Federal University of São Carlos
Exact Sciences and Technology Center
Department of Statistics

Models for inflated data applied to credit risk analysis

Mauro Ribeiro de Oliveira Júnior

São Carlos, SP, Brasil
September 27th, 2016

Mauro Ribeiro de Oliveira Júnior

Models for inflated data applied to credit risk analysis

A DOCTORAL DISSERTATION SUBMITEED TO THE DEPARTMENT OF STATISTICS OF THE FEDERAL UNIVERSITY OF SÃO CARLOS - DES/UFSCAR IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR IN STATISTICS.

Advisor:

Prof. Dr. Francisco Louzada Neto

São Carlos, SP, Brasil

September 27th, 2016

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária UFSCar
Processamento Técnico
com os dados fornecidos pelo(a) autor(a)

Oliveira Júnior, Mauro Ribeiro de
O48m Models for inflated data applied to credit risk analysis / Mauro Ribeiro de Oliveira Júnior. -- São Carlos : UFSCar, 2016.
103 p.

Tese (Doutorado) -- Universidade Federal de São Carlos, 2016.

1. Análise de Sobrevivência. 2. Modelo mistura. 3. Modelo Tempo Promoção. 4. Risco de Crédito. 5. Gestão de Risco de Crédito. I. Título.

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Estatística

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Tese de Doutorado do candidato Mauro Ribeiro de Oliveira Júnior, realizada em 27/09/2016:

Prof. Dr. Francisco Louzada Neto
USP

Prof. Dr. Adriano Kamimura Suzuki
USP

Profa. Dra. Maria Aparecida de Paiva Franco
Consultora

Prof. Dr. Marinho Gomes de Andrade Filho
USP

Profa. Dra. Vera Lucia Damasceno Tomazella
UFSCar

Disclaimer

The views and opinions expressed in this thesis are solely those of the author and do not necessarily reflect the official policy or position of any of his present or past employers.

Acknowledgments

I am deeply grateful for the guidance and friendship of my doctoral thesis advisor: Francisco Louzada Neto.

My thanks go to Mr. Edson Luiz de Carvalho Barbosa, who helped me to obtain a special license from Caixa Econômica Federal for exclusive dedication to studies. I am eternally grateful to Caixa Econômica Federal for the full support for this thesis.

My thanks extend to CAPES (process number: BEX 10583/14-9), whose financial support provided me with the sandwich doctorate period at the Credit Research Centre, University of Edinburgh Business School, United Kingdom. I am grateful to Dr. Fernando Moreira who welcomed me in the University of Edinburgh Business School and helped me in the development of this research. Literally, Edinburgh's weather was essential to the completion of this research.

Most importantly, I thank my family and my friends for their love and encouragement. I could not have written this without their support. This doctoral dissertation is dedicated to them. Finally, for the anonymous Brazilian taxpayer, I offer my most sincere thanks.

When I find myself in times of trouble, Mother Mary comes to me

Speaking words of wisdom, let it be

And in my hour of darkness she is standing right in front of me

Speaking words of wisdom, let it be

Let it be, let it be, let it be, let it be

Whisper words of wisdom, let it be

And when the broken hearted people living in the world agree

There will be an answer, let it be

For though they may be parted, there is still a chance that they will see

There will be an answer, let it be

Let it be, let it be, let it be, let it be

There will be an answer, let it be

Let it be, let it be, let it be, let it be

Whisper words of wisdom, let it be

Let it be, let it be, let it be, let it be

Whisper words of wisdom, let it be

And when the night is cloudy there is still a light that shines on me

Shine until tomorrow, let it be

I wake up to the sound of music, Mother Mary comes to me

Speaking words of wisdom, let it be

Let it be, let it be, let it be, yeah, let it be

There will be an answer, let it be

Let it be, let it be, let it be, yeah, let it be

Whisper words of wisdom, let it be

Composers: John Lennon / Paul Mccartney

Abstract

In this thesis, we introduce a methodology based on zero-inflated survival data for the purposes of dealing with propensity to default (credit risk) in bank loan portfolios. Our approach enables us to accommodate three different types of borrowers: (i) individual with event at the starting time, i.e., default on a loan at the beginning; (ii) non-susceptible for the event of default, or (iii) susceptible for the event. The information from borrowers in a given portfolio is exploited through the joint modeling of their survival time, with a multinomial logistic link for the three classes. An advantage of our approach is to accommodate zero-inflated times, which is not possible in the standard cure rate model introduced by Berkson & Gage (1952). The new model proposed is called zero-inflated cure rate model. We also extend the promotion cure rate model studied in Yakovlev & Tsodikov (1996) and Chen *et al.* (1999), by incorporating excess of zeros in the modelling. Despite allowing to relate covariates to the fraction of cure, the current approach does not enable to relate covariates to the fraction of zeros. The new model proposed is called zero-inflated promotion cure rate model. The second part of this thesis aims at proposing a regression version of the inflated mixture model presented by Calabrese (2014) to deal with multimodality in loss given default data. The novel methodology is applied in four retail portfolios of a large Brazilian commercial bank.

Resumo

Nesta tese de doutorado, introduzimos uma metodologia baseada em dados de sobrevivência inflacionados em zero com o objetivo de lidar com propensão à inadimplencia (ou seja, risco de crédito) em carteiras de empréstimos bancários. Nossa abordagem permite acomodar (extrair informações de) três tipos diferentes de clientes bancários: (i) indivíduo com empréstimo inadimplente logo no início; (ii) cliente não suscetível ao evento de inadimplência, ou (iii) cliente suscetível ao evento de inadimplir. A informação dos empréstimos em um determinado portfólio é explorada através da modelagem conjunta do seu tempo de sobrevivência, com uma ligação logística multinomial para as três classes. Uma vantagem da nossa abordagem é acomodar tempos inflados em zero, o que não é possível no modelo de fração de cura padrão introduzido por Berkson & Gage (1952). Também estendemos o modelo com fração de cura estudado por Yakovlev & Tsodikov (1996) e Chen *et al.* (1999), incorporando excesso de zeros na modelagem. Apesar de permitir relacionar covariáveis à fração de cura do modelo, a abordagem padrão não permite relacionar covariáveis com a proporção de zeros dos dados. A segunda parte desta tese visa propor uma versão de regressão do modelo de mistura inflada apresentada por Calabrese (2014), visando extrair informações referentes a multimodalidade apresentada em dados relacionados à perda dado a inadimplência (LGD). A nova metodologia é aplicada em quatro carteiras de empréstimo de varejo de um grande banco comercial brasileiro.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

More often than not, banks and financial institutions completely lose contact with customers as soon as their loans are granted and, therefore, all amount lent is lost. This group of borrowers, arguably, is the most costly for the bank. Here, they are the primary concern, so we particularly defined them as straight-to-default customers, or STD customers for short. The term "default", used throughout this doctoral dissertation, means the event of interest in credit risk analysis. It happens when borrowers lose the creditworthiness to meet their commitments with loans. The default criterion may vary from bank to bank by conservative reasons. Generally, a bank declare a default condition if a customer has not been paying any instalments for more than three consecutive months. Henceforward, it is the definition that we assume in order to declare that a customer has defaulted on a loan.

There is also another group of problematic customers, the usual ones. Those no longer can afford their loan instalments, but, unlike STD customers, they manage keep up to date with their debts for a while. Mostly of the time, by private financial reasons, they cannot afford anymore their debts with the bank and default on their loans. Fortunately, to ensure the survival of bankers, there are good customers, actually, most of them. Those who always keep up to date with their obligations and, therefore, there will not be records of events of default. Therefore, for the survival of bankers and mainly for the maximization of profits, they must seek to maintain high rate of non-defaulting loans, while the rates of STD and defaulted customers must be very low.

Also in accordance with regulations already established by international supervisory bodies, such as the Basel Committee on Banking Supervision, (BCBS, 2006), we have maintained in this doctoral dissertation that the event of default of a loan happens when it has three consecutive months without any repayment. Therefore, the default rate, or the

non-performing loan (NPL) rate, is computed as the ratio of loans past due in excess of 90 days (three months) relative to the total outstanding within a corresponding category of loan. For example, if a financial institution has a student loan portfolio of 1.000 loans and 100 of those have delinquent payments greater than three months, then the default rate (also known as delinquency rate) of the student loan portfolio is 10%.

Figure 1.1 illustrates the historical default rates of the two largest Brazilian private banks, respectively, Itaú Unibanco S.A and Banco Bradesco S.A. These historical data sets are available once both companies have shares listed on stock exchanges. Their balance sheets are also quarterly reported to the public in general. Following, we illustrate the NPL historical rates from these public data, which were followed for eight quarters in addition to be segregated by type of borrower, i.e., loans to individuals, small and medium enterprises (SME), corporate, and finally, the total portfolio (without any segregation).

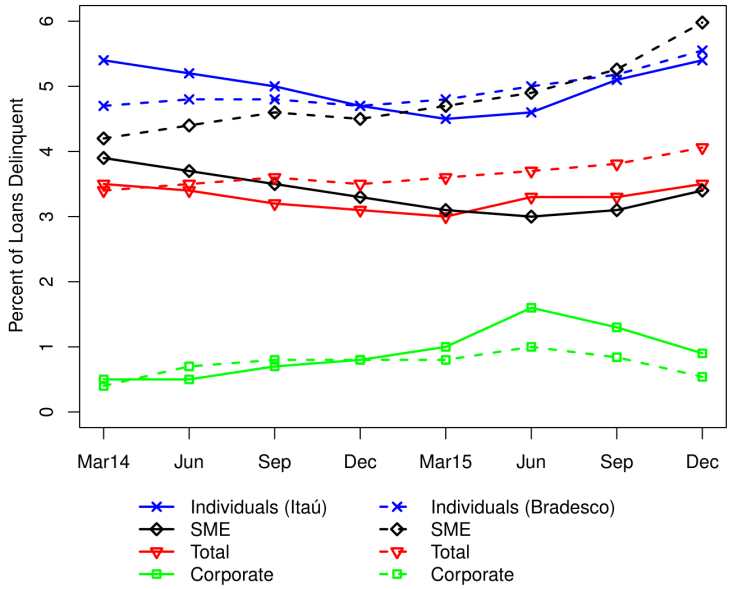

Figure 1.1: Percent of Loans Delinquent with 90+ days past due (Quarterly Data)

All banking customers described initially, STD's, defaulters and non-defaulter customers, and their credit risk behavior, motivate the models we first propose in this doctor dissertation. Thereby, the dataset analysis comprises customers who, in one way or another, have not honoured their contractual obligations with the bank, either by willingness to do not pay even from the beginning (by STD customers), or by the loss of creditworthiness over time (by usual defaulters), along with good customers, who have always honoured their obligations and, therefore, have never experienced the event of default (non-defaulters).

Such data analysis must be addressed to make a holistic risk management of the

banking loan portfolio, that is, dealing with control of default and ensuring the customer loyalty growth within the group of customers non-susceptible to default. As we see in the next Figure 1.2, these considerations delimit the data we will cover first in this dissertation: a set of zeros, positives and unrecorded banking loan survival times.

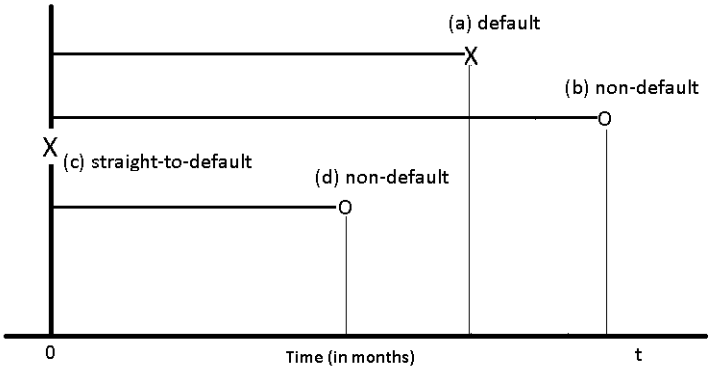

Figure 1.2: Loan survival time data.

In the daily routine of banks, borrowers and loans are monitored from their granting date, but, especially, the first three months are essential for marking as straight-to-default customers. Generally, the follow-up period ranging from 12, 24, 36 months, or even more, depending on the loan portfolio terms. As to register the event of interest, i.e., the event of default, it takes at least three months of follow-up, because one needs at least three months without payments, so, 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 $t - 3$. So, the origin point in the chart above (t = 0) takes place from the third month after the loans have been granted. Therefore, in Figure 1.2, the (c) survival time equal to zero comes from STD customer, those who went through the first three months without any payment; positive default times as in (a) are from usual defaulters, those who defaulted during the normal loan repayment term. In (d), the absence of registration is due early repayment, while in (b) the loan is still under payment at the end of the follow-up period.

Thus, this doctoral dissertation firstly aims at proposing a model that jointly accommodate three types of time-to-default data present in banking loan portfolios. It leads to a new framework that overcomes the standard cure rate model, introduced by Berkson & Gage (1952), with regard to the accommodation of zero-inflated data in survival analysis modeling.

The second kind of data analysed in this dissertation is associated with losses incurred by the bank due to the realization of a event of default. The term used in the literature for

this random variable is loss given default, shortly referred as LGD. LGD (%) is the ratio between the net amount of lost funds, that is lost by a bank when a borrower defaults on a loan, and the overdue debt at the exact moment of the default. Its distribution ranges in the closed real interval $[0, 1]$.

To illustrate the concepts involved in implementing a LGD methodology, consider the following example: the customer has defaulted on a collateralized loan with an outstanding debt of $£10,000$. If the bank is able to sell the collateral for a net price of $£6,000$, including all costs related to the repossession process, then $£4,000$, or 40% of the value exposed at the moment of default are lost, and thus the LGD is 40%. From the banking practice, it is known that some loans have strong guarantees and are easier to repossess, as mortgage loans for instance. In these cases, in the LGD dataset, there will be a large concentration of zeros in the LGD distribution. This excess of zeros is result of a successful recovery process of all amounts that were overdue. On the other hand, unfortunately by the bank side, might appear excess of ones in the database, which comes from the complete failure in the recovery process.

Therefore, it is very important manage overall factors that lead into a successful recovery process within defaulted loan portfolios. For example, figuring out borrower features, loan characteristics and details of the entire collection process that triggers a smaller monetary loss to the bank. According to the datasets available for study in this dissertation, the LGD distribution may present a bimodality within the $(0, 1)$ interval, due, mainly, to the concentration in partial recovery peaks. For this reason, the LGD approach proposed in this dissertation is concerned with loss given default modeling in the presence of bimodality in the $(0, 1)$ interval and along with the aforementioned presence of zeros and ones excess.

This subject is developed in the fourth chapter of the dissertation, where we present a zero-and-one inflated mixture model to deal with it. In the next section, we present the data sets that motivated the two main objectives of this dissertation, i.e., survival data inflated with zeros from bank loan portfolios, and bimodal LGD data inflated with excess of zeros and ones from defaulted bank loan portfolios. Following to the dataset presentations, we proceed a literature review on the related topics and, finally, we present our objectives regarding to the gaps in the literature we propose fill in with this doctoral dissertation.

1.1 Real data sets

This section presents the databases made available by a major Brazilian bank. It is important to note that the presented datasets, amounts, rates and levels of the available covariates, do not necessarily represent the actual condition of the financial institution's portfolio. 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.

1.1.1 Loan survival time data

The first analysed portfolio was collected from customers who have taken a personal loan over a 60-month period, between the years 2010 and 2015. Table 1.1 shows the customer's quantitative frequencies of the loan portfolio provided by the bank. It is composed of 5733 time-to-default (in months), with an approximate 80% rate of censored data, that is, a high rate of non-default loans. Our objective is to assess if customer characteristics are associated with consumer propensity of being STD, defaulter or non-defaulter customers.

	Number of customers	Number of STD ($T = 0$)	Number of defaulters (T>0)	Number of censored (T>0)
Total	5733	321 (5.60%)	810 (14.13%)	4602 (80.27%)

Table 1.1: 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 1.2 shows the quantitative frequency according to the available covariates.

Figure 1.3 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.

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 1.2: Quantity of the available covariates.

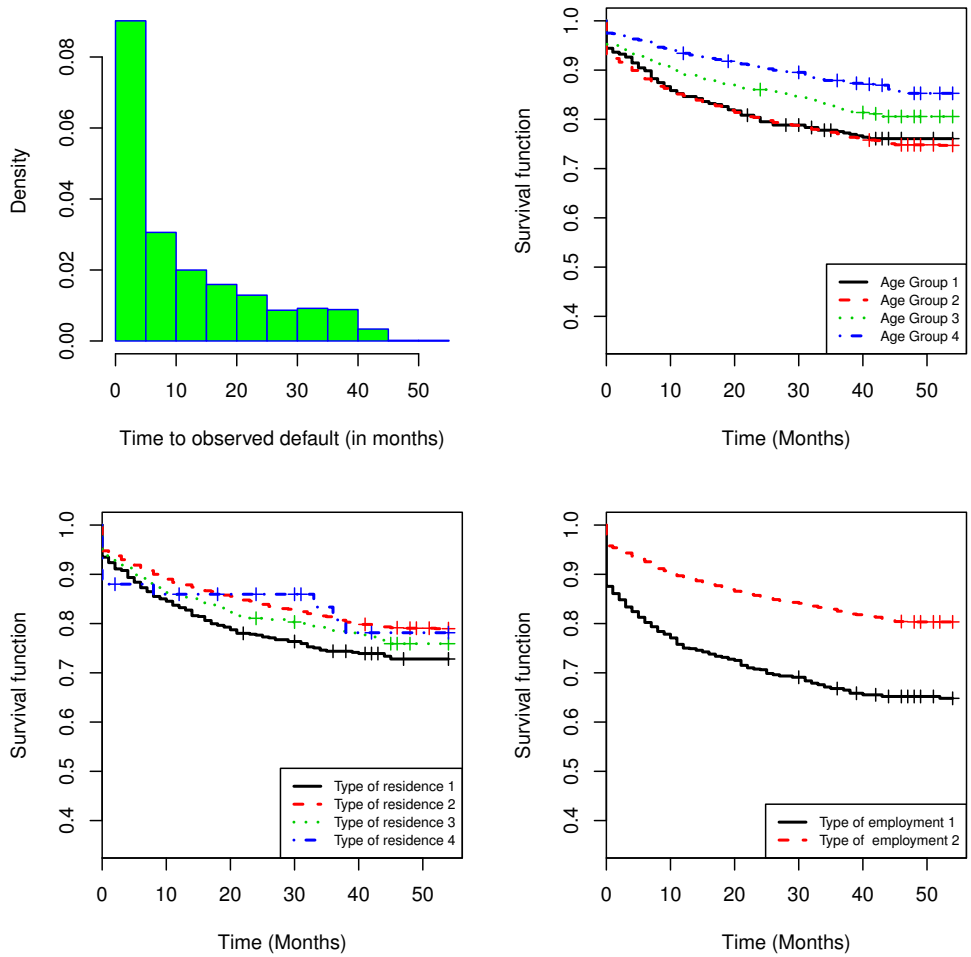

Figure 1.3: Brazilian bank loan portfolio data. Top panel, shows a histogram for the observed time-to-default variable of interest (left) and Kaplan-Meier survival curves stratified by age group (right). Bottom panel, Kaplan-Meier survival curves stratified by type of residence (left) and Kaplan-Meier survival curves stratified by type of employment (right).

1.1.2 Loss given default data

To motivate the LGD modeling proposed in the sixth and seventh chapters of this dissertation, we analyse four retail portfolios of defaulted loans made available by a large Brazilian commercial bank. Each portfolio is grouped according to the type of guarantees offered in the loan, or even the complete lack thereof. Of course, loan contract characteristics affect directly the presented shapes of the LGD distributions. For data confidentiality reasons, we do not explain the features of each loan making up each portfolio, we can only mention these are retail exposures, as defined in BCBS (2006), paragraph 231. The all data set comprises 41.677 defaulted retail loans, as summarized separately in the Table 1.3.

Portfolio	Quantity	Mean	Median	SD	Number of 0's	Number of 1's
1	15.295	0.52195	0.7272	0.4746	5.722	6.634
2	22.951	0.59814	0.9093	0.4596	8.349	8.398
3	440	0.32945	0.7466	0.4004	232	44
4	2.991	0.72060	0.9175	0.3810	510	265

Table 1.3: Summary of observed LGD data.

Its whole LGD distribution, that is, considering all 41.677 defaulted loans together, is presented in Figure 1.4, where it shows a five-modal distribution.

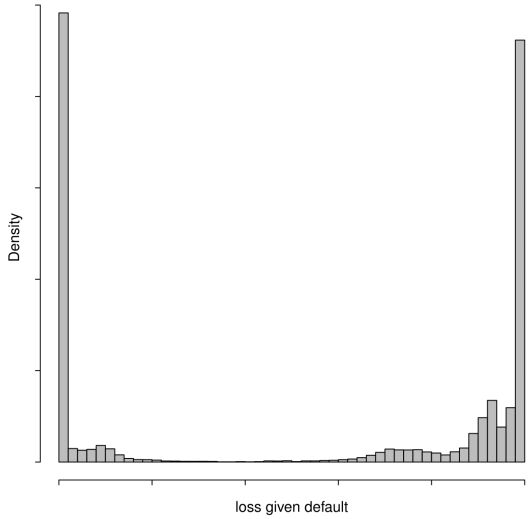

Figure 1.4: Multimodal LGD distribution.

In the following figures, are shown separately the four portfolios, each one of them presenting quite different shape of bimodality. Note that in these next figures, for clarity

in data visualization, the zeros and ones are excluded, however, they are accounted in the proposed parameter estimation procedure. The zero and ones amounts (#) were presented in the right columns of the Table 1.3.

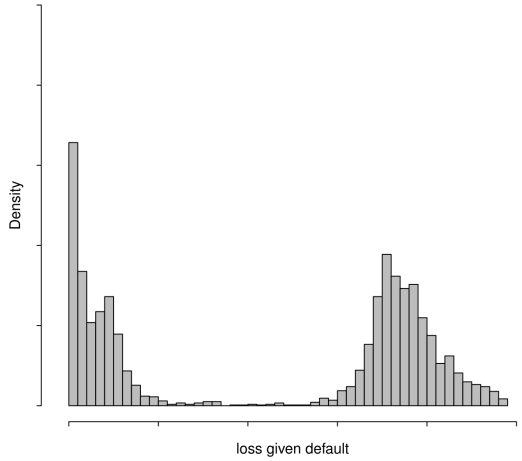

Figure 1.5: Portfolio 1: Loss given default distribution.

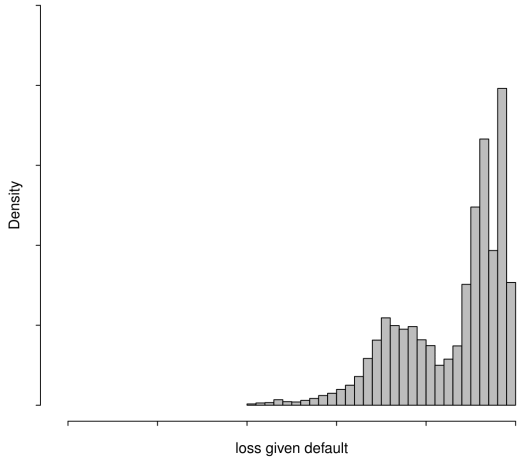

Figure 1.6: Portfolio 2: Loss given default distribution.

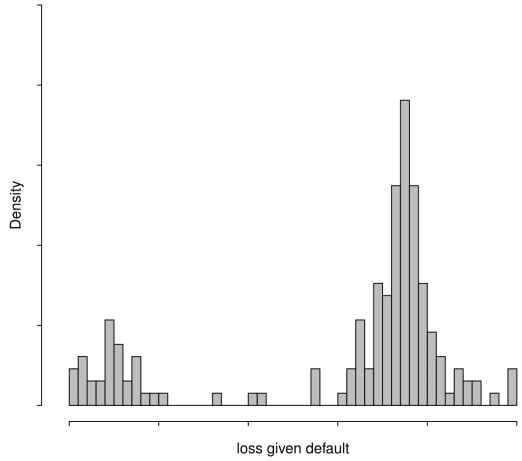

Figure 1.7: Portfolio 3: Loss given default distribution.

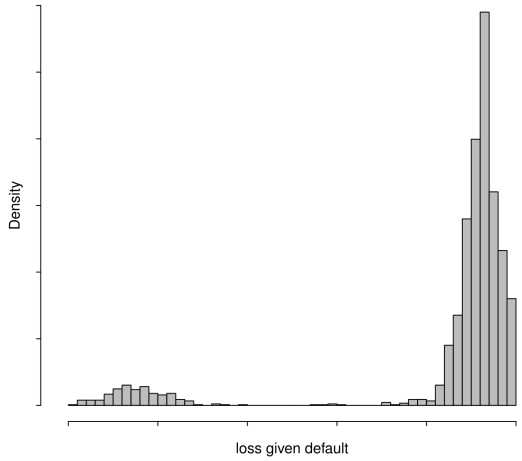

Figure 1.8: Portfolio 4: Loss given default distribution.

Next, we present a review of literature for the first topic of this dissertation, i.e., about the zero-inflated survival models applied to credit risk analysis, which will be developed in the next chapters. Following the survival analysis review, we present some concepts and a review of literature for the second studied subject, i.e., the loss given default modeling in the presence of zeros and ones excess, which will be developed in the last chapters of this dissertation.

1.2 Literature review

To use survival analysis techniques in credit risk settings, we must consider the modeling outcome of interest (event of interest) as the survival time after the loan approval, also mentioned here as customer or loan survival time. It is represented by the time until the occurrence of the event of default. Consequently, as already mentioned, we can say that bankers expect it to be rarely recorded within loan portfolios. In order to deal with this type of survival data, it is generally modeled by a continuous probability distribution, with support on the real non-negative interval $[0, +\infty)$. Such approach has been applied in different papers, through the application of different survival frameworks.

Banasik *et al.* (1999) list a series of advantages of using survival analysis in credit risk modeling. The main is due it is concerned with estimating over time the probability of default, allowing banks to re-evaluate provision against expected losses during the maturity of their portfolios. The authors raised that survival analysis can also bring more accurate information on the portfolio profitability over time, thus providing information on how profitable loans are according with different portfolio terms. The authors also envisioned

the possibility of being incorporated macroeconomic variables in the modeling.

In Abad *et al.* (2009), analysing a portfolio of 25000 consumer loans from a Spanish bank, the authors applied a semi-parametric approach, a Cox's proportional hazard model, to compare its performance among other models. The second approach used was generalized linear models under censoring and the third one was based on non-parametric kernel estimation. According to the authors, a reasonable improvement in the results could be achieved when using the survival analysis to modeling credit quality in terms of "lifetime of loans". Stepanova & Thomas (2002), and Bellotti & Crook (2009) more recently, also applied survival analysis based on Cox's proportional hazard models. The latter, built model for time-to-default on a large UK portfolio of credit cards, taking into account the presence of time-varying covariates. The authors found survival analysis competitive in comparison with the standard logistic regression as a credit scoring method for prediction, and, the most important part, the inclusion of macroeconomic variables may gives a statistically significant improvement in its predictive performance.

Abreu (2004) and Tong *et al.* (2012) presented different frameworks to deal with modeling time to default on loan portfolios. While Abreu (2004) dealt with simulated data, the latter authors modeled time to default on a real UK personal loan portfolio. In these works, the authors estimated mixture cure models and compared its performance to the Cox proportional hazards method and to the standard logistic regression. They all found the three approaches were competitive regarding discrimination performance, furthermore, they showed the mixture cure models can offer additional insights. Mainly due to the possibility of estimating the parameters that determine susceptibility to default, in addition to parameters that influence time to default of a borrower. Such conclusion is base on the mixture cure model's ability to distinguish between two sub-populations, according with its susceptibility to default on a loan or not. However, in the newer mixture cure approach presented in Tong *et al.* (2012), it was required a proportional hazard structure, once the latency model component is based on a proportional hazard survival model.

In Louzada *et al.* (2014), the authors considered a parametric Weibull mixture cure model for time to default on a Brazilian personal loan portfolio. In that data set, taking into account presence of covariates, it was evident the presence of disproportionate hazard rate among covariate levels. Such an approach, as appeared previously in Abreu (2004), can be seen as a complement approach to the modeling framework presented in Tong

et al. (2012), which requires a proportional hazard structure. Louzada *et al.* (2014) found that their proposed mixture cure models were more appropriate for non proportional scenarios, which by the way has not been claimed in recent articles that brings proportional hazard survival approaches for credit scoring purposes. According to the authors, such misspecification may lead to erroneous measurements, as under or overestimated expected losses, which is financially damaging to the bank.

The application of cure rate models in credit risk, alongside other modeling techniques, is that it can accommodate censored data. The analysis of such information is not supported in credit scoring techniques purely based on good and bad customer classification. Thus, more information can be gathered when building a survival model, see for instance Abreu (2004), Hand & Henley (1997) and Lessmann *et al.* (2015). In the credit risk context, censoring occurs when a loan is still under repayment at the moment of data collection, i.e., still being a good loan. 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, thus, we cannot conclude on whether the customer is a good or a bad customer at the end of the follow-up period.

In other areas, as in medicine for instance, censoring 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 in the literature. The first cure rate model was introduced by Berkson & Gage (1952). In Barriga *et al.* (2015), 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 use of positive continuous distributions in the cure rate framework is already considered an usual modeling practice, as it can well accommodate time-to-event occurrences, which primarily contains non-negative (or censored data), see for instance Cordeiro *et al.* (2010) and Ortega *et al.* (2009). However, it cannot fit excess of zeros that may make up time-to-default data set of loan portfolios, for example. Unlike survival data analysis, in other areas we can observe most commonly the existence of non-negative data with presence of zeros, sometimes with excess.

Usually, the excess of zeros occurs in count data studies, as analysed in Lambert

(1992), Barry & Welsh (2002), Lord *et al.* (2005), Conceição *et al.* (2013). In Vieira *et al.* (2000) and Ospina & Ferrari (2012), the authors dealt with zero-inflated proportion data models. Therefore, it is already a commonplace the expression "zero-inflated data". In Liu *et al.* (2015), the occurrence of zeros excess is exploited within 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 tumours in a soft tissue sarcoma study. Zero-inflated data also appears in the context of left censored data. In Blackwood (1991), for example, left censored data are generate in experiments related to the presence of toxic products in the environment. Due to the inaccuracy of the tools used for measurement, it is not always possible to fully observe some results and only a lower limit is recorded.

Also dealing with the presence of left censored data, Braekers & Grouwels (2016) reviewed a laboratory experiment with mice conducted by Markel *et al.* (1995), where the outcome of interest is the induced sleeping time, measured after ingestion of a dose of ethanol. As some mice present immunity for the administered dose of ethanol, the analysed data set contains a proportion of sleeping time equal to zero. In the statistical approach proposed to re-analyse the data obtained in the earlier conducted experiment, i.e., in order to reinvestigate the influence of covariates on the outcome of interest, Braekers & Grouwels (2016) 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 aim at considering human survival times equal to zero in clinical trials or medical studies. That is, seeing that clinical trials may lead to the event of interest at the starting time (zero time), it will mean that there are instantaneous deaths of patients undergoing such an experimental procedure. Hence, it might be why, to the best of our knowledge, we have not found study that is willing to account for zero-inflated data in the medical specialized literature which aims to analyse the survival of human patients under disease treatment. However, the same sense of respect expected in clinical trials, in some way, does not seem to be required when dealing with credit risk events. On the contrary, information about zero-inflated time should be taken into account in credit risk analysis, and must be useful for identifying customers who apply for loans only for the purpose of defrauding the bank by, since the beginning, not honouring its obligations under the granted credits.

Thus, the first objective of this doctoral dissertation is to propose a way to accom-

modate zero-inflated survival date in a credit risk setting, in a way that has not yet been incorporated into the statistical literature. For that, we extend the standard cure rate model introduced by Berkson & Gage (1952), by incorporating excess of zeros in the modeling through the inclusion of a multinomial logistic link for the three classes of available loan survival times: the zero-inflated ones due to straight-to-default loans; the positive time-to-default due to defaulted loans; and finally, the class of censored observations due to the high non-default rate shown in the data.

Next, we present some concepts and a review of literature for the second topic of this dissertation, i.e., the loss given default modeling in the presence of zeros and ones excess, which will be developed in the last chapters.

The development of new statistical methodologies for credit risk analysis was pushed by the recommendations of the Basel Committee on Banking Supervision, strongly favouring the development of new models. Since the Basel II publications in the mid-2000s, recommending central banks to allow banks to use internal data to calculate credit risk measures of their portfolios, much has been proposed in the literature on probability of default, loss given default and exposure at default. See, for example, Valvonis (2008), Engelmann & Rauhmeier (2011), Loterman *et al.* (2012), Yashkir & Yashkir (2013), Leow & Crook (2014) , Leow & Crook (2016) and Tong *et al.* (2016). The importance is justified as these parameters comprise the main ingredients of regulatory capital calculation, that is, what banks must set aside to cope with unexpected losses from credit loan portfolios.

According to Basel II rules for corporate, sovereign and bank exposures, see BCBS (2006) in the paragraphs 286 and 297, the loss given default (LGD) is measured as the proportion of unrecovered debt, compared to total counterparty overdue debt. That is, given that default has occurred and the process of recovery has been finalized, the fraction of all debit past due which is not recovered by the bank, in relation to the total amount of the overdue debt at moment of the default, is defined as loss given default. Thus, the support of LGD distribution lies in the $[0, 1]$ real closed interval.

