Loss Distributions in Consumer Credit Risk:

Macroeconomic Models for Expected and Unexpected Loss

Thesis Presented for the Degree of

MASTER OF COMMERCE IN MATHEMATICAL STATISTICS

In the Department of Finance

UNIVERSITY OF CAPE TOWN

By

Musa Malwandla

Student Number: MLWMUS002

May 2016

Supervisors:

Kanshukan Rajaratnam

Allan Clark

Declaration:

This work has not been previously submitted in whole, or in part, for the award of any degree. It is my own work. Each significant contribution to, and quotation in, this dissertation from the work, or works, of other people has been attributed, and has been cited and referenced.

Signature: Musa Malwandla

Date: 05 May 2016

Abstract

This thesis focuses on modelling the distributions of loss in consumer credit arrangements, both at an individual level and at a portfolio level, and how these might be influenced by loan-specific factors and economic factors. The thesis primarily aims to examine how these factors can be incorporated into a credit risk model through logistic regression models and threshold regression models.

Considering the fact that the specification of a credit risk model is influenced by its purpose, the thesis considers the IFRS 7 and IFRS 9 accounting requirements for impairment disclosure as well as Basel II regulatory prescriptions for capital requirements. The thesis presents a critique of the unexpected loss calculation under Basel II by considering the different ways in which loans can correlate within a portfolio.

Two distributions of portfolio losses are derived. The Vašíček distribution, which is the assumed in Basel II requirements, was originally derived for corporate loans and was never adapted for application in consumer credit. This makes it difficult to interpret and validate the correlation parameters prescribed under Basel II. The thesis re-derives the Vašíček distribution under a threshold regression model that is specific to consumer credit risk, thus providing a way to estimate the model parameters from observed experience. The thesis also discusses how, if the probability of default is modelled through logistic regression, the portfolio loss distribution can be modelled as a log-log-normal distribution.

The first chapter of the thesis introduces the topic of credit risk modelling by discussing the different types of credit risk models that are used in retail banks. This chapter also describes the accounting standards and the Basel requirements as they pertain to consumer credit risk.

Chapter 2 of the thesis offers a critique of the statistical and quantitative methods that are used in consumer credit modelling for the different purposes: acquisition scoring, accounting provisioning, capital provisioning

and stress testing. The critique highlights some of the shortcomings of the current approaches to loss aggregation and methods for incorporating economic information into loss estimation.

Chapter 3 of the thesis focuses on the theory of developing a credit risk model, leading to a derivation of the Vašíček distribution and the log-log-normal distribution for portfolio losses. The chapter also provides an overview of relevant statistical and quantitative methods that are used in credit risk modelling.

Chapter 4 of the thesis uses the discussions from Chapter 3 to develop a number of credit risk models for a South African bank's fixed-rate loan portfolio.

Chapter 5 of discusses the use of the models for determining capital requirements. It compares the capital requirements under the models to those required under Basel II. The thesis offers a critique of the Basel II requirements, and enumerates some of its major deficiencies.

The thesis, therefore, provides two comprehensive modelling approaches that allow both loan-specific and economic factors to be included in a regression model for default rates. These allow the default rate model to be applied to both expected loss provisioning and stress testing.

Tables of Contents

Chapters and Sections

Contents

Tables

Figures

Acknowledgements

I thank my grandmother and my parents for their support throughout my education. I also thank Kanshukan and Allan for their patient supervision.

Chapter 1: Introduction

1.1. Credit Risk Analysis

Consumer credit arrangements account for a significant portion of the services offered by retail banks today. A typical consumer credit agreement involves the consumer borrowing money from the bank. The consumer is then required to repay the principle sum of money, with interest, over a set period through regular level instalments. The main risk is the potential of the consumer being unable to make payments as they fall due. Therefore, at any point, the bank is at risk of losing part or the entire outstanding amount of the loan. The extent of the risk is influenced by factors such as whether the loan is secured or unsecured, the economic conditions and the customer's financial situation (Thomas, 2000).

The second Basel accord has increased the importance of understanding the risk associated with such arrangements and provisioning for this risk through capital requirements (Basel II, 2004). Accounting practice also requires banks to be in full view of their credit risk for the purpose of financial reporting. Understanding credit risk allows the bank to put measures in place to manage loan applications, impaired debt as well as to determine the appropriate interest rate and terms to charge applicants.

In recent years it has become convention for banks offering consumer loans to have statistical models in place to assess and manage the different elements of credit risk. Models are used at loan application stage to assist in deciding whether a particular application should be approved or declined. Banks also make use of models to determine the appropriate interest rate to charge for different loan arrangements, according to the level of risk associated with the loan. The interest rate charged on a loan is the effective price of the loan agreement – it thus needs to be high enough to meet the

banks profit criteria, while being affordable and competitive from the consumer's perspective.

Provisioning models are a different type of credit risk models. These are used to determine the value of reserves that should be set aside with respect to existing loan contracts. These models vary in specification and sophistication, according to their purpose. Impairment models are used to determine the expected loss on existing loan arrangements on a realistic basis. The provisions held for expected loss feature on the financial statements of the bank, and thus need to be objective and simplistic in a way that allows shareholders to compare the bank's results to its peers.

Capital models are used to determine provisions to hold above the realistic provisions for variations in credit experience. These additional provisions capture the fact that there will be statistical variation around the expected loss amount and aim to ensure that the bank has adequate provisions for loss scenarios that are more extreme. Stress testing models go further by attempting to capture the fact that credit risk is affected in a major way by economic conditions. Stress testing models formulate scenarios for economic conditions and assess the impact of these scenarios on reserves.

The following sections provide an overview of consumer credit modelling by discussing the different models used in consumer credit in more detail.

1.1.1. Credit Scorecards and Pricing

Consumer credit scoring models (scorecards) are used to assess the credit risk associated with a customer. The probability scores produced by these models are either called application scores or behavioural scores, depending on the purpose of the model. A common type of credit scoring model is an application scoring model, which estimates the probability that a loan applicant will default on the account being applied for, and that this default will ultimately lead to a loss. An inverse scaling is often applied to the

predicted probability, so that the higher the credit score, the lower the risk of loss on the account.

A similar type of credit scoring model is a behavioural scoring model. This produces a general estimate of the default risk associated with a customer or an account. At a customer level, this estimates the probability that at least one of the customer's accounts will move into default over a given event horizon (e.g. over the next year). At an account level, the score produced is a scaled estimate of the probability of default on the specific account. The distinction between an application score and a behavioural score is that application scores are used in the origination phase of an account, while behavioural scores are used throughout the life of a credit account and customer.

One of the most common statistical techniques used in application and behavioural scoring is logistic regression (Hand, 2005). However, other techniques such as stochastic processes (e.g. Crook and Bellotti, 2010) and artificial neural networks (e.g. Baesens, van Gestel, Stepanova, van den Poel and Vanthienen, 2005) are becoming more common.

Application scores can be used in simplistic pricing models to produce an interest rate that will be charged on a loan. The higher the risk associated with a customer, the higher the rate charged on the loan. Although this form of risk-based pricing is common in the insurance industry, Thomas (2000) notes how slowly it is being developed in consumer credit risk. The interest rate charged is sometimes linked to a baseline interest rate in what are called variable-rate loan agreements (Thomas, Oliver and Hand, 2005). This is typical in long term loans, such as residential mortgage loans and vehicle finance arrangements. The baseline interest rate is set by some central financial institution within the country, and changes according to the government's monetary policy (*In South Africa, the prime overdraft rate is*

used by banks as a basis for loan pricing and is set by the South African Reserve Bank).[1]

The fact that future changes in this baseline interest rate are uncertain at loan origination brings an additional level of risk. For example, if the baseline rate increases dramatically, the loan repayment amount on variable-rate loans would also increase, making it harder for the consumer to repay the loan. This is an example of the influence of economic conditions on credit risk. Mileris (2012) discusses interest rates as being but one of the many macroeconomic influences on credit risk.

1.1.2. Impairment Modelling

Once the loan is issued, focus shifts to estimating the deterioration of credit quality, determining the amount of provisions to be set aside to cover the expected losses arising from this deterioration and determining portfolio-level capital requirements. The level of deterioration in credit quality of a book of contracts is often measured as the level of impairment provisions, or expected losses.

For the purpose of accounting, under the sevenths International Financial Reporting Standard (IFRS 7, 2005), the bank is required to disclose expected losses on loans that are deemed to be impaired as at the reporting date, i.e., credit loss provisions are required for *impairment events* that have already occurred, not necessarily for future events. An impairment event is taken to mean any event that occurs in the life of a credit account that compromise the ability of the accountholder to make payments.

For the purposes of IFRS 7 disclosure, the bank will aim to estimate the following with respect to accounts that exist on its books.

[1] See: "The Role of the Prime Rate and the Prime-Repurchase Rate Spread in the South African Banking System"
https://www.resbank.co.za/Lists/News%20and%20Publications/Attachments/4279/Summary%20of%20the%20main%20conclusion.pdf

1. The bank will require an estimate of the proportion of the book that is impaired, i.e., that has experienced an impairment event. Impairment events include occurrences such as loss of employment, death, insolvency, sickness and bankruptcy. The bank is typically unable to determine when these events have taken place in the life of an accountholder. The total size of the impaired population thus needs to be estimated. The estimated proportion can be divided into *identified impairment (II)* and *unidentified impairment (UI)*. The II portion consists of accounts that have an objective evidence of impairment, i.e., where a credit event is known to have occurred. The fact that an account is in arrears is often taken to be objective evidence of impairment. The UI portion will be the estimated probability that an account that currently has no objective evidence of impairment is actually impaired e.g. the accountholder has become unemployed but, through savings, has kept the loan up-to-date.
2. The bank will require an estimate of the loss from the accounts that are impaired. In statistical terms, this can be understood as the expected loss given impairment. The impairment is then calculated as the total exposure (the total amount at risk of loss) on impaired accounts (II and UI) multiplied by the expected loss ratio.

Therefore, a possible model for calculating the impairment of a credit book is as follows.

$$Impairment = Exposure \times [UI + II] \times PD \times LGD$$

where:

- Exposure is the outstanding balance on the book of contracts,
- PD is the estimated probability of default on the impaired accounts,
- II and UI are the proportion of exposure where an impairment event is either identified or expected, respectively, and
- LGD is the expected loss ratio (of the exposure) on an impaired account, if default were to occur.

This impairment model is essentially a 3-state models which classifies accounts into “Performing”, “Impaired” and “Default”. Thus, the elements of the model can be estimated via maximum likelihood methods. An example variation of this basic impairment model is described by Kelly (2011).

The ninth International Financial Reporting Standard (IFRS 9, 2014) aims to achieve the same broad aim of IFRS 7, which is to disclose expected losses on a credit portfolio, in what is considered a more comprehensive way. IFRS 9 makes use of what is sometimes referred to as a *three bucket* approach.

In the three bucket approach, bucket one (low risk) contains accounts that have the same level of estimated as at loan origination. For these accounts, a 12-month expected loss calculation is performed. Accounts that have experienced a significant deterioration in credit risk will transition from bucket one to bucket two (medium risk), where a lifetime expected loss calculation will be performed. The calculation of expected loss at this stage can be calculated for a group of loans as a collective. After further deterioration, or in the case where losses have been realised, accounts move to bucket three (high risk), where the lifetime expected losses are determined on each individual account.

The utility of the IFRS 9 approach over the IFRS 7 approach, at least philosophically, can be seen when considering the concept of *temporary* or *express* loans in the South African market. These are priority loans on the accountholder’s next pay-check, to be immediately repaid when the accountholder’s salary is deposited into his/her bank account. Arguably, applicants of such loans would have typically already experienced an impairment event, such as the loss of employment, since the main contingency from the bank’s perspective is whether the accountholder will receive his/her regular salary over the next month.

Under IFRS 7 philosophy, all such express loans would be considered already impaired. However, express loans will typically be priced much higher than regular loans, to account for the additional risk associated with the typical applicant. Therefore, only if the associated risk is considerably

worse than initially expected (and priced for) should a loan be considered impaired. IFRS 9 overcomes this difficulty by allowing transitions between low risk, medium risk and high risk statuses when the credit risk has deteriorated significantly to what it was understood to be when pricing the loans.

Additionally, IFRS 9 requires that economic conditions and forecasts be allowed for in determining the loss provisions. Therefore, a credit risk model that allows for the influence of economic conditions is required. This is particularly important for the determination of lifetime losses on mortgages, where the economic forecasts will generally span a longer period of time. IFRS 9 is due for implementation in 2018.

1.1.3. Capital Modelling and Stress Testing

The purpose of impairment modelling is to estimate the expected loss on the current credit book. The loss is a random variable with an associated loss distribution. Thus, there is a non-zero probability that the observed loss will exceed the expected loss. If the loss distribution is symmetric (so that the mean equals the median) then the probability that the loss exceeds the expected loss is 0.5, which mean that there would be a 50% chance that the provisions are inadequate.

Capital modelling aims to estimate the required capital to *reasonably* cover the losses in excess of expected losses, i.e., the capital required to cover what is often called *unexpected loss*. The capital requirement is typically calculated as a percentile point of the portfolio loss distribution, i.e., the unexpected loss provisions ensure that there is a α% chance that capital held will be sufficient to cover actual losses. In this way, the capital determination is equivalent to the value-at-risk (VaR) measure from

financial risk management (Philippe, 2006). The unexpected loss provision can thus be determined as the excess of VaR over expected loss[2].

Methods for calculating unexpected loss are well-prescribed by the second Basel accord (Basel II, 2004). Basel II prescribes two approaches that retail banks can follow in calculating capital requirements: the foundations approach and the advanced approach (Thomas et al., 2005). Under the foundations approach, the bank calculates its own probability of default (PD) parameter, while the other parameters (Loss Given Default – LGD – and Exposure at Default – EAD) are provided by the bank's regulator. Under the advanced approach all the parameters are set by the bank. The PD estimates must be adjusted for seasonal effects, while the LGD must be based on downturn economic conditions. Basel II also prescribes a 90-day default definition, where accounts more than 90 days past-due are regarded as defaults. This may differ from the default definition used for the purpose of impairment provisioning.

The modelling approaches used in capital provisioning and impairment provisioning are generally very similar, barring a few technical differences between the requirements (e.g. default definitions and seasonal adjustment to PDs). However, the focuses of the two exercises have a fundamental difference. Since an impairment model is only concerned with expected loss, it is unimportant to consider the possibility of correlation between losses[3]. Meanwhile, when considering the entire portfolio loss distribution, different levels and forms of correlation can lead to different values for VaR (e.g. Dhaene, Denuit, Goovaerts, Kaas and Vyncke, 2002).

The unexpected loss equation in Basel II assumes that portfolio losses are distributed according to the Vašíček (1987) distribution. The specific correlation structure that underlies this assumption is discussed at length in Chapters 2 and 3.

[2] Although, theoretically, expected loss should be the same as the impairment provision, differences may arise due to differences in the requirements of an impairment model (as dictated by accounting standards) as compared to those of a capital model (as prescribed by the Basel Committee on Banking Supervision).

[3] If $A = \sum_k X_k$ is the aggregation of individual loan losses (X_k), then $E[A] = \sum_k E[X_k]$, so that the dependence structure between the different X_ks is irrelevant.

For both impairment and capital modelling, the assumed or implied loss distribution is typically a best estimate from past experience. For this reason, there is a chance that economic environments may have changed in a way that alters the properties or nature of the loss distribution. Stress testing thus aims to determine the unexpected loss based on adverse assumptions about the loss distribution. The process of stress testing involves fitting a macroeconomic model to a particular credit risk metric. A stress scenario is then formulated and expressed in terms of the fitted macroeconomic variables (Foglia, 2009). This allows the impact of the stress scenario on the loss distribution to be tested. This intern produces *stressed* provisions for expected and unexpected losses.

The stress testing model can be based on any of a wide range of available credit risk metrics, including the nonperforming loan ratio (e.g. Blaschke, Jones, Majnoni and Peria, 2001) and the default rate (e.g. Simon and Rolwes, 2009). Figure 1 illustrates the difference between expected loss and unexpected loss, as well as the stressed loss distribution.

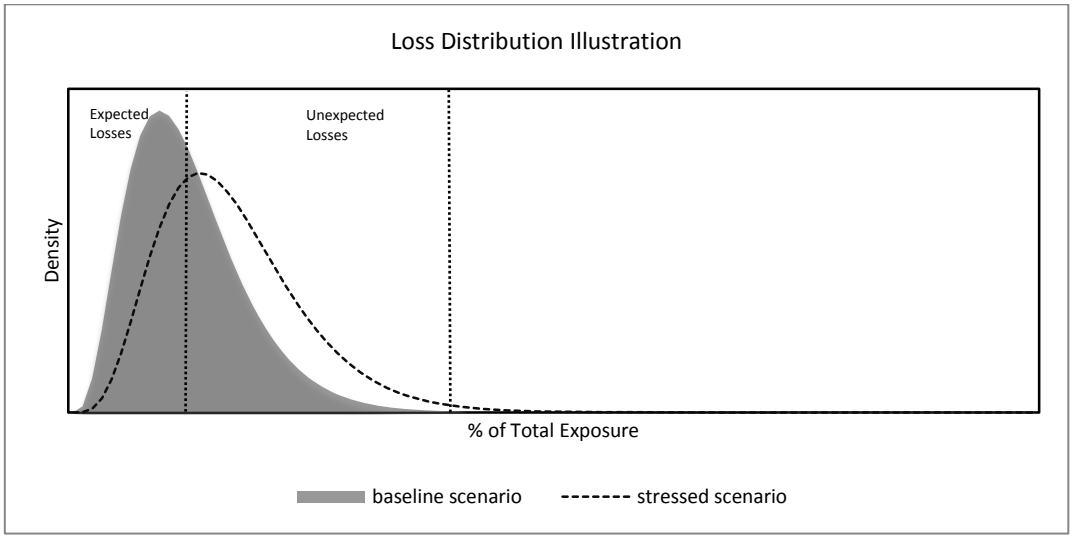


Figure 1: Illustration of Expected Loss, Unexpected Loss

1.1.4. Forecasting and Model Deterioration

As with any statistical model, the credit risk models discussed above are developed based on historical data. As such, once the trends, patterns or behaviour observed in the model development period cease to hold, the model will require redevelopment. Thomas (2000) estimates that a credit risk model in use in 1999 was likely to have been developed in 1998. Such lags in development and implementation cause a problem in fast changing risk environment. If economic conditions are considered the main cause of changes in credit behaviour then the average credit risk model would perform fairly poorly in times of rapid economic changes e.g. the onset of an economic recession.

Crook, Hamilton and Thomas (1992) illustrate the sensitivity of credit risk models to the model development period. Two applications scorecards were developed based on data from two consecutive years (1989 and 1990). Both scorecards were used to score both sets of customers. The findings were that about 25% of the customers who would be rejected under one scorecard would be accepted under the other.

Thus, if economic conditions are the major influence on changes in credit risk, it makes sense to build models that take into account changes in economic conditions. We would expect such model to require less frequent recalibrations or redevelopments. Indeed, even if economic conditions are not the primary influence of credit risk changes, the mere fact that modelling and forecasting techniques of economic conditions are well-studied (see Roos, 1955; Clemen and Winkler, 1986) forms a motivation of building such models. Coupling credit risk modelling and economic forecasting would allow the bank to forecast credit risk metrics such as default rates and loan recovery rates. This in turn allows for more robust stress testing. These forms would also assist in the implementation of IFRS 9 impairment provisioning requirements and assist in the seasonal adjustment of PDs under Basel II capital provisioning.

1.2. Research Question

Loss distributions form the central role in credit risk modelling. In application and behavioural scorecards the aim is to estimate the probability of zero loss, in impairment modelling it is to determine the first moment of the loss distribution while in capital modelling the aim is to determine a particular percentile of the loss distribution. Stress testing and risk forecasting aim to test the influence of economic conditions on this loss distribution.

The research aims to determine the loss distribution on consumer credit risk models, allowing both for factors that affect individual loans and those that affect the entire population of borrowers. Categorically, the research aims are as follows.

1. The primary aim of the research is to estimate the loss distribution on credit contracts, conditioned on account-level and customer-level information.
2. The secondary aim is to use the loss distributions to determine the expected and unexpected loss provisions for a book of contracts.
3. The tertiary aim is to determine how the influence of economic conditions can be incorporated into a loss model.

Specifically, the following contributions are made by the research.

1. A method for deriving a conditional loss distribution is discussed, where the parameters of the distribution are estimated as functions of account-level and economic covariates via regression analysis. The overall loss distribution is specified analytically, via simulation and through approximations.
2. The dependence of credit risk (particularly default rates) on economic conditions is discussed and modelled via two techniques, the first of which is logistic regression with an adjustment for economic conditions. The second technique is threshold regression, which

borrows the idea of credit deterioration being modelled as a stochastic process from the Merton (1974) model.

3. Large homogenous portfolio (LHP) approximations for the aggregate loss distribution are derived for the logistic regression and the threshold regression model. The analysis finds that the threshold regression model leads to the Vašíček model for portfolio losses, while the logistic regression model leads to a log-log-normal distribution. This is accompanied by an assessment of parameters prescribed by Basel II for capital determination.

1.3. Structure of the Paper

The remainder of the paper is structured as follows. In Chapter 2, a review and critique of current literature in credit risk modelling is provided. This aims to provide justification for the thesis, and position the contributions of the thesis amongst contemporary literature.

Chapter 3 discusses the principles, methods and models that are used this paper. The chapter also contains derivations for the portfolio loss distribution under different scenarios. Chapter 4 describes the process followed in developing and implementing the models described in Chapter 3.

Chapter 5 collates the different credit models discussed in Chapter 4 into loss distributions and addresses the ways in which these can be aggregated into portfolio losses. Chapter 6 concludes the paper by summarising the findings of the paper and the extent to which the aims of the research have been addressed. Chapter 7 contains references.

1.4. Contributions to the Field

The thesis attempts to make a number of contributions to literature on consumer credit risk modelling. Firstly, the thesis offers two approaches for modelling the influence of economic conditions on account-level default probabilities: through survival analysis and through logistic regression. The logistic regression approach can be applied in a number of areas in credit risk management, including impairment modelling, application scoring, pricing and capital provisioning. The approaches offered overcome the scorecard reactiveness issue identified by Crook, Hamilton and Thomas (1992).

The thesis explores the use of threshold regression in credit risk modelling, which is one of the first applications of this approach in credit risk modelling. The model introduced is shown to lead to convenient expressions for the portfolio loss distribution and has close relationship to the Merton model and the Vašíček distribution.

By leveraging off the original work by Vašíček (1987), the thesis derives a general approach for modelling the portfolio loss distribution. The derivations adds to the otherwise limited number of portfolio loss distribution functions available in consumer credit modelling literature. Particularly, the thesis offers an alternative to the Vašíček distribution, but also offers alternative ways of estimating and interpreting the parameters of the Vašíček distribution.

Finally, the thesis provides a comprehensive critique of the Basel capital requirements, adding to some of the critique already covered authors such as Thomas et al. (2005).

Chapter 2: Literature Review

This chapter provides a literature review of contemporary credit risk modelling and identifies some of the gaps in literature that are addressed by this thesis. It also elaborates on the research questions given in Chapter 1.

2.1. Contemporary Credit Risk Modelling

2.1.1. Loan-Specific Models

Quantitative analysis in consumer credit risk is a fairly new adoption. Historically, credit decisions like which loan applications to accept or reject were based on the subjective judgement of bankers. Although the decisions were not quantitative, it was well-understood that factors such as the value of the loan, the amount of collateral as well as the prevailing economic situation are important determinants of the amount of risk associated with the loan (Thomas, 2000). The advent of computers, coupled with developments in operations research and mathematical statistics, has led to a formalisation of these analyses into a quantitative framework.

