Artigo Original

DOI:10.5902/2179460X24118

Ciência e Natura, Santa Maria v.39 n.2, 2017, Mai - Ago, p. 244 – 258
Revista do Centro de Ciências Naturais e Exatas - UFSM
ISSN impressa: 0100-8307 ISSN on-line: 2179-460X

CIÊNCIA e NATURA

Bayesian and maximum likelihood inference for the defective Gompertz cure rate model with covariates: an application to the cervical carcinoma study

Inferência Bayesiana e de máxima verossimilhança para o modelo defectivo Gompertz com fração de cura e covariáveis: uma aplicação ao estudo do carcinoma cervical

Milene Regina dos Santos, Jorge Alberto Achcar, Edson Zangiacomi Martinez

Department of Social Medicine, University of São Paulo, Ribeirão Preto, Brazil
milene.santos@usp.br; achcar@fmrp.usp.br; edson@fmrp.usp.br

Resumo

A análise de sobrevivência é uma classe de métodos estatísticos usados para estudar o tempo até a ocorrência de um evento. Os métodos usuais assumem que todos os indivíduos sob estudo são sujeitos ao evento de interesse. Entretanto, há situações em que este pressuposto não é real. Por exemplo, em uma pesquisa clínica, uma proporção de pacientes pode responder favoravelmente ao tratamento sob investigação e consequentemente não morrer devido à doença. Modelos baseados em distribuições defectivas são adequados para analisar dados com estas características. Neste artigo, apresentamos inferências Bayesianas e de máxima verossimilhança para o modelo de fração de cura baseado na distribuição defectiva de Gompertz, incluindo covariáveis. Um exemplo de aplicação à sobrevida livre de doença de mulheres tratadas para câncer cervical é usado para ilustrar a metodologia. Na análise Bayesiana, distribuições a posteriori dos parâmetros foram estimadas por métodos Monte Carlo em cadeias de Markov (MCMC). Códigos R, SAS e OpenBUGS são apresentados em um apêndice no final do artigo aos leitores que desejarem usar o método para conduzir suas próprias análises.

Palavras-chave: *Estimação de máxima verossimilhança, Inferência Bayesiana, Distribuições defectivas, Análise de sobrevida, Distribuição Gompertz modificada.*

Abstract

Survival analysis is a class of statistical methods to study the time until the occurrence of a specified event. The usual methods assume that all individuals under study are subjects to the event the interest. However, there are situations where this case is unrealistic. For example, in a clinical research, a proportion of patients could respond favourably to the treatment under investigation and consequently they would not die from the disease. Models based on defective distributions are a suitable way to analyse data with these characteristics. In this paper, we present Bayesian and maximum likelihood inference for the defective Gompertz cure rate model in presence of covariates. An example with application to disease-free survival of women treated for cervical carcinoma is used to illustrate the proposed methodology. In the Bayesian analysis, posterior distributions of parameters are estimated using the Markov chain Monte Carlo (MCMC) method. R, SAS and OpenBUGS codes are provided in an appendix at the end of the paper so that reader can carry out their own analysis.

Keywords: *Maximum likelihood estimation, Bayesian inference, Defective distributions, Survival analysis, Modified Gompertz distribution.*

Recebido: 23/09/2016 Aceito: 10/02/2017

1 Introduction

Statistical lifetime methods are widely used in studies considering the analysis of data sets in the health area, when the variable of interest is related to the time until the occurrence of an event (Klein and Moeschberger, 2005). This event can be, for example, the first recurrence of a disease after treatment, death due to a specific cause or discharge after a hospital admission. While many methods of survival analysis are readily available in statistical computer packages, most of these techniques assume that all individuals are susceptible to the occurrence of the event of interest. However, this assumption can be unrealistic in some specific situations, when there is a subpopulation of individuals that is immune or not susceptible to the occurrence of the event of interest. As an example, in a clinical research, a proportion of patients can respond favourably to the treatment under investigation and be regarded as "cured". A number of statistical tools have been developed to handle such data, especially the cure fraction models (Lambert et al., 2007).

These models include as a special case the mixture model (Farewell, 1982), that explicitly includes a parameter accounting for the cure rate. The mixture model assumes that the probability of the time-to-event to be greater than a specified time t is given by the survival function

$$S(t) = P(T > t) = p + (1 - p)\, S_0(t),$$

where p is a parameter which represents the proportion of "cured patients", regarding the event of interest ($0 < p < 1$), and $S_0(t)$ is the baseline survival function for the susceptible individuals (Maller and Zhou, 1996). Common choices for $S_0(t)$ are the Gompertz, exponential and Weibull distributions. The probability density function for the lifetime T is given by

$$f(t) = \frac{dF(t)}{dt} = (1 - p)\, f_0(t),$$

where $F(t) = 1 - S(t)$ and $f_0(t)$ is the baseline probability density function for the susceptible individuals.

Models based on defective distributions are alternatives to the mixture model. A defective distribution is defined as a distribution that is not normalized to one for some values of their parameters, and it can be useful to fit data including both immune and susceptible individuals without explicitly including the parameter p. The use of defective distribution in survival analysis have been considered by a number of authors, including Balka et al. (2011), Cancho and Bolfarine (2001) and Rocha et al. (2014).

In this paper we propose Bayesian and frequentist approaches to a model based on the modified Gompertz distribution, considering survival data in presence of a cure fraction. More recently a number of new families of defective distributions have been proposed in the literature, such as those based on the Kumaraswamy Rocha et al. (2015) and the Marshall–Olkin families Rocha et al. (2017). However, in the present study, the modified Gompertz distribution was preferred because of its computational simplicity and the ease of interpretation of their parameters. Our methodology is motivated by a real data set that arises from the cervical carcinoma study by Brenna et al. (2004) and it extends the model described by Rocha et al. (2014) to the situation in which covariates are present. The computational codes used in this article are provided in the Appendix.

