MODELOS DE SOBREVIVÊNCIA COM LONGA DURAÇÃO: UMA APLICAÇÃO A GRANDES BANCOS DE DADOS FINANCEIROS

Daniele Cristina Tita GRANZOTTO[1]
Francisco LOUZADA NETO[2]
Gleici da Silva Castro PERDONÁ[3]

▪ RESUMO: Os modelos de análise de sobrevivência com longa duração acomodam a heterogeneidade de duas populações (susceptíveis e imunes ao evento de interesse). Para exemplificar a aplicabilidade de tais modelos a grande bancos de dados da área de finanças, consideramos o modelo proposto por [3], assumindo as distribuições de sobrevivência Weibull e log-logística para os tempos de duração. Um estudo de simulação foi realizado com o propósito de testar a medida de distância entre curvas de Kaplan-Meier e a ajustada como alternativa à métricas usuais. Também para verificar o comportamento de tais em diferentes situações de percentuais de censura e tamanhos de amostras. Verificamos que, uma métrica simples como a medida de distância entre curvas, é capaz de selecionar o modelo mais apropriado aos dados na presença de longa-duração tanto para pequenas quanto para grandes carteiras de clientes, mesmo na presença de censuras.

▪ PALAVRAS-CHAVE: Sobrevivência; dados financeiros; longa duração; critérios de seleção de modelos.

1 Introdução

A fidelização e retenção de clientes têm papel fundamental nas empresas que hoje atuam em mercados altamente competitivos, principalmente naquelas ligadas à área de finanças: bancos, financiadoras, seguradoras etc.

[1]Universidade Estadual de Maringá - UEMA, Departamento de Estatística, CEP: 87020-900, Maringá, PR, Brasil. E-mail: *dctgranzotto@uem.br*
[2]Universidade Federal de São Carlos - UFSCar, Departamento de Estatística, CEP: 18052-780, São Carlos, SP, Brasil. E-mail: *dfln@ufscar.br*
[3]Universidade de São Paulo - USP, Faculdade de Medicina de Ribeirão Preto - FMRP, Departamento de Medicina Social, CEP: 14049-900, Ribeirão Preto, SP, Brasil. E-mail: *pgleici@fmrp.usp.br*

Para a manutenção ou acréscimo da lucratividade é primordial que estas instituições identifiquem antecipadamente clientes com alto potencial de ruptura de relacionamento, possibilitando ações preventivas que evitem a perda de tais clientes.

Neste contexto, técnicas estatísticas usadas em análise de sobrevivência e confiabilidade podem ser aplicadas e desenvolvidas [11]. Uma característica importante presente em carteiras de clientes é a fração de clientes fidelizados, essa característica exige, além das técnicas usuais, a utilização de modelos com longa-duração.

Os modelos de análise de sobrevivência com longa duração, possuem vantagem com relação aos modelos de sobrevivência usuais por incorporem a heterogeneidade de duas subpopulações (susceptíveis e imunes ao evento de interesse) e são conhecidos também como modelos com fração de imunes (*cure rate models*). Estes modelos podem ser utilizados quando existe a possibilidade da não ocorrência do evento de interesse para uma porcentagem dos indivíduos de uma população. Em experimentos biomédicos, uma porcentagem dos pacientes podem não experimentar a ocorrência de um evento de interesse, por exemplo, de um determinado tipo de doença. Em dados industriais, um equipamento pode ter duração maior do que seu tempo de missão. Em dados financeiros, um cliente pode não se tornar inadimplente etc.

Vários modelos tem sido formulados para acomodar dados de sobrevivência na presença de longa duração, sendo o mais antigo o de [3]. A partir deste artigo, muitos outros foram escritos considerando o modelo proposto não somente para dados relacionados a experimentos na área da saúde e industrial, como também para experimentos na área de finanças, sinistros em seguradoras dentre outras áreas, ver por exemplo [7], [15] e [8]. Outros modelos para dados de longa duração podem ser encontrados em, por exemplo, [5].

Na área financeira, que é o foco deste artigo, admite-se que os indivíduos podem ser classificados como imunes. Neste caso dizemos "bons"ou fidelizados (sem possibilidade de apresentar o evento de interesse), com probabilidade p, ou serem susceptíveis, neste caso dizemos ¨maus¨ ou não fidelizados, com probabilidade $q = 1 - p$. A cada indivíduo associamos uma variável aleatória T, representando o tempo até a ocorrência do evento de interesse ou até a censura.

Desta forma, dada uma função de sobrevivência, $S(t)$, temos que $\lim_{t\to\infty} S(t) = p$, onde p é a proporção de não ocorrência do evento de interesse na população.

O modelo proposto por [3] é caracterizado pela função de sobrevivência dada por

$$S(t) = P\left(T > t\right) = p + (1 - p) \times S_0\left(t\right) \tag{1}$$

em que $S_0\left(\cdot\right)$ é a função de sobrevivência para indivíduos não fidelizados, de tal forma que para $t \to +\infty$, $S_0(t) \to 0$ e assim, $\lim_{t\to\infty} S(t) = (1 - p)(> 0)$. A função $S_0(t)$ pode ser especificada por funções de sobrevivência de modelos de sobrevivência usuais, tais como o modelo de Weibull, o log-logístico, log-normal, entre outros, ver por exemplo [12], [10].

