A zero-inflated non-default rate regression model for credit scoring data

A zero-inflated non-default rate regression model for credit scoring data

Francisco Louzada

Institute of Mathematical Science and Computing, University of São Paulo, Brazil

Fernando F. Moreira

Credit Research Centre, University of Edinburgh Business School, UK

Mauro Ribeiro de Oliveira Jr.[1]

Caixa Econômica Federal, Brazil

Abstract

The aim of this paper is to propose a survival credit risk model that jointly accommodates three types of time-to-default found in bank loan portfolios. It leads to a new framework that extends the standard cure rate model introduced by Berkson & Gage [8] regarding the accommodation of zero-inflations. In other words, we propose a new survival model that takes into account three different types of individuals which have so far not been jointly accounted for: (i) an individual with an event at the starting time (zero time); (ii) non-susceptible for the event, or (iii) susceptible for the event. Considering this, the zero-inflated Weibull non-default rate regression models, which include a multinomial logistic link for the three classes, are presented using an application for credit scoring data. The parameter estimation is reached by the maximum likelihood estimation procedure and Monte Carlo simulations are carried out to assess its finite sample performance.

Keywords: `non-default rate models, portfolios, survival, zero-inflated, Weibull`

1. Introduction

More often than not, banks and financial institutions completely lose contact with their customers as soon as their loans are granted and, therefore, all the amount lent is lost. Arguably, this group of borrowers is the most expensive for the bank. Here, they are defined as straight-to-default[2] customers or STD customers for short. There is also another group of problematic customers, which are commonplace. They can no longer afford their installments, but unlike STD customers, they manage to keep up to date with their debts for a while. Most of the time, for private financial reasons, they cannot afford their debts and default on their loans. To ensure the survival of banks, fortunately, there are good customers, which are in fact most of them. They who always keep up to date with their obligations and, therefore, will not have a default record. Therefore, for this reason and mainly to maximize profits, they should try to maintain a high rate of non-defaulting loans, while the STD rates and defaulted loans should be very low.

This type of banking client, the STD customer, lives on the threshold of being a fraudster. Since we only have information on the occurrence of default, we can not infer here whether there was any type of fraud,

[1]Corresponding author: mauroexatas@gmail.com. The views and opinions expressed in this paper are solely those of the author and do not necessarily reflect the official policy or position of any of his present or past employers.

[2]The term "default", used throughout this paper, means the event of interest in credit risk analysis. It happens when clients lose the creditworthiness to meet their commitments with bank loans, for instance. The default criterion may vary from bank to bank for conservative reasons. Generally, a bank declares a default condition if a customer has not been paying any installments for more than three consecutive months. This is the definition that we have assumed in this paper.

whether it is documental fraud or internal bank fraud. Usually, it is expected that the frequency of STD customers will be greater in lines of credit where the concession process is massified and automated, i.e., the customer does not pass through the individual screening of a credit analyst. This is not the case of real estate credit lines, where a more rigorous granting process is expect. It must result in a very low occurrence of STDs. In the personal loan portfolio that we will analyze in Section 5, we have that this rate amounts to about 5% of the portfolio. The way to mitigate this rate is by identifying the most likely profiles and thus taking action to improve the lending process, either by avoiding the groups of customers with more credit risk or those more likely to commit fraud in the credit application process.

To use survival analysis techniques in credit risk settings, we must define the outcome of interest. The lifetime of a loan (event of interest) is the span time for the occurrence of the event of default. Consequently, as already mentioned, it can be said that bankers expect it to be rarely recorded within loan portfolios. This has been addressed in different papers, such as Abad *et al.* [1], Banasik *et al.* [4], Barriga *et al.* [5], Bellotti & Crook [7], Leow & Crook [20], Louzada *et al.* [24], Stepanova & Thomas [37] and Tong *et al.* [38]. The reason for the widespread use of survival analysis in credit risk rather than other modeling techniques, besides monitoring the loan portfolio credit risk over time, is that it can accommodate censored data, which is not supported. An example of this can be found in credit scoring techniques, which are purely based on good and bad client classification, see for instance Abreu [2], Hand & Henley [16], Lessmann *et al.* [21] and Louzada-Neto [25].

In the credit risk context, censoring occurs when a loan is still under repayment at the moment of data collection, i.e., it is still a good loan. In other areas, as in medicine for instance, it happens when there is no information about the event of interest, such as the patient has not experienced the recurrence of a disease or is still alive at the end of the treatment. From these cases based on clinical studies, models that accommodate cure fractions of the data events, known as cure rate models, were introduced into the literature. In Barriga *et al.* [5], the authors used a different terminology in order to clarify its use in a credit scoring setting. They denoted cure by non-default, leading to what they called by non-default rate models.

The lack of default information in credit risk setting also happens when the borrowers anticipate paying the debt before the end of the follow-up period, known as early repayment. If the default has not occurred or the loan term has been anticipated, conclusions cannot be drawn as to whether the client is a good or a bad client at the end of the follow-up period. For instance, in Figure 1, the (c) survival time equal to zero comes from STD clients; positive default times as in (a) are from defaulters. In (d), the absence of registration is due to early repayment, while in (b) the loan is still under payment at the end of the follow-up period. All loans are monitored from the granting time, therefore the initial time $t = 0$ is the start date of the loan. Generally, the follow-up period ranges from 12, 24, 36 months, or even more, depending on the loan portfolio features.

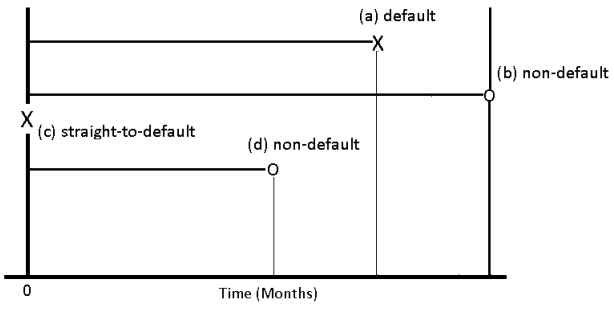

Figure 1: Loan survival time data.

To register the default, it takes at least three months of follow-up, because s at least three months without payments is needed. Therefore, it would have to occur from $t = 3$. In order to introduce the methodology based in zero-inflated data, we brought all the data to $t - 3$. As mentioned above, and explained in the next section, there are many customers who do not pay any installments of the loan, who are defined as straight-to-default clients. Furthermore, there is clear evidence that most of the loans are granted to people who

actually want to pay it back and will, therefore there are customers who can be considered immune to default. Hence, the methodology to be introduced in this paper is intended for credit risk analysis and generalizes the usual cure rate model due to Berkson & Gage [8], by taking into account a non-zero probability of failure at time zero by STD borrowers, together with the (usual) non-zero probability of surviving up to any time t. As in Barriga *et al.* [5], we maintain the notation of the non-default rate to make reference to the cured rate, leading to what we call the zero-inflated non-default rate model.

1.1. Organization

The remainder of this paper is organized as follows. In Section 2, we present a brief review of the literature and preliminary concepts related to the standard survival analysis already used to deal with the credit risk. In Section 3, we formulate our proposed model and present the approach for parameter estimation. A study based on Monte Carlo simulations using a variety of parameters is presented in Section 4. An application to a real data set of a Brazilian bank loan portfolio is presented in Section 5. Some general remarks are presented in Section 6.

2. Literature review

In order to use survival analysis techniques in credit risk settings, we first need to consider the modeling outcome of interest (event of interest) as the survival time shortly after the loan has been granted, a concept that is also known as the loans survival time. It is represented by the time span to an occurrence of an event of default. In order to perform such an approach, survival data are generally modeled by a continuous probability distribution, supported on the real non-negative interval $[0, +\infty)$.

The use of positive continuous distributions in the cure rate framework is already considered an usual modeling practice, as it can accommodate time-to-event occurrences well, which primarily contains non-negative (or censored data), see for instance Cordeiro *et al.* [15] and Ortega *et al.* [28]. However, it cannot fit an excess of zeros that may make up a time-to-default data set of loan portfolios, for example. Unlike survival data analysis, in other areas we can most commonly observe the existence of non-negative data with the presence of zeros, sometimes with an excess.

Usually, the excess of zeros occurs in count data studies, as analyzed in Barry & Welsh [6], Conceição *et al.* [13], Lambert [19] and Lord *et al.* [23]. In Ospina & Ferrari [30] and Vieira *et al.* [39], the authors dealt with zero-inflated proportion data models. Therefore, the expression "zero-inflated data" is already commonplace. In Liu *et al.* [22], the occurrence of excess zeros was investigated in two longitudinal medical follow-ups. In the first one, a SIDA study, the zero data comes from records of non-recurrence of opportunistic diseases, while in the second study, zero data are recorded as the number of non-recurrent tumors in a soft tissue sarcoma study. Zero-inflated data also appears in the context of left censored data. In Blackwood [9], for example, left censored data are generated in experiments related to the presence of toxic products in the environment. Due to the inaccuracy of the tools used for measuring, it is not always possible to fully observe some results and only a lower limit is recorded.

