2015-04

A power series beta Weibull regression model for predicting breast carcinoma

Statistics in Medicine, Malden, Ma : John Wiley and Sons, v. 34, n. 8, p. 1366-1388, Abr. 2015
http://www.producao.usp.br/handle/BDPI/51339

Statistics
in Medicine

Research Article

Received 4 November 2013, Accepted 17 December 2014 Published online 26 January 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6416

A power series beta Weibull regression model for predicting breast carcinoma

Edwin M. M. Ortega,[a*†] Gauss M. Cordeiro,[b] Ana K. Campelo,[c] Michael W. Kattan[d] and Vicente G. Cancho[e]

The postmastectomy survival rates are often based on previous outcomes of large numbers of women who had a disease, but they do not accurately predict what will happen in any particular patient's case. Pathologic explanatory variables such as disease multifocality, tumor size, tumor grade, lymphovascular invasion, and enhanced lymph node staining are prognostically significant to predict these survival rates. We propose a new cure rate survival regression model for predicting breast carcinoma survival in women who underwent mastectomy. We assume that the unknown number of competing causes that can influence the survival time is given by a power series distribution and that the time of the tumor cells left active after the mastectomy for metastasizing follows the beta Weibull distribution. The new compounding regression model includes as special cases several well-known cure rate models discussed in the literature. The model parameters are estimated by maximum likelihood. Further, for different parameter settings, sample sizes, and censoring percentages, some simulations are performed. We derive the appropriate matrices for assessing local influences on the parameter estimates under different perturbation schemes and present some ways to assess local influences. The potentiality of the new regression model to predict accurately breast carcinoma mortality is illustrated by means of real data. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: beta Weibull distribution; breast cancer; cure fraction; likelihood function; long-term survivor; mastectomy; power series distribution

1. Introduction

Cancer is an important public health concern around the world. Statistics on cancer are crucial for estimating its prevalence, incidence, and mortality/survival rates. An overview of descriptive epidemiological data on this disease is a first step to appreciate control measures and preventive interventions in a global context of progressive cancer burden. It is noteworthy that breast cancer is the most common malignancy as well as the primary cause of death among women globally. In particular, the American Cancer Society's estimates for breast cancer in the USA for 2013 are as follows. Approximately 232,340 new invasive cases of breast cancer will be diagnosed in women and 2240 in men. The women diagnosed with breast cancer in its earliest stages actually have a 5-year survival rate of over 98%. One in eight women (12%) are expected to have this diagnosis in her lifetime. It is also likely that there will be a number of 64,640 new cases of carcinoma in situ, which is noninvasive and is the earliest form of breast cancer. Additional information relating to mortality predicts 39,620 deaths of US women compared to 410 US men because of breast cancer. There are more than 2.8 million 'breast cancer survivors' in the USA. It is often taken that 'cure' is related to survival beyond 5 years. Therefore, prevalence is generally related to the number of people alive who have had cancer diagnosed within the last 5 years. This latter statistic is harder to obtain because it depends on medical practice, and it includes people who are still being treated or,

[a]*Department of Exact Sciences, University of São Paulo, Piracicaba, Brazil*
[b]*Department of Statistics, Federal University of Pernambuco, Recife, Brazil*
[c]*Department of Economics, Federal University of Pernambuco, Recife, Brazil*
[d]*Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH, U.S.A.*
[e]*Department of Applied Mathematics and Statistics, University of São Paulo, São Carlos, Brazil*
**Correspondence to: Edwin M. M. Ortega, Department of Exact Sciences, University of São Paulo, Piracicaba, São Paulo, 13418-900, Brazil.*
[†]*E-mail: edwin@usp.br*

at least, being followed up medically for the disease. Surgery is the most common treatment for breast cancer. There are several kinds of surgery. The surgeon usually removes one or more lymph nodes from under the arm to check for cancer cells. If cancer cells are found in the lymph nodes, other cancer treatments will be needed. At any stage of disease, care is available to control pain and other symptoms, to relieve the side effects of treatment, and to ease emotional concerns.

Regression models for survival data with a *surviving fraction* (also known as cure rate models or long-term survival models) play an important role in reliability and survival analysis. These models typically assume that all units under study are susceptible to an event of interest and will eventually experience it if its follow-up is sufficiently long. However, there are situations for which a fraction of individuals is not expected to experience the event of interest, that is, those individuals are cured or insusceptible. For example, researchers may be interested in analyzing the recurrence of a disease. Many individuals may never experience a recurrence; therefore, a *cured fraction* of the population exists. Cure rate models have been applied to estimate the possibility of a cured fraction.

The literature on the subject is by now vast and growing rapidly. The books by Maller and Zhou [1] and Ibrahim *et al.* [2] and the papers by Tsodikov *et al.* [3], Cooner *et al.* [4], Tournoud and Ecochard [5], Zhao *et al.* [6], de Castro *et al.* [7], Ortega *et al.* [8], Rodrigues *et al.* [9], Rodrigues *et al.* [10], and de Castro *et al.* [11] and Perdona and Louzada-Neto [12] represent only some examples. Two formulations of cure rate models have received attention in the literature, namely the mixture cure model [13, 14] and the promotion time cure model [15]. There is a basic distinction between them. While in the mixture cure modeling, the unknown number of causes of the event of interest is assumed to be a binary random variable on $\{0, 1\}$, in the promotion time cure modeling, this number follows a Poisson distribution. In a biological context, the idea behind these assumptions lies within a latent competing cause structure, in the sense that the event of interest can be the death of a patient or a tumor recurrence, which can happen because of unknown competing causes. These latent competing causes can be assigned to metastasis-component tumor cells left active after an initial treatment [11]. A metastasis-component tumor cell is a tumor cell having the potential of metastasizing [15]. If death or tumor recurrence did not occur, one can consider the patient to be cured. In this paper, the unknown number of tumor cells left active after the surgery having potential for metastasizing is modeled by a discrete power series (PS) distribution, whereas the cancer recurrence time for each cell follows the beta Weibull (BW) distribution [16–18]. This distribution is very suitable to model the four most common types of the hazard rate function (HRF). Then, the proposed cure rate model can be seen as a general model encompassing the mixture cure and promotion time cure models. Also, an advantage of our modeling is that the PS distribution is very flexible, including some important special cases, such as the Bernoulli (the mixture model), geometric, logarithmic, negative binomial, and Poisson (the promotion time cure model), which can be tested for choosing the best fitted special model of the proposed model.

The effectiveness of a parametric model depends on the proper choice of the parametric distribution. Traditionally, the exponential model with cure fraction, which takes an exponential distribution for the time to event, has been among the most popular choices, partly because of its simplicity. However, the Weibull cure fraction model, a generalization of the exponential, is more flexible and has been shown to fit the data well in many applications. We define a new wider regression model called the *power series beta Weibull* (PSBW) model with cure fraction to predict survival time in women who had been treated with mastectomy and axillary lymph node dissection. The proposed regression model includes the traditional cure models as special cases. Further, we examine statistical inference aspects using asymptotic likelihood theory and formulate the PSBW model with prognostically significant explanatory variables.

After modeling, it is important to check some assumptions in the model and conduct a robustness study in order to detect influential or extreme observations that can cause distortions in the results of the analysis. Numerous approaches have been proposed in the literature to detect influential or outlying observations. An efficient way to detect influential observations was proposed by Cook [19]. He suggested that more confidence can be put in a model that is relatively stable under small modifications. We develop a similar methodology to detect influential subjects in the PSBW regression model with long-term survivors.

The paper is organized as follows. In Section 2, we define the new model. Special cases are discussed in Section 3. We provide in Section 4 an extended characterization of the new model as a link to Laplace and probability generating function (PGF) forms of the model. Parametric inference is addressed in Section 5 based on standard likelihood techniques. We perform two simulation studies in Section 6 to examine the accuracy of the maximum likelihood estimates (MLEs) and the rejection rates of the likelihood ration (LR) statistics in the Poisson beta Weibull (PBW) model. Diagnostic methods are investigated

in Section 7. An application to women with breast carcinoma [20] is discussed in Section 8. Further, some conclusions are mentioned in Section 9. Finally, some mathematical properties of the noncured population distribution are derived in Appendix A.

2. The power series beta Weibull model

Our model can be derived as follows. Let M be an unobserved positive integer-valued random variable denoting the unknown number of breast tumor cells at the end of the mastectomy that can produce a detectable cancer. We assume that the probability mass function of M is given by

$$P(M = m; \theta) = \frac{a_m \theta^m}{A(\theta)}, m = 0, 1, 2, \ldots, \tag{1}$$

where $a_m > 0$ for $m \geqslant 0$, θ is called the power parameter, and $A(\theta) > 0$. The PGF of M is $P(z) = E(z^M) = A(z\theta)/A(\theta)$. Five important distributions belong to Equation (1): the Bernoulli, Poisson, logarithmic, negative binomial, and geometric distributions. The Bernoulli distribution with success parameter $\theta \in (0, 1)$ is a PS distribution with $A(\theta) = (1 + \theta)$. If $A(\theta) = e^{\theta}$, we have the Poisson distribution with mean parameter θ. The logarithmic distribution with success probability $\theta \in (0, 1)$ corresponds to $A(\theta) = -\log(1 - \theta)$. If we are observing a sequence of independent Bernoulli trials, each trial having a success probability $\theta \in (0, 1)$, the random number of successes until $k > 0$ failures has occurred follows the negative binomial distribution with $A(\theta) = (1 - \theta)^{-k}$. Finally, the geometric is a special case of the negative binomial distribution when $k = 1$.

Let $\mu'_r = E(M^r)$ be the rth ordinary moment of M for $r \geqslant 0$. The moments of M satisfy the recurrence equation $\mu'_{r+1} = \mu'_1 \mu'_r + \theta d\mu'_r/d\theta$ (for $r \geqslant 0$). In particular, $\mu'_1 = \theta A'(\theta)/A(\theta)$ and $Var(M) = \theta^2 \{A''(\theta)/A(\theta) - [A'(\theta)/A(\theta)]^2\}$. The moment generating function and cumulant generating function of M are $M(t) = E(e^{tM}) = A(\theta e^t)/A(\theta)$ and $K(t) = \log[A(\theta e^t)/A(\theta)]$, respectively. The cumulants of M obey a simple recurrence equation $\kappa_{r+1} = \theta\, d\kappa_r/d\theta$ for $r \geqslant 1$.

The time for the jth breast carcinoma cell to metastasize is denoted by Z_j, $j = 1, \ldots, M$. We assume that, conditional on M, the random variables Z_j's are independent identically distributed with cumulative distribution function (CDF) $F(t)$ and survival function $S(t) = 1 - F(t)$. Further, we consider that the Z_j's are independent of M. The observable lifetime of the mastectomized women can be defined by

$$T = \min\{Z_1, \ldots, Z_M\}, \tag{2}$$

under the hypothesis that $P(T = \infty | M = 0) = 1$. The BW distribution is considered for the random variables Z_j's. The assumption of independence and identical distribution to the Z_j's is surely a strong one, but it favors simplicity and analytical tractability at the expense of a more general formulation, as remarked by Yakovlev and Tsodikov [15]. Despite this shortcoming, these models have been proven useful in applications to cancer data analysis.

