Credit price optimisation using survival analysis

M Smuts

orcid.org 0000-0002-3485-7761

Thesis accepted in fulfilment of the requirements for the degree *Doctor of Philosophy in Science with Business Mathematics* at the North-West University

Promoter: Prof SE Terblanche

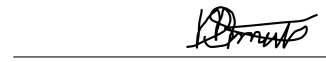
Co-promoter: Prof JS Allison

Graduation October 2020

22996168

Declaration

I, the undersigned, declare that the work contained in this thesis is my own work, except for references specifically indicated in the text, and that I have not previously submitted it elsewhere for degree purposes.



M. Smuts

2020/06/02

Date

Abstract

The competitive nature of the financial industry requires the effective use of prescriptive models to assist with strategic decision-making. One of the challenges in managing consumer credit portfolios is determining the optimal prices (*i.e.* interest rates) that maximise both the loan take-up probability of a potential borrower and the expected net present interest income (NPII) to the lender, while still adhering to certain risk distribution constraints on the portfolio. According to Phillips (2013) the miss-allocation and miss-pricing of consumer credit may have a severe impact on the global economy as seen in the 2008 global financial crisis. This impact can mainly be attributed to the high cost associated with an unexpectedly high number of defaults.

Traditionally, risk-based pricing has been used to determine the price of consumer credit. For this type of credit, the price included a risk premium which is dependent on the risk category of the borrower (or customer). However, this approach does not account for the demand of the customers *i.e.* the willingness of the customers to pay for a product or service. Hence, in recent years, pricing methodologies have moved away from risk-based pricing towards price optimisation (Phillips, 2013). In price optimisation, the willingness of a customer to pay for a product (or demand) is mathematically represented by a price response function, where the demand is expressed as a function of price (Terblanche and De la Rey, 2014).

In this study, a price response model that not only relates loan take-up probabilities to price but also to loan-to-value (LTV), is presented. This allows one to relate the demand of a borrower to both a change in price and LTV. Furthermore, by including the LTV in the credit price optimisation problem, constraints can be imposed to limit the proportion of loans with a high LTV, since these loans are considered more risky as apposed to loans with lower LTVs (see Phillips, 2013 and Caufield, 2012). Two different approaches, namely a non-linear and a piece-wise linear approximation approach, were used to simultaneously determine the optimal price and LTV, when the objective is to maximise the expected value of the NPII. Although both approaches yield similar results, the latter provides proven optimal solutions and, in addition to this, it allows for the inclusion of binary decision variables, which facilitate logical decision-making capability on a portfolio level. The results indicate that, when constraints are imposed on the risk distribution to limit the take-up proportion of high risk customers, a higher average price is offered to customers deemed more risky in conjunction with a lower average LTV. Conversely, the low and medium risk customers are offered a lower average price together with a

higher average LTV, subsequently increasing the take-up proportion of these risk gradings. In addition to this, when constraints are imposed on the loans with a high LTV, the average LTV of the high risk customers were substantially lower when compared to the low and medium risk customers.

To make provision that a borrower may default during a loan, a parametric mixture cure model (which is a generalisation of the well-known Cox Proportional Hazard model) was used to estimate the probability that a borrower is still repaying the loan (not defaulting on the loan). The use of the mixture cure model permits one to take into account the relatively large proportion of customers that are not susceptible to default, when solving the optimisation problem. This newly developed optimisation model was applied to two simulated data sets. The results demonstrate that a clear interaction between price, LTV, take-up and survival probabilities exist. On average the price offered to the low risk customers were lower than the price offered to the high risk customers. On the other hand, the LTV proposed to the low risk customers was, on average, higher than that of the high risk customers. As a result of including logical decision making variables, customers with lower survival probabilities (or higher default probabilities), were excluded from the portfolio when the price offered to these customers were too high.

The literature study on survival analysis, and more specifically on the parametric families of distributions present in survival analysis, led to the development of a new goodness-of-fit test for exponentiality. The newly proposed test performed favourably in terms of powers relative to several existing tests. The test was also applied to a simulated data set to determine whether the parametric assumptions underlying the Cox Proportional Hazard model were violated.

Keywords: price optimisation, survival analysis, loan-to-value, credit risk

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost I would like to thank our Heavenly Father for giving me the opportunity, strength and guidance to complete my PhD.

I would like to acknowledge and express my gratitude to the following people for their immense contribution throughout this journey. Firstly, to my promoter, Prof. Fanie Terblanche, and co-promoter, Prof. James Allison, thank you for your continued support and guidance as well as the time you invested in this study. I sincerely appreciate all your support and I am very grateful for your assistance. Next, I would also like to thank Dr. Jaco Visagie for his time and assistance.

To my parents, Marius and Hanlie Smuts, and in-laws, Piet and Anita Roos; thank you for you belief, encouragement and support from start to finish.

Lastly, I would like to thank my wife, Anke Smuts, for being by my side every step of the way. Your love and support kept me positive and motivated, especially during tough times. I love you, always.

List of Abbreviations

BFS	Basic Feasible Solution
CPH	Cox Proportional Hazard
ILP	Integer Linear Programming
LGD	Loss Given Default
LP	Linear Programming
LTV	Loan-To-Value
MILP	Mixed Integer Linear Programming
NCA	National Credit Act
NIACC	Net Income After Cost of Capital
NII	Net Interest Income
NLP	Non-Linear Programming
NPII	Net Present Interest Income
PVNI	Present Value of the Net Interest
RAROC	Risk Adjusted Return On Capital
ROE	Return on Equity
SARB	South African Reserve Bank
SOS1	Special Ordered Set of type 1

Chapter 1

Introduction

1.1 Overview

The miss-pricing and miss-allocation of consumer credit can have a severe impact on the financial institution providing the credit or even on the global economy as witnessed in the 2008 financial crisis (Phillips, 2013). Simply increasing the price (interest rate) of loans may result in higher net present interest income (NPII) generated from these loans. However, this could potentially increase the probability that borrowers will default on the loan and/or decrease the probability of borrowers taking up a loan (take-up probability) at the higher price, ultimately decreasing the expected NPII generated from these loans. Hence, the interaction between price and risk is of great importance in the consumer credit market. According to Phillips (2013) this interaction has not yet been fully addressed in credit price optimisation.

The competitive nature of the financial industry requires the effective use of prescriptive models to assist with strategic decision-making. One of the challenges in consumer credit portfolios is to determine the optimal prices that maximise both the loan take-up probability of a potential borrower and the expected net present interest income to the lender, while still adhering to certain risk distribution constraints on the portfolio.

Traditionally, risk-based pricing was used to determine the price for consumer credit. For this type of credit, the price included a risk premium which was dependent on the risk category of the borrower (or customer). However, in recent years pricing methodologies moved away from risk-based pricing towards demand (or profit) based pricing (see, *e.g.*, Phillips, 2013 and Terblanche and De la Rey, 2014). In demand based pricing, the demand of a potential borrower is mathematically captured by a price elasticity model (price response model) where the demand is expressed as a function of price. In consumer credit the demand refers to the probability that the potential borrower will take up a loan at a quoted price.

In this study a price response model that not only relates loan take-up probabilities to price but also to loan-to-value (LTV), is investigated. This allows one to relate the demand of a borrower not

only to a change in price, but also to a change in LTV. Furthermore, by including the LTV in a credit price optimisation problem, constraints can be imposed to limit the proportion of loans with a high LTV, since these loans are considered more risky than loans with lower LTVs (see Phillips, 2013 and Caufield, 2012). A piece-wise linear approximation approach is followed to simultaneously determine the optimal price and LTV for a potential borrower, while adhering to the risk distribution and also LTV constraints on the portfolio. This simultaneous optimisation of price and LTV is something that has not been addressed in the literature to date.

There are various risks associated with the repayment of a loan, namely default, early settlement and prepayment. Caufield (2012) suggest that risks, more specifically default, should be modelled using a more forward looking perspective. One way of modelling these risks (*e.g.* the probability of default) is by using survival probabilities. More specifically, Dirick et al. (2017) suggested that an accurate estimate of the probability that a borrower is still repaying a loan (not defaulting on the loan) at every time instant of the loan, can be obtained using different models that originated in the field of survival analysis.

The question now arises, how does one maximise the expected NPII while not only taking into account the simultaneous effect between price and LTV, but also making provision that a borrower may default during the loan? The question is addressed in this thesis.

1.2 Objectives

The main aim of this study is to build an optimisation model that maximises the expected value of the NPII by finding the right balance between price and LTV according to a price response model. This model must also be able to incorporate the effect of price and LTV on the survival behaviour of the customers during the loan.

The primary objectives of this thesis can be summarised as follows:

- Review the existing literature on different pricing methodologies.
- Review the existing literature on mathematical optimisation and the various methods used to solve optimisation problems.
- Develop a new optimisation model that determines the optimal price and LTV that maximises the expected NPII by using a piece-wise linear approximation approach.
- Discuss the existing literature on survival analysis with emphasis on different survival models, which include the Cox Proportional Hazards (CPH) model and the mixture cure model (a more general alternative to the CPH).

- Develop a new optimisation model that maximises the expected NPII by finding the right balance between price and LTV while incorporating estimated survival probabilities obtained from a mixture cure model.
- Evaluate and investigate the performance of the newly developed optimisation model by implementing the model on two simulated data sets.

A secondary objective that originated from the literature review on survival models was to develop and investigate a new goodness-of-fit test for exponentiality, that can ultimately be used to determine the adequacy of fit of a parametric CPH model.

1.3 Thesis outline

In Chapter 2, different pricing methodologies are discussed and the credit price optimisation problem is formulated. Chapter 3 contains an overview of optimisation theory and some methods that can be used to solve optimisation problems. The focus is on solving non-linear programming problems using a piece-wise linear approximation approach. In Chapter 4, a new optimisation model is developed. This model makes use of a piece-wise linear approximation approach to determine the optimal price and LTV, to quote a borrower, that maximises the expected NPII. The model is applied to data obtained from a financial institution to investigate the model behaviour for different constraints imposed on the risk distribution and on LTV. In Chapter 5, an overview of some of the basic concepts of survival analysis are presented. The focus is specifically on the CPH model and the mixture cure model, which is a more general alternative to the CPH model. The estimation of these models (mainly parametrically) and how they can be used to estimate the survival probability of a customer during the loan term, are discussed. In Chapter 6, a new optimisation model, which incorporates the estimated survival probabilities obtained from a mixture cure model, is developed. This model again makes use of a piece-wise linear approximation approach to determine the optimal price and LTV by taking into account the effect of these variables on the survival behaviour of the customers during the loan. The steps to simulate data from a parametric mixture cure model assuming different possible baseline distributions are outlined. The model is implemented on two data sets, where the survival times are simulated from a parametric mixture cure model using the covariates obtained from a financial institution. The thesis concludes in Chapter 7 with some final remarks and suggestions on possible future research in the area of credit price optimisation and goodness-of-fit testing.

Appendix A contains the published paper “A new goodness-of-fit test for exponentiality based on a conditional moment characterisation”. In Appendix B, the R-code used to maximise the log-likelihood in a mixture cure model, for both the exponential and Weibull baseline distributions, is provided. Appendix C contains the tables and figures of the model implemented on the survival data generated from a mixture cure model with a Weibull baseline distribution.

Chapter 2

Pricing methodologies

2.1 Introduction

"Price is what you pay. Value is what you get." – the famous words of Warren Buffet. The way in which prices are set, evaluated, updated and managed often vary considerably from one industry to another and even from company to company within an industry. Pricing is considered as one of the greatest levers to increase profitability (Chapter 34 of Özer et al., 2012). The use of consumer credit is a rising practice and the pricing thereof was once considered to be simple and straightforward, but as seen in the recent financial crisis, the miss-pricing and miss-allocation of consumer credit can have a severe impact on the global economy (Phillips, 2013).

Phillips (2013) refers to consumer credit as lines and loans extended to individuals as apposed to businesses or entities in the government. Consequently, consumer credit can take on several forms. It may either be secured (collateralised) as is the case with home loans (mortgage), home equity and auto loans, or unsecured as is the case with credit cards, personal loans and overdrafts for example (Chapter 8 of Özer et al., 2012). Even though the terms, also referred to in relevant literature as characteristics, and conditions of consumer credit may vary, one of the most important elements that control the risk and performance of a loan is the interest rate, which is frequently referred to as the price of a loan.

Until the 1980's, financial institutions and banks charged a similar interest rate to all borrowers for a given type of credit *i.e.* personal loans, a mortgages or even the interest rate charged on credit cards (Chapter 3 of Thomas, 2009). Hence, there was no segmentation of the population in the consumer credit market. This price was set to cover all possible costs and expenses, including the cost of funds, operating expenses and expected credit losses incurred by these credit products. However, in the early 1990's, banks and finance companies noticed that profitability in the consumer credit industry could be increased by differentiating between customers (note that the word customer and borrower will be used interchangeably). This lead to a strategy where prices were no longer similar for all the customers, but instead, varied according to the perceived level of risk associated with the loan or customer and different loan terms. Hence, a risk-based pricing approach, which according to Özer et al. (2012) is

simply the consumer credit industry's name for cost-plus pricing, was adopted, accounting for the fact that the expected credit losses were different among the different types of customers. This approach has enabled lenders to maintain the same margin and profit when lowering the prices for customers with lower expected credit losses than the average.

While the development and adoption of a risk-based pricing approach was of great importance in the consumer credit industry, the approach took no account of the demand of the customers *i.e.* the willingness of the customers to pay for a product or service. In recent years, pricing methodologies have moved away from risk-based pricing towards demand-based pricing (Skugge, 2011). In demand-based pricing, the demand of a potential customer is mathematically represented by a response model (price response function), where the demand is expressed as a function of price (Terblanche and De la Rey, 2014). This enables the lender to take the customer's willingness to pay for a credit product into account when determining the price. Furthermore, Caufield (2012) argues that by adopting a profit-based pricing (price optimisation) approach, which is a combination of risk-based pricing and demand-based pricing, an increase in expected profits of between 10 and 25 percent can be achieved. Therefore, according to Caufield (2012), pricing needs to become a specialist function within lending practices due to the accelerated adoption of profit-based pricing (price optimisation) and the advances in the modelling of customer behaviour and losses.

According to Phillips (2013), the use of explicit price optimisation approaches in the lending industry is a relatively new concept when compared to retail and passenger airline industries. Caufield (2012) suggests that, just like these other industries, the consumer credit industry should use pricing to achieve their objectives relating to profit and revenue. In conjunction with these objectives, constraints have to be imposed on the credit portfolio, such as limits on the proportion of customers with a specified level of risk. In addition to these constraints, Caufield (2012) suggests that lenders also want to limit the proportion of home loans with a high loan-to-value (LTV) as these loans are considered high risk.

Therefore, the words of Warren Buffet are without a doubt true, the value you attach to a product or service reflects in the price you are willing to pay for the product or service.

The remainder of this chapter is organised as follows: in Section 2.2 an overview of the history of interest and where it all started is provided. Then, in Section 2.3 the traditional risk-based pricing approach is first considered where after the focus shifts to the modern profit-based pricing approach in Section 2.4 that includes income and response functions used in this approach. This chapter concludes in Section 2.4.3 where the current credit price optimisation approach, that takes into account the customer's willingness to pay for a consumer credit product when determining the optimal price, is discussed.

2.2 The history of interest

Earl Wilson once said: "Modern man drives a mortgaged car over a bond financed highway on credit card gas." The practice of lending dates way back into the mists of time and has pretty much become second nature in the common era (Redden, 2014). Moreover, from the beginning of civilisation, there were those who lent and those who borrowed (Ecoinomic, 2018). In the most general sense, lending, also known as financing, is the temporary provisioning of money, property or other material goods to another person with the expectation that it will be repaid (see Murray, 2019). The first loans were in the form of seeds in the agricultural society. For this society, the possession of seeds was of great value, since a single seed could yield a crop of hundreds of seeds. Similar to seeds, live stock could also reproduce themselves. Hence, the acquisition of seeds and live stock with the purpose to reproduce themselves lead to the justification of interest. However, in consumer lending, the price of a loan refers to the interest rate charged by the lender to the borrower (Thomas, 2009), since the interest rate is essentially the price paid for lending money. According to Guardia (2002) the definition of credit in the private sector is, "lending to households and private businesses to carry out transactions in the private sector of the economy." Hand and Henley (1997) refer to credit as the amount of money that is loaned to a consumer by a financial institution, whereafter the loaned money has to be repaid, with interest and in installments, that is, equal payments spread over a period of time that was agreed upon. Furthermore, Guardia (2002) states that credit to the household sector comprises of consumer credit and mortgage credit.

Retail banks offer a large number of credit products. Some of these products include mortgages, loans, debit cards, credit cards, telephone banking, internet banking, savings accounts, etc (see Hand, 2001). Retail credit (lending) products include credit cards, auto loans, student loans, personal loans, and loans secured by an individual's residence, including home improvement loans, home equity and first mortgage. Each of these products has their own unique features and properties, making the retail banking market a complex environment. That is, the interest rate of a loan may be fixed or variable over the life of the loan and for secured loans like mortgages, the term may be fixed, whereas in credit cards and lines of credit the term is usually indefinite. Furthermore, several forms of price exists *i.e.* interest rate and fees to name a few. The most revenue generated to the lender of credit is however, in the form of interest, but as mentioned, there are also fees associated with credit (Özer et al., 2012). In this thesis, the term price is used to refer to interest rate only.

In consumer credit, the pricing challenges are quite different compared to other industries. The approaches used to determine the price for loans have changed over the years with the advances in computer technology. Even as late as the early 1990s, the conventional lenders simply posted a "house rate" for each loan type, with most high risk borrowers being rejected a loan offer (Johnson, 1992). Similarly, Thomas (2009) states that it is surprising that comparable prices were charged by lenders for secured loans and unsecured loans, or even the interest rates that they applied to credit cards. However,

during the early 1990s, banks noticed that their profitability could be improved by offering different loan terms to different segments of the population. At first, the use of credit scoring technology to segment loan applicants into different risk categories, mainly affected the decision whether or not to issue a loan to the applicant (Walke et al., 2018). Hereafter, these credit scoring technologies were increasingly being used by banks and credit unions, not only to decide whether or not to issue a loan, but also to assign different prices to borrowers based on their credit risk (Edelberg, 2006). This pricing approach, where the prices that are offered to individual borrowers vary according to the borrower's risk category, is known as risk-based pricing.

2.3 Risk-based pricing

Thomas (2009) suggests that even though risk-based pricing is the strategy of offering different loan terms (interest rates) to different borrowers, the price of a loan in consumer lending is rarely just set by the riskiness of the loan, but rather factored into the costs of the loan. Furthermore, the price of a loan is set by banks to encompass their objectives like maximising expected profit, market share or return on capital.

Thomas (2009) states that since the early 1990s, banks have noticed that risk-based pricing could increase their profitability. According to Edelberg (2006), risked-based pricing is the most common approach used for pricing consumer credit in the United States and that lenders increasingly used this approach for the pricing of interest rates during the mid 1990s. Edelberg (2006) established that even a slight increase in the probability of default leads to a corresponding increase in interest rate for first mortgages, automobile loans and second mortgages that triple, double and increase six-fold, respectively. Walke et al. (2018) found that the adopters of risk-based pricing in the United States increased the availability of loans, however, primarily to lower risk borrowers rather than high risk borrowers as initially expected. Magri and Pico (2011) found that, consistent with the adoption of credit scoring, there is evidence in their estimates that show that Italian lenders increasingly priced mortgage interest rates in accordance to household credit risk *i.e.* risk-based.

Phillips (2013) suggests that the rationale behind risk-based pricing is clear, a high risk customer should pay a higher price to offset the higher default probability of the customer and costs to the lender. However, empirical evidence suggests that high risk customers are less sensitive to a change in price than low risk customers and subsequently this lead to a tendency that charged them even higher rates (Phillips, 2013). Furthermore, in addition to this price sensitivity, adverse selection is likely to play a significant role in the pricing of retail credit as it is an important characteristic within this industry (Stiglitz and Weiss, 1981). Various definitions of adverse selection can be found in literature, with a distinction being made between adverse selection on hidden (indirect) and observable (direct) information. Here, adverse selection on observable information is applicable and refers to the instance where low risk customers are more sensitive to a change in price as apposed to high risk customers.

Thomas (2009) suggests that, since adverse selection influences the interaction between the probability of customers taking up a credit product and the quality of the customer, it should be taken into account when following a risk-based pricing approach. Terblanche and De la Rey (2014) illustrate the existence and effect of both price sensitivity and adverse selection in retail credit products as they find evidence that lower take up rates of a loan are associated with an increase in the price of the loan, and that high risk customers are more likely to take up a loan at the same quoted price compared to low risk customers. Furthermore, Park (1997) finds that in the credit card industry, an increase in price is associated with a decrease in the take up of this credit products with Karlan and Zinman (2008) finding similar evidence for credit cards in less developed economies, specifically South Africa.

Risk-based pricing was of great importance and has played a key role in the consumer credit industry. However, Caufield (2012) suggests that risk-based pricing is just the credit industry's version of cost-plus pricing and according to Skugge (2011) cost-plus pricing is an inside-out approach that neither takes into account the value of the product to the customer nor the willingness of the customer to pay for the product. Hence, in recent years there has been a significant shift in the pricing of credit products, with more and more lenders following a profit-based pricing approach (a combination between risk-based pricing and demand-based pricing). That is, lenders increasingly began to adopt price optimisation approaches, not only taking into account the risk of a customer but also the willingness of a customer to pay for the product.

2.4 Profit-based pricing

To understand the impact of price sensitivity on demand and expected profitability, the willingness of a customer to pay for a loan has to be taken into account when determining the price of the loan. In consumer credit this is referred to as price elasticity and is commonly captured using a price response function. In recent years, lenders began to adopt a price optimisation approach (see Phillips, 2013) that takes into account the willingness of the customer to pay for a loan. Terblanche and De la Rey (2014) refer to the price optimisation approach as demand-based pricing whereas Skugge (2011) regard it as an "outside-in" or value-based approach to pricing. Terblanche and De la Rey (2014) argue that in demand-based pricing, the demand of a potential customer is mathematically captured by a price elasticity model, in order to express the demand as a function of price. Furthermore, Caufield (2012) describes the price optimisation approach as a profit-based pricing approach *i.e.* an approach that combines risk-based pricing and demand-based pricing with the intention to maximise profits. In this thesis, the terms profit-based pricing and credit price optimisation will be used interchangeably when referring to price optimisation.

In credit pricing, the demand refers to the probability that a potential customer will take up a loan at the quoted price, also referred to as the take-up probability. The take-up probability is modelled using a price response function. Moreover, the price response function enables the lender to relate the take-up

probability to not only the price of a loan, but also the customer's risks, characteristics and the price of the loan. Examples of commonly used price response functions are the linear, constant-elasticity and logit price response functions. Below various types of price response functions are considered together with their properties.

2.4.1 The price response function

A price response function plays a key role in determining the price of a product or loan. The price response function commonly refers to how the demand for a product varies as the price varies. In credit pricing, this demand can be interpreted as the probability that a customer will take up a loan, that is, accept a loan offered at a given price (interest rate) and given the characteristics of the customer and the loan. Thomas (2009) suggests that in the same way as scoring the customer's chance of repaying a loan, *i.e.* not defaulting, one could score the likelihood of the customer taking up the loan or not. Here, the response score allows the lender to estimate the take-up probability of a loan as a function of the customer's characteristics**.** Consequently, the take-up probability is a function of the price charged for the loan and the characteristics of the customer with one characteristic typically being the probability that the customer will not repay the loan *i.e.* default on the loan. Banks and financial institutions can estimate the price response functions if they record whether or not the customers took up the loan as a result of offering different loan prices (interest rates) to customers. Together with the take-up, various other characteristics of the customers are usually available to the lender, for instance the credit score, age, geography, number of accounts, first time buyer etc.

First, consider the situation where the take-up probability (response probability) of a loan, denoted by $R(.)$, is only dependent on the price x charged for the loan *i.e.* $R(x)$. Thus, other characteristics of the customer (*e.g.* probability of default demographics *etc*.) and loan characteristics (amount, term *etc.*), do not affect the probability of loan take-up. Furthermore, assume that the price, x, charged for the loan can take on any value between 0 and ∞, even though constraints can be added such as $x_L \leq x \leq x_M$ where x_L denotes the minimum price to be charged *i.e.* the repurchase rate (interest rate at which banks borrow money from the South African Reserve Bank (SARB)) and x_M denotes the maximum allowable price to be charged for a loan. Note that these minimum and maximum limits can be set by the banks, but are usually determined by the National Credit Act (NCA). The following properties can be deduced for the price response function considering the above mentioned:

- $0 \leq R(x) \leq 1$, since $R(x)$ represents a probability. Further, by assuming that $\lim_{x\to\infty} R(x) \to 0$, the lender can always find a price, x, where the customer would not take up the loan, even though this price is highly unlikely to be charged due to maximum allowable prices.

- $R(x)$ is a non-increasing monotonic function of x, as a result of the inverse relationship between price and take-up probability. That is, an increase in price relates to a take-up probability that either stays the same or decreases.

- $R(x)$ is a continuous function in x if x is also continuous. This assumption and the previous assumption simplifies the analysis of the models, even though banks and financial institutions will in many cases set a finite number of possible prices to charge for a given loan product.

- $R(x)$ is a continuously differentiable function with respect to x and in conjunction with the monotone non-increasing property this suggests that

$$\frac{dR(x)}{dx} = R'(x) \leq 0.$$

Given the above properties of the price response function, it is often worthwhile to have a characterisation of the price sensitivity implied by the price response function at a specific price. The slope and the elasticity of the price response function are the two most common measures of price sensitivity. Hence, consider the characterisations of the price sensitivity of the price response function below.

Characterisation. Let x_1 and x_2 denote two different prices with $0 \leq x_1 < x_2$ and let $R(x_1)$ and $R(x_2)$ denote the demand at prices x_1 and x_2 respectively. A measure of how the demand changes as a result of a change in the price is given by the slope of the price response function, *i.e.* the difference in the demand divided by the change in the price. That is,

$$\delta(x_1,x_2) = \frac{R(x_1) - R(x_2)}{x_1 - x_2}. \tag{2.1}$$

Hence, by the non-increasing property of the price response function, $\delta(x_1,x_2)$ will always be less than or equal to 0 *i.e.* $\delta(x_1,x_2) \leq 0$. Note that if two prices are specified, the slope of the price response function will be constant across all prices only if the price response function is linear. It is however, common to determine the slope of the price response function at a single price, x_1. Here it can be computed as the limit of (2.1), that is

$$\begin{aligned}\delta(x_1) &= \lim_{h\to 0}\frac{1}{h}\left[R(x_1+h) - R(x_1)\right] \\ &= R'(x_1),\end{aligned}$$

where $R'(x_1)$ is the derivative of the price response function at the price x_1. By the differentiation and non-increasing property of the price response function, this derivative exists and is less than or equal to zero. Furthermore, the change in the demand as a result of a small change in the price can also be given by the slope, that is,

$$R(x_1) - R(x_2) \approx \delta(x_2)(x_1 - x_2).$$

This implies that for large negative slope the demand is more responsive to the price than for a smaller negative slope. Another measure of price sensitivity and perhaps the most common measure is the elasticity of the price response function.

Characterisation. Let x_1 and x_2 denote two different prices and let $R(x_1)$ and $R(x_2)$ denote the demand at prices x_1 and x_2 respectively. A measure of the sensitivity of demand to price is the so called elasticity, defined as the percentage change in the demand relative to the proportion change in the price given by

$$\varepsilon(x_1,x_2) = -\frac{100\{[R(x_2)-R(x_1)]/R(x_1)\}}{100\{(x_2-x_1)/x_1\}} = -\frac{[R(x_2)-R(x_1)]x_1}{[x_2-x_1]R(x_1)}. \tag{2.2}$$

The non-increasing property of $R(x)$ guarantees that the change in the demand is in the opposite direction to the change in the price, hence the negative sign is added to the right hand side of (2.2) to ensure that $\varepsilon(x_2,x_1) \geq 0$. This elasticity as defined in (2.2) is sometimes called the arc elasticity, since it depends on both the old and the new price to calculate fraction of the percentage change in the demand to the percentage change in the price. In general, the percentage increase (decrease) in demand as a result of a 1% decrease (increase) in price will not be the same, therefore to characterise the elasticity in full, both prices need to be specified. Furthermore, by taking the limit of (2.2), the point elasticity can be derived in a similar way as the slope. That is the limit as x_2 tends to x_1,

$$\varepsilon(x_1) = -\lim_{x_2\to x_1}\frac{[R(x_2)-R(x_1)]x_1}{[x_2-x_1]R(x_1)} = -\frac{R'(x_1)x_1}{R(x_1)}. \tag{2.3}$$

The point elasticity as calculated per equation (2.3), will be greater than or equal to zero, since $R'(x_1) \leq 0$. Hence, the point elasticity gives an estimate of the percentage decrease in the demand of a loan as a result of a 1% increase in the price of a loan. Moreover, the point elasticity indicates the relative change in the demand of the loan as a result of a unit relative change in the price of the loan.

Willingness to pay

Banks and financial institutions estimate the price response functions using information about whether or not a customer took up a loan at the quoted price. Therefore, the price response function is built on assumptions about the behavior of customers. Considering this, another way of viewing the take-up probability of a loan, is to describe the proportion of potential customers that would take up the loan at various prices. Hence, to determine whether the price response function is based on appropriate assumptions regarding the application thereof, it is useful to understand the assumptions about the behavior of customers underlying the price response functions. The willingness to pay for a loan is the most important assumption inherent in models of customer behavior. This assumes that each potential customer has a maximum price that they are willing to pay for a loan, *i.e.* a maximum price at which they are willing to take up the loan (the maximum price is sometimes referred to as the reservation price). The customer will therefore only take up the loan if and only if the price of the loan is below the maximum price they are willing to pay. Hence, let $m(x)$ denote the density function of the maximum

willingness to pay of the population of customers, then

$$\int_{x_1}^{\infty} m(x)\,dx \coloneqq \text{the proportion of the population willing to pay a price of } x_1 \text{ or more for the loan. This proportion is defined as } \mathrm{R}(x_1). \tag{2.4}$$

From (2.4) the maximum willingness to pay distribution can be derived using the price response function,

$$m(x) = -R'(x). \tag{2.5}$$

If the price goes up by an infinitesimal amount, those customers who have a maximum willingness to pay of exactly x, will turn down the offer.

Typical price response functions

Thomas (2009) suggests, based on observed numerical evidence, that the price response function has a reverse S-shape. This implies that when the price for a loan is very low, the probability of a customer taking up the loan is high, whereas if the price of the loan is very high, the probability of a customer taking up the loan is low. The sub-interval of prices between the very low and high prices, is where the elasticity of the price response function is very high. Hence, a small change in the price could lead to a large change in the take-up probability of the loan. Consider the different price response functions together with their respective properties.

Linear price response function. The linear price response function is the simplest function with some of the properties outlined above and is given by

$$R(x) = \max\{0, 1 - b(x - x_L)\} \text{ for } 0 \leq x_L \leq x \leq x_M, \tag{2.6}$$

where b denotes the slope of the price response function and x the quoted price. Recall that x_L and x_M denote the minimum and maximum allowable prices to be quoted. Furthermore, the quoted price, x, can be bounded above by $x_M = x_L + \frac{1}{b}$ since $R(x_M) = 0$ and therefore no customer will take the loan at that price.

The linear price response function is considered a simple function that vaguely resembles the reverse S-shape as displayed in Figure 2.1. However, this function is not a very realistic price response function due to the form of the elasticity. Consider the point elasticity for the linear price response

function defined by (2.3)

$$\varepsilon(x) = \begin{cases} -\frac{R'(x)x}{R(x)} = \frac{bx}{1-b(x-x_L)} & \text{if } x_L \leq x \leq x_M \\ 0 & \text{elsewhere,} \end{cases}$$

from which it is clear that there are discontinuities in the elasticity at the prices x_L and x_M. For the linear price response function, the maximum willingness to pay is uniformly distributed between x_L and x_M since, between this interval the density function as derived from (2.5) and (2.6), is constant, that is

$$m(x) = \begin{cases} -R'(x) = b & \text{if } x_L \leq x \leq x_M \\ 0 & \text{elsewhere.} \end{cases}$$

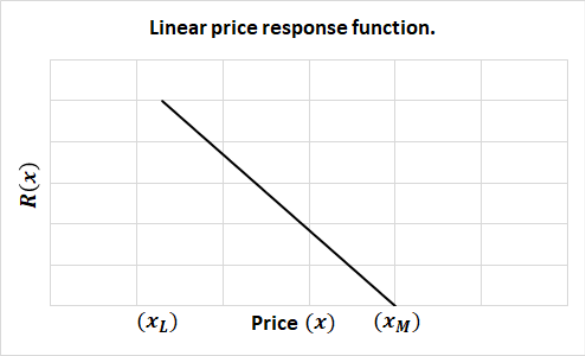


Figure 2.1: Linear price response function.

Constant-elasticity price response function. The constant-elasticity price response function has a constant point elasticity at all the prices, as its name suggests. That is, the elasticity for this price response function as by (2.3) is

$$\varepsilon(x) = \frac{-R'(x)x}{R(x)} = \varepsilon \ \ \forall \ \ x \geq x_L > 0, \tag{2.7}$$

where $\varepsilon > 0$ denotes the elasticity that is constant for all the prices. The corresponding price response function relating to (2.7) is

$$R(x) = K\left(\frac{x}{x_L}\right)^{-\varepsilon} \ \ \forall \ \ x \geq x_L > 0, \tag{2.8}$$

where $K > 0$ is a parameter (which is equal to 1 if it is assumed that $R(x_L) = 1$). Furthermore, the slope

of the price response function given by (2.8) is

$$R'(x) = -K\varepsilon \left(\frac{x}{x_L}\right)^{-\varepsilon-1} \left(\frac{1}{x_L}\right)$$

which is negative, since $\varepsilon > 0$ and $K > 0$. The constant-elasticity price response function is also downward sloping due to the negative slope. However, for the constant-elasticity price response function the demand never reaches 0 as the price increases, as can be seen in Figure 2.2. Hence, this price response function is neither finite nor adequate and for these reasons not a typical global presumed price response function. Furthermore, the constant-elasticity price response function has a corresponding willingness to pay distribution given by

$$m(x) = -R'(x) = K\varepsilon \left(\frac{x}{x_L}\right)^{-\varepsilon-1} \left(\frac{1}{x_L}\right) \quad \forall\, x \geq x_L > 0. \tag{2.9}$$

From (2.9), it can be concluded that the willingness to pay distribution drops steadily and approaches but never reaches zero as the price increases, yet another limitation of this price response function.

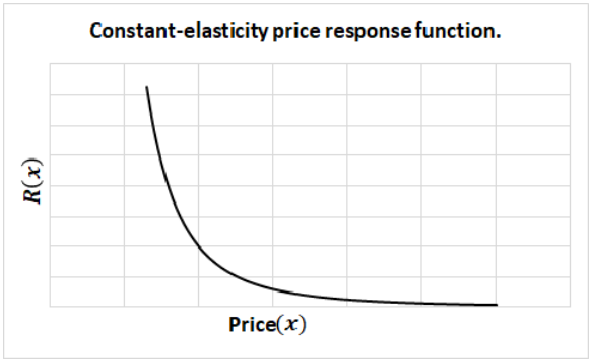


Figure 2.2: Constant-elasticity price response function.

The linear and constant-elasticity price response functions both have their limitations when considering their measures of sensitivity and willingness to pay distributions. For the linear price response function the corresponding willingness to pay distribution is uniform between x_L and x_M, whereas for the constant-elasticity price response function, the willingness to pay distribution gradually decreases and approaches, but never reaches zero, making both these price response functions unrealistic when modeling customer behavior.

A well-known function that is used to estimate the default probability of a customer is the logistic function. This function can also be used to estimate the demand of potential customers as a function of the quoted price. Hence, consider the logistic price response function together with the properties of this function.

Logistic price response function. In reality, when considering the price and demand of a loan, it is expected that the demand for a loan will be high (low) if the quoted price for the loan is low (high). Furthermore, it is also expected that the demand will be changing slowly with small price changes at very low (or high) prices and somewhere in between these, an interval of prices where the elasticity of the price response function is very high, as displayed in Figure 2.3. A price response function exhibiting this behaviour in demand, has a kind of a reversed S-shape as seen in Figure 2.3 and can be modelled using the logistic price response function. The logistic (or logit) price response function is given by

$$R(x) = \frac{e^{a-bx}}{1+e^{a-bx}} = \frac{1}{1+e^{-a+bx}} \Longleftrightarrow \ln\left(\frac{R(x)}{1-R(x)}\right) = a - bx, \tag{2.10}$$

where a and b are parameters with $b > 0$.

Considering the logistic price response function in (2.10), the log odds of the take-up probability, $R(x)$, versus the non take-up probability, $1-R(x)$, is a linear function of the quoted price x. Hence, by considering the log odds as a price response score, the price response score is linear in terms of the price response function with the gradient being $-b$ with $b > 0$ since, the take-up probability decreases with an increase in the price. The elasticity and willingness to pay distribution corresponding to the logistic price response function, as derived from (2.3) and (2.5) is given by

$$\varepsilon(x) = \frac{-R'(x)x}{R(x)} = \frac{b\left(e^{-a+bx}\right)x}{\left(1+e^{-a+bx}\right)^2} \times \frac{1+e^{-a+bx}}{1} = \frac{bx\left(e^{-a+bx}\right)}{1+e^{-a+bx}} = bx\left[1-R(x)\right] \tag{2.11}$$

and

$$m(x) = -R'(x) = \frac{b\left(e^{-a+bx}\right)}{\left(1+e^{-a+bx}\right)^2} = \frac{b\left(1-R(x)\right)}{1+e^{-a+bx}} = bR(x)\left[1-R(x)\right], \tag{2.12}$$

respectively. From these equations it is clear that the price sensitivity of the logistic price response function is influenced by the parameter b, since larger values of b relate to a higher sensitivity to the price x. Furthermore, the parameter a is linked to the demand of a loan at a price of 0%, since for a price of 0% the following is true,

$$R(0) = \frac{1}{1+e^{-a+b(0)}} = \frac{1}{1+e^{-a}}.$$

The distribution of the willingness to pay, has a similar bell-shape to that of the normal distribution, but with heavy tails and is therefore called the logistic distribution. The mode of the willingness to pay distribution for the logistic price response function is at the price $x = \frac{a}{b}$, with the slope also being the steepest at this price.

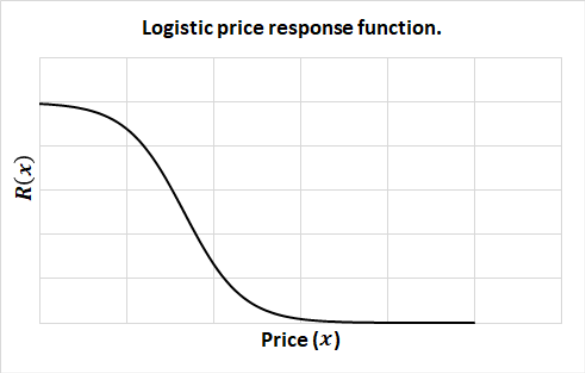


Figure 2.3: Logistic price response function.

