A Distribuição Gama Generalizada na Modelagem de Dados de Tempo Até Reincidência de uma Multa

Fabio Douglas Soares Bezerra[1], Daniel Valentins de Lima[2], Evandro Mariano Barros da Silva[3], Marcelino Alves Rosa de Pascoa[4], Graziela Dutra Rocha Gouvea[5]

Introdução

A distribuição Gama Generalizada (GG) proposto por Stacy em (1962), contém a característica de três parâmetros sendo eles todos positivos, sendo eles: γ, k e α. Sendo uma distribuição que modela com flexibilidade dados de sobrevivência, pois acomoda inúmeras variações da função de risco (COLOSIMO; GIOLO, 2006).

Através da distribuição GG também é possível obter outras distribuições tais como: Weibull, Gama e Log-Normal. Para uma variável aleatória T com distribuição de GG tem-se a função de densidade de probabilidade, de sobrevivência e de risco dadas, respectivamente por:

$$g(t)=\frac{\tau}{\Gamma(k)\alpha^{\tau k}}t^{\tau k-1}\exp\left\{-\left(\frac{t}{\alpha}\right)^{\tau}\right\},\ t>0.$$

Em que $\Gamma(k)$ é a função gama definida por:

$$\Gamma(k)=\int_0^{\infty}t^{k-1}e^{-t}dt\,.$$

Se T é uma variável aleatória positiva e tem distribuição de probabilidade GG com parâmetros α, τ e k, então denota-se T ~ GG(α,τ,k). As funções de distribuição acumulada $G(t)$, de sobrevivência $S(t)$ e de taxa de falha $h(t)$ são escritas, respectivamente por:

$$G(t)=P[T\le t]=\frac{1}{\Gamma(k)}\int_0^{\left(\frac{t}{\alpha}\right)^{\tau}}u^{k-1}e^{-u}du=\left\{\gamma_1\left[k,\left(\frac{t}{\alpha}\right)^{\tau}\right]\right\},$$

$$S(t)=1-G(t)=1-\left\{\gamma_1\left[k,\left(\frac{t}{\alpha}\right)^{\tau}\right]\right\}\text{ e}$$

[1] UFMT. email: *email*.fabiodouglas.soares@gmail.com
[2] UFMT. email: *email*.dvalentins@outlook.com
[3] UFMT. email: *email*.evandromarina_barros@live.com
[4] UFMT. email: *email*. marcelino.pascoa@gmail.com
[5] UFOP. email: *email*. gragouvea@gmail.com

$$h(t)=\frac{g(t)}{S(t)}=\frac{t^{\tau k-1}\exp\left\{-\left(\frac{t}{\alpha}\right)^{\tau}\right\}}{\int_0^{\infty}x^{\tau k-1}\exp\left\{-\left(\frac{x}{\alpha}\right)^{\tau}\right\}dx}$$

Sendo $\gamma(k,x)=\int_0^x u^{k-1}e^{-u}du$ é a função gama incompleta e $\gamma_1(k,x)=\frac{\gamma(k,x)}{\Gamma(k)}=\frac{1}{\Gamma(k)}\int_0^x w^{k-1}e^{-w}dw$ é a razão da função gama incompleta.

Objetivo

O objetivo deste trabalho foi através do método de máxima verossimilhança, estimar os parâmetros da distribuição de GG, utilizando dados de tempo até a reincidência de uma multa e comparar seu ajuste com o estimador de Kaplan-Meier.

Metodologia

Os dados foram disponibilizados pela Secretaria de Mobilidade Urbana, oriundos de registros de multas originadas através de radar eletrônico e registros manuais realizados por agentes de trânsito no ano de 2017. Foram avaliadas 34614 ocorrências onde o infrator cometeu uma nova infração de trânsito, ou seja, a variável resposta no estudo foi o tempo até a primeira reincidência após o indivíduo sofrer a primeira multa.

