A distribuição logística generalizada tipo I, com três, dois e um parâmetro e a distribuição logística, no ajuste de dados de sobrevivência.

Tiago Viana Flor de Santana - IMECC, UNICAMP [1]

Resumo: *Neste trabalho, estimou-se os parâmetros das distribuições logística generalizada tipo I com três, dois e um parâmetro e da distrição logística pelo método de máxima verossimilhança para dados censurados baseado em uma amostra completa, utilizando dados assimétricos à direita de pacientes com cancer de bexiga, dependentes químicos e de crianças expostas ao HIV. O ajuste de cada distribuição aos conjuntos de dados foi comparado entre as demais distribuições apresentadas no trabalho, verificando a influência de cada parâmetro no ajuste dos modelos. Conclui-se que a distribuição logística generalizada tipo I com três parâmetros, que tem como caso particular as distribuições logística generalizada tipo I, com dois e um parâmetro e a distribuição logística, apresenta melhor ajuste para os dados em estudo (menor AIC, BIC, CAIC), em relação as distribuições comparadas. O parâmetro de locação μ, confere a distribuição maior flexibilidade, melhorando o ajuste aos dados principalmente quando o conjunto de dados é fortemente assimétrico a direita. O parâmetro b proporciona um ajuste mais acurado comparado a distribuição logística.*

Palavras-chaves: *Distribuição Logística Generalizada Tipo I, Máxima Verossimilhança, Parâmetro de Locação, Análise de Sobrevivência.*

1 Introdução

A distribuição e função logística tem sido grandemente utilizada em várias áreas de aplicação desde do fim do século XIX, tais como em estudos de demografia, de crescimento de populações humanas e de organismos biológicos, em econômia no estudo de distribuição de renda e na área de saúde pública (Balakrishnan, 1992).

Balakrishnan e Leng (1988) definiram as distribuições logística generalizadas tipo I, com um parâmetro, tipo II e tipo III como generalizações da distribuição logística padrão e observaram que estas distribuições são assimétricas. Em particular a distribuição logística generalizada tipo I (LGI) é uma família de assimetria positiva com coeficientes de curtose maior que a distribuição logística. Essa distribuição modela dados cuja função de risco tem forma crescente, sendo mais flexível que a distribuição logística e tem como submodelo a distribuição logística padrão. Alkasasbeh e Raqab (2009) consideram a estimação de máxima verossimilhança de diferentes

[1]Doutorando em Estatística - IMECC, UNICAMP. Contato: *tiagodesantana@yahoo.com.br*

parâmetros da distribuição LGI, com dois parâmetros, bem como outros cincos procedimentos de estimação comparando suas performances atráves de extensiva simulação numérica, obtendo resultados satisfatórios para o método de máxima verossimilhança devido as suas vantangens computacionais.

Dessa forma, neste trabalho, estimou-se os parâmetros da distribuição LGI, incluindo os parâmetros de locação (μ) e escala (σ) e denotada por LGI3(b, μ, σ), pelo método de máxima verossimilhança, utilizando três conjuntos reais de dados de sobrevivência na presença de censura, e comparou-se o ajuste dos modelos com o modelo logístico com parâmetros de escala σ e de locação μ, e com os modelos provenientes das distribuições LGI com um parâmetro, que será denotado por LGI1(b) e LGI com dois parâmetros que será denotado por LGI2(b,σ). A função densidade de probabilidade (fdp), de sobrevivência e de risco para distribuição LGI3 são dadas pelas expressões:

$$f(y) = \frac{b}{\sigma}\frac{exp(-bz)}{[1+exp(-z)]^{b+1}} \quad , \quad S(y) = \frac{exp(-bz)}{[1+exp(-z)]^{b}} \quad \text{e} \quad h(y) = \frac{b}{\sigma[1+exp(z)]} \quad ,$$

em que $z = \frac{y-\mu}{\sigma}$, $-\infty < y < \infty$, $-\infty < \mu < \infty$, $\sigma > 0$ e $b > 0$.

