Universidade de Brasília

Instituto de Exatas

Departamento de Estatística

Análise de Sobrevivência Aplicada ao Risco de Crédito:

Ajuste de Modelos Paramétricos Contínuos a Dados de Tempo Discretos

Thiago Morais de Carvalho

Brasília

2011

Thiago Morais de Carvalho[1] **- 10/07700**

Análise de Sobrevivência Aplicada ao Risco de Crédito:

Ajuste de Modelos Paramétricos Contínuos a Dados de Tempo Discretos

Relatório apresentado à disciplina Estágio Supervisionado II do curso de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientação: *Prof.º Dr. Afrânio Márcio Corrêa Vieira*

Brasília - DF
2011

[1]thiagotouya@gmail.com

Dedicatória

À minha família:
meus pais, Maria de Lourdes e João,
meu irmão Jhonathan e minha cunhada Graciana
e aos meus sobrinhos Lucas e Isabela.

Thiago Morais de Carvalho

Agradecimentos

Agredeço primeiramente a Deus pelo dom da vida, pelo dom das possibilidades. Por ter me dado uma excelente família, grandes amigos e pela força diária.

Agradeço aos meus pais por terem, da melhor forma, me oferecido até agora uma excelente formação, tanto com relação a valores pessoais quanto à educação intelectual.

Agradeço ao meu irmão Jhonathan pelo exemplo e por mostrar que é possível mudar de situação por meio dos estudos.

Agradeço à minha grande amiga Maísa por toda a força e exemplo ao longo do curso e principalmente nos últimos dois semestres, quando da monografia.

Agradeço aos meus grandes amigos: Cristiano (Monange), João Paulo (Piru), Mastrangelo (Mastrangelo), Padre Zacarias, Rafael (Fungo), Thiago (Angel) e Wesley (Cego) por toda a força ao longo dessa vida. Valew galera!

Finalmente agradeço à todos os professores e funcionários do Departamento de Estatística da UnB, principalmente aos professores Afrânio (orientador), Eduardo Nakano e Juliana pela grandiosa ajuda na construção dessa monografia e pelo conhecimento em termos de programação em R, Análise de Sobrevivência e da área financeira que consegui apreender por meio do contato com eles.

Epígrafe

"A vida é um universo de possibilidades."
Touya

Sumário

Resumo

Neste trabalho serão analisados dados simulados para um determinado plano de financiamento, com vistas à análise de crédito sob a ótica da análise de sobrevivência, com o intuito de explorar principalmente o conceito de **fração de cura**. O objetivo principal será a utilização desse conceito, interpretado na área médica como a proporção de pacientes que respondem bem a um determinado tratamento e que se tornam imunes aos sinais e sintomas da doença, sendo assim considerados curados, [9].

Essa abordagem de fração de cura no contexto de análises financeiras é algo ainda pouco explorado, e portanto, este estudo objetiva também explorar este método para se conhecer a sua performance no contexto de uma grande base de dados.

Em nosso contexto, análise dos dados de financiamentos à pessoas físicas, a fração de cura teria como interpretação a quantidade de tomadores de empréstimos que quitaram sua dívida antes de findo o prazo e/ou que quitaram todo o plano sem entrar em inadimplência. Lembrando que destes curados, aqueles que quitaram o financiamento antes de findo o prazo, devem continuar sendo acompanhados até o tempo de truncamento, pois deve-se ter em mente um dos conceitos de fração de cura, indivíduos de longa duração.

Este tipo de informação pode ser útil às instituições financeiras, por exemplo, na hora de construir modelos de avaliação de risco de crédito. Analisar dados discretos por meio de modelos contínuos trará informações valiosas à área, tendo em vista que os tempos de falha, considerados em meses, são discretos e que dessa forma, simplificações substanciais são conseguidas nas análises por meio de modelos contínuos.

Introdução

As técnicas de análise de sobrevivência têm grande aplicação em diversas áreas como, por exemplo, em segurança pública, onde pode ser usada no acompanhamento de presos em regime de liberdade condicional, sendo a reicidiva do detento em algum tipo de crime o evento de interesse, tendo como tempo de estudo o período desses indivíduos no regime de liberdade condicional; em estudos clínicos (onde nasceram tais técnicas), por exemplo, para o acompanhamento de pacientes por um dado período de tempo para avaliar a eficácia de um certo tratamento, onde o evento de interesse reside na reicidiva da doença; em análises financeiras tendo, por exemplo, um plano de pagamentos mensais de um título de capitalização ou, como será abordado neste trabalho, tendo tais técnicas aplicadas na avaliação de pagamentos de planos de financiamento, sendo a falta de pagamento de qualquer das parcelas (inclusive no caso do título de capitalização) o evento de interesse, isto é, a falha, indicando que o cliente do plano de financiamento, a partir daquele momento, se encontra em condição de inadimplemento.

Em análise de sobrevivência, a variável resposta é, geralmente, o tempo até a ocorrência de um evento de interesse. Este tempo é denominado **tempo de falha**. Neste trabalho, a falha significará a mudança da condição de **adimplente** para a condição de **inadimplente** do indivíduo contratante do plano simulado de financiamento.

Em [7], modelos de sobrevivência para tempo discreto foram aplicados a uma base de dados de uma instituição financeira para um produto de crédito parcelado. Foi identificado que, dado o fato da maioria dos contratos serem finalizados sem atrasos superiores a 60 dias, uma considerável proporção de clientes dos planos de financiamento analisados se encaixava no conceito de fração de curados, isto é, aqueles que não falharam ao longo do período de financiamento. Assim, uma análise dos dados tendo como foco uma modelagem que leve em conta essa fração de curados se torna de grande interesse, pois se configura como algo novo no contexto de análises financeiras, visto que não existem modelos para tempo discreto na presença de fração de curados, e pode servir às instituições que oferecem planos de financiamento ou qualquer tipo de empréstimo como uma ferramenta para a geração de modelos de risco de crédito.

Sendo assim, este trabalho tem como objetivo principal a aplicação das técnicas de análise de sobrevivência na área de risco de crédito com a utilização do conceito de **fração de cura**. E como objetivos específicos, sob os dados simulados, um ajustamento por meio

de três modelos contínuos (Weibull, Gama e Log-Normal). Tendo em vista que os dados são gerados por meio de uma distribuição de probabilidade discreta, **binomial negativa**, este trabalho objetiva a avaliação dos modelos contínuos para tal ajustamento e também procura avaliar a performance destes.

Fundamentação Teórica

2.1 Crédito

Crédito significa confiança. Confiança em uma pessoa (física ou jurídica) que se compromete hoje a cumprir uma obrigação futura. As obrigações envolvendo dinheiro, por meio do crédito, agilizam as atividades econômicas, principalmente pelo fato de se poder satisfazer hoje uma necessidade ou desejo, pagando seu preço somente no futuro.

Numa instituição financeira bancária, as operações de crédito constituem o próprio negócio da empresa. Dessa forma, o banco empresta dinheiro ou financia bens aos seus clientes, funcionando com um 'intermediário financeiro', o que de fato é, pois os recursos obtidos por meio da captação através dos depósitos, por exemplo, é que são usados para o empréstimo via crédito a outros clientes.

A concessão de crédito num banco consiste em emprestar dinheiro, isto é, colocar a disposição do cliente determinado valor em determinado momento, mediante promessa de pagamento futuro. A taxa de juros aplicada à operação de crédito é a retribuição por essa prestação de serviço cujo recebimento poderá ser antecipado, periódico ou mesmo ao final do período, juntamente com o principal emprestado.

O crédito pode fazer com que empresas aumentem seu nível de atividade, pode estimular o consumo influenciando na demanda, pode cumprir uma função social ajudando as pessoas a obterem moradia, bens e até alimentos. A tudo isso, por outro lado, deve-se acrescentar que o crédito pode tornar empresas e pessoas físicas altamente endividadas, assim como pode ser parte componente de um processo inflacionário e também gerar o fenômeno da inadimplência, onde o tomador não honra o compromisso assumido com a instituição financeira.

Portanto, às instituições bancárias cabe a difícil decisão de conceder ou não o crédito aos tomadores. Essa é uma decisão onde a incerteza sempre estará presente. Conforme [10], neste tipo de evento sempre haverá, por parte das instituições financeiras, a possibilidade de perda. O ideal seria quantificar essa possibilidade de perda em probabilidade, permitindo assim uma melhor decisão sobre a concessão do crédito. Essa probabilidade de perda é também conhecida como 'risco de crédito'.

2.2 Risco de Crédito

Como vimos anteriormente,

RISCO DE CRÉDITO ⇒ PROBABILIDADE DE PERDA

É importante frisar que a estimativa dessa probabilidade de perda é função das características do solicitante (tomador).

O risco de uma solicitação de crédito pode ser avaliado de forma subjetiva ou medido de forma objetiva utilizando metodologia quantitativa. A avaliação subjetiva, apesar de incorporar a experiência do analista, não quantifica a risco de crédito. Dizer que uma empresa é de alto risco não é suficiente para estimar de maneira precisa as perdas ou ganhos esperados com a operação e, consequentemente, tomar a decisão mais adequada.

Segundo [10], medir o risco de crédito de forma objetiva, utilizando técnicas quantitativas apresenta como vantagens, por exemplo, consistência nas decisões, decisões rápidas, decisões adequadas, dentre outras. Além de também permitir a verificação do grau com que a instituição atende aos requesitos dos órgãos reguladores, de permitir o estabelecimento de uma linguagem comum entre os decisores de crédito e de permitir a definição de níveis de alçada para a concessão do crédito.

Metodologia

3.1 Análise de Sobrevivência

Análise de Sobrevivência pode ser caracterizada por um conjunto de técnicas estatísticas que têm como objetivo principal a análise de tempos até a ocorrência de um determinado evento de interesse, onde as observações são acompanhadas ao longo de períodos de tempo. Estamos falando da aplicação de tais técnicas em estudos do tipo longitudinal, onde as mesmas unidades observacionais são analisadas, tendo as características de interesse sendo medidas periodicamente para avaliação do evento de interesse, a falha.

