\mathbf{z})]p_{SRWD}(t;\theta) \\ &= \left(\frac{1}{1+e^{\boldsymbol{\phi}'\mathbf{z}}}\right)\left(q^{t^\beta} - q^{(t+1)^\beta}\right), \end{aligned} \tag{3.1}$$

em que $\Phi = (\boldsymbol{\phi},\ \theta = (q,\beta))$, $\mathbf{z}' = (\mathbf{z}_1,\ldots,\mathbf{z}_k)$ é o vetor de covariáveis observadas e $\boldsymbol{\phi}' = (\phi_0,\phi_1,\ldots,\phi_k)$ o vetor de k parâmetros associados às covariáveis, tal que $-\infty < \phi_0,\phi_1,\ldots,\phi_k < \infty$.

As funções de sobrevivência e de risco são, respectivamente, definidas por:

$$\begin{aligned} S_{FCWD}(t;\Phi|\mathbf{z}) &= \pi(\boldsymbol{\phi},\mathbf{z}) + [1-\pi(\boldsymbol{\phi},\mathbf{z})]S_{SRWD}(t;\theta) \\ &= \frac{e^{\boldsymbol{\phi}'\mathbf{z}}}{1+e^{\boldsymbol{\phi}'\mathbf{z}}} + \left(\frac{1}{1+e^{\boldsymbol{\phi}'\mathbf{z}}}\right)(q^{(t+1)^\beta}) \\ &= \frac{e^{\boldsymbol{\phi}'\mathbf{z}} + q^{(t+1)^\beta}}{1+e^{\boldsymbol{\phi}'\mathbf{z}}} \qquad \text{e} \end{aligned} \tag{3.2}$$

$$\begin{aligned} \lambda_{FCWD}(t;\Phi|\mathbf{z}) &= \frac{p_{FCWD}(t;\Phi|\mathbf{z})}{S_{FCWD}(t-1;\Phi|\mathbf{z})} \\ &= \frac{\left(\frac{1}{1+e^{\boldsymbol{\phi}'\mathbf{z}}}\right)\left(q^{t^\beta} - q^{(t+1)^\beta}\right)}{\left(\frac{e^{\boldsymbol{\phi}'\mathbf{z}}+q^{t^\beta}}{1+e^{\boldsymbol{\phi}'\mathbf{z}}}\right)} \\ &= \frac{(q^{t^\beta} - q^{(t+1)^\beta})}{e^{\boldsymbol{\phi}'\mathbf{z}} + q^{t^\beta}}. \end{aligned} \tag{3.3}$$

Para encontrar as definições dessas funções para o modelo de regressão geométrico

com fração de cura, basta considerar $\beta = 1$:

$$\begin{aligned} p_{FCG}(t;\boldsymbol{\phi},q|\mathbf{z}) &= [1-\pi(\boldsymbol{\phi},\mathbf{z})]p_{SRG}(t;q) \\ &= \left(\frac{1}{1+e^{\boldsymbol{\phi}'\mathbf{z}}}\right)\left(q^t - q^{(t+1)}\right), \end{aligned}$$

$$\begin{aligned} S_{FCG}(t;\boldsymbol{\phi},q|\mathbf{z}) &= \pi(\boldsymbol{\phi},\mathbf{z}) + [1-\pi(\boldsymbol{\phi},\mathbf{z})]S_{SRG}(t;q) \\ &= \frac{e^{\boldsymbol{\phi}'\mathbf{z}}}{1+e^{\boldsymbol{\phi}'\mathbf{z}}} + \left(\frac{1}{1+e^{\boldsymbol{\phi}'\mathbf{z}}}\right)(q^{(t+1)}) \\ &= \frac{e^{\boldsymbol{\phi}'\mathbf{z}} + q^{(t+1)}}{1+e^{\boldsymbol{\phi}'\mathbf{z}}} \qquad \text{e} \end{aligned}$$

$$\begin{aligned} \lambda_{FCG}(t;\boldsymbol{\phi},q|\mathbf{z}) &= \frac{p_{FCG}(t;\boldsymbol{\phi},q|\mathbf{z})}{S_{FCG}(t-1;\boldsymbol{\phi},q|\mathbf{z})} \\ &= \frac{\left(\frac{1}{1+e^{\boldsymbol{\phi}'\mathbf{z}}}\right)\left(q^t - q^{(t+1)}\right)}{\left(\frac{e^{\boldsymbol{\phi}'\mathbf{z}}+q^t}{1+e^{\boldsymbol{\phi}'\mathbf{z}}}\right)} \\ &= \frac{q^t - q^{(t+1)}}{e^{\boldsymbol{\phi}'\mathbf{z}} + q^t}. \end{aligned}$$

3.1 Função de verossimilhança

Inicialmente, considere as variáveis indicadoras δ_i e c_i, sendo δ_i o indicador de falha definido em (2.1) e c_i o indicador de não cura apresentado em (2.5). Se $\delta_i = 0$ (tempo censurado), c_i pode assumir valores 0 ou 1, pois o tempo censurado pode ser proveniente de uma observação curada e que certamente não virá a sofrer o evento de interesse, ou ainda ser proveniente de uma observação que está suscetível ao evento, mas que não veio a falhar no período observado. Já no caso em que $\delta_i = 1$ (tempo de falha), necessariamente c_i assume valor 1, pois uma observação que veio a falhar não faz parte do grupo de curados.

Considere uma amostra aleatória de tamanho n com fração de curados, tempos de sobrevivência censurados à direita provenientes de uma distribuição Weibull discreta e vetor de covariáveis $\mathbf{z}$. De acordo com Aljawadi *et al* (2011), considerando que os dados estão completos com δ_i e c_i conhecidos, o logaritmo da verossimilhança completo para estes dados é:

$$
\begin{aligned}
l_c(\Phi) &= \log \prod_{i=1}^{n} [p_{FCWD}(t_i;\Phi|\mathbf{z})]^{\delta_i} [S_{FCWD}(t_i;\Phi|\mathbf{z})]^{1-\delta_i} \qquad (3.4)\\
&= \log \prod_{i=1}^{n} [(1-\pi_i(\boldsymbol{\phi},\mathbf{z}))p_{SRWD}(t_i;\theta)]^{\delta_i} [\{\pi_i(\boldsymbol{\phi},\mathbf{z})\}^{1-c_i} + \\
&\quad \{(1-\pi_i(\boldsymbol{\phi},\mathbf{z})S_{SRWD}(t_i;\theta)\}^{c_i}]^{1-\delta_i} \\
&= \sum_{i=1}^{n} \delta_i \log(1-\pi_i(\boldsymbol{\phi},\mathbf{z})) + \sum_{i=1}^{n} \delta_i \log[p_{SRWD}(t_i;\theta)] + \\
&\quad \sum_{i=1}^{n} (1-\delta_i) \log[\{\pi_i(\boldsymbol{\phi},\mathbf{z})\}^{1-c_i} + \{(1-\pi_i(\boldsymbol{\phi},\mathbf{z}))S_{SRWD}(t_i;\theta)\}^{c_i}],
\end{aligned}
$$

sendo $\Phi = (\boldsymbol{\phi}, q, \beta)$.

O estimador de Φ é obtido através da maximização de $l_c(\Phi)$. Acontece que neste caso estamos lidando com dados incompletos, visto que c_i é desconhecido quando $\delta_i = 0$. Por este motivo, as estimativas de máxima verossimilhança de Φ serão obtidas por meio do algoritmo EM.

O intervalo de confiança dos parâmetros com $(1-\alpha)\times 100\%$ de confiança é definido por $\hat{\gamma} \pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\gamma})}$, em que $\hat{\gamma}$ é o parâmetro para o qual o intervalo está sendo construído e $z_{\alpha/2}$ é o quantil da distribuição normal padrão. As estimativas das variâncias dos parâmetros são obtidas pelo inverso da matriz de Informação de Fisher.

3.2 Maximização via algoritmo EM

O algoritmo EM (*Expectation Maximization*) é um método computacional utilizado para se obter o estimador de máxima verossimilhança (EMV) de forma iterativa e se tornou bastante aplicado na Estatística. De acordo com Casella e Berger (2010) o seu uso se disseminou após o trabalho de Dempster *et al* (1977), no qual foi apresentada uma abordagem geral do algoritmo juntamente com algumas aplicações. Seu extenso uso se deve à sua simples implementação e abrangência de campo de aplicação.

Trata-se de um método que é geralmente utilizado em dois casos. O primeiro é quando os dados são incompletos e, então, o algoritmo usa os dados observados como informação para os que estão faltando. O segundo é quando a função de verossimilhança é bastante complexa e difícil de ser maximizada analiticamente. Em ambos os casos o EM permite encontrar estimativas de máxima verossimilhança dos parâmetros. Neste trabalho será

considerado o primeiro caso, pois, como foi apresentado, os dados que informam se a observação é curada ou sob risco não são completamente observados, sendo incompletos quando se trata de observações censuradas.

Na Seção 3.2.1 a seguir é feita uma formulação geral do algoritmo EM. Na Seção seguinte é apresentada a sua aplicação para a estimação de máxima verossimilhança dos parâmetros do modelo de regressão Weibull discreto com fração de cura.

3.2.1 Algoritmo EM

Considere o caso em que se deseja estimar um conjunto de parâmetros θ com base nos dados $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ de X. A função de densidade de $\mathbf{X}$ é dada por $f(x|\theta)$ e a sua função de verossimilhança é:

$$\mathcal{L}(\theta; \mathbf{X}) = \prod_{i=1}^{n} f(x_i; \theta).$$

Denota-se por $\ell_X(\theta; \mathbf{X})$ o logaritmo da verossimilhança de $\mathbf{X}$:

$$\ell_X(\theta; \mathbf{X}) = \log \mathcal{L}(\theta; \mathbf{X}) = \sum_{i=1}^{n} f(x_i; \theta).$$

Acontece que $\mathbf{X}$ é um conjunto de dados incompletos, sendo necessário completá-los. Para isto considere $\mathbf{Y} = (y_1, y_2, \ldots, y_n)$ o conjunto de dados que estão faltando em $\mathbf{X}$ e $\mathbf{V} = (\mathbf{X}, \mathbf{Y})$ o vetor dos dados completos. Então, a função densidade de probabilidade de $\mathbf{V}$ é:

$$\begin{aligned} f(v; \theta) &= f(x, y; \theta) \\ &= f(x; \theta) f(y; x, \theta), \end{aligned}$$

E o logaritmo da verossimilhança é:

$$\begin{aligned} \ell_V(\theta; \mathbf{V}) &= \log \mathcal{L}(\theta; \mathbf{V}) \\ &= \log \mathcal{L}(\theta; \mathbf{X}, \mathbf{Y}) \end{aligned} \tag{3.5}$$

O algoritmo EM tem como objetivo encontrar o EMV $\hat{\theta}$ de θ. Assim, com as notações citadas, a aplicação do EM consiste em basicamente dois passos:

- **Passo E**: Obter a esperança condicional do logaritmo da verossimilhança dado por (3.5) que depende de $\mathbf{Y}$:

$$\begin{aligned} Q(\theta, \theta^{(k)}) &= E[\ell_V(\theta; \mathbf{V})|\mathbf{X}, \theta^k] \\ &= E[\log f(\mathbf{X}, \mathbf{Y}; \theta)|\mathbf{X}, \theta^{(k)}] \\ &= \sum_{y_i} \log[f(\mathbf{X}, y_i; \theta)] f(y_i; \mathbf{X}, \theta^{(k)}). \end{aligned}$$

  Temos que $\mathbf{V}$ é uma combinação dos dados $\mathbf{X}$ e $\mathbf{Y}$, logo é necessário ponderar os possíveis valores não observados de $\mathbf{X}$ através das suas probabilidades. Sendo assim:

$$f(y_i; \mathbf{X}, \theta^k) = \frac{f(\mathbf{X}, y_i; \theta^k)}{f(\mathbf{X}; \theta^k)} = P(Y = y_i; \mathbf{X}, \theta^k)$$

- **Passo M**: Maximizar Q com respeito a θ, usando um método iterativo.

Os passos do algoritmo são aplicados sucessivamente até que algum critério de convergência pré-estabelecido seja atingido. Por exemplo: $|Q(\theta, \theta^{(k)}) - Q(\theta, \theta^{(k+1)})| < \epsilon$, com $\epsilon \to 0$.

Considere agora que a função densidade de X possa ser escrita na forma de mistura de duas componentes:

$$f(x; \theta) = p f_0(x; \theta_0) + (1 - p) f_1(x; \theta_1), \tag{3.6}$$

sendo $0 \leq p \leq 1$ e $\theta = (p, \theta_0, \theta_1)$.

A função de verossimilhança dos dados completos pode ser escrita como:

$$\begin{aligned} \ell_V(\theta; \mathbf{V}) &= \log f(\mathbf{X}, \mathbf{Y}; \theta) \\ &= \sum_{i=1}^{n} \log[f(x_i; \theta) f(y_i; x_i, \theta)]. \end{aligned}$$

Seja Y_i uma variável aleatória com distribuição Bernoulli que indica se a observação x_i pertence à componente f_0 ou f_1, respectivamente. As probabilidades associadas à variável

Y_i são definidas por:

$$P(Y_i = 0|x_i, \theta^{(k)}) = \frac{pf_0(x_i;\theta_0)}{f(x_i;\theta)}$$

$$P(Y_i = 1|x_i, \theta^{(k)}) = \frac{(1-p)f_1(x_i;\theta_1)}{f(x_i;\theta)}.$$

Logo, no passo E, $Q(\theta, \theta^{(k)})$ pode ser escrita como:

$$\begin{aligned} Q(\theta, \theta^{(k)}) &= \sum_{y_i} \log[f(\mathbf{X}, y_i; \theta)] f(y_i; \mathbf{X}, \theta^{(k)}) \\ &= \sum_{l=0}^{1} \sum_{i=1}^{n} \log[f(\mathbf{X}, y_i; \theta)] P(\mathbf{Y} = l|x_i, \theta^{(k)}) \qquad (3.7) \\ &= \sum_{i=1}^{n} \log(p) P(\mathbf{Y} = 0; x_i, \theta^{(k)}) + \sum_{i=1}^{n} \log[f_0(x_i;\theta_0)] P(\mathbf{Y} = 0|x_i, \theta^{(k)}) + \\ &\quad \sum_{i=1}^{n} \log(1-p) P(\mathbf{Y} = 1|x_i, \theta^{(k)}) + \sum_{i=1}^{n} \log[f_1(x_i;\theta_1)] P(\mathbf{Y} = 1|x_i, \theta^{(k)}) \end{aligned}$$

No passo M é encontrado θ^{k+1} que maximiza $Q(\theta, \theta^{(k)})$. Assim, θ^{k+1} será utilizado para obter $\hat{\theta}^{k+2}$ na atualização de $Q(\theta, \theta^{(k+1)})$ que será maximizada. Desta forma, estes passos irão se repetir até que o critério de convergência adotado seja atingido.

Segundo Casella e Berger (2010), "uma das vantagens do algoritmo EM é que as condições para a convergência para os dados incompletos são conhecidas". Wu (1983) cita que a demonstração de convergência das estimativas mostrada em Dempster *et al* (1977) contém um erro e, então, apresenta um teorema que assegura que as estimativas convergem monotonicamente para um ponto estacionário, que pode ser um máximo local ou ponto de sela. Como a convergência para um ponto estacionário, máximo local ou máximo global depende do chute inicial dos parâmetros, a sua recomendação é que o algoritmo EM seja implementado a partir de diferentes valores iniciais dos parâmetros. Outras demonstrações de convergência podem ser encontradas em Boyles (1983).

