Volume 36 (2), pp. 89–110
http://orion.journals.ac.za

ORiON
ISSN 0259–191X (print)
ISSN 2224–0004 (online)
©2020

An overview of survival analysis with an application in the credit risk environment

M Smuts* J Allison†

Received: 30 October 2020; Revised: 11 December 2020; Accepted: 11 December 2020

Abstract

Survival analysis has become a popular technique to more accurately model the probability of default in the credit risk environment with the ultimate goal of finding the optimal price for credit. In this paper we present an overview of some of the basic concepts of survival analysis. The focus is specifically on the Cox Proportional Hazards (CPH) model and the mixture cure model, which is a general alternative to the CPH model. A detailed algorithm that can be used to simulate survival times (default times) from a mixture cure model is provided. A parametric CPH and mixture cure model are fitted to a simulated data set and the benefits of fitting the latter model are illustrated.

Key words: Cox proportional hazards model, mixture cure model, probability of default, survival analysis.

1 Introduction

The competitive nature of the financial industry requires the effective use of prescriptive models to assist with strategic decision making. In the credit environment, one of the main challenges is to determine the optimal price (*i.e.*, interest rate) that not only maximises the net interest income to the lender, but also takes into consideration that a customer may default during the loan term.

Analysing or trying to predict when a customer is likely to default is similar to time-to-event modelling in, *e.g.*, medical sciences and engineering. In these fields the use of survival analysis has been well established and proven to produce models with desirable properties for time-to-event data (see *e.g.*, [36], [26] and [34]). Survival analysis was first introduced in the credit environment by Narain [28], where they made use of a parametric accelerated failure time (AFT) model to predict the event of default. Since then a number of authors have started to use more advanced survival analysis models. Some of these include Banasik *et al.* [3], which extended the use of the AFT model and also included a non-parametric Cox Proportional Hazards (CPH) model

*Corresponding author: School of Mathematical and Statistical Sciences, North-West University, Potchefstroom, South Africa, email: smuts.marius@nwu.ac.za

†School of Mathematical and Statistical Sciences, North-West University, Potchefstroom, South Africa, email: james.allison@nwu.ac.za

http://dx.doi.org/10.5784/36-2-690

and Bellotti & Crook [4], which allowed for time varying covariates in the CPH model. Tong *et al.* [35], Dirick *et al.* [10] and Dirick *et al.* [11] introduced the mixture cure model as a more general alternative to the CPH model for modelling data in the credit risk environment. Zhang *et al.* [37] took it a step further and developed a new mixture cure model under different competing risks in order to score online consumer loans.

In this paper, an overview of some of the basic concepts of survival analysis is presented. The focus is specifically on the CPH model and the mixture cure model, which is a general alternative to the CPH model. In many practical applications (*e.g.*, the banking industry), it is still customary to fit the more well known CPH model than the mixture cure model. The reason for this can be attributed to the fact that, since its introduction by Cox [8], the CPH model became the golden standard of models in survival analysis when covariates are present. However, if we are in a situation where a specific event is only ever experienced by a small fraction of the population (*e.g.*, in a loan book of a bank, the event of 'default' will only be experienced by a small minority), fitting a CPH model may lead to estimated survival probabilities which are lower than what is actually experienced. In these cases a mixture cure model should be used to produce more accurate predictions. This will be illustrated by a simple example.

One of the main advantages of using survival analysis (compared to using *e.g.*, logistic regression) is that one can predict not only whether borrowers will default on their loans, but also when they are likely to default. That is, using survival analysis the probability that a borrower will still be repaying a loan at every time instant of the survival curve, *e.g.*, every month, can be accurately estimated. The remainder of the article is organised as follows. In §2 some basic survival analysis concepts are introduced. Estimating the survival function, both parametrically and non-parametrically, is discussed in §3. In §4 the semi-paramteric and parametric CPH model are reviewed. The mixture cure model, which is a generalisation of the CPH model, is discussed in detail in §5. In §6 we provide a detailed algorithm that can be used to simulate survival times (default times) from a mixture cure model. A CPH and mixture cure model are fit on a simulated data set and the benefits of rather fitting the latter model are illustrated. The article concludes in §7 with some final remarks and a possible goodness-of-fit test for mixture cure models with covariates as future research.

2 Basic concepts

Survival analysis is used to analyse data in which the time until an event happens is of interest. In our case the time until a borrower defaults is the event of interest. This event time will be represented by a non-negative continuous random variable, denoted by Y. The distribution function and survival function of Y are given by

$$F(t) = P\left(Y \leq t\right)$$

and

$$S(t) = P\left(Y > t\right) = 1 - F(t),$$

respectively. The survival function represents the probability that a borrower will survive (*i.e.*, not default) up to time t. Another function of interest is the hazard function, $h(t)$, which is the instantaneous risk of defaulting at time t given that the borrower did not default before time t, *i.e.*,

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P\left(t < Y \leq t + \Delta t \mid Y > t\right)}{\Delta t} \\ &= \frac{f(t)}{S(t)}, \end{aligned}$$

where $f(t)$ is the density function of Y. Depending on the distribution of Y the hazard rate can have many different shapes (constant, increasing, decreasing and non-monotone) and is therefore a very useful tool to summarise survival data. Figure 1 depicts the different shapes mentioned.

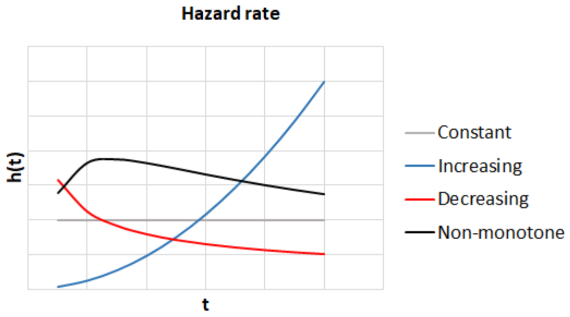

Figure 1: *Hazard rate shapes.*

The cumulative hazard function,

$$H(t) = \int_0^t h(u)\ du,$$

is an increasing function on $(0, \infty)$ and represents the accumulated risk up to time t. Figure 2 shows the cumulative hazard rates corresponding to the hazard rates in Figure 1.

The following relations exist between these various functions $f(t) = -\frac{d}{dt}S(t)$, $h(t) = -\frac{d}{dt}\log S(t)$, $H(t) = -\log S(t)$ and $S(t) = e^{-H(t)}$.

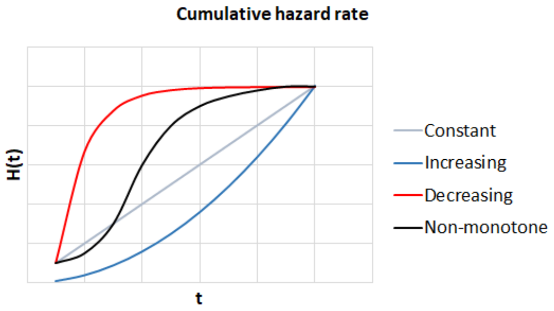

Figure 2: *Cumulative hazard rate shapes.*

In survival analysis a certain proportion of the individuals are censored. This means that some information about the individual's event time (default time) is known, but the exact event time (default time) is not known. The second definition of censoring in [11] will be used, which states that a customer who did not experience default by the moment of data gathering, corresponds to a censored case. In this definition mature cases and early settlement cases are considered censored since the only event of interest is default. Mature cases refer to loans that are repaid in time over the full term whereas early settlement cases refer to loans that are repaid before the end of the full term. Therefore, according to this definition of censoring, only two possible states are considered; default and censored. This censoring scheme is referred to as right censoring and is graphically depicted in Figure 3. In this figure, Λ denotes the censoring time, *e.g.*, the time of data gathering.

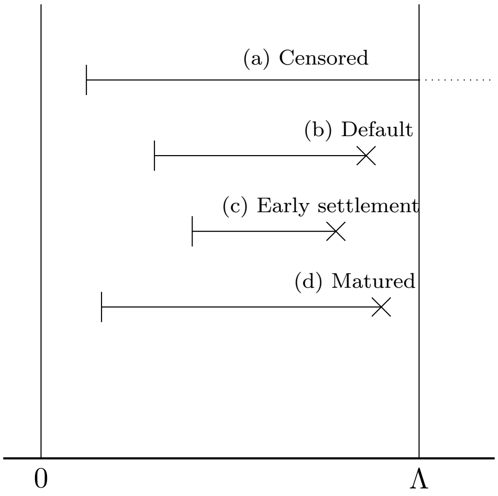

Figure 3: *Right censoring scheme.*

In right censoring there are two latent random variables Y and C. Here, as before, Y is the default time and C is the censoring time. The random variable which is observed is (T, δ), where

$$T = \min(Y, C)$$

and

$$\delta = \begin{cases} 1 & \text{if } Y \leq C \\ 0 & \text{if } Y > C \end{cases}.$$

The following two assumptions are made:

Assumption 1 *Assume that Y and C are independent (this is a standard assumption in survival analysis, see, e.g., [19]).*

Assumption 2 *Assume that the censoring is non-informative, i.e., the distribution of C does not depend on the parameters of interest related to $S(t)$ (also a standard assumption).*

Now, consider a random sample of size n, denoted by (T_i, δ_i), $i = 1, 2, \ldots, n$ with

$$T_i = \min(Y_i, C_i)$$

and

$$\delta_i = \begin{cases} 1 & \text{if } \; Y_i \leq C_i \\ 0 & \text{if } \; Y_i > C_i, \end{cases}$$

where $Y_1, Y_2, \ldots, Y_n$ are the default times and $C_1, C_2, \ldots, C_n$ are the censoring times.