The aim of most credit risk analysis can be reduced to an assessment of the likelihood that a given loan will end up in default, although this is sometimes accompanied by an estimation of potential loss given default. Different quantitative methods exist for estimating this likelihood, with one of the most popular one being logistic regression. This produces a simple estimate of the probability that an account will default over some predefined period. This could be, for example, a 1-year horizon or the entire account lifetime. The covariates associated with the loan at the observation point are used as inputs into the model.

Logistic regression has few assumptions and produces results and relationships that are easy to interpret. If p_i is the probability of default on account i and x_i is the vector of covariates associated with the account then the main assumption of logistic regression is that a linear relationship exists between $f(p_i)$ and x_i, where f(x) is an appropriate link function. The link function is a transformation of the target variable that aims to ensure linearity with the covariates (see Hosmer, Hosmer, Le Cessie and Lemeshow, 1997). The most common link functions are the logit function, the probit function and the complementary log-log function.

Logistic regression has been applied to different elements of credit risk, ranging from risk scoring (Whittaker, Whitehead and Somers, 2007) to loss calculations (Ingolfsson and Elvarsson, 2010).

The fact that logistic regression leads to simplistic models means that it is relatively incapable of capturing more complex patterns (see Kumar, Roa and Soni, 1995). Its two main weaknesses are the limited choice of model structures (i.e., link-functions) and the limited amount of insight it produces. There are only three well-studied link-functions in logistic regression which means that only three model forms are generally available for the typical logistic regression model.

Other techniques have since emerged that overcome the inflexibility of logistic regression. A common alternative to logistic regression is decision-tree analysis (e.g. Matuszyk, Mues and Thomas, 2010; Chan and Loh, 2004). Although this is a more complex form of analysis, it overcomes the inflexibility of logistic regression while remaining simple to interpret. A decision-tree model starts off with an entire population of accounts exposed to the risk of default. At each branch of the tree, a set of decisions dictating how the population can be segmented into “good risk” and “bad risk” (low default probability and high default probability, respectively) are made. Each decision uses the available covariates to segment the population (or sub-population) into two sub-populations. A hypothetical decision tree is illustrated in Figure 2.

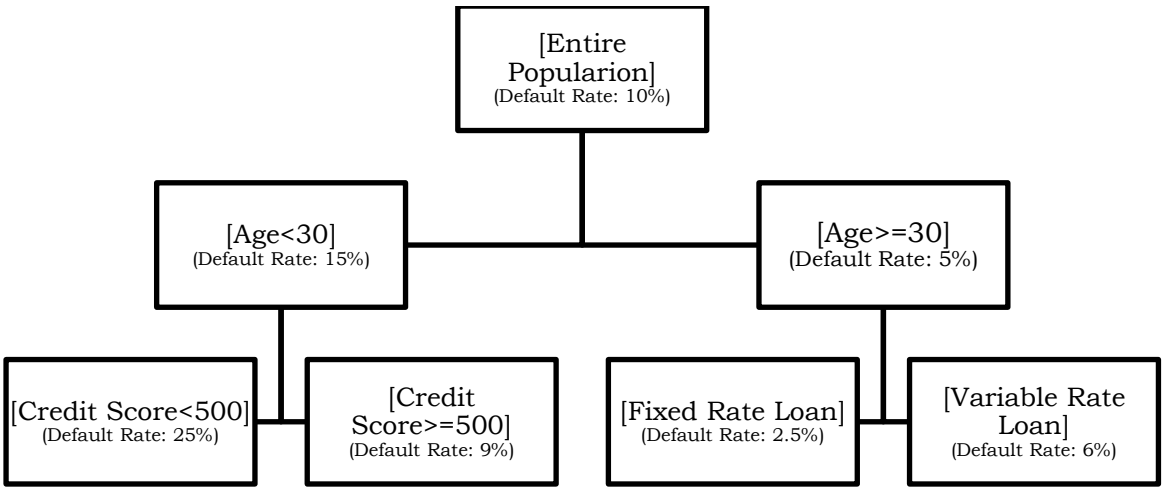


Figure 2: Illustration of a Decision-Tree Model

In the illustration, the primary determinant of risk within the portfolio is the age of the accountholder. For younger accountholders (younger than 30) the main risk determinant is the credit score, while for older accountholders it is the type of rate charged on the loan.

Other techniques that overcome the structural disadvantages of logistic regression exist. For example, machine learning techniques generally assume more complex model structures (e.g. Virag and Kristof, 2005; Van Gestel, 2005). Although such models have the advantage of producing very good fit, their complexity often raises questions of parsimony and the risk of overfitting. They are also generally difficult to interpret and intuitively justify. A comparison of machine learning techniques against logistic regression is summarised in the Table 1, adapted from Kumar, Roa and Soni (1995).

Attribute	**Neural Network**	**Logistic Regression**
Parsimony	Good	Fair
Classification Accuracy	Good	Fair
Solution Methodology	Fair	Good

Interpretability	Poor	Good
Intuitive Appeal	Poor	Good
Complex Interaction	Good	Poor
Statistical Testing	Fair	Good
Interpolating	Good	Fair
Extrapolating	Fair	Good
Interpretation of Importance Weights	Fair	Good

Table 1: A Comparison of Neural Networks to Logistic Regression (Kumar et al., 1995)

The second disadvantage of logistic regression is the limited amount of insight it produces. In some cases one is interested in more than just the probability of default. For example, we may be interested in the amount that we stand to lose in the event of default (i.e., exposure at default). In such a case, we may wish to model the distribution of time-to-default. Survival analysis is a technique that can produce such additional insight.

Survival analysis is concerned with estimating the waiting-time until the event of interest (i.e., default) occurs. An interesting application of survival analysis is to the analysis of defaults on corporate debt. In this setting, credit risk relates to the event that the issuer of the bond will default on its repayment. A default occurs if the issuer fails to make payments, typically due to bankruptcy or insolvency. This either means that none or only a part of the payment promised will be made. It could also mean that some of the payments may be deferred to a later date.

The Merton (1974) model, which models the issuer's assets as a geometric Brownian motion, is a common structural model for credit risk. Here it is assumed that default occurs at maturity only if the face value of the bond is greater than the value of the issuer. An extension to this model is the Black and Cox (1976) model, which defines default to have occurred at the first instance when the value of the firm falls below the face value of the bond. The distribution of the instance of default can be derived or estimated via survival analysis.

This form of survival analysis allows for the fact the default of the issuer is not a peculiarity, but something that is preceded by a period of poor financial performance. The waiting time until a Brownian motion process reaches a particular threshold follows an inverse Gaussian distribution, which means that the time-to-default under the Black and Cox (1976) model has an inverse Gaussian distribution (Lee and Whitmore, 2006). Threshold regression is the name of the form of survival analysis that models the process-degradation that leads to the event of interest. This form of regression works, producing tractable models, only under certain assumptions for the underlying process. This is a major limitation to its applicability. An overview and application of threshold regression with Brownian motion are given by Lee and Whitmore (2006).

A more general and common form of survival analysis is the proportional hazard regression – specifically, the Cox proportional hazard model. This assumes that the likelihood of a subject experiencing the event of interest is a function that can be expressed as a proportion of some baseline hazard function $h(t)$. The baseline hazard function $h(t)$ can be modelled either parametrically or non-parametrically. This fact is the main contributor to popularity of this approach. The main limitation is the requirement that event rates must be proportional at any point. There are many applications of Cox regression in credit risk analysis, including Malik and Thomas (2009) and Bellotti and Crook (2009).

Actuarial science generally focuses on certain financial risk that span a long period of time, such as longevity risk and long-term investment risk. Although the field mainly focused on insurance, there is growing involvement of actuarial thinking in retail banking. Booth, Chadburn, Haberman, James, Khorasanee, Plumb and Rickayzen (2005) mention some of the analogies in credit management that already exist between insurance and credit such as stress testing, survival analysis and VaR modelling.

Actuarial analysis has been applied to assist in modelling some of the complexities of that are difficult to model through parametric approaches.

For instance, Booth and Walsh (1998) discuss a cashflow model for the pricing of credit contracts. The paper goes beyond what typical research in the field attempts to achieve by showing how risk influences pricing strategy. The main advantage of cashflow techniques over parametric techniques is flexibility. The cashflow model is capable of allowing for ideas such as early repayment of loans, loan expense charges and sensitivity of the interest rate to the term of the loan in arriving at an interest rate to charge loan applicants.

2.1.2. Portfolio Models

Logistic regression, decision-tree analysis, survival analysis, machine learning and cashflow modelling are techniques particularly suitable for modelling the distribution of loss on individual accounts. A small combination of these modelling techniques would generally be sufficient for impairment modelling. However, in order to model unexpected losses, additional assumptions about the correlation within the portfolio are often required. This is required to produce a VaR measure for the portfolio of loans.

The VaR of a portfolio is defined over a defined horizon h for a particular loss percentage α, to be such that:

$$\alpha = F[A_h > VaR_h(\alpha)]$$

$$= 1 - F_h[VaR_h(\alpha)],$$

where A_h is the portfolio loss over the horizon h and $F_h(x)$ is the distribution function of A_h (Philippe, 2006). The unexpected loss over horizon h is then given by:

$$UL_h = 1 - F_h[VaR_h(\alpha)] - E[A_h].$$

where $E[A_h]$ is the expected loss. Under Base II, $\alpha = 0.999$, $h = 12$ and $F_h(x)$ is the distribution function of the Vašíček distribution:

$$F_h(x) = W \times \Phi\left[\frac{\sqrt{\rho}\Phi^{-1}(x)+\mu_h}{\sqrt{1-\rho}}\right],$$

where W is a scaling factor (equal to the LGD), ρ is the correlation factor under the Vašíček distribution and $\mu_h = \Phi^{-1}(PD)$ is the drift, $\Phi^{-1}(x)$ is the inverse of the Gaussian distribution function and PD is the loan portfolio default rate.

The derivation of the Vašíček distribution has a number of noteworthy assumptions. The distribution was derived specifically for a portfolio of corporate loans and is applicable to the default definition under the Merton (1974) model. The first assumption is that all loans have equal concentration within the portfolio. In consumer credit, this is analogous to assuming that all loans have the same outstanding balance. It is further assumed that, in the event of default, the loss ratio is non-random and equal across all loans, i.e., the LGD is constant and deterministic.

The third assumption of the Vašíček model is the value of the firm evolves according to a geometric Brownian motion, as under the Merton (1974) model (this paper discusses an analogy for this assumption for consumer credit).

The fourth assumption is that all portfolios are homogenous in risk, i.e., they have the same default rate. This is analogous to the assumption that all loans have the same PD.

The fifth assumption is that the firms value are independent, except for a dependence to the general economic environment (this paper discusses an analogy for this assumption for consumer credit). This dependence is assumed to be identical, as represented by the correlation parameter ρ.

The sixth assumption is that the portfolio is large enough for the Law of Large Numbers (see Golberg, 1984) to apply.

Due to the first and fourth assumption (homogeneity) and the sixth assumption (Law of Large Numbers), the Vašíček distribution is often called a large portfolio approximation (LHP). Attempts to relax the requirements for homogeneity and size exist. Of note, Pimbley (2011) relaxes both assumptions by using factorial approximations to analytically solve for the distribution. His results show that the Vašíček distribution produces poor fit for small portfolios (size=100) but quickly achieves accuracy as the size increases (with a near-perfect fit for portfolios large than 1000).

2.1.3. Systemic Risk Models

The Vašíček model, and subsequent refinements to the LHP assumptions, generally account for the fact that credit risk is influenced by some exogenous process. IFRS 9 and Basel II, as well as academic literature (e.g. Mileris, 2012; Thomas, 2000), share the broad consensus that the process is highly correlated to (or representative of) economic conditions.

A number of attempts to explicitly link credit risk to economic conditions exist. Rajaratnam, Beling and Overstreet (2010) consider the inclusion of economic outlook into credit scoring decisions. Their work is theoretical, focusing on the task of creating *efficient* credit portfolios in the case when the bank is faced with only two possible economic outlooks, each with a known probability of occurring. Although theoretical, the work addresses the problem initially tested by Crook et al. (1992), where it was found that the outcome of a credit scorecard is heavily influenced by the circumstances under which the development sample was observed.

The concept of hidden Markov models (HMM) can be considered a generalisation on the idea of binary economic outlook in Rajaratnam et al (2010). A HMM, as applied to credit risk, assumes that defaults rates are driven by the *state* of circumstances to which the credit portfolio is exposed. Banachewicz, Lucas and van der Vaart (2008) explore a particular type of

HMM where the transition between the different states is influenced by economic indicators such as the growth gross domestic product and interest rates. One of the weaknesses of the HMM is that it only allows for a finite number of states. HMMs are also more complex and intractable compared to conventional approaches and would not easily allow for loan-specific covariates.

Breaking away from finite-state modelling, Malik and Thomas (2009) and Bellotti and Crook (2009) apply Cox proportional hazard regression models with macroeconomic factors to predict default rates. These have the advantage of allowing for loan-specific covariates, as well as economic indicators. Both these studies find evidence for the influence of economic conditions on default rates. Both papers are accompanied by the complexities of using time-varying covariates within the Cox proportional hazard framework. However, the inclusion of macroeconomic variable results in an improvement in fit (compared to conventional logistic regression) and allows the models to be used for stress testing as well.

The influence of economic conditions on credit risk has also been modelled via time series analysis. This involves considering certain risk metrics, such as default rate, as a time series whose evolution is influenced by economic factors. A common approach for this form of modelling is the vector autoregressive model (e.g. Marcucci, and Quagliariello, 2008). Other forms of time series regression are surveyed by Foglia (2009). Time series modelling in credit risk is most appropriate for stress-testing, since is only focuses on the aggregate behaviour. This would not be appropriate for, for example, IFRS 9 modelling, where there is a requirement for loan-specific expected loss estimates to be made for certain high risk loans.

2.2. Research Areas

Although methods of credit risk analysis can be seen to be rich and varied, this form of analysis is still in its infancy, with many potential arrears of improvement and development. Thomas et al. (2005) identify a number of research areas and areas of development in consumer credit risk modelling. This paper is directly concerned with addressing two of these items.

According to Thomas et al. (2005), there is a need to develop a method for forecasting default rates, using economic data, without introducing bias into the ability of the model to rank-order risk at an individual customer level. The need for economic forecast methods in credit risk is also mentioned by Thomas (2000) as being an important area for development within consumer credit risk modelling.

Another area of research, according to Thomas et al. (2005), is the evaluation of the validity of the use of models for corporate credit risk in consumer credit lending. This point is specifically referring to the use of the Vašíček distribution in the Basel Accord's portfolio unexpected loss calculation. There are a number of noteworthy problems with the Basel Accord's assumptions for unexpected loss:

1. The Vašíček distribution is derived specifically for corporate loans, with little adaption attempted for consumer credit. This can lead to confusions around how the distribution ought to be applied. For example, the derivation uses a Merton (1974) approach, where default is only observed at maturity, while consumer credit defaults probabilities are more comparable to a Black and Cox (1976), where default occurs once a particular trigger event has occurred e.g. missing a certain number of payments. The question of the appropriateness of the Vašíček distribution concerns whether the parametric form of the distribution is appropriate for consumer loan portfolios.
2. The Basel Accord assumes different correlation parameter estimates for the distribution of loss according to the nature of the loan portfolio (e.g. whether the portfolio consists of mortgages or revolving loans).

However, little justification is available for the assumed correlation levels. This makes it difficult to validate the appropriateness of the model. Thomas et al. (2005) also note that, although we would expect extra information about default rates to decrease unexpected loss (since correlation between defaults would decrease), the application of the Basel model does not ensure this. The question of the chosen correlation levels concerns whether the parameters often chosen for the Vašíček distribution are appropriate for the loan portfolio at hand.

According to Thomas et al. (2005), the deficiency or, at least, lack of transparency of the Basel model provides motivation for investigations into new models for losses on credit portfolios.

Chapter 3: Default Rate Models

This chapter describes proposes two approaches for modelling credit defaults at an account-level and at portfolio level.

3.1. Inverse Gaussian Random Effects Model Survival

The first approach we describe is the inverse Gaussian survival model. The inverse Gaussian distribution was introduced by Schrödinger (1915) as the distribution of first-passage time of a Brownian motion process. It is one of the most commonly-used distributions in threshold regression.

We begin our discussion of the inverse Gaussian model by describing the derivation of the Vašíček distribution.

3.1.1. Vašíček Distribution

The Vašíček distribution was derived to model the distribution of losses on a portfolio of bonds, under the Merton (1974) default framework. Adopting the work of Campolongo, Jönsson and Schoutens (2012), we show below the derivation of the Vašíček distribution for a portfolio of corporate bonds.

Consider a portfolio of N zero-coupon bonds issued by N respective firms with the same outstanding term T. Let $B_k, k = 1,2,\dots,N$ be the face-value of the k^{th} bond in the portfolio, and V_t^k be value of the issuer of the k^{th} bond. Let L_T be the random loss on the portfolio of bonds.

To begin, a number of assumptions are made about the behaviour of the issuing firms. It is assumed that, in the event of a default on the k^{th} bond, an amount $LGD_k \times B_k$ will be lost per B_k of face-value. We assume that

default occurs at maturity time T if $V_T^k < B_k$. Further, we assume that the value of each firm V_t^k evolves stochastically according to a geometric Brownian motion, so that the value of the firm at time T is given by the following random function:

$$ln\, V_T^k = ln\, V_0^k + \left[\mu_k - \frac{\sigma_k^2}{2}\right] T + \sigma_k \sqrt{T} \varepsilon_k + v_k \sqrt{T} Z_T,$$

where V_0^k is the current value of the firm, ε_k is a standard normal random variable, μ_k is the drift parameter in the value for the firm and σ_k^2 is the volatility parameter. The variable Z_T is a standard Brownian motion, assumed to represent economic conditions, that affects all firms within the economy, as represented by the parameter v_k.

With these assumptions, the conditional probability that the k^{th} bond is in default at maturity is as follows:

$$\begin{aligned} Prob\left[ln\, V_T^k \le ln\, B_k \middle| Z_T = z_T\right] &= Prob\left[ln\, V_0^k + \left[\mu_k - \frac{\sigma_k^2}{2}\right] T + v_k \sqrt{T} z_T + \sigma_k \sqrt{T} \varepsilon_k \le ln\, B_k\right] \\ &= Prob\left[\varepsilon_k \le \frac{(ln\, B_k - ln\, V_0^k) - \left[\mu_k - \frac{\sigma_k^2}{2}\right] T - v_k z_T}{\sigma_k \sqrt{T}}\right] \\ &= \Phi\left[-\frac{\frac{ln\, V_0^k}{B_k} + \left[\mu_k - \frac{\sigma_k^2}{2}\right] T + v_k \sqrt{T} z_T}{\sigma_k \sqrt{T}}\right], \end{aligned}$$

where $\Phi(x)$ is the standard normal distribution function.

To proceed, we make additional assumptions about the nature of the portfolio. We assume that risk is homogenous within the portfolio, so that $\mu_k = \mu$, $\sigma_k = \sigma$, $v_k = v$ and $ln\frac{V_0^k}{B_k} = ln\frac{V_0}{B}$ for all $k = 1, \dots, N$. As a result, we have:

$$Prob\left[ln\, V_T^k \le ln\, B_k \middle| \mathrm{Z_T} = z_T\right] = p(T, z_T),$$

for all $\mathrm{k} = 1, \dots, \mathrm{N}$.

Secondly, we assume that the portfolio is sufficiently large (essentially infinite), allowing the Law of Large Numbers to apply for the portfolio default

rate, i.e., letting $ODR_T(z_T)$ be the observed rate of default within the portfolio, we assume that the portfolio size is such that $ODR_T(z_T) \cong p(T, z_T)$. We call these assumptions, that the portfolio is large and homogenous, the large homogenous portfolio (LHP) assumptions.

With these assumptions, the probability function of the observed default rate is as follows:

$$P[ODR_T(z_T) \leq x] = Prob\left[\Phi\left[-\frac{ln\frac{V_0}{B}+\left[\mu-\frac{\sigma^2}{2}\right]T+v\sqrt{T}z_T}{\sigma\sqrt{T}}\right] \leq x\right]$$

$$= Prob\left[-z_T \leq \frac{\sigma\sqrt{T}\Phi^{-1}(x)+ln\frac{V_0}{B}+\left[\mu-\frac{\sigma^2}{2}\right]T}{v\sqrt{T}}\right]$$

$$= \Phi\left[\frac{\sigma\sqrt{T}\Phi^{-1}(x)+ln\frac{V_0}{B}+\left[\mu-\frac{\sigma^2}{2}\right]T}{v\sqrt{T}}\right].$$

In addition, the correlation between the natural logarithms of any two firm values is as follows:

$$\rho(T) = \frac{Cov\left[ln\, V_T^k, ln\, V_T^j\right]}{\sqrt{Var\left[ln\, V_T^k\right]Var\left[ln\, V_T^j\right]}}$$

$$= \frac{v^2}{v^2+\sigma^2} = \rho,$$

i.e., the correlation between (logged) firm values is equivalent to the proportion of contribution of systemic volatility (v^2) to total volatility ($v^2 + \sigma^2$). Finally, we make the standardisation requirement that the total volatility be equal to one: $v^2 + \sigma^2 = 1$, which leads to $v = \sqrt{\rho}$ and $\sigma = \sqrt{1-\rho}$. The distribution function of portfolio default rate thus becomes:

$$\Phi\left[\frac{\sqrt{1-\rho}\Phi^{-1}(x)+ln\frac{V_0}{B}+\left[\mu-\frac{\sigma^2}{2}\right]T}{\sqrt{\rho}}\right].$$

The associated α^{th} percentile is thus as follows:

$$k_\alpha = \Phi\left[\frac{\sqrt{\rho}\Phi^{-1}(\alpha) - ln\frac{V_0}{B} - \left[\mu - \frac{\sigma^2}{2}\right]T}{\sqrt{1-\rho}}\right].$$

The portfolio loss is then given by:

$$L_T = N \times B \times LGD \times ODR_T(z_T).$$

By our assumptions, the only random component of the loss equation is $ODR_T(z_T)$. Therefore, the value-at-risk of the portfolio loss is as follows:

$$VaR_T(\alpha) = N \times B \times LGD \times \Phi\left[\frac{\sqrt{\rho}\Phi^{-1}(\alpha) - DD(T)}{\sqrt{1-\rho}}\right],$$

where $DD(T) = ln\frac{V_0}{B} + \left[\mu_k - \frac{\sigma_k^2}{2}\right]T$ is called the non-standardised distance-to-default.

3.1.2. Threshold Regression

We now discuss how threshold regression can be used to derive a similar model for consumer credit risk portfolios.

Threshold regression is a contemporary alternative to the more common proportional hazard regression (for a comparison, see Lee and Whitmore, 2010). This form of regression assumes that the occurrence of the target event is preceded by a degradation of some underlying, often latent, variable. For example, the transition of a patient's health status from HIV to AIDS may be assumed to be preceded by a decrease in the patient's CD4 cell count (see Lee, DeGruttola and Schoenfeld, 2000). Threshold regression models the waiting time until a particular process reaches some predetermined critical threshold (e.g. in HIV research, the waiting time until the CD4 cell count process reaches a lower barrier is of interest). Another example: in the Black and Cox (1976) model for corporate credit risk, threshold regression can be applied to model the waiting time until the firm's value drops below the outstanding loan value. In order to apply a

threshold model to consumer credit, in a way that is similar to the Black and Cox (1976) model, we make the following assumptions.

1. We assume that the savings of an accountholder follow a Brownian motion process $W(t)$ with drift parameter μ and volatility parameter σ.
2. We assume that in periods when the accountholder has insufficient disposable income, he/she draws on his/her savings account to make the monthly repayments. The accountholder starts missing payments when all savings are diminished and disposable income is unable to cover the repayment amount. At this point, savings become negative.
3. We assume that default occurs when the customer has missed θ payments. If p is the monthly instalment amount, default occurs at the first point when $W(t) < -\theta p$, i.e., $S_{HD} = \min\{t: W(t) < -\theta p\}$. The graph below shows the evolution of the savings process for five customers, where one customer's account moves into default.

