Revista da Estatística UFOP, Vol III(3), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS
Trabalhos apresentados na versão poster.

A distribuição beta exponencial generalizada na modelagem de dados em análise de sobrevivência

Rondiny Moreira Carneiro [1]
Marcelino Alves Rosa de Pascoa [2]
Anderson Castro Soares de Oliveira [3]
Neuber José Segri [4]
Tiago Almeida de Oliveira [5]

1 Introdução

A distribuição beta exponencial generalizada (BEG), proposta recentemente por Barreto-Souza *et al.* (2010), é útil para analisar dados assimétricos. Este modelo tem quatro parâmetros e os seus submodelos são as distribuições beta exponencial (BE), exponencial generalizada (EG) e exponencial. Estes possuem comprovada adequação a dados envolvendo várias situações práticas, como na área média, industrial e econômica. Uma importante vantagem da distribuição BEG, é acomodar não apenas funções de risco monótonas, isto é, funções crescentes, decrescentes ou constantes, mas também as funções de risco não-monótonas, como a forma unimodal.

A função densidade de probabilidade, de sobrevivência e de risco são dadas, respectivamente por

$$f(x) = \frac{\alpha\lambda}{B(a,b)}\, e^{-\lambda x}\left(1-e^{-\lambda x}\right)^{\alpha a-1}\left\{1-\left(1-e^{-\lambda x}\right)^{\alpha}\right\}^{b-1},$$

$$S(x) = 1 - \frac{1}{B(a,b)}\int_0^{\left(1-e^{-\lambda x}\right)^{\alpha}} w^{a-1}(1-w)^{b-1}dw,$$

$$h(x) = \frac{\alpha\lambda e^{-\lambda x}\left(1-e^{-\lambda x}\right)^{\alpha a-1}\left\{1-\left(1-e^{-\lambda x}\right)^{\alpha}\right\}^{b-1}}{B(a,b)I_{1-\left(1-e^{-\lambda x}\right)^{\alpha}}(a,b)},$$

em que $x > 0$; $a > 0$, $b > 0$ e $\alpha > 0$ são parâmetros de forma e $\lambda > 0$ é parâmetro de escala.

[1]Graduando em estatística - UFMT. e-mail: *rondinycarneiro@gmail.com*
[2]Dest - UFMT. e-mail: *marcelino.pascoa@gmail.com*
[3]Dest - UFMT. e-mail: *andersoncso@gmail.com*
[4]Dest - UFMT. e-mail: *professor.neuber@gmail.com*
[5]UEPB. e-mail: *tiagoestatistico@gmail.com*

A finalidade deste trabalho foi utilizar o método de máxima verossimilhança para estimar os parâmetros da distribuição BEG, utilizando dados de tempo de permanência no Japão e comparar seu ajuste com o das distribuições BE e EG, utilizando as estatísticas AIC (Critério de Informação de Akaike), BIC (Critério de Informação Bayesiano) e CAIC (Critério de Informação Akaike Consistente) e o teste da razão de verossimilhança (TRV).

2 Material e métodos

Os dados foram obtidos por meio de uma pesquisa eletrônica (e-survey) Babbie (1999), que buscou obter de características, ações ou opiniões do grupo de alunos utilizando a internet como ferramenta. A pesquisa foi realizada no primeiro semestre de 2010, por meio de um site reservado, apenas para acesso dos acadêmicos, em que foram obtidos 246 questionários. Deste apenas 150 foram utilizados para analise, pois haviam alunos de outras nacionalidades. Foi considerado como variável de estudo o tempo de permanência no Japão, sendo este contado a partir da chegada pela primeira vez até o presente momento, foram censurados alunos que tenham retornado para o Brasil ao menos uma vez.

O comportamento da função risco foi observado por meio da construção do gráfico do tempo total em teste (curva TTT), proposto por Aarset (1987). A curva TTT é obtida construindo um gráfico de

$$G\Big(\frac{r}{n}\Big) = \frac{\sum_{i=1}^{r} T_{i:n} + (n-r)T_{r:n}}{\sum_{i=1}^{n} T_{i:n}} \quad \text{por} \quad \frac{r}{n},$$

em que n é o tamanho da amostra, $r = 1,\ldots,n$ e $T_{i:n}, i = 1,\ldots,n$ são estatísticas de ordem da amostra.