3.2.2 Algoritmo EM para modelo de regressão Weibull discreto com fração de cura

Neste trabalho está sendo estudado o modelo apresentado em (3.1), (3.2) e (3.3) e desejamos encontrar a estimativa de $\Phi = (\boldsymbol{\phi}, q, \beta)$ a partir de uma amostra aleatória de

tamanho n de T. Considere que os dados estão na forma $(t_i, \delta_i, c_i, \mathbf{z}_i)$ e que o i-ésimo indivíduo possui tempo de falha ou censura t_i e $\mathbf{z}_i$ representa seu vetor de covariáveis. Considere ainda que existem m tempos de falha, sendo, então, $(n-m)$ tempos de censura, não necessariamente ordenados. Sendo assim, temos que os dados que foram observados são: os tempos de falha t_i; o indicador de falha, que assume $\delta_i = 0$ se $i = (m+1), \ldots, n$, e $\delta_i = 1$ se $i = 1, \ldots, m$; o indicador de não cura $c_i = 1$ quando $i = 1, \ldots, m$; e o vetor de covariáveis $\mathbf{z}_i$ quando $i = 1, \ldots, n$. Já o dado denominado como desconhecido, incompleto, ou ainda, não observado é o valor de c_i quando o indivíduo é censurado, ou seja, para $i = (m+1), \ldots, n$ (ALAJAWADI *et al*, 2011).

Com base nas informações anteriores, é possível escrever o logaritmo da verossimilhança apresentado em (3.4) da seguinte forma:

$$\begin{aligned} l_c(\Phi) &= \sum_{i=1}^{m} \delta_i \log(1-\pi_i(\boldsymbol{\phi}, \mathbf{z})) + \sum_{i=(m+1)}^{n} \delta_i \log(1-\pi_i(\boldsymbol{\phi}, \mathbf{z})) + \\ &\sum_{i=1}^{m} \delta_i \log[p_{SRWD}(t_i;\theta)] + \sum_{i=(m+1)}^{n} \delta_i \log[p_{SRWD}(t_i;\theta)] + \\ &\sum_{i=1}^{m} (1-\delta_i) \log[\{\pi_i(\boldsymbol{\phi}, \mathbf{z})\}^{1-c_i} + \{(1-\pi_i(\boldsymbol{\phi}, \mathbf{z}))S_{SRWD}(t_i;\theta)\}^{c_i}] + \\ &\sum_{i=(m+1)}^{n} (1-\delta_i) \log[\{\pi_i(\boldsymbol{\phi}, \mathbf{z})\}^{1-c_i} + \{(1-\pi_i(\boldsymbol{\phi}, \mathbf{z}))S_{SRWD}(t_i;\theta)\}^{c_i}], \end{aligned} \tag{3.8}$$

em que $\theta = (q, \beta)$.

Seja $\mathbf{X}$ o vetor $(1 \times (n-m))$ dos tempos censurados (variável incompleta) e $\mathbf{C}$ o vetor $(1 \times (n-m))$ dos dados que estão faltando em $\mathbf{X}$ e, consequentemente, $\mathbf{V} = (\mathbf{X},\mathbf{C})$ o vetor de dados completos. As observações de $\mathbf{X}$ são censuradas, logo a sua distribuição é descrita pela função de sobrevivência, dada por:

$$g(x;\Phi) = \pi(\boldsymbol{\phi}, \mathbf{z}) + (1-\pi(\boldsymbol{\phi}, \mathbf{z}))S_{SRWD}(t;\theta),$$

que é uma mistura como a apresentada em (3.6).

De acordo com Peng (2003), para aplicar o algoritmo EM é necessário substituir c_i por uma estimativa da sua esperança condicional, dada a estimativa de Φ:

$$p_i = E(c_i|\Phi) = P(c_i = 1|\Phi),$$

na qual $\Phi = (\boldsymbol{\phi}, q, \beta)$.

Como já foi citado no início do capítulo, o valor que c_i assume depende de δ_i, ou seja, se a observação é referente à um tempo de falha ou não. Sendo assim, podemos encontrar as seguintes probabilidades condicionais de c_i:

$$\begin{aligned} P(c_i = 1|\Phi, \delta_i = 1) &= \frac{P(\delta_i = 1|c_i = 1)P(c_i = 1)}{P(\delta_i = 1|c_i = 0)P(c_i = 0) + P(\delta_i = 1|c_i = 1)P(c_i = 1)} \\ &= \frac{P(\delta_i = 1|c_i = 1)(1 - \pi_i(\boldsymbol{\phi}, \mathbf{z}))}{0 \times \pi_i(\boldsymbol{\phi}, \mathbf{z}) + P(\delta_i = 1|c_i = 1)(1 - \pi_i(\boldsymbol{\phi}, \mathbf{z}))} = 1 \end{aligned}$$

$$\begin{aligned} P(c_i = 1|\Phi, \delta_i = 0) &= \frac{P(\delta_i = 0|c_i = 1)P(c_i = 1)}{P(\delta_i = 0|c_i = 0)P(c_i = 0) + P(\delta_i = 0|c_i = 1)P(c_i = 1)} \\ &= \frac{P(T > t_i|c_i = 1)(1 - \pi_i(\boldsymbol{\phi}, \mathbf{z}))}{P(T > t_i|c_i = 0)\pi_i(\boldsymbol{\phi}, \mathbf{z}) + P(T > t_i|c_i = 1)(1 - \pi_i(\boldsymbol{\phi}, \mathbf{z}))} \\ &= \frac{S_{SRWD}(t_i; \theta)(1 - \pi_i(\boldsymbol{\phi}, \mathbf{z}))}{\pi_i(\boldsymbol{\phi}, \mathbf{z}) + (1 - \pi_i(\boldsymbol{\phi}, \mathbf{z}))S_{SRWD}(t_i; \theta)} \end{aligned}$$

Esses resultados mostram que se uma observação falhou ($\delta_i = 1$) ela só pode não estar curada, logo $P(c_i = 1|\Phi, \delta_i = 1) = 1$. Já para o caso das observações censuradas ($\delta_i = 0$) é possível que elas sejam provenientes do grupo das que não estão curadas e estejam sob risco ou provenientes do grupo das que estão curadas, sendo possível atribuir as probabilidades de pertencer a cada um dos grupos:

$$P(c_i = 0|\Phi, \delta_i = 0) = \frac{\pi_i(\boldsymbol{\phi}, \mathbf{z})}{\pi_i(\boldsymbol{\phi}, \mathbf{z}) + (1 - \pi_i(\boldsymbol{\phi}, \mathbf{z}))S_{SRWD}(t_i; \theta)}$$

$$P(c_i = 1|\Phi, \delta_i = 0) = \frac{(1 - \pi_i(\boldsymbol{\phi}, \mathbf{z}))S_{SRWD}(t_i; \theta)}{\pi_i(\boldsymbol{\phi}, \mathbf{z}) + (1 - \pi_i(\boldsymbol{\phi}, \mathbf{z}))S_{SRWD}(t_i; \theta)}$$

Dessa forma, podemos definir a estimativa da esperança condicional de c_i, dada a estimativa de Φ, como:

$$\begin{aligned} p_i &= E(c_i|\Phi) \\ &= \delta_i + (1 - \delta_i)\frac{(1 - \pi_i(\boldsymbol{\phi}, \mathbf{z}))S_{SRWD}(t_i; \theta)}{\pi_i(\boldsymbol{\phi}, \mathbf{z}) + (1 - \pi_i(\boldsymbol{\phi}, \mathbf{z}))S_{SRWD}(t_i; \theta)} \end{aligned} \tag{3.9}$$

Substituindo c_i pela estimativa da esperança de c_i condicional a Φ em (3.8), obtemos:

$$\begin{aligned} l_c(\Phi) &= \sum_{i=1}^{m} \delta_i \log(1-\pi_i(\boldsymbol{\phi},\mathbf{z})) + \sum_{i=(m+1)}^{n} \delta_i \log(1-\pi_i(\boldsymbol{\phi},\mathbf{z})) + \\ & \sum_{i=1}^{m} \delta_i \log[p_{SRWD}(t_i;\theta)] + \sum_{i=(m+1)}^{n} \delta_i \log[p_{SRWD}(t_i;\theta)] + \\ & \sum_{i=1}^{m} (1-\delta_i) \log[\{\pi_i(\boldsymbol{\phi},\mathbf{z})\}^{1-p_i} + \{(1-\pi_i(\boldsymbol{\phi},\mathbf{z}))S_{SRWD}(t_i;\theta)\}^{p_i}] + \\ & \sum_{i=(m+1)}^{n} (1-\delta_i) \log[\{\pi_i(\boldsymbol{\phi},\mathbf{z})\}^{1-p_i} + \{(1-\pi_i(\boldsymbol{\phi},\mathbf{z}))S_{SRWD}(t_i;\theta)\}^{p_i}]. \end{aligned} \tag{3.10}$$

Partindo do princípio de que nos tempos de falha $\delta_i = 1$, temos que $p_i = 1$ para $i = 1, \ldots, m$. Já para os casos de tempos censurados, $i = (m+1), \ldots, n$, temos que $p_i = \frac{(1-\pi_i(\boldsymbol{\phi},\mathbf{z}))S_{SRWD(t_i;\theta)}}{\pi_i(\boldsymbol{\phi},\mathbf{z})+(1-\pi_i(\boldsymbol{\phi},\mathbf{z}))S_{SRWD}(t_i;\theta)}$. Considerando estas informações, podemos definir:

$$\begin{aligned} w_{1i}(\mathbf{z}) &= 1 - p_i = P(c_i = 0 | T > t_i) = \frac{\pi_i(\boldsymbol{\phi},\mathbf{z})}{g(x;\Phi)} \\ &= \frac{\pi_i(\boldsymbol{\phi},\mathbf{z})}{\pi_i(\boldsymbol{\phi},\mathbf{z}) + (1-\pi_i(\boldsymbol{\phi},\mathbf{z}))S_{SRWD}(t_i;\theta)} \end{aligned}$$

$$\begin{aligned} w_{2i}(\mathbf{z}) &= p_i = P(c_i = 1 | T > t_i) = \frac{(1-\pi_i(\boldsymbol{\phi},\mathbf{z}))S_{SRWD}(t_i;\theta)}{g(x;\Phi)} \\ &= \frac{(1-\pi_i(\boldsymbol{\phi},\mathbf{z}))S_{SRWD}(t_i;\theta)}{\pi_i(\boldsymbol{\phi},\mathbf{z}) + (1-\pi_i(\boldsymbol{\phi},\mathbf{z}))S_{SRWD}(t_i;\theta)} \end{aligned}$$

A partir desses resultados e definições, podemos obter uma forma mais simplificada do logaritmo da verossimilhança dos dados completos apresentado em (3.10):

$$
\begin{aligned}
l_c(\Phi) &= \sum_{i=1}^{m} \log[p_{SRWD}(t_i;\theta)] + \sum_{i=1}^{m} \log[1-\pi_i(\boldsymbol{\phi},\mathbf{z})] + \\
&\sum_{i=(m+1)}^{n} w_{1i}(\mathbf{z}) \log[\pi_i(\boldsymbol{\phi},\mathbf{z})] + \sum_{i=(m+1)}^{n} w_{2i}(\mathbf{z}) \log[1-\pi_i(\boldsymbol{\phi},\mathbf{z})] + \\
&\sum_{i=(m+1)}^{n} w_{2i}(\mathbf{z}) \log[S_{SRWD}(t_i;\theta)]
\end{aligned}
$$

Essa expressão de $l_c(\Phi)$ é similar à apresentada por Kannan *et al* (2010).

Com o intuito de facilitar o processo de estimação, o logaritmo da verossimilhança $l_c(\Phi)$ pode ainda ser escrito como:

$$
l_c(\Phi) = Q(\Phi,\Phi^{(k)}) = g_1(\boldsymbol{\phi}) + g_2(q,\beta), \tag{3.11}
$$

na qual

$$
g_1(\boldsymbol{\phi}) = \sum_{i=1}^{m} \log[1-\pi_i(\boldsymbol{\phi}^{(k)},\mathbf{z})] + \sum_{i=(m+1)}^{n} w_{1i}(\mathbf{z})^{(k)} \log[\pi_i(\boldsymbol{\phi}^{(k)},\mathbf{z})] +
$$

$$
\sum_{i=(m+1)}^{n} w_{2i}(\mathbf{z})^{(k)} \log[1-\pi_i(\boldsymbol{\phi}^{(k)},\mathbf{z})]
$$

e

$$
g_2(q,\beta) = \sum_{i=1}^{m} \log[p_{SRWD}(t_i;\theta^{(k)})] + \sum_{i=(m+1)}^{n} w_{2i}(\mathbf{z})^{(k)} \log[S_{SRWD}(t_i;\theta^{(k)})].
$$

$Q(\Phi,\Phi^{(k)})$ é a esperança do logaritmo da verossimilhança dos dados completos e no passo M é maximizada em relação aos parâmetros desconhecidos, considerando valores fixos de $w_{1i}(\mathbf{z})$ e $w_{2i}(\mathbf{z})$. A partir de (3.11), as estimativas de máxima verossimilhança de $\boldsymbol{\phi}$ podem ser obtidas separadamente das estimativas de q e β, pois g_1 depende apenas de $\boldsymbol{\phi}$ e g_2 depende só de β e q. Assim, se $\boldsymbol{\phi}^{(k)}, \beta^{(k)}$ e $q^{(k)}$ são estimativas de $\boldsymbol{\phi}, \beta$ e q, respectivamente, na k-ésima iteração, então $\boldsymbol{\phi}^{(k+1)}$ é obtida pela maximização de $g_1(\boldsymbol{\phi})$, enquanto que $\beta^{(k+1)}$ e $q^{(k+1)}$ são obtidas pela maximização de $g_2(\beta,q)$, todos considerando valores fixos para $w_{1i}(\mathbf{z})$ e $w_{2i}(\mathbf{z})$. Para realizar as maximizações de $g_1(\boldsymbol{\phi})$ e $g_2(\beta,q)$ no passo $(k+1)$, $w_{1i}(\mathbf{z})$ e $w_{2i}(\mathbf{z})$ são obtidos da seguinte forma a partir de $\boldsymbol{\phi}^{(k)}$ e $\theta^{(k)} = (\beta^{(k)}, q^{(k)})$:

$$w_{1i}(\mathbf{z})^{(k+1)} = \frac{\pi_i(\boldsymbol{\phi}^{(k)}, \mathbf{z})}{\pi_i(\boldsymbol{\phi}^{(k)}, \mathbf{z}) + (1 - \pi_i(\boldsymbol{\phi}^{(k)}, \mathbf{z}))S_{SRWD}(t_i; \theta^{(k)})} \tag{3.12}$$

$$w_{2i}(\mathbf{z})^{(k+1)} = \frac{(1 - \pi_i(\boldsymbol{\phi}^{(k)}, \mathbf{z}))S_{SRWD}(t_i; \theta^{(k)})}{\pi_i(\boldsymbol{\phi}^{(k)}, \mathbf{z}) + (1 - \pi_i(\boldsymbol{\phi}^{(k)}, \mathbf{z}))S_{SRWD}(t_i; \theta^{(k)})} \tag{3.13}$$

De modo geral, o passo E é a etapa do algoritmo EM na qual são atribuídas as probabilidades de cura para cada indivíduo, π_i, que são dadas pelas estimativas das esperanças condicionais p_i, que foram reescritas na forma de $w_{1i}(\mathbf{z})$ e $w_{2i}(\mathbf{z})$. Já o passo M consiste na maximização da função de verossimilhança, com base em $w_{1i}(\mathbf{z})$ e $w_{2i}(\mathbf{z})$, para obter as estimativas dos parâmetros $\boldsymbol{\phi}$, q e β. Deste modo, ao implementar o algoritmo, serão obtidas as estimativas dos parâmetros β e q associados à distribuição dos tempos de sobrevivência e de um vetor $\boldsymbol{\phi}$ de parâmetros relacionado à probabilidade de cura π_i.