Under this setup, the survival function of T is given by

$$\begin{aligned}
S_{\text{pop}}(t) = P(T \geqslant t) &= \sum_{m=1}^{\infty} P(T > t | M = m) \\
&= P(M = 0) + \sum_{m=1}^{\infty} P[min\{Z_1, \ldots, Z_m\} > t | M = m]\, P(M = m) \\
&= P(M = 0) + \sum_{m=1}^{\infty} P[Z_1 > t, \ldots, Z_m > t | M = m]\, P(M = m) \\
&= P(M = 0) + \sum_{m=1}^{\infty} P[Z_1 > t]^m\, P(M = m) = \sum_{m=0}^{\infty} S(t)^m\, P(M = m) = G_N(S(t)),
\end{aligned}$$

where $G_N(\dot{})$ is the PGF of the M, latent random variable with distribution given in Equation (1), so that the survival function of T is given by

$$S_{\text{pop}}(t) = \frac{A[\theta S(t)]}{A(\theta)}, \; t > 0. \tag{3}$$

Hereafter, the model (3) is called the *PS cure rate* model. The cure fraction is given by

$$p_0 = \lim_{t\to\infty} S_{\text{pop}}(t) = \frac{A(0)}{A(\theta)} = \frac{a_0}{A(\theta)} > 0.$$

The probability density function (PDF) corresponding to Equation (3) can be expressed as

$$f_{\text{pop}}(t) = -\frac{dS_{\text{pop}}(t)}{dt} = \frac{A'[\theta S(t)]}{A(\theta)}\,\theta f(t), \tag{4}$$

where $A'(\theta S(t)) = A'(z)\,|_{z=\theta S(t)}$ and $f(t) = -dS(t)/dt$ denotes the (proper) density function of the cancer recurrence time Z in Equation (4). Note that $f_{\text{pop}}(t)$ is not a proper PDF, because $S_{\text{pop}}(t)$ is not a proper survival function. Further, the HRF corresponding to Equation (4) is given by

$$h_{\text{pop}}(t) = \frac{A'[\theta S(t)]}{A[\theta S(t)]}\,\theta f(t). \tag{5}$$

In model (5) solely, when M has a Poisson distribution with parameter θ, we can verify that $h_{\text{pop}}(t)$ is multiplicative in θ and $f(t)$ and thus has the proportional hazards structure when the covariates are modeled trough θ.

The proper survival function for the noncured population (i.e., the individuals at risk at time t), say $S_{\text{nc}}(t)$, is given by $S_{\text{nc}}(t) = P(T > t \mid N \geqslant 1)$. Then, it can be expressed as

$$S_{\text{nc}}(t) = \frac{A(\theta S(t)) - a_0}{A(\theta) - a_0},\ t > 0. \tag{6}$$

We note that $S_{\text{nc}}(0) = 1$ and $S_{\text{nc}}(\infty) = 0$ so that $S_{\text{nc}}(t)$ is a proper survival function.

There is a mathematical relationship between the model (3) and the mixture cure rate model [13, 14]. We can write

$$S_{pop}(t) = \frac{a_0}{A(\theta)} + \left(1 - \frac{a_0}{A(\theta)}\right) S_{\text{nc}}(t),$$

where $S_{\text{nc}}(t)$ is given by Equation (6). Thus, $S_{\text{pop}}(t)$ is a mixture cure rate model with cure fraction equal to $p_0 = A(\theta(1-p))/A(\theta)$ and survival function $S_{\text{nc}}(t)$ for the noncured population. These results imply that every mixture cure rate model corresponds to some model of the form (3) for any θ and any survival function $S(\cdot)$.

By differentiating Equation (6), the proper density function for the noncured population reduces to

$$f_{\text{nc}}(t) = \frac{A'(\theta S(t))}{A(\theta) - a_0}\,\theta f(t) = J(\theta) f_{\text{pop}}(t),\ t > 0, \tag{7}$$

where $J(\theta) = A(\theta)/[A(\theta) - a_0]$. We derive expansions for the new proper density function which do not depend on complicated functions. In Appendix A, we obtain explicit expressions for the moments.

Standard lifetime distributions usually present very strong restrictions to produce bathtub curves and thus appear to be unappropriate for interpreting data with this characteristic. In the last years, new extended classes of the Weibull distribution for modeling bathtub HRFs were developed. Mudholkar *et al.* [21] pioneered the exponentiated Weibull, Lai *et al.* [22] introduced the modified Weibull, Famoye *et al.* [16] and Lee *et al.* [17] presented the BW, and Silva *et al.* [23] defined the beta modified Weibull distributions. Further, Cordeiro *et al.* [18] investigated several mathematical properties of the BW distribution, which is a highly flexible lifetime distribution because it takes lower and higher skewness and kurtosis values. This distribution can accommodate the most important types of the HRF (i.e., increasing, decreasing, unimodal, and bathtub) depending on its parameter values, and thus, it becomes an important model to be fitted to a wide variety of lifetime data.

In order to provide more flexibility to model (3), we consider that the time Z to the recurrence of the breast cancer has the BW density function given by

$$f(z;\boldsymbol{\gamma}) = \frac{c}{\tau^c\, B(a,b)}\, z^{c-1}\, \exp\left[-b\left(\frac{z}{\tau}\right)^c\right]\left\{1-\exp\left[-\left(\frac{z}{\tau}\right)^c\right]\right\}^{a-1}, \tag{8}$$

where $\boldsymbol{\gamma}^\top = (a,b,c,\tau)^\top$ is the parameter vector; $a>0$, $b>0$, and $c>0$ are shape parameters; $\tau>0$ is a scale parameter; and $B(a,b)$ denotes the beta function. The CDF corresponding to Equation (8) is given by

$$F(z;\boldsymbol{\gamma}) = I_{1-\exp\left[-\left(\frac{z}{\tau}\right)^c\right]}(a,b), \tag{9}$$

where $I_y(a,b) = B_y(a,b)/B(a,b)$ denotes the incomplete beta function ratio and $B_y(a,b)$ is the incomplete beta function.

The main motivation for using the BW model (8) is that it extends some important distributions previously considered in the literature. In particular, it contains as special cases the exponentiated Weibull [21] for $b=1$, beta exponential [24] for $c=1$, exponentiated exponential [25] for $b=c=1$, generalized Rayleigh [26] for $b=1$ and $c=2$, Weibull for $a=b=1$, among others.

It is important to emphasize that most BW mathematical properties are manageable using modern computer programs with numerical and analytic capabilities. Therefore, they may turn into adequate tools comprising the arsenal of applied statisticians [18]. For example, the rth generalized moment of Z for a real noninteger $a>0$ is given by

$$E(Z^r) = \frac{\tau^r\,\Gamma(r/c+1)}{B(a,b)}\sum_{j=0}^{\infty}\frac{(-1)^j}{(b+j)^{r/c+1}}\binom{a-1}{j}. \tag{10}$$

If $a>0$ is an integer, the index j in the sum stops at $a-1$. For $a=b=1$, Equation (10) gives precisely the rth Weibull moment.

The moment generating function of Z can be expressed in closed form for special cases using the Meijer-G and Wright generalized hypergeometric functions [18]. Extreme values, mean deviations, Bonferroni and Lorenz curves, moments of order statistics and L moments, reliability, and Rényi and Shannon entropies for the BW distribution are also derived by Cordeiro *et al.* [18].

Inserting the BW survival function in model (3), we obtain the PSBW model with long-term survivors given by

$$S_{\text{pop}}(t;\theta,\boldsymbol{\gamma}) = \frac{A\left\{\theta\left[1-I_{1-\exp\left[-\left(\frac{t}{\tau}\right)^c\right]}(a,b)\right]\right\}}{A(\theta)},\ t>0. \tag{11}$$

The PDF corresponding to Equation (11) reduces to

$$\begin{aligned} f_{\text{pop}}(t;\theta,\boldsymbol{\gamma}) = {} & \frac{c\,\theta\, t^{c-1}\, A'\left\{\theta\left[1-I_{1-\exp\left[-\left(\frac{t}{\tau}\right)^c\right]}(a,b)\right]\right\}}{\tau^c\, B(a,b)\, A(\theta)} \\ & \times \exp\left[-b\left(\frac{t}{\tau}\right)^c\right]\left\{1-\exp\left[-\left(\frac{t}{\tau}\right)^c\right]\right\}^{a-1}. \end{aligned} \tag{12}$$

Equation (12) refers to the *PSBW model with long-term survivors* in competing-risk structure. It is very flexible to accommodate increasing, decreasing, bathtub, and unimodal hazard rates. The PSBW model is also suitable for testing goodness of fit of some special models.

Several methods of introducing one or more parameters to generate new distributions that provide adequate fits to real data have been investigated recently in the statistical literature. Among these methods, the compounding of some discrete and important lifetime distributions has been in the vanguard of lifetime modeling. We can propose some new distributions for long-term survivors from Equations (3) and (4).

3. Special power series beta Weibull models

Here, we present some specific compounding models that arise from our general formulation. Particularly, we consider situations for which M follows the binomial, Poisson, logarithmic, negative binomial, and geometric distributions.

- *The Bernoulli beta Weibull (BeBW) model*
  If $A(\theta) = 1+\theta$, then M has the Bernoulli distribution with success parameter $\theta/(1+\theta)$. The Bernoulli exponentiated Weibull model is a submodel when $b = 1$. For $c = 1$, we obtain the Bernoulli beta exponential model. If $b = 1$, in addition to $c = 1$, it yields the Bernoulli exponentiated exponential model. For $a = b = 1$ and $a = b = c = 1$, it gives the Bernoulli Weibull and Bernoulli exponential models, respectively. The cure rate is $p_0 = (1+\theta)^{-1}$ for $\theta > 0$. The BeBW density function is given by

$$f_{\text{pop}}(t;\theta,\boldsymbol{\gamma}) = \frac{\theta}{(1+\theta)}\frac{c\,t^{c-1}}{\tau^c\,B(a,b)}\exp\left[-b\left(\frac{t}{\tau}\right)^c\right]\left\{1-\exp\left[-\left(\frac{t}{\tau}\right)^c\right]\right\}^{a-1}. \tag{13}$$

- *The PBW model*
  If $A(\theta) = \mathrm{e}^{\theta}$, then M has the Poisson distribution with mean θ, and the PBW survival function becomes

$$S_{\text{pop}}(t;\theta,\boldsymbol{\gamma}) = \exp\left\{-\theta\, I_{1-\exp\left[-\left(\frac{t}{\tau}\right)^c\right]}(a,b)\right\}. \tag{14}$$

  The Poisson exponentiated Weibull model is also a submodel of Equation (14) when $b = 1$. For $c = 1$, we obtain the Poisson beta exponential model. If $b = 1$, in addition to $c = 1$, it reduces to the Poisson exponentiated exponential model. The Poisson Weibull (PW) model pioneered by Chen *et al.* [27] arises when $a = b = 1$. For $a = b = c = 1$, we obtain the Poisson exponential model. The cure fraction is given by $p_0 = \mathrm{e}^{-\theta}$. The PBW density function reduces to

$$\begin{aligned} f_{\text{pop}}(t;\theta,\boldsymbol{\gamma}) = {} & \frac{\theta\,c}{\tau^c\,B(a,b)}\,t^{c-1}\,\exp\left[-b\left(\frac{t}{\tau}\right)^c\right]\left\{1-\exp\left[-\left(\frac{t}{\tau}\right)^c\right]\right\}^{a-1} \\ & \times\exp\left\{-\theta\left[1-I_{1-\exp\left[-\left(\frac{t}{\tau}\right)^c\right]}(a,b)\right]\right\}. \end{aligned} \tag{15}$$

  After some algebraic developments, the proper density function of the population at risk for the PBW model (15) can be expressed as

$$\begin{aligned} f_{\text{nc}}(t;\theta,\boldsymbol{\gamma}) = {} & \frac{c\,\theta\exp\left\{1-\theta\left[1-I_{1-\exp\left[-\left(\frac{z}{\tau}\right)^c\right]}(a,b)\right]\right\}\,t^{c-1}}{\tau^c\,B(a,b)[1-\exp(-\theta)]} \\ & \times\exp\left[-b\left(\frac{t}{\tau}\right)^c\right]\left\{1-\exp\left[-\left(\frac{t}{\tau}\right)^c\right]\right\}^{a-1}, \end{aligned} \tag{16}$$

  where $\boldsymbol{\gamma}^{\top} = (a,b,c,\tau)^{\top}$. Equation (16) can be widely used for modeling survival data. Plots of the population density functions (13) and (15) for some parameter values are displayed in Figure 1(a) and (b), respectively.
- *The logarithmic beta Weibull (LBW) model*
  If M has the logarithmic distribution, $A(\theta) = -\log(1-\theta)$, then we obtain the LBW survival function. The logarithmic exponentiated Weibull model is a special model when $b = 1$. For $c = 1$, we obtain the logarithmic beta exponential model. If $b = 1$, in addition to $c = 1$, it reduces to the logarithmic exponentiated exponential model. For $a = b = 1$ and $a = b = c = 1$, we obtain the logarithmic Weibull and logarithmic exponential distributions, respectively.

