Journal homepage http://revistas.unitru.edu.pe/index.php/SSMM

SELECCIONES MATEMÁTICAS
Universidad Nacional de Trujillo
ISSN: 2411-1783 (Online)
Vol. 06(01): 84 - 97 (2019)

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

Discrete Weibull regression model with cure fraction in survival data.*

Carolina A. Silva †, Cira E. G. Otiniano ‡, and Eduardo Y. Nakano §

Received, Feb. 20, 2019 *Accepted, Jun. 21, 2019*

DOI: http://dx.doi.org/10.17268/sel.mat.2019.01.11

Resumo

Este trabalho apresenta a formulação de um modelo de regressão para dados de tempo discretos com fração de cura. Para tanto, foi considerado um 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. Por se tratar de um modelo de mistura a estimação dos parâmetros do modelo é feita via o algoritmo EM. O comportamento do algoritmo de estimação é testado com vários experimentos de simulação Monte Carlo e uma aplicação em dados reais foi adicionada.

Palavras chave. Modelo de mistura; regressão; distribuição Weibull discreta; algoritmo EM.

Abstract

This paper presents a regression model for discrete time data with cure fraction. For this purpose, we considered a mixture model, in which times are modeled through the discrete Weibull distribution and the cure probability modeled with covariates by using the logit link function. Considering that the model is a mixture, the estimation of the parameters was performed by EM algorithm. The behavior of the estimating algorithm was tested with Monte Carlo simulation experiments and an application for real data was added.

Keywords. Mixture model; regression; Discrete Weibull distribution; EM algorithm.

1. Introdução. Em análise de sobrevivência o tempo de ocorrência de um evento de interesse (falha ou morte) é uma variável aleatória T. Em dados de pacientes com certas doenças, observa-se que depois de seguir um certo tratamento nem todos os pacientes são totalmente curados. Os pacientes que são curados são chamados immunes, enquanto os doentes restantes que desenvolvem uma recorrência da doenças são chamados de susceptíveis. A população de interesse pode assim ser considerada como uma mistura destes dois tipos de pacientes. O modelo de mistura proposto por Berkson e Gage (1952) propõe a construção de uma função de sobrevivência com fração de cura, dada por

$$\begin{aligned} S(t;\pi) &= P(T>t) \\ &= P(C=0)P(T>t|C=0) + P(C=1)P(T>t|C=1) \\ &= \pi + (1-\pi)S_1(t), \end{aligned} \tag{1.1}$$

ao considerar uma variável aleatória C com distribuição Bernoulli$(1-\pi)$, cuja probabilidade do paciente ser susceptível é $P(C=1) = (1-\pi)$ e $P(C=0) = \pi$ é a probabilidade de cura. Nesse modelo $S(t;\pi) = S_1(t)$ quando $\pi = 0$, $\lim_{t\to\infty} S(t;\pi) = \pi$ e $\lim_{t\to 0} S(t;\pi) = 1$.

* Coordenacão de Aperfeicõamento de Pessõal de Nível Superior - Brazil (CAPES); FAHUB-UnB
†Dept. of Statistics, University of Brasília, Brasília- DF, 70910-900 (carolina-andrade5@hotmail.com),
‡**ID ORCID:** https://orcid.org/0000-0002-5619-0478, Dept. of Statistics, University of Brasília, Brasília- DF, 70910-900 (cira@unb.br),
§**ID ORCID:** https://orcid.org/0000-0002-9071-8512, Dept. of Statistics, University of Brasília, Brasília- DF, 70910-900 (nakano@unb.br).

Informações intrínsecas a cada observação em estudo podem ser consideradas ao incluir covariáveis no modelo de sobrevivência. Isto se justifica, por exemplo, pelo fato de que a probabilidade de cura pode ser diferente em pessoas de sexos diferentes, em grupos que estão sujeitos a diferentes tipos de medicação ou ainda quando pessoas têm idades distintas. 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 de por exemplo a função de ligação logito,

$$\pi(\boldsymbol{\phi}, \mathbf{z}) = \frac{e^{\boldsymbol{\phi}'\mathbf{z}}}{1 + e^{\boldsymbol{\phi}'\mathbf{z}}}, \tag{1.2}$$

em que $\boldsymbol{\phi}' = (\phi_0, \phi_2, \ldots, \phi_k)$ é o vetor de parâmetros que representam os efeitos das covariáveis, tal que $-\infty < \phi_0, \ldots, \phi_k < \infty$ e $\mathbf{z}' = (1, \mathbf{z}_1, \mathbf{z}_2 \ldots, \mathbf{z}_k)$ é o vetor de covariáveis observadas.

Assim, as proporções de observações curadas e não curadas são dadas por

$$P(C=0) = \pi(\boldsymbol{\phi}, \mathbf{z}) = \frac{e^{\boldsymbol{\phi}'\mathbf{z}}}{1 + e^{\boldsymbol{\phi}'\mathbf{z}}} \quad \text{e} \quad P(C=1) = 1 - \pi(\boldsymbol{\phi}, \mathbf{z}) = \frac{1}{1 + e^{\boldsymbol{\phi}'\mathbf{z}}}. \tag{1.3}$$

Uma das variáveis aleatórias mas utilizadas em análise de sobrevivência é a Weibull contínua. As funções de densidade de probabilidade (fdp) e de sobrevivência de uma variável aleatória T com distribuição Weibull contínua são

$$f_1(t; \alpha, \beta) = \frac{\beta}{\alpha^\beta} t^{\beta-1} \exp\left\{ -\left(\frac{t}{\alpha}\right)^\beta \right\}, \quad t \geq 0 \tag{1.4}$$

e

$$S_1(t; \alpha, \beta) = \exp\left\{ -\left(\frac{t}{\alpha}\right)^\beta \right\}, \quad t \geq 0, \quad \alpha > 0 \quad \text{e} \quad \beta > 0, \tag{1.5}$$

respectivamente, sendo β o parâmetro de forma e α de escala. Desde o trabalho de Berkson e Gage (1952) o modelo (1.2) com a distribuição Weibul contínua tem sido bastante utilizado para modelar dados de sobrevivência. Por outro lado, utilizar um modelo contínuo para analisar dados discretos de sobrevivência pode levar a resultados pouco satisfatórios, (ver Nakagawa e Osaki (1975), Nakano e Carrasco (2006), Fernandes (2013), Brunello e Nakano (2014)).