Despite the simplicity in setting it, according to Schuermann (2004), there are distinctions about the treatment that should be given to different types of portfolios. For instance, in case of the aforementioned portfolios, corporate, sovereign and bank exposures, banks must provide an individual estimate of LGD for each exposure and, for that reason, a different approach that has been applied to loan retail portfolios. In fact, as retail exposures typically represent majority of loan portfolios of commercial banks, it would be

impossible to give an individualized treatment for each exposure. That is why, even before Basel II recommendations, bank risk managers relied on automated scoring models. This means that, mostly, modeling credit risk involves estimating parametric statistical models, see for example Thomas *et al.* (2002), Crook *et al.* (2007), Porath (2011) and Lessmann *et al.* (2015).

However, as expected, an exaggerated dependence on complex statistical models may lead to new sources of risks, in this case, the model risk. In other words, the risk of not choosing the best model in the light of the available data. An attempt to draw attention to model risk and encourage mitigation of this source of risk has already been addressed by the Basel Committee, as stated in BCBS (2015), p. 2, "Supervisors should be cautious against over-reliance on internal models for credit risk management and regulatory capital. Where appropriate, simple measures could be evaluated in conjunction with sophisticated modeling to provide a more complete picture".

Regarding the modeling of LGD, Basel II also recommends that its calculation must consider all relevant factors that impact in the loss triggered by the event of default. According its paragraph 460, BCBS (2006) strongly recommends the calculation must include all material discount effects and all material direct and indirect costs associated with collection process on the defaulted loan portfolio. Since it is known that LGD has considerable impact on the regulatory capital amount, according to Gürtler & Hibbeln (2013) and Yao *et al.* (2015), small differences can lead to major distortions in its calculation. For this reason, when dealing with large retail portfolios without sufficient evidence of the impact of each direct and indirect recovery cost, in order to proceed an reliable estimate, we must opt for models that bring a extra dose of conservatism.

Another challenge in modeling LGD lies in the fact that there are an excesses of zeros and ones in the data. These excesses are expected, since LGD equal to zero means that the default event has not incurred any loss given its realization, i.e., the bank was able to recover the overdue amounts. For example, appropriating the assets given as collateral, see for example Leow & Mues (2012), Loterman *et al.* (2012) and Oliveira & Louzada (2014b), or by some other action, such as renegotiation of the debt under the national bankruptcy protection law. On the other hand, excess of LGD equal to 1 has very unfavourable meaning for the bank, since it means that 100% of the overdue debts of defaulting borrowers have not been recovered and, therefore, the bank fully assumes the loss with the defaulted loan.

When dealing with mortgage loans, it is very common the property be subsequently repossessed after the customer incur in default and the sold price fully covered the loan balance at default. In such cases, obviously, the LGD data will contain zero excesses due to the total recoveries. In Tong *et al.* (2013), the authors proposed a model based on a zero-adjusted gamma distribution for a mortgage loss given default data, in order to account for the presence of high excesses of zeros. But, rather than fitting the rate of loss given default, the authors accounted for the lost values in GBP resulting in the occurrence of a mortgage default.

In addition to the excess of zeros and ones, LGD data sets may present another type of bimodality, now referring to data included between the extremes of the interval $[0, 1]$. As presented in the section 1.1.2, the LGD data sets studied in this dissertation, presents a LGD concentration both close to zero as close to one, hence, presenting a multimodality which is inherent to the behavior of the data made available by the Bank for studying, mainly due to the concentration of partial recovery peaks, in addition to peaks at 100% and 0% recovery. In the foregoing context, i.e., concerning to the loss given default modeling in the presence of zeros and ones excess and the mentioned bimodality within the interval $(0, 1)$, we contribute to the literature by extending the established framework already proposed by Calabrese (2014), where we propose a regression version of the zero-and-one inflated mixture beta model presented by that author to accommodate such multimodality in the LGD distribution.

Although Calabrese (2014) has dealt with excess of zeros and ones in a real LGD data on Italian bank loans, the author has not presented real situations of multimodality. Instead, it was assumed an arbitrary mixture of two betas to encourage the forecasting of two distinct periods, one with higher and another with lower LGD average. Furthermore, the author is not intended to study the relation between covariates and the outcome of interest. In this sense, we complement the work made in Calabrese (2014) by applying our methodology in a variety of real bank loan portfolios within a regression model version. In addition, we perform a simulation study to assess estimation performance of the inflated mixture regression model proposed, which was not carried out in that referred paper. We also complement the work done by Hlawatsch & Ostrowski (2011), which, despite dealing with simulated bimodality, do not address the occurrence of zeros and ones excesses in its bimodal LGD data.

1.3 Objectives

The first objective of this dissertation is supported by a need to accommodate zero-inflated survival date in a credit risk setting. For that, we extend the standard cure rate model introduced by Berkson & Gage (1952) and the promotion cure rate model studied in Yakovlev & Tsodikov (1996) and Chen *et al.* (1999), by incorporating excess of zeros in the modeling of survival data with cure rate. To exemplify the application of the proposed approach, we analyse a portfolio of personal loans made available by a large Brazilian commercial bank. Our goal is to assess whether a borrower is more likely to go straight to default, i.e., presetting a survival time zero. The jointly modeling also allow get the information if she or he will (or will not) become a defaulter within a survival analysis context, i.e., analysing the probability of being a non-default customer. This is why we propose a methodology based on augmented cure rate survival model, for the purposes of dealing with the problem of assessing the propensity to default at the begging in bank loan data, with excess of zeros and with a high rate of censored data.

Furthermore, notwithstanding the bimodality and zeros and ones excess has been partially accounted for in the recent literature, as in Hlawatsch & Ostrowski (2011), Tong *et al.* (2013) and Calabrese (2014), to the best of our knowledge, the full regression configuration of the aforestated LGD model has not been wholly incorporated into any framework. Hence, in the second part of this dissertation we fill a gap in the literature by introducing a simple statistical tool for credit risk managers deal, as effectively as possible, with loss given default in multi-shape data. Thereby, is presented in the last chapters of this dissertation an inflated mixture regression model by assuming a mixed of degenerate distributions to handle all zeros and ones excess, together with a mixture of distributions to account for bimodal losses. Along with the already mentioned variety of real applications, we carried out Monte Carlo simulation studies to check the finite sample performance of the all regression estimation procedures proposed.

1.4 Overview

This doctoral dissertation is organized as follows. In chapter 2, we present a methodology based on zero-inflated survival modeling to account for zero inflated data in bank loan portfolios, where we extend the standard cure rate model introduced by Berkson & Gage (1952). A study based on Monte Carlo simulations with a variety of parameters

is presented in section 2.4. An application to a real data set of a Brazilian bank loan portfolio is presented in section 2.5. Conclusions are presented in the section 2.6.

In chapter 3, we formulate the zero-inflated promotion cure rate, which is proposed to extend the model studied in Yakovlev & Tsodikov (1996) and Chen *et al.* (1999), by incorporating excess of zeros in its modeling. A Monte Carlo simulation studies with a variety of parameters is presented in section 3.3. An application to a real data set of a Brazilian bank loan portfolio is presented in section 3.4 and the results are compared with the obtained in the chapter 2. Conclusions are presented in the section 3.5.

In chapter 4, we formulate the inflated mixture model based on the model introduced by Calabrese (2014). Section 4.3 we present a maximum likelihood estimation procedure. A simulation study with different vector parameters is presented in section 4.4. An application to a real variety of retail portfolios of a large Brazilian bank is presented in section 4.5. Conclusions are presented in the section 4.6.

In chapter 5, we present the final conclusions and proposals for future work. Note that, to allow the chapters be read (almost) independently of one another, some concepts may appear repeatedly in some parts of the text.

Chapter 2

The Zero-inflated Non-default Rate Model

In this chapter, we propose a new non-default rate model for taking into account three different types of individuals: (i) individual with event at the starting time (zero time); (ii) non-susceptible for the event, or (iii) susceptible for the event. With respect to the survival analysis framework, this approach accommodate zero-inflated times, which is not possible in the standard cure rate model introduced by Berkson & Gage (1952). To illustrate the proposed method, a real dataset of loan survival times is fitted by the zero-inflated Weibull non-default rate model. The parameter estimation is reached by maximum likelihood estimation and Monte Carlo simulations are carried out to assess its finite sample performance.

2.1 Introduction

In survival analysis, the random variable T of interest is the time elapsed until the occurrence of an expected event, i.e., the event of interest. 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 time until to 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 time up to the occurrence of a loan default. Obviously, in all cases T is

non-negative and, generally, is treated as a continuous random variable.

According to Colosimo & Giolo (2006), there are several functions which completely specify the distribution of a random variable in survival analysis, since they are mathematically equivalent functions. They are the probability density function (PDF), cumulative distribution function (CDF), complementary cumulative distribution function (CCDF), the hazard function, the cumulative hazard funciton and, finally, the mean residual life function. Within a survival analysis context, the complementary cumulative distribution function (CCDF) is known as 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. In that way, it cannot properly accommodate the proportion of customer who are not susceptible to default on a loan given its approval. This follows to the fact 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 contemplated. For instance, returning to examples in the medical field, there are patients suffering from disease who, once submitted to treatment, recover completely. 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 in default. In this financial context, they are also known as non-defaulting clients or long-term clients. Therefore, when it is needed to consider the presence of cure or long-term data, the traditional survival analysis is not at all suitable for modelling failure time. In those cases, where there are immunity to the occurrence of failures, new statistical tools have been proposed.

To handle the aforementioned challenge, Berkson & Gage (1952) 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, the susceptible group and the non-susceptible group to the occurrence of the event of interest:

$$S(t) = p + (1 - p)S_0(t), \qquad t \geq 0, \tag{2.1}$$

where S_0 is the survival baseline function of the individuals susceptible to failure and $p > 0$ is the proportion of the individuals immune to failure (cured). This model is called standard cure rate model or long-term survival model. Unlike S_0, S is an improper survival

function, since it satisfies $\lim_{t\to\infty} S(t) = p > 0$.

Another attribute of the standard cure rate model, according to Othus *et al.* (2012), among others authors, is that it can bring to light more informations about the event in study since it allows associate covariates in both parts of the model. Indeed, it allows covariates to have different influence 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.

2.1.1 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 (default on a loan) at the beginning of the considered study, i.e., with survival time equal to zero. As aforementioned, we used a different terminology in order to clarify its use in a credit scoring setting. So, from now on, we denote cure by non-default, leading to what we call by non-default rate models. In this sense, 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 default, i.e., related to non-defaulters.

Thus, we propose the following expression for the improper survival function of all possible loan survival times:

$$S(t) = p_1 + (1 - p_0 - p_1)S_0(t), \qquad t \geq 0, \tag{2.2}$$

where S_0 is the baseline survival function related to the $(1-p_0-p_1)$ proportion of subjects susceptible to default, p_1 is the proportion of subjects immune to default and finally, p_0 is the proportion of straight-to-default (STD) individuals. This model in (2.2) is called zero-inflated non-default rate model. The fact that differentiates the zero-inflated non-default rate version from the standard non-default rate approach in (2.1), since 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 $S(0) = 1 - p_0 < 1$.

The above properties can be viewed in the plot of the proposed survival function expression. Note that, if $p_0 = 0$, i.e., without the excess of zeros, we have the non-default rate model of Berkson & Gage (1952).

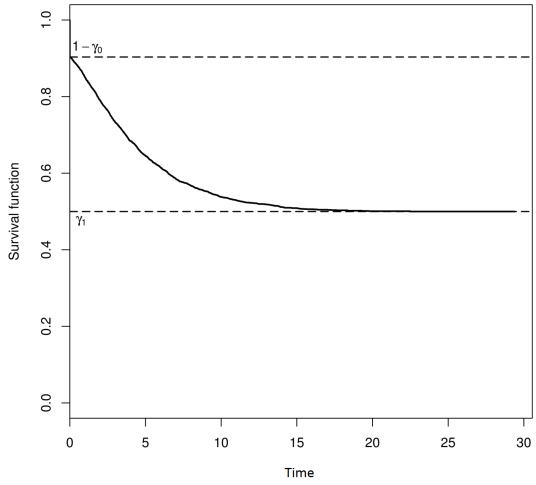

Figure 2.1: The (improper) survival function of the zero-inflated survival model.

Justification

In this dissertation, 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. To reach this goal, we propose a jointly modelling of zero-inflated time in loan survival data with non-default rate. To exemplify the application of the proposed approach, we analyse a portfolio of loans made available by a large Brazilian commercial bank.

Organization

The chapter is organized as follows. In Section 2.2, we formulate the model and present the approach for parameter estimation. A study based on Monte Carlo simulations with a variety of parameters is presented in Section 2.4. An application to a real data set of a Brazilian bank loan portfolio is presented in Section 2.5. Some general remarks are presented in Section 2.6.

2.2 Model specification

In what follows, we consider the zero-inflated non-default rate model as defined in (2.2), with baseline function $f_0(t|\boldsymbol{\phi})$, indexed by a vector of parameters $\boldsymbol{\phi}$ of size k, to be freely chosen according with the data analysed. The associated (improper) cumulative distribution function (CDF) and probability density function (PDF) are given by

$$F(t|p_0, p_1, \boldsymbol{\phi}) = p_0 + (1 - p_0 - p_1)F_0(t|\boldsymbol{\phi}), \qquad t \geq 0 \tag{2.3}$$

and

$$f(t|p_0, p_1, \boldsymbol{\phi}) = \begin{cases} p_0, & \text{if} \quad t = 0, \\ (1 - p_0 - p_1)f_0(t|\boldsymbol{\phi}), & \text{if} \quad 0 < t, \end{cases} \tag{2.4}$$

where the parameters p_0 and p_1 are as defined in Section 2.1.1. F_0 and f_0 are, respectively, the cumulative distribution function and probability density function underpinning the $(1 - p_0 - p_1)$ proportion of subject susceptible to failure.

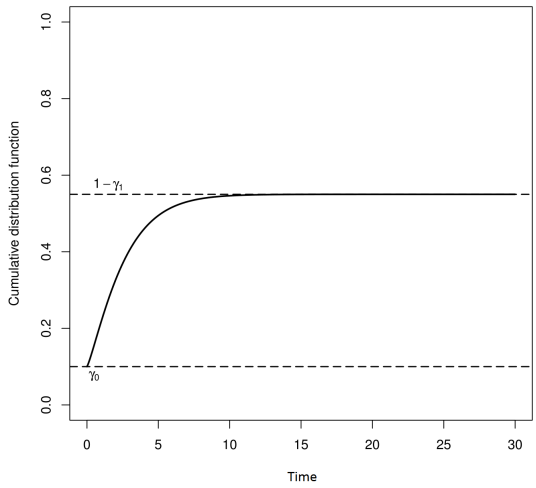

Figure 2.2: The (improper) cumulative distribution function (CDF) of the zero-inflated survival model.

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

$$\lim_{t \to \infty} F(t|p_0, p_1, \boldsymbol{\phi}) = 1 - p_1.$$

2.2.1 Likelihood function

The zero-inflated non-default rate model proposes to distinguish between three sub-populations of banking borrowers: (i) a segment of those who will not honour any instalment of the loan, i.e., STD borrowers with failure time zero; (ii) a segment of those are susceptible to default; and (iii) 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, that is, the borrower has defaulted during the monitoring of the loan; or information about the default time is right censored, that is, either the customer will probably become a defaulter if given enough time, or she is really a good payer and will never default, regardless of the monitoring period term.

So, 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-group of customers. Therefore, the likelihood contribution of each time-to-default t_i, obtained from Section 2.2 and all considerations we have done above, must assume three different values:

$$\begin{cases} p_0, & \text{if the } i\text{-}th \text{ subject is an STD,} \\ (1-p_0-p_1)f_0(t_i|\boldsymbol{\phi}), & \text{if the } i\text{-}th \text{ subject is not censored,} \\ p_1+(1-p_0-p_1)S_0(t_i|\boldsymbol{\phi}), & \text{if the } i\text{-}th \text{ subject is censored.} \end{cases} \tag{2.5}$$

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 $\boldsymbol{\phi}$ denote the parameter vector associated with the f_0 baseline distribution and, finally, (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 $\vartheta=(p_0,p_1,\boldsymbol{\phi})$, is based on a sample of n independent and identically distributed observations, $\mathcal{D}=\{t_i,\delta_i\}$, hence, following Klein & Moeschberger (2003), we write the likelihood function $L(\vartheta;\mathcal{D})$ under non-informative censoring as:

$$L(\vartheta;\mathcal{D})=\prod_{i:\ t_i=0} p_0 \prod_{i:\ t_i>0}\left\{[(1-p_0-p_1)f_0(t_i|\boldsymbol{\phi})]^{\delta_i}[p_1+(1-p_0-p_1)S_0(t_i|\boldsymbol{\phi})]^{1-\delta_i}\right\}. \tag{2.6}$$

2.2.2 Classic parameter estimation

Parameter estimation can be performed by straightforward use of maximum likelihood estimation (MLE), where, as we will see, its simple application is supported by our simulation studies. Hence, the maximum likelihood estimator $\hat{\vartheta}$, regarding to the parameter vector ϑ, are obtained through maximization of $L(\vartheta;\mathcal{D})$ or $\ell(\vartheta;\mathcal{D}) = \log\{L(\vartheta;\mathcal{D})\}$. The MLE is, then, obtained through solving the non-linear system of equations $U(\vartheta) = \dfrac{\partial l(\vartheta)}{\partial \vartheta} = 0$. We use the free software **R** to solve them numerically by means of iterative techniques. There are various routines available to approximate the parameter estimate. We choose the method "BFGS", see details in R Core Team (2015).

Following Migon *et al.* (2014, p. 176), the asymptotic distribution of the maximum likelihood estimator, $\hat{\vartheta}$, is a multivariate normal with mean vector ϑ and covariance matrix $\mathbf{I}^{-1}(\vartheta)$, where $\mathbf{I}(\vartheta) = -E[\partial U(\vartheta)/\partial\vartheta]$ is the Fisher information. Since it is not possible to compute the Fisher information matrix, due to the censored observations, instead of it, according with Migon *et al.* (2014, p. 178), is possible to use the observed information matrix $\mathbf{J}(\vartheta) = \{-\partial^2\ell(\vartheta)/\partial\vartheta\partial\vartheta^T\}$, i.e., is the negative of the second derivative (the Hessian matrix) of the logarithm of the likelihood function, to make large sample inference about $\hat{\vartheta}$. Thereafter, let J^{ii} be the ith diagonal element of the inverse of $\mathbf{J}$, evaluated in $\hat{\vartheta}$. An approximate $100(1-\alpha)\%$ confidence interval for $\hat{\vartheta}_i$, is given by $\left(\hat{\vartheta}_i - z_{\alpha/2}\sqrt{J^{ii}},\ \ \hat{\vartheta}_i + z_{\alpha/2}\sqrt{J^{ii}}\right)$, where z_α denotes the $100(1-\alpha)$ percentile of the standard normal random variable. In the application section we set $\alpha = 0.05$, where we get a 95% confidence interval for each ML estimation.

In the application section, we compare the proposed model configured with different covariates. The comparison of the models was realized through by the criteria for models selection: Akaike Information Criterion (AIC), proposed by Akaike (1974). The criterion is defined by $\text{AIC} = -2\log(L) + 2k$, where k is the number of estimated parameters 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. Another criterion, as the Bayesian information criterion ($\text{BIC} = -2\log(L) + k\log(n)$, where n is the sample size), is also a criterion for model selection among a finite set of models. Due it is closely related to the Akaike information criterion (AIC), we decided to use only the AIC as the selection model criterion, although the BIC is also calculated in some cases in this thesis.

2.2.3 Bayesian parameter estimation

The Bayesian approach has some advantages over classical inference, because it does not depend on the asymptotic theory and codes are easy to be implemented, for instance, through the use of free software as openBUGS or winBUGS. Therefore, in this thesis, in addition to the MLE approach for model parameter estimation, i.e., via classical approach, we also proceed parameter estimation via Monte Carlo Markov Chains (MCMC) algorithms, through use of the software openBUGS. The results obtained from both classical approach and Bayesian approach are compared within the simulation studies and, finally, in the application sections of this thesis.

Considering a Bayesian approach, the parameters are treated as random variable. Hence, a prior distribution $\pi(p_0, p_1, \boldsymbol{\phi})$ for $p_0, p_1, \boldsymbol{\phi}$ must be assigned. The joint posterior distribution for p_0, p_1 and $\boldsymbol{\phi}$ is given by

$$\begin{aligned}\pi(p_0, p_1, \boldsymbol{\phi}|\mathcal{D}) \propto \pi(p_0, p_1, \boldsymbol{\phi}) \prod_{i:\ t_i=0}^{n} p_0 \prod_{i:\ t_i>0}^{n} [(1-p_0-p_1) f_0(t_i|\boldsymbol{\phi})]^{\delta_i} \times \\ \times \prod_{i:\ t_i>0}^{n} [p_1 + (1-p_0-p_1) S_0(t_i|\boldsymbol{\phi})]^{1-\delta_i}.\end{aligned} \tag{2.7}$$

The selection of $\pi(\cdot)$ is not an easy task. In the absence of prior information, we will be interested in specify a prior distribution in which the dominant information in the posterior distribution is provided by the data, such priors are known as non-informative prior. A well-known non-informative prior was introduced by Jeffreys (1939) and can be obtained through the Fisher information matrix. Another important non-informative reference prior was introduced by Bernardo (1979), with further developments by Berger & Bernardo (1989), Berger & Bernardo (1992). However, such priors depend on the Fisher information matrix and even considering common baseline PDF. such as Weibull, Gamma and log-normal, the Fisher information matrix does not have closed form. Therefore, a simple form for the joint prior distribution could be given as $\pi(p_0, p_1, \boldsymbol{\phi}) = \pi(p_1|p_0)\pi(p_0)\pi(\boldsymbol{\phi})$, where p_0 and p_1 are independent of $\boldsymbol{\phi}$. For instance, we could consider an independent Jeffreys prior for $p_0 \propto \frac{1}{\sqrt{p_0(1-p_0)}}$.

Since $p_1 \in (0, 1-p_0)$, one prior distribution for p_1 could be an Uniform$(0, 1-p_0)$. The prior distribution for $\boldsymbol{\phi}$ must be selected depending on the parametric space of $f(t|\boldsymbol{\phi})$. Moreover, conjugate priors should be used when it is possible.

2.2.4 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 also has served as motivation for the proposal of various types of generalizations, see for example, Cooner *et al.* (2007), Rinne (2008), Rodrigues *et al.* (2009), Ortega *et al.* (2012) and Cancho *et al.* (2013). Then, let the Weibull distribution represents the survival behavior of the non-negative random variable T_0. The CDF of the Weibull distribution is given by $F_0(t) = 1 - e^{-\left(\frac{t}{\theta}\right)^{\alpha}}, \quad t \geq 0$, where $\alpha > 0$ and $\theta > 0$ are, respectively, shape and scale parameters. The PDF of the Weibull distribution is obtained as $f_0(t) = \frac{d}{dt}F_0(t) = \frac{\alpha}{\theta}\left(\frac{t}{\theta}\right)^{\alpha-1} e^{\left(-\frac{t}{\theta}\right)^{\alpha}}, \quad t \geq 0$.

The log-likelihood function $\log\{L(\vartheta; \mathcal{D})\}$, corresponding to the observed data, and 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:

$$
\begin{aligned}
\log\{L(\vartheta; \mathcal{D})\} &= \sum_{i:\ t_i=0} \log(p_0) + \sum_{i:\ t_i>0} \log\left\{[(1-p_0-p_1)f_0(t_i)]^{\delta_i}\right\} \\
&\quad + \sum_{i:\ t_i>0} \log\left\{[p_1 + (1-p_0-p_1)S^*(t_i)]^{1-\delta_i}\right\} \\
&= \sum_{i:\ t_i=0} \log(p_0) + \sum_{i:\ t_i>0} \delta_i \log(1-p_0-p_1) \\
&\quad + \sum_{i:\ t_i>0} \delta_i \log[f_0(t_i)] + \sum_{i:\ t_i>0} (1-\delta_i) \log[p_1 + (1-p_0-p_1)S^*(t_i)] \\
&= \sum_{i:\ t_i=0} \log(p_0) + \sum_{i:\ t_i>0} \delta_i \log(1-p_0-p_1) \\
&\quad + \sum_{i:\ t_i>0} \delta_i \log\left[\frac{\alpha}{\theta}\left(\frac{t_i}{\theta}\right)^{\alpha-1} e^{\left(-\frac{t_i}{\theta}\right)^{\alpha}}\right] \\
&\quad + \sum_{i:\ t_i>0} (1-\delta_i) \log[p_1 + (1-p_0-p_1)e^{-\left(\frac{t_i}{\theta}\right)^{\alpha}}]
\end{aligned}
$$

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

$$\frac{\partial l(\vartheta)}{\partial p_1} = -\frac{\sum\limits_{i:\ t_i>0}\delta_i}{(1-p_0-p_1)} + \sum_{i:\ t_i>0}\frac{(1-\delta_i)(1-e^{-\left(\frac{t_i}{\theta}\right)^{\alpha}})}{p_1+(1-p_0-p_1)e^{-\left(\frac{t_i}{\theta}\right)^{\alpha}}}$$

$$\frac{\partial l(\vartheta)}{\partial \alpha} = \sum_{i:\ t_i>0}\delta_i\left[\frac{1}{\alpha}+\left(-\frac{t_i}{\theta}\right)^{\alpha}\log\left(-\frac{t_i}{\theta}\right)+\left(\frac{t_i}{\theta}\right)\right]$$

$$-\sum_{i:\ t_i>0}(1-\delta_i)\frac{(1-p_0-p_1)\left(\frac{t_i}{\theta}\right)^{\alpha}\log\left(\frac{t_i}{\theta}\right)}{1-p_0-p_1+p_1e^{\left(\frac{t_i}{\theta}\right)^{\alpha}}}$$

$$\frac{\partial l(\vartheta)}{\partial \theta} = \sum_{i:\ t_i>0}\delta_i\left[-\frac{\alpha}{\theta}-\frac{\alpha\left(-\frac{t_i}{\theta}\right)^{\alpha}}{\theta}\right]$$

$$+\sum_{i:\ t_i>0}(1-\delta_i)\left[\frac{\alpha(1-p_0-p_1)\left(\frac{t_i}{\theta}\right)^{\alpha}}{\theta\left(1-p_0-p_1+p_1e^{\left(\frac{t_i}{\theta}\right)^{\alpha}}\right)}\right]$$

2.3 The Zero-inflated Non-default Regression Rate Model

Here, we link covariates with the parameters set in the general zero-inflated cure rate model. This allows us to determine the effect of available covariates on the zero-inflated times, on the cure rate and on the observable events. Therefore, we propose to relate the set parameters $\{p_0, p_1, \boldsymbol{\phi}\}$, respectively, proportion of zeros, proportion of cure, the parameters of the baseline hazard distribution, with a set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{k+2}\}$. These covariate vectors, as occurs in practice, may be the same, i.e., $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}_3 = \ldots, \mathbf{x}_{k+2}$.

The regression version of the zero-inflated cure rate regression model are defined by 2.2 up to 2.6, and by the following systematic components:

$$\begin{cases} H(p_{0i}, p_{1i}) = (\zeta_{0i}, \zeta_{1i}), \\ g_j(\phi_{j,i}) = \eta_{ji}, \qquad \text{for } j = 1, \ldots, k \end{cases} \tag{2.8}$$

where $\zeta_{0i} = x_{1i}^{\top}\beta_1$, $\zeta_{1i} = x_{2i}^{\top}\beta_2$ and $\eta_{ji} = x_{(j+2)i}^{\top}\beta_{j+2}$, $j = 1, \ldots, k$ are linear predictors, and β_j's are j+2 vectors of unknown regression coefficients to be estimated. The link function

H and g_j provide the relationship between the linear predictor and the parameters of the distribution function. Following the setting made in Pereira *et al.* (2013), H is set as the multinomial logistic regression (Hosmer & Lemeshow, 2000, p. 261), that is,

$$H(p_{0i}, p_{1i}) = \left(\log\left(\frac{p_{0i}}{1 - p_{0i} - p_{1i}} \right), \log\left(\frac{p_{1i}}{1 - p_{0i} - p_{1i}} \right) \right).$$

Different link functions can be used depending on the support of distribution (McCullagh & Nelder, 1989). For instance, considering the most common lifetime distributions, such as Weibull, Gamma and lognormal the parameters are $\phi_1 > 0$ and $\phi_2 > 0$ and the g_1 and g_2 link functions can be chosen as $g_1(\phi_{1i}) = \log(\phi_{1i})$ and $g_2(\phi_{2i}) = \log(\phi_{2i})$. Therefore, $\phi_{1i} = e^{x_{3i}^\top \beta_3}$ and $\phi_{2i} = e^{x_{4i}^\top \beta_4}$. These are the most convenient links because $g_1(\cdot)$ and $g_2(\cdot)$ are link functions strictly monotonic and twice differentiable that map $\mathbb{R}^+$ into $\mathbb{R}$.

Note that, as required, the component link function H ensures that $0 < \gamma_{0i} < 1$, $0 < \gamma_{1i} < 1$ and $0 < 1 - \gamma_{0i} - \gamma_{1i} < 1$ hold. Indeed, it is always satisfied since $(p_{0i}, p_{1i}) = \left(\frac{e^{x_{1i}^\top \beta_1}}{1+e^{x_{1i}^\top \beta_1}+e^{x_{2i}^\top \beta_2}}, \frac{e^{x_{2i}^\top \beta_2}}{1+e^{x_{1i}^\top \beta_1}+e^{x_{2i}^\top \beta_2}} \right)$. 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 $C = \{(p_{0i}, p_{1i}) | 0 < p_{0i} < 1, 0 < p_{1i} < 1 - p_{0i}\}$ (Pereira *et al.*, 2013, p. 128).

Following Migon *et al.* (2014), as aforementioned, approximate $(1-\alpha)$ 100% confidence intervals for the regression vector parameters, $\boldsymbol{\beta} = \{\beta_j, j = 1, \ldots, k+2\}$, presented in the simulation studies and in the application section, are given by $\widehat{\beta}_j \pm \xi_{\alpha/2} \sqrt{Var(\widehat{\beta}_j)}$, where $\xi_{\alpha/2}$ is the upper $\alpha/2$ percentile of standard Normal distribution and $j = 1, \ldots, k+2$.

In the Bayesian estimation context, with these above assumptions, we can rewrite the posterior distribution expression, (2.7), as

$$\begin{aligned} \pi(\boldsymbol{\beta}|\mathcal{D}, \boldsymbol{x}) \propto \pi(\boldsymbol{\beta}) \prod_{i:\ t_i=0}^{n} \frac{e^{x_{1i}^\top \beta_1}}{1 + e^{x_{1i}^\top \beta_1} + e^{x_{2i}^\top \beta_2}} \prod_{i:\ t_i>0}^{n} \left(\frac{f_0(t_i|\eta(\boldsymbol{\beta}), \boldsymbol{x})}{1 + e^{x_{1i}^\top \beta_1} + e^{x_{2i}^\top \beta_2}} \right)^{\delta_i} \times \\ \times \prod_{i:\ t_i>0}^{n} \left(\frac{e^{x_{1i}^\top \beta_1} + S_0(t_i|\eta(\boldsymbol{\beta}), \boldsymbol{x})}{1 + e^{x_{1i}^\top \beta_1} + e^{x_{2i}^\top \beta_2}} \right)^{1-\delta_i}. \end{aligned} \tag{2.9}$$

The prior distribution adopted for the parameters is

$$\begin{aligned} \pi(\boldsymbol{\beta}) &\propto \pi(\beta_1)\pi(\beta_2)\ldots\pi(\beta_{k+2}), \\ \beta_j &\sim N(0, b_j), \quad \text{for} \quad j = 1, \ldots, k+2 \end{aligned} \tag{2.10}$$

where b_j are larger values that control the variance to produce flat priori.

2.4 Simulation studies

We proceed a parameter estimation based on both classical and Bayesian approach. The estimation via maximum likelihood principle is carried out with the use of the method of maximization "BFGS" of the R routine optim(). In order to assess the behavior of the asymptotic theory for increasing sample size, we performed simulations 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 with increasing sample size as expected.

Regarding to the application of Bayesian approach for parameter estimation, the confidence intervals, with a 95% confidence, were obtained trough the empirical quantiles of the marginal posterior distributions, obtained via straightforward use of algorithms MCMC, with the openBUGS software.

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 in this dissertation. So, 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 simulation, we let x be a random variable that represents a consumer characteristic. The description of sample generation, i.e., all details of the simulated survival time distribution, and results obtained regarding to the proposed estimation method are described in the next sections.

2.4.1 Parameter scenarios

The model parameters are linked on a single covariate x, according to the following expressions: $p_{0i} = \frac{e^{\beta_{10}+x_i\beta_{11}}}{1+e^{\beta_{10}+x_i\beta_{11}}+e^{\beta_{20}+x_i\beta_{21}}}$, $p_{1i} = \frac{e^{\beta_{20}+x_i\beta_{21}}}{1+e^{\beta_{10}+x_i\beta_{11}}+e^{\beta_{20}+x_i\beta_{21}}}$, $\alpha_i = e^{\beta_{30}+x_i\beta_{31}}$, $\theta_i = e^{\beta_{40}+x_i\beta_{41}}$. 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 x as a binary covariate with values drawn from a Bernoulli distribution with parameter 0.5.

