UNIVERSITY OF SOUTHAMPTON

FACULTY OF LAW, ARTS & SOCIAL SCIENCES

School of Management

Modelling examples of Loss Given Default and Probability of Default

by

Jie Zhang

Thesis for the degree of Doctor of Philosophy

January 2011

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF LAW, ARTS AND SOCIAL SCIENCES

SCHOOL OF MANAGEMENT

Doctor of Philosophy

Modelling examples of Loss Given Default and Probability of Default

By Jie Zhang

The Basel II accord regulates risk and capital management requirements to ensure that a bank holds enough capital proportional to the exposed risk of its lending practices. Under the advanced internal ratings based (IRB) approach, Basel II allows banks to develop their own empirical models based on historical data for probability of default (PD), loss given default (LGD) and exposure at default (EAD). This thesis looks at some examples of modelling LGD and PD.

One part of this thesis investigates modelling LGD for unsecured personal loans. LGD is estimated through estimating Recovery Rate (RR, RR=1-LGD). Firstly, the research examines whether it is better to estimate RR or Recovery Amounts. Linear regression and survival analysis models are built and compared when modelling RR and Recovery Amount, so as to predict LGD. Secondly, mixture distribution models are developed based on linear regression and survival analysis approaches. A comparison between single distribution models and mixture distribution models is made and their advantages and disadvantages are discussed.

Thirdly, it is examined whether short-term recovery information is helpful in modelling final RR. It is found that early payment patterns and short-term RR after default are very significant variables in final RR prediction models. Thus, two-stage models are built. In the stage-one model short-term Recoveries are predicted, and then the predicted short-term Recoveries are used in the final RR prediction models. Fourthly, macroeconomic variables are added in both the short-term Recoveries models and final RR models, and the influences of macroeconomic environment on estimating RR are looked at.

The other part of this thesis looks at PD modelling. One area where there is little literature of PD modelling is in invoice discounting, where a bank lends a company a proportion of the amount it has invoiced its customers in exchange for receiving the cash flow from these invoices. Default here means that the invoicing company defaults, at which point the bank cannot collect on the invoices. Like other small firms, the economic conditions affect the default risk of invoicing companies. The aim of this research is to develop estimates of default that incorporate the details of the firm, the current and past position concerning the invoices, and also economic variables.

Contents

List of figures

List of tables

Acknowledgements

First and foremost, I would like to express my deep and sincere gratitude to my supervisor, Professor Lyn C. Thomas. Thanks for his help, guidance and encouragement in the whole course of my PhD study, from the PhD application to the completion of this thesis. His wide knowledge and logical way of thinking have been of great value for me. His humour and jokes have always lighted up the atmosphere of many research meetings and seminars which I attended. It is a great honour for me to become one of his students.

I would like to thank the two anonymous financial institutes which provided data for this study. Without these valuable data, this research cannot be completed.

I am very grateful to all the staff members in the Department of Management Science, Lancaster University Management School. During the MSc course, they taught me the useful knowledge of operational research and management science. These knowledge and skills constructed the solid foundation for my PhD study.

I also should thank School of Management, University of Southampton, which funded me to attend three international conferences, where I presented part of my research work.

Lastly, and most importantly, I have to thank my parents who financially and spiritually support me for all long time. Without their support, it would have been impossible for me to finish this work.

Chapter 1

Introduction

1.1 Basel Accord

Basel, Switzerland, is the home of the Bank for International Settlements (BIS), an international organization that works for cooperation to obtain monetary and financial stability. The Basel Committee on Banking Supervision was established by the Central-Bank Governors of the Group of Ten (G 10) countries in 1974. It provides a forum for regular cooperation on banking supervisory matters and its objective is to enhance understanding of key supervisory issues and improve the quality of banking supervision worldwide. (Gup 2004)

1.1.1 1988 Basel Capital Accord (Basel I)