From straight forward calculations and using Assumptions 1 and 2 one obtains the very well known likelihood of (T_i, δ_i), $i = 1, 2, \ldots, n$ given by

$$\begin{aligned} L &= \prod_{i=1}^{n} f(T_i)^{\delta_i} S(T_i)^{1-\delta_i} && (1) \\ &= \prod_{i=1}^{n} S(T_i) h(T_i)^{\delta_i}. && (2) \end{aligned}$$

The likelihood in (2) can be interpreted as follows:

- If $\delta_i = 1$, the loan is considered a default case and the likelihood is simply the probability of surviving until time T_i multiplied by the instantaneous risk of defaulting at time T_i. That is, the borrower survived until time T_i and defaulted immediately after time T_i.
- If $\delta_i = 0$, the loan is considered a censored case and the likelihood is simply the probability of the borrower surviving until time T_i.

The corresponding log-likelihood is given by

$$\log L = \sum_{i=1}^{n} log\left[S(T_i)\right] + \sum_{i=1}^{n} \delta_i log\left[h(T_i)\right]. \tag{3}$$

The estimation of the survival probability, $S(t)$, parametrically and non-parametrically, will now be considered in the presence of censoring.

Any distribution that is defined on $t \in [0, \infty)$, can potentially serve as a lifetime distribution for Y. However, through the years, some distributions were observed to be able to more accurately model the properties of realised survival data (see, *e.g.*, [18]). These include the exponential, Weibull, Gamma, log-normal and linear failure rate distributions.

3 Estimating the survival function

If the assumption is made that every individual has the same survival function (*i.e.*, no individual specific covariates or other individual differences are considered), then the survival function $S(t)$ can easily be estimated either parametrically or non-parametrically. These techniques are discussed below.

3.1 Non-parametric estimation of $S(t)$

If it were possible to completely observe the values of $Y_1, Y_2, \ldots, Y_n$, then $S(t)$ could simply be estimated by the empirical survival function,

$$S_n(t) = \frac{1}{n}\sum_{i=1}^{n} I\left(Y_i > t\right),$$

where $I(\cdot)$ is the indicator function. That is, by using a discrete step function with 'jumps' of size $\frac{1}{n}$ made at each of the observed data points.

However, the observed data in this case are not complete, but rather censored data of the form (T_i, δ_i), $i = 1, 2, \ldots, n$, meaning that the censoring will need to be taken into account when estimating $S(t)$. Kaplan & Meier [17] introduced the product limit estimator to address this issue when working with right censored data. Let $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}$ and $T_{(1)} < T_{(2)} < \cdots < T_{(n)}$ denote the order statistics of $Y_1, Y_2, \ldots, Y_n$ and $T_1, T_2, \ldots, T_n$, respectively, and where $\delta_{(i)}$ is the corresponding censoring indicator variable associated with $Y_{(i)}$ and $T_{(i)}$. The Kaplan-Meier estimator is given by

$$\hat{S}_n(t) = \begin{cases} 1 & \text{if } t \leq T_{(1)} \\ \prod_{j=1}^{k-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_{(i)}} & \text{if } T_{(k-1)} < t \leq T_{(k)},\ k = 2, 3, \ldots, n. \end{cases}$$

This estimator is also a discrete step function, but the jumps now occur only at the event times (in our case the default times). If the largest observation $T_{(n)}$ is censored then $\hat{S}_n(t)$ does not attain 0. Various solutions have been proposed in the literature for this (see *e.g.*, [12]), namely set $\hat{S}_n(t) = 0$ for $t \geq T_{(n)}$ or set $S_n(t) = \hat{S}_n(T_{(n)})$ for $t \geq T_{(n)}$. Figure (4) displays an example of this estimator using a censored data set. The Kaplan-Meier estimator has been studied in detail by many authors in the literature, including Efron [12] and Breslow & Crowley [5]. This estimator is also available in most of the well-known statistical software packages, *e.g.*, the *'survival'* package in R (see [33]).

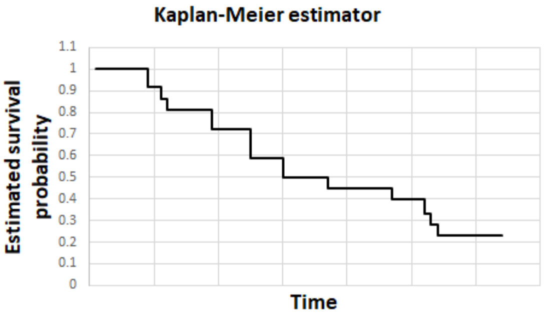

Figure 4: *An example of the Kaplan-Meier estimate of the survival function.*

3.2 Parametric estimation of $S(t)$

An alternative approach to estimating $S(t)$ non-parametrically, is to assume a specific parametric form of the survival function and then estimate the unknown parameters, using maximum likelihood. This idea will be illustrated by an example.

Example 1 *Suppose the assumption is made that the event times are exponentially distri- buted with unknown parameter λ, i.e.,*

$$S(t) = e^{-\lambda t}, \ t > 0.$$

By using the log-likelihood given in (3), λ can now be estimated;

$$\begin{aligned} l(\lambda) = log\left[L(\lambda)\right] &= \sum_{i=1}^{n} log\left[S(T_i)\right] + \sum_{i=1}^{n} \delta_i log\left[h(T_i)\right] \\ &= -\lambda \sum_{i=1}^{n} T_i + log\lambda \sum_{i=1}^{n} \delta_i. \end{aligned}$$

Maximising the log-likelihood yields the estimator

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} T_i}.$$

By the invariance principle of maximum likelihood estimators, the parametric maximum likelihood estimator for $S(t)$ is

$$\hat{S}_n(t) = e^{-\hat{\lambda} t}.$$

By specifying a parametric form for $S(t)$, a smooth function for estimating the survival distribution is obtained. Furthermore, if these parametric models fit the data well, the models tend to give more precise estimates of the quantities of interest (see, *e.g.*, [18]). However, some form of goodness-of-fit testing should be conducted after fitting the parametric models to ensure that the distribution fits to the data. Popular distributions for estimating survival distributions include those discussed in §1. The following are some of the available survival distributions in the *'survival'* package in R: Weibull, exponential, logistic, lognormal and loglogistic.

Both the non-parametric and parametric estimates are only based on event and censoring times. In the vast majority of times, the event time (say default) is a function of one or more covariates *e.g.*, income, risk category, loan amount and interest rate. There exists various models that are able to take these covariates into account when modelling the survival functions. Two of the most widely used models are the accelerated failure time (AFT) model and the Cox Proportional Hazards (CPH) model. In the next section the CPH model will be discussed. For some detail in the AFT model, the interested reader is referred to Chapters 2 and 12 of Klein & Moeschberger [18].

4 Cox Proportional Hazards model

The assumption that has been made thus far is that the survival functions of the individual borrowers are identical. However, in many circumstances this assumption is not reasonable. If, for example, in the medical sciences, there is an interest in the time required for an individual to be cured from a certain disease, factors (called covariates) can influence the time required for an individual to be cured. These covariates may include age, fitness level and smoking status. Similarly, if interested in when a borrower is likely to default on a loan or the time until default, various covariates can influence the time to default. These covariates may include loan amount, risk category, interest rate, loan-to-value, whether or not the borrower is employed, duration at current employment and debt to income ratio. It is thus important that the probability of survival (or probability of default) is modelled, taking into account these covariates. Cox [8] proposed a proportional hazards model which does exactly this, and has become the most commonly used regression model for survival data.

The CPH model is given by

$$h(t|\underline{w}) = h_0(t)e^{\underline{\beta}^T\underline{w}}, \tag{4}$$

where $h(t|\underline{w})$ is the conditional hazard function, $h_0(t)$ is the baseline hazard function and $\underline{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)^T$ is the vector of unknown regression parameters associated with the vector of covariates $\underline{w} = (w_1, w_2, \ldots, w_m)^T$.

