Estimation in Generalized Log-gamma Regression Model for Interval-censored Data

Elizabeth M. Hashimoto[1] **Edwin M. M. Ortega**[2]
ESALQ/USP - Departamento de Ciências Exatas
Gauss M. Cordeiro[3]
UFRPE - Departamento de Estatística e Informática
Vicente G. Cancho[4]
ICMC/USP - Departamento de Matemática Aplicada

Abstract

Interval-censored survival data, in which the event of interest is not observed exactly but it is only known to occur within some time interval, appear very frequently. In this study, we shall be concerned only with parametric forms, so a location-scale regression model based on the generalized log-gamma distribution is proposed for modeling interval-censored data. We also show that the model proposed with interval-censored data is interesting as it represents a parametric family of models that has, as special sub-models, other regression models which are broadly used in lifetime data analysis. Assuming interval-censored data, we consider a frequentist analysis, a Jackknife estimator and non-parametric bootstrap, for the parameters of the model.

Keywords: Interval-censored data; Generalized gamma distribution; Regression model; Estimation.

1 Introduction

In several studies, survival response can be interval-censored, such that the event of interest is not observed exactly but, it is only known to occur within some time intervals that may overlap and vary in length. The literature presents many applications of survival models for interval-censored data considering the Weibull family of distributions (see, e.g., Lawless, 2003). This family is very suitable in situations where the failure rate function is constant or monotone. However, it is not suitable in situations where the failure rate function presents a bathtub or unimodal shape. To cope with these situations, several distributions derived from the Weibull distribution to exhibit bathtub-shaped or unimodal failure rate functions, were developed, one of which is the generalized gamma (GG) distribution (Stacy, 1962).

In some situations, the times of the events of interest T may only be known to have occurred within an interval of time, say $[U, V]$, where $U \leq T \leq V$. This can occur in a clinical trial, for example, when patients are assessed only at pre-scheduled visits. If the event has not occurred at one visit (at time U) but has by the following visit (at time V), T is known only to be within the interval $[U, V]$. These are known as interval-censored data. Note that exactly observed, right and left-censored data are special cases of interval-censored data, with $U = V$ for exactly observed data, $V = \infty$ for right-censored and $U = 0$ for left-censored observations.

This study examines the statistical inference aspects for modeling interval-censored data by using the generalized log-gamma (GLG) regression model. The inferential part was carried out using the asymptotic distribution of the maximum likelihood estimators, which for situations where the sample size is small, normality is more difficult to be justified. As an alternative for frequentist analysis, we explored the use of Jackknife estimator for the GLG regression model for interval-censored data. A punctual and interval estimation methodology based on bootstrap re-sampling methods is also proposed.

[1]Aluna de doutorado da PPG em Estatística e Experimentação Agronômica - ESALQ/USP - Bolsista CNPq.
[2]Docente do departamento de Ciências Exatas - ESALQ/USP.
[3]Docente do departamento de Estatística e Informática - UFRPE.
[4]Docente do departamento de Matemática Aplicada - ICMC/USP.

2 The generalized log-gamma regression models for interval-censored data

We assume that the random variable T follows a GG distribution with parameters $(\alpha, \tau, k)^\top$. The probability density function (pdf) of the GG distribution is given by

$$f(t;\alpha,\tau,k) = \frac{\tau}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{\tau k-1}\exp\left[-\left(\frac{t}{\alpha}\right)^{\tau}\right], \quad t,\alpha,\tau,k>0, \tag{1}$$