Assim, neste trabalho, fazemos uma adaptação do modelo de Nakagawa e Osaki (1975) para a distribuição **Weibull discreta** com fração de cura incorporando covariáveis. Esse modelo é obtido ao discretizar a variável T associada as funções (1.4) e (1.5). Á variável Weibull discreta com parâmetros $q, \beta, \boldsymbol{\phi}$ a denotamos por $T \sim Wd(q, \beta, \boldsymbol{\phi})$, cuja função de probabilidade é $p(t) = P(T = t) = P(T \leq t) - P(T \leq t-1)$. Ao utilizar o modelo (1.2) suas funções de probabilidade e de sobrevivência são dadas por

$$\begin{aligned} p(t; q, \beta, \boldsymbol{\phi}') &= [1 - \pi(\boldsymbol{\phi}, \mathbf{z})] p_1(t; q, \beta) \\ &= \left(\frac{1}{1 + e^{\boldsymbol{\phi}'\mathbf{z}}}\right) \left(q^{t^\beta} - q^{(t+1)^\beta}\right) \end{aligned} \tag{1.6}$$

e

$$\begin{aligned} S(t; q, \beta, \boldsymbol{\phi}') &= \pi(\boldsymbol{\phi}, \mathbf{z}) + [1 - \pi(\boldsymbol{\phi}, \mathbf{z})] S_1(t; q, \beta) \\ &= \frac{e^{\boldsymbol{\phi}'\mathbf{z}} + q^{(t+1)^\beta}}{1 + e^{\boldsymbol{\phi}'\mathbf{z}}}, \quad t = 0, 1, \ldots \end{aligned} \tag{1.7}$$

sendo $0 < q = \exp\{-\frac{1}{\alpha^\beta}\} < 1$, $\mathbf{z}_i' = (1, \mathbf{z}_1, \ldots, \mathbf{z}_k)$ o vetor de covariáveis observadas e $\boldsymbol{\phi}' = (\phi_0, \phi_1, \ldots, \phi_k)$ o vetor de $k+1$ parâmetros associados às covariáveis, tal que $\infty < \phi_0, \phi_1, \ldots, \phi_k < \infty$.

No modelo em estudo consideramos a presença de censura à direita, definida quando o evento de interesse não ocorre até o último instante em que o indivíduo foi observado. Diante da presença da censura, 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,

$$\delta_i = \begin{cases} 1, & \text{se} \quad t_i \quad \text{é um tempo de falha,} \\ 0, & \text{se} \quad t_i \quad \text{é um tempo de censura.} \end{cases} \tag{1.8}$$

Como o modelo de sobrevivência (1.7) é uma mistura de distribuições com dados incompletos, na Seção 3 utilizamos o algoritmo EM para obter de forma iterativa as estimativas de máxima verossimilhança dos seus parâmetros. Os resultados do desempenho do algoritmo avaliado por meio de simulação Monte Carlo são apresentados na Seção 3. Além disso, o modelo é aplicado em um conjunto de dados reais e os resultados são mostrados na Seção 4.

2. Estimação via algoritmo EM.

Nesta seção, utiliza- se o algoritmo EM para encontrar a estimativa de máxima verossimilhança dos parâmetros $\boldsymbol{\phi}$, q e β do modelo de mistura (1.7). Teoria e aplicações do algoritmo EM podem ser consultadas no livro de McLachlan and Peel (2000).

Ao considerar (1.6), (1.7) e (1.8), com 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 censura t_i, indicador de falha δ_i, indicador de cura c_i e $\mathbf{z}_i$ vetor de covariáveis da proporção de cura.

Se $\mathbf{T}'$ o vetor de dados incompletos, a função de log- verossimilhança dos dados incompletos é

$$\begin{aligned} l_c(\Phi) &= \sum_{i=1}^{n} \delta_i \log\left\{(1-\pi_i(\boldsymbol{\phi},\mathbf{z}))\left[q^{{t_i}^{\beta}} - q^{(t_i+1)^{\beta}}\right]\right\} \\ &+ \sum_{i=1}^{n}(1-\delta_i)\log\left\{[\pi_i(\boldsymbol{\phi},\mathbf{z})]^{1-c_i} + \left\{[1-\pi_i(\boldsymbol{\phi},\mathbf{z})]q^{(t_i+1)^{\beta}}\right\}^{c_i}\right\}, \end{aligned} \tag{2.1}$$

sendo $\pi_i(\boldsymbol{\phi},\mathbf{z}) = \pi(\boldsymbol{\phi},\mathbf{z}_i)$ e $\Phi = (\boldsymbol{\phi}, q, \beta)$. Sem perda de generalidade, considere que os primeiros m tempos, t_i, são de falha e os $(n-m)$ tempos restantes, t_i, são de censura. Assim, para os primeiros m indivíduos se tem que cada indivíduo com covariável $\mathbf{z}$ corresponde a $c_i = 1$ e o valor de c_i é desconhecido para os $n-m$ indivíduos restantes. Com estas informações, a função de log-verossimilhança do modelo é re-escrita por

$$\begin{aligned} l_c(\Phi) &= \sum_{i=1}^{m} \log\left\{(1-\pi_i(\boldsymbol{\phi},\mathbf{z}))\left[q^{{t_i}^{\beta}} - q^{(t_i+1)^{\beta}}\right]\right\} \\ &+ \sum_{i=(m+1)}^{n} \log\left\{[\pi_i(\boldsymbol{\phi},\mathbf{z})]^{1-c_i} + \left\{[1-\pi_i(\boldsymbol{\phi},\mathbf{z})]q^{(t_i+1)^{\beta}}\right\}^{c_i}\right\}, \end{aligned} \tag{2.2}$$