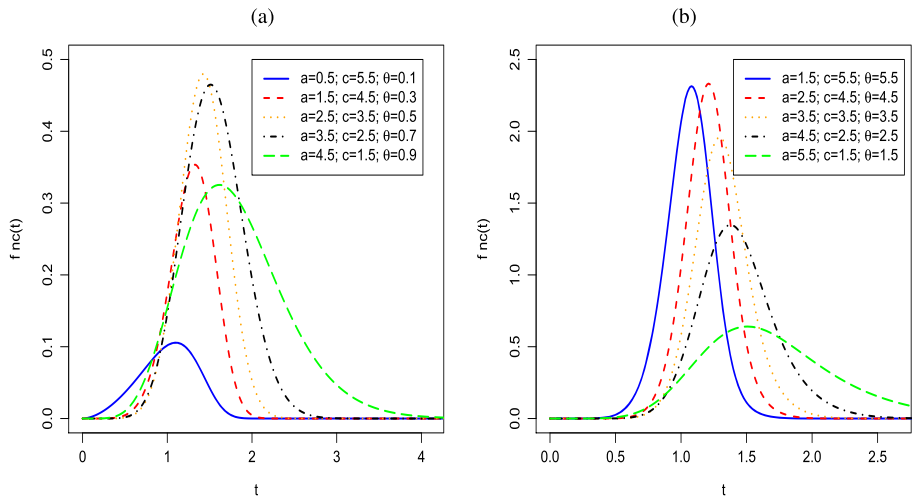

Figure 1. (a) Plots of the Bernoulli beta Weibull population density function for some values of a, c, and θ with $b = 2.0$ and $\tau = 1.5$. (b) Plots of the Poisson beta Weibull population density function for some values of a, c, and θ with $b = 2.0$ and $\tau = 1.5$.

The cure fraction is $p_0 = 1/\log(1-\theta)$, and the LBW density function reduces to

$$f_{\text{pop}}(t;\theta,\boldsymbol{\gamma}) = \frac{-\theta\, c\, t^{c-1} \exp\left[-b\left(\frac{t}{\tau}\right)^c\right]\left\{1-\exp\left[-\left(\frac{t}{\tau}\right)^c\right]\right\}^{a-1}}{\tau^c\, B(a,b)\log(1-\theta)\left\{1-\theta\left[1-I_{1-\exp\left[-\left(\frac{t}{\tau}\right)^c\right]}(a,b)\right]\right\}}. \tag{17}$$

- *The negative binomial beta Weibull (NBiBW) model*
  For the negative binomial distribution, $A(\theta) = (1-\theta)^{-k}$, and then, we obtain the NBiBW model. The negative binomial exponentiated Weibull model is a special model when $b = 1$. For $c = 1$, we obtain the negative binomial beta exponential model. If $b = 1$, in addition to $c = 1$, it gives the negative binomial exponentiated exponential distribution. For $a = b = 1$ and $a = b = c = 1$, we obtain the negative binomial Weibull and negative binomial exponential distributions, respectively.

  The cure fraction is given by $p_0 = (1-\theta)^k$ for $0 \leqslant \theta < 1$. The NBiBW density function can be expressed as

$$f_{\text{pop}}(t;k,\theta,\boldsymbol{\gamma}) = \frac{k\theta(1-\theta)^k c\, t^{c-1} \exp\left[-b\left(\frac{t}{\tau}\right)^c\right]\left\{1-\exp\left[-\left(\frac{t}{\tau}\right)^c\right]\right\}^{a-1}}{\tau^c\, B(a,b)\left\{1-\theta\, I_{1-\exp\left[-\left(\frac{t}{\tau}\right)^c\right]}(a,b)\right\}^{k+1}}. \tag{18}$$

  Plots of the population density functions (17) and (18) for some parameter values are displayed in Figure 2(a) and (b), respectively. These plots indicate great flexibility of these distributions.
- *The geometric beta Weibull (GBW) model*
  For the geometric distribution, $A(\theta) = (1-\theta)^{-1}$, which leads to the GBW model. The geometric exponentiated Weibull model is also a special case when $b = 1$. For $c = 1$, we obtain the geometric beta exponential model. If $b = 1$, in addition to $c = 1$, it reduces to the geometric exponentiated exponential distribution. For $a = b = 1$ and $a = b = c = 1$, we have the geometric Weibull and geometric exponential distributions, respectively. The cure fraction is given by $p_0 = 1-\theta$. The GBW density function is given by

$$f_{\text{pop}}(t;\theta,\boldsymbol{\gamma}) = \frac{\theta(1-\theta)c\, t^{c-1} \exp\left[-b\left(\frac{t}{\tau}\right)^c\right]\left\{1-\exp\left[-\left(\frac{t}{\tau}\right)^c\right]\right\}^{a-1}}{\tau^c\, B(a,b)\left\{1-\theta\, I_{1-\exp\left[-\left(\frac{t}{\tau}\right)^c\right]}(a,b)\right\}^{2}}.$$

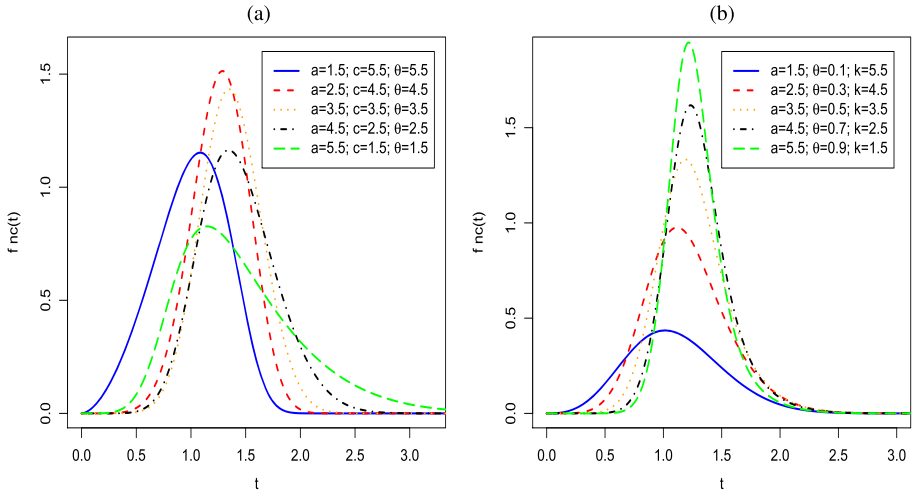

Figure 2. (a) Plots of the logarithmic beta Weibull population density function for some values of a, c, and θ with $b = 2.0$ and $\tau = 1.5$. (b) Plots of the negative binomial beta Weibull population density function for some values of a, θ, and k with $b = 2.0$, $c = 2.5$, and $\tau = 1.5$.

The problems related to identifiability in survival models with cure fraction have been discussed by several authors, among them, [28–30]. Based on these authors and following the proof of Theorem 6.2 in [30], we can conclude that the PSBW model with explanatory variables associated to the parameter θ is identifiable as well as its special cases.

4. Relationship between the long-term survivor models and frailty models

There is a relationship between the long-term survivor models and frailty models, which is now addressed. The frailty model assumes a proportional hazards structure conditional on the random effect, W. Let W denote a nonnegative continuous frailty variable, that is, a random variable indicating the individual level of risk. Then, the frailty model is basically specified by the following HRF (given the frailty):

$$h(t|W) = W\, h_B(t), \tag{19}$$

where $h_B(t)$ is the baseline HRF that can been equal to $h_B(t)\exp(\mathbf{x}^T\boldsymbol{\beta})$ in case of the proportional hazard model. The corresponding survival function, conditional on Z, is given by

$$S(t|W) = P(T > t|W) = \exp\left\{-WH_B(t)\right\} = S_B(t)^W, \tag{20}$$

where $H_B(t) = \int_0^t h_B(u)du$ is the baseline cumulative hazard function and $S_B(t)$ is the corresponding survival function. Many published studies on these models often assume that W is a nonnegative continuous random variable. Thus, the unconditional survival function can be expressed as

$$S(t) = \int S(t|W)\pi(w)dw = \int \exp\left\{-WH_B(t)\right\}\, \pi(w)dz = \mathbf{L}(H_B(t)),$$

where $\pi(w)$ is the density function of W and $\mathbf{L}(s)$ is the Laplace transform of W. The most common frailty distribution are the gamma and other positive stable distributions [31].

In some cases, it may be appropriate to consider discrete frailty distributions. Survival data containing experimental units in the event of interest have not happened even after a long period of observation. In this situation, these units have zero frailty, and survival models induced by frailty with continuous distribution would not be appropriate (see [30, 31]).

Consider that the frailty variable W is a discrete random variable taking values in $\{0,1,\ldots\}$ and having probability mass function $P(W = w)$. The unconditional survival function $S(t)$ can be obtained by

summing in Equation (20) over the support of the distribution of Z, and then, it is given by

$$S(t) = \sum_{w=0}^{\infty} S_B(t)^w \, P[W = w] = E[S_B(t)^W] = G_W(S_B(t)), \tag{21}$$

where $G_W(s)$ is the PGF of W. The model (21) is the same survival model with cure fraction obtained in the competing causes scenarios by Rodrigues *et al.* [9] and Tsodikov *et al.* [3]. Also, because $\lim_{t\to\infty} S_B(t) = 0$ and $\lim_{t\to\infty} S(t) = G_W(0) = P(W = 0) > 0$, Equation (21) is not a proper survival function. This characterize the survival models with a cure fraction, where $p_0 = P(W = 0)$ denotes the proportion of cured individuals. If W is a random variable having the PS distribution and $h_B(t)$ is the HRF of the BW distribution, we obtain the PSBW model given in Equation (12).

5. Inference methods

We consider the situation where the survival time of the mastectomized women is not completely observed and is subjected to right censoring. Let C_i denote the censoring time. We then observe $t_i = \min\{T_i, C_i\}$ and $\delta_i = \mathrm{I}(T_i \leqslant C_i)$, where $\delta_i = 1$ if T_i is the observed time to the event defined before and $\delta_i = 0$ if it is right censored, for $i = 1, \ldots, n$. Let $\boldsymbol{\gamma}$ denote the parameter vector of the distribution function of the time-to-event $F(t)$. From n pairs of times and censoring indicators $(t_1, \delta_1), \ldots, (t_n, \delta_n)$, the observed likelihood function under noninformative censoring can be expressed as [11]

$$L(\theta, \boldsymbol{\gamma}) = \prod_{i=1}^{n} [f_{\text{pop}}(t_i; \theta, \boldsymbol{\gamma})]^{\delta_i} \, [S_{\text{pop}}(t_i; \theta, \boldsymbol{\gamma})]^{1-\delta_i}, \tag{22}$$

where $S_{\text{pop}}(t_i; \theta, \boldsymbol{\gamma})$ and $f_{\text{pop}}(t_i; \theta, \boldsymbol{\gamma})$ are given in Equations (11) and (12), respectively.

In many medical problems, the lifetimes are affected by explanatory variables such as the cholesterol level, blood pressure, and weight. Parametric models to estimate univariate survival functions for censored data regression problems are widely used. Now, we link the parameter θ_i in Equation (11) to the explanatory variable vector $\mathbf{x}_i$ by defining $\theta_i = g(\mathbf{x}_i, \boldsymbol{\beta})$, where $g(\cdot)$ is a suitable link function, for $i = 1, \ldots, n$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ denotes the vector of regression coefficients. Let $\boldsymbol{\psi} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$ be the model parameters.