Uma forma empírica de determinar o comportamento da função risco se da 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

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

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, logo considere uma amostra $X = (x_1, \ldots, x_n)$ de n observações independentes, e seja F o conjunto do logaritmo dos tempos de vida e C o conjunto do logaritmo dos tempos de censura. Assim, o logaritmo da função de verossimilhança para o vetor de parâmetros $\theta=(\alpha,\tau,k)^T$ para o modelo (1) tem a forma $l(\theta)=\sum_{i\in F}l_i(\theta)+\sum_{i\in C}l_i^{(c)}(\theta)$, em que $l_i(\theta)=\log\left[f(x_i)\right]$,

$l_i^{(c)}(\boldsymbol{\theta}) = \log\left[S(x_i)\right]$, $f(x_i)$ é a função densidade de probabilidade (1) e $S(x_i)$ é a função de sobrevivência (2). Dessa forma, o logaritmo da função de verossimilhança para θ é

$$l(\boldsymbol{\theta}) = r\log(\tau) - r\log[\alpha\Gamma(k)] + (\tau k - 1)\sum_{i\in F}\log\left(\frac{t_i}{\alpha}\right) - \sum_{i\in F}\left(\frac{t_i}{\alpha}\right)^{\tau} + \sum_{i\in C}\log\left(1 - \left\{\gamma_1\left[k, \left(\frac{t_i}{\alpha}\right)^{\tau}\right]\right\}\right)$$

$$+ \sum_{i\in C}\log\left(1 - \left\{\gamma_1\left[k, \left(\frac{t_i}{\alpha}\right)^{\tau}\right]\right\}\right), \qquad (1)$$

em que r é o número de falhas, *F* e *C* denotam o conjunto de observações não censuradas e censuradas, respectivamente.

As derivadas de primeira ordem de (1) em relação aos parâmetros em $\boldsymbol{\theta}$ são dadas por

$$U_\alpha(\boldsymbol{\theta}) = -\frac{r\tau k}{\alpha} + \frac{\tau}{\alpha}\sum_{i=1}^{r} u_i + \sum_{i=r+1}^{n}\frac{-\tau(u_i)^k\exp(-u_i)}{\alpha\Gamma(k)\{1-[\gamma_1(k,u_i)]\}},$$

$$U_\tau(\boldsymbol{\theta}) = \frac{r}{\tau} + \frac{k}{\tau}\sum_{i=1}^{r}\log(u_i) - \frac{1}{\tau}\sum_{i=1}^{r}u_i\log(u_i) + \sum_{i=r+1}^{n}\frac{(u_i)^k\exp(-u_i)\log(u_i)}{\tau\Gamma(k)\{1-[\gamma_1(k,u_i)]\}},$$

$$U_k(\boldsymbol{\theta}) = -r\psi(k) + \sum_{i=1}^{r}\log(u_i) + \sum_{i=r+1}^{n}\left(\frac{-\psi(k)\,\gamma_1(k,u_i)}{\{1-[\gamma_1(k,u_i)]\}} + \frac{[\dot{\gamma}(\mathrm{k},\mathrm{u_i})]_\mathrm{k}}{\Gamma(k)\{1-[\gamma_1(k,u_i)]\}}\right),$$

em que, $\psi(.)$ é a função digama, $u_i = \left(\frac{t_i}{\alpha}\right)^{\tau}$, $[\dot{\gamma}(\mathrm{k},\mathrm{u_i})]_\mathrm{k} = \int_0^{u_i} w^{k-1}\exp(-w)\log(w)dw$ e n é o tamanho da amostra.