2 Modelos de longa-duração

Nesta seção apresentamos dois modelos de longa duração usuais: o modelo Weibull e o modelo log-logístico. De forma geral, outros modelos de sobrevivência poderiam ser considerados, mas os modelos Weibull e log-logístico são aqui focados por apresentarem características complementares com relação ao comportamento de suas funções de risco. Enquanto o modelo Weibull pode acomodar dados de sobrevivência com funções de risco constante, crescente e decrescentes, o modelo log-logístico acomoda dados de sobrevivência com função de risco unimodal.

Inferência para os parâmetros do modelo são baseadas na função de verossimilhança, a qual, contemplando a presença de dados censurados a direita, é dada por

$$L(\theta) = \prod_{i=1}^{n} [f_p(t_i;\theta)]^{\delta_i} [S_p(t_i;\theta)]^{1-\delta_i}, \tag{2}$$

onde $S_p(t)$ representa a função de sobrevivência de longa-duração do modelo a ser estimado, $f_p(t)$ representa a função densidade de probabilidade de longa-duração e θ os parâmetros a serem estimados, sendo os estimadores de máxima verossimilhança obtidos via maximização direta de (2).

2.1 Modelo Weibull

O modelo Weibull, proposto originalmente por [16], é muito importante em análise de sobrevivência por apresentar uma grande variedade de formas para a função de risco, todas com uma única propriedade em comum: a sua taxa de falha é monótona, ver [9].

Usando o modelo proposto por [3] dado em (1), o modelo Weibull de sobrevivência com longa duração é dado por

$$S_p(t) = P(T > t) = p + (1-p) \times \exp\left[-\left(\frac{t}{\mu}\right)^{\beta}\right], \tag{3}$$

onde $\mu > 0$ é o parâmetro de escala da distribuição Weibull, p é o percentual de clientes fidelizados na população em estudo e β é o parâmetro de forma da distribuição (se $\beta < 1$ a função de risco do modelo é monótona decrescente; se $\beta = 1$ é constante e; se $\beta > 1$ é monótona crescente).

Considerando a equação (2), $t_1, t_2, \ldots, t_n$ os tempos em estudo e δ_i a variável indicadora de censura do tipo I, temos que, para caso Weibull (3), a função de verossimilhança dada por

$$\begin{aligned} L(p,\mu,\beta|\mathbf{t},\delta) &= \prod_{i=1}^{n} \left[(1-p)\frac{\beta}{\mu}\left(\frac{t_i}{\mu}\right)^{\beta-1} \exp\left[-\left(\frac{t_i}{\mu}\right)^{\beta}\right]\right]^{\delta_i} \\ &\quad \left[p + (1-p)\exp\left[-\left(\frac{t_i}{\mu}\right)^{\beta}\right]\right]^{1-\delta_i}. \end{aligned} \tag{4}$$

Para obter as estimativas dos parâmetros, [12] propuseram uma reparametrização onde $\mu_0 = \ln(\mu)$, $\beta_0 = \ln(\beta)$ e $p_0 = \ln\left(\frac{p}{1-p}\right)$, garantindo assim que os parâmetros estimados, μ e β, sejam sempre positivos e que o parâmetro estimado p esteja no intervalo $[0, 1]$.

2.2 Modelo Log-Logístico

Usando novamente o modelo proposto por [3] dado em (1), o modelo log-logístico de sobrevivência com longa duração é dada por,

$$S_p(t) = p + \frac{1-p}{1+e^{\mu}t^{\sigma}}, \tag{5}$$

onde $-\infty < \mu < \infty$, $\sigma > 0$ e $0 \leq p \leq 1$.

Assim, de (2), temos a função de verossimilhança para o caso log-logístico dado por,

$$L(p, \mu, \sigma|\mathbf{t}, \delta) = \prod_{i=1}^{n}\left[(1-p)\frac{e^{\mu}\sigma t_i^{\sigma-1}}{[1+e^{\mu}t_i^{\sigma}]^2}\right]^{\delta_i}\left[p + \frac{1-p}{1+e^{\mu}t_i^{\sigma}}\right]^{1-\delta_i} \tag{6}$$

Da mesma forma que para o caso Weibull, para garantir a positividade do parâmetro σ e para que o parâmetro p esteja no intervalo $[0, 1]$, consideramos a reparametrização $\mu_0 = \ln(\mu)$, $\sigma_0 = \ln(\sigma)$ e $p_0 = \ln\left(\frac{p}{1-p}\right)$.

3 Critérios de seleção de modelos

Para avaliar o ajuste do modelo, utilizamos três critérios diferentes. O primeiro método de seleção de modelo que será apresentado aqui, é o critério AIC (*Akaike Information Criterio*) definido por

$$\Delta AIC = -2\ln\left[\frac{\sup_{M_1} f(x|\theta_1, M_1)}{\sup_{M_2} f(x|\theta_2, M_2}\right] - 2(d_2 - d_1), \tag{7}$$

onde, d_i, $i = 1, 2$, representa o número de parâmetros de cada modelo [1].

