Zero-inflated cure rate model: an application to HIV/AIDS patients data in Mato Grosso do Sul

Bueno, M. V. [*] Rossi, R. M. [†] Peres, M. V. O. [‡]

Revista Eletrônica de Ciências Exatas e Tecnológicas

Submitted: 18 jan.2023. **Approved**: 20 abr.2023. **Published**: 18 fev.2024.

Edition: 1ª. **Volume**: 5º.

ABSTRACT

In this study, it was aimed to adjust some probability distributions to describe survival time in HIV/AIDS patients in Mato Grosso do Sul, Brazil, followed between 2009 and 2018. In the distributions discussed, it was necessary to implement parameters to model "zero" survival times (zero inflation) and the long duration times (cure rate). Based on Akaikes Criterion, the Kumaraswamy Generalized Gamma distribution was the best fit to the data, and was then used in a regression model that contained the explanatory variables: sex, race, and education. Based on this adjustment, female gender, white race, and education of more than eight years were associated with longer survival time. Based on these interpretations, one can discuss the need for HIV prevention and early diagnosis policies focused on specific groups associated with lower survival.

Keywords: Survival analysis, Censored data, Zero-inflated models, Long-term survival models.

RESUMO

Neste trabalho, objetivou-se ajustar algumas distribuições de probabilidade para descrever o tempo de sobrevida em pacientes com HIV/AIDS no Mato Grosso do Sul, Brasil, acompanhados entre 2009 e 2018. Nas distribuições abordadas, fez-se necessária a implementação de parâmetros para modelar os tempos "zerados" (inflação de zeros) e os de longa duração (fração de cura). Com base no Critério de Akaike, a distribuição Kumaraswamy Gama Generalizado foi a que melhor se ajustou aos dados, e então foi utilizada em um modelo de regressão que continha as variáveis explicativas: sexo, raça e escolaridade. Baseado nesse ajuste, o sexo feminino, a raça branca, e a escolaridade superior a oito anos foram associados a um maior tempo de sobrevida. Com base nessas interpretações, pode-se discutir a necessidade de políticas de prevenção e diagnóstico precoce de HIV focadas para grupos específicos associados à menor sobrevida.

Palavras-chave: Análise de sobrevivência, Dados censurados, Modelos inflacionados de zeros, Modelos de longa duração.

Contents

INTRODUCTION

Human immunodeficiency syndrome (AIDS) is a chronic disease, being the third stage of human immunodeficiency virus (HIV) infection, which has been present in Brazil since the 1980s. According to Brito, Castilho e Szwarcwald (2001), initially the disease was present in large urban centers and affected mainly male homosexuals and hemophiliacs. According to Melo, Almeida e Donalisio (2021), the epidemiological profile of HIV infection has undergone changes over the decades, no longer being mostly concentrated in large urban centers, spreading this way to countryside municipalities. Besides the dissemination to small and medium-sized municipalities, the involvement of more socially vulnerable populations and

ISSN: 2763-8855

[*] Marcos Vinicius Bueno. Possui graduação em Matemática pela Universidade Estadual do Paraná (UNESPAR), Mestrado em Bioestatística pela Universidade Estadual de Maringá (UEM). Filiação: Secretária de Estado e Educação do Paraná (SEED). E-mail: bueno.vinicius201@gmail.com

[†] Robson Marcelo Rossi. Possui graduação em Matemática pela Universidade Estadual de Maringá (UEM), Mestrado em Estatística pela Universidade Federal de São Carlos (UFSCar) e Doutorado em Zootecnia pela Universidade Estadual de Maringá (UEM). Filiação: Departamento de Estatística (DES) - Universidade Estadual de Maringa (UEM). rmrossi@uem.br

[‡] Marcos Vinicius de Oliveira Peres. Possui Graduação em Matemática pela Universidade Estadual do Paraná (UNESPAR), Mestrado em Bioestatística pela Universidade Estadual de Maringá (UEM) e Doutorado em Saúde na Comunidade pela Universidade de São Paulo (USP). Filiação: Colegiado de Matemática - Universidade Estadual do Paraná (UNESPAR), Campus de Paranavaí. marcos.peres@ies.unespar.edu.br

Table 1 – Probability distributions considered for the data fits.

Model	Probability density function	Author
Weibull (W)	$f(t) = \frac{\tau}{\alpha^{\tau}} t^{\tau-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\tau}\right\}$	Weibull (1939)
Burr XII (BXII)	$f(t) = \frac{\gamma\alpha t^{\alpha-1}}{\theta^{\alpha}} \left[1 + \left(\frac{t}{\theta}\right)^{\alpha}\right]^{-(\gamma+1)}$	Burr (1942)
Log-Normal (LN)	$f(t) = \frac{1}{\sqrt{2\pi}t\sigma} \exp\left\{-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^{2}\right\}$	Aitchison e Brown (1957)
Beta-Weibull (BW)	$f(t) = \frac{\gamma t^{\gamma-1}}{B(\alpha,\beta)\lambda^{\gamma}} \exp\left[-\beta\left(\frac{t}{\lambda}\right)^{\gamma}\right]$ $\times \left\{1 - \exp\left[-\left(\frac{t}{\lambda}\right)^{\gamma}\right]\right\}^{\alpha-1}$	Lambert et al. (2007)
Kumaraswamy Weibull (KW)	$f(t) = \frac{\lambda\varphi\tau}{\alpha} \left(\frac{t}{\alpha}\right)^{\tau-1} \exp\left[-\left(\frac{t}{\alpha}\right)^{\tau}\right]$ $\times \left\{\gamma_1\left[1, \left(\frac{t}{\alpha}\right)^{\tau}\right]\right\}^{\lambda-1}$ $\times \left(1 - \left\{\gamma_1\left[1, \left(\frac{t}{\alpha}\right)^{\tau}\right]\right\}^{\lambda}\right)$	Cordeiro, Ortega e Nadarajah (2010)
Kumaraswamy Generalized Gamma (KGG)	$f(t) = \frac{\lambda\varphi\tau}{\alpha\Gamma(k)} \left(\frac{t}{\alpha}\right)^{\tau k-1} \exp\left[-\left(\frac{t}{\alpha}\right)^{\tau}\right]$ $\times \left\{\gamma_1\left[k, \left(\frac{t}{\alpha}\right)^{\tau}\right]\right\}^{\lambda-1}$ $\times \left(1 - \left\{\gamma_1\left[k, \left(\frac{t}{\alpha}\right)^{\tau}\right]\right\}^{\lambda}\right)$	Pascoa, Ortega e Cordeiro (2011)
Kumaraswamy Burr XII (KBXII)	$f(t) = \alpha\gamma\beta\lambda\theta^{-\alpha} t^{(\alpha-1)} \left[1 + \left(\frac{t}{\theta}\right)^{\alpha}\right]^{-(\gamma+1)}$ $\times \left\{1 - \left[1 + \left(\frac{t}{\theta}\right)^{\alpha}\right]^{-\gamma}\right\}^{(\lambda-1)}$ $\times \left[1 - \left\{1 - \times \left[1 + \left(\frac{t}{\theta}\right)^{\alpha}\right]^{-\gamma}\right\}^{\lambda}\right]^{(\beta-1)}$	Paranaíba et al. (2013)
Exponentiated Power Generalized Weibull (EPWG)	$f(t) = \alpha\beta\lambda\gamma t^{\gamma-1} (1 + \lambda t^{\gamma})^{\alpha-1}$ $\times \frac{\exp\left[1 - (1 + \lambda t^{\gamma})^{\alpha}\right]}{\left\{1 - \exp\left[1 - (1 + \lambda t^{\gamma})^{\alpha}\right]\right\}^{1-\beta}}$	Pena-Ramirez et al. (2018)