Implementação do algoritmo EM

Dados $\boldsymbol{\phi}^{(k-1)}, \beta^{(k-1)}$ e $q^{(k-1)}$, as etapas da iteração k do algoritmo EM podem ser descritas da seguinte forma:

1. (Passo E) A partir de $\boldsymbol{\phi}^{(k-1)}, \beta^{(k-1)}$ e $q^{(k-1)}$, obter $w_{1i}(\mathbf{z})^{(k)}$ e $w_{2i}(\mathbf{z})^{(k)}$ utilizando (3.12) e (3.13), respectivamente, e calcular $Q(\Phi, \Phi^{(k)})$ conforme foi apresentada em (3.11);

2. (Passo M) Atualizar os parâmetros $\boldsymbol{\phi}, \beta$ e q. $\boldsymbol{\phi}^{(k)}$ é obtido através da maximização de $g_1(\boldsymbol{\phi})$ com respeito a $\boldsymbol{\phi}$, enquanto $\beta^{(k)}$ e $q^{(k)}$ são obtidos através da maximização de $g_2(q, \beta)$ com respeito a β e q, respectivamente;

3. Declarar convergência se algum critério de parada for atingido.

Capítulo 4

Simulações

Neste capítulo serão apresentadas as simulações computacionais e seus resultados, que foram obtidos através do software `R` na versão 3.2.2. O objetivo foi gerar dados de sobrevivência com fração de cura e com a presença de covariáveis a fim de obter as estimativas dos parâmetros do modelo apresentado em (3.1), (3.2) e (3.3).

Foram considerados três tamanhos de amostra, sendo $n = 250$, $n = 500$ e $n = 1.000$. Os tempos foram gerados a partir da distribuição Weibull discreta utilizando o pacote DiscreteWeibull do `R` (BARBIERO, 2015) e, para possibilitar a ilustração da fração de cura, o tempo das observações curadas foi dado pela parte inteira de $1,5\times$TM, em que TM é o maior tempo gerado. O mecanismo de censura utilizado foi o de censura à direita

A variação dos valores do vetor de parâmetros $\Phi = (\boldsymbol{\phi}, q, \beta)$ permite a criação de cenários com diferentes características dos dados: os valores de $\boldsymbol{\phi}$ influenciam na probabilidade de cura; o valor de q interfere no tamanho dos tempos gerados e principalmente no percentual de tempos iguais a zero (quanto menor o q, menores os valores dos tempos e maior o número de tempos iguais a zero); enquanto que β é inversamente proporcional aos tempos de modo que quanto menor o valor de β, maiores os valores dos tempos e, assim, há uma quantidade maior de tempos distintos.

Os resultados das simulações apresentados neste trabalho consideraram uma única covariável dicotômica, Z, no modelo, que foi gerada a partir de uma distribuição Bernoulli com probabilidade de sucesso $p = 0,6$. Simulações com covariáveis não dicotômicas também foram realizadas, considerando uma distribuição Poisson, e os resultados obtidos quanto às estimativas foram similares. Optou-se pela apresentação dos resultados com covariável dicotômica pelo fato de ser possível visualizá-los graficamente através das estimativas das curvas de sobrevivência.

No total foram considerados quatro cenários (que diferem quanto à quantidade total de censura nos dados) e 18 valores de Φ. O percentual de censura das observações sob risco foi igual a 10% nos três primeiros cenários. Em todos os cenários temos $\Phi_2 = (\boldsymbol{\phi}, q = 0,5, \beta = 1)$ e $\Phi_5 = (\boldsymbol{\phi}, q = 0,9, \beta = 1)$ que representam dados de tempos que seguem distribuição Geométrica, visto que em ambos os casos temos $\beta = 1$.

Para cada Φ em cada um dos cenários foram realizadas 1.000 simulações e o erro quadrático médio (EQM) de cada uma das estimativas foi calculado. No que se refere ao algoritmo EM, as maximizações das funções $g_1(\boldsymbol{\phi})$ e $g_2(q, \beta)$ foram feitas utilizando a função optim do `R` (R Core Team, 2015). O critério de parada utilizado foi quando o maior valor absoluto da diferença entre as estimativas dos respectivos parâmetros nas iterações k e $k + 1$ ficasse menor que 1×10^{-100}.

A Tabela 4.1 apresenta os valores dos parâmetros utilizados na geração do Cenário 1, que representa dados com baixo percentual de censura (em média, 26%). Este percentual de censura refere-se ao total de censura de todas as observações, estando sob risco ou curadas. O percentual médio de valores de tempos iguais a zero foi: 40%, quando q = 0,5; e 8%, quando q = 0,9. Além disto, as probabilidades de cura condicionadas aos valores da covariável Z das observações geradas neste cenário são: $\pi_i(\boldsymbol{\phi}, Z = 0) = P(c_i = 0|Z_i = 0) = 11,9\%$ e $\pi_i(\boldsymbol{\phi}, Z = 1) = P(c_i = 0|Z_i = 1) = 23,1\%$. Os resultados das estimativas obtidas juntamente com cada EQM são mostrados na Tabela 4.2.

Tabela 4.1: Valores dos parâmetros utilizados na simulação do Cenário 1. Censura média: 26%

Φ	q	β	ϕ_0	ϕ_1
1	0,50	0,50	-2,00	0,80
2	0,50	1,00	-2,00	0,80
3	0,50	2,00	-2,00	0,80
4	0,90	0,50	-2,00	0,80
5	0,90	1,00	-2,00	0,80
6	0,90	2,00	-2,00	0,80

Com base nos resultados da Tabela 4.2, nota-se que as médias das estimativas dos parâmetros estão próximas aos seus verdadeiros valores em todas os casos. Nota-se também que há uma tendência de subestimar os valores de ϕ_0 e de ϕ_1 e de superestimar os parâmetros q e β, mas sem prejudicar a qualidade dos resultados.

Tabela 4.2: Resultados das médias e EQM das estimativas de máxima verossimilhança dos parâmetros obtidas através de 1.000 réplicas de Monte Carlo do algoritmo EM em dados com baixo percentual de censura (Cenário 1 - Censura média: 26%).

Φ	n	q	$EQM(\hat{q})$	β	$EQM(\hat{\beta})$	ϕ_0	$EQM(\hat{\phi_0})$	ϕ_1	$EQM(\hat{\phi_1})$
1		0,528	2×10^{-3}	0,511	0,002	-1,783	0,150	0,786	0,138
2		0,528	2×10^{-3}	1,030	0,008	-1,734	0,164	0,744	0,141
3	250	0,522	2×10^{-3}	2,026	0,030	-1,644	0,214	0,695	0,135
4		0,906	3×10^{-4}	0,504	0,001	-1,802	0,140	0,778	0,138
5		0,908	3×10^{-4}	1,021	0,004	-1,790	0,143	0,782	0,143
6		0,906	3×10^{-4}	2,025	0,017	-1,751	0,159	0,753	0,135
1		0,528	1×10^{-3}	0,512	0,001	-1,738	0,114	0,754	0,063
2		0,528	1×10^{-3}	1,025	0,004	-1,714	0,132	0,731	0,074
3	500	0,523	1×10^{-3}	2,018	0,015	-1,630	0,181	0,687	0,074
4		0,907	2×10^{-4}	0,505	0,001	-1,771	0,099	0,758	0,062
5		0,907	2×10^{-4}	1,016	0,002	-1,756	0,105	0,755	0,064
6		0,907	2×10^{-4}	2,028	0,009	-1,719	0,126	0,735	0,069
1		0,529	1×10^{-3}	0,514	0,001	-1,734	0,094	0,745	0,036
2		0,528	1×10^{-3}	1,024	0,002	-1,706	0,109	0,728	0,039
3	1000	0,523	8×10^{-4}	2,016	0,008	-1,632	0,158	0,686	0,043
4		0,907	1×10^{-4}	0,505	0,000	-1,761	0,080	0,748	0,035
5		0,908	1×10^{-4}	1,016	0,001	-1,753	0,085	0,752	0,034
6		0,907	1×10^{-4}	2,020	0,004	-1,717	0,104	0,738	0,038

As estimativas de q e β são mais precisas do que as de ϕ por apresentarem valores de EQM menores. Ainda, é possível notar que o $EQM(\hat{q})$ é menor nos casos em que o percentual de valores de tempos iguais a zero é menor. Já o $EQM(\hat{\beta})$ diminui quando o parâmetro β assume valores menores, ou seja, os tempos gerados são grandes e com mais tempos distintos. Nota-se que os valores de q e β influenciam também as estimativas de ϕ_0, visto que quando o percentual de tempos iguais a zero é maior e o número de tempos distintos é pequeno (ou seja, quando $q = 0,5$ e $\beta = 2$) o $EQM(\hat{\phi}_0)$ é maior do que nos outros casos. Por fim, temos que os valores do EQM para todos os parâmetros diminui ao passo em que a amostra aumenta, entretanto estas diferenças são pequenas e aparecem nas casas decimais, o que mostra o bom resultado das estimativas mesmo em amostras menores.

A Figura 4.1 ilustra as funções de sobrevivência estimadas para cada um dos valores de Φ, considerando amostras simuladas de tamanho $n = 500$. É possível perceber o bom ajuste obtido pelo modelo de regressão Weibull discreto com fração de cura.

O Cenário 2 foi caracterizado por dados com moderado percentual de censura (em média 50%) e os valores dos parâmetros utilizados para gerar estes dados estão presentes

na Tabela 4.3. Neste caso, o percentual médio de tempos iguais a zero para $q = 0,5$ foi igual a 28%, e para $q = 0,9$ foi igual a 6%. As probabilidades de cura para os dois grupos criados a partir dos valores de Z foram: $\pi_i(\boldsymbol{\phi}, Z = 0) = P(c_i = 0|Z_i = 0) = 73,1\%$ e $\pi_i(\boldsymbol{\phi}, Z = 1) = P(c_i = 0|Z_i = 1) = 26,9\%$. Os resultados das estimativas com seus respectivos EQM são mostrados na Tabela 4.4.

Tabela 4.3: Valores dos parâmetros utilizados na simulação do Cenário 2. Censura média: 50%.

Φ	q	β	ϕ_0	ϕ_1
1	0,50	0,50	1,00	-2,00
2	0,50	1,00	1,00	-2,00
3	0,50	2,00	1,00	-2,00
4	0,90	0,50	1,00	-2,00
5	0,90	1,00	1,00	-2,00
6	0,90	2,00	1,00	-2,00

Tabela 4.4: Resultados das médias e EQM das estimativas de máxima verossimilhança dos parâmetros obtidas através de 1.000 réplicas de Monte Carlo do algoritmo EM em dados com moderado percentual de censura (Cenário 2 - Censura média: 50%).

Φ	n	q	$EQM(\hat{q})$	β	$EQM(\hat{\beta})$	ϕ_0	$EQM(\hat{\phi_0})$	ϕ_1	$EQM(\hat{\phi_1})$
1		0,520	2×10^{-3}	0,507	3×10^{-3}	1,125	0,073	-1,9327	0,098
2		0,519	2×10^{-3}	1,025	1×10^{-2}	1,145	0,081	-1,934	0,098
3	250	0,514	2×10^{-3}	2,033	4×10^{-2}	1,140	0,081	-1,914	0,106
4		0,905	4×10^{-4}	0,506	2×10^{-3}	1,125	0,073	-1,951	0,096
5		0,905	4×10^{-4}	1,018	7×10^{-3}	1,138	0,076	-1,964	0,095
6		0,904	4×10^{-4}	2,033	2×10^{-2}	1,145	0,084	-1,945	0,107
1		0,518	1×10^{-3}	0,508	1×10^{-3}	1,132	0,046	-1,942	0,052
2		0,519	1×10^{-3}	1,019	5×10^{-3}	1,126	0,044	-1,915	0,055
3	500	0,515	1×10^{-3}	2,038	2×10^{-2}	1,145	0,049	-1,900	0,056
4		0,904	2×10^{-4}	0,504	7×10^{-4}	1,135	0,045	-1,967	0,047
5		0,905	2×10^{-4}	1,013	3×10^{-3}	1,126	0,044	-1,947	0,049
6		0,906	2×10^{-4}	2,035	1×10^{-2}	1,133	0,044	-1,934	0,046
1		0,519	9×10^{-4}	0,509	7×10^{-4}	1,138	0,033	-1,941	0,025
2		0,518	8×10^{-4}	1,019	3×10^{-3}	1,125	0,030	-1,910	0,030
3	1000	0,515	8×10^{-4}	2,022	1×10^{-2}	1,141	0,035	-1,898	0,032
4		0,904	1×10^{-4}	0,503	3×10^{-4}	1,125	0,029	-1,938	0,026
5		0,905	1×10^{-4}	1,011	2×10^{-3}	1,131	0,031	-1,944	0,026
6		0,905	1×10^{-4}	2,023	7×10^{-3}	1,125	0,030	-1,919	0,030

Os resultados da Tabela 4.4 mostram que os valores do EQM das estimativas de todos os parâmetros diminuem em amostras maiores. Observa-se que em dados com maior percentual de valores de tempos iguais a zero, ou seja, com $q = 0,5$, os valores de $EQM(\hat{q})$ e $EQM(\hat{\beta})$ são maiores. Já o parâmetro β interfere nos valores de $EQM(\hat{\beta})$

e $EQM(\hat{\phi}_0)$, que crescem ao passo em que β também cresce. Em todos os casos as médias das estimativas se aproximaram dos valores dos parâmetros e apresentaram EQM pequeno. Neste cenário houve superestimação de q, β e ϕ_0 e subestimação de ϕ_1, mas sem afetar a qualidade dos resultados. A Figura 4.2 ilustra as estimativas das curvas de sobrevivência deste cenário obtidas via Kaplan-Meier e por meio do modelo de sobrevivência proposto.

Já o Cenário 3 foi gerado de modo que os dados tivessem alto percentual de censura (em média 70%). As probabilidades de cura condicionadas aos valores gerados da covariável Z são: $\pi_i(\boldsymbol{\phi}, Z = 0) = P(c_i = 0|Z_i = 0) = 42,6\%$ e $\pi_i(\boldsymbol{\phi}, Z = 1) = P(c_i = 0|Z_i = 1) = 84,6\%$. Os percentuais médios de tempos iguais a zero dependem do valor de q e foram iguais a: 16%, para $q = 0,5$; e 3% para $q = 0,9$. A Tabela 4.5 apresenta os valores dos parâmetros utilizados para gerar os dados deste cenário e a Tabela 4.6 apresenta os resultados das médias das estimativas e do EQM, obtidos através das simulações.

Tabela 4.5: Valores dos parâmetros utilizados na simulação do Cenário 3. Censura média: 70%.

Φ	q	β	ϕ_0	ϕ_1
1	0,50	0,50	-0,30	2,00
2	0,50	1,00	-0,30	2,00
3	0,50	2,00	-0,30	2,00
4	0,90	0,50	-0,30	2,00
5	0,90	1,00	-0,30	2,00
6	0,90	2,00	-0,30	2,00

Analisando os resultados apresentados na Tabela 4.6, nota-se um desempenho semelhante aos citados nos Cenários 1 e 2. O aumento do tamanho das amostras faz com que haja uma redução no EQM das estimativas, e os valores de q e β, usados para gerar os tempos, impactam diretamente nas suas características, o que faz com que as estimativas também sejam influenciadas. Novamente, há uma leve subestimação dos valores de ϕ_0 e ϕ_1 enquanto que os parâmetros q e β são superestimados. Mesmo com cerca de 70% das observações sendo censuradas, de modo geral foram obtidos bons resultados, que são ilustrados na Figura 4.3 através das estimativas das curvas de sobrevivência.