The Basel Capital Accord (Basel I) was finalised in 1988. Its purpose is to standardize international bank capital regulations in order to maintain the health and stability of the international banking system and create an equally competitive playing field among international banks. The Accord describes a set of rules on the minimum capital levels a bank should hold so as to defend the financial markets from unexpected losses due to financial risks. However, Basel I is a simple 'one-size-fits-all' standard, and it does not distinguish between the risks that banks face and sets the minimum capital requirement to 8% of risk weighted assets. So, Basel I does not distinguish well between commercial loans of different risk degrees. This leads to 'regulatory arbitrage', where the less risky loans are moved off the balance sheet and only the riskier lending retained. Because of this non-sensitivity to risk, the Committee developed a new Basel Accord during the first few years of this millennium. (Gup 2004)

1.1.2 The New Basel Capital Accord (Basel II)

The New Basel Capital Accord was proposed in January 2001, and after a series of consultative papers and surveys, the final accord was published in June 2006:

International Convergence of Capital Measurement and Capital Standards, A Revised Framework, Comprehensive Version.

The major difference between the two capital accords is that Basel II provides for more flexibility and risk sensitivity than Basel I. The new capital adequacy accord is based upon 3 mutually reinforcing pillars depicted in Figure 1.1: Pillar 1, minimum capital requirement; Pillar 2, supervisory review process; and Pillar 3, market discipline and public disclosure. Pillar 1 will be discussed in the following paragraphs. Pillar 2 provides qualitative approaches for the supervisory review of a bank's capital adequacy and internal risk assessment processes. Pillar 3 serves to catalyze prudential risk management by market mechanisms and corresponds to mainly reporting and disclosure (Van Gestel, Baesens 2009).

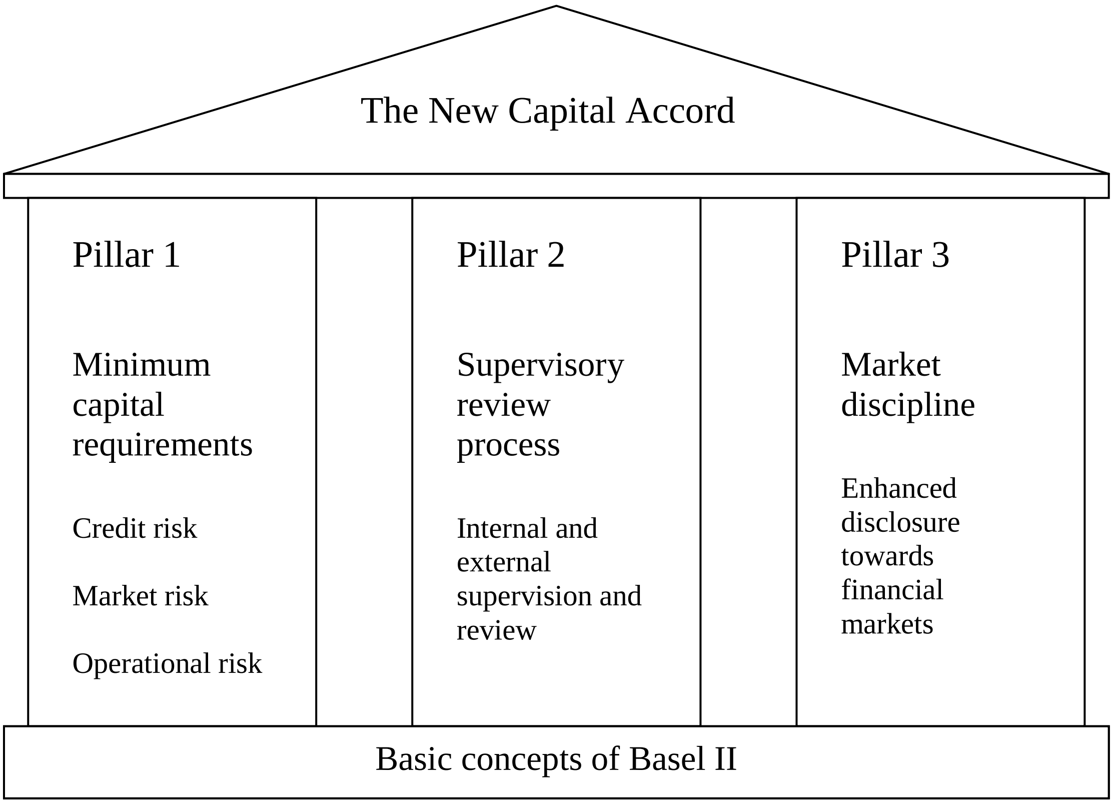
The New Capital Accord
Pillar 1
Minimum capital requirements
Credit risk
Market risk
Operational risk
Pillar 2
Supervisory review process
Internal and external supervision and review
Pillar 3
Market discipline
Enhanced disclosure towards financial markets
Basic concepts of Basel II