O método de Máxima Verossimilhança estima os parâmetros, baseados nos dados amostrais, e define qual a distribuição, entre todas as possíveis definidas pelos valores de seus parâmetros, tem maior probabilidade de gerar a amostra em discussão. A contribuição de cada observação, não censurada, na estimação dos parâmetros é a sua função de densidade porém, no caso de observações censuradas a contribuição é a sua função de sobrevivência, pois estas observações informam apenas que o tempo de falha é maior que o tempo de censura observado (Colosimo e Giolo, 2006).

O objetivo deste trabalho, portanto, foi através do método de máxima verossimilhança para dados censurados, estimar os parâmetros da distribuição LGI3, utilizando três conjuntos dados reais de sobrevivência, considerando censura, com diferentes graus de assimétria a direita, e comparar o ajuste dos modelos com as distribuições LGI2, LGI1 e logística comparando o efeito da inclusão de cada parâmetro.

2 Metodologia

O primeiro conjunto de dados, refere-se ao tempo até a recorrência do primeiro tumor de oitenta e cinco pacientes com câncer de bexiga, obtidos de Collet (2003). O experimento foi controlado com placebo e conduzido pelo VACURG (Veterans Administration Cooperative Urological Research Group), onde pacientes com tumores superficial de bexiga primeiramente tiveram seus tumores removidos trans-uretralmente e em seguida foram aleatorizados para receber placebo, ou um agente quimioterápico, denominado thiotepa. Das $n = 85$ observações, 47 receberam placebo, 38 receberam o agente quimioterápico thiotepa e 38 observações foram censuradas.

O segundo conjunto de dados foi fornecido pela Associação Mãe Admirável, situada na cidade de Caratinga, MG. Foram avaliados 141 residentes, dependentes químicos, no período de 2000 a 2005. A variável resposta foi o tempo de permanência na comunidade até a desistência do tratamento, considerando que cada residente permanece na Comunidade por um período máximo de 270 dias, sem qualquer contato com as drogas e quem alcança esta meta foi considerado, neste

trabalho, como um dado censurado. Estes dados foram analisados por Pascoa (2008).

O último conjunto de dados refere-se a dados de tempo até a soro-reversão de 143 crianças expostas ao HIV por via vertical, nascidas no Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto no período de 1995 a 2001, onde as mães não foram tratadas, retirados de Silva (2004) e Perdoná (2006).

Na Figura 1 apresenta-se a distribuição de frequência para cada conjunto de dados. Observa-se que todos os conjuntos de dados são assimétricos à direita. Como os dados em análise de sobrevivência são positivos, pois trata-se do tempo até a ocorrência do envento em estudo (falha), foi feito a reparametrização $y = log(t)$ onde $t > 0$.

Os softwares utilizados para estimativa dos parâmetros das distribuições foram o R Development Core Team (2009) e o procedimento nlmixed do SAS - Statistical Analysis System (Institute, 2004).