Por fim, o Cenário 4 foi gerado com os mesmos parâmetros do Cenário 2 (Tabela 4.3),

Tabela 4.6: Resultados das médias e EQM das estimativas de máxima verossimilhança dos parâmetros obtidas através de 1.000 réplicas de Monte Carlo do algoritmo EM em dados com alto percentual de censura (Cenário 3 - Censura média: 70%).

Φ	n	q	$EQM(\hat{q})$	β	$EQM(\hat{\beta})$	ϕ_0	$EQM(\hat{\phi_0})$	ϕ_1	$EQM(\hat{\phi_1})$
1		0,514	3×10^{-3}	0,501	0,005	-0,145	0,067	1,969	0,101
2		0,511	3×10^{-3}	1,027	0,020	-0,137	0,071	1,964	0,101
3	250	0,512	4×10^{-3}	2,097	0,164	-0,115	0,076	1,968	0,102
4		0,902	7×10^{-4}	0,502	0,003	-0,167	0,062	1,995	0,106
5		0,901	7×10^{-4}	1,012	0,010	-0,151	0,063	1,980	0,093
6		0,905	8×10^{-4}	2,067	0,051	-0,129	0,071	1,960	0,102
1		0,511	2×10^{-3}	0,504	0,002	-0,145	0,046	1,970	0,049
2		0,511	2×10^{-3}	1,020	0,009	-0,128	0,051	1,953	0,052
3	500	0,509	2×10^{-3}	2,043	0,043	-0,108	0,056	1,937	0,050
4		0,902	4×10^{-4}	0,503	0,001	-0,153	0,044	1,985	0,053
5		0,903	4×10^{-4}	1,010	0,005	-0,147	0,044	1,972	0,053
6		0,904	3×10^{-4}	2,033	0,021	-0,134	0,047	1,963	0,053
1		0,511	1×10^{-3}	0,507	0,001	-0,132	0,038	1,954	0,025
2		0,511	1×10^{-3}	1,015	0,005	-0,130	0,040	1,958	0,027
3	1000	0,511	1×10^{-3}	2,036	0,012	-0,112	0,046	1,930	0,028
4		0,903	2×10^{-4}	0,503	0,001	-0,147	0,034	1,961	0,026
5		0,903	2×10^{-4}	1,009	0,002	-0,151	0,033	1,974	0,024
6		0,903	2×10^{-4}	2,022	0,011	-0,135	0,038	1,954	0,026

com a diferença de que neste há censura apenas nos indivíduos curados. O percentual médio de censura é igual a 46%. A Tabela 4.7 apresenta os resultados das estimativas e do EQM obtidos através das simulações.

Com base nos resultados da Tabela 4.7 é possível observar o impacto positivo da ausência de censura nas observações sob risco na qualidade das estimativas. Ao comparar com os resultados obtidos no Cenário 2, nota-se que o EQM das estimativas no Cenário 4 é menor, principalmente com tamanho de amostra pequeno. Diferentemente dos demais cenários, neste não houve um padrão de sub ou superestimação dos parâmetros. Mas, é possível perceber os mesmos comportamentos de impactos nas estimativas a depender das características dos tempos gerados e do tamanho de amostra. A Figura 4.4 ilustra as curvas de sobrevivência estimadas pelo método de Kaplan-Meier e por meio do modelo proposto e, através delas, podemos ver o bom ajuste do modelo de regressão Weibull Discreto com fração de cura.

O modelo apresentou bons resultados ao ser aplicado em dados de sobrevivência com tempos discretos, presença de fração de cura e covariáveis, mesmo em situações com diferentes tamanhos de amostra e diferentes percentuais de tempos iguais a zero e de

Tabela 4.7: Resultados das médias e EQM das estimativas de máxima verossimilhança dos parâmetros obtidas através de 1.000 réplicas de Monte Carlo do algoritmo EM em dados com censura apenas nas observações curadas (Cenário 4 - Censura média: 46%).

Φ	n	q	$EQM(\hat{q})$	β	$EQM(\hat{\beta})$	ϕ_0	$EQM(\hat{\phi_0})$	ϕ_1	$EQM(\hat{\phi_1})$
1		0,499	2×10^{-3}	0,493	2×10^{-3}	1,013	0,050	-2,027	0,083
2		0,501	2×10^{-3}	1,007	9×10^{-3}	1,016	0,055	-2,043	0,089
3	250	0,500	2×10^{-3}	2,037	6×10^{-2}	0,998	0,049	-2,005	0,087
4		0,898	4×10^{-4}	0,498	1×10^{-3}	1,000	0,052	-2,022	0,091
5		0,900	4×10^{-4}	1,008	6×10^{-3}	1,001	0,054	-2,008	0,090
6		0,900	4×10^{-4}	2,022	2×10^{-2}	1,016	0,051	-2,019	0,086
1		0,500	8×10^{-4}	0,498	1×10^{-3}	0,999	0,027	-2,009	0,044
2		0,499	9×10^{-4}	1,004	4×10^{-3}	0,999	0,028	-2,002	0,046
3	500	0,501	9×10^{-4}	2,017	2×10^{-2}	1,008	0,024	-2,013	0,042
4		0,899	2×10^{-4}	0,499	6×10^{-4}	0,997	0,026	-2,001	0,045
5		0,901	2×10^{-4}	1,006	3×10^{-3}	1,009	0,026	-2,011	0,043
6		0,899	2×10^{-4}	2,004	1×10^{-2}	1,005	0,026	-2,003	0,042
1		0,499	4×10^{-4}	0,499	6×10^{-4}	0,997	0,013	-2,001	0,022
2		0,501	4×10^{-4}	1,005	2×10^{-3}	1,008	0,013	-2,011	0,022
3	1000	0,500	5×10^{-4}	2,008	1×10^{-2}	1,007	0,014	-2,014	0,023
4		0,899	1×10^{-4}	0,499	3×10^{-4}	1,001	0,013	-2,000	0,021
5		0,899	1×10^{-4}	1,001	1×10^{-3}	1,003	0,013	-2,004	0,021
6		0,899	1×10^{-4}	2,003	5×10^{-3}	0,999	0,013	-1,999	0,020

dados censurados. Entretanto, ao realizar os estudos de simulação, observou-se que é necessário ter cautela ao aplicar este modelo em dados com percentual de censura muito baixo, principalmente em amostras pequenas. Isto se deve ao fato de que sob estas condições é possível que, ao dividir a amostra em grupos de acordo com os valores das covariáveis categóricas, um deles (ou mais) não tenha indivíduos curados, ou seja, todos os indivíduos com características semelhantes sofreram o evento de interesse e a função de sobrevivência assumiu valor zero. Isto causa prejuízos nos resultados das estimativas de $\boldsymbol{\phi}$, visto que nesta situação estará sendo feita a modelagem com o intuito de estimar a probabilidade de cura ($\pi_i(\boldsymbol{\phi}, \mathbf{z}) = \mathbf{P}(\mathbf{c_i} = \mathbf{0}|\mathbf{Z})$) utilizando uma amostra na qual todos os indivíduos são não curados (o maior tempo observado é de falha).

Diante do exposto, recomenda-se que antes da aplicação do modelo de regressão com fração de cura seja construído o gráfico com as estimativas da curva de sobrevivência via Kaplan-Meier. Com este resultado será possível verificar se nos dados há ou não algum grupo no qual todos os indivíduos tenham vindo a sofrer o evento de interesse.

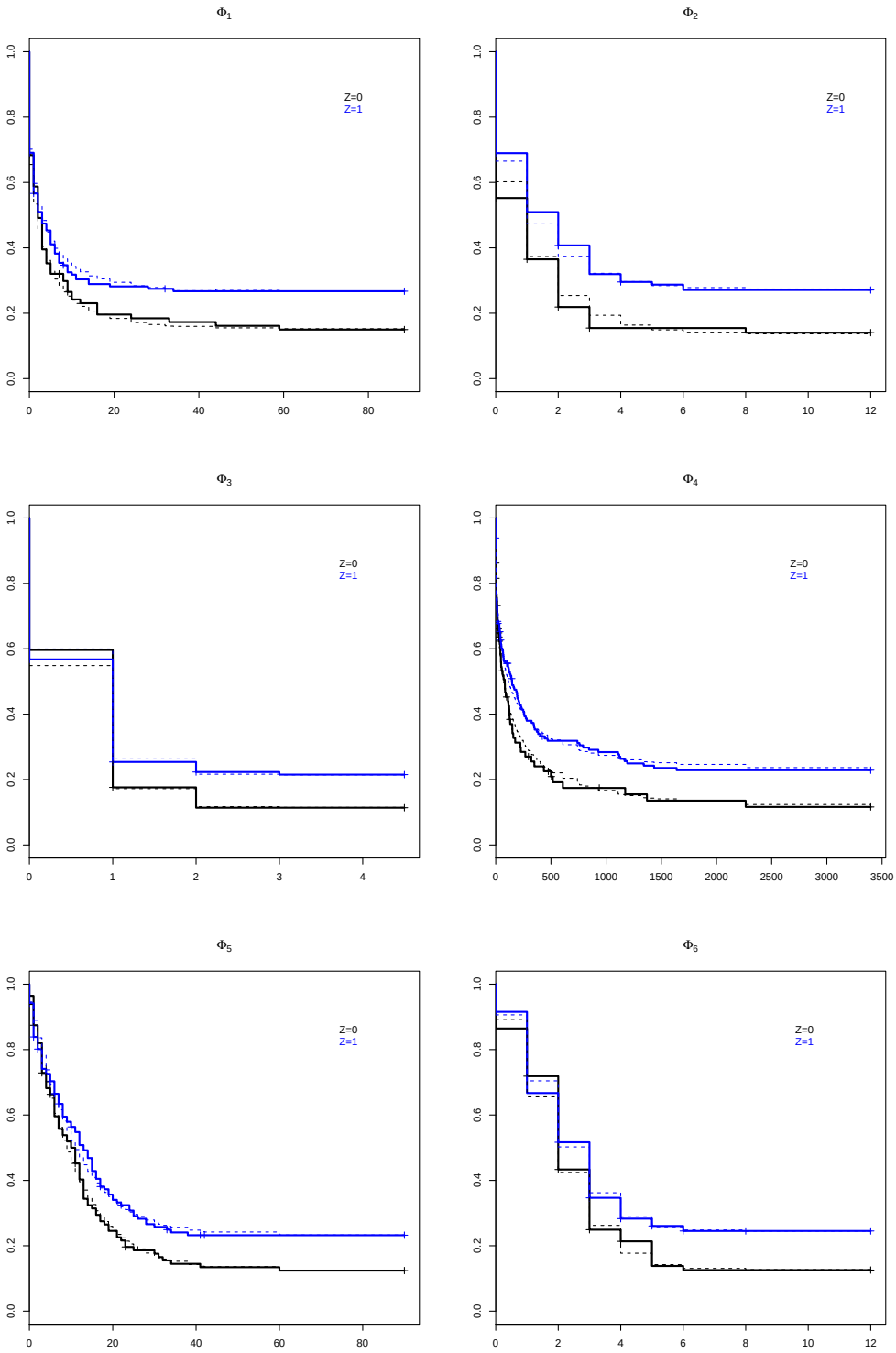

Figura 4.1: Funções de sobrevivência estimadas para o Cenário 1 considerando uma amostra de tamanho n=500. As linhas sólidas representam as estimativas de Kaplan-Meier e as linhas tracejadas as estimativas pelo modelo proposto.

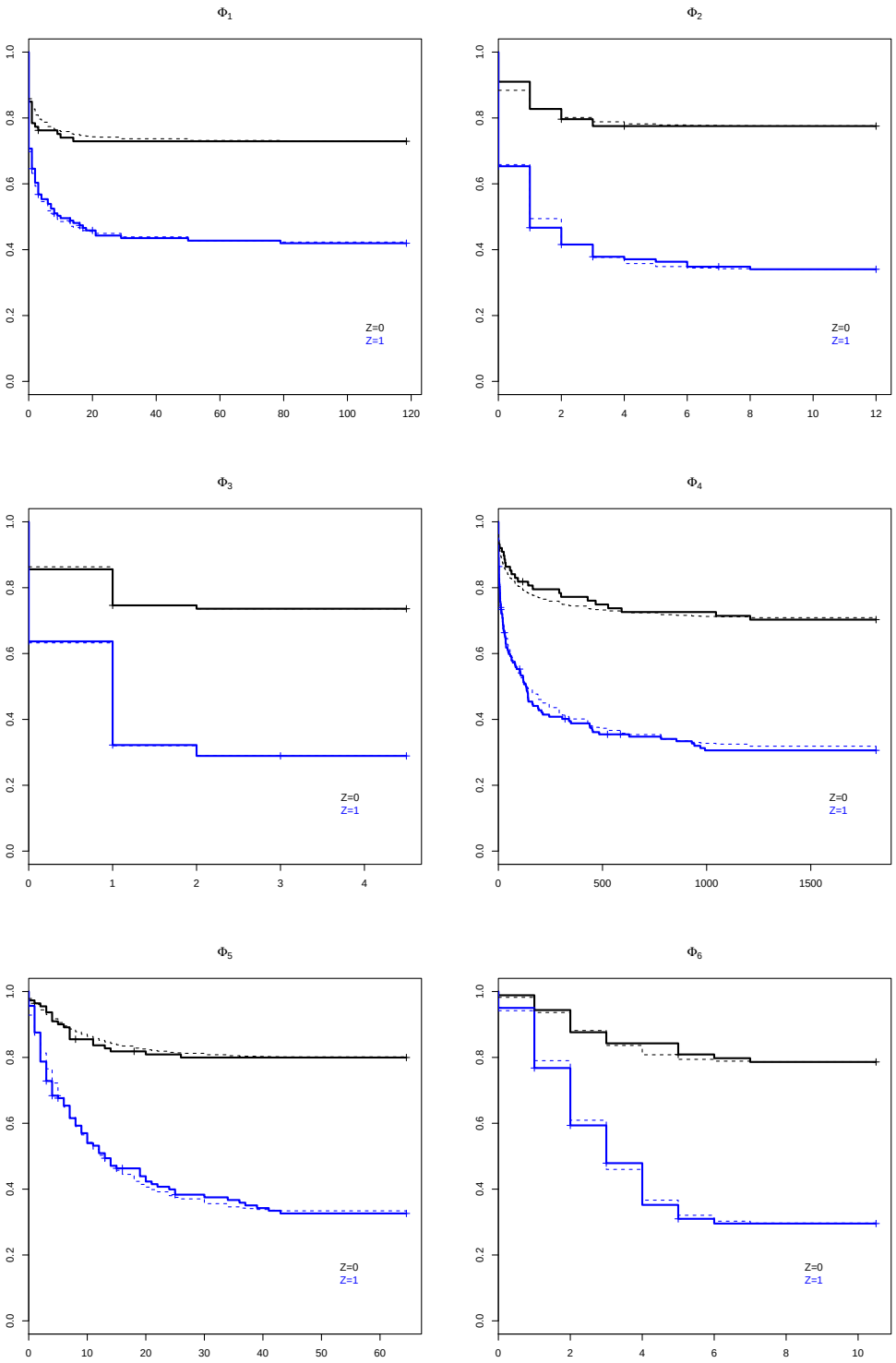

Figura 4.2: Funções de sobrevivência estimadas para o Cenário 2 considerando uma amostra de tamanho n=500. As linhas sólidas representam as estimativas de Kaplan-Meier e as linhas tracejadas as estimativas pelo modelo proposto.