O segundo método em questão, é critério BIC (*Bayesian Information Criterio*). Assim, para o cálculo, temos que

$$\Delta BIC = -2\ln\left[\frac{\sup_{M_1} f(x|\theta_1, M_1)}{\sup_{M_2} f(x|\theta_2, M_2}\right] - (d_2 - d_1)\ln n, \tag{8}$$

onde, n é o tamanho da amostra em questão e d_i, $i = 1, 2$, é o número de parâmetros de cada modelo [14].

Os dois critério apresentados, AIC e BIC, têm como objetivo introduzir a complexidade do modelo no critério de seleção, pois são critérios que penalizam a função de verossimilhança [13], [4], [2].

O terceiro critério a ser usado para verificar o ajuste do modelo, alternativo aos métodos conhecidos apresentados anteriormente, é o de medir a distância entre pontos da curva ajustada e da curva empírica (Kaplan-Meier) do modelo em questão, usando para isto, a norma Euclidiana.

Optaremos sempre pela curva que tiver mais proximidade, ou seja, menor valor da distância que é dada por

$$d(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}, \tag{9}$$

onde a representa o valor da curva estimada pela empírica de sobrevivência, b o valor estimado pela sobrevivência do modelo paramétrico em teste e n o número de valores plotados.

Este terceiro critério não é comumente usado para seleção de modelos como será empregado neste estudo.

4 Estudo de simulação

Nesta seção reportamos os resultados de um estudo de simulação realizado com o objetivo de comparar os diferentes critérios de comparação de modelos apresentados na seção anterior.

4.1 Especificações gerais

Quando se trabalha com carteiras de clientes para a análise de crédito, financiamento ou dados de seguradoras, ou seja, grandes bancos de dados, muitos dos clientes, apesar de serem expostos ao evento de interesse não serão acometidos por ele, ou seja são fidelizados. Quando isto ocorre, existe uma grande quantidade censuras à direita, o que justificaria a utilização de modelos apropriados, por exemplo, o uso de modelos com longa duração.

Para comparar modelos e verificar o ajuste, algumas métricas usuais para análise de adequabilidade do modelo tais como, AIC e BIC são utilizadas, mas além dessas métricas usuais propomos a utilização da norma Euclidiana como forma alternativa para a análise da adequabilidade do modelo estimado.

Neste contexto, o procedimento de reamostragem Bootstrap [6], é utilizado para verificar quais das métricas é a mais adequada, considerando diferentes tamanhos de amostras e porcentagens de censura.

Grosso modo, este processo de reamostragem tenta realizar o que seria desejável na prática, isto é, repetir a experiência da amostragem B vezes. Além disso, o procedimento trata a amostra observada como se esta representasse exatamente toda a população. Sendo assim, para $t = (t_1, t_2, \ldots, t_n)$ uma amostra aleatória contendo n tempos de sobrevivência disponíveis para análise, com $\delta_i = 1$ para os tempos exatamente observados e, $\delta_i = 0$, para os tempos censurados a direita, $i = 1, 2, \ldots, n$. O processo de reamostragem *Bootstrap* consiste

em reamostrar B amostras $T^{*(1)}, T^{*(2)}, \ldots, T^{*(B)}$, independentes e identicamente distribuídas, cada uma de tamanho n, onde T é o conjunto de dados disponíveis dado por $T = (t, \delta)$.

Utilizando o procedimento Bootstrap apresentado, foram geradas amostras de duas distribuições: Weibull e log-logística. As amostras foram geradas de tamanhos e com valores de parâmetros pré-estabelecidos. Para estas amostras, que trataremos como "amostras originais", geradas para específicos casos (apresentados na Figura 1), o processo de reamostragem Bootstrap é empregado e $B = 1.000$ reamostras foram retiradas para cada caso em estudo.

Fizemos então um estudo de má especificação de modelo. Estimamos os parâmetros do modelo através do qual foi gerada a amostra, bem como estimamos os parâmetros da outra distribuição que foi considerada. Por exemplo, se a amostra original for gerada da distribuição Weibull, estimamos os parâmetros do modelo Weibull e também do modelo log-logístico. Obtivemos assim, 1.000 estimativas para cada um dos parâmetros da distribuição e uma média destas estimativas foi calculada. Posteriormente, calculamos os valores das métricas (AIC, BIC e norma Euclidiana), para cada uma dessas reamostras, bem como seus valores médios.

Para comparar os resultados, consideramos a razão destas métricas, ou seja, para sua obtenção, quando a amostra foi gerada a partir da distribuição Weibull, as estimativas das métricas para este modelo foram divididas pelas estimativas obtidas para os parâmetros do modelo log-logístico. O processo inverso também foi considerado quando geramos amostras a partir do modelo log-logístico.