For scenario 1, β_{10} assumes -3 and β_{11} assumes 1. β_{20} assumes -2 and β_{21} assumes 0.75. Given that the assumed values of x are 0 and 1, we have that p_0 assumes, respectively, 4.20% and 9.51%, while p_1 assumes 11.41% and 20.15%. Compared to the other scenarios

2 and 3, scenario 1 has the characteristic of having a **low rate of STD and non-default**. Regarding to the Weibull parameters, β_{30} assumes 0.5, β_{31} assumes 0.5, β_{40} assumes 1.5 and β_{41} assumes 2. This implies that the Weibull parameter α can assume 1.64 or 2.71 values, while θ assumes 4.48 or 33.11.

For scenario 2, β_{10} assumes -2 and β_{11} assumes 1.5. β_{20} assumes -1.25 and β_{21} assumes 1. Given that the assumed values of x are 0 and 1, we have that p_0 assumes, respectively, 9.51% and 25.42%, while p_1 assumes 20.15% and 32.64%. Compared to the other scenarios 1 and 3, scenario 2 has the characteristic of having a **moderate rate of STD and non-default**. Regarding to the Weibull parameters, β_{30} assumes -0.5, β_{31} assumes 1.5, β_{40} assumes -0.75 and β_{41} assumes 3. This implies that the Weibull parameter α can assume 0.60 or 2.71 values, while θ assumes 0.47 or 9.48.

For scenario 3, β_{10} assumes -1 and β_{11} assumes 1. β_{20} assumes -1 and β_{21} assumes 1. Given that the assumed values of x are 0 and 1, we have that p_0 assumes, respectively, 21.20% and 33.33%, while p_1 assumes 20.20% and 33.33%. Compared to the other scenarios 1 and 2, scenario 3 has the characteristic of having a **high rate of STD and non-default**. Regarding to the Weibull parameters, β_{30} assumes -0.75, β_{31} assumes 1, β_{40} assumes 1.25 and β_{41} assumes 1. This implies that the Weibull parameter α can assume 0.42 or 1.28 values, while θ assumes 3.49 or 9.48.

The following Kaplan-Meier plots show the survival distinction between the three scenarios set for the regression parameters, trying to simulate, not at all, a range of scenarios consistent with a probable current condition of a real loan portfolio.

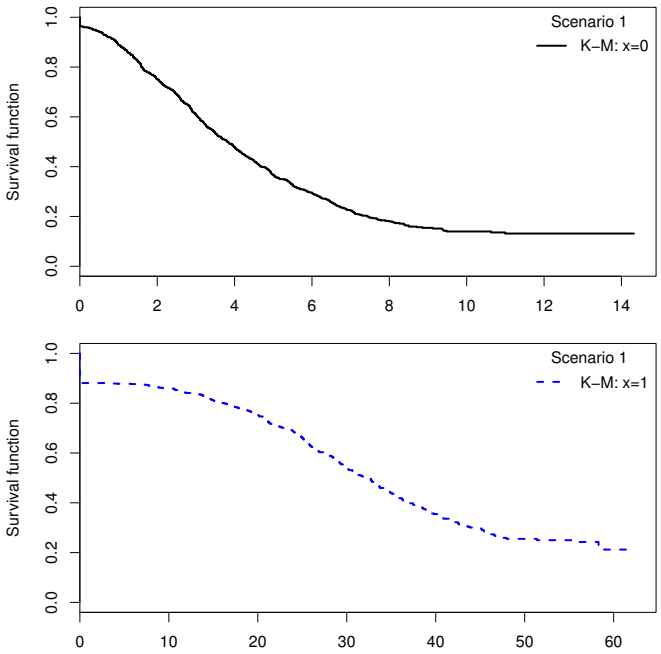

Figure 2.3: Kaplan-Meier (K-M) survival curves of the simulated survival data according to the parameter scenario 1.

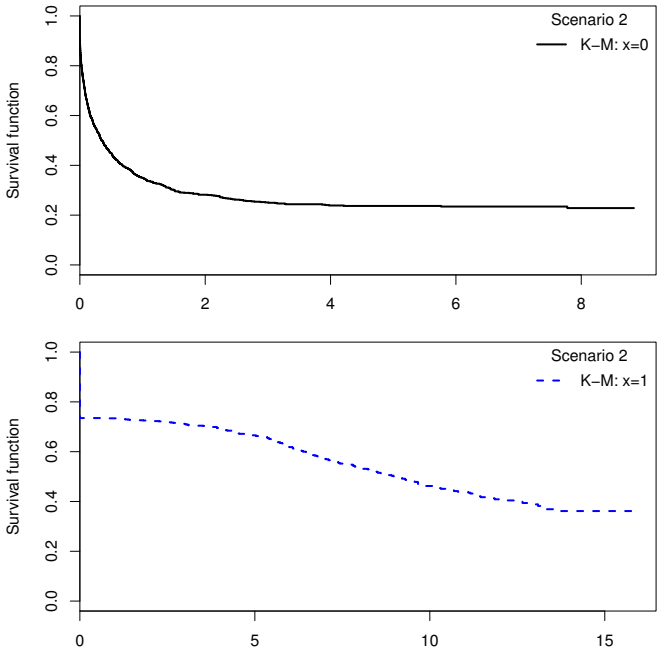

Figure 2.4: Kaplan-Meier (K-M) survival curves of the simulated survival data according to the parameter scenario 2.

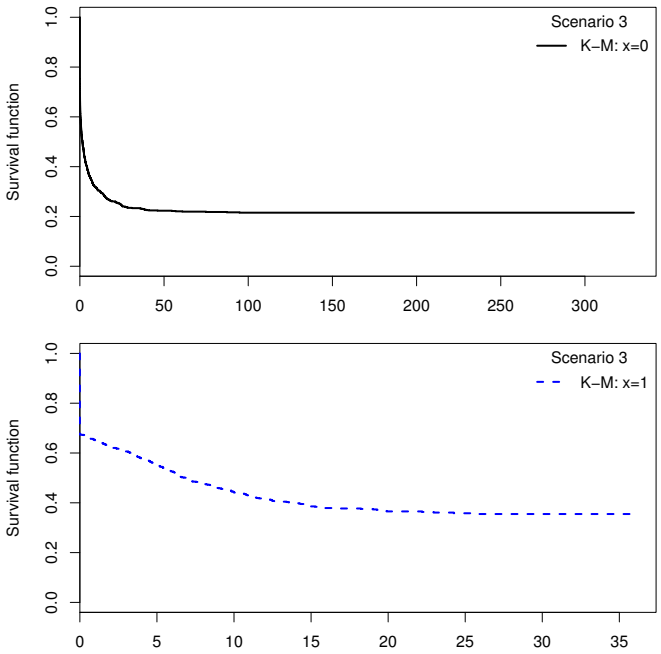

Figure 2.5: Kaplan-Meier (K-M) survival curves of the simulated survival data according to the parameter scenario 3.

The following histograms show the data distribution of the three scenarios set for the regression parameters.

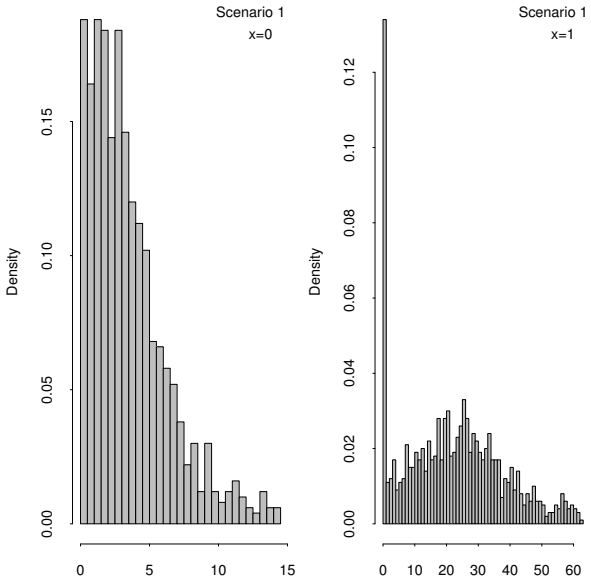

Figure 2.6: Histogram of simulated loan survival data according with the parameter scenario 1.

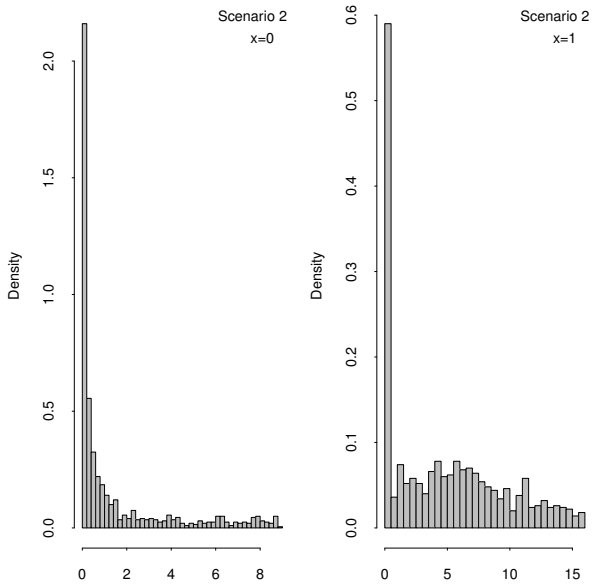

Figure 2.7: Histogram of simulated loan survival data according with the parameter scenario 2.

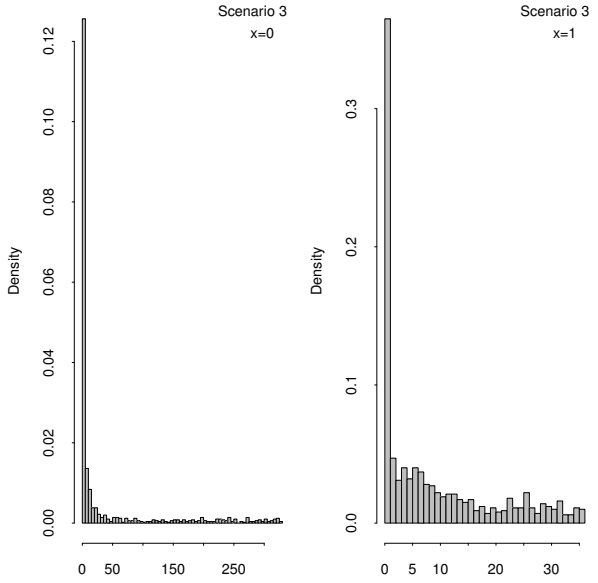

Figure 2.8: Histogram of simulated loan survival data according with the parameter scenario 3.

The description of sample generation, i.e., all details of the simulated survival time distribution, and results obtained regarding to the proposed estimation methods for the model parameters are described in the next sections.

2.4.2 Simulation algorithm

Suppose that the time of occurrence of an event of interest has the improper cumulative distribution function $F(t)$ given by 2.3, 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 2.8, with an x covariate drawn from a Bernoulli distribution with parameter 0.5, representing a consumer characteristic.

1. Set β_{10} and β_{11} related to the value of the desired proportion of zero-inflated times, p_0, along with β_{20} and β_{21} related to the value of the desired non-default fraction, p_1; finally, set the Weibull parameters β_{30} and β_{31} related to α, as well as, β_{40} and β_{41} related to θ;

2. Drawn x_i from $x \sim$ Bernoulli(0.5) and calculate p_{0i}, p_{1i}, α_i 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 U$(p_{0i}, 1 - p_{1i})$ and take s_i as the root of $F(t) - v_i = 0$, where $F(t)$ is given as in 2.3;

7. Generate w_i from a uniform 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 get the desired amount of sample (t_i, δ_i).

Note that the censoring distribution chosen is a uniform distribution with limited range in order to keep the censoring rates reasonable, see Rocha *et al.* (2015), p.12.

2.4.3 Results of Monte Carlo simulations

The followings figures describe the simulation results for the three simulated scenarios of parameters, where the sample size varies as n = 100, 250, 500, 750 and 1000, and

considering both the classical (Figures 2.9, 2.10 and 2.11) and the Bayesian estimation approach (Figures 2.12, 4.7 and 4.8). For the purpose of simulation, we let x be a random variable that represents a consumer characteristic. Hence, the link configuration of the eight parameters (β_{10}, β_{11}, β_{20}, β_{21}, β_{30}, β_{31}, β_{40}, β_{41}) to be estimated is given by the following expressions:

$$\begin{aligned} \gamma_{0i} &= \frac{e^{\beta_{10}+x_i\beta_{11}}}{1+e^{\beta_{10}+x_i\beta_{11}}+e^{\beta_{20}+x_i\beta_{21}}}, \\ \gamma_{1i} &= \frac{e^{\beta_{20}+x_i\beta_{21}}}{1+e^{\beta_{10}+x_i\beta_{11}}+e^{\beta_{20}+x_i\beta_{21}}}, \\ \alpha_i &= e^{\beta_{30}+x_i\beta_{31}}, \\ \theta_i &= e^{\beta_{40}+x_i\beta_{41}}. \end{aligned} \tag{2.11}$$

The parameter values are selected in order to assess the ML estimation performance under different shape and scale parameters (β_{30}, β_{31}, β_{40} and β_{41}, related to the Weibull time-to-default distribution), and also under a composition of different proportions of zero-inflated data (β_{10} and β_{11}) and non-defaulters rates (β_{20} and β_{21} related to censored data). It can be seen from the Figures 2.9 to 4.8, that:

1. in general, the maximum likelihood estimation on average, MLEA, is close to the parameters set in the simulated parameter scenarios, see Figure 2.11. However, in scenarios 1 and 2, the parameters $\widehat{\beta}_{11}$ and $\widehat{\beta}_{21}$ need a larger sample size (from at least n=500 for β_{21}) to achieve convergence.

2. in general, according to Figures 2.10 and 2.11, biases and root mean square errors decrease as 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), due, of course, to greater presence of zeros and censored data;

4. on the other hand, in the scenario with the fewer 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, performs better compared to others scenario, due to greater presence of observed time-to-default data;

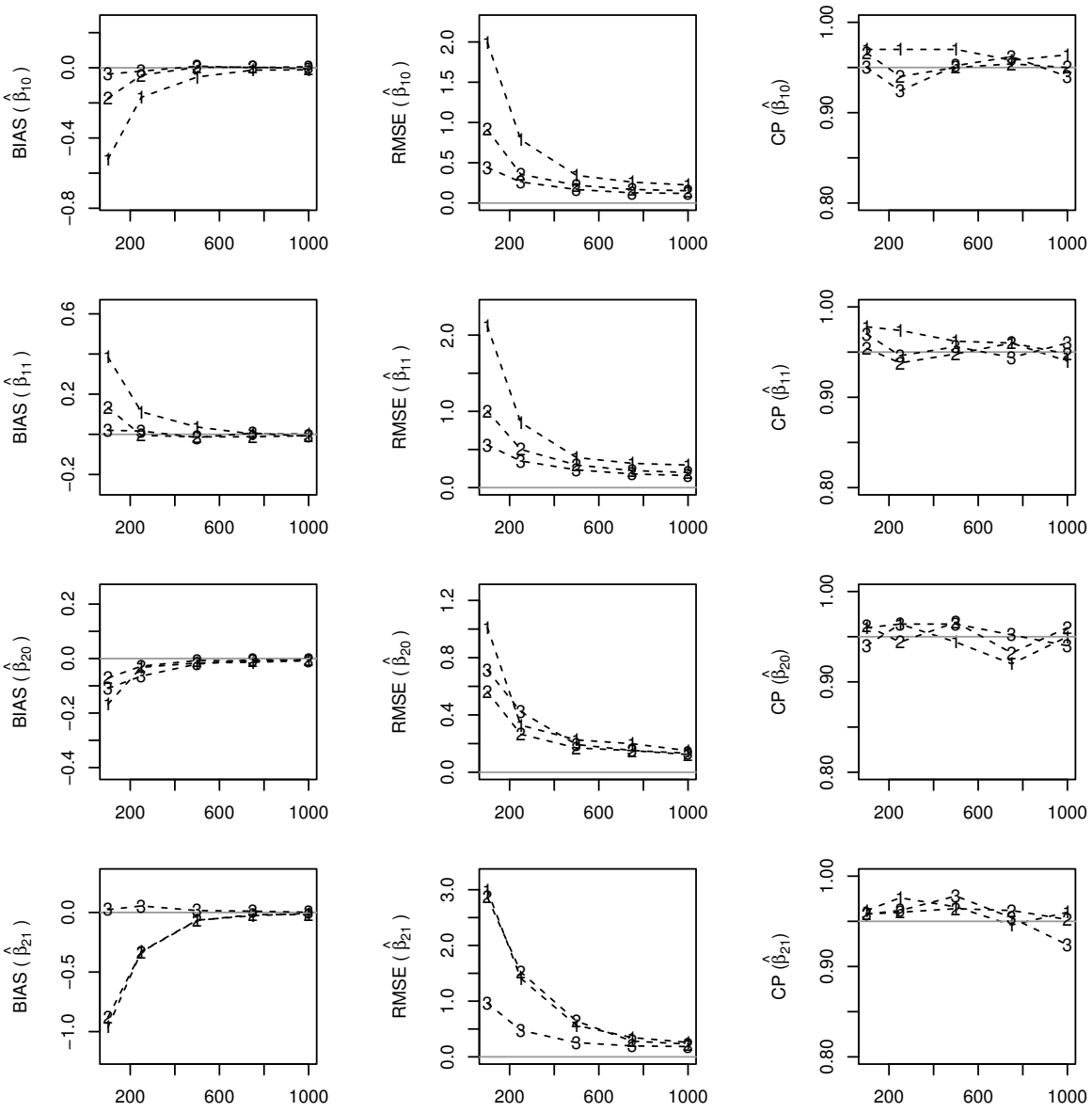

Figure 2.9: Bias, square root of mean squared error and coverage probability (CP) of **the maximum likelihood estimation** ($\widehat{\beta}_{10}$, $\widehat{\beta}_{11}$, $\widehat{\beta}_{20}$, $\widehat{\beta}_{21}$) of zero-inflated Weibull non-default rate regression model for simulated data under the three scenarios of parameters, obtained from Monte Carlo simulations with 1000 replications and increasing sample size (n).

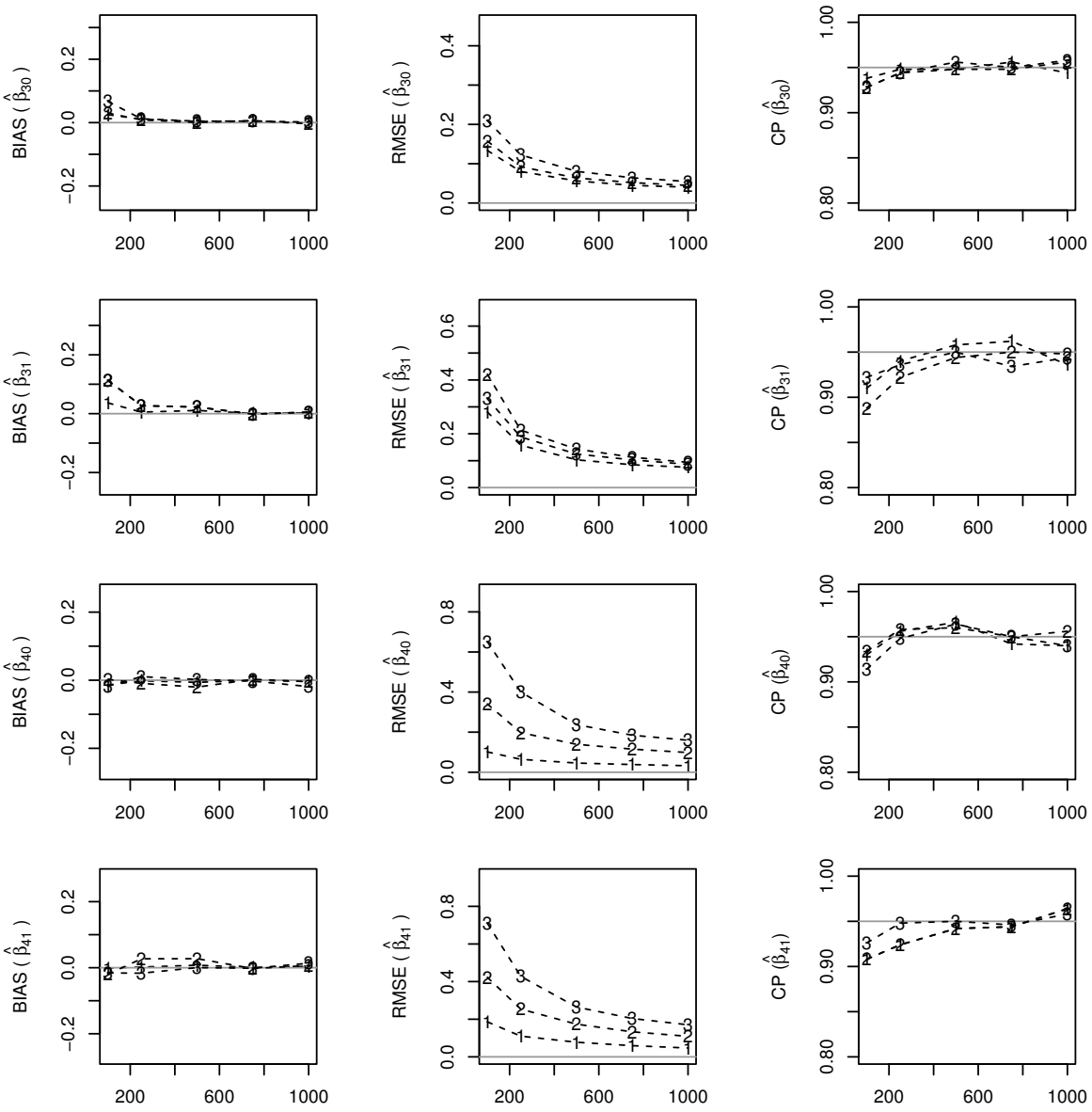

Figure 2.10: Bias, square root of mean squared error and coverage probability (CP) of **the maximum likelihood estimation** ($\widehat{\beta}_{30}$, $\widehat{\beta}_{31}$, $\widehat{\beta}_{40}$, $\widehat{\beta}_{41}$) of zero-inflated Weibull non-default rate regression model for simulated data under the three scenarios of parameters, obtained from Monte Carlo simulations with 1000 replications and increasing sample size (n).

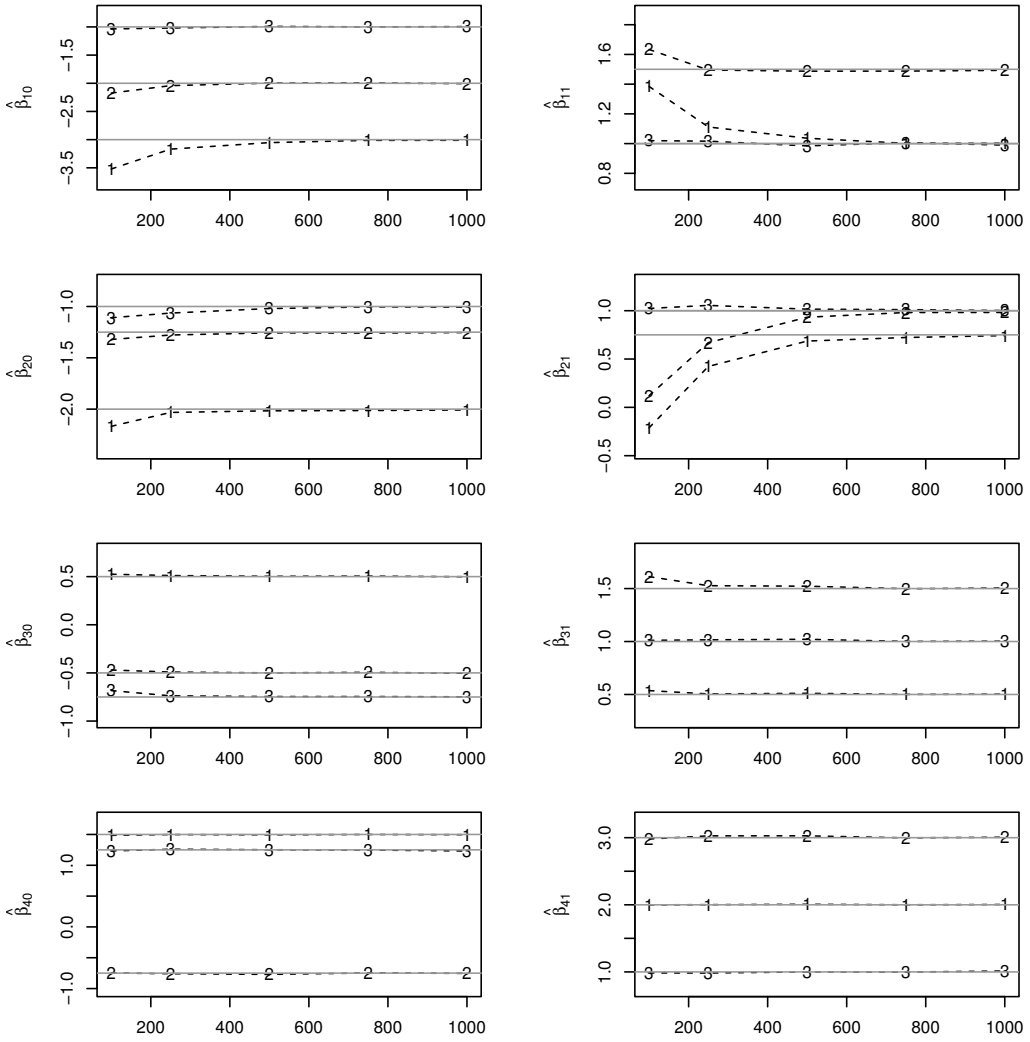

Figure 2.11: MLEA, **maximum likelihood estimation** on average of the parameters ($\widehat{\beta}_{10}$, $\widehat{\beta}_{11}$, $\widehat{\beta}_{20}$, $\widehat{\beta}_{21}$, $\widehat{\beta}_{30}$, $\widehat{\beta}_{31}$, $\widehat{\beta}_{40}$), $\widehat{\beta}_{41}$ of zero-inflated Weibull non-default rate regression model for simulated data under the three scenarios of parameters, obtained from Monte Carlo simulations with 1000 replications and increasing sample size (n).

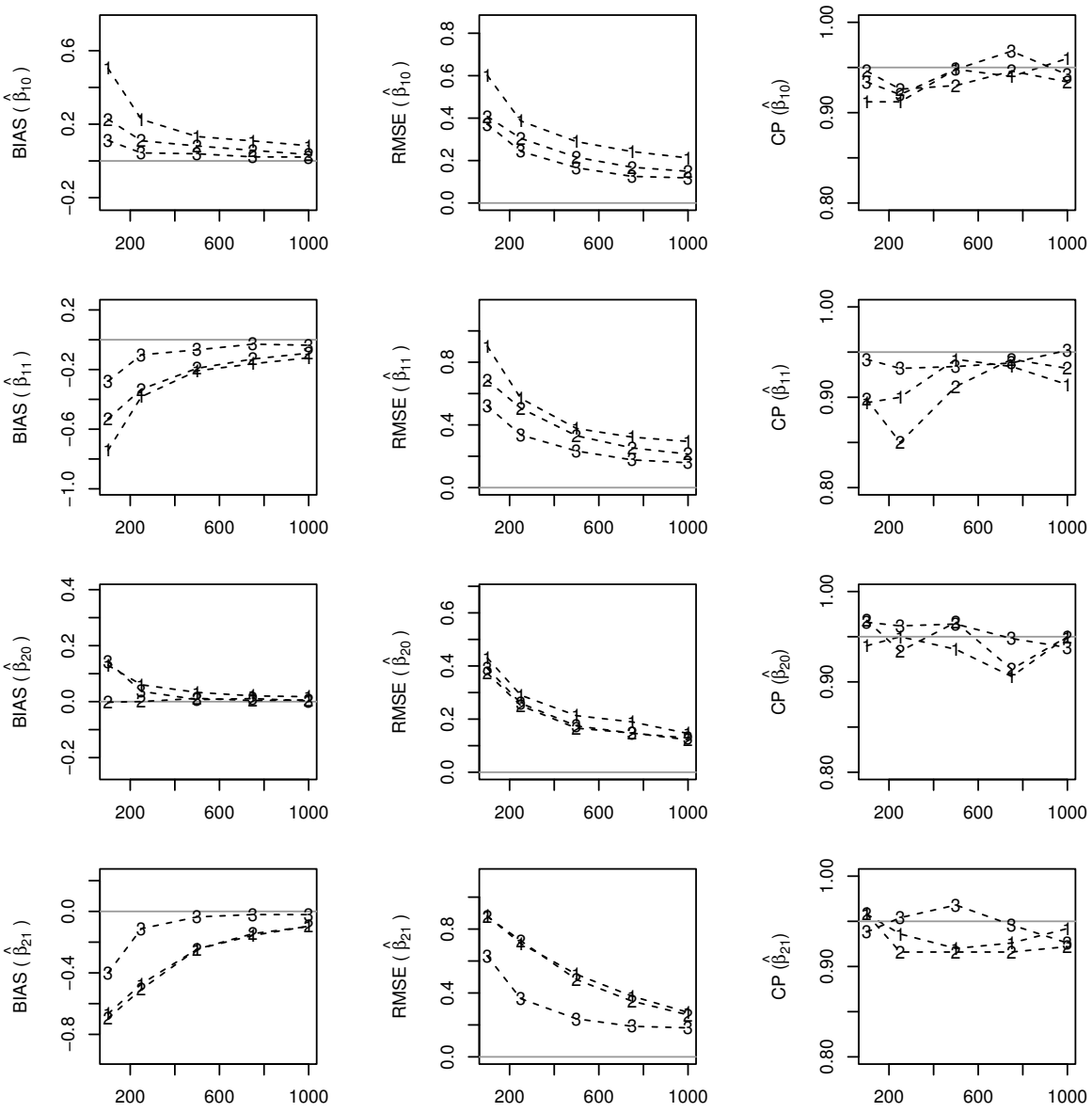

Figure 2.12: Bias, square root of mean squared error and coverage probability (CP) of **the Bayesian parameter estimations** ($\widehat{\beta}_{10}$, $\widehat{\beta}_{11}$, $\widehat{\beta}_{20}$, $\widehat{\beta}_{21}$) of zero-inflated Weibull non-default rate regression model for simulated data under the three scenarios of parameters, obtained from Monte Carlo simulations with 1000 replications and increasing sample size (n).

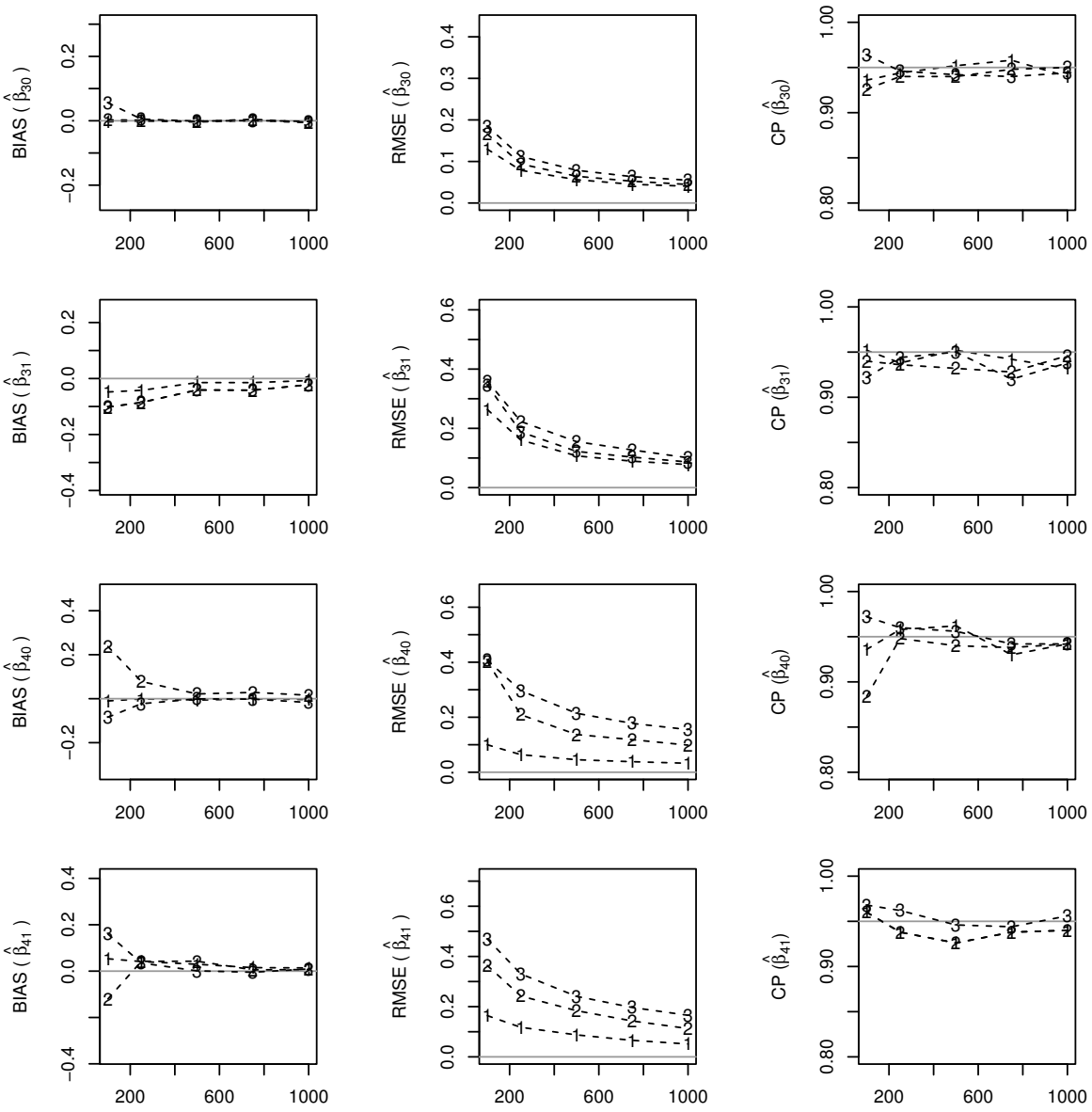

Figure 2.13: Bias, square root of mean squared error and coverage probability (CP) of **the Bayesian parameter estimations** ($\widehat{\beta}_{30}$, $\widehat{\beta}_{31}$, $\widehat{\beta}_{40}$, $\widehat{\beta}_{41}$) of zero-inflated Weibull non-default rate regression model for simulated data under the three scenarios of parameters, obtained from Monte Carlo simulations with 1000 replications and increasing sample size (n).