Substituting Equations (11) and (12) into Equation (22) yields the log-likelihood function

$$\begin{aligned} l(\boldsymbol{\psi}) = {} & r \log\left[\frac{c}{\tau^c \, B(a,b)}\right] + \sum_{i=1}^{n} \delta_i \, g(\mathbf{x}_i, \boldsymbol{\beta}) + (c-1) \sum_{i=1}^{n} \delta_i \, \log(t_i) - b \sum_{i=1}^{n} \delta_i \left(\frac{t_i}{\tau}\right)^c \\ & + (a-1) \sum_{i=1}^{n} \delta_i \, \log\left\{1 - \exp\left[-\left(\frac{t_i}{\tau}\right)^c\right]\right\} \\ & + \sum_{i=1}^{n} \delta_i \log\left\{\frac{A'\left\{g(\mathbf{x}_i, \boldsymbol{\beta})\,[1 - I_{1-\exp\left[-\left(\frac{t_i}{\tau}\right)^c\right]}(a,b)]\right\}}{A(g(\mathbf{x}_i, \boldsymbol{\beta}))}\right\} \\ & + \sum_{i=1}^{n} (1-\delta_i) \, \log\left\{\frac{A\left\{g(\mathbf{x}_i, \boldsymbol{\beta})[1 - I_{1-\exp\left[-\left(\frac{t_i}{\tau}\right)^c\right]}(a,b)]\right\}}{A(g(\mathbf{x}_i, \boldsymbol{\beta}))}\right\}, \end{aligned} \tag{23}$$

where r is the observed number of failures and $A(\cdot)$ and $A'(\cdot)$ are given in Section 2.

The MLE of the parameter vector $\boldsymbol{\psi}$ can be calculated by maximizing the log-likelihood function (23). We use the procedure NLMixed in SAS to compute the MLE $\hat{\boldsymbol{\psi}}$. Under suitable regularity conditions, it can be shown that the asymptotic distribution of the MLE $\hat{\boldsymbol{\psi}}$ is multivariate normal with mean vector $\boldsymbol{\psi}$

and covariance matrix $-\ddot{\mathbf{L}}^{-1}(\boldsymbol{\psi})$, that is,

$$\left(\widehat{\boldsymbol{\beta}}^{\top}, \widehat{\boldsymbol{\gamma}}^{\top}\right)^{\top} \sim \mathrm{N}_{(p+4)}\left\{(\boldsymbol{\beta}^{\top}, \boldsymbol{\gamma}^{\top})^{\top}, -\ddot{\mathbf{L}}^{-1}(\boldsymbol{\psi})\right\},$$

where $-\ddot{\mathbf{L}}(\boldsymbol{\psi}) = \left\{\frac{\partial^2 l(\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \boldsymbol{\psi}^{\top}}\right\}$ is the $(p+4)\times(p+4)$ observed information matrix

$$\ddot{\mathbf{L}}(\boldsymbol{\psi}) = \begin{pmatrix} \mathbf{L}_{\beta_j \beta_{j'}} & \mathbf{L}_{\beta_j \gamma_k} \\ \cdot & \mathbf{L}_{\gamma_k \gamma_{k'}} \end{pmatrix},$$

where $j, j' = 1, \ldots, p$, $k, k' = 1, \ldots, 4$. The submatrices of $\ddot{\mathbf{L}}(\boldsymbol{\psi})$ can be computed numerically. We investigate the asymptotic distribution of the MLEs using a simulation study; see Section 6.

Besides estimation, hypothesis tests can be taken into account. Let $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ be proper disjoint subsets of $\boldsymbol{\psi}$. We aim to test $H_0 : \boldsymbol{\psi}_1 = \boldsymbol{\psi}_1^{(0)}$ against $H_1 : \boldsymbol{\psi}_1 \neq \boldsymbol{\psi}_1^{(0)}$, $\boldsymbol{\psi}_2$ unspecified. Let $\hat{\boldsymbol{\psi}}_0$ maximize $L(\boldsymbol{\psi})$ constrained to H_0 and define the LR statistic by $w = 2[l(\hat{\boldsymbol{\psi}}) - l(\hat{\boldsymbol{\psi}}_0)]$, where $l(\cdot)$ is the log-likelihood. Under the null hypothesis H_0 and some regularity conditions, the statistic w converges in distribution to a chi-square distribution with $\dim(\boldsymbol{\psi}_1)$ degrees of freedom.

6. Simulating the Poisson beta Weibull model

Here, we evaluate the performance of the MLEs of the parameters of the PBW regression model with cure rate given by Equation (14) by means of a simulation study. For each individual $i = 1, \ldots, n$, an unknown number of breast tumor cells at the end of the mastectomy that can produce a detectable cancer, say M_i, is generated from a Poisson distribution with parameter $\theta_i = \exp(\beta_1 + \beta_1 x_i)$, where the covariate x_i is generated from a Bernoulli with probability 0.5. We take $\beta_0 = -0.5$ and $\beta_1 = 0.7$ so that the cure fractions for the two levels of the covariate are $p_0^{(0)}$ and $p_1^{(0)}$, respectively. We generate the event times Z_{ij}, for $j = 1, \ldots, M_i$, from a BW distribution with parameters $a = 2$, $b = 2$, $c = 2$, and $\tau = 10$. The censoring times are sampled from the uniform $(0, \rho)$, where ρ is set in order to control the proportion of censored observations on average to be approximately 60%.

Table I. Maximum likelihood estimates in average (MLEA), standard deviations (SDs), root of mean squared errors (RMSE), and empirical coverage probabilities (CPs) of the parameters of the Poisson beta Weibull regression model.

n		a	b	c	τ	$\hat{\beta}_0$	$\hat{\beta}_1$	$p_0^{(0)}$	$p_{(1)}^{0}$
100	MLEA	1.688	1.524	2.921	9.479	−0.495	0.705	0.539	0.293
	SD	1.171	0.898	1.207	1.802	0.263	0.326	0.085	0.074
	BIAS	−0.312	−0476	0.921	−0.521	0.005	0.005	−0.006	−0.002
	RMSE	1.211	1.016	1.518	1.875	0.263	0.326	0.085	0.0074
	CP	0.926	0.917	0.928	0.969	0.948	0.960	0.939	0.945
200	MLEA	1.852	1.633	2.599	9.424	−0.499	0.705	0.543	0.294
	SD	1.163	0.888	1.004	1.639	0.178	0.225	0.058	0.053
	BIAS	−0.148	−0.367	0.599	−0.576	0.005	0.001	−0.002	−0.001
	RMSE	1.172	0.961	1.168	1.737	0.178	0.225	0.058	0.053
	CP	0.956	0.958	0.962	0.946	0.937	0.958	0.949	0.955
400	MLEA	2.063	1.844	2.242	9.546	−0.505	0.707	0.546	0.294
	SD	1.025	0.780	0.679	1.329	0.126	0.155	0.041	0.037
	BIAS	0.063	−0.156	0.242	−0.454	−0.005	0.007	0.001	0.000
	RMSE	1.026	0.795	0.721	1.404	0.126	0.155	0.041	0.037
	CP	0.954	0.948	0.952	0.946	0.937	0.959	0.949	0.952
800	MLEA	2.111	1.954	2.106	9.722	−0.507	0.708	0.547	0.295
	SD	0.873	0.645	0.475	1.038	0.089	0.113	0.029	0.027
	BIAS	0.111	−0.046	0.106	−0.278	−0.007	0.008	0.002	0.000
	RMSE	0.879	0.647	0.486	1.074	0.089	0.114	0.029	0.027
	CP	0.951	0.949	0.953	0.946	0.945	0.951	0.952	0.953

We take sample sizes of $n = 100, 200, 400$, and 800. For each sample size, we perform 1000 simulations and calculate the average of the MLEs, root of mean squared errors of the MLEs, and empirical coverage probabilities of 95% confidence intervals for the parameters in model (14). The simulation results are given in Table I. We note that the averages of the MLEs are close to the true parameter values, the root of mean squared errors decrease as the sample size increases, and the empirical coverage probabilities are closer to the nominal levels when the sample size increases. Further, we examine the distribution of the MLEs of a, b, c, β_0, β_1, $p_0^{(0)}$, and $p_1^{(0)}$ and provide the plots of the empirical distributions of $\hat{\beta}_0$, $\hat{\beta}_1$,

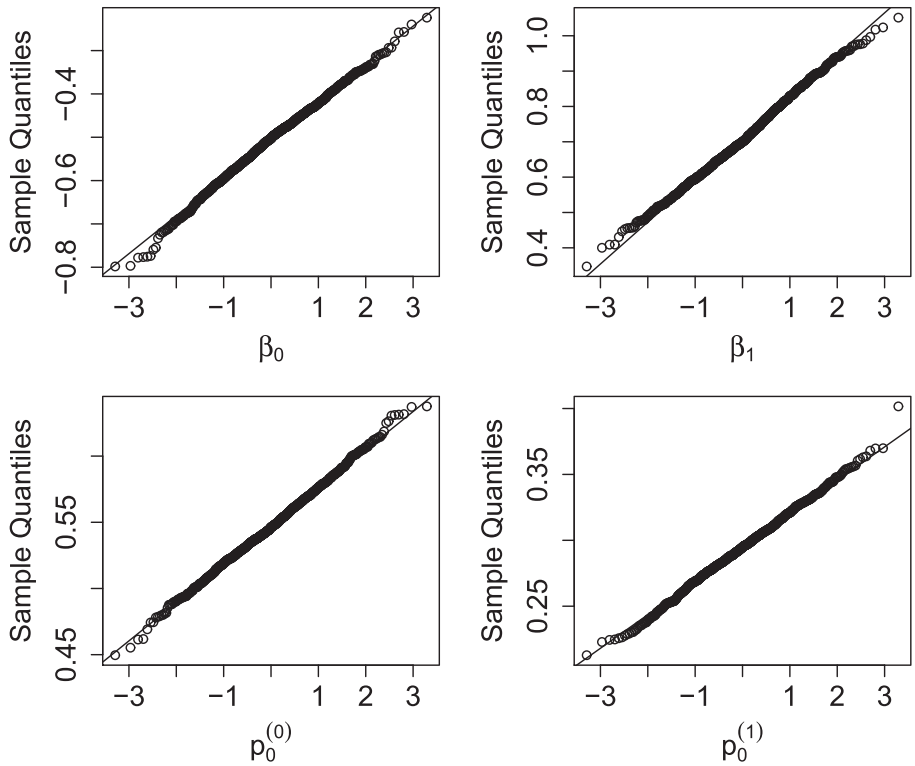

Figure 3. QQ-normal plots for the maximum likelihood estimates of β_0, β_1, p_0, and p_1 sample size $n = 200$.

Table II. Empirical rejection rates of the null hypothesis $H_0 : a = b = 1$ at a nominal significance level of 5%.

		n			
Censored	(a, b)	100	200	400	800
30%	(1.0,1.0)	0.039	0.048	0.052	0.050
	(1.5,2.0)	0.189	0.417	0.608	0.857
	(2.0,3.0)	0.572	0.801	0.878	0.901
	(3.0,3.0)	0.867	0.903	0.921	0.938
	(6.0,4.0)	0.905	0.941	0.950	0.968
	(6.0,8.0)	0.923	0.967	0.979	0.979
	(10.0,10.0)	0.948	0.973	0.963	0.983
	(15.0,15.0)	0.990	0.999	0.999	0.999
60%	(1.0,1.0)	0.028	0.046	0.048	0.051
	(1.5,2.0)	0.175	0.326	0.576	0.745
	(2.0,3.0)	0.443	0.726	0.836	0.885
	(3.0,3.0)	0.810	0.887	0.892	0.898
	(6.0,4.0)	0.835	0.915	0.919	0.929
	(6.0,8.0)	0.915	0.905	0.949	0.929
	(10.0,10.0)	0.936	0.943	0.951	0.955
	(15.0,15.0)	0.967	0.989	0.992	0.998

$\hat{p}_0^{(0)}$, and $\hat{p}_1^{(0)}$ (for fixed sample size $n = 200$). The plots are displayed in Figure 3. They indicate that the normal distribution provides a reasonable approximation to the distribution of the estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{p}_0^{(0)}$, and $\hat{p}_1^{(0)}$.

