Tilburg University

A Methodology for Including the Expected Loss of the Current Default Portfolio into Loss Given Default Measurements

Author:
Janneke van Schijndel (605798)
BSc Tilburg University

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Quantitative Finance and Actuarial Science

Tilburg School of Economics and Management
Tilburg University

Supervised by:
Prof. dr. J.H.J. Einmahl (Tilburg University)
ir. E. Heikens (Rabobank Nederland N.V.)

Second Reader:
dr. G. Nieuwenhuis (Tilburg University)

Public Version
September 18, 2014

Abstract

For different purposes, it is important for the bank to estimate its credit risk. One important risk factor used for determining credit risk is Loss Given Default. This is the percentage of the defaulted exposure that will be a loss for the bank, if a client can not pay the loan anymore. The estimation of this risk factor can be evaluated by comparing the estimated losses by the realized losses. Only at the end of a default process a realized loss is observed. Because most default processes last for several years, for loss given default measurement we use defaults that started several years ago. Therefore, recent trends are not included, which is especially an issue in case of bad financial times. Bad financial times imply an increase of the number of defaults and a possible different character of the occurred defaults. In this thesis a methodology is developed for including the expected loss of the current default portfolio into loss given default measurement. In this way unresolved cases are also incorporated in the process. Recent trends can be noticed and also the possible different character of the current default portfolio can be incorporated partly.

Preface

This thesis is written to obtain my master's title for the master Quantitative Finance and Actuarial Science at Tilburg University. I wrote my thesis during an internship at Rabobank at the Credit team of the department Model Validation, which focuses on validating credit risk models.

I would like to tank all people assisted and supported me to obtain this final result. Especially, I would like to thank Esme Heikens and Erik Winands for guiding me during the internship. Furthermore, I would like to thank Leonie van den Berge and Sebastian Marban for helpful comments and discussion. From Tilburg University, I was supervised by John Einmahl, to whom I am very grateful. Next I want to thank Gert Nieuwenhuis for his participation in the graduation committee. Last, but not leas, I would like to thank my family and friends for their great support.

Janneke van Schijndel

Contents

Chapter 1

Introduction

To complete my Master of Science in Quantitative Finance and Actuarial Science, I have to write a master thesis in this field. I have chosen to write an applied master thesis at Rabobank, one of the biggest banks in the Netherlands. Rabobank is a cooperative bank and focuses on broad financial services. One of the most important services is lending. The bank provides loans to individuals, for example mortgages, and to businesses. Providing these financial services introduces risks for the Rabobank, such as market risk, credit risk and interest rate risk, out of these risks, credit risk is the dominant source of risk for the Rabobank and banks in general. Bank risks are the subject of strict regulatory oversight and policy debate. Credit risk is commonly defined as the loss resulting from failure of obligors to honor their payments. In other words credit risk is the risk that the counter party can not pay back the loan or interest on the moment agreed with the bank. This risk is modeled in several risk models based on different clients, countries and purposes of the loans.

The Basel Committee on Banking Supervision issued three Basel Accords. These consist of recommendations on banking laws and regulations. The first two accords are fully implemented, the deadline of implementation for Basel III is 31 March 2019. Basel II uses a 'three pillars' concept: the first pillar consists of minimum capital requirements, the second of supervisory review and the third of market discipline. In pillar 1 the calculation of minimum capital requirements for three major risks are presented. Credit risk is part of pillar 1. The other components of pillar 1 are operational risk and market risk. The credit risk component can be calculated in two ways: the standardized approach and the IRB approach. IRB stands for Internal Rating-Based Approach. The Basel Accord allows banking organizations, like Rabobank, to calculate their credit risk capital requirements using the IRB approach. Banks that use the IRB approach are supposed to use their own quantitative models to estimate the Probability of Default (PD), Exposure at Default (EaD), Loss Given Default (LGD) and other parameters required for calculating the Risk-Weighted Asset (RWA). These factors together are also known as the Expected Loss (EL) framework.

Now we give a short introduction to the EL framework. To determine the risk on loans the expected loss is calculated. The EL is the product of three factors: the PD, the EaD and the LGD. The PD is a probability and therefore a number between 0 and 1. The EaD is the outstanding amount at default and therefore an amount in euros. The LGD is the economic loss at default expressed as a percentage of EaD. In figure 1.1 the potential credit loss is presented. A distinction is made between three types of potential credit loss: EL, unexpected loss (UL) and Stress loss (SL). The three factors PD, EaD and LGD are used to derive the capital the bank has to hold for EL and UL. The SL is the potential unexpected loss against which it is judged to be too expensive to hold capital against. Unexpected losses of this extent lead to insolvency. The EL are the normal costs of doing business and this loss is covered by provisioning and pricing policies. For EL the formula PD × EaD × LGD is used. The UL is the potential unexpected loss for which capital should be held. The calculation of the unexpected loss (UL) is based on the Vasicek formula. The three factors mentioned are used as input for this formula.

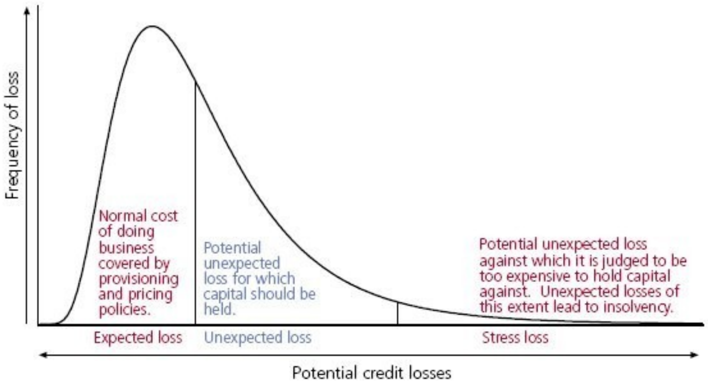
Frequency of loss
Normal cost of doing business covered by provisioning and pricing policies.
Potential unexpected loss for which capital should be held.
Potential unexpected loss against which it is judged to be too expensive to hold capital against. Unexpected losses of this extent lead to insolvency.
Expected loss
Unexpected loss
Stress loss
Potential credit losses

Figure 1.1: The Expected, Unexpected and Stress Loss

The EL is covered by provisioning. For UL capital is held. The bank uses two definitions of capital: Regulatory Capital (RC) and Economic Capital (EC). RC is the amount of capital a bank or other financial institution has to hold as required by its financial regulator. Economic Capital is the amount of risk capital, assessed on a realistic basis, which a firm requires to cover the risks that it is running. Both capital definitions are based on a one year horizon. This means that the capital is derived such that it will cover the UL that can occur in one year.

As stated above LGD is one of the key parameters needed in order to estimate EL and UL. While the PD techniques have been well developed in recent decades, LGD has attracted little attention before 2000. Since the first Basel II consultative papers being published there has been an increasing amount of research on LGD estimation techniques. One of the first papers on the subject is Schuermann (2004) and it provides that the realized LGD is either relatively high (around 70-80%) or low (20-30%). The loss distribution is said to be bimodal (two-humped). Schuermann (2004) also states that recoveries are systematically lower in recessions, and the difference can be dramatic: about one-third lower. Also, the current financial crises has a huge impact on the realized LGD of defaulted loans.

There are several possibilities for the development of the default process of a defaulted loan. A defaulted loan can be cured, restructured or liquidated. If a defaulted loan cures all arrears and interest arrears is paid back. In this case there is no loss for the bank. An other possibility is restructuring. Then the conditions of the loan will be changed in such a way that the counterparty can meet its obligations. This restructuring implies a decrease of the payed interest or a delay of the cashflows. Therefore restructuring implies an economic loss for the bank. However, restructuring does not lead to a direct loss. There is no write off. Although restructuring leads to an economic loss, it is treated if there is no loss. Both a cure and a restructuring cause that a defaulted loan leaves the default portfolio and enters the performing portfolio again. If cure will not occur anymore and restructuring is not an option, the defaulted loan will be liquidated. The collateral of the loan will be sold. If the proceeds of the sale are high enough to cover all arrears, there is no loss. If the proceeds are lower than the arrears, there is a loss for the bank. If this loss is expressed as a percentage of the EaD, it is the realized LGD for this defaulted loan. Notice that if there is no loss, it can still be expressed as a percentage of the LGD, in this case the realized LGD is zero. If a loan is liquidated, all collateral is sold and the loss is written off, the financing agreement is terminated.

In order to estimate the LGD, different models exists. Most LGD models make use of a cure rate (CR) and a Loss Given Loss (LGL). In one of the models the CR is defined as the probability that the defaulted loan cures or is restructured within 12 months after default. This definition is not for all models the same. Some models make use of a CR that gives the probability that a loan cures within 24 months, 36 months or even a lifetime cure. Which definition is used is a decision made by the developers and is based on the purpose of the model and the available data. The CR is used to cover the loans that cure or can be restructured and therefore have a loss (and realized LGD) of zero. The LGL is used to cover the loans that are not cured or restructured, but are liquidated. The definition of the LGL is the percentage of EaD that has to be written off given that the loan will be liquidated. Notice that it is still possible that the LGL is zero, if the collateral value is more than the outstanding amount and the costs of liquidation. In most models, the LGD is modeled as a combination of the CR and the LGL: $\text{LGD} = (1 - \text{CR})\text{LGL}$. Both CR and LGL can be estimated to obtain an estimate of the LGD. The estimated LGD is compared with this realized LGD to test the performance of the estimation.

This thesis is organized as follows. In chapter 2, the current methodology of including expected loss of the current default portfolio into Loss Given Default Measurements of the portfolio is provided. This section also includes the issues on this methodology and a solution is proposed. In chapters 3 and 4 the theoretical framework will be formulated. Firstly, in chapter 3 an introduction on survival analysis is provided. Also the random censorship model and the concept of competing risks are introduced. Chapter 4 the used estimate and its (asymptotic) properties are described. To apply the theory to the formulated problem, some reformulations have to be made. In chapter 5, the translation from theory to application is provided. In chapter 6, the data are described and some assumptions will be checked. Chapter 7 is the combination of the applied theory of chapter 5 and the description of the data of chapter 6. In this chapter the results regarding to our dataset are presented. Chapters 8 consists of two checks regarding two assumptions. Conclusions and further discussion will be given in the final chapter 10.

Chapter 2

Estimation of Expected Loss of the Current Default Portfolio

2.1 Estimation of the Expected Loss for the Portfolio

2.2 Conditional Loss Rate Model

In figure **??**, it is presented that there are two options of resolving: a defaulted case can cure or it can be liquidated. If the case will be liquidated there are two possibilities regarding the loss. It is possible that the value of the underlying collateral is high enough to cover the exposure and the liquidation costs, then the loss is zero. It is also possible that the value of the underlying collateral of a liquidated case is not high enough to pay back the loan, interest, arrears and costs made for liquidation. Then the loss is greater than zero. In table **??** the different release options are presented. If the loss is zero, which happens in case of cure and in case of liquidation with a higher value of collateral than exposure, the release will be 100%. The states in figure **??** can be regrouped. The two options where release is 100% can be combined in one state: the zero loss state. All other options can be grouped in the other state: the loss state. In this state it is known that there will be a loss, but it is not known which amount has to be written off. Therefore, it is also not known what the provision release will be in the loss state. However, it is supposed that the determined provision for a defaulted loan is a good estimate in case the loan will be liquidated with a loss.

In table **??** the provision release rate presented in case of a loss. The provision release rate is in the interval $[0\%, 100\%)$. However, if the loss is higher than the provision determined there is extra loss, this is not taken into account in this table. Therefore, we introduce a slightly adjusted definition of the provision release rate to cover this situation. First introduce the loss provision ratio, denoted by LPR. The LPR for a certain case is the loss made on the loan divided by the total provision determined for this case. If we assume that the provision is not zero, the provision release rate is defined as:

$$1 - \frac{\text{Loss}}{\text{Provision}} \tag{2.1}$$

Notice that in this case the provision release rate can become negative. Now the cases where loss is higher than the determined provision are also taken into account. The provision release rate for liquidated loans with a loss higher than zero can be statistically estimated. However, this not in the scope of this master thesis. As stated above, it is supposed that the determined provisions for the defaulted loans are a good estimated in case the loans will be liquidated with a loss. Therefore, we assume in this thesis, that on a portfolio level the provisions for the liquidated cases with loss are equal to the loss for these cases. Notice that under this assumption the provision release rate for liquidated case with loss is zero. This assumption is considered in section 8.1. Notice that every case will become resolved and will end up in one the two states: loss or zero loss. Using survival analysis (see section 3) the probability to end up in the zero loss state can be estimated and is depended on time t. This probability is called the Zero Loss Rate and is denoted by $ZLR(t)$. Subsequently, the probability to end up in the loss state is $1 - ZLR(t)$. See figure 2.1.