Consequentemente, o estimador de máxima verossimilhança (EMV) $\widehat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$é obtido numéricamenteapartir das equações não lineares

$$U_\alpha(\boldsymbol{\theta}) = U_\tau(\boldsymbol{\theta}) = U_k(\boldsymbol{\theta}) = 0.$$

Para a estimação de intervalo e testes de hipóteses sobre os parâmetros do modelo é necessário obter a matriz 3 × 3 de informação observada

$$J=J(\boldsymbol{\theta})=\begin{pmatrix} j_{\alpha,\alpha} & j_{\alpha,\tau} & j_{\alpha,k} \\ j_{\tau,\alpha} & j_{\tau,\tau} & j_{\tau,k} \\ j_{k,\alpha} & j_{k,\tau} & j_{k,k} \end{pmatrix}$$

O ajuste da distribuição GG, foi comparado com o da distribuição Exponencial, por meio do teste da razão de verossimilhança (TRV). Pode-se usar a estatística TRV para checar se o ajuste usando a distribuição GG é estatisticamente "superior" ao ajuste usando a distribuição Exponencial para um determinado banco de dados. Em qualquer caso, o teste de hipóteses do tipo H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus H_1: $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, onde $\boldsymbol{\theta}_0$ é um vetor especificado, pode ser realizado usando a estatística TRV. O teste de H_0: $k = \tau = 1$ *versus* H_1: H_0 *não é verdade*, é equivalente a comparar a distribuição GG com a distribuição Exponencial para o qual a estatística TRV reduz a

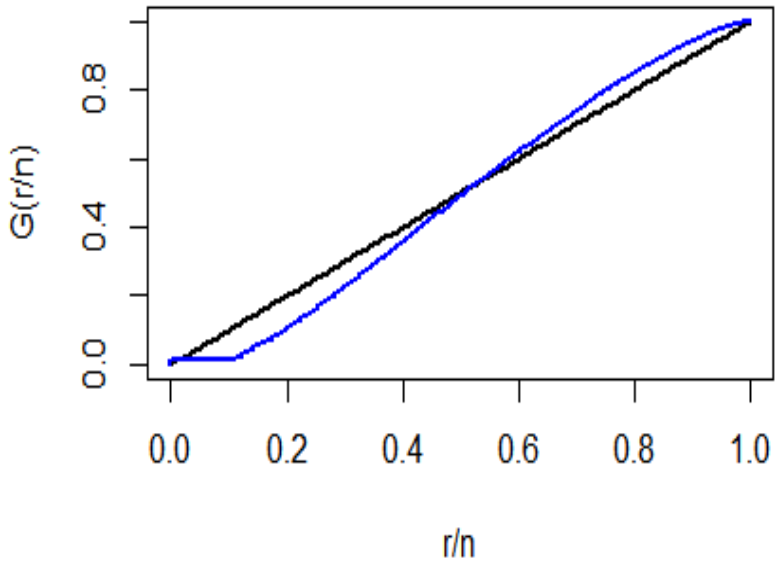
$$w = 2\left[\ell\left(\hat{\alpha},\hat{\tau},\hat{k}\right) - \ell(\tilde{\alpha},1,1)\right],$$

em que $\hat{\alpha}$, $\hat{\tau}$ e $\hat{k}$ são os estimadores de máxima verossimilhança sobre H_1 e $\tilde{\alpha}$ e $\tilde{\tau}$, são os estimadores sobre H_0. Foram utilizadas as estatísticas AIC, BIC e CAIC para a seleção dos modelos. As análises foram implementadas no *software* estatístico R.

Resultados e Discussões

A Curva TTT para o conjunto de dados de tempo até a ocorrência de uma nova multa encontra-se na Figura 1 e indica uma função risco na forma de "U". Assim, como a distribuição GG modela tal função de risco ela é apropriada para analisar esse conjunto de dados.

Figura 1: Curva TTT para dados de tempo até reincidência de multa.



Na Tabela 1, podem ser vistos as estimativas de máxima verossimilhança (e os correspondentes erros-padrão que estão entre parênteses) dos parâmetros e os valores das estatísticas dos modelos, Exponencial e GG. Os resultados indicam que o modelo GG tem os menores valores das estatísticas AIC (Critério de Informação de Akaike), BIC (Critério de Informação Bayesiano) e CAIC (Critério de Informação Akaike Consistente) entre os modelos ajustados, portanto, o modelo GG é o mais adequado para os dados de tempo até a reincidência de uma multa. O teste da razão de verossimilhança é apresentado na Tabela 2. Os resultados nessa tabela sugerem que o modelo GG produz um ajuste mais adequado a esses dados quando comparado com a distribuição exponencial.