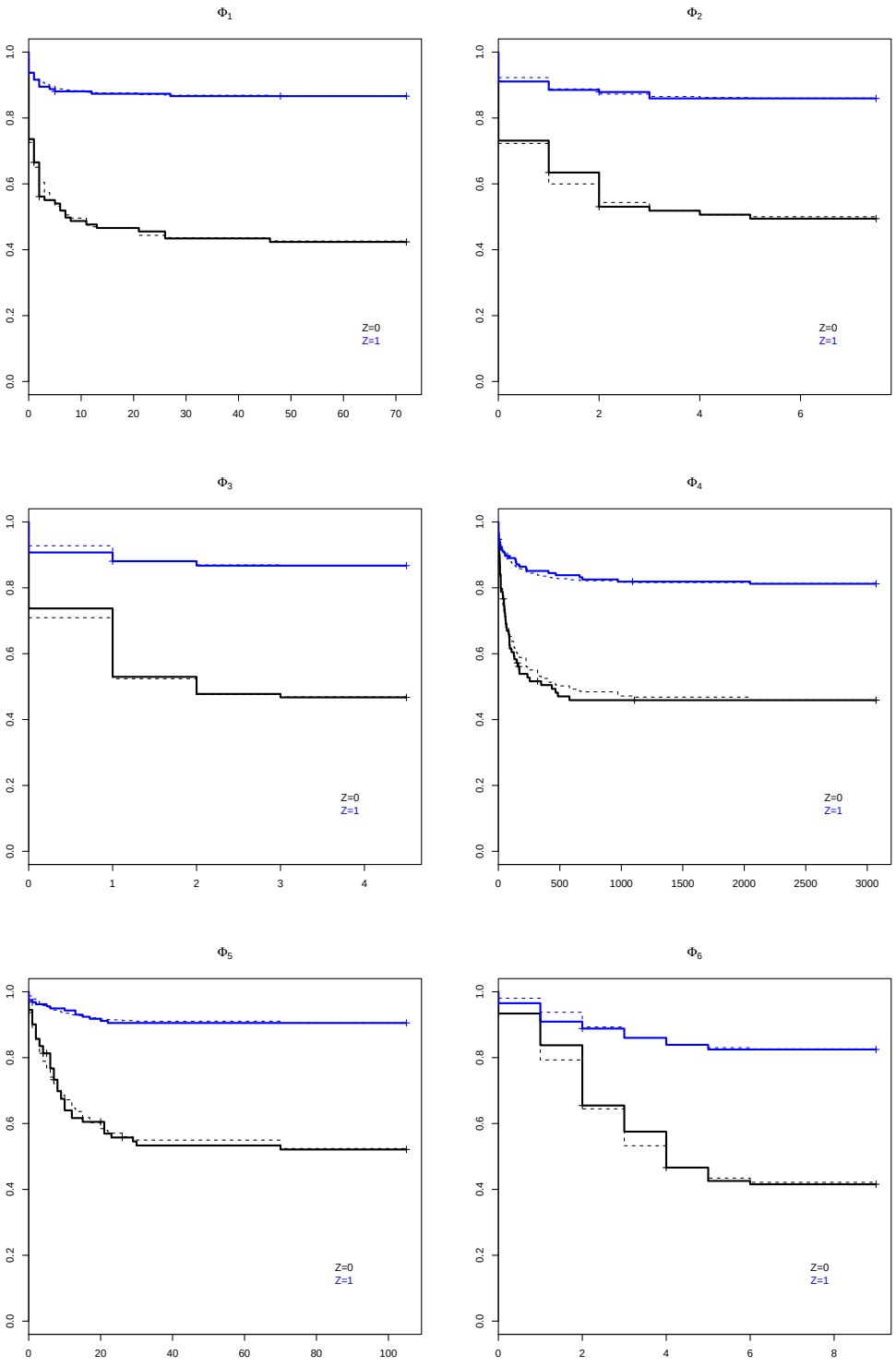

Figura 4.3: Funções de sobrevivência estimadas para o Cenário 3 considerando uma amostra de tamanho n=500. As linhas sólidas representam as estimativas de Kaplan-Meier e as linhas tracejadas as estimativas pelo modelo proposto.

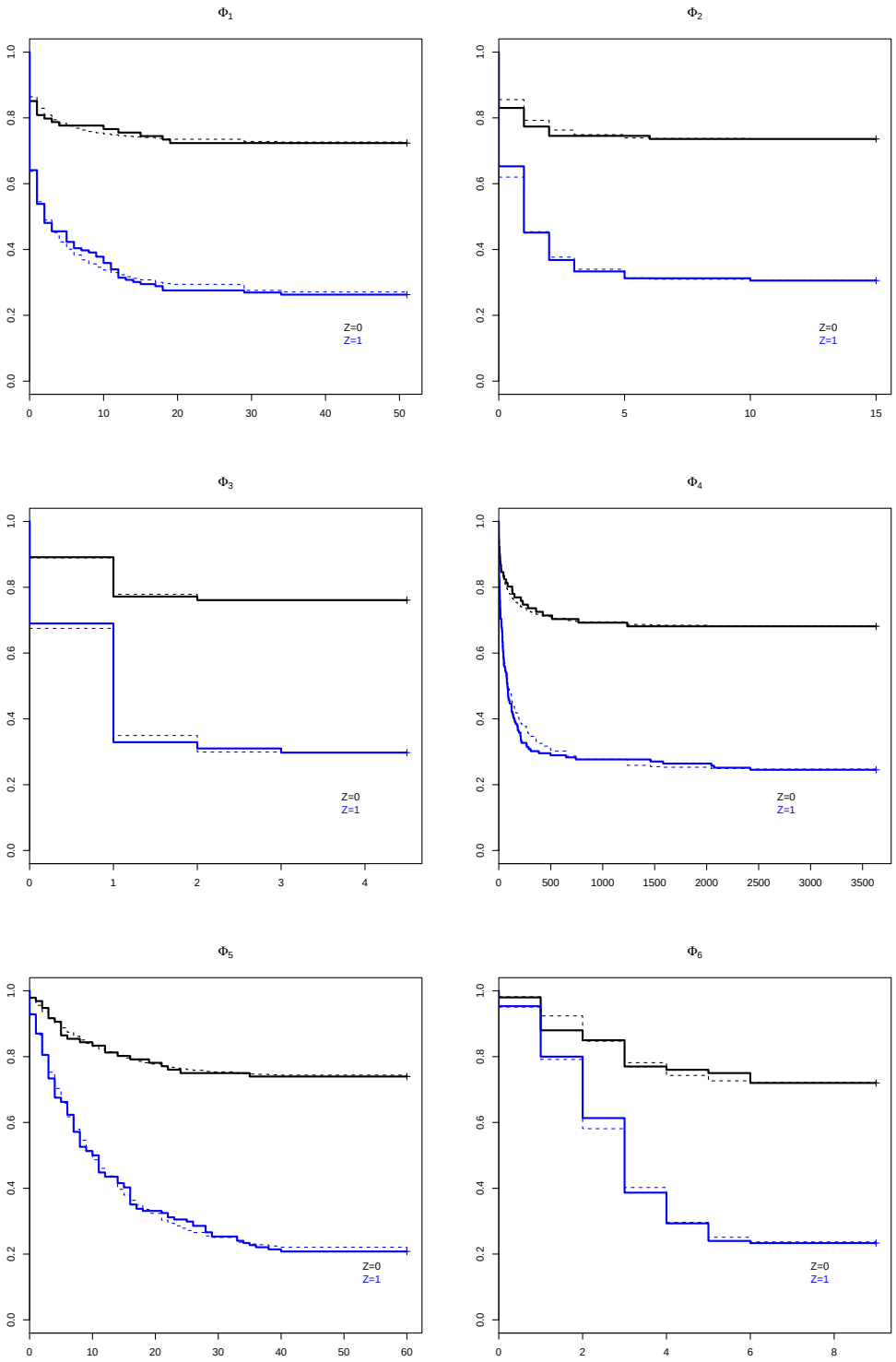

Figura 4.4: Funções de sobrevivência estimadas para o Cenário 4 considerando uma amostra de tamanho n=500. As linhas sólidas representam as estimativas de Kaplan-Meier e as linhas tracejadas as estimativas pelo modelo proposto.

Capítulo 5

Aplicações em dados reais

O objetivo deste capítulo é mostrar duas aplicações do modelo de regressão Weibull discreto com fração de cura em dados reais. A primeira aplicação é feita em um conjunto de dados de um estudo sobre o tempo até a morte de pacientes submetidos ao transplante de medula óssea (TMO) para tratamento de leucemia mielóide crônica (LMC). Já a segunda é sobre o tempo até o óbito de pacientes diagnosticados com Síndrome da Imunodeficiência Adquirida (AIDS).

5.1 Aplicação 1

Os dados da aplicação 1 são uma adaptação do estudo apresentado por Byington (1999). Trata-se de uma coorte de 96 pacientes portadores de LMC que foram submetidos ao TMO no Centro de Transplantes de Medula Óssea do Instituto Nacional do Câncer (CEMO/INCa) entre junho de 1986 e abril de 1998. O TMO é visto como o único tratamento que dá perspectivas de cura para pacientes com esta patologia.

O evento de interesse é o óbito do paciente, logo a variável resposta T é o tempo, em meses completos, desde o transplante até o óbito do paciente, censura ou o fim do estudo. Foi considerada uma covariável dicotômica (Z) no modelo que indica se houve a ocorrência de doença enxerto aguda contra o hospedeiro.

Durante todo o estudo houve um total de 47 censuras, o que corresponde a 48,9% dos pacientes. O maior tempo de óbito registrado foi igual a 22 meses e o menor foi 1 mês. No caso deste banco de dados o tempo máximo observado foi 33 meses, ou seja, entre 22 e 33 meses só houve a ocorrência de indivíduos censurados e estas censuras correspondem a 28% de todos os pacientes. Estas características nos dados dão indícios da existência de

uma parcela de pacientes curados na amostra em estudo. A Figura 5.1 mostra a estimativa de Kaplan-Meier da curva de sobrevivência e através dela é possível reafirmar a existência de uma fração de curados, visto que $S(t)$ permanece constante e diferente de zero pelo período de aproximadamente um ano.

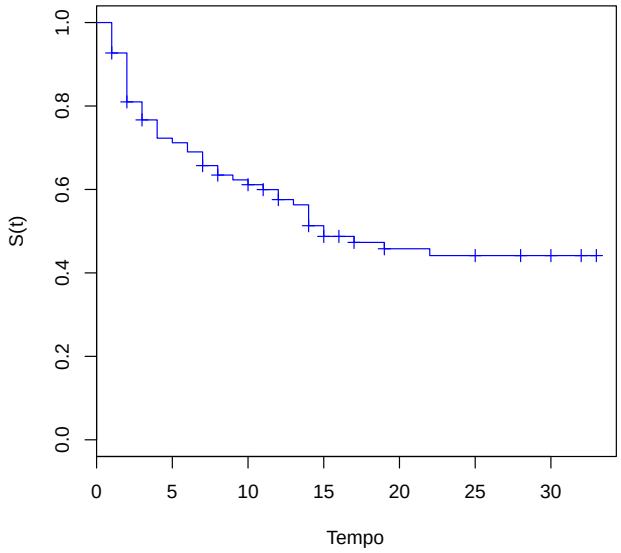

Figura 5.1: Função de sobrevivência estimada pelo método de Kaplan-Meier de pacientes com leucemia mielóide crônica submetidos ao transplante de medula óssea.

Sabe-se que 36,4% dos pacientes estudados tiveram doença enxerto aguda crônica contra o hospedeiro. A partir da inclusão desta covariável no modelo é possível verificar se a ocorrência de doença enxerto aguda influencia na probabilidade de cura ou não do paciente. Para tanto foi realizado o ajuste do modelo de regressão Weibull discreto com fração de cura e as estimativas pontuais e intervalares dos seus parâmetros são apresentados na Tabela 5.1. O nível de confiança considerado foi de 95%.

Tabela 5.1: Estimativas dos parâmetros do modelo de regressão Weibull discreto com fração de cura aplicado aos dados de transplante de medula óssea.

Parâmetros	**Estimativas**	**Intervalos de confiança**
$\mathbf{q}$	0,938	(0,894;0,982)
$\boldsymbol{\beta}$	1,274	(0,993;1,554)
$\boldsymbol{\phi_0}$	0,365	(-0,146;0,875)
$\boldsymbol{\phi_1}$	-2,114	(-3,177;-1,051)

Como pode ser observado nos resultados da Tabela 5.1, a covariável Z é estatisticamente significativa para modelar a fração de cura dos indivíduos em estudo, visto que o valor zero não está contido no intervalo de confiança do parâmetro ϕ_1. Ao calcular os valores de $\pi_i(\boldsymbol{\phi}, Z = 0)$ e $\pi_i(\boldsymbol{\phi}, Z = 1)$ temos que são iguais a 59% e 15%, respectivamente, ou seja, a probabilidade de cura dos indivíduos que não tiveram doença enxerto aguda é maior do que os que tiveram. Com base na *odds ratio* dada por OR=$\exp(-2,114) = 0,120$, vemos que os pacientes que tiveram doença enxerto aguda têm a chance de cura reduzida em quase 90%.

Ao analisar os resultados do parâmetro β, nota-se que a função de risco é crescente ($\beta > 1$) e que este parâmetro não é estatisticamente diferente de 1, o que sugere que os dados possam ser modelados considerando a distribuição geométrica. O alto valor da estimativa de $q = 0,938$ é justificável pela característica dos próprios dados, pois não existe nenhum tempo igual a zero, ou seja, nenhum paciente veio a óbito antes que completasse um mês após o transplante.

A Figura 5.2 ilustra os resultados do modelo por meio da estimativa da curva de sobrevivência. É possível perceber que o modelo proposto apresentou bom ajuste aos dados, visto que a curva de sobrevivência estimada por meio do modelo proposto ficou bem próxima à estimativa de Kaplan-Meier.

Foi realizado também o ajuste do modelo Weibull contínuo a estes dados discretos para ver se neste caso este modelo também se ajusta bem. Os resultados das estimativas pontuais e intervalares dos parâmetros modelo de regressão Weibull contínuo estão apresentados na Tabela 5.2. O nível de confiança considerado também foi de 95%.

Tabela 5.2: Estimativas dos parâmetros do modelo de regressão Weibull contínuo com fração de cura aplicado aos dados de transplante de medula óssea.

Parâmetros	**Estimativas**	**Intervalos de confiança**
$\boldsymbol{\alpha}$	8,007	(5,988;10,026)
$\boldsymbol{\beta}$	1,142	(-0,889;1,394)
$\boldsymbol{\phi_0}$	0,349	(-0,161;0,858)
$\boldsymbol{\phi_1}$	-2,175	(-3,260;-1,089)

A partir dos resultados da Tabela 5.2 observa-se que as estimativas dos parâmetros do modelo Weibull contínuo foram bem próximas àquelas obtidas no modelo Weibull discreto, inclusive do parâmetro q ($q = \exp\{-\frac{1}{\alpha^\beta}\} = 0,911$). O AIC do modelo Weibull discreto foi igual a -384,21, enquanto que o do modelo Weibull contínuo foi -379,07.