Furthermore, dealing with the presence of left censored data, Braekers & Grouwels [10] reviewed a laboratory experiment using mice conducted by Markel *et al.* [26], where the outcome of interest is the induced sleep time measured after injecting a dose of ethanol. As some mice present immunity for the administered dose of ethanol, the analyzed data set contains a proportion of sleep time equal to zero. In the statistical approach proposed to re-analyze the data obtained from the conducted experiment, i.e., in order to re-investigate the influence of covariates on the outcome of interest, Braekers & Grouwels [10] proposed a logistic regression model for the probability of a zero outcome value and the Cox regression model for the non-zero outcomes.

Perhaps it is unhelpful, or cruelly insensitive, if we consider human survival times equal to zero in clinical trials or medical studies. Hence, it might be why, to the best of our knowledge, we have not found a study that is willing to account for zero-inflated data in the medical specialized literature and that aims to analyze human patient survival time. However, the same sense of respect expected in clinical trials, to a certain extent, does not seem to be required when we deal with credit risk events. On the other hand, information

about zero-inflated time should be taken into account in credit risk analysis, which would be useful for identifying customers who apply for loans only for the purpose of obtaining free advantages from the bank so that once they obtain the loan they do not pay the corresponding installments from the beginning.

2.1. Preliminary

In survival analysis, the random variable T of interest is the time until the occurrence of an expected event. Depending on the context in which it appears, T might be called lifetime or failure time. In industry, it is customarily associated with the time up to failure of a machine. In the medical area, for example, it can be associated with the recurrence of a disease under treatment, or even the death of a patient. The focus of interest in credit risk setting is the failure time related to the occurrence of a loan default. Obviously, in all cases T is non-negative and is generally treated as a continuous random variable.

According to Colosimo & Giolo [12] and Rinne [34], there are several functions which completely specify the distribution of a random variable in survival analysis as they are mathematically equivalent functions. They are the Probability Density Function (PDF), the Cumulative Distribution Function (CDF), the Complementary Cumulative Distribution Function (CCDF), the hazard rate, the cumulative hazard rate and, finally, the mean residual life function. Within a survival analysis context, the Complementary Cumulative Distribution Function (CCDF) is known as the survival function and is commonly denoted by $S(\cdot)$. The downside of considering the standard survival analysis in credit risk is the mathematical fact that the survival function is a proper survival function, i.e., goes to zero as time progresses indefinitely. This means that the survival function, $S(t) = P(T > t)$, satisfies $\lim_{t\to\infty} S(t) = 0$.

Unlike what happens in many real situations, in this standard framework the presence of immunity to the effects that lead to the occurrence of the concerned event is not considered. Indeed, returning to examples in the medical field, there are patients suffering from diseases who, after undergoing treatment, completely recover. They are known as cured or long-term survivors. Similarly, in credit risk studies on loan portfolios of financial institutions, most customers never experience the condition of being delinquent (defaulter). In this financial context, they are also known as non-defaulting customers. Therefore, when the presence of cure needs to be considered, the traditional survival analysis is not at all suitable for modeling failure time. In these cases, where there is immunity to the occurrence of failures, new statistical tools are proposed. To handle the aforementioned challenge, Berkson & Gage [8] proposed a simple way that added the fraction of cured ($p > 0$) into the survival function. The authors have introduced the following survival expression based on two sub-populations of individuals susceptible and non-susceptible to the occurrence of the event of interest

$$S(t) = p + (1 - p)S_0(t), \qquad t \geq 0, \tag{1}$$

where S_0 is the baseline survival function of the individuals susceptible to failure and $p > 0$ is the proportion of the individuals immune to failure (cured). This model is called the cure rate model. Unlike S_0, S is an improper survival function as it satisfies: $\lim_{t\to\infty} S(t) = p > 0$.

Mixture cure models as presented in Berkson & Gage [8] were initially proposed in medical setting to model long-term survival in terms of two distinct sub-populations of patients according to their responses regarding the treatment against cancer. According to Othus *et al.* [31], Tong *et al.* [38], among several others authors, the advantage of the cure rate model is that it allows for associate covariates in both parts of the model. Indeed, it enable covariates to have different influences on cured patients, linking covariates with p, and on patients who are not cured, i.e., susceptible to the event, linking covariates with the parameters of the proper survival function S_0. From Tong *et al.* [38], this technique begins to be applied in credit risk modeling. According to the authors, the large proportion of costumer who are not defaulters on bank loans can be similarly modeled as the proportion of patients that is not susceptible to the event of interest in medical studies. Consequently, through the interpretation of the estimated parameters, the risk manager can now gather more information about credit risk, i.e, in addition to the standard binary analysis of *if* it may or not may happen, the mixture cure interpretation also focuses on *when* default on a loan will occur.

2.2. Proposal

To the best of our knowledge, there is no credit risk literature considering a cure rate model that accounts for the excess of individuals who have already experimented the event of interest at the beginning of the considered study, i.e., with a survival time equal to zero. Taking this into account and focusing on the portfolio credit risk context, we define the following proportions to be accommodated in our new proposed model

- p_0: the proportion of zero-inflated times, i.e., related to straight-to-default borrowers;
- p_1: the proportion of immune to failure, i.e., related to non-defaulters.

Thus, we propose the following expression for the improper survival function of a dataset comprised by all possible loan survival times

$$S(t) = p_1 + (1 - p_0 - p_1)S_0(t), \qquad t \geq 0, \tag{2}$$

where S_0 is the baseline survival function related to the $(1 - p_0 - p_1)$ proportion of subjects susceptible to failure, p_1 is the proportion of subjects immune to failure and finally, p_0 is the proportion of STD individuals. This model in (2) is called the zero-inflated non-default rate model. The important fact that differentiates the inflated non-default rate version from the standard non-default rate approach in (1), given that they share the fact that both are based on improper survival functions, is expressed in the second of the following satisfied properties: $\lim_{t\to\infty} S(t) = p_1 > 0$ and $S(0) = 1 - p_0 < 1$. Note that if $p_0 = 0$, i.e., without the excess of zeros, we have the cure rate model of Berkson & Gage [8].

2.3. Justification

In this paper, we justify the need for the zero-inflated non-default rate model based on a credit risk setting. The purpose is to deal with assessing the propensity to immediately default on a loan, in terms of estimating the rate of zero inflated data according to the available characteristics of all customers. To reach this goal, we propose jointly modeling zero-inflated time in loan survival data with a non-default rate, where we link together covariates in all parts of the proposed model. To exemplify the application of the proposed approach, we analyze a portfolio of loans made available by a large Brazilian commercial bank.

3. Model specification

In what follows, we consider the zero-inflated non-default rate model as defined in expression (2), with baseline hazard functions to be freely chosen according to the analyzed data. The associated CDF and PDF are given by

$$F(t) = p_0 + (1 - p_0 - p_1)F_0(t), \qquad t \geq 0 \tag{3}$$

and

$$f(t) = \begin{cases} p_0, & \text{if} \quad t = 0, \\ (1 - p_0 - p_1)f_0(t), & \text{if} \quad 0 < t, \end{cases} \tag{4}$$

where parameters p_0 and p_1 are defined in Section 2.2. F_0 and f_0 are, respectively, the cumulative distribution function and probability density function underpinning the $(1-p_0-p_1)$ proportion of the subject susceptible to failure.

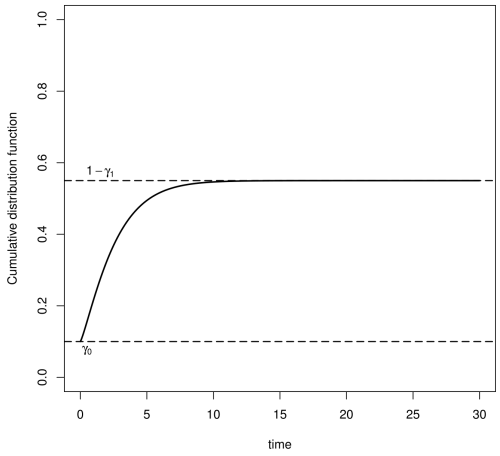

Figure 2: The CDF of the zero-inflated non-default rate model.

Note that, the CDF of the zero-inflated non-default rate model, $F(t)$, has the property of accommodating the excess of zeros, p_0, as it satisfies: $F(0) = p_0$. Moreover, it accounts for the fraction of non-defaulters, p_1, as it also satisfies: $\lim_{t\to\infty} F(t) = 1 - p_1$.

3.1. Likelihood function

The zero-inflated non-default rate model proposes to distinguish between three sub-populations of banking borrowers: a segment of those who will not honor any installment of the loan, i.e., STD borrowers with failure time zero; a segment of those are susceptible to default; and a segment of those who are not susceptible to default. Consequently, as in the standard non-default rate modeling, there are two possibilities for the customer who is not an STD customer: information about the default time (event of interest) is fully observed, i.e., the borrower defaulted while the loan was being monitored; or information about the default time is right censored, i.e., either the customer will probably become a defaulter if given enough time or he/she is actually a good customer and will never default, regardless of the monitoring period term.

Therefore, for the likelihood contribution of a survival time t_i of a customer i, we should pay attention to the fact that there are different sub-groups of customers. The likelihood contribution of each time-to-default t_i, obtained from Section 2, and all that we have considered above must assume three different values:

1. p_0, if subject i is an STD,
2. $(1 - p_0 - p_1)f_0(t_i)$, if subject i is not censored,
3. $p_1 + (1 - p_0 - p_1)S_0(t_i)$, if subject i is censored.