Status	Options	Release
Cured		100 %
Liquidation	value of collateral $>$ exposure	100 %
	value of collateral $<$ exposure	0 %

Table 2.1: Provision Release Possibilities for defaults

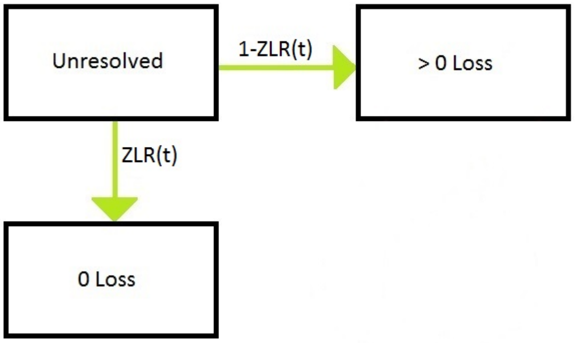

Figure 2.1: The conditional loss rate model considers two final states 0 loss or > 0 loss.

The model presented above is a solution for the problems mentioned in the previous subsection. The liquidated cases with zero loss are not taken into account for the provision release rate. By regrouping the states from cure and liquidation to loss and zero loss, this problem is solved. By applying survival analysis it is also possible to derive a provision release rate conditional on the number of months the case is already in default. This is explained in the next chapter.

An extra complication of the application of survival analysis on the model described above is the presence of competing risks (see section 3.4). Because it is supposed that it is not possible that a liquidated loan would ever cure if the liquidation process was not started, for the conditional cure rate the problem of competing risks is not involved. However, for the zero loss rate we look at zero loss events (cure or liquidation without loss) and loss events (liquidation with loss). In very few cases it would be possible that a case ends up with zero loss if the liquidation process, that gives a loss, was not finished yet. For example, wait longer with the sale of the collateral to get a better price and therefore a no loss. This gives some complications, which are considered in the application of the theory in section 6.1.

Chapter 3

Survival Analysis

A possible approach for deriving the conditional cure rate is survival analysis. In this section, the focus will be on the theory of survival analysis. Survival analysis can be used to derive a lifetime estimate for the cure rate that depends on the number of months in default. The point-in-time estimate does not depend on time. An estimation of the cure rate dependent on time is an advantage of survival analysis compared to the point-in-time estimation, because the subjects of the current default portfolio are already in default for several months, this can be taken into account. The first section will give some explanation about the concepts of survival analysis. An other advantage of survival analysis is that censored data can be incorporated in the model, this is explained in section 3.2. As mentioned before, there are a lot of unresolved cases, these can be treated as censored data. For the use of survival analysis, the data have to be presented in a specific framework, this is presented in section 3.3. For correct application of the survival framework, some assumption are needed. Therefore, the random censorship model is introduced in Section 3.4.

3.1 Concept of Survival Analysis

In survival analysis, the focus lies on a group or groups of individuals for each of whom (or which) there is a defined point event, often called failure, occurring after a length of time called the failure time. The object of primary interest is the survival function. The survival function captures the probability that an observation survives beyond a specified time. Failure can occur at most once for each individual. Examples of failure times include the lifetimes of machine components in industrial reliability, the duration of strikes or periods of unemployment in economics and the time to cure for a default. To determine failure time precisely, there are three requirements:

- a time origin must be unambiguously defined
- a scale for measuring the passage of time must be agreed

- the event of failure must be entirely clear.

The time origin should be precisely defined for each individual. It is also desirable that, subject to any known differences on explanatory variables, all individuals should be as comparable as possible at their time origin. Notice that the time origin need not be and usually is not at the same calender time for each individual. Most projects which use survival analysis, have staggered entry, so cases enter over a substantial time period. Each cases failure time is usually measured from his own date of entry. If we consider the origin time of default, it is clear that the default date is a proper origin time.

Often the 'scale' for measuring time is clock time (real time), although other possibilities certainly arise, such as the use of operating time of a system. If two or more different ways of measuring are available, it may be possible, having selected the most appropriate timescale, to use other 'times' as explanatory variables. In this research, clock time is used.

Finally, the event of failure must be defined precisely. For example in medical work, failure could mean death or death from a specific cause. In the case of the default portfolio, failure could mean cure or become resolved, which is a wider definition. Resolved cases could be cured or liquidated. To apply survival analysis this has to defined precisely. The choice depends on the purpose of the study.

Now the three requirements for survival analysis are defined, it is known what has to be observed for a subject to define the survival time. For every subject, there is one observation, which starts at the start point and which ends at the last time period this subject is observed. Notice that the last time period the subject is observed is not necessarily the time period the event of interest occur. See the next section for the explanation of the mechanisms censoring and truncation.

3.2 Censoring and Truncation

One of the most important differences between the outcome variables modeled via standard linear and logistic regression analyzes and the time variable in the current example is the fact that we may only observe the survival time partially. Subjects for which only partial survival time is observed are censored or truncated, depended on the conditions. For example, in a medical study, a patient could move out of town and therefore can no longer be followed. It is also possible that the patient dies in a car accident or the study could end before death from the disease of interest is observed. These are all examples of censoring in medical studies.

There exists two types of observations: complete and incomplete observations. Incomplete observations can not be discussed until a complete observation is carefully defined. This may seem trivial, but in applied settings confusion about censoring and truncation may be the result of an unclear definition of survival time. As mentioned in Section 3.1, the observation of survival time has two components which must be unambiguously defined: a beginning point where $t = 0$ and a reason or cause for the observation of time to end. In some applications it is difficult to find the best starting point. For example, there are multiple choices or the best choice is not available. For example in this research, the preferred starting time is the real default time. Unfortunately, we are limited by the administered default time. It is possible that there pass some time between the actual default time and the time the bank notices the default. A common example of this problem is found in medical studies. The best start point might be the infection date; another choice might be the date of diagnosis and enrollment in the study. The choice depends whether the first option is available. In a medical study an observation may end at the time when a subject literally 'dies' from the disease of interest, or it may end upon the occurrence of some other, non-fatal, well-defined, condition such as meeting clinical criteria for remission of cancer [Hosmer and Lemeshow (1999)]. Thus, it is also important to define the endpoint unambiguously. The survival time is the distance on the time scale between these points.

Now incomplete observations can be specified. There are two mechanisms that can lead to incomplete observation of time: censoring and truncation.

- *Censoring*: A censored observation is one whose value is incomplete due to random factors for each subject. Such as a patient that can no longer be followed, because he moved out of town.

- *Truncation*: A truncated observation is one which is incomplete due to a selection process inherent in the study design. For example an insurance company that only pays a claim up that is above the amount x and therefore all clients with a claim below the amount x will not be noticed by the insurance company.

The main difference between censoring and truncation is that censoring occurs due to random effects and truncations occurs due to selection criteria. This difference is important and causes that both mechanisms that can lead to incomplete observations should be treated differently.

Also within the mechanisms censoring and truncation there are different definitions. Within the mechanism censoring there exists left-censoring and right-censoring. The most common forms of truncation are left-truncation and right-truncation.

3.2.1 Censoring

A censored observation is one whose value is incomplete due to random factors for each subject. There exists different kinds of censoring. right-censoring and left-censoring are the most common forms.

- *Right-Censoring*: An observation is right-censored if the event of interest has not occurred when the observation ends (see figure 3.1).

- *Left-Censoring*: An observation is left-censored if the event of interest has already occurred when observation begins (see figure 3.2).

The most commonly encountered form of a censored is one in which observation begin at the defined time $t = 0$ and terminates before the outcome of interest is observed. Since the incomplete nature of the observation occurs in the right tail of the time axis, such observations are said to be right-censored. Figure 3.1 is a line plot for five subjects in a hypothetical follow-up study. Subject one, two and three enter the study after the start of the study. These subjects also end before the end of the study and therefore the endpoint is known. Obviously, for these subjects the real survival time is known. Subject four starts and ends after the end of the study and is not included in the study. Subject five starts before the end of the study and ends after the end of the study. The real survival time is not observed, but the minimum survival time is observed. The data is right-censored.

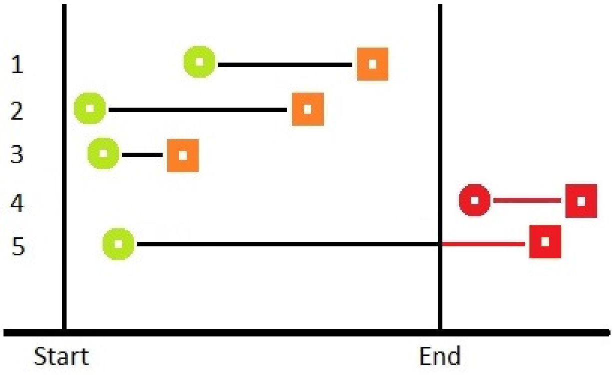

Figure 3.1: Line plot for five subjects in a hypothetical follow-up study, subject five is right-censored

Left-censoring occurs when the event of interest has already occurred before enrollment. This is very rarely encountered. An example is a medical study where the event of interest is experiencing chickenpox. Suppose all subjects enters the study at age of two. If the subject enters at this age, and has already experienced chickenpox, the subject's time is left censored. Figure 3.2 is the example of left-censoring explained in a line plot. This figure contains a line plot for five subjects in a hypothetical follow-up study. All five subjects enter the study at the same time. For example if all subjects are ten years old. The event of interest for the subjects 1,2,3 and 5 occurs before the end of the study and therefore the endpoint is known. Obviously, for these subjects the real survival time is known, because the start and end time is observed. The event of interest of subject four already occurred before the subject enrolled in the study. The available information is that the event already occurred, but it is not known when. Subject four is left-censored.

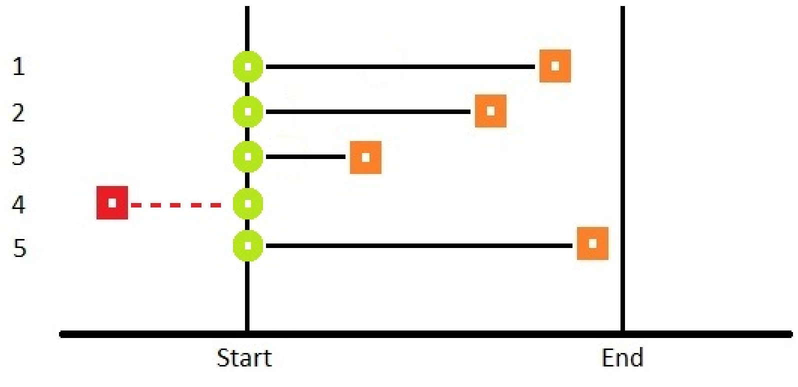

Figure 3.2: Line plot for five subjects in a hypothetical follow-up study, subject four is left-censored

3.2.2 Truncation

Another mechanism that can lead to incomplete observations is truncation. A truncated observation is one which is incomplete due to a selection process inherent in the study design. There are two types of truncation: left-truncation and right-truncation.

- *Right-Truncation*: Occurs when the entire study population has already experienced the event of interest before the start of the study.

- *Left-Truncation*: An observation is left-truncated when the subject has been at risk before entering the study.

An example of right-truncation is a medical study of risk factors for time to diagnosis of a type of cancer among subjects in a cancer registry with this diagnosis. Being in the cancer registry means that the diagnosis of cancer has already been made. This represents the selection process. This selection process must be taken into account in the analysis. However, right-truncation is rarely encountered.

The most common form of truncation is left-truncation. A left-truncated subject was already 'at risk' before it entered the study. An example of a study that encounters left-truncation is a default study, where the event of interest is a cure of a default. If there is a difference between the start of the default and the notation of the default, there is a period before the start of the study that the subject was 'at risk' of cure. If the default would be cured before it is noted, it would not have been noted. Only defaults that are at least in default for the period that is needed to administer the default are selected.

This is presented in figure 3.3 left-truncation is presented in a line plot. Subject 1,2 and 3 are all complete observations, with a known start and end point. Subject 4 started and ended before the end of the study and is therefore not incorporated in the study. Only subjects that survived the period before the start of the study are selected.