2 Methods

2.1 The modified Gompertz model

The modified Gompertz hazard function was first used by Cantor and Shuster (1992) for estimation of cure rates from paediatric clinical trials and after extended by Gieser et al. (1998) to include covariate effects. This model assumes a survival function given by

$$S(t) = \exp\left\{-\frac{\alpha}{\beta}\left[1 - \exp\left(-\beta t\right)\right]\right\},$$

where $t > 0$, $\alpha > 0$ is a shape parameter, $\beta > 0$ is a scale parameter, and the corresponding cure rate η is given by

$$\lim_{t \longrightarrow \infty} S(t) = \exp\left(-\frac{\alpha}{\beta}\right) = \eta,$$

where $0 < \eta < 1$. In addition, the corresponding probability density function is given by

$$f(t) = \alpha \exp\left(-\beta t\right) \exp\left\{-\frac{\alpha}{\beta}\left[1 - \exp\left(-\beta t\right)\right]\right\}$$

and the hazard function is given by

$$h\left(t\right) = \frac{f(t)}{S(t)} = \alpha \exp\left(-\beta t\right).$$

2.2 Maximum likelihood estimation

Considering a random sample (t_i,δ_i) of size n, $i = 1,...,n$, the contribution of the ith subject for the likelihood function is given by

$$L_i = [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i} = S(t_i) [h(t_i)]^{\delta_i},$$

where δ_i is a censoring indicator variable, that is, $\delta_i = 1$ for an observed lifetime and $\delta_i = 0$ for a censored lifetime.

Assuming the modified Gompertz model, the likelihood function for $\boldsymbol{\theta} = (\alpha,\beta)$ is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \alpha^{\delta_i} \exp\left\{-\frac{\alpha}{\beta}[1 - \exp(-\beta t_i)] - \beta\delta_i t_i\right\}$$

and the corresponding log-likelihood function is

$$l(\boldsymbol{\theta}) = (\ln\alpha)\sum_{i=1}^{n}\delta_i - \frac{\alpha}{\beta}\sum_{i=1}^{n}[1 - \exp(-\beta t_i)] - \beta\sum_{i=1}^{n}\delta_i t_i.$$

By deriving the log-likelihood function with respect to α and β, we have the following equations:

$$\frac{\partial l(\theta)}{\partial\alpha} = \frac{1}{\alpha}\sum_{i=1}^{n}\delta_i - \frac{1}{\beta}\sum_{i=1}^{n}\left(1 - e^{-\beta t_i}\right)$$

and

$$\frac{\partial l(\theta)}{\partial\beta} = \frac{\alpha\left[\sum_{i=1}^{n}\left(1 - e^{-\beta t_i}\right) - \beta\sum_{i=1}^{n}e^{-\beta t_i}t_i\right]}{\beta^2} - \sum_{i=1}^{n}\delta_i t_i.$$

Setting these expressions equal to zero, we get the corresponding score equations whose numerical solution leads to the maximum likelihood estimators (MLE). Although we cannot obtain explicit expressions for the MLEs for the parameters α and β, they can be estimated numerically using iterative algorithms such as the Newton-Raphson method and its variants (Rocha et al., 2014). R (R Development Core Team, 2009) and SAS (Littell et al., 2006) codes for implementing these procedures are presented in the Appendix.

The asymptotic variances of MLEs are given by the elements of the inverse of the Fisher's information matrix $I(\alpha,\beta)$. The second partial derivatives of the maximum likelihood function are given as follows:

$$\frac{\partial^2}{\partial\alpha^2}l(\theta) = -\frac{1}{\alpha^2}\sum_{i=1}^{n}\delta_i,$$

$$\frac{\partial^2}{\partial\beta^2}l(\theta) = \frac{\alpha}{\beta^2}\left[\sum_{i=1}^{n}\frac{e^{-\beta t_i}(2 + \beta t_i) - 2}{\beta} + \sum_{i=1}^{n}t_i e^{-\beta t_i}(1 + \beta t_i)\right]$$

and

$$\frac{\partial^2}{\partial\alpha\beta}l(\theta) = -\sum_{i=1}^{n}\frac{e^{-\beta t_i}(1 + \beta t_i) - 1}{\beta^2}.$$

Therefore, the expected Fisher's information matrix is given by

$$I(\alpha,\beta) = \begin{bmatrix} -E\left[\frac{\partial^2}{\partial\alpha^2}l(\theta)\right] & -E\left[\frac{\partial^2}{\partial\alpha\beta}l(\theta)\right] \\ -E\left[\frac{\partial^2}{\partial\alpha\beta}l(\theta)\right] & -E\left[\frac{\partial^2}{\partial\beta^2}l(\theta)\right] \end{bmatrix}.$$

In practical applications we could use the observed Fisher's information matrix given the difficulties to get the expected matrix in the determination of the asymptotical normality of the MLEs used to construct confidence intervals and hypotheses tests for the parameters of the model. Wald-type 95% confidence intervals for the parameters can be thus obtained from the respective estimates of the standard errors.

2.3 Model with covariates

In order to include covariates, the parameter α in the likelihood function $L(\boldsymbol{\theta})$ can be replaced by a function $\alpha(\mathbf{x}_i)$ such as

$$\ln \alpha(\mathbf{x}_i) = \mathbf{x}_i \boldsymbol{\alpha}^*,$$

where $\mathbf{x}_i = (1,x_{1i},x_{2i},...,x_{pi})$ is a vector containing observations on p independent variables and $\boldsymbol{\alpha}^* = (\alpha_0,\alpha_1,...,\alpha_p)$ is a vector of unknown parameters. Analogously, the parameter β in $L(\boldsymbol{\theta})$ can be replaced by an function $\beta(\mathbf{w}_i)$ such as

$$\ln \beta(\mathbf{w}_i) = \mathbf{w}_i \boldsymbol{\beta}^*,$$

where $\mathbf{w}_i = (1,w_{1i},w_{2i},...,w_{qi})$ is a vector which may or may not be equal to $\mathbf{x}_i$ and $\boldsymbol{\beta}^* = (\beta_0,\beta_1,...,\beta_q)$ is a vector of unknown parameters.

Comparisons between different model formulations can be based on the Akaike's information criterion (AIC) (Akaike, 1973). The model with the lowest AIC value suggest a better fit to the data.

2.4 Bayesian analysis

The Bayesian method considers that prior distributions of the parameters of the model are used to derive their corresponding posterior densities (Gelman et al., 2013). According to the Bayes theorem, we can write the joint posterior density by combining the joint prior distribution with the likelihood function for α and β. In this case, Markov chain Monte Carlo (MCMC) is a useful method to sample from posterior probability distributions by means of simulation.

For a Bayesian analysis of the model without covariates, we assume inverse gamma (IG) prior distributions for the parameters α and β, considering that these parameters are real and positive numbers. Thus,

$$\alpha \sim IG(a_1,b_1)$$

and

$$\beta \sim IG(a_2,b_2),$$

where a_1, b_1, a_2 and b_2 are known hyperparameters. The mean and variance of a random variable following the inverse gamma distribution with parameters a and b are given by

$$\frac{b}{a-1} \quad \text{and} \quad \frac{b^2}{(a-1)^2(a-2)},$$

respectively. Therefore, an improper prior distribution is considered if $a \leq 2$. An improper prior distribution is not a true probability distribution in that it does not integrate to 1. In Bayesian inference, an improper prior distribution is acceptable, in the sense that the corresponding posterior distribution is proper.