where $\Gamma(.)$ is the gamma function, α and τ are shape parameters and k is the scale parameter. The survival function reduces to $S(t;\alpha,\tau,k) = 1 - Q\left[k,\left(\frac{t}{\alpha}\right)^{\tau}\right]$, where $Q(k,x) = \frac{1}{\Gamma(k)}\int_0^x u^{(k-1)}\mathrm{e}^{-u}du$, is the incomplete gamma integral. The GG family is very flexible and includes several well-known models as sub-models (see, Johnson et al., 1994). The sub-models of the GG distribution thus considered in the literature are: exponential ($k=\tau=1$), gamma for ($\tau=1$) and Weibull for ($k=1$). The lognormal distribution is also obtained as a limiting distribution when $k\to\infty$. By letting $\tau=2$ we obtain a sub-model of the GG distribution which is known as the generalized normal distribution, GN$(2k,\alpha)$. The GN distribution is itself a flexible family and includes half-normal ($k=0.5$), Rayleigh ($k=1$), Maxwell-Boltzmann ($k=3/2$), and Chi ($k=\nu/2, \nu=1,2,\ldots$). The hazard function is given simply by $h(t;\alpha,\tau,k) = f(t;\alpha,\tau,k)/S(t;\alpha,\tau,k)$. The great flexibility of this model to fit lifetime data is due to the different forms that the hazard function can take, that is, (i) if $\tau>1$ and $k=1$, then the hazard function is monotonically increasing; (ii) if $\tau<1$ and $k=1$, then the hazard function is monotonically decreasing; (iii) if $1<\tau<1/k$ and $k<1$, then the hazard function is bathtub-shaped and, (iv) if $1/k<\tau<1$ and $k>1$, then we have a unimodal hazard function (Ortega et al., 2009).

Let T be a random variable following the GG density function (1). The random variable $Y=\log(T)$ follows a GLG distribution with density function parameterized in terms of $\mu = \log(\alpha) + \tau^{-1}\log(\lambda^{-2})$, $\sigma = \frac{1}{\tau\sqrt{k}}$ and $\lambda = \frac{1}{\sqrt{k}}$ is given by

$$f(y;\lambda,\sigma,\mu) = \begin{cases} \frac{|\lambda|}{\sigma\Gamma(\lambda^{-2})}(\lambda^{-2})^{\lambda^{-2}}\exp\left\{\lambda^{-2}\left[\left(\frac{y-\mu}{\sigma}\right)\lambda - \exp\left\{\left(\frac{y-\mu}{\sigma}\right)\lambda\right\}\right]\right\}, & \text{if} \quad \lambda\neq 0 \\ \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right], & \text{if} \quad \lambda = 0. \end{cases} \tag{2}$$

where $-\infty<y<\infty$, $-\infty<\lambda<\infty$ is the shape parameter, $\sigma>0$ is the scale parameter and $-\infty<\mu<\infty$ is the location parameter.

The standardized random variable $Z=(Y-\mu)/\sigma$ has density function

$$f(z;\lambda,\sigma,\mu) = \begin{cases} \frac{|\lambda|(\lambda^{-2})^{\lambda^{-2}}}{\Gamma(\lambda^{-2})}\exp\left[\lambda^{-1}z - \lambda^{-2}\exp(\lambda z)\right], & \text{if} \quad \lambda\neq 0, \\ \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right), & \text{if} \quad \lambda = 0. \end{cases} \tag{3}$$

The extreme-value standard distribution corresponds to the particular choice $\lambda=1$.

In many practical applications, the lifetimes are affected by explanatory variables such the cholesterol level, blood pressure and many others. Let $\mathbf{x}=(x_1,\ldots,x_p)^\top$ be the explanatory variable vector associated with the response variable y. Based on the GLG density, we can construct a linear regression model linking the response variable y_i and the explanatory variable vector $\mathbf{x}_i$ as follows

$$y_i = \mathbf{x}_i^\top\boldsymbol{\beta} + \sigma z_i, \ i=1,\ldots,n, \tag{4}$$

where the random error z_i has the distribution (3), $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)^T$, $\sigma>0$ and $\lambda>0$ are unknown parameters and $\mathbf{x}_i^T = (x_{i1},\ldots,x_{ip})$ is the explanatory variable vector modeling the location parameter μ_i. Hence, the location parameter vector $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_n)^T$ of the GLG regression model can be expressed as a linear model $\boldsymbol{\mu}=\mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}=(\mathbf{x}_1,\ldots,\mathbf{x}_n)^\top$ is a known model matrix.