Where: $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ (Beta function), such that $\Gamma(z) = \int_0^{\infty} t^{z-1}e^{-t}dt$ (Gamma function) and $\gamma_1(k,x) = \gamma(k,x)/\Gamma(k) = \left(\int_0^{x} w^{k-1}e^{-w}dw\right)/\Gamma(k)$.

the growing number of women infected by the virus are marks of the changes that have occurred over time in the country (CANDIDO et al., 2021).

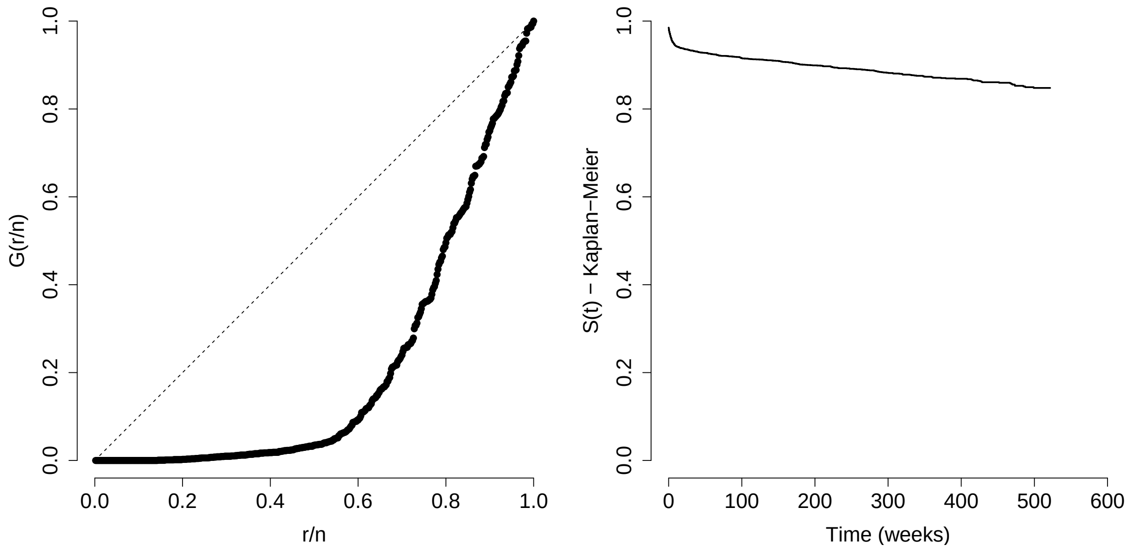
An exploratory analysis of spatial data on AIDS incidence for each Brazilian state between the years 1992 and 2017 is presented by Ribeiro, Fonseca e Pereira (2019), allowing to observe the evolution of the incidence rate of the disease over the years. According to the authors, even though the epidemic is more evenly distributed throughout the country in current times, there is still considerable incidence in the states of Rio de Janeiro, Santa Catarina, and Rio Grande do Sul, and the Northern region has become part of the epidemic's areas of concentration.

According to data from the 2020 Epidemiological Bulletin of HIV/AIDS, (BRASIL, 2020), in Brazil, between 2007 and June 2020, a total of 342,459 cases of HIV infection, of which 44.4%, 20%, 19%, 9% and 7.6% belong to the Southeast, South, Northeast, North and Center-West regions, respectively.

Survival analysis consists of the area of Statistics responsible for studying the time until the occurrence of a certain event of interest. In this context, several papers can be found in the literature analyzing HIV/AIDS data as in Colosimo e Vieira (1996), Melo, Donalisio e Cordeiro (2017), Medeiros et al. (2017), Zaslavsky, Goulart e Ziegelmann (2019), Müller e Borges (2020), Melo et al. (2021), and several others. However, most papers using data in this context, start from a semi-parametric approach through the Cox (1972) regression

ISSN: 2763-8855
RΣCΣT Revista Eletrônica de Ciências Exatas e Tecnológicas

Figure 1 – TTT-plot and Kaplan-Meier estimated survival function, respectively.



model, without the need to assume of a probability distribution for the response variable.

Although scarcer, some work on survival analysis in the parametric context involving HIV/AIDS data can be seen in Nakhaee e Law (2011), Pascoa, Ortega e Cordeiro (2011), Million (2018) and Bueno e Rossi (2021).