Important remarks of the CPH model:

1. The vector of covariates may include continuous factors (*e.g.*, the interest rate and the loan amount), discrete factors (*e.g.*, employment status and risk category) and possible interaction terms (*e.g.*, interest rate and risk category interaction).
2. $h_0(t)$ is called the baseline hazard function, because it denotes the hazard function for borrowers with all covariates equal to 0. That is, $h(t|\underline{w} = \underline{0}) = h_0(t)$, where $\underline{w} = (w_1, w_2, \ldots, w_m)^T = \underline{0} = (0, 0, \ldots, 0)^T$.
3. The CPH model can also be formulated in terms of the conditional cumulative hazard rate function, as follows
$$H(t|\underline{w}) = H_0(t)e^{\underline{\beta}^T\underline{w}},$$
where $H_0(t) = \int_0^t h_0(u)du$ is the baseline cumulative hazard rate function.
4. Using the relationships between the hazard rate, cumulative hazard rate and survival function, the conditional survival function can be written as
$$S(t \mid \underline{w}) = e^{-H(t|\underline{w})} = e^{-H_0(t)e^{\underline{\beta}^T\underline{w}}} = S_0(t)^{e^{\underline{\beta}^T\underline{w}}},$$
where $S_0(t)$ is the baseline survival function (with associated hazard function $h_0(t)$ and associated cumulative hazard function $H_0(t)$).

Previously, the observed data was $(T_i, \delta_i),\ i = 1, 2, \ldots, n$, when no covariates were present. The observed data now consist of the 'triplet' $(T_i, \delta_i, \underline{w}_i),\ i = 1, 2, \ldots, n$, where $\underline{w}_i = (w_{i1}, w_{i2}, \ldots, w_{im})^T$ represents the values of observed covariates.

From (2), the likelihood function based on the data $(T_i, \delta_i, \underline{w}_i)$ and conditional on $\underline{w}_1, \underline{w}_2, \ldots, \underline{w}_n$ is

$$L\left(\underline{\beta} \mid T_i, \delta_i, \underline{w}_i\right) = \prod_{i=1}^{n} S\left(T_i \mid \underline{w}_i\right)\left[h\left(T_i \mid \underline{w}_i\right)\right]^{\delta_i}, \tag{5}$$

where

$$S\left(T_i \mid \underline{w}_i\right) = S_0(T_i)^{e^{\underline{\beta}^T \underline{w}_i}}.$$

Using the relationship

$$h(T_i | \underline{w}_i) = h_0(T_i) e^{\underline{\beta}^T \underline{w}_i},$$

(5) becomes

$$\begin{aligned} L\left(\underline{\beta} \mid T_i, \delta_i, \underline{w}_i\right) &= \prod_{i=1}^{n} S_0(T_i)^{e^{\underline{\beta}^T \underline{w}_i}} \left[h_0(T_i) e^{\underline{\beta}^T \underline{w}_i}\right]^{\delta_i} \\ &= \prod_{i=1}^{n} e^{-H_0(T_i) e^{\underline{\beta}^T \underline{w}_i}} \left[h_0(T_i) e^{\underline{\beta}^T \underline{w}_i}\right]^{\delta_i}. \end{aligned}$$

The corresponding log-likelihood is given by

$$l(\underline{\beta}) = \log\left[L\left(\underline{\beta} \mid T_i, \delta_i, \underline{w}_i\right)\right] = -\sum_{i=1}^{n} H_0(T_i) e^{\underline{\beta}^T \underline{w}_i} + \sum_{i=1}^{n} \delta_i \log h_0(T_i) + \sum_{i=1}^{n} \delta_i \underline{\beta}^T \underline{w}_i. \tag{6}$$

It is clear that the likelihood function contains an unknown (or unspecified) baseline hazard as well as unknown parameters $\underline{\beta}$. The estimation of these unknown quantities, semi-parametrically as well as parametrically, will now be discussed.

4.1 Semi-parametric estimation of the CPH model

The CPH is considered a semi-parametric model if no parametric assumptions are made about the form of the baseline hazard function $(h_0(t))$ but a parametric assumption is made regarding the effect of the covariates on the hazard rate function (in this case the functional form is $e^{\underline{\beta}^T \underline{w}}$). In the semi-parametric setting, Cox [8] proposed a partial likelihood approach to estimate the vector of parameters $\underline{\beta}$, without specifying $h_0(t)$ (see, *e.g.*, [9]). The baseline hazard function (or cumulative baseline hazard function) can then be estimated non-parametrically by the estimator introduced by Breslow [6]. This estimator is obtained by maximising the likelihood of $h_0(t)$ in which the parameters $\underline{\beta}$ are replaced by the maximum partial likelihood estimators $\underline{\hat{\beta}} = \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_m\right)^T$. Denote this estimate by $\hat{h}_0(t)$ and the corresponding estimate for $H_0(t)$ by $\hat{H}_0(t)$ (for more detail on this estimator the interested reader is referred to Breslow [6] and Kalbfleisch & Prentice [16]).

The semi-parametric estimated CPH model is given by

$$\hat{h}(t \mid \underline{w}) = \hat{h}_0(t) e^{\underline{\hat{\beta}}^T \underline{w}}.$$

From Remark 4, the following semi-parametric estimate of the conditional survival function is obtained,

$$\begin{aligned} \hat{S}(t \mid \underline{w}) &= \hat{S}_0(t)^{e^{\underline{\hat{\beta}}^T \underline{w}}} \\ &= e^{-\hat{H}_0(t) e^{\underline{\hat{\beta}}^T \underline{w}}}. \end{aligned} \tag{7}$$

As can be seen from (7), if one estimates the model by making use of the semi-parametric approach, then the estimated conditional survival function will not be able to be expressed using a simple, closed-form (and smooth) expression (this follows from the fact that $\hat{H}_0(t)$ is a discrete non-parametric estimate for $H_0(t)$). However, when the model makes use of the parametric approach, the estimated conditional survival function permits a concise closed-form. While the parametric method requires additional distributional assumptions, the benefit of this approach is that it only requires the estimation of the model parameters (and not the entire hazard function using non-parametric methods), with the result that these models are comparatively parsimonious, easy to understand and can be readily implemented in practice (*e.g.*, in a banking environment). Indeed, if the parametric assumptions hold, then it is well-known that a parametric model will produce more accurate and reliable answers than the semi-parametric (or fully non-parametric) approach.

The parametric estimation of the CPH model and how to test these parametric assumptions, will now be discussed.

4.2 Parametric estimation of the CPH model

In the parametric CPH model an additional distribution assumption is made regarding the baseline hazard function. The assumption is that the baseline distribution is a known lifetime distribution (*e.g.*, one of the distributions discussed in §1), but with unknown parameters. A popular choice in the literature for the baseline distribution is the Weibull function (see *e.g.*, [22] and [37]), specifically in the credit environment. The log-likelihood given in (6) can be used to estimate the CPH model. The following example illustrates this idea.

Example 2 *Suppose a parametric CPH model is to be fitted to the observed data $(T_i, \delta_i, \underline{w}_i)$, $i = 1, 2, \ldots, n$, and the assumption is made that the baseline distribution is Weibull with unknown parameters $\lambda > 0$ and $\alpha > 0$ i.e., $S(t) = e^{-\lambda t^{\alpha}}$. From (6), the log-likelihood is given by*

$$l(\underline{\beta}, \lambda, \alpha) = -\sum_{i=1}^{n} \lambda T_i^{\alpha} e^{\underline{\beta}^T \underline{w}_i} + \sum_{i=1}^{n} \delta_i log\left(\lambda \alpha T_i^{\alpha-1}\right) + \sum_{i=1}^{n} \delta_i \underline{\beta}^T \underline{w}_i.$$

By maximising this log-likelihood (either implicitly or by numerical methods) the maximum likelihood estimators $\underline{\hat{\beta}}$, $\hat{\alpha}$ and $\hat{\lambda}$ are obtained. It then easily follows that the parametric estimate of the conditional survival function is given by

$$\hat{S}(t \mid \underline{w}) = \hat{S}_0(t)^{e^{\underline{\hat{\beta}}^T \underline{w}}} = e^{-\hat{H}_0(t) e^{\underline{\hat{\beta}}^T \underline{w}}} = e^{-\hat{\lambda} t^{\hat{\alpha}} e^{\underline{\hat{\beta}}^T \underline{w}}}.$$

It is clear from this example that a closed-form expression for the conditional survival function is obtained, which is both easy to implement and understand. However, using a parametric instead of a semi-parametric approach requires an additional distributional assumption and it is therefore important to test the validity of this assumption before implementing the model in practise. Diagnostic (or goodness-of-fit) tests for this assump- tion are available in the form of simple graphical techniques, like plotting the martingale, deviance or Schoenfeld residuals (see, *e.g.*, [18]), as well as formal hypothesis tests. One such formal test is based on the Cox-Snell residuals (defined below) which approximately follow a standard exponential distribution when the model is correctly specified. The reason for this is as follows:

Recall that the parametric CPH model for the $i'th$ individual, $i = 1, 2, \ldots, n$ is given by

$$H(t|\underline{w}_i) = -\log S(t \mid \underline{w}_i) = H_0(t) e^{\underline{\beta}^T \underline{w}_i} = H_0(t; \underline{\theta}) e^{\underline{\beta}^T \underline{w}_i}, \tag{8}$$

where $H_0(t;\theta)$ is the notation used to denote the cumulative baseline hazard rate function (assumed to be known) with unknown parameters $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$.