ao utilizar (1.8). Como as últimas $n-m$ observações de (2.2) são tratadas como dados faltantes, considere $\mathbf{C}'$ o vetor de dados faltantes e $V = (\mathbf{T}', \mathbf{C}')$ o vetor de dados completos, então a densidade dos dados completos é $f(\mathbf{T}', \mathbf{C}'; \Phi)$. O algoritmo EM consiste em alternar iterativamente dois passos, passo E e passo M. No passo E se calcula a esperança do logaritmo da função de verossimilhança dos dados completos dado os dados observados e o vetor de parâmetros do passo anterior. Isto é, na k - ésima iteração do passo E (expectation) do algoritmo EM, calcula-se a função

$$\begin{aligned} Q(\Phi, \Phi^{(k)}) &= E[\log f(\mathbf{T}', \mathbf{C}'; \Phi^{(k)})|\mathbf{T}', \Phi^{(k)}] \\ &= \sum_{i=1}^{m} \log\left\{(1-\pi_i(\boldsymbol{\phi^{(k)}},\mathbf{z}))\left[{q^{(k)}}^{{t_i}^{\beta^{(k)}}} - {q^{(k)}}^{(t_i+1)^{\beta^{(k)}}}\right]\right\} \\ &+ \sum_{i=m+1}^{n} \log[\pi_i(\boldsymbol{\phi^{(k)}},\mathbf{z})]^{1-c_i} P(\mathbf{c_i}=0|T>t_i, \Phi^{(k)}) \\ &+ \sum_{i=m+1}^{n} \log\left\{[1-\pi_i(\boldsymbol{\phi^{(k)}},\mathbf{z})]{q^{(k)}}^{(t_i+1)^{\beta^{(k)}}}\right\}^{c_i} P(\mathbf{c_i}=1|T>t_i, \Phi^{(k)}) \\ &= \sum_{i=1}^{m} \log[(1-\pi_i(\boldsymbol{\phi^{(k)}},\mathbf{z}))\left[{q^{(k)}}^{{t_i}^{\beta^{(k)}}} - {q^{(k)}}^{(t_i+1)^{\beta^{(k)}}}\right] \\ &+ \sum_{i=m+1}^{n} \log[\pi_i(\boldsymbol{\phi^{(k)}},\mathbf{z})]w_1(\mathbf{z}_i; t_i) \\ &+ \sum_{i=m+1}^{n} \log[(1-\pi_i(\boldsymbol{\phi^{(k)}},\mathbf{z})){q^{(k)}}^{(t_i+1)^{\beta^{(k)}}} w_2(\mathbf{z}_i; t_i), \end{aligned} \tag{2.3}$$

sendo $\pi_i(\boldsymbol{\phi^{(k)}}, \mathbf{z}) = \pi(\boldsymbol{\phi^{(k)}}, \mathbf{z}_i)$ calculado conforme (1.3). As funções proporção $w_1(\mathbf{z}_i, t_i)$, $w_2(\mathbf{z}_i, t_i)$ são obtidas a partir de (1.6) e (1.7) por

$$w_1(\mathbf{z}_i, t_i) = \frac{e^{\boldsymbol{\phi^{(k)}}\mathbf{z}_i}}{e^{\boldsymbol{\phi^{(k)}}\mathbf{z}_i} + q^{(k)^{(t_i+1)^{\beta^{(k)}}}}} \tag{2.4}$$

e

$$w_2(\mathbf{z}_i, t_i) = \frac{q^{(k)^{(t_i+1)^{\beta^{(k)}}}}}{e^{\boldsymbol{\phi^{(k)}}\mathbf{z}_i} + q^{(k)^{(t_i+1)^{\beta^{(k)}}}}}. \tag{2.5}$$

O passo M consiste em maximizar a função (2.3) em relação aos parâmetros ϕ, q, β, sendo primeiro atualizados $w_1(\mathbf{z}_i, t_i)$ e $w_2(\mathbf{z}_i, t_i)$. Para facilitar a maximização de (2.3), considera-se

$$Q(\Phi, \Phi^{(k)}) = g_1(\boldsymbol{\phi}) + g_2(q, \beta), \tag{2.6}$$

sendo

$$\begin{aligned} g_1(\boldsymbol{\phi}) = &\sum_{i=1}^{m} \log[1 - \pi_i(\boldsymbol{\phi}^{(k)}, \mathbf{z}_i)] + \sum_{i=(m+1)}^{n} w_1(\mathbf{z}_i, t_i) \log[\pi_i(\boldsymbol{\phi}^{(k)}, \mathbf{z}_i)] \\ &+ \sum_{i=m+1}^{n} w_2(\mathbf{z}_i, t_i) \log[1 - \pi_i(\boldsymbol{\phi}^{(k)}, \mathbf{z}_i)] \end{aligned} \tag{2.7}$$

e

$$\begin{aligned} g_2(q, \beta) = &\sum_{i=1}^{m} \log \left[q^{(k)^{t_i^{\beta^{(k)}}}} - q^{(k)^{(t_i+1)^{\beta^{(k)}}}} \right] \\ &+ \sum_{i=m+1}^{n} w_2(\mathbf{z}_i, t_i) \log \left[q^{(k)^{(t_i+1)^{\beta^{(k)}}}} \right]. \end{aligned} \tag{2.8}$$

Com isso, $\boldsymbol{\phi}^{(k+1)}$ é obtida pela maximização de $g_1(\boldsymbol{\phi})$, enquanto que $\beta^{(k+1)}$ e $q^{(k+1)}$ são obtidas pela maximização de $g_2(\beta, q)$. Estes passos irão se repetir até que $|Q(\Phi, \Phi^{(k+1)}) - Q(\Phi, \Phi^{(k)})| < 10^{-3}$.