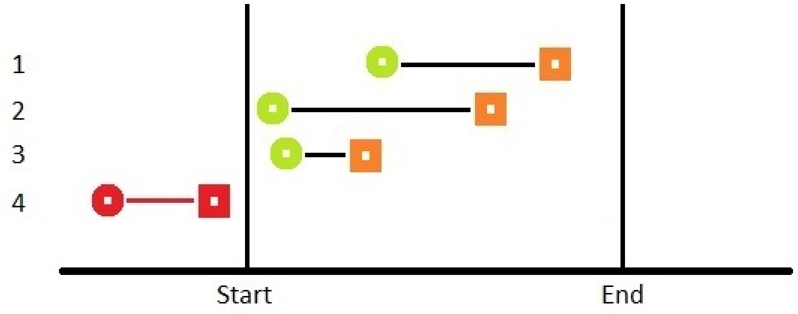

Figure 3.3: Line plot for five subjects in a hypothetical follow-up study. The data are left-truncated. Subject 4 is not selected.

3.3 Data Description

As mentioned in the introduction of this section, a dataset can consist of two types of data, complete and incomplete observations. Survival analysis is perfect to deal with right-censored data. In this section we suppose that there is a dataset with complete observations and right-censored observations. There is a difference in the meaning of an observation for both types. For an uncensored observation, the true survival time is observed. The event of interest is observed and the time between the start point and the endpoint is the true survival time. In this research this situation occurs, if a cure (or zero loss event) is observed. For a censored observation, the censoring time is observed. The true survival time is at least as long as the censoring time. The endpoint is not known (yet). Breslow and Crowley (1974) express this mathematically (see section 3.3.1), but it can also be described in a grouped framework(see section 3.3.2). The second description is often used to derive the parameters of the applied model.

3.3.1 Data Framework

Breslow and Crowley (1974) express the observation of an individual in a mathematical way. This mathematical description can be generalized for the whole set of observations. Suppose there is a set with N individuals. Let $X_1^*, ..., X_N^*$ denote the true survival times for the N individuals included in our data. The period of observation, or follow-up, for the ith individual will typically limited by an amount Y_i. Formally speaking, the X_i^* is censored on the right by the Y_i. This influences the observations. For every observation, one observes only the pair of variables (X_i, δ_i), where:

$$X_i = \min(X_i^*, Y_i) \tag{3.1}$$

$$\delta_i = I_{[X_i^* \leq Y_i]}, \tag{3.2}$$

where I is the indicator function. Thus, δ_i indicates whether X_i is censored ($\delta_i = 0$) or not ($\delta_i = 1$). One observes only the real survival time or the censoring time, not both. It is also known which of the two is observed. Note that both X^* and Y are random.

Now the underlying framework of the observations is presented the random censorship model can be applied. The random censorship model has three properties:

- Let $X_1^*, ..., X_N^*$ denote the true survival times for the N individuals included in the data. These have to be independent random variables having a common distribution:

  $$F^*(t) = \mathbb{P}[X_i^* \leq t] \text{ such that } F^*(0) = 0$$

- Recall that observation i will be limited by the amount Y_i. The censoring variables Y_i $(i = 1, ..., N)$ are also assumed to be independent random variables having a distribution:

  $$H(t) = \mathbb{P}[Y_i \leq t] \text{ such that } H(0) = 0$$

- The censoring variables Y_i $(i = 1, ..., N)$ have to be independently drawn of the true survival probabilities X_i^*. In particular, Y_i and X_i^* are independent.

Hence, if the properties above hold, the observed X's form a random sample from the distribution F given by

$$\begin{aligned} F(t) &= \mathbb{P}[X_i \leq t] \\ &= 1 - \mathbb{P}[X_i > t] \\ &= 1 - \mathbb{P}[\min(X_i^*, Y_i) > t] \\ &= 1 - \mathbb{P}[X_i^* > t \text{ and } Y_i > t] \\ &= 1 - \mathbb{P}[X_i^* > t]\mathbb{P}[Y_i > t] \\ &= 1 - (1 - F^*(t))(1 - H(t)) \end{aligned}$$

or described as:

$$1 - F = (1 - F^*)(1 - H)$$

The probability to have an observation at time period t is the probability that the observation survived until this time period times the probability that it is not censored until this time period.

3.3.2 Grouped Data Framework

Classical life table estimates are calculated from grouped data arising from a partition. The observations are grouped by survival time. Assume there is a sample of N independent observations denoted by

$$(X_i, \delta_i), i = 1, 2, ..., N \tag{3.3}$$

where X is the underlying observation time variable and δ is the censoring indicator variable. See subsection 3.3.1 for the precise definition of these two variables. Assume that among the N observations there are $M \leq N$ recorded times of failure (cure). Notice that for the censored cases there is no recorded time of failure. According to the introduced notation in the last subsection, Y_i is observed. We denote the ordered (true) survival times as

$$X^*_{(1)} < X^*_{(2)} < ... < X^*_{(M)} \tag{3.4}$$

$X^*_{(1)}$ is the smallest observed failure time and $X^*_{(M)}$ is the largest observed failure time. Notice that these quantities are random and depend on N. A small set of observations gives probably an other smallest and longest failure time than a large set, even if both have the same distribution. The largest observed failure time ($X^*_{(M)}$) is random. However, this does not mean that $X^*_{(M)}$ can increase to infinity. In the limit there is a largest observed real failure time is a certain value called ω, with $\omega < \infty$. In other words:

$$\operatorname*{plim}_{N \to \infty} X^*_{(M)} = \omega \tag{3.5}$$

Unfortunately, in a study the number of observations (N) is finite. This means that the largest observed true failure time $X_{(M)}$ might not be equal to the longest (possible) failure time ω. In some studies a longest failure time, that is not based on a stochastic, is needed. In that case an ω can be determined based on the data and the expertise of experts in this field. Obviously, this longest failure time can never be smaller than the longest observed failure time.

An observation is at least one time period observed. Notice that the description is based on time periods, this implies a discrete time scale. For example survival time and failure time can be measured per month or per week. Thus, all possible failure times lie in the range of $1, .., \omega$.

Now the possible failure times are determined, the data can be described by a three statistics per time period. n_t represents the number of individuals alive at the beginning of time t. The number of individuals alive at the beginning of the time period can also be seen as the number of individuals 'at risk' of death just before time period t. d_t represents the number known to have died at time period t and w_t is the number 'withdrawn alive' in time period t. Introduce the following three statistics for all $t(= 1, ..., \omega)$:

$$d_t = \sum_{i=1}^{N} I_{[X_i=t,\delta_i=1]} \tag{3.6}$$

$$w_t = \sum_{i=1}^{N} I_{[X_i=t,\delta_i=0]} \tag{3.7}$$

$$n_t = \sum_{i=1}^{N} I_{[X_i \geq t]} \tag{3.8}$$

3.4 Competing Risks

Above, standard survival analysis is described. An observation ends because the event of interest occurs or because a censoring event occurs. It is assumed that the event of interest and all censoring events are independent. However, in many contexts it is likely that the time to censoring is somehow correlated with the time to the event of interest. In general, there are different events (death, relapse, relocation, etc) which can be related. If one is interested in more events or if there is one event of interest correlated with another event, there are competing risks. An example of two types of failure that are correlated: following patients after a bone marrow transplantation to evaluate leukemia-free survival. There are two types of failures: leukemia relapse and non-relapse deaths. Because leukemia patients have a weakened immune system, it is likely that both events are correlated. An example where there are more events of interest originates from actuarial analysis, there the time to death is observed, but some may want to provide separate estimates of hazards for each cause of death. Thus the different causes of death are events of interest.

The data description presented in the previous section will slightly change if competing risks are involved. Suppose we have a study where we have a true survival time, the survival time until dying from the event of interest, denoted by X^*. We have also the survival time, which is defined as the time until dying from the competing risk Z. The competing risk Z is for example an other cause of death like a car accident or pneumonia. Besides we have a censoring time Y, which is for example the time to censoring due to the end of the study.

Simulare to "normal" survival time, we observe the pair of variables (X_i, δ_i), but X_i and δ_i are defined differently:

$$X_i = \min(X_i^*, Y_i, Z_i)$$

$$\delta_i = \begin{cases} 1 & \text{if } X^* \leq Y \text{ and } X^* \leq Z \\ 2 & \text{if } Y < X^* \text{ and } Y < Z \\ 3 & \text{if } Z < X^* \text{ and } Z \leq Y \end{cases}$$

In the previous section also the grouped data framework is presented. The pair of variables for all observation is summarized by a few statistics (n_t, d_t and w_t). This will differ in case of competing risks. The study presented above can be summarized by the four statistics. Again n_t represents the number of individuals alive at the beginning of time t. The number of individuals alive at the beginning of the time period can also be seen as the number of individuals 'at risk' of death just before time period t. d_t represents the number known to have died at time period t from the event of interest and w_t is the number 'withdrawn alive' in time period t. c_t is the number of individuals died from the competing risk in time period t. This gives the following formulas:

$$d_t = \sum_{i=1}^{N} I_{[X_i=t,\delta_i=1]} \tag{3.9}$$

$$w_t = \sum_{i=1}^{N} I_{[X_i=t,\delta_i=2]} \tag{3.10}$$

$$c_t = \sum_{i=1}^{N} I_{[X_i=t,\delta_i=3]} \tag{3.11}$$

$$n_t = \sum_{i=1}^{N} I_{[X_i \geq t]} \tag{3.12}$$

To find a way to threat these competing risks, first it is important to consider whether the risks are independent or dependent. The rest of this section considers the types of competing risks and the possible problems that can occur by applying the survival analysis framework described in the previous section. Also some approaches are presented.

3.4.1 Independent Risks

If there are competing risks, it is important to know whether these risks are independent or dependent. In this subsection the independent risks are considered. In the next section dependent risks are considered. A common example of independent competing risks is a follow-up study of patients diagnosed with cancer (or any other life threatening disease). One might be interested in the time a patient lives from the diagnoses to the death of cancer. In this example dying from cancer is the event of interest. But it is also possible that the patient gets a car accident. In this example dying from a car accident is a competing risk. If a patient died from a car accident, the survival time of cancer can not be observed anymore. However, one might not expect that the survival time of a patient diagnosed with cancer influences the probability to die from a car accident. Survival time of cancer and dying from a car accident are independent. We conclude that these competing risks are independent.

In the case of independent competing risk, the data framework presented above can be reduced to the data framework presented in the previous section. Threat a subject that died from the competing risk (not the risk of interest) as a right-censored observation. Because Z and Y both are independent of the event of interest, both can be incorporated in the same group, denoted by Y^* and $Y_i^* = \min(Z_i, Y_i)$. One observes:

$$X_i = \min(X_i^*, Y_i^*)$$
$$\delta_i = I_{(X_i^* \leq Y_i^*)}$$

Notice that for a censored observation it is assumed that, if censoring did not take place, the probability of dying from the event of interest would be equal to the probability of dying for uncensored observations with equal covariates. Referring to the example above, the probability of dying of cancer is assumed to be the same for a patient that died from a car accident, if the accident would not have occurred, as for another patient with the same characteristics. Notice also that we loose information about the relation between Y and Z. However, we are not interested in this relation. Therefore, loosing information gives no problems.

A requirement for this approach is that X^* and Y should be independent. Tsiatis (1975) showed that observing (X, δ) does not provide enough information to estimate the joint distribution of (X^*, Z). We cannot check directly whether all assumptions necessary to apply the random censorship model are valid. Ibrahim (2005) presents two arguments that have to be fulfilled to make it reasonable to assume that the risks are independent:

- when it is reasonable to assume random censoring if censoring occurs because of the study ends, or because the subject moves to a different state

- Beside that there should be no trend over time.

Ibrahim (2005) states that under these arguments, it is valid to use the approach presented above.

3.4.2 Dependent Risks

In the previous subsection two arguments are presented that have to be fulfilled to make it reasonable to assume that the risks are independent. One can think of many examples that do not fulfill both arguments. If we recall the example of the previous section about a follow-up study that is interested in the survival time of patients diagnosed with cancer. Now consider the competing risk dying from pneumonia. It is likely that a patient that died from pneumonia have such a poor immune system that the probability to die from cancer, if the patient would not have died from pneumonia, was higher than for other patients. In this case the risks are dependent. Therefore, the approach presented in the previous section is not valid.

There has been a lively debate in the literature about the best way to attack this problem. The approaches can basically be divided in two groups:

- Cause-specific hazard functions: this approach focuses on what the observed survival is due to a certain cause of failure, acknowledging that there are other types of failures operating at the same time.

- Latent variable models: this approach attempts to estimate what the survival associated with a certain failure time would have been, if other types of failures had been removed.

Both models have disadvantages. It is possible to estimate the cause-specific hazard functions, since we can observe whether each subject is still alive or not (and cause of death). But unfortunately, we can not estimate the marginal hazard function when the risks are dependent, since we can not observe when for a subject, a certain event would occur if a different event occurred first. Cause-specific hazard functions are also difficult to interpret, especially in the presence of highly dependent risks. In the latent variable approach, one needs to make a lot of untestable assumptions.

Ibrahim (2005) proposes another approach, which can be applied in specific cases. In this approach an upper and lower bound for the true marginal survival function, in the absence of competing risks is derived. Suppose we have two risks with failure time for the event of interest denoted by X^* and failure time for the competing risk denoted by Z. As noted previously, it is not possible to estimate $S_{X^*}(t) = \mathbb{P}(X^* \geq t)$ if the failure times are dependent. However, we may be able to say something about the range of $S_{X^*}(t)$ by finding upper and lower bounds that contain $S_{X^*}(t)$. Peterson (1976) obtained bounds based on the minimal and maximal dependence structure. The disadvantage of this approach is that the bounds can become very wide. Slud and Rubinstein (1983) and Klein and Moeshberger (1988) obtained tighter bounds on $S_{X^*}(t)$ by using additional information, but require the user to specify reasonable bounds on the dependence.

Ibrahim (2005) presents an example of these approaches. In this example data from promotion from Assistant Professor to Associate Professor are used to asses whether there were differences between males and females. Obviously, a promotion is the event of interest. If an Assistant Professor leaves, this is considered as a right-censored observation. This gives a problem. It is very unlikely that an Assistant Professor will leave his position prior to getting promoted. In this case censoring is negatively correlated with the event of interest. To illustrate the approach above, four options are proposed:

- Those who left would not have been promoted. Censoring and the event of interest are depended.

- Censoring and the event of interest are assumed to be independent: Those who leaved would have been promoted at the same rate as those who stayed.

- There is some dependence between censoring and the event of interest: For example it is assumed that 50% of those who departed would not have been promoted, and the other 50% would have been promoted at the same rate as those who stayed.

- Those who left would have been immediately promoted, if they not would not have left. This option is very unlikely, given the example above.

Peterson (1976) proposed to present the results of the first and the last option. This would give a very wide confidence interval. Some assumptions on the correlation will help to narrow the confidence bounds. In method 3, the 50% of the Assistant Professors who have left, but would have been promoted at the same rate as the others, will be treated as censored observations. For the other 50% time to promotion can be set equal to infinity, since they would not have been promoted. Slud and Rubinstein (1983) introduced a general approach. They introduce a variable $p(t)$, defined as follows:

$$p_t = \lim_{\epsilon \to 0} \frac{\mathbb{P}(X^* < t + \epsilon | Z \leq t, X^* > t)}{\mathbb{P}(X^* < t + \epsilon | Z > t, X^* > t)} \tag{3.13}$$

This fraction consists of two probabilities. The probability in the numerator is the probability that the observation dies immediately from the event of interest, given that the competing risk already occurred. The probability in the denominator is the probability that the subject dies immediately from the event of interest, given that the subject survived until the current time period. Every non-negative value of p is possible. Three special cases are defined

$p = 1$ X^* and Z are independent.

$p = 0$ the minimal value for p. If the competing risk occurred, a subject will never die from the event of interest.

$p \approx \infty$ the maximal value for p. If the competing risk occurred, a subject will die immediately form the event of interest.

Peterson (1976) proposed to use $p = 0$ and $p = \infty$ to derive the bounds. If we can make some reasonable assumptions on the dependence structure, as Slud and Rubinstein (1983) and Klein and Moeshberger (1988) propose, other values of p can be choosen. Slud and Rubinstein (1983) showed that each p leads to a well defined estimator for the survival function, called $\hat{S}_p$ generalizing Kaplan-Meier; bounds for $p(t)$ give consistently estimated bounds on S. The results of Slud and Rubinstein (1983), regarding the estimator for S_p are presented in section 4.4

Chapter 4

The Kaplan-Meier Estimator

In this chapter the used estimator for the survival function is described. This non-parametric estimator is called the Kaplan-Meier estimator, founded by Kaplan and Meier (1985). This estimator has nice and useful asymptotic properties. These properties are also described in the next subsection. First the basic concept of the estimator is described. Thereafter the asymptotic properties and the derivation of the confidence intervals of the survival function are presented.

The survival function, $S = 1-F^*$, captures the probability that a subject 'survives' beyond a specified time. The Kaplan-Meier estimator of the survival function [Kaplan and Meier (1958)], also called *the product limit* estimator, is the most used estimator in non-parametric survival analysis. This estimator incorporates information from all of the observations available, both uncensored and censored, by considering survival to any point in time as a series of steps defined by survival and censored times [Hosmer and Lemeshow (1999)]. We conclude that the Kaplan-Meier estimate for the survival function is a step function with jumps at the observed failure times. See figure 4.1 for an hypothetical example of a survival function.

4.1 Kaplan-Meier Estimator of the Survival Function

The Kaplan-Meier estimator makes use of statistics presented in section 3.3.2: deaths d_t, withdraws w_t and number of cases at risk n_t. These statistics are defined for every time period $t = 1, ..., \omega$, three statistics are introduced: Under the random censorship model presented in section 3.3, the conditional probability of dying in time period t given that the subject was alive just before can be estimated by:

$$\frac{d_t}{n_t} \tag{4.1}$$

The estimated conditional probability of surviving time period t, denoted by $\hat{\pi}_t$, is the complement:

$$\hat{\pi}_t = 1 - \frac{d_t}{n_t} \tag{4.2}$$

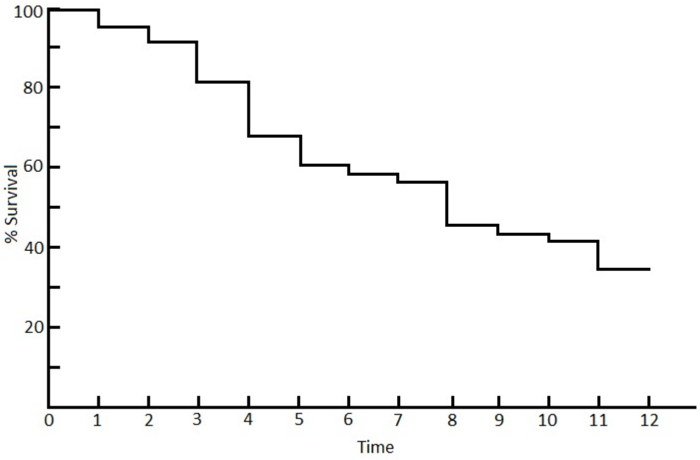

Figure 4.1: Survival Function of an Hypothetical dataset

The overall unconditional probability of surviving to t is obtained by multiplying the conditional probabilities for all relevant times up to t. This is the product limit estimator or the Kaplan-Meier estimator of the survival function:

$$\hat{S}(t) = \prod_{i \leq t} \hat{\pi}_i = \prod_{i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{4.3}$$

with the convention that

$$\hat{S}(0) = 1.$$

Conclude that $\hat{S}(t)$ is the Kaplan-Meier estimator of the survival function and is defined on the interval (0,∞). The interval $(-\infty, 0)$ is not relevant. On the interval $[0, \omega]$, $\hat{S}(t)$ is a non-increasing step function, with jumps on the occurred true failure times. On the interval (ω, ∞) the survival function is constant.

4.2 Asymptotic Properties

In the previous subsection the Kaplan-Meier estimator is derived. In this subsection the asymptotic properties of this estimator are presented. First, consistency is considered. Under the random censorship model, the Kaplan-Meier estimator of the survival function is consistent and the Central Limit Theorem can be applied.

$\hat{S}$ is the Kaplan-Meier estimator for the survival function S. The survival function is a step function on the interval $\{0, 1, .., \omega\}$. Peterson (1977) and Gill(1983) show that the Kaplan-Meier estimator $\hat{S}(t)$ is uniformly consistent:

$$\operatorname*{plim}_{N \to \infty} \sup_{t \leq \omega} |\hat{S}(t) - S(t)| = 0 \tag{4.4}$$

where plim means the probability limit. This means that the Kaplan-Meier estimator will converge in the limit to the real survival function. Recall that $X^*_{(M)}$ is the largest observed true failure time.

Notice that the survival function $S(t)$ for all $t = 1, ..., \omega$ can be written as the vector $\mathbf{S}$. Breslow and Crowley (1974) derived the limiting distribution for the Kaplan-Meier estimator of the survival function under the random censorship model. Applying the δ-method and the Central Limit Theorem gives:

$$\sqrt{N}(\hat{\mathbf{S}} - \mathbf{S}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma) \tag{4.5}$$

The limiting distribution is multivariate normal. See for the derivation of the limiting distribution above appendix A. Miller (1981) presents the asymptotic covariance matrix Σ. The covariance matrix term (s, t) of Σ:

$$S(s)S(t) \int_0^{t \wedge s} \frac{1}{(1 - F(u))^2} \, \mathrm{d}F^*_u(u) \tag{4.6}$$

Notice that $t \wedge s$ is the minimum of s and t. Recall F* is the distribution of the real survival times X^*, H is the distribution of the censoring times Y and F is the distribution of the observed survival times X. F^*_u is the distribution of the real survival times X^* of the uncensored observations: $F^*_u(t) = \mathbb{P}[X \leq t \text{ and } \delta = 1]$. Miller (1981) and Andersen at all (1986) propose to use the Greenwood's formula [Greenwood (1926)] to derive an estimator for the covariance matrix of $\hat{\mathbf{S}}$. Conditional on the data d_t and n_t for all $t \in \{1, .., \omega\}$, the covariance matrix term (s, t) is:

$$S(s)S(t) \sum_{i=1}^{s \wedge t} \frac{d_i}{n_i(n_i - d_i)} \tag{4.7}$$

$S(s)$ and $S(t)$ can be estimated by $\hat{S}(s)$ and $\hat{S}(t)$. See appendix A.3 for the application of Greenwood's Formula to derive an estimator for the covariance matrix. Also the performance of this estimator is discussed.

4.3 Pointwise Confidence Intervals for the Survival Function

After obtaining the estimated survival function and the limiting distribution, pointwise confidence intervals can be obtained. The Kaplan-Meier estimator of the survival function is asymptotically normally distributed for each t. Thus, pointwise confidence intervals can be derived by adding and subtracting the product of the estimated standard error times a quantile of the standard normal distribution. This gives:

$$\hat{S}(t) \pm z_{1-\frac{\alpha}{2}} \widehat{SE}\{\hat{S}(t)\} \tag{4.8}$$

where $z_{1-\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ quantile of the standard normal distribution and $\widehat{\mathrm{SE}}$ represents the estimated standard error of the argument in parentheses.

This theory can be applied directly to the Kaplan-Meier estimator. this approach, however, can lead to confidence interval endpoints that are less than zero or greater than one. To address these problems, Kalbfleisch and Prentence (1980) suggest that confidence interval should be based on the function:

$$\ln[-\ln(\hat{S}(t))] \tag{4.9}$$

This function is called the log-log survival function. This function has a possible range from minus to plus infinity. The endpoints of a $100(1-\alpha)$ percent confidence intervals based on the log-log survival function are

$$\exp[-\exp(\hat{c}_u)] \text{ and } \exp[-\exp(\hat{c}_l)]. \tag{4.10}$$

where $\hat{c}_u$ is

$$\ln\left[-\ln\left(\hat{S}(t)\right)\right] + z_{1-\frac{\alpha}{2}}\widehat{\text{SE}}\left\{\ln\left[-\ln\left(\hat{S}(t)\right)\right]\right\}, \tag{4.11}$$

and where $\hat{c}_l$ is

$$\ln\left[-\ln\left(\hat{S}(t)\right)\right] - z_{1-\frac{\alpha}{2}}\widehat{\text{SE}}\left\{\ln\left[-\ln\left(\hat{S}(t)\right)\right]\right\}, \tag{4.12}$$

where $z_{1-\frac{\alpha}{2}}$ is again the upper $\frac{\alpha}{2}$ quantile of the standard normal distribution and $\widehat{\text{SE}}$ represents the estimated standard error of the argument in parentheses.