Let the data take the form $\mathcal{D} = \{t_i, \delta_i\}$, where $\delta_i = 1$ if t_i is an observable time to default, and $\delta_i = 0$ if it is right censored, for $i = 1, 2, \cdots n$. Let Φ denote the parameter vector associated with the f_0 baseline distribution and, finally, let (p_0, p_1) be the parameters associated, respectively, with the proportion of STD (inflation of zeros) and the proportion of non-default. The likelihood function of the zero-inflated non-default rate model, with a vector of parameters $\vartheta = (p_0, p_1, \Phi)$, is based on a sample of n observations, $\mathcal{D} = \{t_i, \delta_i\}$. Therefore, following Klein & Moeschberger [18], we write the likelihood function $L(\vartheta; \mathcal{D})$ under non-informative censoring, where we estimate the parameters by maximizing its log function.

$$L(\vartheta; \mathcal{D}) = \prod_{i:\ t_i=0} p_0 \prod_{i:\ t_i>0} \left\{ [(1 - p_0 - p_1)f_0(t_i)]^{\delta_i} [p_1 + (1 - p_0 - p_1)S_0(t_i)]^{1-\delta_i} \right\}. \tag{5}$$

3.2. The zero-inflated Weibull non-default rate model

In this section, we associate the Weibull distribution as the probability density function for the subjects susceptible to failure. We choose the Weibull function since it has been widely used to model survival data, and has also been a motivation for the proposal of various types of generalizations, see for example, Cooner *et al.* [14], Rinne [34], Rodrigues *et al.* [36], Ortega *et al.* [29] and Cancho *et al.* [11]. Then, let the Weibull distribution represent the survival behavior of the non-negative random variable T_0, which denotes the time-to-default for the susceptible subjects. The CDF of the Weibull distribution is given by

$$F_w(t) = 1 - e^{-\left(\frac{t}{\theta}\right)^{\alpha}}, \qquad t \geq 0, \tag{6}$$

where $\alpha > 0$ and $\theta > 0$ are, respectively, shape and scale parameters. The PDF of the Weibull distribution and the survival function are, respectively, obtained from (6) as

$$f_w(t) = \frac{d}{dt}F_w(t) = \frac{\alpha}{\theta}\left(\frac{t}{\theta}\right)^{\alpha-1} e^{\left(-\frac{t}{\theta}\right)^{\alpha}}, \quad S_w(t) = 1 - F_w(t) = e^{-\left(\frac{t}{\theta}\right)^{\alpha}}, \qquad t \geq 0. \tag{7}$$

The log-likelihood function for $\vartheta = (p_0, p_1, \alpha, \theta)$, corresponding to the observed data and the likelihood function as in 5, is given by

$$\begin{aligned}
\log\{L(\vartheta;\mathcal{D})\} &= \sum_{i:\ t_i=0} \log(p_0) + \sum_{i:\ t_i>0} \log\left\{[(1-p_0-p_1)f_w(t_i)]^{\delta_i}\right\} \\
&\quad + \sum_{i:\ t_i>0} \log\left\{[p_1 + (1-p_0-p_1)S_w(t_i)]^{1-\delta_i}\right\} \\
&= \sum_{i:\ t_i=0} \log(p_0) + \sum_{i:\ t_i>0} \delta_i \log(1-p_0-p_1) \\
&\quad + \sum_{i:\ t_i>0} \delta_i \log[f_w(t_i)] + \sum_{i:\ t_i>0} (1-\delta_i)\log[p_1 + (1-p_0-p_1)S_w(t_i)] \\
&= \sum_{i:\ t_i=0} \log(p_0) + \sum_{i:\ t_i>0} \delta_i \log(1-p_0-p_1) \\
&\quad + \sum_{i:\ t_i>0} \delta_i \log\left[\frac{\alpha}{\theta}\left(\frac{t_i}{\theta}\right)^{\alpha-1} e^{\left(-\frac{t_i}{\theta}\right)^{\alpha}}\right] \\
&\quad + \sum_{i:\ t_i>0} (1-\delta_i)\log[p_1 + (1-p_0-p_1)e^{-\left(\frac{t_i}{\theta}\right)^{\alpha}}]
\end{aligned}$$

The components of the score function $U(\vartheta) = U(p_0, p_1, \alpha, \theta) = \left(\frac{\partial l(\vartheta)}{\partial p_0}, \frac{\partial l(\vartheta)}{\partial p_1}, \frac{\partial l(\vartheta)}{\partial \alpha}, \frac{\partial l(\vartheta)}{\partial \theta}\right)$, are given as follows.

$$
\begin{aligned}
\frac{\partial l(\vartheta)}{\partial p_0} &= \frac{n_0}{p_0} - \frac{\sum\limits_{i:\ t_i>0} \delta_i}{(1-p_0-p_1)} - \sum_{i:\ t_i>0} \frac{(1-\delta_i)e^{-\left(\frac{t_i}{\theta}\right)^{\alpha}}}{p_1 + (1-p_0-p_1)e^{-\left(\frac{t_i}{\theta}\right)^{\alpha}}} \\
\frac{\partial l(\vartheta)}{\partial p_1} &= -\frac{\sum\limits_{i:\ t_i>0} \delta_i}{(1-p_0-p_1)} + \sum_{i:\ t_i>0} \frac{(1-\delta_i)(1-e^{-\left(\frac{t_i}{\theta}\right)^{\alpha}})}{p_1 + (1-p_0-p_1)e^{-\left(\frac{t_i}{\theta}\right)^{\alpha}}} \\
\frac{\partial l(\vartheta)}{\partial \alpha} &= \sum_{i:\ t_i>0} \delta_i \left[\frac{1}{\alpha} + \left(-\frac{t_i}{\theta}\right)^{\alpha} \log\left(-\frac{t_i}{\theta}\right) + \left(\frac{t_i}{\theta}\right)\right] \\
&\quad - \sum_{i:\ t_i>0} (1-\delta_i) \frac{(1-p_0-p_1)\left(\frac{t_i}{\theta}\right)^{\alpha} \log\left(\frac{t_i}{\theta}\right)}{1-p_0-p_1+p_1 e^{\left(\frac{t_i}{\theta}\right)^{\alpha}}} \\
\frac{\partial l(\vartheta)}{\partial \theta} &= \sum_{i:\ t_i>0} \delta_i \left[-\frac{\alpha}{\theta} - \frac{\alpha\left(-\frac{t_i}{\theta}\right)^{\alpha}}{\theta}\right] \\
&\quad + \sum_{i:\ t_i>0} (1-\delta_i) \left[\frac{\alpha(1-p_0-p_1)\left(\frac{t_i}{\theta}\right)^{\alpha}}{\theta\left(1-p_0-p_1+p_1 e^{\left(\frac{t_i}{\theta}\right)^{\alpha}}\right)}\right]
\end{aligned}
$$

The maximum likelihood estimates $\hat{\vartheta} = (\hat{p_0}, \hat{p_1}, \hat{\alpha}, \hat{\theta})$ can be obtained by solving the non-linear system of equations $U(\vartheta) = \dfrac{\partial l(\vartheta)}{\partial \vartheta} = 0$. We use the free statistical software R to solve them numerically using iterative techniques, such as the Newton-Raphson algorithm. The computational code is available from the authors upon request.

Following Migon *et al.* [27] and Ospina & Ferrari [30], large sample inference for the parameters is based on the matrix of second derivatives of the log likelihood by using the observed information matrix, $\mathbf{I}(\vartheta) = \{-\partial^2 \ell(\vartheta)/\partial\vartheta\partial\vartheta^T\}^{-1}$, evaluated at $\vartheta = \hat{\vartheta}$. The approximate $(1-\alpha)$ 100% confidence intervals for the parameters p_0, p_1, α, θ are given by $\widehat{p}_0 \pm \xi_{\alpha/2}\sqrt{Var(\widehat{p}_0)}$, $\widehat{p}_1 \pm \xi_{\alpha/2}\sqrt{Var(\widehat{p}_1)}$, $\widehat{\alpha} \pm \xi_{\alpha/2}\sqrt{Var(\widehat{\alpha})}$ and $\widehat{\theta} \pm \xi_{\alpha/2}\sqrt{Var(\widehat{\theta})}$, where $\xi_{\alpha/2}$ is the upper $\alpha/2$ percentile of the standard Normal distribution.

In the application section, we compare the proposed model configured with different covariates. A comparison of the models was made using the selection criterion known as the Akaike Information Criterion (AIC), proposed by Akaike Akaike [3]. The criterion is defined by AIC $= -2\log(L) + 2k$, where k is the number of estimated parameters, n the sample size and L is the maximised value of the likelihood function. The model with the smallest value is chosen as the preferred for describing a given dataset among all models considered.

3.3. The regression model

Here, we introduce a way to link the covariates with the parameters set in the zero-inflated Weibull non-default rate model. This modeling can determine the effect of the covariates all at once on the zero-inflated times on the non-default rate and on the failure times. Therefore, we propose to relate the set of four parameters $\{p_0, p_1, \alpha, \theta\}$, respectively, the proportion of zeros, the proportion of non-defaulters, the scale and shape parameters of the Weibull distribution, with a set of four-covariate vectors $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$. These covariate vectors, as occurs in practice, may be the same, i.e., $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}_3 = \mathbf{x}_4$.