Implementação do algoritmo EM. As etapas para a implementação do algoritmo EM podem ser resumidas e descritas da seguinte forma:

1. Inicializar o contador das iterações $k = 0$ e os parâmetros $\boldsymbol{\phi}^{(k)}$, $\beta^{(k)}$ e $q^{(k)}$ representam os chutes iniciais desses parâmetros ;
2. **Passo E**. A partir de $\boldsymbol{\phi}^{(k)}, \beta^{(k)}$ e $q^{(k)}$, obter $w_1(\mathbf{z}_i, t_i)^{(k+1)}$ e $w_2(\mathbf{z}_i, t_i)^{(k+1)}$ utilizando (2.4) e (2.5), respectivamente, e calcular $Q(\Phi, \Phi^{(k+1)})$ conforme foi apresentada em (2.6);
3. **Passo M**. $\boldsymbol{\phi}^{(k+1)}$ é obtido através da maximização de (2.7) com respeito a $\boldsymbol{\phi}$, enquanto $\beta^{(k+1)}$ e $q^{(k+1)}$ são obtidos através da maximização de (2.8) com respeito a β e q, respectivamente;
4. Atualizar o contador de (k) para $(k + 1)$;
5. Repetir os passos 2, 3 e 4 até atingir a convergência das estimativas de $\boldsymbol{\phi}, \beta$ e q.

3. Resultados das simulações.

Para avaliar a performance do algoritmo EM na estimação dos parâmetros do modelo $T \sim Wd(q, \beta, \boldsymbol{\phi})$, nesta seção, obtemos as estimativas dos parâmetros com amostras de tamanhos $n = 250$ e $n = 500$, via simulação Monte Carlo.

Para simular o mecanismo de censura à direita, o tempo das observações curadas foi dado pela parte inteira de $1,5\times$TM, em que TM é o maior tempo gerado .

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 enquanto que β é inversamente proporcional aos tempos, de modo que quanto menor o valor de β maiores os valores dos tempos e em consequência há uma quantidade maior de tempos distintos.

Para as simulações optou-se por considerar uma única covariável dicotômica, Z, Bernoulli com probabilidade de sucesso 0,6, pelo fato de ser possível visualizá-los graficamente através das estimativas das curvas de sobrevivência.

Foram considerados quatro cenários que diferem quanto ao total de censura. O percentual de censura das observações sob risco foi igual a 10 % nos três primeiros cenários e no quarto cenário cosidera-se apenas censura nos curados. Em cada cenário definimos seis vetores de parâmetros para Φ e nos casos em que $\beta = 1$ esses vetores correspondem a um modelo geométrico discreto com fração de cura com covariáveis.

Para cada Φ foram realizadas 1.000 simulações de tamanho n 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 10^{-3}.

A Tabela 3.1 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: $\pi(\boldsymbol{\phi}, Z = 0) = P(c = 0|Z = 0) = 11,9\,\%$ e $\pi(\boldsymbol{\phi}, Z = 1) = P(c = 0|Z = 1) = 23,1\,\%$. Os resultados das estimativas obtidas juntamente com cada EQM são mostrados na Tabela 3.2.

CUADRO 3.1

Tabela 3.1. 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

CUADRO 3.2

Tabela 3.2. 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

Com base nos resultados da Tabela 3.2, 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.

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 A.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 3.3. 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: $\pi(\boldsymbol{\phi}, Z = 0) = P(c = 0|Z = 0) = 73,1\,\%$ e $\pi(\boldsymbol{\phi}, Z = 1) = P(c = 0|Z = 1) = 26,9\,\%$. Os resultados das estimativas com seus respectivos EQM são mostrados na Tabela 3.4.

CUADRO 3.3

Tabela 3.3.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

CUADRO 3.4

Tabela 3.4. 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 3.4 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 A.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: $\pi(\boldsymbol{\phi}, Z = 0) = P(c = 0|Z = 0) = 42,6\,\%$ e $\pi(\boldsymbol{\phi}, Z = 1) = P(c = 0|Z = 1) = 84,6\,\%$. 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 3.5 apresenta os valores dos parâmetros utilizados para gerar os dados deste cenário e a Tabela 3.6 apresenta os resultados das médias das estimativas e do EQM, obtidos através das simulações.

Analisando os resultados apresentados na Tabela 3.6, 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

CUADRO 3.5
Tabela 3.5. 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

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 A.3 através das estimativas das curvas de sobrevivência.

CUADRO 3.6
Tabela 3.6. 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 2	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}	0,5068	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	0,05	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

Por fim, o Cenário 4 foi gerado com os mesmos parâmetros do Cenário 2 (Tabela 3.3), com a diferença de que neste há censura apenas nos indivíduos curados. O percentual médio de censura é igual a 46 %. A Tabela 3.7 apresenta os resultados das estimativas e do EQM obtidos através das simulações.

CUADRO 3.7
Tabela 3.7. 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

Com base nos resultados da Tabela 3.7 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 A.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.

4. Aplicações em dados reais. Nesta seção mostra-se uma aplicação do modelo de regressão Weibull discreto com fração de cura em dados reais. Os dados correspondem ao tempo até a morte de pacientes submetidos a um transplante de medula óssea (TMO) para tratamento de leucemia mielóide crônica (LMC). Os dados 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 ($Z = 0$ se não houve a ocorrência da doença e $Z = 1$ se houve a ocorrência da doença).

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 4.1 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.