Como quanto menor o valor obtido pela métrica, mais adequado é o modelo em questão, esperamos que o valor da métrica, para o modelo a partir do qual gerou-se os dados, seja menor do que do outro modelo e, a razão obtida deverá ser menor do que 1. Quando isto não ocorre significa que o critério de seleção, métrica, não foi capaz de identificar o modelo que gerou os dados.

No estudo de simulação, vários tamanhos de amostras, percentuais de censura e parâmetros dos modelos foram considerados. Estes casos são apresentados na Figura 1.

5 Resultados

Os resultados provenientes do estudo de simulação descrito anteriormente foram condensados em box-plots.

Como exemplo, consideramos o caso onde as amostras foram obtidas de uma distribuição Weibull com parâmetro de forma $\beta = 0,5$. Para este caso temos as Figuras 2, 3 e 4 que seguem.

A Figura 2 apresenta as razões de normas euclidianas para os quatro tamanhos de amostras estudados: $n = 100$, $n = 5.000$, $n = 15.000$ e $n = 30.000$; considerando os quatro percentuais de censura: $p = 10\%$,$p = 25\%$, $p = 50\%$ e $p = 75\%$. Podemos verificar que quanto menor o tamanho da amostra, maior a dispersão das razões das Normas Euclidianas. Esta dispersão diminui à medida que aumenta o tamanho da amostras. Também podemos verificar que, à medida que aumenta o percentual de

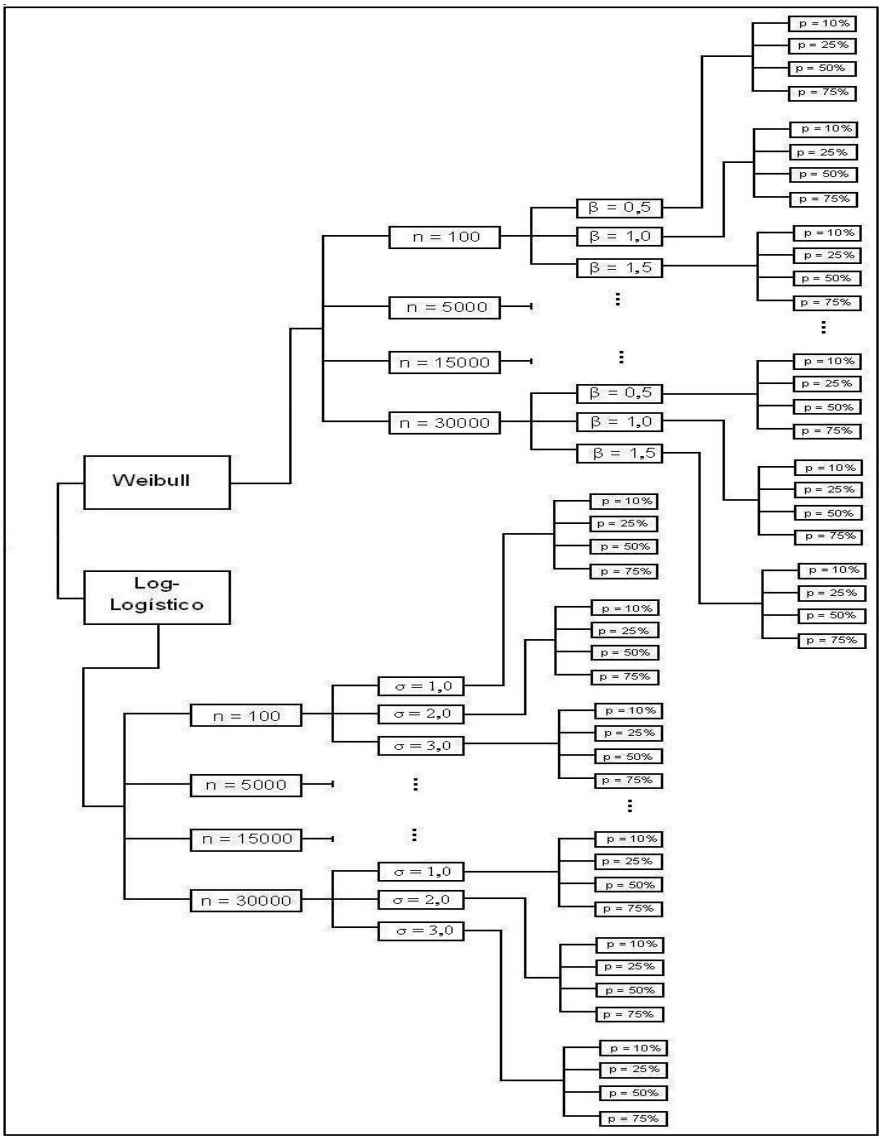

Figura 1 - Delineamento do estudo de simulação.

censura, mais os valores obtidos para estas razões se aproximam 1. Contudo, mesmo com um aumento da dispersão e dos valores das razões se aproximarem de 1 para alguns casos citados, a Norma Euclidiana se mostrou uma métrica satisfatória para selecionar modelos com a presença de longa duração para amostras da distribuição Weibull.