The regression version of the zero-inflated Weibull non-default rate model is defined by (2) and by the

following systematic components:

$$\begin{cases} H(p_{0i}, p_{1i}) &= (\zeta_{0i}, \zeta_{1i}), \\ g_1(\alpha_i) &= \eta_{1i}, \\ g_2(\theta_i) &= \eta_{2i}, \end{cases} \tag{8}$$

where $\zeta_{0i} = \mathbf{x}_{1i}^\top \boldsymbol{\beta}_1$, $\zeta_{1i} = \mathbf{x}_{2i}^\top \boldsymbol{\beta}_2$, $\eta_{1i} = \mathbf{x}_{3i}^\top \boldsymbol{\beta}_3$ and $\eta_{2i} = \mathbf{x}_{4i}^\top \boldsymbol{\beta}_4$ are linear predictors, and $\boldsymbol{\beta}_j$'s are four vectors of unknown regression coefficients to be estimated. The link function H, g_1 and g_2 provide the relationship between the linear predictor and the parameters of the distribution function. Following the setting made in Pereira *et al.* [32], H is set as the multinomial logistic regression [17, p. 261], i.e., $H(p_{0i}, p_{1i}) = \left(\log\left(\frac{p_{0i}}{1-p_{0i}-p_{1i}}\right), \log\left(\frac{p_{1i}}{1-p_{0i}-p_{1i}}\right)\right)$. Since $\alpha > 0$ and $\theta > 0$, the g_1 and g_2 link functions are chosen as $g_1(\alpha_i) = \log(\alpha_i)$ and $g_2(\theta_i) = \log(\theta_i)$. Therefore, $\alpha_i = e^{\mathbf{x}_{3i}^\top \boldsymbol{\beta}_3}$ and $\theta_i = e^{\mathbf{x}_{4i}^\top \boldsymbol{\beta}_4}$. These are the most convenient link functions because $g_1(\cdot)$ and $g_2(\cdot)$ are strictly monotonic link functions and twice differentiable that map $\mathbb{R}^+$ into $\mathbb{R}$.

Note that, as required, the component link function H ensures that $0 < p_{0i} < 1$, $0 < p_{1i} < 1$ and $0 < 1 - p_{0i} - p_{1i} < 1$ hold. Indeed, it is always satisfied since $(p_{0i}, p_{1i}) = \left(\frac{e^{\mathbf{x}_{1i}^\top \boldsymbol{\beta}_1}}{1+e^{\mathbf{x}_{1i}^\top \boldsymbol{\beta}_1}+e^{\mathbf{x}_{2i}^\top \boldsymbol{\beta}_2}}, \frac{e^{\mathbf{x}_{2i}^\top \boldsymbol{\beta}_2}}{1+e^{\mathbf{x}_{1i}^\top \boldsymbol{\beta}_1}+e^{\mathbf{x}_{2i}^\top \boldsymbol{\beta}_2}}\right)$. In addition, H is a bijective link function and twice differentiable that maps C into $\mathbb{R}^2$, where C is a subspace of $\mathbb{R}^2$ defined as $C = \{(p_{0i}, p_{1i}) | 0 < p_{0i} < 1, 0 < p_{1i} < 1 - p_{0i}\}$ [32, p. 128].

Following Migon *et al.* [27], Ospina & Ferrari [30], as aforementioned, approximate $(1-\alpha)$ 100% confidence intervals for the regression vector parameters, $\boldsymbol{\beta}_j$, presented in the simulation studies and in the application section, are given by $\widehat{\boldsymbol{\beta}}_j \pm \xi_{\alpha/2}\sqrt{Var(\widehat{\boldsymbol{\beta}}_j)}$, where $\xi_{\alpha/2}$ is the upper $\alpha/2$ percentile of standard Normal distribution and $j = 1, 2, 3, 4$. Note that, in the regression model version, the components of the score function changes to $U(\vartheta) = U(\boldsymbol{\beta}_j)$, once we are now maximizing against the betas.

There is no feasible analytical expression for the score function $U(\vartheta) = U(\beta_1, \beta_2, \beta_3, \beta_4)$, therefore, numerical maximization of the log-likelihood function $\log\{L(\vartheta; \mathcal{D})\}$ is accomplished by using existing software. There are various routines available for numerical maximization. We chose the routine `optim` in the R software for numerical maximization, see the manual [33] for `optim` for details. In the application section 5, as well as in the simulation section 4, the method of maximization was chosen to be "BFGS" as we did not face numerical problems such as lack of convergence by using this. We also emphasize that in section 5 different initial points were considered in the maximization algorithm and we always obtained similar results. The computational code is available from the authors upon request.

4. Simulation studies

We proceed a parameter estimation based on a maximum likelihood principle and use the method of maximization "BFGS" of the R routine optim() for that. In order to check if the maximum likelihood estimator is well-behaved and its convergence rates, we performed a simulation study to examine the coverage probabilities of the 95% confidence intervals for the MLEs. The simulation study also provides the results for bias and root mean square errors for the estimated parameters to ensure that they decrease as expected with increasing sample sizes.

(1) that the maximum likelihood estimator is well-behaved, and (2) its convergence rates.

The simulation study is based on 1000 sample replications, where the sample size increases according to the nature of the real data sets in which the model has been applied. . Therefore, we perform Monte Carlo simulations where the sample size varies as $n = 100$, 250, 500, 750 and 1000. Three simulation studies are performed for the proposed zero-inflated Weibull non-default rate regression model. For the purpose of the simulation, we let x be a random variable that represents a consumer characteristic. The description of the sample generation, i.e., all details of the simulated survival time distribution and results obtained regarding the proposed estimation method are described in the next sections.

The model parameters are linked to a single covariate $\boldsymbol{x}$, according to the following expression: $p_{0i} = \frac{e^{\boldsymbol{\beta}_{10}+\boldsymbol{x}_i\boldsymbol{\beta}_{11}}}{1+e^{\boldsymbol{\beta}_{10}+\boldsymbol{x}_i\boldsymbol{\beta}_{11}}+e^{\boldsymbol{\beta}_{20}+\boldsymbol{x}_i\boldsymbol{\beta}_{21}}}$, $p_{1i} = \frac{e^{\boldsymbol{\beta}_{20}+\boldsymbol{x}_i\boldsymbol{\beta}_{21}}}{1+e^{\boldsymbol{\beta}_{10}+\boldsymbol{x}_i\boldsymbol{\beta}_{11}}+e^{\boldsymbol{\beta}_{20}+\boldsymbol{x}_i\boldsymbol{\beta}_{21}}}$, $\alpha_i = e^{\boldsymbol{\beta}_{30}+\boldsymbol{x}_i\boldsymbol{\beta}_{31}}$, $\theta_i = e^{\boldsymbol{\beta}_{40}+\boldsymbol{x}_i\boldsymbol{\beta}_{41}}$. Considering the

parameters established in the regression model defined above, we set three different scenarios of parameters for the simulation studies performed here. Playing the role of covariate, we assume $\boldsymbol{x}$ as a binary covariate with values drawn from a Bernoulli distribution with parameter 0.5.

For scenario 1, $\boldsymbol{\beta}_{10}$ assumes -3 and $\boldsymbol{\beta}_{11}$ assumes 1. $\boldsymbol{\beta}_{20}$ assumes -2.5 and $\boldsymbol{\beta}_{21}$ assumes 0.3. Given that the average value of $\boldsymbol{x}$ is 0.5, we have that p_0 assumes on average a value of 0.0697, and p_1 assumes on average a value of 0.0809. Compared to the other scenarios 2 and 3, scenario 1 has the characteristic of having a **low rate of STD and non-default**, respectively, 6.97% and 8.09%. Regarding the Weibull parameters, $\boldsymbol{\beta}_{30}$ assumes 0.5, $\boldsymbol{\beta}_{31}$ assumes 0.5, $\boldsymbol{\beta}_{40}$ assumes 1.5 and $\boldsymbol{\beta}_{41}$ assumes 2. Given that the average value of $\boldsymbol{x}$ is 0.5, this implies that the Weibull parameters α and θ on average are, respectively, equal to 2.11 and 12.18.

For scenario 2, $\boldsymbol{\beta}_{10}$ assumes -2 and $\boldsymbol{\beta}_{11}$ assumes 2. $\boldsymbol{\beta}_{20}$ assumes -1.5 and $\boldsymbol{\beta}_{21}$ assumes 1.5. Given that the average value of $\boldsymbol{x}$ is 0.5, we have that p_0 assumes on average a value of 0.1999, and p_1 assumes on average a value of 0.2566. Compared to the other scenarios 1 and 3, scenario 2 has the characteristic of having a **moderate rate of STD and non-default**, respectively, 19.99% and 25.66%. Regarding the Weibull parameters, $\boldsymbol{\beta}_{30}$ assumes -0.5, $\boldsymbol{\beta}_{31}$ assumes 1.5, $\boldsymbol{\beta}_{40}$ assumes -0.5 and $\boldsymbol{\beta}_{41}$ assumes 3. This implies the Weibull parameters α and θ on average are, respectively, equal to 1.28 and 2.71.