3.1.1 Censura

Um conceito de extrema importância em análise de sobrevivência é a **censura**. Chamamos dados censurados àqueles onde se tem apenas a observação parcial da resposta, isto é, apenas uma observação parcial do tempo de falha. Esse tipo de dado pode ser gerado por uma infinidade de circunstâncias, como por exemplo, num estudo da área de saúde, a saída do paciente do estudo por algum motivo diferente do evento de interesse como, por exemplo, a mudança de residência inviabilizando sua participação no estudo ou a morte do paciente por alguma razão diferente da esperada. Dessa forma, o que se tem é apenas a informação de que o tempo de falha daquele paciente é maior do que o tempo observado, isto é, o tempo de censura. Alguém poderia pensar em simplesmente retirar esse tipo de dado da amostra para se fazer uso das técnicas clássicas de análise estatística como, por exemplo, análise de regressão e planejamento de experimentos, mas sem dúvidas, agindo dessa forma, o estudo ficaria viciado e incompleto, haja vista que mesmo sendo dados incompletos, estes fornecem informações sobre o tempo de vida das unidades observacionais.

Existem alguns tipos de censura, tais como **censura à esquerda**, **censura intervalar** e **censura à direita**. Neste trabalho daremos ênfase à censura à direita, que diz respeito à observação parcial da resposta quando o tempo de ocorrência do evento de interesse está à direita do tempo registrado. Esse tipo de censura é caracterizado por outros três tipos: censura do tipo 1, censura do tipo 2 e censura do tipo aleatória.

A censura do tipo 1 consiste num estudo onde há uma determinação temporal para seu

fim e onde se poderá verificar a presença de indivíduos censurados ao final, isto é, depois de passado o período determinado para a consecução do estudo ocorre a verificação de indivíduos que não chegaram a experimentar o evento de interesse.

A censura do tipo 2 é definida, num estudo, como aqueles indivíduos que não observaram o evento do interesse após um número k de observações terem falhado, ou seja, o fim do experimento foi condicionado a observação da falha em num número pre-estabelecido de indivíduos e os indivíduos que não falharam são ditos censurados pelo tipo 2 de censura.

A censura aleatória ocorre quando, num estudo, se observa que indivíduos deixam de fazer parte da pesquisa por quaisquer outros motivos que não a observação do evento de interesse, isto é, antes do experimento acabar alguns indivíduos deixam de participar do estudo, restando ao pesquisador apenas a certeza de que o tempo de falha deste indivíduo é maior do que o tempo registrado pois ele deixa o experimento por razões diferentes das que estão sob estudo.

3.1.2 Estimador de Kaplan-Meier

Uma maneira de interpretar tempos de sobrevivência e de se obter estimativas não-paramétricas é por meio do gráfico da função de sobrevivência empírica, que na ausência de censuras, numa amostra de tamanho **n**, pode ser estimada através da seguinte expressão:

$$\hat{S}(t) = \frac{\#\ de\ obs\ \ \geq t}{n}, \qquad t \geq 0. \tag{3.1}$$

Expressão esta de uma função degrau onde cada degrau tem tamanho igual a $1/n$, numa amostra cujos tempos observados não se repetem. No caso de verificada a repetição de valores dos tempos, os degrais no gráfico desta função terão tamanho r/n, onde r se refere a quantidade de vezes que um determinado tempo se repetiu na amostra.

Na presença de dados censurados (uma observação censurada informa que o tempo até a falha é maior do que aquele que foi registrado, [2]), algumas modificações devem ser feitas em (3.1) para que esta nova informação seja acomodada na estimação da função de sobrevivência. Kaplan & Meier, em 1958, desenvolveram um método que considera a presença de censura nos dados. Esta estimativa não-paramétrica é também conhecida como estimador produto-limite e se resume numa adaptação da função de sobrevivência empírica, que considera (como descrito em [2]) tantos intervalos de tempo quantos forem o número de falhas distintas. Tem-se também que os limites dos intervalos de tempo são os tempos de falha da amostra.

Considere uma amostra aleatória de tempos de vida onde existam k, $(k < n)$ tempos distintos, $t_1 < t_2 < \cdots < t_k$ nos quais o evento de interesse ocorre. Dessa forma, tem-se que é natural a existência de mais de um evento de interesse num dado tempo t_j. Considere também d_j como sendo o número de falhas no tempo t_j. De posse dessas definições podemos expressar o estimador produto-limite de Kaplan-Meier como:

$$\hat{S}(t) = \prod_{j:t_j<t} \frac{n_j - d_j}{n_j}, \tag{3.2}$$

em que n_j é o número de indivíduos sob risco em t_j.

Este estimador apresenta-se como não-viciado, para grandes amostras; fracamente consistente e como estimador de máxima verossimilhança para a função de sobrevivência $S(t)$. A estimativa não muda nos tempos censurados, o efeito dos tempos censurados é, entretanto, sentido nos valores de n_j, portanto, nos tamanhos dos degrais em $\hat{S}(t)$.

3.1.3 Fração de Cura

No contexto das diversas aplicações das técnicas estatísticas de análise de sobrevivência surge um conceito muito valioso informativamente, denominado fração de curados ou indivíduos de longa duração, que consiste daqueles indivíduos (ou objetos de estudo, a depender da área de aplicação) que não sofrerão o evento de interesse, ou seja, que não falharão ao longo do estudo. Para explicar melhor este conceito, voltemos aos exemplos de aplicação mostrados na introdução:

- para o caso da segurança pública, onde estão sendo avaliados detentos em regime de liberdade condicional, a fração de curados consistirá na parcela dos indivíduos daquela população que não irá experimentar o evento de interesse, isto é, a parcela de prisineiros que não incorrerá novamente em algum crime durante o período de semi-liberdade, e espera-se, após este;

- no caso do acompanhamento de pacientes para a avaliação de um tratamento, os indivíduos de longa duração consistirão da parcela da população em estudo que responderá positivamente ao tratamento, não voltando a apresentar a doença até o fim do estudo, ou melhor, não voltando a apresentar a doença por um período suficientemente grande de tempo;

- na área financeira, mais especificamente, no estudo de planos de financiamento, onde a falha consiste na falta de pagamento de pelo menos uma parcela, a fração de curados se mostra como a fração de indivíduos, do total de contratantes do plano, que não falhará, pagando sempre corretamente as parcelas do plano ou quitando-o antes do tempo previsto para o fim das parcelas, isto é, se caracterizando como indivíduos que não mais poderão falhar com relação àquela operação.

A importância do conceito de indivíduos de longa duração é muito grande. Para ilustrar essa relevância, tome este último exemplo, da área financeira, onde estão sendo estudados vários planos de financiamento e o evento de interesse se resume na falta de

pagamento de pelo menos uma parcela por pelo menos 61 dias. Nesse caso, como já fora falado, a fração de curados consistirá dos indivíduos do estudo que se manterão adimplentes até o fim do estudo, e sendo assim, o conceito em voga servirá à instituição financeira fornecendo informações a respeito do sucesso ou fracasso daquele modelo de financiamento, àquele público-alvo, ou seja, as informações obtidas por meio da análise de sobrevivência com o uso do conceito de fração de cura poderão subsidiar a modelagem dos produtos de crédito afim de minimizar as perdas com inadimplencia e aumetar a eficiência do processo de concessão de crédito consequentemente gerando maiores receitas à instituição, dentre outros ganhos.

Abaixo é mostrado um gráfico onde se verifica a fração de cura. O gráfico consiste na função de sobrevivência estimada pelo método de Kaplan-Meier, também conhecido como estimador **produto-limite**. Este exemplo foi tirado de [2]. Consiste de dados de um estudo para investigar o efeito da terapia com esteróide no tratamento da hepatite viral aguda.

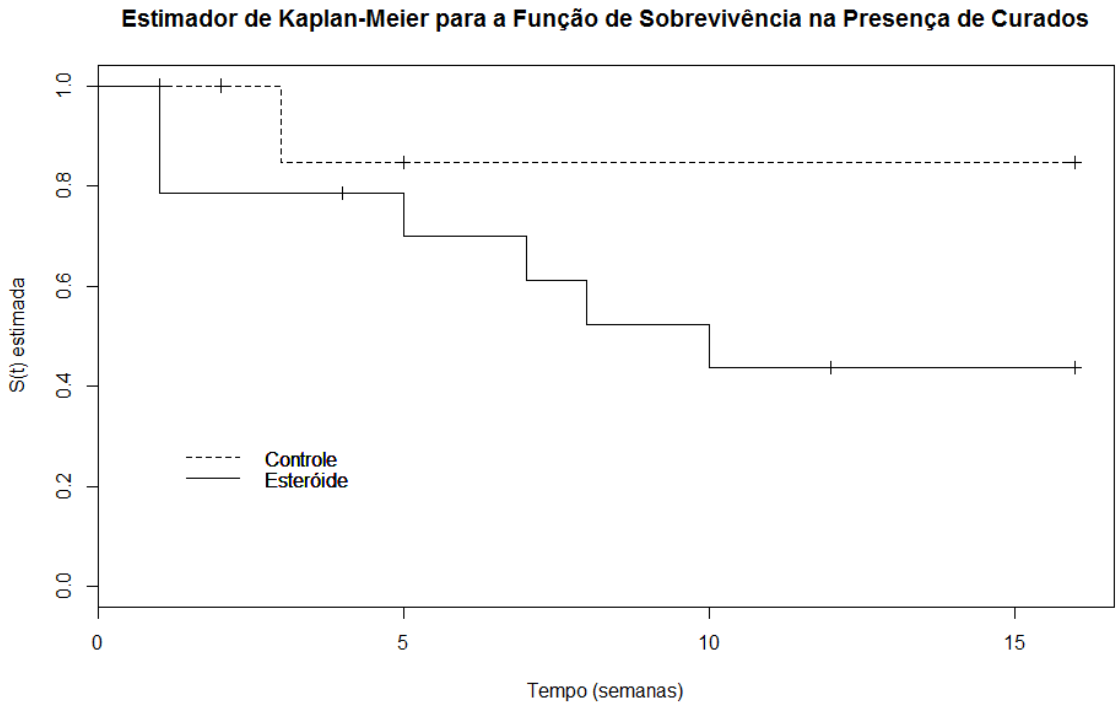


Figura 3.1: Estimativas de Kaplan-Meier para os grupos controle e esteróide dos 29 dados de hepatite. Os tempos representados por + mostram onde ocorreram censuras em cada grupo.