Figure 1.1 Basel II framework with 3 mutually reinforcing pillars

Pillar 1 Minimum Capital Requirements

Pillar 1 describes the rules to calculate the minimum regulatory capital standards. It separates the risks involved in lending into credit risk, market risk, and operational risk. Compared to Basel I, Basel II aims to better allocate economic and regulatory capital requirements and reduce incentives for regulatory arbitrage. (Van Gestel, Baesens 2009)

The minimum capital requirement is calculated by Equation (1.1)

$$\frac{Capital}{Risk\,Weighted\,Assets\,of\,(Credit\,Risk + Market\,Risk + Operational\,Risk)} \geq 8\% \quad (1.1)$$

According to Clementi (2000), Basel II focuses on modernising the risk-weighted-assets (RWA) denominator, but with no attention paid to the capital numerator, and the minimum percentage is left unchanged at the level of Basel I.

Market risks arise from on- and off-balance sheet positions due to changes in market prices. Operational risk refers to losses resulting from inadequate or failed internal processes, people and systems, or from external events. Credit risk will be discussed in detail in the following paragraphs.

Credit Risk

Credit risk is the risk of default by a creditor or counterparty. Banks must allocate risk weights to on- and off-balance sheet items that produce a sum of risk-weighted asset values. Credit risk can be measured using 2 approaches: standardized approach, and internal-rating-based (IRB) approach.

Standardized Approach

Standardized approach is a further sophistication of Basel I Capital Accord with a finer classification of the credit risk (Van Gestel, Baesens 2009). The credit quality is measured by External Credit Assessment Institutions, such as Moody’s, Standard & Poor’s, and Fitch. The risk weights for the standardized approach for different asset grades are set and given by the accord. For the retail portfolio, most retail exposures have a weight of 75% (i.e. ¾ of 8%, or 6%), residential mortgages weight at 35%, and 90 days overdue loans weight at 150%.

Internal-Rating-Based (IRB) Approach

The IRB approach is a highly mathematical ‘value at risk’ (VaR) approach, where the risk weights are (partially) derived based on the banks’ own measurements of the risk components. In this approach, the exposures are split into 5 categories: corporate, sovereign, banks, retail, and equity. The treatment for each category may vary.

There are two possible IRB approaches:

Foundation — Banks calculate their own probability of default (PD), and the other parameters are supplied by the regulator.

Advanced — Banks are allowed to provide internal estimates for the probability of default (PD), loss given default (LGD), exposure at default (EAD), and maturity (M).

The expected loss (EL) is calculated by formula EL=PD*LGD*EAD, and it should be covered by returns and provisions of the loans. The unexpected loss (UL) should be covered by regulatory capital, which Basel II defines as

$$EAD \cdot \left(LGD \cdot \left\{ N\left[\left(\frac{1}{1-R} \right)^{1/2} \cdot N^{-1}(PD) + \left(\frac{R}{1-R} \right)^{1/2} \cdot N^{-1}(0.999) \right] - PD \right\} \right) \left(\frac{1+(M-2.5)b}{1-1.5b} \right) \tag{1.2}$$

Where N is cumulative standard normal distribution, N^{-1} is inverse cumulative standard normal distribution and R is asset correlation.

For corporate exposures,

$$b = (0.11852 - 0.05478 \cdot \ln(PD))^2$$

$$R = 0.12 \cdot \left(\frac{1-e^{-50PD}}{1-e^{-50}} \right) + 0.24 \cdot \left(1 - \frac{1-e^{-50PD}}{1-e^{-50}} \right)$$

For retail exposures,
Residential mortgage exposures R=0.15
Qualifying revolving exposures R=0.04

Other retail exposures

$$R = 0.03\left(\frac{1-e^{-35PD}}{1-e^{-35}}\right) + 0.16\cdot\left(1-\frac{1-e^{-35PD}}{1-e^{-35}}\right)$$

There is no maturity adjustment (M=1) for retail exposures, thus the maturity term disappears.