In the case of the model with covariates, consider the following prior distributions for the parameters:

$$\alpha_j \sim N(c_j,d_j), \quad j = 0,1,...,p,$$

and

$$\beta_k \sim N(e_k,f_k), \quad k = 0,1,...,q,$$

where $N(c,d)$ denotes a normal distribution with mean c and variance d, and c_j, d_j, e_k and f_k ($j = 0,1,...,p$ and $k = 0,1,...,q$) are known hyperparameters.

Posterior summaries of interest will be obtained in the examples section from simulated samples for the joint posterior distribution using standard MCMC procedures, as the Gibbs sampling. We will generate 1,005,000 samples for each parameter of interest. The 5,000 first simulated samples will be discarded as a burn-in period, which is usually used to minimize the effect of the initial values. The posterior summaries of interest will be based on 10,000 samples, taking every $100th$ sample to have approximately uncorrelated values. We assume prior independence between all model parameters. The Bayes estimates of the parameters will be obtained as the mean of samples drawn from the joint posterior distribution. In addition, 95% credible intervals for the parameters are given by the $0.025th$ and $0.975th$ percentiles of the respective posterior distributions. Convergence of the MCMC algorithm will be monitored by usual time series plots for the simulated samples and also using some existing Bayesian convergence methods.

Posterior summaries of interest will be obtained using the OpenBUGS software (Lunn et al., 2000), that only requires the specification of the distribution for the data and the prior distributions for the parameters. See Appendix for details about the OpenBUGS codes used.

The Deviance Information Criterion (DIC) will be used to compare the fit of different model formulations (Spiegelhalter et al., 2014). Smaller values correspond to models that provide better fit to the data.

2.5 A simulation method for a random variable with a modified Gompertz distribution

The algorithm used to simulate a sample of size n from the modified Gompertz distribution with right-censored data follows the steps:

Step 1. Fix values of α and β.

Step 2. Generate n random samples from

$$M_i \sim Bernoulli(0,1-\eta),$$

where η is the corresponding cure rate given by

$$\eta = \exp\left(-\frac{\alpha}{\beta}\right).$$

Step 3. Consider

$$t_i' = \begin{cases} \infty \text{ if } M_i = 0 \\ F_Y^{-1}(U_i) \text{ if } M_i = 1 \end{cases}$$

where

$$F_Y^{-1}(U_i) = -\frac{1}{\beta}\ln\left[1+\frac{\beta}{\alpha}\ln\left(1-U_i\right)\right]$$

and

$$U_i \sim Uniform(0,1-\eta).$$

Step 4. Generate n random samples from

$$u_i' \sim Uniform(0,max(t_i')),$$

considering only the finite t_i'.

Step 5. Calculate $t_i = min(t_i',u_i')$.

Step 6. Pairs of values (t_1,δ_1), (t_2,δ_2),...,(t_n,δ_n) are thus obtained, where $\delta_i = 1$ if $t_i < u_i'$ and $\delta_i = 0$ if $t_i \geq u_i'$, $i = 1,...,n$.

This algorithm is similar to that presented by Rocha et al. (2017). An R function for generating random samples based on these steps is presented in the Appendix.

3 Results

3.1 Simulation study

A brief simulation study of the performance of the maximum likelihood method of estimation was carried on based on simulated samples of sizes $n = 25$, 50, 75, 100, 150 and 200. Each sample was replicated 5000 times. The variance of the cure rate η was estimated using the delta method. The results from the simulation study are presented in Table 1, considering a nominal confidence coefficient of 95 per cent and two sets of arbitrary values for the parameters, given by $(\alpha,\beta) = (0.4,0.2)$ and $(\alpha,\beta) = (0.3,0.7)$.

The results in Table 1 show that the coverage probability of the Wald-type confidence intervals for the parameters α and β is quite close to the nominal confidence level of 95%. Furthermore, the coverage probability of the confidence intervals for η approaches 95% as the sample size is increased. In all simulations, the biases and the mean squared errors (MSE) always approached zero as the sample size increased.

3.2 An application to real data

Let us consider the data from a study that included a total of 148 women diagnosed and treated for invasive cervical carcinoma between 1992 and 2002 (Brenna et al., 2004). For the purposes of the present study, it was considered a subsample of 118 women who received the standard treatment recommended by the International Federation of Gynecology and Obstetrics (FIGO). Let us define the disease-free survival (DFS) as the time in complete months from the date of surgery to the first event of disease recurrence. Nearly 48% of the data are censored observations.

Figure 1 is a contour plot of the log of the likelihood function for these data, considering the parametric model based on the modified Gompertz distribution. The plot indicates that the maximum of the log-likelihood function is at $(\alpha,\beta) = (0.04178, 0.03995)$, as showed in the Table 2. Table 2 shows the maximum likelihood estimates of α, β and the cure rate η, as well as the estimates of their standard errors and confidence intervals. Figure 2 shows plots of the Kaplan-Meier estimates for the survival function and the predicted values obtained from the parametric model against times (months). From Figure 2, it is possible to note that the predicted values obtained from the model based on the modified Gompertz distribution are closest to the empirical values, suggesting that this model is well fitted for the data.

Table 1: Results from the simulation study: coverage probabilities of the Wald-type confidence intervals for the parameters α, β and η, biases and mean squared errors (MSE).

