Distribuição *Logistic-sinh* para aplicações em dados de Sobrevivência

Juliana Gardelli [1] e Josmar Mazucheli[2]

[1,2] Programa de Bioestatística / Universidade Estadual de Maringá

RESUMO

Dados altamente assimétricos negativamente com observações extremas não são muito comuns em análise de sobrevivência. As distribuições de probabilidade usadas em dados com essa característica consideram estas observações como *outliers* e isso pode levar a uma má estimação das pequenas probabilidades nas caudas. Na prática, estes quantis mal estimados podem levar a falhas estruturais nas construções, ou falha prematura em componentes mecânicos. A distribuição estudada, a logistic-sinh, é derivada da distribuição logística e apresenta caudas grossas para assimetria negativa. Neste trabalho, dado que a distribuição logistic-sinh é apropriada para a análise de dados com assimetria a esquerda, avaliamos as propriedades dos estimadores de máxima verossimilhança via simulação Monte Carlo. Uma aplicação em dados de resistência de uma fibra de vidro é apresentada para comparar o ajuste desta distribuição com a Weibull de 3 parâmetros.

Palavras chave: Distribuição Logistic-sinh, Análise de sobrevivência, Simulação de Monte Carlo.

1 INTRODUÇÃO

Neste trabalho foi estudada a distribuição Logistic-sinh, proposta Kahadawala Cooray em 2005. É uma distribuição de dois parâmetros (λ, de forma e θ de escala), obtida a partir de uma transformação na distribuição Logística, substituindo um termo exponencial por um termo seno hiperbólico. A principal característica desta distribuição é que ela foi obtida para modelar dados altamente assimétricos negativamente, com observações extremas, consequentemente, ela possui caudas grossas para assimetria negativa. Segundo o autor estes tipos de dados são comuns em Análise de Sobrevivência.

As distribuições como a Gompertz, sinh-normal, *exponentiated Weibull, generalized gamma* e Weibull são utilizadas para modelar dados assimétricos negativamente, mas elas ignoram as observações extremas na cauda direita ou possuem caudas finas para assimetria negativa. Isso leva a uma má estimação das pequenas probabilidades encontradas nas caudas. A necessidade de aumentar a precisão na estimação das observações das caudas motivou o autor a explorar novos modelos. A importância de se estimar bem as caudas é porque quantis mal estimados podem levar a sérias consequências como: falha estrutural nas construções e falha prematura em componentes mecânicos. Por isso estas pequenas probabilidades das caudas devem ser incorporadas no modelo.

São apresentadas as funções densidade, acumulada, sobrevivência e risco. E ainda, a log-verossimilhança, da qual são estimados os dois parâmetros através da função escore. São apresentados também os resultados da Simulação de Monte Carlo e da aplicação da distribuição em dados de resistência a pressão de uma determinada fibra de vidro.

2 METODOLOGIA

A distribuição *half logistic* apresenta distribuição acumulada dada por

$$F(x) = 1 - (1 + 0.5(exp(x) - 1))^{-1}; \quad 0 \leq x < \infty$$

A distribuição *logistic-sinh* é obtida substituindo o termo $[exp(x)-1]$ da *half logistic*, por $sinh[exp(x)-1]$ obtendo então a seguinte distribuição acumulada

$$F(x;\lambda,\theta) = 1 - \left(1 + \lambda\, sinh\left(exp\left(\frac{x}{\theta}\right) - 1\right)\right)^{-1}, \quad 0 < x, \lambda, \theta < \infty \tag{1}$$

Esta substituição na distribuição acumulada produz uma função densidade assimétrica negativamente, que é dada por

$$f(x \mid \lambda,\theta) = \left(\frac{\lambda}{\theta}\right) exp\left(\frac{x}{\theta}\right) cosh\left(exp\left(\frac{x}{\theta}\right) - 1\right)\left(1 + \lambda\, sinh\left(exp\left(\frac{x}{\theta}\right) - 1\right)\right)^{-2} \tag{2}$$