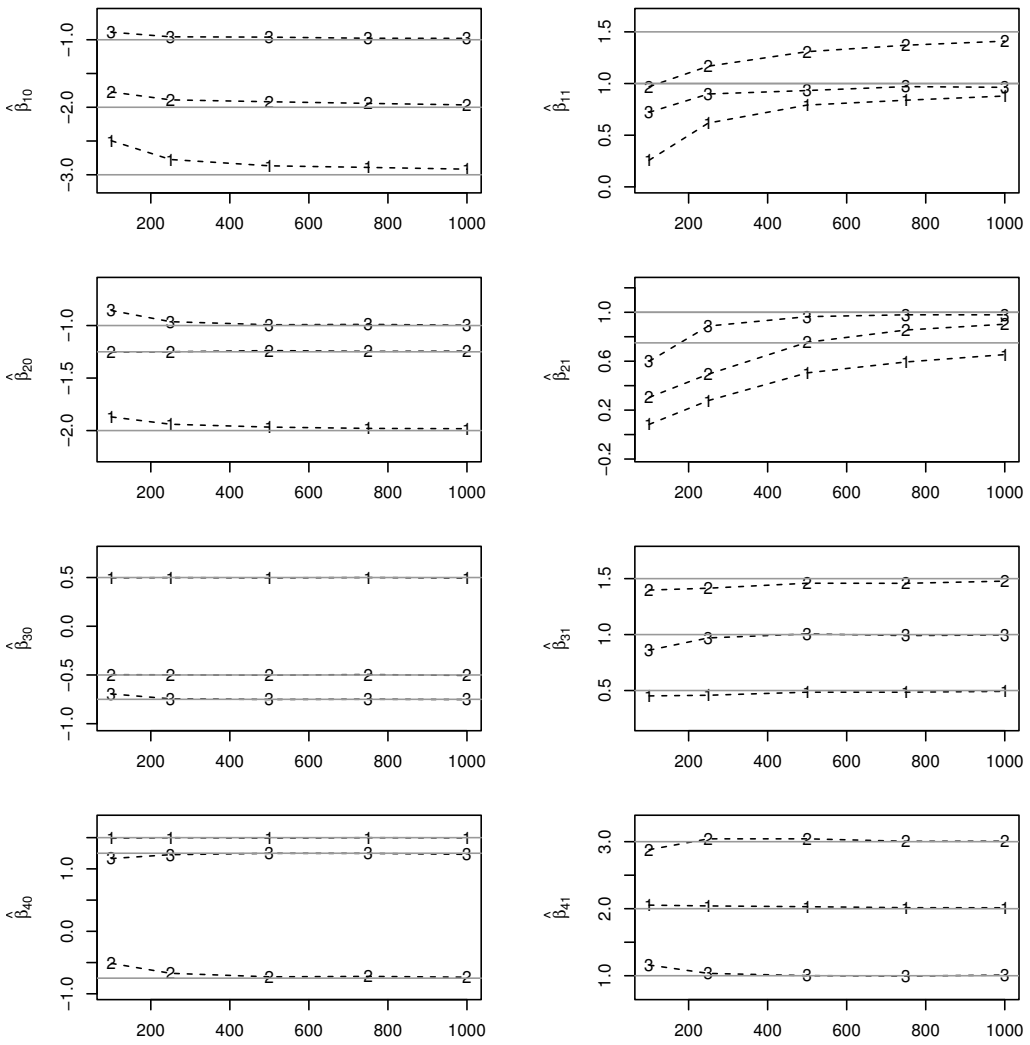

Figure 2.14: **The Bayesian parameter estimations** on average of the parameters ($\widehat{\beta}_{10}$, $\widehat{\beta}_{11}$, $\widehat{\beta}_{20}$, $\widehat{\beta}_{21}$, $\widehat{\beta}_{30}$, $\widehat{\beta}_{31}$, $\widehat{\beta}_{40}$, $\widehat{\beta}_{41}$) of zero-inflated Weibull non-default rate regression model for simulated data under the three scenarios of parameters, obtained from Monte Carlo simulations with 1000 replications and increasing sample size (n).

2.5 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 bank. 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.

As described in the section 1.1.1, the portfolio was collected from customers who have

taken a personal loan over a 60-month period, between the years 2010 and 2015. It is composed of 5733 time-to-default (in months), with an approximate 80% rate of censored data, that is, a high rate of non-default loans. In order to proceed the model fit, we have considered dummy covariates for all levels of the available covariates. So, including all the intercepts, we might have up to thirty two (32=4x4x2) regression parameters to be estimated.

Henceforth, we are concerned whether the use of covariates explains better the distribution of the time-to-default than assuming that the observations are identically distributed. The fitted model without any covariate has AIC of 12768.71 (and BIC of 12795.33, $l\{\hat{p_0}, \hat{p_1}, \hat{\alpha}, \hat{\theta}\} = -6380.355,\ p = 4$) and the model with all the dummy covariates has AIC of 12809.21 (and BIC of 13022.14, $l\{\hat{p_0}, \hat{p_1}, \hat{\alpha}, \hat{\theta}\} = -6372.604,\ p = 32$). To reach the final model, variables were selected in a backward way using the p-values of the Wald test and AIC. The final model is summarized in Table 2.1, which has AIC of 12602.52 (and BIC of 12669.06, $l\{\hat{\beta}_{10}, \hat{\beta}_{11}, \hat{\beta}_{12}, \hat{\beta}_{13}, \hat{\beta}_{20}, \hat{\beta}_{21}, \hat{\beta}_{22}, \hat{\beta}_{30}, \hat{\beta}_{40}, \hat{\beta}_{41}\} = -6291.259,\ p = 10$).

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

Table 2.1: The Zero-Inflated Non-default Regression Model for time-do-default on a Brazilian Bank Loan Portfolio. Notes: (1) Related regression parameter to be estimated; (2) Standard error; (3) Exp(estimated parameter).

Based on the last column of Table 2.1, 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 1.3), can again be ratified. For example, the odds of a customer within the age group 4 be an STD borrower, decreases by 56%, with all other independent covariates held constant. 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 by 22,5% the odds of non-default on the loan,

while the group within type of residence 4 increases by 139% the odds of being an STD customer, with all other independent covariates held constant.

The selected dummy covariates (Table 2.1) allowed us to split the portfolio within twelve (12) different group of borrowers (segmentations). Next, we present the estimated survival curves for the most representative group of borrowers (5544 out of 5733), considering the following segmentation: the **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; the **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; the **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; the **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; and, finally, the **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.

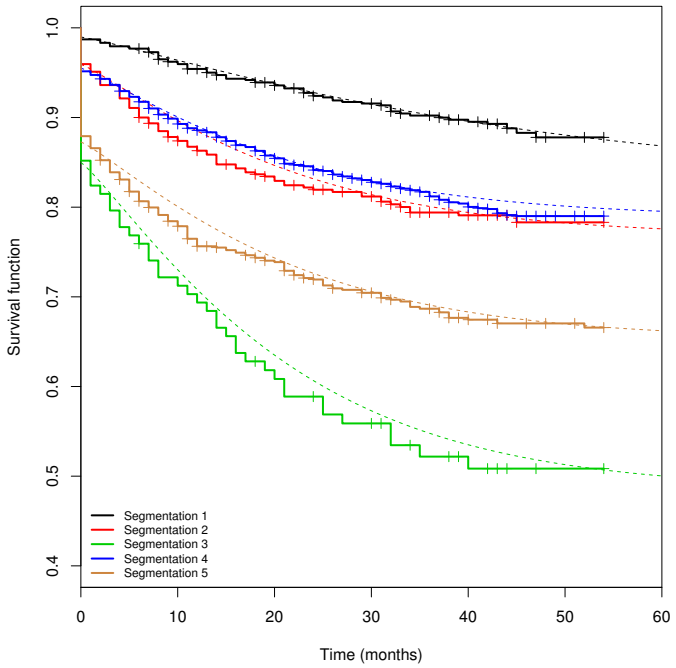

Figure 2.15: Brazilian bank loan portfolio. Kaplan-Meier survival curves stratified through the covariate selection given by the final model presented in the Table 2.1.

For the Bayesian parameter estimation, were adopted the *priori* distributions as denoted in (2.3). Hence, $\beta_j \sim N(0, b_j)$, for $j = 1, \ldots, 10$. Using the openBUGS, has been generated a chain of size 2400, with the first 800 discarded as *burn-in* iterations, and

considered n.this set as 60. Follow, we present the 95% credible interval for each estimation along with the punctual estimation, which was set as the mode of the *posterior* generated samples. Appendix A presents the convergence plots for the estimated parameters, i.e., the trace for the estimated parameters via MCMC algorithm, the approximate marginal density *a posteriori* and, finally, the autocorrelation function plot for each parameter.

The final Bayesian model is summarized in Table 3.2, which has AIC of 12602.65 ($l\{\hat{\beta}_{10}, \hat{\beta}_{11}, \hat{\beta}_{12}, \hat{\beta}_{13}, \hat{\beta}_{20}, \hat{\beta}_{21}, \hat{\beta}_{22}, \hat{\beta}_{30}, \hat{\beta}_{40}, \hat{\beta}_{41}\} = -6291.324,\ p = 10$).

Parameter	Estimated parameter	95% Credible interval Lower	Upper
$\hat{\beta}_{10}$	-0.6769	-0.9216	-0.4212
$\hat{\beta}_{11}$	-0.8208	-1.2390	-0.4030
$\hat{\beta}_{11}$	0.7411	-0.3084	1.5210
$\hat{\beta}_{12}$	-0.6335	-0.9483	-0.3383
$\hat{\beta}_{20}$	0.9146	0.7096	1.1170
$\hat{\beta}_{21}$	-0.2706	-0.4795	-0.0601
$\hat{\beta}_{22}$	0.6370	0.4312	0.8371
$\hat{\beta}_{30}$	0.1452	0.0697	0.2177
$\hat{\beta}_{40}$	3.0984	2.9849	3.2510
$\hat{\beta}_{41}$	0.6974	0.4006	0.9749

Table 2.2: Parameters obtained via Bayesian estimation using the software openBUGS.

We note that the Bayesian estimated parameters are close to the results obtained from the classical approach, i.e., via maximum likelihood estimation. For this reasons, in this application section, we have discussed only the practical outcomes obtained by the MLE approach, since the parameter estimation are quite similar with the Bayesian ones.

2.6 Conclusion

We have presented a methodology in which we modify the standard cure rate model introduced by Berkson & Gage (1952) to a credit risk setting. It allowed 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 improper 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 instalments 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 who are equal to zero. To illustrate the proposed method, a 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 showed satisfactory.

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

Finally, we pointed out the importance of the jointly 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, those who are more likely to miss their payments at the beginning of the relationship with the bank.

Chapter 3

The zero-inflated promotion cure rate model

In this chapter we extend the promotion cure rate model studied in Yakovlev & Tsodikov (1996) and Chen *et al.* (1999), by incorporating excess of zeros in the modeling. Despite allowing to relate covariates to the fraction of cure, the current approach, which is based on a biological interpretation of the causes that trigger the event of interest, does not enable to relate covariates to the fraction of zeros. The presence of zeros in survival data, unusual in medical studies, can frequently occur in banking loan portfolios, as presented in the earlier chapter 2, where we dealt with propensity to credit risk in lending loans in a major Brazilian bank. To illustrate the new cure rate survival method, the same real dataset analysed in the chapter 2 is again fitted and the results are compared.

3.1 Introduction

The cure rate model has overcome the disadvantage of the standard survival model using for loan credit risk analysis, where there are individuals who are not susceptible to the occurrence of the event of interest. This problem was handled in Berkson & Gage (1952), where the authors proposed a simple model that add the fraction of cured ($p > 0$) into the survival analysis, getting the following expressions for the survival and density

functions:

$$S(t) = p + (1-p)S_0(t), \qquad t \geq 0, \tag{3.1}$$

$$f(t) = (1-p)f_0(t), \qquad t \geq 0, \tag{3.2}$$

where S_0 is the baseline survival function of the subjects susceptible to failure, f_0 is its density probability function, and p is the proportion of subjects immune to failure (cured). This model is called cure rate model, or long-term survival model. S is an improper survival function, unlike S_0, since it satisfies: $\lim_{t\to\infty} S(t) = p > 0$.

The advantage of the cure rate model is that it allows to associate covariates in both parts of the model, i.e., it allows covariates to have different influence on cured patients, linking them with p, and on patients who are not cured, linking them with parameters of the proper survival function S_0.

To accommodate the presence of zero excess, which is impossible in the cure rate model, in the earlier chapter 2, we proposed a zero-inflated cure rate model, with survival function given by:

$$S(t) = p_1 + (1 - p_0 - p_1)S_0(t), \qquad t \geq 0, \tag{3.3}$$

where, S_0 is the survival function related to the $(1 - p_0 - p_1)$ proportion of subject susceptible to failure, p_0 is the proportion of zero-inflated survival times, and p_1 is the proportion of subjects immune to failure (cured or long-term survivors).

Thus, it is now possible to link together the influence of the covariates in the three parts of the model, i.e., to the proportion of zero-inflated survival times, whose we have identified in a credit risk context as borrowers who do not pay any instalment after the loan approval, along with the usual sub-populations of susceptible and non-susceptible to the event of interest. As we will see in the application section, the event of interest concerned here is related to the time until the occurrence of default on bank loan portfolios.

The fact that differentiates the zero-inflated cure version from the standard cure approach is highlighted in the second of the following satisfied properties:

$$\lim_{t\to\infty} S(t) = p_1 > 0. \tag{3.4}$$

$$S(0) = 1 - p_0 < 1. \tag{3.5}$$

Note that, if $p_0 = 0$, i.e., without the excess of zeros, we have the cure rate model of

Berkson & Gage (1952).

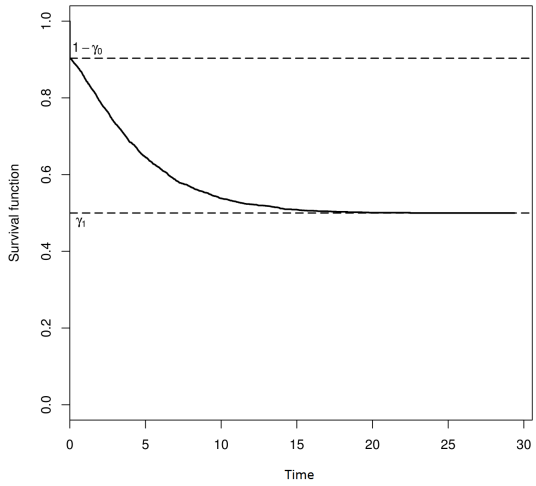

Figure 3.1: Survival function of the zero-inflated cure rate model as presented in Louzada *et al.* (2015).

3.1.1 Preliminaries

In this section we shall briefly describe the promotion cure rate model studied in Yakovlev & Tsodikov (1996) and Chen *et al.* (1999), further extended by Rodrigues *et al.* (2009) among others authors, wherein we follow the same notations. This model also incorporates the presence of immune individuals to the event of interest, but still has the disadvantage of not accommodating zero time excess in its framework.

This survival model with fraction of cure, according to Chen *et al.* (1999), is based on a biological interpretation of the causes that trigger (promote) a cancer disease relapse. As described by the authors, the process that leads to formation of a detectable cancer mass is triggered by a set of N competitive underlying causes, biologically represented by the number of carcinogenic cells that the individual has left active after the initial treatment. In their paper, it is assumed that N follows a Poisson distribution with mean θ.

Regarding to the time until the relapse of the cancer under treatment, the authors Chen *et al.* (1999) have let Z_i be the random time for the ith carcinogenic cells to produce a detectable cancer mass, i.e., the incubation time for the ith (out of N) carcinogenic cell. The random variables Z_i, $i = 1, 2, \cdots$, are assumed to be iid, with a common distribution function $F(t) = 1 - S(t)$, and are independent of N.

In order to include those individuals who are not susceptible to the event of cancer

relapse, i.e, the individuals with the initial number of cancer cells, N, equals to 0 and, theoretically, with infinity survival time, it is assumed that $P(Z_0 = \infty) = 1$.

Finally, the time to the relapse of cancer is defined by the random variable $T = min\{Z_i, 0 \leq i \leq N\}$, and therefore, the survival function of T, for the entire population, is given by

$$\begin{aligned} S_p(t) &= P(T > t, N \geq 0) \\ &= P(N = 0) + P(Z_1 > t, \cdots, Z_N > t, N \geq 1) \\ &= \exp(-\theta) + \sum_{k=1}^{\infty} S(t)^k \frac{\theta^k}{k!} \exp(-\theta) \\ &= \exp(-\theta + \theta S(t)) = \exp(-\theta F(t)). \end{aligned} \tag{3.6}$$

The density function corresponding to 3.6 is given by

$$f_p(t) = -\frac{d}{dt} S_p(t) = \theta f(t) \exp(-\theta F(t)), \tag{3.7}$$

We notice that, S_p and f_p are not, properly, survival function and density function, respectively. In fact, note that, $P(Z_0 = \infty) = 1$, leads to the cure proportion

$$\lim_{t \to \infty} S_p(t) \equiv S_p(\infty) \equiv P(N = 0) = \exp(-\theta) > 0,$$

which comes from the population of individuals who are not susceptible to the occurrence of cancer relapse (cured). Moreover, the fraction of cure is very flexible, i.e., it has the property to accommodate a wide variety of cases, since as $\theta \to \infty$, the proportion of cured tends to 0, whereas as $\theta \to 0$, the proportion of cured tends to 1.

In the situation where we consider the model formulation taking into account only susceptible individuals, that is, when it is present in all individuals a number of initial cancer cells greater than zero, $N \geq 1$, we have a slightly modified expression for the survival function:

$$S_p^*(t) = P(T > t, N \geq 1) = \frac{\exp(-\theta F(t)) - \exp(-\theta)}{1 - \exp(-\theta)}. \tag{3.8}$$

According to this formulation, we figure out now that $S_p^*(t)$ is a proper survival function, since the following conditions are satisfied: $S_p^*(0) = 1$ and $S_p^*(\infty) = 0$. Still following the model presentation as done by Chen *et al.* (1999), we come to the probability density

function of individuals who are susceptible to recurrence of the considered event:

$$f_p^*(t) = -\frac{d}{dt}S^*(t) = \left(\frac{\exp(-\theta F(t))}{1-\exp(-\theta)}\right)\theta f(t). \tag{3.9}$$

Finally, we come to the mathematical relation between the cure rate model, as presented by Berkson & Gage (1952), see expression (3.1), and the biological based model studied by Chen *et al.* (1999), among others, in the expression (3.6):

$$S_p(t) = \exp(-\theta) + (1-\exp(-\theta))S_p^*(t), \qquad t \geq 0, \tag{3.10}$$

$$f_p(t) = (1-\exp(-\theta))f_p^*(t), \qquad t \geq 0, \tag{3.11}$$

where, S_p^* and f_p^* are the proper survival function and the proper density function as given in 3.8 and 3.9, respectively. Thus, we see that Chen *et al.* (1999) model can be rewritten as a cure rate model, with cure rate equal to $p = \exp(-\theta)$.

Although the promotion model be formulated within a biological context, it has also been applied in other areas, such as credit risk analysis of bank loan portfolios. In these new developments, the number N is related to the number of risks that compete to the occurrence of a particular financial event of interest, i.e., default or non-performing of loans. Therefore, the formulation admits generalizations in various ways, such as done in Barriga *et al.* (2015), where the authors studied the time until the event of default on a Brazilian personal loan portfolio, and where the authors let N follows a geometric distribution, and $F(t)$ be a cumulative density function of the inverse Weibull distribution.

Also in the area of credit risk modeling, in Oliveira & Louzada (2014b) the authors applied the model given by (3.6) to study the time until the full recovery of non-performing loans in a portfolio of personal loans.

In Oliveira & Louzada (2014a) the authors compare the parameters θ obtained from two follow-up studies of a set of non-performing loans. The first follow-up is related to the time until the default occurrence, while the second one is related to the time until the full recovery of the related loan. The authors found a significant relationship between default and recovery processes. The paper suggests that in times of higher risk of default, it is also likely to have a decrease in the recovery rates of non-performing loans.

Identifiability issues of the cure rate model in (3.1) and the promotion cure model (3.6) are discussed in Li *et al.* (2001). According to Mateluna (2014), the authors concluded

that in both cases, it is necessary to include covariates in the cure fractions to make them identifiable. From Peng & Zhang (2008), it can be ensured identifiability for the promotion cure model when covariates are included on both parameters related to the fraction of susceptible and the fraction of cured individuals, see (Mateluna, 2014, p. 28).

Although we have included one more parameter in both models mentioned above, identifiability issues will not be discussed in this thesis. This important subject is intended to be addressed in future research.

3.1.2 Proposal

To accommodate zero excess in a survival analysis of loan portfolios, in the earlier chapter 2 we have proposed a modification in the survival function of the cure rate model, which has led to the improper survival function given in 3.3, also labelled as zero-inflated cure rate model. In this scenario, information from credit risk in loan applications is exploited through the joint modeling of the zero survival times, along with the survival times of the remaining group of borrowers.

The purpose of this chapter 3 is to propose a way of incorporating the fraction of zeros into the biological based promotion cure model. Such an approach leads the credit risk management to a complete overview of the risk factors involved in lending, that is, dealing with likelihood to default on a loan since the loan approval, the non-performing loan control and ensure customer loyalty among long-term survival customers. To exemplify the application of the proposed approach, we re-analyse the portfolio of loans made available by a large Brazilian commercial bank that had been studied in the earlier chapter.

This chapter is organized as follows. In Section 3.2, we formulate the new model named zero-inflated promotion cure rate, where we present the approach for parameter estimation. Simulation studies are presented in the Section 3.3. An application to a real data set is presented in Section 3.4. Some general remarks are presented in Section 3.5.

3.2 Model specification

In what follows, we consider the promotion cure rate model as defined in expression (3.10). Hence, we propose a new (improper) survival function as follows:

$$S_p(t) = p_1 + (1 - p_0 - p_1)S_p^*(t), \qquad t \geq 0, \tag{3.12}$$

where S_p^* is given by 3.8, and the parameters p_0 and p_1 are defined as follows: $p_0 = \exp(-\kappa)$ and $p_1 = \exp(-\theta)$, with $\kappa > 0$ and $\theta > 0$.

To ensure that p_0, p_1, and $(1 - p_0 - p_1) \in (0, 1)$, we propose to link two vector of covariates, x_1 and x_2, into the parameters related to zero inflation and cure rate, respectively, as follows: $\kappa_i = -\log\left(\frac{e^{x_{1i}^\top \beta_1}}{1 + e^{x_{1i}^\top \beta_1} + e^{x_{2i}^\top \beta_2}}\right)$, and $\theta_i = -\log\left(\frac{e^{x_{2i}^\top \beta_2}}{1 + e^{x_{1i}^\top \beta_1} + e^{x_{2i}^\top \beta_2}}\right)$, where β_1, is a vector of regression coefficients to be estimated, that relates the influence of the covariates into the excess of zeros, while β_2, is a vector of regression coefficients that relates the influence of the covariates into the fraction of cured.

To complete the configuration of the model, i.e., to determine the parametric form of S_p^*, we let $f(t)$ and $F(t)$ be, respectively, the density probability function and the cumulative probability function of the Weibull distribution. The Weibull distribution is a continuous probability distribution, commonly applied in survival analysis and reliability. It has two parameters, $\alpha_1 > 0$ and $\alpha_2 > 0$, respectively, the shape and scale parameters. Therefore, we link the Weibull parameters as follows: $\alpha_{1i} = e^{x_{3i}^\top \beta_3}$ and $\alpha_{2i} = e^{x_{4i}^\top \beta_4}$. These are the most convenient links, as mentioned in the section 2.3, for non-negative parameters. Finally, we present the following framework for the zero inflated promotion cure rate model:

$$\begin{aligned}
S_p(t) &= \exp(-\theta) + (1 - \exp(-\kappa) - \exp(-\theta))S_p^*(t), \\
S_p^*(t) &= \tfrac{\exp(-\theta F(t)) - \exp(-\theta)}{1 - \exp(-\theta)}, \\
f_p^*(t) &= \left(\tfrac{\exp(-\theta F(t))}{1 - \exp(-\theta)}\right)\theta f(t), \\
F(t) &= 1 - e^{-\left(\frac{t}{\theta}\right)^\alpha} \text{ and} \\
f(t) &= \tfrac{\alpha}{\theta}\left(\tfrac{t}{\theta}\right)^{\alpha-1} e^{\left(-\frac{t}{\theta}\right)^\alpha}.
\end{aligned} \tag{3.13}$$

3.2.1 Likelihood function

The construction of the likelihood function follows the same logic presented in the previous chapter 2. Thus, regarding to the likelihood contribution of each customer for the likelihood function, we must note that there are different sub-group of customers: (i) individual with event at the starting time (zero time); (ii) non-susceptible for the event, or (iii) susceptible for the event. The expression (3.14) presents the likelihood contribution of each time-to-default t_i:

$$\begin{cases}
p_{0i}, & \text{if } t_i = 0, \\
(1 - p_{0i} - p_{1i}) f_p^*(t_i), & \text{if } t_i \text{ is fully observed} \\
p_{1i} + (1 - p_{0i} - p_{1i}) S_p^*(t_i), & \text{if } t_i \text{ is right censored.}
\end{cases} \tag{3.14}$$

Let the data take the form $\mathcal{D} = \{t_i, \delta_i, x = \{x_{1i}, x_{2i}, x_{3i}, x_{4i}\}\}$, where $\delta_i = 1$ if t_i is an observable time to default, $\delta_i = 0$ if it is right censored, for $i = 1, 2, \cdots n$, and x is vector of covariates associated with a consumer i. As we shall see in the application section, the covariate vectors can be the same, i.e., $x_1 = x_3 = x_2 = x_4$. Let (α_1, α_2) denote the parameter vector of the Weibull distribution and, finally, let $(\beta_\kappa, \beta_\theta)$ be the regression parameters associated, respectively, with the proportion of inflation of zeros and the proportion of long-term survivors (cure rate).

The likelihood function of the proposed new zero-adjusted cure rate survival model, with a parameter vector, $\vartheta = (\alpha_1, \alpha_2, \beta_\kappa, \beta_\theta)$, to be estimated via MLE approach or Bayesian approach, likewise we have proceeded as in sections 2.2.2 and section 2.2.3, is based on a sample of n observations, $\mathcal{D} = \{t_i, \delta_i, x\}$. Finally, we write the likelihood function, under non-informative censoring, as

$$L(\vartheta; \mathcal{D}) \propto \prod_{t_i=0} \{p_0\} \prod_{t_i>0} \left\{ \Big[(1 - p_0 - p_1) f_p^*(t_i)\Big]^{\delta_i} \Big[p_1 + (1 - p_0 - p_1) S_p^*(t_i)\Big]^{1-\delta_i} \right\} \quad (3.15)$$

3.3 Simulation studies

We proceed a parameter estimation in the same way it was made in the section 2.4 of the previous chapter 2, i.e., based on both classical and Bayesian approach. Thus, the estimation via maximum likelihood principle is carried out with the use of the method of maximization "BFGS" of the R routine optim(). In order to assess the behavior of the asymptotic theory for increasing sample size, in the same way, we performed simulations 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 with increasing sample size as expected.

Regarding to the application of Bayesian approach for parameter estimation, the confidence intervals, with a 95% confidence, were obtained trough the empirical quantiles of the marginal posterior distributions, obtained via straightforward use of algorithms MCMC, with the openBUGS software.

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 in this dissertation. So, 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 simulation, we let x be a random variable that represents a consumer characteristic. The description of sample generation, i.e., all details of the simulated survival time distribution, and results obtained regarding to the proposed estimation method are described in the next sections.

The parameter scenarios are based in the same framework as done in the section 2.4.1, which description we refer the reader to consult the aforementioned section.

3.3.1 Results of Monte Carlo simulations

The followings figures describe the simulation results for the three simulated scenarios of parameters, where the sample size varies as n = 100, 250, 500, 750 and 1000, and considering both the classical (Figures 3.2, 3.3 and 3.4) and the Bayesian estimation approach (Figures 3.5, 3.6 and 3.7). For the purpose of simulation, we let x be a random variable that represents a consumer characteristic. Hence, the link configuration of the eight parameters (β_{10}, β_{11}, β_{20}, β_{21}, β_{30}, β_{31}, β_{40}, β_{41}) to be estimated is given by the following expressions:

$$\begin{aligned} \kappa_i &= -\log\left(\frac{e^{\beta_{10}+x_i\beta_{11}}}{1+e^{\beta_{10}+x_i\beta_{11}}+e^{\beta_{20}+x_i\beta_{21}}}\right), \\ \theta_{1i} &= -\log\left(\frac{e^{\beta_{20}+x_i\beta_{21}}}{1+e^{\beta_{10}+x_i\beta_{11}}+e^{\beta_{20}+x_i\beta_{21}}}\right), \\ \alpha_{1i} &= e^{\beta_{30}+x_i\beta_{31}}, \\ \alpha_{2i} &= e^{\beta_{40}+x_i\beta_{41}}. \end{aligned} \tag{3.16}$$

The parameter values are selected in order to assess the ML estimation performance under different shape and scale parameters (β_{30}, β_{31}, β_{40} and β_{41}, related to the Weibull time-to-default distribution), and also under a composition of different proportions of zero-inflated data (β_{10} and β_{11}) and non-defaulters rates (β_{20} and β_{21} related to censored data). Similarly to the conclusions reached in section 2.4.3, it can be seen from the Figures 3.2 to 3.7, that:

1. in general, the maximum likelihood estimation on average, MLEA, is close to the parameters set in the simulated parameter scenarios, see Figure 3.4. However, in scenarios 1 and 2, the parameters $\widehat{\beta}_{11}$ and $\widehat{\beta}_{21}$ need a larger sample size (from at least n=500 for β_{21}) to achieve convergence.

2. in general, according to Figures 3.3 and 3.4, biases and root mean square errors

decrease as 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 $p_0 = \exp(-\kappa)$ and $p_1 = \exp(-\theta)$, performs better compared to scenario 1 (Low), due, of course, to greater presence of zeros and censored data;

4. on the other hand, in the scenario with the fewer 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_1 and α_2, performs better compared to others scenario, due to greater presence of observed time-to-default data;

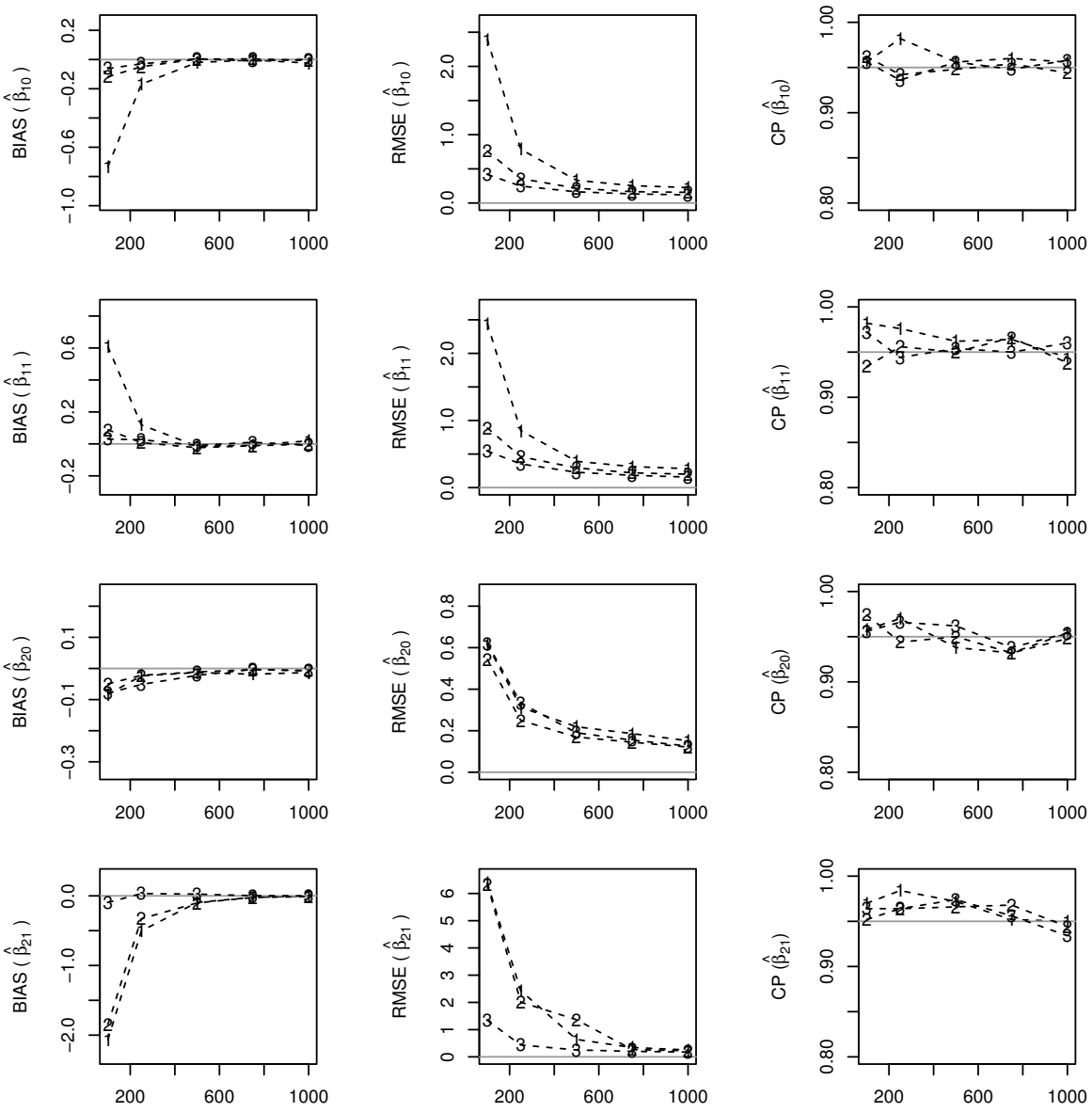

Figure 3.2: Bias, square root of mean squared error and coverage probability (CP) of **the maximum likelihood estimation** ($\widehat{\beta}_{10}$, $\widehat{\beta}_{11}$, $\widehat{\beta}_{20}$, $\widehat{\beta}_{21}$) of zero-inflated Promotion Cure rate regression model for simulated data under the three scenarios of parameters, obtained from Monte Carlo simulations with 1000 replications and increasing sample size (n).