A fração de curados numa população pode ser analisada por meio de duas abordagens. A primeira, segundo Berkson e Gage (1952), consiste na modelagem da fração de cura considerando a função de sobrevivência populacional construída na forma de mistura e é conhecida como função de sobrevivência imprópria [3]. A segunda, proposta por Yakovlev e Tsodikov (1996) e Chen (1999), considera tempos de sobrevivência infinitos para os

indivíduos curados, permitindo assim que os tempos de sobrevivência de pacientes curados e não curados possam ser expressos em uma única fórmula [11]. Esta segunda abordagem consiste numa classe de modelos de mistura com estrutura de riscos competitivos [9]. Neste trabalho, utilizaremos apenas a primeira abordagem.

3.1.4 Modelo de Mistura

Berkson e Gage (1952) propuseram um modelo onde a população em estudo deve ser dividida em duas subpopulações bem definidas: uma constituída de indivíduos não-suscetíveis à falha, e a outra formada por indivíduos sob risco ao longo do período de estudo.

Este modelo consiste de uma função de sobrevivência populacional imprópria $S_{pop}(t)$, isto é, num gráfico da função de sobrevivência empírica pelo tempo (estimador produto-limite de Kaplan-Meier) cuja cauda da função tende para um valor diferente de zero ao longo de um tempo suficientemente grande, há evidências da existência de uma possível parcela de curados da população.

A modelagem consiste numa função de sobrevivência própria $S(t)$ com probabilidade $(1-\phi)$ para a parcela da população cujos indivíduos se encontram sob risco; e para a outra parte da população, correspondente aos indivíduos curados, associa-se uma probabilidade ϕ, $\phi \in (0,1)$, e somente ela, visto que o tempo de falha para estes é suposto infinito implicando numa função de sobrevivência igual a 1, para todo tempo t. Dessa forma, tem-se o seguinte modelo de sobrevivência com fração de cura:

$$S_{pop}(t) = \phi + (1-\phi)S(t), \tag{3.3}$$

com as seguintes propriedades:

$$\lim_{t\to\infty} S_{pop}(t) = \phi$$

e

$$\lim_{t\to 0} S_{pop}(t) = 1.$$

Com relação às probabilidades supracitadas, suponha uma variável binária C_i, determinando a condição de cada observação. Tem-se $C_i = 0$ para os não-suscetíveis, e $C_i = 1$ para o i-ésimo indivíduo sob risco. Dessa forma, define-se $P(C_i = 0) = \phi$, a probabilidade de um indivíduo ser não-suscetível e $P(C_i = 1) = 1 - \phi$, a probabilidade do i-ésimo indivíduo estar sob risco.

3.2 Modelagem

3.2.1 Weibull

O Modelo Probabilístico

A distribuição Weibull se tornou muito popular por apresentar uma grande variedade de formas e todas com a importante propriedade de que sua função taxa de falha é monótona, isto é, ela é crescente, decrescente ou constante.

Uma variável aleatória T que segue esta distribuição tem a seguinte função densidade de probabilidade:

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\}, \quad t \geq 0, \tag{3.4}$$

em que γ, o parâmetro de forma, e α, o parâmetro de escala, são ambos positivos. O parâmetro α tem a mesma unidade de medida de t e γ não tem unidade de medida.

As funções de sobrevivência e de taxa de falha são, respectivamente

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\} \tag{3.5}$$

e

$$\lambda(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1}, \tag{3.6}$$

para $t \geq 0$, α e γ maiores que zero.

Como características da função taxa de falha $\lambda(t)$, temos que ela é estritamente crescente para $\gamma > 1$, estritamente decrescente para $\gamma < 1$ e constante para $\gamma = 1$.

As expressões da média e variância da Weibull são mostradas abaixo.

$$E(T) = \alpha\Gamma[1 + (1/\gamma)] \tag{3.7}$$

e

$$Var(T) = \alpha^2[\Gamma[1 + (2/\gamma)] - \Gamma[1 + (1/\gamma)]^2], \tag{3.8}$$

sendo

$$\Gamma(\beta) = \int_0^\infty x^{\beta-1} \exp\{-x\} dx. \tag{3.9}$$

3.2.2 Gama

O Modelo Probabilístico

A função densidade da distribuição gama, caracterizada por dois parâmetros, β e α, com $\beta > 0$, o parâmetro de forma e $\alpha > 0$, o de escala, tem a seguinte expressão:

$$f(t) = \frac{1}{\Gamma(\beta)\alpha^{\beta}} t^{\beta-1} \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}, \quad t > 0. \tag{3.10}$$

As respectivas funções de sobrevivência e de risco desta distribuição são dadas por:

$$S(t) = \int_t^{\infty} \frac{1}{\Gamma(k)\alpha^{k}} u^{k-1} \exp\left\{-\left(\frac{u}{\alpha}\right)\right\} du \tag{3.11}$$

e

$$\lambda(t) = \frac{f(t)}{S(t)}. \tag{3.12}$$

A distribuição Gama apresenta algumas particularidades, tais como:

- Para valores de β maiores do que 1, sua densidade apresenta um único pico em $t = (k-1)/\alpha$;
- A função taxa de falha apresenta padrão crescente ou decrescente convergindo para um valor constante quando t cresce de 0 a infinito;
- Para $\beta = 1$ tem-se a distribuição exponencial, um caso particular da distribuição Gama.

3.2.3 Log-Normal

O Modelo Probabilístico

A função de densidade de uma variável aleatória T com distribuição log-normal é dada por:

$$f(t) = \frac{1}{\sqrt{2\pi t\sigma}} \exp\left\{-\frac{1}{2}\left(\frac{log(t)-\mu}{\sigma}\right)^2\right\}, \quad t > 0 \tag{3.13}$$

em que μ é a média do logaritmo do tempo de falha, e σ é o desvio-padrão.

As funções de sobrevivência e de taxa de falha de uma variável log-normal não apresentam uma forma analítica explícita, como verificado em [2], e são, desse modo, representadas, respectivamente, por:

$$S(t) = \Phi\left(\frac{-log(t)+\mu}{\sigma}\right) \tag{3.14}$$

e

$$\lambda(t) = \frac{f(t)}{S(t)}, \tag{3.15}$$

em que Φ é a função de distribuição acumulada de uma normal-padrão.

É importante notar que a função taxa de falha da log-normal, diferentemente do que ocorre na Weibull, não é monótona. Seu comportamento é basicamente composto da seguinte forma: no início, crescente, até atingir um valor máximo e depois decrescente, portanto, unimodal.

3.2.4 Estimação dos Parâmetros do Modelo

O método de estimação utilizado neste trabalho foi o método de **máxima verossimilhança** com restrição nos parâmetros, pois este consegue incorporar os dados referentes às censuras no processo de estimação e apresenta propriedades ótimas para grandes amostras.

O método da máxima verossimilhança, conforme [2], trata o problema da estimação dos parâmetros se baseando nos resultados obtidos pela amostra para a definição da distribuição, entre as várias possíveis, com maior possibilidade de ter gerado a amostra em estudo.

O que ocorre neste método é que a função de verossimilhança, $L(\theta)$, informa que a contribuição de cada observação não-censurada é dada pela sua função densidade de probabilidade e que a contribuição das observações censuradas é dada por sua função de sobrevivência, pois estas observações censuradas informam apenas que seus tempos são maiores do que os tempos de censura observado.

De [2], tem-se que uma expressão para a função de verossimilhança é dada por

$$L(\theta) \propto \prod_{i=1}^{n} [f_{pop}(t_i;\theta)]^{\delta_i} [S_{pop}(t_i;\theta)]^{1-\delta_i} \tag{3.16}$$

onde δ_i é o vetor censura dos dados e θ é o vetor dos parâmetros a serem estimados.