Also in the context of survival analysis, there may exist within the population under study a proportion of individuals who are not likely to present the event of interest, so that common probabilistic models are not able to estimate this proportion of individuals. Thus, it is necessary to use models with a cure rate, where these are usually immune or cured of the outcome of interest. Martinez et al. (2013), Peres, Santos e Oliveira (2020), Moraes, Previdelli e Silva (2021), among others, are examples of works using this approach.

Another situation, not so common in practice, is the occurrence of observations of null times (or zeros) and, in high frequency, are called zero-inflated data, and it is then necessary to use distributions that accommodate such a situation, as observed in Hashimoto (2013), Júnior, Moreira e Louzada (2017), Calsavara et al. (2019), among others.

In this context, the objective of this work was to perform parametric modeling, considering different distributions for the response of survival time of HIV/AIDS patients, in the presence of cure rate and zero inflation for fitting the model to the data.

MATERIAL AND METHODS

According to Carvalho et al. (2011), survival analysis is a branch of Statistics used when time is the object of interest, which can be interpreted as the time until the occurrence of an event of interest or the risk of occurrence of an event of interest for a given unit of time. Survival, hazard, and cumulative hazard functions are some characteristic functions in this area of study, so a discussion of these can be seen in Lee e Wang (2003). The distinction between commonly seen statistical techniques and those employed in survival analysis lies in the presence of what is referred to as “censoring”. Censoring data occurs when the information regarding failure time is incomplete for all individuals under study. Censoring can be right-censoring (failure events occur after the end of the study or are unknown) or left-censoring (failure events occur before the start of the study). Survival analysis is capable of handling these censoring situations and utilizing the available partial information.

Colosimo e Giolo (2006), suggest initially performing a descriptive analysis of the data by estimating the survival function, and from there, estimate the other statistics of interest. There are several techniques for estimating the survival function in the presence of censored data, however, the most usual is through the Kaplan-Meier estimator (or product-limit estimator). Proposed by Kaplan e Meier (1958), the estimator is defined as:

$$\hat{S}(t) = \prod_{j:t_j<t} (1 - d_j/n_j),$$

so that d_j is the number of failures and n_j the number of individuals at risk at each time t_j. An important measure to describe survival data with censoring is the restricted mean survival time (RMST). According to Kim,

ISSN:
2763-8855
REΣET Revista Eletrônica de Ciências Exatas e Tecnológicas

Table 2 – Descriptive summary for the explanatory variables.

Variables	Classes	Frequency	%	Censoring (%)	Zeros (%)	RMST$_{se}$
Sex	Male	3,004	63.14	89.04	1.70	460.25 (3.19)
	Female	1,754	36.86	90.25	1.48	471.27 (3.66)
Race	White	2,094	44.01	90.97	1.24	473.78 (3.22)
	Non-white	2,664	55.99	88.32	0.98	456.80 (3.44)
Education (in years)	> 8	2,008	42.20	93.38	1.10	483.62 (3.21)
	≤ 8	2,750	57.80	86.65	2.00	452.39 (3.36)

RMST$_{se}$: restricted mean survival time and standard error.

Uno e Wei (2017), RMST is a widely recognized but often overlooked measure that represents the average duration of event-free survival up to a specific, clinically significant time point. It can be understood as the area under the Kaplan-Meier curve from the beginning of the study up to that particular point. The difference in RMST indicates the increase or decrease in event-free survival time attributed to the treatment compared to the control group within this timeframe.

Probabilistic models in survival analysis start from the assumption that the response variable follows a probability distribution. These models provide a statistical framework for understanding the underlying probability distribution of survival times and capturing the dynamic nature of event occurrence. By incorporating covariates and time-dependent factors, probabilistic models allow for the estimation of survival probabilities and hazard rates over time. They also enable the assessment of the impact of various risk factors on survival outcomes. With their ability to handle censored data and account for time-varying covariates, probabilistic models offer valuable insights into the prediction, understanding, and interpretation of survival data in diverse fields such as medical research, epidemiology, and social sciences. The total time-to-test plot (TTT-plot), originally proposed by Barlow e Campo (1975) and generalized by Aarset (1985), allows the shape of the hazard function to be identified, making it possible to consider models best suited to the data in question. Table 1 shows the distributions considered in this study.

It is known that these originally do not have support for observations of time equal to 0 and therefore, statistical analyses on data characterized by the presence of excess zeros are performed using the inflated models, as presented in Hashimoto (2013). Considering T a random variable representing the time until the occurrence of the event of interest, with excess or inflation of zeros, according to Calsavara et al. (2019), the survival function is given as follows:

$$S(t) = (1 - p_0)S_0(t),$$

where $0 < p_0 < 1$ is the parameter that models the proportion of zero times and $S_0(t)$ is the of eigen-survival.

Thus, the probability density function $f(t)$ is given as follows:

$$f(t) = \begin{cases} p_0, & \text{if} \quad t = 0 \\ (1 - p_0)f_0(t), & \text{if} \quad t \neq 0. \end{cases}$$

Cure rate survival models can be used to model survival data where observations have a long survival time, so this happens when part of the individuals do not receive the event of interest, or it is not observed during the study time. In this approach, it is assumed that the population under study is composed of individuals who are susceptible to experiencing the event of interest, as well as individuals who are not susceptible (MALLER; ZHOU, 1996).

In this approach the population survival function is given by the standard mixture model:

$$S(t) = p_1 + (1 - p_1)S_0(t),$$

where $0 < p_1 < 1$ is the proportion of individuals who are not susceptible to the occurrence of the event of interest (cure rate), and $S_0(t)$ is the eigen-survival function.