Additionally, we conduct a simulation to investigate the null distribution of the likelihood ratio statistic (w) to test the hypotheses $H_0 : a = b = 1$ (PW cure rate model) versus $H_1 : a > 0$ or $b > 0$ (PBW cure rate model). In this simulation study, we consider 30% and 60% of proportion of censored observations approximately. Table II summarizes the results of the simulation study considering different sample sizes. The rejection rates are closer to the 5% nominal level for large sample sizes. Also, the chi-squared approximation for w deteriorates when the proportion of censored observations increases. Further, the power of the LR test increases as the sample size increases.

7. Local influence

Influence diagnostic is an important step in the analysis of data, because it provides an indication of bad model fit or of influential observations. Because regression models are sensitive to the underlying model assumptions, generally performing a sensitivity analysis is strongly advisable. Another approach suggested by Cook [19] considers small perturbations represented by the vector $\boldsymbol{\omega}$ instead of removing observations and is related to a particular perturbation scheme. Local influence calculation can be carried out for models (11) and (12). If the likelihood displacement $LD(\boldsymbol{\omega}) = 2\{l(\hat{\boldsymbol{\psi}}) - l(\hat{\boldsymbol{\psi}}_{\boldsymbol{\omega}})\}$ is used, where $\hat{\boldsymbol{\psi}}_{\boldsymbol{\omega}}$ is the MLE under the perturbed model, the normal curvature for $\boldsymbol{\psi}$ at the direction $\mathbf{d}$, where $\|\mathbf{d}\| = 1$, is given by $C_{\mathbf{d}}(\boldsymbol{\psi}) = 2|\mathbf{d}^T\boldsymbol{\Delta}^T\left[\ddot{\mathbf{L}}(\boldsymbol{\psi})\right]^{-1}\boldsymbol{\Delta}\mathbf{d}|$. Here, $\boldsymbol{\Delta}$ is a $(p+4)\times n$ matrix, which depends on the perturbation scheme and whose elements are given by $\Delta_{ji} = \partial^2 l(\boldsymbol{\psi}|\boldsymbol{\omega})/\partial\psi_j\partial\omega_i$ ($i = 1,\ldots,n$ and $j = 1,\ldots,p+4$) evaluated at $\hat{\boldsymbol{\psi}}$ and $\boldsymbol{\omega}_0$, where $\boldsymbol{\omega}_0$ is the no perturbation vector (see [19, 32, 33]). For the PSBW model with long-term survivors, we calculate the normal curvatures $C_{\mathbf{d}}(\boldsymbol{\beta})$ and $C_{\mathbf{d}}(\boldsymbol{\gamma})$ to perform various index plots, for instance, the index plot of the eigenvector $\mathbf{d}_{max}$ corresponding to the largest eigenvalue $C_{\mathbf{d}_{max}}$ of the matrix $\mathbf{B} = -\boldsymbol{\Delta}^T\left[\ddot{\mathbf{L}}(\boldsymbol{\psi})\right]^{-1}\boldsymbol{\Delta}$ and the index plots of $C_{\mathbf{d}_i}(\boldsymbol{\beta})$ and $C_{\mathbf{d}_i}(\boldsymbol{\gamma})$, the so-called total local influence (see, for example, [34]), where $\mathbf{d}_i$ is an $n\times 1$ vector of zeros with one at the ith position. Thus, the curvature at direction $\mathbf{d}_i$ takes the form $C_i = 2|\boldsymbol{\Delta}_i^T\left[\ddot{\mathbf{L}}(\boldsymbol{\psi})\right]^{-1}\boldsymbol{\Delta}_i|$, where $\boldsymbol{\Delta}_i^T$ denotes the ith row of $\boldsymbol{\Delta}$. It is usual to point out those cases such that $C_i \geqslant 2\bar{C}$, where $\bar{C} = \frac{1}{n}\sum_{i=1}^n C_i$. Another influence measure for the ith observation is $U_i = \sum_{k=1}^{n_1}\lambda_k e_{ki}^2$, where $\{(\lambda_k, \mathbf{e}_k)|k = 1,\ldots,n\}$ are the eigenvalue–eigenvector pairs of $\mathbf{B}$ with $\lambda_1 \geqslant \cdots \geqslant \lambda_{n_1} \geqslant \lambda_{n_1+1} = \cdots = \lambda_n = 0$, and $\{\mathbf{e}_k = (e_{k1},\ldots,e_{kn})^T\}$ is the associated orthonormal basis. Zhu and Zhang [35] studied the influence measure u_i systematically under a case-weight perturbation. Thus, this influence measure expresses local sensitivity to the log-likelihood of the perturbations.

Consider the vector of weights $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_n)^T$. From the log-likelihood (23), under three perturbation schemes, we derive the matrix

$$\boldsymbol{\Delta} = (\boldsymbol{\Delta}_{ji})_{(p+4)\times n} = \left(\frac{\partial^2 l(\boldsymbol{\psi}|\boldsymbol{\omega})}{\partial\psi_i\boldsymbol{\omega}_j}\right)_{(p+4)\times n}, \quad j = 1,\ldots,p+4 \quad \text{and} \quad i = 1,\ldots,n.$$

- *Case-weight perturbation*

  In this case, the log-likelihood function has the form

$$\begin{aligned} l(\boldsymbol{\psi}|\boldsymbol{\omega}) = {} & \sum_{i=1}^n \omega_i\delta_i \log\left[\frac{c}{\tau^c B(a,b)}\right] + \sum_{i=1}^n \delta_i g(\mathbf{x}_i,\boldsymbol{\beta}) + (c-1)\sum_{i=1}^n \omega_i\delta_i\log(t_i) \\ & - b\sum_{i=1}^n \omega_i\delta_i\left(\frac{t_i}{\tau}\right)^c + (a-1)\sum_{i=1}^n \omega_i\delta_i\log[q(t_i)] \\ & + \sum_{i=1}^n \omega_i\delta_i\log\left\{\frac{A'\left\{g(\mathbf{x}_i,\boldsymbol{\beta})[1-I_{q(t_i)}(a,b)]\right\}}{A(g(\mathbf{x}_i,\boldsymbol{\beta}))}\right\} \\ & + \sum_{i=1}^n \omega_i(1-\delta_i)\log\left\{\frac{A\left\{g(\mathbf{x}_i,\boldsymbol{\beta})[1-I_{q(t_i)}(a,b)]\right\}}{A(g(\mathbf{x}_i,\boldsymbol{\beta}))}\right\}, \end{aligned}$$

where $q(t_i) = 1 - \exp\left[-\left(\frac{t_i}{\tau}\right)^c\right]$, $0 \leqslant \omega_i \leqslant 1$, $\boldsymbol{\omega}_0 = (1, \ldots, 1)^T$, and δ_i, $g(.)$, $A(.)$ and $A'(.)$ are defined in Equation (23). Here, $\boldsymbol{\Delta} = \left(\boldsymbol{\Delta}_{\boldsymbol{\beta}}^T, \boldsymbol{\Delta}_{\boldsymbol{\gamma}}^T\right)^T$ can be calculated numerically.

- *Response perturbation*

  Now, we consider that each t_i is perturbed as $t_{iw} = y_i + \omega_i S_t$, where S_t is a scale factor that may be estimated by the standard deviation of the observed response y and $\omega_i \in \Re$. The perturbed log-likelihood function can be expressed as

$$\begin{aligned} l(\boldsymbol{\psi}|\boldsymbol{\omega}) = r\log\left[\frac{c}{\tau^c B(a,b)}\right] + \sum_{i=1}^{n}\delta_i g(\mathbf{x}_i,\boldsymbol{\beta}) + (c-1)\sum_{i=1}^{n}\delta_i \log\left(t_i^*\right) - b\sum_{i=1}^{n}\delta_i\left(\frac{t_i^*}{\tau}\right)^c \\ + (a-1)\sum_{i=1}^{n}\omega_i\delta_i\log\left[q\left(t_i^*\right)\right] + \sum_{i=1}^{n}\delta_i\log\left\{\frac{A'\left\{g(\mathbf{x}_i,\boldsymbol{\beta})[1-I_{q(t_i^*)}(a,b)]\right\}}{A(g(\mathbf{x}_i,\boldsymbol{\beta}))}\right\} \\ + \sum_{i=1}^{n}(1-\delta_i)\log\left\{\frac{A\left\{g(\mathbf{x}_i,\boldsymbol{\beta})[1-I_{q(t_i^*)}(a,b)]\right\}}{A(g(\mathbf{x}_i,\boldsymbol{\beta}))}\right\}, \end{aligned}$$

  where $t_i^* = t_i + \omega_i S_y$, $q\left(t_i^*\right) = 1 - \exp\left[-\left(\frac{t_i^*}{\tau}\right)^c\right]$ and $\boldsymbol{\omega}_0 = (0, \ldots, 0)^T$. The matrix $\boldsymbol{\Delta} = \left(\boldsymbol{\Delta}_{\boldsymbol{\beta}}^T, \boldsymbol{\Delta}_{\boldsymbol{\gamma}}^T\right)^T$ is obtained numerically.

- *Explanatory variable perturbation*

  Now, consider an additive perturbation on a particular continuous explanatory variable, say X_q, by setting $x_{iq\omega} = x_{iq} + \omega_i S_q$, where S_q is a scale factor and $\omega_i \in \Re$. The perturbed log-likelihood function has the form

$$\begin{aligned} l(\boldsymbol{\psi}|\boldsymbol{\omega}) = r\log\left[\frac{c}{\tau^c B(a,b)}\right] + \sum_{i=1}^{n}\delta_i g\left(\mathbf{x}_i^{**},\boldsymbol{\beta}\right) + (c-1)\sum_{i=1}^{n}\delta_i\log(t_i) - b\sum_{i=1}^{n}\delta_i\left(\frac{t_i}{\tau}\right)^c \\ + (a-1)\sum_{i=1}^{n}\delta_i\log[q(t_i)] + \sum_{i=1}^{n}\delta_i\log\left\{\frac{A'\left\{g\left(\mathbf{x}_i^{**},\boldsymbol{\beta}\right)[1-I_{q(t_i)}(a,b)]\right\}}{A\left(g\left(\mathbf{x}_i^{**},\boldsymbol{\beta}\right)\right)}\right\} \\ + \sum_{i=1}^{n}(1-\delta_i)\log\left\{\frac{A\left\{g\left(\mathbf{x}_i^{**},\boldsymbol{\beta}\right)[1-I_{q(t_i)}(a,b)]\right\}}{A\left(g\left(\mathbf{x}_i^{**},\boldsymbol{\beta}\right)\right)}\right\}, \end{aligned}$$

  where $\mathbf{x}_i^{**T}\boldsymbol{\beta} = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_q(x_{iq} + \omega_i S_q) + \cdots + \beta_p x_{ip}$ and $\boldsymbol{\omega}_0 = (0, \ldots, 0)^T$. The matrix $\boldsymbol{\Delta} = \left(\boldsymbol{\Delta}_{\boldsymbol{\beta}}^T, \boldsymbol{\Delta}_{\boldsymbol{\gamma}}^T\right)^T$ is obtained numerically.

8. Application: breast carcinoma data

Among the several proposed risk classification schemes for predicting survival in women with breast carcinoma, one of the most commonly used is the Nottingham Prognostic Index (NPI). The NPI is an index, derived under a retrospective multivariate study, that is able to predict survival in patients with breast cancer. The index is based on tumour size, lymph node stage, and histological grade and allows the stratification of patients into three different prognostic groups. The NPI allows us to accurately predict prognosis, and we advocate its more common use. For more details, see, for example, [36–38].

Here, the goal is to use the PSBW regression model with long-term survivors to predict breast carcinoma mortality more accurately.