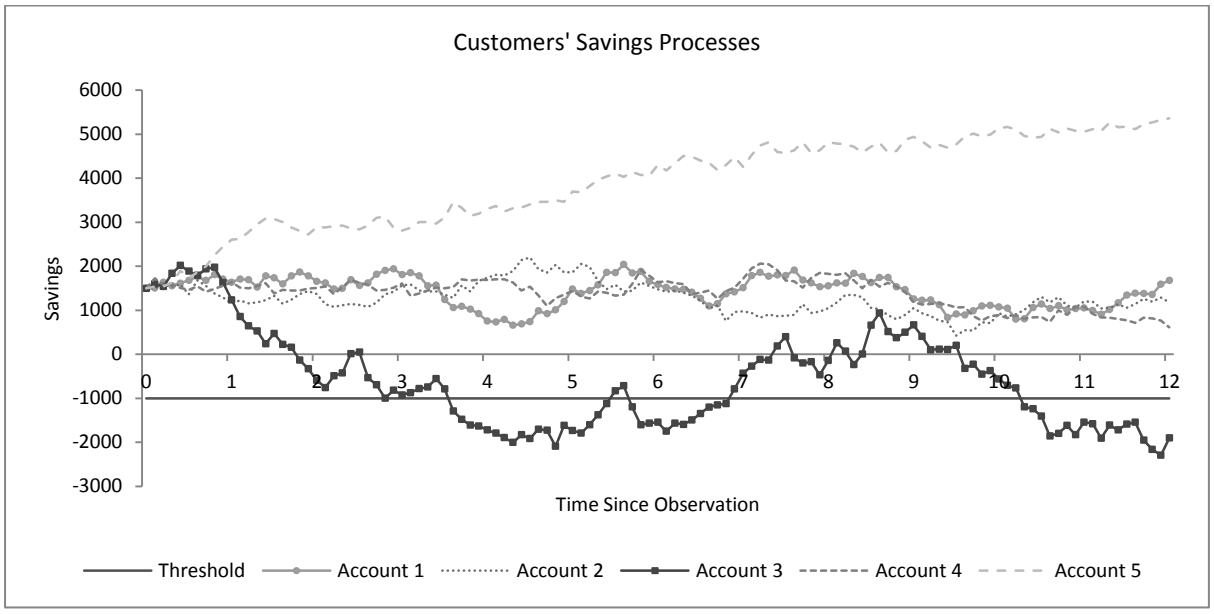


Figure 3: Illustration of the Filtration of a Customer's Savings Process

For the Brownian motion process described above, the waiting time until the account moves into default has an inverse Gaussian distribution, with the following density function (Lee and Whitmore, 2006):

$$f(t,\gamma,\sigma,\mu,\lambda) = \frac{\gamma+\lambda}{\sigma t\sqrt{2\pi t}} e^{-\frac{(\mu t-(\gamma+\lambda))^2}{2\sigma^2 t}},$$

where $\gamma = \theta p$ is the default threshold and λ is the accountholder's initial savings. The value of $\frac{\gamma+\lambda}{\sigma}$ can be seen as representing a standardised *distance-to-default*, i.e., how far a customer is, in monetary units, from the default point. The distribution function of the waiting time is as follows:

$$F(t, \delta, \sigma, \mu) = \Phi\left(\frac{\mu t - \delta}{\sigma\sqrt{t}}\right) + \Phi\left(-\frac{\mu t + \delta}{\sigma\sqrt{t}}\right) e^{\frac{2\delta\mu}{\sigma^2}},$$

where $\delta = (\gamma + \lambda)$. The two risk parameters of this model are the *distance-to-default* $\frac{\delta}{\sigma}$, which represents how secure the loan is initially, and the drift $\frac{\mu}{\sigma}$, which represents the speed with which the account moves towards or away from default.

The inverse Gaussian distribution is a survival distribution, like any other, although the associated hazard function is less tractable when compared to the more commonly-assumed survival functions. Since the volatility parameter merely scales the parameter estimates, it can be used to standardised the distance-to-default and drift e.g. we may set the $\sigma = 100$, so that the drift and distance-to-default are expressed in currency units of 100.

Threshold regression involves obtaining a regression formula for each the model parameters δ and μ based on account and customer covariates. However, unlike typical proportional hazards regression, threshold regression allows the target event to be changed, without having to refit the model, i.e., the default threshold θ can be changed without affecting the other model parameters. This is useful since different default definitions may be required for different modelling applications.

An important advantage that threshold regression approach has over the proportional hazard model is in how it allows for long-term survivorship. A survival model is subject to long-term survivorship if the population to which it applies consists of members that are not susceptible to the event of interest. These are called long-term survivors (see Roman, Louzada, Cancho and Leite, 2012). The presence of long term survival in the model population

is evidenced by a survival functions that has a non-zero horizontal asymptote. This asymptote represents the proportion of long-term survivors in the population. In the inverse Gaussian threshold model, the long term survivors are present when $\mu < 0$. In this case, the proportion of long-term survivors is given by $1 - e^{-\frac{\mu\delta}{2\sigma^2}}$ (Lee and Whitmore, 2006).

The task of fitting a threshold regression model consists of first specifying the model and then estimating the parameters of the model. Suppose that the drift parameter can be modelled as $\mu = \boldsymbol{\beta}^T \boldsymbol{Y}$, the initial savings level as $\delta = e^{\boldsymbol{\alpha}^T \mathbf{X}}$ and the volatility parameter as a constant, where $\mathbf{X}$ is a vector of covariates. We can estimate the parameters of the model $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ via maximum likelihood estimation. Consider a set n of observed accounts. Let μ_j, δ_j and T_j be the drift, initial savings parameter and observed time before default of the j^{th} account. Define C_j to be the default indicator:

$$C_j = \begin{cases} 1 & \text{if account j has not defaulted time } T_j \\ 0 & \text{if account j has defaulted time } T_j \end{cases}.$$

The parameters can be estimated by maximising the following log-likelihood function:

$$l(\sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{j=1}^{n} (1 - C_j) \ln f(T_j, \delta_j, \sigma, \mu_j) + \sum_{j=1}^{n} C_j \ln[1 - F(T_j, \delta_j, \sigma, \mu_j)].$$

Numerical methods can be used to find the maximum likelihood estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, given σ.

3.1.3. Macroeconomic Random Effects

In keeping with contemporary credit risk modelling objectives, our aim is to model the distribution of time-to-default as being influenced by an account's specific particulars, as well as macroeconomic factors. We define $\boldsymbol{X}_{k,s} = \{X_{k,s,1}, \dots, X_{k,s,p}\}$ as the covariate vector for account k observed in the credit

portfolio at calendar month s, and $\boldsymbol{Y}(s) = \{Y_1(s), Y_2(s), \dots, Y_q(s)\}$ as a vector of macroeconomic variables observable at calendar month s.

We assume that account k's distance-to-default is influenced by the account-specific information, as follows:

$$\delta_{k,s} = e^{\alpha_0 + \sum_{j=1}^{p} \alpha_j X_{k,s,j}},$$

while its drift is influenced by economic conditions as follows:

$$\mu_{k,s} = \mu_s = \sum_{j=1}^{q} \beta_j Y_j(s - l_j) + v\tilde{\varepsilon}_s$$

$$= \tilde{\mu}_s + v\tilde{\varepsilon}_s,$$

where $\varepsilon_s \sim N(0,1)$ is the random effect and l_j is the lag or the j^{th} macroeconomic variable.

The inclusion of the random factor ε_s implies that the drift applying to a particular credit portfolio can be modelled as a random function of observable macroeconomic variables, i.e., macroeconomic variables do not perfectly represent the portfolio drift at any calendar time. This is done without considerable loss of tractability, since the distribution function of a single loan would still have a closed form (Peng and Tseng, 2009):

$$F(t, \delta, \sigma, \mu, v) = \Phi\left(\frac{\mu t - \delta}{\sqrt{v^2 t^2 + \sigma^2 t}}\right) + \Phi\left(-\frac{2v^2\delta t + \sigma^2(\mu t + \delta)}{\sigma^2\sqrt{v^2 t^2 + \sigma^2 t}}\right) e^{\frac{2\delta\mu}{\sigma^2} + \frac{2v^2\delta^2}{\sigma^4}}.$$

Here, the original distribution function is obtained by letting $v \to 0$.

3.1.4. Loss Aggregation

The fact that the threshold regression model discussed above resembles the traditional model for corporate default allows us to derive the Vašíček distribution for aggregate loss in a consumer loan portfolio. Consider a portfolio of n_s homogenous customers at calendar time s whose savings satisfy the assumptions of the threshold regression model described above.

Let μ_s and σ be the common drift and volatility of the savings processes. Suppose that all customers have a current savings value of λ_s and that default occurs when $s_{j,s}(t) < -\gamma$, where $s_{j,s}(t)$ is the value of savings at time calendar time $s+t$ for the j^{th} customer in the portfolio at time s. The evolution of savings $s_j(t)$ is as follows:

$$s_{j,s}(t) = \lambda_s + \mu_s t + \varepsilon_j \sigma\sqrt{t},$$

where $\varepsilon_j \sim N(0,1)$. Note that, although all customers have the same risk parameters, the value of their individual savings evolve differently. We assume that ε_j and ε_k are independent for any $j \neq k$, so that savings processes evolve independently. Therefore, the probability that the j^{th} customer is in default at time h is as follows.

$$\begin{aligned} p_s(h) &= Prob\left[s_{j,s}(t) < -\gamma\right] \\ &= Prob\left[\lambda_s + \mu_s t + \varepsilon_j \sigma\sqrt{t}, < -\gamma\right] \\ &= \Phi\left[-\frac{\delta_s + \mu_s h}{\sigma\sqrt{h}}\right], \end{aligned}$$

where Φ is the distribution function of the standard normal distribution. It should be noted that, since defaults are only observed at time h, this probability excludes accounts that default and *cure* by time h. As described above, we model the drift as $\mu_s = \tilde{\mu}_s + v\tilde{\varepsilon}_s$. Substituting this into $s_{j,s}(t)$ we have:

$$s_{j,s}(t) = \lambda_s + \hat{\mu}_s t + \tilde{\varepsilon}_s v t + \varepsilon_j \sigma\sqrt{t}.$$

The total savings volatility is thus given by:

$$Var\left[s_{j,s}(t)\right] = v^2 t^2 + \sigma^2 t,$$

where $\sigma^2 t$ represents idiosyncratic savings volatility specific to the j^{th} customer, $v^2 t^2$ is the systemic savings volatility that remains uncaptured in the estimate for the drift. Thus, a better model for the drift produces a lower value for v – with the limiting case being $v = 0$ when the drift model is deterministic. The correlation coefficient between any two customers'

savings at time t, $s_{j,s}(t)$ and $s_{k,s}(t)$ for $j \neq k$, is equal to the ratio of systemic volatility to total volatility, as follows:

$$\rho(s,t) = \frac{Cov\left(s_{j,s}(t), s_{k,s}(t)\right)}{\sqrt{Var\left(s_{j,s}(t)\right)}\sqrt{Var\left(s_{k,s}(t)\right)}}$$

$$= \frac{\nu^2 t^2}{\sqrt{\nu^2 t^2 + \sigma^2 t}\sqrt{\nu^2 t^2 + \sigma^2 t}}$$

$$= \frac{\nu^2 t}{\nu^2 t + \sigma^2}.$$

Generally, $\nu > 0$ so that accounts savings processes are correlated, which means that defaults will typically be correlated. The probability of default on a single account can be restated as:

$$p_s(h, \tilde{\varepsilon}_s) = \Phi\left[-\frac{\delta_s + [\tilde{\mu}_s + \nu\tilde{\varepsilon}_s]h}{\sigma\sqrt{h}}\right].$$

We are now interested in calculating the distribution the proportion of the portfolio observed at calendar time s that will be in default at calendar time $s + h$. Let $d_{j,s}(h)$ be the default indicator for the j^{th} account at event horizon h:

$$d_{j,s}(h) = \begin{cases} 0 & \text{if the account is not in default at time } h \\ 1 & \text{if the account is in default at time } h \end{cases},$$

so that the observed default rate is as follows:

$$\tilde{p}_s(h) = \frac{1}{n_s}\sum_{j=1}^{n_s} d_{j,s}\,(h).$$

The expected value of $\tilde{p}_s(h)$ is given by $E[\tilde{p}_s(h)] = p_s(h, \tilde{\varepsilon}_s)$, since all contracts are homogenous. By the Law of Large Numbers, as $n_s \to \infty$ we have $\tilde{p}_s(h) \to p_s(h, \tilde{\varepsilon}_s)$. But $p_s(h, \tilde{\varepsilon}_s)$ is a random variable in its own right, since it is influenced by the random effect $\tilde{\varepsilon}_s$.

Therefore, the asymptotic distribution function of $\hat{p}_s(h)$ is as follows:

$$Prob[p_s(h, \tilde{\varepsilon}_s) \leq x] = Prob\left[\Phi\left[-\frac{\delta_s + [\tilde{\mu}_s + \nu\tilde{\varepsilon}_s]h}{\sigma\sqrt{h}}\right] \leq x\right]$$

$$= Prob\left[\tilde{\varepsilon} \geq -\frac{(\delta_s + \tilde{\mu}_s h) + \Phi^{-1}(x)\sigma\sqrt{h}}{\nu h}\right]$$

$$= \Phi\left[\frac{(\delta_s+\tilde{\mu}_s h)+\Phi^{-1}(x)\sigma\sqrt{h}}{\nu h}\right].$$

It is noteworthy that, by relaxing the normality assumption for the error $\tilde{\varepsilon}_s$, we would obtain a more general distribution function for $\hat{p}_s(h)$:

$$Prob[p_s(h,\tilde{\varepsilon}_s) \leq x] = G\left[\frac{(\delta_s+\tilde{\mu}_s h)+\Phi^{-1}(x)\sigma\sqrt{h}}{\nu h}\right],$$

where G is would be the distribution function of $\tilde{\varepsilon}_s$.

If the model parameters are calibrated for $h = 1$ such that $\sigma + \nu = 1$ then, letting $DD_s = (\delta_s + \tilde{\mu}_s)$ and $\rho(1) = \rho$, we retrieve the Vašíček distribution:

$$Prob[p_s(1,\tilde{\varepsilon}_s) \leq x] = \Phi\left[\frac{\Phi^{-1}(x)\sigma+DD_s}{\nu}\right]$$

$$= \Phi\left[\frac{\sqrt{1-\rho}\Phi^{-1}(x)+DD_s}{\sqrt{\rho}}\right],$$

since $\nu = \sqrt{\rho}$ and $\sigma = \sqrt{1-\rho}$ when $\nu^2 + \sigma^2 = 1$. The value-at-risk for a credit portfolio is thus given by:

$$VaR(\alpha) = n_s \times E_s \times LGD_s \times \Phi\left[\frac{\sqrt{\rho}\Phi^{-1}(\alpha)-DD_s}{\sqrt{1-\rho}}\right],$$

where E_s is the amount outstanding on each account and LGD_s is the loss-given-default on each account. The non-standardised distance-to-default, DD_s, has a special relationship with the probability of default on a single loan, given the drift:

$$p_{s|\tilde{\mu}_s} = P\left[s_{j,s}(t) < -\gamma \middle| \mu = \tilde{\mu}_s\right]$$

$$= \Phi\left[-\frac{DD_s}{\sigma}\right],$$

Therefore, the non-standardised distance-to-default can be expressed as a function of the probability of default, conditional on the drift:

$$DD_s = -\sigma\Phi^{-1}\left(p_{s|\mu}\right).$$

3.1.5. Comparison to Basel Capital Requirements

The capital requirements for consumer loans under Basel II are specified in terms of the value-at-risk, assuming the Vašíček distribution holds for portfolio default rate. The value-at-risk is given by:

$$VaR(\alpha) = n_s \times E_s \times LGD_s \times \Phi\left[\frac{\sqrt{\rho}\Phi^{-1}(\alpha)+\Phi^{-1}(\bar{p})}{\sqrt{1-\rho}}\right],$$

where ρ is the asset correlation coefficient and $\bar{p}$ is the through-the-cycle probability of default. The Basel requirement prescribes the value of ρ for different type of credit portfolios. In the Basel II requirements $-\Phi^{-1}(\bar{p})$ is analogous to the non-standardised distance-to-default.

Through looking at the model we derived in comparison to the Basel capital requirements model, there are a number of comments that can be made. Firstly, both models are dependent on the LHP assumptions:

1. The portfolio is homogenous (i.e. all accounts within the portfolio have identical risk profiles and outstanding balances).
2. The portfolio is large (i.e. the Law of Large Numbers provides a reasonable approximation to the portfolio default rate).

In practice, the homogeneity assumption can be reasonable ensured by segmenting the portfolio by risk profile, although this does not necessarily guarantee homogeneity. However, the extent to which this is possible is limited by the desire for each segment to be large enough to satisfy the assumption of a large portfolio. The assumption that the portfolio is large essentially removes any requirement to hold capital with respect to sampling (statistical) error, which may lead to overly optimistic capital reserves on small portfolios with large exposures.

The portfolio value-at-risk calculation essentially treats the loss given default as a non-random component of the loss distribution. In reality, the losses that result from default are random variables. Therefore, any

uncertainty that may arise from the loss-given-default component would be unaccounted for in the capital requirement.

The Vašíček distribution arises out of the assumption that defaults are related to the financial position of the borrower, and that the borrower's financial position evolves according to a Brownian motion process. If these assumptions are unmet, there is a possibility that the Vašíček distribution could lead to a gross misrepresentation of the true loss distribution.

From the derivation of the Vašíček distribution, ρ represent the correlation coefficient between the respective financial positions of individual borrowers. The correlation arises from a joint dependence on the idiosyncratic movements of a systemic credit risk cycle. Specifically, ρ is the ratio of systemic volatility to asset-specific volatility. Thus, the larger the idiosyncrasy of the system (i.e. the volatility of the credit risk cycle), the greater the value of ρ should be.

In our derivation, we remove some of volatility of the credit risk index by attempting to match its movements through macroeconomic variables. This adjusts the distance-to-default component to closely match the observed (or point-in-time) default rate for the given month, and reduces the value of the correlation coefficient. The Basel requirements, on the other hand, are based on a through-the-cycle approach: the distance-to-default matches the long-run default rate, and asset correlation coefficient is expected to be larger, i.e., in our derivation the correlation coefficient captures movements in the credit risk index left unexplained by macroeconomic variables while the through-the-cycle methodology captures the full volatility of the credit risk index.

The through-the-cycle methodology is expected to lead to more stable capital requirements, since the distance-to-default is designed to correspond to the long-run default rate. However, the methodology is based on the assumption that the credit risk index is cyclical, so that a long-run default probability exists. This means that capital requirements would fail to capture credit risk experience that falls outside of the presumed credit cycle.

In this thesis, we consider extent to which the LHP assumptions are invalidated in a given portfolio of loans, and how this affects the estimated capital requirement. We also consider the extent to which the assumption that borrower's assets follow a Brownian motion are met in the portfolio as well as the effect of a non-constant LGD.

The paper does discuss the extent of applicability of the through-the-cycle methodology in the loan portfolio. However, to the extent that the credit risk cycle is assumed to be produced by the business cycle, we note that the cyclicality of the credit risk index may be contentious (Mankiw, 1989). Especially in developing countries that may have a volatile or irregular economic cycle, the through-the-cycle methodology may lead to capital inadequacy. Here, a better approach may be to hold capital to match the prevailing economic circumstance. At the least, economic capital provisions set on a through-the-cycle approach should be tested for adequacy on a point-in-time basis.

A common weakness shared by both Basel II capital model and the economic capital model described in this paper is that they both rely on two sources of randomness: one idiosyncratic, and one systemic. In some instances, one may require more factors. Fok, Yan and Yao (2014) discuss a hierarchical approach for corporate loans where three factors are used: an idiosyncratic factor, a sector factor (specific to the sector that the corporate operates in) and a systemic factor (general to the entire economy). However, the extent to which additional factors would benefit the analysis for consumer loans needs to be understood.

3.2. Logistic Random Effects Model

The second approach we discuss is based on logistic regression, which is one of the most widely-used methods in consumer credit modelling. Logistic

regression is part of the class of generalised linear models, generally focused on modelling the distribution of binary random variables (Cox, 1958).

3.2.1. Account-Level Model

Consider a sequence of Bernoulli random variables $\{D_k\}_{k=1}^{n}$ with associated covariate vectors $\{\boldsymbol{X}_k\}_{k=1}^{n}$, where $\boldsymbol{X}_k = \{X_{k,1}, X_{k,2}, \dots, X_{k,p}\}$. Logistic regression aims to estimate the parameter p_k of the distribution of D_k as a function of $\boldsymbol{X}_k$, where:

$$p_k = Prob[D_k = 1].$$

The probability mass function of D_k can thus be written as:

$$f(d, p_k) = (1 - p_k)e^{d\, ln\left(\frac{p_k}{1-p_k}\right)}, \text{ for } d = 0,1.$$

The probability mass function is in the form of the exponential family of distribution, with natural parameter $Q(p) = ln\left(\frac{p}{1-p}\right)$. Therefore, it is common to model p_k as:

$$\ln\left(\frac{p_k}{1-p_k}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{k,j},$$

where $\boldsymbol{\beta} = \{\beta_0, \beta_1, \beta_2, \dots, \beta_p\}$ is the vector-parameter of the model, i.e., the logit transformation of p_k is assumed to be linearly related to the covariate vector. Common alternatives to the logit transformation include the probit transformation:

$$\Phi^{-1}(p_k) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{k,j},$$

where Φ is the cumulative density function of the standard normal distribution, and the complementary-log-log transformation:

$$\ln(-\ln(1 - p_k)) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{k,j}.$$

If, in general, we say:

$$g^{-1}(p_k) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{k,j},$$

then we can estimate **β** by maximising the following log-likelihood function with respect to **β**:

$$l(\boldsymbol{\beta}) = \sum_{k=1}^{n} \left[D_k \ln g\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{k,j}\right) + (1 - D_k) \ln\left(1 - g\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{k,j}\right)\right)\right].$$

$\widehat{\boldsymbol{\beta}}$, the maximum likelihood estimate for **β** has no closed-form solution. However, estimation can be done numerically, through Fisher-scoring (Jennrich and Sampson, 1976) or nonlinear optimisation techniques such as the Nelder-Mead method (Nelder and Mead, 1965). A general discussion on the application and interpretation of logistic regression is done by Lottes, Adler and DeMaris (1996).

[4]For our purposes, we define $\{D_{k,s}(h)\}_{k=1}^{n_s}$ to represents the default indicators on loans in a portfolio of size n_s held at calendar time s, where default is observed over some chosen event horizon of length h, i.e:

$$D_{k,s} = \begin{cases} 1 & \text{if the loan moves into default at any time before time } t+h \\ 0 & \text{otherwise} \end{cases},$$

so that $p_k(s)$ is the default rate of loan k. $\boldsymbol{X}_{k,s} = \{X_{k,s,1}, \ldots, X_{k,s,p}\}$ represents the set of loan-specific covariates associated with the loan as at observation (calendar time s).

3.2.2. Macroeconomic Factors

The covariate vectors $\{\boldsymbol{X}_k\}_{k=1}^{n}$ contain representations of only loan-specific information. We now further assume that the distributions of $\{D_{k,s}\}_{k=1}^{n}$ are further influenced by a vector of time series $\boldsymbol{Y}(s) = \{Y_1(s), Y_2(s), \ldots, Y_q(s)\}$, where $Y_j(s)$ is the value of macroeconomic variable Y_j at time t. For this, we will consider two model forms:

[4] $D_{k,s}$ is used throughout as the shorthand of $D_{k,s}(h)$.

