Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.12785/jsap/030301

Modeling Time to Default on a Personal Loan Portfolio in Presence of Disproportionate Hazard Rates

Francisco Louzada[1,*], *Vicente G. Cancho*[1], *Mauro R. Oliveira*[2] *and Bao Yiqi*[2]

[1] Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, Brazil
[2] Department of Statistics, University of São Carlos, São Carlos, Brazil

Received: 10 Jun. 2014, Revised: 2 Sep. 2014, Accepted: 6 Sep. 2014
Published online: 1 Nov. 2014

Abstract: In this paper we consider a parametric Weibull mixture cure model for modeling time to default on a personal loan portfolio in presence of disproportionate hazard rate. The main contribution of this paper is to evidence that mixture cure models are appropriate for non proportional sceneries, which has not been claimed in recent articles that brings survival analysis approach for credit scoring modeling. A straight comparison with well known proportional hazard mixture cure model presented in Peng and Dear (2000), provides evidence that risk measurements derived from this framework can be greatly affected if required proportional conditions are not satisfied. In fact, taking into account presence of covariates, if covariates levels do not have proportional hazards rate over time, adjustment with models that assume proportional hazard rates will not be appropriate, and then, erroneous measurements may be derived, i.e., under or overestimate expected losses of a portfolio can be observed. Our approach can be seen as a complement to modeling framework presented in Tong et al. (2012) for credit scoring purposes, which require a proportional hazard structure. A credit data from a Brazilian commercial bank illustrates the procedure.

Keywords: Survival Analysis, Disproportionate Hazard Rate, Cox Proportional Hazards Models, Mixture Cure Model

1 Introduction

Statistical models from survival analysis are being proposed in financial risk management as alternative tools to model internal risk parameters, according to development improvements demanded by the Basel II Accord. Due, mainly, by reason that survival techniques can do continuous monitoring of risk over time in a financial institution (Louzada-Neto, 2006; Malik and Thomas, 2009; Thomas, 2010; Bonini and Caivano, 2013).

Indeed, these models have been considered as an evolution compared to traditional methods of credit risk analysis, which are based on a dichotomous response variable (good or bad behavior). In this novel setting, survival models predict not only whether the customer will be in default, but also the time when the default is more likely.

While borrowers might not honor obligations and become bank debtors, losses associated with such bad consumers may be estimated over time from results of such survival analysis techniques. These tools can represent an important improvement in formulation of interest rates that will be charged by banks in their financial activities, and it can become a competitive strategic to earn market share in the actual competitive scenario.

Survival analysis has shown be very useful for modeling banking data, as well as, clinical trial, relapse of diseases, client churning, among others, once they handle characteristically from long term survivors, which are present in a large proportion in loan portfolio data. The most considered models are standard mixture cure model (Berkson and Gage, 1952; Farewell, 1982), Cox PH regression model (Cox, 1972; Breslow, 1975), and finally, Cox proportional hazard (PH) mixture cure model (Peng and Dear, 2000; Sy and Taylor, 2000). Although, in theory, standard PH regression methods assume survival function reaching zero, we showed Cox PH regression may have well performance modeling time to default with long term survivors. For standard mixture cure models are not requested to satisfy PH assumption, which is indispensable to the Cox PH one.

According to our purpose, the aim of this paper is about need of take correctly assumptions on modeling financial data in order to obtain most appropriate risk parameters and, thereafter, effective losses measures associated with bad

* Corresponding author e-mail: louzada@icmc.usp.br

borrowers over time. From calculations based on real data, we reached amounts that revealed Weibull mixture cure model has better measurements, in general, because financial data have not always taken PH assumption satisfied.

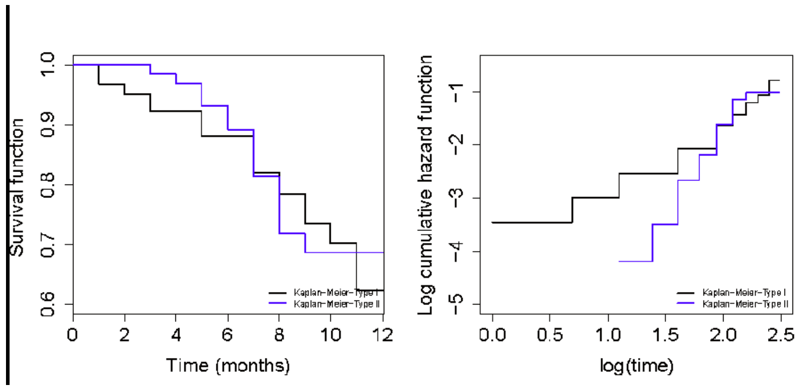

Fig. 1: Kaplan-Meier estimates of surviving, cumulative hazard and log cumulative hazard functions for first sample data stratified by "age range".