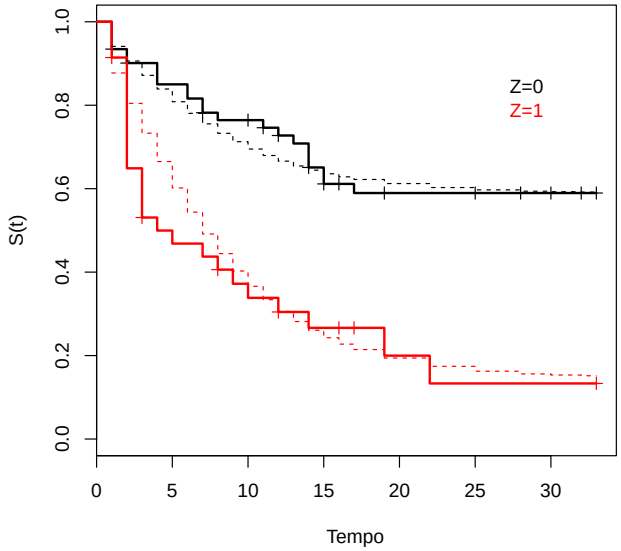

Figura 5.2: Funções de sobrevivência estimadas para pacientes que fizeram transplante de medula óssea. Se $Z = 1$ significa que o paciente teve doença enxerto aguda. A linha contínua representa a estimativa obtida via estimador de Kaplan-Meier e a linha tracejada foi estimada através do modelo de regressão Weibull discreto com fração de cura.

As Figuras 5.3 e 5.4 mostram as estimativas das curvas de sobrevivência e o gráfico P-P plot, respectivamente, comparando as funções de sobrevivência e acumulada obtidas a partir dos modelos Weibull discreto e contínuo. Nota-se que o desempenho dos dois modelos analisados é bem similar. De modo geral, tanto o modelo de regressão Weibull discreto quanto o modelo de regressão Weibull contínuo se ajustaram bem aos dados. Porém o AIC do modelo Weibull discreto foi menor, o que indica melhor ajuste deste modelo.

Nesta aplicação havia 24 tempos discretos distintos e isto pode ter contribuído para que os dados pudessem ser modelados por uma distribuição contínua sem prejudicar estimativas, pois ao passo em que o número de tempos distintos aumenta, é possível aproximar a distribuição dos dados discretos por uma distribuição contínua. Entretanto, como visto em Nakano e Carrasco (2006), em situações em que este número é pequeno pode haver perda da qualidade das estimativas ao utilizar uma distribuição contínua, pois a aproximação pode não ser possível. Além disso, a presença de uma única observação não censurada igual a zero no banco de dados prejudica a obtenção das estimativas de máxima veros-

similhança do modelo Weibull contínuo, visto que a função de verossimilhança será nula quando $\beta > 1$ e tenderá a ∞ quando $0 < \beta < 1$. A Aplicação 2 ilustra essa situação, pois contém tempos discretizados iguais a zero, e neste caso somente a distribuição discreta foi utilizada.

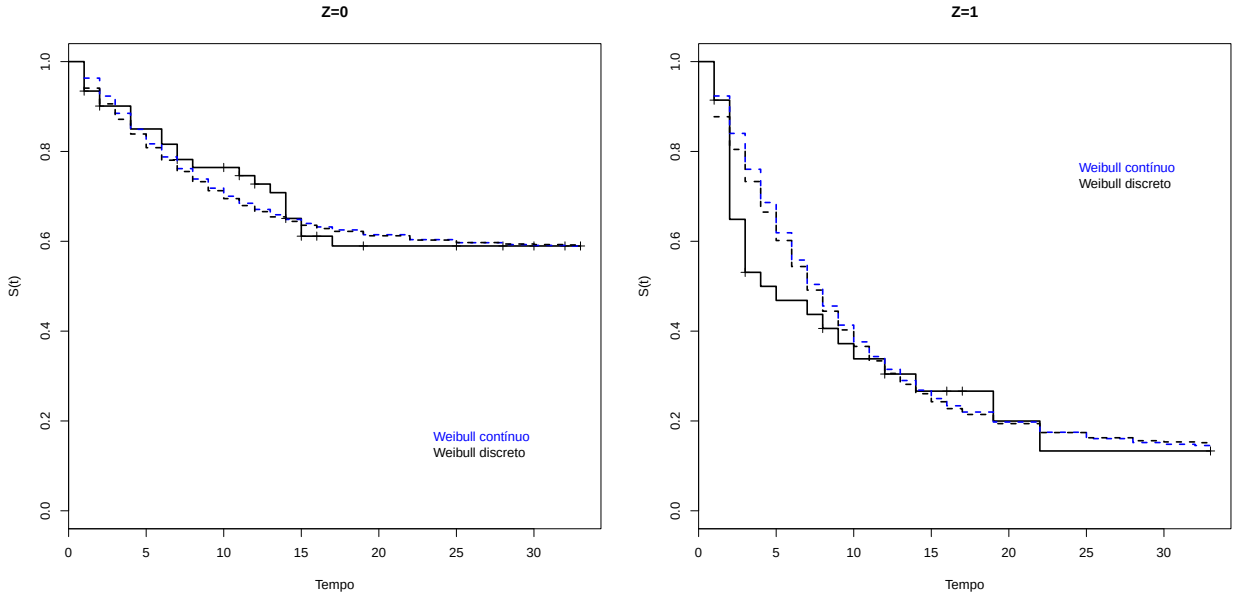

Figura 5.3: Funções de sobrevivência estimadas a partir dos modelos de regressão Weibull discreto e Weibull contínuo com fração de cura para pacientes que fizeram transplante de medula óssea. Se $Z = 1$ significa que o paciente teve doença enxerto aguda. A linha contínua representa a estimativa obtida via estimador de Kaplan-Meier e as linhas tracejadas foram estimadas através dos modelos de regressão Weibull discreto e contínuo com fração de cura.

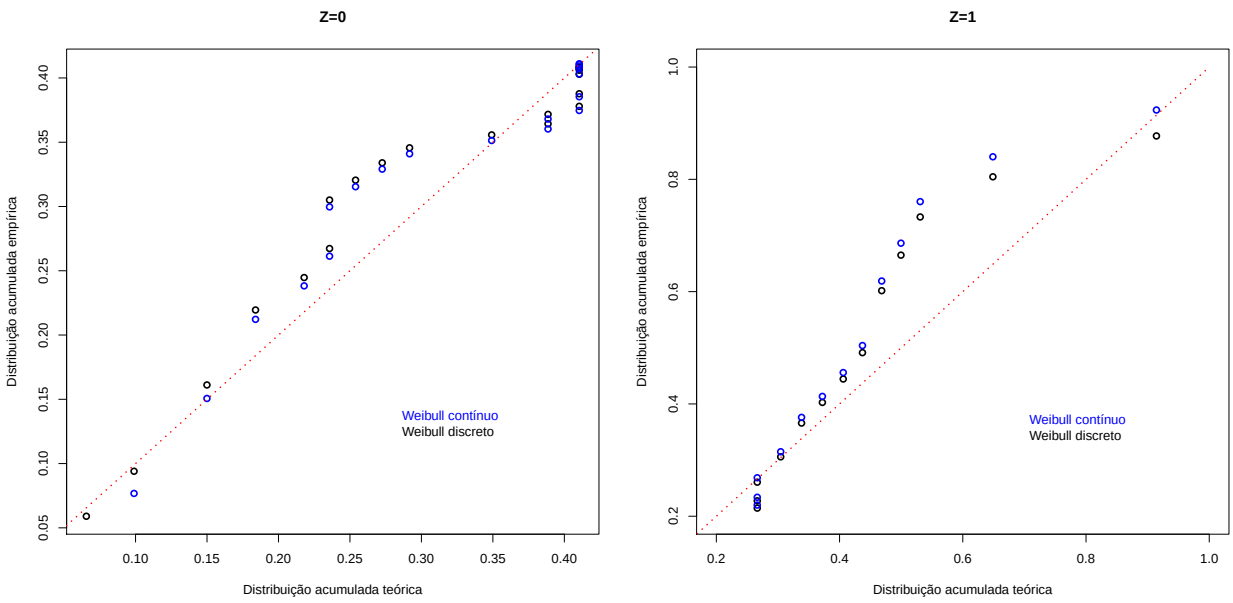

Figura 5.4: P-P plot dos modelos de regressão Weibull discreto e Weibull contínuo com fração de cura para pacientes que fizeram transplante de medula óssea. Se $Z = 1$ significa que o paciente teve doença enxerto aguda.

5.2 Aplicação 2

A segunda aplicação do modelo de regressão Weibull discreto foi feita em um conjunto de dados de um estudo sobre AIDS (Síndrome da Imunodeficiência Adquirida) realizado em São Francisco no estado da Califórnia (Selvin, 2008). Este estudo foi realizado com uma amostra aleatória 1.034 homens solteiros que tinham entre 25 e 54 anos. Os dados utilizados aqui são de um recorte de 174 homens que entraram no estudo durante o primeiro ano. O evento de interesse é o óbito do indivíduo, ou seja, a variável T representa o tempo, em meses completos, desde o diagnóstico até o óbito do indivíduo, censura ou o fim do estudo. No caso em que $t = 0$, significa que o óbito ocorreu antes de completar 1 mês desde o diagnóstico.

Dos 174 indivíduos, 19 foram censurados, o que corresponde a 10,9% de dos dados. Cerca de 65% dos homens morreram antes de completar 22 meses desde o diagnóstico. Além disto, há uma concentração de censuras ao final do estudo, com 9 censuras após 56 meses. A última morte ocorreu com 60 meses de acompanhamento e, desde então até completar os 107 meses de estudo, só houve censura. Estas informações nos fornecem indícios de que há um percentual de observações curadas nos dados, ou seja, homens que

não estão propensos a sofrer o evento de interesse. A Figura 5.5 ilustra a estimativa de Kaplan-Meier da curva de sobrevivência para estes dados e por ela é possível notar que S(t) permanece estável e diferente de zero por um período de quase quatro anos. Apesar do baixo percentual de censura, Fernandes(2013) mostrou que a fração de cura neste caso é significativa.

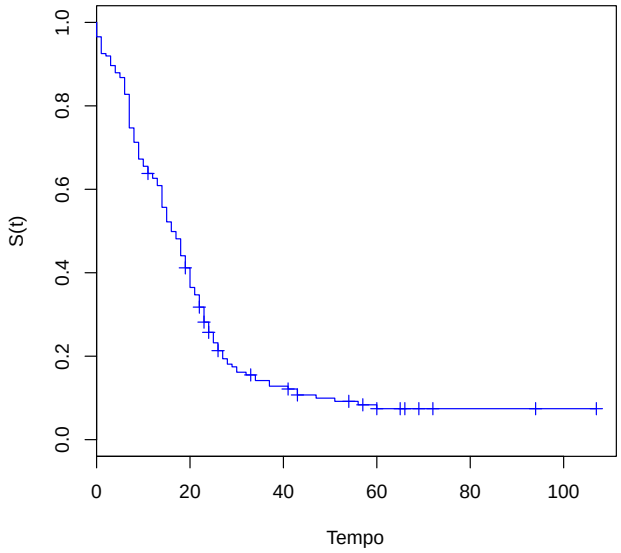

Figura 5.5: Função de sobrevivência estimada pelo método de Kaplan-Meier de homens com AIDS.

Uma informação que existe nos dados é um indicador que mostra se o indivíduo fuma ou não. A maior parte dos indivíduos estudados, 63,2%, não fuma. Esta informação pode ser incluída no modelo como uma covariável Z que será utilizada para modelar a probabilidade de cura. O interesse é analisar se o fato de fumar influencia na probabilidade de cura do indivíduo.

Neste estudo houve 6 homens que vieram a óbito antes de que completassem um mês desde o diagnóstico, o que fez com que existissem 6 tempos discretos de falha iguais a zero. Como citado anteriormente, isto prejudica a aplicação do modelo utilizando a distribuição Weibull contínua para modelar os tempos. Deste modo, foi aplicado o modelo de regressão Weibull discreto com fração de cura e os resultados das estimativas pontuais e intervalares estão presentes na Tabela 5.3. As curvas de sobrevivência estimadas são apresentadas na

Figura 5.6. O nível de confiança adotado foi de 95%.

Tabela 5.3: Estimativas dos parâmetros do modelo de regressão Weibull discreto com fração de cura aplicado aos dados de AIDS.

Parâmetros	**Estimativas**	**Intervalos de confiança**
$\boldsymbol{q}$	0,982	(0,973;0,992)
$\boldsymbol{\beta}$	1,369	(1,201;1,539)
$\boldsymbol{\phi_0}$	-2,863	(-3,946;-1,779)
$\boldsymbol{\phi_1}$	0,412	(-0,873;1,697)

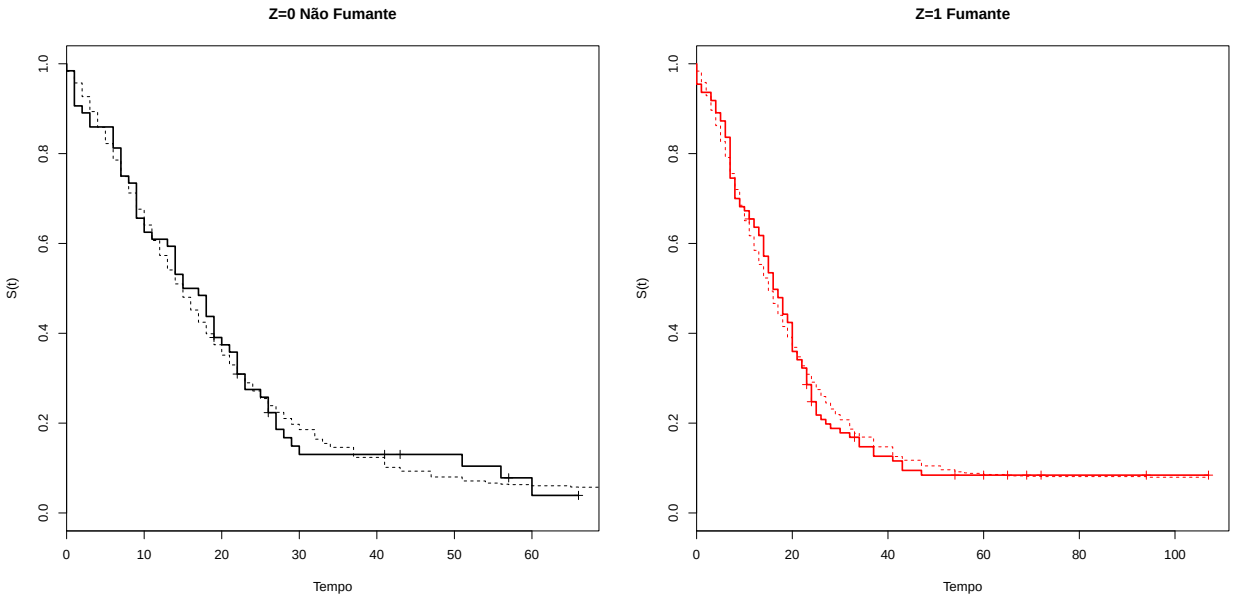

Figura 5.6: Funções de sobrevivência estimadas para os dados de homens com AIDS. A linha contínua representa a estimativa obtida via estimador de Kaplan-Meier e a linha tracejada foi estimada através do modelo de regressão Weibull discreto com fração de cura.

Com base nos resultados da Tabela 5.3, nota-se que a função de risco é crescente ($\beta > 1$) e que o parâmetro β é estatisticamente diferente de 1, o que não sugere o ajuste do modelo utilizando a distribuição geométrica. O percentual de tempos iguais zero nos dados é pequeno (3,4%) e há valores grandes de tempos, o que justifica o alto valor estimado para o parâmetro q. Além disto, é importante destacar que o parâmetro ϕ_1 é estatisticamente igual a zero, visto que este valor está contido no intervalo de confiança do respectivo parâmetro, o que nos leva a concluir que a covariável Z não é significativa para explicar a probabilidade de cura dos homens com AIDS estudados. Apesar disto,

nota-se pela Figura 5.6 que o modelo se ajustou bem aos dados. O AIC deste modelo foi igual a -1.245,35.