Notice that the estimated lower endpoint in the log-log survival function confidence interval ($\hat{c}_l$) yields the upper endpoint of the estimated survival function confidence interval and vice versa. The derivations of the pointwise confidence intervals based on the log-log survival function are presented in appendix A.4

Notice that this approach narrows the confidence intervals compared to the standard approach. For this approach a reasonably large number of observations is needed. The confidence interval is valid only for values of time over which the Kaplan-Meier estimator is defined, which is basically the observed range of survival times. Borgan and Leistøl (1990) studied this confidence interval and found that it performed well for sample sizes as small as 25 observations with up to 50 percent right-censored observations. This can be used as rule of thumb.

4.4 The Survival Function in Case of Dependent Censoring

In section 3.4 we introduced the approach used by Slud and Rubinstein (1983), regarding two dependent competing risks (X^*, Z). In their paper, they first introduce the variable $p(t)$ for the dependence structure. Recall:

$$p_t = \lim \epsilon \to 0 \frac{\mathbb{P}(X^* < t + \epsilon | Z \leq t, X^* > t)}{\mathbb{P}(X^* < t + \epsilon | Z > t, X^* > t)} \tag{4.13}$$

Notice that the event of interest and the competing risk are positive correlated if $p > 1$ and the event of interest and the competing risk are negative correlated if $0 \leq p < 1$. If $p = 1$, the event of interest and the competing risk are independent.

Slud and Rubinstein (1983) state that each p leads to well-defined estimator $\hat{S}_p$ generalizing Kaplan-Meier. Slud and Rubinstein (1983) present the following estimator for S_p:

$$\hat{S}_p(t) = \frac{1}{N}\left(n_t + \sum_{k=0}^{t-1} c_k \prod_{i=k+1}^{t-1} \frac{n_i - d_i}{n_i + d_i(p_i - 1)}\right) \tag{4.14}$$

This formula can be divided in two parts. The left part are all cases that survived until time period t. The right part are all cases that are censored before time period t multiplied by the probability that they would survive the period between censoring and t. By using the formula $c_t = n_t - n_{t+i} + d_t$, Slud and Rubinstein derived the following intuitive formula for $\hat{S}_p(t)$

$$\hat{S}_p(t) = \prod_{i=1}^{t} \frac{n_i - d_i}{n_i + d_i(p_i - 1)} + \frac{1}{N}\sum_{k=1}^{t} d_k(p_k - 1) \prod_{i=k}^{t} \frac{n_i - d_i}{n_i + d_i(p_i - 1)} \tag{4.15}$$

In particular, if $p = 1$, $\hat{S}_p$ is exactly the Kaplan-Meier estimator. Slud and Rubinstein (1983) conclude that $\hat{S}_p(t)$ is a generalized maximum likelihood nonparametric estimator for S(t), under the assumption that $p(t)$ is chosen equal to equation 4.13. Slud and Rubinstein (1983) also conclude that we can derive the same asymptotic properties as for $\hat{S}$ presented in section 4.2.

In the estimator for S_p presented above, only possible events of failure (Z, X^*) are incorporated. However, it is likely that their are more risks involved. For example a study as described in section 3.4,where we have an event of failure X^*, a (dependent) competing risk Z and an (independent) censoring Y, due to the end of the study. If there are more than two risks involved equation (4.14) should be adjusted. The example above, with X^*, Y and Z, gives the following formula:

$$\hat{S}_p(t) = \frac{1}{N}\left(n(t) + \sum_{k=0}^{t-1} c_k \prod_{i=k+1}^{t-1} \frac{n_i - d_i}{n_i + d_i(p_i - 1)} + \sum_{k=0}^{t-1} w_k \prod_{i=k+1}^{t-1} 1 - \frac{d_i}{n_i}\right) \tag{4.16}$$

Notice that we assume Y is independent of X^*.

Chapter 5

Application of Survival Analysis to Default Data

In order to apply the introduced the methodology we need to define a start point, end point and scale and check whether these variables are available in the data. Then we have to check for competing risks and the dependence structure of these risks. We have also to check whether the assumptions presented in section 3.3.1. If all conditions are met, we can apply the introduced methodology and derive the estimator for the conditional cure rate and conditional provision release rate.

As mentioned in section 3, for the application of the survival framework a clear start point and event of interest have to be specified. Also the time scale has to be chosen. In our study we have to make some remarks on the available data for the start point, end point and scale.

- *Start point*: The preferred start point is the default date. However, the presence of the default date is depending on the administration of the default. If a client is technically in default, it may take some time for the bank to notice this default. However, it is assumed that the Rabobank knows his clients very well and that a possible default is already observed before the actual default take place. In this way a number of defaults will be prevented and the occurred defaults can be administrated immediately. If this assumption is valid the starting time is unambiguously defined. Notice that this means that there is no left-truncation. If this assumption is not valid, some cases that cure after a very short time, will not be noticed as default. Therefore, the cure rate will be underestimated.

- *End point*: Again there is dependence of the administration of the bank regarding defaults. It is assumed that the Rabobank has a lot of contact with its clients. Therefore, a cure or a liquidation can be noticed immediately. If this assumption is not valid, a defaulted case that is already resolved but not administrated as resolved case will be denoted as a censored observation. This can happen for a cure and for a liquidation, therefore it is not predetermined whether this will lead to an under or overestimation.

- *Scale*: The used scale is months. Every month the progress in the default process is administrated very precisely. Therefore this is the natural scale.

We assume that the data needed for the definition of the start point, end point and scale are available. In the first subsection we will define these points for the cure rate. We will also check for competing risks and present the applied data framework. We will do the same for the provision release rate in section 5.2. The assumptions presented in section 3.3.1, will be checked after the description of the data in section **??**. In subsections 5.3, 5.4, 5.5 and 5.6 the expression, estimates and asymptotic properties are presented, using the Kaplan-Meier estimator. The derivation of the estimate is the same for the cure rate and the provision release rate. Therefore, the estimate and asymptotic properties are only presented once.

5.1 Application Survival Analysis to Cure Rate

As mentioned in the introduction of this chapter we will present the applied data framework for the cure rate in this section. First we have to check for competing risks. Thereafter, we have to specify the start en end points of the variables in the data framework.

For the cure rate the problem of competing risks occurs. In section 3.4, we make a distinction between independent and dependent competing risks. We assume that the event of interest and censoring, because of the end of the study, are independent. However, we assume that liquidation and cure are dependent competing risks. The approach these competing risks is presented further in this section.

First, we will specify the variables in our data framework and specify what we will perceive. There are three variables in our data framework: the real survival time X^*, Z is the survival time from the dependent competing risk, liquidation and Y the censoring time. All three variables define a time period between a start point and an end point. These start points and end points are as follows:

Variable	start point	end point
X^*	start default	cure
Z	start default	liquidation
Y	start default	end of the study

Now the start points and end points are defined we can present the pair of variables (X_i, δ_i)

observed:

$$X_i = \min(X_i^*, Y_i, Z_i)$$
$$\delta_i = \begin{cases} 1 & \text{if } X^* \leq Y \text{ and } X^* \leq Z \\ 2 & \text{if } Y < X^* \text{ and } Y < Z \\ 3 & \text{if } Z < X^* \text{ and } Z \leq Y \end{cases}$$

To apply the random censorship model we have to consider the dependence structure of the competing risks. We assumed that the censoring time Y and cure X^* are independent. We assumed that liquidation Z and cure X^* are dependent competing risks. However, this is a special case of dependent competing risks. Before, we stated that for a case that is liquidated never a cure would have been occurred if the liquidation process was not started. This means that if we observe a liquidation ($\delta = 3$), the event of interest will never occur ($X^* = \infty$). This relation can be used to transform the data into a standard survival analysis problem without competing risks. For the standard survival problem without competing risks see equations 3.1 and 3.2. In this standard framework we have the survival time X and a binary δ, which is one if the event of interest occurs and zero otherwise.

For the cured cases, there are no changes regarding this transformation. The observed survival time X is the real survival time X^* and $\delta = 1$. For the censored cases, there is only a small notation change regarding this transformation. The observed survival time X is the censoring time Y and $\delta = 0$. For the liquidated cases we have a change regarding this transformation. In the framework described above, a liquidated case is denoted as $X = \min(X^*, Y, Z) = Z$. In the standard survival analysis framework without competing risks Z is not used. However, above is argued why X^* is known, namely $X = \infty$, if the variable Z is observed. Therefore, we can use this known X^* in the standard framework. In section 3.3 is stated that the largest possible failure time is ω. Therefore, it is not relevant to follow a subject longer than ω months. Subsequently, Y_i is set equal to ω, which is equal to the observed survival time. This means that a liquidated case has the following pair of observed variables (X_i, δ_i):

$$X_i = \omega$$
$$\delta_i = 0$$

To apply the random censorship model and derive conclusions on the base of survival analysis, the conditions presented in section 3.3 have to be satisfied. This is checked in section 6.

Notice that the approach presented above is equal to approach of Slud and Rubinstein (1983) explained in sections 3.4 and 4.4, for $p = 0$.

5.2 Application Survival Analysis to Provision Release Rate

As mentioned in the introduction of this chapter we will present the applied data framework for the provision release rate in this section. First we have to check for competing risks. Thereafter, we have to specify the start en end points of the variables in the data framework.

For the cure rate the problem of competing risks occurs. In section 3.4, we make a distinction between independent and dependent competing risks. We assume that the event of interest and censoring, because of the end of the study, are independent. However, we assume that cure and liquidation without loss and liquidation with loss are dependent competing risks. The approach these competing risks is presented further in this section.

First, we will specify the variables in our data framework and specify what we will perceive. There are three variables in our data framework: the real survival time X^*, Z is the survival time from the dependent competing risk, liquidation with loss and Y the censoring time. All three variables define a time period between a start point and an end point. These start points and end points are as follows:

Variable	start point	end point
X^*	start default	cure or liquidation without loss
Z	start default	liquidation with loss
Y	start default	end of the study

Now the start points and end points are defined. The pair of observed variables (X_i, δ_i) are the same as presented in the previous section for the cure rate.

To apply the random censorship model we have to consider the dependence structure of the competing risks. In contrast to the cure rate, we will not assume that the event of interest (cure or liquidation without loss) cannot occur anymore if the competing risk already occurred (liquidation with loss). We still assume that for a liquidated case never a cure would have been occurred, if the liquidation process was not started. But under some circumstances it is possible that a liquidation without loss would have been occurred, if the liquidation process with loss was not finished. This could be the case if the value of the underlying collateral of a loan is fluctuating. It is possible that in bad financial times capital goods are worth much less than in a recovering economy. Then waiting can be rewarded. Notice that this case is very rare, because usually arrears are increasing. However, if we include this possibility, we cannot assume $p = 0$, from the Slud and Rubinstein (1983) approach.

The use of the default dataset for the application of the approach of Slud and Rubinstein (1983) on the provision release rate is straightforward and mainly explained above. For the application of equation (4.16), we need the quantities d_t, w_t, c_t and n_t. d_t is defined as the sum of all cured cases and liquidated cases without loss, resolved in time period t. w_t is defined as the sum of all unresolved cases censored due to the end of the study at month t. c_t is defined as the sum of the liquidated cases with loss, that become resolved at time period t. Finally, n_t

is defined as all cases 'at risk' at time t.

The only unknown factor is p. Equation (3.13) can be used to determine this factor. However, we observe only $\min(X^*, Y, Z)$. Therefor we can not derive p. We can do some assumptions about p. First, we assume that the effect of competing risks is equal over time. This means that $p(t) = p$ for all t. This simplify our problem, we only need one value for p. Above we argued that the probability that a liquidation without loss would happen if the loss liquidation would not have finished is very small. The correlation is definitely negative and therefore it sure that $p < 1$. The approach of Slud and Rubinstein (1983) is used to derive an upper and lower bound for the survival function, by choosing an upper and lower bound for p. It is natural to choose $p = 0$ as lower bound. The choice for an upper bound for p is an expert-based decision. See section 7.2.

5.3 Expression for the Unconditional Cure Rate and Unconditional Provision Release Rate

The unconditional cure rate (UCR) and the unconditional provision release ($UPRR$) rate follow easily from the survival functions derived for both. The expression is derived for the UCR. the expression for the $UPRR$ is exactly the same. Notice that we only need other quantities to derive the survival function for provision release. The survival function at time period t ($S(t)$) is defined as the probability that a subject survives for a time period t. The (UCR) is defined as the probability that a subject will fail (cure). Notice that it is not important when a subject fails, therefore UCR does not depend on time. It is known that a subject can only fail in the time periods $\{1, ..., \omega\}$. Thus, for the cure rate we are looking for:

$$\mathbb{P}[X^* \leq \omega] \tag{5.1}$$

This is the probability that a cure will occur. Obviously, this is the complement of the probability that no cure will occur. This is the same as the probability that a subject survives until time period ω. This gives:

$$UCR = 1 - S(\omega) \tag{5.2}$$

5.4 Estimation of the Unconditional Cure Rate and Unconditional Provision Release Rate

Notice that the estimation of the UCR and $UPRR$ is exactly the same. Only the quantities used to derive the survival function are different. In section 4 the Kaplan-Meier estimator for the survival function is derived. Also the limiting distribution is presented. In the last subsection, it is shown that the cure rate (UCR) only depends on the survival function in the last time period $S(\omega)$. Thus, the Kaplan-Meier estimator can be used for estimating the cure rate $\widehat{UCR}$.

$$\widehat{UCR} = 1 - \hat{S}(\omega) \tag{5.3}$$

The Kaplan-Meier estimator is a consistent estimator for the survival function. So the estimator above is also a consistent estimator for the cure rate.

$$\begin{aligned}
\operatorname*{plim}_{N\to\infty} |\widehat{UCR} - UCR| &= \operatorname*{plim}_{N\to\infty} |(1 - \hat{S}(\omega) - (1 - S(\omega)))| && (5.4)\\
&= \operatorname*{plim}_{N\to\infty} | - (\hat{S}(\omega) - S(\omega))| && (5.5)\\
&= \operatorname*{plim}_{N\to\infty} |(\hat{S}(\omega) - S(\omega))| && (5.6)\\
&= 0 && (5.7)
\end{aligned}$$

By applying the theory of Section 4 the limiting distribution can be derived from the cure rate. Under the random censorship model the CLT can be applied:

$$\sqrt{N}(\widehat{UCR} - UCR) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \tag{5.8}$$

where

$$\sigma^2 = S(\omega)^2 \int_0^\omega \frac{1}{(1 - F(u))^2}\, \mathrm{d}F_u^*(u) \tag{5.9}$$

This variance can be estimated by applying the Greenwood's formula, see Appendix A.3. The estimated variance of $\widehat{UCR}$ is

$$\hat{S}(\omega)^2 \sum_{t=1}^{\omega} \frac{d_t}{n_t(n_t - d_t)}. \tag{5.10}$$

Also the confidence interval can be obtained for the cure rate. Again the confidence interval is based on the log-log survival function. The unconditional cure rate is a linear transformation of the survival function at time ω: $UCR = 1 - S(\omega)$. The endpoints of a $100(1-\alpha)$ percent confidence interval for the cure rate can be derived directly from the results obtained in section 4.3. The $100(1-\alpha)$ percent confidence interval of the survival function at time ω, using the log - log survival function, is denoted $(\hat{c}_l, \hat{c}_u)$. Then $100(1-\alpha)$ percent confidence interval of the conditional cure rate is:

$$(1 - \hat{c}_u, 1 - \hat{c}_l) \tag{5.11}$$

5.5 Expression for the Conditional Cure Rate and Conditional Provision Release Rate

Notice that the estimation of the conditional cure rate and conditional provision release rate is exactly the same. Only the quantities used to derive the survival function are different. The conditional cure rate follows also from the survival function. However, there are more steps needed than for the unconditional cure rate. The conditional cure rate CCR is defined as the probability that a subject will cure (fail) ($X^* \leq \omega$) given that it already survived for a time period t ($X^* > t$). Notice that the conditional cure rate depends on time in contrast to the conditional cure rate. To derive the CCR, first, notice that X is observed and that $X = \min(X^*, Y)$. This implies that if the observed X is greater or equal to t, then X^* is also greater than or equal to t. Therefore, it is possible to condition on X^*, while only X is observed. The conditional cure rate is expressed by:

$$\mathbb{P}[X^* \leq \omega | X^* > t] \tag{5.12}$$

This probability can be elaborated

$$\begin{aligned}
CCR(t) &= \mathbb{P}[X^* \leq \omega | X^* > t] && (5.13)\\
&= \frac{\mathbb{P}[X^* \leq \omega \text{ and } X^* > t]}{\mathbb{P}[X^* > t]} && (5.14)\\
&= \frac{\mathbb{P}[t < X^* \leq \omega]}{\mathbb{P}[X^* > t]} && (5.15)\\
&= \frac{\mathbb{P}[t < X^*] - \mathbb{P}[\omega < X^*]}{\mathbb{P}[X^* > t]} && (5.16)\\
&= 1 - \frac{\mathbb{P}[\omega < X^*]}{\mathbb{P}[X^* > t]} && (5.17)\\
&= 1 - \frac{\mathbb{P}[X^* > \omega]}{\mathbb{P}[X^* > t]} && (5.18)\\
&= 1 - \frac{S(\omega)}{S(t)} && (5.19)
\end{aligned}$$

In the last step, the two probabilities in this expression are substituted by the survival function on a certain month. Notice that the conditional cure rate depends on t. Therefore, the conditional cure rate can be derived for all $t = 1, ..., \omega$.

5.6 Estimation of the Conditional Cure Rate and Conditional Provision Release Rate

Notice that the estimation of the conditional cure rate and conditional provision release rate is exactly the same. Only the quantities used to derive the survival function are different. In the previous subsection an expression for the conditional cure rate is presented. This expression depends on two points of the survival function. These points in the survival function can be estimated by the Kaplan-Meier estimator. Since the Kaplan-Meier estimator is consistent for all points in the survival function, both the numerator and denominator can be estimated consistently. Thus, $S(\omega)$ is estimated by $\hat{S}(\omega)$ and $S(t)$ is estimated by $\hat{S}(t)$. Both consistent estimators can be substituted in the expression 5.13. To show that $1 - \frac{\hat{S}(\omega)}{\hat{S}(t)}$ is a consistent estimator for the CCR at any time period t, apply Slutsky's Theorem, see Appendix B for the derivations. This yields

$$\widehat{CCR}(t) = 1 - \frac{\hat{S}(\omega)}{\hat{S}(t)} \xrightarrow{p} 1 - \frac{S(\omega)}{S(t)} = CCR(t) \tag{5.20}$$

Anderson et al. (1993) showed that under the random censorship model:

$$\sqrt{N}\left(\widehat{CCR}(t) - CCR(t)\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2(t)) \tag{5.21}$$

with

$$\sigma^2(t) = \left(\frac{S(\omega)}{S(t)}\right)^2 \int_t^{\omega} \frac{1}{(1-F(u))^2}\,\mathrm{d}F_u^*(u) \tag{5.22}$$

Recall that F* is the distribution of the real survival times X^*, H is the distribution of the censoring times Y and F is the distribution of the observed survival times X. F_u^* is the distribution of the real survival times X^* of the uncensored observations: $F_u^*(t) = \mathbb{P}[X \leq t \text{ and } \delta = 1]$.

Anderson et al. (1993) propose to derive an estimator for the covariance matrix in the same way as for the Kaplan-Meier estimator of the survival function $\hat{S}$ by using Greenwood's formula:

$$\widehat{\mathrm{Var}}(\widehat{CCR}(t)) = \left(\frac{\hat{S}(\omega)}{\hat{S}(t)}\right)^2 \sum_{i=t}^{\omega} \frac{d_i}{n_i(n_i - d_i)} \tag{5.23}$$

See appendix B.2 for the derivations.

Also the pointwise confidence intervals can be obtained for the conditional cure rate. Since the conditional cure rate depends on t, a confidence interval is derived for every $t = \{0, ..., \omega\}$. Again the pointwise confidence intervals are based on the log-log survival function. The endpoints of a $100(1-\alpha)$ percent confidence interval for the conditional cure rate are

$$\exp[-\exp(\hat{c}_u)] \text{ and } \exp[-\exp(\hat{c}_l)]. \tag{5.24}$$

where $\hat{c}$ is

$$\ln\left[-\ln(\widehat{CCR}(t))\right] \pm z_{1-\frac{\alpha}{2}} \widehat{\mathrm{SE}}\left\{\ln\left[-\ln(C\hat{C}R(t))\right]\right\} \tag{5.25}$$

where $z_{1-\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ quantile of the standard normal distribution and $\widehat{SE}$ represents the estimated standard error of the argument in parentheses. See Appendix B for the derivations.

Chapter 6

Data Description

Chapter 7

Results Regarding the Dataset

In this chapter the derived estimators and limiting distributions of chapter 5 will be applied to the dataset described in chapter 6. First the results for the (conditional) cure rate will be derived. The estimated survival function $\hat{S}(t)$, unconditional cure rate $\widehat{UCR}$ and conditional cure rate $\widehat{CCR}(t)$ will be presented. With the estimated $\widehat{CCR}(t)$, an estimate for the cure rate of the current default portfolio can be derived. In the second section the results for the (conditional) provision release rate will be derived. The estimated survival function $\hat{S}(t)$, unconditional zero loss rate ($\widehat{ZLR}$) and conditional zero loss rate $\widehat{CZLR}(t)$ will be presented. Different levels of p will be chosen and different estimates of $\widehat{CZLR}(t)$ will be presented. With the $\widehat{CZLR}(t)$ the conditional provision release rate can be derived and the provision release rate for the current default portfolio can be estimated.

7.1 Conditional Cure Rate

7.2 Conditional Provision Release Rate

Chapter 8

Check of Assumptions

Chapter 9

Conclusions and Further Discussion

In this thesis a methodology is proposed to include the expected loss of the current default portfolio into LGD measurement method used by Rabobank. This is an important adjustment to the current LGD measurement, where only resolved cases, are used. Due to the high number of unresolved cases the influence is not negligible and the unresolved cases possibly have a different character than resolved cases. The current methodology is not precise enough. The two main points that have to be included are:

- The probability that a default will cure (or become liquidated without a loss) is lower for cases that are longer in the default portfolio than for cases that are shorter in the default portfolio. Therefore, the probability that the bank makes no loss on a certain case decreases if it is longer in the default portfolio.

- It is possible that a case becomes liquidated without a loss. This must be taken into account. Therefore, the cure rate does not suffice as an approximation of the provision release rate.

The new methodology has the following improvements:

- By using the survival analysis framework, a conditional cure rate and conditional provision release rate can be derived, in stead of a point-in-time estimate. The advantage of this adjustment is that we can give a cure rate or provision release rate on the bases of the time the case is already in default.

- It groups the defaults by loss and no loss for the bank instead of by cure and liquidation, to obtain the provision release rate.

- Competing risks are taken into account.

Besides these improvements, the new methodology has also some weaknesses:

- It is assumed that the provision taken for the cases that will be liquidated with loss is equal to the loss for these cases. Of course there is uncertainty involved. This is not taken into account.

- For using the survival analysis framework the assumptions of the random censorship model have to be checked. On the used dataset it is difficult to check these assumptions. We cannot reject the hypothesis that the censoring time and the real survival times are dependent on 99% significance level. However, we can reject that the distribution of the cured cases and liquidated cases have the same distribution.

- We cannot derive the correlation of the dependent competing risks. We made some broad assumptions on the dependence structure of competing risks and derive the provision release rate for different values.

In general, it seems reasonable to assume that these assumptions are valid. Therefore, the methodology seems reasonable. However, for some of these weaknesses extra research is recommended.

The conditional cure rate and the provision release rate can be applied to derive the cure rate and provision release rate of the current default portfolio. This gives the following results:

Unfortunately, it was not possible to analyze all points of the proposed model. The following points further research is recommended:

- The loss provision rate is assumed to be fixed to simplify the model. This can be an improvements for the future.

- There are broad bounds assumed for the dependence structure of the competing risks. This dependence structure can be researched in the future. Further research will probably lead to a better choice of p.

- It is interesting to see what happens if also non-crisis data are incorporated. This can be investigated in the future.

Although several points can be used for further research, we would recommend to implement the methodology derived in this master thesis. The adjustments made compared to the current methodology are a big improvement. Also the duration of a default process is a good predictor for the loss for a certain case.

Bibliography

[1] Andersen P.K., Borgan Ø., Gill R.D. and Keiding N.,(1993), Statistical Models Based on Counting Processes. New York: Springer.

[2] Basel Committee on Banking Supervision, (2004), International Convergence of Capital Measurement and Capital Standards. http://www.bis.org/publ/bcbs107.pdf, Bank for International Settlements.