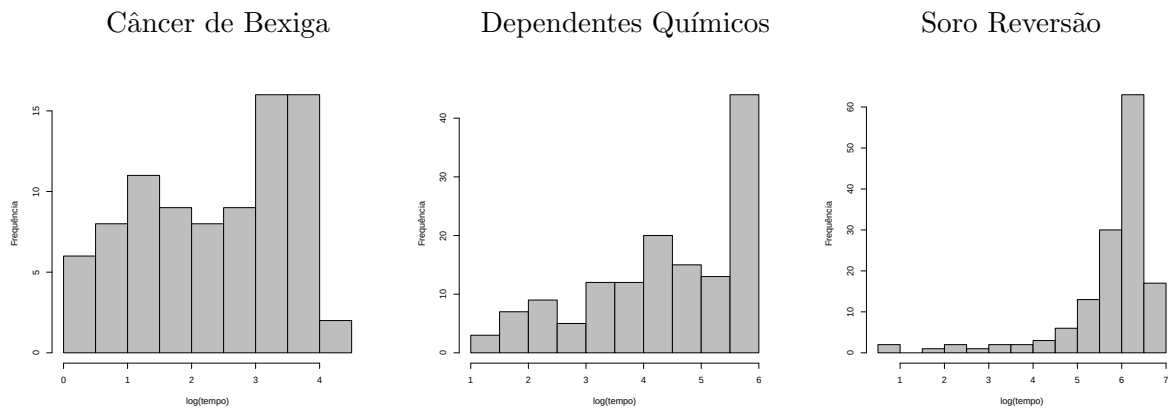

Figura 1: Histogramas referente aos conjuntos de dados de pacientes com cancer de bexiga, dependentes químicos e crianças expostas ao HIV.

A estimação dos parâmetros da distribuição LGI3, foi feita pelo método de máxima verossimilhança para dados censurados. Para que fosse possível realizar inferências fundamentadas no modelo, foi necessário obter a função de verossimilhança, que é expressa por, $L(b, \mu, \sigma; y) = \prod_{i=1}^{r} f(y_i; b, \mu, \sigma) \prod_{i=r+1}^{n} S(y_i; b, \mu, \sigma)$, onde $f(y_i; b, \mu, \sigma)$ e $S(y_i; b, \mu, \sigma)$ são a fdp e a função de sobrevivência da distribuição LGI3, $y = (y_1, ..., y_n)$ é uma amostra completa da LGI3, r é o número de falhas, F e C representam as observações sem censuras e censuradas respectivamente. Seja $\boldsymbol{\theta} = (b, \mu, \sigma)^T$, o logaritmo da função de verossimilhança do modelo é dado por:

$$l(\boldsymbol{\theta}) \quad = \quad r \log\left(\frac{b}{\sigma}\right) - b\sum_{i\in F} z_i - (b+1)\sum_{i\in F} \log\left(1 + e^{-z_i}\right) - b\sum_{i\in C} z_i - b\sum_{i\in C} \log\left(1 + e^{-z_i}\right),$$

Os componentes do vetor score $U = \left(\frac{\partial l}{\partial b}, \frac{\partial l}{\partial \mu}, \frac{\partial l}{\partial \sigma}\right)$ são obtidos por diferenciação dos parâmetros, logo:

$$
\begin{aligned}
U_a(\boldsymbol{\theta}) &= \frac{r}{b} - \sum_{i\in F} z_i - \sum_{i\in F} \log\left(1+e^{-z_i}\right) - \sum_{i\in C} z_i - \sum_{i\in C} \log\left(1+e^{-z_i}\right) \\
U_\mu(\boldsymbol{\theta}) &= \frac{nb}{\sigma} - \frac{(b+1)}{\sigma}\sum_{i\in F}\frac{e^{-z_i}}{1+e^{-z_i}} - \frac{b}{\sigma}\sum_{i\in C}\frac{e^{-z_i}}{1+e^{-z_i}} \\
U_\sigma(\boldsymbol{\theta}) &= -\frac{r}{\sigma} + \frac{b}{\sigma}\sum_{i\in F} z_i - \frac{1}{\sigma}\sum_{i\in F} z_i\frac{e^{-z_i}}{1+e^{-z_i}} - \frac{b}{\sigma}\sum_{i\in C} z_i - \sum_{i\in C} z_i\frac{e^{-z_i}}{1+e^{-z_i}}
\end{aligned}
$$

As estimativas dos parâmetros são soluções simultâneas das equações $U_b(\boldsymbol{\theta}) = 0$, $U_\mu(\boldsymbol{\theta}) = 0$ e $U_\sigma(\boldsymbol{\theta}) = 0$. O ajuste da distribuição LGI3, foi comparado com o ajuste das distribuições LGI2, LGI1 e a distribuição logística, por meio dos critérios AIC (Critério de Informação de Akaike), BIC (Critério de Informação Bayesiano) e CAIC (Critério de Informação Akaike Consistente). As análises foram implementadas no software estatístico R (R Development Core Team, 2009).