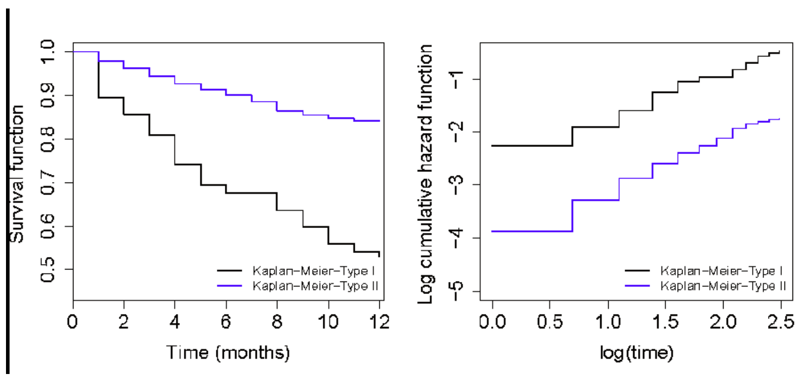

Fig. 2: Kaplan-Meier estimates of surviving, cumulative hazard and log cumulative hazard functions for second sample data stratified by "civil status".

The Figures 1 and 2 show Kaplan-Meier estimate of survival and log cumulative hazard functions from two representative samples of a loan personal portfolio formed by 40.115 customers from a Brazilian commercial bank. In order to preserve confidentiality of costumer characteristics, we refer to them as Type 1 and Type 2 levels.

The data analyzed in this work comes from a real data of a Brazilian bank. To address the proposed theme, we consider two representative samples of clients. In the first sample, it is stratified by age range levels and does not satisfy a PH assumption; on other hand, the second data is stratified by civil status and there is satisfied hazard proportionality condition. The curves of logarithm of cumulative hazard are not parallel for first data sample, which same does not happen with second date. Thus, we have two examples of samples where we shall show measurement errors when assumptions required by model are not noticed.

In this case, a traditional Cox PH regression model or even a PH mixture cure model, as considered in Tong et al. (2012), are not appropriate. In this case, an option is to use a parametric mixture cure model. Here, for simplicity of calculation, we followed a Weibull based structure, what we see is more efficient to fit a data set in presence of proportionate or disproportionate hazard rates.

In summary, our results show that Weibull mixture cure model is sensitive to changes in the behavior of risk of customers over time, fact that may be common in characteristics of applications of bank loans. Therefore, the model can be applied to model both cases displayed by Figure 1 and Figure 2. In addition, we show PH models are not sensitive to these sudden changes in risk behavior, and that models can be use only in conditions as showed in Figure 2.

The organization of remainder of this paper is as follows. In Section 2, we outlined statistical models used in this study. Statistical inference to estimate model parameters are described in Section 3. The illustrations of compared approaches, using a Brazilian banking data, are presented in the Section 4. Section 5 concludes the paper summarizing the results reached.

2 Survival Analysis in Credit Risk Modeling

Since the Basel Accord was revised in 2004 by Basel Committee for Banking Supervision (BCBS, 2004), and replaced by Basel Accord II, it has been used by bank regulatory agencies around world to improve their ability as supervisor of banking environment. The document received several updates since then; it went through financial crisis that began with subprime in 2006 and recently was extended and effectively replaced by Basel III Accord (BCBS, 2010).

The major recommendations contained in the rules issued by Basel II, and maintained in the new Basel III, have demanded that central banks improve their ability to measure systemic risk and incite more prudence and conservatism in risk management setting. It has required banks best risk measurement tools in order to make them keep saving enough capital to safeguard its solvency, eventually to cover losses from risks arising from their banking activities.

In order to calculate how much capital banks need put aside to guard against risks of financial activities, in the first pillar of Basel II, is required capital from different components of risk that a bank faces, ie, credit risk, operational risk, liquidity risk and market risk. For credit risk, which is concern of this paper, three different ways were proposed to be chosen according various degrees of sophistication, namely, in increasing order, Standardized Approach, Foundation IRB Approach and Advanced IRB Approach. The acronym IRB stands for “Internal Rating-Based”.

As has already happened, banking groups have agreed to adopt Basel II to improve their measures of capital at risk adopting advanced IRB approach (Thomas, 2010; Bonini and Caivano, 2013). With this method, banks are allowed to develop internal models to determine parameters that are necessary to calculate the minimum capital required in the Basel II first pillar. These parameters are labelled PD (probability of default), LGD (loss given default), EAD (exposure at default) and, finally, M (maturity).

To model credit risk parameters, methods of survival analysis have been introduced into credit risk area since late 80 (Lane et al., 1986). Survival analysis has showed be complete than standard approaches to predict probability of default. In fact, survival methods may find out more information about risk, showing not only if the borrower will default, but also when the default can occur (Louzada-Neto, 2006; Tong et al., 2012). Therefore, banks can calculate how much of their portfolio is at risk of lost over time. Otherwise, usual methods for modeling a dichotomous variable, usually logistic type regressions, are tools for classification loan application according to the probability of customer become a bad or a good payer (Lee et al., 2002; Lim and Sohn, 2007). And based on this classification the bank can take decision whether the loan will be granted, but will not have subsidies to know when this initial classification of risk would deteriorate over time.

The Cox proportional hazard (Cox PH) regression, which is one particularly kind of method in survival analysis and certainly the mostly chosen modeling in medical research, was firstly applied in credit risk area by Lane et al. (1986) and by Whalen (1991), for modeling time to bankruptcy of institutions in the U.S. financial market. Narain (1992) applied survival methods in a loan bank portfolio to predict the time to default of borrowers. In Witzany et al. (2010) the authors applied Cox PH regression models to modeling the LGD and showed that this method had better performance than a standard logistic regression.

Survival analysis methods such as the mixture cure model and the Cox PH regression have been their performance discussed and compared against standard approaches in credit risk setting in different works, for instance, in Tong et al. (2012), Thomas (2010), Stepanova and Thomas (2002), Abreu (2004) and Banasik et al. (1999). But, our concern in this paper is to show that we can fall into error if we do not take precautions as to check the validity of assumptions required by the PH models.

Survival analysis aims model time to occurring specific events. For example, in the medical area it has been used to model the survival times of patients in cancer clinical trials. In credit risk area, the interesting event is when a bank borrower will not pay his obligations in some time after the loan be granted, in another instance, the event of interest may be the time that a customer will abandon a financial product as canceling a credit card.

In standard survival analysis, the survival function $S_0(t)$ is the probability of observing a survival time greater than some stated value t, which is formulated by $S_0(t) = P(T > t)$, where a random variable T is defined as survival time of an individual. Hence this formulation presumes that $S_0(t)$ tends to zero as time extends and that all observations will experience the event of interest. However, there are examples in credit scoring, where a substantial proportion of costumers may not experience event of credit default during loan lifetime; therefore they are good borrowers, i.e., costumers are default-free. They have been called as long term survivors, for those standard survival methods are not suitable to be used.

Mixture cure models are an extension of standard survival models which have been used in the medical area to model survival times in data set where there is a large proportion of long term survivors, ie, patients that will not experience the event of interesting, referred to as the cured ones (Sy and Taylor, 2000). The proposed models can be classified as parametric and semi parametric methods. This modeling assumes a binary distribution to model patients who have susceptibility to occurs the interested event, namely incidence model component, and it has also a parametric or semi parametric time to event distribution namely latency model component. This theory was originally proposed by Berkson and Gage (1952) and the method was further developed in the medical area by Farewell (1982). Kuk and Chen (1992) extended this model by using Cox PH regression for conditional survival function, that is, for latency model component. This model is now known as the Cox PH mixture cure model.

Long-term survivors are common in data set from commercial banks. Tong et al. (2012) have introduced a mixture cure model to this area predicting time to default on a UK commercial bank, where there was a large proportion of long term survivors. In other words, the total borrowers were formed by two kinds of sub populations: those who are good borrowers and will not be defaulters along with those who are susceptible costumers whose will not honor their obligations. In that paper, the authors compared the Cox PH mixture cure model performance to the Cox PH regression model and standard logistic regression. They showed that, for both survival methods, there were good performances displayed by the marginal survival function against Kaplan-Meier estimates stratified by the covariate levels available in the study.

In this paper, we also fit both survival methods, a Cox PH mixture cure model and Cox PH regression model, on a personal loans portfolio at a different account levels, to modeling time to default in the data from a Brazilian commercial bank. Nevertheless, we call attention to the fact that, in general, we can not use the Cox PH method because it is common, in that kind of financial data, account levels do not have proportional failure rates and therefore the method is not applicable. In this case, the Weibull mixture cure model shows to better predict the time to default. The Cox PH and the mixture cure models are outlined as following.

2.1 Cox Proportional Hazards Regression Model

A classic model for analysis of survival data is the PH model introduced by Cox (1972). This is a semi-parametric model which is composed by the product of one nonparametric component and other parametric one. The PH model is given by

$$h(t|\mathbf{x}) = h_0(t)\exp\{\mathbf{x}'\boldsymbol{\beta}\} = h_0(t)\exp\{\beta_1\boldsymbol{x}_1+\cdots+\beta_k\boldsymbol{x}_k\}, \tag{1}$$

where $\mathbf{x} = (\boldsymbol{x}_1,\cdots,\boldsymbol{x}_k)$ is a vector of k covariates, h_0 is a baseline hazard function, it is not specified, which describe how hazard changes over time at baseline levels of covariates, and $h(t|x)$ is the hazard function at time t given the vector of covariates $\mathbf{x}$.

In this model, we can measure different impacts for each covariate in event risk. To illustrate, consider two observations i and i' that differ in their x-values, with the corresponding linear predictors: $\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$ and $\eta_i' = \beta_1 x_{i'1} + \beta_2 x_{i'2} + \cdots + \beta_k x_{i'k}$.

The hazard ratio for these two observations is given by

$$\frac{h_i(t)}{h_{i'}(t)} = \frac{h_0(t)\exp\{\eta_i\}}{h_0(t)\exp\{\eta_{i'}\}} = \exp\{\eta_i - \eta_{i'}\},$$

is independent of time t. Consequently, this mean that the Cox PH modeling has the PH property.

The Cox model can be estimated by the method of partial likelihood (Cox, 1972). The PH model is implement in the survival package developed in R by Therneau (2013) and Terry M. Therneau and Patricia M. Grambsch (2000).

2.2 Mixture Cure Model

To build a mixture cure model we considere two sub populations of accounts based on the susceptibility to default: a segment that will not experience the event default during the loan term and another segment for those that will eventually default. Then it is considered a binary Bernoulli random variable Y, where $Y = 0$ if the account is non-susceptible to default while $Y = 1$ states that account is susceptible and will default at some time point T, though it may be censored in dataset. Let δ be a censoring indicator, with $\delta = 1$ denoting non-censored accounts and $\delta = 0$ denoting censored accounts. So it is easy to see that $Y = 1$ if $\delta = 1$, and Y is unknown if $\delta = 0$. There are then three possible states for the data: defaulted accounts at some time point T ($Y = 1$, $\delta = 1$); accounts that are right-censored at the end of period of study and would eventually default given sufficient exposure time ($Y = 1$, $\delta = 0$); and accounts that do not default in exposure period of study either will not default in future ($Y = 0$, $\delta = 0$).

Let T be a random variable which is defined as the account survival time and S_{pop} denotes a marginal survival function of T for the entire population. The mixture cure model is goven as follows,

$$\begin{aligned} S_{\text{pop}}(t|\mathbf{x},\mathbf{z}) &= P(T>t|\mathbf{x},\mathbf{z}) \\ &= P(Y=0|\mathbf{z})P(T>t|Y=0,\mathbf{x}) + P(Y=1|\mathbf{z})P(T>t|Y=1,\mathbf{x}) \\ &= 1-\pi(\mathbf{z}) + \pi(\mathbf{z})S(t|Y=1,\mathbf{x}), \end{aligned} \tag{2}$$

where $\pi(\mathbf{z}) = P(Y=0|\mathbf{z})$ denotes the proportion of accounts which are susceptible, hence $1-\pi(\mathbf{z})$ denote proportion of accounts non susceptible (or the cured fraction), given a covariate vector $\mathbf{z} = (\mathbf{z}_1,\ldots,\mathbf{z}_p)$ and $S(t|\mathbf{x}) = P(T>t|Y=1,\mathbf{x})$

is the latency or survival function conditional on account being susceptible to default given a covariates vector $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q)$ which may or may not comprise the same covariates as $\mathbf{z}$. From equation (2), we can note that $S_{\text{pop}}(t|\mathbf{x}) \to 1 - \pi(\mathbf{z})$ as $t \to \infty$. The susceptible fraction $\pi(\mathbf{z})$ can be modeled by a logistic regression as proposed by Farewell (1982),

$$\pi(\mathbf{z}) = \frac{\exp\{\mathbf{z}'\boldsymbol{\beta}\}}{1 + \exp\{\mathbf{z}'\boldsymbol{\beta}\}} = \frac{\exp\{\beta_0 + \mathbf{z}_1\beta_1 + \cdots + \mathbf{z}_p\beta_p\}}{1 + \exp\{\beta_0 + \mathbf{z}_1\beta_1 + \cdots + \mathbf{z}_p\beta_p\}}, \tag{3}$$

where $\boldsymbol{\beta}$ is the vector of regression parameters associated with the covariate vector $\mathbf{z}$.

There are a number of parametric model which can be used to model the survival function $S(t|Y = 1)$. In this paper, the Weibull distribution proposed by Weibull et al. (1951) is considered once it is one of most important distribution in survival analysis area and indeed it can accommodate a great variety of forms for the hazard function, all with a single property in common: a monotonous hazard rate (see Lawless (2011)).

2.2.1 Weibull Mixture Cure Model

Considering the mixture cure model given in (2), the survival function of the Weibull mixture cure distribution (WMC) is given by

$$S_{\text{WMC}}(t) = 1 - \pi(\mathbf{z}) + \pi(\mathbf{z}) \exp\left\{-\left(\frac{t}{\lambda}\right)^a\right\}, \tag{4}$$

where $1 - \pi(\mathbf{z})$ is the cure rate (ie, proportion of non-susceptible accounts) given a covariate vector $\mathbf{z}$, $\lambda > 0$ and $a > 0$ are scale and shape parameters, respectively. If $a < 1$ the hazard function is monotonously decreasing, if $a = 1$ it is constant and if $a > 1$ it is monotonously increasing. Note that $\lim_{t\to 0} S_{\text{WMC}}(t) = \pi(\mathbf{z})$, then $S_{\text{WMC}}(t)$ is not proper survival function.

2.2.2 Cox PH Mixture Cure Model

We consider an extended model proposed by Kuk and Chen (1992) by using the Cox PH regression for the latency. Considering the mixture cure model given in (2), the semi-parametric model has conditional survival function given by

$$\begin{aligned} S(t|Y = 1, \mathbf{x}) &= S_0(t|Y = 1)^{\exp\{\mathbf{x}'\boldsymbol{\beta}\}} \\ &= \exp\left\{-\exp(\mathbf{x}'\boldsymbol{\beta}) \int_0^t h_0(u|Y = 1)du\right\}, \end{aligned} \tag{5}$$

where $S_0(t|Y = 1)$ and $h_0(u|Y = 1)$ are the conditional baseline survival and hazard functions, respectively.

Hence, the survival function of the Cox PH with mixture cure rate (CPHMC) is given by

$$S_{\text{CPHMC}}(t) = 1 - \pi(\mathbf{z}) + \pi(\mathbf{z}) \exp\left\{-\exp(\mathbf{x}'\boldsymbol{\beta}) \int_0^t h_0(u|Y = 1)du\right\}. \tag{6}$$

3 Inference

Let us consider the situation where the time-to-default U is not completely observed, and it is subject to right censoring. Let C denote a censored time-to-default. We define time-to-default random variable given by $T = \min\{U, C\}$, such that, in a sample of size n, we then observe $t_i = \min\{u_i, c_i\}$ and $\delta_i = I(y_i \leq c_i)$ is an censor indicator, ie, when $\delta_i = 1$ if t_i is a time-to-default and $\delta_i = 0$ if it is right censored, for $i = 1, \cdots, n$. In this paper, we propose to relate the parameters p (non-default fraction), a (shape parameter) and λ (scale parameter) of the distribution Weibull with cure rate to the covariates vector $\mathbf{z}$, $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. The following link functions were chosen,

$$\log\left(\frac{p_i}{1 - p_i}\right) = \mathbf{z}_i^\top\boldsymbol{\beta} \quad , \quad \log(\lambda_i) = \mathbf{x}_{1i}^\top\boldsymbol{\gamma} \quad , \quad \log(a_i) = \mathbf{x}_{2i}^\top\boldsymbol{\theta}, \tag{7}$$

$i = 1, \cdots, n$, where $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ are vectors of coefficients associated with covariate vectors $\mathbf{z}$, $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. These covariate vectors may be the same, ie, $\mathbf{z} = \mathbf{x}_1 = \mathbf{x}_2$.

From (7), the likelihood function under non-informative censoring is given by

$$L(\boldsymbol{\vartheta}; \mathscr{D}) \propto \prod_{i=1}^n f_{\text{WMC}}(t_i; \boldsymbol{\vartheta})^{\delta_i} S_{\text{WMC}}(t_i; \boldsymbol{\vartheta})^{1-\delta_i}, \tag{8}$$

where $\vartheta = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta})^\top$ is the parameter vector, $\mathscr{D} = (\boldsymbol{t}, \boldsymbol{\delta}, \mathbf{z}, \mathbf{x}_1, \mathbf{x}_2)$ are observed data, with vector of times-to-default $\boldsymbol{t} = (t_1, \ldots, t_n)^\top$, the censoring vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^\top$, the covariate vectors $\mathbf{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^\top$, $\mathbf{x}_j = (\boldsymbol{x}_{j1}, \cdots, x_{jn})^\top$, and $f_{\mathrm{WMC}}(\cdot; \boldsymbol{\vartheta})$ and $S_{\mathrm{WMC}}(\cdot; \boldsymbol{\vartheta})$ are the Weibull density and improper survival function given in (4). The maximum likelihood estimates (MLEs) of the parameter vector $\boldsymbol{\vartheta}$ are obtained by direct maximization of $L(\boldsymbol{\vartheta}; \mathscr{D})$ or $\ell(\boldsymbol{\vartheta}; \mathscr{D}) = \log\{L(\boldsymbol{\vartheta}; \mathscr{D})\}$. In this paper, the software R (see, R Development Core Team, 2009) was used to determine the MLEs numerically. The computational code is available from the authors upon request. Cox PH model estimates are easily obtained by considering the procedure presented in Tong et al. (2012) and will not be discussed further here.

Under suitable regularity conditions, the asymptotic distribution of the MLE, $\hat{\boldsymbol{\vartheta}}$, is a multivariate normal with the mean vector $\boldsymbol{\vartheta}$ and the covariance matrix, which can be estimated by $\{-\partial^2 \ell(\boldsymbol{\vartheta}) / \partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}^T\}^{-1}$ evaluated at $\boldsymbol{\vartheta} = \hat{\boldsymbol{\vartheta}}$, where the required second derivatives are computed numerically.

4 Application

In this application section, we consider two samples of real personal loan portfolio, originally composed by 40.115 loans from a Brazilian bank. These loans were granted between January and February 2009 with a short loan term of 12 months. A costumer was considered defaulter if his had a period of 90 days without loan repayment. We fitted data with the three survival methods: Weibull mixture cure model, Cox PH mixture cure model and Cox PH regression.

The first sample data considered is a representative set stratified into ages range equals to Type 1 or Type 2. The second sample was stratified into civil status equals to Type 1 and Type 2. As stated before, it is common in data set from commercial banks covariates levels do not have proportional hazards rate and erroneous measurements may be derived. We showed in this section an example of them.

4.1 Application with Presence of Disproportionate Hazard

In the introduction of this paper, we saw the Figure 1 which showed Kaplan-Meier estimates of the surviving and log cumulative hazard functions levels of the covariates ages range for the first data set. We can note that the curves of the survival function of the Type I and Type II intersect each other and their curves on the logarithm of cumulative hazard are not parallel, so the proportion hazard rate condition are not satisfied in this situation.

The survival function estimates of the Cox PH mixture cure model, Cox PH regression and Weibull mixture cure model, are shown in the Figure 3 and it shows that the curves of survival function estimates of the Cox PH mixture cure model and Cox PH regression are not intersect each other, which causes the survival function estimates from the Cox PH modeling become far from the Kaplan-Meier estimates after the intersection points. On the other hand, Weibull mixture cure model can capture when the curves of the survival function of the covariates intersect.

Table 2, in the Appendix A, presents the MLE, standard error (SE), lower and upper limit of the confidence interval (LI and UI) and the p-value of the parameters of the Weibull mixture cure survival model and estimate, standard error (SE) and the p-value of the parameters for both Cox PH models.

4.2 Application with Presence of PH

Here we showed results for model fitting in a second data sample which was stratified into two proportional hazard levels: costumers that have Type 1 civil status and costumers that have Type 2 civil status. The Figure 2 shows Kaplan-Meier estimates of the surviving according to different levels of the civil status. We can note that the curves of the survival function and the cumulative hazard function of Type I and Type II do not intersect, and their curves on the logarithm of the cumulative hazard are almost parallel, so in this case, the proportion hazard rate condition is satisfied.

Figure 4 shows the survival function estimates of the Cox PH mixture cure model, Cox PH regression and Weibull mixture cure model. A good fit is observed for both Cox PH modeling as well as Weibull mixture cure model.

Table 3, in the Appendix A, presents the MLE, SE, LI and UI and the p-value of the parameters of Weibull mixture cure survival model and estimate, SE and the p-value of the parameter of both Cox PH models.

4.3 Expected Loss Function Over Time

To illustrate erroneous measurements that can be derived from the results of the estimated models when we do not take into account the presence of disproportional hazard rate, as appeared in the first sample data set, we defined the expected

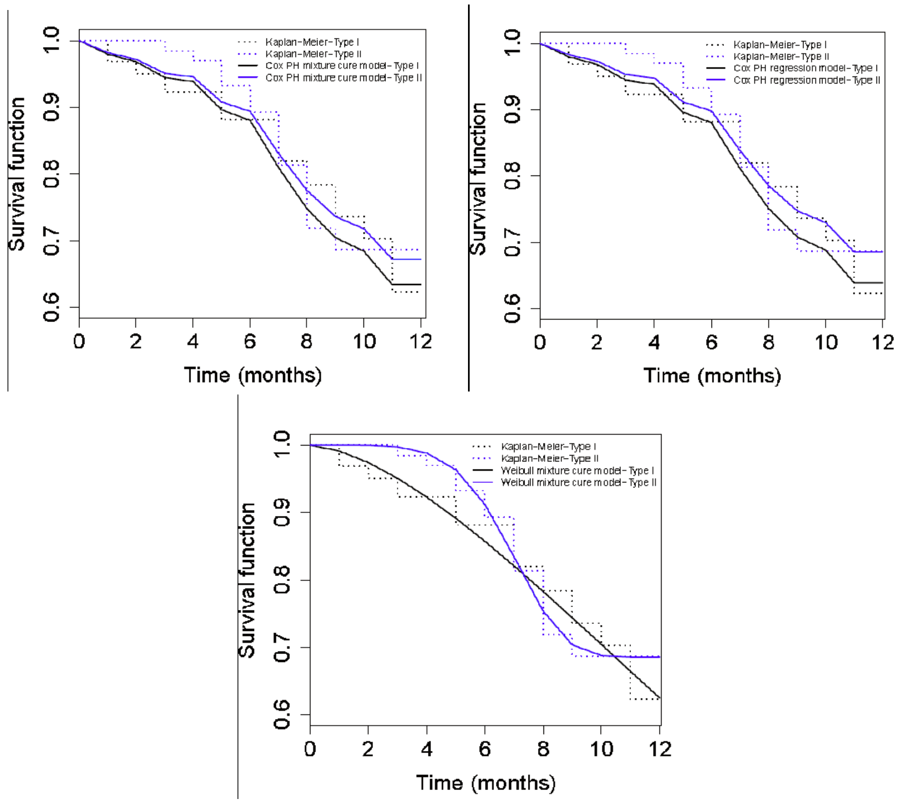

Fig. 3: Survival functions estimated by (a): Cox PH mixture cure, (b:) Cox PH regression and (c): Weibull mixture cure model superimposed on Kaplan-Meier estimates for the first data set.

loss function below, in (9). We show how losses estimated by these models in Section 4.1 are underestimated when the Cox PH models are considered.

Firstly, to define an expected loss (EL) function over time, we made some assumptions to simplify: here, a loan at default was not recovered, ie, the Loss Given Default equals 1. We considered that Exposition at Default (EAD) is a sample of a normal distribution with mean 10.000 and standard deviation equal to 1.500. Thus, the EL function is set to result from one minus the estimated survival function (from the considered models) times the value of the EAD, that is,

$$EL_{\text{expect}}(t) = (1 - S_{\text{estimated}}(t)) * EAD. \tag{9}$$

This approach was applied in the first sample data, where covariate levels do not have proportional hazard rate over time. The results are shown in Table 1. We can see that observed loss is better approximated by the EL obtained from the Weibull mixture cure (WMC) model rather than from the Cox PH ones, showed to be the most appropriate to model the time to default on a personal loan portfolio on the presence of disproportionate hazard rates.

5 Conclusion

Survival analysis techniques have been considered in financial risk management as alternative tools for financial institutions calculate risk measurements over time. Here, we considered time to default modeling found in use today for financial data set, ie, the Cox proportional hazard mixture cure model, Cox proportional hazard regression and Weibull mixture cure model. We evidenced that measurement errors can run if the modeling do not satisfy the PH conditions expected in the Cox PH modeling. These situation are common in data set from commercial banks, and in this case, the Weibull mixture cure model showed to be a feasible option to be used without requiring proportional condition.

To carry out this work, we showed an application of these models in a data base from a Brazilian commercial bank, in order to modeling time up to default of a personal loan portfolio. In that data set was evidenced that stratified samples

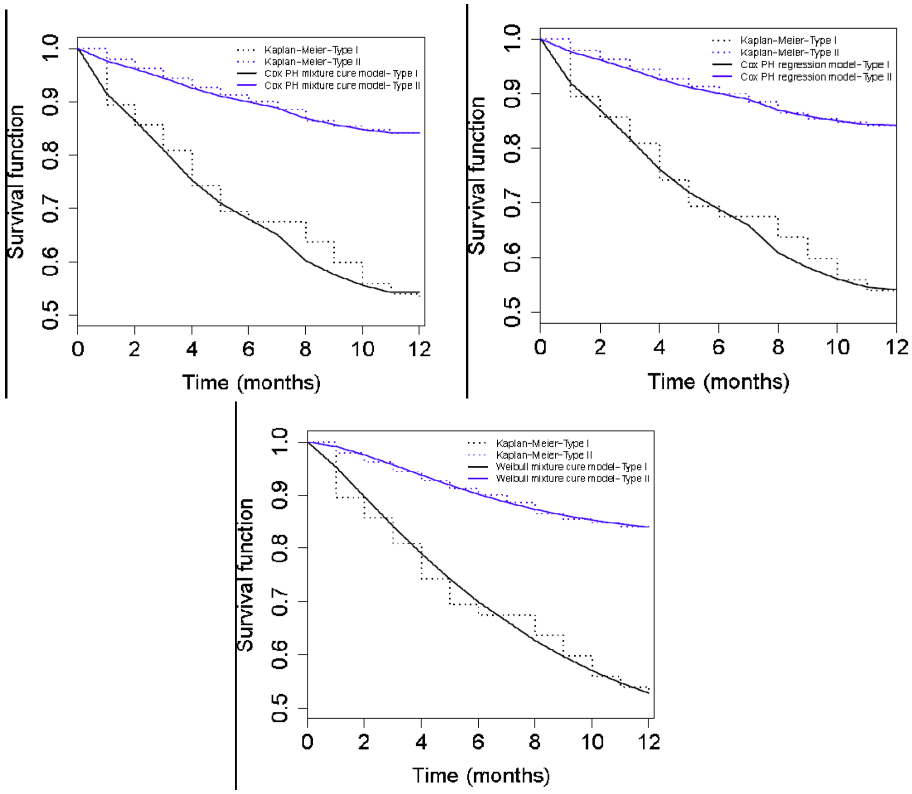

Fig. 4: Survival functions estimated by (a): Cox PH mixture cure, (b:) Cox PH regression and (c): Weibull mixture cure model superimposed on Kaplan-Meier estimates for the second data set.

Table 1: Expected loss function for the first sample (disproportionate hazard rates case).

Ages range = Type 1				
Time	Observed Loss	Expected Loss		
(months)		WMC	Cox PH regression	Cox PH misture cure
6	129.937	158.443	86.153	85.975
7	190.134	198.864	135.869	136.978
8	225.139	241.022	179.303	181.442
9	270.344	284.345	210.085	212.748
10	300.401	328.325	224.556	227.255
11	370.324	372.512	260.088	263.689
12	370.324	416.513	260.088	263.689
Relative difference (%)	-	8.68%	12.39%	13.55%
Ages range = Type 2				
Time	Observed Loss	Expected Loss		
(months)		WMC	Cox PH regression	Cox PH misture cure
6	69.911	63.160	73.378	76.286
7	120.418	119.058	116.426	121.809
8	180.232	177.075	154.506	161.688
9	200.288	213.332	181.784	189.892
10	200.288	224.696	194.696	203.002
11	200.288	226.101	226.654	236.053
12	200.288	226.152	226.654	236.053
Relative difference (%)	-	8.15%	8.64%	8.93%

have proportional hazard rates over time, as well as, others may do not have, and thus, risk measures could be affected if requirements of models are not verified.

The survival functions, estimated considering all models in sample data where they do not have proportional hazard rates between covariate levels, showed that curves of survival function estimates from the Cox PH mixture cure model and from the Cox PH regression do not intersect each other, which causes the survival function estimates of the Cox PH modeling become far from the Kaplan-Meier estimates after the intersect points. On the other hand, the Weibull mixture cure model can reproduce all the Kaplan-Meier curve intersections in a natural way, not requiring the present of a PH structure.

In this sense, with what we claimed previously, banks should continually check the validity of requirements for use of the available models in order to avoid underestimating the risk, enabling have a more appropriated measures of future losses.

Finally, in the context of survival analysis applied to credit risk setting, our approach can be seen as an adequate complement to modeling financial bank data to the modeling framework presented by Tong et al. (2012).

Acknowledgement

The research was sponsored by CNPq and FAPESP, Brazil.
The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

Appendix A

The appendix presents the tables of the MLE, standard error (SE), lower and upper limit of the confidence interval (LI and UI) and the p-value of the parameters of Weibull mixture cure survival model and estimate, standard error (SE) and the p-value of the parameter of both Cox PH models for both data sets.

Table 2: Summaries of the WMC, Cox PH regression model and Cox PH Mixture Cure models fitted to the first data set.

Weibull mixture cure						
Parameter	MLE	SE	LI	UI	$\|Est\|/SE$	p-value
$\beta_{\text{intercept}}$	10.8108	0.0956	10.6235	10.9981	113.1397	<0.0001
β_{civil}	-11.5917	0.0956	-11.7790	-11.4044	121.3120	<0.0001
$\gamma_{\text{intercept}}$	2.9534	0.0962	2.7649	3.1419	30.7120	<0.0001
γ_{civil}	-0.9531	0.1010	-1.1511	-0.7551	9.4355	<0.0001
$\theta_{\text{intercept}}$	0.4767	0.1076	0.2657	0.6876	4.4285	<0.0001
θ_{civil}	1.2011	0.1704	0.8672	1.5350	7.0497	<0.0001
Cox PH regression						
Covariate	Estimate	exp(Estimate)	SE	$\|Est\|/SE$	p-value	
β_{civil}	-0.1703	0.8434	0.1963	0.8680	0.3860	
Cox PH mixture cure						
Covariate	Estimate	exp(Estimate)	SE	$\|Est\|/SE$	p-value	
β	-0.0322	0.9682	0.1582	0.2038	0.8385	
$b_{\text{Intercept}}$	0.4902	-	0.1383	3.5454	0.0004	
b_{civil}	-0.2228	-	0.2177	1.0236	0.3060	

where β is a parameter of the Cox PH regression model and $\boldsymbol{b}$ is a parameter vector of Logistic regression of the Cox PH Mixture Cox Model.