Nominal values	n		Coverage probability	Bias	MSE
$\alpha = 0.4$ $\beta = 0.2$	25	α	0.9381	0.0201	0.0223
		β	0.9591	0.0123	0.0255
		η	0.7800	0.0045	0.0134
	50	α	0.9485	0.0090	0.0095
		β	0.9541	0.0022	0.0088
		η	0.8756	-0.0018	0.0064
	75	α	0.9493	0.0046	0.0059
		β	0.9545	-0.0007	0.0046
		η	0.9158	-0.0034	0.0039
	100	α	0.9478	0.0045	0.0043
		β	0.9476	0.0007	0.0031
		η	0.9248	-0.0022	0.0028
	150	α	0.9480	0.0013	0.0027
		β	0.9496	-0.0004	0.0017
		η	0.9344	-0.0016	0.0017
	200	α	0.9482	0.0022	0.0019
		β	0.9538	0.00004	0.0012
		η	0.9426	-0.0014	0.0012
$\alpha = 0.3$ $\beta = 0.7$	25	α	0.9158	0.0928	1.3430
		β	0.9638	0.1241	0.4927
		η	0.7311	-0.1127	0.0978
	50	α	0.9300	0.0245	0.0202
		β	0.9675	0.0886	0.2825
		η	0.9036	-0.0469	0.0338
	75	α	0.9408	0.0192	0.0125
		β	0.9582	0.0585	0.1344
		η	0.9316	-0.0243	0.016
	100	α	0.9434	0.0095	0.0078
		β	0.9543	0.0284	0.0742
		η	0.9404	-0.0146	0.0087
	150	α	0.9428	0.0075	0.0049
		β	0.9570	0.0225	0.0393
		η	0.9508	-0.0048	0.0034
	200	α	0.9424	0.0069	0.0037
		β	0.9510	0.0185	0.0254
		η	0.9520	-0.0027	0.0021

The cure rate is estimated by

$$\widehat{\eta} = \exp\left(-\frac{\widehat{\alpha}}{\widehat{\beta}}\right) = 0.35147,$$

and the respective standard error was obtained by the Delta method as a first-order approximation of a Taylor-series expansion (Table 2).

Table 2: Maximum likelihood estimates

Parameter	Estimate	Std. error	Wald-type 95% CI
α	0.04178	0.00747	(0.0271 , 0.0564)
β	0.03995	0.00795	(0.0243 , 0.0555)
η	0.35147	0.05440	(0.2446 , 0.4584)

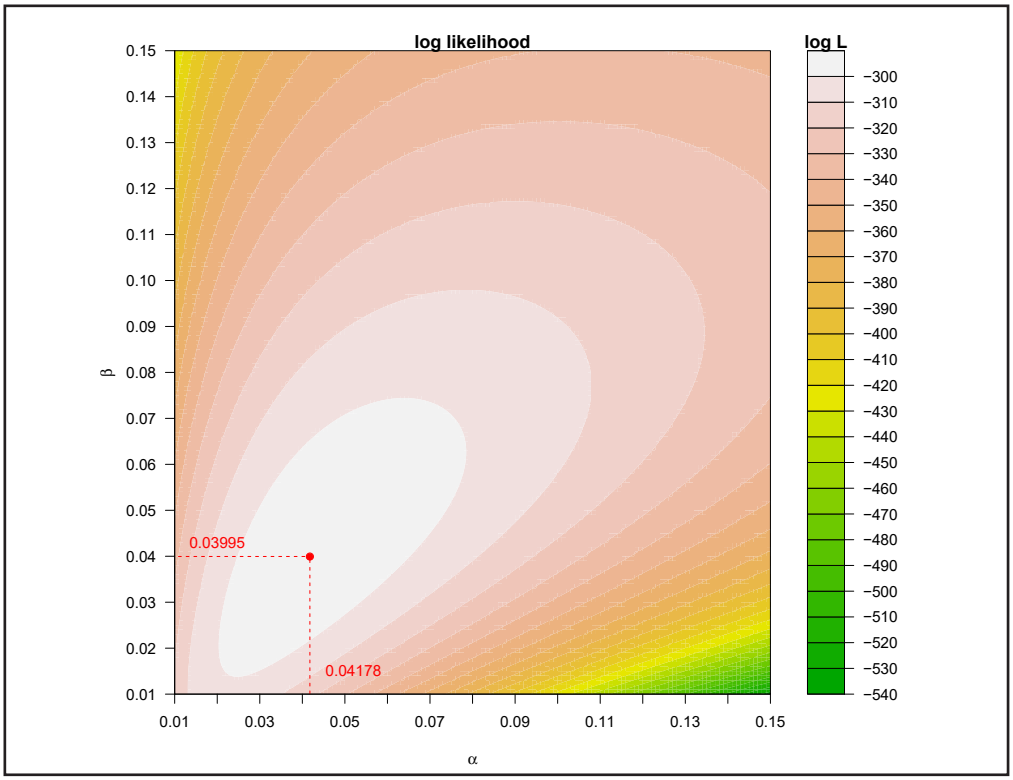

Figure 1: Contour plot of the log-likelihood function, considering the data from the cervical carcinoma study. The maximizing point is $(\alpha,\beta) = (0.04178, 0.03995)$.

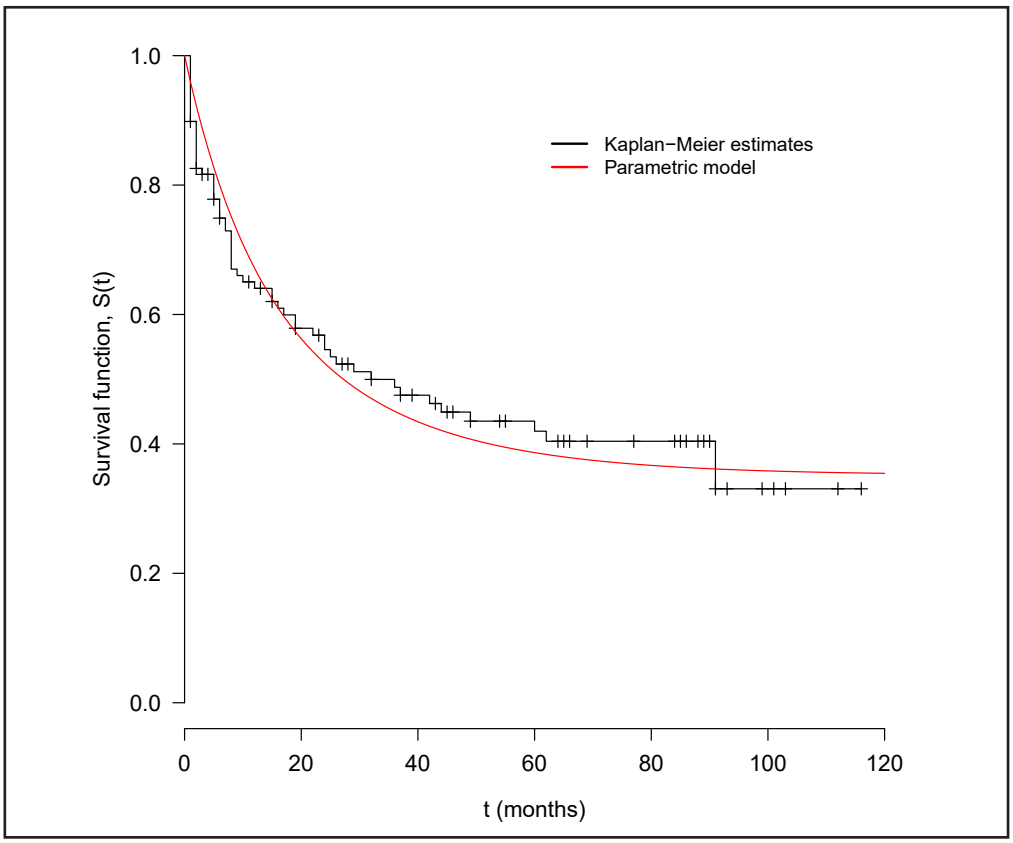

Figure 2: Plots of the disease-free survival functions estimated from the Kaplan-Meier method and the parametric model based on the modified Gompertz distribution. Censored observations are marked by ticks on the survival curve.