A total of $n = 284$ women who had been treated with mastectomy and axillary lymph node dissection at Memorial Sloan-Kettering Cancer Center (New York, NY) between 1976 and 1979 met the following requirements for study inclusion: confirmation of the presence of invasive mammary carcinoma, no receipt of neoadjuvant or adjuvant systemic therapy, no previous history of malignancy, and negative lymph node status as assessed on routine histopathologic examination.

The data collected by Kattan [20] represent the survival times (T) until the patient's death (73 died of the disease) or the censoring times at the end of the study. Some explanatory variables are associated with pathologic characteristics of the tumor. The tumor grading was performed using the standard modified Bloom–Richardson system. The lymphovascular invasion was obtained using morphologic criteria. The lymph node status was measured according to immunohistochemistry (IHC) and hematoxylin and eosin (H&E) stains. The explanatory variables for each woman ($i = 1, \ldots, 284$) are described as follows:

- t_i: observed time (in years);
- δ_i: censoring indicator (0: right censored, 1: time observed);
- x_{i1}: age (in years);
- x_{i2}: multifocality (0: no, 1:yes);
- x_{i3}: tumor size (in cm);
- x_{i4}: tumor grading (0: I, 1: II, II and lobular);
- x_{i5}: lymphovascular invasion (0: no, 1: yes);
- x_{i6}: lymph node status (0: IHC+ IHC- and H&E-, 1: IHC+ and H&E+).

There are 74% censored observations corresponding to the women who died from other causes or were still alive at the end of the study. Figure 4 displays the estimated Kaplan–Meier survival function with a well-pronounced plateau, which, according to Yakovlev and Tsodikov [15], may be thought of as an indication of the presence of a proportion of patients for whom the breast carcinoma will never recur, and the patients can be considered as cured.

The explanatory variables are related with the parameter θ according to the following structures. For the PBW cure rate model, we consider ($i = 1, \ldots, 284$):

$$\theta_i = g(\mathbf{x}_i\boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}).$$

For the BeBW, LBW, NBiBW, and GBW models, we consider

$$\theta_i = g(\mathbf{x}_i\boldsymbol{\beta}) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6})}.$$

We determine the MLEs of the model parameters using the NLMixed procedure in SAS. The values of the global deviance, Akaike information criterion, and consistent Akaike information criterion statistics are listed in Table III. The lowest values of the information criteria correspond to the PBW model, which provides a better fit to the current data than the other models.

Further, we calculate the maximum unrestricted and restricted log-likelihoods and the LR statistics for testing some submodels. For example, the LR statistic for testing the hypotheses H_0: $a = b = 1$ versus H_1: H_0 *is not true*, that is, to compare the PBW and PW regression models with long-term survivors, is $w = 2\{-337.10 - (-339.45)\} = 4.70$ (p-value < 0.05), which yields favorable indications toward the PBW regression model with long-term survivors.

Table IV lists the MLEs for the fitted PBW regression model. At a 5% significance level, all regression coefficients are significant except those for the explanatory variables: age (x_1), multifocality (x_2), and lymphovascular invasion (x_5).

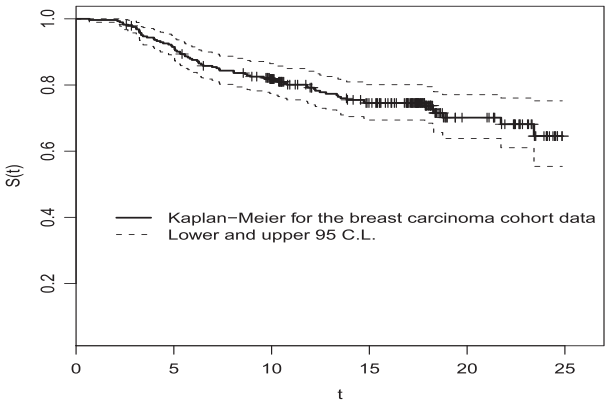

Figure 4. Kaplan–Meier curves for the breast carcinoma cohort data.

Table III. The global deviance, Akaike information criterion, and consistent Akaike information criterion statistics for some models.

Model PSBW	GD	AIC	CAIC
BeBW	718.5	740.5	741.5
BeW	727.4	745.4	746.0
PBW	674.2	696.2	697.2
PW	678.9	696.9	697.5
LBW	693.8	715.8	716.8
LW	694.7	711.8	712.5
NBiBW	673.1	697.1	698.3
NBiW	678.9	698.9	699.7
GBW	675.5	697.5	698.5
GW	680.2	698.2	698.9

GD, global deviance; AIC, Akaike information criterion; CAIC, consistent Akaike information criterion; PSBW, power series beta Weibull; BeBW, Bernoulli beta Weibull; BeW, Bernoulli Weibull; PBW, Poisson beta Weibull; PW, Poisson Weibull; LBW, logarithmic beta Weibull; LW, logarithmic Weibull; NBiBW, negative binomial beta Weibull; NBiW, negative binomial Weibull; GBW, geometric beta Weibull; GW, geometric Weibull.

Table IV. Maximum likelihood estimates for the full Poisson beta Weibull regression model with cure rate fraction fitted to the breast carcinoma data.

Parameter	Estimate	Standard error	p-value
a	6.2829	0.5850	–
b	4.1988	0.4689	–
c	0.5221	0.0801	–
τ	12.2495	4.5505	–
β_0	−4.2533	1.1925	0.0004
β_1	0.0008	0.0108	0.9378
β_2	0.4520	0.3560	0.2053
β_3	0.2850	0.0974	0.0037
β_4	2.6132	1.0096	0.0101
β_5	0.4476	0.2771	0.1074
β_6	1.2660	0.3069	<0.0001

As suggested by a referee, we also present the results of the fit of a semiparametric mixture cure model. These analyses are based on the paper by Corbiere and Joly [39], who used a SAS macro to estimate this model with covariates. The macro is applied to the breast carcinoma data, whose results are given in Table V.

The parameter estimates from the PSBW model with long-term survivors and the semiparamteric mixture cure model are quite similar (Table V). Moreover, the standard errors of the MLEs from the fitted PSBW model by using the inverse of the Hessian matrix at the parameter estimates are smaller than those obtained from the Weibull mixture model. This fact indicates that the estimates of the PSBW model are more precise than those of the semiparametric mixture cure model. A difference exists regarding the significance of the covariate x_5; because x_5 is insignificant in the PSBW model, it becomes significant at the 5% level in the semiparametric mixture model. However, we note that in the final model

Table V. Estimates, standard errors, and p-values for the breast carcinoma data from the semiparametric mixture cure model.

Parameter	Estimate	Standard error	p-value
β_0	−5.6896	1.2929	<0.0001
β_1	−0.0021	0.0127	0.8685
β_2	0.6375	0.4972	0.1998
β_3	0.5386	0.1544	0.0005
β_4	3.7188	1.0444	0.0004
β_5	1.0938	0.3896	0.0050
β_6	2.4069	0.5804	<0.001

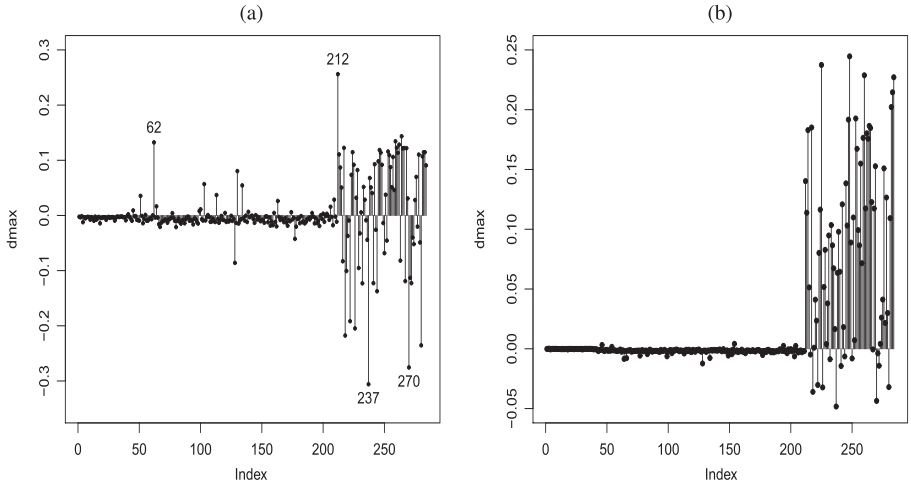

Figure 5. Index plot of $\mathbf{d}_{max}$ for $\boldsymbol{\psi}$ for the breast carcinoma data. (a) Case-weight perturbation; (b) Response perturbation.

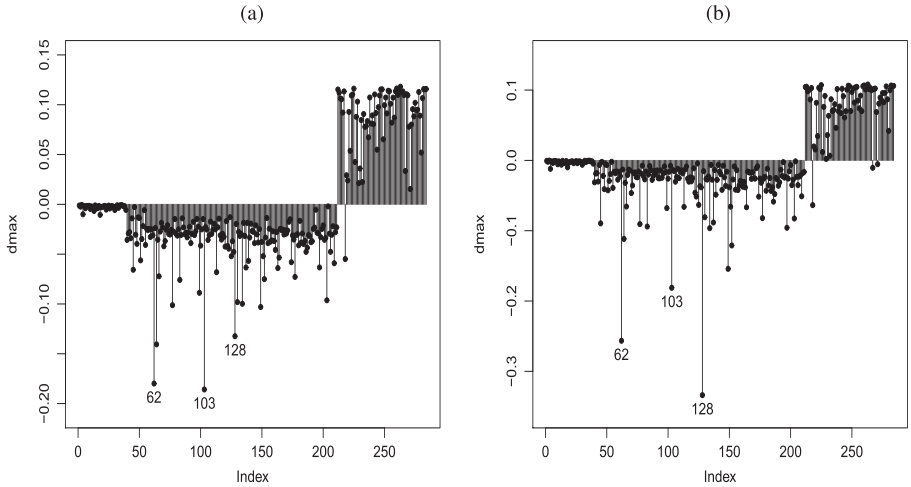

Figure 6. Index plots for $\boldsymbol{\psi}$ on the breast carcinoma data (explanatory variable perturbation): (a) $\mathbf{d}_{max}$ (*age*); (b) $\mathbf{d}_{max}$ (*tumor size*).

presented in Table VIII, the covariate x_5 is significant at the 6% significance level. Future work may be addressed to compare advantages and disadvantages of the PSBW model with long-term survivors and the semiparamteric mixture cure model because both models are very important for the analyses.

Table VI. Relative changes [%], new estimates, and the corresponding p-values in parentheses.

Dropped observation	a	b	c	τ
None	–	–	–	–
	6.2829	4.1988	0.5221	12.2495
	(–)	(–)	(–)	(–)
#62	[1.2]	[1.4]	[0.2]	[0.0]
	6.3562	4.2571	0.5210	12.2543
	(–)	(–)	(–)	(–)
#103	[−0.5]	[1.6]	[1.1]	[−0.1]
	6.3136	4.1313	0.5162	12.2604
	(–)	(–)	(–)	(–)
#128	[0.4]	[−1.0]	[−1.0]	[0.0]
	6.2578	4.2417	0.5271	12.2457
	(–)	(–)	(–)	(–)
#212	[−7.1]	[−8.5]	[4.9]	[−6.1]
	6.7282	4.5574	0.4964	12.9970
	(–)	(–)	(–)	(–)
#237	[0.1]	[−8.1]	[−5.2]	[0.4]
	6.2788	4.5403	0.5495	12.1945
	(–)	(–)	(–)	(–)
#270	[−5.7]	[−12.8]	[−1.1]	[0.0]
	6.6421	4.7369	0.5281	12.2473
	(–)	(–)	(–)	(–)
Set $\mathcal{A}$	[3.8]	[−10.2]	[−12.2]	[0.8]
	6.0438	4.6285	0.5858	12.1546
	(–)	(–)	(–)	(–)

8.1. Diagnostics analysis

In what follows, we shall apply the local and global influence methods developed in the previous sections for the purpose of identifying influential observations in the new regression model fitted to the dental caries data.