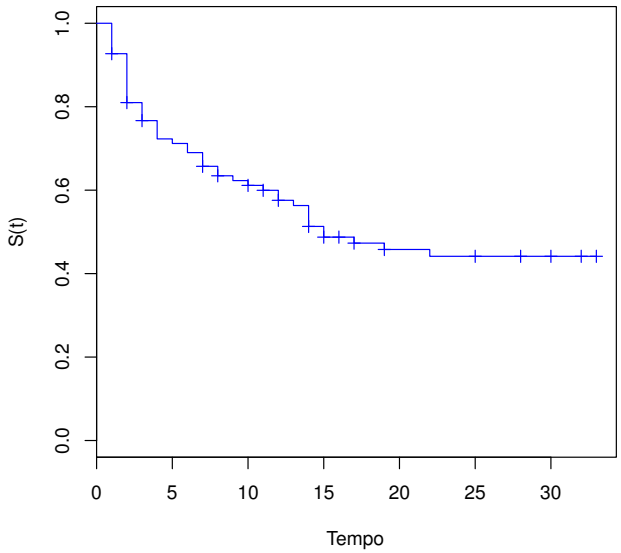

FIGURA 4.1. *Função de sobrevivência estimada pelo método de Kaplan-Meier de pacientes com leucemia mielóide crônica submetidos ao transplante de medula óssea.*

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 4.1. O nível de confiança considerado foi de 95 %.

Como pode ser observado nos resultados da Tabela 4.1, 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

CUADRO 4.1

Tabela 4.1.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 95 % 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)

do parâmetro ϕ_1. Ao calcular os valores de $\pi(\boldsymbol{\phi}, Z = 0)$ e $\pi(\boldsymbol{\phi}, Z = 1)$ 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 4.2 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.

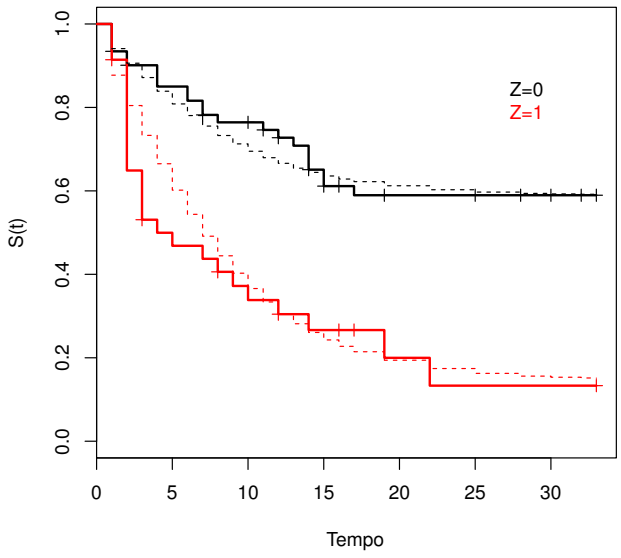

FIGURA 4.2. *Funções de sobrevivência estimadas para pacientes que fizeram transplante de medula óssea. Se $Z = 1$ significa que o paciente teve doença enxerto aguda. A linha contínua representa a estimativa obtida via estimador de Kaplan-Meier e a linha tracejada foi estimada através do modelo de regressão Weibull discreto com fração de cura.*

Foi realizado também o ajuste do modelo Weibull contínuo a estes dados discretos.

O AIC do modelo Weibull discreto foi igual a -384,21, enquanto que o do modelo Weibull contínuo foi -379,07, com o que pode-se concluir que o modelo discreto teve melhor desempenho.

Nesta aplicação haviam 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 não permitirá a 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 $0 < \beta < 1$.

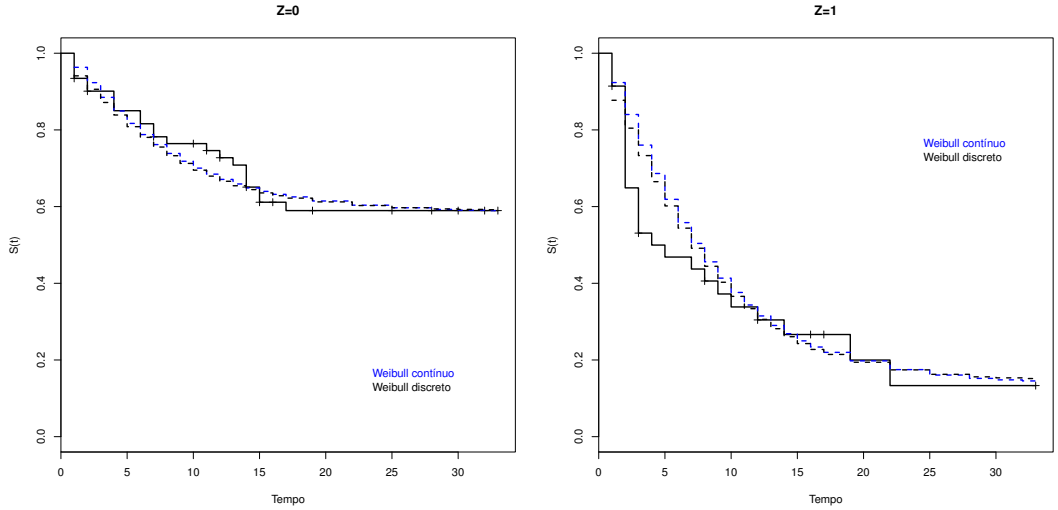

FIGURA 4.3. *Funções de sobrevivência estimadas a partir dos modelos de regressão Weibull discreto e Weibull contínuo com fração de cura para pacientes que fizeram transplante de medula óssea. Aqui, $Z = 1$ significa que o paciente teve doença enxerto aguda. A linha contínua representa a estimativa obtida via estimador de Kaplan-Meier e as linhas tracejadas foram estimadas através dos modelos de regressão Weibull discreto e contínuo com fração de cura.*

5. Conclusões. 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 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 ϕ, visto que nesta situação estará sendo feita a modelagem com o intuito de estimar a probabilidade de cura ($\pi(\phi, \mathbf{z}) = \mathbf{P}(\mathbf{c} = \mathbf{0}|\mathbf{Z})$) 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.

Referencias

[1] Berkson J and Gage, R. P. *Survival Curve for Cancer Patients Following Treatment*, Journal of the American Statistical Association, 4 (1952), pp. 501–515.