A função de sobrevivência é definida como a probabilidade de uma observação sobreviver a um certo tempo t, em termos probabilísticos ela é escrita como $S(t) = P(T \geq t)$. A distribuição acumulada é definida por $F(t) = P(T \leq t)$, que probabilisticamente equivale a $F(t) = 1 - S(t)$, logo, $S(t) = 1 - F(t)$. Coma a utilização destes artifícios matemáticos a função de sobrevivência desta distribuição é dada por

$$S(x \mid \lambda,\theta) = \left(1 + \lambda\, sinh\left(exp\left(\frac{x}{\theta}\right) - 1\right)\right)^{-1} \tag{3}$$

A função de risco é mais informativa do que a função de sobrevivência, pois mede o risco associado a um indivíduo/componente no tempo t. Dado que o indivíduo/componente sobreviveu até t, a probabilidade dele falhar no próximo t é dada por

$$h(x \mid \lambda,\theta) = \left(\frac{\lambda}{\theta}\right) exp\left(\frac{x}{\theta}\right) cosh\left(exp\left(\frac{x}{\theta}\right) - 1\right)\left(1 + \lambda\, sinh\left(exp\left(\frac{x}{\theta}\right) - 1\right)\right)^{-1} \tag{4}$$

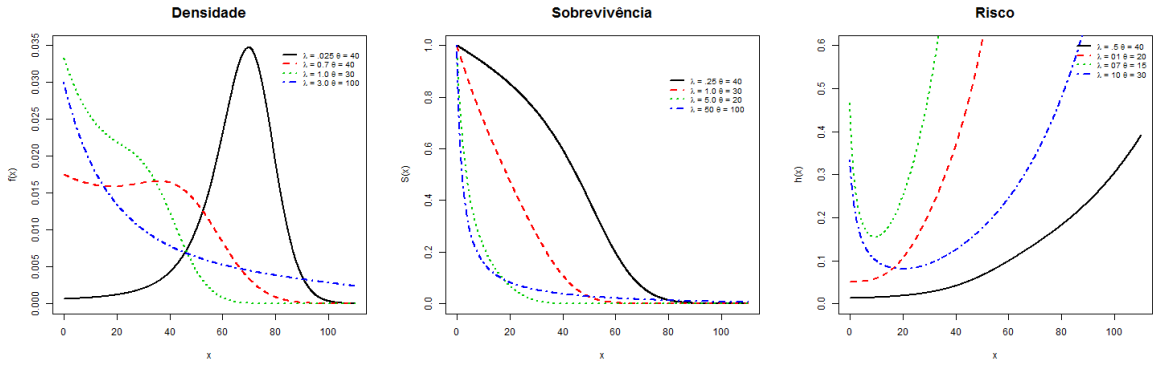

Figura 1: Funções de densidade, sobrevivência e risco da distribuição *Logistic-sinh*.

Os estimadores de máxima verossimilhança são os valores obtidos para λ e θ maximizando a função de log-verossimilhança, calculando as derivadas parciais por meio da função escore e igualando a zero. A função de máxima verossimilhança é definida por

$$L(\lambda,\theta \mid X_i) = \prod_{i=1}^{n}\left[\left(\frac{\lambda}{\theta}\right) exp\left(\frac{x}{\theta}\right) cosh\left(exp\left(\frac{x}{\theta}\right) - 1\right)\left(1 + \lambda\, sinh\left(exp\left(\frac{x}{\theta}\right) - 1\right)\right)^{-2}\right]$$

A função de log-verossimilhança da família logistic-sinh é dada por

$$l(\lambda,\theta) = \sum_{i=1}^{n} ln\left(\frac{\lambda}{\theta}\right) + \frac{x_i}{\theta} + ln(cosh(e^{\frac{x_i}{\theta}} - 1)) - 2\sum_{i=1}^{n} ln(1 + \lambda senh(e^{\frac{x_i}{\theta}} - 1)) \tag{5}$$