The case-weight perturbation is used, and we obtain the value of the maximum curvature $C_{\mathbf{d}_{max}} = 1.3782$. Figure 5(a) plots the eigenvector corresponding to $\mathbf{d}_{max}$ and indicates that the observations ♯62, ♯212, ♯237, and ♯270 are again very distinct from the others. The influence of perturbations on the observed survival times is now analyzed (response variable perturbation). The value of the maximum curvature is $C_{\mathbf{d}_{max}} = 13.8180$. Figure 5(b) displays the plots of $\mathbf{d}_{max}$ versus the observation index, where we can verify that there is no observation highlighted.

The perturbation of vectors for each continuous explanatory variable (x_1 and x_3) is now investigated. The values for the maximum curvature are $C_{\mathbf{d}_{max}} = 1.87$, $C_{\mathbf{d}_{max}} = 1.52$, $C_{\mathbf{d}_{max}} = 0.0453$, and $C_{\mathbf{d}_{max}} = 4.3832$ for x_1 and x_3, respectively. Then, plots of $\mathbf{d}_{max}$ against the index of the observations are displayed in Figure 6(a) and (b). We note that the observations ♯62, ♯103, and ♯128 are very distinct from the others.

We now identify the characteristics of each observation in the sample. The observations ♯62 and ♯103 have the largest censoring times, and the censored observation ♯128 has the large tumor size (x_3). The observation ♯212 corresponds to the minimum lifetime, and the observations ♯237 and ♯270 have the maximum failure times. We note that each observation detected in the influence analysis has a typical characteristic very different from the others.

8.2. Impact of the detected influential observations

We conclude that the diagnostic analysis detected, as potentially influential observations, the following six cases: #62, #103, #128, #212 #237, and #270. In order to reveal the impact of the six observations on the parameter estimates, Tables VI and VII present the relative changes (RC)—in percentage—in

Table VII. Relative changes [%], new estimates, and the corresponding p-values in parentheses.

Dropped observation	β_0	β_1	β_2	β_3	β_4	β_5	β_6
None	–	–	–	–	–	–	–
	−4.2533	0.0008	0.4520	0.2850	2.6132	0.4476	1.2660
	(0.0004)	(0.9378)	(0.2053)	(0.0037)	(0.0101)	(0.1074)	(<0.0001)
#62	[−7.7]	[394.9]	[30.3]	[10.7]	[8.9]	[17.0]	[2.1]
	−3.9260	−0.0025	0.5891	0.3156	2.3801	0.5235	1.2396
	(0.0004)	(0.8196)	(0.0924)	(0.0011)	(0.0090)	(0.0569)	(<0.0001)
#103	[1.3]	[228.4]	[6.9]	[0.1]	[−2.2]	[−17.8]	[−8.0]
	−4.1987	−0.0011	0.4207	0.2847	2.6718	0.5271	1,3675
	(0.0005)	(0.9202)	(0.2391)	(0.0038)	(0.0088)	(0.0562)	(<0.0001)
#128	[−3.1]	[8.8]	[18.0]	[−37.5]	[3.6]	[7.1]	[−2.1]
	−4.3832	0.0008	0.3705	0.3918	2.5202	0.4157	1.2922
	(0.0002)	(0.9428)	(0.3050)	(0.0014)	(0.0108)	(0.1347)	(<0.0001)
#212	[−298.2]	[−141.0]	[−4.4]	[4.5]	[−485.6]	[11.1]	[3.2]
	−16.9349	0.0020	0.4719	0.2722	15.3026	0.3981	1.2261
	(0.9766)	(0.8513)	(0.1872)	(0.0062)	(0.9788)	(0.1604)	(0.0001)
#237	[−2.8]	[−72.2]	[−1.9]	[2.1]	[−0.4]	[−3.4]	[0.1]
	−4.3710	0.0014	0.4607	0.2790	2.6232	0.4627	1.2648
	(0.0003)	(0.8939)	(0.1976)	(0.0049)	(0.0104)	(0.0964)	(<0.0001)
#270	[−0.9]	[46.8]	[−1.3]	[0.5]	[0.6]	[−3.1]	[0.1]
	−4.2931	0.0004	0.4578	0.2835	2.5985	0.4614	1.2653
	(0.0004)	(0.9671)	(0.2003)	(0.0040)	(0.0104)	(0.0973)	(<0.0001)
Set $\mathcal{A}$	[−14.0]	[601.8]	[−16.5]	[−46.4]	[−16.9]	[−20.6]	[−5.8]
	−4.8470	−0.0042	0.5268	0.4172	3.0546	0.5398	1.3398
	(0.0008)	(0.7007)	(0.1367)	(0.0010)	(0.0183)	(0.0528)	(<0.0001)

Table VIII. Maximum likelihood estimates of the parameters for the reduced Poisson beta Weibull regression model with cure fraction fitted to the breast carcinoma data.

Parameter	Estimate	Standard error	p-value
a	117.83	0.6867	–
b	69.4666	0.6032	–
c	0.1082	0.0183	–
τ	12.2310	3.6079	–
β_0	−4.1218	1.0468	0.0001
β_3	0.2984	0.0942	0.0017
β_4	2.6146	1.0104	0.0102
β_5	0.5102	0.2711	0.0609
β_6	1.2050	0.3012	<0.0001

the parameter estimates of the systematic components for the mean and dispersion, respectively, after dropping one of the six cases with outstanding influence and also when all of them are dropped at once (represented by the set $\mathcal{A} = \{62, 103, 128, 212, 237, 270\}$) and the new parameter estimates. We also present the corresponding p-values (in parentheses) for the new estimates in these tables. The RC of each estimate is defined as $\mathrm{RC}_{\theta_j} = [\widehat{\theta}_j - \widehat{\theta}_{j(\mathrm{I})}]/\widehat{\theta}_j \times 100\%$, where $\widehat{\theta}_{j(\mathrm{I})}$ denotes the MLE of θ_j, for $j = 1, \ldots, k$ (where k is the total number of parameters), after the set I of observations has been removed.

We have no convergence problems to compute the figures of these two tables: the MLEs of the parameters for the BW model (Table VI) and the estimates of the regression parameters (Table VII). We note that the results are very close with the simulation results given in Section 6. The figures in these tables reveal that the significance (or insignificance) of the model parameters is not modified when the observations #22, #119, #220, #431, and #458 or the set $\mathcal{A}$ is removed from the data; that is, these cases do not

change the inference on the parameters of the PBW regression model. We can conclude that this model with long-term survivors is robust to potential influential points.

Further, we seek a parsimonious regression model for these data. The MLEs of the parameters for the reduced PBW regression model with long-term survivors fitted to these data are listed in Table VIII.

We now estimate the cure rate (p_0). Note that

$$\hat{\theta} = \frac{1}{417}\sum_{i=1}^{417}\hat{\theta}_i = 0.4491,$$

where

$$\hat{\theta}_i = \exp(-4.1218 + 0.2984x_{i3} + 2.6146x_{i4} + 0.5102x_{i5} + 1.2050x_{i6}), \tag{24}$$

and then $\hat{p}_0 = \mathrm{e}^{-\hat{\theta}} = 0.6382$.

The regression structure (24) and the p-values in Table VIII suggest that the tumor size, tumor grading, lymphovascular invasion, and lymph node status as defined earlier are significant prognostic variables for assessing mortality risk in women with breast carcinoma. In order to assess if the model is appropriate, Figure 7 displays the empirical survival function and the estimated marginal survival functions given by (14) from the fitted reduced PBW model with long-term survivors.

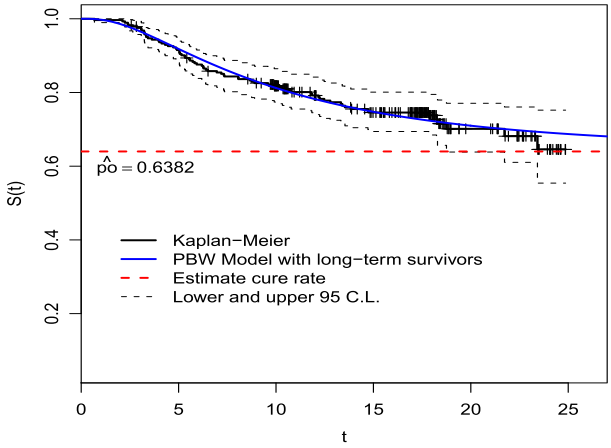

Figure 7. Kaplan–Meier curves (solid lines), the estimated Poisson beta Weibull survival functions and the estimated cure fraction for the breast carcinoma data.

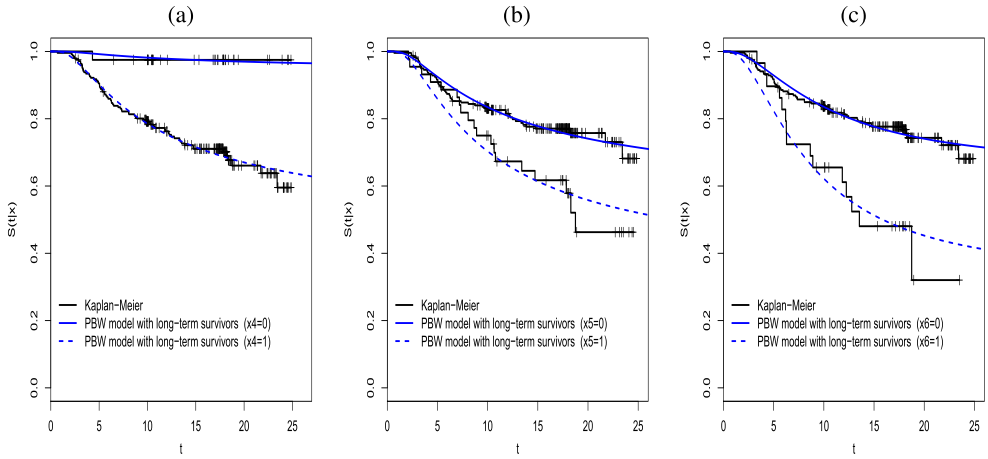

Figure 8. Kaplan–Meier curves stratified by explanatory variable and estimated survival functions to the breast carcinoma data: (a) $x4$ (*tumor grading*); (b) $x5$ (*lymphovascular invasion*); (c) $x6$ (*lymph node status*).

Table IX. Estimated cure probabilities for four mastectomized women.

Patient	Tumor size	Tumor grade	Lymphovascular	Staining	$\hat{p}_0$
A	2.2	0	1	0	0.94
B	2.2	0	1	1	0.84
C	2.2	1	1	0	0.49
D	2.2	1	0	1	0.24

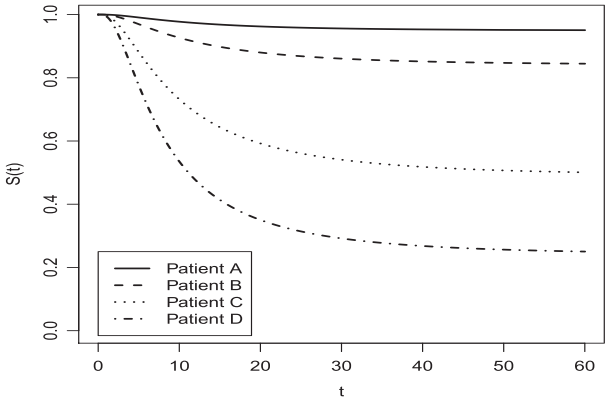

Figure 9. Estimates of recurrence free probability curves for the patients A, B, C, and D.

We fit the PBW regression models with long-term survivors for each explanatory variable. In Figure 8(a)–(c), we plot the empirical survival function and the estimated survival function (14) for each explanatory variable. We conclude that the PBW regression model provides a good fit to these data.