[2] Brunello, G. H. V. and E. G. Nakano, *Inferência bayesiana no modelo Weibull discreto em dados com presença de censuras*, Tendências em Matemática Aplicada e Computacional, 16(2015), pp. 97–110.

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

[4] Fernandes, L. M. *Inferência bayesiana em modelos discretos com fração de cura*, Dissertação de Mestrado em Estatística - Departamento de Estatística, Universidade de Brasília, 2015.

[5] McLachlan, G. and Peel, D. *Finite Mixture Models*, Wiley & Sons, Canada, 2000.

[6] Nakagawa, T. and Osaki, S. *The discrete weibull distribution*, IEEE Transactions on Reliability, 24(1975), 300 – 301.

[7] Nakano, E. Y. and Carrasco, C. G. *Uma avaliação do Uso de um Modelo Contínuo na Análise de Dados Discretos de Sobrevivência*, Tendências em Matemática Aplicada e Computacional, 7(2006), 91 – 100.

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

Apéndice A. Alguns perfis adicionais.

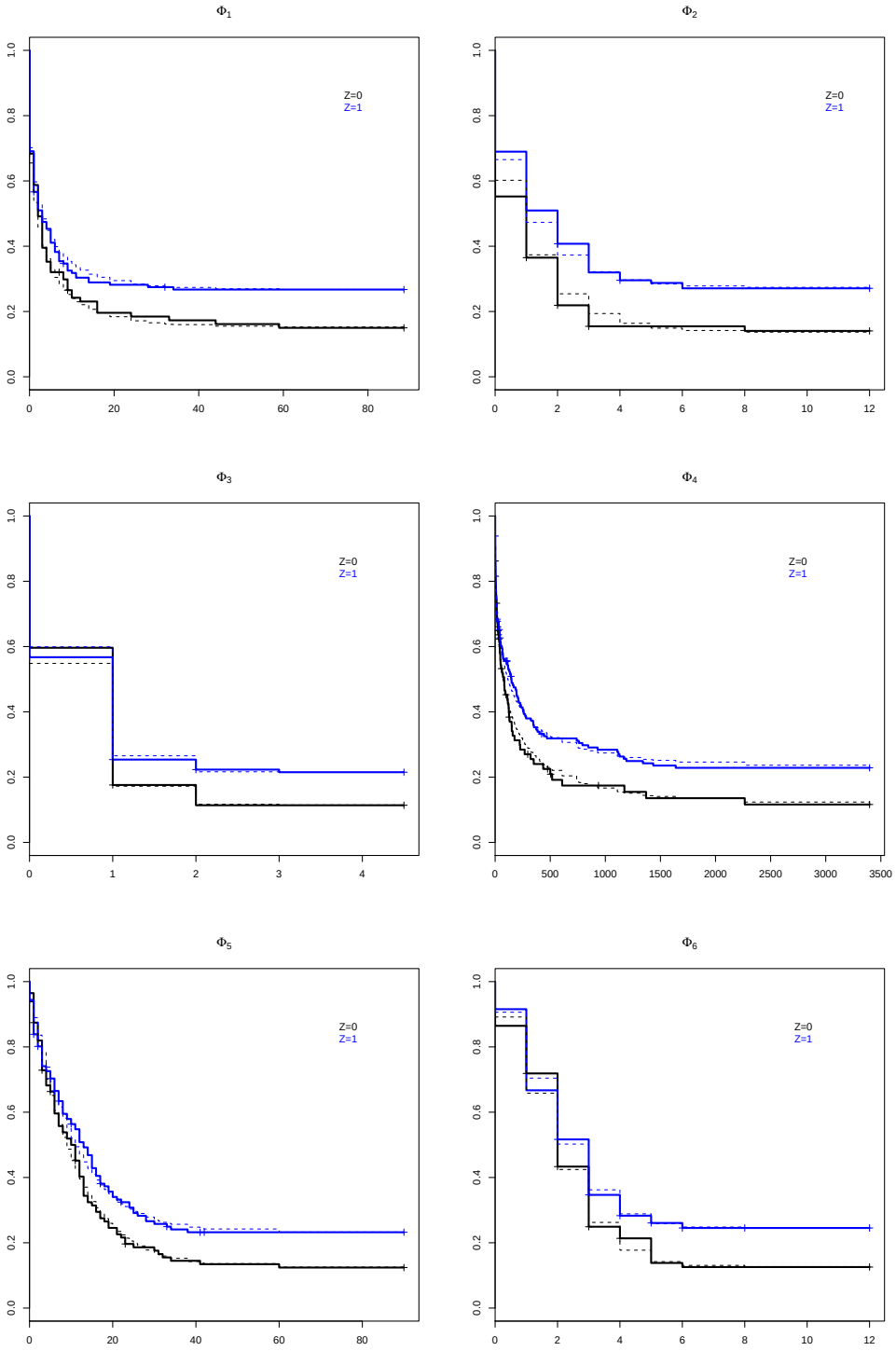

FIGURA A.1. *Funções de sobrevivência estimadas para o Cenário 1 considerando uma amostra de tamanho n=500. As linhas sólidas representam as estimativas de Kaplan-Meier e as linhas tracejadas as estimativas pelo modelo proposto.*