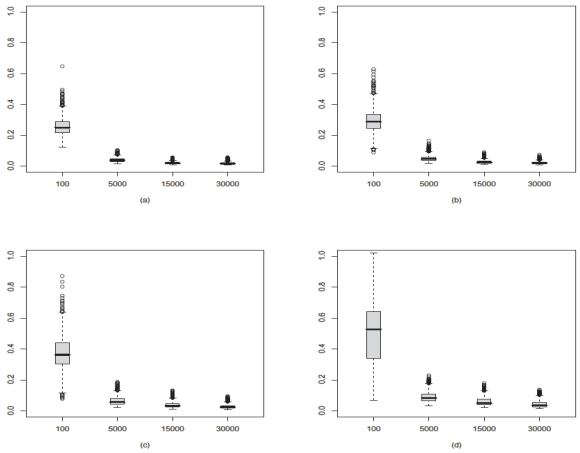

Figura 2 - Box-Plot das razões de normas Euclidianas do modelo Weibull com relação ao modelo log-logístico, para amostra gerada de uma Weibull com $\beta = 0,5$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

Nas Figuras 3 e 4, temos as razões de AIC's e BIC's, respectivamente. Podemos verificar que, para amostras pequenas, o mesmo comportamento apresentado na Figura 2 é obtido, ou seja, uma grande dispersão foi observada. À medida que o tamanho da amostra aumenta, esta dispersão tende a ser menor. Outra observação é com relação ao comportamento gráfico quando aumenta o percentual de censura na amostra. Quanto maior este percentual, mais as razões destas métricas se aproximam de 1.

Assim, podemos evidenciar que, para o caso onde as amostras foram geradas da distribuição Weibull, com risco decrescente, na presença de longa duração, as três métricas em questão (Norma Euclidiana, AIC e BIC), demonstraram ser boas métricas para verificar o ajuste do modelo. Similarmente, para os casos estudados, onde as amostras foram geradas da distribuição Weibull, com risco constante e crescente, as três métricas em estudo também podem ser consideradas adequadas para seleção dos modelos.

Agora, consideremos os caso onde as amostras geradas são provenientes da distribuição log-logística. Como exemplo, temos o caso onde as amostras foram geradas de uma log-logística com parâmetro $1/\sigma = 2$.

As Figuras 5, 6 e 7 apresentam, respectivamente, as razões de Normas Euclidianas, AIC's e BIC's para estas amostras. Verificamos que, à medida que o tamanho da amostra aumenta, diminui a dispersão das razões de métricas, isto considerando todos os casos estudados e percentuais de censura, comportamento similar ao mostrado anteriormente.

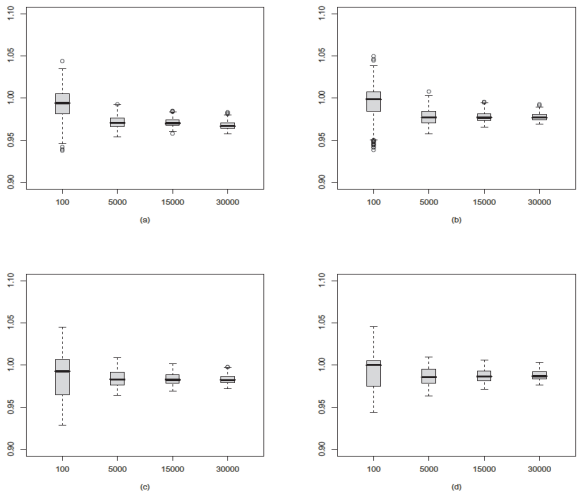

Figura 3 - Box-Plot das razões de AIC's do modelo Weibull com relação ao modelo log-logístico, para amostra gerada de uma Weibull com $\beta = 0,5$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

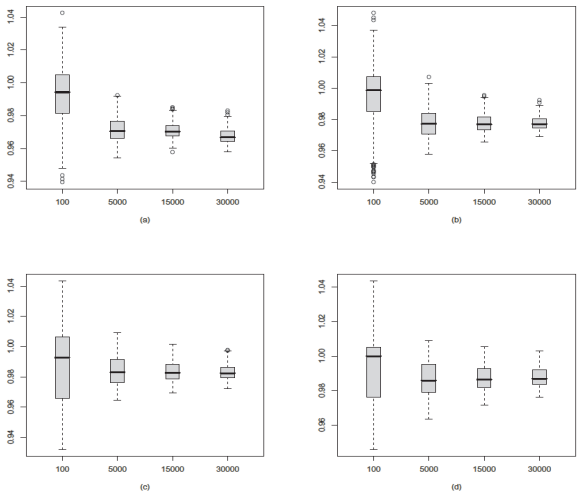

Figura 4 - Box-Plot das razões de BIC's do modelo Weibull com relação ao modelo log-logístico, para amostra gerada de uma Weibull com $\beta = 0,5$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