Finally, for scenario 3, $\boldsymbol{\beta}_{10}$ assumes -0.5 and $\boldsymbol{\beta}_{11}$ assumes 0.75. $\boldsymbol{\beta}_{20}$ assumes -0.35 and $\boldsymbol{\beta}_{21}$ assumes 1.75. Given that the average value of $\boldsymbol{x}$ is 0.5, we have that p_0 assumes on average a value of 0.2469, and p_1 assumes on average a value of 0.6832. Compared to the other scenarios 1 and 2, scenario 3 has the characteristic of having a **high rate of STD and non-default**, respectively, 24.69% and 68.32%. Regarding the Weibull parameters, $\boldsymbol{\beta}_{30}$ assumes -1, $\boldsymbol{\beta}_{31}$ assumes 2, $\boldsymbol{\beta}_{40}$ assumes 1.25 and $\boldsymbol{\beta}_{41}$ assumes 3.5. This implies that the Weibull parameters α and θ on average are, respectively, equal to 1 and 20.08.

4.1. Simulation algorithm

Suppose that the time of occurrence of an event of interest has the cumulative distribution function $F(t)$ given by (3), i.e.: $F(t) = p_0 + (1 - p_0 - p_1)F_0(t), \qquad t \geq 0.$

We aim to simulate random samples of size n posing as loan survival times, where each sample comprises a proportion p_0 of zero-inflated times, a non-default fraction of p_1 and with a proportion $(1 - p_0 - p_1)$ of failure times drawn from a Weibull distribution with α and θ parameters. The following step-by-step algorithm is proposed for this purpose, which is based on the link functions (8), with an $\boldsymbol{x}$ covariate drawn from a Bernoulli distribution with parameter 0.5, representing a consumer feature.

1. Set $\boldsymbol{\beta}_{10}$ and $\boldsymbol{\beta}_{11}$ related to the value of the desired proportion of zero-inflated times, p_0, along with $\boldsymbol{\beta}_{20}$ and $\boldsymbol{\beta}_{21}$ related to the value of the desired non-default fraction, p_1; finally, set the Weibull parameters $\boldsymbol{\beta}_{30}$ and $\boldsymbol{\beta}_{31}$ related to α, $\boldsymbol{\beta}_{40}$ and $\boldsymbol{\beta}_{41}$ related to θ;
2. Draw $\boldsymbol{x}_i$ from $\boldsymbol{x} \sim$ Bernoulli(0.5) and calculate p_{0i}, p_{1i}, α and θ_i;
3. Generate u_i from a uniform distribution U(0,1);
4. If $u_i \leq p_{0i}$, set $s_i = 0$;
5. If $u_i > 1 - p_{1i}$, set $s_i = \infty$;
6. If $p_{0i} < u_i \leq 1 - p_{1i}$, generate v_i from a uniform distribution $\mathrm{U}(p_{0i}, 1 - p_{1i})$ and take s_i as the root of $F(s_i) - v_i = 0$, where $F(\cdot)$ is given as in (3);
7. Generate w_i from a uniform $\mathrm{U}(0, max(s_i))$, considering only finites s_i;
8. Calculate $t_i = min(s_i, w_i)$, if $t_i < w_i$, set $\delta_i = 1$, otherwise, set $\delta_i = 0$.
9. Repeat as necessary from step 2 until you obtain the desired amount of sample (t_i, δ_i).

Note that the censoring distribution chosen is a uniform distribution with a limited range in order to keep the censoring rates reasonable, see Rocha *et al.* [35, p. 12].

4.2. Results of Monte Carlo simulations

Figures 3, 4 and 5 describe the simulation results for the three simulated scenarios of parameters, where the sample size varies as n = 100, 250, 500, 750 and 1000. For the purpose of the simulation, we let $\boldsymbol{x}$

be a random variable that represents a consumer characteristic. Hence, the link configuration of the eight parameters ($\boldsymbol{\beta}_{10}$, $\boldsymbol{\beta}_{11}$, $\boldsymbol{\beta}_{20}$, $\boldsymbol{\beta}_{21}$, $\boldsymbol{\beta}_{30}$, $\boldsymbol{\beta}_{31}$, $\boldsymbol{\beta}_{40}$, $\boldsymbol{\beta}_{41}$) to be estimated is given by the following expressions:

$$\begin{array}{lcl} p_{0i} & = & \frac{e^{\boldsymbol{\beta}_{10}+\boldsymbol{x}_i\boldsymbol{\beta}_{11}}}{1+e^{\boldsymbol{\beta}_{10}+\boldsymbol{x}_i\boldsymbol{\beta}_{11}}+e^{\boldsymbol{\beta}_{20}+\boldsymbol{x}_i\boldsymbol{\beta}_{21}}}, \\ p_{1i} & = & \frac{e^{\boldsymbol{\beta}_{20}+\boldsymbol{x}_i\boldsymbol{\beta}_{21}}}{1+e^{\boldsymbol{\beta}_{10}+\boldsymbol{x}_i\boldsymbol{\beta}_{11}}+e^{\boldsymbol{\beta}_{20}+\boldsymbol{x}_i\boldsymbol{\beta}_{21}}}, \\ \alpha_i & = & e^{\boldsymbol{\beta}_{30}+\boldsymbol{x}_i\boldsymbol{\beta}_{31}}, \\ \theta_i & = & e^{\boldsymbol{\beta}_{40}+\boldsymbol{x}_i\boldsymbol{\beta}_{41}}. \end{array} \tag{9}$$

The parameter values are selected in order to assess the ML estimation performance under different shape and scale parameters ($\boldsymbol{\beta}_{30}$, $\boldsymbol{\beta}_{31}$, $\boldsymbol{\beta}_{40}$ and $\boldsymbol{\beta}_{41}$, related to the Weibull time-to-default distribution), and also under a composition of different proportions of zero-inflated data ($\boldsymbol{\beta}_{10}$ and $\boldsymbol{\beta}_{11}$) and non-defaulters rates ($\boldsymbol{\beta}_{20}$ and $\boldsymbol{\beta}_{21}$ related to censored data). It can be seen from Figures 3 to 5 that:

1. in general, the maximum likelihood estimation on average, MLEA, is close to the parameters set in the simulated parameter scenarios, see Figure 5. However, in scenarios 1 and 2, the parameters $\widehat{\boldsymbol{\beta}}_{11}$ and $\widehat{\boldsymbol{\beta}}_{21}$ need a larger sample size (from at least n=500 for $\boldsymbol{\beta}_{21}$) to achieve convergence.
2. in general, according to Figures 4 and 5, biases and root mean square errors decrease as the sample size increases; we also observe that, in general, the coverage probability, i.e., the proportion of the time that the interval contains the true value of interest, is close to 95%, as expected;
3. in the scenarios with the greatest presence of non-default and zeros, i.e., scenario 2 (Moderate) and 3 (High), the MLEA, and the measures of RMSE, Bias and CP of the estimated regression parameters related to p_0 and p_1, performs better compared to scenario 1 (Low), of course, due to a greater presence of zeros and censored data;
4. on the other hand, in the scenario with less presence of zeros and non-default and , i.e., scenario 1 (Low), the MLEA, and the measures of RMSE, Bias and CP of the estimated regression parameters related to α and θ, perform better compared to other scenarios due to the greater presence of observed time-to-default data.