According to Júnior, Moreira e Louzada (2017), considering models that address both zero inflation and cure rate, their survival and probability density functions are defined, respectively by:

$$S(t) = p_1 + (1 - p_0 - p_1)S_0(t)$$

and

$$f(t) \begin{cases} p_0, & \text{if} \quad t = 0 \\ (1 - p_0 - p_1)f_0(t), & \text{if} \quad t \neq 0. \end{cases}$$

The parameter estimates, considering the maximum likelihood method, in the context of survival analysis, with the presence of zero inflation and cure rate, are found by maximizing the likelihood function given by:

ISSN: 2763-8855

RΣCΣT Revista Eletrônica de Ciências Exatas e Tecnológicas

Table 3 – Estimates of the parameters of the considered distributions.

Distribution (Parameters)	Estimates (Standard Error)							AIC
ZCW(p_0, p_1, τ, α)	0.015 (0.001)	0.747 (0.020)	0.435 (0.017)	989.606 (1.438)				7,641.44
ZCBXII$(p_0, p_1, \theta, \alpha, \gamma)$	0.013 (0.001)	0.567 (0.039)	1.202 (0.312)	1.466 (0.153)	0.028 (0.003)			7,650.73
ZCLN$(p_0, p1, \mu, \sigma)$	0.015 (0.002)	0.674 (0.087)	7.343 (1.168)	3.976 (0.403)				7,617.01
ZCBW$(p_0, p_1, \lambda, \alpha, \beta, \gamma)$	0.040 0.004	0.596 0.087	0.990 0.142	0.863 0.327	0.035 0.013	0.420 0.103		7,769.48
ZCKW$(p_0, p_1, \tau, \alpha, \lambda, \varphi)$	0.015 (0.001)	0.725 (0.052)	0.411 (<0.001)	0.081 (<0.001)	1.524 (<0.001)	0.018 (0.005)		7,631.06
ZCKGG$(p_0, p_1, \tau, \alpha, \lambda, \varphi, k)$	0.015 (0.002)	0.709 (0.054)	0.303 (0.004)	0.024 (0.002)	16.041 (0.005)	0.030 (0.007)	0.696 (0.003)	**7,596.49**
ZCKBXII$(p_0.p_1.\lambda.\beta.\alpha.\gamma.\theta)$	0.014 (0.002)	0.583 (0.035)	1.351 (0.160)	0.081 (0.009)	1.124 (0.090)	0.394 (0.040)	0.285 (0.056)	7,684.43
ZCEPWG$(p_0.p_1.\alpha.\beta.\lambda.\gamma)$	0.015 (0.002)	0.646 (0.090)	0.104 (0.040)	1.291 (0.186)	0.0820 (0.402)	0.720 (0.138)		7,617.64

$$L(p_0, p_1, \mathrm{D}, \phi) = \prod_{i:t_i=0} p_0 \prod_{i:t_i>0} \left[(1-p_0-p_1) f_0(t_i|\phi)\right]^{\delta_i} \left[p_1 + (1-p_1-p_0) S_0(t_i|\phi)\right]^{1-\delta_1},$$

where ϕ is the parameter vector, $\mathrm{D} = \{t_i, \delta_i\}$, such that $\delta_i = 1$ indicates an observation that presented the event of interest and $\delta_i = 0$ indicates censoring.

The data analyzed in this paper come from an epidemiological study presented by Werle (2021). The database originally contains information regarding 9,021 individuals diagnosed with HIV/AIDS in the state of Mato Grosso do Sul, Brazil, observed between 2009 and 2018. We removed 4,263 observations that contained missing values, so the data set used contained observations from 4,758 patients. The study originated from data that complied with the national and international norms of ethics in research with human beings, and was approved with register number 3.789.678. In the present study, the National Health Council (CNS) and National Research Ethics Committee (CONEP) review based on Resolution 466/2012 CEP-CONEP is waived.

After estimating the parameters of the models, done with the R statistical environment (R Core Team, 2022), the best one was selected based on Akaike's criterion (AIC). In addition, the Cox e Snell (1968) residuals were used to check the quality of the fit, where, according to Lawless (2003), for the model to present an adequate fit, the Cox-Snell residuals should follow a standard Exponential distribution.

RESULTS AND DISCUSSION

Considering HIV/AIDS data set, in Table 2 is shown a descriptive summary of the explanatory variables considered is presented, as well as absolute and relative frequencies, and the proportion of censoring. The table reveals a substantial percentage of censoring across all variables. Additionally, there is a significant number of zero survival times, justifying the use of the zero-inflated survival model. Furthermore, it is evident that the RMST differs significantly for each category of the variables, indicating that these covariates may have a significant impact on patients' survival time. The high percentage of censoring is directly linked to treatment advancements, enabling a larger number of individuals with illness to avoid experiencing the event under study (in this case, death due to HIV/AIDS).

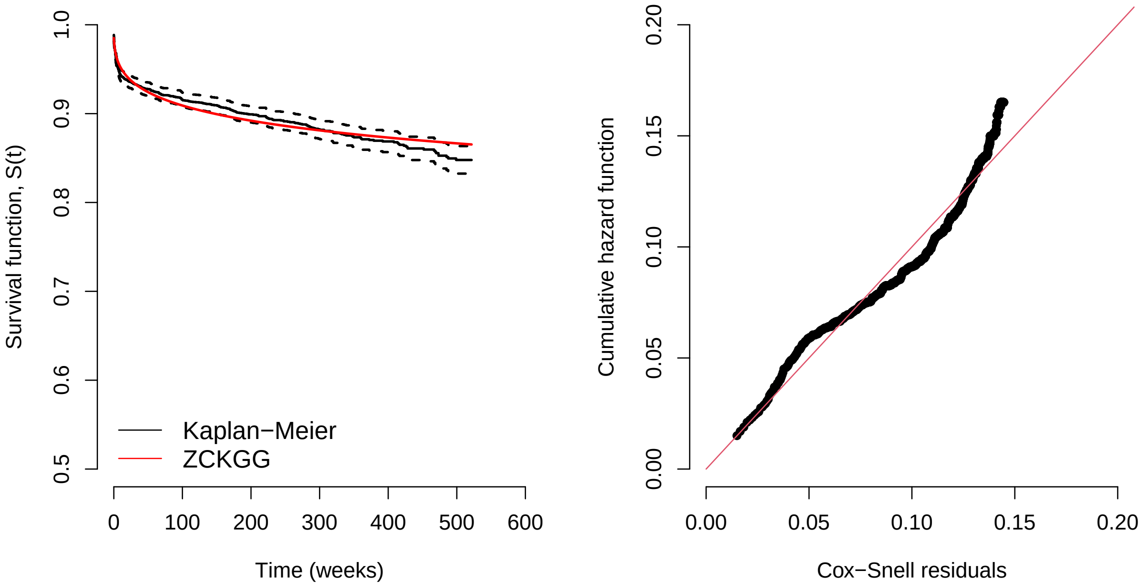
The presence of the cure rate is observed in the data through the Kaplan-Meier estimates (Figure 1). Note that it is not possible to observe a drop in the KM plot, and there is still a noticeable plateau trend at the far right of the graph. The TTT chart (Figure 1) indicates that the hazard function presents decreasing behavior.