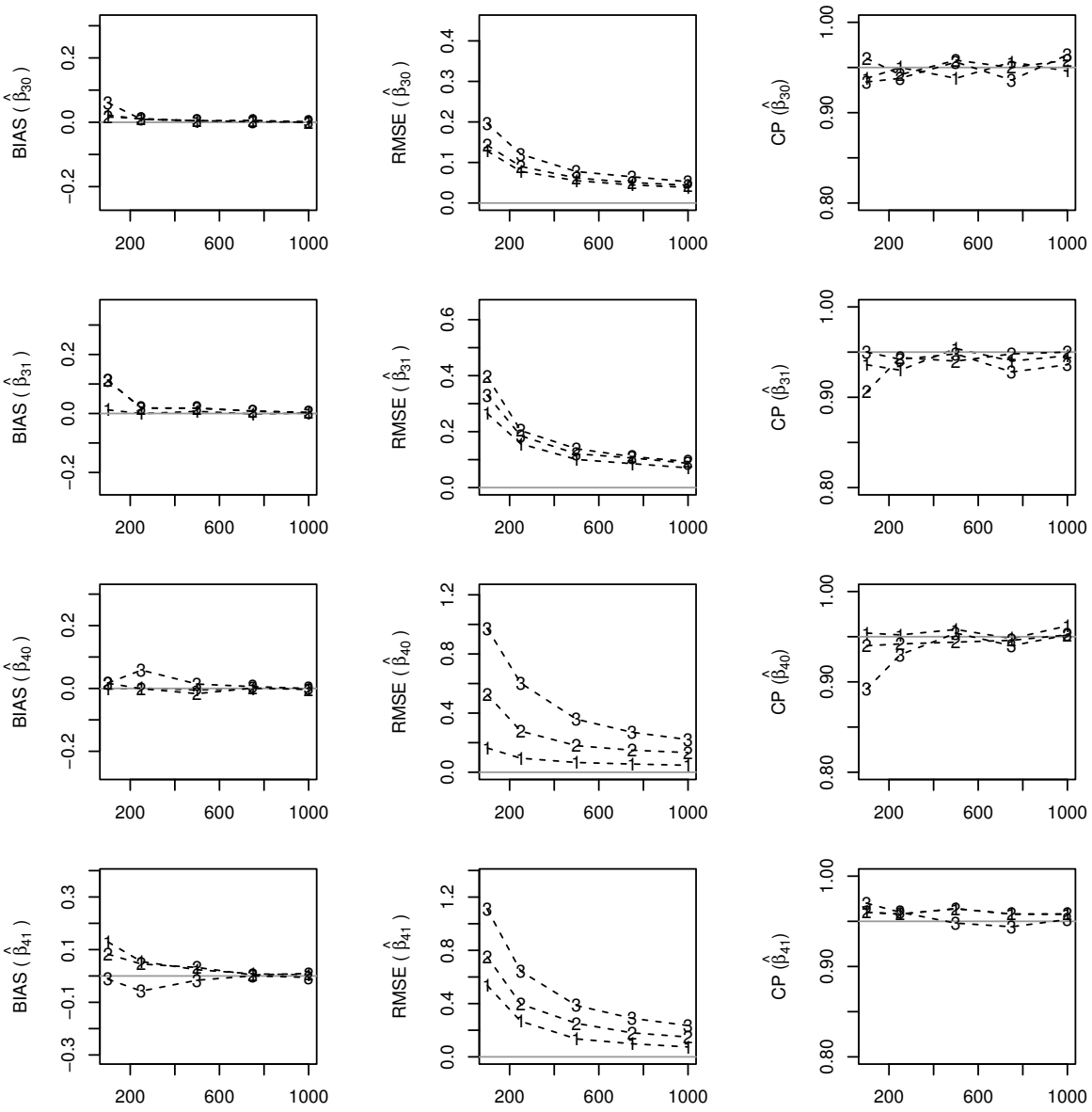

Figure 3.3: Bias, square root of mean squared error and coverage probability (CP) of **the maximum likelihood estimation** ($\widehat{\beta}_{30}$, $\widehat{\beta}_{31}$, $\widehat{\beta}_{40}$, $\widehat{\beta}_{41}$) of zero-inflated Promotion Cure rate regression model for simulated data under the three scenarios of parameters, obtained from Monte Carlo simulations with 1000 replications and increasing sample size (n).

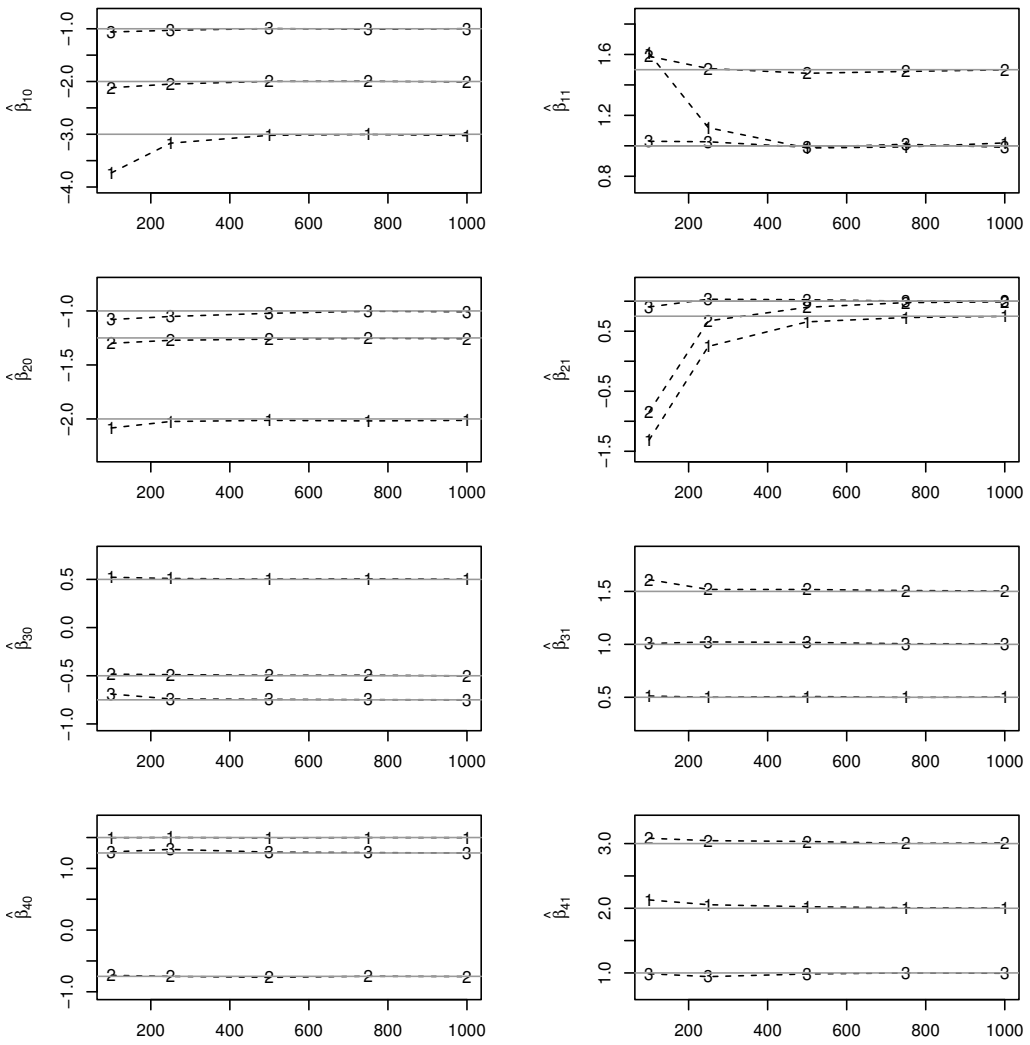

Figure 3.4: MLEA, **maximum likelihood estimation** on average of the parameters ($\widehat{\beta}_{10}$, $\widehat{\beta}_{11}$, $\widehat{\beta}_{20}$, $\widehat{\beta}_{21}$, $\widehat{\beta}_{30}$, $\widehat{\beta}_{31}$, $\widehat{\beta}_{40}$), $\widehat{\beta}_{41}$ of zero-inflated Promotion Cure rate regression model for simulated data under the three scenarios of parameters, obtained from Monte Carlo simulations with 1000 replications and increasing sample size (n).

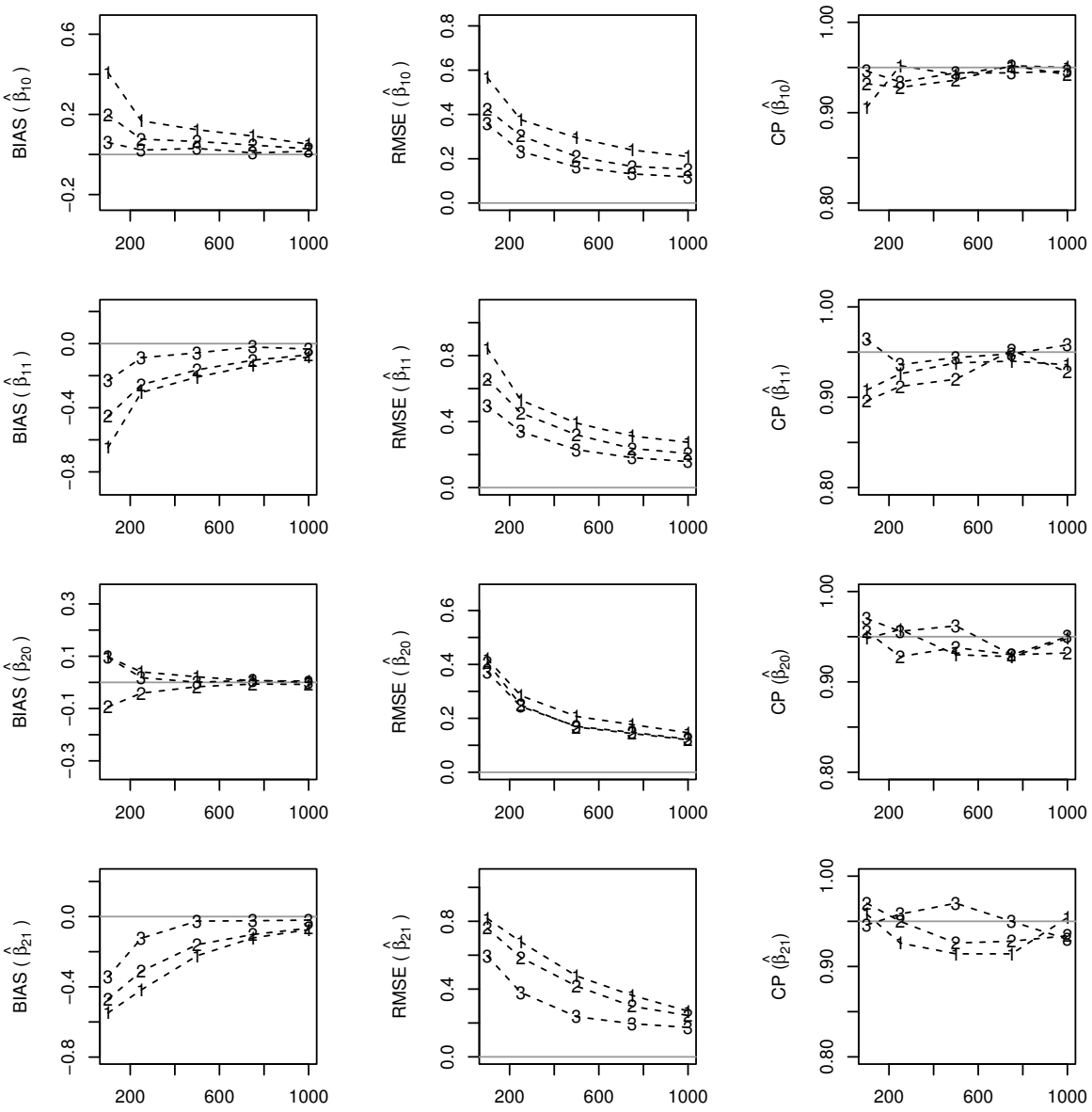

Figure 3.5: Bias, square root of mean squared error and coverage probability (CP) of **the Bayesian parameter estimations** ($\widehat{\beta}_{10}$, $\widehat{\beta}_{11}$, $\widehat{\beta}_{20}$, $\widehat{\beta}_{21}$) of zero-inflated Promotion Cure rate regression model for simulated data under the three scenarios of parameters, obtained from Monte Carlo simulations with 1000 replications and increasing sample size (n).

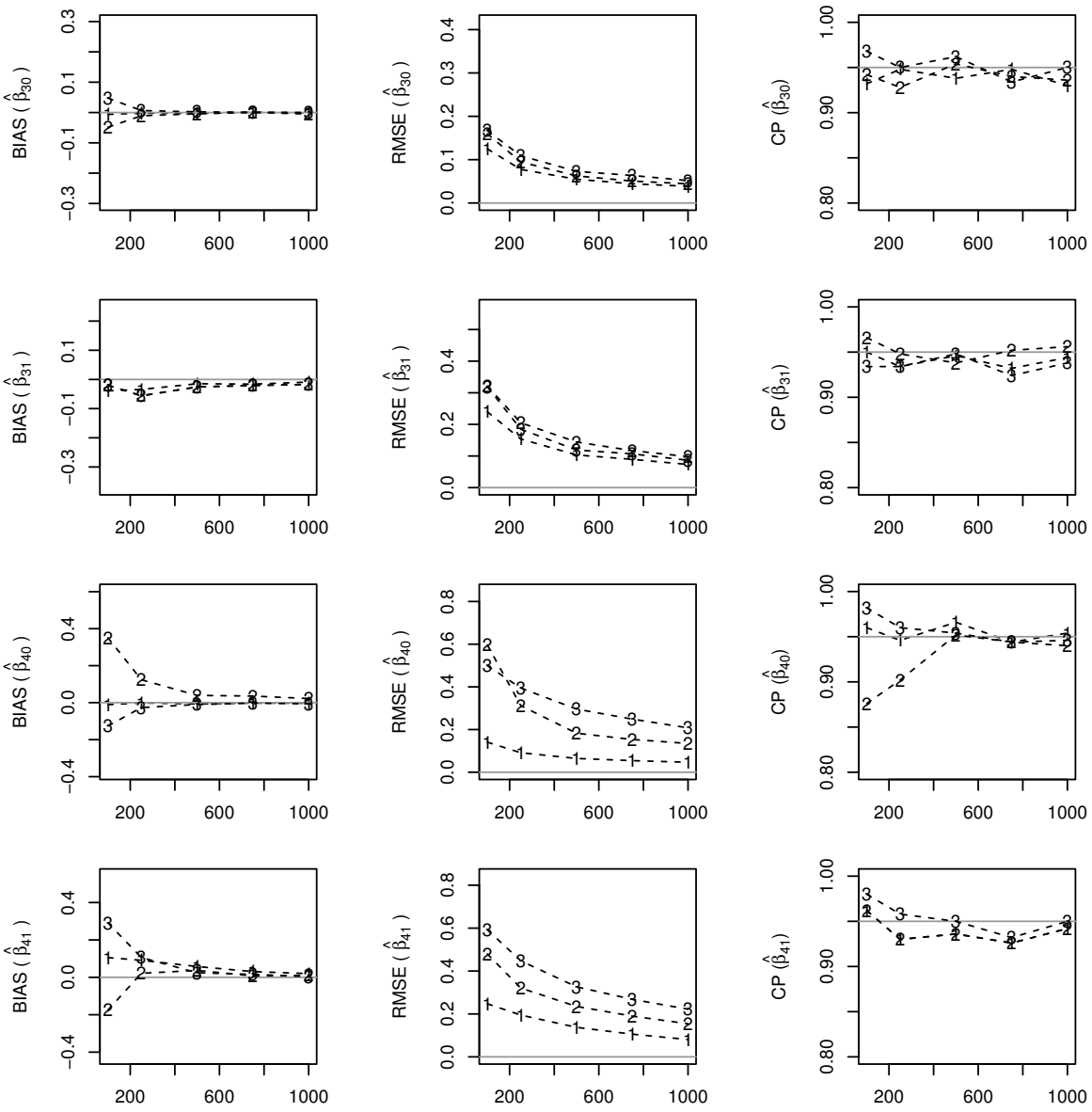

Figure 3.6: Bias, square root of mean squared error and coverage probability (CP) of **the Bayesian parameter estimations** ($\widehat{\beta}_{30}$, $\widehat{\beta}_{31}$, $\widehat{\beta}_{40}$, $\widehat{\beta}_{41}$) of zero-inflated Promotion Cure rate regression model for simulated data under the three scenarios of parameters, obtained from Monte Carlo simulations with 1000 replications and increasing sample size (n).

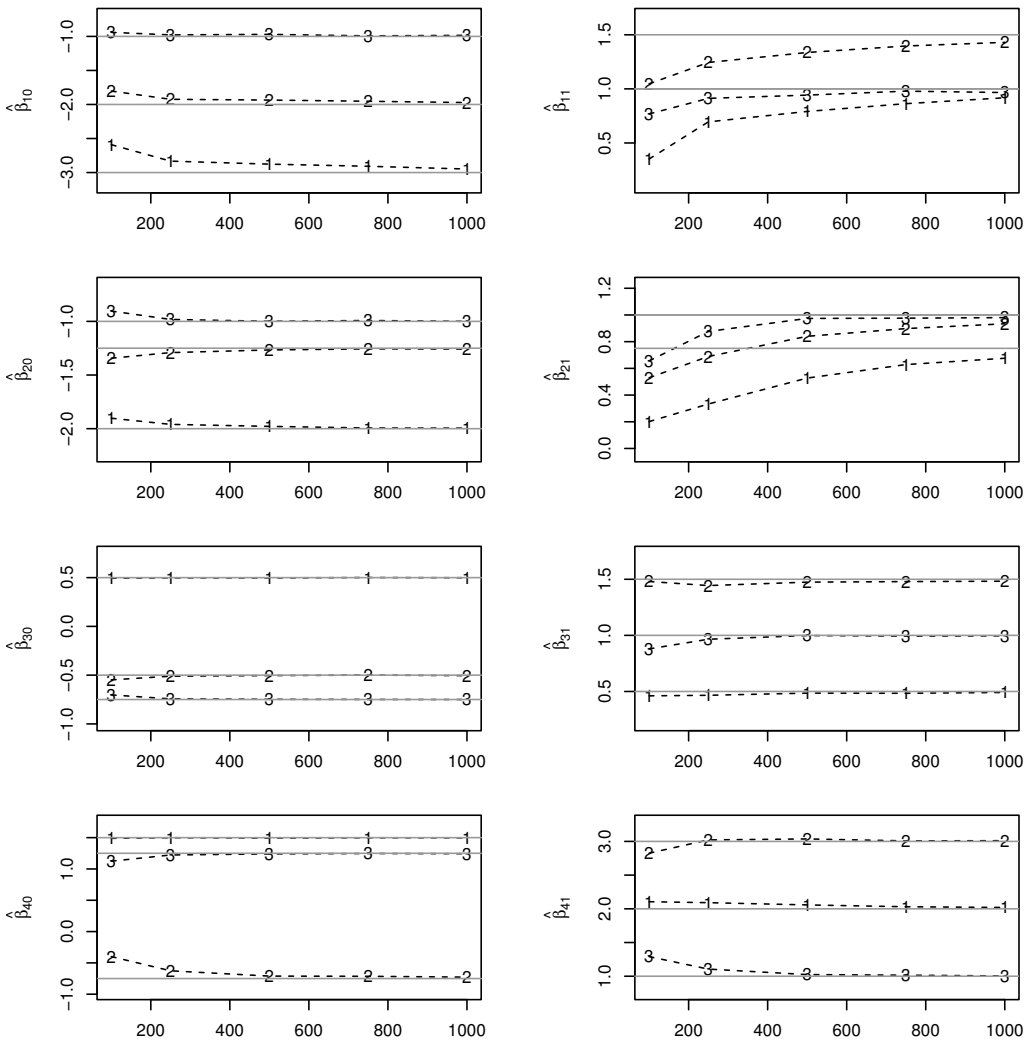

Figure 3.7: **The Bayesian parameter estimations** on average of the parameters ($\widehat{\beta}_{10}$, $\widehat{\beta}_{11}$, $\widehat{\beta}_{20}$, $\widehat{\beta}_{21}$, $\widehat{\beta}_{30}$, $\widehat{\beta}_{31}$, $\widehat{\beta}_{40}$, $\widehat{\beta}_{41}$) of zero-inflated Promotion Cure rate regression model for simulated data under the three scenarios of parameters, obtained from Monte Carlo simulations with 1000 replications and increasing sample size (n).

3.4 Application: Brazilian bank loan portfolio

In this section we present the application of the zero-inflated promotion cure rate regression model introduced in Section 3.2. For that, the same real dataset analysed in the chapter 2 is again fitted and the results are compared. Therefore, we fit the model with the same selected dummy covariates showed in Table 2.1.

Table 3.1 summarizes the estimated parameters via MLE approach for the regression parameters of the zero-inflated promotion cure rate regression model. The final model has AIC of 12596.26 (and BIC of 12662.8, $l\{\hat{\beta}_{10}, \hat{\beta}_{11}, \hat{\beta}_{12}, \hat{\beta}_{13}, \hat{\beta}_{20}, \hat{\beta}_{21}, \hat{\beta}_{22}, \hat{\beta}_{30}, \hat{\beta}_{40}, \hat{\beta}_{41}\} = -6288.128,\ p = 10$). The Bayesian model is summarized in Table 3.2, which has AIC of 12596.38 ($l\{\hat{\beta}_{10}, \hat{\beta}_{11}, \hat{\beta}_{12}, \hat{\beta}_{13}, \hat{\beta}_{20}, \hat{\beta}_{21}, \hat{\beta}_{22}, \hat{\beta}_{30}, \hat{\beta}_{40}, \hat{\beta}_{41}\} = -6288.189,\ p = 10$). We see

that the model estimated by MLE approach is slightly better than its Bayesian version in terms of AIC selection criteria.

Parameter	Dummy covariate ($\text{param}_{(1)}$)	Estimate	S.E.$_{(2)}$	P-value	Exp.$_{(3)}$
p_0	Intercept (β_{10})	-0.6690	0.1213	0.0000	0.5122
	Age group =4 (β_{11})	-0.8187	0.2231	0.0002	0.4409
	Type of residence =4 (β_{12})	0.8653	0.4736	0.0677	2.3757
	Type of employment=2 (β_{13})	-0.6473	0.1434	0.0000	0.5234
p_1	Intercept (β_{20})	0.9123	0.0957	0.0000	2.4901
	Type of residence =1 (β_{21})	-0.2905	0.1028	0.0047	0.7478
	Type of employment=2 (β_{22})	0.6331	0.0970	0.0000	1.8834
α	Intercept (β_{30})	0.1730	0.0376	0.0000	1.1889
θ	Intercept (β_{40})	3.1855	0.0697	0.0000	24.1817
	Age group =4 (β_{41})	0.6895	0.1435	0.0000	1.9927

Table 3.1: The Zero-Inflated Promotion Cure Regression Model for time-do-default on a Brazilian Bank Loan Portfolio. Notes: (1) Related regression parameter to be estimated; (2) Standard error; (3) Exp(estimated parameter).

Parameter	Estimated parameter	95% Credible interval Lower	Upper
$\hat{\beta}_{10}$	-0.6681	-0.9044	-0.4387
$\hat{\beta}_{11}$	-0.8202	-1.2500	-0.4011
$\hat{\beta}_{11}$	0.7264	-0.2316	1.4920
$\hat{\beta}_{12}$	-0.6589	-0.9354	-0.3803
$\hat{\beta}_{20}$	0.9105	0.7162	1.0940
$\hat{\beta}_{21}$	-0.2883	-0.4925	-0.0867
$\hat{\beta}_{22}$	0.6347	0.4451	0.8180
$\hat{\beta}_{30}$	0.1684	0.0967	0.2385
$\hat{\beta}_{40}$	3.1840	3.0630	3.3440
$\hat{\beta}_{41}$	0.6844	0.4025	0.9666

Table 3.2: Parameters obtained via Bayesian estimation using the software openBUGS.

Figure 3.8 shows the adjusted survival curves according to the parameters obtained via classical approach. We can see that they are very similar to the curves obtained in the previous section 2.5, see K-M curves in 2.15.

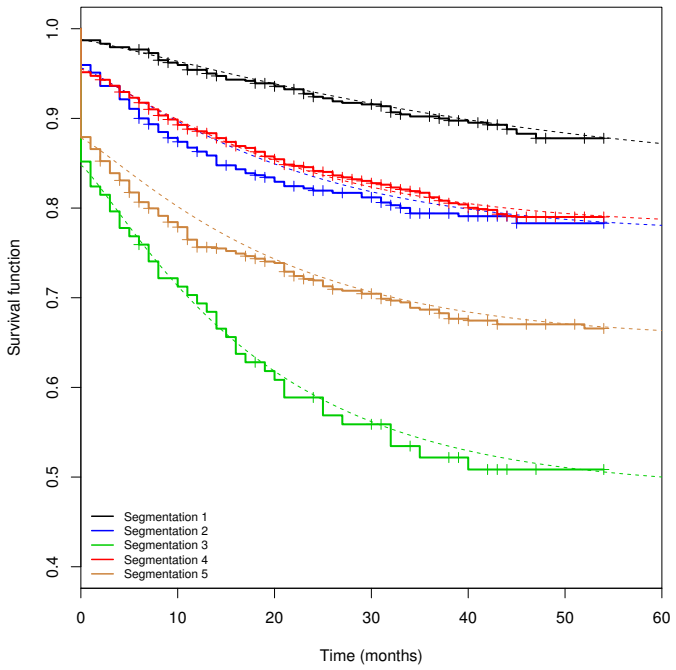

Figure 3.8: Brazilian bank loan portfolio. Kaplan-Meier survival curves stratified through the
in the Table 3.2.

The parameters obtained in this section are consistent with the parameters obtained in chapter 2, that is, have the same order of magnitude and signs. We can thus, as already discussed in the previous section 2.5, ratify risk behaviors presented in the initial graphical analysis given by the K-M curves in 1.3.

3.5 Concluding remarks

We introduced a methodology based on a zero-inflated survival data that extends the model studied in Yakovlev & Tsodikov (1996) and Chen *et al.* (1999). In this sense, an advantage of our approach is to accommodate zero-inflated times, which is not possible in the standard cure rate model. To illustrate the methodology presented here, we re-analysed a bank loan survival data, in order to assess the propensity to default in loan applications. In this scenario, informations from borrowers are exploited through the joint modeling of the zero survival time, along with the survival times of the remaining portfolio. The results showed the new model performed very well.

Despite the new zero-inflated promotion cure model presenting slightly better results in terms of AIC, 12596.26 compared with AIC of 12602.52 of the zero-inflated cure rate model presented in the chapter 2, it is important to note that the actual performance of

novel models will be measured through its daily use by the bank and with the use of a wider variety of available covariates, since the model allows the use of as many covariates as needed, whether continuous or categorical.

Chapter 4

An inflated mixture of beta models with applications to Loss Given Default

In this chapter, we propose an inflated mixture model to deal with multimodality in loss given default data. We propose a mixed of degenerate distributions, to handle zeros and ones excess, with a mixture of beta distributions for non-zeros and non-ones proportions. By applying the methodology in four retail portfolios of a large Brazilian commercial bank, we show that the inflated mixture of beta distributions plays better role minimizing model risk in fitting an inadequate model, in comparison with others considered competitive models. We explore the use of maximum likelihood estimation procedure. Monte Carlo simulations are carried out in order to check its finite sample performance.

4.1 Introduction

Since the Basel II publications in the mid-2000s, recommending central banks to allow banks to use internal data to calculate credit risk measures of their portfolios, much has been proposed in the literature on probability of default, loss given default and exposure at default. See for example Engelmann & Rauhmeier (2011), Loterman *et al.* (2012), and Yashkir & Yashkir (2013). The importance is justified as these parameters comprise the main ingredients of regulatory capital calculation, what banks must set aside to cope with

unexpected losses from credit portfolios.

According to Basel II rules for corporate, sovereign and bank exposures (see BCBS (2006), paragraphs 286 and 297), loss given default (LGD) is measured as the proportion of unrecovered debt, compared to total counterparty overdue debt. However, despite the simplicity in setting it, there are distinctions about the treatment that should be given to different types of portfolios (see Schuermann (2004)). For instance, in case of the aforementioned portfolios, corporate, sovereign and bank exposures, banks must provide an individual estimate of LGD for each exposure and, for that reason, a different approach that has been applied to retail portfolios.

In fact, as retail exposures typically represent majority of loan portfolios of commercial banks, it would be impossible to give an individualized treatment for each exposure. That is why, even before Basel recommendations, bank risk managers relied on automated scoring models. This means that, mostly, modeling credit risk involves estimating parametric statistical models, see (Porath, 2011). However, as expected, an exaggerated dependence on complex statistical models may lead to new sources of risks. In this case, the model risk, in other words, the risk of not choosing the best model in the light of the available data.

An attempt to draw attention to model risk, and encourage mitigation of this source of risk, has already been addressed by the Basel Committee, as stated in (BCBS, 2015), p. 2, "Supervisors should be cautious against over-reliance on internal models for credit risk management and regulatory capital. Where appropriate, simple measures could be evaluated in conjunction with sophisticated modeling to provide a more complete picture".

Regarding to the modeling of LGD, BCBS (2006) also recommends that its calculation must consider all relevant factors that impact in the loss triggered by the event of default. Strictly following the paragraph 460, the calculation must include all material discount effects and all material direct and indirect costs associated with collecting on the defaulted loan.

Since it is known that LGD has considerable impact on the regulatory capital amount, small differences can lead to major distortions in its calculation. For this reason, when dealing with large retail portfolios without sufficient evidence of the impact of each direct and indirect recovery cost, in order to proceed an reliable estimate, we must opt for models that bring a extra dose of conservatism.

In the foregoing context, i.e., concerning to the loss given default modeling, in this

chapter we aim to extend the established framework already proposed to accommodate multimodality in LGD data. Therefore, basically, we attempt to minimize model risk in fitting an inadequate distribution to LGD data. For that, we complement the work done by Hlawatsch & Ostrowski (2011), which, despite dealing with simulated bimodality, do not address the occurrence of zeros and ones excesses in bimodal LGD data.

Here, beyond to accommodate multimodality of LGD, we also account for the high evidence of excesses of zeros and one, as shown in Figure 4.1.

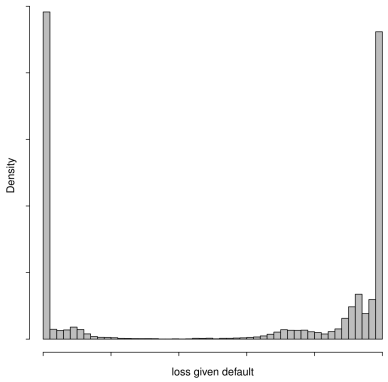

Figure 4.1: Multimodal LGD.

4.1.1 Brazilian bank non-performing retail data

To illustrate the LGD modeling proposed in this chapter, we analyse four retail portfolios of non-performed loans from a large Brazilian commercial bank. The whole data set comprises 41,677 retail loans, as summarized separately in the Table 4.1. Its whole LGD distribution is presented in Figure 4.1, where it shows a five-modal distribution. In Figures 4.12, 4.13, 4.14 and 4.15 are shown separately the 4 portfolios, all presenting quite different forms of four-modality. Note that in these figures, for clarity in data visualization, the zeros and ones are excluded, however, they are counted in the parameter estimation. Their zero and ones amounts are also presented in the Table 4.1.

Each portfolio is grouped according to the type of guarantees offered, or even the complete lack thereof. Of course, contract characteristics affect directly the presented shapes. For data confidentiality reasons, we do not explain the features of each loan making up each portfolio, we can only mention these are retail exposures, as defined in (BCBS, 2006), paragraph 231.