1. Linear-Logistic: $\ddot{p}_k(s, \varepsilon_s) = \Phi\big(\alpha_0 + \sum_{j=1}^{p} \alpha_j X_{k,s,j} + \sum_{j=1}^{q} \beta_j Y_j\big(s - l_j\big) + \varepsilon_s\big)$
2. Log-Logistic: $\ddot{p}_k(s, \varepsilon_s) = \Phi\big(\alpha_0 + \sum_{j=1}^{p} \alpha_j X_{k,s,j}\big)^{e^{\sum_{j=1}^{q} \beta_j Y_j\left(s-l_j\right)+\varepsilon_s}}$

where $\varepsilon_s \sim N(0, v^2)$ is the random effect and l_j is the lag of the j^{th} macroeconomic variable. The presence of the random effect leads to the interpretation that the macroeconomic index that influences default rates is captured imperfectly by the macroeconomic variables included. This is a reasonable assumption, since macroeconomic variables are always an aggregation of economic conditions, and would rarely ever perfectly represent the exact situation of a given credit portfolio.

In the linear-logistic model, the random effect is linear with both macroeconomic and account-level covariates. This could confuse the source of the randomness: an alternative interpretation of this model could be that the influence of certain account-level covariates is random, from one month to the next, so that the random effect is not purely macroeconomic. The log-logistic model overcomes this issue, since the random effect is linear only with the macroeconomic variables. It should be noted that the linear-logistic model is very similar to some binary adaptations of the Merton model for consumer loans. As such, it is comparable to the model considered by Crook and Bellotti (2012).

The parameters from these random effects models are estimated by maximising the following likelihood function:

$$L(\boldsymbol{X}, \boldsymbol{Y}) = \prod_s E_{\varepsilon_s}\Big[\prod_{k=1}^{n_s} \ddot{p}_k(s, \varepsilon_s)^{D_{k,s}(h)}[1 - \ddot{p}_k(s, \varepsilon_s)]^{1-D_{k,s}(h)}\Big].$$

The inclusion of a random effect requires that the log-likelihood function be expressed as an expectation of the probability mass function of $D_{k,s}(h)$, taken over the distribution of ε_s. This expectation generally has no closed-form solution. However, numerical integration techniques can be used in conjunction with conventional optimisation techniques to obtain parameter estimates.

3.2.3. Correction Factor Approach

The log-logistic model described above can be arrived at using a two-stage modelling approach. Let $\ddot{p}_k$ be a default rate model for an account with covariate vector $\boldsymbol{X}_k$:

$$\bar{p}_{k,s} = g\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{k,s,j}\right),$$

where g is the link function. Since $\bar{p}_{k,s}$ includes no macroeconomic factors, it can be interpreted as a through-the-cycle predicted rate of default (i.e. an average predicted default rate across the entire economic cycle). The prediction $\ddot{p}_{k,s}$ can be adjusted to a particular stage in the cycle by introducing the macroeconomic factors that are thought to represent the credit risk cycle. Thus, the first step in reproducing the log-logistic model is to identify the macroeconomic variables that represent the credit risk cycle for the portfolio being modelled. We assume the following parametric form for the relationship between the true default rate $\ddot{p}_k(s)$ and the through-the-cycle default rate $\bar{p}_{k,s}$:

$$\ddot{p}_k(s, \varepsilon_s) = {\bar{p}_{k,s}}^{e^{C_s}},$$

where C_s is the credit index. The value of C_s can be estimated from the entire book of accounts observed at time s by maximising the following likelihood:

$$l_s(C_s) = \sum_{k=1}^{n_s} \left[D_{k,s}\, ln\, {\bar{p}_{k,s}}^{e^{C_s}} + \left(1 - D_{k,s}\right) \ln\left(1 - {\bar{p}_{k,s}}^{e^{C_s}}\right)\right].$$

The maximum likelihood estimates of C_s form an estimated time series of the credit risk index. Therefore, time series approaches could be used to model the credit risk index, and thus identify the macroeconomic factors that drive its value. Assume that the following model is obtained:

$$C_s = \sum_{j=1}^{q} \beta_j Y_j\left(s - l_j\right) + \varepsilon_s,$$

where $\varepsilon_s \sim N(0, v^2)$ is the random effect and l_j is the lag of the j^{th} macroeconomic variable. The model assumes that the macroeconomic

variables model the credit risk index with a standard error of v. This suggests a random effect model. Therefore, the final fitting can be re-performed as a complementary-log-log random effect regression model, with the following model equation:

$$\ln\left(-\ln\left(\ddot{p}_k(s,\varepsilon_s)\right)\right) = \ln(-\ln(\ddot{p}_k)) + \sum_{j=1}^{q}\beta_j Y_j\left(s-l_j\right) + \varepsilon_s.$$

This is an example of the correction factor approach to macroeconomic modelling described by Crook and Bellotti (2010).

3.2.4. Moments of $\mathbf{D_k(s)}$

The probability of default calculated on a single account can be found by taking the expectation of $\ddot{p}_k(s,\varepsilon_s)$ over the random effect ε_s as follows:

$$\ddot{p}_k(s) = E_{\varepsilon_s}[\ddot{p}_k(s,\varepsilon_s)].$$

The expectation generally has no analytical solution. It can be approximated using a first-order Taylor expansion as follows:

$$[f(X)] \approx f(\mu_X).$$

This approximation works well for approximately linear functions, $f(X)$, or for small variance v^2 (Ang and Tang, 2007). With this approximation, the expectation for the linear-logistic random effect model is given by:

$$\ddot{p}_k(s) \approx \Phi\left(\delta\left(\boldsymbol{X}_{k,s}\right) + \mu_s\right),$$

which is equivalent to a simple (non-random) logistic regression model with account-level and macroeconomic covariates included, where $\delta\left(\boldsymbol{X}_{j,s}\right) = \alpha_0 + \sum_{j=1}^{p}\alpha_j X_{k,s,j}$ and $\mu_s = \sum_{j=1}^{q}\beta_j Y_j\left(s-l_j\right)$. The log-logistic random effect model yields the following approximation:

$$\ddot{p}_k(s) \approx \Phi\left(\delta\left(\boldsymbol{X}_{k,s}\right)\right)^{e^{\mu_s+\frac{1}{2}v^2}}.$$

We note that a closed-form solution can be found for the linear-logistic model, the derivation of which is given in the appendices (see Appendix 7.2). This expectation is given by:

$$\ddot{p}_k(s) = \Phi\left[-\frac{\alpha_0+\sum_{j=1}^{p}\alpha_j X_{k,s,j}+\sum_{j=1}^{q}\beta_j Y_j(s-l_j)}{\sqrt{v^2+1}}\right].$$

The introduction of random effects into the logistic regression model introduces dependence between the default events on individual accounts. Consider the default indicator on two accounts observed in calendar month s: $D_{j,s}$ and $D_{k,s}$. The covariance coefficient of these two random variables is obtained as follows:

$$Cov\left(D_{j,s}, D_{k,s}\right) = E\left[D_{j,s}D_{k,s}\right] - E\left[D_{j,s}\right]E\left[D_{k,s}\right]$$

$$= E_{\varepsilon_s}\left[\ddot{p}_j(s,\varepsilon_s)\ddot{p}_k(s,\varepsilon_s)\right] - E_{\varepsilon_s}\left[\ddot{p}_j(s,\varepsilon_s)\right]E_{\varepsilon_s}[\ddot{p}_k(s,\varepsilon_s)].$$

For $D_{j,s}$ and $D_{k,s}$ to be uncorrelated, we require that $Cov\left(D_{j,s}, D_{k,s}\right) = 0$, which is only met on the condition that:

$$E_{\varepsilon_s}\left[\ddot{p}_j(s,\varepsilon_s)\ddot{p}_k(s,\varepsilon_s)\right] = E_{\varepsilon_s}\left[\ddot{p}_j(s,\varepsilon_s)\right]E_{\varepsilon_s}[\ddot{p}_k(s,\varepsilon_s)],$$

Since $\ddot{p}_j(s,\varepsilon_s)$ and $\ddot{p}_k(s,\varepsilon_s)$ are both functions of ε_s, this condition is not satisfied.

One of the interests in credit risk modelling is the determination of the distribution of the number of defaults on a credit portfolio:

$$D_s = \sum_{k=1}^{n_s} D_{k,s},$$

and, equivalently, the distribution of the portfolio default rate:

$$p_s = \frac{D_s}{n_s}.$$

Finding an exact formula for the distribution of the portfolio default rate for even small portfolios can be intractable. A convention exists to approximate this distribution by assuming that all accounts in the portfolio have similar characteristics (i.e. the homogeneity assumption).

If $\ddot{p}(s, \varepsilon_s)$ is the probability of default on all accounts within a homogenous portfolio at calendar time s then, for a given realisation of the random effect ε_s, the distribution of the number of defaults follows a binomial distribution, with a mass function given by (Pimbley, 2011):

$$f(x, \varepsilon_s) = \binom{n_s}{x} \ddot{p}(s, \varepsilon_s)^x [1 - \ddot{p}(s, \varepsilon_s)]^{1-x}.$$

An additional assumption is often made, that the size of the portfolio approaches infinity (i.e. $n_s \to \infty$). Under this assumption, the loss distribution would be approximated by the Gaussian distribution, through the central limit theorem. However, the central limit theorem only holds when defaults are uncorrelated (see Hilhorst, 2009).

3.2.5. Loss Aggregation

It is possible to derive a distribution for portfolio losses that takes the default correlation into account, by making use of the large homogenous portfolio assumptions (i.e. that all accounts have equal probability of default and $n_s \to \infty$). This in turn allows us to define an expression for economic capital, one which is comparable to that which is used for Basel regulatory capital requirements.

We are interested in finding the distribution of losses that arise from defaults occurring within a single h-month horizon. We define the random loss on account k in the portfolio in calendar month s as:

$$L_{k,s} = EAD_{k,s} \times LGD_{k,s} \times D_{k,s},$$

where $EAD_{j,s}$ is the exposure-at-default and $LGD_{k,s}$ is the loss-given-default (i.e. $EAD_{j,s}$ is the amount at risk and $LGD_{k,s}$ is the proportion of this amount that is lost as a result of default). The portfolio loss is thus given by:

$$L_s = \sum_{k=1}^{n_s} L_{k,s}.$$

We are interested in defining the portfolio loss distribution:

$$F(x) = Prob[L_s \leq x].$$

This is done under the assumption of homogeneity, that $EAD_{k,s} \times LGD_{k,s} = W_s$ and $\boldsymbol{X}_{k,s} = \boldsymbol{X}_s$ for all k and s (the vector $\boldsymbol{X}_s = (X_{s,1}, \dots, X_{s,p})$ is a *notional average* of the covariates within the portfolio at calendar time s; its exact value is discussed below - see Table 3). For simplicity, we further assume that W_s is a constant, so that the number of defaults is the only source of randomness in the loss model. Therefore, the total portfolio loss simplifies to:

$$L_s = W_s D_s.$$

Under these assumptions, the distribution function of L_s is given by:

$$F(x) = Prob[W_s D_s \leq x]$$

$$= G\left(\frac{x}{n_s W_s}\right),$$

where $G(x)$ is its distribution function of the portfolio default rate p_s. Letting $n_s \to \infty$, we have:

$$p_s \to \ddot{p}(s, \varepsilon_s),$$

which follows from the Law of Large Numbers, i.e., as the sample size approaches infinity, the sample mean p_s approaches the expected value $\ddot{p}(s, \varepsilon_s)$. However, since $\ddot{p}(s, \varepsilon_s)$ is itself a random value (since it is a function of ε_s), we can derive an expression for the distribution function $G(x)$ as follows:

$$G(x) = Prob[\ddot{p}(s, \varepsilon_s) \leq x].$$

The analytical solution of $G(x)$ will differ for the two models of $\ddot{p}(s, \varepsilon_s)$. Under the linear-logistic model, we have:

$$G(x) = Prob[\Phi(\delta_s + \mu_s + \varepsilon_s) \leq x]$$

$$= \Phi\left(\frac{\Phi^{-1}(x) - \delta_s - \mu_s}{v}\right),$$

where $\delta_s = \delta(\boldsymbol{X}_s) = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_{s,j}$. Similarly for the log-logistic model, we have:

$$G(x) = Prob\left[\Phi(\delta_s)^{e^{\mu_s+\varepsilon_s}} \leq x\right]$$

$$= \Phi\left(-\frac{\ln\left[\frac{\ln(x)}{\ln \Phi(\delta_s)}\right]-\mu_s}{v}\right),$$

which is in the form of a log-log-normal distribution. From $G(x)$ we are interested in deriving the percentiles of the distribution of the default rate and, ultimately, the portfolio loss. The α^{th} default rate percentile is given by:

$$r_s(\alpha) = G^{-1}(\alpha),$$

and the α^{th} portfolio loss percentile is given by:

$$l_s(\alpha) = r_s(\alpha) W_s.$$

The analytical formulae for $r_s(\alpha)$ for the two models are given in Table 2.

Linear-Logistic VaR	Log-Logistic VaR
$\Phi\left(\boldsymbol{\delta_s} + \boldsymbol{\mu_s} + \boldsymbol{v\Phi^{-1}(\alpha)}\right)$	$\Phi(\delta_s)^{e^{\mu_s - v\Phi^{-1}(\alpha)}}$

Table 2: Formulae for the Portfolio Value-at-Risk

The portfolio average δ_s should not be necessarily taken as linear averages of $\delta(\boldsymbol{X}_{j,s})$, since $\delta(\boldsymbol{X}_{j,s})$ is not a linear function of the default rate. We propose estimating δ_s from the Taylor approximations of the portfolio default rate. If $\bar{p}_s = \sum_{k=1}^{n_s} \ddot{p}_k(s)$ is the Taylor approximation for the portfolio default rate, we estimate δ_s from:

$$\bar{p}_s \approx \Phi(\beta_0 + \delta_s + \mu_s),$$

under the linear-logistic model, and:

$$\bar{p}_s \approx \Phi(\delta_s)^{e^{\mu_s + \frac{1}{2}v^2}},$$

under log-logistic model. The estimates obtained from solving these equations are given in Table 3.

Linear Logistic	Logarithmic Logistic
$\Phi^{-1}(\bar{\boldsymbol{p}}_{\boldsymbol{s}}) - \boldsymbol{\mu}_{\boldsymbol{s}}$	$\Phi^{-1}\left((\bar{p}_s)^{e^{-\mu_s-\frac{1}{2}v^2}}\right)$

Table 3: Notional Portfolio Average Risk Levels, δ_s

The loss distributions derived from these approximations offer an alternative to the Vašíček distribution, which is used for Basel capital requirements. Notably, the Vašíček distribution is derived under similar assumptions to those used in this paper.

We called the assumption that the portfolio is (infinitely) large and consists of accounts that are homogenous in risk the large homogenous portfolio (LHP) assumption. The different distributions of the portfolio default rate that result from this assumption under different models we call the LHP approximations. The LHP approximations under the log-logistic model and the linear-logistic model are summarised, in the form of the value-at-risk, in Table 2.

Further comparison between the Vašíček distribution and linear-logistic regression model can be made. The Vašíček distribution is derived from the Merton model for defaults on corporate bonds (Vašíček, 1987). Under this model, a firm's default probability is given by:

$$\Phi\left[\frac{\delta-\mu}{\sigma}\right],$$

where δ is the default threshold, μ is the tendency of the firm to move towards default and σ is the volatility of the firm's value. The Vašíček distribution is obtained by introducing a systemic random effect, leading to the following portfolio default rate distribution:

$$G(x) = \Phi\left[\frac{\Phi^{-1}(x)\sqrt{\rho}+\delta-\mu}{\sqrt{1-\rho}}\right],$$

where ρ is a measure of the correlation introduced by the random effect (see section 3.1). This distribution function bears close similarity to the distribution function derived under the linear-logistic model. This is expected, since the default rate under the Merton model is similar to the

default rate under the linear-logistic model. In fact, it would be possible to use the linear-logistic model to produce the Vašíček distribution as an LHP approximation by restating the regression model as:

$$\ddot{p}_k(s, \varepsilon_s) = \Phi\left(\frac{\delta(\boldsymbol{X}_{j,s}) + [\mu(\boldsymbol{Y}_s) + \varepsilon_s]}{\sigma}\right),$$

although σ would be treated as a nuisance parameter. This idea has been pursued, to an extent, by (Crook and Bellotti, 2012).

Chapter 4: Model Development

In this chapter, the two approaches proposed in Chapter 3 are applied to build default rate models for a portfolio of personal loans. The chapter begins by describing the portfolio and how the development sample is constructed.

4.1. Portfolio Summary

We consider a portfolio of unsecured fixed-rate personal loans issued by a South African bank over the period from 2006 to 2013. The loans are denominated in South African Rand (ZAR) and issued to within the South African economy.

The period selected for model development is considered long enough to cover a full economic cycle. Particularly, the period covers the economic downturn experienced between 2008 and 2009. This is particularly important when modelling loss distributions on a through-the-cycle basis. Also, when building macroeconomic time series models, one requires a long enough period to determine how the target variable is influenced by macroeconomic factors.

The accounts were separated into a development sample of size 1,800,000, with 97,323 defaults over a 12 month horizon, and a validation sample of size 450,000, with 24,616 defaults over a 12 month horizon.

4.2. Sample Construction

An account is defined to be in default if it is three or more payments in arrears or is escalated to the bank's legal department for loss mitigation. An

account would be escalated to the legal department if the accountholder is deemed unable to make further repayments. Examples of what might lead to this include the event where the accountholder becomes deceased or is declared bankrupt.

Not all accounts in the study eventually default. For the purpose of the analysis, these accounts will be right-censored. This includes cases where the accountholder repays the full contractual amount and accounts that have still not defaulted after C months, where C is the predetermined point of censoring.

For the inverse Gaussian model, the following are calculated for each account:

- $T_{j,s}$: the length of time account j observed in the portfolio at calendar time s was observed in the study prior to default or censorship (i.e. the observed survival time)
- $C_{j,s}$: an indicator of whether account j observed in the portfolio at calendar time s was censored at time $T_{j,s}$, i.e.,

$$C_{j,s} = \begin{cases} 1 & \text{if account j has not defaulted time } T_{j,s} \\ 0 & \text{if account j has defaulted time } T_{j,s} \end{cases}.$$

- $D_{j,s}(h)$: the h-month default indicator for account j observed in the portfolio at calendar time s, i.e.,

$$D_{k,s} = \begin{cases} 1 & \text{if the loan moves into default at any time before time } t+h \\ 0 & \text{otherwise} \end{cases}.$$

- $\boldsymbol{X}_{k,s}$: covariate vector for account j observed in the portfolio at time s.
- $\boldsymbol{Y}(s)$: macroeconomic vector observable at time s, obtained from the South African Reserve Bank website .

The default indicator $D_{j,s}(h)$ is used at the target variable in the logistic regression models, with horizon $h = 12$. The survival time $T_{j,s}$ and censorship indicator $C_{j,s}$ are used in the inverse Gaussian survival model. The vectors $\boldsymbol{X}_{k,s}$ and $\boldsymbol{Y}(s)$ are used as model inputs.

4.3. Macroeconomic Analysis

The first step in model development is to determine the optimal set of macroeconomic variables to include in the model, as well as the lag at which the effect of each variable on the default rate is maximal. For this, a correlation analysis was carried out.

A set of 13 variables were selected as the starting point of the analysis. These were chosen to be general indicators of economic conditions and customer credit affordability. Some of these variables commonly feature in macroeconomic credit risk models (e.g. Bellotti and Crook, 2009, Malik and Thomas, 2010). In order to avoid spurious correlation, we ensure that all variables possess an explicable influence (positive or negative) on the default rate before including them into the model. The variables considered as well as the explanation of what influence would be expected, are summarised in Table 4.

Variable	**Interpretation**	**Expected Influence**
Coincident_Indicator	The yearly change in the South African Reserve Bank's coincident economic index.	We expected a negative relationship, since positive economic conditions would lead to less difficulty servicing debt.
CPI	The yearly change in the consumer price index.	We expect a positive relationship, since rising prices would leave lower levels of disposable income.
Debt_Cost_to_Income	The ratio of household debt service cost to disposable income of households.	We expect a positive relationship, since rising debt costs would lead to greater likelihood of being unable to service the debt.
Debt_to_Income	The ratio of debt to household disposable income.	We expect a positive relationship, since rising debt would lead to rising debt servicing costs.
Disposable_Income	The yearly change in disposable household income.	We expect a negative relationship, since rising disposable income would mean greater ease in servicing debt.
Emp_Compensation	The yearly change in	We expect a negative relationship,

	employee compensations.	since increasing compensation would increase disposable income.
GDP	The yearly change in the gross domestic produces	We expect a negative relationship, since a growing economic would generally result an improvement of per-capita income.
HH_Consumption	The yearly change in aggregate household consumptions	We expect a positive relationship, since an increase in consumption would decrease disposable income.
Leading_Indicator	The yearly change in the South African Reserve Bank's lagging economic index.	We expected a negative relationship, since positive economic conditions would lead to less difficulty servicing debt.
Prime	The yearly change in the South African Reserve Bank's leading economic index.	We expect a positive relationship, since increasing interest rates would lead to increasing debt servicing costs.
Res_Compensation	The yearly change in compensations of residents.	We expect a negative relationship, since increasing compensation would increase disposable income.
Savings_to_Income	The ratio of household savings to household disposable income.	We expect a negative relationship, since savings can be drawn upon to augment disposable income where necessary.
Unemployment	The official South African unemployment rate.	We expect a positive relationship, since rising unemployment would signify downturn conditions.

Table 4: Macroeconomic Variables Used for Modelling

The process followed in arriving at the optimal set of variables is as follows.

1. Each variable was tested for stationarity via the single-mean Augmented Dicky-Fuller (ADF) test (Fuller, 1976). Variables showing significant evidence for non-stationary at a 5% level of significance were differenced. The same was done for the observed portfolio default probability. The probability values of the ADF test are summarised in Table 5.

Variable	**Non-Differenced**	**Differenced**
default_rate	*0.5451*	*0.0008*

coincident_indicator	0.2476	0.0187
CPI	0.0008	
debt_cost_to_income	0.2735	0.0372
debt_to_income	0.2057	0.0067
disposable_income	0.0608	
emp_compensation	0.7300	0.0008
GDP	0.0166	
hh_consumption	0.0214	
leading_indicator	0.1869	0.0052
Prime	0.9217	0.0008
res_compensation	0.0324	
savings_to_income	0.3456	0.0035
unemployment	0.0501	0.0008

Table 5: ADF Test Results

2. In order to determine the lag at which the influence of each macroeconomic variable is highest, the correlation between the default rate and each variable (differenced where necessary, as indicated in Table 5) was calculated for a 3-month lag, 6-month lag, 9-month lag and 12-month lag. The lag with the highest correlation coefficient was taken as the lag where the influence is maximal. The results are summarised in the Table 6.

Variable	**Lag**	**Correlation Coefficient**
coincident_indicator	6	0.2203
cpi	12	-0.2370
debt_cost_to_income	3	0.1253
debt_to_income	12	0.2329
disposable_income	9	0.1314
emp_compensation	6	-0.2366
gdp	3	0.2843
hh_consumption	3	0.3120
leading_indicator	12	0.1029
prime	12	0.1187
res_compensation	12	-0.1466
savings_to_income	9	-0.2997
unemployment	12	-0.1161

Table 6: Final Macroeconomic Variables and Lags

3. In order to limit the effect of multicollinearity, a correlation matrix of the variables at the selected lags was constructed. For each pair of variables with an absolute correlation coefficient of more than 60%, the variable with a lower correlation to the default rate was removed.
4. Variables that had a counter-intuitive association with the default rate were also removed, to ensure the absence of spurious correlation. The correlation matrix on the final set of variables is given in Table 7.

	debt_to_income	**disposable_income**	**emp_compensation**	**hh_consumption**	**leading_indicator**	**prime**	**res_compensation**	**savings_to_income**
debt_to_income		19.75%	-21.47%	27.31%	3.94%	17.08%	8.21%	-28.54%
disposable_income	19.75%		6.96%	37.44%	47.67%	-29.00%	-44.10%	-8.42%
emp_compensation	-21.47%	6.96%		-13.19%	-4.89%	-17.24%	-2.74%	18.46%
hh_consumption	27.31%	37.44%	-13.19%		54.14%	-14.23%	-56.11%	-45.87%
leading_indicator	3.94%	47.67%	-4.89%	54.14%		-17.06%	-29.35%	-6.45%
prime	17.08%	-29.00%	-17.24%	-14.23%	-17.06%		42.24%	-11.99%
res_compensation	8.21%	-44.10%	-2.74%	-56.11%	-29.35%	42.24%		13.14%
savings_to_income	-28.54%	-8.42%	18.46%	-45.87%	-6.45%	-11.99%	13.14%	

Table 7: Macroeconomic Variable Correlation Matrix

This final set of variables is taken to be representative of all macroeconomic influences on default rates. These variables were later used in the development of default rate models.