[3] Borgan Ø. and Leistøl K., (1990), A note on confidence bands for the survival curve based on transformations. *Scandanavian Journal of Statistics*, 17: 35-41.

[4] Breslow N. and Crowley J.,(1974), A Large Sample Study of the Life Table and Product Limit Estimates under Random Censorship. *The Annals of Statistics*, 2: 437-453.

[5] Brown Jr, B.W., Hollander, M. and Korwar, R.M., (1974), Nonparametric tests of independence for censored data with application to heart transplant studies. *Reliability and Biometry*, SIAM, Philadelphia.

[6] Cox, D.R. and Oaks, D., (1984), Analysis of Survival Data. Londen, U.K.: Chapman Hall.

[7] Daniels H.E., (1944), The relation between measures of correlation in the universe of sample permutations. *Biometrika*, 2: 129-135.

[8] Gill, R.D., (1983), Large sample behavior of the product limit estimator on the whole line. *Annals Statistics*, 9: 853-860.

[9] Greenwood M. (1926), The natural duration of cancer. *Reports on Public Health and Medical Subjects*, 33: 1-26.

[10] Hosmer D.W. and Lemeshow S.,(1999), Applied Survival Analysis: Regression Modeling of Time to Event Data. New York: John Wiley & Sons.

[11] Ibrahim J.G., (2005), Applied Survival Analysis, 21st Annual Summer Workshop of the Northeastern Illinois Chapter of the American Statistical Association.

[12] Kalbfleisch, J.D. and Prentice, R.L. (1980), The Statistical Analysis of Failure Time Data. New York: John Wiley & Sons.

[13] Kaplan E.L. and Meier P.,(1958), Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53: 457-481.

[14] Kendall, M.G., (1970), Rank correlation methods, London: Griffin.

[15] Miller G.R., (1981), Survival Analysis. New York: John Wiley & Sons.

[16] Moeschberger M.L. and Klein J.P., (1988), Bounds on net survival probabilities for dependent competing risks *Biometrics*, 44, 529-538.

[17] Peterson A.V., (1976), Bounds for a joint distribution function with fixed sub-distribution functions: Application to competing risks *Proceedings of the National Academy of Sciences*, 73, 11-13.

[18] Peterson A.V., (1977), Expressing the Kaplan-Meier Estimator as a Function of Empirical Subsurvival Functions. *Journal of the American Statistical Association*, 72: 854-858.

[19] Rao C.R., (1965), Linear Statistical Inference and its Applications. New York: Wiley.

[20] Schuermann T., (2004), What Do We Know about Loss Given Default? *Credit Risk Models and Management*,Londen, Risk Books, 2004.

[21] Shorack G.R. and Wellner J.A., (1986), Empirical Processes with Applications to Statistics. New York: Wiley.

[22] Slud E.V. and Rubinstein L.V., (1983), Dependent Competing Risks and Summary Survival Curves *Biometrika*, 70, 643-649.

[23] Thomas D.R. and Grunkemeier G.L., (1975), Confidence interval estimation of survival probabilities for censored data *Journal of the American Statistical Association*, 70, 865-871.

[24] Tsiatis A., (1975), A non identifiability aspect of the problem of competing risks *Proceedings of the National Academy of Sciences*, 72, 20-22.

[25] Wang, J.G., (1987), A Note on the Uniform Consistency of the Kaplan-Meier Estimator. *Annals Statistics*, 15:1313-1316.

[26] Witzany J., Rychnovsky M. and Charamza P., (2012), Survival Analysis in LGD Modeling. *European Financial and Accounting Journal*, 7:6-27.

Appendix A

Derivations of Asymptotic Properties of the Kaplan-Meier Estimator

In this Appendix all derivation of Asymptotic Properties of the Kaplan-Meier estimator, that are presented in Section 4, have been worked out. First, consistency is discussed. Then the derivation of the limiting distribution is explained. For the limiting distribution a covariance matrix is presented. Using Greenwood's formula, an estimator for this covariance matrix is derived. The confidence intervals for the Kaplan-Meier estimator of the survival function are bases on the log-log survival function. The log-log survival function for the confidence is especially relevant for the Kaplan-Meier estimator, since this approach prevents confidence interval endpoints lower than zero or higher than one.

A.1 Consistency of the Kaplan-Meier Estimator

In section 4, it is stated that the Kaplan-Meier estimator is consistent under the random censorship model. This is proved by Peterson (1977). The result is:

$$\operatorname*{plim}_{N\to\infty} \sup_{t\leq\omega} |\hat{S}(t) - S(t)| = 0 \tag{A.1}$$

Peterson (1977) derived this result by using Peterson's representation of $S(t)$.

A.2 Derivation of the limiting distribution of the Kaplan-Meier estimator

In section 4 the asymptotic properties of the Kaplan-Meier estimator for the survival function are presented. It is stated that, under the random censorship model, the Kaplan-Meier estimator for the survival function has a multivariate normal limiting distribution. This result is proved by Breslow and Crowley (1974), they present:

$$\sqrt{N}[\hat{S}(t) - S(t)] \overset{d}{\to} Z(t) \text{ as } N \to \infty \tag{A.2}$$

where $Z(t)$ is a Gaussian process with moments:

$$\mathbb{E}[Z(t)] = 0 \tag{A.3}$$

and

$$\text{Cov}[Z(t), Z(s)] = S(t)S(s) \int_0^{t\wedge s} \frac{1}{(1 - F(u))^2} \, \mathrm{d}F_u^*(u) \tag{A.4}$$

$$= S(t)S(s) \int_0^{t\wedge s} \frac{1}{(1 - F^*(u))(1 - F(u))} \, \mathrm{d}F^*(u) \tag{A.5}$$

where

$$F_u^*(t) = \mathbb{P}[X \leq t, \delta = 1] = \int_0^t (1 - H(u))\mathrm{d}F^*(u) \tag{A.6}$$

and

$$1 - F = (1 - F^*)(1 - H) \tag{A.7}$$

Recall $F*$ is the distribution of the real survival times X^*, H is the distribution of the censoring times Y and F is the distribution of the observed survival times X.

A.3 Estimation of the Covariance Matrix of Kaplan-Meier Estimator of the Survival Function

In the previous subsection the limiting distribution of the Kaplan-Meier estimator of the survival function is presented. The Kaplan-Meier estimator of the survival function is normally distributed with covariance matrix Σ, see equation (A.4). Because the distribution function of F, F^* and H are not known, the covariance matrix must be estimated. Miller (1980) proposes the following distraction to obtain an estimator for the asymptotic variance of $\hat{S}(t)$. Recall that $F_u^*(t) = \mathbb{P}[X \leq t \text{ and } \delta = 1]$, let

$$\mathrm{d}\hat{F}_u^*(t) = \frac{d_t}{N}. \tag{A.8}$$

Estimate the probability of failure at time t by the number of deaths at time t with respect to the total number of subjects. For $F(t) = \mathbb{P}[X \leq t]$, let

$$1 - \hat{F}(t) = 1 - \frac{\sum_{i=1}^{t-1} w_i - \sum_{i=1}^{t} d_i}{N} \tag{A.9}$$

$$= \frac{N - \sum_{i=1}^{t-1} w_i - \sum_{i=1}^{t-1} d_i - d_t}{N} \tag{A.10}$$

$$= \frac{n_t - d_t}{N}. \tag{A.11}$$

Estimate the probability that an observed survival time is smaller or equal than t by the number of subjects that has a smaller or equal observed survival time (died or withdrawn) with respect to the total number of subjects. Notice that we assume that subjects that withdraw at time t (w_t), will withdraw between time t and $t+1$. For $F(t-) = \mathbb{P}[X < t]$, where $t-$ means just before time period t, let

$$1 - \hat{F}(t-) = 1 - \frac{\sum_{i=1}^{t-1} w_i - \sum_{i=1}^{t-1} d_i}{N} \tag{A.12}$$

$$= \frac{n_t}{N}. \tag{A.13}$$

Estimate the probability that an observed survival time is smaller than t by the number of subjects that has a smaller observed survival time (died or withdrawn) with respect to the total number of subjects.

Replace $(1 - F(u))^2$ in equation (A.4) by $(1 - F(u))(1 - F(u-))$. Substitution of the above estimates gives:

$$\widehat{\mathrm{Cov}}(\hat{S}(t), \hat{S}(s)) = \frac{S(t)S(s)}{N} \sum_{i \le t \wedge s} \frac{\frac{d_i}{N}}{\left(\frac{n_i - d_i}{N}\right)\left(\frac{n_i}{N}\right)} \tag{A.14}$$

$$= S(t)S(s) \sum_{i \le t \wedge s} \frac{d_i}{n_i(n_i - d_i)} \tag{A.15}$$

which is precisely the Greenwood's formula [Greenwood (1926)]. Notice that $S(t)$ can be estimated by $\hat{S}(t)$ and $S(s)$ can be estimated by $\hat{S}(s)$.

Anderson et all (1993) show that the Greenwood's formula, equation (A.15) is a consistent estimate for the covariance matrix, equation (A.4) . Notice that $\hat{S}(s)$ and $\hat{S}(t)$ are consistent estimators for $S(s)$ and $S(t)$. Thus, it is only needed to prove that $\sum_{i \le t} \frac{d_i}{n_i(n_i - d_i)}$, denoted by $\hat{\sigma}^2(t)$ is a consistent estimator for $\frac{1}{N} \int_0^t \frac{1}{(1-F(u))^2} \, \mathrm{d}F_u^*(u)$, denoted by $\frac{1}{N}\sigma^2(t)$. Anderson et all (1993) showed:

$$\underset{N \to \infty}{\mathrm{plim}} \sup_{t \le \omega} |N\hat{\sigma}^2(t) - \sigma^2(t)| = 0 \tag{A.16}$$

Conclude that the Greenwood's formula is a consistent estimator for the variance of the Kaplan-Meier estimator for the survival function. Anderson et all (1993) propose several other estimators for the covariance matrix. Anderson et all (1993) show that all these alternatives are more biased than the Greenwood's formula.

A.4 Pointwise Confidence Intervals for the Kaplan-Meier Estimator

Above is presented that the Kaplan-Meier estimator of the survival function is asymptotically normally distributed. The mean and variance can be use to derive confidence intervals.

Pointwise confidence intervals are obtained by adding and subtracting the product of the estimated standard error times a quantile of the standard normal distribution.

$$\hat{S}(t) \pm z_{1-\frac{\alpha}{2}} \widehat{SE}\{\hat{S}(t)\} \tag{A.17}$$

where $z_{1-\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ quantile of the standard normal distribution and $\widehat{SE}$ represents the estimated standard error of the argument in parentheses. However, [Thomas and Grunkemeier (1975)] argued that this interval is not satisfactory for small sample sizes and can lead to confidence interval endpoints that are less than zero or greater than one. Therefore, Kalbfleisch and Prentice (1980) suggest the log-log survival function:

$$\ln[-\ln(\hat{S}(t))] \tag{A.18}$$

The advantage of this function is that it has a range from minus to plus infinity. The endpoints of a 100(1-α) percent confidence interval for $\ln[-\ln(\hat{S}(t))]$ are given by the expression

$$\exp[-\exp(\hat{c}_u)] \text{ and } \exp[-\exp(\hat{c}_l)].$$

where $\hat{c}_u$ is

$$\ln[-\ln(\hat{S}(t))] + z_{1-\frac{\alpha}{2}} \widehat{\mathrm{SE}}\{\ln[-\ln(\hat{S}(t))]\},$$

and where $\hat{c}_l$ is

$$\ln[-\ln(\hat{S}(t))] - z_{1-\frac{\alpha}{2}} \widehat{\mathrm{SE}}\{\ln[-\ln(\hat{S}(t))]\},$$

where $z_{1-\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ quantile of the standard normal distribution and $\widehat{\mathrm{SE}}$ represents the estimated standard error of the argument. Notice that the lower endpoint in the log-log survival function confidence interval $\hat{c}_l$ yields the upper endpoint of the estimated survival function confidence interval and vice versa.

Hosmer and Lemeshow (1999) present the following estimated variance for $\ln[-\ln(\hat{S}(t))]$ based on Greenwood's formula:

$$\widehat{\mathrm{Var}}\left[\ln[-\ln(\hat{S}(t))]\right] = \frac{1}{[\ln(\hat{S}(t))]^2} \sum_{i=1}^{t} \frac{d_i}{n_i(n_i - d_i)}. \tag{A.19}$$

From this estimated variance, the estimated standard error $\widehat{\mathrm{SE}}$ can be obtained.