Using the log-linear model in (4), the survival function of $Y_i|\mathbf{x}$ can take three different forms

$$S(y_i|\mathbf{x}) = \begin{cases} Q\left\{\lambda^{-2},\lambda^{-2}\exp\left[\lambda\left(\frac{y_i-\mathbf{x}_i^\top\boldsymbol{\beta}}{\sigma}\right)\right]\right\}, & \text{if} \quad \lambda>0, \\ 1-Q\left\{\lambda^{-2},\lambda^{-2}\exp\left[\lambda\left(\frac{y_i-\mathbf{x}_i^\top\boldsymbol{\beta}}{\sigma}\right)\right]\right\}, & \text{if} \quad \lambda<0, \\ 1-\Phi\left(\frac{y_i-\mathbf{x}_i^\top\boldsymbol{\beta}}{\sigma}\right), & \text{if} \quad \lambda=0, \end{cases}$$

where $\Phi(.)$ is the standard normal cumulative distribution and $Q(.)$ is the incomplete gamma integral.

For interval-censored data, the observed data consist of an interval $(\log(u_i), \log(v_i))$ for each individual, where such intervals are known to include $y_i = \log(t_i)$ with probability one, i.e. $P\big[\log(u_i) \leq y_i \leq \log(v_i)\big] = 1$ and if $\log(v_i) = \infty$, then it is a right-censored time for y_i.

2.1 Estimation by maximum likelihood

Given a set of interval-censored observations and explanatory variables $(\log(u_1), \log(v_1), \mathbf{x}_1), \ldots, (\log(u_n), \log(v_n), \mathbf{x}_n)$ of n observations, where $(\log(u_i), \log(v_i))$ is the observed data, $\mathbf{x}_i$ is the explanatory variable vector, the observed full log-likelihood function for the parameter vector $\boldsymbol{\theta} = (\lambda, \sigma, \boldsymbol{\beta}^\top)^\top$ is given by

$$l(\boldsymbol{\theta}) = \sum_{i \in F} l_1(\lambda, zu_i, zv_i) + \sum_{i \in C} l_2(\lambda, zu_i), \tag{5}$$

where

$$l_1\big(\lambda, zu_i, zv_i\big) = \begin{cases} \log\Big\{Q\big\{\lambda^{-2}, \lambda^{-2}\exp[\lambda(zu_i)]\big\} - Q\big\{\lambda^{-2}, \lambda^{-2}\exp[\lambda(zv_i)]\big\}\Big\}, & \text{if } \lambda > 0, \\ \log\Big\{Q\big\{\lambda^{-2}, \lambda^{-2}\exp[\lambda(zv_i)]\big\} - Q\big\{\lambda^{-2}, \lambda^{-2}\exp[\lambda(zu_i)]\big\}\Big\}, & \text{if } \lambda < 0, \\ \log\big[\Phi(zv_i) - \Phi(zu_i)\big], & \text{if } \lambda = 0, \end{cases}$$

and

$$l_2\big(\lambda, zu_i, zv_i\big) = \begin{cases} \log\Big\{Q\big\{\lambda^{-2}, \lambda^{-2}\exp[\lambda(zu_i)]\big\}\Big\}, & \text{if } \lambda > 0, \\ \log\Big\{1 - Q\big\{\lambda^{-2}, \lambda^{-2}\exp[\lambda(zu_i)]\big\}\Big\}, & \text{if } \lambda < 0, \\ \log\big[1 - \Phi(zu_i)\big], & \text{if } \lambda = 0, \end{cases}$$