Foi realizado o ajuste do modelo Weibull discreto com fração de cura sem covariáveis. Os valores das estimativas dos parâmetros juntamente com os intervalos de confiança estão presentes na Tabela 5.4. A Figura 5.7 ilustra a estimativa da função de sobrevivência através do modelo. O nível de confiança adotado também foi de 95%.

Tabela 5.4: Estimativas dos parâmetros do modelo Weibull discreto com fração de cura aplicado aos dados de AIDS.

Parâmetros	**Estimativas**	**Intervalos de confiança**
$\boldsymbol{q}$	0,982	(0,973;0,992)
$\boldsymbol{\beta}$	1,373	(1,204;1,542)
$\boldsymbol{\phi_0}$	-2,564	(-3,141;-1,987)

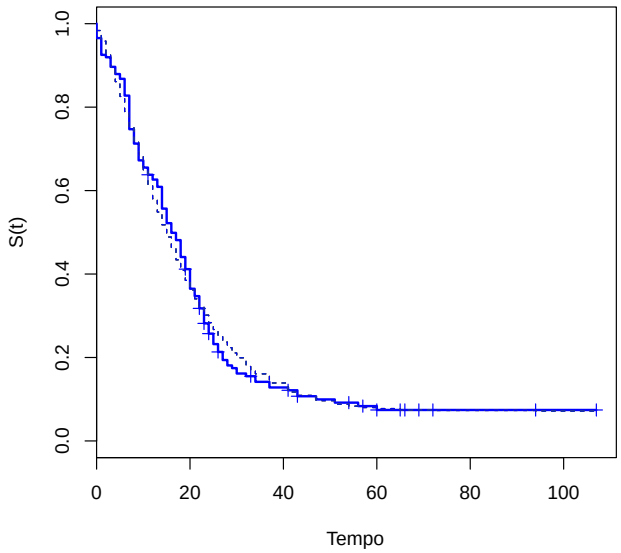

Figura 5.7: Funções de sobrevivência estimadas para os dados de homens com AIDS. A linha contínua representa a estimativa obtida via estimador de Kaplan-Meier e a linha tracejada foi estimada através do modelo Weibull discreto com fração de cura.

Os resultados da Tabela 5.4 mostram que as estimativas dos parâmetros não sofreram tanta influência após a retirada da covariável. Além disto, o AIC do modelo sem a covariável foi menor que o do modelo anterior (-1.247,64), o que indica que este se ajustou melhor aos dados. A partir de ϕ_0 é possível obter a probabilidade de cura, que é apro-

ximadamente 7%. A Figura 5.7 confirma o bom ajuste do modelo Weibull discreto com fração de cura sem considerar covariáveis.

Com esta aplicação foi possível ilustrar uma outra situação na qual o modelo proposto neste trabalho apresentou bons resultados. Este estudo em específico exemplifica um caso no qual não foi possível aplicar o modelo considerando a distribuição Weibull contínua devido às características dos dados, sendo, então, uma boa alternativa a modelagem dos dados a partir da distribuição Weibull discreta.

Capítulo 6

Considerações finais

Neste trabalho foi descrito o processo de inclusão da fração de cura e de covariáveis no modelo Weibull discreto no contexto de análise de sobrevivência. Como a probabilidade de cura de um indivíduo, π_i, pode variar de acordo com as suas características, o efeito das covariáveis foi inserido para modelar este parâmetro.

Numa população na qual existe uma parcela de indivíduos considerados curados, a informação do indicador de não cura c_i é parcialmente desconhecida, ou seja, estamos lidando com dados incompletos e, além disto, o modelo trabalhado é em forma de mistura. Sendo assim, descrevemos também o processo de estimação dos parâmetros do modelo de regressão Weibull discreto com fração de cura por meio do algoritmo EM.

No Capítulo 4, usamos simulação Monte Carlo para testar no software `R` o algoritmo descrito no Capítulo 3, definindo quatro cenários ao variar os valores dos parâmetros do modelo apresentado em (3.1), (3.2) e (3.3), com amostras de tamanhos 250, 500 e 1.000. De modo geral, todos os cenários apresentaram bons resultados. As características dos dados tais como tamanho de amostra, percentual de tempos iguais a zero e quantidade de tempos distintos influenciaram nas estimativas dos parâmetros. Entretanto este impacto foi pequeno, visto que em todos os casos analisados nos quatro cenários as médias das estimativas estavam muito próximas dos verdadeiros valores dos parâmetros e os valores de EQM eram bem pequenos. É importante ressaltar que o modelo apresentado deve ser usado apenas em dados em que há indícios de fração de curados, inclusive quando as observações são agrupadas de acordo com os valores da covariável. Recomenda-se a estimação das funções de sobrevivência via Kaplan-Meier antes do ajuste do modelo a fim de verificar se todas as sobrevivências estimadas são diferentes de zero no último tempo observado. Caso isto não aconteça em pelo menos uma delas, as estimativas do modelo

poderão ser prejudicadas.

As aplicações apresentadas no Capítulo 5 permitiram ilustrar o uso do modelo de regressão Weibull discreto com fração de cura em dados reais, ambos os casos considerando que o tempo está sendo medido em meses completos. A primeira aplicação mostrou um caso em que a covariável Z foi significativa para explicar a probabilidade de cura dos indivíduos submetidos ao transplante de medula óssea e, com isto, foi possível estimar a fração de cura de cada grupo. A probabilidade de cura dos pacientes que não tiveram doença enxerto aguda é maior do que os que tiveram (59% e 15%, respectivamente). A *odds ratio* mostra que os pacientes que tiveram doença enxerto aguda têm chance de cura reduzida em quase 90%. O modelo de regressão Weibull contínuo com fração de cura também foi aplicado e as estimativas dos parâmetros foram muito próximas às do modelo discreto. Entretanto, o valor do AIC mostrou que o modelo utilizando a distribuição discreta apresentou melhor ajuste. O fato de neste conjunto de dados haver uma quantidade razoável de número de tempos distintos justifica o bom desempenho do modelo contínuo no ajuste de dados discretos, visto que quando este número é grande é possível aproximar para uma distribuição contínua. Entretanto, nem sempre isto é possível. Nakano e Carrasco (2006) mostraram que em situações em que este número é pequeno (há muitos empates nos valores dos tempos) pode haver perda da qualidade das estimativas ao utilizar uma distribuição contínua.

A segunda aplicação mostrou uma situação na qual só foi aplicado o modelo discreto. Quando nos dados há pelo menos um tempo não censurado igual a zero há prejuízo na obtenção das estimativas de máxima verossimilhança do modelo Weibull contínuo, visto que a função de verossimilhança será nula quando $\beta > 1$ e tenderá a ∞ quando $0 < \beta < 1$. Como nos dados de AIDS havia indivíduos que vieram a óbito antes de completar um mês desde o diagnóstico e o tempo discreto registrado para eles é zero, foi aplicado somente o modelo de regressão utilizando a distribuição discreta. A covariável Z não foi significativa para modelar a fração de cura dos homens com AIDS e o modelo sem covariável apresentou menor AIC quando comparado com o modelo com covariável, logo o modelo final considerado foi o modelo Weibull discreto com fração de cura. A probabilidade de cura dos homens com AIDS foi igual a 7%.

Diante do que foi exposto, nota-se que o modelo proposto é bastante eficaz quando utilizado em dados com tempos discretos e com fração de curados. Algumas sugestões para

trabalhos futuros são: a inclusão de covariáveis para modelar também o parâmetro q, visto que as características dos indivíduos podem interferir também nos tempos de sobrevida. Neste caso, no processo de estimação serão obtidas as estimativas do parâmetro de forma β e de dois vetores de parâmetros, sendo um associado à π e outro associado ao parâmetro q da distribuição dos tempos de sobrevivência; o estudo de métodos de avaliação do ajuste do modelo de regressão Weibull discreto quando a covariável é contínua, pois neste caso não é possível a estimação das curvas de sobrevivência para que seja possível a comparação entres as estimativas de Kaplan-Meier e as estimadas pelo modelo; o estudo de um modelo mais robusto que permita o uso em dados nos quais pelo menos uma curva de sobrevivência assume valor zero; e, por fim, o uso da metodologia aqui descrita considerando outras distribuições discretas para modelar os dados.

Apêndice A

Dados utilizados nas aplicações do Capítulo 5

Tabela A.1: Dados utilizados na Aplicação 1 do estudo de 96 pacientes portadores de leucemia que foram submetidos ao transplante de medula óssea. Censura=0 (o tempo é censurado); Censura=1 (o tempo é de falha). Doença enxerto aguda=0 (Não); Doença enxerto aguda=1 (Sim). Fonte: Byington (1999).

Indivíduo	**Tempo (meses)**	**Censura**	**Doença enxerto aguda**
1	33	0	0
2	1	1	1
3	14	1	1
4	2	1	1
5	22	1	1
6	3	1	1
7	33	0	0
8	13	1	0
9	8	1	1
10	2	1	0
11	33	0	1
12	33	0	0
13	33	0	1
14	11	1	0
15	15	1	0
16	9	1	1
17	33	0	0
18	6	1	0
19	33	0	0
20	2	1	1

21	33	0	0
22	33	0	0
23	15	1	0
24	33	0	0
25	3	1	1
26	10	1	1
27	33	0	0
28	4	1	0
29	14	1	0
30	2	1	1
31	7	1	0
32	6	1	0
33	4	1	0
34	2	1	1
35	33	0	0
36	4	1	1
37	33	0	0
38	2	1	1
39	33	0	0
40	33	0	0
41	1	1	0
42	3	1	1
43	5	1	1
44	7	1	0
45	2	1	1
46	33	0	0
47	33	0	0
48	33	0	0
49	33	0	0
50	33	0	0
51	19	0	0
52	2	1	1
53	33	0	0
54	1	1	1
55	30	0	0
56	7	1	1
57	1	1	0
58	28	0	0
59	4	1	0
60	32	0	0
61	10	0	0
62	2	1	1
63	1	1	1

64	25	0	0
65	28	0	0
66	25	0	0
67	17	1	0
68	1	1	0
69	19	0	0
70	2	1	1
71	15	0	0
72	14	1	0
73	15	0	0
74	16	0	0
75	19	1	1
76	17	0	1
77	2	0	0
78	14	1	0
79	17	0	1
80	12	1	1
81	16	0	1
82	14	0	0
83	8	0	1
84	12	0	0
85	12	1	0
86	8	1	0
87	12	0	1
88	2	1	0
89	7	0	0
90	7	0	0
91	3	1	1
92	3	0	1
93	1	0	1
94	1	0	0
95	1	1	0
96	11	0	0

Tabela A.2: Dados utilizados na Aplicação 2 do estudo de 174 homens portadores da AIDS. Censura=0 (o tempo é censurado); Censura=1 (o tempo é de falha). Fumante=0 (Não); Fumante=1(Sim).
Fonte: Selvin (2008).

Indivíduo	**Tempo (meses)**	**Censura**	**Fumante**
1	0	1	1
2	0	1	1
3	0	1	0
4	0	1	1
5	0	1	1
6	0	1	1
7	1	1	0
8	1	1	0
9	1	1	0
10	1	1	1
11	1	1	0
12	1	1	0
13	1	1	1
14	2	1	0
15	3	1	0
16	3	1	1
17	3	1	0
18	3	1	1
19	4	1	1
20	4	1	1
21	4	1	1
22	5	1	1
23	5	1	1
24	6	1	1
25	6	1	0
26	6	1	1
27	6	1	1
28	6	1	0
29	6	1	1
30	6	1	0
31	7	1	1
32	7	1	0
33	7	1	0
34	7	1	1
35	7	1	1
36	7	1	0
37	7	1	1

38	7	1	0
39	7	1	1
40	7	1	1
41	7	1	1
42	7	1	1
43	7	1	1
44	7	1	1
45	8	1	1
46	8	1	1
47	8	1	0
48	8	1	1
49	8	1	1
50	8	1	1
51	9	1	0
52	9	1	0
53	9	1	0
54	9	1	1
55	9	1	0
56	9	1	0
57	9	1	1
58	10	1	0
59	10	1	0
60	10	1	1
61	11	1	0
62	11	1	1
63	11	1	1
64	11	0	1
65	12	1	1
66	12	1	1
67	13	1	1
68	13	1	1
69	13	1	0
70	14	1	1
71	14	1	1
72	14	1	0
73	14	1	1
74	14	1	0
75	14	1	1
76	14	1	0
77	14	1	0
78	14	1	1
79	15	1	1
80	15	1	0

81	15	1	1
82	15	1	0
83	15	1	1
84	15	1	1
85	16	1	1
86	16	1	1
87	16	1	1
88	16	1	1
89	17	1	1
90	17	1	1
91	17	1	0
92	18	1	1
93	18	1	1
94	18	1	1
95	18	1	0
96	18	1	0
97	18	1	0
98	18	1	1
99	19	1	0
100	19	1	0
101	19	1	1
102	19	1	0
103	19	0	0
104	19	1	1
105	20	1	0
106	20	1	1
107	20	1	1
108	20	1	1
109	20	1	1
110	20	1	1
111	20	1	1
112	20	1	1
113	21	1	0
114	21	1	1
115	21	1	1
116	22	1	0
117	22	1	1
118	22	1	0
119	22	1	1
120	22	0	0
121	22	1	0
122	23	1	1
123	23	0	1

124	23	1	0
125	23	1	1
126	23	1	0
127	23	1	1
128	23	1	1
129	24	1	1
130	24	1	1
131	24	0	1
132	24	1	1
133	24	1	1
134	25	1	1
135	25	1	0
136	25	1	1
137	25	1	1
138	26	1	1
139	26	1	0
140	26	0	0
141	26	1	0
142	27	1	1
143	27	1	0
144	27	1	0
145	28	1	1
146	28	1	0
147	29	1	0
148	30	1	1
149	30	1	0
150	32	1	1
151	33	0	1
152	34	1	1
153	34	1	1
154	37	1	1
155	37	1	1
156	41	1	1
157	41	0	0
158	43	1	1
159	43	1	1
160	43	0	0
161	47	1	1
162	51	1	0
163	54	0	1
164	56	1	0
165	57	0	0
166	60	0	1

167	60	0	1
168	60	1	0
169	65	0	1
170	66	0	0
171	69	0	1
172	72	0	1
173	94	0	1
174	107	0	1