Table 3: Bayesian estimates

Parameter	Estimate	95% credible intervals
α	0.04134	(0.0282 , 0.0574)
β	0.03928	(0.0244 , 0.0557)
η	0.3474	(0.2380 , 0.4554)

Table 3 shows the Bayesian estimates of α, β and the cure rate η, with their respective 95% credible intervals. In this analysis, it was considered the hyperparameters $a_1 = b_1 = a_2 = b_2 = 0.001$ in order to obtain noninformative priors, that is, $\alpha \sim IG(0.001,0.001)$ and $\beta \sim IG(0.001,0.001)$. We can note that the Bayesian estimates for the parameters are relatively close to those obtained by the maximum likelihood method (Table 2).

3.3 Model with covariates

In order to illustrate an application of the regression model based on the modified Gompertz distribution, in this analysis we consider the following variables:

- Age of the patients at start of follow-up (x_1), classified as less than 50 years ($x_1 = 0$) versus greater or equal to 50 years ($x_1 = 1$).
- Clinical stage of the disease at the start of treatment, classified as I, II, or III. This variable is represented in the regression model by using two dummy variables (x_2 and x_3), where $x_2 = 0$ and $x_3 = 0$ if stage I, $x_2 = 1$ and $x_3 = 0$ if stage II, and $x_2 = 0$ and $x_3 = 1$ if stage III.

Thus, the regression model for the data considers that

$$\ln \alpha(\mathbf{x}) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_{12} x_1 x_2 + \alpha_{13} x_1 x_3$$

and

$$\ln \beta(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3,$$

where α_{12}, α_{13}, β_{12} and β_{13} are interaction terms. Let us consider the following model formulations:

- **Model 1:** Model 1 considers only the effect of age (x_1). In this case, α_2, α_3, α_{12}, α_{13}, β_2, β_3, β_{12} and β_{13} are considered equal to zero.
- **Model 2:** Model 2 considers only the effect of the stage of the disease (dummy variables x_2 and x_3). In this case, α_1, β_1 and the interaction terms α_{12}, α_{13}, β_{12} and β_{13}, are considered equal to zero.
- **Model 3:** Model 3 considers the effects of age (x_1) and the stage of the disease (x_2 and x_3) but it does not includes the interaction terms α_{12}, α_{13}, β_{12} and β_{13}.
- **Model 4:** Model 4 considers the effects of age (x_1) and the stage of the disease (x_2 and x_3) and it includes the interaction terms α_{12}, α_{13}, β_{12} and β_{13}.

For all regression coefficients, we assumed normal prior distributions with mean 0 and large variance. Both Bayesian and maximum likelihood estimates are shown in Table 4. The fit of the Models 1 and 2 also considered the estimation of the cure fraction (η) for each level of the respective independent variable. In the frequentist estimation, standard errors for the cure fraction were calculated using the Delta method. As an illustration, Figure 3 shows the survival curves estimated by the Kaplan-Meier method and by the Model 2 using the maximum likelihood approach.

Table 4 shows that the maximum likelihood and the Bayesian estimates are fairly close to each other. Model 1 suggests that the age do not have a significant effect on the disease-free survival, since the confidence and credible intervals for the parameters α_1 and β_1 include the value 0. Model 2 suggests a significant effect of the stage of the disease on the disease-free survival, given that the confidence and credible intervals for the shape parameters α_2 and α_3 do not include the zero value (Figure 3).