4.4. Inverse Gaussian Model

We start by discussing the development of a threshold regression model for the probability of default.

4.4.1. Testing the Inverse Gaussian Assumption

We are interested in fitting the inverse Gaussian threshold regression model, which assumes that the occurrence of default is influenced by an underlying Brownian motion process. Since the underlying process is latent, we cannot directly test whether it is a Brownian motion. However, we can test whether the assumption is met by assessing whether the distribution of time-to-default is inverse Gaussian. We do this by constructing an empirical default hazard function, as the Kaplan-Meier estimator (see Kaplan and Meier, 1958), and comparing it to the inverse Gaussian hazard function, with the drift and distance-to-default estimated via maximum likelihood methods.

In general applications, there is no reason to expect the underlying process to follow a Brownian motion. However, in some cases, even if the underlying process is not a Brownian motion, it may be possible to obtain a close approximation by applying a running-time transformation i.e. model the survival function $S(\tau)$ instead of $S(t)$, where $\tau = g(t)$ and g is a monotonic increasing function (see Whitmore and Schenkelberg, 1997). We attempt the transformation $\tau = \sqrt{t}$. Figure 4 compares the inverse Gaussian model and the inverse Gaussian model with the running-time transformation to the empirical hazard.

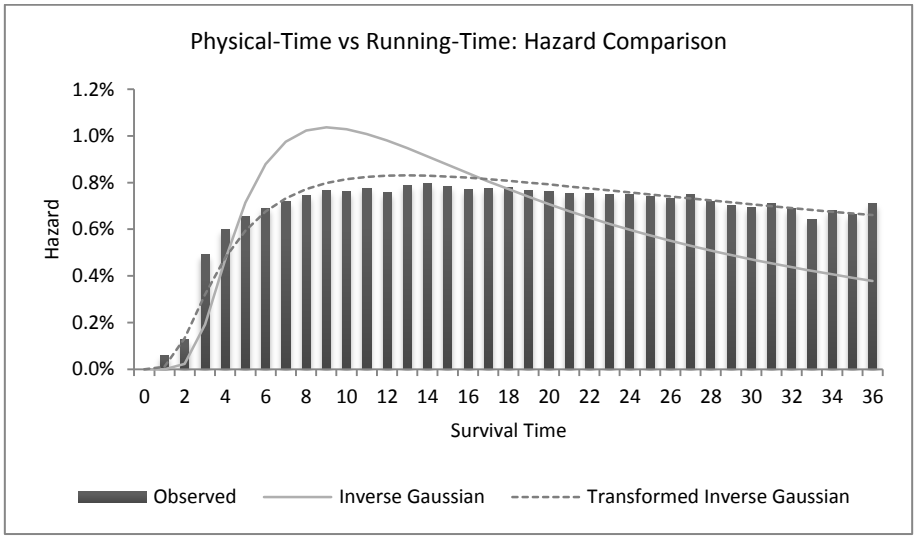


Figure 4: The Fitness of the Inverse Gaussian Hazard Function

The graph shows that the inverse Gaussian model with a running-time transformation improves the model considerably. However, we see that the model over-predicts defaults in the first two months and under predicts from month three to month six.

4.4.2. Including Account-Level Covariates

We begin the regression process by introducing account-specific covariates to estimate the distance-to-default:

$$\delta_{k,s} = e^{\alpha_0 + \sum_{j=1}^{p} \alpha_j X_{k,s,j}}.$$

The process of finding an optimal model (offering a good balance between parsimony and fitness) would typically involve performing variable selection (e.g. stepwise, forward or backward selection) and comparing information criteria, e.g. AIC or BIC (see Burnham and Anderson 2002). However, since most software packages do not have special procedures for threshold, given a large number of covariates this would be computationally intensive. A heuristic approach for doing this would be conduct variable selection through fitting a logistic model for the probability of default within h months, for a sufficiently long h (e.g. $h = 12$).

4.4.3. Including Macroeconomic Variables

Once all account-level covariates are allowed for, we wish to introduce macroeconomic variables to estimate the drift:

$$\mu_s = \sum_{j=1}^{q} \beta_j Y_j(s - l_j) + v\tilde{\varepsilon}_s.$$

Three things are of note about this model. Firstly, the model assumes that the drift applying to a portfolio observed at time s will be constant over the

horizon h to which the model will be applied. This is unlikely to be the case in reality, especially over long horizons. It thus makes sense to make the horizon to be as short as is necessary. This means that accounts that survive for more than h months should be right-censored, where h is a predetermined horizon.

Secondly, the model assumes that the drift is constant across different levels of risk within a portfolio. This assumption can be relaxed by introducing account-level covariates into the drift model:

$$\mu\left(\boldsymbol{X}_{j,s}, \boldsymbol{Y}_s\right) = \alpha_0 + \sum_{k=1}^{p} \beta_k X_{j,s,k} + \sum_{k=1}^{q} \alpha_k Y_{k,s-l_k} + v\tilde{\varepsilon}_s.$$

This can be further generalised to allow for interactions between the macroeconomic variables and account-level covariates.

Finally, the model assumes that the true drift is unobserved, and that the macroeconomic variables only offer an unbiased approximation – the error of this approximation is assumed to be normally distributed with variance v^2. Consequently, this becomes a random effect model. We estimate the parameters using SAS, with the NLMIXED procedure (see Wolfinger, 1999).

Given a large set of macroeconomic variables, in the interest of parsimony, we would need to select an optimal set of macroeconomic variables. This selection needs to also allow for the fact that a macroeconomic variable $Y_{k,s}$ may affect default patterns with a lag l_k. A heuristic approach to obtaining the optimal set of macroeconomic variable is to reduce the default rates of a given portfolio into a time series $\hat{p}_s$ and fit a time-series regression via stepwise selection.

The final fitted parameters for the regression model are summarised in Table 8.

Type	Variable 1	Variable 2	Variable Level 1	Variable Level 2	Estimate
Intercept					6.2015
Covariate	pl_mmsinc1plus		1		-0.2868
Covariate	pl_mmsinc1plus		2		-0.1945
Covariate	pl_mmsinc1plus		3		-0.1169

Covariate	paid_down_ratio		2		0.0242
Covariate	relative_interest_rate		2		-0.0832
Covariate	relative_interest_rate		4		-0.1375
Covariate	fb_hr_indicator		2		-0.4561
Covariate	pl_wpp_1y		1		-0.1600
Interaction	pl_mmsinc1plus	FB_HR_INDICATOR	1	2	0.3270
Interaction	pl_mmsinc1plus	FB_HR_INDICATOR	2	2	0.3391
Interaction	pl_mmsinc1plus	FB_HR_INDICATOR	3	2	0.3529
Interaction	pl_mmsinc1plus	relative_interest_rate	1	2	0.0317
Interaction	pl_mmsinc1plus	relative_interest_rate	1	4	0.0379
Interaction	pl_mmsinc1plus	relative_interest_rate	2	2	0.0314
Interaction	pl_mmsinc1plus	relative_interest_rate	2	4	0.0258
Interaction	pl_mmsinc1plus	relative_interest_rate	3	2	0.0350
Interaction	pl_mmsinc1plus	relative_interest_rate	3	4	0.0277
Interaction	paid_down_ratio	relative_interest_rate	1	2	0.0591
Interaction	paid_down_ratio	relative_interest_rate	2	2	0.0511
Interaction	paid_down_ratio	relative_interest_rate	2	4	0.0763
Macroeconomic	debt_to_income_12				-0.0897
Macroeconomic	disposable_income_9				1.1239
Macroeconomic	emp_compensation_6				0.3919
Macroeconomic	hh_consumption_3				0.4208
Macroeconomic	leading_indicator_12				-0.1239
Macroeconomic	prime_12				1.4833
Macroeconomic	savings_to_income_9				-3.6132
Random Error	v				1.8348

Table 8: Parameter Estimates of the Macroeconomic Inverse Gaussian Model

4.4.4. Accuracy Assessment

There are a number of dimensions over which the accuracy of the model can be assessed. Here we are interested in two: the ability of the model to predict 12-month cumulative default rates over time and the ability of the model to predict 12-month non-cumulative default rates over time.

Cumulative Default Rate Assessment

We first assess the model's ability to predict the overall 12-month cumulative default rates. The 12-month cumulative default rates are observations from the time series of $\tilde{q}_s(12)$, where $\tilde{q}_s(t)$ is the Kaplan-Meier estimate of the probability of an account observed in the portfolio in calendar month s defaulting within t months. These are compared to the average predicted default probability:

$$\hat{q}_s(12) = 1 - \frac{1}{n_s}\sum_j^{n_s} S\big(12, \delta_{j,s}, \sigma, \mu_s, v\big),$$

where $S\big(t, \delta_{j,s}, \sigma, \mu_s, v\big)$ is the predicted survival function for account j in month s:

$$S\big(t, \delta_{j,s}, \sigma, \mu_s, v\big) = \Phi\left(\frac{\mu_s t - \delta_{j,s}}{\sqrt{v^2t^2+\sigma^2 t}}\right) + \Phi\left(-\frac{2v^2\delta_{j,s}t + \sigma^2(\mu_s t + \delta_{j,s})}{\sigma^2\sqrt{v^2t^2+\sigma^2 t}}\right) e^{\frac{2\delta_{j,s}\mu_s}{\sigma^2} + \frac{2v^2{\delta_{j,s}}^2}{\sigma^4}}.$$

This comparison is shown in Figure 5. In order to demonstrate the effect of the presence of macroeconomic variables in the model, the graph also shows the prediction $\hat{q}_s(12)$ from a model fitted without macroeconomic variables.

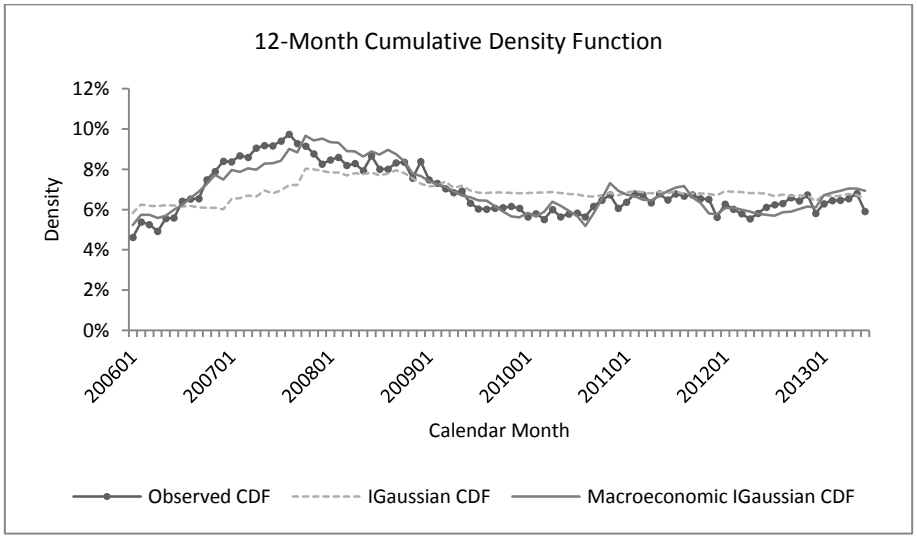


Figure 5: Fitness of the 12-Month Cumulative Density Function

We see that the inclusion of macroeconomic variables improves the model's ability to predict changes in default rates over calendar time.

Non-Cumulative Default Rate Assessment

The Vašíček distribution was derived as an approximation to the probability of a firm defaulting at the date of the maturity of debt. This corresponds to a non-cumulative probability of default. Under the model, the h-month non-cumulative probability of default is the probability that an account is in default at time h, which is given by:

$$p_{j,s}(h, \tilde{\varepsilon}_s) = \Phi\left[-\frac{\delta_{j,s}+[\tilde{\mu}_s+v\tilde{\varepsilon}_s]h}{\sigma\sqrt{h}}\right],$$

where:

$$\tilde{\mu}_s = \sum_{j=1}^{q} \beta_j Y_j(s - l_j).$$

Since $\tilde{\varepsilon}_s$ is unobserved, $p_{j,s}(h, \tilde{\varepsilon}_s)$ is a random variable. The expected value of $p_{j,s}(h, \tilde{\varepsilon}_s)$ is given by (the proof is provided in Appendix 7.2):

$$p_{j,s}(h) = \Phi\left[-\frac{\delta_{j,s}+\tilde{\mu}_s h}{\sqrt{v^2h^2+\sigma h}}\right].$$

We calculate the predicted portfolio default rate as:

$$\hat{p}_s = \frac{1}{n_s}\sum_{j=1}^{n_s} p_{j,s}(h).$$

This is compared to the empirical non-cumulative default rate:

$$\tilde{p}_s(h) = \frac{1}{n_s}\sum_{j=1}^{n_s} D_{j,s}(h).$$

This comparison is given in Figure 6.

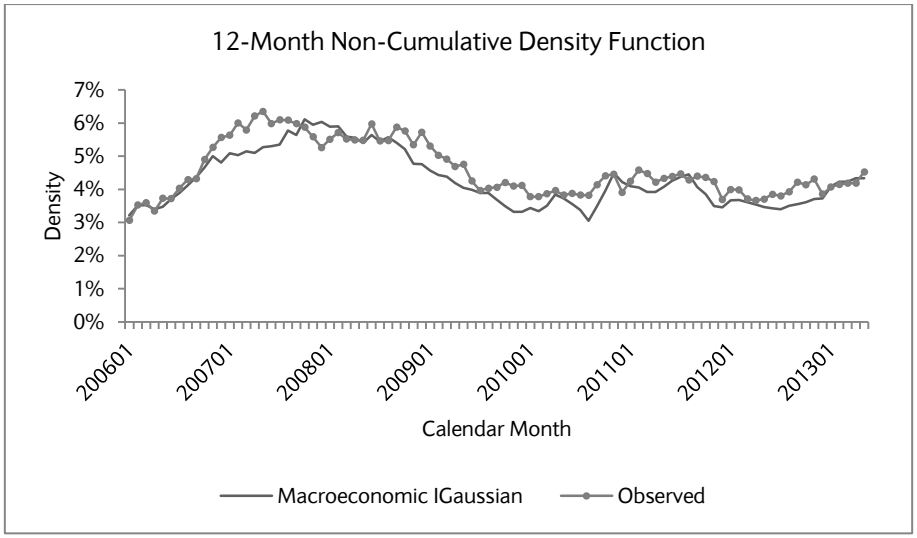


Figure 6: Fitness of the 12-Month Non-Cumulative Density Function

The graph shows that the predicted default rate aligns well with the observed default rate, but the model tends to under-predict the default rate. This may be a result of the fact that the Brownian motion assumption is not fully met by the default experience.

4.5. Logistic Regression Default Model

In this section we consider the use logistic regression to create a simple model for account defaults over a 12-month time horizon, taking into account a small number of covariates.

4.5.1. Default Rate Models

Three default rate models were developed. The first model was developed using only account-level information. The second and third models are the linear-logistic and log-logistic models, respectively, which both incorporate macroeconomic covariates. The set of covariates considered in the first model, without macroeconomic variables, are summarised in Table 9.

Variable	Description	Gini Statistic (Monthly Average)
pl_mmsinc1plus	Number of months since the account was more than one payment in arrears.	23%
paid_down_ratio	Proportion of loan paid off since inception.	11%
relative_interest_rate	Interest rate on account relative to other accounts within portfolio.	16%
fb_hr_indicator	Forbearance & High Risk Indicator.	18%
pl_wpp_1y	Worst payment position in the last 12 months.	23%

Table 9: Account-Level Covariates

The fitting results for the linear-logistic and the log-logistic model are given in the appendices (Logistic Regression Parameter Estimates).

The first model was fitted with a Gini statistic of about 40% (see Mair, Reise, Bentler, 2008, for a description of the Gini statistic). The Gini statistic here is used as a measure of a model's discriminatory power. This model was also assessed using the Hosmer-Lemeshow test (see Hosmer and Lemeshow, 2000, for a description of the test), with a p-value of 29%. Therefore, at a 5% level of significance, there is no evidence that, in aggregate, the actual default probabilities on accounts in the portfolio are different from those predicted by the model.

We also wish to test whether the model is able to predict with a stable level of accuracy across business cycle. Firstly, we plot the Gini statistic (in Figure 7) over time, to measure the discriminatory power of the model across different time periods. Secondly, we plot the time series of predicted portfolio default rate along with the actual portfolio default rate (in Figure 8).

Looking at Figure 7, we see that the model maintains its ability to associate high risk accounts with higher probabilities of default. However, from Figure 8 we see it fails to predict the correct default rate over certain periods. This is expected, since the model does not incorporate any macroeconomic information.

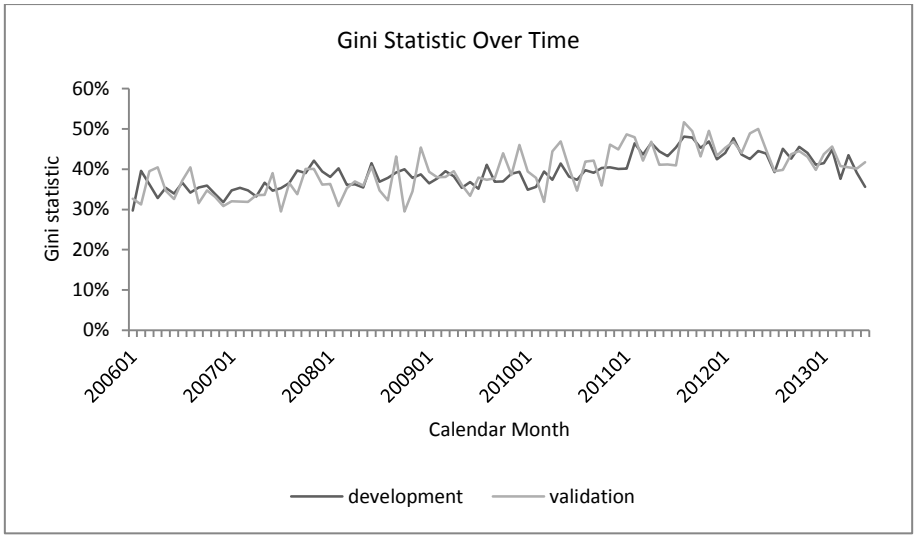


Figure 7: Discriminatory Power across Time

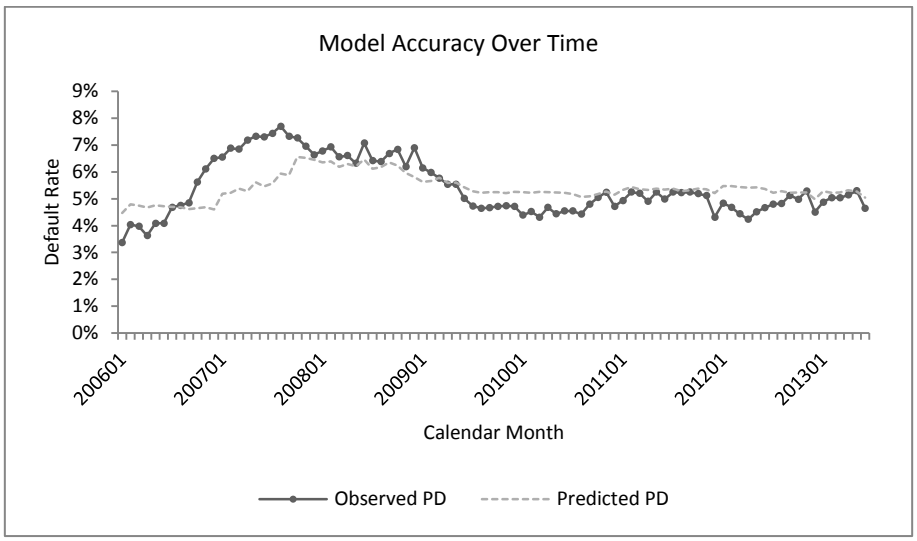


Figure 8: Logistic Model Accuracy across Time

The linear-logistic random effect model was fitted with a Gini statistic of 41.73%. The log-logistic model showed similar results, with a Gini statistic of 41.63%. From this, we conclude macroeconomic variables do not add a considerable amount to a models ability to discriminate between risks. However, plotting the model's predictions against observed default rates, we see in Figure 9 below that the inclusion of macroeconomic conditions increases the consistency of the model's predictions across time.

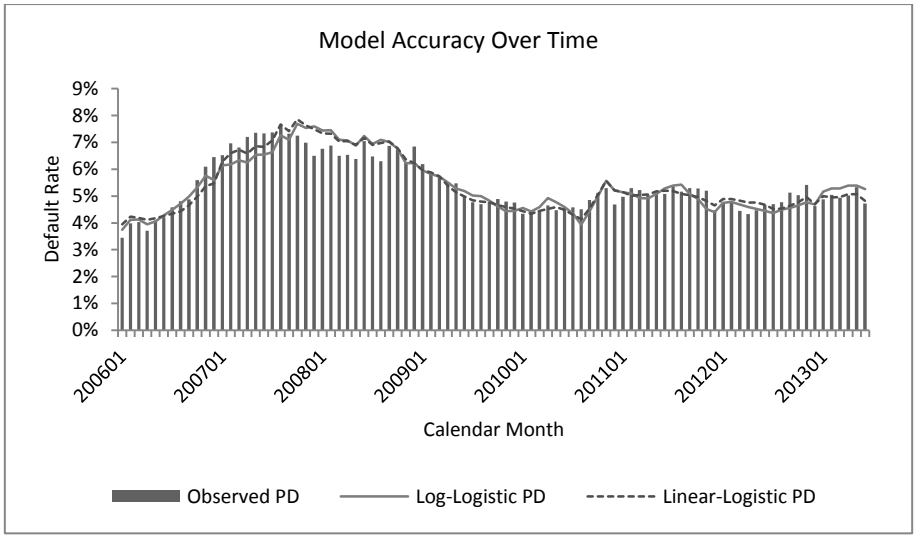


Figure 9: Random Effect Model Accuracy across Time

Chapter 5: Economic Capital

In Chapter 4 we discussed various statistical models for the loss distribution. Two models for the probability of default were fitted. We discussed an inverse Gaussian model for the distribution of time-to-default and how, through random effects threshold regression, the model can be adjusted to accommodate macroeconomic factors. A logistic regression model was fitted as an alternative to the inverse Gaussian model.

In Chapter 3 we showed that the inverse Gaussian model approximates the distribution of the non-cumulative default rate of a portfolio as a Vašíček distribution (see section 3.1.4), while the logistic regression model produces either the log-log-normal distribution or the Vašíček distribution as an approximation for the portfolio default rate (see section 3.2.5), depending on the chosen model link function. We refer to these distributions as LHP approximations to the portfolio default rate.

In this chapter we discuss how these distributions are used to determine economic capital by using the models developed in Chapter 4.

5.1. Portfolio Default Rate Confidence Intervals

The derivation of the LHP distributions makes further assumptions that those that are required for the regression models, i.e., the assumption that the portfolio is large and homogenous. The assumptions for the regression models were tested via conventional methods, such as the Homser-Lemeshow test and the Gini statistic. The LHP assumptions can be tested by a binomial test.