Table 3: Summaries of the WMC, Cox PH regression model and Cox PH Mixture Cure models fitted to the second data set

Weibull mixture cure						
Parameter	MLE	SE	LI	UI	$\lvert Est\rvert/SE$	p-value
$\beta_{\text{intercept}}$	0.3390	0.5238	-0.6876	1.3656	0.6471	0.5177
β_{age}	-1.8814	0.5360	-2.9319	-0.8309	3.5104	0.0005
$\gamma_{\text{intercept}}$	2.0637	0.3549	1.3682	2.7592	5.8156	0.0000
γ_{age}	-0.1415	0.3702	-0.8670	0.5840	0.3823	0.7023
$\theta_{\text{intercept}}$	0.1780	0.1716	-0.1583	0.5143	1.0374	0.2998
θ_{age}	0.2555	0.1962	-0.1290	0.6400	1.3025	0.1930
Cox PH regression						
Covariate	Estimate	exp(Estimate)	SE	$\lvert Est\rvert/SE$	p-value	
β_{civil}	-1.2700	0.2808	0.1657	-7.6620	$<$0.0001	
Cox PH mixture cure						
Covariate	Estimate	exp(Estimate)	SE	$\lvert Est\rvert/SE$	p-value	
β	-0.9862	0.3729	0.0839	11.7514	$<$0.0001	
$b_{\text{Intercept}}$	1.1453	-	0.2263	5.0599	$<$0.0001	
b_{age}	-0.9749	-	0.2359	4.1320	$<$0.0001	

References

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

Banasik, J., Crook, J. N., Thomas, L. C., 1999. Not if but when will borrowers default. Journal of the Operational Research Society, 1185–1190.

BCBS, 2004. International convergence of capital measurement and capital standards: a revised framework. Basel Committee on Banking Supervision.

BCBS, 2010. Basel lll: A global regulatory framework for more resilient banks and banking systems. Basel Committee on Banking Supervision, Basel.

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

Bonini, S., Caivano, G., 2013. The survival analysis approach in basel ii credit risk management: modeling danger rates in the loss given default parameter. Journal of Credit Risk Volume 9 (1), 101–118.

Breslow, N. E., 1975. Analysis of survival data under the proportional hazards model. International Statistical Review/Revue Internationale de Statistique, 45–57.

Cox, D. R., 1972. Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological), 187–220.

Farewell, V., 1982. The use of mixture models for the analysis of survival data with long-term survivors. Biometrics, 1041–1046.

Kuk, A. Y., Chen, C.-H., 1992. A mixture model combining logistic regression with proportional hazards regression. Biometrika 79 (3), 531–541.

Lane, W. R., Looney, S. W., Wansley, J. W., 1986. An application of the cox proportional hazards model to bank failure. Journal of Banking & Finance 10 (4), 511–531.

Lawless, J. F., 2011. Statistical models and methods for lifetime data. Vol. 362. John Wiley & Sons.

Lee, T.-S., Chiu, C.-C., Lu, C.-J., Chen, I.-F., 2002. Credit scoring using the hybrid neural discriminant technique. Expert Systems with applications 23 (3), 245–254.

Lim, M. K., Sohn, S. Y., 2007. Cluster-based dynamic scoring model. Expert Systems with Applications 32 (2), 427–431.

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

Malik, M., Thomas, L. C., 2009. Modelling credit risk of portfolio of consumer loans. Journal of the Operational Research Society 61 (3), 411–420.

Narain, B., 1992. Survival analysis and the credit granting decision. Credit Scoring and Credit Control. Oxford University Press: Oxford, 109–121.

Peng, Y., Dear, K. B., 2000. A nonparametric mixture model for cure rate estimation. Biometrics 56 (1), 237–243.

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

Sy, J. P., Taylor, J. M., 2000. Estimation in a cox proportional hazards cure model. Biometrics 56 (1), 227–236.

Terry M. Therneau, Patricia M. Grambsch, 2000. Modeling Survival Data: Extending the Cox Model. Springer, New York.

Therneau, T. M., 2013. A Package for Survival Analysis in S. R package version 2.37-4. URL http://CRAN.R-project.org/package=survival

Thomas, L. C., 2010. Consumer finance: challenges for operational research. Journal of the Operational Research Society 61 (1), 41–52.

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.

Weibull, W., et al., 1951. A statistical distribution function of wide applicability. Journal of applied mechanics 18 (3), 293–297.

Whalen, G., 1991. A proportional hazards model of bank failure: an examination of its usefulness as an early warning tool. Federal Reserve Bank of Cleveland Economic Review 27 (1), 21–31.

Witzany, J., Rychnovsky, M., Charamza, P., 2010. Survival analysis in lgd modeling. Available at SSRN 1574452.