Table 4: Maximum likelihood and Bayesian estimates, model with covariates

Parameter	Maximum likelihood estimates				Bayesian estimates		
	Estimate	Std. error	Wald-type 95% CI	AIC	Estimate	95% credible interval	DIC
Model 1				592.9			593.3
α_0 (intercept)	-3.041	0.2493	(-3.532 , -2.551)		-3.091	(-3.629 , -2.585)	
α_1 (age ≥ 50 vs. < 50)	-0.258	0.3578	(-0.961 , 0.445)		-0.285	(-1.032 , 0.455)	
β_0 (intercept)	-3.020	0.2637	(-3.539 , -2.501)		-3.084	(-3.711 , -2.580)	
β_1 (age ≥ 50 vs. < 50)	-0.386	0.4036	(-1.178 , 0.407)		-0.461	(-1.524 , 0.424)	
η_1 (cure fraction, age ≥ 50)	0.376	0.0756	(0.227 , 0.524)		0.371	(0.214 , 0.523)	
η_2 (cure fraction, age < 50)	0.329	0.0794	(0.173 , 0.485)		0.311	(0.121 , 0.471)	
Model 2				565.2			563.4
α_0 (intercept)	-4.825	0.5266	(-5.859 , -3.790)		-4.850	(-5.746 , -3.972)	
α_2 (stage II vs. I)	1.814	0.6071	(0.621 , 3.007)		1.734	(0.636 , 2.820)	
α_3 (stage III vs. I)	2.325	0.5789	(1.187 , 3.462)		2.281	(1.258 , 3.285)	
β_0 (intercept)	-4.093	0.8831	(-5.828 , -2.358)		-4.269	(-6.409 , -3.070)	
β_2 (stage II vs. I)	0.948	0.9590	(-0.936 , 2.832)		0.902	(-0.939 , 3.153)	
β_3 (stage III vs. I)	0.803	0.9348	(-1.034 , 2.639)		0.839	(-0.665 , 3.085)	
η_1 (cure fraction, stage I)	0.618	0.1726	(0.278 , 0.957)		0.560	(0.043 , 0.814)	
η_2 (cure fraction, stage II)	0.319	0.1026	(0.117 , 0.520)		0.291	(0.039 , 0.506)	
η_3 (cure fraction, stage III)	0.110	0.0554	(0.001 , 0.219)		0.106	(0.014, 0.231)	
Model 3				563.7			560.8
α_0 (intercept)	-5.490	0.369	(-6.226 , -4.753)		-5.486	(-6.244 , -4.828)	
α_1 (age ≥ 50 vs. < 50)	0.297	0.340	(-0.396 , 0.991)		0.260	(-0.433 , 0.898)	
α_2 (stage II vs. I)	2.529	0.452	(1.638 , 3.421)		2.387	(1.571 , 3.255)	
α_3 (stage III vs. I)	2.602	0.415	(1.778 , 3.424)		2.527	(1.765 , 3.366)	
β_0 (intercept)	-12.674	2.443	(-15.654 , -9.795)		-13.020	(-17.140 , -9.101)	
β_1 (age ≥ 50 vs. < 50)	2.012	0.994	(-0.168 , 4.192)		2.853	(0.575 , 6.468)	
β_2 (stage II vs. I)	8.713	2.680	(5.166 , 12.359)		8.071	(4.553 , 11.790)	
β_3 (stage III vs. I)	7.322	2.771	(3.414 , 11.331)		6.602	(2.946 , 10.420)	
Model 4				562.7			561.9
α_0 (intercept)	-4.408	0.7133	(-5.806 , -3.010)		-3.445	(-4.928 , -2.464)	
α_1 (age ≥ 50 vs. < 50)	-0.586	0.8465	(-2.244 , 1.073)		-1.558	(-2.834 , 0.037)	
α_2 (stage II vs. I)	1.472	0.7731	(-0.043 , 2.987)		0.559	(-0.5678 , 2.039)	
α_3 (stage III vs. I)	2.152	0.8377	(0.509 , 3.793)		1.142	(-0.1288 , 2.633)	
α_{12} (age vs. stage II)	0.477	1.1565	(-1.789 , 2.744)		1.041	(-1.089 , 2.889)	
α_{13} (age vs. stage III)	0.237	1.0077	(-1.737 , 2.212)		1.123	(-0.5672 , 2.725)	
β_0 (intercept)	-3.028	0.6244	(-4.252 , -1.804)		-3.015	(-9.219 , -1.836)	
β_1 (age ≥ 50 vs. < 50)	-8.734	1.8052	(-12.272 , -5.196)		-8.243	(-12.19 , -4.538)	
β_2 (stage II vs. I)	-1.186	1.1913	(-3.521 , 1.148)		-1.074	(-3.464 , 4.537)	
β_3 (stage III vs. I)	0.008	0.8434	(-1.644 , 1.661)		-0.075	(-2.014 , 5.786)	
β_{12} (age vs. stage II)	10.867	2.1123	(6.727 , 15.008)		9.907	(5.959 , 13.85)	
β_{13} (age vs. stage III)	8.359	1.9524	(4.532 , 12.186)		7.568	(3.585 , 11.62)	

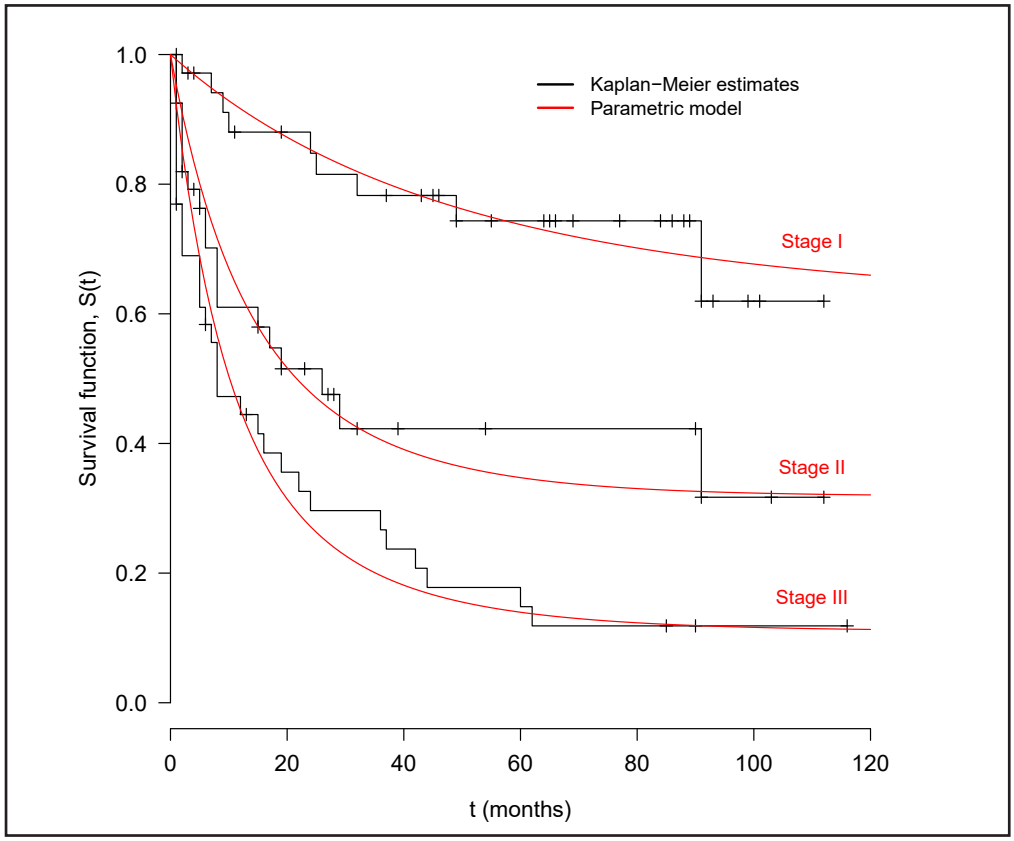

Figure 3: Kaplan–Meier and parametric survival curves of disease-free survival function according to the stage of the disease at the start of treatment. The parametric curves are based on the Model 2, with parameters estimated by the maximum likelihood approach. Censored observations are marked by ticks on the survival curves.

Although the Model 1 does not evidences a significant effect of the age on the disease-free survival, Model 4 suggest that the interaction between age and the clinical stages is important to understand the role of the age on the time to progression of the disease. The confidence and credible intervals for the interaction terms β_{12} and β_{13} do not include the zero value, suggesting a significant effect of interaction. In fact, the survival curves showed in the Figure 4 help us to understand this effect. The parametric curves in Figure 4 were obtained based on the Model 4, and shown that the age has a strong effect on the disease-free survival of patients in clinical stage II, but the age is not an important factor influencing the survival of patients in clinical stage III given that the survival curves showed in the panel (c) are close to each other.