where F represents the set of individuals with interval censoring, that is, $y_i \in \big(\log(u_i), \log(v_i)\big]$, C represents the set of individuals with direct censoring, that is, $y_i \in \big(\log(u_i), +\infty\big)$, $zu_i = \big[\log(u_i) - \mathbf{x}_i^\top \boldsymbol{\beta}\big]/\sigma$ and $zv_i = \big[\log(v_i) - \mathbf{x}_i^\top \boldsymbol{\beta}\big]/\sigma$. The maximum likelihood estimates (MLEs) of the parameter vector $\boldsymbol{\theta}$ can be obtained by maximizing the likelihood function. We use the software Ox (MAXBFGS subroutine) (Doornik, 2001) to compute the MLEs. The estimate of the covariance matrix of the MLEs $\widehat{\boldsymbol{\theta}}$ can also be obtained through the Hessian matrix. Confidence intervals and hypothesis testing can be conducted by employing a large sample distribution of the MLEs, which has a multivariate normal distribution with a covariance matrix given by the inverse of the information matrix since regularity conditions are satisfied.

2.2 Jackknife estimator

Suppose that $T_1, \ldots, T_n$ is a random sample of n values and $\bar{T} = \sum_{i=1}^{n} T_i/n$ is the sample mean used to estimate the mean of the population.

The sample mean calculated with the lth observation missing out is

$$\bar{T}_{-l} = \frac{\sum_{i=1}^{n} T_i - T_l}{n-1},$$

for which

$$T_l = n\hat{T} - (n-1)\bar{T}_{-l}. \tag{6}$$

In a general situation, consider that θ is a parameter estimated by $\hat{E}(T_1, \ldots, T_n)$, and for ease of notation we drop $(T_1, \ldots, T_n)$. Thus, $\hat{E}_{-l}$ is calculated, which is obtained with the T_l observation missing out. It follows, from equation (6), that pseudo-values can be determined $\hat{E}_l^* = n\hat{E} - (n-1)\hat{E}_{-l}, \quad l = 1, \ldots, n$. The average of the pseudo-values is the Jackknife estimate of θ given by $\hat{E}^* = \sum_{l=1}^{n} \hat{E}_l^*/n$.

Manly (1997) suggests that an approximate $100(1-\alpha)\%$ confidence interval for θ is given by $\hat{E}^* \pm t_{\alpha/2, n-1} s/\sqrt{n}$, where s is the standard deviation of the pseudo-values, $t_{\alpha/2,n-1}$ is the upper $(1-\alpha/2)$ point of the t distribution with $(n-1)$ degrees of freedom, which has the effect of removing the bias of order $1/n$.

2.3 Bootstrap re-sampling method

The bootstrap re-sampling method was proposed by Efron (1979). The method treats the observed sample as if it represented the population. From the information obtained from such a sample, B bootstrap samples of similar size to that of the observed sample are generated, from which it is possible to estimate various characteristics of the population, such as mean, variance, percentiles and so on.

According to the literature, the re-sampling method may be non-parametric or parametric. In this study, the non-parametric bootstrap method is addressed, according to which the distribution function F can be estimated by empirical distribution $\hat{F}$.

Let $\mathbf{T}=(T_1,\ldots,T_n)$ be an observed random sample and $\hat{F}$ is empirical distribution of $\mathbf{T}$. Thus, a bootstrap sample $\mathbf{T}^*$ is constructed by re-sampling with replacement of n elements of the sample $\mathbf{T}$. For the B bootstrap samples generated, $T_1^*,\ldots,T_B^*$, the bootstrap replication of the interest parameter for the b-th sample is given by $\hat{\boldsymbol{\theta}}_b^* = s(T_b^*)$, that is, the value of $\hat{\boldsymbol{\theta}}$ for sample T_b^*, $b=1,\ldots,B$.

The bootstrap estimator of the standard error (Efron and Tibshirani, 1993) is the standard deviation of these bootstrap samples; it is denoted by $\hat{EP}_B$ and obtained by the following expression

$$\hat{EP}_B = \left[\frac{1}{(B-1)}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \bar{\theta}_B\right)^2\right]^{1/2},$$

in which $\bar{\theta}_B = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$. For further details on bootstrap intervals, see, Efron and Tibshirani (1993), DiCiccio and Efron (1996) and Davison and Hinkley (1997).

BCa bootstrap interval

The bootstrap interval based on the BCa method assumes that the percentiles used in delimitating the bootstrap confidence intervals depend on the corrections to tendency $\hat{a}$ and acceleration $\hat{z}_0$.