Verossimilhança Considerando a Fração de Cura para a Weibull

As funções de densidade e de sobrevivência populacionais para este modelo são dadas pelas seguintes expressões:

$$\begin{aligned} f_{pop}(t) &= (1-\phi) f(t) \\ &= (1-\phi) \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \exp\left\{ -\left(\frac{t}{\alpha}\right)^{\gamma} \right\} \end{aligned} \tag{3.17}$$

e

$$\begin{aligned} S_{pop} &= \phi + (1-\phi)S(t) \\ &= \phi + (1-\phi)\exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\} \end{aligned} \tag{3.18}$$

De posse delas pode-se escrever a verossimilhança do modelo da seguinte forma:

$$L(\theta) = \prod_{i=1}^{n}\left[(1-\phi)\frac{\gamma}{\alpha^{\gamma}}t^{\gamma-1}\exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}\right]^{\delta_i}\left[\phi + (1-\phi)\exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}\right]^{1-\delta_i} \tag{3.19}$$

Verossimilhança Considerando a Fração de Cura para a Gama

As funções de densidade e de sobrevivência populacionais para este modelo são dadas pelas seguintes expressões:

$$\begin{aligned} f_{pop}(t) &= (1-\phi)f(t) \\ &= (1-\phi)\frac{1}{\Gamma(\beta)\alpha^{\beta}}t^{\beta-1}\exp\left\{-\left(\frac{t}{\alpha}\right)\right\} \end{aligned} \tag{3.20}$$

e

$$\begin{aligned} S_{pop} &= \phi + (1-\phi)S(t) \\ &= \phi + (1-\phi)\int_{t}^{\infty}\frac{1}{\Gamma(k)\alpha^{k}}u^{k-1}\exp\left\{-\left(\frac{u}{\alpha}\right)\right\}du \end{aligned} \tag{3.21}$$

De posse delas pode-se escrever a verossimilhança do modelo da seguinte forma:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n}\left[(1-\phi)\frac{1}{\Gamma(\beta)\alpha^{\beta}}t^{\beta-1}\exp\left\{-\left(\frac{t}{\alpha}\right)\right\}\right]^{\delta_i}\text{x} \\ &\left[\phi + (1-\phi)\int_{t}^{\infty}\frac{1}{\Gamma(k)\alpha^{k}}u^{k-1}\exp\left\{-\left(\frac{u}{\alpha}\right)\right\}du\right]^{1-\delta_i} \end{aligned} \tag{3.22}$$

Verossimilhança Considerando a Fração de Cura para a Log-Normal

As funções de densidade e de sobrevivência populacionais para este modelo são dadas pelas seguintes expressões:

$$\begin{aligned} f_{pop}(t) &= (1-\phi)f(t) \\ &= (1-\phi)\frac{1}{\sqrt{2\pi t\sigma}}\exp\left\{-\frac{1}{2}\left(\frac{log(t)-\mu}{\sigma}\right)^2\right\} \end{aligned} \tag{3.23}$$

e

$$\begin{aligned} S_{pop} &= \phi + (1-\phi)S(t) \\ &= \phi + (1-\phi)\Phi\left(\frac{-log(t)+\mu}{\sigma}\right) \end{aligned} \tag{3.24}$$

De posse delas pode-se escrever a verossimilhança do modelo da seguinte forma:

$$\begin{aligned} L(\theta) = \prod_{i=1}^{n}\left[(1-\phi)\frac{1}{\sqrt{2\pi t\sigma}}\exp\left\{-\frac{1}{2}\left(\frac{log(t)-\mu}{\sigma}\right)^2\right\}\right]^{\delta_i} \text{x} \\ \left[\phi + (1-\phi)\Phi\left(\frac{-log(t)+\mu}{\sigma}\right)\right]^{1-\delta_i} \end{aligned} \tag{3.25}$$

3.2.5 Kolmogorov-Smirnov

O teste Kolmogorov-Smirnov como pode ser visto em [4], KS, é usado para decidir sobre a igualdade de duas populações com relação a uma distribuição específica. Este teste observa a máxima diferença absoluta entre a função de distribuição acumulada assumida para os dados, F(X), e a função de distribuição empírica dos dados, G(X).

Nesse teste os dados consistem em **n** observações independentes de uma variável aleatória X, associadas a alguma função de distribuição desconhecida F(X).

Hipóteses a serem testadas

Seja $F^*(X)$ uma função de distribuição completamente especificada. A hipótese geral do teste é

$H_0 : F(X) = F^*(X)$: Os dados seguem a distribuição especificada
$H_1 : F(X) \neq F^*(X)$: Os dados não seguem a distribuição especificada

Essa hipótese geral pode se tornar mais refinada, gerando as seguintes outras hipóteses a serem testadas

1. $H_0 : F(X) \geq F^*(X)$
   $H_1 : F(X) < F^*(X)$

2. $H_0 : F(X) \leq F^*(X)$
   $H_1 : F(X) > F^*(X)$

Estatística do teste

Seja S(X) a distribuição empírica baseada na amostra aleatória. A estatística do teste compara $G(X)$ com $F^*(X)$ e é definida diferentemente para os três conjuntos de hipóteses (Geral, 1 e 2).

- Para a hipótese geral: **teste bilateral:** Seja KS a maior distância vertical, em valor absoluto, entre $G(X)$ e $F^*(X)$. Denotada por

$$KS = \sup_X |F^*(X) - G(X)|$$

- Para a hipótese 1: **teste unilateral:** Seja KS^+ a maior distância vertical de $F^*(X)$ sobre $G(X)$. Denotada por

$$KS^+ = \sup_X [F^*(X) - G(X)]$$

- Para a hipótese 2: **teste unilateral:** Seja KS^- a maior distância vertical de $G(X)$ sobre $F^*(X)$. Denotada por

$$KS^- = \sup_X [G(X) - F^*(X)]$$

Regra de Decisão

Rejeita-se H_0 ao nível α de significância, se KS excede o quantil $1 - \alpha$.

3.3 Simulação dos Dados

Os dados serão simulados via R, *software* estatístico, usando como distribuição de probabilidade a Binomial Negativa com parâmetros n e p, que apresenta a seguinte função de probabilidade:

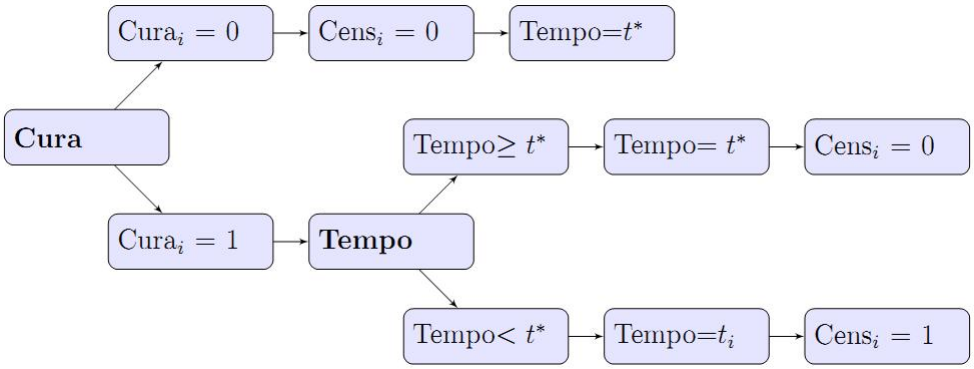
$$P(X = n) = \binom{x + n - 1}{x} p^n (1 - p)^x, \quad sendo \quad n = 1, 2, 3, \cdots \quad e \quad x = 0, 1, 2, \cdots \tag{3.26}$$

A escolha desta distribuição para a geração do banco de dados partiu da necessidade de se gerar tempos de falha que se conportassem de forma parecida com a realidade de um plano de financiamento, isto é, poucas observações do evento de interesse no início do plano e poucas no fim, tendo como principal período de grandes falhas observadas a parte central do tempo do financiamento. Outro ponto importante foi a necessidade de geração dos dados por meio de uma distribuição de probabilidade discreta, tendo em vista as características da variável que precisava ser gerada.

Aplicação

4.1 Banco de Dados

O banco de dados utilizado neste trabalho foi simulado via software estatístico R. Para a simulação utilizou-se como base o seguinte algoritmo:



O primeiro ponto observado na construção do banco de dados foi o vetor **cura** (onde a cura é sinalizada através do valor **0**. O valor **1** indica indivíduos suscetíveis ao evento de interesse), pois quando este tem valor **0**, necessariamente o vetor **censura** (cujo valor **0** indica censura e o valor **1**, a não-censura) também tem valor **0** e o vetor **tempo**, dos tempos gerados por meio de uma binomial negativa, tem valor igual a $\mathbf{t}^*$, um valor previamente definido como o **tempo-limite**, tempo que delimita a janela de observação dos dados. Se o vetor **cura** tem valor **1**, um tempo é gerado no vetor **tempo**, e se este tempo é menor que o tempo-limite, então o vetor **tempo** o recebe e ao mesmo tempo é atribuído ao vetor **censura** o valor **1**, indicando que aquele tempo foi observado. Caso contrário, isto é, quando o valor gerado no vetor **tempo** é maior que $\mathbf{t}^*$, tem-se que o valor recebido pelo vetor é o tempo-limite e ao mesmo tempo é atribuído ao vetor **censura** o valor **0**, indicando que aquele indivíduo é censurado.

Para a geração do vetor **cura** foi utilizada a distribuição de probabilidade de Bernoulli de parâmetro p, onde esse parâmetro indicava a proporção de não-curados, haja vista que

nessa distribuição o valor indicativo de sucesso é 1, o que no caso do vetor **cura** indica a suscetibilidade. Para o vetor **tempo** utilizou-se a distribuição binomial negativa pois a intenção era gerar dados que se comportassem de forma semelhante à realidade de um plano de financiamentos, isto é, poucas pessoas entrando em inadimplência no início e ao fim do plano e uma concentração de indivíduos experimentando o evento de interesse na parte mais central do período. Ao vetor **censura** foram atribuídos valores (**0** ou **1**) de acordo com a combinação de valores dos vetores **cura** e **tempo**, ou seja, se o vetor **cura** tinha valor **0**, necessariamente e independente do valor de **tempo**, o vetor **censura** recebia o valor **0**, mas se o valor do vetor indicativo de cura era **1**, o vetor **tempo** é que determinaria o valor da censura, ou seja, se o tempo em **tempo** fosse menor que o tempo-limite, a observação não seria censurada, caso contrário, vetor **censura**=0.

Com relação ao tempo-limite da janela de observação, o significado de sua existência reside no fato de que em planos de financiamentos médios ou longos o ato de se observar todo o período para depois se proceder à construção de um modelo, se torna inviável, primeiro pelo tempo de espera para a obtenção dos dados, e também porque nos planos que se quer avaliar os créditos conceditos, os dados devem ser suficientemente recentes para conterem as características dos clientes atuais e, ao mesmo tempo, suficientemente antigos para que se possa observar a performance/comportamento desses clientes.