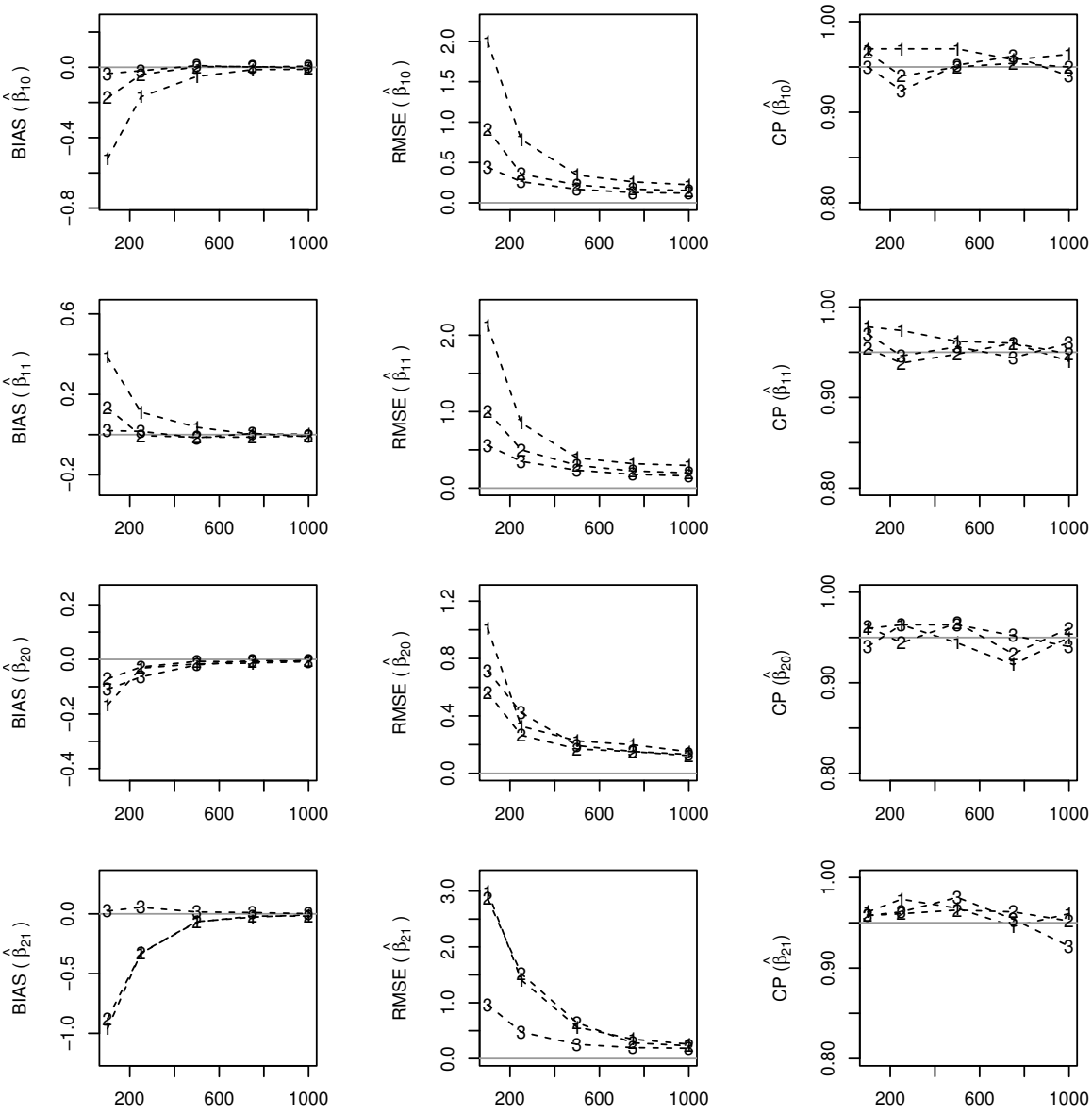

Figure 3: Bias, square root of mean squared error and coverage probability (CP) of **the maximum likelihood estimation** ($\widehat{\boldsymbol{\beta}}_{10}$, $\widehat{\boldsymbol{\beta}}_{11}$, $\widehat{\boldsymbol{\beta}}_{20}$, $\widehat{\boldsymbol{\beta}}_{21}$) of zero-inflated Weibull non-default rate regression model for simulated data under the three scenarios of parameters (the numbers 1, 2 and 3 on the graphs refer to the three scenarios), obtained from Monte Carlo simulations with 1000 replications and an increasing sample size (n).

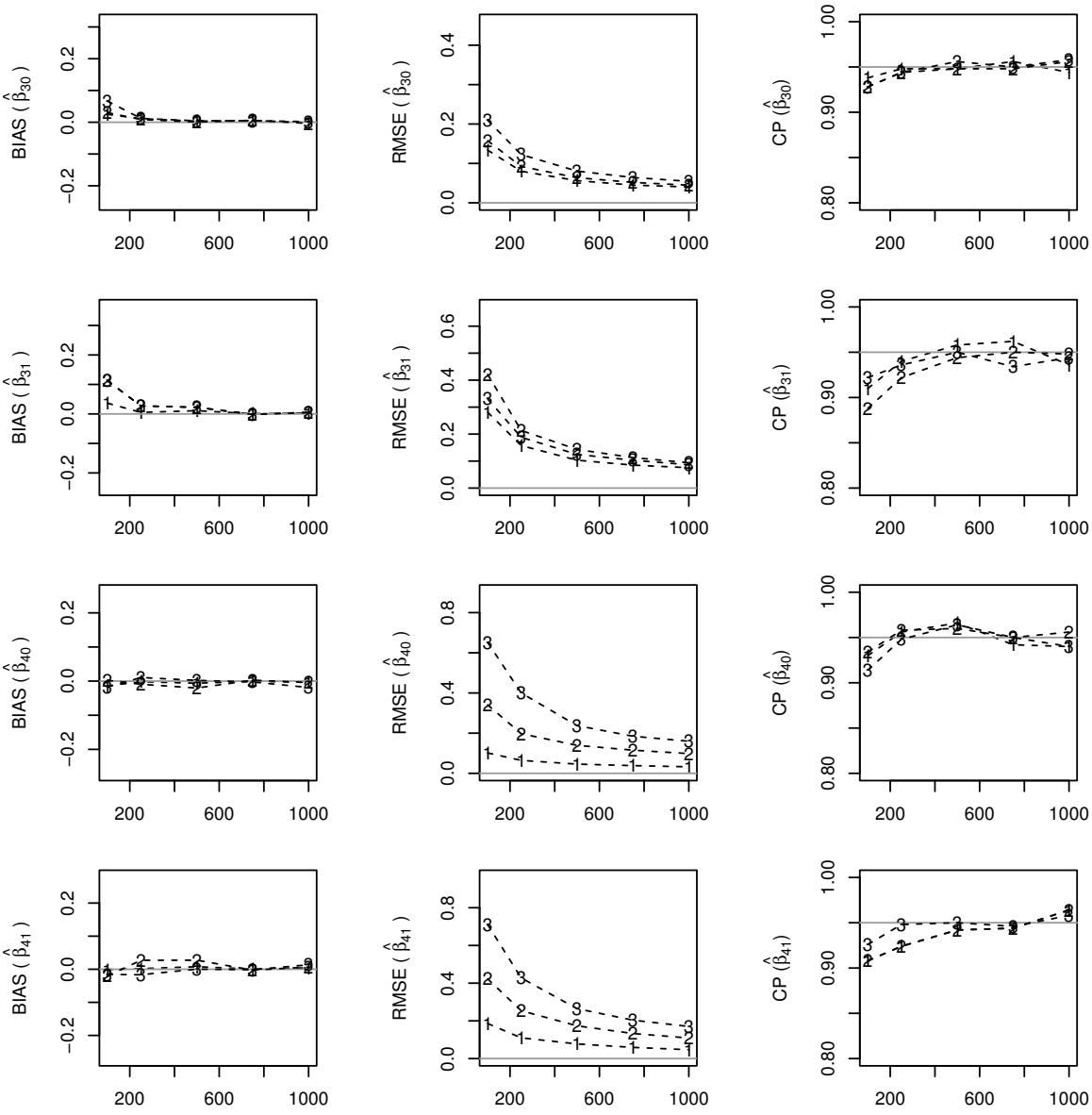

Figure 4: Bias, square root of mean squared error and coverage probability (CP) of **the maximum likelihood estimation** ($\widehat{\boldsymbol{\beta}}_{30}$, $\widehat{\boldsymbol{\beta}}_{31}$, $\widehat{\boldsymbol{\beta}}_{40}$, $\widehat{\boldsymbol{\beta}}_{41}$) of zero-inflated Weibull non-default rate regression model for simulated data under the three scenarios of parameters (the numbers 1, 2 and 3 on the graphs refer to the three scenarios), obtained from Monte Carlo simulations with 1000 replications and an increasing sample size (n).

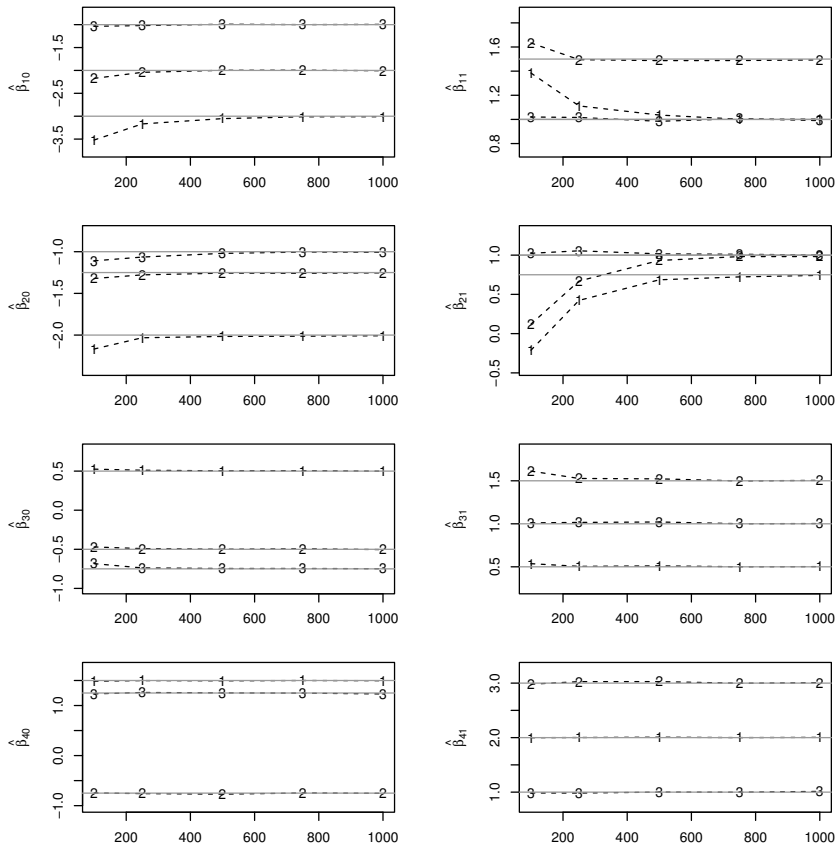

Figure 5: MLEA, **maximum likelihood estimation** on average of the parameters ($\widehat{\boldsymbol{\beta}}_{10}$, $\widehat{\boldsymbol{\beta}}_{11}$, $\widehat{\boldsymbol{\beta}}_{20}$, $\widehat{\boldsymbol{\beta}}_{21}$, $\widehat{\boldsymbol{\beta}}_{30}$, $\widehat{\boldsymbol{\beta}}_{31}$, $\widehat{\boldsymbol{\beta}}_{40}$), $\widehat{\boldsymbol{\beta}}_{41}$ of zero-inflated Weibull non-default rate regression model for simulated data under the three scenarios of parameters (the numbers 1, 2 and 3 on the graphs refer to the three scenarios), obtained from Monte Carlo simulations with 1000 replications and an increasing sample size (n).