Portfolio	Qtd	Mean	Median	SD	#0	#1
1	15,295	0.52195	0.7272	0.4746	5722	6634
2	22,951	0.59814	0.9093	0.4596	8349	8398
3	440	0.32945	0.7466	0.4004	232	44
4	2,991	0.72060	0.9175	0.3810	510	265

Table 4.1: Summary of observed LGD data.

Although Calabrese (2014) has dealt with excess of zeros and ones in a real LGD data on Italian bank loans, the author has not presented real situations of multimodality (without considering the excess of zeros and ones), assuming an arbitrary mixture of two beta distributions to encourage the forecasting of two distinct periods, one with higher and another with lower mean intensity for the variable LGD. In this sense, we complement the work made in Calabrese (2014) by applying our methodology in a variety of real bank loan portfolios. In addition, we perform a simulation study to assess estimation performance of the inflated mixture model, what was not carried out in that referred paper.

Notwithstanding the bimodality and zeros and ones excess has been partially accounted for in recent literature, (Hlawatsch & Ostrowski, 2011; Tong *et al.*, 2013; Calabrese, 2014), to the best of our knowledge, the full configuration of the aforestated data has not been wholly incorporated into any model. Thereby, by assuming a mixed of degenerate distributions to handle all zeros and ones excess, together with a mixture of distributions to account for multimodal losses, along with the already mentioned variety of real applications and simulation studies, here, we fill this gap by introducing a simple statistical tool for risk managers deal, as effectively as possible, with loss given default multi-shape data.

Summing up, the first novelty of this chapter is to present an application of the inflated mixture models on a wide range of multimodality shapes, accompanied by a simulation study, which, at our knowledge, has not been presented in the literature. The second novelty is the presentation of the model considering the influence of a set of covariates, i.e., presenting a regression model version, thus, allowing to measure how customer and loan features impact on LGD results.

The chapter is organized as follows. In Section 4.2, we formulate the inflated mixture model and its regression version. Section 4.3 introduces maximum likelihood estimation. A simulation study with different vector parameters (with and without covariates) is presented in Section 4.4. An application to a real variety of retail portfolios of a large Brazilian bank is presented in Section 4.5. General remarks are presented in Section 4.6.

4.2 Model specification

Here, we propose an inflated mixture model to handle multimodality in the loss given default framework. To present in a more didactic way, without compromising general understanding of ideas, we only deal with beta distributions and consider the sum of two distributions in the mixture model. Other mixtures beyond the number of two, and the use of others bounded distributions that appear in the application section, i.e., Kumaraswamy, truncated normal and logit-normal distributions, can be easily implemented computationally, since there are a lot of statistical packages available in R and, also, a comprehensive amount of academic materials about them in the statistical literature.

4.2.1 The Inflated mixture of beta distributions

Inflated models are a way to incorporate degenerate points that do not belong to original distribution, assign to them probability to occur. Thereby, we firstly define what is mixture of two beta distributions, and consequently, it leads us naturally to our main definition, the inflated mixture of beta distributions. The well-known beta distribution, as appears in Ferrari & Cribari-Neto (2004), has mean parameter $\mu \in (0,1)$, and precision parameter $\phi > 0$. Defined only for $y \in (0,1)$, beta distribution has density function as follows:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{(\mu\phi-1)}(1-y)^{(1-\mu)\phi-1}, \tag{4.1}$$

where $\Gamma(.)$ is gamma function.

Given $f_1(.\,;\mu_1,\phi_1)$ and $f_2(.\,;\mu_2,\phi_2)$, beta distributions, we set a mixture of two beta distributions, with a 5-parameter density function given by $f_{m_2b} = \pi f_1 + (1-\pi) f_2$. The parameter π is commonly known as mixing proportion. Roughly, this means for the probability of $y \in (0,1)$ be more suitable accommodate by f_1, while $1-\pi$ stands for the probability of $y \in (0,1)$ be more accommodate by f_2. Now, let Y be a random variable with support in $\{0,1\} \cup (0,1)$. The Y distribution is said to be a inflated (in zeros and ones) mixture of two beta distributions, with a 7-parameter $\vartheta = (\delta_0, \delta_1, \pi, \mu_1, \phi_1, \mu_2, \phi_2)$, if its density function is given by:

$$f_{im_2b}(y;\vartheta) = \begin{cases} \delta_0, & \text{if} \quad y = 0 \\ (1-\delta_0-\delta_1) f_{m_2b}, & \text{if} \quad 0 < y < 1 \\ \delta_1, & \text{if} \quad y = 1, \end{cases} \tag{4.2}$$

where f_{m_2b} follows a mixture of two beta distributions. Note that $\alpha = \delta_0 + \delta_1$ is a mixing proportion and, mathematically, the parameters, δ_0, δ_1 and $(1-\alpha)$, account for, respectively, $P[y=0]$, $P[y=1]$ and $P[y \in (0,1)]$.

As we will see in the application section, the average estimate is an important decision-making criterion for the best LGD model. In fact, according to the Basel II agreement, it is advised to avoid a non-conservative estimate given practical circumstances of lack of informations (see (BCBS, 2006), paragraph 460). What is our case, once the Bank has made available a database records of accounting losses rather than economic losses. For that, we present the first moment (mean) of Y, given by $E[Y] = (1-\delta_0-\delta_1)(\pi\mu_1 + (1-\pi)\mu_2) + \delta_1$ (Bussab & Morettin, 2005, p. 208)

4.2.2 The inflated mixture of beta regression model

Here, we introduce an approach to accommodate covariates in a regression setting. In the application section, we discuss the model application to a real retail portfolios. Therefore, we propose to connect the set of seven parameters, $\vartheta = (\delta_0, \delta_1, \pi, \mu_1, \phi_1, \mu_2, \phi_2)$, with a set of 7-covariate vectors, $x = (x_1, x_2, x_3, x_4, x_5, x_6, x_7)$.These covariate vectors, as occurs in practice, may be the same, i.e., $x_1 = x_2 = x_3 = x_4 = x_5 = x_6 = x_7$. Following Pereira *et al.* (2013), the regression version of the inflated mixture of beta distributions is defined by 4.2, and by the following components, also known as link functions:

$$\begin{array}{rcl} (\delta_{0i}, \delta_{1i}) & = & \left(\frac{e^{x_{1i}^\top \beta_1}}{1+e^{x_{1i}^\top \beta_1}+e^{x_{2i}^\top \beta_2}}, \frac{e^{x_{2i}^\top \beta_2}}{1+e^{x_{1i}^\top \beta_1}+e^{x_{2i}^\top \beta_2}} \right), \\ \pi_i & = & \frac{e^{x_{3i}^\top \beta_3}}{1+e^{x_{3i}^\top \beta_3}}, \\ \mu_{1i} & = & \frac{e^{x_{4i}^\top \beta_4}}{1+e^{x_{4i}^\top \beta_4}}, \\ \phi_{1i} & = & e^{x_{5i}^\top \beta_5}, \\ \mu_{2i} & = & \frac{e^{x_{6i}^\top \beta_6}}{1+e^{x_{6i}^\top \beta_6}} \text{ and} \\ \phi_{2i} & = & e^{x_{7i}^\top \beta_7}, \end{array} \tag{4.3}$$

where the β_j's are seven vectors of regression coefficients to be estimated. Note that the inflated mixture of beta regression model can be viewed as an extension of the inflated beta regression model introduced by Martínez (2008) and Ospina & Ferrari (2012).

4.3 Parameter estimation

Parameter estimation is performed by straightforward use of maximum likelihood estimation (MLE) approach. Despite the existence of different computational strategies, see for example expectation-maximization (EM) algorithm in Ji *et al.* (2005) and Calabrese (2014, p. 274), our simulation studies support the simple application of the MLE approach regarding to the asymptotic behavior of the error measures as Biases and MSE's.

The likelihood function of the inflated mixture model f_{im_2b}, with a vector of 7-parameter $\vartheta = (\delta_0, \delta_1, \pi, \mu_1, \phi_1, \mu_2, \phi_2)$, linked with a covariate vector x, is based on a sample of n observations, $\mathcal{D} = \{y_i; x\}$:

$$L(\vartheta; \mathcal{D}) \propto \prod_{y_i} f_{im_2b}(y_i; \delta_{0i}, \delta_{1i}, \pi_i, \mu_{1i}, \phi_{1i}, \mu_{2i}, \phi_{2i}).$$

The maximum likelihood estimation $\hat{\vartheta}$, of the parameter vector ϑ, is obtained through maximization of $L(\vartheta; \mathcal{D})$ or $\ell(\vartheta; \mathcal{D}) = \log\{L(\vartheta; \mathcal{D})\}$. According to Migon *et al.* (2014, p. 176), the asymptotic distribution of the maximum likelihood estimates (MLEs), $\hat{\vartheta}$, is a multivariate normal with mean vector ϑ and covariance matrix, which can be estimated by $\{-\partial^2\ell(\vartheta)/\partial\vartheta\partial\vartheta^T\}^{-1}$, evaluated at $\vartheta = \hat{\vartheta}$, where the required second derivatives are computed numerically. There are many software and routines available for numerical maximization. We use the software **R** and the method "BFGS" for maximizing the log-likelihood function.

Different models can be compared by using the Akaike information criterion. The model with the smallest value of AIC is commonly chosen as the preferred for describing a given dataset. In the application section, we compare the proposed model configured with four different density functions and, through the application in four different real portfolios, the combination of the best results of AIC, along with the most conservative estimates of LGD, will decide which model better meet the Basel II conservative recommendations.

4.4 Simulation Studies

We proceed a parameter estimation based on a maximum likelihood principle and use the R routine optim() for that. In order to assess the performance of the maximum likelihood estimation with respect to sample size, we perform Monte Carlo simulations, where each sample is replicate 1000 times and the sample size varies as $n =$

250, 500, 750, 1000, 1250, 1500, 1750, 2000, 2500, 3000. Three simulation studies for the regression version introduced by the link functions in 4.3.

We define the following sets of parameters $\mathcal{B}_1$, $\mathcal{B}_2$ and $\mathcal{B}_3$ to perform the study simulations:

1. $\mathcal{B}_1 = (-3, 1, -2.5, 0.3, -0.8, 0.5, -0.8, 0.5, 0.5, 2, 1, 0.6, 1, 3)$,

2. $\mathcal{B}_2 = (-1.75, 0.1, -1.5, 0.05, -0.9, 0.6, -0.8, 0.5, 1, 3, 1, 0.6, 2, 2)$ and

3. $\mathcal{B}_3 = (-1, 0.5, -1.5, 1, -0.7, 0.4, -0.8, 0.5, -1, 3, 1, 0.6, 1, 1)$.

The parameters were chosen so as to have different proportions of ones and zeros in the three scenarios of LGD. The following histograms show the data distribution of the three scenarios set for the regression parameters.

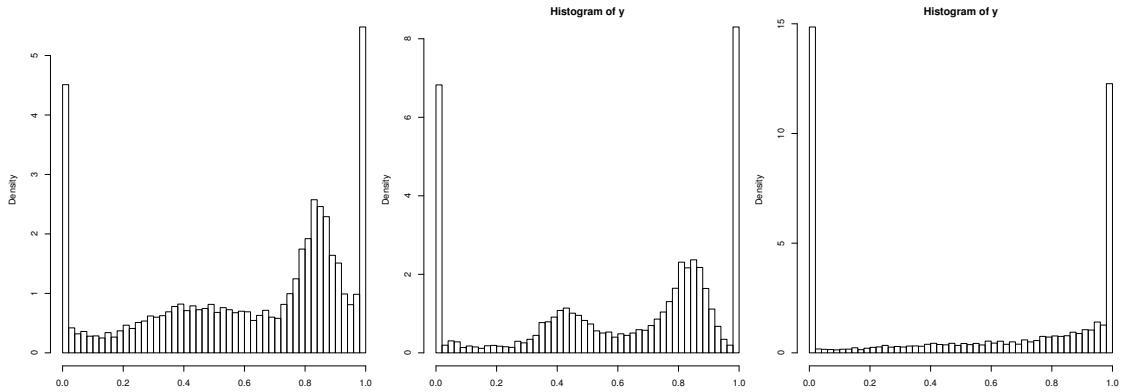

Figure 4.2: Histogram of simulated loss given default data: Left panel, lgd distribution according with the parameter scenario 1. Central panel, lgd distribution according with the parameter scenario 2. Right panel, lgd distribution according with the parameter scenario 3.

Finally, we present the estimates of bias, root mean square error (RMSE) and the average parameter estimation of the inflated mixture of beta regression model, obtained from Monte Carlo simulations, with 1000 replications and increasing sample size. For covariate simulation, we consider an intercept covariate $x_1 = 1$, and we assume x_2 as a binary covariate with values drawn from a Bernoulli distribution with parameter 0.5.

The following graphs 4.3 to 4.6 show the decrease to zero of the biases and RMSE for the three different parameter settings, each with around 10%, 25% and 50% of excess of zeros and ones. These settings are chosen as representing the reality of available data regarding to the amount of excess zeros and ones.

As expected, we can observe, in general, that the biases and root mean square errors of all parameters decrease as sample size increases. In particular, the biases and root mean square errors of the parameters $\widehat{\beta}_{10}$, $\widehat{\beta}_{11}$, $\widehat{\beta}_{20}$, $\widehat{\beta}_{21}$ decrease faster in the scenario 3 as sample size increases due to greater presence of zeros and ones.

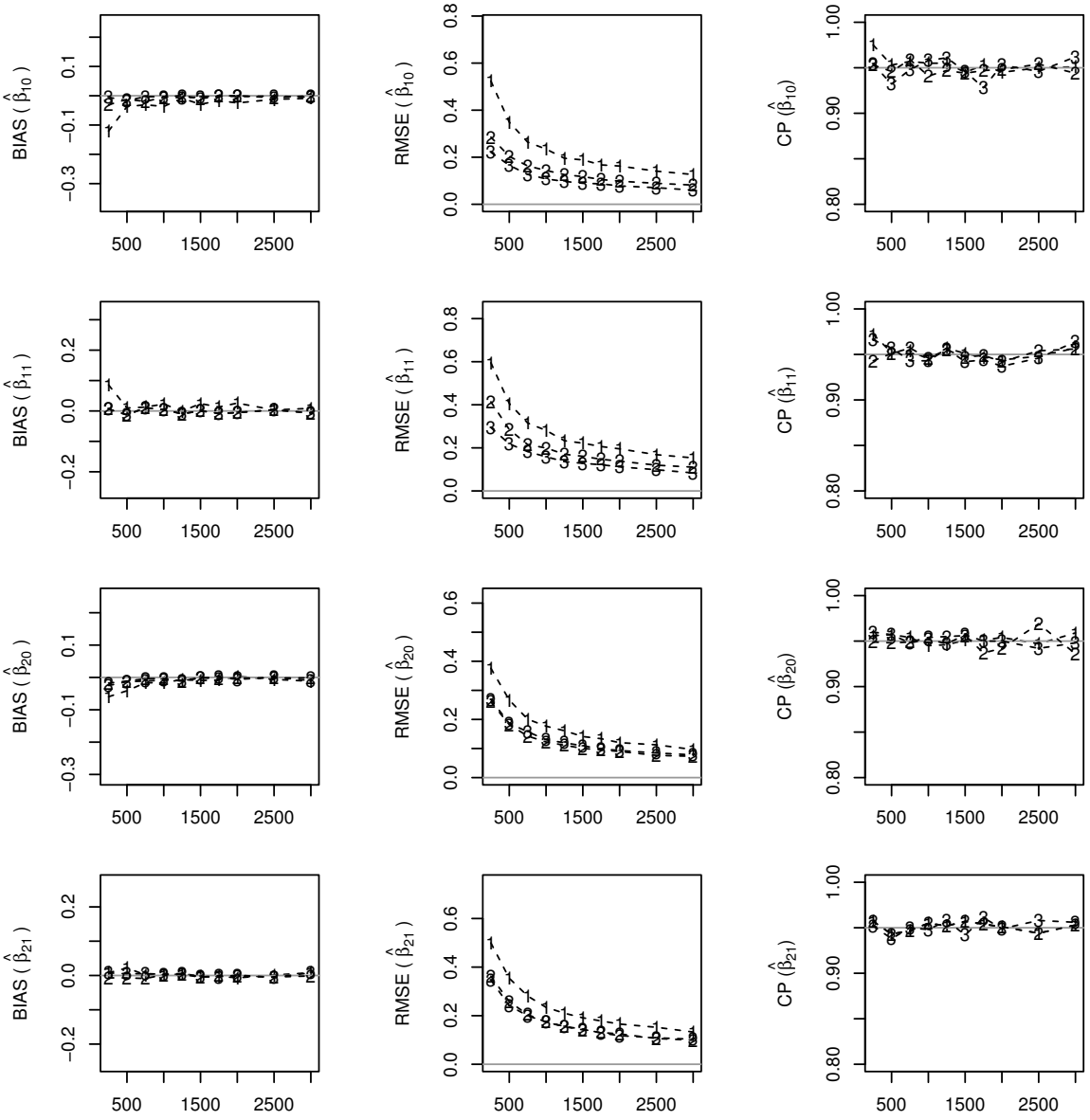

Figure 4.3: Bias, square root of mean squared error and coverage probability (CP) of **the maximum likelihood estimation** ($\widehat{\beta}_{10}$, $\widehat{\beta}_{11}$, $\widehat{\beta}_{20}$, $\widehat{\beta}_{21}$) of inflated mixture of beta regression model for simulated data under the three scenarios of parameters, obtained from Monte Carlo simulations with 1000 replications and increasing sample size.

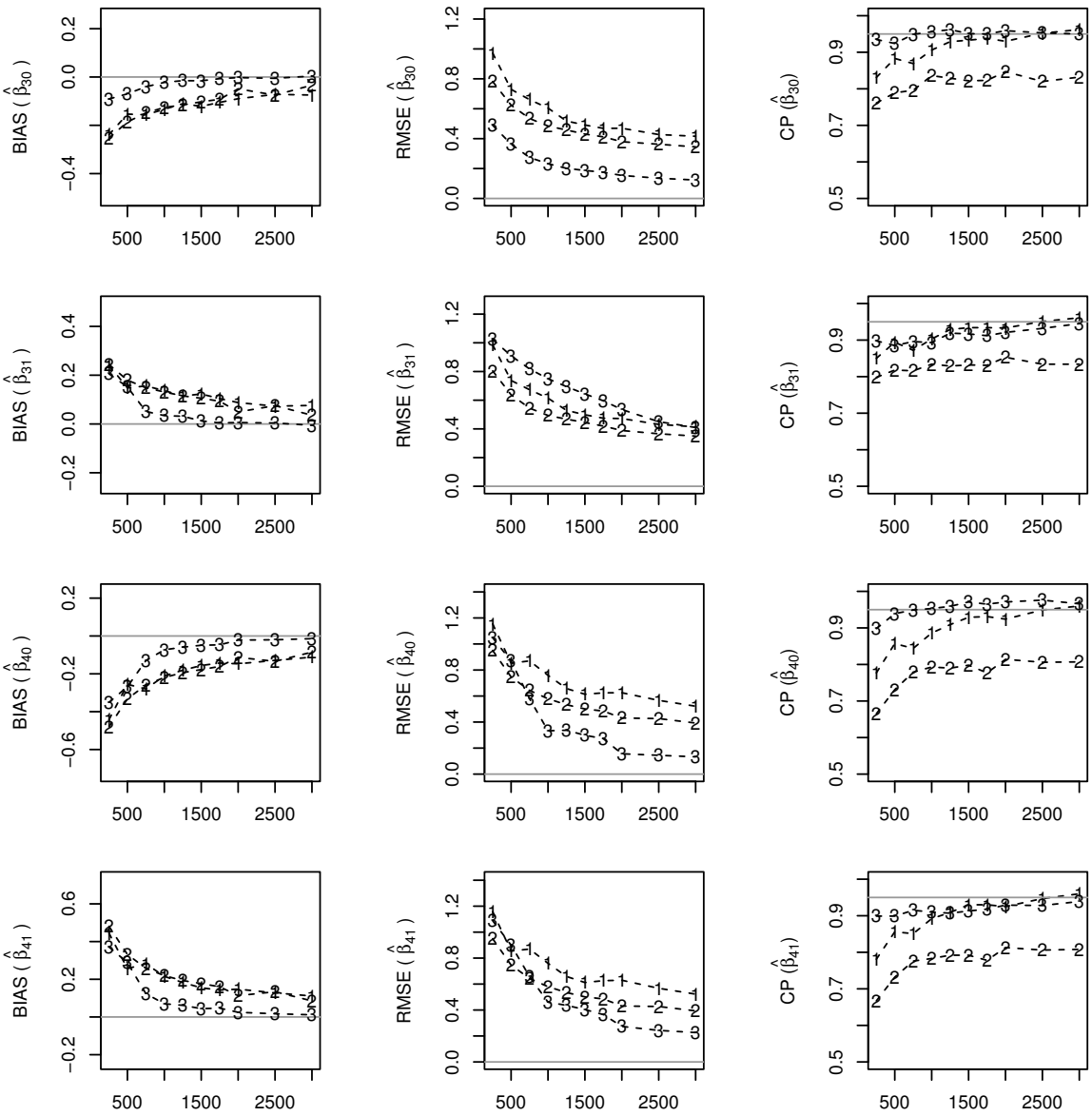

Figure 4.4: Bias, square root of mean squared error and coverage probability (CP) of **the maximum likelihood estimation** ($\widehat{\beta}_{30}$, $\widehat{\beta}_{31}$, $\widehat{\beta}_{40}$, $\widehat{\beta}_{41}$) of inflated mixture of beta regression model for simulated data under the three scenarios of parameters, obtained from Monte Carlo simulations with 1000 replications and increasing sample size.

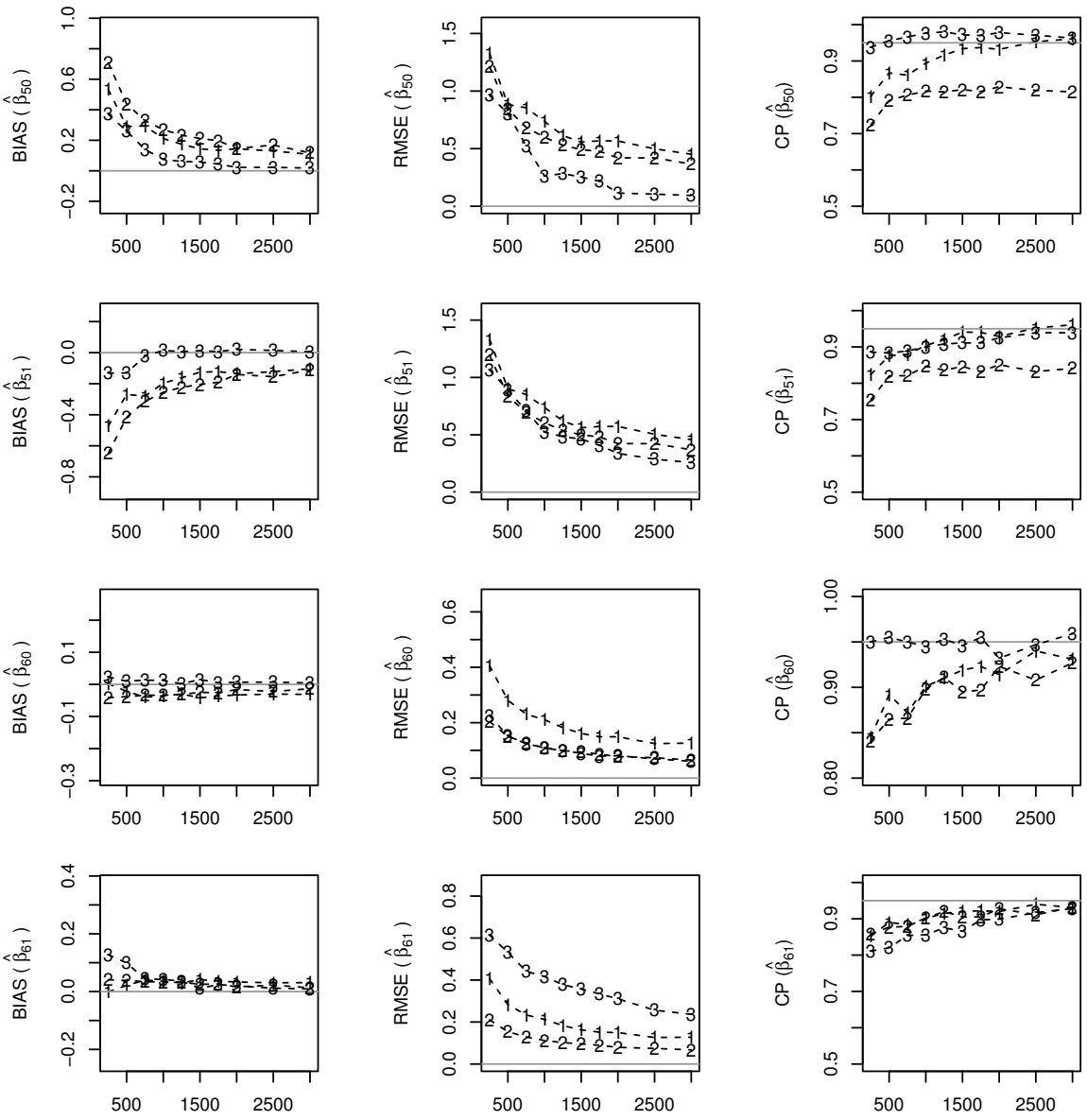

Figure 4.5: Bias, square root of mean squared error and coverage probability (CP) of **the maximum likelihood estimation** ($\widehat{\beta}_{50}$, $\widehat{\beta}_{51}$, $\widehat{\beta}_{60}$, $\widehat{\beta}_{61}$) of inflated mixture of beta regression model for simulated data under the three scenarios of parameters, obtained from Monte Carlo simulations with 1000 replications and increasing sample size.

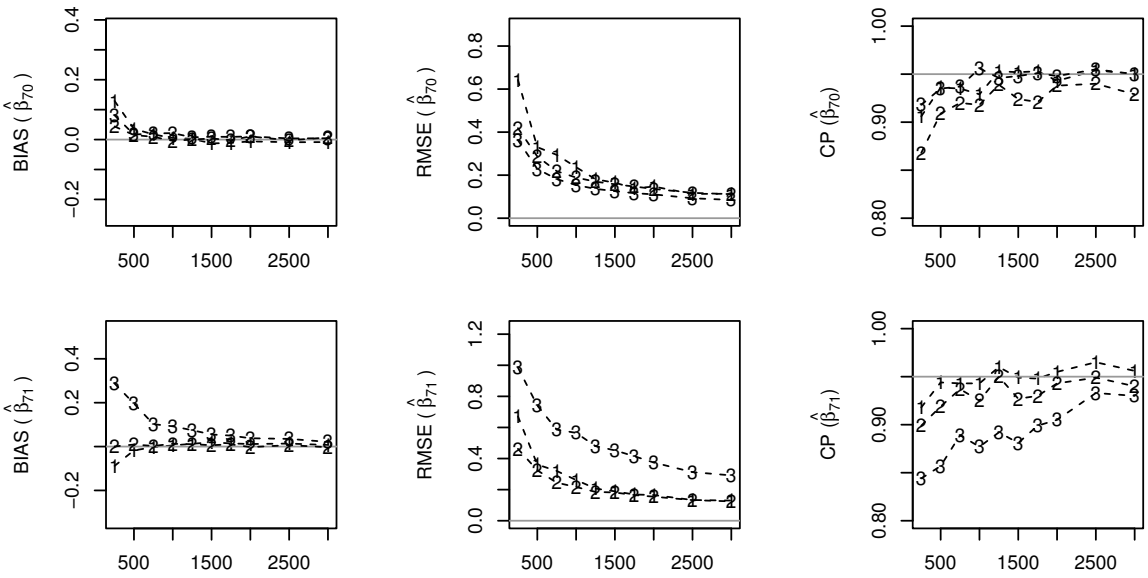

Figure 4.6: Bias, square root of mean squared error and coverage probability (CP) of **the maximum likelihood estimation** ($\widehat{\beta}_{70}$, $\widehat{\beta}_{71}$) of inflated mixture of beta regression model for simulated data under the three scenarios of parameters, obtained from Monte Carlo simulations with 1000 replications and increasing sample size.

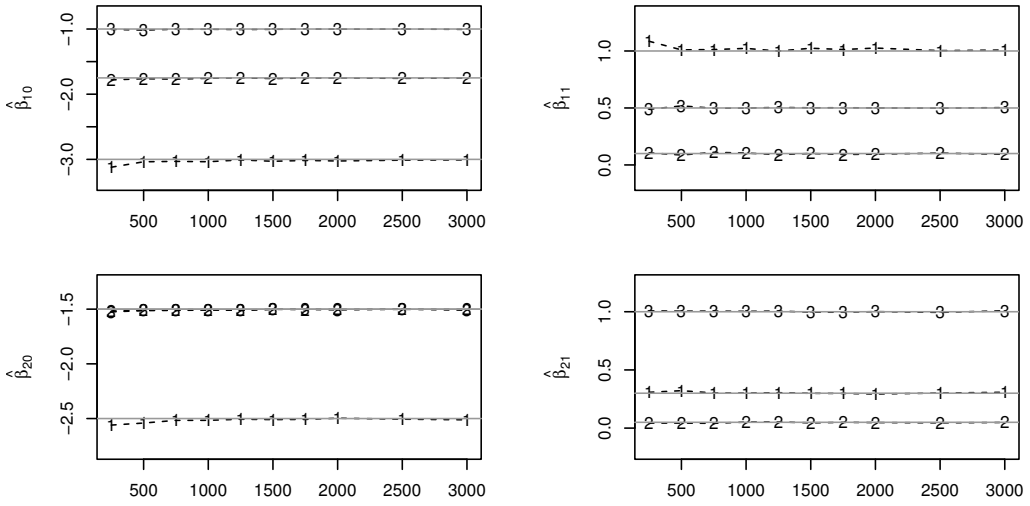

Figure 4.7: MLEA, **maximum likelihood estimation** on average of the parameters ($\widehat{\beta}_{10}$, $\widehat{\beta}_{11}$, $\widehat{\beta}_{20}$, $\widehat{\beta}_{21}$) of inflated mixture of beta regression model for simulated data under the three scenarios of parameters, obtained from Monte Carlo simulations with 1000 replications and increasing sample size.

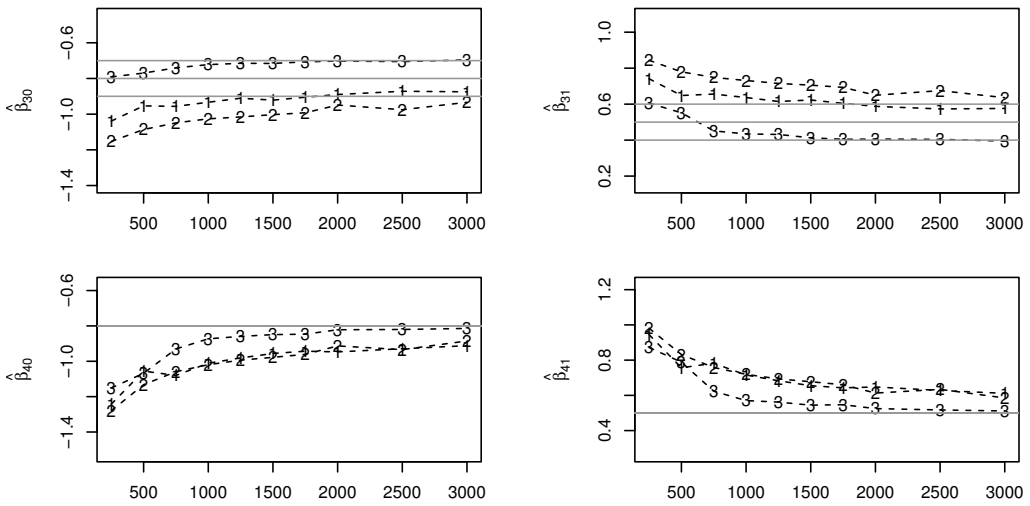

Figure 4.8: MLEA, **maximum likelihood estimation** on average of the parameters ($\widehat{\beta}_{30}$, $\widehat{\beta}_{31}$, $\widehat{\beta}_{40}$, $\widehat{\beta}_{41}$) of inflated mixture of beta regression model for simulated data under the three scenarios of parameters, obtained from Monte Carlo simulations with 1000 replications and increasing sample size.

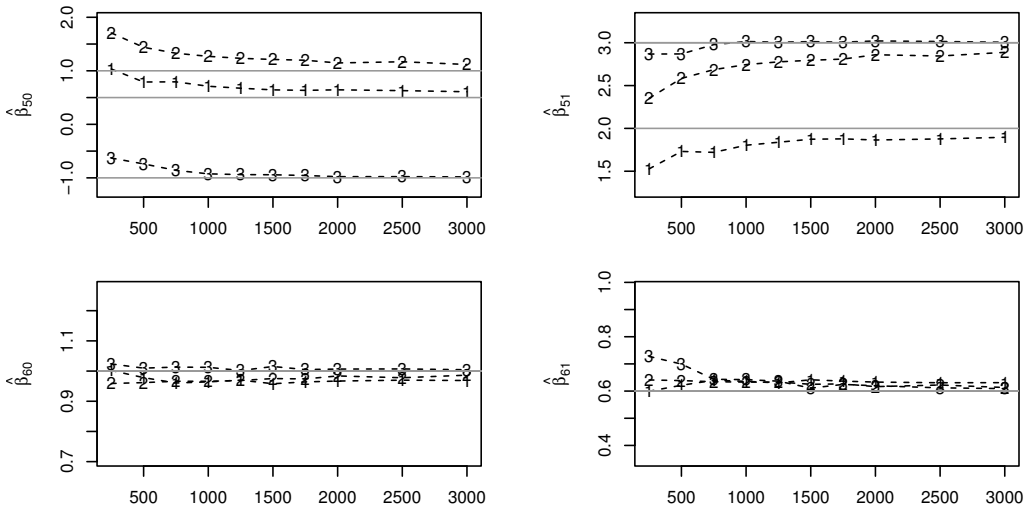

Figure 4.9: MLEA, **maximum likelihood estimation** on average of the parameters ($\widehat{\beta}_{50}$, $\widehat{\beta}_{51}$, $\widehat{\beta}_{60}$, $\widehat{\beta}_{61}$) of inflated mixture of beta regression model for simulated data under the three scenarios of parameters, obtained from Monte Carlo simulations with 1000 replications and increasing sample size.