Tabela 1: Ajuste final dos modelos comparados, para os dados de tempo até a reincidência de uma multa.

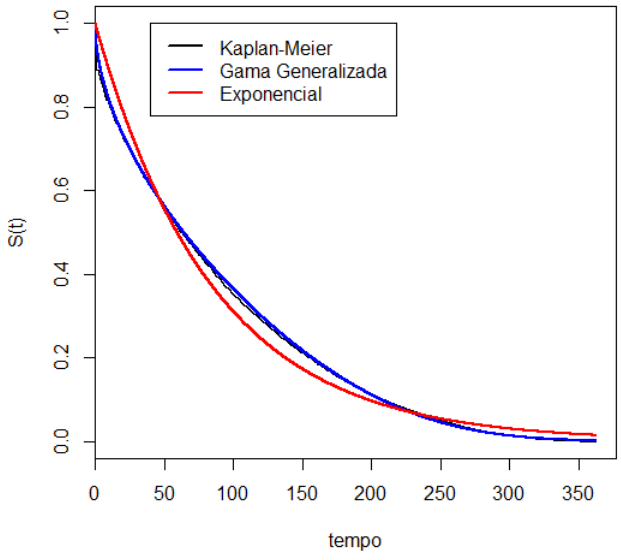
Modelo	α	τ	k	AIC	CAIC	BIC
GG	258,5869	3,6567	0,1490	371290,5	371290,5	371315,9
	(1,7178)	(0,0894)	(0,0042)			
Exponencial	85,5140	1	1	377203,3	377203,3	377211,7
	(0,4596)	(-)	(-)			

Tabela 2: Teste da razão de verossimilhança, para os dados de tempo até o desmame de suínos

Modelo	Estatística do Teste	Valor p
GG *vs*Exponencial	5.916,8	< 0,0001

A Figura 2 apresenta a comparação das estimativas da função de sobrevivência segundo Kaplan-Meier e segundo os modelos GG e exponencial, para os dados de tempo até reincidência de multa. Observa-se pela figura que a distribuição GG nos fornece um ajuste satisfatório para os dados em estudo. Logo, conclui-se que em média o tem até a reincidência de uma multa é de 87 dias. Com 200 dias 89% dos carros já haviam sido multados.

Figura 2: Estimativas da função de sobrevivência segundo Kaplan-Meier e segundo os modelos GG e Exponencial, para os dados de tempo até reincidência de multa.



Conclusão

A distribuição Gama Generalizada proposta por Stacy (1962), apresentou melhor ajuste para os dados em estudo, segundo o teste da razão de verossimilhança e as estatísticas AIC, BIC e CAIC, quando comparada com a distribuição exponencial, se mostrando mais flexível. Com uma estimativa em relação ao tempo médio de ocorrência de uma eventual multa e um estudo sobre qual tipo de agravamento de trânsito é mais recorrente, será possível agir com políticas de transito a fim de haja uma redução em tais infrações.

Agradecimentos

Agradeço primeiramente a Deus e a minha esposa pela motivação a apreço ao estudo, bem como ao professor Marcelino e a Secretaria de Mobilidade Urbana na pessoa do servidor Michel Diniz que nos concedeu a oportunidade de realizar este trabalho.

Referências Bibliográficas

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

COLOSIMO, E. A.; GIOLO, S.R., **Análise de Sobrevivência Aplicada.** Projeto Fisher – ABE. São Paulo: Edgard Blucher Ltda., 1º Edição 89 p., 2006.

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. ISBN 3-900051-07-0, URL http://www.R-project.org/.

STACY, E.W. A generalization of the gamma distribution. The Annals of Mathematical Statistics, New York, v.33, p.409-419, 1962.