Obtém-se o seguinte sistema não linear, também conhecido como função escore:

$$\frac{\partial l(\lambda,\theta)}{\partial \lambda} = \frac{n}{\lambda} - 2\sum_{i=1}^{n} \frac{sinh(e^{\frac{x_i}{\theta}} - 1)}{1 + \lambda\, sinh(e^{\frac{x_i}{\theta}} - 1)} = 0$$

$$\frac{\partial l(\lambda,\theta)}{\partial \theta} = -\frac{n}{\theta} - \sum_{i=1}^{n} \frac{x_i}{\theta^2} - \sum_{i=1}^{n} \frac{sinh(e^{\frac{x_i}{\theta}} - 1)x_i e^{\frac{x_i}{\theta}}}{\theta^2 cosh(e^{\frac{x_i}{\theta}} - 1)} + 2\sum_{i=1}^{n} \frac{\lambda cosh(e^{\frac{x_i}{\theta}} - 1)x_i e^{\frac{x_i}{\theta}}}{\theta^2(1 + \lambda sinh(e^{\frac{x_i}{\theta}} - 1))} = 0$$

Duas propriedades são desejáveis para um estimador, que ele seja assintoticamente: não viesado $E(\widehat{\theta}) = \theta$ e tenha variância mínima, para a qual se avalia o Erro Quadrático Médio (MSE): $MSE(\widehat{\theta}) = E(\widehat{\theta} - \theta)^2$.

3 RESULTADOS E DISCUSSÕES

A Simulação de Monte Carlo é capaz de produzir um fluxo sem fim de variáveis aleatórias para distribuições novas ou já conhecidas. Com isto, é possível avaliar se os parâmetros estimados pelo método da máxima verossimilhança para esta distribuição apresentam as propriedades assintóticas descritas.

Foram atribuídos os seguintes valores para os parâmetros: $\lambda = (0.01, 0.1, 1, 10, 50)$, e $\theta = (10, 20, 50)$. A combinação destes valores gerou 15 diferentes pares, tomados 7 tamanhos de amostras n= (20,50,80,110,140,170,200). Cada combinação foi simulada com os 7 tamanhos diferentes, gerando um total de 105 situações. Cada situação é representada por exemplo (n=20, $\lambda = 0.01$ e $\theta = 10$). Foram geradas 10.000 valores para cada uma das 105 situações e calculada a média $\widehat{\theta}$, e esta foi comparada com o valor atribuído inicialmente para verificar o Viés e o MSE.

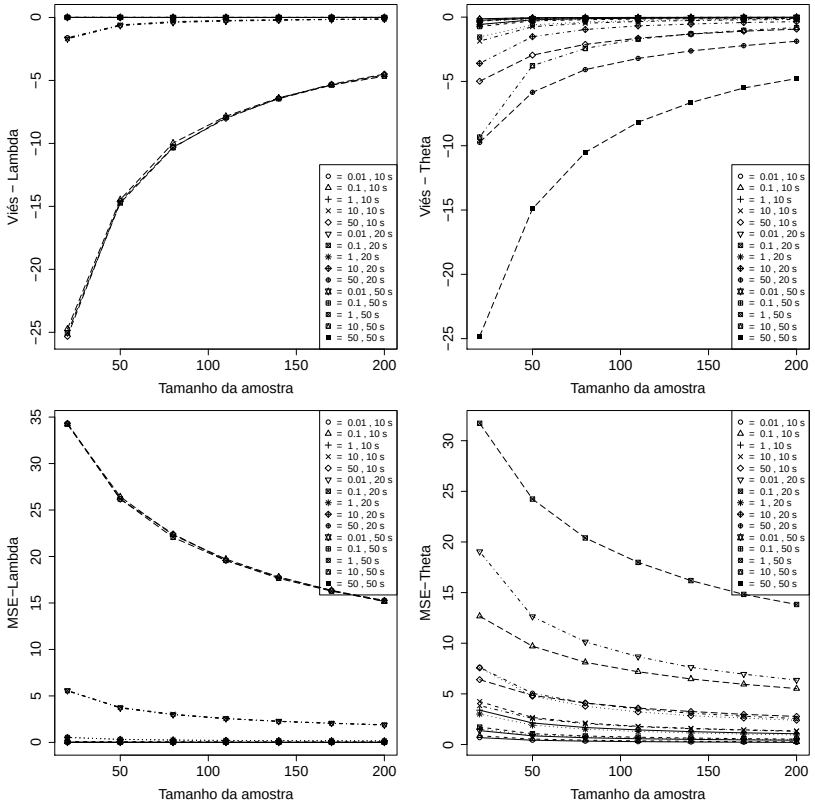

Figura 2: Vieses e Erro Quadrático Médio.