3 Resultados

Na Tabela 1, 2 e 3 são apresentadas as estimativas de máxima verossimilhança de cada parâmetro com os respectivos erros padrões (entre parenteses) e os valores das estatísticas AIC, BIC e CAIC, para as distribuições comparadas.

Tabela 1: Ajuste dos modelos comparados, para os dados de câncer de bexiga.

Modelo	b	μ	σ	AIC	BIC	CAIC
LGI3	0,0725 (0,0368)	0,4629 (0,2030)	0,2508 (0,0932)	235,5	242,8	235,7
LGI2	0,0268 (0,0039)	-	0,1145 (0,0018)	237,6	242,5	238,8
LGI1	0,2215 (0,0323)	-	-	258,5	261,0	258,6
Logístico	-	2,9537 (0,2236)	1,0651 (0,1280)	247,2	252,1	247,4

Tabela 2: Ajuste dos modelos comparados, para os dados de dependentes químicos.

Modelo	b	μ	σ	AIC	BIC	CAIC
LGI3	0,4363 (0,1888)	3,4784 (0,5501)	0,7893 (0,1515)	473,1	481,9	473,2
LGI2	0,0272 (0,0027)	-	0,1594 (0,0009)	561,0	566,9	561,1
LGI1	0,1617 (0,0162)	-	-	569,7	572,7	569,7
Logístico	-	4,6244 (0,1540)	1,0290 (0,0866)	473,5	479,4	473,6

Tabela 3: Ajuste dos modelos comparados, para os dados de soro reversão.

Modelo	b	μ	σ	AIC	BIC	CAIC
LGI3	18,5417 (16,7206)	7,4712 (0,4264)	0,4096 (0,0318)	207,7	216,6	207,8
LGI2	0,0301 (0,0027)	-	0,2029 (0,0010)	700,5	706,5	700,6
LGI1	0,1458 (0,0133)	-	-	700,5	7003,4	700,5
Logístico	-	6,1200 (0,4524)	0,3042 (0,0239)	225,3	231,2	225,4

Observa-se, nas Tabelas 1, 2 e 3, que a distribuição LGI3 apresentou-se melhor entre as distribuições comparadas para o ajuste dos três conjuntos de dados em estudo, com menor valores de AIC, BIC e CAIC. Para os dados de câncer de bexiga a distribuição LGI1 obteve os

maiores valores das estatísticas seguida da distribuição logística. A distribuição LGI2 obteve valores de AIC, BIC e CAIC bem próximos da distribuição LGI3 tendo valor de BIC menor que da distribuição LGI3. A distribuição LGI3 obteve valores bem próximos da distribuição logística no ajuste dos dados de dependentes químicos sendo que a distribuição logística teve menor valor de BIC. Novamente o modelo LGI1 obteve os maiores valores para as estatísticas em análise seguido da distribuição LGI2. E para o dados de soro reversão as distribuições LGI1 e LGI2 obtiveram valores próximos para as estatísticas porém altos.

Na Figura 2, são apresentados os gráficos da função de sobrevivência empírica e da função de sobrevivência estimada calculada para as quatros distribuições. Observa-se na Figura 2 que a distribuição LGI3 apresentou melhor ajuste para os três conjuntos de dados, confirmando os resultados expostos na Tabela 1, 2 e 3.

Observa-se que o parâmetro de locação μ dá maior flexibilidade a distribuição melhorando o ajuste em relação as distribuições LGI1 e LGI2 e o parâmetro b proporciona um ajuste mais refinado comparando a distribuição logística.