A estimação dos parâmetros foi feita pelo método de máxima verossimilhança. Para que fosse possível realizar inferências fundamentadas no modelo, foi necessário obter a função de verossimilhança, que é expressa por, $L(\alpha,\lambda,a,b;x) = \prod_{i\in F} f(x_i;\alpha,\lambda,a,b)\prod_{i\in C} S(x_i;\alpha,\lambda,a,b)$, em que $f(x_i;\alpha,\lambda,a,b)$ e $S(x_i;\alpha,\lambda,a,b)$ são a função densidade de probabilidade e de sobrevivência da distribuição BEG, F representa as observações que falharam e C representa as observações que foram cencuradas. Sendo $\theta = (\alpha,\lambda,a,b)^T$, o logaritmo da função de verossimilhança do modelo paramétrico para uma única observação x de X é representado por

$$\begin{aligned} L(\theta) &= \log(\alpha) + \log(\lambda) - \log[(a,b)] - \lambda x + (\alpha a - 1)\log\left(1 - e^{-\lambda x}\right) \\ &+ (b-1)\log\left\{1 - \left(1 - e^{-\lambda x}\right)^{\alpha}\right\}, \qquad x > 0. \end{aligned}$$

Os componentes do vetor score $U = \left(\frac{\partial l}{\partial a}, \frac{\partial l}{\partial b}, \frac{\partial l}{\partial \lambda}, \frac{\partial l}{\partial \alpha}\right)$ são obtidos por diferenciação de θ em

relação aos parâmetros, logo

$$\begin{aligned}
U_a(\theta) &= \alpha \log\left(1-e^{-\lambda x}\right) - \psi(a) + \psi(a+b), \\
U_b(\theta) &= \log\left\{1-\left(1-e^{-\lambda x}\right)^{\alpha}\right\} - \psi(a) + \psi(a+b), \\
U_\lambda(\theta) &= \frac{1}{\lambda} - x + \frac{(\alpha a - 1)xe^{-\lambda x}}{1-e^{-\lambda x}} - \frac{\alpha(b-1)e^{-\lambda x}\left(1-e^{-\lambda x}\right)^{\alpha-1} x}{1-\left(1-e^{-\lambda x}\right)^{\alpha}}, \\
U_\alpha(\theta) &= \frac{1}{\alpha} + \alpha \log\left(1-e^{-\lambda x}\right) - \frac{(b-1)\left(1-e^{-\lambda x}\right)^{\alpha} \log\left(1-e^{-\lambda x}\right)}{1-\left(1-e^{-\lambda x}\right)^{\alpha}},
\end{aligned}$$

$\psi(.)$ é a função digama.

Consequentemente, o estimador de máxima verossimilhança (EMV) $\widehat{\theta}$ de θ é obtido numericamente a partir das equações não lineares,

$$U_a(\theta) = U_b(\theta) = U_\lambda(\theta) = U_\alpha(\theta) = 0.$$

O ajuste da distribuição BEG, foi comparado com o das distribuições BE e EG, por meio das estatísticas AIC, BIC, CAIC e pelo TRV. As análises foram implementadas no *software R* (*R Development Core Team*, 2013).

3 Resultados e discussões

A Curva TTT para o conjunto de dados de tempo de permanência no Japão, encontra-se na Figura 1(a) e indica uma função risco na forma crescente. Assim, para analisar esse conjunto de dados, a distribuição BEG pode ser utilizada.

Na Tabela 1, podem ser vistos as EMVs (e os correspondentes erros-padrão que estão entre parênteses) dos parâmetros e os valores das estatísticas dos modelos, BEG, BE e EG. Os resultados indicam que o modelo BEG tem os menores valores de AIC, BIC e CAIC entre os modelos ajustados, portanto, o modelo BEG é o mais adequado para os dados de tempo de permanência no Japão. O TRV é apresentado na Tabela 2. Os resultados nessa tabela sugerem que o modelo BEG produz um ajuste mais adequado a esses dados quando comparado com as outras duas distribuições.

A Figura 1(b) apresenta a comparação das estimativas da função de sobrevivência segundo Kaplan-Meier e segundo os modelos BEG, BE e EG, para os dados de tempo de permanência no Japão. Observa-se pela figura que a distribuição BEG nos fornece um ajuste satifatório para os dados em estudo.