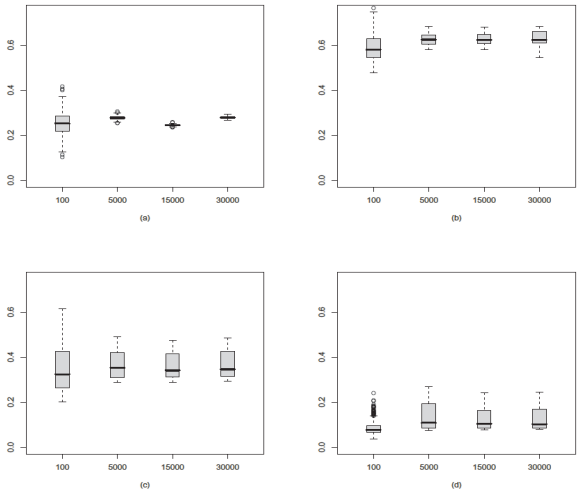

Figura 5 - Box-Plot das razões de normas Euclidianas do modelo log-logístico com relação ao modelo Weibull, para amostra gerada de uma log-logística com $1/\sigma = 2$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

Para este caso, de amostras geradas da distribuição log-logística com parâmetro $1/\sigma = 2$, na presença de longa duração, também as três métricas em questão (Norma Euclidiana, AIC e BIC) demonstraram ser boas métricas para verificar o ajuste do modelo.

Analogamente, para os outros casos estudados e demonstrados no esquema (Figura 1), verificamos resultados similares. Ou seja, em nenhum caso, houve problema quanto a identificar o modelo que gerou os dados. Como estamos interessados em analisar, modelar e finalmente selecionar modelos na presença de longa duração, podemos optar por utilizar qualquer uma das métricas apresentadas. Porém, como temos o interesse de testar e apresentar uma forma de seleção mais simples e menos usual, a Norma Euclidiana demonstrou ser uma excelente alternativa.

6 Aplicação

Nesta seção, consideramos a aplicação da metodologia descrita nas seções anteriores em um conjunto de dados reais fornecidos por uma instituição financeira brasileira. Os dados são compostos por de 65.535 cadastros de clientes, onde o interesse é observar o tempo até que o cliente deixar a instituição, isto é deixar de ter relacionamento com a empresa. Quando o cliente deixa a instituição, o seu tempo sobrevivência é dito observado. Para o caso em que o cliente é ainda ativo, o seu tempo é dito censurado.

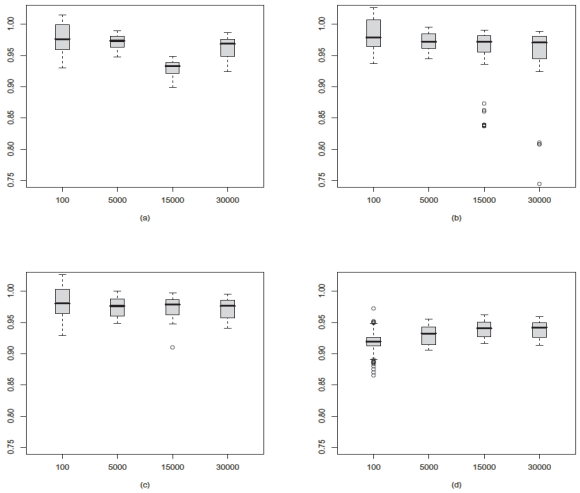

Figura 6 - Box-Plot das razões de AIC's do modelo log-logístico com relação ao modelo Weibull, para amostra gerada de uma log-logística com $1/\sigma = 2$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

Para os dados fornecidos temos a presença de 41.787 censuras, ou seja, $63,76\%$ dos clientes têm seus tempos censurados, isto é, podemos assumir que são clientes fidelizados. O tempo máximo observado no estudo foi de 201 meses e o mínimo 0 meses. Os tempos iguais a zero foram considerados clientes que não iniciaram um relacionamento com a instituição e, desta forma, foram descartados da análise (em um total de 5 clientes).

A Figura 8 (painel esquerdo) apresenta o TTT-Plot dos dados da instituição.A Figura 8 indica que a forma da função de risco é monótona crescente. Sendo assim, um possível modelo para ajuste deste dados é o modelo Weibull com parâmetro de forma maior que 1. Outra distribuição candidata para ajustar os tempos até o não pagamento dos empréstimos é a função log-logística que também acomoda esta forma de função de risco. O modelo de [3] para dados de sobrevivência na presença de longa duração, apresentado neste artigo, é então ajustado aos dados assumindo as distribuições Weibull e log-logística para os tempos de sobrevivência. Temos na Tabela 1, as estimativas dos parâmetros para estes modelos, os erros padrões das estimativas e os p-valores (representam o resultado da hipótese nula, obtido do teste bilateral baseado na estatística t-Student, medindo o quão significativo é o valor estimado para o parâmetro em questão).