If the model in (8) is correct and $S(t \mid \underline{w}_i)$ is the true conditional survival function of Y_i, $i = 1, 2, \ldots, n$, then, $S(Y_i \mid \underline{w}_i)$ follows a uniform $(0,1)$ distribution. This follows from the well-known probability integral transform which states that, if a random variable X has distribution function F and survival function S, then $F(X) = U$, where U has a uniform $(0,1)$ distribution. Since $S(t) = 1 - F(t)$, the same holds true for $S(X)$. From straight forward calculations it follows that $H(Y_i|\underline{w}_i) = -\log S(Y_i \mid \underline{w}_i)$ has a standard exponential distribution.

However, $H(t|\underline{w}_i)$ will be unknown and is required to be estimated (by making use of the parametric approach discussed earlier in this section). The fitted model is given by

$$H(T_i|\underline{w}_i) = H_0(T_i;\hat{\underline{\theta}})e^{\hat{\underline{\beta}}^T \underline{w}_i},$$

which can be expressed as

$$\hat{\varepsilon}_i = H_0(t;\hat{\underline{\theta}})e^{\hat{\underline{\beta}}^T \underline{w}_i},$$

where $\hat{\varepsilon}_i$, $i = 1, 2, \ldots, n$, are the so-called Cox-Snell residuals (*i.e.*, the fitted values) and $\hat{\underline{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p\right)$ is the vector of estimators for the parameters of the cumulative baseline hazard function.

If the cumulative baseline hazard is correctly specified, then the Cox-Snell residuals should approximately follow a standard exponential distribution (see, *e.g.*, Chapter 11 of Klein & Moeschberger [18] for more detail). One can use any of the multitude tests for exponentiality to test whether this is the case.

5 Mixture cure models

In survival analysis it often happens that, for a certain subgroup of the subjects involved in the study, the event of interest does not happen during the time interval under consideration. This subgroup is referred to as the "cured" group. The population of subjects therefore consists of two subgroups; the "cured" group which is not susceptible to the event of interest as well as a group which is susceptible to the event of interest. As an example, consider a hypothetical loan portfolio of a very small financial institution. The portfolio contains the data of 80 customers that were all awarded home loans on the 1st of January 1999. The event of interest here is whether or not the customer will default on their loan. Upon studying the portfolio, it is found that only 19 customers defaulted on their loans. Figure 5 show the survival probabilities (*i.e.*, the probability of not defaulting on the loan) based on this study.

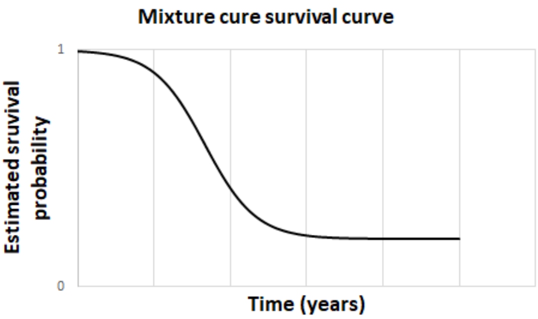

Figure 5: *Illustration of mixture cure survival curve.*

From this figure one can see that there are two distinct subgroups of customers, the group of long term "survivors" in the sense that they "survived" long enough not to experience default during the lifetime of the loan (*i.e.*, not susceptible to default) and the group that did not "survive" (*i.e.*, the group that was susceptible to default and experienced default during the loan term).

Mixture cure models provide a widely used alternative to CPH models (see, *e.g.*, [2]) if the survival trend of the data is similar to the trend shown in Figure 5 (*i.e.*, a certain proportion of the population is not expected to experience the event of interest). As the name suggests, the mixture cure model is used to model the survival probability using a mixture of the two subgroups in the population (*i.e.*, the "cured" and the "not cured" subgroups). Crudely stated, the mixture cure model represents the following survival calculation,

$$P(\text{being alive at time } t) = P(\text{being cured}) + P(\text{not being cured}) \times P(\text{being alive at time } t \text{ if not cured}),$$

or in terms of default as the event of interest,

$$\begin{aligned} P(\text{not defaulting by time } t) &= P(\text{not being susceptible to default}) + P(\text{being susceptible to default}) \times \\ &\quad P(\text{not defaulting by time } t \text{ if susceptible to default}) \\ &= [1 - P(\text{being susceptible to default})] + P(\text{being susceptible to default}) \times \\ &\quad P(\text{not defaulting by time } t \text{ if susceptible to default}). \end{aligned}$$

The model therefore incorporates two components, namely

1. An incidence model to predict which subjects are susceptible. This is typically modelled by making use of a logistic regression model.
2. A latency model that is used to predict the subjects' survival times conditional on the fact that they are susceptible. The CPH model is commonly used to model this component (either with the use of a semi-parametric or parametric approach).

Mixture cure models were first introduced to the field of medical statistics by Farewell [13], whereafter Kuk & Chen [21] and Sy & Taylor [31] generalised some of the results. In addition, Tong *et al.* [35] and Dirick *et al.* [10] recently applied mixture cure models to the credit scoring environment. Other applications of cure models can be found in [30], [23], [24] and [25]. For a comprehensive review of mixture cure models, see the overview paper of Amico & Van Keilegom [2].

Below the model is mathematically introduced and parameter estimation for this model is discussed.

5.1 The model formulation

To model the survival probability in the setting where a large proportion of the population are not susceptible to the event of interest (*e.g.*, a large proportion of the customers in a home loan portfolio are not expected to default), a susceptibility indicator variable is introduced. Let Υ be a random variable that can take on the values 0 and 1, where $\Upsilon = 1$ indicates the customer is susceptible to default and $\Upsilon = 0$ indicates that the customer is not susceptible to default. Similar to before, let Y denote the default time and C the censoring time. The observed random variable is (T, δ), where

$$T = \min(Y, C)$$

and

$$\delta = \begin{cases} 1 & \text{if } Y \leq C \\ 0 & \text{if } Y > C \end{cases}.$$

Now, considering the combination of the susceptibility indicator, Υ, and censoring indicator, δ, the following states are obtained,

1. $\Upsilon = 1$ and $\delta = 1$, indicate that the customer is susceptible and uncensored and therefore the event of interest took place, *i.e.*, the customer defaulted on the loan during the observation period.
2. $\Upsilon = 1$ and $\delta = 0$, indicate that the customer is susceptible and censored and therefore no event took place, *i.e.*, the customer did not default during the observation period but will eventually default.
3. $\Upsilon = 0$ and $\delta = 0$, indicate that the customer is not susceptible and censored and therefore no event will take place, *i.e.*, the customer did not default during the observation period and will not default.
4. $\Upsilon = 0$ and $\delta = 1$, this event cannot be observed since default is impossible if the customer is not susceptible.

It is however important to note that Υ is only observed when $\delta = 1$ and latent otherwise, whereas T and δ are fully observed. In this case, the unconditional survival function of the mixture cure model is given by

$$S\left(t \mid \underline{v}, \underline{w}\right) = \pi(\underline{v}) S\left(t \mid \Upsilon = 1, \underline{w}\right) + \left[1 - \pi(\underline{v})\right], \tag{9}$$

where $\pi(\underline{v})$ denotes the incidence model component with covariate vector $\underline{v} = \left(v_1, v_2, \ldots, v_k\right)^T$ and $S\left(t \mid \Upsilon = 1, \underline{w}\right)$ denotes the latency model component with covariate vector $\underline{w} = \left(w_1, w_2, \ldots, w_m\right)^T$. The incidence model component represents the probability of being susceptible to default and is modelled using logistic regression,

$$\pi(\underline{v}) = P\left(\Upsilon = 1 \mid \underline{v}\right) = \frac{1}{1 + e^{-\underline{\eta}^T \underline{v}}} = \left(1 + e^{-\underline{\eta}^T \underline{v}}\right)^{-1}, \tag{10}$$

where $\underline{\eta} = \left(\eta_1, \eta_2, \ldots, \eta_k\right)^T$ is the vector of unknown parameters associated with the covariates of this component. The latency model component represents the conditional survival probability of the susceptible cases and is modelled using a CPH model,

$$\begin{aligned} S\left(t \mid \Upsilon = 1, \underline{w}\right) &= S_0(t \mid \Upsilon = 1)^{e^{\underline{\beta}^T \underline{w}}} \\ &= e^{-H_0(t \mid \Upsilon = 1) e^{\underline{\beta}^T \underline{w}}}, \end{aligned}$$

using the same notation as in §4.