A amostra utilizada neste trabalho consistiu de 10 mil observações. Foram feitos os ajustes dos dados nos três modelos (Weibull, Gama e Log-Normal) em 3 cenários diferentes, para planos de financiamento de 15 e 36 meses. O que mudava dentro de cada plano de financiamento era a proporção de curados e censurados. Foram analisados cenários com 5%, 15% e 35% de cura com respectivamente 25%, 25% e 15% de censura (censura unicamente, haja vista que os indivíduos curados também recebem classificação de censurados). As proporções de curados, nos cenários estudados, foram definidas de acordo com o parâmetro da distribuição Bernoulli. Já as censuras, em sua totalidade, foram definidas tendo como base o parâmetro de média da distribuição binomial negativa.

A tabela mostrada abaixo, exibe os valores dos parâmetros, tanto da distribuição binomial negativa quanto da distribuição de Bernoulli, usados na simulação do banco de dados, lembrando que o parâmetro **p** se refere à distribuição de Bernoulli e que os parâmetros **size** e **média**, se referem à distribuição binomial negativa.

Cura	p	t^*	Size	Média
5%	0.95	12	15	9.3
15%	0.85	12	15	9.7
35%	0.65	12	15	8.97

Tabela 4.1: Tabela para o plano de 15 meses com os valores dos parâmetros usados na geração dos dados.

Cura	p	t^*	Size	Média
5%	0.95	18	36	14.1
15%	0.85	18	36	12.5
35%	0.65	18	36	14.5

Tabela 4.2: Tabela para o plano de 36 meses com os valores dos parâmetros usados na geração dos dados.

Abaixo é mostrado um exemplo de um dos bancos de dados gerados neste trabalho. É composto pelos valores dos vetores **cura**, **censura**, **tempo** e **estado da observação**.

Obs	Cura	Censura	Tempo	Estado
9982	1	1	8	Falha
9983	1	0	12	Censura
9984	1	1	6	Falha
9985	1	1	9	Falha
9986	1	1	11	Falha
9987	1	0	12	Censura
9988	0	0	12	Cura
9989	1	0	12	Censura
9990	1	1	11	Falha
9991	0	0	12	Cura

Tabela 4.3: Parte de um dos bancos de dados criados neste estudo

Resultados

5.1 Estimativas dos Parâmetros

Neste trabalho, os parâmetros dos modelos foram estimados via software estatístico R (versão 2.14.0) através do método de máxima verossimilhança. As expressões das verossimilhanças, considerando a fração de cura (modelagem de Berkson & Gage) usadas para cada modelo são mostradas na seção 3.2.4 (Estimação dos parâmetros do modelo).

Abaixo são mostradas tabelas com as estimativas para cada parâmetro em cada cenário para cada um dos três modelos. Os parâmetros estimados são o de forma γ, o parâmetro de escala α e a proporção de indivíduos curados ϕ.

Tabela 5.1: Estimativas dos parâmetros de forma (γ), escala (α) e fração de cura (ϕ) para os cenários do plano de 15 meses.

Prazo	%de Cura	Distribuição	Forma (γ)	Escala (α)	Cura (ϕ)
		$Weibull$	3.1337	8.8324	0.2469
15	5%	$Gama$	4.6510	0.4910	0.0729
		$LogNormal$	2.2169	0.5591	0.00000001
		$Weibull$	3.2454	9.0886	0.3503
15	15%	$Gama$	4.7160	0.4700	0.1623
		$LogNormal$	2.3550	0.5920	0.00000001
		$Weibull$	3.0564	8.6977	0.4690
15	35%	$Gama$	4.5577	0.4944	0.3587
		$LogNormal$	2.4272	0.6734	0.07445

Algo interessante a se observar nessa primeira tabela diz respeito primeiramente às estimativas do parâmetro de cura. Com relação ao modelo Weibull é notável que as estimativas para o parâmetro ϕ se aproximam muito dos valores de censura para cada cenário, haja vista que para 5% de cura temos 30% de censura, para 15% de cura temos 40% de censura e para 35% de cura temos 50% de censura. Já para o modelo gama, temos o que se parece com a estimativa da proporção de cura propriamente dita, e não da cura com a censura, como visto para a Weibull. Verificamos isso observando a proximidade das estimativas de ϕ com as porcentagens de cura estabelecidas em cada cenário. Já o modelo lognormal para o plano de 15 meses, especialmente nos cenários de 5% e 15% de

cura não apresenta um valor significativo para o parâmetro ϕ, e para o cenário com 35% de cura ele reconhece uma pequena proporção de curados, mas muito aquém do valor estabelecido. Este fato pode ter como razão dois fatores: o plano de pagamentos é muito curto, e consequentemente, a janela de observação é muito pequena, e nos dois primeiros cenários a fração de curados é muito baixa.

Nesta segunda tabela, mostrada abaixo, de certa forma, confirmamos o que foi dito sobre os modelos na tabela anterior. O comportamento da estimativa do parâmetro de cura continua tendo a mesma caracterização para a Weibull, isto é, estimando algo proximo do valor total de censura. Para a gama acontece o mesmo da primeira tabela, ela estima os valores de ϕ bem próximos da fração de cura considerada. Com relação à lognormal verificamos que o que foi dito anteriormente sobre o parâmetro ϕ parece ter coerência haja vista que agora o plano considerado é de 36 meses e a janela de observação tem 6 meses a mais do que o verificado no plano de 15 meses. Com isso, foi possível encontrar valores mais razoáveis para a cura estimada e também foi verificado que para proporções muito baixas de cura a lognormal não mostra boas estimativas.

Tabela 5.2: Estimativas dos parâmetros de forma (γ), escala (α) e fração de cura (ϕ) para os cenários do plano de 36 meses.

Prazo	%de Cura	Distribuição	Forma (γ)	Escala (α)	Cura (ϕ)
		$Weibull$	4.6664	14.2984	0.2648
36	5%	$Gama$	9.6733	0.6601	0.1065
		$LogNormal$	2.7009	0.3789	0.00000001
		$Weibull$	4.7127	14.5073	0.3616
36	15%	$Gama$	9.5250	0.6316	0.2008
		$LogNormal$	2.7883	0.4090	0.0007
		$Weibull$	4.4521	14.2549	0.4775
36	35%	$Gama$	9.1372	0.6273	0.3698
		$LogNormal$	2.7326	0.4087	0.2442

5.2 Ajuste do Modelo

Nesta seção são apresentadas as tabelas do teste de Kolmogorov-Smirnov. Cada tabela é constituída de todas as distâncias entre a curva de Kaplan-Meier e as curvas dos modelos testados para cada tempo. O teste foi feito dessa forma (manual), pois trata-se de um modelo com fração de cura, o que não seria levado em consideração se o teste estatístico fosse feito de maneira usual no R, por exemplo, por meio do comando *ks.test*.

São apresentados também os gráficos com as curvas dos modelos e a curva de sobrevivência de Kaplan-Meier. Por meio desses gráficos podemos ter uma idéia de quão bom ficaram os ajustes e por meio dos valores do teste de Kolmogorov-Smirnov podemos selecionar os modelos que melhor se ajustaram para cada situação.

Os conjuntos (tabela-gráfico) serão apresentados em sequência, de acordo com a proporção de curados para cada plano de pagamentos, isto é, os três primeiros conjuntos se referem ao plano de pagamentos de 15 meses, na seguinte ordem: 5%, 15% e 35% de indivíduos curados. Para o plano de 36 meses, o mesmo acontece.

Tabela 5.3: Teste de Kolmogorov-Smirnov para o cenário com 5% de fração de cura considerando o plano de 15 meses.

	KaplanMeier	Weibull	Erro	Gama	Erro	LogNormal	Erro
1	1.00	1.00	0.00	1.00	0.00	1.00	0.00
2	0.98	0.99	0.01	0.99	0.01	1.00	0.01
3	0.95	0.97	0.02	0.97	0.02	0.98	0.02
4	0.91	0.94	0.03	0.93	0.02	0.93	0.02
5	0.84	0.88	0.04	0.87	0.03	0.86	0.02
6	0.76	0.81	0.04	0.79	0.03	0.78	0.01
7	0.67	0.71	0.04	0.70	0.03	0.69	0.01
8	0.57	0.61	0.04	0.61	0.04	0.60	0.03
9	0.47	0.51	0.04	0.52	0.05	0.51	0.05
10	0.38	0.42	0.04	0.44	0.06	0.44	0.06
11	0.30	0.35	0.05	0.36	0.06	0.37	0.07

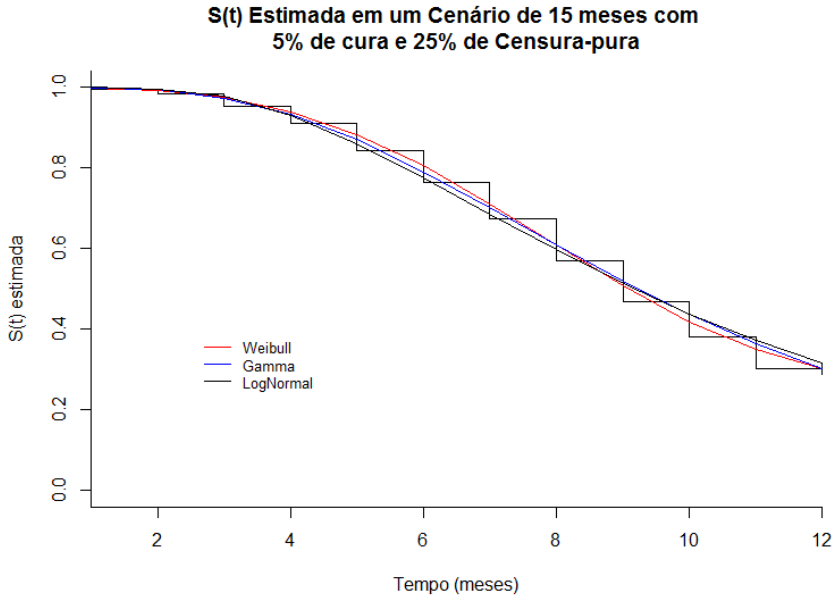


Figura 5.1: Kaplan-Meier da sobrevivência e curvas dos modelos contínuos num cenário com 5% de cura

Tabela 5.4: Teste de Kolmogorov-Smirnov para o cenário com 15% de fração de cura considerando o plano de 15 meses.

	KaplanMeier	Weibull	Erro	Gama	Erro	LogNormal	Erro
1	1.00	1.00	0.00	1.00	0.00	1.00	0.00
2	0.99	1.00	0.01	1.00	0.01	1.00	0.01
3	0.97	0.98	0.01	0.98	0.01	0.98	0.01
4	0.93	0.96	0.02	0.95	0.02	0.95	0.02
5	0.88	0.91	0.03	0.90	0.02	0.90	0.01
6	0.82	0.85	0.04	0.84	0.02	0.83	0.01
7	0.74	0.77	0.03	0.76	0.02	0.76	0.01
8	0.65	0.69	0.03	0.69	0.03	0.68	0.03
9	0.57	0.60	0.03	0.61	0.04	0.61	0.04
10	0.48	0.52	0.04	0.53	0.05	0.54	0.05
11	0.41	0.45	0.05	0.46	0.06	0.47	0.07

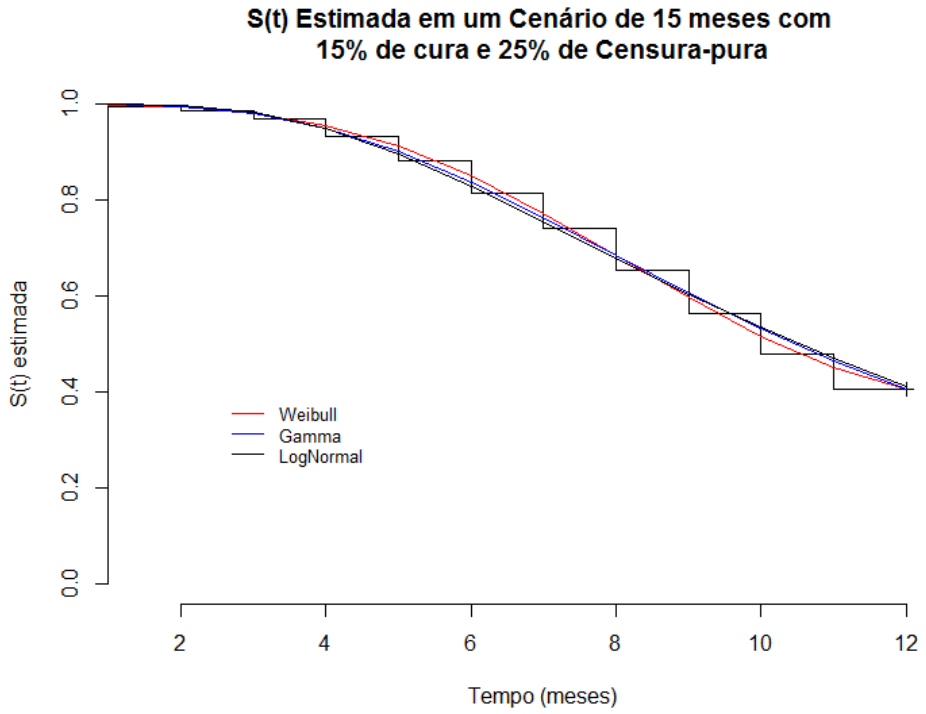


Figura 5.2: Kaplan-Meier da sobrevivência e curvas dos modelos contínuos num cenário com 15% de cura

Tabela 5.5: Teste de Kolmogorov-Smirnov para o cenário com 35% de fração de cura considerando o plano de 15 meses.

	KaplanMeier	Weibull	Erro	Gama	Erro	LogNormal	Erro
1	1.00	1.00	0.00	1.00	0.00	1.00	0.00
2	0.99	0.99	0.01	1.00	0.01	1.00	0.01
3	0.97	0.98	0.01	0.98	0.01	0.98	0.01
4	0.93	0.95	0.02	0.95	0.02	0.94	0.01
5	0.88	0.91	0.03	0.90	0.02	0.90	0.01
6	0.82	0.85	0.03	0.84	0.02	0.84	0.02
7	0.76	0.79	0.03	0.78	0.02	0.78	0.02
8	0.69	0.71	0.03	0.71	0.03	0.72	0.03
9	0.62	0.64	0.03	0.65	0.03	0.66	0.04
10	0.56	0.58	0.03	0.60	0.04	0.61	0.05
11	0.51	0.54	0.03	0.55	0.04	0.55	0.05

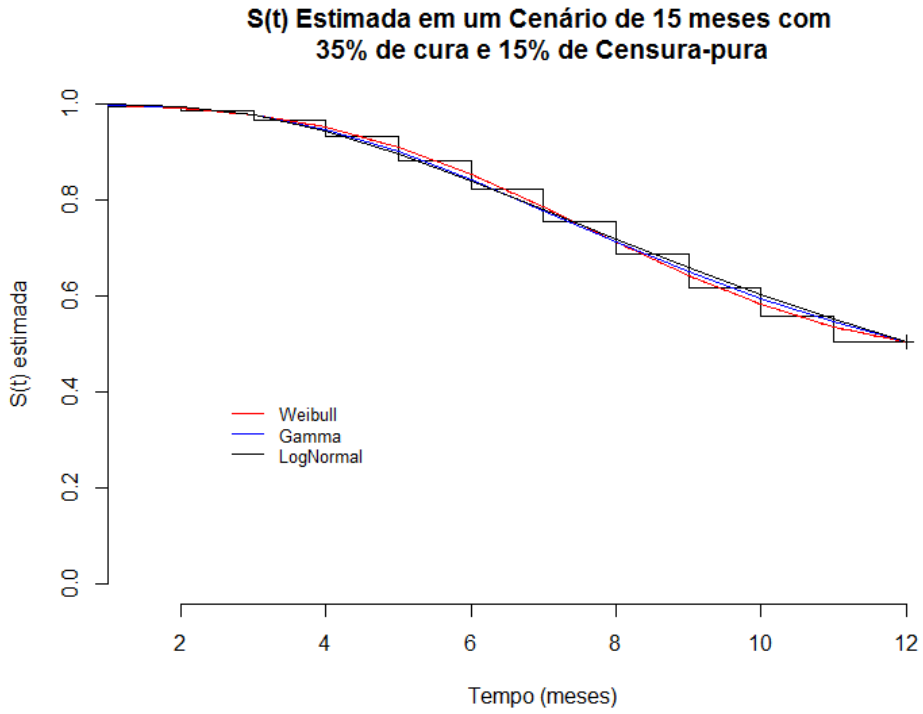


Figura 5.3: Kaplan-Meier da sobrevivência e curvas dos modelos contínuos num cenário com 35% de cura

Tabela 5.6: Teste de Kolmogorov-Smirnov para o cenário com 5% de fração de cura considerando o plano de 36 meses.

	KaplanMeier	Weibull	Erro	Gama	Erro	LogNormal	Erro
1	1.00	1.00	0.00	1.00	0.00	1.00	0.00
2	1.00	1.00	0.00	1.00	0.00	1.00	0.00
3	1.00	1.00	0.00	1.00	0.00	1.00	0.00
4	1.00	1.00	0.00	1.00	0.00	1.00	0.00
5	0.99	0.99	0.00	1.00	0.01	1.00	0.01
6	0.98	0.99	0.01	0.99	0.01	0.99	0.01
7	0.96	0.97	0.01	0.98	0.01	0.98	0.01
8	0.93	0.95	0.02	0.95	0.02	0.95	0.02
9	0.89	0.92	0.03	0.91	0.02	0.91	0.01
10	0.84	0.87	0.03	0.86	0.02	0.85	0.01
11	0.78	0.81	0.04	0.80	0.02	0.79	0.01
12	0.70	0.74	0.04	0.72	0.02	0.72	0.02
13	0.61	0.65	0.04	0.64	0.03	0.64	0.03
14	0.53	0.56	0.03	0.56	0.04	0.56	0.04
15	0.44	0.48	0.03	0.49	0.05	0.49	0.05
16	0.37	0.40	0.03	0.42	0.05	0.42	0.05
17	0.30	0.34	0.04	0.36	0.05	0.36	0.06

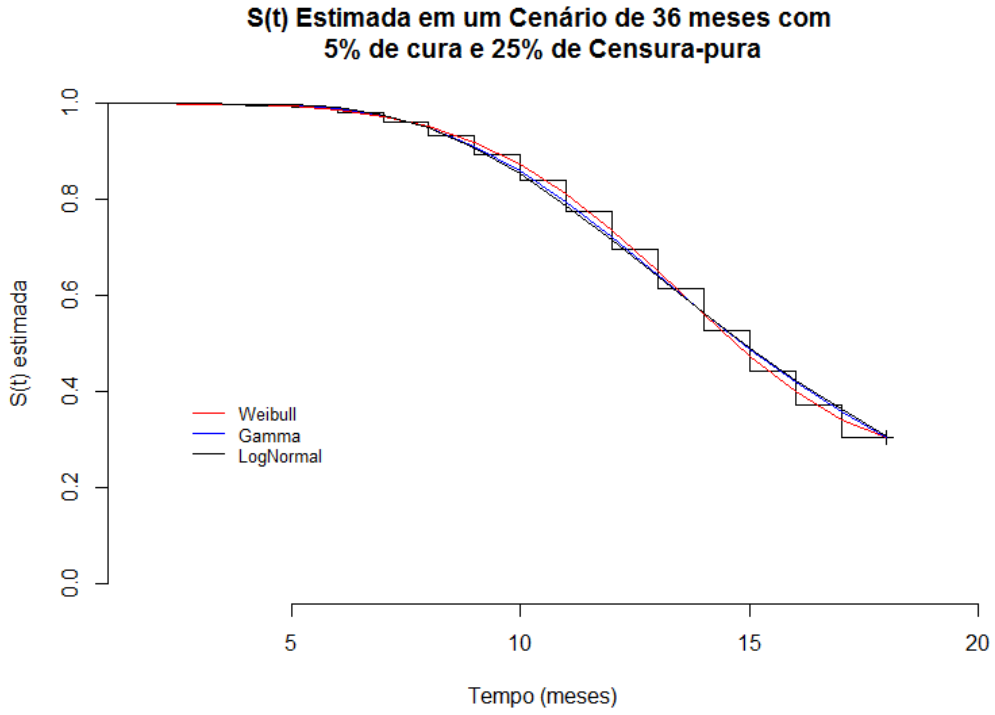


Figura 5.4: Kaplan-Meier da sobrevivência e curvas dos modelos contínuos num cenário com 5% de cura

Tabela 5.7: Teste de Kolmogorov-Smirnov para o cenário com 15% de fração de cura considerando o plano de 36 meses.

	KaplanMeier	Weibull	Erro	Gama	Erro	LogNormal	Erro
1	1.00	1.00	0.00	1.00	0.00	1.00	0.00
2	1.00	1.00	0.00	1.00	0.00	1.00	0.00
3	1.00	1.00	0.00	1.00	0.00	1.00	0.00
4	1.00	1.00	0.00	1.00	0.00	1.00	0.00
5	0.99	1.00	0.00	1.00	0.00	1.00	0.00
6	0.99	0.99	0.00	0.99	0.01	0.99	0.01
7	0.97	0.98	0.01	0.98	0.01	0.98	0.01
8	0.95	0.96	0.01	0.96	0.01	0.96	0.01
9	0.91	0.94	0.02	0.93	0.02	0.93	0.01
10	0.87	0.90	0.03	0.89	0.02	0.88	0.01
11	0.82	0.85	0.03	0.83	0.02	0.83	0.01
12	0.75	0.79	0.03	0.77	0.02	0.77	0.02
13	0.68	0.71	0.03	0.71	0.02	0.71	0.03
14	0.61	0.64	0.03	0.64	0.03	0.64	0.04
15	0.53	0.56	0.03	0.57	0.04	0.58	0.04
16	0.46	0.49	0.03	0.51	0.04	0.52	0.05
17	0.40	0.44	0.04	0.45	0.05	0.46	0.05

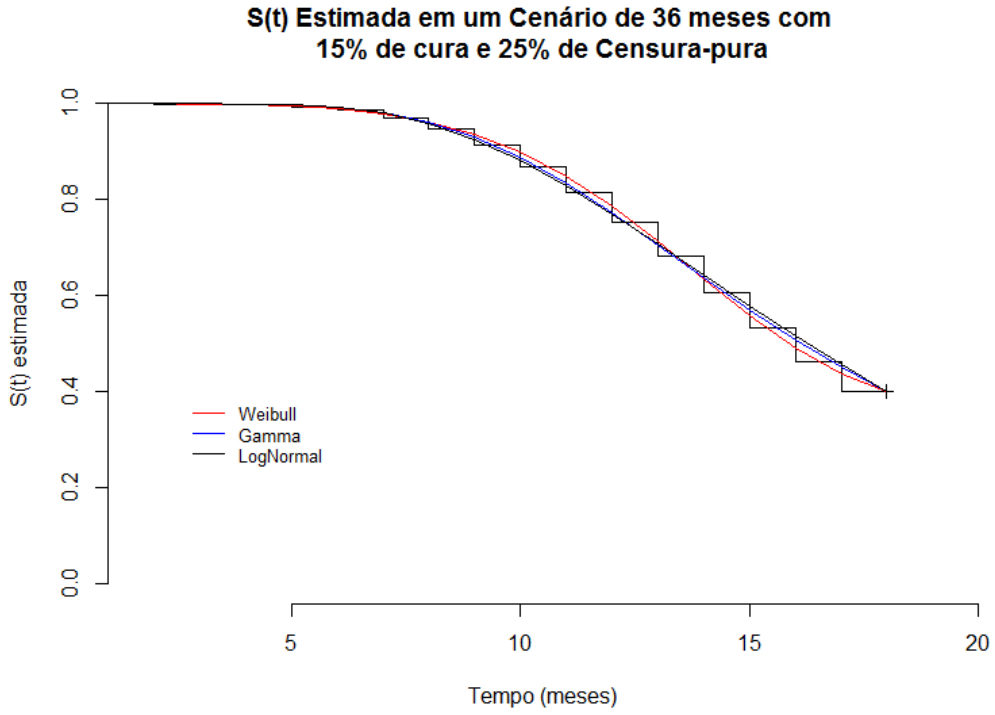


Figura 5.5: Kaplan-Meier da sobrevivência e curvas dos modelos contínuos num cenário com 15% de cura

Tabela 5.8: Teste de Kolmogorov-Smirnov para o cenário com 35% de fração de cura considerando o plano de 36 meses.

	KaplanMeier	Weibull	Erro	Gama	Erro	LogNormal	Erro
1	1.00	1.00	0.00	1.00	0.00	1.00	0.00
2	1.00	1.00	0.00	1.00	0.00	1.00	0.00
3	1.00	1.00	0.00	1.00	0.00	1.00	0.00
4	0.99	1.00	0.00	1.00	0.01	1.00	0.01
5	0.98	1.00	0.01	1.00	0.01	1.00	0.01
6	0.97	0.99	0.02	0.99	0.02	0.99	0.02
7	0.95	0.98	0.03	0.98	0.03	0.98	0.03
8	0.92	0.96	0.04	0.96	0.04	0.96	0.04
9	0.88	0.94	0.06	0.93	0.05	0.93	0.05
10	0.83	0.90	0.07	0.89	0.06	0.89	0.06
11	0.78	0.86	0.08	0.85	0.07	0.84	0.07
12	0.72	0.81	0.08	0.80	0.07	0.79	0.07
13	0.66	0.75	0.09	0.74	0.08	0.74	0.08
14	0.61	0.69	0.08	0.69	0.08	0.69	0.08
15	0.56	0.63	0.07	0.64	0.08	0.64	0.08
16	0.51	0.58	0.07	0.59	0.08	0.59	0.08

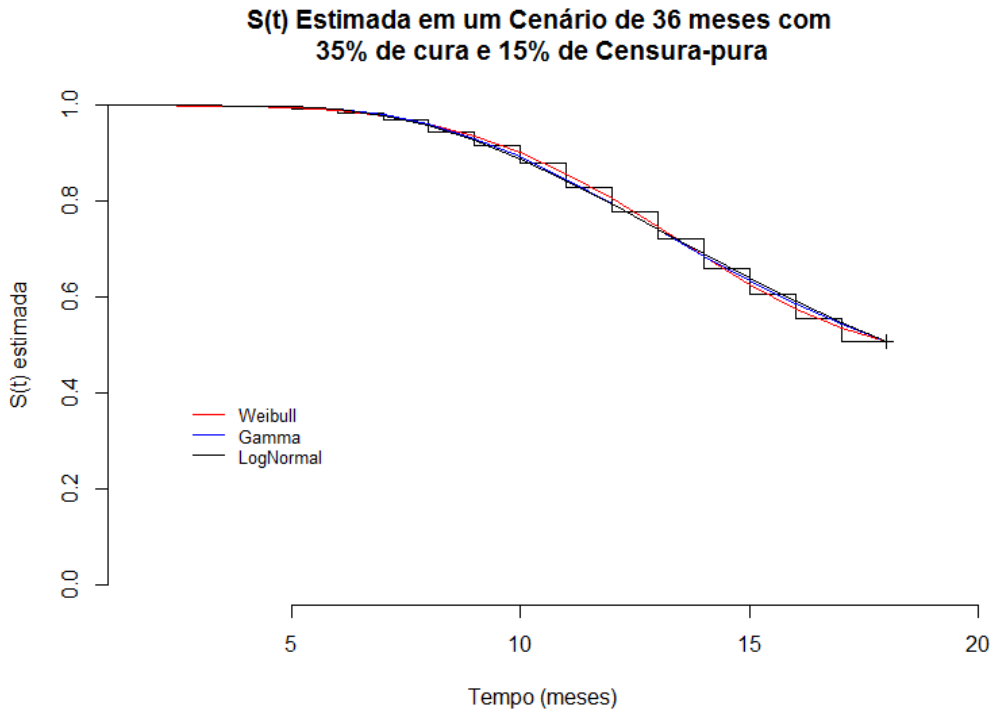


Figura 5.6: Kaplan-Meier da sobrevivência e curvas dos modelos contínuos num cenário com 35% de cura

Análise dos Resultados

Os gráficos mostrados anteriormente informam que os modelos testados se ajustaram bem aos dados de tempo discretos, mas, por meio deles, não é possivel conhecer o modelo que melhor ajuste apresentou em cada uma das seis situações. Por isso, uma análise dos valores apresentados pelo teste de Kolmogorov-Smirnov foi feita para cada cenário.

Para o primeiro cenário (5% de cura num plano com 15 meses de pagamento) a tabela com os valores do teste K-S mostra, na coluna "Erro", que o modelo que apresentou menor diferença com relação à curva de sobrevivência de K-M foi o Weibull, e logo em seguida com uma diferença de 1% o modelo gama. O modelo Lognormal foi o que se mostrou mais distante, embora a diferença seja de apenas 2% do modelo mais bem ajustado. Portanto, nesta situação o modelo com melhor ajuste foi o Weibull.

Para o segundo cenário (15% de cura num plano com 15 meses de pagamento) temos exatamente a mesma situação verificada anteriormente, para 5% de cura em 15 meses. O erro máximo observado foi de 7% para o modelo Lognormal e o modelo com melhor ajuste foi o Weibull.

Para o terceiro cenário (35% de cura num plano com 15 meses de pagamento) verificamos, para os três modelos, um ajuste melhor do que o visto anteriormente. Aqui o erro máximo foi de 5%, associado ao modelo Lognormal. Novamente o modelo Weibull apresentou melhor ajuste, seguido bem de perto pelo modelo gama.

Passando agora para os últimos três cenários, para o plano de 36 meses, temos na primeira situação (5% de cura) que o modelo que melhor se ajustou foi o Weibull com um erro máximo de 4%, seguido pelo modelo gama com um erro máximo de 5%. Por último o modelo Lognormal com um erro máximo de 6%.

Para o quinto cenário (15% de cura num plano com 36 meses de pagamento) verifica-se que o modelo Weibull apresenta um ajuste melhor do que o apresentado pelos outros modelos. Os modelos gama e lognormal apresentam, nesta situação, um empate, ambos apresentando erro máximo de 5%. A vantagem, neste empate, fica para o modelo gama que apresenta erro máximo de 5% apenas no tempo 17, enquanto que o modelo lognormal apresenta o mesmo erro para os tempos 16 e 17.

No sexto e último cenário temos algo inesperado. O modelo Weibull apresenta o pior ajuste com um erro máximo de 9% para o tempo 13, enquanto que o erro máximo observado, tanto para o modelo gama quanto para o modelo lognormal foi de 8%.

Conclusão

Neste trabalho foi tentado o ajuste de modelos paramétricos contínuos a dados de tempos discretos, por meio de um banco de dados simulado, considerando a presença de indivíduos curados. Os modelos Weibull, Gama e LogNormal foram avaliados em seis situações, três em cada plano de pagamento, com 15 e 36 meses. Em cada plano com 5%, 15% e 35% de fração de cura.

Ao final do presente trabalho foi possível observar que os modelos paramétricos contínuos Weibull, Gama e Lognormal utilizados, conseguem se ajustar de forma satisfatória aos dados de tempo discretos.

Para os modelos Weibull e gama percebeu-se um ajuste melhor, tendo como base os resultados do teste de Kolmogorov-Smirnov, do que o verificado pelo modelo lognormal. Esses dois modelos conseguem estimar o parâmetro ϕ com mais eficiência, embora na modelagem Weibull a estimativa desse parâmetro se mostre muito superestimada, se aproximando do valor total de censurados, o que pode significar que ele não consegue identificar dentro dos indivíduos censurados aqueles que são curados (embora, num banco de dados real realmente não se consiga fazer essa identificação). Já o modelo com distribuição gama parece conseguir estimar a cura unicamente e não a censura total.

Com relação ao modelo lognormal o que chamou a atenção foi o fato de que no plano de 15 meses em quase todos os cenários ele não considerou a fração de curados, assim como para o cenário com 5% de cura para o plano de 36 meses, isto é, sua estimativa para esse parâmetro foi proxima de zero, mas verificou-se que quando se tratava de 35% de cura no plano de 15 meses ou 15% e 35% de cura no plano de 36 meses sua estimativa já se apresentava relativamente considerável, embora subestimada com relação ao valor real do parâmetro. Esse comportamento sugere que para janelas de observação pequenas, como é o caso do plano de 15 meses cuja janela de observação foi de 12 meses, e para proporções pequenas de fração de curados, o modelo lognormal não seja bom na estimação do parâmetro ϕ.

Dessa forma, a escolha do modelo que melhor se ajustou aos dados simulados, fica um pouco complicada, pois analisando os gráficos temos que os tês mostraram-se muito bem ajustados. Passamos então à análise dos valores obtidos pelo teste de Kolmogorov-Smirnov. Esse teste, indica, por uma diferença muito pequena, que, na maior parte das seis situações observadas, o modelo que melhor se ajustou foi o Weibull, com diferença

máxima com relação ao modelo gama de 1%. Porém, busca-se um modelo com fração de cura onde os parâmetros estimados tenham valores o mais próximo possível dos valores reais, e portanto, ainda que com uma pequena diferença pra mais no teste de Kolmorogov-Smirnov, o modelo que melhor se apresenta com relação a essa necessidade, é o modelo gama.

Fica como sugestão para trabalhos futuros a construção de modelos de análise de sobrevivência, com fração de cura, paramétricos para dados discretos com a inclusão de covariáveis. Seria muito valiosa a aplicação desses modelos a dados reais de operações de crédito.

Referências Bibliográficas

[1] BERKSON, J.; GAGE, R.P. **Survival Curve for Cancer Patients Follwing Treatment**. Journal of the American Statistical Association, Alexandria, v. 47, p. 501-515, 1952.

[2] COLOSIMO, E. A. e GIOLO, S. R. (2006). **Análise de Sobrevivência Aplicada**. São Paulo: Edgard Blucher.

[3] FACHINI, J. B. **Modelos de regressão para dados bivariados com e sem fração de cura**. Tese (Doutorado em Estatística e Experimentação Agronômica)- Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2011.

[4] GOVINDARAJULU, Z. **Nonparametric Inference**. 2007. USA. 1st ed. World Scientific.

[5] LAWLESS, J.F. **Statistical Models and Methods for Lifetime Data**. USA. 2nd ed. Wiley

[6] LICHTENSTEIN, P.; WIENKE, A.; YASHIN, A. (2003) **A Bivariate Frailty Model a Cure Fraction for Modeling Familial Correlations in Diseases.** Biometrics 59, 1178-1183.

[7] MACHADO, A. R. **Modelos Estatísticos para Avaliação de Risco em Produtos de Crédito Parcelados.** 2010. 92 fl.. Monografia (Graduação em Estatística) - Departamento de Estatística, Universidade de Brasília, Brasília.

[8] MELLO, M.P, e PETERNELLI, L.A. (2007). **Conhecendo o R - Uma visão Estatística**. Viçosa: Ed. UFV

[9] RIZZATO, F. B. **Modelos de Regressão Log-Gama Generalizado com Fração de Cura.** 2006. 74 fl.. Dissertação (Mestrado em Agronomia) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, São Paulo.

[10] SICSU, A.L. **Credit Scoring: desenvolvimento, implantação e acompanhamento**. São Paulo: Blucher, 2010.

[11] SILVA, R. A. **Modelos de Fração de Cura com Fatores Latentes Competitivos e Fragilidade.** 2011. 76 fl.. Dissertação (Mestrado em Ciências) - Universidade de São Paulo, São Paulo.

[12] YAKOVLEV, A. Y. e TSODIKOV, A. D. (1996). **Stochastic Models of Tumor Latency and Their Bioestatistical Applications**, first edn, World Scientific, Singapore.

Geração do Banco de Dados

```
n <- *tamanho da amostra*
p0 <- *1-prob(cura)*
t <- *tempo de truncamento*
tam <- *k*
med <- *média usada na binomial negativa*
set.seed(156565)
bn<-cens<-cura<-numeric()
for (i in 1:n) {
   cura[i] <- rbinom(1,1,p0)
   if (cura[i]==0) {
      bn[i] <- t
      cens[i] <- 0
      }
   if (cura[i]==1) {
      bn0<-rnbinom(1,size=tam,mu=med)
      if (bn0<t) {
         bn[i] <- bn0
         cens[i] <- 1
      }
      if (bn0>=t) {
         bn[i] <- t
         cens[i] <- 0
      }
   }
}
Estado <- rep(1L, n)
Estado[!cens & cura] <- 2L # Indica a falha
Estado[!cens & !cura] <- 3L  # Indica a Cura
labels = c('Falha', 'Censura', 'Cura')
bd15.1 <- data.frame(cura=cura,cens=cens,
bn=bn,Estado=labels[Estado])
```

Estimação dos Parâmetros

```
# Usando a Weibull

# Estimativa dos parametros (Regressão Weibull)
ajuste1 <- survreg(tempo~1,dist="weibull")
alpha <- exp(ajuste1$coefficients[1])
gamma <- 1/ajuste1$scale
estima1 <- cbind(gamma,alpha)
gamma1 <- estima1[1]
alpha1 <- estima1[2]

# Valores Iniciais
vi <- c(gamma1,alpha1,1-p0)
delta <- cens
t0 <- bn

# Função de Verossimilhança1 - com a Weibull
flv <- function(param,t0,delta){
g <- param[1]
a <- param[2]
p <- param[3]

lv <- delta* log((1-p)*dweibull(t0,g,a)) +
(1-delta) * log( p + (1-p)*(1-pweibull(t0,g,a) ) )
sum(-lv)
}

estw <- nlminb(vi,flv,lower=c(10^-10,10^-10,10^-10),upper=c(10^10,10^10,1),
t0=t0,delta=delta)
estw
```

Gráficos

```
# Gráficos
tempos <- sort(t0)
plot(estim,conf.int="F",bty="l",lwd=1,ylim=c(0,1),
xlab="Tempo (meses)",ylab="S(t) estimada",
main="S(t) Estimada em um Cenário de 15 meses com
5% de cura e 25% de Censura-pura",
xlim=c(1,12))

# com a Weibull
NM <- (estw$par[3])+(1-estw$par[3])*(1-pweibull(tempos,estw$par[1],
estw$par[2]))
lines(tempos,NM,lty=1, col=2, type="l")

# com a Gamma
NM2 <- (estg$par[3])+(1-estg$par[3])*(1-pgamma(tempos,estg$par[1],
estg$par[2]))
lines(tempos,NM2,lty=1, col=4, type="l")

# com a LogNormal
NM3 <- (ln$par[3])+(1-ln$par[3])*(1-plnorm(tempos,ln$par[1],
ln$par[2]))
lines(tempos,NM3,lty=1, col=6, type="l")

legend(2.5,0.4,lty=c(1,1),c("Weibull","Gamma"),bty="n",cex=0.8,
col=c(2,4))
```

Kolmogorov-Smirnov

```
#Usando a Weibull

# Medindo o erro de ajustamento do modelo
sksw <- summary(estim)
km <- sksw$surv
temp <- c(1:11)

# Weibull
ksw <- (estw$par[3])+(1-estw$par[3])*(1-pweibull(temp,estw$par[1],
estw$par[2]))
difw <- abs(km-ksw)
kstw <- cbind(km,ksw,difw)
kstw
xtable(kstw)
```