Aplicação em dados não censurados para comparar o ajuste das distribuições *Logistic-Sinh* e Weibull de 3 parâmetros. Os dados analisados são da resistência de uma fibra de vidro de 1.5 cm e 15 cm, provenientes do Laboratório Nacional de física, na Inglaterra. Uma inspeção preliminar dos dados revela possíveis outliers na extremidade inferior das amostras. Em geral valores de resistência experimental têm grande dispersão, então distribuições de caudas mais longas são mais apropriadas para modelar esses dados.

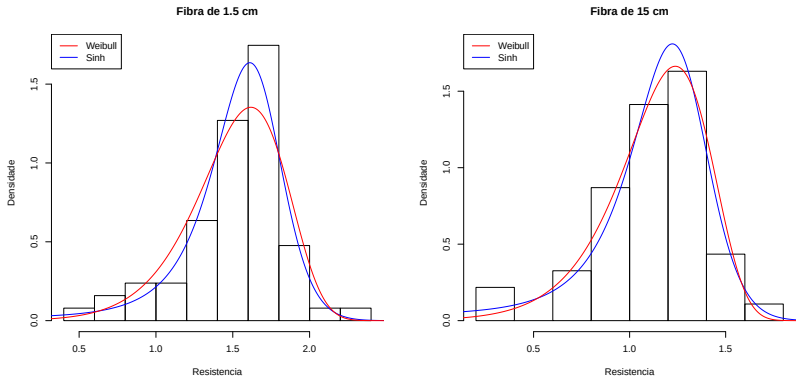

Figura 3: Comparação de ajustes da Logistic-Sinh e Weibull de 3 parâmetros.

Na Figura 3 observa-se que as curvas da Logistic-Sinh ajusta um pouco melhor os dados, ela capta uma pouco melhor a assimetria à esquerda. Outro critério de comparação é o AIC, este tenta equilibrar a necessidade de um modelo que ajuste bem os dados, com o menor número possível de parâmetros, o critério de decisão é quanto maior, melhor. É baseado no valor $l(\widehat{\theta})$ e no número de parâmetros. A Amostra 1 apresentou $r = -1.485$ para LS e $r = -3.341$ para Weibull. A Amostra 2 apresentou $r = -3.026$ para LS e $r = -5.266$ para Weibull. Em ambas as amostras o valor de r foi maior para a Logistic-Sinh, indicando que ela ajusta melhor as duas amostras. O terceiro critério de comparação foi o Teste KS Kolmogorov Smirnov, analisa a maior distância entre as acumuladas e compara com a distribuição de referência, H_0 : X e Y têm a mesma distribuição. A estatística de teste resultou $D = 0.52381$, $p - valor = 6.222e - 08$, logo rejeita-se a hipótese de que ambas as curvas seguem a mesma distribuiçã, como era de se esperar.

4 CONCLUSÃO

Conclui-se através do estudo de simulação e análise gráfica que os vício tanto de λ quanto de θ convergem para zero, comprovando que os estimadores são não tendenciosos. E com a aplicação conclui-se que para os dados analisados, a distribuição Logistic-Sinh ajusta-se um pouco melhor, captando a assimetria à esquerda.

Referências

[1] KAHADAWALA, Cooray. Analyzing lifetime data with long-tailed skewed distribution: the logistic-sinh family.Statistical Modelling. v.5, p.343. 2005. Disponivel em $< http : //smj.sagepub.com/cgi/content/abstract/5/4/343 >$. Acesso em Novembro de 2017.

[2] COLOSIMO, Enrico Antônio; GIOLO, Suely Ruiz. Análise de sobrevivência aplicada. In: **ABE-Projeto Fisher**. Edgard Blücher, 2006.