Considering the properties described above, the logistic price response function is a more realistic price response function than the linear and constant-elasticity price response functions. A disadvantage of using the logistic price response function is that the demand never reaches 1, even if the price is 0 and the demand will also never fall to 0, even if the price is very high. However, to ensure that the demand is 1 at some price, say x_L, the translated logistic price response function (as it is referred to by Thomas, 2009) can be used, that is,

$$R(x) = \begin{cases} (1+e^{-a})\left(\frac{e^{a-b(x-x_L)}}{1+e^{a-b(x-x_L)}}\right) = \frac{e^{-b(x-x_L)}+e^{a-b(x-x_L)}}{1+e^{a-b(x-x_L)}} = \frac{1+e^{-a}}{1+e^{-a+b(x-x_L)}} & \text{for } x \geq x_L \\ 1 & \text{otherwise.} \end{cases}$$

The translated logistic price response function has the property that $R(x_L) = 1$. The shape of this price response function is similar to that of the logistic price response function seen in Figure 2.3, specifically in the region above x_L, but rescaled by a multiplicative factor. The elasticity and willingness to pay distribution corresponding to the translated logistic price response function are similar to that of the ordinary logistic price response function with the obvious conversion from 0 to x_L and with the multiplicative factor. The elasticity and willingness to pay distribution are respectively given by

$$\begin{aligned} \varepsilon(x) &= \frac{-R'(x)x}{R(x)} = \frac{b(1+e^{-a})\left(e^{-a+b(x-x_L)}\right)x}{\left(1+e^{-a+b(x-x_L)}\right)^2} \times \frac{1+e^{-a+b(x-x_L)}}{(1+e^{-a})} \\ &= \frac{bx\left(e^{-a+b(x-x_L)}\right)}{1+e^{-a+b(x-x_L)}} \\ &= bx\left[1-\frac{R(x)}{(1+e^{-a})}\right] \qquad \text{for } r \geq r_L \end{aligned} \tag{2.13}$$

and

$$m(x) = -R'(x) = \frac{b(1+e^{-a})\left(e^{-a+b(x-x_L)}\right)}{\left(1+e^{-a+b(x-x_L)}\right)^2} = bR(x)\left[1-\frac{R(x)}{(1+e^{-a})}\right]. \tag{2.14}$$

Considering these properties of the logistic price response functions *i.e.* the shape, the measures of price sensitivity (slope and elasticity) and the willingness to pay distribution, this price response function is far more realistic for modelling the demand for a loan than the other price response functions.

However, assuming the demand for a loan, $R(x)$, is only a function of the quoted price, x, suggests that the potential customers applying for a loan are homogeneous in their response to a price. Moreover, if the same price is given to low and high risk customers, the low risk customers are less likely to take up the loan compared to the high risk customers. This can be attributed to adverse selection, *i.e.* low risk customers are more price sensitive as opposed to high risk customers. Therefore, characteristics of the customer, in particular the probability of default of the customer, should also be considered when determining the take-up probability of a loan. Other characteristics of the customer to consider might include age, residential status, occupation, first secured loan etc. Apart from the characteristics of the customer that play a role in the take-up probability and price of the loan, there are also other features of a loan that could vary. Specifically, for loans, the loan amount, duration, type of interest rate (fixed/variable) etc. can vary. Consequently, the price response function is not only a function of the price of a loan, but also the characteristics of the customer and other features of the loan. Hence, there is an obvious extension of the logistic price response function introduced in (2.10) that makes it possible to relate the take-up probability not only to price, but also to the characteristics of the customer and other features of the loan.

Suppose y and z denote vectors of the characteristics of the customer and other features of the loan, respectively. Then the extension of the logistic price response function from (2.10) which incorporates the customer characteristics and the other feature of the loan is given by

$$R(x,y,z) = \frac{e^{a-bx+c\cdot y+d\cdot z}}{1+e^{a-bx+c\cdot y+d\cdot z}} = \frac{1}{1+e^{-a+bx-c\cdot y-d\cdot z}} \iff \ln\left(\frac{R(x,y,z)}{1-R(x,y,z)}\right) = a-bx+c\cdot y+d\cdot z. \tag{2.15}$$

Note that the log odds of the take-up probability versus the non take-up probability is also a linear function of the variables. Hence, by using the extended logistic price response (2.15), it is possible to model the take-up probability as a function of price, customer characteristics and features of the loan. This price response function also has the reverse S-shape, which according to the experience of Thomas (2009), is the appropriate form.

Moreover, this price response function establishes the relationship between demand, the price of a loan, customer characteristics and loan features. Skugge (2011) suggests that one of the most challenging parts of pricing outside-in, or demand-based pricing as referred to by Terblanche and De la Rey (2014), is to understand the relationship between price and demand. This extended logistic price response function establishes that relationship, but by using any form of the logistic price response function, it is clear from the reverse S-shape that an increase in the price of a loan relates to a decrease

in the take-up probability. However, in contrast to this, an increase in the price of a loan increases the profitability of the loan. Thus, the problem is to determine the price (*i.e.* interest rate) that maximise both the loan take-up probability of a potential borrower and the expected profit to the lender. Therefore, to determine the expected profit to the lender, loan profitability measures (or income functions) will now be considered.

2.4.2 The income function

Various loan profitability measures are used by lenders including but not limited to *Net Income After Cost of Capital (NIACC), Net Interest Income (NII), Present Value of the Net Interest (PVNI), Return on Equity (ROE), Risk Adjusted Return On Capital (RAROC)* etc. Phillips (2013) uses the PVNI as the appropriate measure of loan profitability, which is in fact a measure of the expected incremental contribution (after tax) of a loan, to the total profitability of the lender. The following components form part of the PVNI:

- *Net Present Interest Income (NPII).* This is the difference between the present value of the interest received during the term of the loan and the interest the lender has to pay on the principle amount *i.e.* the repurchase rate.
- *Present Value of the Expected Non-interest Income.* This income includes the fees associated with a loan.
- *Present Value of operation expenses.* This refers to the costs associated with processing payment, the mailing of statements etc.

The key component of loan profitability when using PVNI is generally NPII, since the income generated by this component is significant compared to the other two components, specifically when considering loans with large principle amounts. Furthermore, it is also reasonable to assume that the income from fees and operational costs of loans are often not dependent on the interest rate, whereas the income generated from NPII, primarily depends on the price (or interest rate). Consider a simple loan with a fixed annual price x, a principle amount (loan amount) of a and a term of n. The price, loan amount and term are considered as the features of the loan. Furthermore, assuming the repayment of the loan occurs on a monthly basis *i.e.* for a 20 year loan $n = 240$, the monthly repayment amount p is given by

$$p(a,r,n) = a\left[\frac{r(1+r)^n}{(1+r)^n - 1}\right], \tag{2.16}$$

where $r = \frac{x}{12}$ denotes the monthly interest rate. Moreover, assume the lender has to pay an annual interest rate of x^0 for the capital borrowed (*i.e.* a monthly interest rate of $r_0 = \frac{x^0}{12}$) also known as the annual repurchase rate or the cost of funding the loan and suppose the internal monthly discount rate

is given by r_d *i.e.* the rate at which the feature payments are discounted. Then, the net present interest income (NPII) for a loan that is guaranteed to go the full term (a risk-less loan) is given by

$$I^m(r, r_0, r_d, n, a) = \sum_{t=1}^{n} a \left(\frac{1}{(1+r_d)^t} \right) \left(\frac{r(1+r)^n}{(1+r)^n - 1} - \frac{r_0(1+r_0)^n}{(1+r_0)^n - 1} \right), \tag{2.17}$$

where the superscript m in $I^m(r, r_0, r_d, n, a)$ stands for maturity, indicating that the loan is certain to go the full term. From (2.17) NPII is simply the sum of the difference between the monthly repayment amount of the borrower and the lender discounted by the internal discount rate.

With most loans, the lender faces the risk (at the time of funding the loan) that the borrower may default at some point during the term of the loan, *i.e.* stop making the monthly repayments. Hence, the probability that a borrower defaults at some point during the term of the loan, should be taken into account when determining the NPII. Let p_t denote the probability that a borrower defaults in month t, then the probability that the borrower will not default (*i.e.* survive or make the repayment) from $t-1$ to t is given by $1-p_t$. Then, the probability that a borrower will not default before month t, that is, the probability that the borrower will make the payment in month t, is given by

$$s_t = \prod_{j=1}^{t} (1 - p_j), \;\; j = 1, 2, \ldots, n. \tag{2.18}$$

By taking into account the probability that a borrower will make payment t, the expected NPII is

$$I(r, r_0, r_d, n, a, s_t) = \sum_{t=1}^{n} a \left(\frac{1}{(1+r_d)^t} \right) \left(\frac{s_t r(1+r)^n}{(1+r)^n - 1} - \frac{r_0(1+r_0)^n}{(1+r_0)^n - 1} \right). \tag{2.19}$$

This is a more realistic representation of NPII, since the lender still has to repay the outstanding capital even if the borrower does in fact default sometime during the term of the loan. Consequently, by taking into account the probability of default of a borrower, the expected NPII can be negative for borrowers that are considered to be high risk. Phillips (2013) suggests that a lender who wishes to maximise expected profits, should consider to either raise the price of the loan, or not extend credit to these type of customers, as there might not be a price at which the loan is expected to be profitable. Phillips (2013) further considers an approximation of the NPII where the following assumptions are made

- $r_d \approx 0$, since the discount rate is being applied on top of the repurchase rate r_0.
- $\frac{(1+r)^n}{(1+r)^n - 1} \approx 1$ and $\frac{(1+r_0)^n}{(1+r_0)^n - 1} \approx 1$ for large values of n.

Thus, by taking these assumptions into account, the expected NPII in (2.19) can be approximated by

(see Phillips, 2013)

$$\begin{aligned} I(r,r_0,r_d,n,a,s_t) &= \sum_{t=1}^{n} a\left(\frac{1}{(1+r_d)^t}\right)\left(\frac{s_t r(1+r)^n}{(1+r)^n-1}-\frac{r_0(1+r_0)^n}{(1+r_0)^n-1}\right) \\ &\approx \sum_{t=1}^{n} a(s_t r - r_0) \\ &= a\left[\sum_{t=1}^{n} s_t r - n r_0\right] \\ &= an(r-r_0) - ar\left[\sum_{t=1}^{n}(1-s_t)\right] \\ &= an(r-r_0) - (1-s_n)ar\left[\sum_{t=1}^{n}\frac{(1-s_t)}{(1-s_n)}\right] \\ &= an(r-r_0) - pa\delta \\ &= an(\frac{x}{12}-\frac{x^0}{12}) - pa\delta, \end{aligned} \tag{2.20}$$

where p denotes the probability of default and δ the loss given default (LGD) *i.e.* the percentage of the loan amount the lender loses when a borrower defaults. This approximation can be decomposed into two parts, the first part given by $an(\frac{x}{12}-\frac{x^0}{12})$ which denotes the approximate income generated by the loan, and the second part given by $pa\delta$ which denotes the expected loss (or cost of risk) as a result of default. The approximation of the expected NPII in (2.20) has the advantage of being linear with respect to the price, x, charged for the loan. In addition to this, the approximation requires only the overall probability of default p instead of the probability of default for each month $p_t, t = 1,2,\ldots,n$ (Phillips, 2013). Given these advantages, (2.20) is commonly used to approximate the expected NPII and often used as measure of loan profitability. Terblanche and De la Rey (2014) generalise the approximation in (2.20) by adopting a customer segmentation approach. This is done by segmenting customers based on their credit score (*i.e.* risk profile), loan amounts and loan terms and subsequently determining the annual price per customer segment. Denote the index set of all customer segments by $\mathscr{C} = \{1,2,...,C\}$ and let x_c represent the mean annual price for a segment $c \in \mathscr{C}$. The approximation of the expected NPII in (2.20) for a segment $c \in \mathscr{C}$ is then given by

$$I(x_c,x^0,a_c,n_c,p_c) = q_c\left[a_c n_c(\frac{x_c}{12}-\frac{x^0}{12}) - p_c a_c \delta\right], \tag{2.21}$$

where q_c denotes the number of customers in the segment and a_c, n_c and p_c denote the mean loan amount, mean loan term and mean probability of default, respectively. However, using only (2.21) as a measure of loan profitability when determining the price of a loan for a segment is unrealistic. This is due to the assumption made that the loan take-up probability for the customers in a segment

with a quoted price x_c is 100%. Moreover, the profit generated by a loan is in fact conditional on whether or not the customer accepts the quoted price for the loan *i.e.* take-up the loan at the quoted price. To relate the take-up probability to the quoted price, a price response function can be fitted to the data. That is, the take-up probability of a loan as a function of the quoted price can be mathematically captured by the price response function *i.e.* a logistic regression model (Terblanche and De la Rey, 2014). Phillips (2013) suggests fitting a logistic regression model to each customer segment $c \in \mathscr{C}$ whereas Terblanche and De la Rey (2014) consider the case where a single logistic regression model is fitted to the segment averages, due to poor predictive power in certain customer segments. Fitting a single logistic regression model requires the use of segment averages as the input variables for the model *i.e.* a_c, n_c and p_c for each segment $c \in \mathscr{C}$. Consequently, the target variable modelled is given by $Y_c = \sum_{j \in \mathscr{G}(c)} \frac{Y_j}{q_c}$, where $\mathscr{G}(c)$ denotes the set of customers belonging to the customer segment $c \in \mathscr{C}$. The average take-up probability for a customer segment $c \in \mathscr{C}$, obtained by fitting a single logistic regression model to segment averages as inputs and the target variable Y_c, is given by

$$R(x_c, x^0, a_c, n_c, p_c) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_c + \beta_2 x^0 + \beta_3 a_c + \beta_4 n_c + \beta_5 p_c)}}. \tag{2.22}$$

This price response function exhibits the properties of an appropriate response function when used to model the take-up probability. Figure 2.4 displays the required reverse S-shape of the price response function, indicating the inverse relationship between price and take-up probability where an increase in price relates to a decrease in take-up probability (price elasticity). In addition to this, Figure 2.4 illustrates the existence of adverse selection captured by the price response functions, as low risk customer (gray line) are more sensitive to a change in price than high risk customers (black line) as discussed in Terblanche and De la Rey (2014).

Figure 2.4: Logistic price response function for high risk (black line) and low risk (grey line) customers.

This price response function is similar to the extended logistic price response function given by (2.15), as the take-up probability is modelled as a function of price, customer characteristics and features of the loan. Hence, given the price response function in (2.22) and the income function in (2.21), the credit price optimisation problem can be formulated.

2.4.3 The credit price optimisation problem

Phillips (2013) considers the price optimisation problem as determining the price to quote each customer segment in order to maximise the total expected profit given by the product of the NPII given by (2.19) and the demand for the loans obtained by fitting a different logistic model for each customer segment. Phillips (2013) discusses various aspects of the suggested approach such as using the non-linear income function (2.19), the uncertainty of the return generated by the loan and the interaction between risk and pricing in the context of the credit price optimisation problem. Terblanche and De la Rey (2014) follow a similar approach, but consider the use of a single logistic regression model as given by (2.22) and the linear form of the NPII given by (2.21). The objective of the credit price optimisation problem is to determine what price to charge each customer segment such that the total expected profit of the lender is a maximum. Hence, the credit price optimisation problem is to maximise the expected value of the approximated NPII given by the product of (2.21) and (2.22) , defined as

$$\max_{x_c \geq 0} \sum_{c \in \mathscr{C}} R(x_c, x^0, a_c, n_c, p_c) I(x_c, x^0, a_c, n_c, p_c), \tag{2.23}$$

where x_c denotes the price quoted for each customer segment $c \in \mathscr{C}$ (see Terblanche and De la Rey, 2014). The credit price optimisation problem in (2.23) is considered an unconstrained problem and can be solved using standard non-linear programming methods. However, constraints can be imposed on the risk distribution to restrict the take-up proportion of a specific risk category. In addition to this, lenders generally also want to limit the proportion of loans with a high loan-to-value (LTV), to a small proportion of the portfolio (see Caufield, 2012) since loans with a higher LTV are considered more risky as apposed to loans with lower LTVs (Phillips, 2013).

Terblanche and De la Rey (2014) incorporate constraints on the risk distribution to limit the take-up proportion of certain risk categories. A concave price response function is also proposed by Terblanche and De la Rey (2014) to overcome the non-convex solution space obtained by the imposed risk distribution constraints, specifically when using the standard logistic price response function. Furthermore, a linear approximation approach is implemented to solve the credit price optimisation problem in (2.23) by using standard linear programming methods. In addition to this, Terblanche and De la Rey (2014) implement a two-stage stochastic linear programming approach to incorporate the uncertainty of future price sensitivity into the credit price optimisation problem. Terblanche and De la Rey (2014) find that the use of a credit price optimisation approach indicate there is a loss in expected profit as a result of opportunities not taken. In addition to this, Terblanche and De la Rey (2014) find that constraints on the risk distribution may be violated due to a change in future price sensitivity *i.e.* more high risk customers are expected to take-up a loan at a higher price than lower risk customers (also known as adverse selection).

The credit price optimisation approach considered by Terblanche and De la Rey (2014) and Phillips (2013) enables the lender to take into account the customer's willingness to pay for a loan (take-

up probability) when determining the optimal price to quote. The take-up probability of a potential customer is modelled using a response model, also known as the price response function. Furthermore, by incorporating the take-up probability in credit pricing, price sensitivity and adverse selection is taken into account when determining the optimal prices that maximise the expected net present interest income to the lender. Compared to other industries, price optimisation is a relatively new approach to credit pricing, with only a limited number of lenders adopting this approach. Hence, there are still various challenges that need to be addressed such as using a non-linear income function, the interaction between pricing and risk, the modelling of risk in a more forward looking perspective (see Caufield, 2012) *etc*. These challenges contribute to the complexity of the credit price optimisation problem. In addition to these challenges, various optimisation methods can be used to solve the credit price optimisation problem *i.e.* non-linear programming methods and linear programming methods. In the next chapter, mathematical optimisation and methods used for solving linear and non-linear programming problems will be discussed.

Chapter 3

Mathematical optimisation

3.1 Introduction to optimisation

Mathematical optimisation, or mathematical programming, as it is also known, is a powerful prescriptive analytics tool that enables individuals or businesses to solve complex real-world problems whilst, making better use of available data and resources. Furthermore, mathematical programming also implies the use of mathematical models, more specifically optimisation models, to assist with strategic decisions.

The process of optimisation does not necessarily result in an optimal solution. The application of heuristics is also considered to be optimisation. Optimisation is a branch of applied mathematics and Operations Research and involves selecting the best or optimal solution. The importance of optimisation is evident from both the wide variety of fields it is applied to, as well as the availability of efficient algorithms for solving those optimisation problems. Over the past decades, intense and innovative research has been done for many classes of optimisation problems, hence the availability of reliable and fast algorithms for solving these problems. Many industries, including the financial industry, use optimisation as an effective management and decision support tool.

Mathematically, optimisation refers to the maximisation (or minimisation) of a certain objective function with several decision variables (unknown variables), while satisfying functional constraints. Typical optimisation models determine the optimal allocation of scarce resources amongst potential alternative uses with the goal to maximise the objective function, such as the total profit (Cornuejols and Tütüncü, 2006). The three essential elements of any optimisation problem are the objective function, the decision variables and the constraints. Optimisation problems with no constraints are called unconstrained optimisation problems, whereas those with constraints are called constrained optimisation problems. Furthermore, optimisation problems that have no objective function are called feasibility problems, whereas other optimisation problems might have multiple objective functions, often referred to as multi-objective optimisation problems. The latter are often solved by reducing the optimisation problem to a single-objective function optimisation problem or a sequence of single-objective function

problems. Optimisation problems where the decision variables are restricted to a discrete set of possible values, or integer values, are called discrete or integer optimisation problems. An optimisation problem is said to be continuous if there are no such restrictions on the decision variables and they may take on a real value. In addition to integer and continuous optimisation problems, there exists some problems that may have a mixture of integer and continuous decision variables, called mixed integer optimisation problems. Considering the different types of optimisation problems that exist, researchers have, over the years, developed problem specific optimisation algorithms, instead of general purpose optimisation algorithms.

The remainder of this chapter is organised as follows: in Section 3.2 the theory behind mathematical optimisation will be considered by defining various concepts related to the formulation of optimisation problems and the solving thereof. In Sections 3.3 and 3.4 linear programming (LP) problems and mixed integer linear programming (MILP) problems are discussed together with the methods and algorithms used to solve these problems. This chapter concludes in Section 3.5 with a discussion of non-linear programming (NLP) problems as well as how these problems can be solved using a piece-wise linear approximation approach.

3.2 Optimisation theory

Mathematically, optimisation refers to the maximisation (or minimisation) of an objective function, subject to (s.t.) constraints on the decision variables. Consider the function $f(x) : \mathbb{R}^n \to \mathbb{R}$ and a set $S \subset \mathbb{R}^n$, then a generic optimisation problem can be defined as the problem of finding a solution $x^* \in \mathbb{R}^n$ that solves

$$\begin{aligned} &\min f(x) \\ &\text{s.t. } x \in S. \end{aligned} \tag{3.1}$$

Here, f refers to the objective function or cost function whereas S refers to the feasible region and x denotes the decision variable(s). Note that, if the region S is empty, the optimisation problem is called infeasible and if it is possible to find a sequence x^k such that

$$f(x^k) \to -\infty \text{ if } k \to +\infty, \tag{3.2}$$

the optimisation problem is said to be unbounded. However, if an optimisation problem is neither unbounded nor infeasible, then there exists a solution $x^* \in S$ for which

$$f(x^*) \leq f(x), \ \forall\, x \in S.$$

In optimisation, it is important to differentiate between a solution $x^* \in S$ being a global or local solution. Consider the following definitions from Cornuejols and Tütüncü (2006), on global and local minimisers.

Definition 3.1. Let $f(x) : \mathbb{R}^n \to \mathbb{R}$ denote a function with $x \in S \subset \mathbb{R}^n$. The solution x^* for the optimisation problem $\min f(x)$, as defined in (3.1), is called a global minimiser if

$$f(x^*) \leq f(x), \ \forall\, x \in S. \tag{3.3}$$

Furthermore, $x^*$is called a strict global minimiser if it satisfies

$$f(x^*) < f(x), \ \forall\, x \in S. \tag{3.4}$$

Definition 3.2. Consider the function $f(x) : \mathbb{R}^n \to \mathbb{R}$ with $x \in S \subset \mathbb{R}^n$. The solution $x^*$for the optimisation problem $\min\ f(x)$ defined in (3.1), is called a local minimiser if

$$f(x^*) \leq f(x), \ \ \forall\, x \in S \cap A_{x^*}(\varepsilon), \tag{3.5}$$

where $A_{x^*}(\varepsilon)$ is an open ball centered at x^* with a radius $\varepsilon > 0$, that is

$$A_{x^*}(\varepsilon) = \{x : \|x - x^*\| < \varepsilon\},$$

with $\|\cdot\|$ denoting the euclidean norm. Note that, although only a minimisation problem was considered above, any maximisation problem can be rewritten as a minimisation problem by simply multiplying the objective function with a negative constant. However, if considering maximisation problems as apposed to minimisation problems, the converse of (3.2)–(3.5) should hold. For example, if it is possible to find a sequence x^k such that

$$f(x^k) \to \infty \ \text{ if } \ k \to +\infty,$$

the optimisation problem is unbounded and $x^* \in S$ is called a global maximiser if

$$f(x^*) \geq f(x), \ \forall x \in S.$$

Furthermore, for optimisation problems, the feasible set S can explicitly be described using the functional constraints, which in turn can be equality and inequality constraints. Therefore, the feasible set S is given by

$$S := \{x : h_i(x) = 0, \ i \in \mathscr{I} \text{ and } h_i(x) \geq 0, \ i \in \mathscr{E}\},$$

where $\mathscr{I}$ *and* $\mathscr{E}$ denote the index sets for the equality and inequality constraints, respectively. The generic optimisation problem as defined in (3.1) can now be rewritten to incorporate the constraints, taking on the form

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & h_i(x) = 0 \quad i \in \mathscr{I}, \\ & h_i(x) \geq 0 \quad i \in \mathscr{E}. \end{aligned} \tag{3.6}$$

There are various factors that can influence whether optimisation problems can be solved in an efficient manner. Typical predictors include the number of decision variables, m, as well as the total number of constraints $|\mathscr{I}| + |\mathscr{E}|$. Cornuejols and Tütüncü (2006) note that additional factors that could add to the difficulty in efficiently solving these optimisation problems include the form of the objective and constraint functions, f and g_i or restricting some of the decision variables to integers values.

Optimisation problems involving linear objective functions and linear constraints functions are easier to solve compared to problems with convex objective functions and convex feasible sets. These concepts will now be defined in the following 4 definitions from Boyd and Vandenberghe (2004).

Definition 3.3. A function $f(x) : \mathbb{R}^n \to \mathbb{R}$ is said to be linear if for all $x,\ y \in \mathbb{R}^n$ and all $\alpha,\ \beta \in \mathbb{R}$ the following equation holds:

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y). \tag{3.7}$$

However, if the function is not linear *i.e.* does not satisfy (3.7), the function is called a non-linear function .

Definition 3.4. A function $f(x) : \mathbb{R}^n \to \mathbb{R}$ is said to be a convex function if for all $x_1,\ x_2 \in \mathbb{R}^n$ and all $\alpha,\ \beta \in \mathbb{R}$ with $\alpha + \beta = 1,\ \alpha \geq 0,\ \beta \geq 0$ the following inequality is satisfied:

$$f(\alpha x_1 + \beta x_2) \leq \alpha f(x_1) + \beta f(x_2). \tag{3.8}$$

Alternatively, if the function $-f$ is convex or if the opposite of (3.8) is true, the function is concave, that is,

$$f(\alpha x_1 + \beta x_2) \geq \alpha f(x_1) + \beta f(x_2). \tag{3.9}$$

Figure 3.1 illustrate the difference between a convex and concave function as given in definition (3.8) and (3.9).

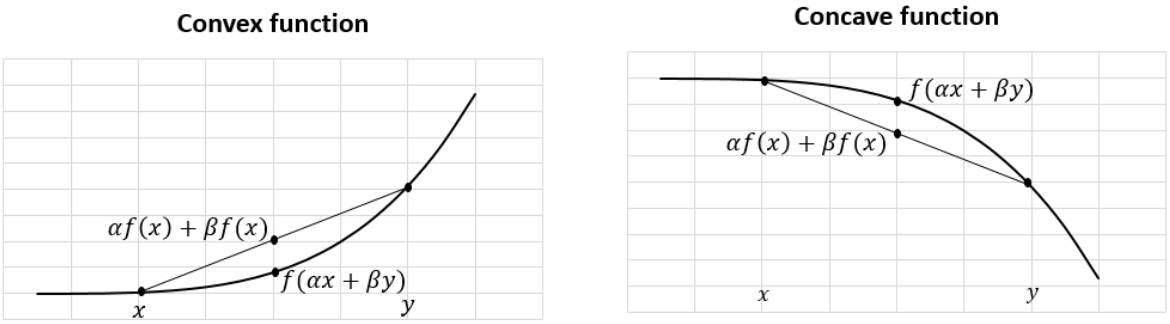
Definition 3.5. A set S is convex if, for any points $x_1,\ x_2 \in S \subset \mathbb{R}^n$, and any $\lambda \in [0,1]$, the following holds,

$$\lambda x_1 + (1 - \lambda) x_2 \in S. \tag{3.10}$$

That is, if the line segment between any two points in the set S, lies within S. Figure 3.2 (a), (b) and (c) illustrate the basic idea of convex and non-convex sets in $\mathbb{R}^2$ with Figure 3.2 (c) also illustrating a set consisting of discrete points.

Definition 3.6. A convex combination of n points, say $x_1, x_2, \ldots, x \in \mathbb{R}^n$, with $\lambda_i \geq 0,\ i = 1, 2, \ldots, n$ and $\sum_{i=1}^{n} \lambda_i = 1$ is given by

$$\sum_{i=1}^{n} \lambda_i x_i. \tag{3.11}$$



(a) (b)

Figure 3.1: (a) Convex function; (b) Concave function.

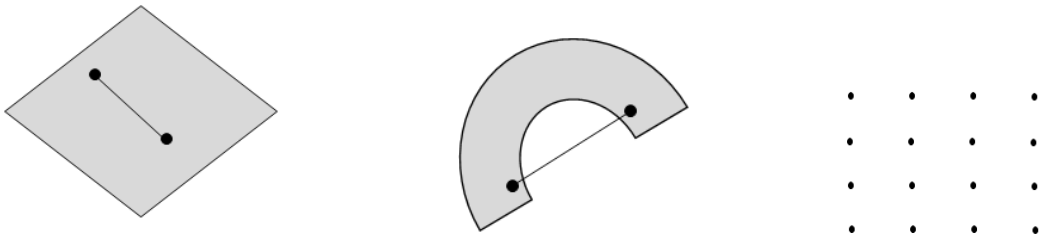
A convex combination of n points should therefore be considered as a mixture or a weighted average of the points x_i, with λ_i representing the weights. Furthermore, the set containing every convex combination of points in a set is called the convex hull of a set. The convex hull of the set of discrete points in Figure 3.2 (c) can be seen in Figure 3.3, with the convex hull comprising of the square with the dark boundary line as well as the shaded part. Formally, the convex hull of a set can be defined as (Boyd and Vandenberghe, 2004):

Definition 3.7. The set of all convex combinations of points in S is called the convex hull of S (denoted *convS*) and is given by

$$convS = \left\{ \sum_{i=1}^{n} \lambda_i x_i | x_i \in S, \lambda_i \geq 0, i = 1, 2, \ldots, n, \sum_{i=1}^{n} \lambda_i = 1 \right\}.$$

In addition to this, a convex hull is always convex, as its name suggests. This statement is formulated in the following theorem, of which the proof can be found in Chapter 3 of Leonard and Lewis (2015).

Theorem 3.8. *Consider the convex combination of points $x_1, x_2, \ldots, x_n \in \mathbb{R}^n$ given by $\sum_{i=1}^{n} \lambda_i x_i$ where $\sum_{i=1}^{n} \lambda_i = 1$. Then, the set S is a convex set if and only if it contains every convex combination of its points, that is, the convex hull of S is convex.*



(a) (b) (c)

Figure 3.2: (a) Convex set; (b) Non-convex set; (c) Non-convex set.

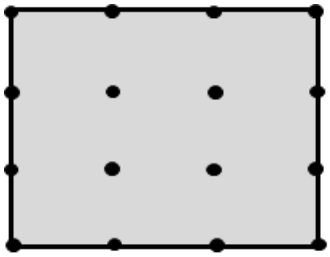


Figure 3.3: Convex hull.

Various types of optimisation problems exists, specifically when considering the different forms of the objective and constraint functions. Consequently, researchers have developed problem-specific optimisation algorithms to accommodate the different types of optimisation problems, rather than using generic algorithms to solve these optimisation problems. Consider below the different types of optimisation problems, together with the methods and algorithms used to find the optimal solution for these problems.

3.3 Linear programming (LP) problems

Optimisation problems are generally classified by the form of the objective function and constraint functions. The most simple and most common problems are linear optimisation problems or linear programming (LP) problems. In LP problems, both the objective function and constraint (equality and inequality) functions are linear, *i.e.* functions f and h_i in (3.6) are all linear.

A generic form of the LP optimisation problem can be written as follows

$$\begin{aligned} \max \quad & c^T x \\ \text{s.t.} \quad & a_i^T x = b_i \quad i \in \mathscr{I}, \\ & a_i^T x \leq b_i \quad i \in \mathscr{E}, \end{aligned} \tag{3.12}$$

where $\mathscr{I}$ and $\mathscr{E}$ denote the index sets for the equality and inequality constraints, respectively, $c \in$

$\mathbb{R}^n$ denotes the cost coefficient vector of the objective function and $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}^n$ denote the constraint coefficient and constraint requirements vectors for the index sets $\mathscr{I}$ and $\mathscr{E}$, respectively. The decision variables (unknown variables to be determined) are given by the vector $x \in \mathbb{R}^n$. Note that the $(1 \times n)$ vector c^T is the transpose of the $(n \times 1)$ vector c.

In order to solve LP problems using algorithmic approaches, it is often necessary to convert the LP problems to a standard form. The standard form of an LP problem can be formulated in matrix notation as follows,

$$\begin{aligned} \max \quad & c^T x \\ \text{s.t.} \quad & Ax = b, \\ & x \geq 0. \end{aligned} \tag{3.13}$$

Here, $c \in \mathbb{R}^n$ denotes the coefficient vector of the objective function, $A \in \mathbb{R}^{m \times n}$ the matrix of constraint coefficients and $b \in \mathbb{R}^m$ the vector of constraint requirements. Furthermore, $x \in \mathbb{R}^n$denotes the vector of decision variables to be determined. Consider the following example of an LP problem that is not in the standard form:

$$\begin{array}{llll} \min & -c_1 x_1 & -c_2 x_2 & \\ \text{s.t.} & a_{11} x_1 & +a_{12} x_2 & \leq b_1 \\ & a_{21} x_1 & +a_{22} x_2 & \leq b_2 \\ & x_1 \geq 0, & x_2 \geq 0. & \end{array} \tag{3.14}$$

To convert this LP problem to the standard form in (3.13), the following transformations can be done:

- for minimisation problems, multiply the objective function with a negative constant (*e.g.* -1)
- for less than or equal constraints add non-negative slack variables,
- for larger than or equal constraints subtract non-negative surplus variables,
- for a variable x unrestricted in sign, create variables $x^+ \geq 0$ and $x^- \geq 0$ such that $x = x^+ - x^-$.

Therefore, the LP in (3.14) converted to the standard form in (3.13), using the above transformations, is given by

$$\begin{array}{llllll} \max & c_1 x_1 & +c_2 x_2 & & & \\ \text{s.t.} & a_{11} x_1 & +a_{12} x_2 & +x_3 & & = b_1 \\ & a_{21} x_1 & +a_{22} x_2 & & +x_4 & = b_2 \\ & x_1 \geq 0, & x_2 \geq 0, & x_3 \geq 0, & x_4 \geq 0. & \end{array}$$

Consequently, the standard LP form is not restrictive and can be obtained through simple transformations. In the following section a method for solving these type of LP problems are considered.

3.3.1 The simplex method

In mathematical programming, the simplex method, as developed by the mathematician George Dantzig in 1948 (Dantzig, 1948), is an algorithm that is used for solving linear optimisation problems or LP problems. The simplex method is considered the most successful and well known method for solving LPs and consequently a key contribution within mathematical programming (see Bixby, 2012). According to Dantzig (1990), the simplex method is the historical reason for the success in the field of linear programming and has been the preferred method for solving LPs for many years. In LP theory, an important result is that, when an LP problem has an optimal solution on the feasible set, it must have an optimal solution that is an extreme point. The extreme points of a feasible set $\{x : Ax \leq b,\ x \geq 0\}$, refer to those points that cannot be expressed as a convex combination of other points in the set.

Furthermore, since the goal is to find an optimal solution for an LP problem, the focus is shifted to the value of the objective function at the extreme points only. Therefore, the search space is reduced from an infinite to finite one since there are only a finite number of extreme points. Hence, the basic idea behind the simplex method, is to move from one extreme point to an adjacent extreme point to find the optimal solution for an LP problem.

To explain the simplex method used to find an optimal solution to an LP problem, consider the following generic LP problem

$$\begin{aligned} \max \quad & cx \\ \text{s.t.} \quad & Ax \leq b, \\ & x \geq 0, \end{aligned} \tag{3.15}$$

where c denotes an n-dimensional row vector (for ease of notation and derivation, c denotes a row vector instead of c^T), x denotes an n-dimensional column vector of decision variables (unknown variables to be determined), A denotes a $m \times n$ matrix of constraint coefficients and b denotes the m-dimensional column vector of constraint requirements. The above mentioned vectors and matrix can be represented in the following manner

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

$$c = \begin{bmatrix} c_1 & c_2 & \cdots & c_n \end{bmatrix}, \text{and } x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

To transform the LP problem in (3.15) to the standard form given in (3.13), introduce m slack variables into the LP problem. Note that the number of slack variables introduced should be the same as the

number of constraints in the LP problem. Therefore, let x_s denote the m-dimensional column vector containing the slack variables introduced into the LP problem in (3.15). That is,

$$x_s = \begin{bmatrix} x_{n+1} \\ x_{n+2} \\ \vdots \\ x_{n+m} \end{bmatrix}.$$

Furthermore, let I denote the $m \times m$ identity matrix given by

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Then, the constraints can be written in augmented form as

$$[A, I] \begin{bmatrix} x \\ x_s \end{bmatrix} = b, \begin{bmatrix} x \\ x_s \end{bmatrix} \geq 0. \tag{3.16}$$

Note that there are many potential solutions to the system given in (3.16). To obtain a unique solution to the system of equations, a square system of equations are required, *i.e.* the same number of variables as equations are required. Moreover, by choosing the decision variable vector to be zero, $x = 0$, a square system of equations is obtained, resulting in a solution $x_s = b$, which does not necessarily satisfy all the inequalities. If all the constraints in (3.15) are $\leq b$, with $b \geq 0$, then it should satisfy all constraints. Otherwise a two-stage approach is followed to find an initial feasible solution (Stojković et al., 2009). This is, however, a good starting point for the simplex algorithm as will be seen at a later stage. Now, consider the partition of the augmented matrix $[A, I]$ such that

$$[A, I] \equiv [B, N], \tag{3.17}$$

where the matrix B denotes a $m \times m$ square matrix that consists of linear independent columns of the matrix $[A, I]$ and are also the coefficients of the basic variables. B is referred to as the basis matrix and this partition is known as the basis partition. The $m \times n$ matrix N denotes the coefficients of the non basic variables. In a similar way, the vector of variables $\begin{bmatrix} x \\ x_s \end{bmatrix}$ can be partitioned such that,

$$\begin{bmatrix} x \\ x_s \end{bmatrix} \equiv \begin{bmatrix} x_B \\ x_N \end{bmatrix}, \tag{3.18}$$

where x_B denotes the basic variables (corresponding to the basis matrix B) and x_N the non basic variables (corresponding to the matrix N). Hence, the partitions (3.17) and (3.18) are substituted into the equality constraint in (3.16), which then becomes

$$\begin{aligned}[A,I]\begin{bmatrix} x \\ x_s \end{bmatrix} &= [B,N]\begin{bmatrix} x_B \\ x_N \end{bmatrix} \\ &= Bx_B + Nx_N = b. \end{aligned} \tag{3.19}$$

Multiplying both sides of equation (3.19) by the inverse of B, B^{-1}, and rewriting the equation in terms of the basic variables x_B yields the following expression for x_B

$$x_B = B^{-1}b - B^{-1}Nx_N. \tag{3.20}$$

Recall that, by choosing $x = x_N = 0$**,** an initial solution for x_B in (3.20) is obtained, that is,

$$x_B = B^{-1}b - B^{-1}N0 = B^{-1}b. \tag{3.21}$$

The vector x_N can thus be seen as the independent variables that can be chosen freely (*i.e.* chosen to be 0), whereafter the vector of dependent variables x_B can be determined uniquely using the chosen values of x_N. The solution for (3.20) is called a basic solution if it has the following form

$$x_N = 0,\ x_B = B^{-1}b. \tag{3.22}$$

In addition to this, if $x_B = B^{-1}b \geq 0$ in (3.22), then the solution $x_N = 0$, $x_B = B^{-1}b$ is called a basic feasible solution (BFS) for the LP problem described in (3.15). Geometrically, basic feasible solutions for the LP problem coincide with extreme points of the feasible set of the LP problem, given by $\{x : Ax \leq b, x \geq 0\}$. Recall that the extreme points of a set refer to those points that cannot be written as a convex combination of another two points in the feasible set.

Now, consider the objective function in (3.15) and denote it by $Z = cx$. By partitioning the cost vector $c = [c_B, c_N]$, the objective function is written as

$$\begin{aligned} &Z = cx \Longleftrightarrow Z - cx = 0, \\ &Z - [c_B, c_N]\begin{bmatrix} x_B \\ x_N \end{bmatrix} = 0, \\ &Z - c_B x_B - c_N x_N = 0. \end{aligned} \tag{3.23}$$

Then, substituting the expression for x_B in (3.20) into (3.23), it follows that

$$\begin{aligned} Z - c_B(B^{-1}b - B^{-1}Nx_N) - c_N x_N &= 0, \\ Z &= c_B B^{-1}b + (c_N - c_B B^{-1}N)x_N, \\ &= z_0 + (c_N - c_B B^{-1}N)x_N, \end{aligned} \tag{3.24}$$

where $z_0 = c_B B^{-1}b$ denotes the value of the objective function of the current basis. Note that (3.24) does not contain the current basic variables and therefore allows us to determine the net effect, of changing the value of a non basic variable x_N, on the objective function. The coefficient vector $(c_N - B^{-1}N)$ of the non basic variables x_N in (3.24) is called the reduced costs, since it contains the cost coefficients, c_N, reduced by $c_B B^{-1}N$, which is the cross effect of the basic variables. Now, by using the equations derived above, the simplex algorithm is formulated.

3.3.2 The simplex algorithm

Recall that, if a LP problem has an optimal solution, the optimal solution must be at an extreme point. The simplex method algorithm therefore solves a LP problem by moving from one extreme point to the next, more specifically, to an adjacent extreme point until the optimal solution is found. Note that the extreme points correspond to basic feasible solutions on the feasible set and therefore the simplex method in effect moves from one basic feasible solution to the next. With that in mind, the simplex algorithm can be summarised for problems like those given in (3.15) by the following steps (Cornuejols and Tütüncü, 2006 and Bazaraa et al., 2011).

1. Choose/ update the basic and non basic variables *i.e.*

$$\begin{bmatrix} x \\ x_s \end{bmatrix} \equiv \begin{bmatrix} x_B \\ x_N \end{bmatrix} \tag{3.25}$$

and

$$[A, I] \equiv [B, N]\,. \tag{3.26}$$

2. Substitute the partitions in (3.25) and (3.26) into (3.20) and choose $x_N = 0$, then a basic feasible solution is obtained,

$$\begin{aligned} x_B &= B^{-1}b - B^{-1}Nx_N \\ &= B^{-1}b - B^{-1}N0 \\ &= B^{-1}b\,. \end{aligned} \tag{3.27}$$

Recall that, if $x_B = B^{-1}b \geq 0$, then x_B is a basic feasible solution of the LP problem.

3. Determine which non basic variable should enter the basis, *i.e.* the variable for which the reduced cost is a maximum positive value. Thus, the non basic variable from the vector x_N corresponding to

$$\max(c_N - B^{-1}N), \text{ with } c_N - B^{-1}N > 0$$

should enter the basis. Denote this non basic variable to enter the basis by x_{N^*}. Increasing this non basic variable would increase the objective function given in (3.24), since for this variable $c_N - B^{-1}N > 0$ and therefore $Z > z_0$. If this is the case, continue to step 4. However, if $c_N - B^{-1}N \leq 0$, the current basic feasible solution x_B, $x_N = 0$ is an optimal solution, since from (3.24) the objective function is given by

$$\begin{aligned} Z &= z_0 + (c_N - c_B B^{-1}N)x_N, \\ &= z_0 + (c_N - c_B B^{-1}N)0, \\ &= z_0, \end{aligned}$$

and therefore by increasing $x_N \geq 0$ when $c_N - B^{-1}N \leq 0$, will imply that:

$$Z \leq z_0.$$

Thus, the algorithm stops when the reduced costs is less than or equal to zero *i.e.* when $c_N - B^{-1}N \leq 0$ with the optimal solution then being the current basic feasible solution $x_N = 0$, $x_B = B^{-1}b \geq 0$ obtained in step 1.

4. Determine which of the basic variables should exit the basis. Recall from (3.20) that

$$x_B = B^{-1}b - B^{-1}Nx_N, \tag{3.28}$$

and consequently, by increasing the term $B^{-1}Nx_N$ in (3.28) the values of the basic variables are reduced. Here, $x_B = B^{-1}b$ denotes a $m \times 1$ vector of values say

$$x_B = \begin{bmatrix} (B^{-1}b)_1 \\ (B^{-1}b)_2 \\ \vdots \\ (B^{-1}b)_m \end{bmatrix}$$

when $x_N = 0$. However, since $x_{N^*} \in x_N$ will enter the basis, the basic variables that will drop to zero first when the value of x_{N^*} is increased (*i.e.* no longer zero but rather increased to say 1)

should be identified, that is,

$$x_N = \begin{bmatrix} (x_N)_1 \\ (x_N)_2 \\ \vdots \\ (x_{N^*})_s \\ \vdots \\ (x_N)_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

where $s \leq n$. Recall that the $m \times n$ matrix N contains the coefficients of the non basic variables, x_N, and is given by

$$N = \begin{bmatrix} (N)_{11} & (N)_{12} & \cdots (N)_{1s} & \cdots & (N)_{1n} \\ (N)_{21} & (N)_{22} & \cdots (N)_{2s} & \cdots & (N)_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (N)_{m1} & (N)_{m2} & \cdots (N)_{ms} & \cdots & (N)_{mn} \end{bmatrix}.$$

Thus, the term Nx_N is now simply a $m \times 1$ vector containing the coefficients of the non basic variable to enter the basis say N^*given by

$$N^* = \begin{bmatrix} (N)_{1s} \\ (N)_{2s} \\ \vdots \\ (N)_{ms} \end{bmatrix}.$$

Now, by considering the ratio of the right hand side terms in (3.28), the first basic variable to drop to zero is identified. The basic variable for which the ratio in (3.29) is a minimum positive value, will be the variable to exit the basis. This ratio is given by

$$\min\left\{\frac{B^{-1}b}{B^{-1}Nx_N}\right\} = \min\left\{\frac{x_B}{N^*}\right\} = \min\left\{\frac{(B^{-1}b)_1}{(N)_{1s}}, \frac{(B^{-1}b)_2}{(N)_{2s}}, \ldots, \frac{(B^{-1}b)_m}{(N)_{ms}}\right\}. \tag{3.29}$$

Note that, only the positive ratios are considered since the negative ratios suggest that by increasing the entering variable the corresponding basic variables will increase as well, subsequently not forcing those basic variables to zero.

5. Update the basic and non basic variables and repeat steps 1–4 until the reduced cost in step 2 is less than or equal to zero, i.e until $c_N - B^{-1}N \leq 0$, with $x_B \geq 0,\ x_N = 0$ at the time being the optimal solution of the LP problem.

As mentioned previously, for the first/initial iteration of the simplex algorithm, the initial basic variables and basis partition could be chosen as

$$x_B = x_S, \tag{3.30}$$

and

$$x_N = x = 0 \tag{3.31}$$

with

$$B = I \tag{3.32}$$

and

$$N = A, \tag{3.33}$$

since this would result in an initial basic solution from (3.27) given by

$$\begin{aligned} x_B &= B^{-1}b \\ &= I^{-1}b \\ &= b. \end{aligned}$$

This is a good starting point for the algorithm, since it can then subsequently proceed to steps 3 and 4, whereafter the basic variables and basis partition can be updated and steps 1–4 can be repeated until the reduced cost in step 2 is less than or equal to zero, i.e until $c_N - B^{-1}N \leq 0$, with $x_B \geq 0, x_N = 0$.

Thus, in summary, the simplex method can be implemented through an algorithm that moves from one extreme point to an adjacent extreme point to find the optimal solution for a standard LP problem, with the only requirement being that the decision variables are positive. Hence, the decision variables can take on any real numbers larger than or equal to zero, that is, $x \in \mathbb{R}^n$, with $x \geq 0$. However, certain problems might require that some or all of the variables in the LP problem are restricted to be integer. These type of optimisation problems are called integer linear programming problems. Moreover, if all the variables are restricted to be integer, the problem is known as a pure integer linear program, whereas if some variables are restricted to integer and some not (continuous or real-valued numbers), the problem is known as a mixed integer linear programming problem (MILP). In the following section integer and mixed integer linear programming problems are considered together with the methods and algorithms used to find the optimal solution for these type of problems.

3.4 Integer and mixed integer linear programming (MILP) problems

Integer linear programming problems are LP problems that require either some or all of the decision variables to be integer. From a modelling point of view, the use of integer variables in LP problems has a wide variety of applications, *e.g.*, the introduction of logical requirements or logical decision making capability and the modelling of fixed costs (see Cornuejols and Tütüncü, 2006).

Consider a pure integer linear program (ILP) with a linear objective function, linear constraints and where the decision variables are restricted to be integer. That is,

$$\begin{aligned} \min \quad & c^T x \\ \text{s.t.} \quad & Ax \geq b, \\ & x \geq 0 \text{ and integer}, \end{aligned} \tag{3.34}$$

where $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ are known and $x \in \mathbb{N}_0$ denotes the vector of integer decision variables to be determined with $\mathbb{N}_0 = \{0, 1, \cdots\}$. However, for certain problems it might be required that the vector of decision variables x represent a binary variable, that is, $x \in \{0, 1\}$. These problems occur surprisingly often and are called pure binary integer linear programs or pure $0-1$ linear programs. In addition to this, if some variables in an optimisation problem are restricted to integer and some not *e.g.* if there are integer constrained decision variables and continuous decision variables, the problem is called a mixed integer linear programming problem. Hence, consider the following MILP,

$$\begin{aligned} \min \quad & c^T x \\ \text{s.t.} \quad & Ax \geq b, \\ & x \geq 0, \\ & x_j \in \mathbb{N}_0 \quad \text{for } j = 1, 2, \ldots, k, \end{aligned} \tag{3.35}$$

where $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ are known. Furthermore, $x = [x_1, x_2, \cdots, x_k, x_{k+1}, \ldots, x_n]^T$ is a n-dimensional column vector containing the continuous and integer decision variables where $x_j \in \mathbb{N}_0$, $j = 1, 2, \ldots, k$ with $1 \leq k < n$ denote the integer decision variables.

The integrality constraint on the decision variables given in (3.34) and (3.35), classify these optimisation problems as ILP problems rather than just a standard LP problem as given in (3.15). In addition to this, the integrality constraints make these types of problems much harder to solve. In what follows a method and algorithm used for solving these ILPs and MILPs is considered.

3.4.1 The Branch and Bound method

In 1958, Gomory proposed an idea to solve MILPs that is based on cutting planes. More specifically, the approach involves the removal of non integer solutions by adding constraints (cuts) to the underlying LP and was historically the first method developed to solve MILPs. In 1960 Land and Doig proposed the branch and bound method where the problem is divided into numerous smaller problems (branch) and consequently solving the underlying LPs (bound) to evaluate their quality. This method is based on a "divide and conquer" approach and has been the most effective method for solving MILPs for many years. However, in recent years, the branch and bound method was combined with the cutting planes method, resulting in the branch and cut method as termed by Padberg and Rinaldi (1987). The structure behind the branch and cut method is similar to that of the branch and bound method, therefore the underlying idea behind the branch and bound method will be discussed first. All of the above mentioned methods require the solving of a number of LPs in an attempt to solve the following MILP problem

$$\begin{aligned} \min \quad & c^T x \\ \text{s.t.} \quad & Ax \geq b, \\ & x \geq 0, \\ & x_j \in \mathbb{N}_0 \quad \text{for } j = 1,2,\ldots,k. \end{aligned} \tag{3.36}$$

Now, consider the following LP where the integrality constraint in (3.36) is relaxed *i.e.* dropped

$$\begin{aligned} \min \quad & c^T x \\ \text{s.t.} \quad & Ax \geq b, \\ & x \geq 0. \end{aligned} \tag{3.37}$$

The problem in (3.37) is known as the LP relaxation of (3.36). Furthermore, since the LP relaxation in (3.37) is less constrained than the MILP in (3.36), the following can be deduced,

- The objective function value corresponding to the optimal solution of the LP in (3.37), denoted by Z_R, is less than or equal to the objective function value corresponding to the optimal solution of the MILP in (3.36), denoted by Z^*. This suggests that by solving the LP relaxation, a lower bound for the MILP is obtained, *i.e.*

  $$Z_R \leq Z^*. \tag{3.38}$$

  For minimisation problems the LP relaxation gives a lower bound for the MILP, whereas for maximisation problems the LP relaxation gives an upper bound.

- If the LP relaxation is infeasible, then the MILP is also infeasible.

- If the optimal solution for the LP relaxation x^*, satisfies the integer constraints x_j^* for $j = 1,2,\ldots,k$, it can be concluded that the solution x^* is also an optimal solution for the MILP.

Therefore, by solving the LP relaxation in (3.37), a bound for the MILP in (3.36) is obtained and for some problems the solution might also be an optimal solution. However, by simply rounding the solutions of the LP relaxation will not necessarily give the optimal solution for the MILP. Note that if all the decision variables in (3.36) were required to be integer, the problem would be classified as a pure integer linear program *i.e.* an ILP.

In order to explain the basic idea behind the branch and bound method, lets consider the following example of a pure ILP

$$\begin{array}{lllll} \max & z = & 3x_1 & +2x_2 & \\ \text{s.t.} & & -2x_1 & +4x_2 & \leq 5 \\ & & 2x_1 & +1x_2 & \leq 19 \\ & & x_1, & x_2 & \geq 0 \\ & & x_1, & x_2 & \text{integer.} \end{array} \tag{3.39}$$

By ignoring the integrality constraint, the LP relaxation problem is obtained, for example, at node 0. Note that each of the LPs solved in this example will correspond to a node in the branch and bound tree, with the branches connecting the nodes. The branch and bound tree will be considered after this example. Thus , the LP problem at node 0 is given by

$$\begin{array}{lllll} \max & z_0 = & 3x_1 & +2x_2 & \\ \text{s.t.} & & -2x_1 & +4x_2 & \leq 5 \\ & & 2x_1 & +1x_2 & \leq 19 \\ & & x_1, & x_2 & \geq 0, \end{array} \tag{3.40}$$

which can be solved using the simplex method explained in Section 3.3.1. The optimal solution for the LP in (3.40) is $x_1 = 7.1$ and $x_2 = 4.8$ with an objective value and upper bound of $z_0 = 30.9$. As mentioned previously, the objective value of the LP relaxation in (3.40) is also an upper bound for the LP problem in (3.39). However, this is not a feasible solution for the problem in (3.39), since the values of the decision variable are not integers. Thus, this specific solution has to be excluded due to its continuous nature, whilst keeping the feasible integral solutions preserved. This can be done by imposing additional constraints on the decision variables (branching) and consequently creating two LPs, the first with the additional constraint $x_1 \geq 8$ at node 1 and the second with additional constraint $x_1 \leq 7$ at node 2. The first LP is given by

$$\begin{array}{lllll} \max & z_1 = & 3x_1 & +2x_2 & \\ \text{s.t.} & & -2x_1 & +4x_2 & \leq 5 \\ & & 2x_1 & +1x_2 & \leq 19 \\ & & & x_1 & \geq 8 \\ & & x_1, & x_2 & \geq 0, \end{array} \tag{3.41}$$

where the optimal solution is $x_1 = 8$ and $x_2 = 3$ (which is also a feasible integer solution) and with an objective value and lower bound of $z_1 = 30$. As a result of the feasible integer solution obtained, no further branching is necessary from this node. The LP at node 2 is given by

$$\begin{array}{lllll} \max & z_2 = & 3x_1 & +2x_2 & \\ \text{s.t.} & & -2x_1 & +4x_2 & \leq 5 \\ & & 2x_1 & +1x_2 & \leq 19 \\ & & & x_1 & \leq 7 \\ & & x_1, & x_2 & \geq 0, \end{array} \tag{3.42}$$

where the optimal solution is $x_1 = 7$ and $x_2 = 4.75$ and with a objective value and upper bound of $z_2 = 30.5$. This objective value of 30.5 is larger than the objective value of the LP problem in (3.41) which was 30, hence further branching from this node is needed by imposing additional constraints on the non-integer solution obtained for x_2. This can once again be done by creating two LPs similar to (3.42), the first with additional constraint $x_2 \geq 5$ at node 3 and the second with additional constraint $x_2 \leq 4$ at node 4. The LP at node 3 is then given by

$$\begin{array}{lllll} \max & z_3 = & 3x_1 & +2x_2 & \\ \text{s.t.} & & -2x_1 & +4x_2 & \leq 5 \\ & & 2x_1 & +1x_2 & \leq 19 \\ & & & x_1 & \leq 7 \\ & & & x_2 & \geq 5 \\ & & x_1, & x_2 & \geq 0. \end{array} \tag{3.43}$$

However, this LP is infeasible due to the additional constraints imposed and consequently no further branching is necessary from this node. The LP at node 4 is given by

$$\begin{array}{lllll} \max & z_4 = & 3x_1 & +2x_2 & \\ \text{s.t.} & & -2x_1 & +4x_2 & \leq 5 \\ & & 2x_1 & +1x_2 & \leq 19 \\ & & & x_1 & \leq 7 \\ & & & x_2 & \leq 4 \\ & & x_1, & x_2 & \geq 0, \end{array} \tag{3.44}$$

where the optimal solution is $x_1 = 7$ and $x_2 = 4$ (which is also a feasible integer solution) and with an objective value and lower bound of $z_4 = 29$. As a result of the feasible integer solution obtained, no further branching is necessary from this node.

As no further branching is necessary at any of the nodes, the feasible integer solution obtained for the LP in (3.41) corresponds to the largest lower bound, hence this solution is also the optimal feasible

solution. Figure 3.4 shows the branch and bound tree in which the sub problems of the branch and bound method are given. Each of the LPs solved correspond to a node in the branch and bound tree, with the branches connecting the nodes. At each node of the branch and bound tree, the enumeration can be stopped (pruned) for any one of the following three different reasons:

- Prune by integrality: this occurs if the optimal solution for the corresponding LP is integral *i.e.* the optimal solution for the LP problem in (3.41) is a feasible integer solution as can be seen in Figure 3.4.

- Prune by bounds: this occurs if a LP at a node has a smaller objective value than the objective value corresponding to the best feasible integer solution thus far *i.e.* any further branching will only result in a smaller objective function.

- Prune by infeasibility: this occurs if the additional constraints added to the LPs when branching results in a LP at a node that is infeasible meaning that it is not possible to meet the constraints of the LP in (3.43) as indicated in the last node of Figure 3.4.

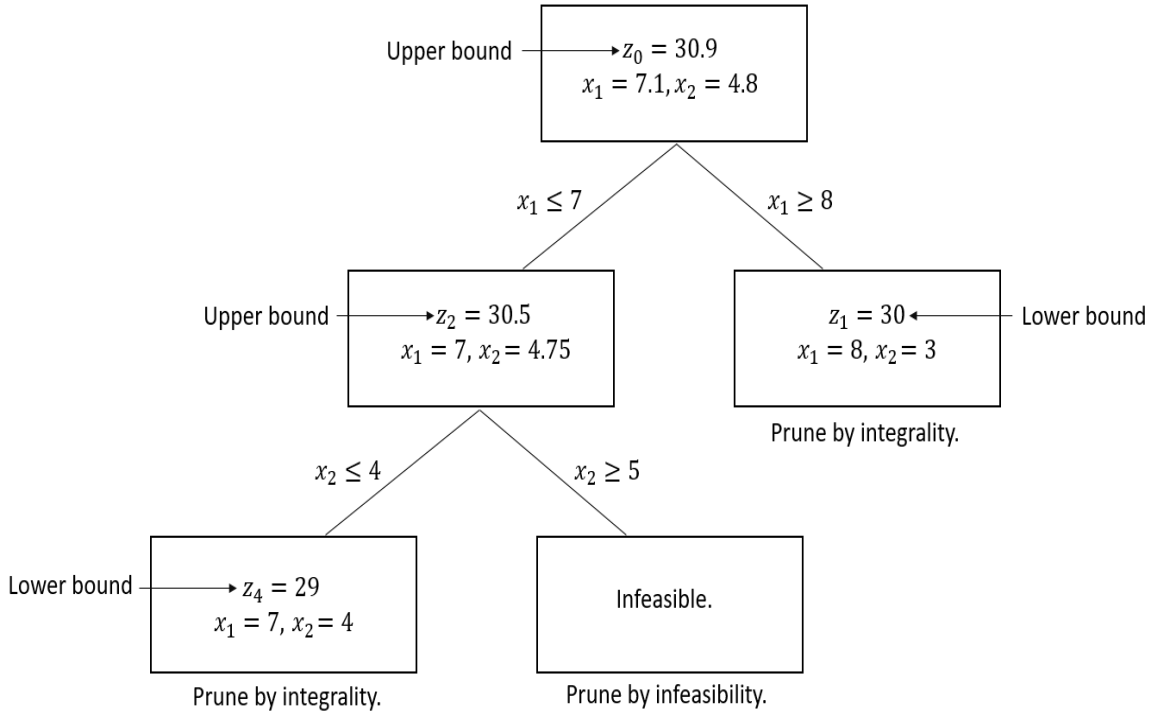


Figure 3.4: The branch and bound tree for the ILP given in (3.39).

The branch and bound method extends naturally to MILP problems like (3.35). For MILP problems, sub problems are created by also relaxing integrality requirements, but then only imposing addi-

tional constraints on the integer decision variables in the form of upper and lower limits until a optimal feasible solution is obtained for the MILP .

3.4.2 Branch and bound algorithm for solving MILPs

Consider a MILP given by

$$\begin{aligned} \max\ z = \quad & c^T x \\ & Ax \leq b, \\ & x \geq 0 \\ & x_j \in \mathbb{N}_0 \quad \text{for } j = 1,2,\ldots,k, \end{aligned} \tag{3.45}$$

where c denotes an n-dimensional row vector, A denotes an $m \times n$ matrix of constraint coefficients, b denotes the m-dimensional column vector of constraint requirements and $1 \leq k \leq n$. Furthermore, let u_j denote the upper bound and l_j the lower bound for the integer decision variables $x_j,\ j = 1,2,\ldots,k$ with $u_j - l_j = 1$. Let L_0 denote the initial LP relaxation corresponding to the root node in the branch and bound tree and let z_0 denote the objective value of the LP relaxation problem at that node. The idea behind the branch and bound algorithm is to create and solve sub problems branching from problem L_0 at the root node, by relaxing integrality requirements and imposing constraints on the integer variables, namely upper bound constraints $x_j \leq u_j$ and lower bound constraints $x_j \geq l_j$ until the optimal feasible solution is obtained. Various strategies exist to determine the order of solving sub problems (node selection strategies) and the order of variable constraining (branching strategies), which may differ depending on the application thereof. These strategies are usually available as options in MILP solvers and therefore will not be discussed in detail. The algorithm (adopted from Taylor, 2009) can be summarised in the following steps.

1. At the root node, find the optimal solution x^* to the initial LP relaxation problem L_0, *i.e.* (3.45) and let z_0 denote the objective value of the LP relaxation and also an upper bound to the MILP.

2. Stop the algorithm if this optimal solution x^* (which contains x_j^*, the optimal solution for the integer decision variables) at the root node satisfies the integer constraints x_j for $j = 1,2,\ldots,k$. However, if L_0 is infeasible the MILP is also infeasible and the algorithm stops (prune by infeasibility). If the algorithm does not stop, go to step 3.

3. Branch: create sub problems by imposing the following upper and lower bound constraints on a x_j (which do not have an integer solution), $x_j \leq \left\lfloor x_j^* \right\rfloor$ and $x_j \geq \left\lceil x_j^* \right\rceil$, where $\left\lfloor x_j^* \right\rfloor$ and $\left\lceil x_j^* \right\rceil$ denote the floor and ceiling of x_j^*, respectively.

4. Create two new nodes, each with a sub problem as a result, of the upper and lower bound constraints in step 3.

5. Solve the sub problems created at each node in step 4.

6. For the solutions obtained in step 5 consider the following options:

   - prune by infeasibility: if the sub problem at any node is infeasible, prune the branch.
   - prune by bounds: if the solution x^* is not a feasible solution to the MILP in (3.45) (the integer constraint for x_j is not satisfied) and the corresponding objective value is smaller than any other feasible integer solution to the MILP, prune the branch, since a better lower bound has been found already.
   - branch: if the solution x^* of a sub problem at any node is not a feasible solution to the MILP in (3.45) i.e., the integer constraint for x_j is not satisfied, but the corresponding objective value is larger than any other feasible solution for the MILP. Note that the corresponding objective value is an upper bound at this node. Go back to step 3.
   - prune by integrality: if the solution x^* is a feasible integer solution for the MILP, prune that branch. The objective value at the nodes pruned by integrality form a lower bound.

7. Stop the algorithm: the feasible integer solution x^* with the largest lower bound for the MILP at any ending node is the optimal integer solution and the algorithm can stop.

Note that for a minimisation problem, the lower and upper bounds are reversed, but the rest of the algorithm remains the same.

Ever since it was proposed by Land and Doig in 1960, the branch and bound method has been the most effective technique used for solving MILPs. However, more recently, the cutting planes method and the branch and bound method have been effectively combined to create the branch and cut method.

3.4.3 The branch and cut method

The cutting plane method proposed by Gomory (1960) involves strengthening the LP-relaxation by adding valid inequality constraints (cuts) for the MILP, attempting to eliminate the non-integer solutions, whilst preserving the feasible integer solutions for the MILP (see Gomory, 1960 and Mitchell, 2002). Initially, this method did not appear to be a competitive method due to the slow convergence of the algorithm. However, in conjunction with the development of polyhedral theory and the introduction of problem specific, strong cutting planes, this method made its reappearance and is now being combined with the branch and bound method in a method called the branch and cut method. According to Cornuejols and Tütüncü (2006) some of the most well-known software packages use the branch and cut method for solving MILPs, with Xpress and Cplex known as two excellent commercial branch and cut codes. This method works similar to the branch and bound method, with the main difference that cuts are generated by adding valid inequality constraints when a node is explored in order to strengthen the formulation, thus improving the bounds of the LP relaxations and approximating

the MILP more accurately. The most common cuts generated include the Gomory mixed integer cuts (GMI cut), lift-and-project cuts, mixed integer rounding cuts (MIR cuts), intersection cuts and split cuts. The interested reader is referred to Cornuéjols (2008) for a discussion on these cuts.

3.5 Non-linear programming (NLP) problems

The above mentioned algorithms and methods i.e., the simplex method, the branch and bound method and the branch and cut method, were considered in the context of solving LP and MILP problems. That is, the objective function and constraint (equality and inequality) functions i.e., functions f and h_i in (3.6), are all linear, with the decision variables being either continuous, integer or a combination of the two. However, if the objective function f or one of the constraint functions h_i are not linear, the corresponding optimisation problem is a non-linear programming (NLP) problem. NLP problems, can be solved effectively by reducing them to MILP problems when using efficient techniques to linearise the non-linear functions (see, *e.g.*, Babayev, 1997 and D'Ambrosio et al., 2010). In addition to this, the increased efficiency of MILP software tools in recent years have encouraged the use of piece-wise linear approximation approaches to linearise non-linear functions and solve NLP problems. A piece-wise linear approximation approach is considered to convert the NLP problems to MILP problems and consequently to solve the MILP problem.

Consider the following generic optimisation problem

$$\begin{aligned} \min \quad & f(x) \\ & h_i(x) = 0 \quad i \in \mathscr{I}, \\ & h_i(x) \geq 0 \quad i \in \mathscr{E}, \end{aligned} \tag{3.46}$$

where $\mathscr{I}$ *and* $\mathscr{E}$ denote the index sets for the equality and inequality constraints, respectively. A NLP problem is an optimisation problem where the objective function f or one of the constraint functions h_i in (3.46) are not linear. Non-linear functions can be approximated using linear interpolation between fixed gridpoints, thus allowing it to be solved using LP methods and a piece-wise linear approximation approach (Misener and Floudas, 2010).

3.5.1 A piece-wise linear approximation approach for solving an NLP problem with a single variable as an MILP problem

Consider a function $f(x) : \mathbb{R} \to \mathbb{R}$ of a single variable $x \in \mathscr{X} \subset \mathbb{R}$. The piece-wise linear approximation of the function $f(x)$ can be obtained by first dividing the domain $\mathscr{X}$ of the variable x into $N+1$ gridpoints/ breakpoints say $x_i,\ i \in \mathscr{I}_0 = \{0,1,\ldots,N\}$ where x_0 and x_N denote the start and endpoint of the domain. The function value $f(x_i)$ can then be calculated at each gridpoint $x_i,\ i \in \mathscr{I}_0$ whereafter the function can be approximated by N linear segments between $(x_{i-1}, f(x_{i-1}))$ and $(x_i, f(x_i))$, where

$i \in \mathscr{I} = \{1,2,\ldots,N\}$. The function value at a point $\tilde{x} \in \mathscr{X}$, between two gridpoints x_{i-1} and x_i can be approximated by a convex combination of the two function values $f(x_{i-1})$ and $f(x_i)$. Note that, if the domain $\mathscr{X}$ is convex, the point $\tilde{x} \in \mathscr{X}$ can be written as a convex combination of the two gridpoints x_{i-1} and x_i (Misener and Floudas, 2010). Thus, if $\lambda \in [0,1]$ and $\tilde{x} \in \mathscr{X}$ lie between the two gridpoints x_{i-1} and x_i, such that

$$\tilde{x} = \lambda x_{i-1} + (1-\lambda)x_i \in \mathscr{X}, \tag{3.47}$$

then the piece-wise linear approximation of the function at the point $\tilde{x}$ is given by

$$\hat{f}(\tilde{x}) = \lambda f(x_{i-1}) + (1-\lambda)f(x_i). \tag{3.48}$$

The approximation in (3.48) is the same as interpolating between the fixed gridpoints $(x_{i-1}, f(x_{i-1}))$ and $(x_i, f(x_i))$ given by

$$\hat{f}(\tilde{x}) = f(x_{i-1}) + \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}(\tilde{x} - x_{i-1}), \tag{3.49}$$

where $\lambda = \frac{x_i - \tilde{x}}{x_i - x_{i-1}}$ in (3.48). Furthermore, according to Definition 3.6, the point $x \in \mathscr{X} \subset \mathbb{R}$ within a line segment, can be written as a convex combination of two adjacent gridpoints (vertices) $x_i,\ i \in \mathscr{I}_0 = \{0,1,\ldots,N\}$ with $\lambda_i \in [0,1],\ \forall\, i \in \mathscr{I}_0$ and $\sum_{i=0}^{N} \lambda_i = 1$. That is,

$$x = \sum_{i \in \mathscr{I}_0} \lambda_i x_i, \tag{3.50}$$

with the corresponding function value in the point $x \in \mathscr{X}$ then approximated by

$$\hat{f}(x) = \sum_{i \in \mathscr{I}_0} \lambda_i f(x_i). \tag{3.51}$$

Note that, if $x = x_i,\ i \in \mathscr{I}_0$, then $\hat{f}(x) = f(x_i)$ which is equal to the actual function value in the gridpoint x_i. To solve an NLP problem using the piece-wise linear approximation approach as described above, additional variables and constraints are required to ensure that only two adjacent gridpoints are activated, such that x lies between them. Therefore, introduce a binary variable $s_i \in \{0,1\},\ i \in \mathscr{I} = \{1,2,\ldots,N\}$ associated with the line segment between the gridpoints (x_{i-1}) and (x_i). Here the set of variables s_i are known as a Special Ordered Set of type 1 (SOS1) as introduced by Beale and Tomlin (1970), since at most one of these variables can take on a non-zero value, ensuring only a single line segment is activated. Furthermore, let $\lambda_i \in [0,1],\ \forall\, i \in \mathscr{I}_0$ denote continuous variables (weights) associated with each gridpoint $x_i,\ i \in \mathscr{I}_0$. Note that, since only the two adjacent gridpoints (vertices) associated with the single activated line segment are allowed to contribute to the function approximation (interpolation), the following constraints are imposed to obtain the function approximation as a

convex combination of two function values

$$\sum_{i\in\mathscr{I}} s_i = 1 \tag{3.52}$$

$$\lambda_0 \leq s_1 \tag{3.53}$$

$$\lambda_i \leq s_i + s_{i+1} \qquad \forall i \in \mathscr{I}/\{N\} \tag{3.54}$$

$$\lambda_N \leq s_N \tag{3.55}$$

$$\sum_{i\in\mathscr{I}_0} \lambda_i = 1 \tag{3.56}$$

$$x = \sum_{i\in\mathscr{I}_0} \lambda_i x_i \tag{3.57}$$

$$\hat{f}(x) = \sum_{i\in\mathscr{I}_0} \lambda_i f(x_i). \tag{3.58}$$

Constraint set (3.52) ensures that only a single line segment is activated, whereas constraints (3.53)–(3.55) guarantee that only the weights associated with the gridpoints (vertices) of the activated line segment (*i.e.* two adjacent gridpoints) can take on values other than 0. Furthermore, constraint set (3.56) ensures that the weights add up to 1 with constraint set (3.57) representing x as a convex combination of two adjacent gridpoints. Lastly, constraint set (3.58) denote the function approximation as a convex combination of two adjacent function values. Consequently, the above piece-wise linear approximation approach, summarised by constraints (3.52)–(3.58), reduces the single variable NLP problem to a MILP problem that can be solved using a MILP solver.

However, for certain NLP problems it might be that f is a function of two variables (x and y) rather than a single variable x. Therefore, the instance where an NLP problem with two variables is reduced to and solved as a MILP, is now considered.

3.5.2 A piece-wise linear approximation approach for solving a NLP problem with two variables as a MILP problem

Similar to the approach followed above, the case where an NLP problem with a function of two variables, say $f(x,y)$, can be solved using a piece-wise linear approximation approach, is considered. The approach followed above for a non-linear function of one variable will be extended to the case where the non-linear function f is a function of two variables. Consider a function $f(x,y) : \mathbb{R}^2 \to \mathbb{R}$ of two variables, $x \in \mathscr{X} \subset \mathbb{R}$ and $y \in \mathscr{Y} \subset \mathbb{R}$. The piece-wise linear approximation of the function $f(x,y)$ can be obtained by first dividing the domain $\mathscr{X}$ into $N+1$ gridpoints $x_i,\ i \in \mathscr{I}_0 = \{0,1,\ldots,N\}$ (where x_0 and x_N denote the start and endpoint of the domain) and divide the domain $\mathscr{Y}$ into $M+1$ gridpoints $y_j,\ j \in \mathscr{J}_0 = \{0,1,\ldots,M\}$ (where y_0 and y_M denote the start and endpoint of the domain). The function $f(x,y) : \mathbb{R}^2 \to \mathbb{R}$ can now be evaluated in each combination of gridpoints $(x_i, y_j) \in \mathbb{R}^2,\ i \in \mathscr{I}_0,\ j \in \mathscr{J}_0$

,i.e., $f(x_i,y_j),\ i \in \mathscr{I}_0,\ j \in \mathscr{J}_0$. To determine the function approximation between two gridpoints, say $(\tilde{x},\tilde{y})$, where $x_{i-1} \leq \tilde{x} \leq x_i$ and $y_{i-1} \leq \tilde{y} \leq y_i$, consider the rectangle with vertices (x_{i-1},y_{j-1}), (x_i,y_{j-1}),(x_{i-1},y_j) and (x_i,y_j) , where $i \in \mathscr{I} = \{1,2,\cdots,N\}$ and $j \in \mathscr{J} = \{1,2,\ldots,M\}$. The function value at the point $(\tilde{x},\tilde{y})$ (*i.e.* $f(\tilde{x},\tilde{y})$) is then approximated by a convex combination of the function values evaluated at the vertices of the rectangle containing the point $(\tilde{x},\tilde{y})$.

Therefore, to solve an NLP problem using the piece-wise linear approximation approach as described above, introduce a set of continuous variables (weights), $\lambda_{ij} \in [0,1]$ where $i \in \mathscr{I}_0$ and $j \in \mathscr{J}_0$ and $\sum_{i\in\mathscr{I}_0}\sum_{j\in\mathscr{J}_0}\lambda_{ij} = 1$, associated with each combination of the gridpoints $(x_i,y_j) \in \mathbb{R}^2,\ i \in \mathscr{I}_0,\ j \in \mathscr{J}_0$. Any point $(x,y) \in \mathbb{R}^2$ where $x \in \mathscr{X} \subset \mathbb{R}$ and $y \in \mathscr{Y} \subset \mathbb{R}$, can be written as a convex combination of the gridpoints and subsequently, the function can be approximated by a convex combination of the function evaluated in the convex combination of gridpoints. Here, the convex combination of gridpoints are given by

$$x = \sum_{i\in\mathscr{I}_0}\sum_{j\in\mathscr{J}_0}\lambda_{ij}x_i \tag{3.59}$$

$$y = \sum_{i\in\mathscr{I}_0}\sum_{j\in\mathscr{J}_0}\lambda_{ij}y_j \tag{3.60}$$

with the corresponding function value in the point (x,y) then approximated by

$$\hat{f}(x,y) = \sum_{i\in\mathscr{I}_0}\sum_{j\in\mathscr{J}_0}\lambda_{ij}f(x_i,y_j). \tag{3.61}$$

Furthermore, to ensure that only the vertices of the single activated rectangle contribute to the linear interpolation/ approximation of the function, introduce a binary variable $s_{ij} \in \{0,1\}$ corresponding to the rectangle with vertices (x_{i-1},y_{j-1}), (x_i,y_{j-1}),(x_{i-1},y_j) and (x_i,y_j), where $i \in \mathscr{I} = \{1,2,\ldots,N\}$ and $j \in \mathscr{J} = \{1,2,\ldots,M\}$. Here the set of variables s_{ij} are also known as SOS1 as introduced by Beale and Tomlin (1970), since at most one of these variables can take on a non-zero value, ensuring only a single rectangle is activated. Consequently, the following constraints are imposed to obtain the function approximation as a convex combination of the function values of the vertices associated with the single activated rectangle (see D'Ambrosio et al., 2010, Babayev, 1997 and Misener and Floudas, 2010)

$$\sum_{i\in\mathscr{I}}\sum_{j\in\mathscr{J}}s_{ij} = 1 \tag{3.62}$$

$$\sum_{j\in\mathscr{J}_0}\lambda_{0j} \leq \sum_{j\in\mathscr{J}}s_{1j} \tag{3.63}$$

$$\sum_{j\in\mathscr{J}_0}\lambda_{ij} \leq \sum_{j\in\mathscr{J}}(s_{ij}+s_{(i+1),j}) \qquad \forall i \in \mathscr{I}/\{N\} \tag{3.64}$$

$$\sum_{j\in\mathscr{J}_0}\lambda_{Nj} \leq \sum_{j\in\mathscr{J}} s_{Nj} \tag{3.65}$$

$$\sum_{i\in\mathscr{I}_0}\lambda_{i0} \leq \sum_{i\in\mathscr{I}} s_{i1} \tag{3.66}$$

$$\sum_{i\in\mathscr{I}_0}\lambda_{ij} \leq \sum_{i\in\mathscr{I}} (s_{ij} + s_{i,(j+1)}) \qquad \forall j \in \mathscr{J}/\{M\} \tag{3.67}$$

$$\sum_{i\in\mathscr{I}_0}\lambda_{iM} \leq \sum_{i\in\mathscr{I}} s_{iM} \tag{3.68}$$

$$\sum_{i\in\mathscr{I}}\sum_{j\in\mathscr{J}_0}\lambda_{ij} = 1 \tag{3.69}$$

$$x = \sum_{i\in\mathscr{I}_0}\sum_{j\in\mathscr{J}_0}\lambda_{ij}x_i \tag{3.70}$$

$$y = \sum_{i\in\mathscr{I}_0}\sum_{j\in\mathscr{J}_0}\lambda_{ij}y_j \tag{3.71}$$

$$\hat{f}(x,y) = \sum_{i\in\mathscr{I}_0}\sum_{j\in\mathscr{J}_0}\lambda_{ij}f(x_i,y_j). \tag{3.72}$$

In the above constraint formulation, constraint (3.62) guarantees that only a single rectangle is activated, whereas constraints (3.63)–(3.68) ensure that only the weights associated with the gridpoints (vertices) of the activated rectangle can take on values other than 0. Furthermore, constraint (3.69) ensure that the activated weights add up to 1 with constraint (3.70) and (3.71) representing x and y as a convex combination of the vertices of the activated rectangle. Lastly, constraint (3.72) denotes the function approximation as a convex combination of the function values corresponding to the vertices of the activated rectangle. This piece-wise linear approximation approach of a function with two variables, as described by constraints (3.62)–(3.72), reduces the NLP problem to a MILP problem, which can be solved using a MILP solver. Thus, using the approaches discussed in Sections 3.5.1 and 3.5.2, NLP problems containing non-linear functions of a single variable or two variables can be solved as MILP problems using a MILP solver.

The optimisation methods discussed in this chapter can, therefore, also be used to solve the credit price optimisation problem formulated in Section 2.4.3 of Chapter 2. More specifically, the credit price optimisation problem can be extended to a problem with, not only price as a decision variable, but also loan-to-value (LTV). In the next chapter, the credit price and LTV optimisation problem is considered in conjunction with the optimisation methods used to solve the problem.

Chapter 4

Credit price and LTV optimisation

4.1 Introduction

In a competitive financial industry, one of the challenges faced with secured retail lending products is to determine the optimal prices (*i.e.*, interest rates) that maximise both the loan take-up probability and expected revenue to the lender. Traditionally, risk-based pricing was used to determine the prices of consumer credit products. For these products, the price included a risk premium based on the risk category of the customer. However, in recent years, pricing methodologies moved away from risk-based pricing towards price optimisation or profit-based pricing as discussed in Section 2.4 of Chapter 2. In credit price optimisation, the demand of a potential customer is mathematically captured by a price elasticity model (price response model) where the demand is expressed as a function of price. In secured loans, the demand is referred to as the probability that the potential customer will take up a loan. Secured loans here, refer to home loans (or mortgages). More specifically, in profit-based pricing, the probability of take-up can be related to the change in price using a price response model. In this Chapter, a response model is proposed which relates take-up probabilities with not only price, but also loan-to-value (loan amount expressed as a percentage of the value of the underlying asset), taking into account a customer's willingness to pay for a loan. With the objective to maximise the expected net present interest income (NPII), a piece-wise linear approximation approach is followed to simultaneously determine the optimal price and loan-to-value (LTV) for a potential customer while still adhering to the risk distribution and LTV constraints on the portfolio. Hence, this allows the lender to determine optimal levels of price and LTV through the application of an explicit optimisation model. Furthermore, by following a piece-wise linear approximation approach, logical decision-making capability is introduced into the model, allowing for the exclusion of customers from the portfolio based on a trade-off between risk and profitability. The optimisation model considered in this chapter is based on the work of Phillips (2013), Terblanche and De la Rey (2014) and partially based on the proceedings of Smuts and Terblanche (2019). To the best of our knowledge, price and LTV optimisation had not yet been considered previously in secured loans.

The remainder of this chapter is organised as follows: in Section 4.3 a formulation of the non-linear credit price and LTV optimisation problem is presented and constraints on the risk distribution and the proportion of loans with a high LTV are considered to limit the risk inherent in the portfolio. In Section 4.4 a piece-wise linear approximation approach is considered to solve the non-linear credit price and LTV optimisation problem. The inclusion of logical decision-making capability in the optimisation model as a result of the approach followed is also introduced in this section. This chapter concludes in Section 4.5 where the model behaviour and computational results of the credit price and LTV optimisation problem are discussed.

4.2 Background

The problem that most lenders face is to determine what price (interest rate) they should quote a prospective customer when they apply for a loan. However, the price is not the only decision variable (unknown variable) to be determined, specifically when considering secured loans. For secured loans *i.e.* home loans or auto loans, an additional variable of interest is the LTV. The LTV refers to the loan amount as a percentage of the underlying asset value, *i.e.*, the value of the home to be purchased. Moreover, by including the LTV as a variable in the credit price optimisation problem, constraints can be imposed to limit the proportion of loans with a higher LTV, since loans with a higher LTV are considered more risky as apposed to loans with lower LTVs (see Phillips, 2013 and Caufield, 2012).

According to Phillips (2013), loans may vary at even the most basic level including but not limited to the price (interest rate), the term over which the loan is to be repaid, the loan amount applied for (the principle amount) etc. In addition to this, loans also differ from one country to another. In the United states the most popular home loan is over a term of 35 years, where a fixed interest rate is determined after the application and which remains the same for the duration of the loan *i.e.* the annual interest rate for the home loan does not change over the duration of 35 years. In contrast to this, the most popular home loan in Canada is considered a renewable 5 year home loan. For these type of loans, the repayment term is usually set to 25 years, but, the outstanding capital balance is due after 5 years. However, when the outstanding capital balance is due, the customer can decide whether or not to renew the loan at the same lender or a different lender. This decision will typically depend on the interest rate the lender offers the customer, as it can possibly be different to the initial interest rate. For the fixed interest loan, the lender faces the risk of an interest rate change over the duration of the loan whereas in the case of the 5 year renewable loan, the customer assumes some of the risk due to a possible interest change after 5 years. In South Africa, the most common home loan is a 20 year loan where the interest rate is not fixed. For these loans, the interest rate is linked to the repurchase rate which may vary over the term of the loan, *i.e.* when there is a change in the repurchase rate. Recall that, the repurchase rate refers to the interest rate at which the South African Reserve Bank (SARB) lends money to the banks. There are, however, fixed interest rate home loans where the interest rate remains the same for

the duration of the loan as apposed to the variable interest rate loan. In addition to this, some home loans in South Africa might also have a term of 30 years. Thus, to get a sense of the applicability of credit price and LTV optimisation in the context of secured loans *i.e.* home loans, consider the home loan application process as depicted in Figure 4.1.

If an application for a home loan has been submitted, the banks or financial institutions first access the credit risk associated with the customer or applicant. Credit risk in loans refers to probability that the customer will default, *i.e.*, the probability that the customer will not make their monthly repayments. The probability of default of a customer is estimated at the time of application and is based on the applicants characteristics and any data available on the applicant (Siddiqi, 2012). The method used to estimate the credit risk associated with the loan applicant is called credit scoring (see Hand and Henley, 1997 and Mester et al., 1997) and is based on a score given to the loan applicant. In general, a higher score corresponds to a lower risk of default, whereas a lower score corresponds to a higher risk. The applicant's score is obtained from a credit risk scorecard and is based on their characteristics *i.e.* demographics (age, time at current residence or job etc), existing relationship (time at current bank, number of credit products, previous payment performance etc), data collected from the credit bureau and or any information collected on the applicant. According to Siddiqi (2012) and Hand and Henley (1997) the estimated default probability or score is then used as a basis to assist the lender's decision on whether or not to accept the loan application. Moreover, the initial decision to accept or reject a loan is made by comparing the score to a suitable threshold or cutoff score, depending on the amount of risk the lender is willing to take. Therefore, the application of a customer with a bad credit score (or high default probability) will be rejected as seen in Figure 4.1. The interested reader is referred to Siddiqi (2012) for more information on scorecards and credit scoring.

If the applicant has a good credit score, it does not necessarily imply the application will be approved. The application of a customer with a good credit score might also be considered as non-eligible (see Figure 4.1) and therefore also be rejected. The loan application is generally classified as non-eligible when a customer applies for an unrealistic home loan amount, which results in monthly repayment amount that the customer may not be able to afford. Subsequently, these type of loan applications are also rejected. Applications from customers with a good credit score *i.e.* above a certain threshold and who are also considered eligible, are usually those who get approved. The applications that are approved get offered a price (interest rate) at which they have to repay the loan along with the approved loan amount. Note that the loan amount approved is not necessarily the loan amount the customer applied for. Further, not all the approved applicants accept the loan *i.e.* take up the loan at the offered interest rate. Some customers reject the loan offered, since they generally apply at various banks and therefore, one of them might have offered a more attractive loan *i.e.* a lower interest rate or a higher loan amount.

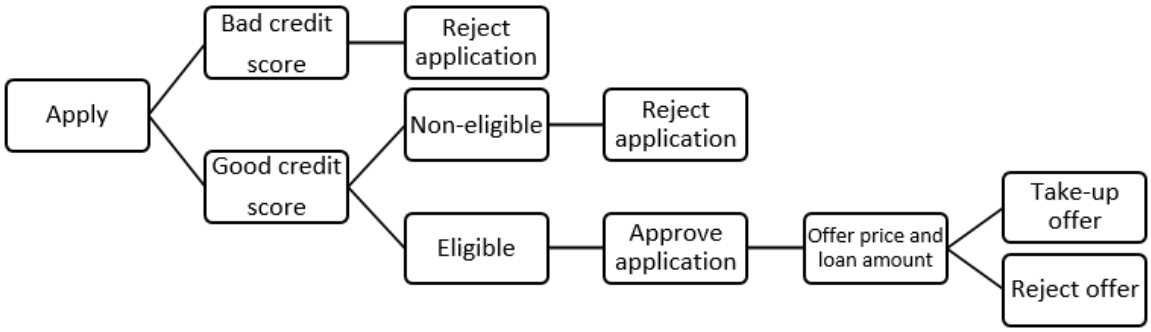


Figure 4.1: Home loan application process.

As a result, banks have data available on whether or not a potential customer took up a loan at the quoted price, *i.e.*, data from the offers that are taken up or rejected as displayed in Figure 4.1. This data can be used to estimate the demand of a potential customer using a price elasticity model (or price response model). Recall that the demand is considered as the probability that the potential customer will take up a loan. In addition to this, the LTV is also available to the banks, hence this can be used together with the interest rate in the price response model to estimate the probability that a potential customer will take up the approved loan amount at the quoted price. Consider the formulation of the credit price and LTV optimisation problem and the approach followed to solve this problem.

4.3 A non-linear approach to credit price and LTV optimisation

To formulate the credit price and LTV optimisation problem, recall from Section 2.4.2 of Chapter 2 that the expected NPII is given by

$$I(r,r_0,r_d,n,a,s_t) = \sum_{t=1}^{n} a\left(\frac{1}{(1+r_d)^t}\right)\left(\frac{s_t r(1+r)^n}{(1+r)^n-1} - \frac{r_0(1+r_0)^n}{(1+r_0)^n-1}\right). \tag{4.1}$$

An approximation of (4.1) was suggested by Phillips (2013) if assuming that $r_d \approx 0$, $\frac{(1+r)^n}{(1+r)^n-1} \approx 1$ and $\frac{(1+r_0)^n}{(1+r_0)^n-1} \approx 1$ for large values of n, that is

$$I(r,r_0,r_d,n,a) \approx an(r-r_0) - pa\delta. \tag{4.2}$$

Note that $r = \frac{x}{12}$ denotes the monthly price (interest rate) and $r_0 = \frac{x^0}{12}$ the monthly repurchase rate, with x and x^0 representing the annual price and annual repurchase rate, respectively.

To consider the credit price and LTV optimisation problem at portfolio level, the approximation of the expected NPII as suggested by Phillips (2013) and Terblanche and De la Rey (2014) will be used. Therefore, suppose a secured credit portfolio consists of a set of customers $\mathscr{C} = \{1,2,...,C\}$ where C denotes the number of customers in the portfolio. Consider Table 4.1 below, in which some of the

parameters used in the credit price and LTV optimisation problem for each customer $c \in \mathscr{C}$ are shown.

Table 4.1: Parameters used in the credit price and LTV optimisation problem.

Symbol	Description
a_c	Loan amount approved
n_c	Loan term
p_c	Probability of default
x^0	Repurchase rate
τ_c	Underlying asset value
l_c	LTV

Furthermore, let δ indicate the loss given default and assume for the remainder of this chapter that $\delta = 1$. Then, an approximation of the expected net present interest income (NPII) for a price x_c, $c \in \mathscr{C}$ is given by

$$I(x_c) := I(x_c|n_c, a_c, x^0, p_c) = n_c a_c \left(\frac{x_c}{12} - \frac{x^0}{12} \right) - a_c p_c \delta. \tag{4.3}$$

An approximation of the probability of take-up (obtained from fitting a single logistic regression model to the data) for a price x_c, $c \in \mathscr{C}$ is given by the following price response function (Terblanche and De la Rey, 2014)

$$R(x_c) := R(x_c|n_c, a_c, l_c, x^0, p_c) = 1/(1 + e^{(-(\beta_0 + \beta_1 a_c + \beta_2 n_c + \beta_3 p_c + \beta_4 x^0 + \beta_5 x_c + \beta_6 l_c))}), \tag{4.4}$$

where the regression coefficients $\beta_0, \beta_1, \ldots \beta_6$ are estimated using maximum likelihood. Stepwise logistic regression was performed when fitting the model to the data (using a significance level of 5%) and a resulting c-statistic of 0.71 was obtained. The c-statistic is an indication of the goodness-of-fit of the logistic regression model fitted to the data. A perfect model will yield a c-statistic of one whereas for a random model the c-statistic will be equal to 0.5. The price response function $R(x_c|n_c, a_c, l_c, x^0, p_c)$ provides an estimate of the probability that a customer will take up the loan at the quoted price x_c, given a loan term n_c, a loan amount a_c, a repurchase rate x_0 and a probability of default p_c. Figure 4.2 illustrates how the take-up probability is expected to decrease with an increase in price when keeping all the other parameters constant, consequently illustrating the inverse relationship between take-up probability and price as established in literature (Terblanche and De la Rey, 2014).

An approximation of the expected value of the NPII is then given by the product of (4.3) and (4.4). That is, the unconstrained credit price optimisation problem given a set of customers $c \in \mathscr{C}$ is to

$$\max \sum_{c \varepsilon \mathscr{C}} R(x_c) I(x_c) \tag{4.5}$$
$$\text{s.t. } x_c \geq 0, \qquad \forall c \in \mathscr{C}$$

where x_c is the only decision variable (unknown variable) to be determined. A unique optimal solution for this credit price optimisation problem is guaranteed provided that $I(x_c)$ is a linear approximation, $R(x_c)$ has the increasing failure rate property (Phillips, 2013) and the constraints define a convex feasible region (Boyd and Vandenberghe, 2004).

Figure 4.2: Relationship between price and take-up probability.

The objective of this study is, however, to maximise the expected NPII by finding the right balance between price and LTV according to the price elasticity model. For this purpose, the decision variable y_c is introduced in the credit price optimisation problem to express LTV in terms of the loan amount (a_c) and the value of the underlying asset (v_c) for each $c \in \mathscr{C}$, that is,

$$y_c = \frac{a_c}{\tau_c}. \tag{4.6}$$

The effect of LTV on the take-up probability is illustrated by the graph in Figure 4.3. The take-up probability is expected to increase with an increase in LTV, especially for first time property buyers who do not necessarily have enough funds available for the required deposit amount. Therefore, by writing (4.6) in terms of the loan amount a_c and then substituting the expression for a_c into (4.3) and (4.4), an approximation of the net NPII and the probability of take-up for a price x_c and LTV y_c, $c \in \mathscr{C}$ is given by

$$I(x_c, y_c) := I(x_c, y_c | n_c, \tau_c, x^0, p_c) = n_c \tau_c y_c \left(\frac{x_c}{12} - \frac{x^0}{12} \right) - \tau_c y_c p_c \delta \tag{4.7}$$

and

$$R(x_c,y_c) := R(x_c,y_c|n_c,\tau_c,x^0,p_c) = 1/(1+e^{(-(\beta_0+\beta_1\tau_c y_c+\beta_2 n_c+\beta_3 p_c+\beta_4 x^0+\beta_5 x_c+\beta_6 y_c))}), \tag{4.8}$$

respectively.

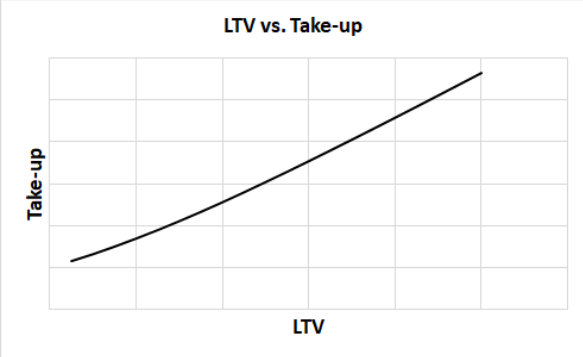


Figure 4.3: Relationship between LTV and take-up probability.

The expected value of the NPII for a customer $c \in \mathscr{C}$ is then given by the product of (4.7) and (4.8). That is, the expected value of the NPII at a price x_c and LTV y_c is given by

$$f(x_c,y_c) := R(x_c,y_c)I(x_c,y_c). \tag{4.9}$$

The credit price and LTV optimisation problem, without any constraints, is to

$$\begin{aligned} &\max \sum_{c\in\mathscr{C}} f(x_c,y_c) = R(x_c,y_c)I(x_c,y_c) \\ &\text{s.t. } x_c,y_c \geq 0 \qquad\qquad \forall c \in \mathscr{C}. \end{aligned} \tag{4.10}$$

According to Phillips (2013), a risky portfolio can be regulated by incorporating constraints on the portfolio's risk distribution. Therefore, let $\mathscr{C}(g)$ denote the set of customers having a risk grading $g \in \mathscr{G} = \{Low, Medium, High\}$ (*i.e.* the union of the sets $\mathscr{C}(g)$, $g \in \mathscr{G}$ is the set $\mathscr{C}$) and let U_g denote the upper bound for the proportion of customers having a risk grading of $g \in \mathscr{G}$ *i.e.* for the unconstrained optimisation problem in (4.10) it follows that $U_g = \{1,1,1\}$. In addition to the risk distribution constraints, the loans with an LTV larger than or equal to, say 90%, can be restricted to a small proportion of the portfolio in a further attempt to reduce the risk in the portfolio. Thus, let U_y denote the upper bound for the proportion of loans with an LTV higher than a user-specified LTV level, y^b. That is, if $U_y = 80\%$ and $y^b = 0.9$, the proportion of loans with an LTV level larger than or equal to 0.9, is limited to less than or equal to 80%. Furthermore, let t_c be an auxiliary variable denoting the take-up

probability of a customer $c \in \mathscr{C}$. The objective of the non-linear price and LTV optimisation model which incorporates the risk distribution constraints and LTV constraints, is to

$$\begin{aligned}
&\max \sum_{c\in\mathscr{C}} f(x_c,y_c) = R(x_c,y_c)I(x_c,y_c) \\
&\text{s.t. } t_c = R(x_c,y_c) && \forall\, c \in \mathscr{C}, \\
&\quad \sum_{c\in\mathscr{C}(g)} t_c \le U_g \sum_{c\in\mathscr{C}} t_c && \forall\, g \in \mathscr{G}, && (4.11) \\
&\quad \sum_{c\in\mathscr{C}} \mathbb{I}\left(y_c \ge y^b\right) \le U_y C && && (4.12) \\
&\quad x_c, y_c \ge 0 && \forall\, c \in \mathscr{C},
\end{aligned}$$

where

$$\mathbb{I}\left(y_c \ge y^b\right) = \begin{cases} 1 & \text{if } y_c \ge y^b \\ 0 & \text{if } y_c < y^b \end{cases}$$

and C denotes the number of customers in the portfolio. Note that, even though the NPII, given by $I(x_c,y_c)$, is a linear function in terms of the variables x_c and y_c, the price response function $R(x_c,y_c)$ is non-linear and consequently also the expected value of the NPII given in (4.9). Hence, the price and LTV optimisation problem is considered as a NLP problem as discussed in Chapter 3 Section 3.5. Thus, a piece-wise linear approximation approach, similar to the approach discussed in Section 3.5.2 of Chapter 3 to solve the non-linear price and LTV optimisation problem as a MILP problem, is considered below.

4.4 A piece-wise linear approximation approach to credit price and LTV optimisation

By using a piece-wise linear approximation approach to solve the price and LTV optimisation problem, proven optimal solutions are obtained and the option of introducing logical decision-making variables is made possible. More specifically, by introducing binary decision variables into the credit price and LTV problem formulation, it is possible to model the exclusion of certain customers from the credit portfolio, based on the required level of risk and the objective of maximising profitability.

The piece-wise linear approximation approach entails dividing the price range into equally spaced intervals $\mathscr{I} = \{1,2,...,I\}$ for each customer $c \in \mathscr{C}$. Furthermore, let x_{ci} denote the price at the end point of the interval $i \in \mathscr{I}_0 = \mathscr{I} \cup 0 = \{0,1,...,I\}$ where $i = 0$ denotes the starting point of interval 1. Similarly, divide the LTV range into equally spaced intervals $\mathscr{J} = \{1,2,...,J\}$ for each customer $c \in \mathscr{C}$. Also, let y_{cj} denote the LTV at the end point of the interval $j \in \mathscr{J}_0 = \mathscr{J} \cup 0 = \{0,1,...,J\}$

where $j = 0$ denotes the starting point of the first interval. An approximation of the expected value of the NPII given in (4.9) at the grid point (x_{ci}, y_{cj}) is

$$f_{cij} := f(x_{ci}, y_{cj}) = R(x_{ci,}y_{cj})I(x_{ci}, y_{cj}) = R_{cij}I_{cij}. \tag{4.13}$$

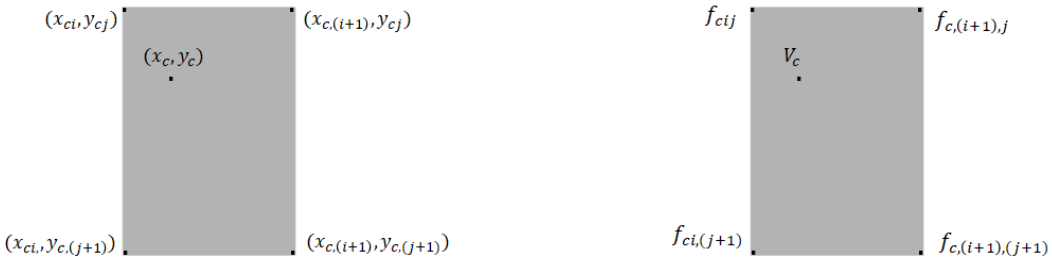
To interpolate between grid points, the price x_c and LTV y_c are expressed as convex combinations of the grid points (x_{ci}, y_{cj}), for all $i \in \mathscr{I}_0$ and $j \in \mathscr{J}_0$. For this purpose, the decision variable $\lambda_{cij} \in [0,1]$, with $\sum_{i\in\mathscr{I}_0}\sum_{j\in\mathscr{J}_0}\lambda_{cij} = 1$ is introduced such that

$$x_c = \sum_{i\in\mathscr{I}_0}\sum_{j\in\mathscr{J}_0} x_{ci}\lambda_{cij}, \tag{4.14}$$

$$y_c = \sum_{i\in\mathscr{I}_0}\sum_{j\in\mathscr{J}_0} y_{cj}\lambda_{cij}, \tag{4.15}$$

$$V_c = \sum_{i\in\mathscr{I}_0}\sum_{j\in\mathscr{J}_0} \lambda_{cij}f_{cij}, \tag{4.16}$$

with V_c the corresponding expected value of the NPII as defined in (4.13). Note that, to obtain an optimal solution when using the piece-wise linear approximation approach, not all of the λ_{cij} variables may take on a value larger than 0. More specifically, only the λ_{cij} variables corresponding to the vertices of the rectangle that contains the point (x_c, y_c) (*i.e.* the vertices of the single activated rectangle) may be allowed to take on a value (Misener and Floudas, 2010). The illustration in Figure 4.4(a) shows the grid points that form the vertices of the rectangle that contains the point (x_c, y_c). The corresponding function value V_c is then approximated by a convex combination of the function values evaluated at the vertices of the rectangle containing the point (x_c, y_c) as shown in Figure 4.4(b).



(a) (b)

Figure 4.4: (a) Convex combination of grid points; (b) Function approximation.

The activation of the appropriate rectangle is facilitated by the binary decision variable $\zeta_{cij} \in \{0,1\}$ where $i \in \mathscr{I}$ and $j \in \mathscr{J}$. For the illustration in Figure 4.4(a) and (b), by letting $\zeta_{c,(i+1),(j+1)} = 1$, the

optimal price, LTV and expected NPII are expressed as the following convex combination of gridpoints

$$x_c = x_{ci}\lambda_{cij} + x_{ci}\lambda_{ci,(j+1)} + x_{c,(i+1)}\lambda_{c,(i+1),j} + x_{c,(i+1)}\lambda_{c,(i+1),(j+1)}, \tag{4.17}$$

$$y_c = y_{cj}\lambda_{cij} + y_{c,(j+1)}\lambda_{ci,(j+1)} + y_{cj}\lambda_{c,(i+1),j} + y_{c,(j+1)}\lambda_{c,(i+1),(j+1)} \tag{4.18}$$

and

$$V_c = f_{cij}\lambda_{cij} + f_{ci,(j+1)}\lambda_{ci,(j+1)} + f_{c,(i+1),j}\lambda_{c,(i+1),j} + f_{c,(i+1),(j+1)}\lambda_{c,(i+1),(j+1)}, \tag{4.19}$$

respectively. In addition to this, the final selection of customers to be included in the credit portfolio is facilitated by introducing the binary decision variable $z_c \in \{0,1\}$, for each customer $c \in \mathscr{C}$. If $z_c = 1$, the customer will be made a loan offer at an interest rate of x_c and an LTV of y_c. However, if $z_c = 0$, the customer will not be made a loan offer and subsequently be excluded from the credit portfolio. Therefore, the NLP credit price and LTV optimisation problem formulated as a MILP problem and incorporating risk distribution constraints, LTV constraints and logical decision making capability is given by

$$\max \sum_{c\in\mathscr{C}} V_c \tag{4.20}$$

$$\text{s.t. } V_c \le \sum_{i\in\mathscr{I}_0}\sum_{j\in\mathscr{J}_0} \lambda_{cij} f_{cij} + (1-z_c)M, \qquad \forall\, c\in\mathscr{C}, \tag{4.21}$$

$$V_c \ge \sum_{i\in\mathscr{I}_0}\sum_{j\in\mathscr{J}_0} \lambda_{cij} f_{cij} - (1-z_c)M, \qquad \forall\, c\in\mathscr{C}, \tag{4.22}$$

$$V_c \le M z_c, \qquad \forall\, c\in\mathscr{C}, \tag{4.23}$$

$$V_c \ge -M z_c, \qquad \forall\, c\in\mathscr{C}, \tag{4.24}$$

$$x_c = \sum_{i\in\mathscr{I}_0}\sum_{j\in\mathscr{J}_0} x_{ci}\lambda_{cij}, \qquad \forall\, c\in\mathscr{C}, \tag{4.25}$$

$$y_c = \sum_{i\in\mathscr{I}_0}\sum_{j\in\mathscr{J}_0} y_{cj}\lambda_{cij}, \qquad \forall\, c\in\mathscr{C}, \tag{4.26}$$

$$t_c \le \sum_{i\in\mathscr{I}_0}\sum_{j\in\mathscr{J}_0} \lambda_{cij} R_{cij} + (1-z_c)M, \qquad \forall\, c\in\mathscr{C}, \tag{4.27}$$

$$t_c \ge \sum_{i\in\mathscr{I}_0}\sum_{j\in\mathscr{J}_0} \lambda_{cij} R_{cij} - (1-z_c)M, \qquad \forall\, c\in\mathscr{C}, \tag{4.28}$$

$$t_c \le M z_c, \qquad \forall\, c\in\mathscr{C}, \tag{4.29}$$

$$t_c \ge -M z_c, \qquad \forall\, c\in\mathscr{C}, \tag{4.30}$$

$$\sum_{j\in\mathscr{J}_0} \lambda_{c0j} \le \sum_{j\in\mathscr{J}} \zeta_{c1j} \qquad \forall\, c\in\mathscr{C}, \tag{4.31}$$

$$\sum_{j\in\mathscr{J}_0} \lambda_{cij} \le \sum_{j\in\mathscr{J}} (\zeta_{cij} + \zeta_{c,(i+1),j}) \qquad \forall\, c\in\mathscr{C}, \forall\, i\in\mathscr{I}/\{I\}, \tag{4.32}$$

$$\sum_{j\in\mathscr{J}_0} \lambda_{cIj} \le \sum_{j\in\mathscr{J}} \zeta_{cIj} \qquad \forall\, c\in\mathscr{C}, \tag{4.33}$$

$$\sum_{i\in\mathscr{I}_0} \lambda_{ci0} \le \sum_{i\in\mathscr{I}} \zeta_{ci1} \qquad \forall\, c\in\mathscr{C}, \tag{4.34}$$

$$\sum_{i\in\mathscr{I}_0} \lambda_{cij} \le \sum_{i\in\mathscr{I}} (\zeta_{cij} + \zeta_{ci,(j+1)}) \qquad \forall\, c\in\mathscr{C}, \forall\, j\in\mathscr{J}/\{J\}, \tag{4.35}$$

$$\sum_{i\in\mathscr{I}_0} \lambda_{ciJ} \le \sum_{i\in\mathscr{I}} \zeta_{ciJ} \qquad \forall\, c\in\mathscr{C}, \tag{4.36}$$

$$\sum_{i\in\mathscr{I}}\sum_{j\in\mathscr{J}} \zeta_{cij} = 1 \qquad \forall\, c\in\mathscr{C}, \tag{4.37}$$

$$\sum_{i\in\mathscr{I}_0}\sum_{j\in\mathscr{J}_0} \lambda_{cij} = 1 \qquad \forall\, c\in\mathscr{C}, \tag{4.38}$$

$$\sum_{c\in\mathscr{C}(g)} t_c \le U_g \sum_{c\in\mathscr{C}} t_c \qquad \forall\, g\in\mathscr{G}, \tag{4.39}$$

$$\sum_{c\in\mathscr{C}} \mathbb{I}\left(y_c \ge y^b\right) \le U_y \sum_{c\in\mathscr{C}} z_c. \tag{4.40}$$

In the optimisation model above, constraints (4.21)−(4.24) incorporate the logical decision making capability by allowing the exclusion of a customer from the credit portfolio through the use of the binary decision variable z_c. When $z_c = 1$, constraints (4.21) and (4.22) bind V_c to a convex combination of points, similar to (4.19) and therefore including the customer in the portfolio. However, when $z_c = 0$, constraints (4.23) and (4.24) bind V_c to 0 and therefore excluding the customer from the portfolio. In order to improve numerical stability, the constant M is assigned the value $\max(f_{cij})$. Furthermore, constraints (4.25) and (4.26) are included to obtain the values of the auxiliary variables x_c and y_c, similar to that of (4.17) and (4.18). Constraints (4.27)−(4.30) ensure that the auxiliary variable for the take-up, t_c, either takes on the value of 0 when the customer is excluded from the portfolio, or the value of a convex combination of points when included in the portfolio. Constraints (4.31)−(4.38) ensure that only a convex combination of weights associated with the single rectangle is activated for each customer $c \in \mathscr{C}$ and that these weights add up to 1. Lastly, constraints (4.39) and (4.40) are included to regulate the risk in the portfolio by setting upper bounds on the risk distribution and the proportion of loans with a LTV higher than a specified level, respectively. Note that, since customers can be excluded from the portfolio, the number of customers in the total portfolio (C) is replaced with the number of customers included in the final portfolio ($\sum_{c\in\mathscr{C}} z_c$) for the LTV constraint as seen in (4.40).

4.5 Model behaviour and computational results

In order to investigate the proposed credit price and LTV optimisation model, an empirical study was performed on a real-world data set from a financial institution. Due to the sensitive nature of the data, no details on the data or the institution itself could be made available. The results reported below are also normalised in order to disguise price and LTV characteristics. That is, the average price per risk category is expressed as a proportion of the risk category with the highest average price.

Furthermore, the results displayed in the tables below are summarised per risk grading $g \in \mathscr{G} = \{Low, Medium, High\}$. The values of the average optimal price per risk grading are expressed as a percentage of the risk grading(s) with highest average optimal price. Tables 4.2 and 4.3 summarise the model behaviour for the unconstrained ($U_g = \{1,1,1\}$) credit price and LTV optimisation problem for both of the optimisation models implemented *i.e.* the non-linear approach and the piece-wise linear approximation approach.

Table 4.2: Computational results (unconstrained): The non-linear approach.

Risk grading	Take-up proportion	Average price	Average LTV
Low	0.28	1.00	1.00
Medium	0.32	0.98	1.00
High	0.40	0.94	1.00

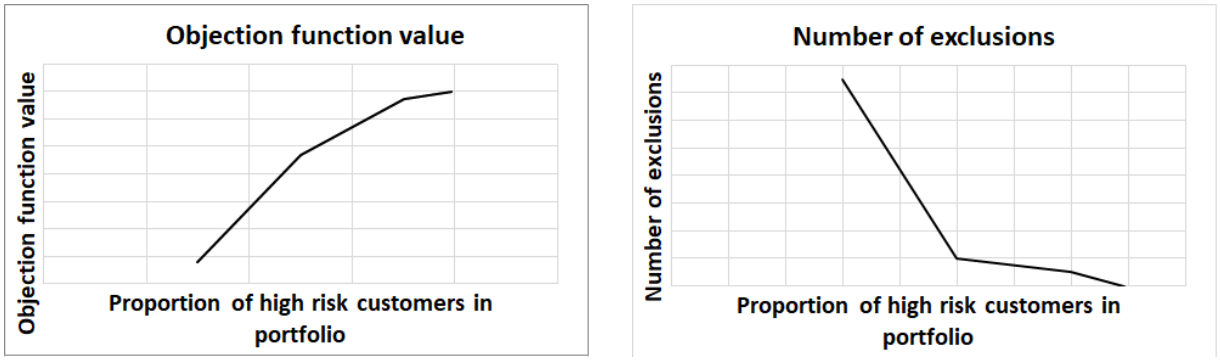
Table 4.3: Computational results (unconstrained): The piece-wise linear approach.

Risk grading	Take-up proportion	Average price	Average LTV
Low	0.27	1.00	1.00
Medium	0.33	0.97	1.00
High	0.40	0.93	1.00

Considering the results in Tables 4.2 and 4.3, it is evident that the average optimal price and LTV per risk grading are similar over the portfolio and also similar for the two different approaches followed. These results serve as a validation for the use of a piece-wise linear approximation approach. Therefore, with the added benefit of including logical decision-making capability and constraints on both the risk distribution and LTV, the piece-wise linear approximation approach is the only approach to be considered henceforth.

In order to illustrate the behaviour of the price and LTV optimisation model when high-risk customers are excluded from the final portfolio, the objective function value and the corresponding number of customers excluded are reported below for a range of high-risk grading constraint values. That is, the take-up proportion of high-risk customers in the portfolio were restricted to 0.05, 0.15, 0.25 and 0.35. The graphs below illustrate the effect of these constraints with respect to objective function values and the number of high-risk customers excluded.

From Figure 4.5(a) and (b) it can be seen that as the proportion of high-risk customers in the portfolio decreases, the objective function value decreases and the number of customers excluded increases. The results in Table 4.4 demonstrate what the impact is on the price and LTV when the constraints $U_g = \{1, 0.3, 0.2\}$ are imposed on the risk distribution.



(a) (b)

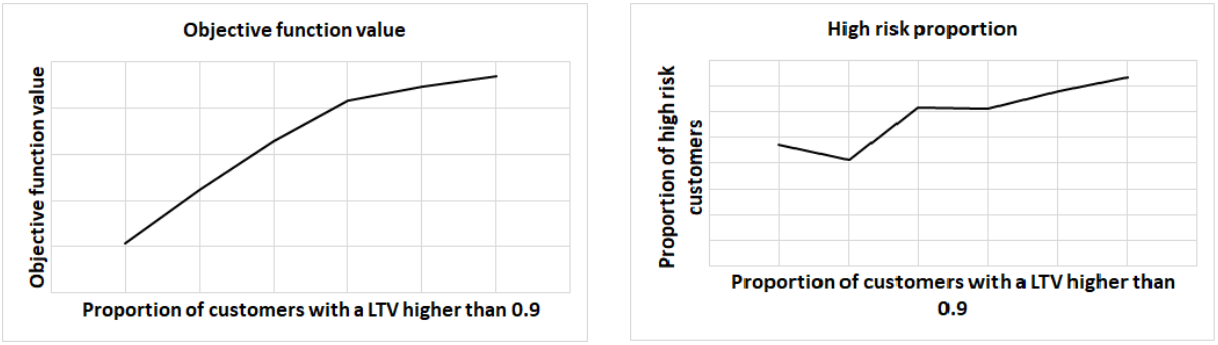
Figure 4.5: (a) Impact of logical decision making capability on objective function value; (b) Impact of logical decision making capability on number of exclusions.

Table 4.4: Computational results (risk constrained): The piece-wise linear approach.

Risk grading	Take-up proportion	Average price	Average LTV
Low	0.50	0.63	1.00
Medium	0.30	0.88	1.00
High	0.20	1.00	0.92

From the results in Table 4.4, it is clear that loans to be offered to high-risk customers will have a higher average price and a lower average LTV compared to the lower and medium risk customers. The expectation is that higher risk customers will be discouraged from taking up a loan if the price is too high and the LTV is too low. Conversely, if lower and medium risk customers are offered lower rates and higher LTV values, it may result in an improved take-up by these customers.

Furthermore, to illustrate the impact of upper bounds on the proportion of loans in the credit portfolio with an LTV larger than or equal to 0.9, the objective function value corresponding to the different upper bounds imposed on the loans with a higher LTV, is investigated. More specifically, the proportion of loans included in the final portfolio with an LTV larger than or equal to 0.9 were restricted to less than or equal to 50%, 60%, 70%,80%, 90% and 100% and the impact of these restrictions on the objective function value is investigated. Since this constraint is included to ultimately reduce the risk in the portfolio, the take-up proportion of high risk customers included in the final portfolio, when limiting the proportion of loans with a high LTV, were investigated. Figure 4.6(a) and (b) display the effect of the imposed LTV constraints on the objective function value and the proportion of high risk customers in the credit portfolio, respectively. Note that, to illustrate the effect of the LTV constraints, no constraints were imposed on the risk distribution.



(a) (b)

Figure 4.6: (a) Impact of LTV constraints on objective function value; (b) Impact of LTV constraints on the proportion of high risk customers;

Recalling the relationship between LTV and take-up probability displayed in Figure 4.3, the objective function value is expected to decrease as the proportion of loans in the credit portfolio with an LTV larger than or equal to 0.9 decreases, which is confirmed when referring to Figure 4.6(a). On the other hand, Figure 4.6(b) illustrates that a decrease in the proportion of higher LTV loans (*i.e.* for this

data set roughly 60%) corresponds to a decrease in portion of high risk customers, hence, reducing the risk in the credit portfolio. The results in Table 4.5 demonstrate the impact of imposing a upper bound of 60% on the proportion of loans included in the final portfolio with a LTV larger than or equal to 0.9.

Table 4.5: Computational results (LTV constrained): The piece-wise linear approach.

Risk grading	Take-up proportion	Average price	Average LTV
Low	0.32	1.00	0.89
Medium	0.35	0.95	0.88
High	0.33	0.92	0.76

The results display a clear reduction in the average LTV in each of the risk categories with the average price remaining similar to that of the unconstrained computational results displayed in Table 4.3 across the risk gradings. Moreover, the average LTV of the high risk customers are significantly lower, compared to that of the low and medium risk customers. In addition to this, the take-up proportion of high risk customers is lower compared to the unconstrained problem, whereas the take-up proportion for the medium and low risk customers are higher, thus confirming the reduction in the risk of the credit portfolio when imposing the LTV constraints.

To illustrate the impact of the constraints on the risk distribution together with the constraints on the LTV, the results displayed in Table 4.6 below, are considered. Here, the constraints $U_g = \{1, 0.3, 0.2\}$ are imposed on the risk distribution whereas the constraints on the LTV are set to $U_y = 60\%$ and $y^b = 0.9$. That is, the take-up proportion in the credit portfolio of the medium and high risk gradings are limited to 0.3 and 0.2, respectively, whereas the proportion of loans with an LTV larger than or equal to 0.9 are limited to 60% in the credit portfolio.

Table 4.6: Computational results (risk and LTV constrained): The piece-wise linear approach.

Risk grading	Take-up proportion	Average price	Average LTV
Low	0.50	0.58	0.98
Medium	0.30	0.92	0.94
High	0.20	1.00	0.72

These results indicate that on average, a higher price is offered for the higher risk customers together with a lower average LTV, thus limiting the take-up proportion of high risk customers by reducing the take-up probability of these customers and subsequently reducing the risk in the portfolio. Conversely, the low and medium risk customers are offered a lower price on average together with a higher average LTV, hence, increasing the take-up probability and ultimately the take-up proportion of these risk gradings.

Two different approaches were used in the simultaneous optimisation of price and LTV when maximising the expected value of the NPII to the lender. Although both approaches yielded similar results

for the unconstrained case, the piece-wise linear approximation approach always provides proven optimal solutions and, in addition to this, it allows for the inclusion of binary decision variables which facilitate logical decision-making capability on a portfolio level. Furthermore, by imposing constraints on the risk distribution and the proportion of loans with higher LTVs, the risk in the portfolio can be managed. The results suggest that by increasing the average price per risk grading in conjunction with decreasing the average LTV, the risk in the credit portfolio is reduced. Hence, there is a clear interaction between price, LTV and the risk inherent in the credit portfolio. Therefore, to regulate the risk in a credit portfolio, the lender should offer a higher price on average for the higher risk customers together with a lower LTV, whereas for the medium and lower risk customers, a lower price should be offered with a higher LTV.

The current approach, however, makes various simplifying assumptions when using the approximation of the expected NPII in (4.2) instead of (4.1). Firstly, the approximation in (4.2) is only dependent on the overall probability of default p of the customer and not the probability that a customer defaults in a specific month t. Hence, the probability that the customer will make the payment in month t, denoted by s_t, is not taken into account, that is, the survival probability until month t is not considered. Furthermore, it is assumed $r_d \cong 0$, $\frac{(1+r)^n}{(1+r)^n-1} \approx 1$ and $\frac{(1+r_0)^n}{(1+r_0)^n-1} \approx 1$ for large values of n and that if the loan is taken up, it will be repaid with probability 1.

In a more realistic setup these assumptions need to be relaxed, since with most loans, the lender faces the risk (at the time of funding the loan) that the customer may default at some point during the term of the loan, *i.e.* stop making the monthly repayments. Therefore in the next chapter, various survival models that may be used to model the probability that a customer will make the payment in a specific month (or not default before a specific month), are discussed.

Chapter 5

Survival analysis models

5.1 Introduction

The current approach to determine the optimal price and LTV that maximises the expected NPII, assumes the loan is to be repaid in full with probability one. However, in practice this is not a realistic assumption. Various events can occur before the loan is fully repaid, and hence before the net present interest income is received by the financial institution. These events include early settlement of the loan, default on the loan or prepayments of the loan that could lead to early settlement. However, the focus will only be on how default can be modelled and incorporated into the credit price and LTV optimisation problem.

Analysing or trying to predict when a customer is likely to default is similar to time-to-event modelling in, *e.g.*, medical sciences and engineering. In these fields the use of survival analysis has been well established and proven to produce models with desirable properties for time-to-event data (see *e.g.* Wei, 1992, Ma and Krings, 2008 and Tolley et al., 2016). Survival analysis was first introduced in the credit environment by Narain (1992), where they made use of a parametric accelerated failure time (AFT) model. Since then a number of authors have started to use more advanced survival analysis models. Some of these include Banasik et al. (1999), which extended the use of the AFT model and also included the non-parametric Cox Proportional Hazard (CPH) model and Bellotti and Crook (2009), which allowed for time varying covariates in the CPH model. Tong et al. (2012), Dirick et al. (2015) and Dirick et al. (2017) introduced the mixture cure model as a more general alternative to the CPH model, for modelling data in the credit risk environment.

One of the main advantages of using survival analysis (compared to using *e.g.* logistic regression) is that one can predict not only whether borrowers will default on their loans, but also when they are likely to default. That is, by using survival analysis the probability that a borrower is still repaying a loan at every time instant of the survival curve, *e.g.* every month, can be accurately estimated.

5.2 Basic concepts

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

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

and

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

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

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

where $f(t)$ is the density function of Y. Depending on the distribution of Y the hazard rate can have many different shapes (constant, increasing, decreasing and non-monotone) and is therefore a very useful tool to summarise survival data. Figure 5.1 depicts examples of all 4 these different shapes.

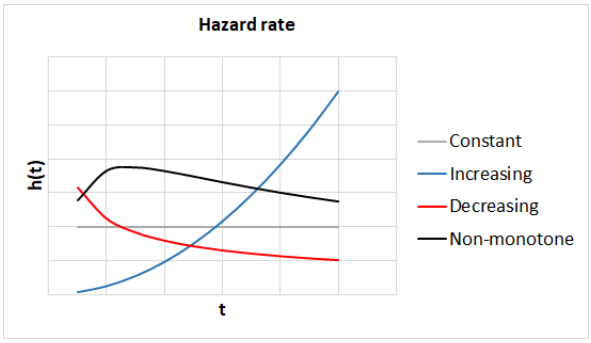


Figure 5.1: Hazard rate shapes.

The cumulative hazard function,

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

is an increasing function on $[0, \infty]$ and represents the accumulated risk up to a time t. Figure 5.2 shows the cumulative hazard rates corresponding to the hazard rates in Figure 5.1.

The following relations exist between these various functions

$$f(t) = -\frac{d}{dt}S(t),$$
$$h(t) = -\frac{d}{dt}\log S(t),$$
$$H(t) = -\log S(t)$$

and

$$S(t) = e^{-H(t)}.$$

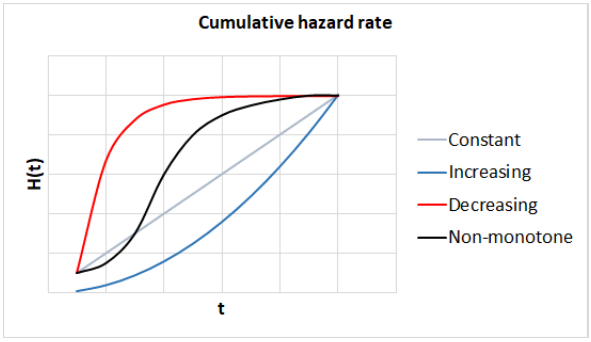


Figure 5.2: Cumulative hazard rate shapes.

In survival analysis a certain proportion of the individuals are censored. This means that some information about the individual's event time (default time) is known, but the exact event time (default time) is not known. The second definition of censoring in Dirick et al. (2017) will be used, which states that a customer who did not experience default by the moment of data gathering, corresponds to a censored case. In this definition mature cases and early settlement cases are considered censored since the only event of interest is default. Mature cases refer to loans that are repaid in time over the full term whereas early settlement cases refer to loans that are repaid before the end of the full term. Therefore, according to this definition of censoring, only two possible states are considered; default and censored. This censoring scheme is referred to as right censoring and is graphically depicted in Figure 5.3. In this figure, Λ denotes the censoring time, *e.g.* the time of data gathering.

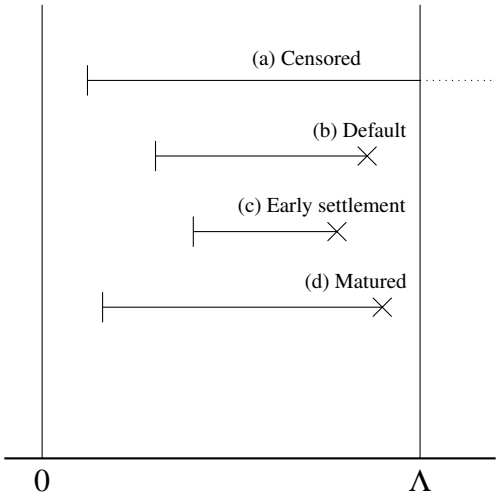


Figure 5.3: Right censoring scheme.

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

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

and

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

The following two assumptions are made:

Assumption 5.1. *Assume that Y and C are independent (this is a standard assumption in survival analysis, see, e.g. Part I of Klein et al., 2016).*

Assumption 5.2. *Assume that the censoring is non-informative, i.e. the distribution of C does not depend on the parameters of interest related to $S(t)$.*

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

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

and

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

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

From straight forward calculations and using Assumptions 5.1 and 5.2 one obtains the very well known likelihood of (T_i, δ_i),$i = 1,2,\ldots,n$ given by

$$L = \prod_{i=1}^{n} f(T_i)^{\delta_i} S(T_i)^{1-\delta_i} \tag{5.1}$$

$$= \prod_{i=1}^{n} S(T_i) h(T_i)^{\delta_i}. \tag{5.2}$$

The likelihood in (5.2) can be interpreted as follows:

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

The corresponding log-likelihood is given by

$$\log L = \sum_{i=1}^{n} log\left[S(T_i)\right] + \sum_{i=1}^{n} \delta_i log\left[h(T_i)\right]. \tag{5.3}$$

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

Any distribution that is defined on $t \in [0,\infty)$, can potentially serve as a lifetime distribution for Y. However, through the years, some distributions were observed to be able to more accurately model the properties of realised survival data (see, *e.g.*, Klein and Moeschberger, 2006). Some of these important distributions will now be reviewed, by giving expressions for the density, survival and hazard functions as well as some other interesting properties (all distributions are defined on $t \geq 0$).

1. Exponential distribution:
   Y is exponentially distributed with parameter $\lambda > 0$, denoted by $Y \sim exp(\lambda)$ if

   $$f(t) = \lambda e^{-\lambda t},$$
   $$S(t) = e^{-\lambda t}$$

   and

   $$h(t) = \lambda.$$

   The exponential distribution is the only distribution that has a constant hazard rate. An immediate consequence of this is the memoryless property of the exponential distribution which states that,

if Y follows an exponential distribution, then

$$P(Y > s+t|Y > s) = P(Y > t),$$

for $s,t > 0$.

2. Weibull distribution:
   Y is Weibull distributed with parameters $\lambda > 0$ and $\alpha > 0$, denoted by $Y \sim W(\lambda, \alpha)$ if

$$\begin{aligned} f(t) =& \alpha\lambda t^{\alpha-1} e^{-\lambda t^{\alpha}}, \\ S(t) =& e^{-\lambda t^{\alpha}} \end{aligned}$$

   and

$$h(t) = \alpha\lambda t^{\alpha-1}.$$

   The hazard rate is increasing for $\alpha > 1$ and decreasing for $0 < \alpha < 1$. For $\alpha = 1$ the exponential distribution is obtained and subsequently a constant hazard rate.

3. Gamma distribution:
   Y is Gamma distributed with parameters $\lambda > 0$ and $\alpha > 0$, denoted by $Y \sim \Gamma(\lambda, \alpha)$ if

$$f(t) = \frac{\lambda^{\alpha} t^{\alpha-1} e^{-\lambda t}}{\Gamma(\alpha)},$$

   where

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx.$$

   Furthermore

$$h(t) = \frac{f(t)}{S(t)},$$

   where

$$S(t) = 1 - \int_0^t f(x) dx.$$

   The hazard rate is increasing for $\alpha > 1$ and decreasing for $0 < \alpha < 1$. For $\alpha = 1$ the exponential distribution is obtained and therefore also a constant hazard rate.

4. Log-normal distribution:
   Y is log-normal distributed with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ denoted by $Y \sim LN(\mu, \sigma^2)$ if

$$f(t) = \frac{1}{\sqrt{2\pi}\,\sigma t} exp\left[-\frac{1}{2} \left(\frac{\ln t - \mu}{\sigma} \right)^2 \right],$$

$$S(t) = 1 - \Phi\left[\frac{\ln t - \mu}{\sigma}\right],$$

where Φ is the cumulative distribution function of the normal distribution and

$$h(t) = \frac{f(t)}{S(t)}.$$

Other well-known survival distributions include the Rayleigh, linear failure rate, Gompertz-Makeham, logistic and log-logistic distributions.

5.3 Estimating the survival function

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

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

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

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

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

However, the observed data in this case are not complete, but rather censored data of the form (T_i, δ_i), $i = 1, 2, \ldots, n$, meaning that the censoring will need to be taken into account when estimating $S(t)$. Kaplan and Meier (1958) introduced the product limit estimator to address this issue when working with right censored data. Let $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}$ and $T_{(1)} < T_{(2)} < \cdots < T_{(n)}$ denote the order statistics of $Y_1, Y_2, \ldots, Y_n$ and $T_1, T_2, \ldots, T_n$, respectively, and where $\delta_{(i)}$ is the corresponding censoring indicator variable associated with $Y_{(i)}$ and $T_{(i)}$. The Kaplan-Meier estimator is given by

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

This estimator is also a discrete step function, but the jumps now occur only at the event times (in our case the default times). If the largest observation $T_{(n)}$ is censored then $\hat{S}_n(t)$ does not attain 0. Various solutions have been proposed in the literature for this (see *e.g.*, Efron, 1967), namely set $\hat{S}_n(t) = 0$ for $t \geq T_{(n)}$ or set $S_n(t) = \hat{S}_n(T_{(n)})$ for $t \geq T_{(n)}$. Figure (5.4) displays an example of this estimator using

a censored data set. The Kaplan-Meier estimator has been studied in detail by many authors in the literature, including Efron (1967) and Breslow and Crowley (1974). This estimator is also available in most of the well-known statistical software packages, *e.g.*, the *'survival'* package in R (see Therneau, 2020).

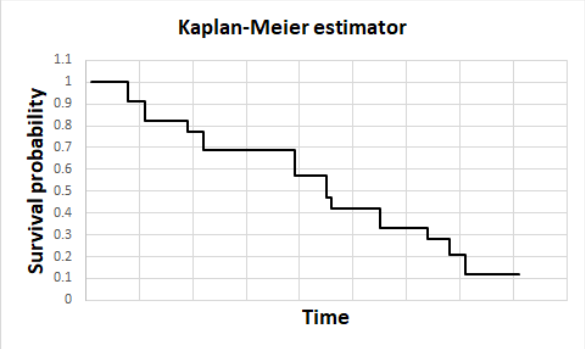


Figure 5.4: The Kaplan-Meier estimator of the survival probability.

5.3.2 Parametric estimation of $S(t)$

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

Example 5.3. Suppose the assumption is made that the event times are exponentially distributed with unknown parameter λ, *i.e.*

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

By using the log-likelihood given in (5.3), λ can now be estimated;

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

Maximising the log-likelihood yields the estimator

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

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

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

By specifying a parametric form for $S(t)$, a smooth function for estimating the survival distribution is obtained. Furthermore, if these parametric models fit the data well, the models tend to give more precise estimates of the quantities of interest. However, some form of goodness-of-fit testing should be conducted after fitting the parametric models to ensure that the distribution fits to the data. Popular distributions for estimating survival distributions include those discussed in Section 5.2. The following are some of the available survival distributions in the *'survival'* package in R: Weibull, exponential, logistic, lognormal and loglogistic.

Both the non-parametric and parametric estimates are only based on event and censoring times. In the vast majority of times, the event time (say default) is a function of one or more covariates *e.g.*, income, risk category, loan amount and interest rate. There exists various models that are able to take these covariates into account when modelling the survival functions. Two of the most widely used models are the accelerated failure time (AFT) model and the Cox Proportional Hazard (CPH) model. In the next section the CPH model will be discussed. For some detail in the AFT model, the interested reader is referred to Chapters 2 and 12 of Klein and Moeschberger (2006).

5.4 Cox Proportional Hazards model

The assumption that has been made thus far is that the survival functions of the individual borrowers are identical. However, in many circumstances this assumption is not reasonable. If, for example, in the medical sciences, there is an interest in the time required for an individual to be cured from a certain disease, factors (called covariates) can have an influence why one person might take longer (or shorter) to be cured. These covariates may include age, fitness level and smoking status. Similarly, if interested in when a borrower is likely to default on a loan or the time until default, various covariates can influence the time to default. These covariates may include loan amount, risk category, interest rate, loan-to-value, whether or not the borrower is employed, duration at current employment and debt to income ratio. It is thus important that the probability of survival (or probability of default) is modelled, taking into account these covariates. Cox (1972) proposed a proportional hazard model which does exactly this, and has become the most commonly used regression model for survival data.

The CPH model is given by

$$h(t|\underline{w}) = h_0(t)e^{\underline{\beta}^T\underline{w}}, \tag{5.4}$$

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

Remarks:

1. The vector of covariates may include continuous factors (*e.g.* the interest rate and the loan amount), discrete factors (*e.g.* employment status and risk category) and possible interaction terms (*e.g.* interest rate and risk category interaction).

2. $h_0(t)$ is called the baseline hazard function, because it denotes the hazard function for borrowers with all covariates equal to 0. That is, $h(t|\underline{w} = \underline{0}) = h_0(t)$, where $\underline{w} = (w_1, w_2, \ldots, w_m)^T = \underline{0} = (0, 0, \ldots, 0)^T$.

3. The CPH model can also be formulated as follows, in terms of the conditional cumulative hazard rate function,
$$H(t|\underline{w}) = H_0(t)e^{\underline{\beta}^T \underline{w}},$$
where $H_0(t) = \int_0^t h_0(u)du$ is the baseline cumulative hazard rate function.

4. Using the relationships between the hazard rate, cumulative hazard rate and survival function, the conditional survival function can be written as
$$\begin{aligned} S(t \mid \underline{w}) &= e^{-H(t|\underline{w})} \\ &= e^{-H_0(t)e^{\underline{\beta}^T \underline{w}}} \\ &= S_0(t)^{e^{\underline{\beta}^T \underline{w}}}, \end{aligned}$$
where $S_0(t)$ is the baseline survival function (with associated hazard function $h_0(t)$ and associated cumulative hazard function $H_0(t)$).

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

From (5.2), the likelihood function based on the data $(T_i, \delta_i, \underline{w}_i)$ and conditional on $\underline{w}_1, \underline{w}_2, \ldots, \underline{w}_n$ is
$$L\left(\underline{\beta} \mid T_i, \delta_i, \underline{w}_i\right) = \prod_{i=1}^{n} S\left(T_i \mid \underline{w}_i\right) \left[h\left(T_i \mid \underline{w}_i\right)\right]^{\delta_i}, \tag{5.5}$$
where
$$S\left(T_i \mid \underline{w}_i\right) = S_0(T_i)^{e^{\underline{\beta}^T \underline{w}_i}}.$$
Using the relationship
$$h(T_i|\underline{w}_i) = h_0(T_i)e^{\underline{\beta}^T \underline{w}_i},$$

(5.5) becomes

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

The corresponding log-likelihood is then given by

$$l(\underline{\beta}) = \log\left[L\left(\underline{\beta} \mid T_i, \delta_i, \underline{w}_i\right)\right] = -\sum_{i=1}^{n} H_0(T_i) e^{\underline{\beta}^T \underline{w}_i} + \sum_{i=1}^{n} \delta_i \log h_0(T_i) + \sum_{i=1}^{n} \delta_i \underline{\beta}^T \underline{w}_i. \tag{5.6}$$

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

5.4.1 Semi-parametric estimation of the CPH model

The CPH is considered a semi-parametric model if no parametric assumptions are made about the form of the baseline hazard function ($h_0(t)$) but a parametric assumption is made regarding the effect of the covariates on the hazard rate function (in this case the functional form is $e^{\underline{\beta}^T \underline{w}}$). In the semi-parametric setting, Cox (1972) proposed a partial likelihood approach to estimate the vector of parameters $\underline{\beta}$, without specifying $h_0(t)$ (see, *e.g.*, Cox, 1975). The baseline hazard function (or cumulative baseline hazard function) can then be estimated non-parametrically by the estimator introduced by Breslow (1972). This estimator is obtained by maximising the likelihood of $h_0(t)$ in which the parameters $\underline{\beta}$ are replaced by the maximum partial likelihood estimators $\underline{\hat{\beta}} = \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_m\right)^T$. Denote the estimate by $\hat{h}_0(t)$ and the corresponding estimate for $H_0(t)$ by $\hat{H}_0(t)$ (for more detail on this estimator the interested reader is referred to Breslow, 1972 and Kalbfleisch and Prentice, 1973).

The semi-parametric estimated CPH model is given by

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

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

$$\hat{S}(t \mid \underline{w}) = \hat{S}_0(t) e^{\underline{\hat{\beta}}^T \underline{w}}$$
$$= e^{-\hat{H}_0(t) e^{\underline{\hat{\beta}}^T \underline{w}}}. \tag{5.7}$$

As can be seen from (5.7), if one estimates the model by making use of the semi-parametric ap-

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

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

5.4.2 Parametric estimation of the CPH model

In the parametric CPH model an additional distribution assumption is made regarding the baseline hazard function. The assumption is that the baseline distribution is a known lifetime distribution (*e.g.* one of the distributions discussed in Section 5.2), but with unknown parameters. The log-likelihood given in (5.6) can be used to estimate the CPH model. The following example illustrates this idea.

Example 5.4. Suppose a parametric CPH model is to be fitted to the observed data $(T_i, \delta_i, \underline{w}_i)$, $i = 1, 2, \ldots, n$, and the assumption is made that the baseline distribution is Weibull with unknown parameters $\lambda > 0$ and $\alpha > 0$. From (5.6), the log-likelihood is given by

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

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

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

It is clear from this example that a closed-form expression for the conditional survival function is obtained, which is both easy to implement and understand. However, the benefits of using a parametric instead of a semi-parametric approach requires an additional assumption. It is therefore impor-

tant to test the validity of this assumption before implementing the model in practise. Diagnostic (or goodness-of-fit) tests for this assumption are available in the form of simple graphical techniques, like plotting the martingale, deviance or Schoenfeld residuals (see, *e.g.*, Klein and Moeschberger, 2006), as well as formal hypothesis tests. One such formal test is based on the Cox-Snell residuals (defined below) which approximately follow a standard exponential distribution when the model is correctly specified. The reason for this is as follows:

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

$$\begin{aligned} H(t|\underline{w}_i) &= -\log S(t \mid \underline{w}_i) \\ &= H_0(t)e^{\underline{\beta}^T \underline{w}_i} \\ &= H_0(t;\underline{\theta})e^{\underline{\beta}^T \underline{w}_i}, \end{aligned} \tag{5.8}$$

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

If the model in (5.8) is correct and $S(t \mid \underline{w}_i)$ is the true conditional survival function of Y_i, $i = 1,2,\ldots,n$, then, $S(Y_i \mid \underline{w}_i)$ follows a uniform $(0,1)$ distribution. This follows from the well-known probability integral transform which states that, if a random variable X has distribution function F and survival function S, then $F(X) = U$,where U has a uniform $(0,1)$ distribution. Since $S(t) = 1 - F(t)$, the same holds true for $S(X)$. From straight forward calculations it follows that $H(Y_i|\underline{w}_i) = -\log S(Y_i \mid \underline{w}_i)$ has a standard exponential distribution.

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

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

which can be expressed as

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

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

If the cumulative baseline hazard is correctly specified, then the Cox-Snell residuals should approximately follow a standard exponential distribution (see, *e.g.*, Chapter 11 of Klein and Moeschberger, 2006 for more detail). One can use any of the multitude tests for exponentiality to test whether this is the case. This study lead to the development of a new test for exponentiality based on a conditional moment characterisation. This new test is applied to testing exponentiality for i.i.d. data and also for testing the exponentiality of the Cox-Snell residuals in a complete sample in the CPH models. More detail can be found in the published article in Appendix A.

5.5 Mixture cure models

In survival analysis it often happens that, for a certain subgroup of the subjects involved in the study, the event of interest does not happen during the time interval under consideration. This subgroup is referred to as the 'cured' group. The population of subjects therefore consists of two subgroups; the 'cured' group which is not susceptible to the event of interest as well as a group which is susceptible to the event of interest. As an example, consider a hypothetical loan portfolio of a very small financial institution. The portfolio contains the data of 80 customers that were all awarded home loans on the 1st of January 1999. The event of interest here is whether or not the customer will default on their loan. Upon studying the portfolio, it is found that only 19 customers defaulted on their loans. Figure 5.5 show the survival probabilities (*i.e.* the probability of not defaulting on the loan) based on this study.

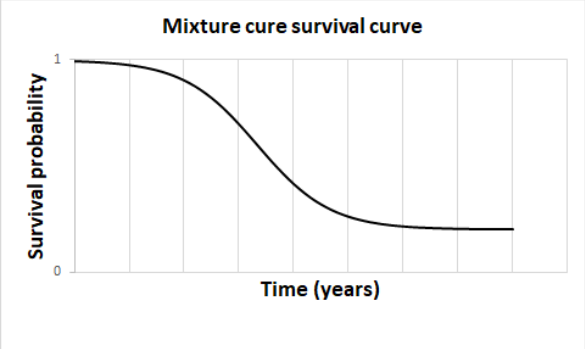


Figure 5.5: Illustration of mixture cure survival curve.

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

Mixture cure models provide a widely used alternative to CPH models (see, *e.g.*, Amico and Van Keilegom, 2018) if the survival trend of the data is similar to the trend shown in Figure 5.5 (*i.e.* a certain proportion of the population is not expected to experience the event of interest). As the name suggests, the mixture cure model is used to model the survival probability using a mixture of the two subgroups in the population (*i.e.* the 'cured' and the 'not cured' subgroups). Crudely stated, the mixture cure model represents the following survival calculation,

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

or in terms of default as the event of interest,

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

The model therefore incorporates two components, namely

1. An incidence model to predict which subjects are susceptible. This is typically modelled by making use of a logistic regression model.

2. A latency model that is used to predict the subjects' survival times conditional on the fact that they are susceptible. The CPH model is commonly used to model this component (either with the use of a semi-parametric or parametric approach).

Mixture cure models were first introduced to the field of medical statistics by Farewell (1982), whereafter Kuk and Chen (1992) and Sy and Taylor (2000) generalised some of the results. In addition, Tong et al. (2012) and Dirick et al. (2015) recently applied mixture cure models to the credit scoring environment. Other applications of cure models can be found in Peng and Dear (2000), Li and Taylor (2002), Liu and Shen (2009) and Ma (2009). For a comprehensive review of mixture cure models, see the overview paper of Amico and Van Keilegom (2018).

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

5.5.1 The model formulation

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

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

and

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

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

1. $\Upsilon = 1$ and $\delta = 1$, indicate that the customer is susceptible and uncensored and therefore the event of interest took place, *i.e.* the customer defaulted on the loan during the observation period.

2. $\Upsilon = 1$ and $\delta = 0$, indicate that the customer is susceptible and censored and therefore no event took place, *i.e.* the customer did not default during the observation period but will eventually default.

3. $\Upsilon = 0$ and $\delta = 0$, indicate that the customer is not susceptible and censored and therefore no event will take place, *i.e.* the customer did not default during the observation period and will not default.

4. $\Upsilon = 0$ and $\delta = 1$, this event cannot be observed since default is impossible if the customer is not susceptible.

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

$$S(t \mid \underline{v}, \underline{w}) = \pi(\underline{v}) S(t \mid \Upsilon = 1, \underline{w}) + [1 - \pi(\underline{v})], \tag{5.9}$$

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

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

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

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

using the same notation as in Section 5.4.

It should be noted that, when using the mixture cure model, the unconditional survival function tends to $1-\pi(\underline{v})$ if $t \to \infty$. If, however, $\pi(\underline{v}) = 1$, the unconditional survival function of the mixture cure model reduces to the standard CPH model as is discussed in Section 5.4.

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

From (5.1), the likelihood function based on the data $(T_i, \delta_i, \underline{v_i}, \underline{w_i})$, conditional on $\underline{v_i}$ and $\underline{w_i}$, is given by

$$L\left(\underline{\eta}, \underline{\beta} \mid T_i, \delta_i, \underline{v_i}, \underline{w_i}\right) = \prod_{i=1}^{n} \{\pi(\underline{v_i}) f(T_i \mid \Upsilon_i = 1, \underline{w_i})\}^{\delta_i} \{\pi(\underline{v_i}) S(T_i \mid \Upsilon_i = 1, \underline{w_i}) + [1-\pi(\underline{v_i})]\}^{1-\delta_i}. \tag{5.10}$$

By using the relationship

$$f(T_i \mid \Upsilon_i = 1, \underline{w_i}) = S(T_i \mid \Upsilon_i = 1, \underline{w_i})\, h(T_i \mid \Upsilon_i = 1, \underline{w_i})$$

the likelihood in (5.10) becomes

$$\begin{aligned} L\left(\underline{\eta}, \underline{\beta} \mid T_i, \delta_i, \underline{v_i}, \underline{w_i}\right) = \prod_{i=1}^{n} &\{\pi(\underline{v_i}) S(T_i \mid \Upsilon_i = 1, \underline{w_i})\, h(T_i \mid \Upsilon_i = 1, \underline{w_i})\}^{\delta_i} \times \\ &\{\pi(\underline{v_i}) S(T_i \mid \Upsilon_i = 1, \underline{w_i}) + [1-\pi(\underline{v_i})]\}^{1-\delta_i}. \end{aligned} \tag{5.11}$$

This likelihood can be interpreted as follows:

- If $\delta_i = 1$, customer i is uncensored (*i.e.* defaulted during the observation period) and therefore susceptible to default and the likelihood contribution is $\pi(\underline{v_i}) S(T_i \mid \Upsilon_i = 1, \underline{w_i})\, h(T_i \mid \Upsilon_i = 1, \underline{w_i})$. The likelihood contribution is thus the probability of being susceptible to default multiplied by the probability of surviving until time T_i multiplied by the instantaneous risk of default immediately after time T_i, given the customer is susceptible to default.

- If $\delta_i = 0$, customer i is censored (*i.e.* did not default during the observation period) and can either be susceptible or not susceptible to default, therefore the likelihood contribution is $\pi(\underline{v_i}) S(T_i \mid \Upsilon_i = 1, \underline{w_i}) + (1-\pi(\underline{v_i}))$. That is, the likelihood contribution is the probability of being susceptible to default multiplied by the probability of surviving until time T_i, given the customer is susceptible to default and the probability that the customer is not susceptible to default times the probability of surviving until time T_i, which is 1, since the customer is not susceptible to default.

From the CPH model the likelihood in (5.11) becomes

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

The corresponding log-likelihood is then given by

$$\begin{aligned} l(\underline{\eta},\underline{\beta}) &= \log\left[L\left(\underline{\eta},\underline{\beta} \mid T_i,\delta_i,\underline{v}_i,\underline{w}_i\right)\right] \\ &= \sum_{i=1}^{n}\delta_i\left\{\log\pi(\underline{v}_i) - H_0(T_i \mid \Upsilon_i=1)e^{\underline{\beta}^T\underline{w}_i} + \log h_0(T_i \mid \Upsilon_i=1) + \underline{\beta}^T\underline{w}_i\right\} + \\ &\quad \sum_{i=1}^{n}(1-\delta_i)\log\left\{\pi(\underline{v}_i)e^{-H_0(T_i|\Upsilon_i=1)e^{\underline{\beta}^T\underline{w}_i}} + [1-\pi(\underline{v}_i)]\right\}, \end{aligned} \tag{5.12}$$

where

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

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

5.5.2 Semi-parametric estimation of the mixture cure model

A mixture cure model is considered a semi-parametric model if no parametric assumptions are made about the form of the baseline hazard function. However, a parametric assumption is made regarding the effect of the covariates on the hazard rate function (in this case the functional form is $e^{\underline{\beta}^T\underline{w}}$) and the effect of the covariates on the probability of being susceptible (with functional form $\log\left(\pi(\underline{v})/(1-\pi(\underline{v}))\right) = \underline{\eta}^T\underline{v}$). In this study, the focus is on the parametric estimation of the mixture cure model, however, Cai et al. (2012) developed a R-package, *'smcure'*, that could be used to estimate the mixture cure model semi-parametrically. This semi-parametric estimation procedure uses the Expectation Maximisation (EM) algorithm to deal with the Υ values that are latent (recall that the Υ values are not observed for the censored cases) and the estimation procedure is based on a partial likelihood method proposed by Peng and Dear (2000) and Sy and Taylor (2000), to estimate the parameters of the conditional survival function without specifying the baseline hazard function. The interested reader is referred to Peng (2003), Sy and Taylor (2000) and Taylor (1995) for more information on the semi-parametric estimation of the mixture cure models.

5.5.3 Parametric estimation of the mixture cure model

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

Below, it is assumed that the baseline distribution is a known lifetime distribution (*e.g.* one of the distributions discussed in Section 5.2) with unknown parameters. In this case, the log-likelihood given in (5.12) can be used to estimate the mixture cure model. The following example illustrates this idea.

Example 5.5. Suppose it is required to fit a parametric mixture cure model to the observed data $(T_i, \delta_i, \underline{v}_i, \underline{w}_i),\ i = 1,2,\ldots,n$, and it is assumed that the baseline distribution is Weibull with unknown parameters $\lambda > 0$ and $\alpha > 0$. From (5.11), the log-likelihood is given by

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

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

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

It is clear from this example that a closed-form expression for the conditional survival function can be obtained using the parametric mixture cure model, which is both easy to implement and to understand, especially for implementing in industry. Amdahl (2019) developed the R-package, *'flexsurvcure'*, to estimate the parameters of a parametric mixture cure model. This package allows covariates to be specified for the latency model, but allows no covariates for the incidence model. This package is therefore not suitable for our purposes, as our incidence model is a function of various covariates. The R-code used to maximise the log-likelihood in (5.12) for both the exponential and Weibull baseline distribution can be found in Appendix B.

Naturally the question again arises, how well does this specific parametric model fit the data? The goodness-of-fit of the parametric form of the survival function in a mixture cure model is still largely an open problem in the literature. Maller and Zhou (1996) proposed a very informal 'test', where they

use the correlation coefficient as test statistic. Very recently, Geerdens et al. (2019) developed a formal goodness-of-fit test based on the Cramér-von Mises distance between a non-parametric estimator of the mixture cure model and the estimated mixture cure model under the null hypothesis (which states that the survival part has a specific parametric form). They derive the asymptotic distribution of the test statistic and also propose a bootstrap procedure to estimate the critical value of the test. However, both of these tests are only applicable for a mixture cure model where both the incidence model component and the latency model component are not dependent on any covariates. In other words, their test are only applicable if the model in (5.9) is given by

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

Testing the goodness-of-fit of a fully parametric mixture cure model, where covariates are present, is thus still an open (and seemingly very difficult) problem in the field of both practical and theoretical statistics. The development of such a test is beyond the scope of this thesis. As a result, in the next chapter where this model is implemented, two different choices of the parametric form are used, and the results are compared.

Chapter 6

The optimisation model incorporating survival probabilities

6.1 Introduction

The miss-pricing and miss-allocation of consumer credit can have a severe impact on the global economy and subsequently on the financial institution providing credit. Simply increasing the price of loans with the expectation of earning higher net interest income, could potentially result in borrowers being unable to repay the loan or unwilling to take up a loan at the higher price, ultimately decreasing the expected NPII generated from these loans. Hence, the interaction between price and risk is of great importance in the consumer credit market. According to Phillips (2013), this interaction has not yet been fully addressed in credit price optimisation.

The credit price and LTV optimisation problem formulated in Chapter 4 determines the optimal price and LTV to quote a prospective customer when the objective is to maximise the NPII to the lender. However, this problem only considers the overall probability of default and not the probability of default during each month (or alternatively, the probability of not defaulting before each month) when calculating the NPII. Subsequently, the probability of survival or the probability of making a payment each month during the term of the loan (or other events/risk that are possible during the loan term) are not taken into account when determining the NPII and ultimately when solving the optimisation problem.

Ma et al. (2010) consider modelling the survival probability and early settlement probability using a standard CPH model when estimating profitability. However, since a large proportion of credit portfolios will usually not experience the event of interest, *i.e.*, default, the mixture cure model will be used in this chapter to estimate the probability that a customer is still repaying a loan during the loan term. Subsequently, the impact of the quoted price and LTV, not only on the take-up probability of the customer, but also the probability that a customer is still repaying the loan, given they took up the loan, can be taken into account. That is, the credit price and LTV optimisation problem will now also in-

corporate the survival probability when determining the optimal price and LTV to quote a prospective customer when maximising the expected value of the NPII.

The remainder of this chapter is organised as follows: in Section 6.2 the formulation of the credit price and LTV optimisation problem incorporating survival analysis will be discussed. In Section 6.3 a piece-wise linear approximation approach to solve the non-linear problem, specifically with the inclusion of survival analysis, is discussed. This chapter concludes in section 6.4 with a discussion on model behaviour and computational results, when the credit price and LTV optimisation model is applied to two simulated survival data sets.

6.2 The formulation of the credit price and LTV optimisation problem incorporating survival analysis

To formulate the credit price and LTV optimisation problem when modelling the probability of making a payment at each month (or not defaulting before each month) using survival analysis, recall from Chapter 2 Section 2.4.2, that the NPII for a risk-less loan is given by

$$I^m(r) \coloneqq I^m(r \mid r_0, r_d, n, a) = \sum_{t=1}^{n} a \left(\frac{1}{(1+r_d)^t} \right) \left(\frac{r(1+r)^n}{(1+r)^n - 1} - \frac{r_0(1+r_0)^n}{(1+r_0)^n - 1} \right), \tag{6.1}$$

where the superscript m stands for maturity, indicating that the loan is certain to go the full term. The NPII of a loan that is to be repaid in full with probability 1 is thus given by (6.1). Note that $r = \frac{x}{12}$ represents the monthly interest rate with x denoting to the annual interest rate. Now, the lender faces the risk (at the time of funding the loan) that the customer (borrower) may default at some point during the term of the loan, *i.e.* stop making the monthly repayments. The lender, however, still has to repay the outstanding capital, even if the customer does in fact default during the term of the loan. Therefore, the probability that a customer defaults at some point during the term of the loan, has to be taken into account when determining the NPII. Let $S(t)$ denote a generic survival function representing the probability that a borrower will make the payment in month t (or not default before month t), then the NPII (when the lender still has to repay the outstanding capital even if the borrower defaults) is given by

$$I^s(r) \coloneqq I^s(r \mid r_0, r_d, n, a, S(t)) = \sum_{t=1}^{n} a \left(\frac{1}{(1+r_d)^t} \right) \left(\frac{S(t)r(1+r)^n}{(1+r)^n - 1} - \frac{r_0(1+r_0)^n}{(1+r_0)^n - 1} \right), \tag{6.2}$$

where the superscript s in $I^s(r)$ represents the survival probability: indicating that the repayment is subject to the survival of the customer (or the customer not defaulting). It is clear that the NPII given in (6.1) and (6.2) are equal if $S(t) = 1$ for all t in (6.2), *i.e.*, if the borrower does not default and consequently repays the loan in full during the loan term with probability 1.

In a credit portfolio it is well-known that a substantial proportion of the customers will not experience the event of interest, *i.e.*, a large proportion will not default and therefore will make each payment. This large proportion of customers are considered the long term survivors and are subsequently not susceptible to default. In contrast to the long term survivors, there is a proportion of customers that are indeed susceptible to default and for which the probability of making each payment is not necessarily one. Therefore, to incorporate the survival behaviour of both sets of customers into the credit price and LTV optimisation problem, the unconditional survival probability will be modelled using a mixture cure model. As discussed in Section 5.5 of Chapter 5, the mixture cure model is able to estimate both the proportion of customers that are non-susceptible (using logistic regression), as well as the survival probability of those customers who are susceptible (using a CPH model). Hence, when using this model to determine the unconditional survival probability, the impact of the quoted price and LTV on the survival probability of the susceptible and non-susceptible customers for the duration of the loan can subsequently be taken into account when determining the optimal price and LTV to quote a prospective customer.

Associated with each customer is a set of covariates (characteristics) that could potentially influence their probability of being susceptible to default as well as their conditional survival probability. The covariates of the incidence model (predicting the probability of being susceptible) and the latency model (predicting the survival probability of the susceptible customers) in the mixture cure model need not necessarily be the same. Henceforth, the notation of the mixture cure model discussed in Section 5.5.1 of Chapter 5 will be used and the covariate vectors of the incidence and latency model are given by $\underline{v} = (v_1, v_2, \cdots, v_k)^T$ and $\underline{w} = (w_1, w_2, \cdots, w_m)^T$, respectively. The probability that a customer with covariates $\underline{v}$ and $\underline{w}$ will make the payment in month t (or the unconditional survival probability) using the mixture cure model, is given by

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

where $\pi(\underline{v})$ denotes the incidence model component and $S(t \mid \Upsilon = 1, \underline{w})$ the latency model component. The incidence model component represents the probability of being susceptible to default and is modelled using logistic regression,

$$\pi(\underline{v}) = P(\Upsilon = 1 \mid \underline{v}) = \frac{1}{1 + e^{-\underline{\eta}^T \underline{v}}} = \left(1 + e^{-\underline{\eta}^T \underline{v}}\right)^{-1}. \tag{6.3}$$

The latency model component represents the conditional survival probability of the susceptible cases and is modelled using a CPH model given by

$$\begin{aligned} S(t \mid \Upsilon = 1, \underline{w}) &= S_0(t \mid \Upsilon = 1) e^{\underline{\beta}^T \underline{w}} \\ &= e^{-H_0(t|\Upsilon=1) e^{\underline{\beta}^T \underline{w}}}. \end{aligned} \tag{6.4}$$

By using the mixture cure model to predict the unconditional survival probability when calculating the NPII, the survival behaviour of the customers susceptible and not susceptible to default are taken into consideration. More specifically, the NPII when modelling the unconditional survival probability using a mixture cure model, is given by

$$\begin{aligned}
I^s(r) &:= I^s(r \mid r_0, r_d, n, a, S(t \mid \underline{v}, \underline{w})) \\
&= \sum_{t=1}^{n} a \left(\frac{1}{(1+r_d)^t} \right) \left(\frac{S(t \mid \underline{v}, \underline{w})\, r(1+r)^n}{(1+r)^n - 1} - \frac{r_0(1+r_0)^n}{(1+r_0)^n - 1} \right) \\
&= \sum_{t=1}^{n} a \left(\frac{1}{(1+r_d)^t} \right) \left(\frac{[\pi(\underline{v}) S(t \mid \Upsilon = 1, \underline{w}) + [1 - \pi(\underline{v})]]\, r(1+r)^n}{(1+r)^n - 1} - \frac{r_0(1+r_0)^n}{(1+r_0)^n - 1} \right) \\
&= \pi(\underline{v}) \sum_{t=1}^{n} a \left(\frac{1}{(1+r_d)^t} \right) \left(\frac{S(t \mid \Upsilon = 1, \underline{w})\, r(1+r)^n}{(1+r)^n - 1} - \frac{r_0(1+r_0)^n}{(1+r_0)^n - 1} \right) + \\
&\quad [1 - \pi(\underline{v})] \sum_{t=1}^{n} a \left(\frac{1}{(1+r_d)^t} \right) \left(\frac{r(1+r)^n}{(1+r)^n - 1} - \frac{r_0(1+r_0)^n}{(1+r_0)^n - 1} \right) \\
&=: \pi(\underline{v}) I^d(r) + [1 - \pi(\underline{v})]\, I^m(r),
\end{aligned} \tag{6.5}$$

where $I^d(r) := I^d(r \mid r_0, r_d, n, a, S(t \mid \Upsilon = 1, \underline{w}))$ denotes the NPII given the customer is susceptible to default and $I^m(r)$ denotes the NPII given the customer is not susceptible to default as defined in (6.1).

To apply the credit price and LTV optimisation problem using survival analysis at a portfolio level, the notation used in Chapter 4 will be adapted. Suppose a secured credit portfolio consists of a set of customers $\mathscr{C} = \{1, 2, ..., C\}$, where C denotes the number of customers in the portfolio and let a_c, n_c and v_c denote the loan amount, loan term and underlying asset value for each customer $c \in \mathscr{C}$. Recall that the LTV can be expressed as the loan amount (a_c) divided by the underlying asset value (τ_c), that is,

$$y_c = \frac{a_c}{\tau_c}. \tag{6.6}$$

The objective of this study is to maximise the expected value of the NPII by finding the right balance between price and LTV according to the price elasticity model and by using survival analysis to predict the survival behaviour of the customers. Therefore, expressing the loan amount in terms of the underlying asset value and the LTV and substituting the expression into (6.5), the NPII for a customer

$c \in \mathscr{C}$ as a function of the quoted price, x_c, and LTV, y_c, is given by

$$
\begin{aligned}
I^s(x_c,y_c) &:= I^s(x_c,y_c \mid r_d,n_c,a_c,S(t \mid \underline{v}_c,\underline{w}_c)) \\
&= \pi(\underline{v}_c)\sum_{t=1}^{n_c} y_c\tau_c\left[\frac{1}{(1+r_d)^t}\right]\left[\frac{S(t \mid \Upsilon_c=1,\underline{w}_c)\left(\frac{x_c}{12}\right)\left(1+\frac{x_c}{12}\right)^{n_c}}{\left(1+\frac{x_c}{12}\right)^{n_c}-1} - \frac{\left(\frac{x^0}{12}\right)\left(1+\frac{x^0}{12}\right)^{n_c}}{\left(1+\frac{x^0}{12}\right)^{n_c}-1}\right] + \\
&\quad [1-\pi(\underline{v}_c)]\sum_{t=1}^{n_c} y_c\tau_c\left[\frac{1}{(1+r_d)^t}\right]\left[\frac{\left(\frac{x_c}{12}\right)\left(1+\frac{x_c}{12}\right)^{n_c}}{\left(1+\frac{x_c}{12}\right)^{n_c}-1} - \frac{\left(\frac{x^0}{12}\right)\left(1+\frac{x^0}{12}\right)^{n_c}}{\left(1+\frac{x^0}{12}\right)^{n_c}-1}\right] \\
&= \left(1+e^{-\underline{\eta}^T\underline{v}_c}\right)^{-1}\sum_{t=1}^{n_c} y_c\tau_c\left[\frac{1}{(1+r_d)^t}\right]\left[\frac{e^{-H_0(t|\Upsilon_c=1)e^{\underline{\beta}^T\underline{w}_c}}\left(\frac{x_c}{12}\right)\left(1+\frac{x_c}{12}\right)^{n_c}}{\left(1+\frac{x_c}{12}\right)^{n_c}-1} - \frac{\left(\frac{x^0}{12}\right)\left(1+\frac{x^0}{12}\right)^{n_c}}{\left(1+\frac{x^0}{12}\right)^{n_c}-1}\right] \\
&\quad + \left[1-\left(1+e^{-\underline{\eta}^T\underline{v}_c}\right)^{-1}\right]\sum_{t=1}^{n_c} y_c\tau_c\left[\frac{1}{(1+r_d)^t}\right]\left[\frac{\left(\frac{x_c}{12}\right)\left(1+\frac{x_c}{12}\right)^{n_c}}{\left(1+\frac{x_c}{12}\right)^{n_c}-1} - \frac{\left(\frac{x^0}{12}\right)\left(1+\frac{x^0}{12}\right)^{n_c}}{\left(1+\frac{x^0}{12}\right)^{n_c}-1}\right] \\
&=: \pi(\underline{v}_c)I^d(x_c,y_c) + [1-\pi(\underline{v}_c)]I^m(x_c,y_c),
\end{aligned}
\tag{6.7}
$$

where $\underline{v}_c = (v_{c1},v_{c2},\cdots,v_{ck})^T$ and $\underline{w}_c = (w_{c1},w_{c2},\cdots,w_{cm})^T$ denote the covariates of the incidence and latency model for each customer $c \in \mathscr{C}$, respectively. Note that the quoted price, x_c, and LTV, y_c, are covariates of both model components of the mixture cure model, that is, $x_c,y_c \in \underline{v}_c$ and $x_c,y_c \in \underline{w}_c$. The NPII in (6.7) is subject to the customer taking up the loan at the quoted price x_c and LTV y_c, which is modelled using the price response function (logistic regression model) given by

$$
R(x_c,y_c) := R(x_c,y_c|n_c,\tau_c,x^0,p_c) = 1/(1+e^{(-(\beta_0+\beta_1\tau_c y_c+\beta_2 n_c+\beta_3 p_c+\beta_4 x^0+\beta_5 x_c+\beta_6 y_c))}). \tag{6.8}
$$

The expected value of the NPII (incorporating survival probabilities) at a price x_c and LTV y_c is simply the product of (6.7) and (6.8),

$$
f^s(x_c,y_c) := R(x_c,y_c)I^s(x_c,y_c). \tag{6.9}
$$

The credit price and LTV optimisation problem without any constraints, but incorporating survival probabilities, is to

$$\begin{aligned} &\max \sum_{c\in\mathscr{C}} f^s(x_c,y_c) = R(x_c,y_c)I^s(x_c,y_c) && (6.10)\\ &\text{s.t. } x_c,y_c \geq 0 && \forall c \in \mathscr{C}. \end{aligned}$$

Constraints of the risk distribution and the proportion of loans with a high LTV in the credit portfolio can now also be added in a similar way as discussed in Section 4.3 of Chapter 4. The optimisation problem is then to

$$\begin{aligned} &\max \sum_{c\in\mathscr{C}} f^s(x_c,y_c) = R(x_c,y_c)I^s(x_c,y_c) && \\ &\text{s.t. } t_c = R(x_c,y_c) && \forall\, c \in \mathscr{C}, \\ &\quad \sum_{c\in\mathscr{C}(g)} t_c \leq U_g \sum_{c\in\mathscr{C}} t_c && \forall\, g \in \mathscr{G}, \qquad (6.11)\\ &\quad \sum_{c\in\mathscr{C}} \mathbb{I}\left(y_c \geq y^b\right) \leq U_y C && \qquad (6.12)\\ &\quad x_c,y_c \geq 0 && \forall\, c \in \mathscr{C}, \end{aligned}$$

where

$$\mathbb{I}\left(y_c \geq y^b\right) = \begin{cases} 1 & \text{if } y_c \geq y^b \\ 0 & \text{if } y_c < y^b \end{cases}.$$

Note that both the NPII and the response function are non-linear functions of price and LTV and, consequently, the expected value of the NPII given in (6.9) is also a non-linear function of these variables. The price and LTV optimisation problem which incorporates survival probabilities is therefore also considered a NLP problem (see Section 3.5 of Chapter 3). The piece-wise linear approximation approach, discussed in Section 4.4 of Chapter 4, will thus be used to solve this optimisation problem. However, this problem now has the added approximation of the survival probability at every month t when approximating the NPII. Naturally, incorporating these survival probabilities makes the optimisation problem more realistic and challenging.

6.3 A piece-wise linear approximation approach to credit price and LTV optimisation incorporating survival probabilities

The piece-wise linear approximation approach for the credit price and LTV optimisation problem incorporating survival probabilities, is similar to the approach followed in Section 4.4 of Chapter 4. An approximation of the expected value of the NPII given in (6.9) at the grid point (x_{ci}, y_{cj}) is given by

$$f^s_{cij} := f^s(x_{ci}, y_{cj}) = R(x_{ci,}y_{cj})I^s(x_{ci}, y_{cj}) =: R_{cij}I^s_{cij}, \tag{6.13}$$

where x_{ci} and y_{cj} denote the price and LTV at the end point of the interval $i \in \mathscr{I}_0 = \mathscr{I} \cup 0 = \{0,1,...,I\}$ and $j \in \mathscr{J}_0 = \mathscr{J} \cup 0 = \{0,1,...,J\}$, respectively. Recall that $i = 0$ and $j = 0$ denotes the starting point of the first interval of the price and LTV range, respectively. To calculate the NPII for each customer $c \in \mathscr{C}$, let $s_{cij}(t) = S\left(t \mid \Upsilon_c = 1, \underline{w}^*_{cij}\right)$ denote the conditional survival probability at month $t \in T = \{1,2,\ldots,n_c\}$, for a price x_{ci} and a LTV y_{cj}. Note that $\underline{w}^*_{cij} = \left(w_{c1}, x_{ci}, y_{cj}, \ldots, w_{cm}\right)^T$ is similar to $\underline{w}_c$, with the only difference being that the price and LTV covariates can take on the different values within their respective ranges. The piece-wise linear approximation of the NPII at a gridpoint (x_{ci}, y_{cj}) is given by

$$\begin{aligned}
I^s_{cij} &:= I^s(x_{ci}, y_{cj}) \\
&= I^s(x_{ci}, y_{cj} \mid n_c, \tau_c, x^0, p_c, s_{cij}(t), \underline{v}^*_{cij}, \underline{w}^*_{cij}) \\
&= \pi(\underline{v}^*_{cij}) I^d(x_{ci}, y_{cj}) + \left[1 - \pi(\underline{v}^*_{cij})\right] I^m(x_{ci}, y_{cj}) \\
&= \left(1 + e^{-\underline{\eta}^T \underline{v}^*_{cij}}\right)^{-1} \sum_{t=1}^{n_c} y_{cj} \tau_c \left[\frac{1}{(1+r_d)^t}\right] \left[\frac{e^{-H_0(t|\Upsilon_c=1)e^{\underline{\beta}^T \underline{w}^*_{cij}}} \left(\frac{x_{ci}}{12}\right)\left(1+\frac{x_{ci}}{12}\right)^{n_c}}{\left(1+\frac{x_{ci}}{12}\right)^{n_c} - 1} - \frac{\left(\frac{x^0}{12}\right)\left(1+\frac{x^0}{12}\right)^{n_c}}{\left(1+\frac{x^0}{12}\right)^{n_c} - 1}\right] \\
&\quad + \left[1 - \left(1 + e^{-\underline{\eta}^T \underline{v}^*_{cij}}\right)^{-1}\right] \sum_{t=1}^{n_c} y_{cj} \tau_c \left[\frac{1}{(1+r_d)^t}\right] \left[\frac{\left(\frac{x_{ci}}{12}\right)\left(1+\frac{x_{ci}}{12}\right)^{n_c}}{\left(1+\frac{x_{ci}}{12}\right)^{n_c} - 1} - \frac{\left(\frac{x^0}{12}\right)\left(1+\frac{x^0}{12}\right)^{n_c}}{\left(1+\frac{x^0}{12}\right)^{n_c} - 1}\right],
\end{aligned} \tag{6.14}$$

where $\underline{v}^*_{cij} = \left(v_{c1}, x_{ci}, y_{cj}, \ldots, v_{ck}\right)^T$ is also similar to $\underline{v}_c$ with the only difference again being that the price and LTV covariates can take on the different values within their respective ranges. Following a piece-wise linear approximation to solve the optimisation problem, with the inclusion of the survival analysis, reduces the NLP problem to an MILP problem, for which proven optimal solutions are obtained and the option of introducing logical decision-making variables are made possible. Moreover, by including survival probabilities in the credit price and LTV optimisation problem, the optimal price and LTV can be determined such that the probability of a customer making each payment during the loan term is taken into account. More specifically, by using a mixture cure model to estimate the unconditional survival probability, the optimisation problem takes into consideration the proportion of customers in a credit portfolio (*i.e.* a home loan portfolio) that are not susceptible to default and likely to make each payment with probability 1, as well as the survival probability of those customers that are indeed susceptible to default. The credit price and LTV optimisation problem using survival analysis is therefore a more comprehensive and realistic approach to credit price and LTV optimisation.

Formulating the credit price and LTV optimisation problem using survival analysis as a MILP problem now has the added benefit that the objective function incorporates the survival probability at each month. For the MILP problem, the notation and variables remain the same as in Section 4.4 of Chapter 4, however with the exception that the expected value of the NPII is now given by f^s_{cij} where the superscript s in f^s_{cij} denotes that the survival probability (*i.e.* the probability of not defaulting), is now taken into account using survival analysis. Therefore, the credit price and LTV optimisation problem using survival analysis and formulated as a MILP problem that incorporates risk distribution constraints, LTV constraints and logical decision making capability is given by

$$\begin{aligned} &\max \sum_{c\in\mathscr{C}} V_c \\ &\text{s.t. } V_c \leq \sum_{i\in\mathscr{I}_0}\sum_{j\in\mathscr{J}_0} \lambda_{cij} f^s_{cij} + (1-z_c)M, && \forall\, c\in\mathscr{C}, \\ &\qquad V_c \geq \sum_{i\in\mathscr{I}_0}\sum_{j\in\mathscr{J}_0} \lambda_{cij} f^s_{cij} - (1-z_c)M, && \forall\, c\in\mathscr{C}, \end{aligned} \tag{6.15}$$

and constraints (4.23)–(4.40) from Section 4.4 of Chapter 4. Note that, the above model is a theoretical representation of the credit price and LTV optimisation model, and the parameters of the response model and mixture cure models have to be estimated first, before implementing the optimisation model.

6.4 Model behaviour and computational results

To consider the model behaviour and computational results, the optimisation model in (6.15) is applied to simulated survival data (default and censoring times). In all the instances, the assumption is made that $r_d \approx 0$, since once again the discount rate is being applied on top of the repurchase rate r_0. Furthermore, a discussion will follow first on how survival data can be simulated from a mixture cure model before the optimisation model is applied to the data. All simulations, estimation and approximations were performed in the statistical software package R (R Core Team, 2020) .

6.4.1 Simulating survival data from a mixture cure model with covariates

A parametric approach is followed to estimate the parameters of the mixture cure model (see Section 5.5.3 of Chapter 5) and since testing the goodness-of-fit of a fully parametric mixture cure model is still an open problem, initially two data sets were generated where the survival times are simulated from a parametric mixture cure model using the covariates of the data obtained from a financial institution (the same data set as used in Chapter 4). For the first data set, default times were generated using an exponential baseline distribution whereas for the second data set, a Weibull baseline distribution was used. The covariates appearing in the data set in Chapter 4 were used for both the simulated survival data sets. More specifically, the loan amount ($\tau_c y_c$) , price (x_c), repurchase rate (x^0) and over-

all probability of default (p_c) were chosen as the covariates for both the incidence and latency model components of the mixture cure model for 100 customers. That is, for the incidence model component $\underline{v}_c = \left(\tau_c y_c, x_c - x^0, p_c\right)^T$, with corresponding coefficients η_0 (for the intercept term), η_1,η_2 and η_3 and for the latency model component $\underline{w}_c = \left(\tau_c y_c, x_c - x^0, p_c\right)^T$, with corresponding coefficients β_1, β_2 and β_3, where $c \in \mathscr{C} = \{1,2,\ldots,100\}$. The true parameter values for both these model components and both data sets are displayed in Table 6.1. For the latency model component, default times were generated assuming an exponential cumulative baseline hazard function with parameter $\lambda = 0.00125$ and Weibull cumulative baseline hazard function with parameters $\lambda = 0.0001$ and $\alpha = 1.7$, respectively. The censoring times for both data sets are assumed to be uniformly distributed between the interval 0 and 100. Using this method when generating the censoring times, the proportion of censored customers in the portfolio was roughly 90%. That is, roughly 10% of the customers in the portfolio defaulted on their loans. The steps followed to generate the survival data from a mixture cure model, where default is the only event of interest, are summarised below.

Table 6.1: True parameter values of the mixture cure models.

Data set	Baseline	Parameters of baseline	η_0	η_1	η_2	η_3	β_1	β_2	β_3
1	Exponential	$\lambda = 0.00125$	-3.5	0.03	12	15	0.0125	5	10
2	Weibull	$\lambda = 0.0001$ and $\alpha = 1.7$	-2.5	0.02	5	15	0.01	15	12

Generating the incidence model component data:

1. Using the relationship in (6.3), where the parameter values of $\underline{\eta}$ are given in Table 6.1, the probability of being susceptible to default, given the covariate values $\underline{v}_c$, is generated for each customer $c \in \mathscr{C}$, *i.e.* $\pi(\underline{v}_c) = P(\Upsilon_c = 1 \mid \underline{v}_c) = \left(1 + e^{-\underline{\eta}^T \underline{v}_c}\right)^{-1}$.

2. Simulate Υ_c, $c \in \mathscr{C}$ from a discrete distribution,

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

Generating the latency model component data:

To generate the event times, recall from Section 5.4.2 of Chapter 5 that if $S(t \mid \Upsilon_c = 1, \underline{w}_c)$ is the true survival function for the susceptible cases, then $S(Y_c \mid \Upsilon_c = 1, \underline{w}_c)$ follows a uniform $(0,1)$ distribution *i.e.*, $S(Y_c \mid \Upsilon_c = 1, \underline{w}_c) = U$, where U denotes an uniform random variable within the interval 0 and 1 and Y_c denotes the default time (month) for customer $c \in \mathscr{C}$. Using the relationship

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

the default times for the susceptible customers can be expressed as

$$Y_c = H_0^{-1}\left[-\log(U)e^{-\underline{\beta}^T \underline{w}_c}\right], \tag{6.16}$$

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

3. Using the relationship in (6.16) with the corresponding parameter values of $\underline{\beta}$ given in Table 6.1, the default times for the customers susceptible to default are generated as $Y_c = H_0^{-1}\left[-\log(U)e^{-\underline{\beta}^T \underline{w}_c}\right]$, $c = 1,2,\ldots 100$.

4. Recall from Assumption 5.1 that the censoring time and default time are independent, hence the censoring times for each customer $c \in \mathscr{C}$ are generated independently from a uniform distribution within the interval 0 and 100 and denoted by C_c, $c = 1,2,\ldots 100$.

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

5. Now, since the true value of Υ_c is known, the customers who are not susceptible to default (*i.e.* the customers for whom $\Upsilon_c = 0$), cannot default and therefore must be censored. Hence, for the customers that are not susceptible to default (*i.e.* the customers that did not default during the observation period and will not default), the event time is the censoring time. Hence, if $\Upsilon_c = 0$, then $T_c = C_c$ and $\delta_c = 0$.

6. The event time and censoring indicator for the customers susceptible to default (*i.e.* the customers that either defaulted on the loan during the observation period or did not default during the observation period but will eventually default) are determined as follow,

$$T_c = \min(Y_c, C_c)$$

and

$$\delta_c = \begin{cases} 1 & \text{if } Y_c \leq C_c \\ 0 & \text{if } Y_c > C_c \end{cases}.$$

Following the steps outlined above, survival data from a mixture cure model can be generated using different baseline distributions and true parameter values. Recall that, if the default time Y is exponentially distributed with parameter $\lambda > 0$, the (baseline) hazard rate is constant *i.e.* $h_0(t) = \lambda$ and the cumulative baseline hazard rate function and its inverse are given by

$$H_0(t) = \int_0^t h_0(u)du = \lambda t \tag{6.17}$$

and

$$H_0^{-1}(t) = \lambda^{-1}t, \tag{6.18}$$

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

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

and

$$H_0^{-1}(t) = \left(\lambda^{-1}t\right)^{1/\alpha}, \tag{6.19}$$

respectively. The survival data for the simulated data sets were subsequently generated using the inverse cumulative baseline hazard functions in (6.18) and (6.19). It is, however, important to note that even though the value of the susceptible indicator Υ_c is known when simulating the survival data from a mixture cure model, the observed data only consists of $(T_c, \delta_c, \underline{v}_c, \underline{w}_c)$ and is therefore the only data available when fitting the mixture cure model.

6.4.2 Parametric estimation of the mixture cure model

To estimate the parameters of the mixture cure model, the method of maximum likelihood, as discussed in Section 5.5.3 of Chapter 5, is applied. This method is used to estimate the values of the parameters that makes the observed data $(T_c, \delta_c, \underline{v}_c, \underline{w}_c)$ most likely. Example 5.5 in Chapter 5 Section 5.5.3, explained how the unknown parameters of a mixture cure model ($\underline{\eta}$ for the incidence model component and $\underline{\beta}$ for the latency model component) are estimated when assuming the baseline distribution is Weibull with unknown parameters $\lambda > 0$ and $\alpha > 0$. Thus, by using these maximum likelihood estimators $\hat{\underline{\eta}}, \hat{\underline{\beta}}, \hat{\alpha}$ and $\hat{\lambda}$, the parametric estimate of the unconditional survival function for the mixture cure model at time (month) t, for a customer $c \in \mathscr{C}$ and assuming a Weibull baseline distribution, is given by

$$\hat{S}(t \mid \underline{v}_c, \underline{w}_c) = \left(1 + e^{-\hat{\underline{\eta}}^T \underline{v}_c}\right)^{-1} e^{-\hat{\lambda} t^{\hat{\alpha}} e^{\hat{\underline{\beta}}^T \underline{w}_c}} + \left[1 - \left(1 + e^{-\hat{\underline{\eta}}^T \underline{v}_c}\right)^{-1}\right]. \tag{6.20}$$

For completeness sake, the case where the baseline distribution is assumed to be exponential with unknown parameter $\lambda > 0$ will also be discussed. Again, consider the log-likelihood of the mixture cure model

$$\begin{aligned} l(\underline{\eta}, \underline{\beta}) = &\sum_{c=1}^{n} \delta_c \left\{ \log \pi(\underline{v}_c) - H_0(T_c \mid \Upsilon_c = 1) e^{\underline{\beta}^T \underline{w}_c} + \log h_0(T_c \mid \Upsilon_c = 1) + \underline{\beta}^T \underline{w}_c \right\} + \\ &\sum_{c=1}^{n} (1 - \delta_c) \log \left\{ \pi(\underline{v}_c) e^{-H_0(T_c \mid \Upsilon_c = 1) e^{\underline{\beta}^T \underline{w}_c}} + [1 - \pi(\underline{v}_c)] \right\}, \end{aligned} \tag{6.21}$$

where

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

Substituting the expressions for the baseline hazard and cumulative baseline hazard rate of the exponential distribution into (6.21), the log-likelihood becomes

$$l(\underline{\eta}, \underline{\beta}, \lambda) = \sum_{c=1}^{n} \delta_c \left\{ \log \pi(\underline{v}_c) - \lambda T_c e^{\underline{\beta}^T \underline{w}_c} + \log \lambda + \underline{\beta}^T \underline{w}_c \right\} + \sum_{c=1}^{n} (1 - \delta_c) \log \left\{ \pi(\underline{v}_c) e^{-\lambda T_c e^{\underline{\beta}^T \underline{w}_c}} + [1 - \pi(\underline{v}_c)] \right\}. \tag{6.22}$$

Maximising the log-likelihood (either implicitly or by numerical methods) the maximum likelihood estimators $\hat{\underline{\eta}}, \hat{\underline{\beta}}$ and $\hat{\lambda}$ are obtained. It then easily follows that the parametric estimate for the mixture cure model at time (month) t, for a customer $c \in \mathscr{C}$, when assuming an exponential baseline distribution, is given by

$$\hat{S}(t \mid \underline{v}_c, \underline{w}_c) = \left(1 + e^{-\hat{\underline{\eta}}^T \underline{v}_c}\right)^{-1} e^{-\hat{\lambda} t e^{\hat{\underline{\beta}}^T \underline{w}_c}} + \left[1 - \left(1 + e^{-\hat{\underline{\eta}}^T \underline{v}_c}\right)^{-1}\right]. \tag{6.23}$$

These estimates of the survival times can now be used in the expression for the NPII (given in (6.7)) to obtain the estimated NPII. Table 6.2 contains the values of the estimated parameters for the mixture cure model for both the simulated data sets (which were simulated using the true parameter values given in Table 6.1).

Table 6.2: Estimated parameter values of the mixture cure models.

Data set	Baseline	Estimated parameters	$\hat{\eta}_0$	$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
1	Exponential	$\hat{\lambda} = 0.00124$	-3.50	0.03	11.99	14.99	0.014	5.00	9.99
2	Weibull	$\hat{\lambda} = 0.000132$, $\hat{\alpha} = 1.7$	-2.49	0.02	4.99	14.99	0.01	14.99	11.99

Furthermore, since the quoted price, x_c, and LTV, y_c, are covariates of both model components of the mixture cure model, the piece-wise linear approximation approach discussed in Section 6.3 is followed using the estimated mixture cure models.

6.4.3 Discussion of the credit price and LTV optimisation model results incorporating the estimated survival probabilities.

Data set 1: Exponential baseline distribution

The results for the data set where default times were generated using an exponential baseline distribution, will be considered first. Naturally, for this simulated data the true value of the susceptibility

indicator Υ_c is known for each $c \in \mathscr{C}$ (and thus the proportion of customers susceptible to default) as well as the censoring proportions. The underlying parameters from which the data were simulated (Table 6.1) were chosen in such a way that these proportions were realistic (*e.g.* a small default proportion) so that one can draw valid and sensible conclusions.

The censoring percentage was 93% of which 19% were susceptible to default and 74% not susceptible to default. Overall, 26% of the customers were susceptible to default, but only 7% defaulted during the observation period. A summary of these percentages can be found in Table 6.3.

Table 6.3: Percentages of censored customers and customers susceptible to default for the simulated data (exponential baseline).

Description	**Proportion**
Censored, $\delta = 0$	93%
Default, $\delta = 1$	7%
Susceptible to default, $\Upsilon = 1$	26%
Not susceptible to default, $\Upsilon = 0$	74%
Susceptible to default and defaulted, $\Upsilon = 1,\ \delta = 1$	7%
Not susceptible to default and censored, $\Upsilon = 0,\ \delta = 0$	74%
Susceptible to default and censored, $\Upsilon = 1,\ \delta = 0$	19%

After fitting the mixture cure model it was found that the estimated average probability of being susceptible to default was 27.8%. This is close to the true susceptible percentage of 26% present in the data (see Table 6.3). Furthermore, the average unconditional estimated survival probability at the last month of the loan term, $t = 240$, was 79.6%, clearly approaching the true percentage of customers not susceptible to default, which is 74%. As was shown in Chapter 5, this is a nice property of the mixture cure model, that if $t \to \infty$, the unconditional survival probability tends to the non-susceptible percentage.

Figure 6.1 depicts the relationship between price and the average estimated take-up for customers in the different risk gradings (low, medium and high). The estimated take-up probabilities for the high risk customers is less compared to the estimated take-up probabilities of the low risk customers for the same price(s). The impact of this will become clear when applying the optimisation model to the data set.

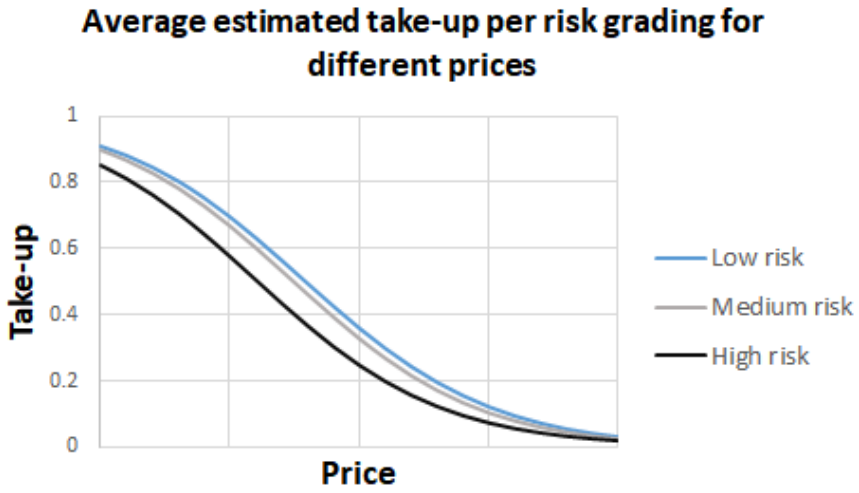


Figure 6.1: Relationship between price and estimated take-up per risk grading.

Figure 6.2 illustrates the relationship between the average estimated take-up and LTV per risk grading, where it can be seen that the average estimated take-up probability is larger for the low risk customers compared to the high risk customers for the same LTV. Clearly, on average, the take-up probability is expected to decrease with an increase in price, whereas an increase in the LTV relates to an increase in the take-up probability. If the converse was true, that is, higher prices related to higher take-up probabilities and higher LTVs related to lower take-up probabilities, the optimisation model would be unrealistic and the prices would be set as large as possible.

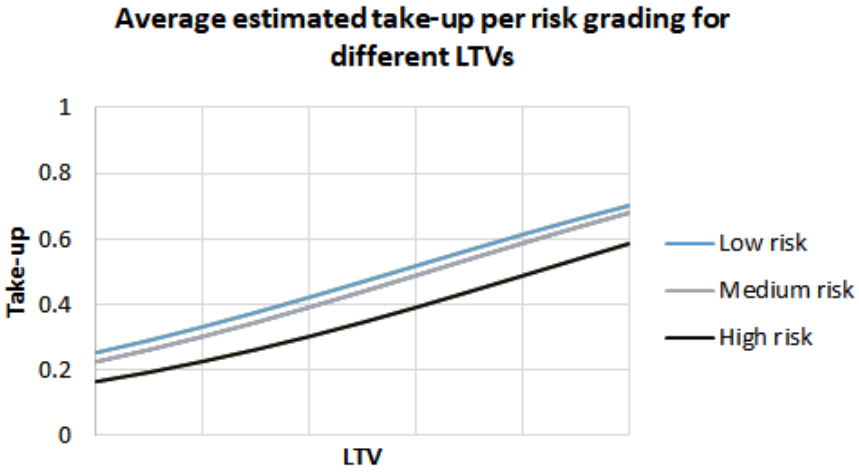
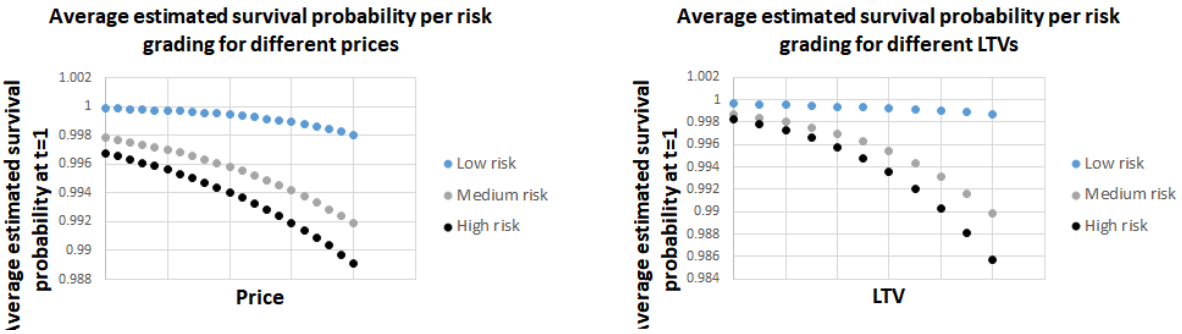


Figure 6.2: Relationship between LTV and take-up per risk grading.

To illustrate the relationship between the estimated survival probabilities and price and the estimated survival probabilities and LTV, the average estimated survival probabilities per risk grading at

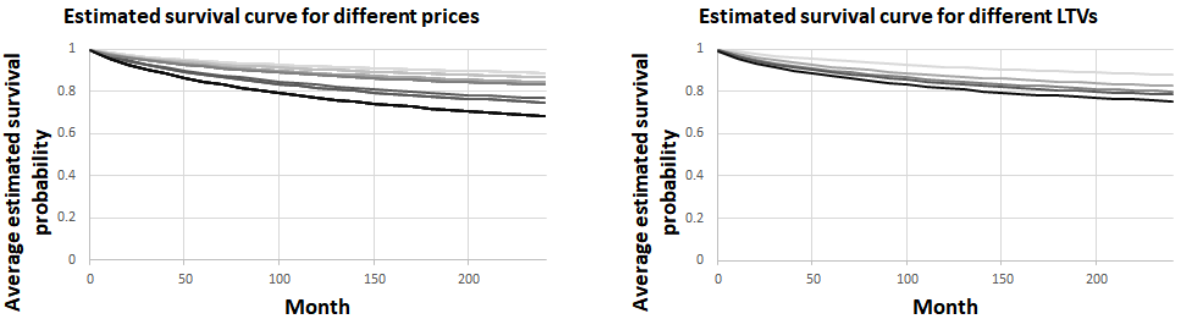
month one ($t = 1$) is considered for different prices and LTVs, respectively. These relationships are depicted in Figures 6.3 (a) and (b) . From both these figures it's clear that the average estimated survival probability in the first month decreases as the price and LTV increases, respectively. Additionally, the average estimated survival probability in the first month is higher for the low risk customers compared to the higher risk customers.



(a) (b)

Figure 6.3: (a) Relationship between price and estimated survival probability in month one; (b) Relationship between LTV and estimated survival probability in month one.

To consider the overall effect of price and LTV on the estimated survival probability *i.e.* probability of making a repayment in each month, the estimated unconditional survival curve for increasing values of both these variables are given in Figures 6.4 (a) and (b), respectively. The graphs from both these figures demonstrate that, by increasing the price and LTV, the average estimated survival probabilities for each month is smaller, when keeping all the other covariates constant. The light grey curves (lines) indicate the estimated survival curve for lower prices and LTVs and the darker grey to black curves (lines) represent the estimated survival curves as the price and LTV increases, respectively. These trends are expected, since if the price and LTV of loans are too high, the monthly repayment amount increases, and subsequently also the probability of default.



(a) (b)

Figure 6.4: (a) Estimated survival curve for different prices; (b) Estimated survival curve for different LTVs.

Similarly, when one considers the average estimated survival curve for the different risk gradings,

the lower risk customers have a higher average estimated survival probability, for each month, compared to the medium and high risk customers.

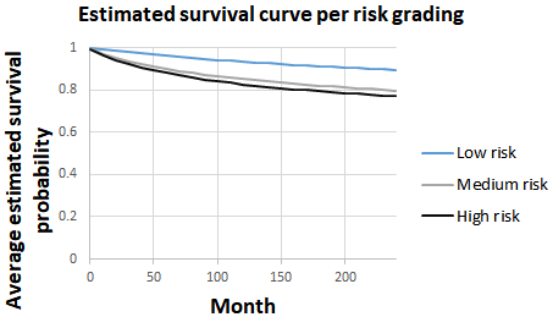


Figure 6.5: Estimated survival curve for different risk gradings.

Given these various relationships between price, LTV, estimated take-up probability and estimated survival probabilities, the objective remains to determine the optimal combination between the price and LTV that maximises the expected value of the NPII, while adhering to risk distribution and LTV constraints. Once again, the results displayed in the tables below are normalised and summarised per risk grading $g \in \mathscr{G} = \{Low, Medium, High\}$. That is, the average price per risk category is expressed as a proportion of the risk category with the highest average price. Table 6.4 summarises the model behaviour for the unconstrained optimisation problem that includes the estimated survival probabilities, when using the piece-wise linear approximation approach. That is, no constraints were imposed on the risk distribution or proportion of customers in the portfolio with an LTV larger than or equal to 0.9.

Table 6.4: Computational results for the unconstrained problem incorporating survival probabilities (exponential baseline).

Risk grading	Take-up proportion	Average price	Average LTV
Low	0.30	1.00	0.97
Medium	0.32	0.95	0.91
High	0.38	0.92	0.92

The results in Table 6.4 indicate that the average price for the high and medium risk customers are lower compared to the lower risk customers, however the low risk customers have a higher average LTV compared to the medium and high risk customers. The higher average price for the low risk customers is a result of these customers having higher average estimated take-up probabilities when quoted the same price in comparison to the higher risk customers (refer to Figure 6.1). The higher average LTV for the low risk customers can be contributed to the higher average estimated take-up probability of these customers compared to the medium and high risk customers when quoted the same LTV (see Figure 6.2). In addition to this, the average estimated survival probability for the lower risk customers are higher compared to the high risk customers for the same LTVs as illustrated in Figure 6.3 (b),

which most probably contribute to the higher average LTV for these customers. In addition to this, for the unconstrained problem including survival probabilities, the average LTV per risk category is significantly lower than 1, whereas for the unconstrained optimisation problem considered in Chapter 4 the average LTV per risk category was 1. This can be attributed to the fact that if the LTV is too high, the average survival probability (or probability of repayment) decreases as illustrated in Figures 6.3(b) and 6.4(b), respectively. Note that, the decrease is more significant for the medium and high risk customers, compared to the low risk customers. Furthermore, the proportion of customers with an LTV larger than or equal to 0.9 is 79%, clearly indicating that the risk in the portfolio is lower, since loans with higher LTVs are considered more risky as apposed to loans with lower LTVs (see Phillips, 2013 and Caufield, 2012). The take-up proportion in the portfolio for the medium and high risk customers is slightly larger compared to the low risk customers. Since the optimisation model allows for the exclusion of customers, almost 43% of the customers that were observed as defaults, were excluded from the portfolio. This can be attributed to the fact that there is no combination of price and LTV for which these customers have a positive expected NPII, specifically when including survival probabilities in the problem.

In order to illustrate the behaviour of the optimisation model when both survival probabilities and risk distribution constraints are present, the proportion of medium and high risk customer were restricted to 30% and 20%, respectively, *i.e.*, $U_g = \{1, 0.3, 0.2\}$. The results in Table 6.5 show the impact of these risk distribution constraints on the average price and LTV.

Table 6.5: Computational results for the risk distribution constrained problem incorporating survival probabilities (exponential baseline).

Risk grading	Take-up proportion	Average price	Average LTV
Low	0.50	0.79	0.98
Medium	0.30	0.95	0.93
High	0.20	1.00	0.89

From the results in Table 6.5, it is clear that loans to be offered to low risk customers will have a much lower average price compared to the medium and high risk customers. This results in higher estimated take-up probabilities as well as higher estimated survival probabilities for these customers due to the relationships between price and take-up, and price and survival probabilities. Furthermore, the low risk customers are also offered a higher average LTV compared to the medium and high risk customers, also resulting in higher take-up and survival probabilities. Moreover, the high risk customers are now offered higher prices on average and lower LTVs, resulting in lower take-up probabilities for these customers. In order to meet the risk distribution constraints, 48% of the high risk customers were excluded from the portfolio, since offering a price that is too high (in an attempt to reduce the take-up proportion), will result in lower survival probabilities (see Figure 6.3 (a)) and subsequently, lower expected NPII.

Another way of limiting the risk in the portfolio is to impose constraints (upper limits) on the proportion of loans in the portfolio that have an LTV larger than or equal to say 90%. The results in Table 6.6 demonstrate the impact of imposing an upper bound of 60% on the proportion of loans included in the final portfolio with an LTV larger than or equal to 0.9.

Table 6.6: Computational results for the LTV constrained problem incorporating survival probabilities (exponential baseline).

Risk grading	Take-up proportion	Average price	Average LTV
Low	0.32	1.00	0.97
Medium	0.35	0.94	0.81
High	0.34	0.93	0.83

From Table 6.6 it is clear that there is a reduction in the average LTV for both the medium and high risk customers. The average price of the medium and high risk customers are almost similar, with the low risk customers still having the highest average price. The estimated take-up proportion for the low and medium risk customers is now higher, whereas the estimated take-up proportion for the high risk customers is much lower compared to the unconstrained computational results, also reducing the risk in the portfolio. Thus, to adhere to the LTV constraint imposed on the portfolio, the average LTV, specifically for the medium and high risk customers, is significantly lower, ultimately reducing the risk in the portfolio.

To illustrate the impact of the constraints on the risk distribution in conjunction with the constraints on the LTV, the results in Table 6.7 are considered. Here, the constraints $U_g = \{1, 0.3, 0.2\}$ are imposed on the risk distribution, whereas the constraints on the LTV are set to $U_y = 60\%$ and $y^b = 0.9$. That is, the take-up proportion in the credit portfolio of the medium and high risk gradings are limited to 0.3 and 0.2, respectively, whereas the proportion of loans with a LTV larger than or equal to 0.9 are limited to 60% in the credit portfolio.

Table 6.7: Computational results for the risk distribution and LTV constrained problem incorporating survival probabilities (exponential baseline)

Risk grading	Take-up proportion	Average price	Average LTV
Low	0.50	0.90	0.97
Medium	0.30	0.95	0.87
High	0.20	1.00	0.80

From Table 6.7 it's clear that the average price for the low risk customers is now lower compared to the higher risk customers and the average LTV for the low risk customers is higher in comparison to the higher risk customers. The average price and LTV per risk category seems much more realistic now (when compared to the corresponding problem's results in Table 4.6 of Chapter 4), with the high risk customers on average receiving loans with much lower LTVs and higher average prices, whereas the

low risk customers are receiving loans with higher LTVs but lower prices on average. Additionally, in order to meet the risk distribution and LTV constraints, 65% of the total number of high risk customers were excluded.

The results presented in Table 6.7 clearly indicate that an interaction between price, LTV, take-up probabilities and survival probabilities exist, specifically if the relationships depicted in Figures 6.1–6.5 are present. Most importantly, there is a clear reduction in the average LTV per risk category, when including the survival probabilities in the credit price and LTV optimisation problem, ultimately reducing the overall risk in the portfolio. That is, the impact of the average estimated survival probabilities decreasing with an increase in the LTV (refer to Figures 6.3 (b) and 6.4 (b)) is clear when considering the average LTVs in Tables 6.4–6.7. Similarly, the impact of higher prices on the average estimated survival probabilities (see Figure 6.3 (a) and 6.4 (a)), is evident when considering the exclusion of high risk customers from the portfolio when constraints are imposed on the risk distribution.

When considering data set 2 (Weibull baseline distribution), the trends and relationships are very similar to those discussed above for the exponential baseline distribution. However, the shape of the estimated survival curve is different and will be given below.

Data set 2: Weibull baseline distribution

The censoring percentage for this data set was 85% of which 15% were susceptible to default and 70% not susceptible to default. Overall, 30% of the customers were susceptible to default of which 15% defaulted during the observation period. A summary of these percentages can be found in Table 6.8.

Table 6.8: Percentages of censored customers and customers susceptible to default for the simulated data (Weibull baseline).

Description	**Proportion**
Censored, $\delta = 0$	85%
Default, $\delta = 1$	15%
Susceptible to default, $\Upsilon = 1$	30%
Not susceptible to default, $\Upsilon = 0$	70%
Susceptible to default and defaulted, $\Upsilon = 1,\ \delta = 1$	15%
Not susceptible to default and censored, $\Upsilon = 0,\ \delta = 0$	70%
Susceptible to default and censored, $\Upsilon = 1,\ \delta = 0$	15%

After fitting a mixture cure model to the survival times generated from a Weibull baseline distribution, it was found that the estimated average probability of being susceptible to default was 31.5%. Clearly this is again close to the true susceptible percentage of 30% (see Table 6.8). Furthermore, the average unconditional estimated survival probability at the last month of the loan term, $t = 240$, was 68.7%, which is close to the true percentage of customers not susceptible to default which was 70%.

As mentioned previously, the average estimated survival curves for the Weibull baseline distribution display a similar trend, that is, for increasing prices and LTVs, the average estimated survival curves decrease, respectively. However, the shape is different and therefore the average estimated survival curve per risk category is shown below to depict the different shape. All the other tables and figures pertaining to this data set can be found in Appendix C. Note that, the same response model was used for this data set, and therefore, the relationships between price, LTV and the estimated take-up probabilities are the same as illustrated in Figures 6.1 and 6.2, respectively.

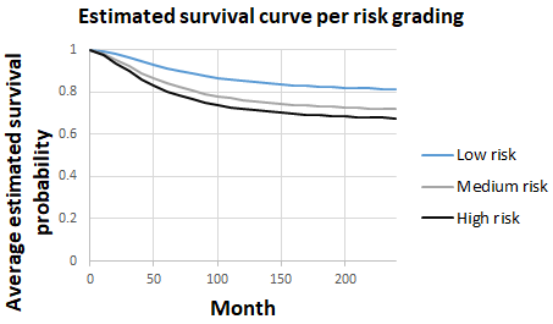


Figure 6.6: Estimated survival curve for different risk gradings.

The average estimated survival curves in Figure 6.6, when assuming a Weibull baseline distribution, now have a S-shape, which is different to the shape when using an exponential baseline distribution, as displayed in Figure 6.5. However, the lower risk customers still have a higher average estimated survival probability, for each month, compared to the medium and high risk customers.

To consider the impact of a different baseline distribution, the optimisation model with constraints on the risk distribution together with the constraints on the LTV, is solved. That is, the constraints $U_g = \{1, 0.3, 0.2\}$ are imposed on the risk distribution whereas the constraints on the LTV are set to $U_y = 60\%$ and $y^b = 0.9$. The corresponding results are displayed in Table 6.9.

Table 6.9: Computational results for the risk distribution and LTV constrained problem incorporating survival probabilities (exponential baseline)

Risk grading	Take-up proportion	Average price	Average LTV
Low	0.50	0.90	0.97
Medium	0.30	1.00	0.88
High	0.20	1.00	0.85

The results in Table 6.9 indicate that the average LTV for the low risk customers is again higher compared to the other risk categories, with the high risk categories having the lowest average LTV. The low risk customers also have the lowest average price, whereas the medium and high risk customers have the same average prices. Generally, the model behaviour is the same for the different baseline distributions considered when constraints are imposed on the risk distribution and LTVs.

However, an interesting question now arises; how will these trends and relationships change if the baseline distribution is incorrectly specified when fitting the parametric mixture cure model. To investigate this, a mixture cure model with an exponential baseline distribution was fitted to data set 2. That is, a mixture cure model with an exponential baseline distribution was fitted to the data set simulated from a mixture cure model with a Weibull baseline distribution. The results for this data set will now be discussed.

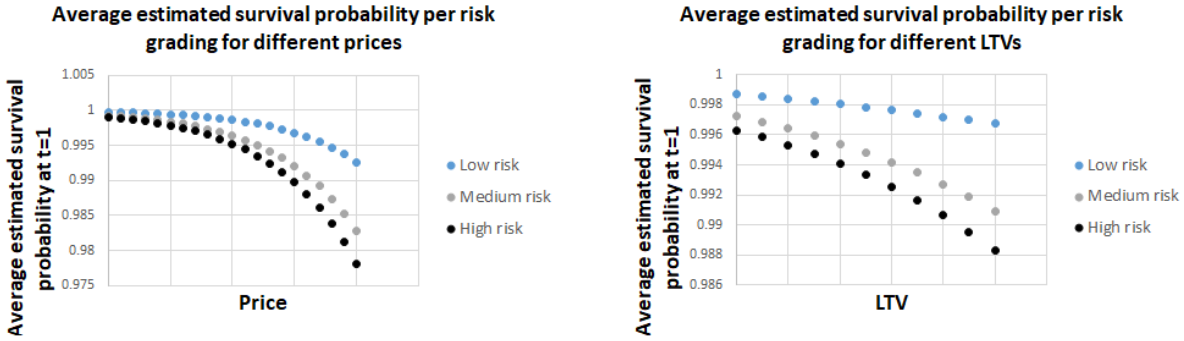
Mixture cure model with exponential baseline distribution fitted to the data set simulated from a Weibull baseline distribution.

The estimated parameter values for the fitted model can be found in Table 6.10 below.

Table 6.10: Estimated parameter values of mixture cure model with exponential baseline fitted to data simulated from Weibull baseline.

Baseline	Estimated parameters	$\hat{\eta}_0$	$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Exponential	$\hat{\lambda} = 0.00114$	-2.49	0.04	5.00	15.00	0.01	14.99	11.99

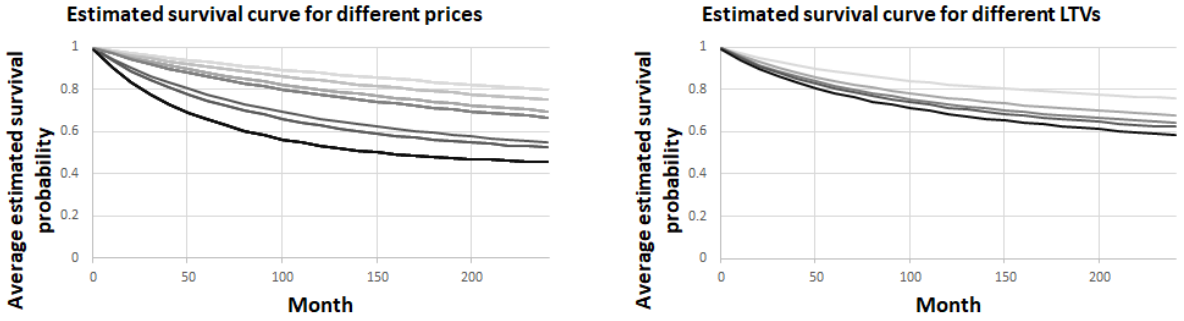
The true proportions for this simulated data set is the same as in Table 6.8. However, after fitting a mixture cure model with exponential baseline to the survival times generated from a Weibull baseline distribution, it was found that the estimated average probability of being susceptible to default was 56.3%. Clearly, this is not close to the true susceptible percentage which was 30% (see Table 6.8). In addition to this, the average unconditional estimated survival probability at the last month of the loan term, $t = 240$, was 62%, which is much lower than the true percentage of customers not susceptible to default that was 70%. To consider the impact of specifying the wrong baseline distribution, the various relationships between price, LTV and the average estimated survival probabilities are given below. The average estimated survival probabilities per risk grading at month one ($t = 1$), for different prices and LTVs, are shown in Figures 6.7 (a) and (b). Both these figures display similar trends compared to Figures 6.3 (a) and (b) for data set 1.



(a) (b)

Figure 6.7: (a) Relationship between price and estimated survival probability in month one; (b) Relationship between LTV and estimated survival probability in month one.

The estimated unconditional survival curve for increasing values of the price and LTV are given in Figures 6.8 (a) and (b), respectively. Once again, for increasing values of the price and LTV, the average estimated survival probabilities for each month are smaller, when all the other covariates are kept constant.



(a) (b)

Figure 6.8: (a) Estimated survival curve for different prices; (b) Estimated survival curve for different LTVs.

Similarly, when the average estimated survival curve for the different risk gradings are considered, the lower risk customers have a higher average estimated survival probability, for each month, compared to the medium and high risk customers.

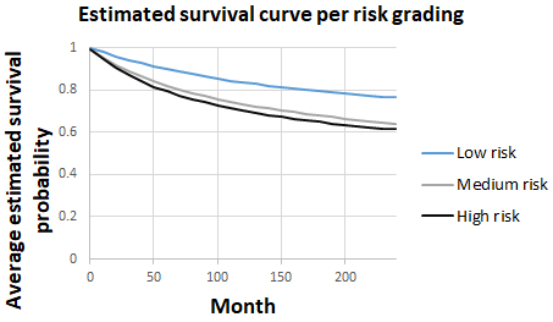


Figure 6.9: Estimated survival curve for different risk gradings.

Now, to consider the impact of fitting a mixture cure model with the wrong (exponential) baseline distribution to data set 2, the optimisation model with constraints on the risk distribution together with the constraints on the LTV, is solved. That is, the constraints $U_g = \{1, 0.3, 0.2\}$ are imposed on the risk distribution whereas the constraints on the LTV are set to $U_y = 60\%$ and $y^b = 0.9$.

Table 6.11: Computational results for the risk distribution and LTV constrained problem incorporating survival probabilities (incorrect baseline distribution)

Risk grading	Take-up proportion	Average price	Average LTV
Low	0.50	0.86	0.89
Medium	0.30	0.95	0.81
High	0.20	1.00	0.80

The results in Table 6.11 reveal the interaction between price, LTV, take-up and survival probabilities are similar when the wrong baseline distribution was fitted to the data. That is, to meet the constraints imposed on the risk distribution and LTVs, the high risk customers are quoted a higher price together with lower LTVs on average, whereas the low risk customers are quoted lower prices in conjunction with higher LTVs on average. This is also confirmed by the trends and relationships displayed in Figures 6.7–6.9. However, the average optimal price and LTV per risk category, obtained when fitting the wrong model, were at least 5% less compared to corresponding risk category's average optimal prices and LTVs, when fitting the correct model to the data. This is, indeed, a confirmation that the average optimal prices and LTVs are affected if the wrong baseline distribution is specified when fitting the mixture cure model.

For the purpose of evaluating scalability, the optimisation model including survival probabilities was also applied to problem instances with larger sample sizes. For all these problems, which include risk distribution and LTV constraints, the model behaviour was similar to that provided in Table 6.7. That is, on average the price offered to the low risk customers were lower than the price offered to the high risk customers whereas the LTV proposed to the low risk customers was, on average, higher than that of the high risk customers. A limit was set on computing time to find the optimal solutions

and subsequently, the optimality gap was observed for each of the larger samples considered. The optimality gap refers to the absolute difference between an incumbent solution *e.g.* the best known solution for the MILP, and the value that bounds the best possible solution. Figure 6.10 displays the optimality gap as a percentage of the best bound for sample sizes of 100, 200, 400, 800 and 1600 customers when the optimisation problem was solved with a time limit of one hour. The computational tests were performed on an HP Compaq Elite computer with 32 GB of RAM operating at 3.4 GHz. The IBM product, CPLEX v12.8 was used for solving the MILP models.

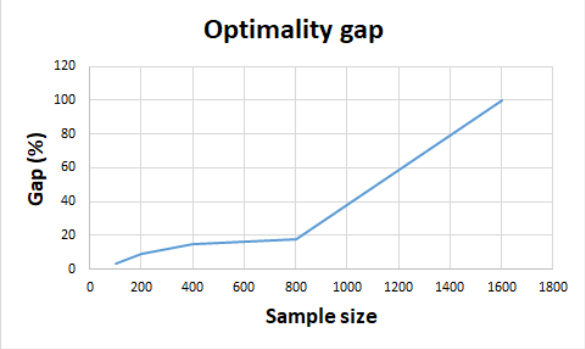


Figure 6.10: Optimality gap for larger samples.

As expected, the optimality gap increases as the sample size increases as this clearly increases the size of the MILP problem. As mentioned, the model behaviour for the larger samples remain similar, the time in which the larger problems are solved just increases and more computing power is needed.

Including the survival probabilities in the price and LTV optimisation problem, permits one to take into account the repayment probability at each month when calculating the expected value of the NPII. More specifically, by using the mixture cure model to estimate the survival probabilities, the proportion of customers that are not susceptible to default can be taken into account when determining the optimal price and LTV to quote prospective customer. In this way the impact of price and LTV on both the estimated take-up probability and estimated survival probability (at each month) can be taken into account. In addition to this, by using a piece-wise linear approximation approach to solve this optimisation problem, logical decision making capability can be introduced and various constraints on the risk distribution and LTV can be imposed, to ultimately reduce the risk in the portfolio. Moreover, as a result of including logical decision making variables, customers with lower survival probabilities (or higher default probability), are excluded from the portfolio, when the price quoted for these customers become too high. Therefore, with the added benefit of modelling the repayment probability in each month, a more realistic problem is formulated when the objective is to determine the optimal price and LTV that maximise the expected value of the NPII.

Chapter 7

Conclusion and future research

7.1 Concluding remarks

In this thesis, an optimisation model was developed that maximises the expected value of the NPII by finding the right balance between price and LTV according to a price response model. This model is able to incorporate the effect of price and LTV on the survival behaviour of the customers during the loan. The model was implemented on two data sets, where the survival times were simulated from a parametric mixture cure model using the covariates of data obtained from a financial institution. From the results obtained it were clear that, when solving the price and LTV optimisation problem that includes survival probabilities, the effect of these variables on the take-up and survival probabilities played a key role. This model can ultimately assist a financial institution to determine the price to quote a customer for the corresponding loan amount.

The main objectives of this study, with their corresponding outcomes can be summarised as follows:

- Review the existing literature on different pricing methodologies.

  A review on different pricing methodologies was conducted in Chapter 2.

- Review the existing literature on mathematical optimisation and the various methods used to solve optimisation problems.

  This was done in Chapter 3 with specific emphasis on a piece-wise linear approximation approach to solve NLP problems, as this was needed to solve the optimisation problem.

- Develop a new optimisation model that determines the optimal price and LTV that maximises the expected NPII by using a piece-wise linear approximation approach.

  This model was developed in Chapter 4, where a piece-wise linear approximation approach was followed to simultaneously determine the optimal price and LTV. This laid the foundation for the final model which incorporates survival probabilities.

- Discuss the existing literature on survival analysis with emphasis on different survival models, which include the Cox Proportional Hazards (CPH) model and the mixture cure model (as a more general alternative to the CPH).

  A review of some basic concepts of survival analysis were presented with the focus specifically on the CPH and mixture cure models.

- Develop a new optimisation model that maximise the expected NPII by finding the right balance between price and LTV and incorporating estimated survival probabilities obtained from a mixture cure model.

  A final model incorporating all the above mentioned aspects was developed and applied to two simulated data sets.

A new goodness-of-fit test for exponentiality, that can ultimately be used to determine the adequacy of fit of a parametric CPH model, was developed, of which the published paper is presented in Appendix A.

We now conclude the thesis by discussing two possible avenues for future research.

7.2 Possible future research

7.2.1 Mixture cure model with multiple events.

In this thesis the assumption was made that default is the only event of interest. However, in consumer credit, various events of interest can occur before the date of gathering the data that will be used to estimate the survival probabilities. These include default, early settlement and maturity. Watkins et al. (2014) proposed a method that provides for the simultaneous modelling of multiple events of interest. A possible way to include the modelling of multiple events in our optimisation problem is to consider the following model,

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

where the superscripts d and e refer to the events default and early settlement, respectively. A number of questions now arises if multiple events are present:

- Should the same baseline distribution be used in modelling the conditional survival probabilities $S^d\left(\cdot \mid \Upsilon^d = 1, \underline{w}\right)$ and $S^e\left(\cdot \mid \Upsilon^e = 1, \underline{w}\right)$?

- How will this influence the parameter estimation of the various model components if some of the events rarely occur?

- How do we test whether the parametric assumptions are violated seeing that the model now consists of many more components?

7.2.2 A possible goodness-of-fit test for the parametric mixture cure model with covariates.

As was discussed in Chapter 5, there is currently no goodness-of-fit test for the parametric mixture cure model where covariates are present. However, such a test is essential to access whether the parametric assumptions of the mixture cure model are violated. We will discuss a potential omnibus goodness-of-fit test for this scenario. Consider the mixture cure model in (5.9), where a parametric assumption is made about the baseline distribution. If the model is correctly specified, then $F(Y \mid \underline{v}, \underline{w}) = 1 - S(Y \mid \underline{v}, \underline{w})$ will be uniformly distributed on the interval $[0, \pi(\underline{v})]$, again where this follows from the well-known probability integral transform. This implies that

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

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

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

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

It is important to note that one will have to use a test for uniformity that is modified to accommodate random censoring. Koziol (1980) and Fleming et al. (1980) discuss extensions of the Kolmogorov-Smirnov and Cramer-Von Mises type tests that can be used for this purpose. Some open questions with regards to this possible new goodness-of-fit test includes:

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

This all needs to be investigated.

References

Amdahl, J. (2019). Flexible parametric cure models. *https://cran.r-project.org/web/packages/flexsurvcure/flexsurvcure.pdf*.

Amico, M. and Van Keilegom, I. (2018). Cure models in survival analysis. *Annual Review of Statistics and Its Application*, 5:311–342.

Babayev, D. A. (1997). Piece-wise linear approximation of functions of two variables. *Journal of Heuristics*, 2(4):313–320.

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

Bazaraa, M. S., Jarvis, J. J., and Sherali, H. D. (2011). *Linear programming and network flows*. John Wiley & Sons.

Beale, E. M. L. and Tomlin, J. A. (1970). Special facilities in a general mathematical programming system for non-convex problems using ordered sets of variables. *OR*, 69:447–454.

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

Bixby, R. E. (2012). A brief history of linear and mixed-integer programming computation. *In Documenta Mathematica, Extra Volume: Optimization Stories. 21st International Symposium on Mathematical Programming*, pages 107–121.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

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(3):437–453.

Breslow, N. E. (1972). Discussion of professor cox's paper. *Journal of the Royal Statistical Society, Series B*, 34:216–217.

Cai, C., Zou, Y., Peng, Y., and Zhang, J. (2012). smcure: An r-package for estimating semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine*, 108(3):1255–1260.

Caufield, S. (2012). Consumer credit pricing. *The Oxford Handbook of Pricing Management, Oxford University Press, Oxford*, pages 138–152.

Cornuéjols, G. (2008). Valid inequalities for mixed integer linear programs. *Mathematical Programming*, 112(1):3–44.

Cornuejols, G. and Tütüncü, R. (2006). *Optimization methods in finance*. Cambridge University Press.

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

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

D'Ambrosio, C., Lodi, A., and Martello, S. (2010). Piecewise linear approximation of functions of two variables in milp models. *Operations Research Letters*, 38(1):39–46.

Dantzig, G. B. (1948). Programming in a linear structure. *Bulletin of the American Mathematical Society*, 54(11):1074–1074.

Dantzig, G. B. (1990). Origins of the simplex method. Technical report, Stanford University.

Dirick, L., Claeskens, G., and Baesens, B. (2015). An akaike information criterion for multiple event mixture cure models. *European Journal of Operational Research*, 241(2):449–457.

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

Ecoinomic (2018). A brief history of loans from antiquity to the present day. *https://medium.com/ecoinomic/a-brief-history-of-loans-from-antiquity-to-the-present-day-a3831b7d52ed*.

Edelberg, W. (2006). Risk-based pricing of interest rates for consumer loans. *Journal of Monetary Economics*, 53(8):2283–2298.

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

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38(4):1041–1046.

Fleming, T. R., O'Fallon, J. R., O'Brien, P. C., and Harrington, D. P. (1980). Modified kolmogorov-smirnov test procedures with application to arbitrarily right-censored data. *Biometrics*, 36(4):607–625.

Geerdens, C., Janssen, P., and Van Keilegom, I. (2019). Goodness-of-fit test for a parametric survival function with cure fraction. *TEST*, pages 1–25.

Gomory, R. (1960). An algorithm for the mixed integer problem. Technical report, Rand Corp Santa Monica Ca.

Gomory, R. E. (2010). Outline of an algorithm for integer solutions to linear programs and an algorithm for the mixed integer problem. In *50 Years of Integer Programming 1958-2008*, pages 77–103. Springer.

Guardia, N. D. (2002). Consumer credit in the european union. *ECRI Research Report*, 1:1–39.

Hand, D. J. (2001). Modelling consumer credit risk. *IMA Journal of Management Mathematics*, 12(2):139–155.

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

Johnson, R. W. (1992). Legal, social and economic issues in implementing scoring in the us. *Credit scoring and credit control*, 19:32.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on cox's regression and life model. *Biometrika*, 60(2):267–278.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Karlan, D. S. and Zinman, J. (2008). Credit elasticities in less-developed economies: Implications for microfinance. *American Economic Review*, 98(3):1040–68.

Klein, J. P. and Moeschberger, M. L. (2006). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G., and Scheike, T. H. (2016). *Handbook of survival analysis*. CRC Press.

Koziol, J. A. (1980). Goodness-of-fit tests for randomly censored data. *Biometrika*, 67(3):693–696.

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

Land, A. H. and Doig, A. G. (2010). An automatic method for solving discrete programming problems. In *50 Years of Integer Programming 1958-2008*, pages 105–132. Springer.

Leonard, I. E. and Lewis, J. E. (2015). *Geometry of convex sets*. John Wiley & Sons.

Li, C. and Taylor, J. M. (2002). A semi-parametric accelerated failure time cure model. *Statistics in Medicine*, 21(21):3235–3247.

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

Ma, P., Crook, J., and Ansell, J. (2010). Modelling take-up and profitability. *Journal of the Operational Research Society*, 61(3):430–442.

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

Ma, Z. and Krings, A. W. (2008). Survival analysis approach to reliability, survivability and prognostics and health management (phm). In *2008 IEEE Aerospace Conference*, pages 1–20. IEEE.

Magri, S. and Pico, R. (2011). The rise of risk-based pricing of mortgage interest rates in italy. *Journal of Banking & Finance*, 35(5):1277–1290.

Maller, R. A. and Zhou, X. (1996). *Survival analysis with long-term survivors*. John Wiley & Sons.

Mester, L. J. et al. (1997). What's the point of credit scoring? *Business Review*, 3(Sep/Oct):3–16.

Misener, R. and Floudas, C. (2010). Piecewise-linear approximations of multidimensional functions. *Journal of Optimization Theory and Applications*, 145(1):120–147.

Mitchell, J. E. (2002). Branch-and-cut algorithms for combinatorial optimization problems. *Handbook of Applied Optimization*, 1:65–77.

Murray, J. (2019). What is lending and types of lenders? *https://www.thebalancesmb.com/what-is-lending-what-are-lenders-398319*.

Narain, B. (1992). Survival analysis and the credit granting decision. *In Thomas, L.C., Crook, J.N. and Edelman, D.B., Eds., Credit Scoring and Credit Control*, pages 109–121.

Özer, Ö., Ozer, O., and Phillips, R. (2012). *The Oxford handbook of pricing management*. Oxford University Press.

Padberg, M. and Rinaldi, G. (1987). Optimization of a 532-city symmetric traveling salesman problem by branch and cut. *Operations Research Letters*, 6(1):1–7.

Park, S. (1997). Effects of price competition in the credit card industry. *Economics Letters*, 57(1):79–85.

Peng, Y. (2003). Fitting semiparametric cure models. *Computational Statistics & Data Analysis*, 41(3-4):481–490.

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

Phillips, R. (2013). Optimizing prices for consumer credit. *Journal of Revenue and Pricing Management*, 12(4):360–377.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Redden, J. (2014). A brief history of loans. *The Calculator Site. https://www.thecalculatorsite.com/articles/finance/history-of-loans.php*.

Siddiqi, N. (2012). *Credit risk scorecards: developing and implementing intelligent credit scoring*. John Wiley & Sons.

Skugge, G. (2011). The future of pricing: Outside-in. *Journal of Revenue and Pricing Management*, 10(4):392–395.

Smuts, M. and Terblanche, S. (2019). The simultaneous optimisation of price and loan-to-value. In *Proceedings of the 48th Annual Conference of the Operations Research Society of South Africa*, pages 88–95.

Stiglitz, J. E. and Weiss, A. (1981). Credit rationing in markets with imperfect information. *The American Economic Review*, 71(3):393–410.

Stojković, N. V., Stanimirović, P. S., and Petković, M. D. (2009). Modification and implementation of two-phase simplex method. *International Journal of Computer Mathematics*, 86(7):1231–1242.

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

Taylor, B. (2009). Integer programming: The branch and bound method. *Introduction to Management Science*.

Taylor, J. M. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics*, 51(3):899–907.

Terblanche, S. and De la Rey, T. (2014). Credit price optimisation within retail banking. *ORiON*, 30(2):85–102.

Therneau, T. M. (2020). *A Package for Survival Analysis in R*. R package version 3.1-12.

Thomas, L. C. (2009). *Consumer credit models: pricing, profit and portfolios*. OUP Oxford.

Tolley, H., Barnes, J., and Freeman, M. (2016). Survival analysis. In *Forensic Epidemiology*, pages 261–284. Elsevier.

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

Walke, A. G., Fullerton Jr, T. M., and Tokle, R. J. (2018). Risk-based loan pricing consequences for credit unions. *Journal of Empirical Finance*, 47:105–119.

Watkins, J. G., Vasnev, A. L., and Gerlach, R. (2014). Multiple event incidence and duration analysis for credit data incorporating non-stochastic loan maturity. *Journal of Applied Econometrics*, 29(4):627–648.

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

Appendix A: A new goodness-of-fit test for exponentiality based on a conditional moment characterisation

Volume 35 (2), pp. 145–160
http://orion.journals.ac.za

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

New goodness-of-fit tests for exponentiality based on a conditional moment characterisation

M Smuts* JS Allison† L Santana‡

Received: 23 August 2019; Accepted: 30 September 2019

Abstract

The exponential distribution plays a key role in the practical application of reliability theory, survival analysis, engineering and queuing theory. These applications often rely on the underlying assumption that the observed data originate from an exponential distribution. In this paper, two new tests for exponentiality are proposed, which are based on a conditional second moment characterisation. The proposed tests are compared to various established tests for exponentiality by means of a simulation study where it is found that the new tests perform favourably relative to the existing tests. The tests are also applied to real-world data sets with independent and identically distributed data as well as to simulated data from a Cox proportional hazards model, to determine whether the residuals obtained from the fitted model follow a standard exponential distribution.

Key words: Characterisation, Cox's proportional hazards model, exponential distribution, goodness-of-fit test.

1 Introduction

The exponential distribution is an important and commonly used statistical model for a multitude of real-life phenomena, such as lifetimes, time to default of loans, and many other *time-to-event* scenarios. As a result, this distribution plays a vital role in the practical application of reliability theory, survival analysis, engineering, and queuing theory

*Subject group Statistics, Department of Statistics, North-West University, Potchefstroom, South Africa email: smuts.marius@nwu.ac.za

†Subject group Statistics, Department of Statistics, North-West University, Potchefstroom, South Africa email: james.allison@nwu.ac.za

‡Corresponding author: Subject group Statistics, Department of Statistics, North-West University, Potchefstroom, South Africa email: leonard.santana@nwu.ac.za

http://dx.doi.org/10.5784/35-2-661

(to name only a few), as the underlying theory governing these applications often assumes an exponential distribution for the data. Therefore, to effectively implement these applications, it is necessary to perform goodness-of-fit tests to determine whether these fundamental distributional assumptions are satisfied or not. Examples where the assumption of exponentiality is necessary (and hence the need to test for this assumption) range from the analysis of queuing networks [26], to cancer clinical trials [15], and the time-to-failure of systems of machines and operators [12]; for further examples of data sets see the papers by Shanker, Fesshaye, and Selvaraj [23, 24].

Suppose that a random variable X follows an exponential distribution with scale parameter λ (written $X \sim Exp(\lambda)$). This random variable has a number of unique distributional properties which include the forms of its cumulative distribution function (CDF), survival function, probability density function, and characteristic function (CF), which are given by

$$F(x) = \mathrm{P}(X < x) = 1 - e^{-\lambda x},$$
$$S(x) = \mathrm{P}(X > x) = 1 - F(x) = e^{-\lambda x},$$
$$f(x) = \lambda e^{-\lambda x},$$

and

$$\phi(x) = \frac{\lambda}{\lambda - ix}, \quad i = \sqrt{-1},$$

respectively, with $x > 0$ and where $\lambda > 0$ is the scale parameter, with $\mathrm{E}(X) = 1/\lambda$. In addition, the exponential distribution also exhibits many other unique distributional properties, called *characterisations*. These characterisations help in the development of tests for exponentiality since, if one can verify that the data has these properties, then one can conclude that the data were obtained from an exponential distribution. One such property is the so-called 'memoryless' property which states that, if X follows an exponential distribution, then we can write

$$\mathrm{P}(X > s + t \,|\, X > s) = \mathrm{P}(X > t), \tag{1}$$

for $s, t > 0$. This property implies that, if X represents the lifetime of a certain component, then the remaining lifetime of that component is independent of its current age. For components that suffer from wear-and-tear (*i.e.*, where the lifetime is dependent on its current age), the exponential distribution would not be an appropriate model. A second property states that the exponential distribution uniquely has the feature that the *hazard rate* is a constant, that is,

$$h(x) = \frac{f(x)}{1 - F(x)} = \lambda.$$

This feature is directly tied to the memoryless property since the failure rate is constant throughout the lifetime of the component.

Suppose now that $X_1, X_2, \ldots, X_n$ are realisations from some random variable X with unknown distribution function F, then the process of testing whether or not this data are realisations from an exponential distribution with parameter λ involves the use of statistical inference via goodness-of-fit tests. The inferential question can be framed in the form of the following composite hypothesis statement:

$$H_0 : \text{The distribution of } X \text{ is } \exp(\lambda), \tag{2}$$

for some $\lambda > 0$, against the alternative hypothesis that the distribution of X is something other than exponential. Note that the tests that will be discussed in this paper all make use of a scaled version of the original data, defined as $Y_j = X_j\widehat{\lambda}$, $j = 1, 2, \ldots, n$, where $\widehat{\lambda}$ denotes the maximum likelihood estimator (MLE) of the parameter λ and is given by $\widehat{\lambda} = 1/\bar{X}_n$ with $\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$. The motivation for the use of this scaling factor primarily comes from the fact that the distribution of exponential random variables is invariant to simple scale transformations, that is, X is exponentially distributed if, and only if, cX is also exponentially distributed for every constant $c > 0$. Therefore, conclusions drawn regarding exponentiality based on the sample $Y_1, Y_2, \ldots, Y_n$ can reasonably be extended to the exponentiality of X (from which $X_1, X_2, \ldots, X_n$ was obtained). Furthermore, many statistics discussed will also employ the order statistics of X_j and Y_j, defined as $X_{(1)}<X_{(2)}<\ldots<X_{(n)}$ and $Y_{(1)}<Y_{(2)}<\ldots<Y_{(n)}$, respectively.

To test the hypothesis in (2), formal test statistics are used, some of which are general statistics that can be applied to test for almost any distribution, whereas others exploit the various unique characteristics of exponentially distributed data, such as the memoryless or constant hazard properties. Examples of general tests employed to test exponentiality include the Kolmogorov-Smirnov test and the Cramér-von Mises test, both of which are based on the same basic principle of measuring the discrepancies between the theoretical CDF of the exponential distribution and its empirical equivalent (see, *e.g.*, Chapter 4 of D'Agostino and Stephens, 1986 [11]). Another class of general tests involves a similar approach, but replaces the CDF with the CF; examples of these include the test by Epps and Pulley [13], and the one by Meintanis, Swanepoel and Allison [21]. In addition, there are many goodness-of-fit tests based on the unique properties of the exponential distribution, which are occasionally desirable as they tend to focus on much more specific aspects of the distribution and are potentially more robust than the more general tests. For example, since the memoryless property uniquely characterises the exponential distribution, this implies that the exponentially distributed random variable X will have this property and conversely, if X exhibits this property it must be exponentially distributed. Therefore, a test based on this property will involve first determining sample estimates of the two probabilities appearing on either side of the expression of (1), and the test can then be designed to measure the equality of these two estimates. For examples of tests based on this characterisation, see [2], [5], and [6].

There are many more such unique characterisations of the exponential distribution and the literature on goodness-of-fit contains numerous test statistics based on these characterisations. For example, for tests based on the mean residual life, see [8], [25], [17], and [9]. For a test based on the Arnold-Villasenor characterisation, see [18], and for a test based on the Rossberg characterisation see [27]. For a comprehensive review of tests for exponentiality, the interested reader is referred to the review papers by Ascher [7], Henze and Meintanis [16], and Allison, Santana, Smit and Visagie [4].

The remainder of the paper is organised as follows: In Section 2 we propose new tests for exponentiality which are based on a conditional second moment characterisation of the exponential distribution and, in Section 3, the results of a brief Monte Carlo simulation are presented to compare the power performance of the newly proposed tests to some commonly used existing tests for exponentiality. The paper concludes in Section

4 where the tests are applied to some real-world data sets with independent and identically distributed random values, as well as to data simulated from a Cox proportional hazards model, to determine whether the residuals obtained from the fitted model follow a standard exponential distribution.

2 New tests for exponentiality based on a characterisation

Consider the following characterisation of the exponential distribution by Afify, Nofal and Ahmed [1]:

Characterisation. *Let X be a non-negative random variable with continuous distribution function F and density f. If $E(X^2) < \infty$, then X has an exponential distribution with parameter λ (that is, $F(x) = 1 - e^{-\lambda x}$) if, and only if, for all $t > 0$*

$$E\left[X^2 | X > t\right] = \frac{2}{\lambda^2} + h(t)\left(\frac{t^2}{\lambda} + \frac{2t}{\lambda^2}\right),$$

where $h(t) = \frac{f(t)}{S(t)}$ is the hazard rate.

From this characterisation we can deduce the following corollary.

Corollary 1 *Let X be a non-negative random variable with continuous distribution function F. If $E(X^2) < \infty$, then X has an exponential distribution with parameter λ if, and only if, for all $t > 0$*

$$E\left[X^2 \,\mathrm{I}(X > t)\right] = S(t) r_\lambda(t),$$

where $r_\lambda(t) := \frac{2}{\lambda^2} + h(t)\left(\frac{t^2}{\lambda} + \frac{2t}{\lambda^2}\right)$ and $\mathrm{I}(\cdot)$ denotes the indicator function.

Proof: Straightforward calculations yield that, for all $t > 0$,

$$E\left[X^2 | X > t\right] = \frac{1}{P(X > t)} E\left[X^2 \,\mathrm{I}(X > t)\right] = \frac{1}{S(t)} E\left[X^2 \,\mathrm{I}(X > t)\right].$$

From the Characterisation, it follows that X has an exponential distribution with parameter λ if, and only if, for all $t > 0$,

$$\frac{1}{S(t)} E\left[X^2 \,\mathrm{I}(X > t)\right] = r_\lambda(t),$$

or equivalently if, and only if, for all $t > 0$,

$$E\left[X^2 \,\mathrm{I}(X > t)\right] = S(t) r_\lambda(t).$$

□

Now, note that if $X \sim Exp(\lambda)$ then

$$Y = \lambda X \sim \exp(1).$$

Based on Y, the characterisation in Corollary 1 can be restated as follows: Y is exponentially distributed if, and only if,

$$E\left[Y^2\,\mathrm{I}(Y>t)\right] = S(t)r_1(t),$$

where $S(t) = P(Y > t)$ and $r_1(t) = 2 + h(t)\left(t^2 + 2t\right)$, with $h(t)$ the hazard rate of Y.

Based on this, a random variable Y has a standard exponential distribution if, and only if,

$$E\left[Y^2\,\mathrm{I}(Y>t)\right] - S(t)r_1(t) = 0,$$

or equivalently if, and only if,

$$\psi(t) \coloneqq E\left[Y^2\,\mathrm{I}(Y>t)\right] - 2S(t) - f(t)\left\{t^2 + 2t\right\} = 0,$$

where $f(t)$ is the density function of Y.

Naturally, $\psi(t)$ is unknown and hence must be estimated from the data $Y_1, Y_2, \ldots, Y_n$, where $Y_j = X_j\widehat{\lambda} = X_j/\bar{X}_n$, $j = 1, 2, \ldots, n$. Define two possible estimators for $\psi(t)$ by

$$\widehat{\psi}_n^{(1)}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i^2\,\mathrm{I}\left(Y_i > t\right) - \frac{2}{n}\sum_{i=1}^{n}\mathrm{I}\left(Y_i > t\right) - e^{-t}\left(t^2 + 2t\right)$$

and

$$\widehat{\psi}_n^{(2)}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i^2\,\mathrm{I}\left(Y_i > t\right) - \frac{2}{n}\sum_{i=1}^{n}\mathrm{I}\left(Y_i > t\right) - \widehat{f}(t)\left(t^2 + 2t\right).$$

Here, $\widehat{f}(t)$ denotes the kernel density estimate of $f(t)$, which is defined as

$$\widehat{f}(t) = \frac{1}{nh}\sum_{i=1}^{n}\phi\left(\frac{t - Y_i}{h}\right),$$

where $\phi(\cdot)$ is the standard normal density function and h is a suitably chosen bandwidth (for an in-depth discussion on kernel density estimators, the interested reader is referred to the monograph by Wand and Jones [28]). The only difference between the estimators $\widehat{\psi}_n^{(1)}$ and $\widehat{\psi}_n^{(2)}$ is that in $\widehat{\psi}_n^{(1)}$ we choose $f(t) = e^{-t}$, the density function specified under the null hypothesis, whilst in $\widehat{\psi}_n^{(2)}$ we estimate f by $\widehat{f}$.

Now, if the observed data originated from an exponential distribution, then both $\widehat{\psi}_n^{(1)}$ and $\widehat{\psi}_n^{(2)}$ should be close to zero. This leads to the following two Cramér-von Mises type test statistics

$$S_n = n\int_0^{\infty}\left[\widehat{\psi}_n^{(1)}(t)\right]^2 w(t)dF_n(t)$$

and

$$T_n = n\int_0^{\infty}\left[\widehat{\psi}_n^{(2)}(t)\right]^2 w(t)dF_n(t),$$

where $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(Y_i \leq t)$ is the empirical distribution function of $Y_1, Y_2, \ldots, Y_n$ and $w(t)$ is a suitable, positive weight function satisfying some standard integrability conditions. For implementation of the proposed test statistics, we will use $w(t) = e^{-at}$, where $a > 0$ is a user-defined tuning parameter.

With this choice of $w(t)$ the following easily calculable form of the proposed test statistics S_n and T_n is obtained.

$$S_{n,a} = \sum_{j=1}^{n} \left[\frac{1}{n} \sum_{i=1}^{n} Y_i^2 \, \mathrm{I}\left(Y_i > Y_j\right) - \frac{2}{n} \sum_{i=1}^{n} \mathrm{I}\left(Y_i > Y_j\right) - e^{-Y_j} \left(Y_j^2 + 2Y_j\right) \right]^2 e^{-aY_j}$$

and

$$T_{n,a} = \sum_{j=1}^{n} \left[\frac{1}{n} \sum_{i=1}^{n} Y_i^2 \, \mathrm{I}\left(Y_i > Y_j\right) - \frac{2}{n} \sum_{i=1}^{n} \mathrm{I}\left(Y_i > Y_j\right) - \widehat{f}(Y_j) \left(Y_j^2 + 2Y_j\right) \right]^2 e^{-aY_j}.$$

Both tests reject the null hypothesis in (2) for large values of the test statistics. The critical values for the test statistics can easily be calculated using the following Monte Carlo procedure.

1. Draw a random sample $X_1, X_2, \ldots, X_n$ from an exponential distribution with parameter 1.
2. Obtain the scaled observations $Y_i = X_i / \bar{X}_n, i = 1, 2, \ldots, n$.
3. Calculate the test statistic, say $S = S_n(Y_1, Y_2, \ldots, Y_n)$.
4. Repeat steps 1–3 a large number of times, say MC times, to obtain MC copies of S denoted $S_1, S_2, \ldots, S_{MC}$.
5. Obtain the order statistics $S_{(1)} \leq S_{(2)} \leq \cdots \leq S_{(MC)}$.
6. The critical value at a $\alpha\%$ significance level is then given by

$$\widehat{C}_n(\alpha) = S_{(\lfloor MC(1-\alpha) \rfloor)},$$

where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x.

3 Simulation study and results

In this section Monte Carlo simulations are used to compare the finite-sample performance of the newly proposed tests $T_{n,a}$ and $S_{n,a}$ to the following existing tests for exponentiality.

- The traditional tests of Kolmogorov-Smirnov (KS_n) and Cramér-von Mises (CM_n), where the test statistics for these tests are given by

$$KS_n = \max\left\{KS_n^+, KS_n^-\right\},$$

where

$$KS_n^+ = \max_{1 \leq j \leq n} \left[\frac{j}{n} - \left(1 - e^{-Y_{(j)}}\right) \right],$$

$$KS_n^- = \max_{1\le j\le n}\left[\left(1-e^{-Y_{(j)}}\right)-\frac{j-1}{n}\right]$$

and

$$CM_n = \frac{1}{12n}+\sum_{j=1}^{n}\left[\left(1-e^{-Y_{(j)}}\right)-\frac{2j-1}{2n}\right]^2.$$

Both of these tests reject the null hypothesis for large values of the test statistics

- A Kolmogorov-Smirnov type test ($\overline{KS}_n$) and a Cramér-von Mises type test ($\overline{CM}_n$) based on the mean residual life, as developed by Baringhaus and Henze [8], with the following test statistics

$$\overline{KS}_n = \sqrt{n}\sup_{t\ge 0}\left|\frac{1}{n}\sum_{j=1}^{n}\min\{Y_j,t\}-\frac{1}{n}\sum_{j=1}^{n}I\left(Y_j\le t\right)\right| = \sqrt{n}\max\left\{\overline{KS}_n^+,\overline{KS}_n^-\right\},$$

where

$$\overline{KS}_n^+ = \max_{j\in\{0,1,\ldots,n-1\}}\left[\frac{1}{n}\left(Y_{(1)}+\cdots+Y_{(j)}\right)+Y_{(j+1)}\left(1-\frac{j}{n}\right)-\frac{j}{n}\right],$$
$$\overline{KS}_n^- = \max_{j\in\{0,1,\ldots,n-1\}}\left[\frac{j}{n}-\frac{1}{n}\left(Y_{(1)}+\cdots+Y_{(j)}\right)-Y_{(j)}\left(1-\frac{j}{n}\right)\right].$$

and

$$\overline{CM}_n = n\int_0^\infty\left[\frac{1}{n}\sum_{j=1}^{n}\min\left\{Y_j,t\right\}-\frac{1}{n}\sum_{j=1}^{n}I\left(Y_j\le t\right)\right]^2 e^{-t}dt$$
$$= \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\left[2-3\exp\left(-\min\{Y_j,Y_k\}\right)-2\min\{Y_j,Y_k\}\left(e^{-Y_j}+e^{-Y_k}\right)+2\exp\left(-\max\{Y_j,Y_k\}\right)\right].$$

Both $\overline{KS}_n$ and $\overline{CM}_n$ reject the null hypothesis for large values.

- The Epps and Pulley test EP_n [13], which is based on the characteristic function, $\phi(x)$, and with the test statistic given by

$$EP_n = \sqrt{48n}\left[\frac{1}{n}\sum_{j=1}^{n}e^{-Y_j}-\frac{1}{2}\right].$$

The null hypothesis is rejected for large values of $|EP_n|$.

3.1 Simulation setting

A significance level of 5% was used throughout the study. Empirical critical values of all the tests were obtained from 10 000 independent Monte Carlo replications using the procedure given at the end of Section 2. Power estimates were calculated for sample sizes $n = 20$ and $n = 30$ using 10 000 independent Monte Carlo replications for the various alternative distributions given in Table 1.

The two new tests in which a tuning parameter appears were evaluated for $a = 0.25$ and $a = 1$. All calculations and simulations were performed in R [22].

3.2 Simulation results

Table 2 contains the percentage of the 10 000 Monte Carlo samples that resulted in the rejection of the null hypothesis in (2) rounded to the nearest integer. For each alternative the top row corresponds to the estimated powers obtained for $n = 20$ whereas the row below corresponds to the estimated powers for $n = 30$.

Alternative	$\boldsymbol{f(x)}$	**Notation**
Gamma	$\frac{1}{\Gamma(\theta)} x^{\theta-1} e^{-x}$	$\Gamma(\theta)$
Weibull	$\theta x^{\theta-1} \exp(-x^{\theta})$	$W(\theta)$
Power	$\frac{1}{\theta} x^{(1-\theta)/\theta}, \quad 0 < x < 1$	$PW(\theta)$
Linear failure rate	$(1+\theta x)\exp(-x-\theta x^2/2)$	$LFR(\theta)$
Exponential logarithmic	$\frac{(\ln\theta)^{-1}(1-\theta)e^{-x}}{(1-\theta)e^{-x}-1}$	$EL(\theta)$
Exponential geometric	$\frac{(1-\theta)e^{-x}}{(1-\theta e^{-x})^2}$	$EG(\theta)$

Table 1: *Alternative distributions considered in the simulation study.*

The highest power for each alternative distribution is highlighted for ease of comparison. From Table 2 it is clear that there is no single test that dominates all of the other tests. However, $S_{n,a}$ outperforms all its competitors for the $EG(\theta)$, $EL(\theta)$ and $\Gamma(0.7)$ alternatives and for both sample sizes. No single test dominates for the majority of the other alternatives with the exception of $T_{n,a}$, which performs well for the alternatives $LF(4)$ and $PW(1)$. Overall, the two newly proposed tests produce estimated powers which are competitive relative to the other tests and, hence, this limited Monte Carlo study shows that they can be used in practice to test whether observed data are realised from an exponential distribution.

4 Practical applications and conclusion

In this section all the tests considered in the simulation study will be applied to both real-world and simulated data sets. The two real-world data sets considered in this study respectively contain the failure times of air conditioning systems and the waiting times of bank customers. For these two data sets, the tests for exponentiality will be used to determine whether the observed values are realisations from an exponential distribution. On the other hand, the remaining two data sets that will be considered are simulated from a Cox proportional hazards (CPH) model and the tests for exponentiality will be used to determine the adequacy of a specific CPH model fitted to the data.

4.1 Practical application to real-world data sets

The first data set contains 30 failure times of the air conditioning system of an airplane as given by Linhart and Zucchini [20], whereas the second data set contains the waiting times

Distribution	$T_{n,0.25}$	$T_{n,1}$	$S_{n,0.25}$	$S_{n,1}$	KS_n	CM_n	$\overline{KS}_n$	$\overline{CM}_n$	EP_n
$EG(0.2)$	5	7	7	**8**	6	6	5	6	6
	6	7	7	**8**	6	6	5	6	6
$EG(0.5)$	10	15	17	**19**	11	12	9	14	15
	16	20	22	**23**	15	16	13	19	19
$EG(0.8)$	37	44	47	**51**	34	39	32	43	45
	53	61	61	**65**	49	55	48	60	60
$EL(0.2)$	14	20	22	**25**	16	18	12	19	20
	23	29	29	**33**	23	26	19	28	29
$EL(0.5)$	6	8	9	**10**	7	7	6	8	8
	8	10	11	**12**	8	8	7	9	9
$EL(0.8)$	5	**6**	**6**	**6**	5	5	5	5	6
	5	5	**6**	**6**	**6**	5	5	6	5
$\Gamma(0.7)$	13	17	19	**20**	15	18	11	18	19
	17	24	24	**27**	22	26	18	26	26
$\Gamma(1)$	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**
	5	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**
$\Gamma(1.5)$	18	15	12	5	17	21	**22**	20	21
	24	19	16	12	25	**28**	27	27	**28**
$\Gamma(2)$	42	40	33	21	41	**48**	46	47	**48**
	59	56	48	42	60	**69**	64	**69**	**69**
$LFR(2)$	30	27	18	14	24	28	**32**	29	30
	44	39	35	24	35	43	46	**47**	46
$LFR(4)$	**45**	40	30	23	34	42	**45**	44	42
	63	57	53	41	51	60	61	62	**63**
$PW(1)$	**84**	73	65	49	54	68	74	71	68
	97	89	87	76	72	86	90	90	86
$W(1.2)$	13	11	8	6	12	14	**16**	13	14
	18	14	11	7	16	18	**20**	19	19
$W(1.5)$	45	40	35	29	40	48	**55**	49	51
	66	59	55	43	58	66	65	70	**72**

Table 2: *Estimated powers for $n = 20$ (top) and $n = 30$ (bottom).*

(in minutes) of 100 bank customers before service as obtained from Ghitany, Atieh and Nadarajah [14]; both data sets can be found in the Appendix in Tables 7 and 8. In Table 3 and 4 a summary of the results of all the different tests for exponentiality can be found. The summary contains the value of each test statistic and associated p-value used to test whether the data originated from an exponential distribution. For the failure time data, all of the tests, except the KS_n test, *do no reject* the null hypothesis of exponentiality using a 5% significance level. In contrast, for the waiting time data, all of the tests *reject* the null hypothesis of exponentiality at the same significance level. This illustrates that the newly proposed test at least agrees with the more traditional tests for exponentiality.

Test	**Test statistic value**	***p*-value**
$T_{n,0.25}$	13.592	0.153
$T_{n,1}$	8.446	0.089
$S_{n,0.25}$	8.336	0.117
$S_{n,1}$	6.032	0.079
KS_n	0.213	0.020
CM_n	0.214	0.053
$\overline{KS}_n$	1.325	0.053
$\overline{CM}_n$	0.397	0.072
EP_n	1.716	0.089

Table 3: *Summary of results for failure times of air conditioning system.*

Test	**Test statistic value**	***p*-value**
$T_{n,0.25}$	57.365	0.004
$T_{n,1}$	30.407	0.005
$S_{n,0.25}$	31.810	0.010
$S_{n,1}$	19.204	0.012
KS_n	0.173	<0.001
CM_n	0.715	<0.001
$\overline{KS}_n$	2.176	<0.001
$\overline{CM}_n$	1.480	<0.001
EP_n	-3.659	0.001

Table 4: *Summary of results for waiting times of bank customers.*

4.2 Practical application to simulated data sets

The following two data sets, given in the Appendix in Tables 9 and 10, contain simulated lifetimes $(t_i, i = 1, 2, \ldots, 100)$ together with a single covariate $(x_i, i = 1, 2, \ldots, 100)$ which can take on the values 0, 1, 2 or 3. The first data set was obtained by simulating data from a CPH model with a Weibull cumulative baseline hazard function, whereas the second data set was simulated from a CPH model with a log-normal cumulative baseline hazard function.

Recall that the cumulative hazard function of the j^{th} individual follows a CPH model with a single covariate if

$$\Lambda_j(t) = e^{\beta x_j} H(t),$$

where $H(\cdot)$ is some unspecified baseline cumulative hazard function, x_j is the value of the covariate of the j^{th} individual and β is an unknown regression parameter.

On the basis of the observed data $(t_j, x_j), j = 1, 2, \ldots, 100$ we wish to test the null hypothesis

$$\mathcal{H}_0 : H(t) = H_0(t; a, b), \tag{3}$$

where $H_0(t; a, b) = \left(\frac{t}{b}\right)^a$ is the Weibull cumulative baseline hazard function with unknown parameters a and b.

We can now estimate the parameters β, a and b by their maximum likelihood estimators $\widehat{\beta}$, $\widehat{a}$ and $\widehat{b}$. Based on these estimators we can obtain the (so-called) Cox-Snell residuals, defined as

$$\widehat{\varepsilon}_j = e^{\widehat{\beta} x_j} H_0(t_j; \widehat{a}, \widehat{b}).$$

If the null hypothesis is true (*i.e.*, if the cumulative baseline hazard was correctly specified as the Weibull cumulative baseline hazard) then the Cox-Snell residuals should (approximately) follow a standard exponential distribution (see, *e.g.*, Chapter 11 of Klein and Moeschberger [19]). Hence any exponential test on the basis of $\widehat{\varepsilon}_j, j = 1, 2, \ldots, 100$ constitutes in effect a goodness-of-fit test for the CPH model itself.

It is, therefore, expected that tests for exponentiality will not reject the null hypothesis for the first simulated data set (recall that this data was generated from a CPH model with a Weibull cumulative baseline hazard), whereas the tests should reject the null hypothesis of exponentiality for the second simulated data set (which was generated from a CPH model with a log-normal cumulative baseline hazard).

The results of all the different tests for exponentiality for the two simulated data sets are summarised in Table 5 and 6, which display both the test statistics and associated p-values used to test whether the residuals originate from a standard exponential distribution, *i.e.*, whether the cumulative baseline hazard is correctly specified as Weibull. Due to the fact that the null hypothesis in (3) involves unknown parameters — and hence must be estimated under $\mathcal{H}_0$ — the p-values had to be obtained using the bootstrap algorithm described in Cockeran, Allison and Meintanis [10]. For the first simulated data set, the MLEs are $\widehat{\beta} = 0.090$, $\widehat{a} = 0.880$ and $\widehat{b} = 0.763$. Table 5 shows that all tests correctly do not reject the null hypothesis, which is expected, as the data was known to be generated using a Weibull cumulative baseline hazard.

The second simulated data set produced the following MLEs: $\widehat{\beta} = 0.008$, $\widehat{a} = 0.854$ and $\widehat{b} = 3.302$, and the resulting p-values displayed in Table 6 indicate that the null hypothesis was rejected by all of the tests. This is not surprising, since a good test for exponentiality should have the ability to detect the mis-specification of the Weibull cumulative baseline hazard when the data originated from a CPH model with a log-normal cumulative baseline hazard.

To ultimately use the two new tests for exponentiality, one would need to make a choice regarding the value of the tuning parameter a, however, from extensive simulation studies

Test	Test statistic value	p-value
$T_{n,0.25}$	2.824	0.549
$T_{n,1}$	0.713	0.692
$S_{n,0.25}$	0.328	0.745
$S_{n,1}$	0.093	0.872
KS_n	0.065	0.358
CM_n	0.040	0.679
$\overline{KS}_n$	0.692	0.447
$\overline{CM}_n$	0.053	0.674
EP_n	0.156	0.337

Table 5: *Summary of results for the first simulated data set (Weibull cumulative baseline hazard).*

Test	Test statistic value	p-value
$T_{n,0.25}$	12.476	0.025
$T_{n,1}$	5.394	0.017
$S_{n,0.25}$	8.955	0.013
$S_{n,1}$	5.183	0.013
KS_n	0.119	0.004
CM_n	0.282	0.001
$\overline{KS}_n$	1.366	0.005
$\overline{CM}_n$	0.483	0.001
EP_n	1.289	0.001

Table 6: *Summary of results for the second simulated data set (log-normal cumulative baseline hazard).*

conducted (not displayed here), it was concluded that $a = 1$ produces satisfactory results. If, however, one would prefer to rather use a data dependent choice of this parameter, one can employ the method outlined in Allison and Santana [3].

Acknowledgements

The second author's work is based on research supported by the National Research Foundation (NRF). Any opinion, finding and conclusion or recommendation expressed in this material is that of the author and the NRF does not accept any liability in this regard.

References

[1] Afify AZ, Nofal ZM & Ahmed AH, 2013, *Two characterizations of gamma distribution in terms of sth conditional moments*, Journal of Advances in Mathematics, **5**, pp. 688–695.

[2] Ahmad I & Alwasel I, 1999, *A goodness-of-fit test for exponentiality based on the memoryless property*, Journal of the Royal Statistical Society Series B, **61(3)**, pp. 681–689.

[3] Allison JS & Santana L, 2014, *On a data-dependent choice of the tuning parameter appearing in certain goodness-of-fit tests*, Journal of Statistical Computation and Simulation, **85(16)**, pp. 3276–3288.

[4] Allison JS, Santana L, Smit N & Visagie IJH, 2017, *An 'apples to apples' comparison of various tests for exponentiality*, Computational Statistics, **32(4)**, pp. 1241–1283.

[5] Alwasel I, 2001, *On goodness of fit testing for exponentiality using the memoryless property*, Journal of Nonparametric Statistics, **13**, pp. 569–581.

[6] Angus JE, 1982, *Goodness-of-fit tests for exponentiality based on a loss-of-memory type functional equation*, Journal of Statistical Planning and Inference, **6**, pp. 241–251.

[7] Ascher S, 1990, *A survey of tests for exponentiality*, Communications in Statistics – Theory and Methods, **19(5)**, pp. 1881–1825.

[8] Baringhaus L & Henze N, 2000, *Tests of fit for exponentiality based on a characterization via the mean residual life function*, Statistical Papers, **41**, pp. 225–236.

[9] Baringhaus L & Henze N, 2008, *A new weighted integral goodness-of-fit statistic for exponentiality*, Statistics and Probability Letters, **78**, pp. 1006–1016.

[10] Cockeran M, Meintanis SG, & Allison JS, 2019, *Goodness-of-fit tests in the Cox proportional hazards model*, Communications in Statistics – Simulation and Computation, pp. 1–12.

[11] D'Agostino RB, & Stephens MA, 1986, *Goodness-of-fit Techniques*. Marcel Dekker, New York.

[12] Davis DJ, 1952, *An analysis of some failure data*, Journal of the American Statistical Association, **47(258)**, pp. 113–150.

[13] Epps TW & Pulley LB, 1986, *A test of exponentiality vs. monotone-hazard alternatives derived from the empirical characteristic function*, Journal of the Royal Statistical Society Series B, **48(2)**, pp. 206–213.

[14] Ghitany ME, Atieh B & Nadarajah S, 2008, *Lindley distribution and its application*, Mathematics and computers in simulation, **78(4)**, pp. 493–506.

[15] Han G, Schell MJ, Zhang H, Zelterman D, Pusztai L, Adelson K & Hatzis C, 2016, *Testing violations of the exponential assumption in cancer clinical trials with survival endpoints*, Biometrics, **73(2)**, pp. 687–695.

[16] Henze N, & Meintanis SG, 2005, *Recent and classical tests for exponentiality: a partial review with comparisons*, Metrika, **36**, pp. 29–45.

[17] Jammalamadaka SR, & Taufer E, 2006, *Use of mean residual life in testing departures from exponentiality*, Journal of Nonparametric Statistics, **18(3)**, pp. 277–292.

[18] Jovanović M, Milošević B, Nikitin YaYu, Obradović M & Volkova KYu, 2015, *Tests of exponentiality based on Arnold-Villasenor characterization and their efficiencies*, Computational Statistics and Data Analysis, **90**, pp. 100–113.

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

[20] Linhart H & Zucchini W, 1986, *Model selection,* John Wiley & Sons.

[21] Meintanis SG, Swanepoel JWH & Allison JS, 2014, *The probability weighted characteristic function and goodness-of-fit testing*, Journal of Statistical Planning and Inference, **146**, pp. 122–132.

[22] R Core Team, 2018, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

[23] Shanker R, Fesshaye H & Selvaraj S, 2015, *On modeling of lifetimes data using exponential and Lindley distributions*, Biometrics & Biostatistics International Journal, **2(5)**, pp. 1/9 – 9/9.

[24] Shanker R, Fesshaye H, & Selvaraj S, 2016, *On modeling of lifetime data using one parameter Akash, Lindley and exponential distributions*, Biometrics & Biostatistics International Journal, **3(2)**.

[25] Taufer E, 2000, *A new test for exponentiality against omnibus alternatives*, Stochastic Modelling and Applications, **3**, pp. 23–36.

[26] Vinod B, & Altiok T, 1986, *Approximating unreliable queueing networks under the assumption of exponentiality*, Journal of the Operational Research Society, **37(3)**, pp. 309–316.

[27] Volkova KYu, 2010, *On asymptotic efficiency of exponentiality tests based on Rossberg's characterization*, Journal of Mathematical Sciences, **167(4)**, pp. 486–494.

[28] Wand MP & Jones MC, 1994, *Kernel smoothing.* Chapman and Hall/CRC.

Appendices

A. Real-world and simulated data sets

This appendix contains the data sets that were used for the practical applications in Section 4.

23	261	87	7	120	14	62	47	225	71
246	21	42	20	5	12	120	11	3	14
71	11	14	11	16	90	1	16	52	95

Table 7: *Failure times of air conditioning system of an airplane.*

0.8	0.8	1.3	1.5	1.8	1.9	1.9	2.1	2.6	2.7
2.9	3.1	3.2	3.3	3.5	3.6	4	4.1	4.2	4.2
4.3	4.3	4.4	4.4	4.6	4.7	4.7	4.8	4.9	4.9
5	5.3	5.5	5.7	5.7	6.1	6.2	6.2	6.2	6.3
6.7	6.9	7.1	7.1	7.1	7.1	7.4	7.6	7.7	8
8.2	8.6	8.6	8.6	8.8	8.8	8.9	8.9	9.5	9.6
9.7	9.8	10.7	10.9	11	11	11.1	11.2	11.2	11.5
11.9	12.4	12.5	12.9	13	13.1	13.3	13.6	13.7	13.9
14.1	15.4	15.4	17.3	17.3	18.1	18.2	18.4	18.9	19
19.9	20.6	21.3	21.4	21.9	23	27	31.6	33.1	38.5

Table 8: *Waiting times of bank customers (in minutes) before service.*

t	0.277	0.171	0.234	0.531	0.319	1.633	0.161	0.373	0.209	0.606
x	1	0	0	1	0	0	3	1	1	2
t	0.470	2.346	0.490	0.565	2.259	0.137	0.502	0.066	0.212	0.448
x	2	0	1	2	1	0	0	3	1	2
t	0.540	1.438	1.650	0.024	0.377	2.456	0.682	0.313	0.697	0.689
x	3	2	2	2	0	0	1	1	3	1
t	0.312	0.188	0.264	0.008	1.400	0.872	1.062	0.006	0.380	0.759
x	1	3	1	3	0	1	2	2	2	2
t	0.920	0.328	0.302	1.210	0.107	1.740	0.792	0.627	0.055	0.567
x	1	1	2	3	0	1	3	3	2	1
t	0.132	0.089	0.068	0.516	2.628	1.325	1.127	0.473	0.051	0.509
x	0	0	0	0	2	1	2	3	1	0
t	0.789	0.029	0.216	2.506	0.021	0.112	0.127	0.167	1.228	0.272
x	1	3	0	0	3	1	2	3	0	2
t	0.144	0.176	0.014	0.269	0.651	0.415	1.525	1.019	0.130	1.152
x	2	1	2	2	3	3	1	3	0	3
t	4.164	0.067	1.297	1.209	0.020	1.072	0.128	1.426	2.085	0.309
x	0	3	2	3	2	3	3	2	2	3
t	0.415	0.121	0.018	1.385	1.880	0.085	0.377	0.009	3.357	0.109
x	0	3	0	1	1	3	0	2	3	0

Table 9: *Simulated data set from a CPH model with a Weibull cumulative baseline hazard.*

160

t	0.454	0.925	0.215	2.831	20.276	0.418	2.076	0.546	1.473	6.037
x	1	3	3	1	0	0	2	2	3	3
t	0.795	0.232	1.423	3.064	5.358	0.630	2.762	2.225	0.339	2.964
x	3	3	1	3	2	0	3	0	1	2
t	9.573	0.289	12.025	1.040	8.959	2.353	8.887	2.544	1.580	0.634
x	3	2	3	3	0	1	0	2	0	3
t	0.434	0.426	0.766	12.508	1.220	1.250	0.284	0.656	1.778	0.736
x	1	1	2	2	2	1	3	1	1	2
t	1.588	13.494	3.873	3.931	0.843	2.385	2.243	1.087	1.583	3.441
x	0	0	3	0	3	0	0	3	1	2
t	1.677	2.294	1.056	1.072	5.101	1.631	1.449	9.263	3.322	0.820
x	0	0	1	2	1	3	1	1	2	0
t	2.267	7.378	10.503	1.043	0.862	0.670	2.078	5.113	2.014	5.540
x	0	1	3	0	0	2	0	1	1	0
t	0.453	2.290	1.065	1.101	0.294	12.021	0.569	0.211	0.277	0.479
x	2	2	3	3	0	1	2	1	3	0
t	2.094	11.347	3.797	27.351	11.561	3.542	0.753	0.479	0.156	5.678
x	1	3	2	2	1	2	2	0	1	0
t	0.375	19.239	1.701	1.210	7.755	4.850	1.830	1.022	11.083	0.563
x	0	2	3	1	1	3	3	2	2	2

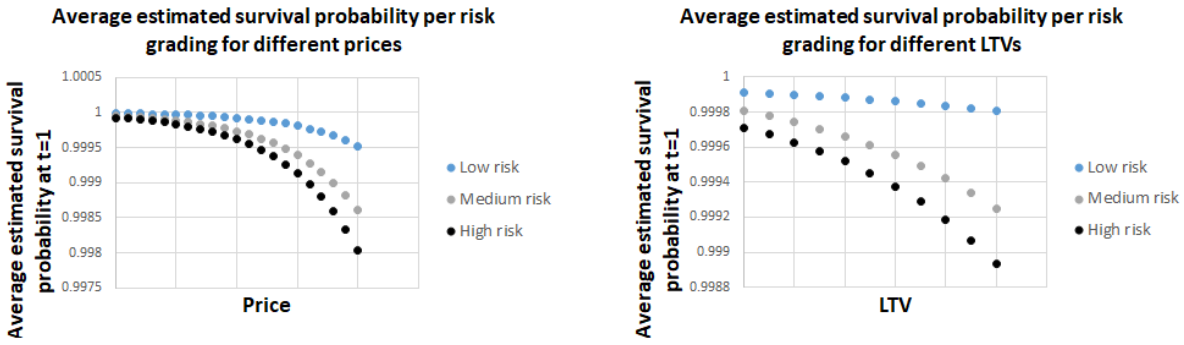
Table 10: *Simulated data set using a CPH model with a log-normal cumulative baseline hazard.*

Appendix B: R-code for parametric mixture cure model estimation

```
#Likelihood function: Mixture cure model with exponential baseline


zi = z_cure       #Cure model covariates
x = z_cure_s      #survival model covariates
t = time_fin      #Event time
d = cens_final    #Censoring indicator
bd = c(b_cure,b_cure_s,lam = lambda)  #Starting value for parameters to be estimated


#Likelihood function
lik = function(ba,z,x,t,d){
  b = ba[1:ncol(z)]
  bet = ba[(ncol(z)+1):(NROW(ba)-1)]
  la = ba[NROW(ba)]
  bz = drop(1/(1+exp(-(crossprod(b,t(z))))))
  bx = drop(exp(crossprod(bet,t(x))))
  l = -(sum(d*log(bz) + d*log(la) +d*log(bx)-d*bx*la*t +(1-d)*(log((1-bz) +(bz)*(exp(-bx*la*t))))))

}

#Call likelihood function
opt = nlm(lik,bd,zi,x,t,d)
```

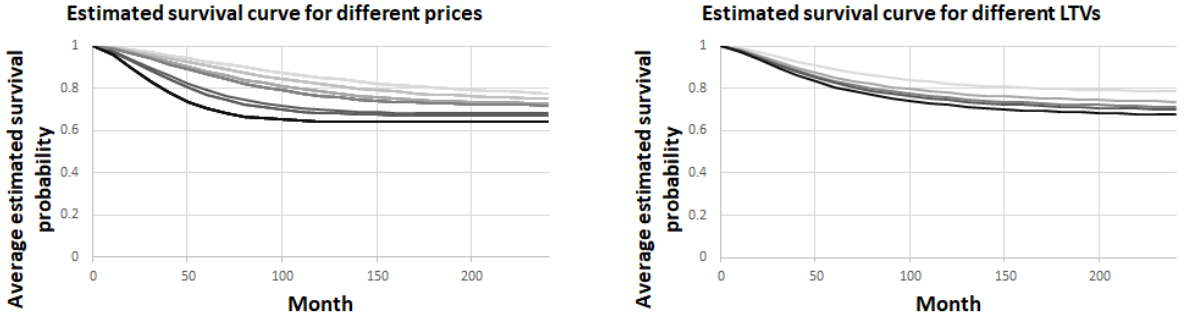
```
#Likelihood function: Mixture cure model with Weibull baseline


zi = z_cure       #Cure model covariates

x = z_cure_s      #survival model covariates

t = time_fin      #Event time

d = cens_final    #Censoring indicator

bd = c(b_cure,b_cure_s,lam = lambda,nu = nu)  #Starting value for parameters to be estimated


#Likelihood function

lik = function(ba,z,x,t,d){

  b = ba[1:ncol(z)]

  bet = ba[(ncol(z)+1):(NROW(ba)-2)]

  la = ba[NROW(ba)-1]

  nu = ba[NROW(ba)]

  bz = drop(1/(1+exp(-(crossprod(b,t(z))))))

  bx = drop(exp(crossprod(bet,t(x))))

  l = -((sum((d*log(bz) + d*log(la)+d*log(nu)+d*(nu-1)*log(t)+d*log(bx)-d*bx*la*t^nu) +(1-d)*(log((1-
bz) +(bz)*(exp(-bx*la*t^nu)))))))


}


# Call likelihood function

opt = nlm(lik,bd,zi,x,t,d)
```

Appendix C: Figures and tables for Weibull baseline data set.



(a) (b)

Figure C.1: (a) Relationship between price and estimated survival probability in month one (Weibull baseline); (b) Relationship between LTV and estimated survival probability in month one (Weibull baseline).



(a) (b)

Figure C.2: (a) Estimated survival curve for different prices (Weibull baseline); (b) Estimated survival curve for different LTVs (Weibull baseline).

Table C.1: Computational results (unconstrained): Weibull baseline distribution.

Risk grading	Take-up proportion	Average price	Average LTV
Low	0.34	1.00	0.97
Medium	0.32	1.00	0.90
High	0.33	1.00	0.91

Table C.2: Computational results (risk constrained): Weibull baseline distribution.

Risk grading	Take-up proportion	Average price	Average LTV
Low	0.50	0.80	0.97
Medium	0.30	0.95	0.89
High	0.20	1.00	0.93

Table C.3: Computational results (LTV constrained): Weibull baseline distribution.

Risk grading	Take-up proportion	Average price	Average LTV
Low	0.38	1.00	0.97
Medium	0.35	1.00	0.90
High	0.27	0.96	0.75