It should be noted that, when using the mixture cure model, the unconditional survival function tends to $1 - \pi(\underline{v})$ if $t \to \infty$. If, however, $\pi(\underline{v}) = 1$, the unconditional survival function of the mixture cure model reduces to the standard CPH model as is discussed in §4.

Since the mixture cure model comprises of two components, the observed data consists of $(T_i, \delta_i, \underline{v}_i, \underline{w}_i)$, $i = 1, 2, \ldots, n$, where $\underline{v}_i = \left(v_{i1}, v_{i2}, \ldots, v_{ik}\right)^T$ and $\underline{w}_i = \left(w_{i1}, w_{i2}, \ldots, w_{im}\right)^T$ are the covariates of the incidence and latency model, respectively.

From (1), the likelihood function based on the data $(T_i, \delta_i, \underline{v}_i, \underline{w}_i)$, conditional on $\underline{v}_i$ and $\underline{w}_i$, is given by

$$\begin{aligned} L\left(\underline{\eta}, \underline{\beta} \mid T_i, \delta_i, \underline{v}_i, \underline{w}_i\right) = &\prod_{i=1}^{n} \left\{\pi(\underline{v}_i) f(T_i \mid \Upsilon_i = 1, \underline{w}_i)\right\}^{\delta_i} \times \\ &\left\{\pi(\underline{v}_i) S\left(T_i \mid \Upsilon_i = 1, \underline{w}_i\right) + \left[1 - \pi(\underline{v}_i)\right]\right\}^{1-\delta_i}. \end{aligned} \tag{11}$$

By using the relationship

$$f(T_i \mid \Upsilon_i = 1, \underline{w}_i) = S\left(T_i \mid \Upsilon_i = 1, \underline{w}_i\right) h\left(T_i \mid \Upsilon_i = 1, \underline{w}_i\right),$$

the likelihood in (11) becomes

$$L\left(\underline{\eta}, \underline{\beta} \mid T_i, \delta_i, \underline{v}_i, \underline{w}_i\right) = \prod_{i=1}^{n} \left\{\pi(\underline{v}_i) S\left(T_i \mid \Upsilon_i = 1, \underline{w}_i\right) h\left(T_i \mid \Upsilon_i = 1, \underline{w}_i\right)\right\}^{\delta_i} \times \left\{\pi(\underline{v}_i) S\left(T_i \mid \Upsilon_i = 1, \underline{w}_i\right) + \left[1 - \pi(\underline{v}_i)\right]\right\}^{1-\delta_i}. \tag{12}$$

This likelihood can be interpreted as follows:

- If $\delta_i = 1$, customer i is uncensored (*i.e.*, defaulted during the observation period) and therefore susceptible to default and the likelihood contribution is $\pi(\underline{v}_i) S\left(T_i \mid \Upsilon_i = 1, \underline{w}_i\right) h\left(T_i \mid \Upsilon_i = 1, \underline{w}_i\right)$. The likelihood contribution is thus the probability of being susceptible to default multiplied by the probability of surviving until time T_i multiplied by the instantaneous risk of default immediately after time T_i, given the customer is susceptible to default.
- If $\delta_i = 0$, customer i is censored (*i.e.*, did not default during the observation period) and can either be susceptible or not susceptible to default, therefore the likelihood contribution is $\pi(\underline{v}_i) S\left(T_i \mid \Upsilon_i = 1, \underline{w}_i\right) + (1 - \pi(\underline{v}_i))$. That is, the likelihood contribu- tion is the probability of being susceptible to default multiplied by the probability of surviving until time T_i, given the customer is susceptible to default and the probability that the customer is not susceptible to default times the probability of surviving until time T_i, which is 1, since the customer is not susceptible to default.

From the CPH model the likelihood in (12) becomes

$$L\left(\underline{\eta}, \underline{\beta} \mid T_i, \delta_i, \underline{v}_i, \underline{w}_i\right) = \prod_{i=1}^{n} \left\{\pi(\underline{v}_i) e^{-H_0(T_i \mid \Upsilon_i = 1) e^{\underline{\beta}^T \underline{w}_i}} h_0(T_i \mid \Upsilon_i = 1) e^{\underline{\beta}^T \underline{w}_i}\right\}^{\delta_i} \times \left\{\pi(\underline{v}_i) e^{-H_0(T_i \mid \Upsilon_i = 1) e^{\underline{\beta}^T \underline{w}_i}} + \left[1 - \pi(\underline{v}_i)\right]\right\}^{1-\delta_i}.$$

The corresponding log-likelihood is then given by

$$\begin{aligned} l(\underline{\eta}, \underline{\beta}) &= \log\left[L\left(\underline{\eta}, \underline{\beta} \mid T_i, \delta_i, \underline{v}_i, \underline{w}_i\right)\right] \\ &= \sum_{i=1}^{n} \delta_i \left\{\log\pi(\underline{v}_i) - H_0(T_i \mid \Upsilon_i = 1) e^{\underline{\beta}^T \underline{w}_i} + \log h_0(T_i \mid \Upsilon_i = 1) + \underline{\beta}^T \underline{w}_i\right\} + \\ &\quad \sum_{i=1}^{n} (1 - \delta_i) \log\left\{\pi(\underline{v}_i) e^{-H_0(T_i \mid \Upsilon_i = 1) e^{\underline{\beta}^T \underline{w}_i}} + \left[1 - \pi(\underline{v}_i)\right]\right\}, \end{aligned} \tag{13}$$

where

$$\pi(\underline{v}_i) = \left(1 + e^{-\underline{\eta}^T \underline{v}_i}\right)^{-1}.$$

It is clear that the log-likelihood function contains an unknown (or unspecified) baseline hazard as well as unknown parameters $\underline{\eta}$ and $\underline{\beta}$ of the respective mixture cure model components. A discussion will now follow on how these unknown parameters can be estimated using either semi-parametric or fully parametric methods.

5.2 Semi-parametric estimation of the mixture cure model

A mixture cure model is considered a semi-parametric model if no parametric assumptions are made about the form of the baseline hazard function. However, a parametric assump- tion is made regarding the effect of the covariates on the hazard rate function (in this case the functional form is

$e^{\underline{\beta}^T\underline{w}}$) and the effect of the covariates on the probability of being susceptible (with functional form $\log\left(\pi(\underline{v})/(1-\pi(\underline{v}))\right) = \underline{\eta}^T\underline{v}$). In this study, the focus is on the parametric estimation of the mixture cure model, however, Cai *et al.* [7] developed a R-package, *smcure*, that could be used to estimate the mixture cure model semi-parametrically. This semi-parametric estimation procedure uses the Expectation Maximisation (EM) algorithm to deal with the Υ values that are latent (recall that the Υ values are not observed for the censored cases) and the estimation procedure is based on a partial likelihood method proposed by Peng & Dear [30] and Sy & Taylor [31], to estimate the parameters of the conditional survival function without specifying the baseline hazard function. The interested reader is referred to [29], [31] and [32] for more information on the semi-parametric estimation of the mixture cure models.

5.3 Parametric estimation of the mixture cure model

Similar to the parametric CPH model, the fully parametric approach to mixture cure models utilises an additional distributional assumption regarding the baseline hazard function in the latency model component.

Below, it is assumed that the baseline distribution is a known lifetime distribution (*e.g.*, one of the distributions discussed in §1) with unknown parameters. In this case, the log-likelihood given in (13) can be used to estimate the mixture cure model. The following example illustrates this idea.