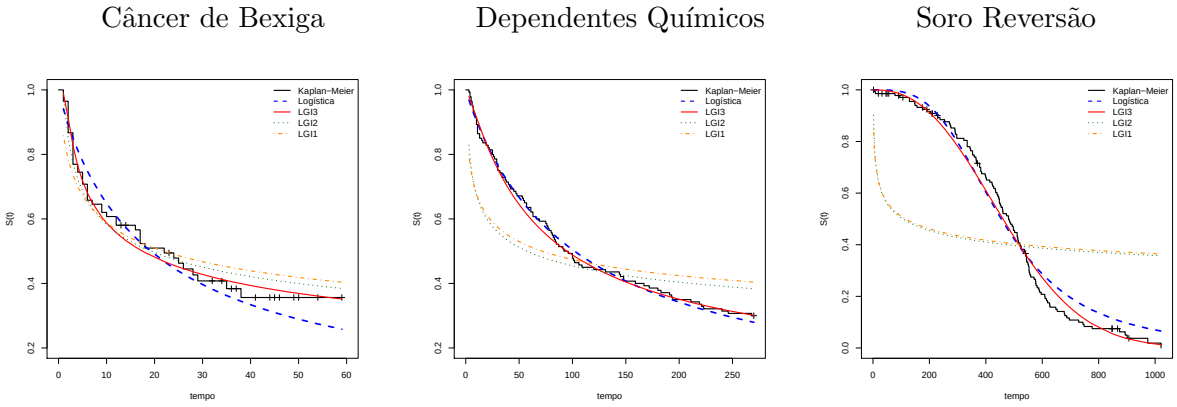

Figura 2: Gráficos da função de sobrevivência empírica e das funções de sobrevivência estimadas pelas distribuições LGI3, LGI2, LGI1 e logística para cada conjunto de dados analisado.

4 Conclusões

A distribuição logística generalizada tipo 1 com três parâmetros, que tem como caso particular as distribuições LGI2, LGI1 e a distribuição logística, apresentou melhor ajuste para os dados em estudo (menor AIC, BIC e CAIC), quando comparada com as demais distribuições apresentadas no trabalho, se mostrando mais flexível em relação a elas.

O parâmetro de locação μ confere a distribuição maior flexibilidade em relação as distribuições LGI1 e LGI2 melhorando grandemente o ajuste dos dados em relação as distribuições LGI1 e LGI2, quando o conjunto de dados é fortemente assimétrico a direita e o parâmetro b proporciona a distribuição LGI3 um ajuste mais refinado comparando a distribuição logística.

A distribuição logística generalizada tipo I com três parâmetros é uma boa candidata para modelar dados com assimetria a direita e apresenta melhor ajuste que a distribuição logística.

Referências

ALKASASBEH, M. R.; RAQAB, M. Z. Estimation of the generalized logistic distribution parameters: Comparative study. **Statistical Methodology**, v.6, p.262–279, 2009.

BALAKRISHNAN, N. **Hand Book of the Logistic Distribution**. New York: Dekker, 1992.

BALAKRISHNAN, N.; LENG, M. Y. Order statistics from the type I generalized logistic distribution. **Communications in Statistical - Simulation and Computation**, v.17, n.1, p.25–50, 1988.

COLLET, D. **Modeling survival data in medical research**. Second edition ed. London: Chapman and Hall, 2003. 320p.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de Sobrevivência Aplicada**. São Paulo: Edgard Blücher, 2006.

INSTITUTE, S. **SAS/STAT User's Guide: Version 9**. Cary: SAS Institute, 2004.

PASCOA, M. A. R. Intervalos de credibilidade para razão de riscos do modelo de cox, considerando estimativas pontuais bootstrap, 2008. Dissertação - Departamento de Ciências Exatas, Universidade Federal de Lavras. Mestrado em Estatística e Experimentação Agropecuária.

PERDONÁ, G. S. C. Modelos de riscos aplicados à análise de sobrevivência. São Carlos - SP, 2006. Tese - Instituto de Ciências Matemática e de Computação - ICMC-USP. Tese apresentada ao Instituto de ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências - Ciências de Computação e Matemática Computacional.

R DEVELOPMENT CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2009.

SILVA, A. N. F. Estudo Evolutivo das Crianças Expostas ao HIV e Notificadas pelo Núcleo de Vigilância Epidemiológica do HCFMRP-USP. Ribeirão Preto, 2004. Dissertação de mestrado - Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo.