The Negative Binomial Beta Prime Regression Model with Cure Rate: Application with a Melanoma Dataset

Jeremias Leão1 · Marcelo Bourguignon2 · Helton Saulo3 · Manoel Santos-Neto4 · Vinícius Calsavara5

Accepted: 1 April 2021 / Published online: 25 May 2021
© Grace Scientific Publishing 2021

Abstract
This paper introduces a cure rate survival model by assuming that the time to the event of interest follows a beta prime (BP) distribution and that the number of competing causes of the event of interest follows a negative binomial distribution. The proposed model provides a novel alternative to the existing cure rate regression models due to its flexibility, as the BP model can exhibit greater levels of skewness and kurtosis than these of the gamma and inverse Gaussian distributions. Moreover, the hazard rate function of this model can have an upside-down bathtub or an increasing shape. We approach both parameter estimation and local influence based on likelihood methods. In special, three perturbation schemes are considered for local influence. Numerical evaluation of the proposed model is performed by Monte Carlo simulations. In order to illustrate the potential for practice of our model, we apply it to the real medical dataset from a population-based study of incident cases of melanoma diagnosed in the state of São Paulo, Brazil.

Keywords Beta prime distribution · Likelihood methods · Local influence · Long-term survival model · Melanoma · Negative binomial distribution

Marcelo Bourguignon
m.p.bourguignon@gmail.com
1 Department of Statistics, Universidade Federal do Amazonas, Manaus, AM, Brazil
2 Department of Statistics, Universidade Federal do Rio Grande do Norte, Natal, RN, Brazil
3 Department of Statistics, Universidade de Brasília, Brasília, DF, Brazil
4 Department of Statistics, Universidade Federal de Campina Grande, Campina Grande, PB, Brazil
5 Biostatistics and Bioinformatics Research Center, Cedars-Sinai Medical Center, Los Angeles, California, USA
1 Introduction

In medical and epidemiological studies, often interest focuses on studying the effect of concomitant information on the time-to-event such as death or recurrence of a disease. When the primary interest is to estimate the covariate effect, the Cox proportional hazards model is commonly used in the analysis of survival time data; see [21]. With the development of medical and health sciences and new treatments, it is expected that a proportion of patients responds favorably to a treatment, consequently improving overall survival/disease-free survival. This proportion of patients is called the cure fraction. Incorporating the cure fraction in survival models leads to cure rate models or long-term survival models. These models have been widely developed in the biostatistics literature. Historically, one of the most famous cure rate models is the mixture cure model introduced by Berkson and Gage [6]. This model has been extensively discussed by several authors, including [11, 25, 26, 32, 39]. Later, Yakovlev and Tsodikov [42] and Chen et al. [17] proposed the promotion time cure model or bounded cumulative hazard model in cancer relapse settings, assuming that a latent biological process of propagation of latent carcinogenic tumor cells is generating the observed failure (relapse). Recently, Cooner et al. [19] generalized this framework to a flexible class of cure models under latent activation schemes. Cancho et al. [15] proposed a flexible cure rate model, which encompasses as special cases the mixture model [6], the promotion time cure model [17], and the cure rate proportional odds model proposed by Gu et al. [28]. The statistical literature for modeling lifetime data in the presence of a cure fraction and latent competing causes is vast and growing rapidly. Interested readers can refer to [8, 13, 14, 16, 19, 37, 41–43], among others.

In this context, our main objective is to introduce the negative binomial beta prime (NBBP) cure rate regression model, conceived inside a latent competing causes scenario with cure fraction, where there is no information about which cause was responsible for the individual death or tumor recurrence, but only the minimum lifetime value among all risks is observed and a part of the population is not susceptible to the event of interest. The BP model has properties that its competitor distributions of the exponential family do not have. For example, the HR function of BP distribution can have an upside-down bathtub or increasing shape depending on the parameter values. Most classical two-parameter distributions such as Weibull and gamma distributions have monotone HR functions. The skewness and kurtosis of the BP distribution can be much larger than these of the gamma and inverse Gaussian distributions, which may be more appropriate in certain practical situations.

Herein, we illustrate the applicability of the proposed model in a real medical dataset from a population-based study of incident cases of melanoma diagnosed in the State of São Paulo, Brazil. Incidence of a tumor is not always related to its severity. For instance, carcinomas of the skin are very common worldwide, but clinical outcomes are among the best ones in oncology. Melanoma is not the most common skin malignancy; however, it is one of the most dangerous
(worse survival) due to its potential for metastatic dissemination. According to the Brazilian National Institute of Cancer (INCA), approximately 6000 new cases were expected in 2018 [20]; whereas according to the International Agency for Research on Cancer (IARC), approximately 7000 new cases were reported [24]. The number of deaths in Brazil due to melanoma is estimated to be approximately 2000 cases per year [20, 24]. In the study from which our dataset is drawn, patients diagnosed with melanoma were enrolled from 2000 to 2014, with follow-up conducted until 2018; death due to cancer is the event of interest [12].

The rest of this paper is organized as follows. In Sect. 2, we introduce the proposed long-term survival model and discuss some of its properties as well as some special models. In Sect. 3, the estimation method for the model parameters is discussed. Finally, in Sect. 4, we illustrate the proposed model through simulation and an application to medical real-world data set.

2 The Negative Binomial Beta Prime Regression Model with Cure Rate

In this section, we will formulate the NBBP regression model with cure rate. The ingredients of the proposed model are: (1) the unified long-term survival model formulated by Rodrigues et al. [37]; (2) the BP distribution for modeling the time to the event of interest; and (iii) the NB distribution for modeling the number of competing causes of the event of interest.

Let $N$ denotes the number of competing causes related to the occurrence of an event of interest, for an individual in the population. Conditional on $N$, we assume that the $Z_j$’s are independent and identically distributed random variables representing the promotion times of the competing causes, for $j = 1, \ldots, n$. Moreover, we assume that $N$ is independent of $Z_1, \ldots, Z_n$ and the observable time-to-event is defined as $T = \min\{Z_1, \ldots, Z_N\}$ for $N \geq 1$, and $T = \infty$ if $N = 0$, which leads to a cured fraction denoted by $p_0$; see [37]. Under this setup, the long-term survival function (SF) of the random variable $T$ is given by

$$S_p(t|\cdot) = P(T \geq t) = P(N = 0) + \sum_{n=1}^{\infty} P(Z_1 > t, \ldots, Z_N > t|N = n)P(N = n)$$

$$= \sum_{n=0}^{\infty} P(N = n)[S_T(t|\cdot)]^n = A_N(S_T(t|\cdot)), \quad t > 0,$$

where $S_T(\cdot|\cdot)$ denotes the SF of the unobserved lifetimes and $A_N(\cdot)$ is the probability generating function of the random variable $N$, which converges when $S_T(t|\cdot) \in [0, 1]$. Various results can be obtained for each choice of $A_N(\cdot)$ and $S_T(\cdot|\cdot)$ considered in (1). Here, we assume that the unobserved (latent) random variable (RV) $N$ has an NB distribution with probability mass function expressed as

$$P(N = n) = \frac{\Gamma(n + \alpha^{-1})}{n!\Gamma(\alpha^{-1})} \left(\frac{\alpha \theta}{1 + \alpha \theta}\right)^n (1 + \alpha \theta)^{-1/\alpha},$$

where $\Gamma(\cdot)$ is the gamma function and $\alpha > 0$, $\theta > 0$.
where \( n = 0, 1, \ldots, \theta > 0, \alpha \geq -1, 1 + \alpha \theta > 0 \) and \( \Gamma \) is the standard gamma function, whereas the corresponding mean and variance are \( \text{E}(N) = \theta \) and \( \text{Var}(N) = \theta + \alpha \theta^2 \), respectively. The NB distribution encompasses some well-known models [6, 37]: (A1) if \( \alpha \to 0 \), the Poisson distribution with mean \( \theta \) is obtained; (A2) if \( \alpha = -1/n \) (with \( n \) integer), we have the binomial distribution, that is, \( N \sim \text{Binomial}(n, \theta/n) \), with \( 0 \leq \theta/n \leq 1 \); (A3) if \( \alpha = -1 \) and \( n \) has two possible outcomes, \( n = 0 \) and \( n = 1 \), namely the Bernoulli distribution is reached, which means that we have the standard mixture cure model [6, 7]; and if \( \alpha = 1 \), we get the geometric distribution with parameter \( \theta/(1 + \theta) \) and mean \( \theta \). The NB distribution is widely used in this context because the model parameters have a biological interpretation; see [40], that is, the mean number of competing causes is represented by \( \theta \), whereas \( \alpha \) is a dispersion parameter. Therefore, considering the number of competing causes to be NB distributed and \( S_T \) being a proper SF, we have that the long-term SF obtained from (1) for the RV \( T \) is given by

\[
S_p(t; \theta, \alpha) = (1 + \alpha \theta(1 - S_T(t)))^{-1/\alpha}, \quad t > 0, \tag{3}
\]

where \( \lim_{t \to \infty} S_p(t; \theta, \alpha) = (1 + \alpha \theta)^{-1/\alpha} = p_0 > 0 \), with \( p_0 \), as mentioned, representing the fraction of cured patients in the population. From this result, note that: (i) as the survival time after diagnosis increases, the SF converges to \( p_0 \); and (ii) \( S_p(t) \) is an improper function.

A RV \( X \) follows the BP distribution with shape parameters \( \kappa > 0 \) and \( \gamma > 0 \), if its cumulative distribution function (CDF) and probability density function (PDF) are given by

\[
F_X(x|\kappa, \gamma) = I_{x/(1+x)}(\kappa, \gamma),
\]

and

\[
f_X(x|\kappa, \gamma) = \frac{x^{\kappa-1}(1+x)^{-(\kappa+\gamma)}}{B(\kappa, \gamma)},
\]

where \( I_{x}(\kappa, \gamma) = B_{x}(\kappa, \gamma)/B(\kappa, \gamma) \) is the incomplete beta function ratio, \( B_{x}(\kappa, \gamma) = \int_0^x u^{\kappa-1}(1-u)^{\gamma-1}du \) is the incomplete function, \( B(\kappa, \gamma) = \Gamma(\kappa)\Gamma(\gamma)/\Gamma(\kappa+\gamma) \) is the beta function and \( \Gamma(\kappa) = \int_0^\infty u^{\kappa-1}\exp(-u)du \) is the gamma function. The BP is related to several models. The interested reader in BP model is referred to [9, 29, 33–35]. These works present reviews and generalizations of the BP model. In this context, to introduced the NBBP model, we are considering the parameterization used in Bourguignon et al. [9], where the PDF of the BP distribution is given by

\[
f_{BP}(t|\mu, \phi) = \frac{\Gamma(\phi+1)+(1+t)^{-(\mu+1+\phi+2)}}{B(\mu(\phi+1), \phi+2)}, \tag{4}
\]

where \( \kappa = \mu(\phi + 1) \) and \( \gamma = \phi + 2 \). In this case, the BP distribution is indexed in terms of the mean (\( \mu \)) and precision (\( \phi \)) parameters.

Consider that the number of competing causes \( N \) follows a NB distribution (particular cases are the Poisson, binomial, Bernoulli and geometric distributions) with
parameters $\alpha$ and $\theta$, for $\theta > 0$ and $\alpha \theta > -1$, and that the time to the event of interest is BP distributed with parameters $\mu$ and $\phi$ as in (4). Then, the long-term SF of cured patients is given by

$$S_p(t|\xi) = \left[1 + \alpha \theta F_{BP}(t|\mu, \phi)\right]^{-1/\alpha}, \quad t > 0,$$

where $\xi = (\alpha, \theta, \mu, \phi)^T$. The corresponding PDF and HR function obtained from (5) are, respectively, expressed as

$$f_p(t|\xi) = \frac{\theta \mu^{\alpha(\phi+1)-1}(1 + t)^{-[\mu(\phi+1)+\phi+2]} B(\mu(\phi + 1), \phi + 2)}{[1 + \alpha \theta F_{BP}(t|\mu, \phi)]^{-1/\alpha - 1}},
$$

$$h_p(t|\xi) = \frac{\theta \mu^{\alpha(\phi+1)-1}(1 + t)^{-[\mu(\phi+1)+\phi+2]} B(\mu(\phi + 1), \phi + 2)}{[1 + \alpha \theta F_{BP}(t|\mu, \phi)]^{-1}}, \quad t > 0.$$

The SF for the non-cured population (or NBBP SF), denoted by $S_{NBBP}$, is given by

$$S_{NBBP}(t|\xi) = \frac{1 + \alpha \theta F_{BP}(t|\mu, \phi)^{-1/\alpha} - (1 + \alpha \theta)^{-1/\alpha}}{1 - (1 + \alpha \theta)^{-1/\alpha}}, \quad t > 0.$$  

(6)

From (6), we have $\lim_{t \to 0} S_{NBBP}(t|\xi) = 1$ and $\lim_{t \to +\infty} S_{NBBP}(t|\xi) = 0$, so $S_{NBBP}$ is a proper SF. The PDF of non-cured population (or NBBP PDF), denoted by $f_{NBBP}(t|\xi)$, is given by

$$f_{NBBP}(t|\xi) = \frac{\theta f_{BP}(t|\mu, \phi)[1 + \alpha \theta F_{BP}(t|\mu, \phi)]^{-(1+1/\alpha)}}{1 - (1 + \alpha \theta)^{-1/\alpha}}, \quad t > 0.$$

### 3 Estimation and Diagnostics

We use the maximum likelihood (ML) method to estimate the model parameters. The model (2) parameterized in the proportion of cured given in (8) was introduced by de Castro et al. [22]. Since then this model has been widely used in several cure rate works, for example, Cancho et al. [15], Leão et al. [31], among others. We assume that the time-to-event is not completely observed and is subject to right censoring. We observe $t_i = \min\{y_i, c_i\}$ and $\delta_i = 1(y_i \leq c_i)$, where $c_i$ is the censoring time, and $\delta_i = 1$ if $y_i$ is a time-to-event and $\delta_i = 0$ if it is right-censored, for $i = 1, \ldots, n$. Then, from $n$ pairs of times and censoring indicators, $(y_1, \delta_1, \ldots, y_n, \delta_n)$ say, the corresponding likelihood function, under uninformative censoring, is given by

$$L(\theta|y) = \prod_{i=1}^n f_{pop}(y_i|\theta)\delta_i S_{pop}(y_i|\theta)^{1-\delta_i},$$

(7)

where $\theta = (\alpha, \mu, \phi, \beta)^T$. 

 Springer
In this paper, we also incorporate explanatory variables in the proposed model through parameter $p_0$, which is a more reasonable approach because it can directly reflect the effect of a treatment. For instance, for some treatment $A$, if the treatment effect is good, then some patients will be cured and the estimate for $p_0$ will be $> 0 \in (0, 1)$; if the treatment is not sufficient, the estimate will be close to zero.

In this sense, explanatory variables are incorporated in the model through the cured fraction $\log[p_{0i}/(1 - p_{0i})] = x_i^\top \beta$, $i = 1, \ldots, n$, where $\beta$ is the vector of regression coefficients. From (8)-(9), the likelihood function in (7) is expressed as

$$L(\theta|y) = \prod_{i=1}^{n} \left[ \prod_{j=1}^{q} \left[ \left( \frac{p_{0i}^{\alpha-1}}{\alpha} \right) f_{BP}(y_i|\mu, \phi) \right]^{\delta_j} \times \left[ 1 + \left( p_{0i}^{\alpha} - 1 \right) F_{BP}(y_i|\mu, \phi) \right]^{-\delta_j-1/\alpha} \right], \quad \text{if } \alpha \neq 0;$$

$$\prod_{i=1}^{n} \left[ - \log \left( p_{0i} \right) p_{0i}^{F_{BP}(y_i|\mu, \phi)} f_{BP}(y_i|\mu, \phi) \right]^{\delta_j}, \quad \text{if } \alpha = 0. \quad (9)$$

The ML estimators will be obtained using numerical methods, since equating the first-order log-likelihood derivatives to zero leads us to a complicated system of non-linear equations. It can be easily performed by using standard non-linear maximization procedures found in most of statistical and data analysis packages.

In order to assess the sensitivity of the ML estimators to atypical cases, we perform local influence analysis which is based on the curvature of the log-likelihood function; the reader is referred to [30] for a convexity plot approach that can be used to visualize the curvature of likelihood function. Recall that $\hat{\theta} = (\alpha, \mu, \phi, \beta)^\top$ and let the vector of perturbations $\omega$ be a subset of $\Omega \subset \mathbb{R}^m$, whereas $\omega_0$ is a non-perturbation vector such that $\ell(\theta|\omega_0) = \ell(\hat{\theta}) = \log(L(\hat{\theta}))$, for all $\theta$. Then, the likelihood distance (LD) is given by $LD(\theta) = 2(\ell(\hat{\theta}) - \ell(\hat{\theta}_{\omega}))$, where $\hat{\theta}_{\omega}$ is the ML estimate of $\theta$ under the perturbed model, and the normal curvature for $\hat{\theta}$, at the direction vector $d$ ($||d|| = 1$), is given by $C_d(\theta) = 2|d^\top \nabla \Sigma(\hat{\theta})^{-1} \nabla d|$, where $\nabla$ is a $(q + 3) \times m$ matrix that depends on the perturbation scheme, whose elements are $\nabla_{ji} = \partial^2 \ell(\theta|\omega)/\partial \theta_j \partial \omega_i$, evaluated at $\theta = \hat{\theta}$ and $\omega = \omega_0$, for $j = 1, \ldots, q + 3$ and $i = 1, \ldots, m$; see [18]. Index plot of the eigenvector $d_{\text{max}}$ associated with the maximum eigenvalue of $B(\theta) = -\nabla^\top \Sigma(\hat{\theta})^{-1} \nabla$, $C_{d_{\text{max}}}(\theta)$ say, evaluated at $\theta = \hat{\theta}$ and $\omega = \omega_0$, can indicate cases which have a high influence on $\text{LD}(\theta)$. Moreover, the
vector $d_i = e_i$ can be considered to detect local influence, where $e_i$ denotes an $m \times 1$ vector of zeros with one at the $i$th position. Thus, the corresponding normal curvature takes the form $C_i(\theta) = 2|b_{ii}(\theta)|$, where $b_{ii}(\theta)$ is the $i$th diagonal element of $B(\theta)$, for $i = 1, \ldots, m$, evaluated at $\theta = \hat{\theta}$ and $\omega = \omega_0$. Then, the case $i$ is potentially influential if $C_i(\hat{\theta}) > \frac{2}{m} \sum_{i=1}^{m} C_i(\hat{\theta})$. In this paper, we consider the following perturbation schemes: case-weight, response and explanatory variable; see [18]. We assess the impact of the detected influential cases on the model inference with the relative change (RC), which is obtained by removing influential cases and re-estimating the parameters and their corresponding standard errors (SEs). The RCs are given by

$$\text{RC}_{\theta(i)} = \left| \frac{\hat{\theta} - \hat{\theta}(i)}{\hat{\theta}} \right| \times 100\%$$

and

$$\text{RC}_{\text{SE}(\theta(i))} = \left| \frac{\text{SE}(\hat{\theta}) - \text{SE}(\hat{\theta})(i)}{\text{SE}(\hat{\theta})} \right| \times 100\%,$$

where $\hat{\theta}(i)$ and $\text{SE}(\hat{\theta})$ are the ML estimate of $\theta$ and its corresponding SE, respectively, after removing the case $i$, for $i = 1, \ldots, n$.

Theoretical results of this paper have been implemented in the R software [36] through the creation of a package called NBBPmodel. The NBBPmodel package provides functions for fitting the proposed cure rate negative binomial beta prime regression model. The current version can be downloaded from Github using the command in R:

```
devtools::install_github("santosneto/NBBPmodel")
```

The function nbbp_fit() fits the model and creates an object of class NBBPmodel which can be applied, for example, in the functions diagnostic() and residuals().

4 Numerical Evaluation

4.1 A Simulation Study

We carry out a MC simulation study to evaluate the performance of the ML estimators for the proposed model. The simulation scenario considers the following: sample sizes $n \in \{200, 400, 600, 800, 1000\}$, $\mu = \{0.50, 1\}$, $\sigma = \{1, 10\}$ and 5,000 MC replications. The cured fraction is $p_{0i} = \exp(\beta_0 + \beta_1 x_i) / [1 + \exp(\beta_0 + \beta_1 x_i)]$. For the simulations, we consider a binary covariate $x$ with values drawn from a Bernoulli distribution with parameter 0.5. We consider $\beta_0 = 0.5$ and $\beta_1 = -1$ such that the cure fraction for the two levels of $x$ are $p_{00} = 0.62$ and $p_{01} = 0.38$,
respectively. The censoring times were samples from the uniform distribution, \( U(a, b) \), where \( a, b > 0 \) were set in order to control the proportion of censored observations. In our study, the proportion of censored observations was on the average obtained for each sample size; see Table 1. Note that, based on the probability integral transform, the NBBP model CDF follows a \( U(0, 1) \) distribution. Then, the NBBP SF is \( U(0, 1) \) distributed as well. Random number generation from the NBBP model is performed following Algorithm 1. In step #3 of this algorithm, we use the function \texttt{uniroot} of the \texttt{R} software to get the root of the equation; see [10]. For each value of the parameter, sample size and censoring proportion, we report the empirical values for the bias and mean squared error (MSE) of the ML estimators in Table 1. According to results, as the sample size increases, the ML estimators become more efficient, as expected.

| \( n \) (Censoring %) | \( \hat{\mu} \) | \( \hat{\phi} \) | \( \hat{\alpha} \) | \( \hat{\rho}_0 \) | \( \hat{\rho}_1 \) | \( \hat{\rho}_{00} \) | \( \hat{\rho}_{01} \) |
|------------------------|------------|------------|------------|-------------|-------------|-------------|-------------|
| True values →          | 0.5        | 1          | 2          | 0.5         | -1          | 0.623       | 0.377       |
| 200 (52.98%)           | 0.544      | 1.371      | 2.401      | 0.441       | -1.048      | 0.608       | 0.357       |
|                        | (0.338)    | (1.215)    | (1.511)    | (0.233)     | (0.293)     | (0.895)     | (0.584)     |
| 400 (52.63%)           | 0.538      | 1.132      | 2.148      | 0.478       | -1.032      | 0.617       | 0.360       |
|                        | (0.210)    | (0.729)    | (0.981)    | (0.160)     | (0.206)     | (0.593)     | (0.376)     |
| 600 (51.83%)           | 0.530      | 1.098      | 2.125      | 0.452       | -1.029      | 0.611       | 0.362       |
|                        | (0.155)    | (0.552)    | (0.766)    | (0.126)     | (0.166)     | (0.460)     | (0.291)     |
| 800 (51.72%)           | 0.520      | 1.078      | 2.113      | 0.470       | -1.022      | 0.615       | 0.365       |
|                        | (0.122)    | (0.453)    | (0.663)    | (0.107)     | (0.143)     | (0.402)     | 0.256       |
| 1000 (51.85%)          | 0.519      | 1.059      | 2.057      | 0.471       | -1.008      | 0.615       | 0.369       |
|                        | (0.110)    | (0.400)    | (0.599)    | (0.095)     | (0.127)     | (0.363)     | (0.233)     |

| \( n \) (Censoring %) | \( \hat{\mu} \) | \( \hat{\phi} \) | \( \hat{\alpha} \) | \( \hat{\rho}_0 \) | \( \hat{\rho}_1 \) | \( \hat{\rho}_{00} \) | \( \hat{\rho}_{01} \) |
|------------------------|------------|------------|------------|-------------|-------------|-------------|-------------|
| True values →          | 1          | 10         | 2          | 0.5         | -1          | 0.623       | 0.377       |
| 200 (65.85%)           | 1.101      | 11.555     | 2.521      | 0.449       | -1.101      | 0.610       | 0.353       |
|                        | (0.295)    | (4.133)    | (1.996)    | (0.305)     | (0.364)     | (1.186)     | (0.766)     |
| 400 (64.12%)           | 1.070      | 10.160     | 2.315      | 0.455       | -1.025      | 0.611       | 0.363       |
|                        | (0.184)    | (2.420)    | (1.322)    | (0.200)     | (0.250)     | (0.781)     | (0.599)     |
| 600 (63.65%)           | 1.070      | 10.081     | 2.149      | 0.457       | -1.018      | 0.612       | 0.364       |
|                        | (0.139)    | (1.902)    | (1.062)    | (0.157)     | (0.197)     | (0.629)     | (0.413)     |
| 800 (63.49%)           | 1.048      | 10.07      | 2.145      | 0.472       | -1.014      | 0.616       | 0.364       |
|                        | (0.111)    | (1.605)    | (0.882)    | (0.134)     | (0.171)     | (0.526)     | (0.340)     |
| 1000 (62.98%)          | 1.026      | 10.033     | 2.023      | 0.485       | -1.006      | 0.618       | 0.366       |
|                        | (0.091)    | (1.079)    | (0.756)    | (0.116)     | (0.151)     | (0.456)     | (0.293)     |
4.2 Application to Real Data

The melanoma dataset is from a retrospective survey of 6749 records of patients diagnosed with melanoma in the State of São Paulo, Brazil, between 2000 and 2014, with follow-up conducted until 2018. It was provided by the Fundação Oncocentro de São Paulo (FOSP), which is responsible for coordinating the Hospital Cancer Registry of the State of São Paulo. The FOSP is a public institution connected to the State Health Secretariat, which assists in the preparation and implementation of healthcare policies in the field of oncology, and serves as an instrument so that oncology hospitals can prepare their own protocols and improve their care practices [1].

The time to the death due to cancer was defined as the period between the date of melanoma diagnosis and death. Those patients who did not die due to melanoma in the follow-up period are characterized as right-censored observations. The covariates recorded when the patient is taken on study include: gender ($x_{i1}$, Male, 3334 patients; Female, 3415 patients); age (in years) ($x_{i2}$, in years, mean=58.04 and standard deviation=16.36), where we consider this covariate categorized by $\leq 45$ years and $> 45$ years, with 1549 and 5200 patients, respectively; Surgery ($x_{i3}$, Surgery, 5978 patients; Not Surgery, 771 patients); and cancer clinical staging ($x_{i4}$, Stage I, 3011 patients; Stage II, 1541 patients, Stage III, 1229 patients and Stage IV, 968 patients). A total of 1912 events occurred during follow-up period. The percentage of censored cases is 71.70%. The maximum observation time was approximately 18.54 years, and the median follow-up time was 5.19 years.

Figure 1a–d shows the Kaplan–Meier (KM) estimate of the SF with the melanoma data for each indicated covariate. Notice that, in all cases, there is a strong
evidence of cure fraction in the population. Among the variables considered in our study, the clinical Stage I has the better prognosis. In general, the estimated curves stabilize at a given value, suggesting that the patients censored at the end of the experiment may be immune to the risk in question or were cured during the experiment. This behavior indicates that models ignoring the possibility of a cured fraction will not be suitable for these data. Thus, it is necessary to consider cure rate models in order to analyze such data.

The shape of HR function is an important point to decide whether a particular distribution is suitable or not for a data set. A manner to characterize the shape of an HR is by means of the scaled total time on test (TTT) function. We can detect the type of HR that the data have and then choose a suitable distribution. See in [2] for different theoretical shapes for the scaled TTT curves. The curve can be built using the function `TTT()` of the package `AdequacyModel` [23]. Figure 2 suggests an increasing HR for the observed lifetimes.

Initially, we fitted four models independently, where we considered only the effect of each single variable in the proposed model. Table 2 provided the obtained

Fig. 1 Fitted SFs a Male and female (solid and dashed lines, respectively); b ≤ 45 and > 45 (solid and dashed lines, respectively); c Stage I, II, III and IV (solid, dashed, dotted and dotdash lines, respectively) d Not Surgery and Surgery (solid and dashed lines, respectively); with the melanoma data
results of the fitted models. According to results, the four covariates considered are important to explain the survival rates. As expected, women have a better prognosis than men ($\beta_1 = 0.293 > 0$); Age greater than 45 years leads to a worse prognosis ($\beta_1 = -0.292 < 0$); patients who have underwent surgery have better prognosis ($\beta_1 = 1.333 > 0$); The Stage is a prognostic factor in the overall survival, as the classification of staging increases worse is the prognosis, as can be seen in the coefficient estimates ($\beta_2, \beta_3, \beta_4 < 0$).

Figure 3 shows the fitted SFs based on the KM estimator and the estimated survival function obtained from NBBP model. The plot of the SFs permits us to compare the empirical and fitted SFs of the data. Moreover, the fitted SFs confirm graphically the good fit of the NBBP regression model for each covariate.

We fitted a full model considering all covariates previously mentioned. The results of the fitted NBBP regression models are summarized in Table 3. All findings in this study are consistent to what it is observed in clinical routine. Gender
and age have already been reported as prognostic factors, suggesting that melanoma can be related to a better prognosis in young patients and women [5, 38]. Clinical staging is also used for prognosis stratification, and the curves shown in this paper are very close to those presented in the three last updates of the AJCC staging system for melanoma [3, 4, 27]. In addition, surgery is usually performed on patients with staging I and II, which is associated with a better prognosis. These patients are normally treated with surgery and the great majority will be alive after 10 years of follow-up [27].

Index plots of $C_i$ under case-weight perturbation are shown in Figure 4. We omit the plots corresponding to response and covariate perturbations as they look very similar. Note that the cases #656, #3035 and #6681 are detected as potential influential observations under the considered perturbation schemes. The cases #656, #3035 and #6681 correspond to patients with small lifetime values, which are 0.2, 0.04 and 0.11, respectively. In addition, all of these patients are female in Stage IV that did not undergo surgery (Not Surgery).

Table 2 ML estimate, SE and selection criteria to fitted models

| Covariate                        | Parameter       | NBBBP | MLE  | SE  | $p$ value |
|----------------------------------|-----------------|-------|------|-----|-----------|
|                                  | NBBBP           |       |      |     |           |
| Dispersion/competing causes (NB) | $\alpha$        | 2.443 | 0.277|     |           |
| Scale (beta prime)               | $\mu$           | 1.699 | 0.057|     |           |
| Shape (beta prime)               | $\phi$          | 0.022 | 0.007|     |           |
| Constant                         | $\beta_0$       | -1.562| 0.181|     |           |
| Gender (female)                  | $\beta_1$       | 0.293 | 0.037| <0.001|          |
| Dispersion/competing causes (NB) | $\alpha$        | 2.268 | 0.386|     |           |
| Scale (beta prime)               | $\mu$           | 1.650 | 0.062|     |           |
| Shape (beta prime)               | $\phi$          | 0.042 | 0.003|     |           |
| Constant                         | $\beta_0$       | -0.927| 0.347|     |           |
| Age (> 45)                       | $\beta_1$       | -0.292| 0.059| <0.001|          |
| Dispersion/competing causes (NB) | $\alpha$        | 2.192 | 0.221|     |           |
| Scale (beta prime)               | $\mu$           | 1.904 | 0.065|     |           |
| Shape (beta prime)               | $\phi$          | 0.058 | 0.023|     |           |
| Constant                         | $\beta_0$       | -2.276| 0.205|     |           |
| Surgery                          | $\beta_1$       | 1.333 | 0.091| <0.001|          |
| Dispersion/competing causes (NB) | $\alpha$        | 1.054 | 0.125|     |           |
| Scale (beta prime)               | $\mu$           | 2.118 | 0.076|     |           |
| Shape (beta prime)               | $\phi$          | 0.018 | 0.004|     |           |
| Constant                         | $\beta_0$       | -1.502| 0.821|     |           |
| Stage II                         | $\beta_2$       | -1.332| 0.143| <0.001|          |
| Stage III                        | $\beta_3$       | -2.344| 0.215| <0.001|          |
| Stage IV                         | $\beta_4$       | -4.144| 0.354| <0.001|          |
Table 4 provides the ML estimates of the cured fraction stratified by clinical stages, gender, age and surgery. In addition, we report the corresponding estimated asymptotic SEs and 95% confidence intervals. The delta method was used to estimate the variance of cure fractions. Notice that the long-term probability decreases more rapidly for higher clinical stage. It is worth mentioning that there is a difference in cured fraction of male patients in Stage I (≤ 45, Not Surgery) in relation to Stages II, III and IV (≤ 45, Not Surgery) male patients at 5% significance level. The same behavior is observed for female. Figure 5 displays the quantile versus quantile (QQ) plot of the normalized (randomized) quantile residuals for the NBBP model. This figure indicates that the normalized quantile

![Graphs showing fitted survival functions (SFs) for different variables: Gender, Age, Surgery, Stage.](image)

**Fig. 3** Fitted SFs (gray) a Male and Female (bold and dashed lines, respectively); b ≤ 45 and > 45 (bold and dashed lines, respectively); c Not Surgery and Surgery (bold and dashed lines, respectively); d II, III and IV (solid, dashed, dotted and dotdash lines, respectively) with the melanoma data
residuals in the NBBP model show a good agreement with the expected standard normal distribution.

The RCs in the parameter estimates and their corresponding estimated SEs are shown in Table 5. Also, \( p \) values from the associated \( t \)-test are shown for the regression coefficients. From this table, note that, in general, the largest RCs are related to the removal of some cases mainly in the \( x_{i1} \) which refers to the gender covariate (female).

In Figure 6, we display the SF estimate from NBBP model considering the combinations of covariates. The surviving probability decreases more rapidly for men or women with age > 45 in Stage IV. As expected, women with age \( \leq 45 \) in Stage I (Surgery) have better prognosis. While men with age > 45 in Stage IV (Not Surgery) have the worst prognostic. For more details of the application, see www.santosnetoce.com.br/application_nbbp.

| Covariate                        | Parameter | NBBP   |
|----------------------------------|-----------|--------|
| Dispersion/competing causes (NB) | \( \alpha \) | 3.896  |
| Scale (beta prime)               | \( \mu \)  | 3.892  |
| Shape (beta prime)               | \( \phi \)  | 0.465  |
| Constant                         | \( \beta_0 \) | 0.545  |
| Gender (female)                  | \( \beta_1 \) | 0.220  |
| Age (> 45)                       | \( \beta_2 \) | −0.203 |
| Surgery                          | \( \beta_3 \) | 0.915  |
| Stage II                         | \( \beta_4 \) | −0.996 |
| Stage III                        | \( \beta_5 \) | −1.533 |
| Stage IV                         | \( \beta_6 \) | −2.536 |
Table 4  ML estimates, SEs and 95% confidence interval of the cured fraction stratified by stage, gender, age and surgery for the NBBP model with melanoma data

| Stage | Gender | Age | Surgery | Estimate | SE  | 95% confidence interval |
|-------|--------|-----|---------|----------|-----|-------------------------|
| 1     | Male   | ≤ 45| Not surgery | 0.730   | 0.012 | (0.701, 0.749) |
|       |        |     | Surgery  | 0.881   | 0.052 | (0.779, 0.983) |
|       |        | > 45| Not surgery | 0.676   | 0.032 | (0.613, 0.739) |
|       |        |     | Surgery  | 0.855   | 0.067 | (0.724, 0.987) |
|       | Female | ≤ 45| Not surgery | 0.772   | 0.037 | (0.700, 0.844) |
|       |        |     | Surgery  | 0.904   | 0.029 | (0.848, 0.960) |
|       |        | > 45| Not surgery | 0.729   | 0.027 | (0.676, 0.781) |
|       |        |     | Surgery  | 0.883   | 0.049 | (0.788, 0.979) |
| 2     | Male   | ≤ 45| Not surgery | 0.460   | 0.033 | (0.396, 0.524) |
|       |        |     | Surgery  | 0.705   | 0.055 | (0.598, 0.814) |
|       |        | > 45| Not surgery | 0.406   | 0.049 | (0.310, 0.501) |
|       | Female | ≤ 45| Not surgery | 0.521   | 0.017 | (0.488, 0.555) |
|       |        |     | Surgery  | 0.755   | 0.019 | (0.719, 0.792) |
|       |        | > 45| Not surgery | 0.465   | 0.072 | (0.325, 0.605) |
|       |        |     | Surgery  | 0.710   | 0.037 | (0.637, 0.783) |
| 3     | Male   | ≤ 45| Not surgery | 0.322   | 0.057 | (0.210, 0.435) |
|       |        |     | Surgery  | 0.565   | 0.060 | (0.448, 0.683) |
|       |        | > 45| Not surgery | 0.277   | 0.028 | (0.222, 0.333) |
|       | Female | ≤ 45| Not surgery | 0.376   | 0.073 | (0.233, 0.519) |
|       |        |     | Surgery  | 0.625   | 0.051 | (0.526, 0.725) |
|       |        | > 45| Not surgery | 0.326   | 0.029 | (0.270, 0.382) |
|       |        |     | Surgery  | 0.570   | 0.051 | (0.470, 0.670) |
| Stage | Gender | Age | Surgery | Estimate | SE  | 95% confidence interval |
|-------|--------|-----|---------|----------|-----|------------------------|
| 4     | Male   | ≤45 | Not surgery | 0.143 | 0.030 | (0.083, 0.203) |
|       |        | >45 | Surgery    | 0.302 | 0.085 | (0.135, 0.470) |
|       | Female | ≤45 | Not surgery | 0.119 | 0.045 | (0.031, 0.208) |
|       |        | >45 | Surgery    | 0.259 | 0.056 | (0.149, 0.369) |
|       | Female | ≤45 | Not surgery | 0.173 | 0.018 | (0.138, 0.208) |
|       |        | >45 | Surgery    | 0.354 | 0.087 | (0.183, 0.525) |
|       | Female | ≤45 | Not surgery | 0.145 | 0.035 | (0.076, 0.214) |
|       |        | >45 | Surgery    | 0.306 | 0.068 | (0.174, 0.439) |
Fig. 4 Index plots of $C_i$ for $\alpha$ (left), $\xi = (\mu, \phi)^T$ (center) and $\beta$ (right) with case-weight perturbation for melanoma data

Fig. 5 QQ-plot of the normalized (randomized) quantile residuals for the NBBP model
Table 5  RC (in %) in ML estimates and their corresponding SEs, and respective \( p \) values in brackets

| Dropped case(s) | \( \hat{\mu} \) | \( \hat{\phi} \) | \( \hat{\sigma} \) | \( \hat{\beta}_0 \) | \( \hat{\beta}_1 \) | \( \hat{\beta}_2 \) | \( \hat{\beta}_3 \) | \( \hat{\beta}_4 \) | \( \hat{\beta}_5 \) | \( \hat{\beta}_6 \) |
|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| #656          | RC\(_{\text{ph}}\) | 6.24           | 117.13         | 6.71           | 58.53          | 123.13         | 122.27         | 51.02          | 8.87           | 11.41          | 3.44           |
|               | RC\(_{\text{SB\(_{\text{ph}}\)}}\) | (5.08)         | (9.59)         | (10.83)        | (88.03)        | (71.04)        | (15.80)        | (6.33)         | (3.64)         | (78.49)        |
| \( p \) value | –              | –              | –              | –              | –              | [<0.001]        | [<0.001]        | [<0.001]       | [<0.001]       | [<0.001]       |
| #3035         | RC\(_{\text{ph}}\) | 22.62          | 26.77          | 18.70          | 2.51           | 50.80          | 105.92         | 57.01          | 71.54          | 28.18          | 16.62          |
|               | RC\(_{\text{SB\(_{\text{ph}}\)}}\) | (41.48)        | (56.85)        | (30.58)        | (3.08)         | (9.16)         | (2.41)         | (7.71)         | (13.14)        | (25.52)        | (22.30)        |
| \( p \) value | –              | –              | –              | –              | –              | [<0.001]        | [<0.001]        | [<0.001]       | [<0.001]       | [<0.001]       |
| #6681         | RC\(_{\text{ph}}\) | 5.71           | 22.94          | 7.49           | 94.53          | 86.51          | 53.18          | 35.95          | 38.41          | 3.61           | 5.92           |
|               | RC\(_{\text{SB\(_{\text{ph}}\)}}\) | (11.99)        | (26.69)        | (5.56)         | (4.50)         | (7.72)         | (5.68)         | (13.29)        | (11.43)        | (6.54)         | (8.66)         |
| \( p \) value | –              | –              | –              | –              | –              | [<0.005]        | [<0.001]        | [<0.001]       | [<0.001]       | [<0.001]       |
| #3035,#6681   | RC\(_{\text{ph}}\) | 10.62          | 3.17           | 4.20           | 64.91          | 97.95          | 153.59         | 39.84          | 63.32          | 22.36          | 13.20          |
|               | RC\(_{\text{SB\(_{\text{ph}}\)}}\) | (23.64)        | (15.22)        | (16.75)        | (14.16)        | (8.69)         | (25.59)        | (13.17)        | (17.25)        | (23.70)        | (19.13)        |
| \( p \) value | –              | –              | –              | –              | –              | [<0.091]        | [<0.001]        | [<0.001]       | [<0.001]       | [<0.0001]      | [<0.001]       |
Fig. 6 Survival functions estimated by NBBP model considering gender (Men [M] and Women [W]), age ($\leq 45$ and $> 45$), surgery (surgery [S] or not surgery [NS]) and clinical staging (I, II, III and IV) for melanoma data
Acknowledgements  The research was partially supported by CNPq and CAPES grants from the Brazilian federal government, by FAPEAM grants from the government of the State of Amazonas, Brazil. The authors thank the Fundação Oncocentro de São Paulo for providing the Melanoma dataset. The data that support the findings of this study are openly available in http://www.fosp.saude.sp.gov.br/publicacoes/downloadarquivos.

Declarations

Conflict of Interest  The authors declare no potential conflict of interests.

References

1. Andrade CTD, Magedanz AMPCB, Escobosa DM, Tomaz WM, Santinho CS, Lopes TO, Lombardo V (2012) The importance of a database in the management of healthcare services. Einstein (São Paulo) 10:360–365
2. Azevedo C, Leiva V, Athayde E, Balakrishnan N (2012) Shape and change point analyses of the Birnbaum–Saunders-t hazard rate and associated estimation. Comput Stat Data Anal 56:3887–3897
3. Balch CM, Buzaid AC, Soong SJ, Atkins MB, Cascinelli N, Coit DG, Fleming ID, Gershenwald JE, Houghton A Jr, Kirkwood JM et al (2001) Final version of the American joint committee on cancer staging system for Cutaneous melanoma. J Clin Oncol 19:3635–3648
4. Balch CM, Gershenwald JE, Soong JF, Thompson MB, Atkins DR, Byrd AC, Buzaid AJ, Cochran DG, Ding CS Final version of et al (2009) AJCC melanoma staging and classification. J Clin Oncol 27:6199–6206
5. Balch CM, Thompson JF, Gershenwald JE, Sj Soong, Ding S, McMasters KM, Coit DG, Eggermont AM, Gimotty PA, Johnson TM et al (2014) Age as a predictor of sentinel node metastasis among patients with localized melanoma: an inverse correlation of melanoma mortality and incidence of sentinel node metastasis among young and old patients. Ann Surg Oncol 21:1075–1081
6. Berkson J, Gage RP (1952) Survival curve for cancer patients following treatment. J Am Stat Assoc 47:501–515
7. Boag JW (1949) Maximum likelihood estimates of the proportion of patients cured by cancer therapy. J R Stat Soc B 11:15–53
8. Borges P, Rodrigues J, Louzada F, Balakrishnan N (2016) A cure rate survival model under a hybrid latent activation scheme. Stat Methods Med Res 25:838–856
9. Bourguignon M, Santos-Neto M, de Castro M (2021) A new regression model for positive random variables with skewed and long tail. Metron 79:33–55
10. Brent RP (1973) Algorithms for minimization without derivatives. Prentice-Hall, New Jersey
11. Calsavara VF, Tomazella VLD, Fogo JC (2013) The effect of frailty term in the standard mixture model. Chilean J Stat 4:95–109
12. Calsavara VF, Milani EA, Bertolli E, Tomazella V (2020) Long-term frailty modeling using a non-proportional hazards model: application with a melanoma dataset. Stat Methods Med Res 29(8):2100–2118
13. Calsavara VF, Rodrigues AS, Tomazella VLD, de Castro M (2017) Frailty models power variance function with cure fraction and latent risk factors negative binomial. Commun Stat Theory Methods 46:9763–9776
14. Cancho VG, Louzada F, Barriga GD (2012) The geometric Birnbaum–Saunders regression model with cure rate. J Stat Plann Inference 142:993–1000
15. Cancho VG, Rodrigues J, de Castro M (2011) A flexible model for survival data with a cure rate: a Bayesian approach. J Appl Stat 38:57–70
16. Castro Md, Cancho VG, Rodrigues J (2009) A Bayesian long-term survival model parametrized in the cured fraction. Biometrical J Math Methods Biosci 51:443–455
17. Chen MH, Ibrahim JG, Sinha D (1999) A new Bayesian model for survival data with a surviving fraction. J Am Stat Assoc 94:909–919
18. Cook RD (1987) Influence assessment. J Appl Stat 14:117–131
19. Cooner F, Banerjee S, Carlin BP, Sinha D (2007) Flexible cure rate modeling under latent activation schemes. J Am Stat Assoc 102:560–572
20. Coordenação de Prevenção e Vigilância (2017) Instituto Nacional de Cancer José Alencar Gomes da Silva. Estimativa 2018: Incidência de Cancer no Brasil. Coordenação de Prevenção e Vigilância - Rio de Janeiro. http://www1.inca.gov.br/estimativa/2018/
21. Cox DR (1972) Regression models and life-tables. J R Stat Soc B 34:187–220
22. De Castro M, Cancho VG, Rodrigues J (2010) A hands-on approach for fitting long-term survival models under the gamlss framework. Comput Methods Programs Biomed 97:168–177
23. Diniz Marinho PR, Bourguignon M, Barros Dias CR (2016) AdequacyModel: adequacy of probabilistic models and general purpose optimization. https://CRAN.R-project.org/package=AdequacyModel, R package version 2.0.0
24. Ervik M, Lam F, Ferlay J, Mery L, Soerjomataram I, Bray F et al (2016) Cancer today lyon, france: international agency for research on cancer. 2016, Cancer Today. http://gco.iarc.fr/today. Accessed 01 Feb 2019
25. Ewell M, Ibrahim JG (1997) The large sample distribution of the weighted log rank statistic under general local alternatives. Lifetime Data Anal 3:5–12
26. Farewell VT (1982) The use of mixture models for the analysis of survival data with long-term survivors. Biometrics 38:1041–1046
27. Gershenwald JE, Scolyer RA, Hess KR, Sondak VK, Long GV, Ross MI, Lazar AJ, Faries MB, Kirkwood JM, McArthur GA et al. (2017) Melanoma staging evidence-based changes in the American joint committee on cancer eighth edition cancer staging manual. CA A Cancer J Clin 67: 472–492
28. Gu Y, Sinha D, Banerjee S (2011) Analysis of cure rate survival data under proportional odds model. Lifetime Data Anal 17:123–134
29. Johnson NL, Kotz S, Balakrishnan N (1995) Continuous univariate distributions, vol 1, 2. Wiley, New York
30. Lambert D, Roeder K (1995) Overdispersion diagnostics for generalized linear models. J Am Stat Assoc 90:1225–1236
31. Leão J, Leiva V, Saulo H, Tomazella V (2018) Incorporation of frailties into a cure rate regression model and its diagnostics and application to melanoma data. Stat Med 37:4421–4440
32. Maller R, Zhou X (1996) Survival analysis with long-term survivors. Wiley, Chichester
33. McDonald JB (1984) Some generalized functions for the size distribution of income. Econometrica 52:647–663
34. McDonald JB (1987) Model selection: some generalized distributions. Commun Stat Theory Meth-ods 16:1049–1074
35. McDonald JB, Butler RJ (1990) Regression models for positive random variables. J Econom 43:227–251
36. R Core Team (2021) R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
37. Rodrigues J, Cancho VG, de Castro M, Louzada-Neto F (2009) On the unification of long-term survival models. Stat Prob Lett 79:753–759
38. Sabel MS, Griffith K, Sondak VK, Lowe L, Schwartz JL, Cimmino VM, Chang AE, Rees RS, Bradford CR, Johnson TM (2005) Predictors of nonsentinel lymph node positivity in patients with a positive sentinel node for melanoma. J Am Coll Surg 201:37–47
39. Stangl DK, Greenhouse JB (1998) Assessing placebo response using Bayesian hierarchical survival models. Lifetime Data Anal 4:5–28
40. Tournoud M, Ecochard R (2007) Application of the promotion time cure model with time-changing exposure to the study of hiv/aids and other infectious diseases. Stat Med 26:1008–1021
41. Tsodikov A, Ibrahim J, Yakovlev A (2003) Estimating cure rates from survival data: an alternative to two-component mixture models. J Am Stat Assoc 98:1063–1078
42. Yakovlev AY, Tsodikov AD (1996) Stochastic models of tumor latency and their biostatistical applications. World Scientific, New Jersey
43. Yin G, Ibrahim JG (2005) Cure rate models: a unified approach. Canad J Stat 33:559–570

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.