Example 3 *Suppose it is required to fit a parametric mixture cure model to the observed data* $(T_i, \delta_i, \underline{v}_i, \underline{w}_i)$, $i = 1, 2, \ldots, n$, *and it is assumed that the baseline distribution is Weibull with unknown parameters* $\lambda > 0$ *and* $\alpha > 0$. *From (12), the log-likelihood is given by*

$$\begin{aligned} l(\underline{\eta}, \underline{\beta}, \lambda, \alpha) = {} & \sum_{i=1}^{n} \delta_i \left\{ log\pi(\underline{v}_i) - \lambda T_i^{\alpha} e^{\underline{\beta}^T\underline{w}_i} + \log\left(\alpha\lambda T_i^{\alpha-1}\right) + \underline{\beta}^T\underline{w}_i \right\} + \\ & \sum_{i=1}^{n} (1-\delta_i) \log\left\{ \pi(\underline{v}_i) e^{-\lambda T_i^{\alpha} e^{\underline{\beta}^T\underline{w}_i}} + [1-\pi(\underline{v}_i)] \right\}. \end{aligned}$$

Maximising this log-likelihood (either implicitly or by numerical methods) the maximum likelihood estimators $\hat{\underline{\eta}} = (\hat{\eta}_0, \hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_k)^T$, $\hat{\underline{\beta}} = \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_m\right)^T$,$\hat{\alpha}$ *and* $\hat{\lambda}$ *are obtained, where* $\hat{\eta}_0$ *denotes the estimated intercept term for the incidence model component. It then easily follows that the parametric estimate of the conditional survival function of the mixture cure model is*

$$\begin{aligned} \hat{S}\left(t \mid \underline{v}, \underline{w}\right) &= \hat{\pi}(\underline{v})\hat{S}\left(t \mid \Upsilon = 1, \underline{w}\right) + [1 - \hat{\pi}(\underline{v})] \\ &= \hat{\pi}(\underline{v}) e^{-\hat{H}_0(t) e^{\hat{\underline{\beta}}^T\underline{w}}} + [1 - \hat{\pi}(\underline{v})] \\ &= \left(1 + e^{-\hat{\underline{\eta}}^T\underline{v}}\right)^{-1} e^{-\hat{\lambda} t^{\hat{\alpha}} e^{\hat{\underline{\beta}}^T\underline{w}}} + \left[1 - \left(1 + e^{-\hat{\underline{\eta}}^T\underline{v}}\right)^{-1}\right]. \end{aligned}$$

It is clear from this example that a closed-form expression for the conditional survival function can be obtained using the parametric mixture cure model, which is both easy to implement and to understand, especially for implementing in industry. Amdahl [1] developed the R-package, *flexsurvcure*, to estimate the parameters of a parametric mixture cure model. This package allows covariates to be specified for the latency model, but allows no covariates for the incidence model. This package is therefore not suitable for our purposes, as our incidence model is a function of various covariates.

Naturally the question again arises, how well does this specific parametric model fit the data? The goodness-of-fit of the parametric form of the survival function in a mixture cure model is still largely an open problem in the literature. Maller & Zhou [27] proposed a very informal 'test', where they use the correlation coefficient as test statistic. Very recently, Geerdens *et al.* [15] developed a formal goodness-of-fit test based on the Cramér-von Mises distance between a non-parametric estimator of the mixture cure model and the estimated mixture cure model under the null hypothesis (which states that the survival part has a specific parametric form). They derive the asymptotic distribution of the test statistic and also propose a bootstrap procedure to estimate the critical value of the test. However, both of these tests are only applicable for a mixture cure model where both the incidence model component and the latency model component are not dependent on any covariates. In other words, their test are only applicable if the model in (9) is given by

$$S(t) = \pi S(t \mid \Upsilon = 1) + (1 - \pi).$$

Testing the goodness-of-fit of a fully parametric mixture cure model, where covariates are present, is thus still an open (and seemingly very difficult) problem in the field of both practical and theoretical statistics. The development of such a test is beyond the scope of this article, but in the last section of the article we propose a possible way this can be done.

6 Comparing the fit of a CPH and mixture cure model

Mixture cure models provide a widely used alternative to CPH models (see, *e.g.*, [2]), specifically if a large proportion of the population do not experience the event of interest, *i.e.*, default when considering a credit portfolio. To compare the results obtained when fitting a CPH model as opposed to fitting a mixture cure model, a data set was generated where the default times were simulated from a parametric mixture cure model with a Weibull baseline distribution and using, as covariates, data obtained from a financial institution. More specifically, the loan amount (a_i), price (x_i), repurchase rate (x^0) and overall probability of default (p_i) were chosen as the covariates for both the incidence and latency model components of the mixture cure model for $i = 1, 2 \ldots, 10\,000$ customers. Based on the overall probability of default (p_i) the customers were also categorised as low, medium or high risk.

A parametric approach is followed to estimate the parameters of both the CPH model and a mixture cure model (see §5.3). For the incidence model component $\underline{v}_i = \left(\tau_i y_i,\ x_i - x^0,\ p_i\right)^T$, with corresponding coefficients η_0 (for the intercept term), η_1, η_2 and η_3 and for the latency model component $\underline{w}_i = \left(\tau_i y_i, x_i - x^0, p_i\right)^T$, with corresponding coefficients β_1, β_2 and β_3, where $i = 1, 2, \ldots, 10\ 000$. The true parameter values from which the default times were generated, for both these model components, are displayed in Table 1. The censoring times are assumed to be uniformly distributed between the interval 0 and 200. Using this censoring distribution when generating the censoring times, the proportion of censored customers in the portfolio was roughly 84%. That is, approximately 16% of the customers in the portfolio defaulted on their loans. For the benefit of the reader, the steps followed to generate the survival data from a parametric mixture cure model, where there is only one event of interest (default) are summarised below.

Baseline	Parameters of baseline	η_0	η_1	η_2	η_3	β_1	β_2	β_3
Weibull	$\lambda = 0.0001$ and $\alpha = 1.7$	-2.5	0.02	5	15	0.01	5	12

Table 1: *True parameter values of the mixture cure model.*

Generating the incidence model component data

1. Using the relationship in (10), where the parameter values of $\underline{\eta}$ are given in Table 1, the probability of being susceptible to default, given the covariate values $\underline{v}_i$, is generated *as* $\pi(\underline{v}_i) = P\left(\Upsilon_i = 1 \mid \underline{v}_i\right) = \left(1 + e^{-\underline{\eta}^T \underline{v}_i}\right)^{-1}$, $i = 1, 2, \ldots, 10\ 000$.
2. Simulate Υ_i, $i = 1, 2, \ldots, 10\ 000$ from a discrete distribution,

$$\Upsilon_i = \begin{cases} 1 & \text{with probability } \pi(\underline{v}_i) \\ 0 & \text{with probability } 1 - \pi(\underline{v}_i). \end{cases}$$

Generating the latency model component data

To generate the event times, recall from §4.2 that if $S\left(t \mid \Upsilon_i = 1, \underline{w}_i\right)$ is the true survival function for the susceptible cases, then $S\left(Y_i \mid \Upsilon_i = 1, \underline{w}_i\right)$ follows a uniform $(0, 1)$ distribu- tion *i.e.*, $S\left(Y_i \mid \Upsilon_i = 1, \underline{w}_i\right) = U$, where U denotes an uniform random variable within the interval 0 and 1 and Y_i denotes the default time (month) for customer $i = 1, 2, \ldots, 10\ 000$. Using the relationship

$$U = S\left(Y_c \mid \Upsilon_c = 1, \underline{w}_c\right) = e^{-H_0(Y_c|\Upsilon_c=1)e^{\underline{\beta}^T \underline{w}_c}},$$

the default times for the susceptible customers can be expressed as

$$Y_i = H_0^{-1}\left[-\log(U)e^{-\underline{\beta}^T \underline{w}_c}\right], \tag{14}$$

where $H_0^{-1}(\cdot)$ denotes the inverse of the baseline cumulative hazard rate function (it is assumed that $H_0(\cdot)$ is strictly increasing).

3. Using the relationship in (14) with the corresponding parameter values of $\underline{\beta}$ given in Table 1, the default times for the customers susceptible to default are generated as

$$Y_i = H_0^{-1}\left[-\log(U)e^{-\underline{\beta}^T \underline{w}_c}\right], \ i = 1, 2, \ldots 10\ 000.$$

4. Recall from Assumption 1 that the censoring time and default time are independent, hence the censoring times for each customer i are generated independently from a uniform distribution within the interval 0 and 200 and denoted by C_i, $i = 1, 2, \ldots\ 10\,000$.

Generating the survival data using the incidence and latency model component data

5. Now, since the true value of Υ_i is known, the customers who are not susceptible to default (*i.e.*, the customers for whom $\Upsilon_i = 0$), cannot default and therefore must be censored. Hence, for the customers that are not susceptible to default (*i.e.*, the customers that did not default during the observation period and will not default), the event time is the censoring time. Hence, if $\Upsilon_i = 0$, then $T_i = C_i$ and $\delta_i = 0$, $i = 1, 2, \ldots 10\ 000$.
6. The event time and censoring indicator for the customers susceptible to default (*i.e.*, the customers that either defaulted on the loan during the observation period or did not default during the observation period but will eventually default) are given by,

$$T_i = \min(Y_i, C_i)$$

with

$$\delta_i = \begin{cases} 1 & \text{if } Y_i \leq C_i \\ 0 & \text{if } Y_i > C_i \end{cases}.$$

> Following the steps outlined above, survival data from a mixture cure model can be generated using different baseline distributions and true parameter values.