Models 2, 3 and 4 have similar AIC and DIC values (Table 4), and these values are lower than the respective values calculated for the Model 1. However, under a clinical point of view, the Model 4 seems to be the most appropriate to the data, given that the Figure 3 shows the importance of considering the interaction terms for the interpretation of the pattern of association between age, clinical stages and the time to progression of the disease.

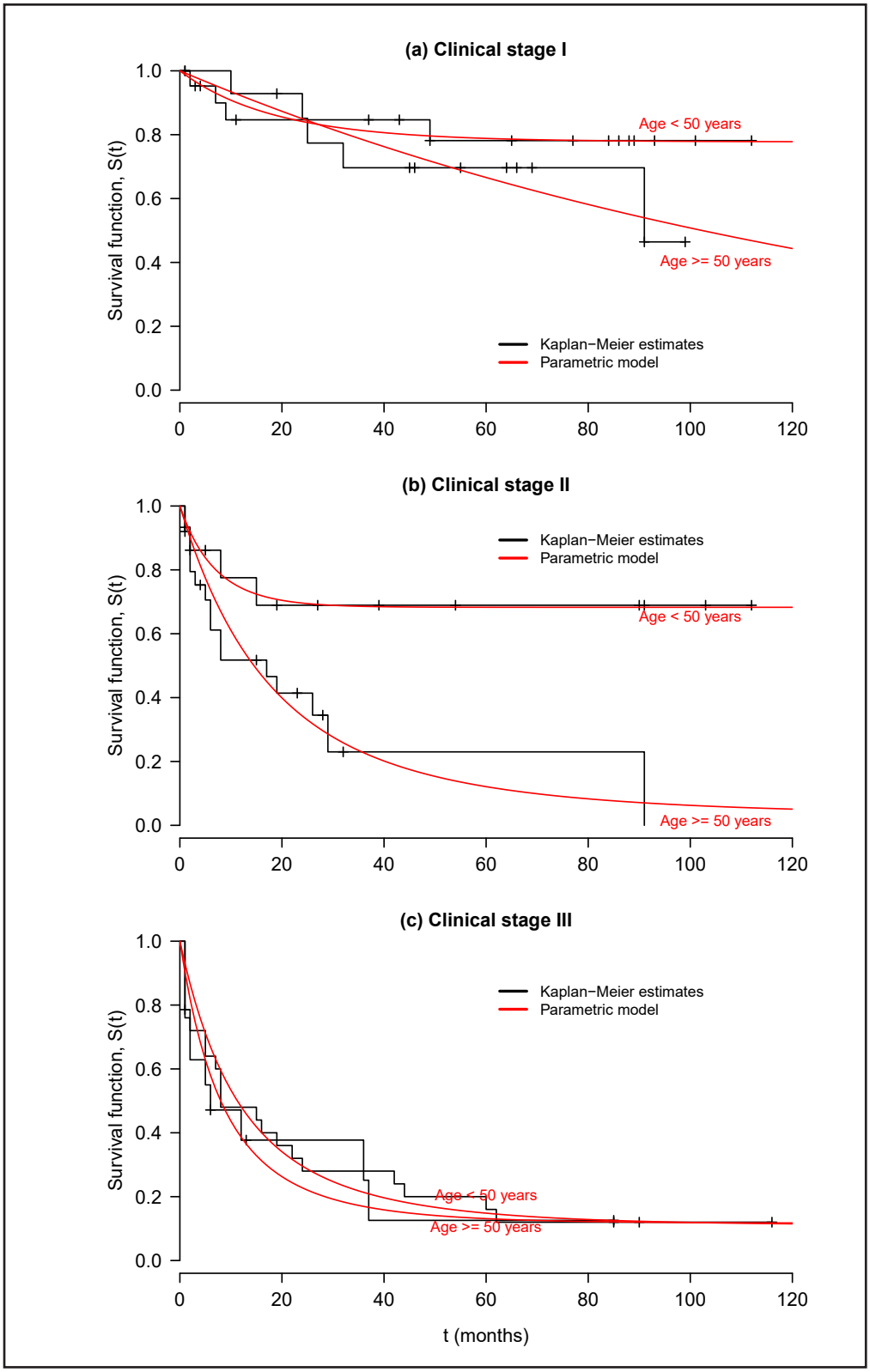

Figure 4: Kaplan–Meier and parametric survival curves of disease-free survival function according to the stage of the disease at the start of treatment and age. The parametric curves are based on the Model 4, with parameters estimated by the maximum likelihood approach. Censored observations are marked by ticks on the survival curves.

4 Conclusions

Time-to-an-event data including a proportion of individuals who are immune to the event of interest are common, especially in medical studies. Basic tools for survival analysis usually consider that the survival function $S(t)$ tends to zero as the time t tends to infinity, and this assumption is unrealistic if immune individuals are present. The use of models based on defective distributions is a suitable way to analyse data in this situation. In this way, the parametric model based on the modified Gompertz distribution allows for the estimation of the cure fraction and allows the insertion of a vector of covariates. Moreover, the model can be easily implemented in computational programs as SAS, R and OpenBUGS, as showed in the Appendix. To illustrate the use of the model we used a real data set from a cervical cancer study. We can note that the model satisfactorily fitted the data.

Acknowledgements

This research was supported by grants from CNPq (Brazil). The authors are grateful to the anonymous reviewers for their valuable comments and suggestions.

Referências

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: *Proceedings of the 2nd International Symposium on Information Theory*, pp. 267–281.

Balka, J., Desmond, A. F., McNicholas, P. D. (2011). Bayesian and likelihood inference for cure rates based on defective inverse Gaussian regression models. *Journal of Applied Statistics*, *38*(1), 127–144.

Brenna, S. M., Silva, I. D., Zeferino, L. C., Pereira, J. S., Martinez, E. Z., Syrjänen, K. J. (2004). Prognostic value of P53 codon 72 polymorphism in invasive cervical cancer in Brazil. *Gynecologic Oncology*, *93*(2), 374–380.

Cancho, V. G., Bolfarine, H. (2001). Modeling the presence of immunes by using the exponentiated-Weibull model. *Journal of Applied Statistics*, *28*(6), 659–671.