We test the following set of hypotheses:

H_0: the LHP distribution is representative of the data
H_1: the LHP distribution is not representative the data

Under the null hypothesis, H_0, the confidence interval for the data should be given by the appropriate quantiles of the log-log-normal distribution and the Vašíček distribution. Therefore, one test is to construct a $100(1-\alpha)\%$ confidence interval for the portfolio default rate and count the number of times the observed portfolio default rate falls outside of this interval, K. Under the null hypothesis, K has a binomial distribution with rate parameter α.

The linear-logistic random effect has an estimated standard deviation of 2.3%. From this, we are able to construct a confidence interval for the portfolio default rate in any given month. Figure 10 is a graph of the LHP confidence interval under the linear-logistic regression model.

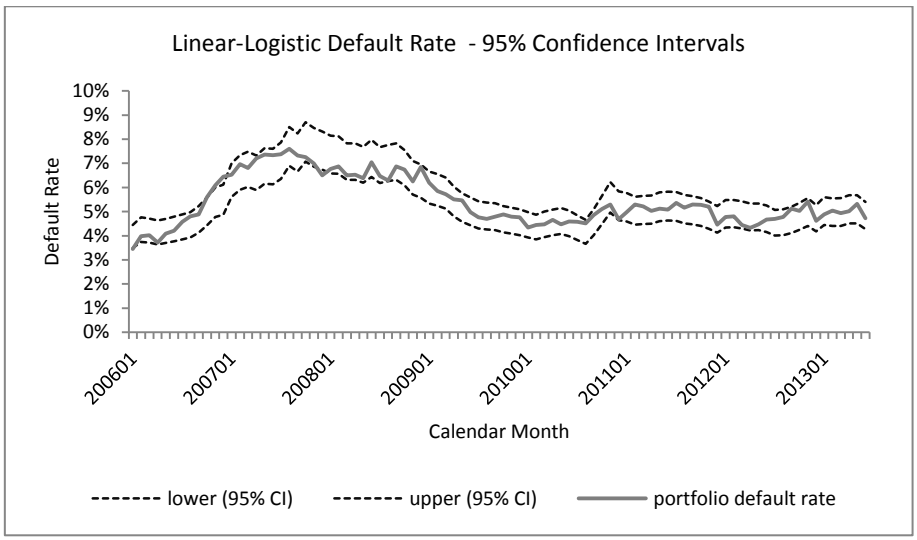


Figure 10: 95% Confidence Interval under Linear-Logistic Model

We see that the observed default rate is well-contained within the confidence intervals, with $K = 5$ observation outside the confidence interval. The associated p-value is 29%, which means we fail to reject the hypothesis that the approximation fits the data at a 95% level of confidence. Similarly, we plot the 95% confidence interval of the log-logistic random effect model, which has an estimated standard deviation of 2.9%, in the Figure 11.

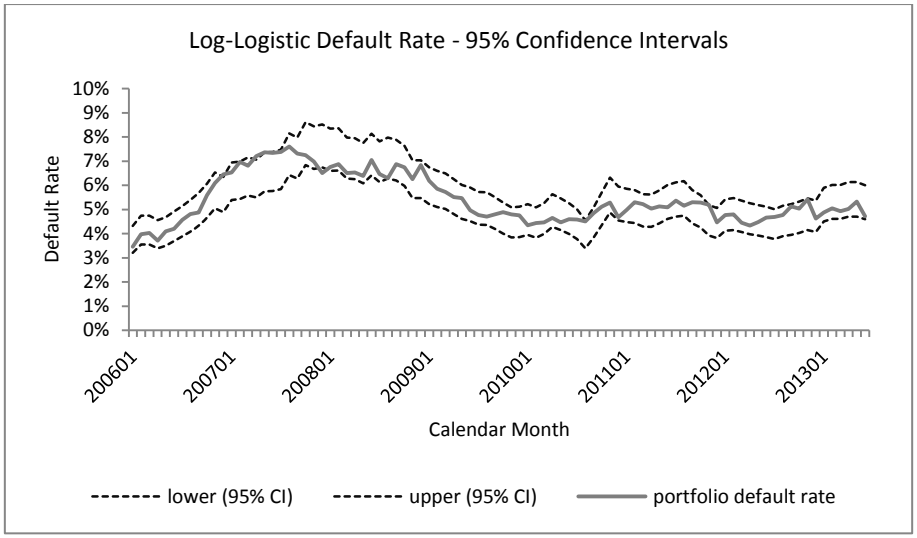


Figure 11: 95% Confidence Interval under Log-Logistic Model

We see similar results for in the logarithmic-logistic model, with K = 4, which yields a p-value of 47%. Therefore, there is a lack of evidence against the LHP approximations. However, we note that breaches in the confidence interval under both models may tend to cluster together, which would suggest the presence of autocorrelation in errors. A possible remedy, if this were the case, would be to include a stronger set of macroeconomic variables in the model e.g. including more macroeconomic variables, or widening the set of macroeconomic variables considered for the model.

5.2. Economic Capital under the Vašíček Model

The inverse Gaussian model is of the same form as that used in determining Basel II capital requirements. The main difference is that the inverse Gaussian model relaxes the assumption of the existence of regular credit risk cycle. Instead, the model assumes a credit risk index that can be reasonably represented by a set of macroeconomic variables, which may well be cyclical. In this way, the inverse Gaussian model can be seen as a generalisation of the Basel II regulatory capital model.

We can thus test some of the assumptions of the Basel II model, through testing the assumptions of the inverse Gaussian model. These assumptions are:

1. The portfolio is infinitely large (or, the Law of Large Numbers applies for the default rates),
2. The portfolio is homogenous in default risk,
3. The portfolio is homogenous in exposure and
4. Accounts are subject to a constant (non-random) LGD.

Large Homogenous Portfolio Assumption

We assess the first two assumptions, that the portfolio is large and homogenous in risk (i.e. the LHP assumption), jointly. Under the LHP assumption, the h-month non-cumulative probability of default of the portfolio has the following distribution function:

$$F_p(x,h) = \Phi\left[\frac{\sqrt{1-\rho(h)}\Phi^{-1}(x)+DD_s}{\sqrt{\rho(h)}}\right],$$

where $\rho(h)$ is the correlation coefficient:

$$\rho(h) = \frac{\nu^2 h}{\nu^2 h+\sigma^2}.$$

In Figure 12, we used this distribution function to test the LHP assumption on the entire sample, by constructing a 95% confidence interval for the true portfolio default rate $p_s(h)$ under the model. We now wish to assess the fitness of the distribution in its entirety, as well as determine how sensitive the fitness is to the size of the sample. In order to assess the fitness on the rest of the sample space of $p_s(h)$, we conduct simulations. The following process is followed:

1. We select a random sample of the desired size k, from a chosen calendar month s.
2. For each account j in the sample, we compute the non-random elements of the model: the expected drift $\tilde{\mu}_s$ and the distance-to-default. $\delta_{j,s}$.

3. We generate a standard normal random variable ω_s, representing the random error from estimating the credit risk index.
4. For each account j in the sample, we calculate the predicted probability of default:

   $$p_{j,s}(h) = \Phi\left[-\frac{\delta_{j,s}+[\tilde{\mu}_s+v\omega_s]h}{\sigma\sqrt{h}}\right].$$

5. For each account j in the sample, we generate a uniform random variable $u_{j,s}$ and simulate a default indicator:

   $$D_{j,s}(h) = \begin{cases} 0 & if\ u_{j,s} > p_{j,s}(h) \\ 1 & if\ u_{j,s} \leq p_{j,s}(h) \end{cases}.$$

6. We average the simulated default indicators to determine the simulated portfolio default rate:

   $$\tilde{p}_s(h) = \frac{1}{k}\sum_{j=1}^{k} D_{j,s}(h).$$

The simulation process is repeated a number of sufficiently large number of times, to produce a simulated empirical probability distribution function for $p_s(h)$. This distribution function is compared to the LHP approximation $F_p(x,h)$.

The simulation process was repeated 2 500 times, for different sample sizes. Since the LHP assumption ignores sampling error, we expect it to provide a poor fit for small samples where sampling error is important. This is confirmed in the distribution plots shown in Figure 12.

These results suggest that in a small portfolio of large risks the LHP approximation produces a significant risk of under-provisioning. Here, we would be better off adopting a more direct approach (e.g. via simulation) or closer approximations (e.g. Pimbley, 2011).

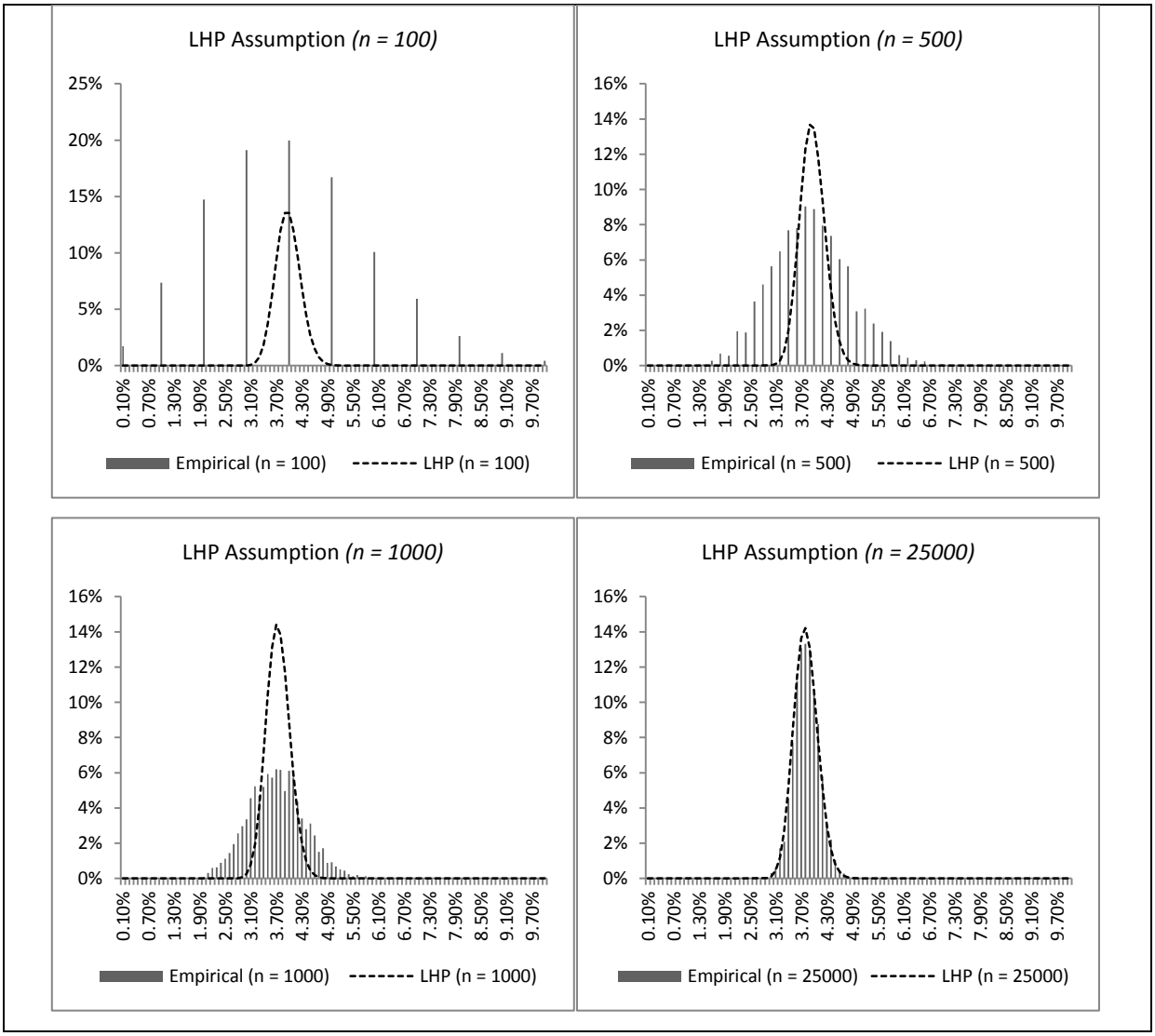


Figure 12: An Assessment of the Large Portfolio Assumption

Homogenous Exposure Assumption

The second assumptions we test are pertain to the idea that all loans are of the same size. Regulatory capital is interested in estimating the value-at-risk of L_s, the random loss arising from accounts in the portfolio observed calendar month s defaulting at time h is. We can write L_s as:

$$L_s = \sum_{j=1}^{n_s} L_{j,s}(h),$$

where $L_{j,s}(h)$ is the random loss arising loan j in calendar month s defaulting at time h:

$$L_{j,s}(h) = E_{j,s}(h) \times LGD_{j,s} \times D_{j,s}(h),$$

$E_{j,s}(h)$ is the exposure after h months of account j observed in calendar month s and $LGD_{j,s}$ is the loss-given default on the account. The portfolio loss ratio under this is given by:

$$r_s(h) = \frac{\sum_{j=1}^{n_s} E_{j,s}(h) \times LGD_{j,s} \times D_{j,s}(h)}{\sum_{j=1}^{n_s} E_{j,s}(h)}.$$

The distribution function of L_s is given by:

$$G(l) = P[r_s(h) \times W_s \leq l]$$

$$= P\left[r_s(h) \leq \frac{l}{W_s}\right],$$

with:

$$W_s = \sum_{j=1}^{n_s} E_{j,s}(h).$$

Thus, the value-at-risk is given by:

$$VaR(\alpha, h) = W_s \times G^{-1}(\alpha).$$

Under the assumption that loans are of the same size (i.e., $E_{j,s}(h) = E_{j,s}$) and have equal and constant loss-given-default (i.e., $LGD_{j,s} = LGD_s$), the portfolio loss ratio becomes:

$$r_s(h) = \frac{E_s(h) \times LGD_s \times \sum_{j=1}^{n_s} D_{j,s}(h)}{n_s \times E_s(h)}$$

$$= LGD_s \tilde{p}_s(h),$$

i.e., the portfolio loss ratio is equal to the portfolio default rate multiplied by the LGD. Therefore, the value-at-risk simplifies to:

$$VaR(\alpha, h) = n_s \times E_s(h) \times LGD_s \times \Phi\left[\frac{\sqrt{\rho(h)}\Phi^{-1}(\alpha) + (\delta_s + \tilde{\mu}_s)}{\sqrt{1-\rho(h)}}\right].$$

Relaxing the assumption of homogenous exposure, the portfolio loss ratio becomes:

$$r_s = \frac{\sum_{j=1}^{n_s} E_{j,s}(h) \times LGD_s \times D_{j,s}(h)}{\sum_{j=1}^{n_s} E_{j,s}(h)}$$

$$= LGD_s \left[\frac{\sum_{j=1}^{n_s} E_{j,s}(h) \times D_{j,s}(h)}{\sum_{j=1}^{n_s} E_{j,s}(h)} \right],$$

i.e., the portfolio loss ratio is equal to the exposure-weighted default rate multiplied by the LGD.

Therefore, to compare the effect of the assumption of constant exposure we compare the simulated empirical distribution of the exposure-weighted portfolio default rate to the LHP approximation of the portfolio default rate. The sample mean of the exposure is ZAR 39 000 and the standard deviation is ZAR 28 000. It was found that the exposure distribution can be reasonably approximated by a Log-Normal distribution. Figure 13 shows the comparison between the LHP approximation and the exposure-weighted portfolio default rate for different assumed standard deviations for the exposure distribution (shown as κ a percentage of the sample standard deviation).

As we would have expected, the larger the variance of the exposure distribution, the less the LHP assumptions are satisfied. Particularly, as the variance increases, the default rate distribution has fatter tails than predicted by the LHP approximation.

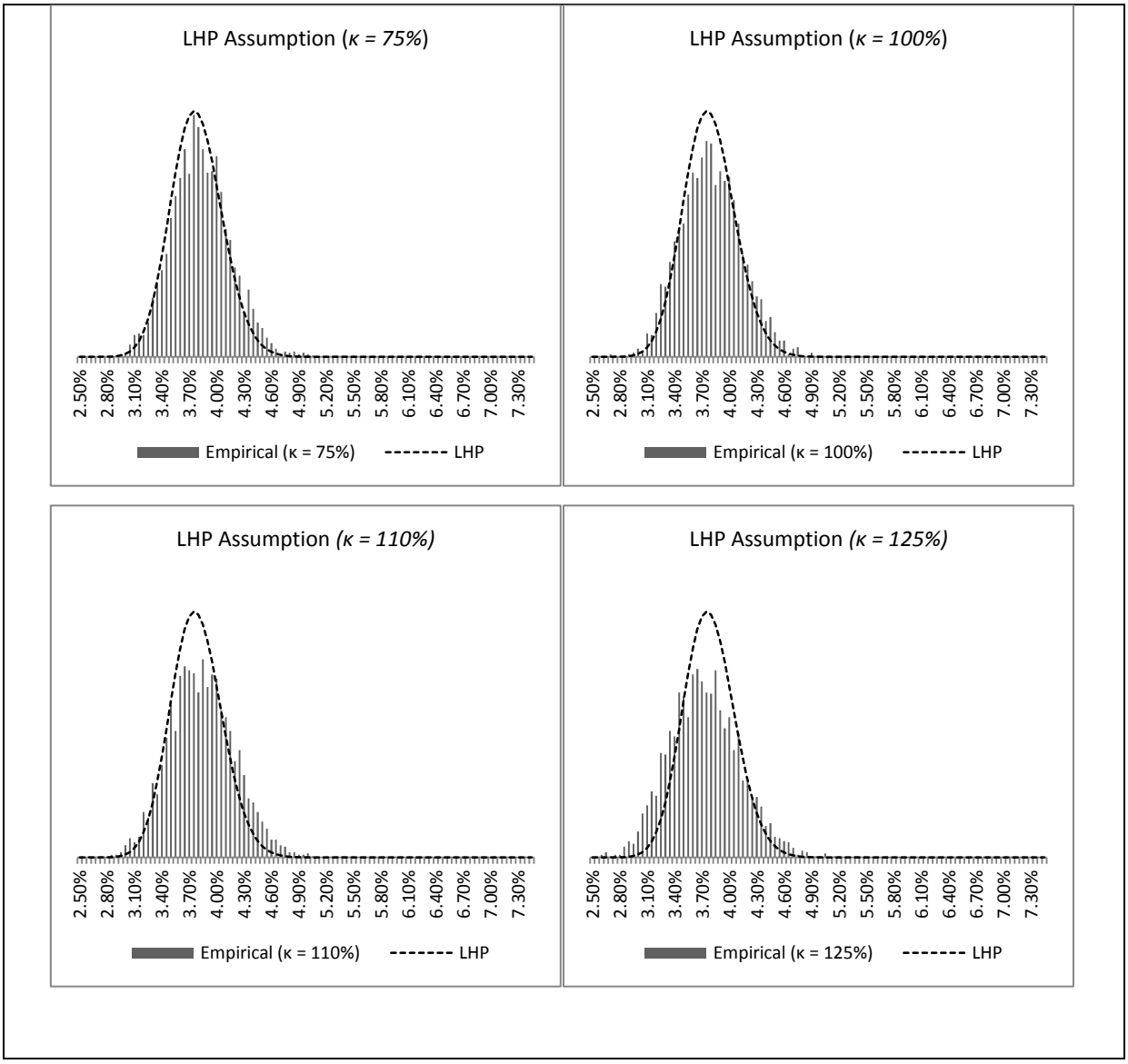


Figure 13: An Assessment of the Uniform Exposure Assumption

Constant LGD Assumption

Finally, we wish to assess the assumption of constant loss-given-default. In the same way that inequality in exposure may lead to poor fit, so can inequality in loss-given-default. We thus assume that all accounts are exposed to the same average loss-given-default of 60%. However, we assume that the loss-given-default follows a beta distribution with a mean of 60%, and allow the variance ς^2 to fluctuate.

We test the impact of the assumption of constant LGD by calculating portfolio loss rate:

$$r_s(h) = \frac{\sum_{j=1}^{n_S} E_{j,s}(h) \times LGD_{j,s} \times D_{j,s}(h)}{\sum_{j=1}^{n_S} E_{j,s}(h)},$$

where $LGD_{j,s}$ is a simulated random variable from the beta distribution. This is compared to the portfolio loss rate under the constant LGD assumption:

$$r_s(h) = LGD_s \left[\frac{\sum_{j=1}^{n_S} E_{j,s}(h) \times D_{j,s}(h)}{\sum_{j=1}^{n_S} E_{j,s}(h)} \right],$$

where $LGD_s = 60\%$.

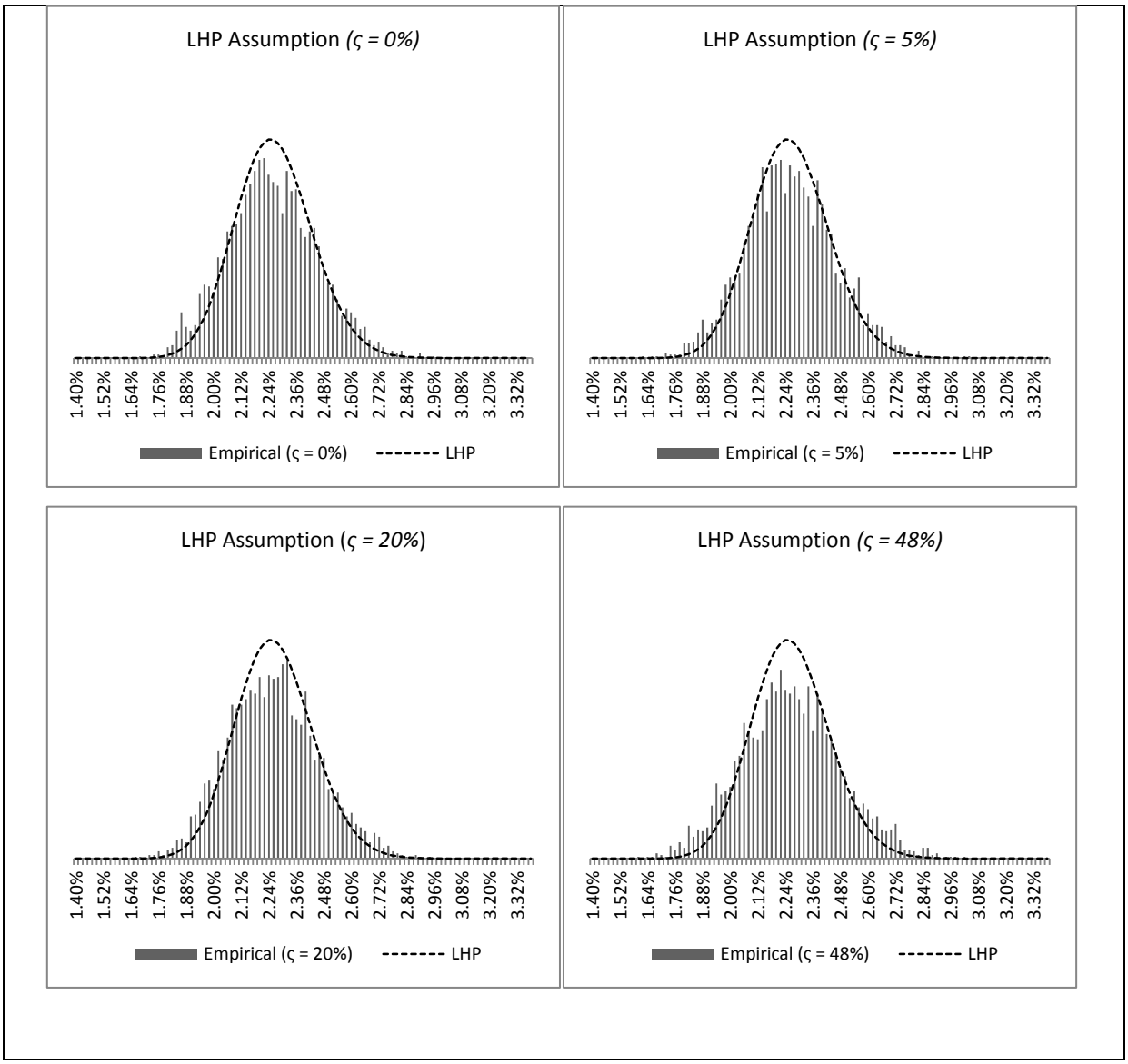


Figure 14: An Assessment of the Constant LGD Assumption

From Figure 14 we see that the portfolio loss rate distribution is not very sensitive to the level of randomness in the LGD. However, as we would expect, a large amount of randomness in the LGD produces slightly fatter tails.