If the default time Y is Weibull distributed with parameters $\lambda > 0$ and $\alpha > 0$, the (baseline) hazard rate is $h_0(t) = \alpha\lambda t^{\alpha-1}$ and the cumulative baseline hazard rate function and its inverse are given by

$$H_0(t) = \int_0^t h_0(u)du = \lambda t^{\alpha}$$

and

$$H_0^{-1}(t) = \left(\lambda^{-1}t\right)^{1/\alpha}, \tag{15}$$

respectively. The survival data for the simulated data set was subsequently generated using the inverse cumulative baseline hazard function in (15). This resulted in a simulated data set with censored and susceptibility proportions summarised in Table 2. The underlying parameters from which the data were simulated (Table 1) were chosen in such a way that these proportions are realistic in a credit environment (*e.g.*, a small default proportion) so that one can draw valid and sensible conclusions.

Description	**Proportion**
Censored, $\delta = 0$	83.86%
Defaulted, $\delta = 1$	16.14%
Susceptible to default, $\Upsilon = 1$	31.11%
Not susceptible to default, $\Upsilon = 0$	68.89%
Susceptible to default and defaulted, $\Upsilon = 1,\ \delta = 1$	16.14%
Not susceptible to default and censored, $\Upsilon = 0,\ \delta = 0$	68.89%
Susceptible to default and censored, $\Upsilon = 1,\ \delta = 0$	14.97%

Table 2: *Percentages of censored customers and customers susceptible to default for the simulated data (Weibull baseline).*

It is, important to note that even though the value of the susceptibility indicator Υ_i is known when simulating the survival data, the observed data only consists of $(T_i, \delta_i, \underline{v}_i, \underline{w}_i)$ and this is therefore the only data available when fitting the CPH and mixture cure model.

6.1 Fitting the parametric CPH and mixture cure model

To estimate the parameters of the CPH and mixture cure model, the method of maximum likelihood is applied, as discussed in §4.2 and §5.3, respectively. This method is used to estimate the values of the parameters that makes the observed data $(T_i, \delta_i, \underline{v}_i, \underline{w}_i)$ most likely. Example 2 and 3 explain how the unknown parameters of the CPH model and the mixture cure model are estimated when assuming the baseline distribution is Weibull with unknown parameters. To avoid confusion, the maximum likelihood estimators of the CPH model will be denoted by$\underline{\hat{\beta}}^{PH}$,$\hat{\alpha}^{PH}$ and $\hat{\lambda}^{PH}$, whereas for the mixture cure model the estimators will be denoted by $\underline{\hat{\eta}}$, $\underline{\hat{\beta}}$,$\hat{\alpha}$ and $\hat{\lambda}$. Tables 3 and 4 contain the values of these estimates for the CPH and mixture cure model, respectively.

Baseline	**Estimated parameters**	$\hat{\beta}_1^{PH}$	$\hat{\beta}_2^{PH}$	$\hat{\beta}_3^{PH}$
Weibull	$\hat{\lambda}^{PH} = 0.000202$, $\hat{\alpha}^{PH} = 1.7924$	0.006	4.579	5.450

Table 3: *Estimated parameter values of the CPH model.*

Baseline	Estimated parameters	$\hat{\eta}_0$	$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Weibull	$\hat{\lambda} = 0.000094$, $\hat{\alpha} = 1.7197$	−2.542	0.020	6.664	15.649	0.001	6.452	10.259

Table 4: *Estimated parameter values of the mixture cure model.*

To consider the overall effect on the estimated survival probabilities of fitting the CPH model instead of the mixture cure model, the average estimated survival curves over a time period of 60 units for all 10 000 customers, are given in Figure 6.

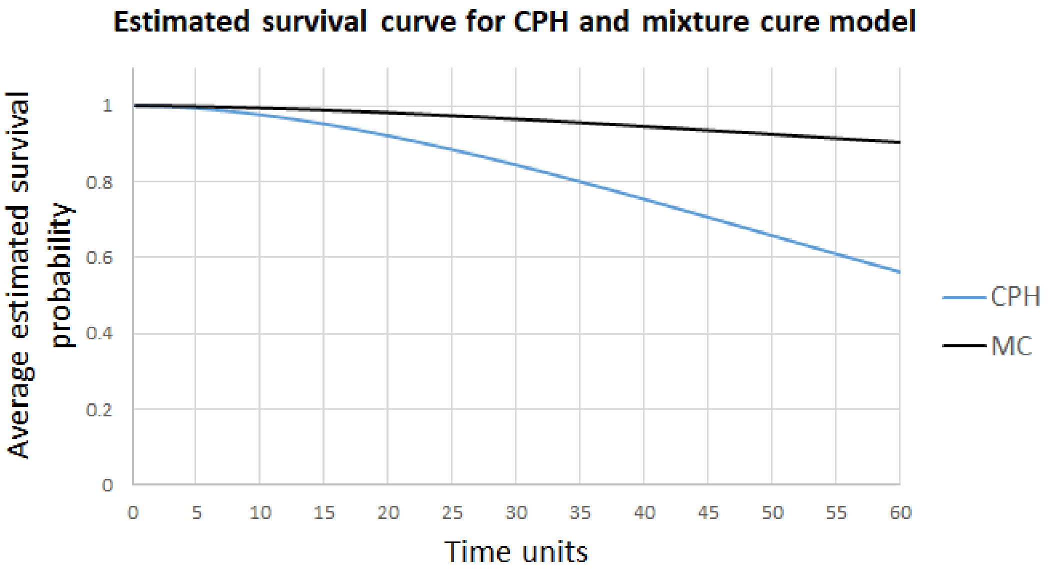

Figure 6: *Average estimated survival curve for the CPH and mixture cure (MC) model.*

After fitting the mixture cure model it was found that the estimated average probability of being susceptible to default is 31.59%. This is close to the true susceptible percentage of 31.11% in the simulated data set (see Table 2). Furthermore, the average unconditional estimated survival probability at the time unit, $t = 200$, was 71.42%, clearly approaching the true percentage of customers not susceptible to default, which is 68.89%. As was shown in §5, this is a nice property of the mixture cure model, that if $t \to \infty$, the unconditional survival probability tends to the non-susceptible percentage. In contrast with this, it was found that the average estimated unconditional survival probability at the time unit, $t = 200$, was 1.01% when fitting the CPH model to the same simulated data set. These findings are also consistent with those depicted in Figure 6. Figure 6 clearly displays that the average estimated survival curve based on the mixture cure model does not tend to zero as $t \to \infty$, but rather to the cure proportion or proportion of customers not-susceptible to default. If we consider Figure 6, *e.g.*, at $t = 35$, one finds that the average estimated probability of default when fitting a CPH model is 20%, whereas for the mixture cure model it is only 4.5%. A higher predicted probability of default will necessarily mean that more economic capital should be held by a bank, which directly impacts their profitability. Taking into consideration that this data set was generated from a mixture cure model, it is obvious that the predicted values of the CPH model are inaccurate.

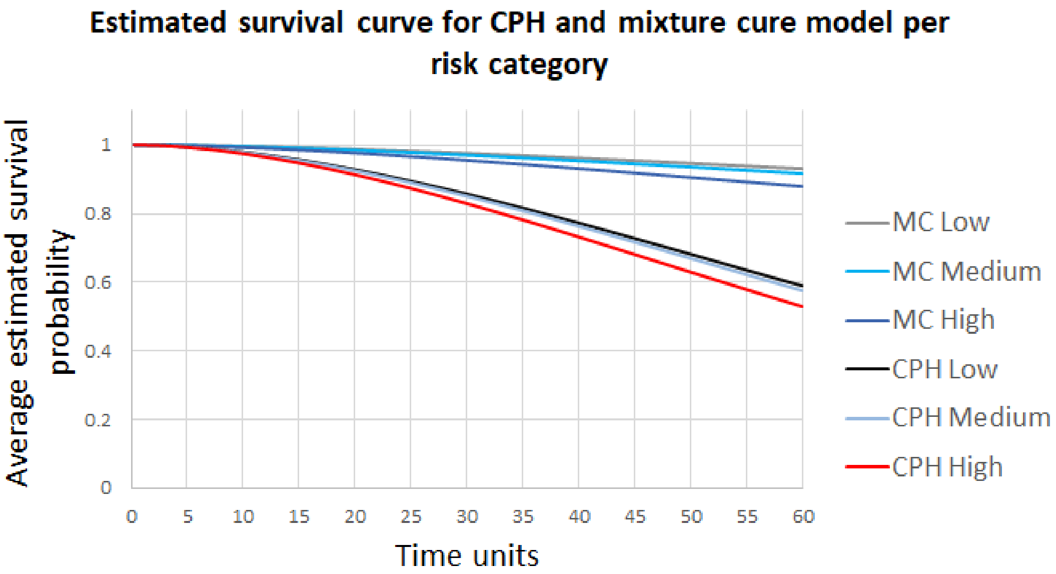

Figure 7: *Average estimated survival curve for the CPH and mixture cure (MC) model per risk category.*

Recall that one of the covariates in the simulated data set was the overall probability of default, which was used to classify an individual as low, medium or high risk. Figure 7 depicts the average estimated survival probability for both the fitted CPH and mixture cure models per risk category. Again it is clear that the average estimated survival probability per risk category for the CPH model is lower than that of the mixture cure model. In addition to this, the average estimated survival curve for the high risk individuals are lower compared to the medium and low risk individuals, as one might expect.