Appendix B

Derivation of the Properties of the Conditional Cure Rate

In chapter 5 the unconditional cure rate and the conditional cure rate are described. Only a few adjustments to the Kaplan-Meier estimator for the survival function are needed to obtain an estimate for the unconditional cure rate and its asymptotic properties. The derivation of an expression and an estimate for the conditional cure rate is less trivial. The results are presented in section 5.5 . This section exists of the derivation of the results.

B.1 Estimation of the Conditional Cure Rate

Recall the expression of the Conditional Cure Rate

$$CCR(t) = 1 - \frac{S(\omega)}{S(t)}$$

Obviously, the numerator and denominator of the fraction can be estimated by the Kaplan-Meier estimator.

$$\operatorname*{plim}_{N \to \infty} \hat{S}(t) = S(t) \ \forall t \in \{1, ..., \omega\} \tag{B.1}$$

Subsequently, Slutsky's theorem is needed to obtain an consistent estimator for the $CC(t)$:

$$\operatorname*{plim}_{N \to \infty} 1 - \frac{\hat{S}(\omega)}{\hat{S}(t)} = 1 - \frac{S(\omega)}{S(t)} \tag{B.2}$$

Notice that this holds if $S(t) \neq 0$.

B.2 Limiting Distribution and Variance for the Conditional Cure Rate

To obtain the limiting distribution and variance first recall the estimator for $S(t) \forall t \in \{1, ..., \omega\}$:

$$\hat{S}(t) = \prod_{i=1}^{t} \left(1 - \frac{d_i}{n_i(n_i - d_i)}\right) \tag{B.3}$$

This expression can be substituted in the consistent estimator of the conditional cure rate $\widehat{CCR}(t)$ at any time t:

$$\widehat{CCR}(t) = 1 - \frac{\hat{S}(\omega)}{\hat{S}(t)} = 1 - \frac{\prod_{i=1}^{\omega} \left(1 - \frac{d_i}{n_i}\right)}{\prod_{i=1}^{t} \left(1 - \frac{d_i}{n_i}\right)} = 1 - \prod_{i=t}^{\omega} \left(1 - \frac{d_i}{n_i}\right) \tag{B.4}$$

Notice that this expression is very similar to the Kaplan-Meier estimator of the survival function $\hat{S}$. Anderson et all (1993) stated that under the random censorship model, the estimator of the $CCR(t)$ for any time period t has a limiting distribution which is asymptotically normal:

$$\sqrt{N}(\widehat{CCR}(t) - CCR(t)) \xrightarrow{d} \mathcal{N}(0, \sigma(t)^2) \tag{B.5}$$

with

$$\sigma^2(t) = \left(\frac{S(\omega)}{S(t)}\right)^2 \int_t^{\omega} \frac{1}{(1 - F(u))^2} \, \mathrm{d}F_u^*(u) \tag{B.6}$$

Recall $F*$ is the distribution of the real survival times X^*, H is the distribution of the censoring times Y and F is the distribution of the observed survival times X. F_u^* is the distribution of the real survival times X^* of the uncensored observations: $F_u^*(t) = \mathbb{P}[X \leq t \text{ and } \delta = 1]$. Because the distribution function of F, F^* and H are not known, the covariance matrix must be estimated. Anderson et all (1993) propose a distraction, similar to the distraction presented in Appendix A.3, to obtain an estimator for the asymptotic variance of $\widehat{CC}(t)$. Let

$$\begin{aligned} \mathrm{d}\hat{F}_u^*(t) &= \frac{d_t}{N} \\ 1 - \hat{F}(t) &= \frac{n_t - d_t}{N} \\ 1 - \hat{F}(t-) &= \frac{n_t}{N}. \end{aligned}$$

Replacement of $(1-F(u))^2$ by $(1-F(u))(1-F(u-))$ in the asymptotic variance and substitution of the above estimates gives

$$\hat{\sigma}(t)^2 = \widehat{\mathrm{Var}}(\widehat{CCR}(t)) = \frac{1}{N}\left(\frac{S(\omega)}{S(t)}\right)^2 \sum_{i=t}^{\omega} \frac{\frac{d_i}{N}}{\left(\frac{n_i-d_i}{N}\right)\left(\frac{n_i}{N}\right)} \tag{B.7}$$

$$= \left(\frac{S(\omega)}{S(t)}\right)^2 \sum_{i=t}^{\omega} \frac{d_i}{n_i(n_i-d_i)} \tag{B.8}$$

Anderson et all (1993) conclude that Greenwood's formula is a consistent estimator for the variance of the estimator of the conditional cure rate. See Appendix A.3 for the reasoning.

B.3 Pointwise Confidence Intervals of the Conditional Cure Rate

The estimator of the conditional cure rate is asymptotically normal distributed and the mean and variance can be use to derive confidence intervals. Pointwise confidence interval are obtained by adding and subtracting the product of the estimated standard error times a quantile of the standard normal distribution. This gives us:

$$\widehat{CCR}(t) \pm z_{1-\frac{\alpha}{2}} \widehat{SE}\{\widehat{CC}(t)\}$$

where $z_{1-\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ quantile of the standard normal distribution and $\widehat{SE}$ represents the estimated standard error of the argument in parentheses.

However, applying this theory directly to the conditional cure rate estimator can lead to confidence interval endpoints that are less than zero or greater than one. Therefore the log-log conditional cure rate is introduced.

$$\ln\left[-\ln(\widehat{CCR}(t))\right] = \ln\left[-\ln\left(1 - \frac{\hat{S}(\omega)}{\hat{S}(t)}\right)\right] \tag{B.9}$$

The advantage of this function is that it has a range from minus to plus infinity. The endpoints of a 100(1-α) percent confidence interval for $\ln[-\ln(\hat{S}(t))]$ are given by the expression

$$\exp[-\exp(\hat{c}_u)] \text{ and } \exp[-\exp(\hat{c}_l)].$$

where $(\hat{c}_u)$ is

$$\ln[-\ln(\hat{S}(t))] + z_{1-\frac{\alpha}{2}} \widehat{\mathrm{SE}}\{\ln[-\ln(\hat{S}(t))]\},$$

and where $(\hat{c}_l)$ is

$$\ln[-\ln(\hat{S}(t))] - z_{1-\frac{\alpha}{2}} \widehat{\mathrm{SE}}\{\ln[-\ln(\hat{S}(t))]\},$$

where $z_{1-\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ quantile of the standard normal distribution and $\widehat{\text{SE}}$ represents the estimated standard error of the argument. Notice that the lower endpoint in the log-log survival function confidence interval ($\hat{c}_l$) yields the upper endpoint of the estimated survival function confidence interval and vice versa.

Hosmer and Lemeshow (1999) present the following estimated variance for $\ln[-\ln(\hat{S}(t))]$ based on Greenwood's formula:

$$\widehat{\text{Var}}\left[\ln[-\ln(\widehat{CCR}(t))]\right] = \frac{1}{\ln\left[\left(\frac{\hat{S}(\omega)}{\hat{S}(t)}\right)\right]^2} \sum_{i=t}^{\omega} \frac{d_i}{n_i(n_i - d_i)}. \tag{B.10}$$

From this estimated variance, the estimated standard error ($\widehat{\text{SE}}$) can be obtained.

Appendix C

Kendall's Rank Correlation Test for Censored Data

There are several tests to test the hypothesis that data consisting of pairs (X_i, Y_i) is independent. Brown et al (1974) proposes to use a modification of Kendall's rank correlation test. In this study, this test is very useful, since it can deal with the fact that we have cases were the event of interest will not occur. First, we will present Kendall's (1970) rank correlation statistic. Then we will explain the modification that has to be made.

Suppose we have data consisting of n pairs (X_1, Y_1), ..., (X_n, Y_n), then the rank correlation statistic is

$$S = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} b_{ij} \tag{C.1}$$

where

$$a_{ij} = \begin{cases} 1 & \text{if } X_i > X_j \\ 0 & \text{if } X_i = X_j \\ -1 & \text{if } X_i < X_j \end{cases} \tag{C.2}$$

and

$$b_{ij} = \begin{cases} 1 & \text{if } Y_i > Y_j \\ 0 & \text{if } Y_i = Y_j \\ -1 & \text{if } Y_i < Y_j \end{cases} \tag{C.3}$$

If, however, either X or Y (or both) is censored, we may be unable to compute certain of the a's and b's. If we consider our pair of data (X_i^*, Y_i) in the cure case, Y_i is known for all i. X^* is known for all cured cases, for all liquidated cases ($X_i^* = \infty$), but X_i^* is not known for unresolved cases, because our observations is censored by Y_i. In this case Brown et al (1974) propose to use the pseudo-conditonal test. In this test the pairs (X_i^*, Y, i) with unknown X_i^* will be filled in by taking a random value from the distribution function defined by

$$\mathbb{P}(X_i > t | X_i > Y_i) \tag{C.4}$$

this can be estimated by

$$\frac{\hat{S}(t)}{\hat{S}(Y)} \tag{C.5}$$

Notice that $\hat{S}$ means the estimated survival function, we derived in previous sections. In this approach we simulate the unknown X_i^* by taking a random value for the estimated distribution function. Let us denote these simulate real survival times by $\hat{X}^*$. In our pseudo-conditional test, we apply the usual test based on S to (X_i^*, Y_i) for i such that $\delta_i = 1$ and to $(\hat{X}_i^*, Y_i)$ for i such that $\delta_i = 0$.

With the method presented above we can derive the statistic S. From this statistic we can derive an intuitive statistic τ:

$$\frac{S}{\frac{1}{2}n(n-1)} \tag{C.6}$$

τ lies between -1 and 1. -1 and 1 imply a very strong correlation between the sets X and Y. If τ is zero or close to zero we can suppose that both sets are independent. To test this hypothesis we use statistic z, which is approximately distributed as a standard normal when the variables are statistically independent:

$$\frac{3S}{\sqrt{\frac{1}{2}n(n-1)(2n+5)}} \tag{C.7}$$

We can obtain a p-value by finding the cumulative probability for a standard normal distribution at $-|z|$. For a two-tailed test, multiply that number by two to obtain the p-value.

Appendix D

List of Abbreviations

In this chapter a list of abbreviations is presented. The meaning and a short description is added.

Abbreviation	Meaning	Description
ABB	Aangesloten Banken Bedrijf	The local banks in the Netherlands that are affiliated with Rabobank
CCR	Conditional Cure Rate	Long-term probability of cure, based on the time a case is already in default, see also UCR and CR
CPRR	Conditional Provision Release Rate	Percentage of the provisions taken for some default that will release based on the time a case is already in default. See also PRR and UPRR.
CR	Cure Rate	Currently used definition of the probability to cure, estimated point-in-time and with one year horizon. See also UCR and CCR.
CZLR	Conditional Zero Loss Rate	Probability to end up in the zero loss state based on the time a case is already in default. See also UZLR.
EaD	Exposure at Default	The exposure of a client at the moment of default.
EC	Economic Capital	The amount of risk capital, assessed on a realistic basis, that a firm requires to cover the risk that it is running.
EL	Expected Loss	The normal costs of doing business.

LGD	Loss Given Default	Percentage of the exposure that will not be paid back given that a client is in default.
LGL	Loss Given Loss	Percentage of the exposure that will not be paid back given that the client is liquidated.
LPR	Loss Provision Rate	The ratio of the write off for a certain case or portfolio and the provision taken for this case or portfolio.
PD	Probability of Default	The probability that a loan will default within a year.
PRR	Provision Release Rate	Currently used percentage of the taken provisions that will release. See also UPRR and CPRR.
RC	Regulatory Capital	The amount of capital a bank has to hold as required by its financial regulator.
RWA	Riks-Weighted Assets	A bank's assets or off-balance-sheet exposures, weighted according to risk.
SE	Standard Error	Square root of the variance, needed to determine confidence interval.
SL	Stress Loss	The potential unexpected loss against which is judged to be too expensive to hold capital against.
UCR	Unconditional Cure Rate	Long-term probability to cure. See also CR and CCR
UL	Unexpected Loss	The potential unexpected loss for which capital should be held.
UPRR	Unconditional Provision Release Rate	Percentage of provisions taken for default that will release. See also PRR and CPRR.
UZLR	Unconditional Zero Los Rate	Probability to end up in the zero loss state. See also CZLR