Apêndice B

Scripts desenvolvidos

O código a seguir foi utilizado no software R na versão 3.2.2 com o objetivo de gerar dados de sobrevivência com fração de cura e a presença de covariável e obter as estimativas dos parâmetros do modelo de regressão Weibull discreto com fração de cura via algoritmo EM.

```
### Pacotes
library(survival)
require(DiscreteWeibull)

set.seed(1)

######## DISTRIBUIÇÃO DE PROBABILIDADES DA WEIBULL DISCRETA
###   q --- parâmetro do modelo
###   b --- parâmetro do modelo
weidc <- function(t,q,b){ ((q)^(t^b)) - (q)^((t+1)^b) }


######## FUNÇÃO DE SOBREVIVÊNCIA DA WEIBULL DISCRETA
###   q --- parâmetro do modelo
###   b --- parâmetro do modelo
sob.weidc <- function(t,q,b){ (q)^((t+1)^b) }
```

```
####### FUNÇÃO PARA GERAR VALORES DA WEIBULL DISCRETA
###   n --- tamanho da amostra
###   q --- parâmetro do modelo
###   b --- parâmetro do modelo
amostra.weidc <- function(n,q,b){ rdweibull(n, q, b, zero = TRUE) }


######## DISTRIBUIÇÃO DE PROBABILIDADES DA WEIBULL DISCRETA COM FRAÇÃO DE CURADOS
###   q --- parâmetro do modelo
###   b --- parâmetro do modelo
###   f --- parâmetro que modela a fração de curados
weidc.fc <- function(t,q,b,f){ (1-f)*(((q)^(t^b)) - (q)^((t+1)^b)) }


######## FUNÇÃO DE SOBREVIVÊNCIA DA WEIBULL DISCRETA COM FRAÇÃO DE CURADOS
###   q --- parâmetro do modelo
###   b --- parâmetro do modelo
###   f --- parâmetro que modela a fração de curados
sob.weidc.fc <- function(t,q,b,f){ f + (1-f)*( (q)^((t+1)^b) )}


#------------------------ INÍCIO DA SIMULAÇÃO ------------------------#

estimacao.EM <- function(n, N.Sim, p.cens, fi0, fi1, q, b,
I.fi0, I.fi1, I.q, I.b){

# n: tamanho da amostra
# N.Sim: Número de simulações
# p.cens: proporção de censura (dos não curados)

N.Par <- 4 # Número de parâmetros
```

```
estimativas <- matrix(NA, nrow = N.Sim, ncol = N.Par, dimnames = list(c(),
                                c("fi0.EM", "fi1.EM", "q.EM", "b.EM")))
Tempos <- matrix(NA, nrow = N.Sim, ncol = n)
Censuras <- matrix(NA, nrow = N.Sim, ncol = n)
X <- matrix(NA, nrow = N.Sim, ncol = n)


   for(j in 1:N.Sim){

######## GERAR VALORES DA WEIBULL DISCRETA COM FRAÇÃO DE CURADOS  ##########

# fi0: parâmetro quando x=0
# fi1: parâmetro quando x=1
# q:   parâmetro do modelo
# b:   parâmetro do modelo

x <- rbinom(n,1,.6) # gerando valores da covariável
X[j,] <- x # vetor com valores da covariável
tempo <- numeric(n) # vetor de tempos
censura <- f <- numeric(n) # vetores de censura e fração de cura

for (i in 1:n){
   f[i] <- exp(fi0 + x[i]*fi1)/(1+exp(fi0 + x[i]*fi1)) # fração de cura
   ind.suscept <- rbinom(1,1,1-f[i]) # O indivíduo é suscetível? P(c=1)=1-f
   if (ind.suscept==1){   # Se o indivíduo está sob risco
      tempo[i] <- amostra.weidc(1,q,b) # Recebe tempo que segue distribuição WD
      censura[i] <- rbinom(1,1,1-p.cens) # P(censura)=p.cens
   }
   if (ind.suscept==0){  # Se o indivíduo é curado
      tempo[i] <- -1 # Recebe tempo negativo
      censura[i] <- 0 }} # É censurado
```

```
tempo.final <- trunc(1.5*max(tempo)) # Tempo máximo de estudo é grande para
                                          # ilustrar a FC

# Os tempos negativos são de curados, logo o maior tempo observado
  # é o último tempo do estudo
for (i0 in 1:n){   if (tempo[i0]==-1) tempo[i0]<- tempo.final  }

Tempos[j,] <- tempo
Censuras[j,] <- censura

################  Estimação dos parâmetos via EM  ########################

### PASSO 1 - inicializar o contador das iterações e os parâmetros iniciais

k <- 1 # Contador

fi0.0 <- I.fi0 # Valor inicial do parâmetro para x=0
fi1.0 <- I.fi1 # Valor inicial do parâmetro para x=1
q.0  <- I.q # Valor inicial do parâmetro do modelo
b.0  <- I.b # Valor inicial do parâmetro do modelo

fi0.EM <- fi1.EM <- q.EM <- b.EM <- numeric()
E <- 1e-100 # Erro
Av <- E+1   # Av:Será o vetor de diferenças entre parâmetros e estimativas
antes <- novo <- NULL
Vfi0 <- Vfi1 <- Vq <- Vb <- NULL; # Vetores com as estimativas

# Iniciando as iterações do EM
while(Av>E){ # Regra de parada

   ### PASSO 2 (Passo E) - Obter w1 e w2 conforme (3.12)
   pi <- exp(fi0.0 + x*fi1.0)/(1+exp(fi0.0 + x*fi1.0))
```

```
w1 <- pi / ( pi + (1-pi)*sob.weidc(tempo,q.0,b.0) )
w2 <- (1-w1)

### PASSO 3 (Passo M) - Maximizar g1 e g2
Vfi0[k] <- fi0.0; Vfi1[k] <- fi1.0; Vq[k] <- q.0; Vb[k] <- b.0

   # Encontrando fi0 e fi1 a partir de g1
   g1 <- function(par1,tempo,censura,x,w1,w2){   # Função g1
    fi0 <- par1[1]
     fi1 <- par1[2]
    pi <- exp(fi0 + x*fi1)/(1+exp(fi0 + x*fi1))
    -1*( sum(censura*log(1-pi)) + sum((1-censura)*w1*log(pi)) +
    sum((1-censura)*w2*log(1-pi)) )        }

   chute.fi<-c(fi0.0,fi1.0)   # Valor inicial dos parâmetros

   max.g1<-optim(chute.fi,g1,tempo=tempo,censura=censura,x=x,w1=w1,w2=w2)
   fi0.EM <-max.g1$par[1] # Estimativas dos parametros fi0 e fi1 no passo k
   fi1.EM <-max.g1$par[2]



  # Encontrando q e b a partir de g2
  g2 <- function(par2,tempo,censura,x,w1,w2){ # Função g2
     q <- par2[1]
     b <- par2[2]
     if ( (q>0) && (q<1) && (b>0)  )
     return ( -1*(sum(censura*log(weidc(tempo,q,b))) +
     sum((1-censura)*w2*log(sob.weidc(tempo,q,b))) )     )
     else return (-Inf)  }

  chute.qb<-c(q.0,b.0) # Valor inicial dos parâmetros
```

```
     max.g2<-optim(chute.qb,g2,tempo=tempo,censura=censura,x=x,w1=w1,w2=w2)
     q.EM <-max.g2$par[1] # Estimativas dos parametros q e b no passo k
     b.EM <-max.g2$par[2]

   # Atualizando os parâmetros
    fi0.0 <-fi0.EM; fi1.0 <-fi1.EM; q.0 <-q.EM; b.0 <-b.EM
    antes <- c(Vfi0[k], Vfi1[k], Vq[k], Vb[k])
    novo <- c(fi0.0, fi1.0, q.0, b.0)
    Av <- max(abs(antes-novo))
    k <- k+1
} # Fecha o while do EM

estimativas[j,] <- c(fi0.EM, fi1.EM, q.EM, b.EM)

} # Fecha o for da simulação

# Avaliação da estimação
est.fi0 <- mean(estimativas[,1])  #Média do vetor fi.0
est.fi1 <- mean(estimativas[,2])  #Média do vetor fi.1
est.q <- mean(estimativas[,3])  #Média do vetor q
est.b <- mean(estimativas[,4])  #Média do vetor b

vies.fi0 <- est.fi0-fi0  #Viés do estimador fi.0
vies.fi1 <- est.fi1-fi1  #Viés do estimador fi.1
vies.q <- est.q-q  #Viés do estimador q
vies.b <- est.b-b  #Viés do estimador b

EQM.fi0 <- sum((estimativas[,1] - fi0) ^2) / N.Sim  #EQM de fi.0
EQM.fi1 <- sum((estimativas[,2] - fi1) ^2) / N.Sim  #EQM de fi.1
EQM.q <- sum((estimativas[,3] - q) ^2) / N.Sim  #EQM de q
EQM.b <- sum((estimativas[,4] - b) ^2) / N.Sim  #EQM de b
```

```
##### GRÁFICO
data <- Surv(tempo,censura)
km <- survfit(data~x)  # Estimador K-M
grafico <- plot(km,conf.int=F,col=c(1,2),xlab="Tempo",ylab="S(t)")

f.est.0<-exp(fi0.EM)/(1+exp(fi0.EM))
f.est.1<-exp(fi0.EM + fi1.EM)/(1+exp(fi0.EM + fi1.EM))
xx<-sort(tempo)
sob.est.0<-sob.weidc.fc(xx, q.EM, b.EM, f.est.0)
sob.est.1<-sob.weidc.fc(xx, q.EM, b.EM, f.est.1)

points(xx,sob.est.0,type="s",col=1,lty=2)
points(xx,sob.est.1,type="s",col=2,lty=2)
legend(5,0.2,c("x=0","x=1"),bty="n",text.col=c(1,2))

Resultado <- list(Estimativa_fi0=est.fi0, Vies_fi0=vies.fi0, EQM.fi0=EQM.fi0,
Estimativa_fi1=est.fi1, Vies_fi1=vies.fi1, EQM.fi1=EQM.fi1,
Estimativa_q=est.q, Vies_q=vies.q, EQM.q=EQM.q,
Estimativa_b=est.b, Vies_b=vies.b, EQM.b=EQM.b,
Estimativas=estimativas,
Tempos=Tempos, Censuras=Censuras, X=X, grafico)
return(Resultado)
} # Fecha a função da simulação

### INTERVALOS DE CONFIANÇA
var.fi <- solve(max.g1$hessian)

LI.fi0 <- fi0.EM - 1.96*sqrt(var.fi[1,1]);LS.fi0 <- fi0.EM + 1.96*sqrt(var.fi[1,1])
LI.fi0;LS.fi0

LI.fi1 <- fi1.EM - 1.96*sqrt(var.fi[2,2]);LS.fi1 <- fi1.EM + 1.96*sqrt(var.fi[2,2])
LI.fi1;LS.fi1
```

```
var.qb <- solve(max.g2$hessian)

LI.q <- q.EM - 1.96*sqrt(var.qb[1,1]);LS.q <- q.EM + 1.96*sqrt(var.qb[1,1])
LI.q;LS.q

LI.b <- b.EM - 1.96*sqrt(var.qb[2,2]);LS.b <- b.EM + 1.96*sqrt(var.qb[2,2])
LI.b;LS.b
```

Referências Bibliográficas

[1] ALJAWADI, B. A. I. et al. Parametric Estimation of the Cure Fraction Based on BCH Model Using Left-Censored Data with Covariates. Modern Applied Science, Vol.5, n.3, p.103-110, Jun.2011.

[2] BARBIERO, A. DiscreteWeibull: Discrete Weibull Distributions (Type 1 and 3). R package version 1.0.1, 2015. http://CRAN.R-project.org/package=DiscreteWeibull.

[3] BASTOS, J.; ROCHA, C. Análise de sobrevivência: Conceitos Básicos. Arquivos de Medicina, Porto, Vol.20, n.5-6, set.2006. Disponível em <http://www.scielo.mec.pt/scielo.php?script=sci_arttext&pid=S0871-34132006000400007&lng=pt&nrm=iso>. Acesso em 18 fev. 2015.

[4] ------. Análise de Sobrevivência Métodos Não Paramétricos. Arquivos de Medicina, Porto, Vol.21, n.3-4, 2007. Disponível em <http://www.scielo.mec.pt/scielo.php?script=sci_arttext&pid=S0871-34132007000300007&lng=pt&nrm=iso>. Acesso em 18 fev. 2015.

[5] BERKSON, J.; GAGE, R. P. Survival Curve for Cancer Patients Following Treatment. Journal of the American Statistical Association, Vol. 47, n.259, p. 501-515, Set.1952.

[6] BOYLES, R. A. On the Convergence of the EM Algorithm. J. Roy. Statist. Soc. Ser. Vol.45. p.47-50, 1983.

[7] BRUNELLO, G. H. V.; NAKANO, E. Y. Inferência bayesiana no modelo Weibull discreto em dados com presença de censuras. TEMA ? Tend. Mat. Apl. Comput., Vol.16, n.2, p. 97-110, 2015.

[8] BYINGTON, M. R. L. Estudo de fatores prognósticos em pacientes submetidos ao transplante de medula óssea para tratamento de leucemia mielóide crônica. Dissertação (Mestrado em Saúde coletiva) - Instituto de medicina social, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 1999.

[9] CARVALHO, M. S. et al. Análise de sobrevivência: Teoria e aplicações em saúde. Rio de Janeiro: Editora Fiocruz, 2011. 434p.

[10] CARVALHO, M. T. Análise de Sobrevivência Aplicada ao Risco de Crédito: Ajuste de Modelos Paramétricos Contínuos a Dados de Tempo Discretos. 2011. 39f. Trabalho de Conclusão de Curso (Graduação em Estatística) - Departamento de Estatística, Universidade de Brasília, Brasília, 2011.

[11] CASELLA, G.; BERGER, R. L. Inferência estatística. São Paulo: Cengage Learning, 2010. 588p.

[12] COLOSIMO, E. A.; GIOLO, S. R. Análise de sobrevivência aplicada. São Paulo: Edgard Blucher Ltda, 2006. 369p.

[13] DEMPSTER, A. P. et al. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Vol.39, n.1, p.1-38, 1977.

[14] FERNANDES, L. M. Inferência bayesiana em modelos discretos com fração de cura. 2013. 62f. Dissertação (Mestrado em Estatística) - Departamento de Estatística, Universidade de Brasília, Brasília, 2013.

[15] GRANZOTTO, D. C. T. et al. Modelo Weibull e log-logístico com longa-duração: uma aplicação a dados reais. 19º SINAPE, São Pedro, 2010.

[16] HOSMER, D. W. et al. Applied survival analysis. 2ed. Wiley-Interscience, 2008.

[17] KANNAN, N. et al. The generalized exponential cure rate model with covariates. Journal of Applied Statistics, Vol. 37, n.10, p. 1625-1636, Set.2010.

[18] LIMA JUNIOR, P.; SILVEIRA, F. L. da; OSTERMANN, F. Análise de sobrevivência aplicada ao estudo do fluxo escolar nos cursos de graduação em física: um exemplo de uma universidade brasileira. Revista Brasileira de Ensino de Física, São Paulo, v.34,

n.1, Mar.2012. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1806-11172012000100014&lng=en&nrm=iso>. Acesso em 19 fev. 2015.

[19] LOUZADA NETO, F. et al. Introdução à análise de sobrevivência e confiabilidade. III Jornada Regional de Estatística e II Semana da Estatística, Maringá, 2002.

[20] MALLER, R.; ZHOU, X. Survival Analysis with Long-Term Survivors. New York: Wiley, 1996. 278p.

[21] NAKAGAWA, T.; OSAKI, S. The discrete weibull distribution. IEEE Transactions on Reliability, Vol.R-24, n.5, p.300-301, Dez.1975.

[22] NAKANO, E. Y.; CARRASCO, C. G. Uma avaliação do Uso de um Modelo Contínuo na Análise de Dados Discretos de Sobrevivência. TEMA - Tend. Mat. Apl. Comput., Vol.7, n.1, p.91-100, 2006.

[23] OLIVEIRA, C. Z. et al. Determinação dos fatores associados à sobrevida de mulheres com câncer de mama via modelos de longa duração Weibull Modificado. 19º SINAPE, São Pedro, 2010.

[24] PAES, A. T. Uso de modelos com fração de cura na análise de dados de sobrevivência com omissão nas covariáveis. 2007. 121f. Tese (Doutorado em Ciências - Estatística) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2007.

[25] PENG, Y. Fitting semiparametric cure models. Computational Statistics and Data Analysis, Vol.41, p.481-490, 2003.

[26] R Core Team. R: A language and environment for statistical computing. Fonte: R Foundation for Statistical Computing, Vienna, Austria, 2015: http://www.R-project.org/.

[27] SELVIN, S. Survival analysis for epidemiologic and medical research: A practical guide. New York: Cambridge University Press, 2008.

[28] WU, C. F. J. On the convergence properties of the EM algorithm. The Annals of Statistics, Vol.11, n.1, p.95-103, Mar.1983.