5. Application: Brazilian bank loan portfolio

In this section, we present an application of the proposed model in a database made available by one of the largest Brazilian banks. Our objective is to assess if customer characteristics are associated with consumer propensity of being STD, defaulter or non-defaulter customers. It is important to note once more that the presented data set, amounts, rates and levels of the available covariates do not necessarily represent the actual condition of the financial institution's customer database. That is, despite being a real database, the bank may have sampled the data in order to change the current status of its loan portfolio.

The portfolio was collected from customers who took out a personal loan over a 60-month period from 2010 to 2015. Table 1 shows the customer's quantitative frequencies of the loan portfolio provided by the bank. It consists of 5733 accounts with its recorded time-to-default (in months), with an approximate 80% rate of censored data, i.e., a high rate of non-default loans. In order to proceed the model fit, we considered dummy covariates for all levels of the available covariates. Therefore, including all the intercepts, we might have up to thirty two (32=4x4x2) regression parameters to be estimated.

	Number of customers	Number of STD	Number of defaulters	Number of censored
Total	5733	321 (5.60%)	810 (14.13%)	4602 (80.27%)

Table 1: Frequency and percentage of the bank loan lifetime data.

The segmentations of customers of the bank was made a priori by the bank. For example, the age group 1 means that customers have been grouped by age from a specified range (determined by the bank). Moreover, the classification of the type of residence and type of employment has not been supplied to our study by confidentiality issues. Table 2 shows the quantitative frequency according to the available covariates.

Covariate	Quantity of customers
Age group 1	503
Age group 2	3088
Age group 3	1220
Age group 4	922
Type of residence 1	629
Type of residence 2	4056
Type of residence 3	998
Type of residence 4	50
Type of employment 1	956
Type of employment 2	4777

Table 2: Quantity of the available covariates.

Figure 6 presents a graphical summary of the survival behavior present in the available covariates: age group, type of residence and type of employment. The histogram shows only the distribution of the observed data, while the censored data is better observed through the KM curves. Notwithstanding, we can see the presence of zero-inflated data in both. We can see from the stratified Kaplan-Meier survival curves that the age group identified as 4 presents lower presence of zero-inflated time (STD borrowers) compared to the others. The group with type of residence 4 shows a higher presence of zero-inflated time (STD borrowers) compared to the borrowers with other type of residence. Type of employment 2 shows clearly a high non-default rate, besides that, it also presents a lower rate of zero-inflated times.

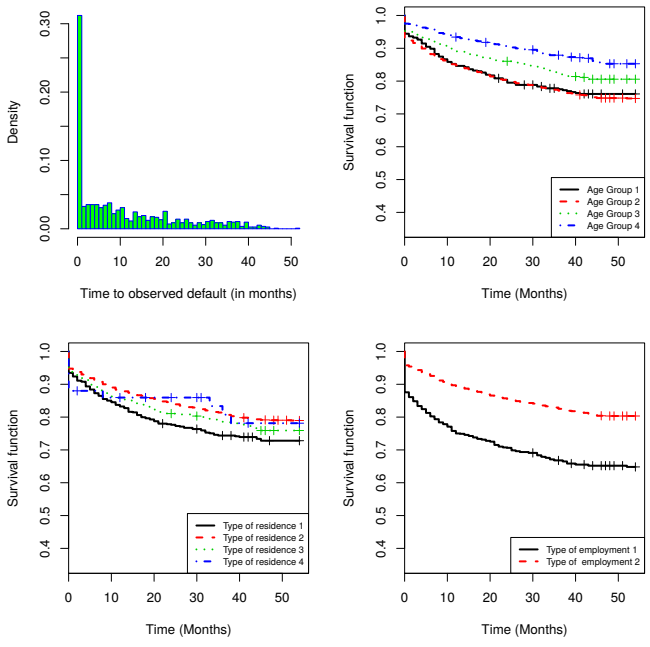

Figure 6: Brazilian bank loan portfolio data. The graphs at the top show histograms for the observed time-to-default variable of interest (left) and Kaplan-Meier survival curves stratified by age group (right). The graphs at the bottom show Kaplan-Meier survival curves stratified by type of residence (left) and Kaplan-Meier survival curves stratified by type of employment (right).

Henceforth, we are concerned about whether the use of covariates explains the distribution of the time-to-default better than assuming that the observations are identically distributed. The fitted model without any covariate has AIC of 12768.71 ($l\{\hat{p_0}, \hat{p_1}, \hat{\alpha}, \hat{\theta}\} = -6380.355,\ p = 4$) and the model with all the dummy covariates has AIC of 12809.21 ($l\{\hat{p_0}, \hat{p_1}, \hat{\alpha}, \hat{\theta}\} = -6372.604,\ p = 32$). To reach the final model, we consider a five-steps backward elimination way to find the best predictors to compose the final model: 1) select a significance level to stay in the model (e.g SL=0.05); 2) fit the full model with all possible predictors; 3) consider the predictor with the highest P-value. If P > SL, go to step 4, otherwise it is finished; 4) remote the predictor; 5) fit model without this variable; go back to step 4 until finished. The final model is summarized in Table 3, which has AIC of 12602.52 ($l\{\hat{\boldsymbol{\beta}}_{10}, \hat{\boldsymbol{\beta}}_{11}, \hat{\boldsymbol{\beta}}_{12}, \hat{\boldsymbol{\beta}}_{13}, \hat{\boldsymbol{\beta}}_{20}, \hat{\boldsymbol{\beta}}_{21}, \hat{\boldsymbol{\beta}}_{22}, \hat{\boldsymbol{\beta}}_{30}, \hat{\boldsymbol{\beta}}_{40}, \hat{\boldsymbol{\beta}}_{41}\} = -6291.259,\ p = 10$).

Parameter	Dummy covariate (param$_{(1)}$)	Estimate	S.E.$_{(2)}$	P-value	Exp.$_{(3)}$
p_0	Intercept ($\boldsymbol{\beta}_{10}$)	-0.6724	0.1208	0.0000	0.5105
	Age group =4 ($\boldsymbol{\beta}_{11}$)	-0.8207	0.2284	0.0003	0.4401
	Type of residence =4 ($\boldsymbol{\beta}_{12}$)	0.8722	0.4751	0.0664	2.3922
	Type of employment=2 ($\boldsymbol{\beta}_{13}$)	-0.6356	0.1431	0.0000	0.5296
p_1	Intercept ($\boldsymbol{\beta}_{20}$)	0.9155	0.0972	0.0000	2.4980
	Type of residence =1 ($\boldsymbol{\beta}_{21}$)	-0.2810	0.1056	0.0078	0.7550
	Type of employment=2 ($\boldsymbol{\beta}_{22}$)	0.6402	0.0998	0.0000	1.8969
α	Intercept ($\boldsymbol{\beta}_{30}$)	0.1507	0.0374	0.0001	1.1626
θ	Intercept ($\boldsymbol{\beta}_{40}$)	3.0967	0.0628	0.0000	22.1248
	Age group =4 ($\boldsymbol{\beta}_{41}$)	0.7039	0.1446	0.0000	2.0216

Table 3: The Zero-Inflated Non-default Regression Model for time-do-default in a Brazilian Bank Loan Portfolio. Notes: (1) Related regression parameter estimated; (2) Standard error; (3) Exp(estimated parameter).

Based on the last column of Table 3, estimates of the relationship among covariates and the time-to-default event of interest, already presented in the graphical analysis (see the K-M survival curves in Figure 6), can be ratified again. For example, the odds of being an STD customer within the age group equal to 4 decrease by 56%, compared to the remaining group. On the order hand, as expected, the group of customers with type of employment 2 shows a 89% higher odds to be non-default customer on a loan. Two dummy covariates related to the covariate type of residence showed to be significant. Type of residence 1 decreases the odds of non-default on the loan by 22.5%, while the group within the type of residence 4 increases the odds of being an STD customer by 139%, with all other independent covariates held constant.

The selected dummy covariates (Table 3) enabled us to split the portfolio between twelve (12) different groups of borrowers (segmentations). In the Figure 7, we present the estimated survival curves (the dotted lines), among with the Kaplan-Meier survival curves, of the most representative group of borrowers (5544 out of 5733), considering the following segmentation: **segmentation 1** comprises 777 borrowers with the following set of attributes: age group equal to 4, type of residence equal to 2 or 3 and type of employment equal to 2, we have that $\hat{p}_0 = 0.02$ and $\hat{p}_1 = 0.80$; **segmentation 2** comprises 470 borrowers with the following set of attributes: age group not equal to 4, type of residence equal to 1 and type of employment equal to 2, we have that $\hat{p}_0 = 0.05$ and $\hat{p}_1 = 0.73$; **segmentation 3** comprises 108 borrowers with the following set of attributes: age group not equal to 4, type of residence equal to 1 and type of employment equal to 1, we have that $\hat{p}_0 = 0.15$ and $\hat{p}_1 = 0.55$; **segmentation 4** comprises 3444 borrowers with the following set of attributes: age group not equal to 4, type of residence equal to 2 or 3 and type of employment equal to 2, we have that $\hat{p}_0 = 0.04$ and $\hat{p}_1 = 0.78$; and, finally, **segmentation 5** comprises 745 borrowers with the following set of attributes: age group not equal to 4, type of residence equal to 2 or 3 and type of employment equal to 1, we have that $\hat{p}_0 = 0.12$ and $\hat{p}_1 = 0.62$.