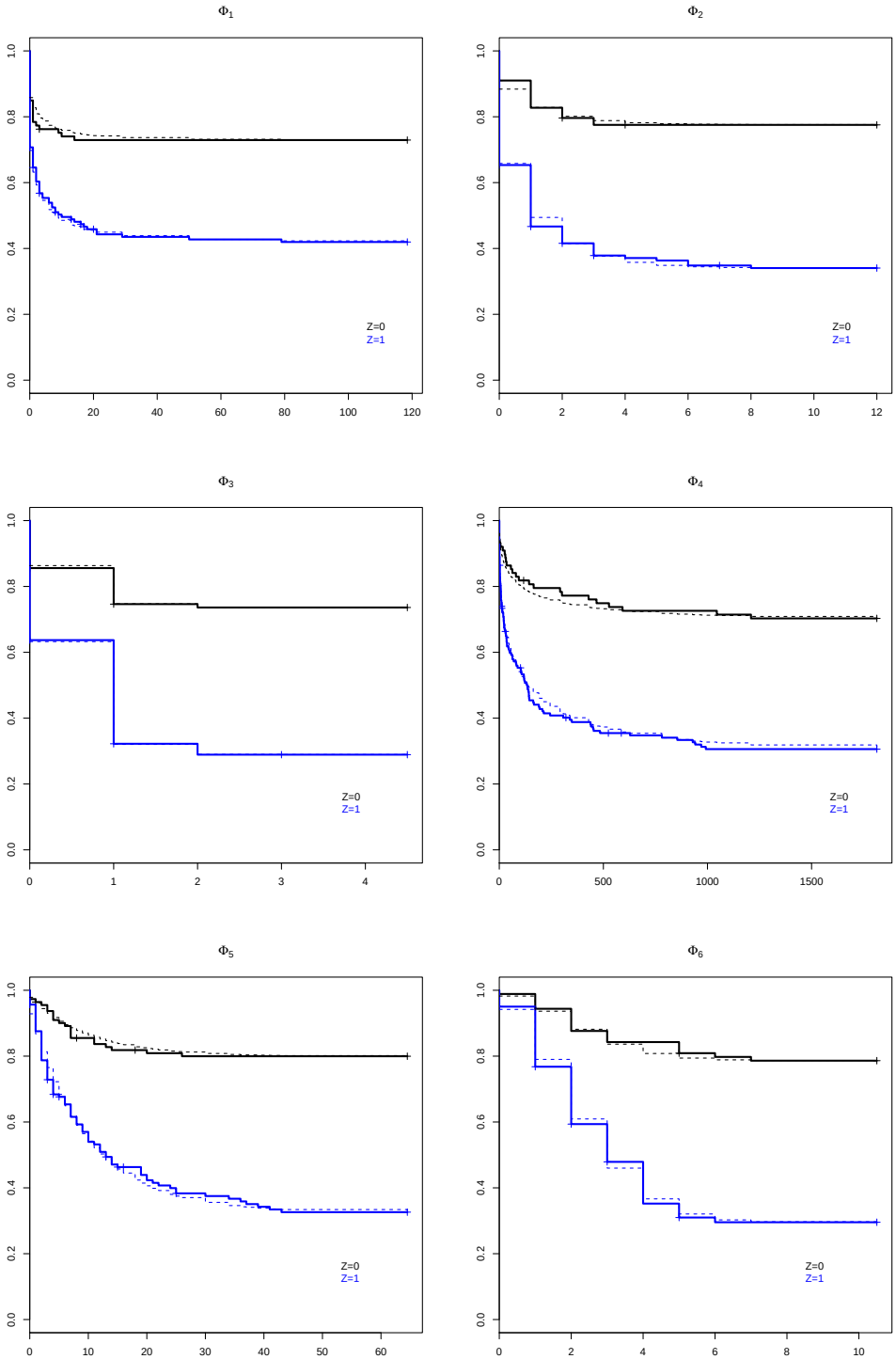

FIGURA A.2. *Funções de sobrevivência estimadas para o Cenário 2 considerando uma amostra de tamanho n=500. As linhas sólidas representam as estimativas de Kaplan-Meier e as linhas tracejadas as estimativas pelo modelo proposto.*