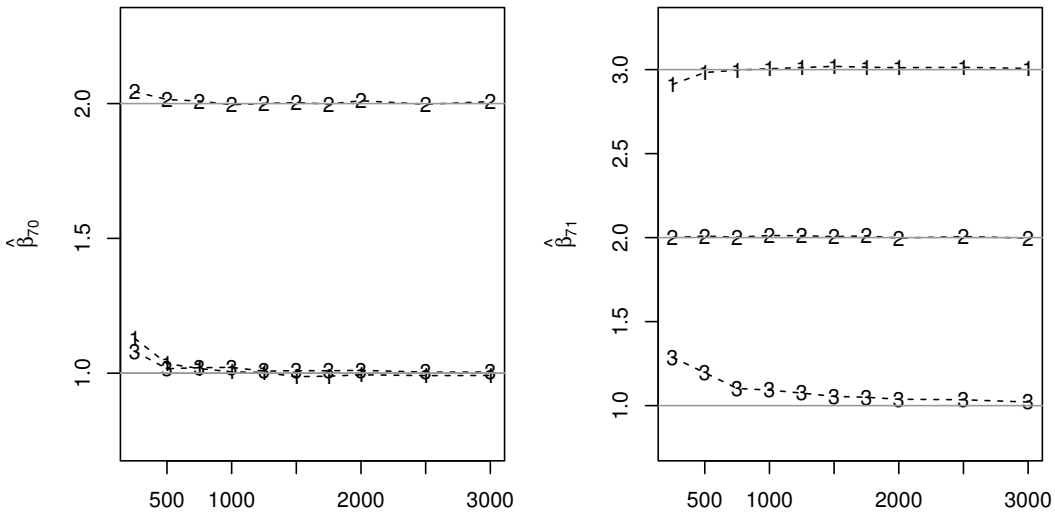

Figure 4.10: MLEA, **maximum likelihood estimation** on average of the parameters $(\widehat{\beta}_{70}, \widehat{\beta}_{71})$ of inflated mixture of beta regression model for simulated data under the three scenarios of parameters, obtained from Monte Carlo simulations with 1000 replications and increasing sample size.

4.5 Application

In this section we illustrate one application of the model presented in this chapter, i.e., the inflated mixture model as defined in (4.3). First, we compare the fit of four competitive inflated mixture models: The inflated mixture of beta distributions (Beckman & Tiet jen, 1978; Gupta & Nadarajah, 2004; Ospina & Ferrari, 2010), the inflated mixture of Kumaraswamy distributions (Kumaraswamy, 1980; Jones, 2009), the inflated mixture of $(0, 1)-$truncated normal distributions (Johnson *et al.*, 1994; Damien & Walker, 2001) and, finally, the inflated mixture of logit-normal distributions (Atchison & Shen, 1980; Frederic & Lad, 2008). These distributions have been chosen for comparison because are limited to the same support.

As mentioned, we opt to use the values of the AIC (Akaike information criterion) for model selection criterion, along with the choice of the most conservative estimates for LGD, in order to decide which model better meet the Basel II conservative recommendations. A goal for future research would be to propose measures of predictive power for the regression models here introduced.

4.5.1 Inflated mixture of beta models

Here, we consider the four portfolios described in the introduction section 4.1.1. As our goal is to select the best modeling LGD portfolio, so we present first the results obtained for the average estimated for each model, without covariates, which we compared with the average appearing in the table 4.1, along with the AIC obtained in adjusting the model to the data.

Given that the relative differences between the mean values estimated with the observed values of LGD, see Table 4.2, the model considering the beta distribution has a slight advantage due to have been presented the most conservative measure.

Portfolio	Expected lgd			
	beta	Kumaraswamy	Truncated normal	Logit-normal
1	0.52205	0.52173	0.52187	0.52272
2	0.59810	0.59817	0.59814	0.59789
3	0.33187	0.33129	0.32993	0.33088
4	0.72072	0.72021	0.72060	0.71841
Relative difference %	**0.1912%**	0.1167%	0.0330%	0.0590%

Table 4.2: Expected mean LGD by portfolio and by model.

Regarding to the measure of AIC, we must define a criterion for choosing the best model taking into account the adjustment of the four scenarios/portfolios. Thus, although the beta model, individually, has not the lowest AIC value acoordint Table 4.3, we see, with the graph 4.11 helps, that the beta model overall performance is better than the other models.

Model	Beta	Kumaraswamy	Truncated normal	Logit-normal
Portfolio 1	27999.30	28147.77	27350.50	28332.28
Portfolio 2	36870.90	36765.73	36903.02	37074.51
Portfolio 3	612.13	613.06	624.38	609.52
Portfolio 4	-1840.22	-1804.26	-1946.01	-1388.19

Table 4.3: AIC values for the fitted distributions.

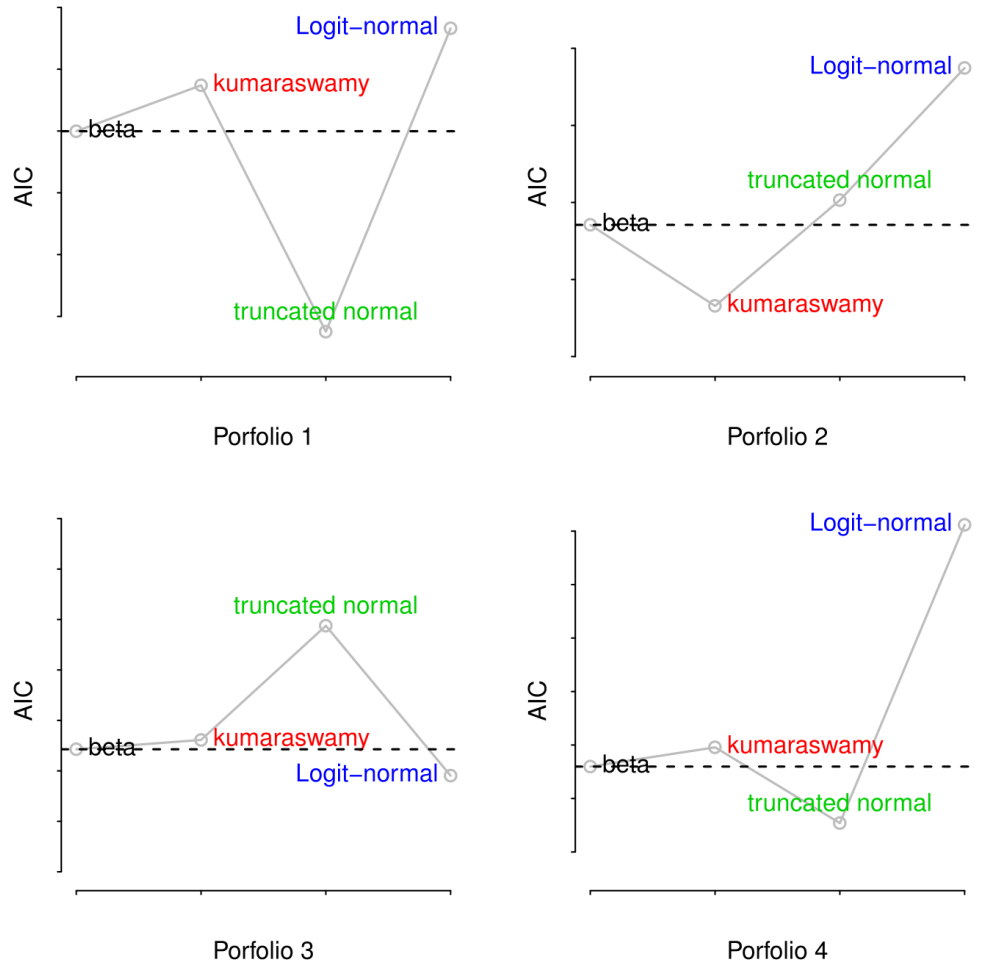

Figure 4.11: AIC values for model selection criterion.

For illustration, following, we present the graphs of the LGD shapes, and their respective adjustments of the each considered model.

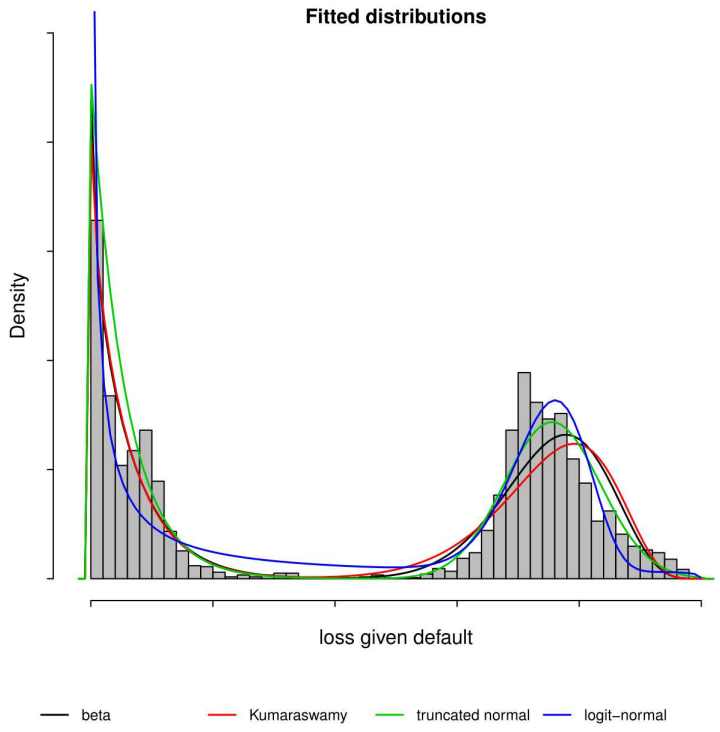

Figure 4.12: Fitted distributions for portfolio 1

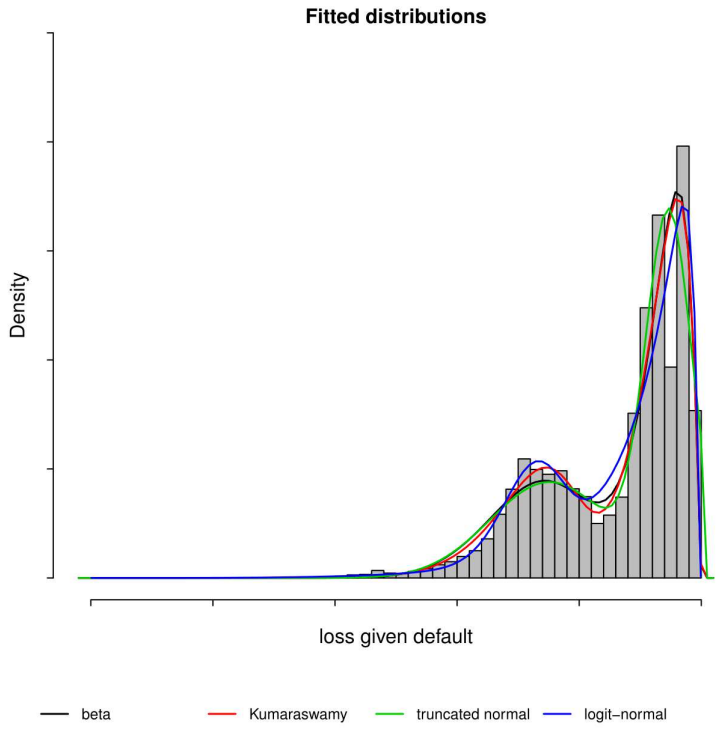

Figure 4.13: Fitted distributions for portfolio 2

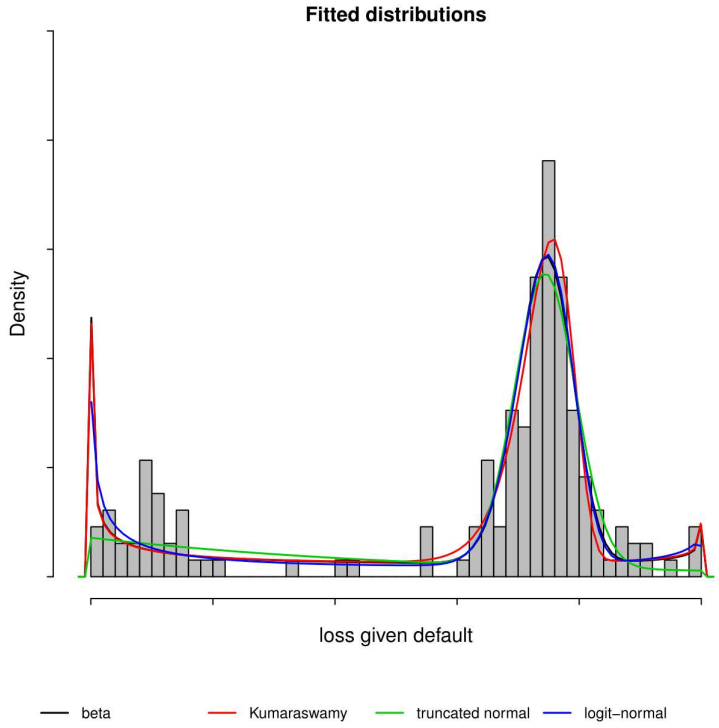

Figure 4.14: Fitted distributions for portfolio 3

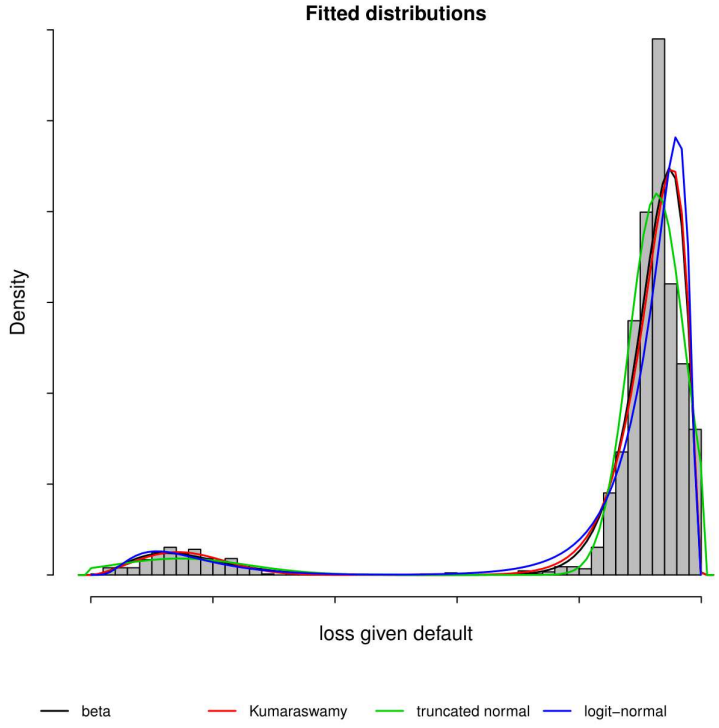

Figure 4.15: Fitted distributions for portfolio 4

4.5.2 Inflated mixture of beta regression models

In this section we illustrate the regression model taking in consideration two covariates made available by the bank. For this purpose, a sample of the portfolio 1, with two covariates and containing 5000 retail loans is considered for modeling purposes.

Let (x_1, x_2, x_3) be the vector of covariates, where x_1 stands for the interceptor parameter, i.e., $x_1 = 1$, and two others are real covariates. Both real covariates are categorized into two classes. The first, x_2 represents two customer groups according to the behavioral risk presented. The bank has its behavior score model and has segregated their customers into two groups, roughly, $x_2 = 0$ to customers with poor credit risk and $x_2 = 1$ with better credit risk. The second covariate is related to the loan characteristics. The loan classified as $x_3 = 0$, represents a group of loans with term relatively shorter than the group with $x_3 = 1$.

After an extensive search of the most significant parameters, through the AIC criterion, only four model parameters were selected to be linked with the covariates in question, δ_0, δ_1, μ_1 and μ_2. Thus, we have the following setting of link functions and parameters to be estimated:

$$\begin{aligned}
\delta_{0i} &= \frac{e^{(\beta_{11}x_{1i}+\beta_{12}x_{2i}+\beta_{13}x_{3i})}}{1+e^{(\beta_{11}x_{1i}+\beta_{12}x_{2i}+\beta_{13}x_{3i})}+e^{(\beta_{21}x_{1i}+\beta_{22}x_{2i}+\beta_{23}x_{3i}),}} \\
\delta_{1i} &= \frac{e^{(\beta_{21}x_{1i}+\beta_{22}x_{2i}+\beta_{23}x_{3i})}}{1+e^{(\beta_{11}x_{1i}+\beta_{12}x_{2i}+\beta_{13}x_{3i})}+e^{(\beta_{21}x_{1i}+\beta_{22}x_{2i}+\beta_{23}x_{3i}),}} \\
\mu_{1i} &= \frac{e^{(\beta_{31}x_{1i}+\beta_{32}x_{2i}+\beta_{33}x_{3i})}}{1+e^{(\beta_{31}x_{1i}+\beta_{32}x_{2i}+\beta_{33}x_{3i}),}} \\
\mu_{2i} &= \frac{e^{(\beta_{41}x_{1i}+\beta_{42}x_{2i}+\beta_{43}x_{3i})}}{1+e^{(\beta_{41}x_{1i}+\beta_{42}x_{2i}+\beta_{43}x_{3i}),}}
\end{aligned} \tag{4.4}$$

Portfolio 1	Subgroups		Quantity	Observed	Estimated
mean lgd (0.5216)	$x_2 = 0$	$x_3 = 0$	1266	0.6325	0.6383
		$x_3 = 1$	1269	0.6324	0.6255
	$x_2 = 1$	$x_3 = 0$	1234	0.4262	0.4265
		$x_3 = 1$	1231	0.3890	0.4101

Table 4.4: Summary of average LGD estimated by the inflated mixture of two beta regression model.

Parameter	Estimate (est)	Standard error (se)	|est|/ se	Exp(est)
$\hat{\beta}_{11}$	0.1890	0.0707	2.6731	1.2080
$\hat{\beta}_{12}$	0.8122	0.0809	10.0333	2.2528
$\hat{\beta}_{13}$	0.0851	0.0802	1.0599	1.0888
$\hat{\beta}_{21}$	0.9171	0.0641	14.3059	2.5020
$\hat{\beta}_{22}$	-0.2448	0.0786	3.1132	0.7828
$\hat{\beta}_{23}$	0.0013	0.0777	0.0173	1.0013
$\hat{\pi}$	0.5784	0.0828	6.9789	1.7831
$\hat{\beta}_{31}$	-0.7320	0.1005	7.2790	0.4809
$\hat{\beta}_{32}$	-0.2577	0.1086	2.3727	0.7728
$\hat{\beta}_{33}$	0.0728	0.1079	0.6750	1.0755
$\hat{\phi}_1$	1.1757	0.0538	21.8377	3.2404
$\hat{\beta}_{41}$	1.1076	0.0293	37.7084	3.0270
$\hat{\beta}_{42}$	-0.1125	0.0349	3.2177	0.8935
$\hat{\beta}_{43}$	-0.0508	0.0351	1.4478	0.9504
$\hat{\phi}_2$	62.8578	0.1110	566.0437	1.98e+27

Table 4.5: Maximum likelihood estimation results for the inflated mixture of two beta regression models.

The results summarized in Table 4.4 corroborate the following findings that loans held by lower credit risk customers have a much lower loss given default that the remaining group. Indeed, the last column shows, from β_{12} and β_{22}, respectively, that better credit scores increases by 125.28% the odds of loss given default be equal to zero while decreases by 21.72% the odd of total losses ($LGD = 1$).

The bimodality showed in the Figure 1.5 comes from some intrinsic aspect of the collection system used by the bank, together with collateral loan characteristics. Thus, we see that partial recoveries accumulate in two peaks, distances from one another, which are modeled by a mixture of two beta distributions. So we have a mixture of two beta distribution, one with average closer to zero while the other is closer to one. From $\hat{\beta}_{32}$ and $\hat{\beta}_{42}$, we see that better credit scored borrowers decrease, respectively, by 22.72% and 10.65%, the average LGD associated with each of the two beta distributions. On the other hand, from $\hat{\beta}_{33}$ and $\hat{\beta}_{43}$, although not statistically significant in the fitted regression model, we can see that longer loan term increases the average LGB by 7.55% in higher recovery scenario, while decreases by 4.96 % in the worst recovery scenario.

4.6 Concluding remarks

We have proposed two novelties in this chapter. The first is present the inflated of zeros and ones mixture models fitting four different real databases, made available by a large Brazilian commercial bank. We also present a simulation study to evaluate the asymptotic performance of the estimation method proposed, that is, the behavior of estimations regarding to the increased sample size.

The second novelty presented is the version regression of the model introduced by Calabrese (2014), where we propose a simple and easy to apply methodology. Our simulation studies and application in a real database show how the model can be useful for application in modeling LGD data sets. A future research aims at proposing measures of predictive power for the regression models here introduced.

Chapter 5

Conclusions

5.1 Concluding remarks

In this doctoral thesis we have proposed models to analyse financial data, mainly related to model portfolio of bank loans and the credit risk involved in granting them. First, we presented a methodology in which we modify the standard cure rate model introduced by Berkson & Gage (1952) to a credit risk setting. It allowed us to estimate the proportions of three sub-populations of borrowers that make up a banking portfolio: straight-to-default customers, defaulters, and non-defaulters. For that, is modified the improper survival function to account for the excess of zeros, which represents the rate of borrowers that do not account for even the first instalments and default on the loan at the beginning, which we called by straight-to-default customers

We also extend the promotion time cure rate model studied in Yakovlev & Tsodikov (1996) and Chen *et al.* (1999), by incorporating excess of zeros in the modeling. The presentation of this new model has enabled an alternative looking for models that consider fraction of cured, opening the possibility for new generalizations, for example, in the same formulation that the model framework has been presented in Rodrigues *et al.* (2009).

Thus, we contributed to the statistical literature regarding to the analysis of zero inflations, which had not been incorporated into the survival analysis that aims at dealing with the risk of default in credit risk setting.

Finally, we have dealt in this thesis with the modeling of inflated data in loss given default datasets, which was made available by a large Brazilian commercial bank. The novelty presented is the regression version of the model introduced by Calabrese (2014), where we propose a simple and easy to implement parameters estimation via maximum

likelihood approach. Our simulation studies and application in a real database show how the model can be useful for application in modeling LGD data sets.

5.2 Further researches

For the zero-inflation extensions of the models with cure fraction, we have considered the Weibull distribution for time-to-default, but different baseline density functions could be considered for that. Thus, further analysis can be conducted in order to propose new zero-inflated mixtures models, with different density functions to accommodate the time to event data.

In the modeling based in biological formulation, as in the promotion time cure rate model, a parametric form for N can be proposed to accommodate presence of survival time equal to zero at the beginning of the study. For instance, Barreto-Souza (2015) proposed a overdispersed distribution for N, in order to have more flexibility in modeling of the cure rate through covariates, however, to the best of our knowledge, there is no literature proposing modification of the N distribution to accommodate event time at the beginning of the study, i.e., zero-inflated data.

Defective distributions also can be proposed as a alternative framework to accommodate zero inflated date. In Rocha *et al.* (2015, 2016), the authors proposed a way to accommodate cured rate through defective distributions, but likewise, it could be useful investigating if such kind of distributions are also able to account for zero-inflated data into the modeling.

Regarding to the loss given default model framework, there are few works considering mixture model to deal with the distribution of LGD. As we have considered Beta distribution to model the loss given default data, different baseline density functions could be considered for such task. Can be proposed new computational strategies, as reversible jump for instance, to better fit the quantity of mixture components. However, from a practical point of view, as we have considered here, the outcomes and estimated parameters of the mixture model of only two Beta distributions can be easier to be analysed and interpreted. Finally, future research can also aims at proposing measures of predictive power for the regression models here introduced for loss given default data.

Appendix A

MCMC simulation graphics for the model applied to the loan survival time dataset.

The Zero-inflated Non-default Rate Model

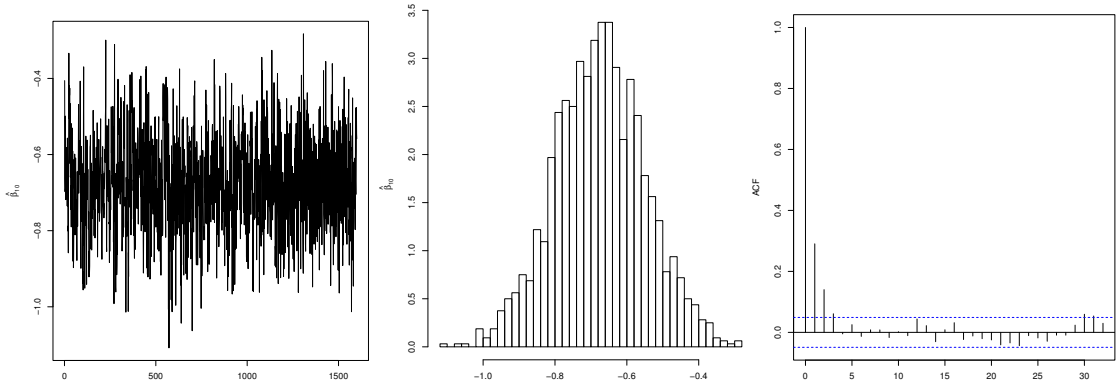

Figure A.1: Checking convergence plots for the estimated parameter $\widehat{\beta}_{10}$: Left panel, the trace for the estimated parameters via MCMC algorithm. Central panel, the approximate marginal density *a posteriori* of the parameter. Right panel, the autocorrelation function plot for the parameter.

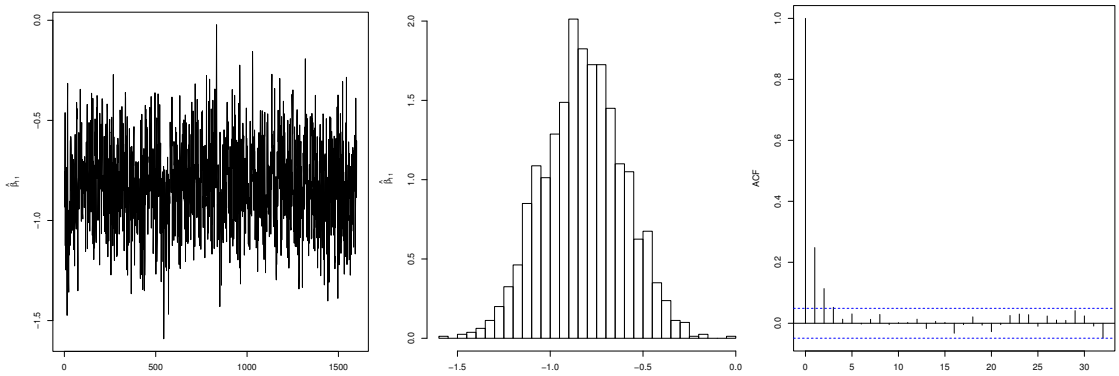

Figure A.2: Checking convergence plots for the estimated parameter $\widehat{\beta}_{11}$: Left panel, the trace for the estimated parameters via MCMC algorithm. Central panel, the approximate marginal density *a posteriori* of the parameter. Right panel, the autocorrelation function plot for the parameter.

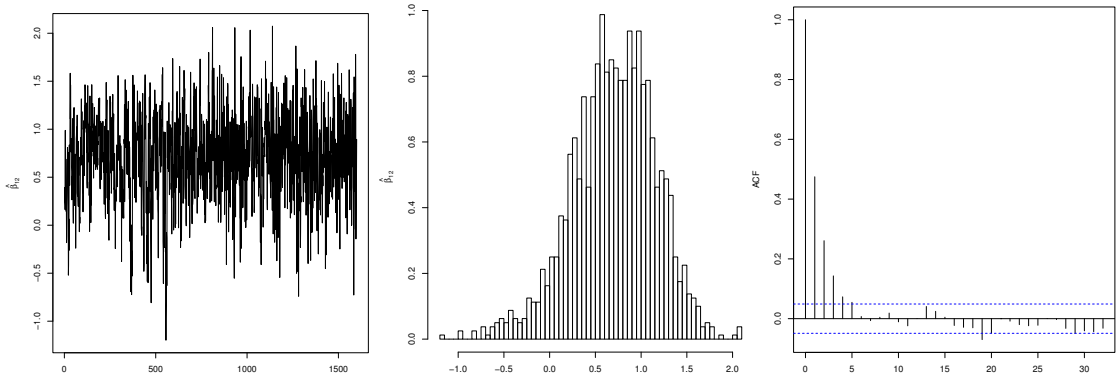

Figure A.3: Checking convergence plots for the estimated parameter $\widehat{\beta}_{12}$: Left panel, the trace for the estimated parameters via MCMC algorithm. Central panel, the approximate marginal density *a posteriori* of the parameter. Right panel, the autocorrelation function plot for the parameter.

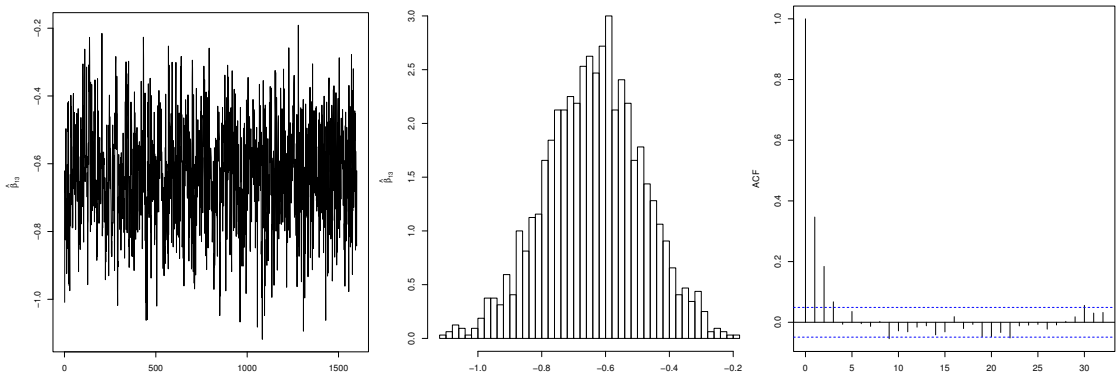

Figure A.4: Checking convergence plots for the estimated parameter $\widehat{\beta}_{13}$: Left panel, the trace for the estimated parameters via MCMC algorithm. Central panel, the approximate marginal density *a posteriori* of the parameter. Right panel, the autocorrelation function plot for the parameter.

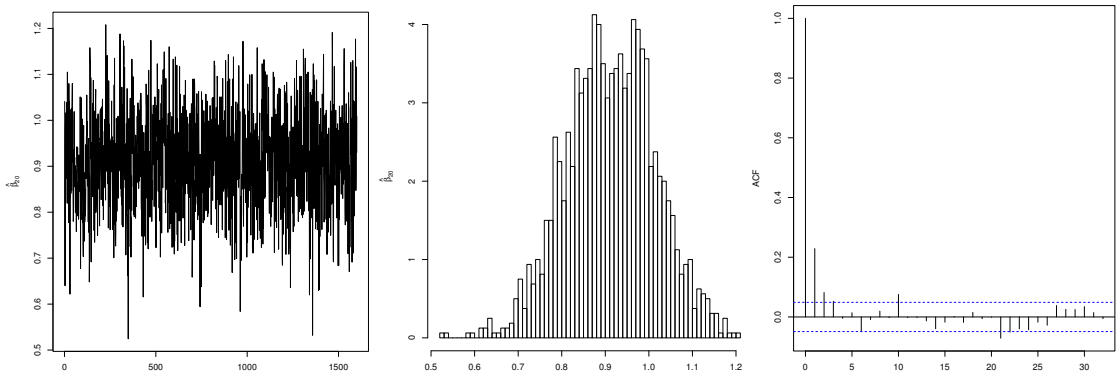

Figure A.5: Checking convergence plots for the estimated parameter $\widehat{\beta}_{20}$: Left panel, the trace for the estimated parameters via MCMC algorithm. Central panel, the approximate marginal density *a posteriori* of the parameter. Right panel, the autocorrelation function plot for the parameter.