Cantor, A. B., Shuster, J. J. (1992) Parametric versus non-parametric methods for estimating cure rates based on censored survival data. *Statistics in Medicine*, *11*(7), 931–937.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, *38*(4), 1041–1046.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B. (2013). *Bayesian Data Analysis*, 3º edn. Chapman and Hall/CRC.

Gieser, P. W., Chang, M. N., Rao, P. V, Shuster, J. J., Pullen, J. (2014). Modelling cure rates using the Gompertz model with covariate information. *Statistics in Medicine*, *17*(8), 831–839.

Henningsen, A., Toomet, O. (2011). maxLik: A package for maximum likelihood estimation in R. *Computational Statistics*, *26*(3), 443–458.

Klein, J. P., Moeschberger, M. L. (2005). *Survival analysis: techniques for censored and truncated data*, 2º edn. Springer Science & Business Media.

Lambert, P. C., Thompson, J. R., Weston, C. L., Dickman, P. W. (2007). Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics*, *8*(3), 576–594.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., Schabenberger, O. (2006). *SAS for Mixed Models*, 2º edn. SAS Institute.

Lunn, D. J., Thomas, A., Best, N., Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325–337.

Maller, R. A., Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*, Wiley.

R Development Core Team (2009). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org`.

Rocha, R., Nadarajah, S., Tomazella, V., Louzada, F., Eudes, A. (2015). New defective models based on the Kumaraswamy family of distributions with application to cancer data sets. *Statistical Methods in Medical Research*, 1–23.

Rocha, R., Nadarajah, S., Tomazella, V., Louzada, F. (2017). A new class of defective models based on the Marshall–Olkin family of distributions for cure rate modeling. *Computational Statistics & Data Analysis*, *107*, 48–63.

Rocha, R. F., Tomazella, V. L. D., Louzada, F. (2014). Bayesian and classic inference for the Defective Gompertz Cure Rate Model. *Revista Brasileira de Biometria*, *32*(1), 104–114.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B Statistical Methodology*, *76*(3), 485–493.

Appendix

The following R function `rMGompertz` is used to generate random samples of size n from a modified Gompertz distribution with parameters α and β.

```
rMGompertz <- function(n,alpha,beta) {
 eta <- exp(-alpha/beta)
 m <- rbinom(n,prob=1-eta,size=1)
 u <- runif(n,0,1-eta)
 y0 <- -(1/beta)*log(1+beta/alpha*log(1-u))
 t0 <- ifelse(m,y0,Inf)
 maxti <- max(y0*m)
 w <- runif(n,0,maxti)
 t <- pmin(t0,w)
 d <- as.numeric(t0<w)
 dados <- data.frame(t,d)
return(dados) }
```

This is the OpenBUGS code for the Bayesian model based on the modified Gompertz distribution without covariates:

```
model {
 for (i in 1:N) {
   S[i] <- exp(-alpha/beta*(1-exp(-beta*t[i])))
   h[i] <- alpha*exp(-beta*t[i])
   L[i] <- S[i]*pow(h[i],d[i])
   logL[i] <- log(L[i])
   zeros[i] <- 0
   zeros[i] ~ dloglik(logL[i])  }
a ~ dgamma(0.001,0.001)
b ~ dgamma(0.001,0.001)
alpha<-1/a
beta<-1/b
eta<-exp(-(alpha/beta))  }
```

In this OpenBUGS code, `S[i]` is the survival function, `h[i]` is the hazard function, `L[i]` is the likelihood function, `t[i]` is the time to event and `d[i]` is a censoring indicator.

Under the frequentist approach, the following R code is used to implement the model using the function `maxLik` of the maxLik package (Henningsen and Toomet, 2011) for the maximization of the likelihood function.

```
log.f <- function(parms){
 alpha <- parms[1]
 beta  <- parms[2]
 if (parms[1]<0) return(-Inf)
 if (parms[2]<0) return(-Inf)
 St <- exp(-alpha/beta*(1-exp(-beta*t)))
 ht <- alpha*exp(-beta*t)
 like <- St * ht^d
 L <- sum(log(like))
 if (is.na(L)==TRUE) {return(-Inf)}
                   else {return(L)}    }
 library(maxLik)
 mle  <- maxLik(logLik=log.f,start=c(.06,.06))
 summary(mle)
```

Alternatively, SAS users can use the following code:

```
data article;
input t d;
cards;
28 0
8 1
116 0
2 1
...
1 1
;;
proc nlmixed data=article df=500;
 parms alpha=0.06 beta=0.06;
 bounds alpha>0, beta>0;
 St = exp(-alpha/beta*(1-exp(-beta*t)));
 ht = alpha*exp(-beta*t);
 eta = exp(-alpha/beta);
 like = St * ht**d;
 loglike = log(like);
 model t ~ general(loglike);
 estimate "eta" eta;
run;
```

The SAS procedure NLMIXED allows for the fitting of nonlinear and generalized linear models with random effects (Littell et al., 2006). However, if random effects are not reported in the SAS code, only fixed effects are included in the model. The `parms` statement describes the names of parameters and specifies initial values. In order to avoid problems of convergence or a Hessian matrix with negative eigenvalues, reasonable initial values should be specified for each parameter. An advantage of the use of SAS procedure NLMIXED is that the `estimate` statement allows to compute directly the standard error and Wald-type confidence limits for the parameter η, based on the delta method.

All the computational codes presented here can be readily adapted to situations in which there are multiple independent variables. For example, under the frequentist approach, the R code for the likelihood function of the Model 4 is the following.

```
log.f <- function(parms){
 alpha0 <- parms[1]
 alpha1 <- parms[2]
 alpha2 <- parms[3]
 alpha3 <- parms[4]
 alpha4 <- parms[5]
 alpha5 <- parms[6]
 beta0  <- parms[7]
 beta1  <- parms[8]
 beta2  <- parms[9]
 beta3  <- parms[10]
 beta4  <- parms[11]
 beta5  <- parms[12]
 alpha <- exp(alpha0 + alpha1*x1 + alpha2*x2
   + alpha3*x3 + alpha4*x1*x2 + alpha5*x1*x3)
 beta  <- exp(beta0 + beta1*x1 + beta2*x2
     + beta3*x3  + beta4*x1*x2 + beta5*x1*x3)
 St <- exp(-alpha/beta*(1-exp(-beta*t)))
 ht <- alpha*exp(-beta*t)
 like <- St * ht^d
 L <- sum(log(like))
 if (is.na(L)==TRUE) {return(-Inf)}
                          else {return(L)}  }
```