Clearly, Equation (24) reveals that the estimates of the cure probability for breast carcinoma in women can be very different depending upon the values of the explanatory variables associated with the tumor biopsy. We consider four hypothetical women A, B, C, and D who underwent postmastectomy with values for the explanatory variables given in Table IX. For example, for woman A, tumor size $=$ 2.2, tumor grade $=$ 0, lymphovascular $=$ 1, and lymph node status $=$ 0; and for woman D, tumor size $=$ 2.2, tumor grade $=$ 1, lymphovascular $=$ 0, and lymph node status $=$ 1, we obtain very different cure probabilities, namely 0.94 for woman A and 0.24 for woman D.

Finally, using Equation (14), we obtain $S_{\text{pop}}(t; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$. Figure 9 displays the plots of the recurrence free probability estimates for the hypothetical women described earlier. These last analyses can only be performed using the PSBW model with long-term survivors, because the semiparametric mixture cure model is much more complicated to do this type of analysis. It represents an advantage of the PSBW model with long-term survivors over the semiparametric mixture cure model.

9. Concluding remarks

In this paper, we propose a new PSBW regression model with cure rate to estimate breast carcinoma mortality versus time and cure probability for women who had been treated with mastectomy and axillary lymph node dissection. The unknown number of competing causes that can influence the survival time is assumed to follow a PS distribution. Further, the time of the left breast carcinoma cells after the mastectomy to metastasize is assumed to have the BW distribution. This distribution is more flexible than other existing models to estimate the effects of the explanatory variables on the survival function and on the proportion of the cure population. Some specific compounding regression models are special cases of our general formulation. Several pathological explanatory variables are taken into account to explain the breast cancer recurrence as, for example, age, multifocality, tumor size, tumor grading, lymphovascular invasion, and lymph node status. The composing model leads to more precise estimates of the effects of the explanatory variables on the recurrence times. The maximum likelihood estimation procedure yields efficient estimators of the regression parameters, and asymptotic tests can be performed for the parameters using likelihood ratio statistics. We conclude that the tumor size, tumor grading, lymphovascular invasion, and lymph node status are significant prognostic variables for determining survival time and mortality risk in women with breast carcinoma.

Appendix A: Some properties

Here, we demonstrate that the noncured density function $f_{\rm nc}(t)$ can be expressed as a linear combination of *exponentiated-G* ('Exp-G' for short) densities, where the weighted coefficients depend only on the probabilities of M. For an arbitrary baseline CDF $G(t) = 1-S(t)$ and a discrete random variable M defined by the PGF $P(z)$ in Equation (1), the unified PDF for the noncured population is given by Equation (7). From now on, a random variable Z_a has the Exp-G distribution with power parameter $a > 0$, say $Z_a \sim$Exp-F(a), if its PDF and CDF are given by

$$h_a(x) = a f(x) F^{a-1}(x) \quad \text{and} \quad H_a(x) = F^a(x),$$

respectively.

The noncured distribution is now represented by the random variable V. By differentiating $A(\theta)$ in Equation (7), expanding the resulting binomial and substituting $\sum_{j=0}^{\infty}\sum_{m=0}^{j}$ by $\sum_{j=0}^{\infty}\sum_{m=j}^{\infty}$, we can write after lengthy algebra

$$f_{\rm nc}(t) = \sum_{j=0}^{\infty} v_j\, h_{j+1}(t), \tag{A.1}$$

where $h_{j+1}(t)$ denotes the Exp-F($j+1$) PDF with power parameter $j+1$ (for $j \geqslant 0$),

$$v_j = \frac{(-1)^j}{(j+1)[A(\theta)-a_0]} \sum_{m=j}^{\infty} (m+1)\, a_{m+1}\, \theta^{m+2} \binom{m}{j}$$

and the a_m's are the probabilities of the discrete distribution (1).

So, several mathematical quantities (such as ordinary and incomplete moments, generating function, and mean deviations) for the noncured population (V) can be obtained from Equation (A.1) by knowing those of the Exp-G distribution. The structural properties of the exponentiated distributions have been studied by many authors in recent years (see, for example, [24]). Equation (A.1) is the main result of this section.

Hereafter, let $Y_j \sim$Exp-Weibull($j+1$). We provide two explicit expressions for the moments of the PSBW distribution. A first formula for the rth moment of V can be immediately obtained from Equation (A.1) as

$$E(V^r) = \sum_{j=0}^{\infty} v_j\, E(Y_j^r).$$

The quantity $E(Y_j^r)$ is determined from Equation (10) with $b = 1$ and $a = j+1$. Then,

$$E(V^r) = \tau^r\, \Gamma(r/c+1) \sum_{n=0}^{\infty} \frac{(-1)^n\, q_n}{(n+1)^{r/c+1}}, \tag{A.2}$$

where $q_n = \sum_{j=0}^{\infty} \frac{v_j}{j+1} \binom{j}{n}$ for $n \geqslant 0$.

A second formula for $E(V^r)$ can be derived in terms of the quantile function of Z, say $Q_{BW}(u)$. We can write from Equation (4)

$$E(V^r) = \sum_{j=0}^{\infty} (j+1)\, v_j\, \tau(r,j), \tag{A.3}$$

where

$$\tau(r,j) = \int_0^{\infty} t^r\, G(x)^j\, dG(x) = \int_0^1 Q_{BW}(u)^r\, u^j du,$$

and $Q_{BW}(u)$ is the inverse function of $F(z;\boldsymbol{\gamma}) = I_{1-\exp\left[-\left(\frac{z}{\tau}\right)^c\right]}(a,b) = u$.

Equations (A.2) and (A.3) are general formulae for the moments of the noncured PSBW distribution.

Acknowledgements

The authors would like to thank the Editor and two referees for careful reading and for their constructive comments, which greatly improved the paper.

References

1. Maller RA, Zhou X. *Survival Analysis with Long-Term Survivors*. Wiley: New York, 1986.
2. Ibrahim JG, Chen MH, Sinha D. *Bayesian Survival Analysis*. Springer: New York, 2001.
3. Tsodikov AD, Ibrahim JG, Yakovlev AY. Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association* 2003; **98**:1063–1078.
4. Cooner F, Banerjee S, Carlin BP, Sinha D. Flexible cure rate modeling under latent activation schemes. *Journal of the American Statistical Association* 2007; **102**:560–572.
5. Tournoud M, Ecochard R. Application of the promotion time cure model with time-changing exposure to the study of HIV/AIDS and other infectious diseases. *Statistics in Medicine* 2007; **26**:1008–1021.
6. Zhao Y, Lee AH, Yau KKW, Burke V. A score test for assessing the cured proportion in the long-term survivor mixture model. *Statistics in Medicine* 2009; **28**:3454–3466.
7. de Castro M, Cancho VG, Rodrigues J. A Bayesian long-term survival model parametrized in the cured fraction. *Biometrical Journal* 2009; **51**:443–455.
8. Ortega EMM, Cancho VG, Paula GA. Generalized log-gamma regression models with cure fraction. *Lifetime Data Analysis* 2009; **15**:79–106.
9. Rodrigues J, Cancho VG, de Castro M, Louzada-Neto F. On the unification of the long-term survival models. *Statistics and Probability Letters* 2009a; **79**:753–759.
10. Rodrigues J, de Castro M, Cancho VG, Balakrishnan N. COM—Poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning Inference* 2009b; **139**:3605–3611.
11. de Castro M, Cancho V, Rodrigues J. A hands-on approach for fitting long-term survival models under the gamlss framework. *Computer Methods and Programs in Biomedicine* 2010; **97**:168–177.
12. Perdona GSC, Louzada-Neto F. A general hazard model for lifetime data in the presence of cure rate. *Journal of Applied Statistics* 2011; **38**:1395–1405.
13. Boag J. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society - Series B* 1949; **11**:15–44.
14. Berkson J, Gage RP. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association* 1952; **88**:1412–1418.
15. Yakovlev A, Tsodikov AD. *Stochastic Models of Tumor Latency and their Biostatistical Applications*, Mathematical Biology and Medicine, Vol, vol. 1. World Scientific: New Jersey, 1996.
16. Famoye F, Lee C, Olumolade O. The beta-Weibull distribution. *Journal of Statistical Theory and Applications* 2005; **4**: 121–136.
17. Lee C, Famoye F, Olumolade O. Beta-Weibull distribution: some properties and applications to censored data. *Journal of Modern Applied Statistical Methods* 2007; **6**:173–186.
18. Cordeiro GM, Nadarajah S, Ortega EMM. General results for the beta Weibull distribution. *Journal of Statistical Computation and Simulation* 2013; **83**:1082–114.
19. Cook RD. Assessment of local influence (with discussion). *Journal of the Royal Statistical Society B* 1986; **48**:133–169.
20. Kattan WM, Giri D, Panageas KS, Hummer A, Cranor M, Zee KJV, Hudis CA, Norton L, Borgen PI, Tan LK. A tool for predicting breast carcinoma mortality in women who do not receive adjuvant therapy. *Cancer* 2004; **101**:2509–2515.
21. Mudholkar GS, Srivastava DK, Friemer M. The exponentiated Weibull family: a reanalysis of the bus-motor-failure data. *Technometrics* 1995; **37**:436–445.
22. Lai CD, Xie M, Murthy DNP. A modified Weibull distribution. *Transactions on Reliability* 2003; **52**:33–37.
23. Silva GO, Ortega EMM, Cordeiro GM. The beta modified Weibull distribution. *Lifetime Data Analysis* 2010; **16**:409–430.
24. Nadarajah S, Kotz S. The beta exponential distribution. *Reliability Engineering and System Safety* 2006; **91**:689–697.
25. Gupta RD, Kundu D. Generalized exponential distributions. *Australian and New Zealand Journal of Statistics* 1999; **41**:173–188.
26. Kundu D, Raqab MZ. Generalized Rayleigh distribution: different methods of estimation. *Computational Statistics and Data Analysis* 2005; **49**:187–200.
27. Chen MH, Ibrahim JG, Sinha D. A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association* 1999; **94**:909–919.
28. Liu CS, Taylor JM. Identifiability of cure models. *Statistics and Probability Letters* 2001; **54**:389–395.
29. Peng Y, Zhang J. Identifiability of a mixture cure frailty model. *Statistics and Probability Letters* 2008; **78**:2604–2608.
30. Tournoud M, Ecochard R. Promotion time models with time-changing exposure heterogeneity: application to infectious diseases. *Biometrical Journal* 2008; **50**:395–407.
31. Hougaard P, Myglegaard P, Borch-Johnsen K. Heterogeneity models of disease susceptibility, with application to diabetic nephropathy. *Biometrics* 1994; **50**:1178–1188.
32. Zhu H, Ibrahim JG, Lee S, Zhang H. Perturbation selection and influence measures in local influence analysis. *The Annals of Statistics* 2007; **35**:2565–2588.
33. Jung KM. Local influence in Generalized estimating equations. *Scandinavian Journal of Statistics* 2008; **35**:286–294.
34. Lesaffre E, Verbeke G. Local influence in linear mixed models. *Biometrics* 1998; **54**:570–582.
35. Zhu H, Zhang H. A diagnostic procedure based on local influence. *Biometrika* 2004; **91**:579–589.

36. Sundquist M, Thorstenson S, Brudin L, Nordenskjold B. Applying the Nottingham Prognostic Index to a Swedish breast cancer population. *Breast Cancer Research and Tretment* 1999; **53**:1–88.
37. D'Eredita G, Giardina C, Martellotta M, Natale T, Ferrarese F. Prognostic factors in breast cancer: the predictive value of the Nottingham Prognostic Index in patients with long-term follow-up that were treated in a single institution. *European Journal of Cancer* 2001; **37**:591–596.
38. Rostgaard K, Mouridsen HT, Vaeth M, Holst H, Olesen K P, Lynge E. A modified Nottingham Prognostic Index for breast cancer patients diagnosed in Denmark 1978–1994. *Acta Oncologica* 2001; **40**:838–843.
39. Corbiere F, Joly P. A SAS macro for parametric and semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine* 2007; **85**:173–180.