The bias correction value $\hat{z}_0$ is generated based on the proportion of estimations of bootstrap samples that are smaller than the original estimation $\hat{\boldsymbol{\theta}}$. The expression of $\hat{z}_0$ is given by $\hat{z}_0 = \Phi^{-1}\left(\frac{\sharp(\hat{\boldsymbol{\theta}}_b^* < \hat{\boldsymbol{\theta}})}{B}\right)$, $\quad b=1,\ldots,B$. Note that $\Phi^{-1}(\cdot)$ is the inverse of the accumulated standard normal distribution; B is the number of generated bootstrap samples; $\hat{\boldsymbol{\theta}}$ is the MLE of the observed sample; and $\hat{\boldsymbol{\theta}}_b^*$ is the MLE of the b-th bootstrap sample.

Let $\hat{\boldsymbol{\theta}}_{(i)}$ be the MLE of the sample without the i-th observation. Then $\hat{a}$ is given by

$$\hat{a} = \frac{\sum_{i=1}^{n}\left[\hat{\boldsymbol{\theta}}_{(\cdot)} - \hat{\boldsymbol{\theta}}_{(i)}\right]^3}{6\left\{\sum_{i=1}^{n}\left[\hat{\boldsymbol{\theta}}_{(\cdot)} - \hat{\boldsymbol{\theta}}_{(i)}\right]^2\right\}^{3/2}},$$

where $\hat{\boldsymbol{\theta}}_{(\cdot)} = \sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_{(i)}/n$ and n is the sample size.

Hence, the BCa bootstrap interval of coverage $100(1-2\alpha)\%$ is given by $\left[\hat{\boldsymbol{\theta}}^*_{(B\alpha_1)}, \hat{\boldsymbol{\theta}}^*_{(B\alpha_2)}\right]$, in which

$$\alpha_1 = \Phi\left\{\hat{z}_0 + \frac{\hat{z}_0 + \Phi^{-1}(\alpha)}{1-\hat{a}\left[\hat{z}_0 + \Phi^{-1}(\alpha)\right]}\right\} \quad \text{and} \quad \alpha_2 = \Phi\left\{\hat{z}_0 + \frac{\hat{z}_0 + \Phi^{-1}(1-\alpha)}{1-\hat{a}\left[\hat{z}_0 + \Phi^{-1}(1-\alpha)\right]}\right\}.$$

Note that α_1 and α_2 are corrections to the bootstrap percentiles; $\Phi(\cdot)$ is an accumulated distribution function of the standard normal distribution; and $\Phi^{-1}(\cdot)$ is the inverse of the accumulated distribution function of the standard normal distribution (Efron and Tibshirani, 1993).

3 Application

We consider a data set on the AIDS cohort study of hemophiliacs discussed in Kim et al. (1993) and Sun et al. (1999). This study consists of individuals with Type A or B hemophilia who were at risk for HIV infection through the contaminated blood factor they received for their treatment. The subjects were classified into two

groups, lightly and heavily treated groups, according to the amount of blood they received, and age indicators that indicate if the age of a subject was below 20 at his or her HIV infection. There were 257 individuals in the original study. In the following, we will focus on 188 subjects who were known to be infected by HIV before the end of the study. Note that, in the original data set, most of the observations for AIDS diagnosis are exact or right censored and the remainder are interval censored with only two time points included in each interval. One objective of the study is to test and estimate the possible difference of the AIDS incubation distributions between the two groups. The covariates considered in the models are: x_{i1}: age (0=below 20, 1=above 20) and x_{i2}: treated group (0=heavily, 1=lightly).

3.1 Estimation

Maximum likelihood and Jackknife estimation

To obtain the MLEs of the parameters in the GLG regression model for interval-censored data, we used the subroutine MAXBFGS in Ox, whose results are given in Table 1. Additionally, in Table 1, we report the Jackknife estimates.