We note one blind spot with our analysis of LGD: it assumes that LGD is not correlated to default rates. In practice we would expect the portfolio PD to correlate strongly with portfolio LGD, since LGD is as much an indicator of credit risk as default rate. This would create a greater effect on the variance of portfolio loss rate.

One approach to address the anticipated correlation between PD and LGD is to explicitly model the correlation. An interesting approach for doing this is provided by Eckert, Jakob and Fischer (2016), where the correlation between individual models for PD, LGD as well as EAD is analysed by assessing the correlation between the model errors.

5.3. Economic Capital under the Log-Log Normal Model

In a similar way to the simulation analysis performed above, we assess the appropriateness of the LHP distributions under the logistic regression models.

Each simulation begins by generating a time series scenario for the credit risk index:

$$C_s = \sum_{j=1}^{q} \hat{\beta}_j Y_j(s - l_j) + \hat{v}\omega_s,$$

where ω_s is a generated standard normal random variable, $\boldsymbol{Y}_s$ is the observed time series vector, $\widehat{\boldsymbol{\beta}} = \{\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_q\}$ is the vector of parameter estimates under the model and $\hat{v}$ is the estimated standard deviation of the random effect. For each account k in the portfolio at each calendar month s, we calculate the predicted default rate as $\ddot{p}_k(s, \omega_s)$.

Default indicators are generated for each account k in the portfolio in calendar month s as Bernoulli random variables as follows:

$$D_k(s) = \begin{cases} 1 & if\ u_{k,s} \leq \ddot{p}_k(s, \omega_s) \\ 0 & if\ u_{k,s} > \ddot{p}_k(s, \omega_s) \end{cases},$$

where $u_{k,s}$ is a $uniform(0,1)$ generated random variable for each account k in the portfolio in calendar month s. The simulated portfolio default rate is then calculated as the mean of the default indicators:

$$\hat{p}(s) = \frac{1}{n_s}\sum_{k=1}^{n_s} D_k(s).$$

Note that the exercise assumes that each model holds when generating each simulation. Therefore, the simulation results can only test the appropriateness of the portfolio distribution approximation (i.e. the LHP approximation), not the validity of the account-level model. The validity of the account-level model was assessed separately in Default Rate Models.

The Figure 15 and Figure 16 show the empirical distributions generated by 25 000 simulation under both the linear-logistic and the log-logistic random effect models, for the month of January 2012. These are plotted on the same axis with the LHP approximation to the distribution.

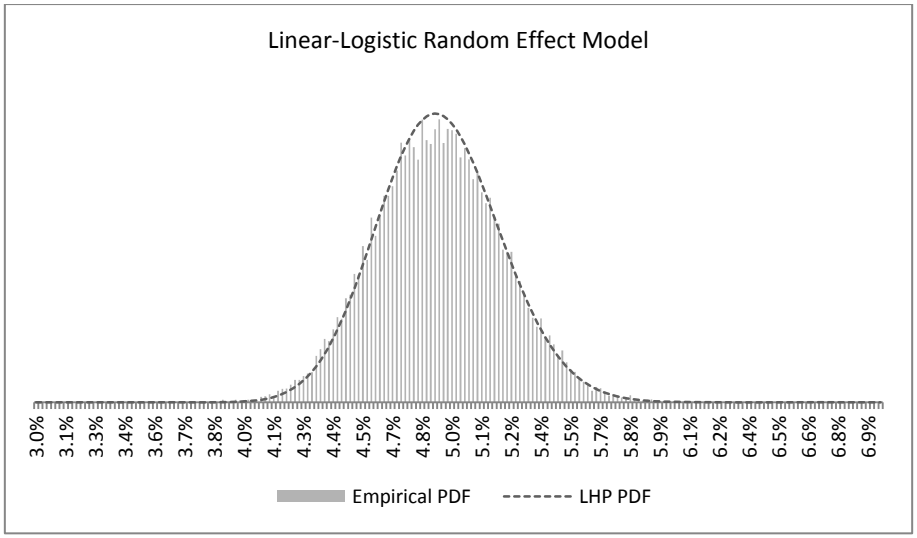


Figure 15: Simulated Portfolio Default Rate Compared to LHP under Linear-Logistic Model

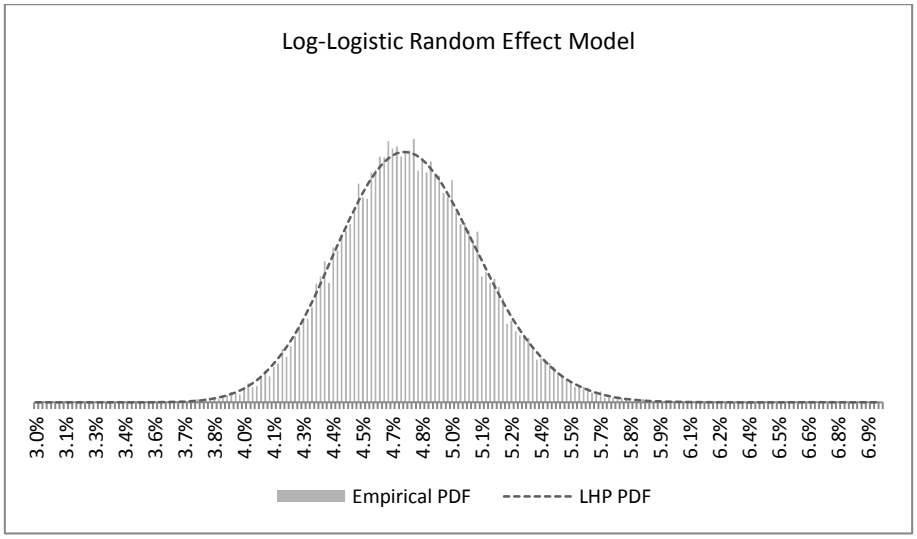


Figure 16: Simulated Portfolio Default Rate Compared to LHP under Log-Logistic Model

The graphs show that the LHP assumption provides a good approximation under both models. We also see that the simulated distribution is fairly symmetric, which means that the Gaussian distribution and binomial distribution may provide a good fit. However, we note that this is a result of

the fact that the macroeconomic variables included in the model provide a close approximation to the credit risk index, i.e., the lower the standard error of the random effect, the lower the correlation between defaults. To illustrate this point, we plot the linear-logistic LHP approximation distribution in the Figure 17 for different values of the standard error.

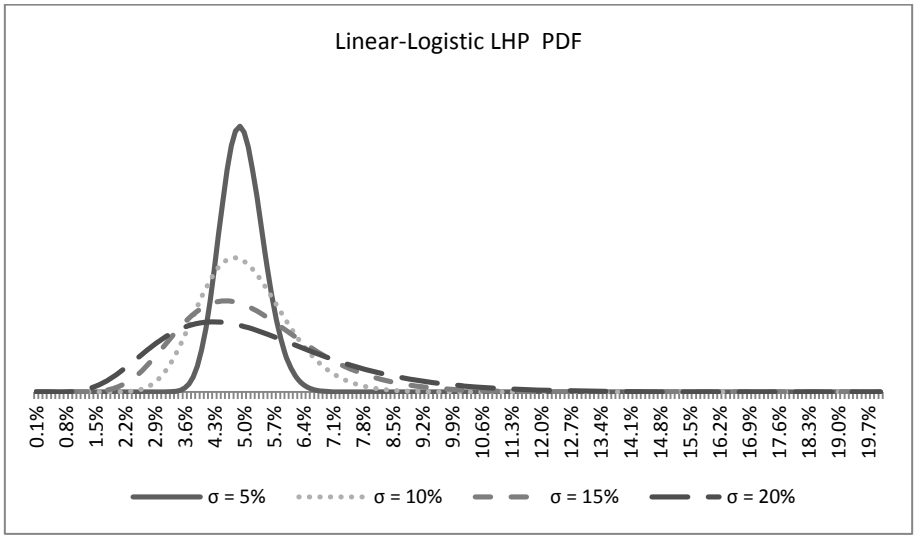


Figure 17: Linear-Logistic LHP for Different Standard Errors

We see from Figure 17 that the symmetry of the LHP distribution is sensitive to the estimated standard error. As the uncertainty around the credit risk index increases, it becomes less plausible to use the Gaussian distribution and binomial distributions as approximations.

Through-the-Cycle LHP Distribution

The portfolio default rate distribution used under Basel II capital requirements is set on a through-the-cycle (TTC) basis, i.e., the expected value of the portfolio default rate distribution is set to reflect the average default rate over an entire credit risk cycle. Therefore, the implied standard error in the Basel II capital requirement (which is represented by the asset

correlation coefficient of the Vašíček distribution) must be large enough to match the fluctuations in default rate across the entire credit risk cycle.

In the models described in this thesis, we can produce a loss distribution that is consistent with the TTC approach by removing macroeconomic factors from the regression modes, while retaining the random effect. The expected value of the portfolio default rate distribution will still fluctuate moderately from one month to the next, as the composition of accounts within the portfolio changes. However, the fluctuations will no longer match those implied by the credit risk index.

The simulated distribution was generated under the TTC linear-logistic model for the month of January 2012. The results are shown in Figure 18.

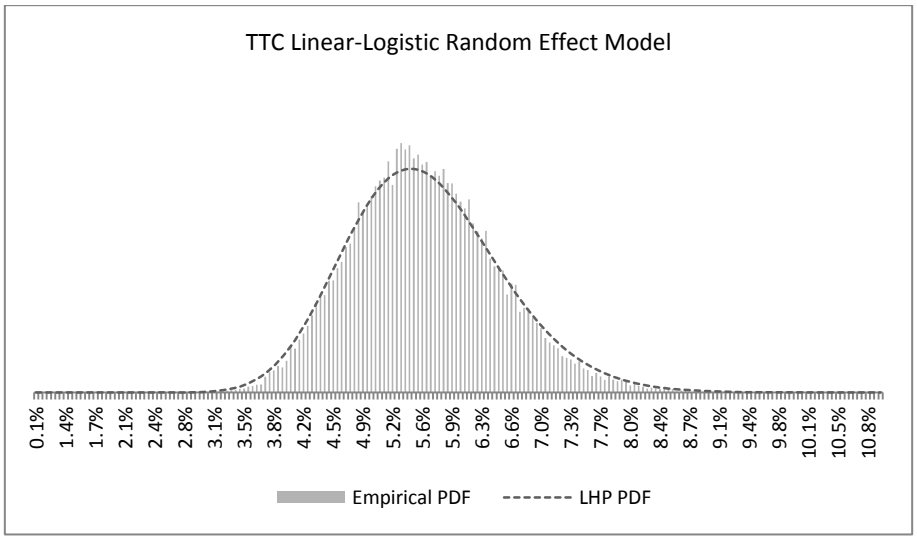


Figure 18: Simulated Portfolio Default Rate Compared to LHP under TTC Linear-Logistic Model

Figure 18 shows that the TTC approach leads to a greater amount of uncertainty around the true values of the portfolio default rate.

Figure 19 below shows the stability of the 95% confidence intervals across time, and how well they contain the observed portfolio default rate, as well as the TTC default rate under the linear-logistic model.

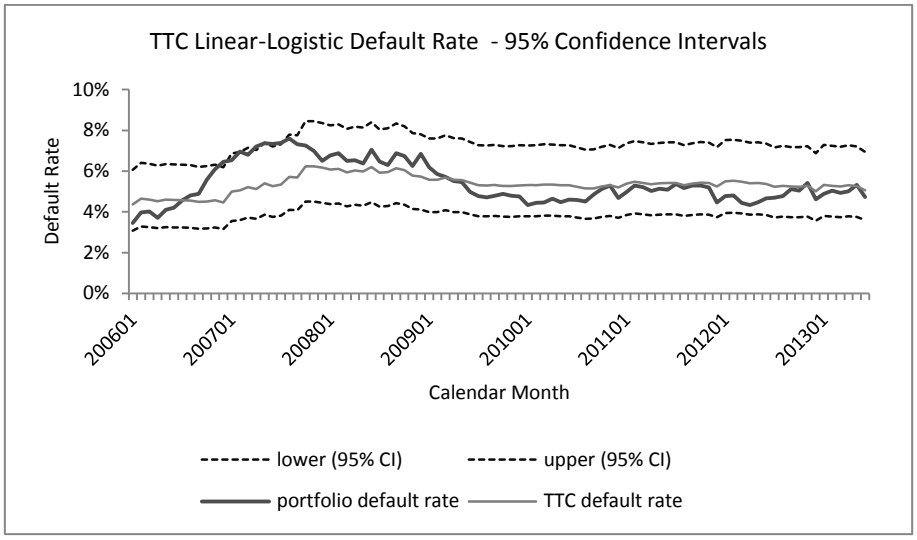


Figure 19: 95% Confidence Interval under TTC Linear-Logistic Model

The graphs show that the confidence intervals under the through-the-cycle models are, indeed, more stable than the ones derived from the models with macroeconomic variables. This would likely lead to more stable capital provisions over time. However, the confidence intervals are less refined in the TTC models. For example, the probability of breaching the confidence intervals over the period between October 2006 and October 2007 is much higher than in the other periods. More generally, the main weakness of the TTC methodology is that it assumes that the credit risk index is cyclical – always being well-confined within a particular range that can be estimated from historic patterns. The inclusion of macroeconomic variables essentially circumvents this assumption and models the credit risk index directly.

The above TTC confidence intervals can be compared to the 95% confidence interval from the Vašíček distribution. The α^{th} of the Vašíček distribution if given by:

$$\Phi\left[\frac{\sqrt{\rho}\Phi^{-1}(\alpha)+\Phi^{-1}(p_{TTC})}{\sqrt{1-\rho}}\right],$$

where p_{TTC} is the through-the-cycle default rate and ρ is the asset correlation coefficient. We use the Basel II requirement's estimate for ρ, which is as follows:

$$\rho = 3\% \times \frac{1-e^{-35p_{TTC}}}{1-e^{-35}} + 16\% \times \left[1 - \frac{1-e^{-35p_{TTC}}}{1-e^{-35}}\right].$$

Notice that the asset correlation coefficient is estimated as a function of default rate, not from data relating to the supposed credit risk cycle. The 95% confidence intervals under the Vašíček distribution are given in the Figure 20 below.

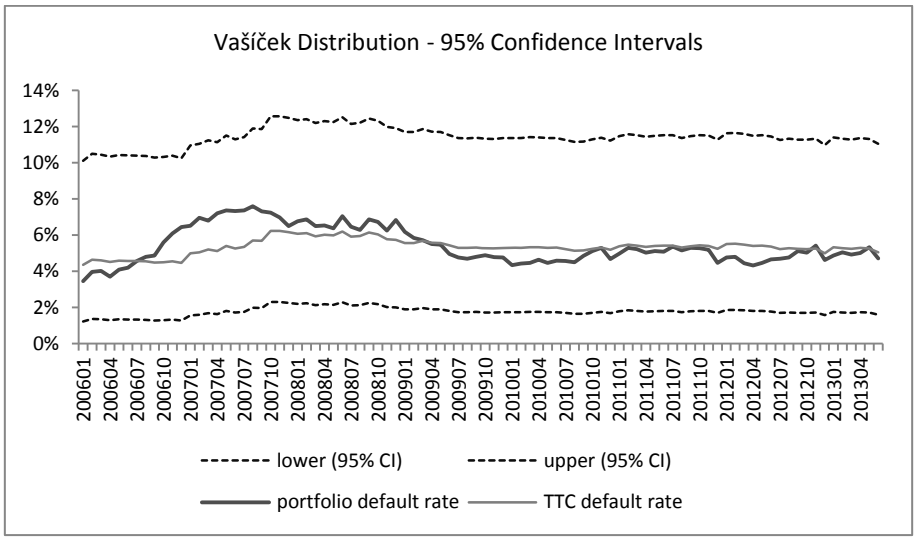


Figure 20: 95% Confidence Interval under TTC Vašíček Distribution

Figure 20 shows that, in this portfolio, the Basel II capital requirements are very conservative. This result coincides with the results found by Crook and Bellotti (2012).

5.4. The Blind Spots of the Basel II Capital Requirement

This chapter has discussed and tested the main assumptions of the Basel II capital requirement. For the portfolio that was analysed, we conclude that the capital requirement is conservative. However, we also note that this is by

coincidence, not by design – the extent to which the requirements are conservative is influenced by:

1. The homogeneity of risk within the portfolio.
2. The homogeneity of exposure within the portfolio.
3. The size of the portfolio.
4. The randomness of LGD.
5. The extent to which LGD is correlated with default rates.

The Basel II capital requirements also assume that the Vašíček distribution is representative of the population being modelled. This is an additional blind spot, as it is possible that the distribution is not appropriate for the portfolio. For example, in the portfolio considered in Chapter 4 we showed that the assumption of a Brownian motion for asset values only fits well when a running-time transformation is applied. Although the linear-logistic model also produces a Vašíček distribution for portfolio losses without making assumptions about evolution of asset values, it makes its own assumptions that need to be tested (e.g. via the Hosmer-Lemeshow test). Thus, regardless of the interpretation assumed, it is important that the assumptions made when deriving the distribution are tested on the portfolio.

We make a final cautionary note that the conservativeness of the Basel II regulatory capital is influenced by the extent to which the credit risk cycle is indeed cyclical. From Figure 20 we see default rates peak below 8%. The existence of a credit risk cycle would imply that, if the period we have analysed contains a full cycle, default rates will always fall below 8%. However, if no such cycle exists, nothing prevents default rates from rising above the 10% mark over a few months. In this regard, a point-in-time approach may be more systematically conservative than a through-the-cycle approach.

5.5. Reporting on Expected and Unexpected Loss

The models discussed in this thesis have two main areas of application in a bank. Firstly, on an account-level, the models can be used to provide estimates of expected loss. The primary use of the expected loss calculation within banking is for published and management accounts, i.e., the estimated expected loss is used for setting impairment provisions. Expected losses can be calculated on an account level as follows:

$$EL_k = PD_k \times EAD_k \times LGD_k,$$

where the PD_k parameter can be estimated from either of the two modelling approaches discussed in the preceding chapters. However, while an approach like this would be suitable under IAS 39, IFRS 9 requires different accounts to be modelled based on different horizons.

IFRS 9 specifies that impairment provisions for accounts that are within the same level of risk as when booked should be based on a 12-month default horizon (these are called Stage 1 Provisions), while provisions for accounts that have increased significantly in risk since origination is to be based on lifetime expected losses (these are called Stage 2 Provisions). This requires a model that can predicted losses over differing time horizons. This creates a limitation for any approach based on logistic regression, since logistic regression requires a fixed outcome period. Survival analysis approaches are more suitable for variable default horizons, which makes the threshold regression approach the more suitable of the two approaches discussed here. Expected losses provisions under IFRS 9 can therefore be calculated as follows:

$$EL_k = PD_k(h_k) \times EAD_k \times LGD_k,$$

where $PD_k(h_k)$ is the probability of default over horizon h_k, and:

$$h_k = \begin{cases} \min(12, \text{Remaining Term}) & \text{for Stage 1 Provisions} \\ \text{Remaining Term} & \text{for Stage 2 Provisions} \end{cases}.$$

The second use of the models described presented in this thesis is for unexpected loss calculation. In Banking, unexpected loss is used for Pillar I regulatory capital calculation, which is described in detail in Chapter 3 (section 3.1.5). It is also used during Pillar II Internal Capital Adequacy Assessment Process (ICAAP), during which a bank determines its own economic capital requirement. Pillar I and Pillar II requirements, along with Pillar III (which focuses on disclosure), are categorisations of the focus areas of Basel II.

For Pillar I purposes, the models described in this thesis can be used to provide inputs into the capital calculation formula, and as a way of validating the assumptions of the capital calculation.

A more important area of application for these models is for Pillar II economic capital calculation. The process for doing this was described earlier in this chapter (sections 5.1 to 5.3). An important contribution of this thesis is that it allows for an independent assessment of capital requirements (an assessment that is independent of the assumptions used in the Pillar I regulatory capital calculation), which speaks to the main aim of the ICAAP. Particularly, the blind spots described in section 5.4 are some of the reasons why the economic capital calculation used in the ICAAP needs to be independent of the regulatory capital assumptions.

Other uses of the models presented are in stress testing and portfolio management. In fact, stress testing is one of the components of the ICAAP. In order to perform a stress test using these models, the economic capital calculation is performed using different macroeconomic inputs (representing different economic scenarios). This results in the so-called stressed-value-at-risk.

Chapter 6: Conclusions

The thesis set out to achieve three goals:

1. The primary aim of the research was to estimate the loss distribution on credit contracts, conditioned on account-level information.
2. The secondary aim was to use the loss distributions to determine the expected and unexpected loss provisions for a book of contracts.
3. The tertiary aim is to determine how the influence of economic conditions can be incorporated into a loss model.

Addressing the First Aim

The first of the listed aims was achieved by building regression models that make use of account-level information to estimate parameters of the distribution of the loss on an individual loan. This was done through logistic regression, to estimate default probability.

Addressing the Second Aim and Third Aims

The macroeconomic inverse Gaussian model and the macroeconomic logistic regression models both provide were discussed as ways to incorporating macroeconomic information in the estimation of default rates. The incorporation of economic information in the process of estimating default rates, and the error with which this is accomplished, were discussed to be the key elements in estimating loss distribution of at portfolio level. In this way, the second and third of the listed aim were addressed concurrently.

In the case when defaults were modelled as being preceded by diminishing savings, which diminish according to a Brownian motion, the portfolio loss distribution was found to be the Vašíček distribution. Since the distribution was originally derived within a corporate credit context, the re-derivation

proposed in this paper offers a more transparent application the distribution in consumer credit risk. The re-derivation also allows gives an estimation approach for the correlation parameter of the distribution, which could be useful to unexpected loss calculation.

A new type of distribution was also derived in the case when the default probabilities are modelled as a logistic regression with a macroeconomic adjustment. The derived loss distribution were compared to the second Basel accord's prescribed loss distribution, where it was found that the process of incorporating macroeconomic factors in to the prediction process reduced the uncertainty around default rates, and thus the level of unexpected loss.

Further Research

The thesis leaves a few areas open for further research. The threshold regression model offered in this paper provides a way of unifying the way in which default rates on corporate bonds and consumer loans are modelled. The fact that this approach allows us to incorporate macroeconomic variables and produce defaults rates for any horizons makes it a good candidate for modelling lifetime probability of default for IFRS 9. However, further work is required on how to include time-varying macroeconomic variables.

The loss aggregation approach described in this paper yielded the log-log normal distribution as a tractable alternative to the Vašíček distribution. The primary difference in the derivation of these two distributions was the dependence structure assumed between default rates and systemic risk. This raises the question of whether there exists other, perhaps more universal, dependence structures that yield tractable loss aggregation formulae. Furthermore, the distribution assumed for systemic risk in deriving both the Vašíček and log-log normal distributions was the normal

distribution. By allowing for distributions other than the normal distribution, we would end up with different aggregate loss distributions.

Finally, the thesis offered a number of assessments and criticisms to the Basel II regulatory requirement. We found that, although the Basel II capital requirements are generally conservative, this may be a coincidence – there are certain situations where we would expect the regulatory capital requirement to be systematically insufficient. Of the identified blind spots of the Basel II capital requirement, two were only discussed briefly. The first of these is the effect of correlation between LGD and default rates on the portfolio loss rate distribution. The second is the impact of the assumption that credit risk is cyclical.