Na Tabela 1, verificamos que o 63^o-percentil da distribuição, estimado através do modelo Weibull, é de aproximadamente 99 meses. Também, a função de risco tem forma crescente uma vez que, o parâmetro de forma estimado pela distribuição

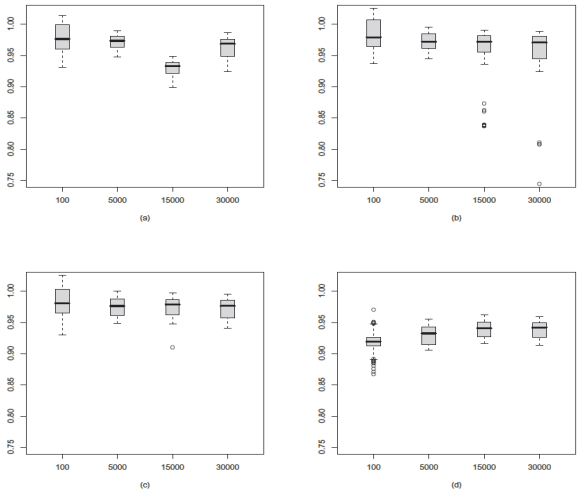

Figura 7 - Box-Plot das razões de BIC's do modelo log-logístico com relação ao modelo Weibull, para amostra gerada de uma log-logística com $1/\sigma = 2$ sendo os percentuais de censura iguais a: (a) 10%; (b) 25%; (c) 50%; (d) 75%.

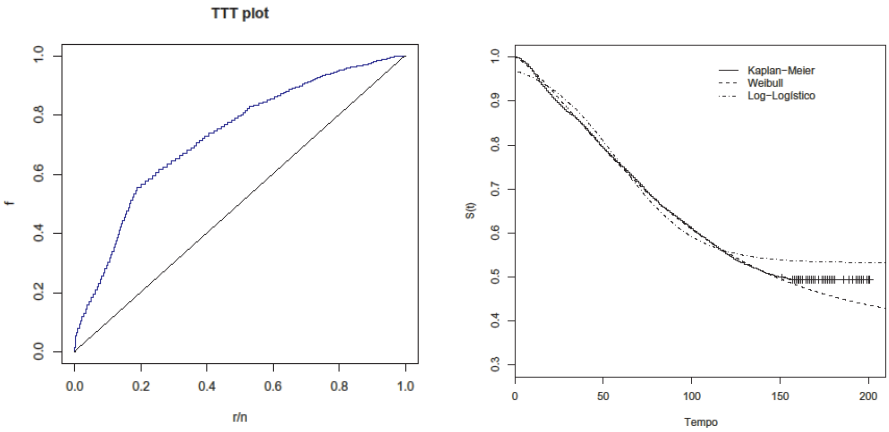

Figura 8 - Painel esquerdo: TTT-Plot para os tempos até o não pagamento dos empréstimos. Painel direito: Curva estimada via Kaplan-Meier e curvas ajustadas através dos modelos Weibull e log-logístico de longa duração.

Tabela 1 - Estimativas dos parâmetros dos modelos Weibull e log-logístico.

	Modelo Weibull de Longa-Duração			
Parâmetros	Estimativas	Erro Padrão	P-valor	Reparametrização
μ_0	$4,596283$	$0,020762$	$< 0,0001$	$99,1152$
β_0	$0,264239$	$0,008086$	$< 0,0001$	$1,3024$
γ_0	$0,465302$	$0,040781$	$< 0,0001$	$0,6143$
	Modelo Log-Logístico de Longa-Duração			
μ_0	$4,062554$	$0,005528$	$< 0,0001$	$58,1226$
σ_0	$3,083985$	$0,006743$	$< 0,0001$	$21,8453$
γ_0	$-0,129251$	$0,011177$	$< 0,0001$	$0,4677$

é de $1,30$. Para o modelo Weibull e log-logístico temos que $\hat{p} = 0,6143$ e $\hat{p} = 0,4677$, respectivamente, ou seja, temos aproximadamente $61,4\%$ e $46,8\%$ de longa duração. O ajuste destes modelos podem ser visualizados através da Figura 8 (painel direito).

Para seleção dos modelos, os métodos apresentados e verificados durante o estudo de simulação serão empregados. Na Tabela 2 temos os valores de AIC, BIC e norma Euclidiana dos dois modelos em questão. Segundo os critérios AIC e BIC, o modelo que mais se adequa aos dados é o modelo Weibull, uma vez que, para este modelo, os valores dos dois critérios são menores do que para o modelo log-logístico. Esta conclusão é também corroborada pela distância entre a curva empírica (Kaplan-Meier) e as curvas ajustadas pelos modelos Weibull e log-logístico, sendo a menor distância apresentada é para o modelo Weibull, mostrando que este é o mais adequado para os dados.

Tabela 2 - Valores de AIC e BIC e norma Euclidiana (NE), para os modelos Weibull e log-logístico.

Modelo	AIC	BIC	NE
Weibull	$300.076,164$	$300.103,436$	$0,224$
Log-Logístico	$307.777,400$	$307.750,100$	$5,084$

Ao analisar graficamente a Figura 8, visualizamos que a curva estimada pelo modelo Weibull realmente tem mais proximidade com a curva empírica do que a curva estimada pelo modelo log-logístico.

Com este exemplo conseguimos verificar na prática o como é funcional utilizar a norma Euclidiana como alternativa na seleção de modelos.

7 Comentários finais

Com o objetivo de analisar dados relativos a carteiras de clientes de financiadoras e empresas ligadas a área de finanças, estudamos os conceitos de análise de sobrevivência, aplicando-os a dados reais. Verificamos que na área financeira é comum encontrarmos duas particularidades: grandes bancos de dados

e a presença de muitas observações censuradas. Diferente do que pode ocorrer na área médica e/ou biológica, onde a maioria dos estudos contemplam amostras de tamanho pequeno e amostras com quantidades pequenas e moderadas de observações censuradas.

Ao analisar os dados, nos deparamos com problemas para verificar adequabilidade do modelo a ser escolhido. Este problema se deve ao tamanho do banco de dados em questão (65.535 observações). Desta forma, alguns modelos que são apropriados aos dados foram ajustados, não havendo problemas em se estimar os parâmetros de tais modelos (na maior parte das vezes houve a convergência das estimativas dos parâmetros dos modelos estimados), devido ao grande número de observações.

Assim, propusemos um estudo de simulação a fim de verificar qual a métrica adequada para seleção de modelos. No estudo de simulação verificamos que as três métricas em estudo (AIC, BIC e Norma Euclidiana) são adequadas para a seleção de modelos na presença de longa duração. A terceira métrica citada, Norma Euclidiana, que não é comumente usada para a finalidade de selecionar modelos, como foi empregada neste estudo, se mostra como uma boa alternativa.

GRANZOTTO, D. C. T.; LOUZADA NETO, F.; PERDONÁ, G. S. C. Survival models with long-term survivals: An Application to large databases in the financial area. *Rev. Bras. Biom.,* São Paulo, v.24, n.4, p.102-116, 2010.

▪ *ABSTRACT: Survival models with long-term survivals accommodate the heterogeneity of two populations (susceptible and immune to the event of interest). To illustrate the applicability of such models to large databases in the financial area we consider the model proposed by [3], assuming Weibull and log-logistic distributions for the lifetimes. A simulation study was performed in order to test the difference between the Kaplan-Meier curve and the fitted one as an alternative to the usual metrics of adjustment, according to different censoring percentages and sample sizes. We observed that the distance between the curves is capable to select the more appropriate model for the data in presence of long-term survivals for small and large client portfolios, even in the presence of censoring.*

▪ *KEYWORDS: Survival; financial data; long-term; criteria for model selection.*

Referências

[1] Akaike, H. Information theory as an extension of the maximum likelihood principle. *2nd International Symposium on InformationTheory*. Akademiai Kiado, Budapest, Hungary, 1973.

[2] Anderson, D. R. & Burnham, K. P. Multimodel Inference Understanding AIC and BIC in Model Selection. *Sociological Methods e Research*, v.33, n.2, 261–304, 2004.

[3] Berkson, J. & Gage, R. P. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*. v.47, p.501–515, 1952.

[4] Carlin, B. P. & Louis, T. A. *Bayes and empirical Bayes methods for data analysis*. Chapman Hall, London. 1997.

[5] Chen, M. H., Ibrahim, J. G. & Sinha, D. Bayesian survival analysis. *Springer Series in Statistics*, 2001.

[6] Efron, B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, v.7, n.1, p.1–26, 1979.

[7] Frankel, P. & Longmate, J. Parametric models for accelerated and long-term survival: a comment on proportional hazards. *Statistics in Medicine*, v.21, 2002.

[8] Lam, K. F., Fong, D. Y. & Tang, O. Y. (2005). Estimating the proportion of cured patients in a censored sample. *Journal of the Royal Statistical Society*, New York, v.24, n.12, p.1865–1879, 2005.

[9] Lawless, J. F. *Statistical models and methods for lifetime data*. John Wiley, New York, 1982.

[10] Lee, E. T. & Wang, J. W. *Statistical Methods for Survival Data Analysis*. John Wiley and Sons, New Jersey, 2003.

[11] Louzada-Neto, F. Modelagem Temporal para Credit Scoring: Uma Nova Alternativa à Modelagem Tradicional Via Análise de Sobrevivência. *Revista de Tecnologias de Crédito*, v.56, p.1–10, 2006.

[12] Maller, R. & Zhou, X. Survival Analysis with Long-Term Survivors. *Wiley Series in Probability and Statistics*. London, 1996.

[13] Paulino, C. D., Turkman, M. A. A. & Murteira, B. *Estatística Bayesiana*. Fundação Calouste Gulbenkian, Lisboa, 2003.

[14] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, v.6, p.416–464, 1978.

[15] Shao, Q. & Zhou, X. A new parametric model for survival data with long-term survivors. *Statistics in Medicine*, v.23, p.3525–3543, 2004.

[16] Weibull, W. (1951). A Statistical distribution function of wide applicability. *Journal of Applied Mechanics*, p.292–297, 1951.

Recebido em 30.03.2010.

Aprovado após revisão em 20.01.2011.