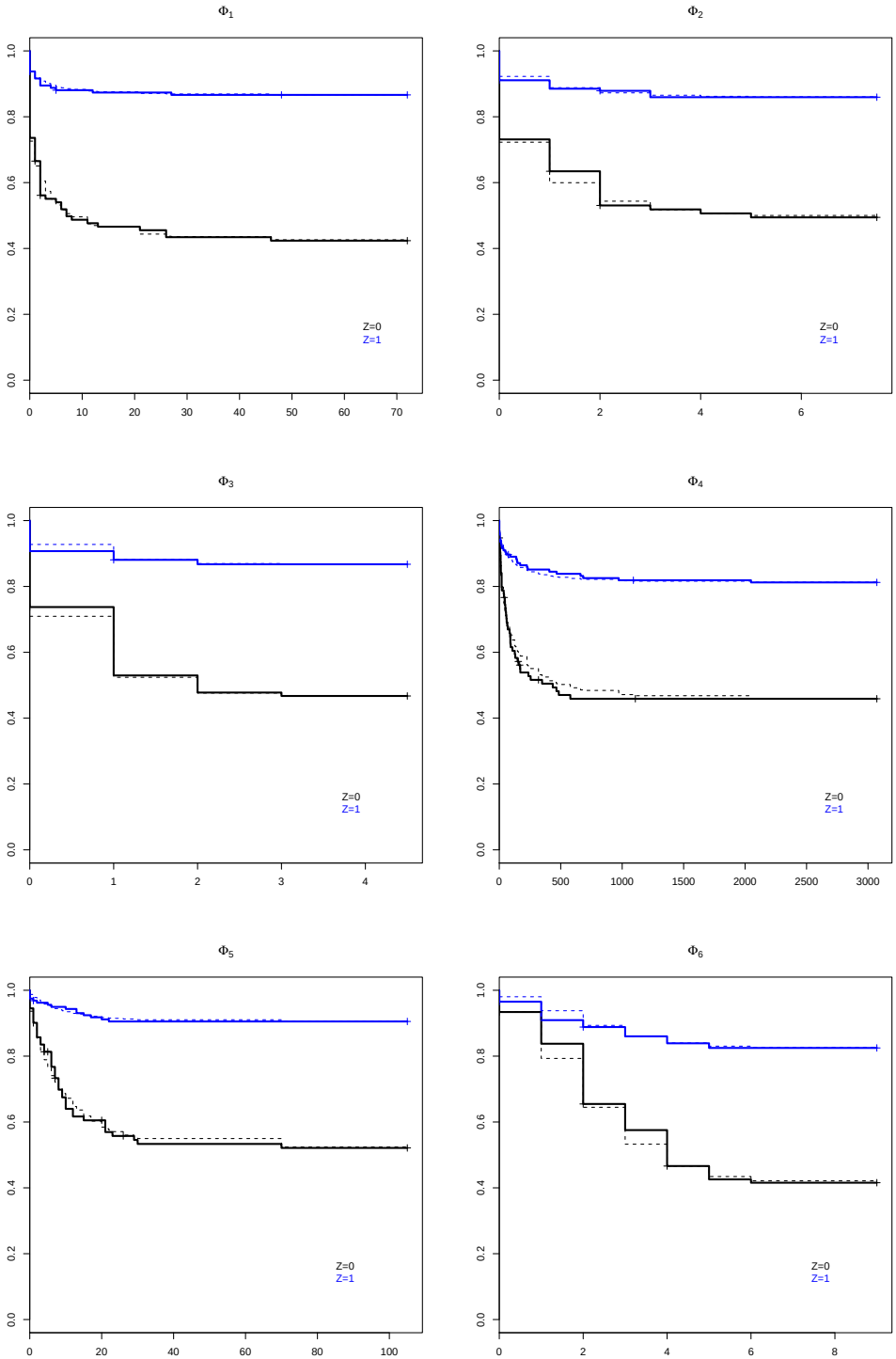

FIGURA A.3. *Funções de sobrevivência estimadas para o Cenário 3 considerando uma amostra de tamanho n=500. As linhas sólidas representam as estimativas de Kaplan-Meier e as linhas tracejadas as estimativas pelo modelo proposto.*

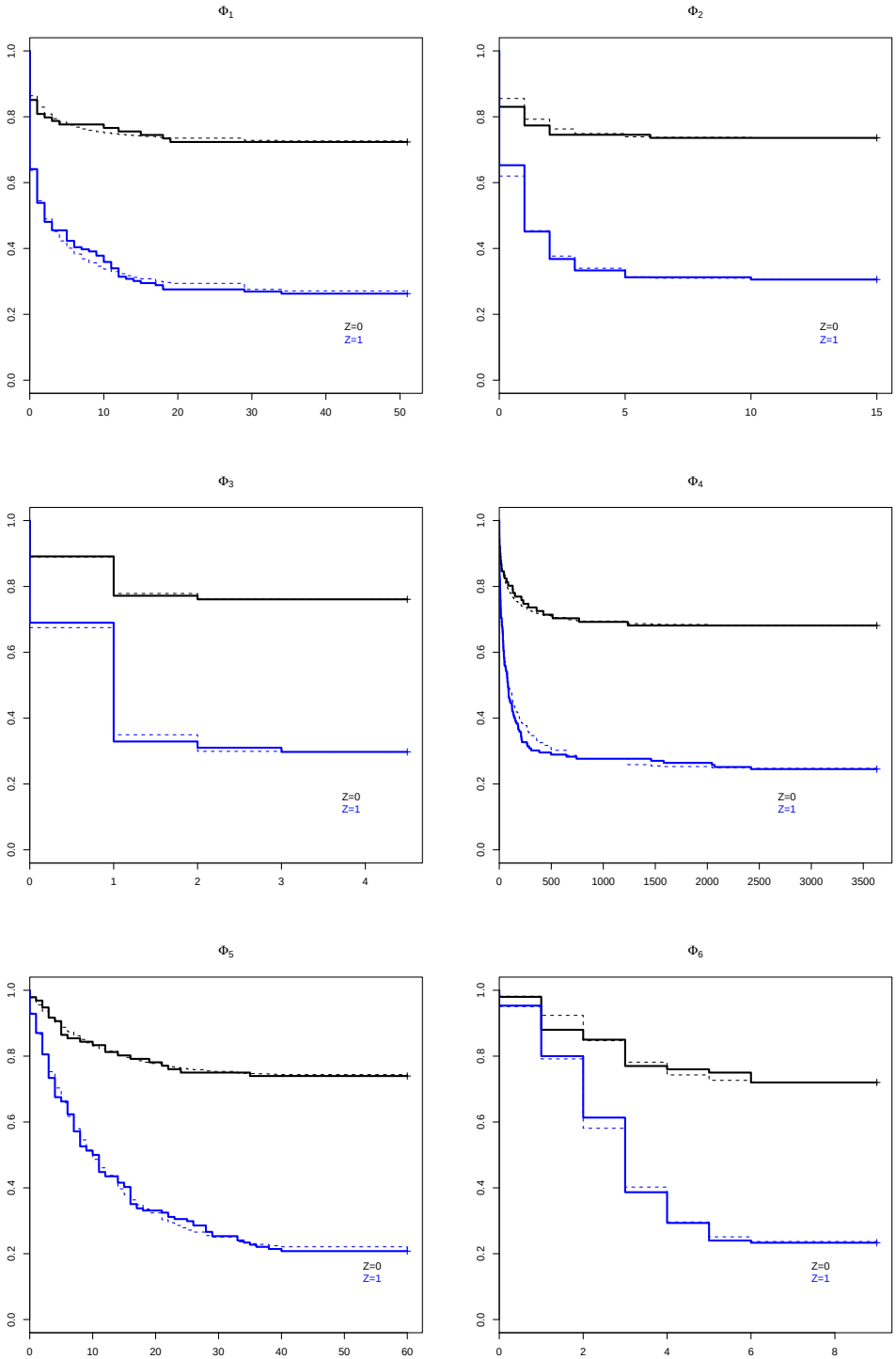

FIGURA A.4. *Funções de sobrevivência estimadas para o Cenário 4 considerando uma amostra de tamanho n=500. As linhas sólidas representam as estimativas de Kaplan-Meier e as linhas tracejadas as estimativas pelo modelo proposto.*