Chapter 7: Appendices

7.1. Logistic Regression Parameter Estimates

7.1.1. Linear-Logistic

Type	Variable 1	Variable 2	Variable 1 Level	Variable 2 Level	Estimate	P-Value
Covariate	intercept				0.254	0.1405
Covariate	pl_mmsinc1plus		1		0.835	<.0001
Covariate	pl_mmsinc1plus		2		0.587	<.0001
Covariate	pl_mmsinc1plus		3		0.372	<.0001
Covariate	paid_down_ratio		1		-0.539	<.0001
Covariate	paid_down_ratio		2		-0.165	<.0001
Covariate	relative_interest_rate		2		0.235	<.0001
Covariate	relative_interest_rate		4		0.393	<.0001
Covariate	fb_hr_indicator		2		1.228	<.0001
Covariate	pl_wpp_1y		1		0.386	<.0001
Interaction	pl_mmsinc1plus	fb_hr_indicator	1	2	-0.925	<.0001
Interaction	pl_mmsinc1plus	fb_hr_indicator	2	2	-0.925	<.0001
Interaction	pl_mmsinc1plus	fb_hr_indicator	3	2	-0.946	<.0001
Interaction	pl_mmsinc1plus	relative_interest_rate	1	2	-0.113	<.0001
Interaction	pl_mmsinc1plus	relative_interest_rate	1	4	-0.178	<.0001
Interaction	pl_mmsinc1plus	relative_interest_rate	2	2	-0.100	<.0001
Interaction	pl_mmsinc1plus	relative_interest_rate	2	4	-0.130	<.0001
Interaction	pl_mmsinc1plus	relative_interest_rate	3	2	-0.119	<.0001
Interaction	pl_mmsinc1plus	relative_interest_rate	3	4	-0.122	<.0001
Interaction	paid_down_ratio	relative_interest_rate	1	2	-0.277	<.0001
Interaction	paid_down_ratio	relative_interest_rate	1	4	-0.190	<.0001
Interaction	paid_down_ratio	relative_interest_rate	2	2	-0.171	<.0001
Interaction	paid_down_ratio	relative_interest_rate	2	4	-0.206	<.0001
Macroeconomic	disposable_income_9				0.009	<.0001
Macroeconomic	emp_compensation_6				-0.044	<.0001
Macroeconomic	hh_consumption_3				0.004	0.004
Macroeconomic	savings_to_income_9				-0.050	<.0001
Macroeconomic Std	sigma				0.001	<.0001

Table 10: Linear-Logistic Model Parameter Estimates

7.1.2. Log-Logistic

Type	Variable 1	Variable 2	Variable 1 Level	Variable 2 Level	Estimate	P-Value
Covariate	intercept				-2.331	<.0001
Covariate	pl_mmsinc1plus		1		0.943	<.0001
Covariate	pl_mmsinc1plus		2		0.664	<.0001
Covariate	pl_mmsinc1plus		3		0.422	<.0001
Covariate	paid_down_ratio		1		-0.607	<.0001
Covariate	paid_down_ratio		2		-0.187	<.0001
Covariate	relative_interest_rate		2		0.267	<.0001
Covariate	relative_interest_rate		4		0.444	<.0001
Covariate	fb_hr_indicator		2		1.378	<.0001
Covariate	pl_wpp_1y		1		0.430	<.0001
Interaction	pl_mmsinc1plus	fb_hr_indicator	1	2	-1.029	<.0001
Interaction	pl_mmsinc1plus	fb_hr_indicator	2	2	-1.031	<.0001
Interaction	pl_mmsinc1plus	fb_hr_indicator	3	2	-1.057	<.0001
Interaction	pl_mmsinc1plus	relative_interest_rate	1	2	-0.119	<.0001
Interaction	pl_mmsinc1plus	relative_interest_rate	1	4	-0.216	<.0001
Interaction	pl_mmsinc1plus	relative_interest_rate	2	2	-0.107	<.0001
Interaction	pl_mmsinc1plus	relative_interest_rate	2	4	-0.157	<.0001
Interaction	pl_mmsinc1plus	relative_interest_rate	3	2	-0.131	<.0001
Interaction	pl_mmsinc1plus	relative_interest_rate	3	4	-0.142	<.0001
Interaction	paid_down_ratio	relative_interest_rate	1	2	-0.326	<.0001
Interaction	paid_down_ratio	relative_interest_rate	1	4	-0.208	<.0001
Interaction	paid_down_ratio	relative_interest_rate	2	2	-0.196	<.0001
Interaction	paid_down_ratio	relative_interest_rate	2	4	-0.227	<.0001
Macroeconomic	disposable_income_9				0.013	<.0001
Macroeconomic	hh_consumption_3				-0.005	<.0001
Macroeconomic	leading_indicator_12				0.002	0.0014
Macroeconomic	prime_12				-0.017	<.0001
Macroeconomic	savings_to_income_9				0.046	<.0001
Macroeconomic Std	std				0.001	<.0001

Table 11: Log-Logistic Model Parameter Estimates

7.2. Derivation: Expectation of the Probit of a Normal Random Variable

Here we provide the proof of the following statement:

$$E_x[\Phi(A+Bx)] = \Phi\left[-\frac{A}{\sqrt{B^2+1}}\right].$$

Consider two independent random variables: $X \sim N\left(-\frac{A}{B}, \frac{1}{B^2}\right)$ and $Y \sim N(0,1)$. We have:

$$Prob[X \leq Y | Y = w] = Prob[X \leq x]$$

$$= \Phi(A+Bx).$$

By the law of total probability, we have:

$$Prob[X \leq Y] = \int_{-\infty}^{\infty} Prob[X \leq Y | Y = x]\phi(x)dx$$

$$= \int_{-\infty}^{\infty} \Phi(A+Bx)\phi(x)dx$$

$$= E_x[\Phi(A+Bx)].$$

However, since $X - Y \sim N\left(-\frac{A}{B}, \frac{1}{B^2}+1\right)$, we also have:

$$Prob[X \leq Y] = Prob[X - Y \leq 0]$$

$$= \Phi\left(\frac{A}{\sqrt{B^2+1}}\right).$$

Therefore:

$$E_x[\Phi(A+Bx)] = \Phi\left[\frac{A}{\sqrt{B^2+1}}\right].$$

In Chapter 3, this was applied to prove that:

$$E_{\varepsilon_s}\Phi\left(\alpha_0 + \sum_{j=1}^{p}\alpha_j X_{k,s,j} + \sum_{j=1}^{q}\beta_j Y_j(s-l_j) + \varepsilon_s\right) = \Phi\left[-\frac{\alpha_0 + \sum_{j=1}^{p}\alpha_j X_{k,s,j} + \sum_{j=1}^{q}\beta_j Y_j(s-l_j)}{\sqrt{v^2+1}}\right].$$

In Chapter 4, this was applied to prove that:

$$E_{\tilde{\varepsilon}_s}\left[\Phi\left(-\frac{\delta_{j,s}+[\tilde{\mu}_s+v\tilde{\varepsilon}_s\,]h}{\sigma\sqrt{h}}\right)\right]=\Phi\left(-\frac{\delta_{j,s}+\tilde{\mu}_s h}{\sqrt{v^2h^2+\sigma h}}\right).$$

7.3. SAS Simulation Code

The figure below provides a SAS macro for performing the simulation described in Chapter 5.

```
%macro distribution (simulations=,sourcefile=);
	data simulations;
		set _null_;
	run;
	%do generation = 1 %to &simulations.;
		/*generate standard RVs*/
		data rann(drop =  j);
			format month yymmn6.;
			do j = 0 to 96;
				month = intnx('month','01jan2006'd,j);
				eta = rannor(0);
				output;
			end;
		run;
		data distribution;
			format eta best8.;
			set &sourcefile.;
			if _n_ = 1 then
				do;
					declare hash raneffect(dataset:'rann');
					raneffect.definekey('month');
					raneffect.definedata('eta');
					raneffect.definedone();
				end;
			rc = raneffect.find();
			/*calculate log-logistic parameter estimates and simulate variables*/
```

```
                    loglogistic = 1-(1-cdf('normal',(delta_log),0,1)**exp(-mu_log + v_log * eta));
                    loglog = rand('binomial',loglogistic,trials);
                    /*calculate linear-logistic parameter estimates and simulate variables*/
                    linlogistic = cdf('normal',(delta_lin + mu_lin + v_lin * eta),0,1);
                    linlog = rand('binomial',linlogistic,trials);
            run;
            proc sql;
                    create table eventrates as
                            select &generation. as generation
                                    ,month
                                    ,sum(loglog) / sum(trials) as loglogistic
                                    ,sum(linlog) / sum(trials) as linlogistic
                            from distribution
                                    group by month;
            quit;
            data simulations;
                    set simulations eventrates;
            run;
    %end;
%mend;
```

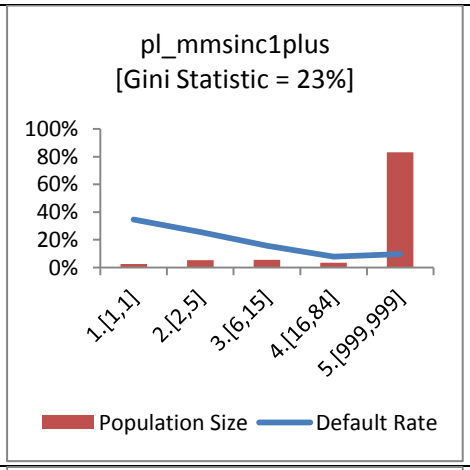
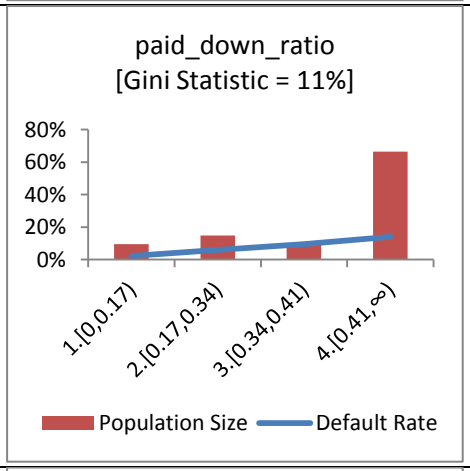
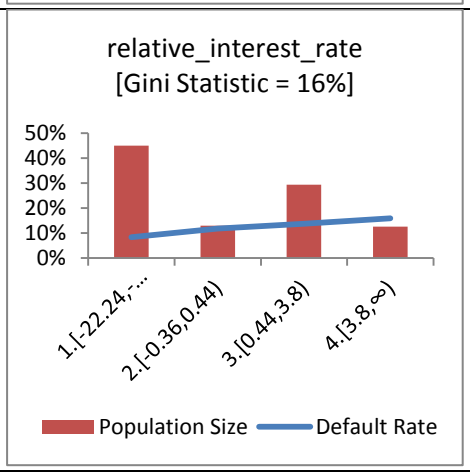
Figure 21: SAS Code for Simulation

In this macro, the *sourcefile*= input should be a SAS dataset with fields:

- *delta_lin* and *delta_log* containing calculated values for $\sum_{j=1}^{p} \hat{\alpha}_j X_{k,s,j}$ under the linear-logistic and log-logistic models, respectively,
- *mu_lin* and *mu_log* containing calculated values for $\sum_{j=1}^{q} \beta_j Y_j(s - l_j)$ under the linear-logistic and log-logistic models, respectively,
- *v_lin* and *v_log* containing the estimates standard error under the linear-logistic and log-logistic models, respectively, and
- *trials* containing the number of observations in a particular group (for grouped data).

7.4. Covariate Descriptions

The tables below provide descriptions of the covariates included in the analysis.

Customer-Level Covariates		
Variable	**Description**	**Grouping**
pl_mmsinc1plus	Number of months since the customer was more than one month in arrears on a personal loan	
paid_down_ratio	Proportion of the original loan that has been paid down	
relative_interest_rate	Interest rate on the loan relative to the average interest rate on the portfolio	

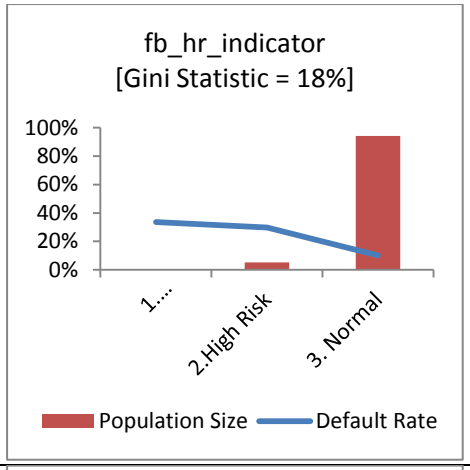
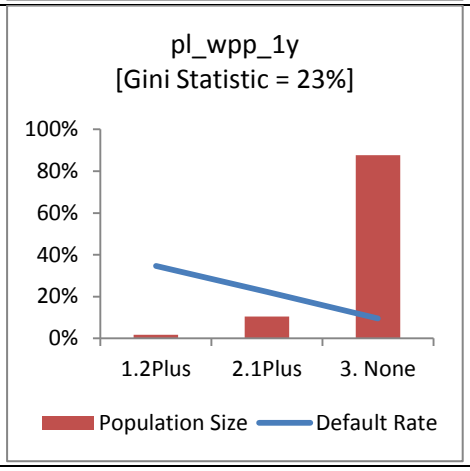
fb_hr_indicator	Indicator of whether the account is in a forbearance programme	
pl_wpp_1y	The customer's worst payment position on a personal loan in the past year	

Table 12: Customer-Level Covariate Details

Macroeconomic Variables					
Variable	**Description**	**Min**	**Mean**	**Std.Dev**	**Max**
debt_to_income_12	Ratio of household debt to household income (lag 12)	52.30	74.03	11.59	88.80
disposable_income_9	The disposable household income (lag 9)	-5.80	3.60	2.99	9.10
emp_compensation_6	The overall employee compensation (lag 6)	47.70	49.68	1.14	52.10
hh_consumption_3	The household consumption (lag 3)	-3.50	3.75	3.38	13.20
leading_indicator_12	The South African Reserve Bank's leading economic indicator (lag 12)	-14.60	2.33	7.29	23.10
prime_12	The Prime Overdraft Rate (lag 12)	8.50	11.52	2.61	17.00
savings_to_income_9	The ratio of household savings to household income (lag 9)	-2.70	-0.75	1.38	2.40

Table 13: Macroeconomic Variable Details

Chapter 8: References

Baesens, B., van Gestel, T., Stepanova, M., van den Poel, D., Vanthienen, J. (2005). *Neural network survival analysis for personal loan data*. Journal of the Operational Research Society. 56: 1089 –1098.

Banachewicz, K., Lucas, A., van der Vaart, A. (2008). *Modelling Portfolio Defaults with Hidden Markov Models with Covariates*. Econometrics Journal. 11: 155-171.

Basel Committee on Banking Supervision [Basel II] (2004). *Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework*. Basel Committee Publications 107: Basel Committee on Banking Supervision.

Belloti, T., Crook, J. (2009). *Credit Scoring with Macroeconomic Variables Using Survival Analysis*. Journal of the Operational Research Society. 60: 1699-1707.

Black, F., Cox, J. (1976). *Valuing Corporate Securities: Some Effects of Bond Indenture Provisions*. Journal of Finance. 31:351-367.

Blaschke, W., Jones, M.T., Majnoni, G., Peria, M. S. (2001). *Stress Testing of Financial Systems: An Overview of Issues, Methodologies and FSAP Experiences*. IMF Working Paper WP/01/88.

Booth, P., Chadburn, R., Haberman, S., James, D., Khorasanee, Z., Plumb, R. H., Rickayzen, B. (2005). *Modern Actuarial Theory and Practice*. CRC Press, 2004

Booth, P., Walsh, D. (1998). *Actuarial Techniques in Risk Pricing and Cashflow Analysis for U.K. Bank Loans*. Journal of Actuarial Practice.

Burnham, K. P., Anderson, D. R. (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach. New York: Springer-Verlag.

Campolongo, F., Jönsson, H., Schoutens, W. (2012). *Quantitative Assessment of Securitisation Deals*. Springer Science & Business Media.

Chan, K. Y., Loh, W. Y. (2004). LOTUS: *An algorithm for building accurate and comprehensible logistic regression trees*. Journal of Computational and Graphical Statistics. 13: 826-852.

Clemen, R. T., Winkler, R. L. (1986). *Combining economic forecasts*. Journal of Business and Economic Statistics. 4: 39-46.

Cox, D.R. (1958). *The Regression Analysis of Binary Sequences (with discussion)*. Journal of the Royal Statistical Society B 20: 215–242.

Crook, J. N., Hamilton, R., Thomas, L. C. (1992). *The degradation of the scorecard over the business cycle*. IMA Journal of Mathematics Applied in Business and Industry. 4: 111–123.

Crook, J., Bellotti, T. (2010). *Time varying and dynamic models for default risk in consumer loans*. Journal of the Royal Statistical Society.173:283-305.

Crook, J., Bellotti, T. (2012). *Asset Correlation for Credit Card Defaults*. Applied Financial Economics. 22: 87-95.

Dhaene, J., Denuit, M., Goovaerts, M.J., Kaas, R., Vyncke, D. (2002) .*The Concept of Comonotonicity in Actuarial Science and Finance: Theory*. Insurance: Mathematics and Economics. 31: 3-33.

Dickey, D. A., Fuller, W. A. (1979). *Distribution of the Estimators for Autoregressive Time Series with a Unit Root*. Journal of the American Statistical Association. 74 (366): 427–431.

Eckert, J., Jakob, K., Fischer, M. (2016). *A Credit Portfolio Framework under Dependent Risk Parameters: Probability of Default, Loss Given Default and Exposure at Default*. Journal of Credit Risk. 12(1): 97-119.

Foglia, A. (2009). *Stress Testing Credit Risk: A Survey of Authorities' Approaches*. International Journal of Central Banking. 5:9-45.

Fok, P-W., Yan, X., Yao, G. (2014). *Analysis of Credit Portfolio Risk Using Hierarchical Multifactor Models*. Journal of Credit Risk, 10(4): 45-70.

Golberg, M. A. (1984). *An Introduction to Probability Theory with Statistical Applications*. Plenum Press, New York.

Hand, D. J. (2005). *Good Practice in Retail Credit Scorecard Assessment*. The Journal of the Operational Research Society. 56: 1109-1117.

Hilhorst, H. J. (2009). *Central Limit Theorems for Correlated Variable: Some Critical Remarks*. Brazilian Journal of Physics. 39: 371-379

Hosmer, D. W., Hosmer, T., Le Cessie, S., Lemeshow,S. (1997). *A Comparison of Goodness of Fit Tests for the Logistic Regression Model*. Statistics in Medicine. 16: 965-980.

Hosmer, D. W., Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley, New York.

Ingolfsson, S., Elvarsson, B. T. (2010). *Cyclical Adjustment of Point-in-Time PD*. Journal of the Operational Research Society. 61: 374 – 380.

International Accounting Standards Board [IFRS 7] (2005). *IFRS 7: Financial Instruments: Disclosures*. London: International Accounting Standards Board.

International Accounting Standards Board [IFRS 9] (2014). *IFRS 9: Financial Instruments*. London: International Accounting Standards Board.

Jennrich, R. I., Sampson, P. F. (1976). *Newton-Raphson and Related Algorithms for Maximum Likelihood Variance Component Estimation*. Technometrics. 18: 11-17.

Kaplan, E. L., Meier, P., (1958). *Nonparametric Estimation from Incomplete Observations*. Journal of the American Statistical Association. 53: 457–481.

Kelly, R. (2011). *The Good, The Bad and The Impaired: A Credit Risk Model of the Irish Mortgage Market*. Central Bank of Ireland Research Technical Paper.

Kumar, A., Rao, V. R., Soni, H. (1995). *An Empirical Comparison of Neural network and Logistic Regression Models*. Marketing Letters. 6: 251-263.

Lee, M. T., Whitmore, G. A. (2006). *Threshold Regression for Survival Analysis: Modeling Event Times by a Stochastic Process Reaching a Boundary*. Statistical Science. 21: 501-513.

Lee, M.-L. T., DeGruttola, V. and Schoenfeld, D. (2000). *A Model for Markers and Latent Health Status*. Journal of the Royal Statistical Society Series on Statistical Methodology. 62:747-762.

Lottes, I. L., Adler, M. A., DeMaris, A. (1996). *Using and Interpreting Logistic Regression: A Guide for Teachers and Students*. Teaching Sociology. 24: 284-298.

Mair, P., Reise, S. P., Bentler, P. M. (2008). *IRT Goodness-of-Fit Using Approaches from Logistic Regression*. Department of Statistics Papers, UCLA.

Malik, M., Thomas, L. C. (2010). *Modelling Credit Risk of Portfolio of Consumer Loans*. The Journal of the Operational Research Society. 61: 411-420.

Mankiw, N. G. (1989). *Real Business Cycles: A New Keynesian Perspective*. Journal of Economic Perspectives. 3(3): 79-90.

Marcucci, J., Quagliariello, M. (2008). *Is Bank Portfolio Riskiness Procyclical? Evidence from Italy using a Vector Autoregression*. Journal of International Financial Markets, Institutions and Money. 18: 46–63.

Matuszyk, A., Mues, C., Thomas, L. C. (2010). *Modelling LGD for Unsecured Personal Loans: Decision Tree Approach*. Journal of the Operational Research Society 61: 393–398.

Merton, R. C. (1974). *On the Pricing of Corporate Debt: The Risk Structure of Interest Rates*. Journal of Finance. 29: 449-470.

Mileris, R. (2012). *Macroeconomic Determinants of Loan Portfolio Credit Risk in Banks*. Inzinerine Ekonomika-Engineering Economics. 23: 496-504.

Nelder, J. A., Mead, R. (1965). *A Simplex Method for Function Minimization*. Computer Journal. 7: 308–313

Philippe, J. (2006). *Value at Risk: The New Benchmark for Managing Financial Risk*. McGraw-Hill

Pimbley, J. (2011). *Portfolio Loss Analysis - Extending the Large Pool Approximation*. Risk Professional. 8: 14 – 22.

Rajaratnam, K., Beling, P., Overstreet, G. (2010). *Scoring Decisions in the Context of Economic Uncertainty*. Journal of Operational Research Society. 61: 421-429.

Roos, C. F. (1955). *Survey of Economic Forecasting Techniques*. Econometrica. 23: 363-395.

Schrödinger, E. (1915*). Zur Theorie der Fall-und Steigversuche an Teilchen mit Brownscher Bewegung*. Physikalische Zeitschrift. 16: 289-295/

Simons, D., Rolwes, F. (2009). *Macroeconomic Default Model and Stress Testing*. International Journal of Central Banking. 5:177-204.

Thomas, L. C., Oliver, R. W., Hand, D. J. (2005). *A Survey of the Issues in Consumer Credit Modelling Research*. The Journal of the Operational Research Society. 56:1006-1015.

Thomas, L.C. (2000). *A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers*. International Journal of Forecasting. 16: 149–172.

van Gestel, T. (2005). *Linear and Non-Linear Credit Scoring by Combining Logistic Regression and Support Vector Machines*. Journal of Credit Risk. 1: 31-60.

Vašíček, O. A. (1987). *Probability of Loss on Loan Portfolio*. KMV Corporation.

Virág, M., Kristóf, T. (2005). *Neural Networks in Bankruptcy Prediction-A Comparative Study on the Basis of the First Hungarian Bankruptcy Model*. Acta Oeconomica. 55: 403-426.

Whitmore, G.A., and Schenkelberg, F. (1997). *Modelling Accelerated Degradation Data Using Wiener Diffusion With A Time Scale Transformation*. Lifetime Data Analysis. 3: 27–45.

Whittaker, J., Whitehead, C., Somers, M. (2007). *A Dynamic Scorecard for Monitoring Baseline Performance with Application to Tracking a Mortgage Portfolio*. Journal of the Operational Research Society. 58: 911–921.