(a) (b)

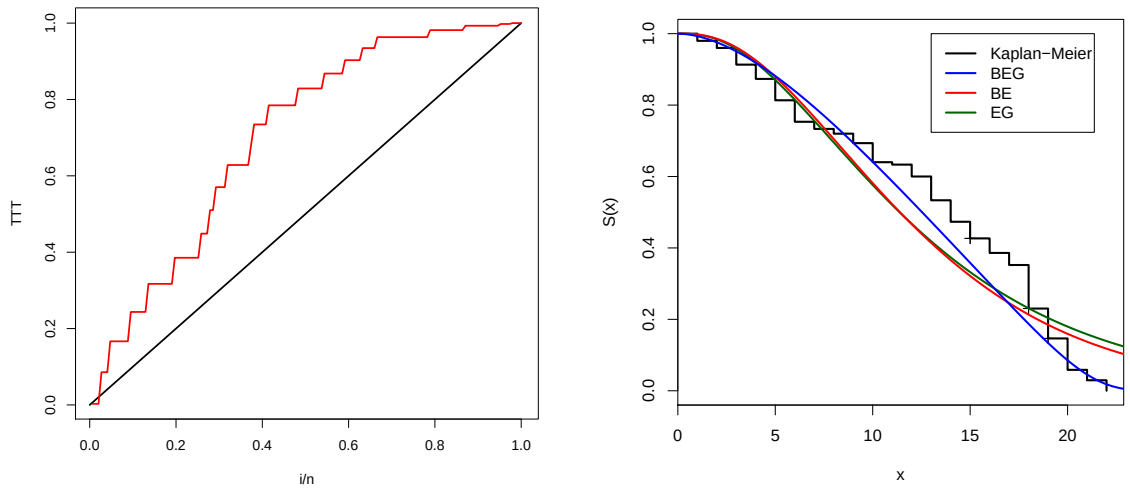

Figura 1: (a) Curva TTT para os dados de tempo de permanência no Japão. (b) Estimativas da função de sobrevivência segundo Kaplan-Meier e segundo os modelos BEG, BE e EG, para os dados de tempo de permanência no Japão.

Tabela 1: Ajuste final dos modelos comparados, para os dados de tempo de permanência no Japão.

Modelo	λ	α	a	b	AIC	BIC	CAIC
BEG	0.0397	24.3520	0.0760	5.0790	915.1	927.2	915.4
	(0.0127)	(9.1363)	(0.0242)	(0.5840)			
BE	0.0027	1	2.8586	81.4000	979.5	988.6	979.7
	(0.0002)	$(-)$	(0.0223)	(0.4057)			
EG	0.1356	2.8823	1	1	994.6	994.7	1000.6
	(0.0110)	(0.3633)	$(-)$	$(-)$			

Tabela 2: Teste da razão de verossimilhança, para os dados de tempo de permanência no Japão.

Modelo	Estatística do teste	Valor p
BEG vs BE	66.4	< 0.0001
BEG vs EG	83.5	< 0.0001

4 Conclusão

Pelos critérios utilizados verificou-se que a distribuição BEG obteve o melhor ajuste, quando comparada com seus submodelos. Sendo assim é possível considerar que a principal vantagem da distribuição BEG é a flexibilidade da sua função risco, de forma a deixar o modelo mais flexível e com a possibilidade de ajustar variados conjuntos de dados em análise de sobrevivência.

Bibliografia

AARSET, M.V. How to identify bathtub hazard rate, **IEEE Transactions on Reliability**, v. 36, p. 106-108, 1987.

BABBIE, E. **Métodos de pesquisas de Survey/Earl Babbie**; tradução de Guilherme Cezarino - Belo Horizonte: Ed. UFMG, 1999. 519p. - (Coleção Aprender) Tradução de: Survey research methods.

BARRETO-SOUZA, W.; SANTOS, A. H. S.; CORDEIRO, G. M. The beta generalized exponential distribution. **Journal of Statistical Computation and Simulation**. v. 80, p. 159-172, 2010.

R DEVELOPMENT CORE TEAM. **R**: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Disponível em: <http://www.R-project.org>. Acesso em: 01 out. 2013.