7 Conclusion and future research

This paper provides an overview of some of the basic concepts of survival analysis with the focus specifically on the CPH model and the mixture cure model. In the previous section, it was illustrated that when one fits a CPH model to survival data where a large proportion of non-susceptible individuals are present, it can lead to inaccurate estimated survival probabilities. If, however, there is a negligible proportion of individuals that are non-susceptible to the event of interest, fitting the mixture cure model will almost be identical to fitting a CPH model, since the estimated susceptible proportion will simply be close to one. Hence, in the credit risk environment it will always be a safe bet to rather fit a mixture cure model than a CPH model, since the former is a generalisation of the latter.

We conclude the article by discussing a possible avenue for future research. As was discussed in §5, there is currently no goodness-of-fit test for the parametric mixture cure model where covariates are present. However, such a test is essential to assess whether the parametric assumptions of the mixture cure model are violated. We will discuss a potential omnibus goodness-of-fit test for this scenario. Consider the mixture cure model in (9), where a parametric assumption is made about the baseline distribution. If the model is correctly specified, then $F(Y \mid \underline{v}, \underline{w}) = 1 - S(Y \mid \underline{v}, \underline{w})$ will be uniformly distributed on the interval $[0, \pi(\underline{v})]$, again where this follows from the well-known probability integral transform. This implies that

$$\frac{F(Y \mid \underline{v}, \underline{w})}{\pi(\underline{v})}$$

will be uniformly distributed on $(0, 1)$. However, both $F(\cdot \mid \underline{v}, \underline{w})$ and $\pi(\cdot)$ are unknown, but can easily be estimated from the data $(T_i, \delta_i, \underline{v}_i, \underline{w}_i)$, $i = 1, 2, \ldots, n$. Denote these estimators by $\hat{F}(\cdot \mid \underline{v}, \underline{w})$ and $\hat{\pi}(\cdot)$, respectively. It then follows that

$$\hat{M}_i = \frac{\hat{F}(Y_i \mid \underline{v}_i, \underline{w}_i)}{\hat{\pi}(\underline{v}_i)}$$

should be approximately uniformly distributed if the model is indeed correctly specified. Hence, any uniform $(0, 1)$ test on the basis of $\hat{M}_i$ constitutes in effect a test for the mixture cure model itself.

It is important to note that one will have to use a test for uniformity that is modified to accommodate random censoring. Koziol [20] and Fleming *et al.* [14] discuss extensions of the Kolmogorov-Smirnov and Cramer-Von Mises type tests that can be used for this purpose. Some open questions with regards to this possible new goodness-of-fit test includes:

- How will one obtain the critical values? A model-based bootstrap method seems to be a possible answer, given the complex nature of the problem.
- What is the asymptotic null distribution of the test statistic?
- Is this test consistent?
- How powerful is this test?

References

[1] Amdahl J, 2019, *Flexible Parametric Cure Models*, Available from https://cran.r-project.org/web/packages/flexsurvcure/flexsurvcure.pdf.

[2] Amico M & Van Keilegom I, 2018, *Cure models in survival analysis*, Annual Review of Statistics and its Application, **5**, pp. 311–342.

[3] Banasik J, Crook JN & Thomas LC, 1999, *Not if but when will borrowers default*, Journal of the Operational Research Society, **50(12)**, pp. 1185–1190.

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

[5] Breslow N & Crowley J, 1974, *A large sample study of the life table and product limit estimates under random censorship*, The Annals of Statistics, **2(3)**, pp. 437–453.

[6] Breslow NE, 1972, *Discussion of Professor Cox's paper*, Journal of the Royal Statistical Society, Series B, **34**, pp. 216–217.

[7] Cai C, Zou Y, Peng Y & Zhang J, 2012, *smcure: An R-Package for estimating semiparametric mixture cure models*, Computer Methods and Programs in Biomedicine, **108(3)**, pp. 1255–1260.

[8] Cox DR, 1972, *Regression models and life-tables*, Journal of the Royal Statistical Society: Series B (Methodological), **34(2)**, pp. 187–202.

[9] Cox DR, 1975, *Partial likelihood*, Biometrika, **62(2)**, pp. 269–276.

[10] Dirick L, Claeskens G & Baesens B, 2015, *An Akaike information criterion for multiple event mixture cure models*, European Journal of Operational Research, **241(2)**, pp. 449–457.

[11] Dirick L, Claeskens G & Baesens B, 2017, *Time to default in credit scoring using survival analysis: a benchmark study*, Journal of the Operational Research Society, **68(6)**, pp. 652–665.

[12] Efron B, 1967, *The two sample problem with censored data*, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, **4**, pp. 831–853.

[13] Farewell VT, 1982, *The use of mixture models for the analysis of survival data with long-term survivors*, Biometrics, **38(4)**, pp. 1041–1046.

[14] Fleming TR, O'Fallon JR, O'Brien PC & Harrington DP, 1980, *Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data*, Biometrics, **36(4)**, pp. 607–625.

[15] Geerdens C, Janssen P & Van Keilegom I, 2019, *Goodness-of-fit test for a parametric survival function with cure fraction*, TEST, pp. 1–25.

[16] Kalbfleisch JD & Prentice RL, 1973, *Marginal likelihoods based on Cox's regression and life model*, Biometrika, **60(2)**, pp. 267–278.

[17] Kaplan EL & Meier P, 1958, *Nonparametric estimation from incomplete observations*, Journal of the American Statistical Association, **53(282)**, pp. 457–481.

[18] Klein JP & Moeschberger ML, 2006, Survival analysis: techniques for censored and truncated data, Springer Science & Business Media.

[19] Klein JP, Van Houwelingen HC, Ibrahim JG & Scheike TH, 2016, Handbook of survival analysis, CRC Press.

[20] Koziol JA, 1980, *Goodness-of-fit tests for randomly censored data*, Biometrika, **67(3)**, pp. 693–696.

[21] Kuk AYC & Chen C, 1992, *A mixture model combining logistic regression with proportional hazards regression*, Biometrika, **79(3)**, pp. 531–541,

[22] Lambrecht B, Perraudin W & Satchell S, 1997, *Time to default in the UK mortgage market*, Economic Modelling, **14(4)**, pp. 485–499.

[23] Li C & Taylor JMG, 2002, *A semi-parametric accelerated failure time cure model*, Statistics in Medicine, **21(21)**, pp. 3235–3247.

[24] Liu H & Shen Y, 2009, *A semiparametric regression cure model for interval-censored data*, Journal of the American Statistical Association, **104(487)**, pp. 1168–1178.

[25] Ma S, 2009, *Cure model with current status data*, Statistica Sinica, **19(1)**, pp. 233–249.

[26] Ma Z & Krings AW, 2008, *Survival analysis approach to reliability, survivability and prognostics and health management (PHM)*, pp. 1–20 of: 2008 IEEE Aerospace Conference, IEEE.

[27] Maller RA & Zhou X, 1996, Survival analysis with long-term survivors, John Wiley & Sons.

[28] Narain B, 1992, *Survival analysis and the credit granting decision*, pp. 109–121 in Thomas LC, Crook JN & Edelman DB (Eds), Credit Scoring and Credit Control, Clarendon Press, Oxford.

[29] Peng Y, 2003, *Fitting semiparametric cure models*, Computational Statistics & Data Analysis, **41(3–4)**, pp. 481–490.

[30] Peng Y & Dear KBG, 2000, *A nonparametric mixture model for cure rate estimation*, Biometrics, **56(1)**, pp. 237–243.

[31] Sy JP & Taylor JMG, 2000, *Estimation in a Cox proportional hazards cure model,* Biometrics, **56(1)**, pp. 227–236.

[32] Taylor JMG, 1995, *Semi-parametric estimation in failure time mixture models*, Biometrics, **51(3)**, pp. 899–907.

[33] Therneau TM, 2020, A Package for Survival Analysis in R, R package version 3.1-12.

[34] Tolley HD, Barnes JM & Freeman MD, 2016, *Survival Analysis*, pp. 261–284 of: Forensic Epidemiology, Elsevier.

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

[36] Wei L 1992, *The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis*, Statistics in Medicine, **11(14-15)**, pp. 1871–1879.

[37] Zhang N, Yang Q, Kelleher A & Si W, 2019, *A new mixture cure model under competing risks to score online consumer loans*, Quantitative Finance, **19(7)**, pp. 1243–1253.