Table 1: Maximum likelihood and Jackknife estimates for the parameters of the GLG regression model for interval-censored fitted to hemophilia data.

	MLEs				Jackknife estimates		
$\boldsymbol{\theta}$	Estimate	S.E.	p-value	95% C.I.	Estimate	S.E.	95% C.I.
λ	1.128	0.261	-	(0.617;1.639)	1.115	0.315	(0.495;1.735)
σ	0.245	0.028	-	(0.191;0.299)	0.249	0.027	(0.196;0.302)
β_0	2.778	0.038	0.000	(2.702;2.852)	2.770	0.044	(2.683;2.857)
β_1	0.047	0.040	0.244	(-0.032;0.125)	0.048	0.044	(-0.039;0.135)
β_2	-0.216	0.039	<0.001	(-0.292;-0.140)	-0.214	0.039	(-0.291;-0.137)

We can observe that the explanatory variable x_2 is significant (at 5%) for the log-survival time. Note that the estimates from the two methods seem to be very similar (Table 1).

Bootstrap re-sampling method

We considered $B = 3000$ bootstrap samples of the GLG regression model with interval censoring. Using the bootstrap method described in Section 2.3, we obtain the estimated bootstrap and the BCa confidence intervals, presented in Table 2.

Table 2: Bootstrap estimate and confidence intervals based on the non-parametric bootstrap re-sampling method for hemophilia data.

	Bootstrap estimates		
$\boldsymbol{\theta}$	Estimate	S.E.	95% C.I.(BCA)
λ	1.164	0.495	(0.621;1.766)
σ	0.240	0.030	(0.205;0.291)
β_0	2.777	0.042	(2.710;2.846)
β_1	0.047	0.045	(-0.026;0.124)
β_2	-0.218	0.034	(-0.272;-0.259)

The estimates from the three methods seem to be very similar. The MLEs appear more conservative (large standard errors). Therefore, since normality for the Jackknife estimator is expected for this sample size ($n = 256$), one can also expect for the MLEs some symmetric distribution with heavy tails. We will continue the analysis by using the MLEs and considering GLG regression models.

4 Concluding remarks

In this study, an GLG regression model for interval-censored data is proposed. We used the Quasi-Newton algorithm to obtain the maximum likelihood estimates and asymptotic tests were performed for the parameters based on the asymptotic distribution of the MLEs. On the other hand, as an alternative analysis, the work discusses the use of the Jackknife estimator and non-parametric bootstrap for the GLG regression model for interval-censored data.

References

Davison, A. C. and Hinkley, D. V. (1997). Bootstrap methods and their application. New York: Cambridge University Press.

DiCiccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals. Statistical Science **11**, 189-228.

Doornik, J. (2001). Ox: Object-oriented matrix programming using Ox. 4th ed. London : Timberlake Consultants Ltd.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. The Annals of Statistics **7**, 1-26.

Efron, B. and Tibshirani, R. J. (1993). An introduction to the bootstrap. New York: Chapman & Hall.

Johnson, N.L, Kotz, S. and Balakrishnan, N. (1994). Continuous univariate distributions. New York: John Wiley and Sons.

Kim, M. Y., DeGruttola, V. G. and Lagakos, S. W. (1993). Analyzing doubly censored data with covariates, with application to AIDS. Biometrics **49**, 13-22.

Lawless, J. F. (2003). Statistical Models and Methods for Lifetime Data. Wiley: New York.

Manly, B. F. J. (1997). Randomization, Bootstrap and Monte Carlo Methods in Biology, 2nd Edition. Chapman and Hall: London.

Ortega, E. M. M., Cancho, V. G. and Paula G. A. (2009). Generalized log-gamma regression models with cure fraction. Lifetime Data Analysis. **15**, 79-106.

Stacy, E. W. (1962). A Generalization of the gamma distribution. Annals of Mathematical Statistics **33**, 1187-1192.

Sun, J., Liao, Q. and Pagano, M. (1999). Regression analysis of doubly censored failure time data with applications to AIDS studies. Biometrics, **55**, 909-914.