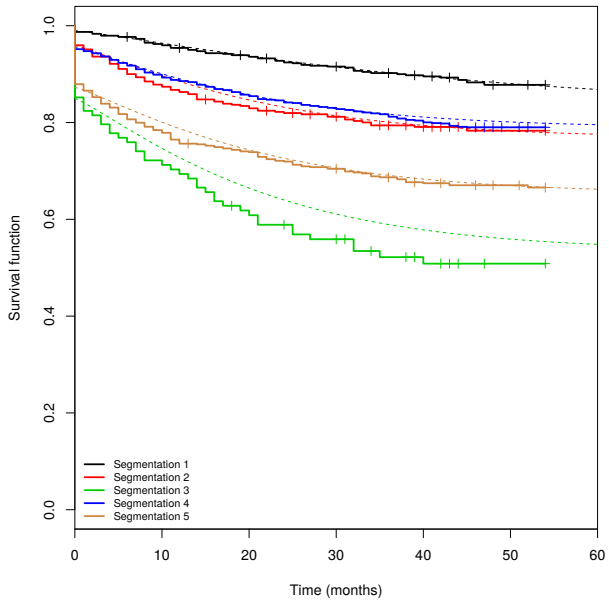

Figure 7: Brazilian bank loan portfolio. Kaplan-Meier survival curves stratified through the covariate selection given by the final model presented in Table 3 and the estimated survival curves (dotted lines)

6. Conclusion

We presented a methodology in which we modify the standard cure rate model introduced by Berkson & Gage [8] to a credit risk setting. It enabled us to estimate the proportions of the following loan applicants in a given portfolio: straight-to-default customers, defaulters and non-defaulters. At the heart of our methodology, the survival function is adapted to account for the excess of zeros, which represents the rate of borrowers that do not account for even the first installments and default on the loan at the beginning. An advantage of our approach is to accommodate zero-inflated times, which is not possible in the standard cure rate model. In this scenario, information from all borrowers can be exploited through the joint modeling of their survival times, even from those who are equal to zero. To illustrate the proposed method, data comprised for loan survival times of a Brazilian bank loan portfolio is modeled. The estimation procedure proposed for the zero-inflated Weibull non-default rate model and the obtained outcomes proved to be satisfactory.

The challenge that we may face using regression models lies in the fact that sometimes we cannot have a set of factors, or covariates, sufficient to explain the risk of default of the portfolio at a very granular level of customers. Furthermore, it is not unusual that the application of regression models can be impaired by the little data available for the study. We believe that in our case, despite the very small number of available covariates, we obtained very useful results. Moreover, we think that if more covariates had been provided by the bank, it could have greatly enriched our model application.

Finally, we pointed out the importance of the joint analysis of zero inflation data with the fraction of non-default, which is the most common scenario for bank portfolios: it can provide credit risk analyst information over the most costly applicants, who are those who are more likely to miss their payments at the beginning of the relationship with the bank.

Acknowledgment

The research was sponsored by CAPES - Process number: BEX 10583/14-9, Brazil.

References

[1] Abad, R. C., Fernández, J. M. V. & Rivera, A. D. (2009). Modelling consumer credit risk via survival analysis. *SORT: Statistics and Operations Research Transactions*, **33**(1), 3–30.

[2] Abreu, H. (2004). Aplicação da análise de sobrevivência em um problema de credit scoring e comparação com a regressão logística. *Biblioteca Digital de Teses e Dissertações da Universidade Federal de São Carlos*.

[3] Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**(6), 716–723.

[4] Banasik, J., Crook, J. N. & Thomas, L. C. (1999). Not if but when will borrowers default. *Journal of the Operational Research Society*, **50**(12), 1185–1190.

[5] Barriga, G. D., Cancho, V. G. & Louzada, F. (2015). A non-default rate regression model for credit scoring. *Applied Stochastic Models in Business and Industry*, **31**(6), 846–861.

[6] Barry, S. C. & Welsh, A. H. (2002). Generalized additive modelling and zero inflated count data. *Ecological Modelling*, **157**(2), 179–188.

[7] Bellotti, T. & Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, **60**(12), 1699–1707.

[8] Berkson, J. & Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**(259), 501–515.

[9] Blackwood, L. G. (1991). Analyzing censored environmental data using survival analysis: single sample techniques. *Environmental monitoring and assessment*, **18**(1), 25–40.

[10] Braekers, R. & Grouwels, Y. (2016). A semi-parametric coxs regression model for zero-inflated left-censored time to event data. *Communications in Statistics-Theory and Methods*, **45**(7), 1969–1988.

[11] Cancho, V. G., de Castro, M., Dey, D. K. *et al.* (2013). Long-term survival models with latent activation under a flexible family of distributions. *Brazilian Journal of Probability and Statistics*, **27**(4), 585–600.

[12] Colosimo, E. A. & Giolo, S. R. (2006). Análise de sobrevivência aplicada. In *ABE-Projeto Fisher*. Edgard Blücher.

[13] Conceição, K. S., Andrade, M. G. & Louzada, F. (2013). Zero-modified poisson model: Bayesian approach, influence diagnostics, and an application to a brazilian leptospirosis notification data. *Biometrical Journal*, **55**(5), 661–678.

[14] Cooner, F., Banerjee, S., Carlin, B. P. & Sinha, D. (2007). Flexible cure rate modeling under latent activation schemes. *Journal of the American Statistical Association*, **102**(478).

[15] Cordeiro, G. M., Ortega, E. M. & Nadarajah, S. (2010). The kumaraswamy weibull distribution with application to failure data. *Journal of the Franklin Institute*, **347**(8), 1399–1429.

[16] Hand, D. J. & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **160**(3), 523–541.

[17] Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression 2nd Edition*. John Wiley and Sons, New York.

[18] Klein, J. & Moeschberger, M. (2003). Survival analysis: statistical methods for censored and truncated data. *Springer, New York*.

[19] Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**(1), 1–14.

[20] Leow, M. & Crook, J. (2016). The stability of survival model parameter estimates for predicting the probability of default: Empirical evidence over the credit crisis. *European Journal of Operational Research*, **249**(2), 457–464.

[21] Lessmann, S., Baesens, B., Seow, H. & Thomas, L. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update. *European Journal of Operational Research*, **247**, 124–136.

[22] Liu, L., Huang, X., Yaroshinsky, A. & Cormier, J. (2016). Joint frailty models for zero-inflated recurrent events in the presence of a terminal event. *Biometrics*, **72**(1), 204.

[23] Lord, D., Washington, S. P. & Ivan, J. N. (2005). Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention*, **37**(1), 35–46.

[24] Louzada, F., Cancho, V. G., Oliveira, M. R. & Yiqi, B. (2014). Modeling time to default on a personal loan portfolio in presence of disproportionate hazard rates. *Journal of Statistics Applications & Probability*, **3**(3), 295–305.

[25] Louzada-Neto, F. (2006). Lifetime modeling for credit scoring: A new alternative to traditional modeling via survival analysis. *Tecnologia de Crédito (Serasa)*, **56**, 8–22.

[26] Markel, P. D., DeFries, J. C. & Johnson, T. E. (1995). Ethanol-induced anesthesia in inbred strains of long-sleep and short-sleep mice: a genetic analysis of repeated measures using censored data. *Behavior genetics*, **25**(1), 67–73.

[27] Migon, H. S., Gamerman, D. & Louzada, F. (2014). *Statistical inference: an integrated approach*. CRC press.

[28] Ortega, E. M., Cancho, V. G. & Paula, G. A. (2009). Generalized log-gamma regression models with cure fraction. *Lifetime Data Analysis*, **15**(1), 79–106.

[29] Ortega, E. M., Cordeiro, G. M. & Kattan, M. W. (2012). The negative binomial–beta weibull regression model to predict the cure of prostate cancer. *Journal of Applied Statistics*, **39**(6), 1191–1210.

[30] Ospina, R. & Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, **56**(6), 1609–1623.

[31] Othus, M., Barlogie, B., LeBlanc, M. L. & Crowley, J. J. (2012). Cure models as a useful statistical tool for analyzing survival. *Clinical Cancer Research*, **18**(14), 3731–3736.

[32] Pereira, G. H., Botter, D. A. & Sandoval, M. C. (2013). A regression model for special proportions. *Statistical Modelling*, **13**(2), 125–151.

[33] R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[34] Rinne, H. (2008). *The Weibull distribution: a handbook*. CRC Press.

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

[36] Rodrigues, J., Cancho, V. G., de Castro, M. & Louzada-Neto, F. (2009). On the unification of long-term survival models. *Statistics & Probability Letters*, **79**, 753–759.

[37] Stepanova, M. & Thomas, L. (2002). Survival analysis methods for personal loan data. *Operations Research*, **50**(2), 277–289.

[38] Tong, E. N., Mues, C. & Thomas, L. C. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research*, **218**(1), 132–139.

[39] Vieira, A., Hinde, J. P. & Demétrio, C. G. (2000). Zero-inflated proportion data models applied to a biological control assay. *Journal of Applied Statistics*, **27**(3), 373–389.