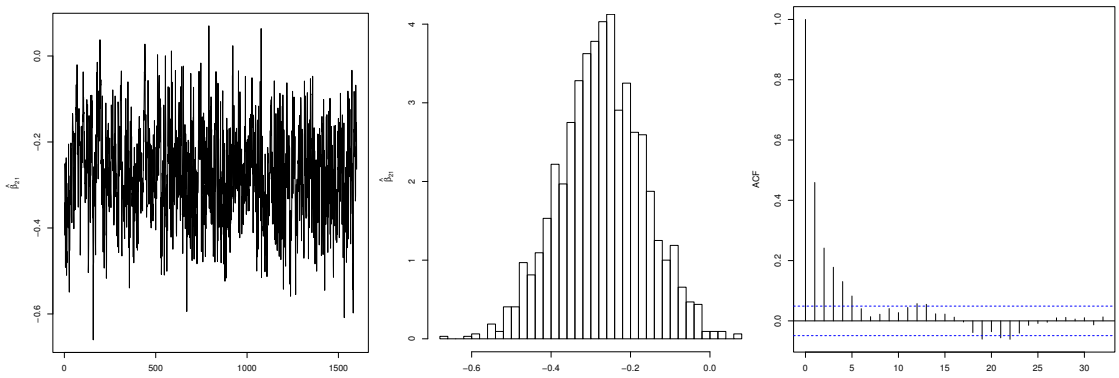

Figure A.6: Checking convergence plots for the estimated parameter $\widehat{\beta}_{21}$: Left panel, the trace for the estimated parameters via MCMC algorithm. Central panel, the approximate marginal density *a posteriori* of the parameter. Right panel, the autocorrelation function plot for the parameter.

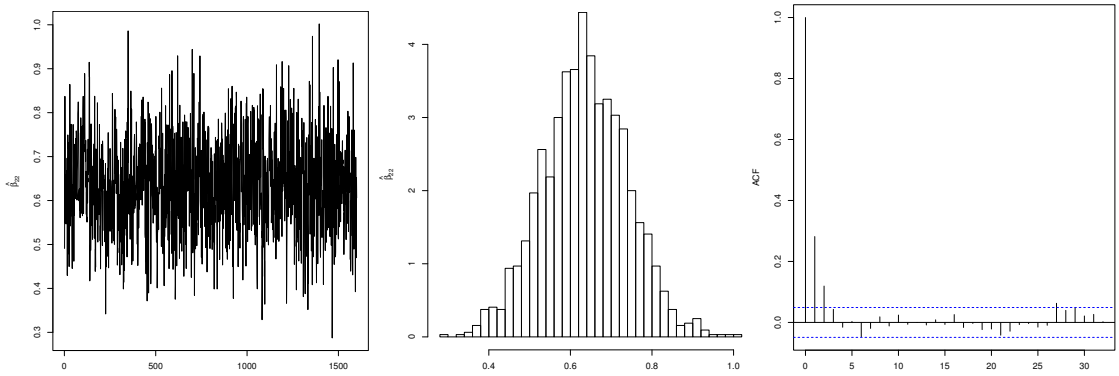

Figure A.7: Checking convergence plots for the estimated parameter $\widehat{\beta}_{22}$: Left panel, the trace for the estimated parameters via MCMC algorithm. Central panel, the approximate marginal density *a posteriori* of the parameter. Right panel, the autocorrelation function plot for the parameter.

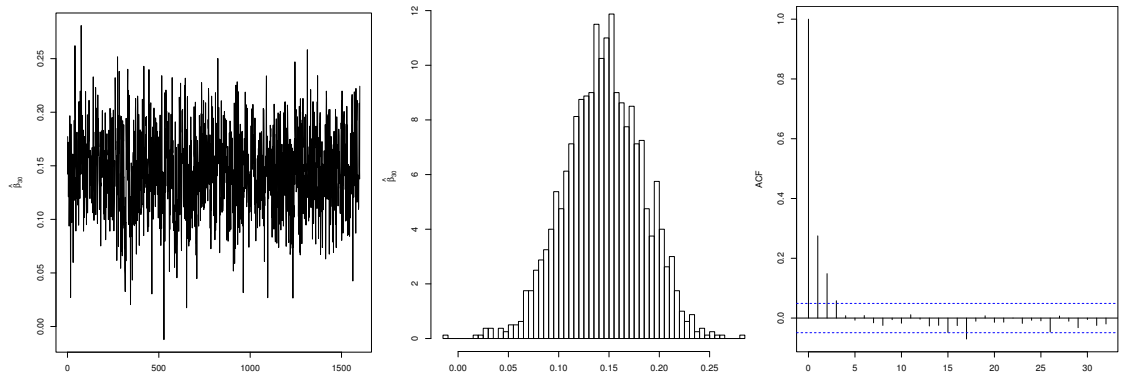

Figure A.8: Checking convergence plots for the estimated parameter $\widehat{\beta}_{30}$: Left panel, the trace for the estimated parameters via MCMC algorithm. Central panel, the approximate marginal density *a posteriori* of the parameter. Right panel, the autocorrelation function plot for the parameter.

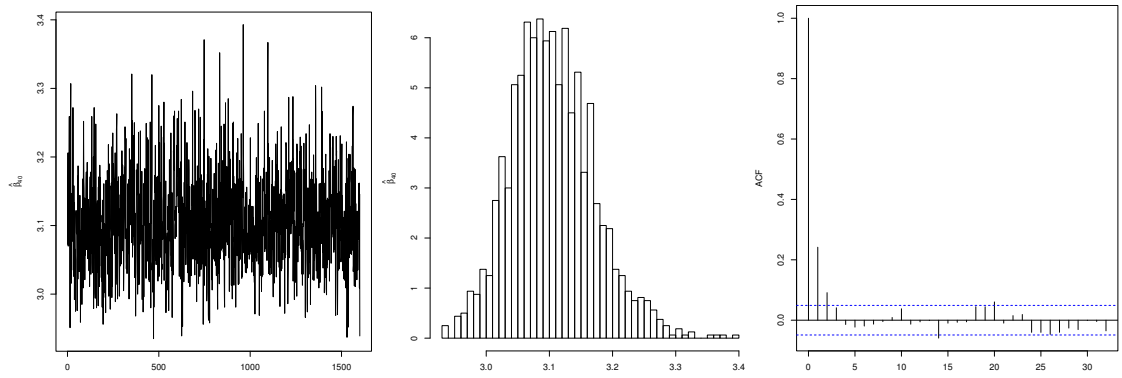

Figure A.9: Checking convergence plots for the estimated parameter $\widehat{\beta}_{40}$: Left panel, the trace for the estimated parameters via MCMC algorithm. Central panel, the approximate marginal density *a posteriori* of the parameter. Right panel, the autocorrelation function plot for the parameter.

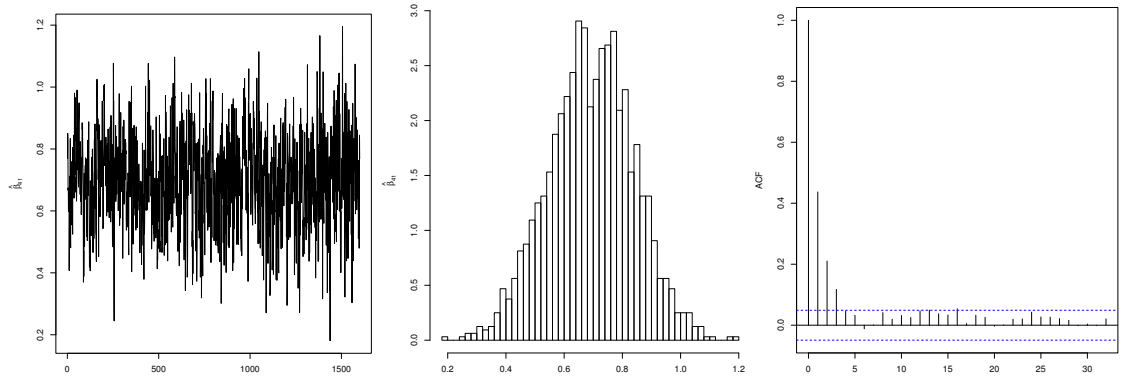

Figure A.10: Checking convergence plots for the estimated parameter $\widehat{\beta}_{41}$: Left panel, the trace for the estimated parameters via MCMC algorithm. Central panel, the approximate marginal density *a posteriori* of the parameter. Right panel, the autocorrelation function plot for the parameter.

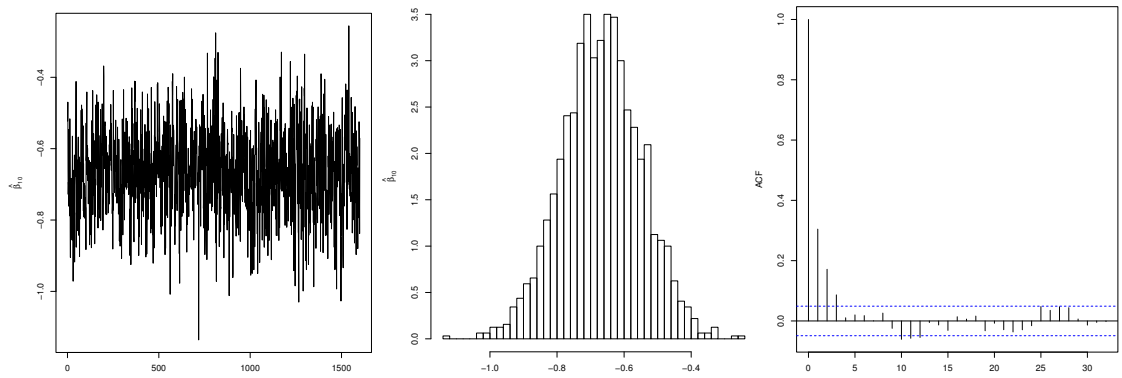

Figure A.11: Checking convergence plots for the estimated parameter $\widehat{\beta}_{10}$: Left panel, the trace for the estimated parameters via MCMC algorithm. Central panel, the approximate marginal density *a posteriori* of the parameter. Right panel, the autocorrelation function plot for the parameter.

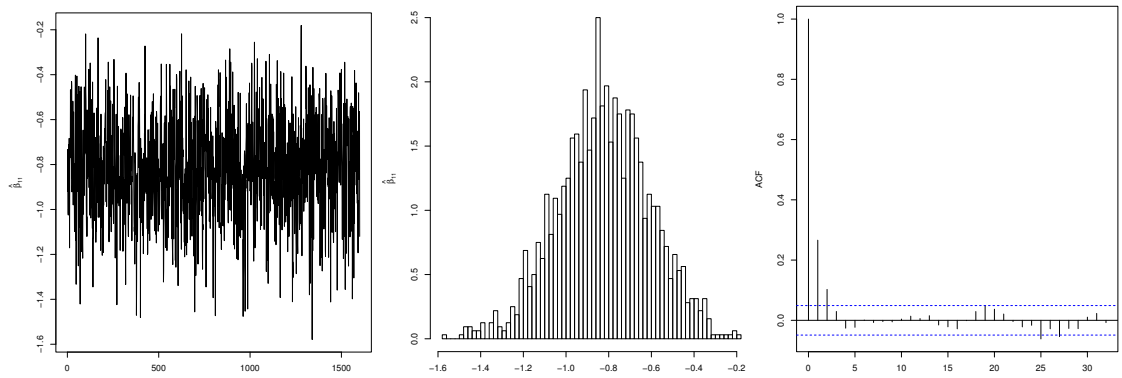

Figure A.12: Checking convergence plots for the estimated parameter $\widehat{\beta}_{11}$: Left panel, the trace for the estimated parameters via MCMC algorithm. Central panel, the approximate marginal density *a posteriori* of the parameter. Right panel, the autocorrelation function plot for the parameter.

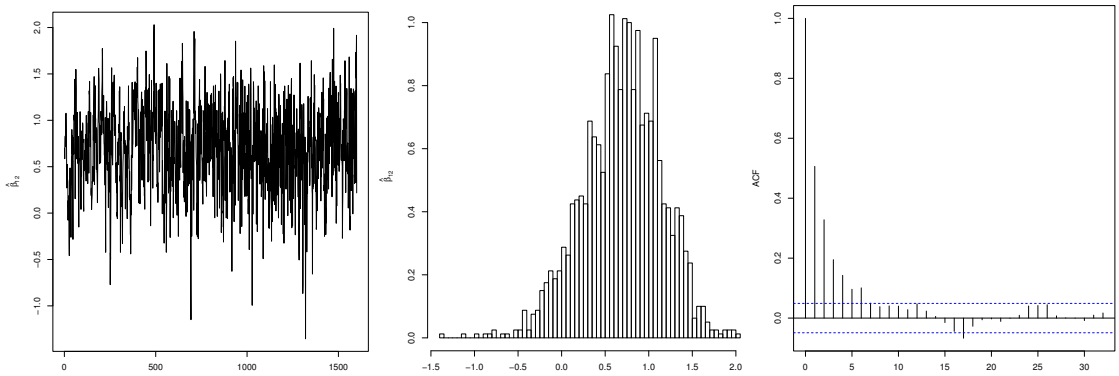

Figure A.13: Checking convergence plots for the estimated parameter $\widehat{\beta}_{12}$: Left panel, the trace for the estimated parameters via MCMC algorithm. Central panel, the approximate marginal density *a posteriori* of the parameter. Right panel, the autocorrelation function plot for the parameter.

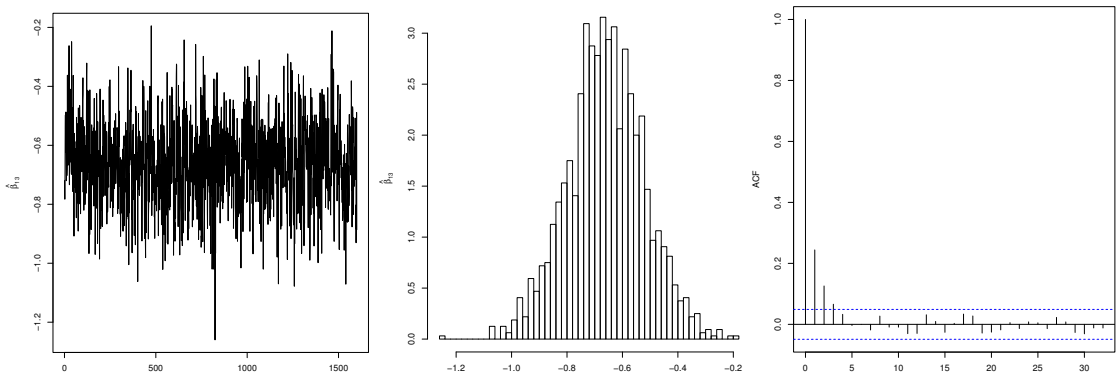

Figure A.14: Checking convergence plots for the estimated parameter $\widehat{\beta}_{13}$: Left panel, the trace for the estimated parameters via MCMC algorithm. Central panel, the approximate marginal density *a posteriori* of the parameter. Right panel, the autocorrelation function plot for the parameter.

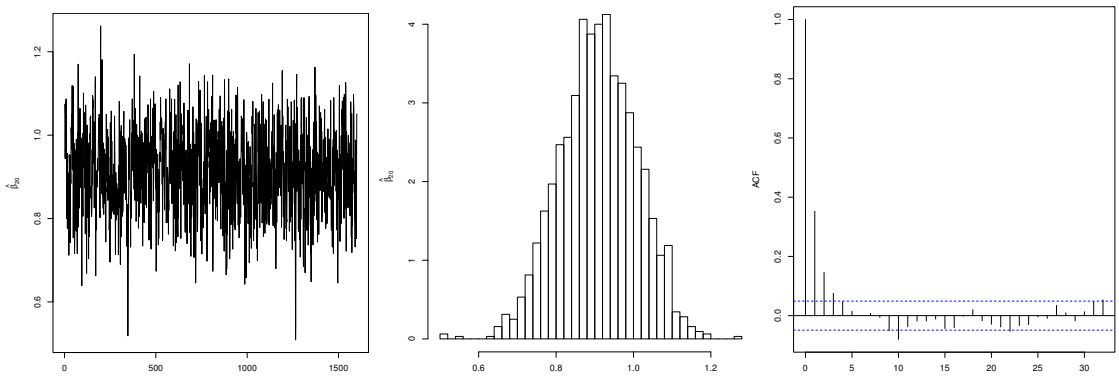

Figure A.15: Checking convergence plots for the estimated parameter $\widehat{\beta}_{20}$: Left panel, the trace for the estimated parameters via MCMC algorithm. Central panel, the approximate marginal density *a posteriori* of the parameter. Right panel, the autocorrelation function plot for the parameter.

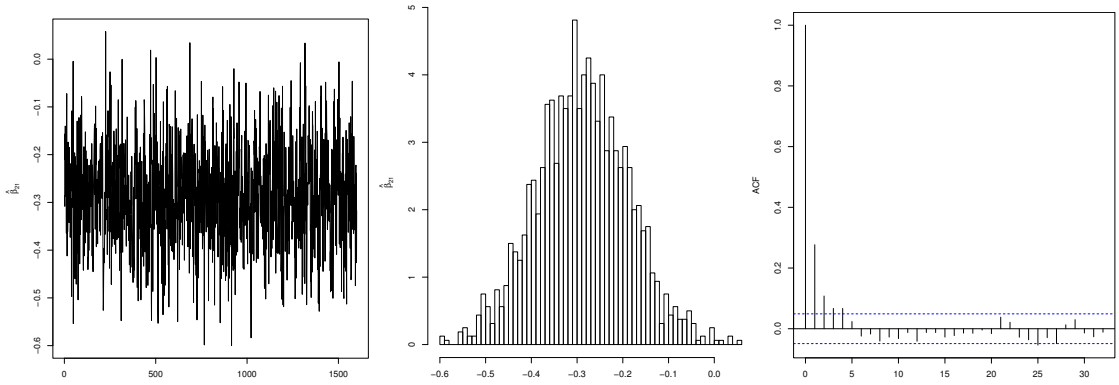

Figure A.16: Checking convergence plots for the estimated parameter $\widehat{\beta}_{21}$: Left panel, the trace for the estimated parameters via MCMC algorithm. Central panel, the approximate marginal density *a posteriori* of the parameter. Right panel, the autocorrelation function plot for the parameter.

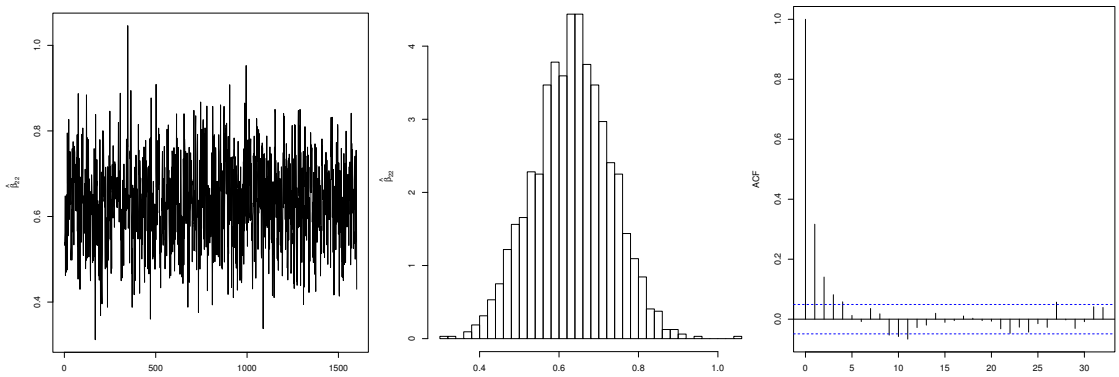

Figure A.17: Checking convergence plots for the estimated parameter $\widehat{\beta}_{22}$: Left panel, the trace for the estimated parameters via MCMC algorithm. Central panel, the approximate marginal density *a posteriori* of the parameter. Right panel, the autocorrelation function plot for the parameter.

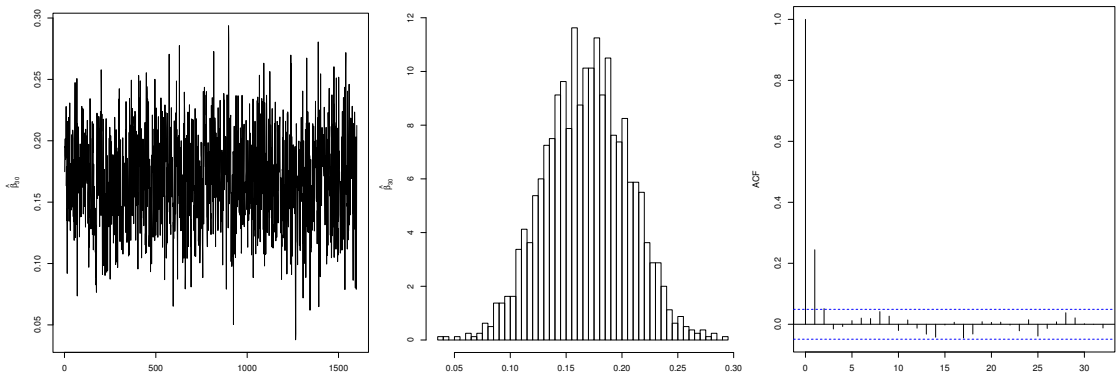

Figure A.18: Checking convergence plots for the estimated parameter $\widehat{\beta}_{30}$: Left panel, the trace for the estimated parameters via MCMC algorithm. Central panel, the approximate marginal density *a posteriori* of the parameter. Right panel, the autocorrelation function plot for the parameter.

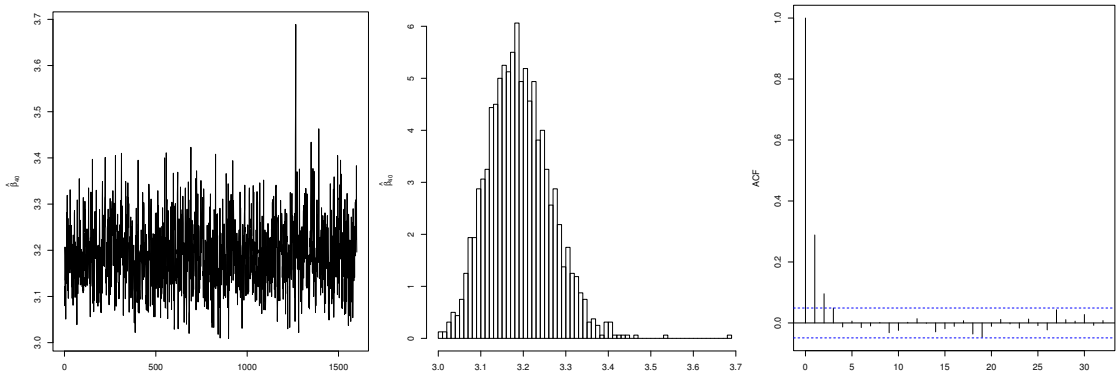

Figure A.19: Checking convergence plots for the estimated parameter $\widehat{\beta}_{40}$: Left panel, the trace for the estimated parameters via MCMC algorithm. Central panel, the approximate marginal density *a posteriori* of the parameter. Right panel, the autocorrelation function plot for the parameter.

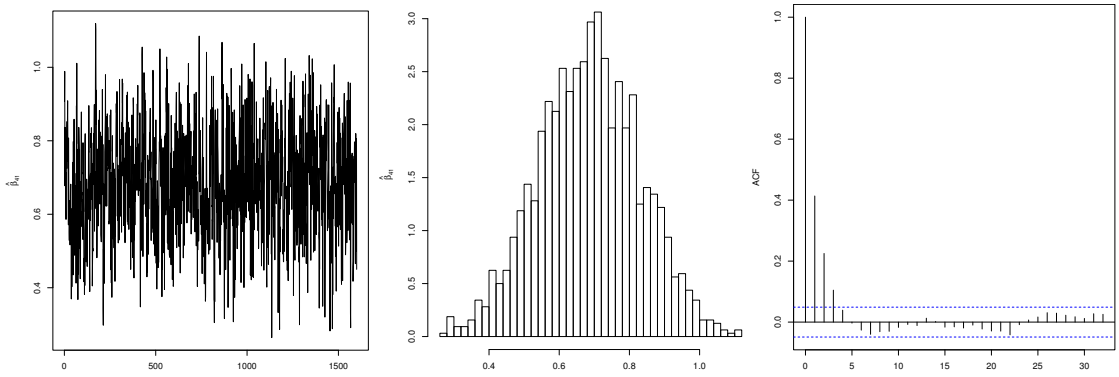

Figure A.20: Checking convergence plots for the estimated parameter $\widehat{\beta}_{41}$: Left panel, the trace for the estimated parameters via MCMC algorithm. Central panel, the approximate marginal density *a posteriori* of the parameter. Right panel, the autocorrelation function plot for the parameter.

References

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. 10

Abreu, H. (2004). Aplicaçao da análise de sobrevivência em um problema de credit scoring e comparaçao com a regressao logıstica. *Biblioteca Digital de Teses e Dissertaç oes da Universidade Federal de Sao Carlos*. 10, 11

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**(6), 716–723. 24

Atchison, J. & Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika*, **67**(2), 261–272. 78

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. 9

Barreto-Souza, W. (2015). Long-term survival models with overdispersed number of competing causes. *Computational Statistics & Data Analysis*, **91**, 51–63. 87

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. 11, 50

Barry, S. C. & Welsh, A. H. (2002). Generalized additive modelling and zero inflated count data. *Ecological Modelling*, **157**(2), 179–188. 12

BCBS (2006). *International Convergence of Capital Measurement and Capital Standards: A Revised Framework, Comprehensive Version (June 2006 Revision)*. Basel Committee on Banking Supervision. 1, 7, 13, 14, 66, 67, 70

BCBS (2015). *Developments in credit risk management across sectors: current practices and recommendations*. Basel Committee on Banking Supervision. 14, 66

Beckman, R. & Tiet jen, G. (1978). Maximum likelihood estimation for the beta distribution. *Journal of Statistical Computation and Simulation*, **7**(3-4), 253–258. 78

Bellotti, T. & Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, **60**(12), 1699–1707. 10

Berger, J. O. & Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *Journal of the American Statistical Association*, **84**(405), 200–207. 25

Berger, J. O. & Bernardo, J. M. (1992). Ordered group reference priors with application to the multinomial problem. *Biometrika*, **79**(1), 25–37. 25

Berkson, J. & Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**(259), 501–515. iv, v, 3, 11, 13, 16, 18, 19, 21, 44, 46, 48, 50, 86

Bernardo, J. M. (1979). Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 113–147. 25

Blackwood, L. G. (1991). Analyzing censored environmental data using survival analysis: single sample techniques. *Environmental Monitoring and Assessment*, **18**(1), 25–40. 12

Braekers, R. & Grouwels, Y. (2016). A semi-parametric Cox's regression model for zero-inflated left-censored time to event data. *Communications in Statistics-Theory and Methods*, **45**(7), 1969–1988. 12

Bussab, W. d. O. & Morettin, P. A. (2005). *Estatística Básica*. Saraiva. 70

Calabrese, R. (2014). Downturn loss given default: Mixture distribution estimation. *European Journal of Operational Research*, **237**(1), 271–277. iv, v, 15, 16, 17, 68, 71, 85, 86

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. 26

Chen, M.-H., Ibrahim, J. G. & Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, **94**, 909–919. iv, v, 16, 17, 46, 48, 49, 50, 63, 86

Colosimo, E. A. & Giolo, S. R. (2006). *Análise de sobrevivência aplicada*. Edgard Blücher. 19

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. 12

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). 26

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. 11

Crook, J. N., Edelman, D. B. & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, **183**(3), 1447–1465. 14

Damien, P. & Walker, S. G. (2001). Sampling truncated normal, beta, and gamma densities. *Journal of Computational and Graphical Statistics*, **10**(2), 206–215. 78

Engelmann, B. & Rauhmeier, R. (2011). *The Basel II Risk Parameters: Estimation, Validation, Stress Testing - with Applications to Loan Risk Management*. Springer Science & Business Media. 13, 65

Ferrari, S. & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**(7), 799–815. 69

Frederic, P. & Lad, F. (2008). Two moments of the logitnormal distribution. *Communications in StatisticsâĂŤSimulation and Computation®*, **37**(7), 1263–1269. 78

Gupta, A. K. & Nadarajah, S. (2004). *Handbook of beta distribution and its applications*. CRC press. 78

Gürtler, M. & Hibbeln, M. (2013). Improvements in loss given default forecasts for bank loans. *Journal of Banking & Finance*, **37**, 2354–2366. 14

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)*, pages 523–541. 11

Hlawatsch, S. & Ostrowski, S. (2011). Simulation and estimation of loss given default. *The Journal of Credit Risk*, **7**(3), 39. 15, 16, 67, 68

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression, 2nd edition*. John Wiley and Sons, New York. 28

Jeffreys, H. (1939). *Theory of Probability*. OUP Oxford. 25

Ji, Y., Wu, C., Liu, P., Wang, J. & Coombes, K. R. (2005). Applications of beta-mixture models in bioinformatics. *Bioinformatics*, **21**(9), 2118–2122. 71

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1994). Continuous univariate distributions , vol. 1john wiley & sons. *New York*, page 163. 78

Jones, M. (2009). Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology*, **6**(1), 70–81. 78

Klein, J. & Moeschberger, M. (2003). Survival analysis: statistical methods for censored and truncated data. *Springer, New York*. 23

Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, **46**(1), 79–88. 78

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**(1), 1–14. 11

Leow, M. & Crook, J. (2014). Intensity models and transition probabilities for credit card loan delinquencies. *European Journal of Operational Research*, **236**(2), 685–694. 13

Leow, M. & Crook, J. (2016). A new mixture model for the estimation of credit card exposure at default. *European Journal of Operational Research*, **249**(2), 487–497. 13

Leow, M. & Mues, C. (2012). Predicting loss given default (LGD) for residential mortgage loans: A two-stage model and empirical evidence for UK bank data. *International Journal of Forecasting*, **28**(1), 183–195. 14

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. 11, 14

Li, C.-S., Taylor, J. M. & Sy, J. P. (2001). Identifiability of cure models. *Statistics & Probability Letters*, **54**(4), 389–395. 50

Liu, L., Huang, X., Yaroshinsky, A. & Cormier, J. N. (2015). Joint frailty models for zero-inflated recurrent events in the presence of a terminal event. *Biometrics*, **72**(1), 204–214. 12

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. 12

Loterman, G., Brown, I., Martens, D., Mues, C. & Baesens, B. (2012). Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*, **28**, 161–170. 13, 14, 65

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. 10, 11

Louzada, F., Oliveira, M. R. & Moreira, F. F. (2015). The zero-inflated cure rate regression model: Applications to fraud detection in bank loan portfolios. *arXiv preprint arXiv:1509.05244*. x, 48

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. 12

Martínez, R. O. (2008). *Modelos de regressao beta inflacionados*. Ph.D. thesis, Universidade de São Paulo. 70

Mateluna, D. I. G. (2014). *Extensões em modelos de sobrevivência com fração de cura e efeitos aleatórios*. Ph.D. thesis, Universidade de São Paulo. 50, 51

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press. 28

Migon, H. S., Gamerman, D. & Louzada, F. (2014). *Statistical inference: an integrated approach*. CRC press. 24, 28, 71

Oliveira, M. R. & Louzada, F. (2014a). An evidence of link between default and loss of bank loans from the modeling of competing risks. *Singaporean Journal of Business Economics and Management Studies*, **3**(1), 30–37. 50

Oliveira, M. R. & Louzada, F. (2014b). Recovery risk: Application of the latent competing risks model to non performing loans. *Tecnologia de Crédito*, **88**, 43–53. 14, 50

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. 11

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. 26

Ospina, R. & Ferrari, S. L. (2010). Inflated beta distributions. *Statistical Papers*, **51**(1), 111–126. 78

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. 12, 70

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. 20

Peng, Y. & Zhang, J. (2008). Identifiability of a mixture cure fraitly model. *Statistics & Probability Letters*, **78**, 2604–2608. 51

Pereira, G. H., Botter, D. A. & Sandoval, M. C. (2013). A regression model for special proportions. *Statistical Modelling*, **13**(2), 125–151. 28, 70

Porath, D. (2011). Scoring models for retail exposures. In *The Basel II Risk Parameters*, pages 25–36. Springer. 14, 66

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 24

Rinne, H. (2008). *The Weibull distribution: a handbook*. CRC Press. 26

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. 34, 87

Rocha, R., Nadarajah, S., Tomazella, V. & Louzada, F. (2016). Two new defective distributions based on the marshall–olkin extension. *Lifetime data analysis*, **22**(2), 216–240. 87

Rodrigues, J., Cancho, V. G., de Castro, M. & Louzada-Neto, F. (2009). On the unification of long-term survival models. *Statistics & Probability Letters*, **79**(6), 753–759. 26, 48, 86

Schuermann, T. (2004). What do we know about loss given default? *SSRN Working Paper Series*. 13, 66

Stepanova, M. & Thomas, L. (2002). Survival analysis methods for personal loan data. *Operations Research*, **50**(2), 277–289. 10

Thomas, L. C., Edelman, D. B. & Crook, J. N. (2002). *Credit scoring and its applications*. Siam. 14

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. 10

Tong, E. N., Mues, C. & Thomas, L. (2013). A zero-adjusted gamma model for mortgage loan loss given default. *International Journal of Forecasting*, **29**(4), 548–562. 15, 16, 68

Tong, E. N., Mues, C., Brown, I. & Thomas, L. C. (2016). Exposure at default models with and without the credit conversion factor. *European Journal of Operational Research*, **252**(3), 910–920. 13

Valvonis, V. (2008). Estimating ead for retail exposures for basel ii purposes. *Journal of Credit Risk*, **4**(1), 79–110. 13

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. 12

Yakovlev, A. Y. & Tsodikov, A. D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. World Scientific, Singapore. iv, v, 16, 17, 46, 48, 63, 86

Yao, X., Crook, J. & Andreeva, G. (2015). Support vector regression for loss given default modelling. *European Journal of Operational Research*, **240**(2), 528–538. 14

Yashkir, O. & Yashkir, Y. (2013). Loss given defult: a comparative analysis. *Journal of Risk Model Validation*, **7**(1), 25–59. 13, 65