The parameter estimates for the distributions considered (with the addition of the zeros inflation and cure rate parameters, indicated by ZC before the acronym of the distribution), as well as the AIC, are presented in Table 3, indicating that the ZCKGG model was the most suitable for the representation of the data, since it presented the lowest value for the AIC (7,596.49). Furthermore, the

ISSN: 2763-8855

RECET Revista Eletrônica de Ciências Exatas e Tecnológicas

Figure 2 – Comparison between empirical and adjusted survival curves, and comparison between Cox-Snell residuals and cumulative hazard function, respectively.



KGG distribution has support for decreasing hazard function, as was observed in the behavior of the TTT-plot (Figure 1).

Several papers using zeros-inflated or cure rate data can be found in the literature, such as in Calsavara et al. (2019), where the defective Gompertz and inverse Gaussian models with zero inflation are introduced, and Moraes, Previdelli e Silva (2021) present a modeling for the cure rate considering a Weibull model, under a Bayesian perspective, for breast cancer data observed between 2004 and 2016 in Paraná, Brazil.

In particular, for data from HIV-infected patients, an example of work using cure rate can be seen in Freitas et al. (2021), where an application of a parametric Weibull Discrete Exponential survival model with cure rate is presented using frequentist and Bayesian approaches.

Some applications using both approaches can be seen in Júnior, Moreira e Louzada (2017) for credit risk data, in Souza et al. (2021) for invasive cervical cancer data, among others.

After choosing the best distribution for the data, a comparison was made between the adjusted and empirical survival function (by Kaplan-Meier), and between the Cox-Snell residuals and the cumulative risk function. Observation of these two graphical representations indicates that the adjusted model is (satisfactorily) adequate to describe the data covered (Figure 2).

To build the ZCKGG regression model, we proposed to add the vector of parameters $(\boldsymbol{\beta})$ associated with the explanatory variables in parameter φ, one of the four shape parameters $(\tau, \lambda, \varphi, k)$ of the distribution. Thus, the regression model is given by:

$$f(t) = (1 - p_0 - p_1)\,\frac{\lambda\varphi(\boldsymbol{x}'\boldsymbol{\beta})\tau}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{k\tau-1}\exp\left[-\left(\frac{t}{\alpha}\right)^{\tau}\right]\left\{\gamma_1\left[k,\left(\frac{t}{\alpha}\right)^{\tau}\right]\right\}^{\lambda-1}\times\left(1-\left\{\gamma_1\left[k,\left(\frac{t}{\alpha}\right)^{\tau}\right]\right\}^{\lambda}\right)^{\varphi(\boldsymbol{x}'\boldsymbol{\beta})-1},$$

where $\varphi(\boldsymbol{x}'\boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)$, which x_p represents the *p*-th explanatory variable.

To build the regression model, it was decided to remove the variables related to the mode of virus infection that were part of the model indicated by (WERLE, 2021), due to strong with relationship with other variables or that did not show improvement in the parametric regression model. The parameter estimates for the ZCKGG regression model are shown in Table 4.

ISSN:
2763-8855
RECET
Revista Eletrônica de Ciências Exatas e Tecnológicas

Table 4 – Parameter estimates (and standard error) of regression models.

Parameter	Classes	Model	
		ZCKGG	ZKGG
p_0	-	0.0151 (0.0018)	0.0126 (0.0015)
p_1	-	0.6316 (0.1431)	-
τ	-	0.1750 (0.5138)	0.1657 (0.0152)
α	-	0.8571 (1.2255)	3.0047 (0.7987)
λ	-	5.0539 (2.6459)	1.3489 (0.1364)
k	-	1.0096 (1.9157)	1.3963 (0.1529)
β_0	-	-1.7556 (9.1419)	-2.5603 (0.287)
β_1: Sex	Male	-	-
	Female	-0.3097 (0.1301)	-1.8194 (0.1584)
β_2: Race	White	-	-
	Non-white	0.2442 (0.1217)	0.2352 (0.0978)
β_3: Education (years)	> 8	-	-
	≤ 8	0.6916 (0.1404)	0.5748 (0.1031)
AIC		7,582.74	7,887.34

In order to verify the suitability of the explanatory variables separately to the ZCKGG model, comparisons between the empirical survival curve (via Kaplan-Meier) and the one estimated via regression models were constructed (Figure 3).

To verify the importance of the cure rate parameter, a new KGG (Kumaraswamy Generalized Gamma) regression model was constructed, incorporating only zero inflation (ZKGG), which estimates are also presented in Table 4. The AIC value is favorable, as it is the lowest value, to the model with the cure rate (AIC = 7,582.45) compared to the second model (AIC = 7,887.34). By replacing the parameter estimates in the regression models, in both, female gender was associated with longer survival, while non-white race and education of 8 years or less were associated with shorter survival.

The effects of the explanatory variables sex, race and education coincide with those presented by the Cox regression model presented by Müller e Borges (2020), for HIV/AIDS data for the Region of Campos Gerais (PR), Brazil.

Considering the AIC values we can state that there is evidence that the model with cure rate is more appropriate to model these data, thus providing an estimate of the proportion of individuals who will not experience the event of interest, interpreted by the parameter p_1. Thus, according to the ZCKGG models, with and without explanatory variables, the proportion of individuals who will not die from HIV/AIDS in the state of Mato Grosso do Sul are 63.16% and 70.90%, respectively. This high proportion is believed to be due to the advances in medicine developed for the treatment of HIV/AIDS in recent decades. Furthermore, the use of the cure rate parameter for the data discussed, as expected, goes against some work in the literature, such as in Varshney et al. (2018) and Freitas et al. (2021).

Once the best regression model (ZCKGG) was chosen, Cox-Snell residuals were constructed to check the quality of the model fit (Figure 4). Although some points showed different behavior than expected, the graphical interpretation of the residuals indicates an acceptable model due to the complexity of the modeling.

Wald's test, built via the maxLik package (HENNINGSEN; TOOMET, 2011) to verify the significance of the explanatory variables present in the final model, at the 5% significance level, whose null hypothesis is that the parameter estimate is equal to 0, indicates that only the intercept (β_0) was not marginally significant among the explanatory variables in the ZCKGG regression model (Table 5). In this model, α is the scale parameter andτ, λ, k, and $\varphi(\boldsymbol{x'\beta}) = \exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3\}$ are responsible for modeling the form of the survival function.

ISSN: 2763-8855

RECET Revista Eletrônica de Ciências Exatas e Tecnológicas

Figure 3 – Comparisons between Kaplan-Meier curves and the estimated survival function are made for each regression model, based on explanatory variables.

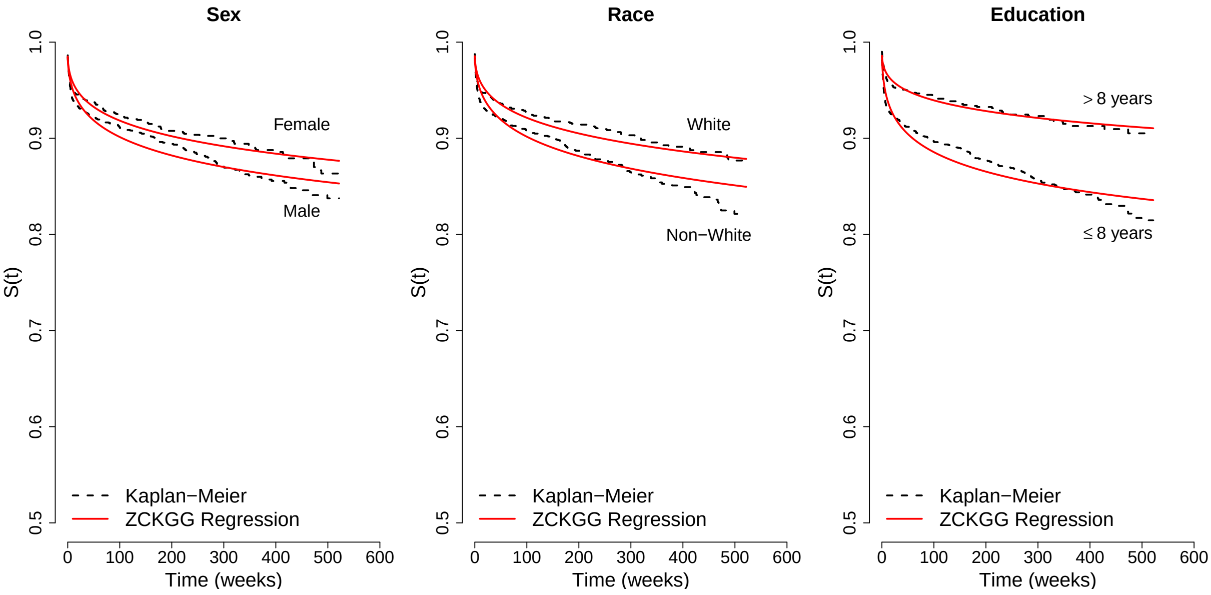


Table 5 – Wald test for the significance of the explanatory variables of the ZCKGG regression model.

Parameter	Statistics test	Value p
β_0	-0.192	0.8477
β_1	-2.381	0.0173
β_2	2.007	0.0448
β_3	4.927	8.36×10^{-7}

The zero-inflation parameter ($p_0 = 0.0151$)gives the probability (approximately 1.51%) that individuals will present with the event of interest zero weeks after diagnosis. The cure rate parameter ($p_1 = 0.6316$) indicates that 63.16% of the study population will not experience the event of interest. Possibly, this high proportion is due to the fact that the individuals maintained their treatment correctly after the HIV-positive diagnosis.

CONCLUSION

Among the probabilistic models assessed, the Generalized Gamma Kumaraswamy distribution with cure rate and zero inflation was the most suitable to represent the data. Furthermore, comparisons between the regression models with and without the cure rate parameter show that the presence of the cure rate resulted in a better fit, corroborating with the survival curve estimated via Kaplan-Meier.

The identification of groups with shorter survival time may help in future policies of prevention and early diagnosis for these groups, aiming to prevent possible deaths or cases of evolution of the viral infection.

ACKNOWLEDGEMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior Brasil (CAPES) Finance Code 001.

Bibliography

AITCHISON, J.; BROWN, J. A. **The Lognormal distribution with special reference to its uses in economics**. Cambridge Univ. Press, 1957. Citado na página 2.

BARLOW, R. E.; CAMPO, R. A. ***Total Time on Test Processes and Applications to Failure Data Analysis***. [S.l.], 1975. Citado na página 4.

BRASIL. ***Boletim epidemiológico: HIV/AIDS***. Brasília: Ministério da Saúde, 2020. Citado na página 2.

BRITO, A. M. d.; CASTILHO, E. A. d.; SZWARCWALD, C. L. **AIDS e infecção pelo HIV no Brasil: uma epidemia multifacetada**. *Revista da sociedade brasileira de medicina tropical*, SciELO Brasil, v. 34, n. 2, p. 207–217, 2001. Citado na página 1.

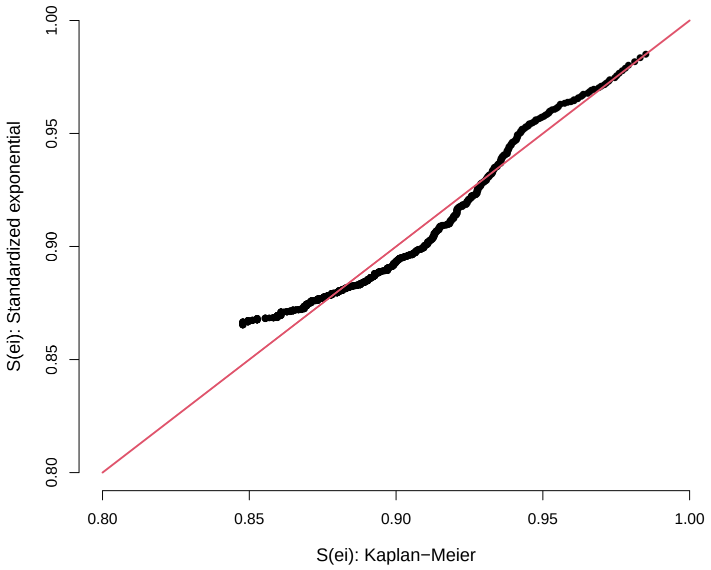
BUENO, M. V.; ROSSI, R. M. **Análise de sobrevivência para dados de HIV/AIDS**: uma abordagem Bayesiana. *Brazilian Journal of Development*, Curitiba, v. 7, n. 6, p. 55797–55805, 2021. Citado na página 3.

BURR, I. W. **Cumulative frequency functions**. *The Annals of Mathematical Statistics*, JSTOR, v. 13, n. 2, p. 215–232, 1942. Citado na página 2.

CALSAVARA, V. F. et al. **Zero-adjusted defective regression models for modeling lifetime data**. *Journal of Applied Statistics*, Taylor & Francis, v. 46, n. 13, p. 2434–2459, 2019. Citado 3 vezes nas páginas 3, 4, and 6.

ISSN: 2763-8855
REΣET Revista Eletrônica de Ciências Exatas e Tecnológicas

Figure 4 – Cox-Snell residual analysis for the ZCKGG regression model.



CANDIDO, P. G. G. et al. **Adherence to antiretroviral therapy among women living with HIV/AIDS in the interior of the Brazilian state of Pará**: cross-sectional study. *São Paulo Medical Journal*, SciELO Brasil, v. 139, n. 2, p. 99–106, 2021. Citado na página 2.

CARVALHO, M. S. et al. ***Análise de sobrevivência****: teoria e aplicações em saúde*. 2. ed. Rio de Janeiro: Fiocruz, 2011. Citado na página 3.

COLOSIMO, E. A.; GIOLO, S. R. ***Análise de sobrevivência aplicada***. São Paulo: Editora Blucher, 2006. Citado na página 3.

COLOSIMO, E. A.; VIEIRA, A. M. C. **O Modelo de Regressão de Cox com Covariável Dependente do Tempo**: Uma Aplicaçao envolvendo Pacientes Infectados pelo HIV. *Revista Brasileira de Estatıstica*, v. 54, n. 57, p. 139–152, 1996. Citado na página 2.

CORDEIRO, G. M.; ORTEGA, E. M.; NADARAJAH, S. **The Kumaraswamy Weibull distribution with application to failure data**. *Journal of the Franklin Institute*, Elsevier, v. 347, n. 8, p. 1399–1429, 2010. Citado na página 2.

COX, D. R. **Regression models and life-tables**. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 34, n. 2, p. 187–202, 1972. Citado na página 2.

COX, D. R.; SNELL, E. J. **A general definition of residuals**. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 30, n. 2, p. 248–265, 1968. Citado na página 5.

FREITAS, B. C. L. et al. **Classical and Bayesian inference approaches for the exponentiated discrete Weibull model with censored data and a cure fraction**. *Pakistan Journal of Statistics and Operation Research*, College of Statistical and Actuarial Sciences, v. 17, n. 2, p. 467–481, 2021. Citado 2 vezes nas páginas 6 and 7.

HASHIMOTO, E. M. ***Modelo de regressão Gama-G em análise de sobrevivência***. 177 p. Tese (Doutorado em Ciências) — Universidade de São Paulo, Piracicaba, 2013. Citado 2 vezes nas páginas 3 and 4.

HENNINGSEN, A.; TOOMET, O. **maxLik**: A package for maximum likelihood estimation in R. *Computational Statistics*, v. 26, n. 3, p. 443–458, 2011. Citado na página 7.

JÚNIOR, M. R. O.; MOREIRA, F.; LOUZADA, F. **The zero-inflated promotion cure rate model applied to financial data on time-to-default**. *Cogent Economics & Finance*, Cogent OA, v. 5, n. 1, p. 1395950, 2017. Citado 3 vezes nas páginas 3, 4, and 6.

KAPLAN, E. L.; MEIER, P. **Nonparametric estimation from incomplete observations**. *Journal of the American Statistical Association*, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958. Citado na página 3.

KIM, D. H.; UNO, H.; WEI, L.-J. **Restricted mean survival time as a measure to interpret clinical trial results**. *JAMA cardiology*, American Medical Association, v. 2, n. 11, p. 1179–1180, 2017. Citado na página 4.

LAMBERT, P. C. et al. **Estimating and modeling the cure fraction in population-based cancer survival analysis**. *Biostatistics*, Oxford University Press, v. 8, n. 3, p. 576–594, 2007. Citado na página 2.

LAWLESS, J. F. **Statistical models and methods for lifetime data**. John Wiley & Sons, Nova Yorkl, v. 362, p. Total de paginas, 2003. Citado na página 5.

LEE, E. T.; WANG, J. ***Statistical methods for survival data analysis***. 3. ed. New Jersey: John Wiley & Sons, 2003. Citado na página 3.

MALLER, R. A.; ZHOU, X. ***Survival analysis with long-term survivors***. New York: Wiley, 1996. Citado na página 4.

MARTINEZ, E. Z. et al. **Mixture and non-mixture cure fraction models based on the generalized modified Weibull distribution with an application to gastric cancer data**. *Computer Methods and Programs in Biomedicine*, Elsevier, v. 112, n. 3, p. 343–355, 2013. Citado na página 3.

MEDEIROS, A. R. C. et al. **Análise de sobrevida de pessoas vivendo com HIV/AIDS**. *Revista de Enfermagem UFPE Online*, v. 11, n. 1, p. 47–56, 2017. Citado na página 2.

ISSN: 2763-8855

RΣCΣT Revista Eletrônica de Ciências Exatas e Tecnológicas

MELO, G. C. d. et al. **Tempo de sobrevida e distância para acesso a tratamento especializado por pessoas vivendo com HIV/Aids no estado de Alagoas, Brasil**. *Revista Brasileira de Epidemiologia*, SciELO Brasil, v. 24, 2021. Citado na página 2.

MELO, M. C. d.; ALMEIDA, V. C. d.; DONALISIO, M. R. **Trend incidence of HIV-AIDS according to different diagnostic criteria in Campinas-SP, Brazil from 1980 to 2016**. *Ciência & Saúde Coletiva*, SciELO Brasil, v. 26, p. 297–307, 2021. Citado na página 1.

MELO, M. C. d.; DONALISIO, M. R.; CORDEIRO, R. C. **Sobrevida de pacientes com AIDS e coinfecção pelo bacilo da tuberculose nas regiões Sul e Sudeste do Brasil**. *Ciência & Saúde Coletiva*, SciELO Public Health, v. 22, p. 3781–3792, 2017. Citado na página 2.

MILLION, W. **Statistical Modeling for the Survival of HIV/AIDS Patients Treated with Highly Active Anti-Retroviral Therapy (HAART)**: A case study at Dilchora Hospital, Dire Dawa, Ethiopia. *Journal of Biometrics & Biostatistics*, v. 9, n. 5, p. 1–10, 2018. Citado na página 3.

MORAES, T. E. N. T. de; PREVIDELLI, I.; SILVA, G. L. da. **A Bayesian Weibull analysis of breast cancer data with long-term survivors in Paraná State, Brazil**. *Revista Brasileira de Biometria*, v. 39, n. 2, p. 293–310, 2021. Citado 2 vezes nas páginas 3 and 6.

MÜLLER, E. V.; BORGES, P. K. de O. **Sobrevida de pacientes HIV/AIDS em tratamento antirretroviral e fatores associados na Região dos Campos Gerais, Paraná**: 2002-2014. *Brazilian Journal of Development*, v. 6, n. 5, p. 28523–28542, 2020. Citado 2 vezes nas páginas 2 and 7.

NAKHAEE, F.; LAW, M. **Parametric modelling of survival following HIV and AIDS in the era of highly active antiretroviral therapy**: data from Australia. *Eastern Mediterranean Health Journal*, v. 17, n. 3, p. 231–237, 2011. Citado na página 3.

PARANAÍBA, P. F. et al. **The Kumaraswamy Burr XII distribution: theory and practice**. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 83, n. 11, p. 2117–2143, 2013. Citado na página 2.

PASCOA, M. A. D.; ORTEGA, E. M.; CORDEIRO, G. M. **The Kumaraswamy Generalized Gamma distribution with application in survival analysis**. *Statistical Methodology*, Elsevier, v. 8, n. 5, p. 411–433, 2011. Citado 2 vezes nas páginas 2 and 3.

PENA-RAMIREZ, F. A. et al. **The exponentiated power generalized Weibull**: Properties and applications. *Anais da Academia Brasileira de Ciências*, SciELO Brasil, v. 90, n. 3, p. 2553–2577, 2018. Citado na página 2.

PERES, M. V. D. O.; SANTOS, F. S. D.; OLIVEIRA, R. P. D. **Estimation of survival and hazard curves of mixture Mirra cure rate model: Application to gastric and breast cancer data**. *Biom Biostat Int J*, v. 9, n. 4, p. 132–137, 2020. Citado na página 3.

R Core Team. ***R***: *A Language and Environment for Statistical Computing*. Vienna, Austria, 2022. Disponível em: <https://www.R-project.org/>. Citado na página 5.

RIBEIRO, R. A.; FONSECA, F. F.; PEREIRA, G. F. M. **Evolução da AIDS no B: uma análise espacial**. *Revista do Seminário Internacional de Estatística com R*, v. 4, n. 2, 2019. Citado na página 2.

SOUZA, H. C. C. de et al. **A Bayesian approach for the zero-inflated cure model: an application in a Brazilian invasive cervical cancer database**. *Journal of Applied Statistics*, Taylor & Francis, p. 1–17, 2021. Citado na página 6.

VARSHNEY, M. et al. **Cure fraction model for the estimation of long-term survivors of HIV/AIDS patients under antiretroviral therapy**. *J Commun Disc*, v. 5, n. 3, p. 1–10, 2018. Citado na página 7.

WEIBULL, W. **A statistical theory of strength of materials**. *IVB-Handl.*, 1939. Citado na página 2.

WERLE, J. E. ***HIV/AIDS em Mato Grosso do Sul***: análise de tendência, distribuição espacial e sobrevida dos casos. 75 p. Dissertação (Mestrado em Enfermagem) — Universidade Federal do Mato Grosso do Sul, Campo Grande, 2021. Citado 2 vezes nas páginas 5 and 6.

ZASLAVSKY, R.; GOULART, B. N. G.; ZIEGELMANN, P. K. **Cross-border healthcare and prognosis of HIV infection in the triple border Brazil-Paraguay-Argentina**. *Cadernos de Saúde Pública*, SciELO Public Health, v. 35, p. e00184918, 2019. Citado na página 2.

ISSN: 2763-8855