Retail exposures are exposures to individual persons or small businesses. For retail exposures, Banks must provide their own estimates of PD, LGD and EAD, if they adopt IRB approach. Hence, there is no foundation approach for retail. All the PD, LGD and EAD should be estimated on a minimum time period of at least 5 years.

The latest version of Basel Accord is called ‘Basel III’, which is a new update to the Basel Accord that is under development. Basel III is essentially a supplement on top of Basel II in order to promote a more resilient banking sector. It adds extra requirements on the minimum size and form of the capital that a bank must hold, but it does not conflict with Basel II. Thus all the requirements in Basel II still stand.

Models to estimate PD and LGD are now vital for all types of lending, including retail lending. This thesis investigates some of the problems in building such models.

1.2 Financial Crisis of 2007-2009

Following the collapse of the sub-prime mortgage market in the United States the global financial system has undergone a period of unprecedented turmoil. Around US$7 trillion has been evaporated from the global stock markets over the course of 2008. New York’s S&P 500 fell 38.5% in the 12 months by the end of December 2008. Japan’s Nikkei 225 fell 42% during the year of 2008, while in the UK the benchmark FTSE 100 index created the worst performance since its launch 24 years ago, down 31.3% compared with 12 months ago (Adair 2009). Several major financial institutions either failed, were acquired under duress, or were subject to government takeover. These included Lehman Brothers, Merrill Lynch, Fannie Mae, Freddie Mac, Washington Mutual, Wachovia, and AIG (Altman 2009). The International Monetary Fund estimated that large U.S. and European banks lost more than $1 trillion on toxic assets and from bad loans from January 2007 to September 2009. These losses are expected to top $2.8 trillion from 2007 to 2010 (reuters.com 2009).

The immediate cause or trigger of the crisis was the bursting of the United States housing bubble which peaked in 2006. Already-rising default rates on ‘sub-prime’ and adjustable rate mortgages (ARM) began to increase quickly thereafter. Borrowers had been encouraged to assume heavy mortgages by attractive initial terms and in the belief that the housing prices would continuously rise and they would be able to quickly refinance at more favourable terms. Along with the housing and credit booms, the number of financial agreements called mortgage-backed securities (MBS) and collateralized debt obligations (CDO), which derived their value from mortgage payments and housing prices, greatly increased. Such financial innovation enabled institutions and investors around the world to invest in the U.S. housing market. However, interest rates began to rise and housing prices started to drop in 2006–2007, refinancing became more difficult. Defaults activity increased dramatically as easy initial terms expired, home prices failed to go up as anticipated, and ARM interest rates reset higher. Falling prices also resulted in homes worthless than the mortgage loan, providing a financial incentive to enter foreclosure. Thus, major global financial institutions and investors which had invested heavily in sub-prime MBS suffered significant losses. Defaults and losses on other loan types also increased significantly as the crisis expanded from the housing market to other parts of the economy. (Lahart 2007, Bernanke 2009, Krugman 2009)

The Financial Services Authority (FSA) in the UK published a supporting Discussion Paper, the Turner Review (2009), in March 2009. It reviewed the underlying causes of the financial crisis. Ross (2009) summarised these fundamental causes to 5 points:

- *"Significant global macro economic imbalances over the last decade; and in particular the building up of large current account surpluses in Asian and oil exporting countries while there were growing current account deficits in the US, UK and other European countries;*
- *Increasing complexity of the securitised credit model; with lower risk-free interest rates leading to an intense search for higher yield and a rapid growth in the complexity of financial products;*
- *Rapid extension of credit in the US and the UK – in the form of mortgage lending to the household sector. This was accompanied by declining credit standards for both the household and corporate sectors. It also led to property a price boom;*
- *Increasing leverage in the banking and shadow banking system, with large positions in securitised credit and related derivatives increasingly held by banks, near banks and shadow banks;*
- *Underestimation of bank and market liquidity risk making the financial system increasingly reliant on the marketability of their assets."*

Triggered by the burst of housing bubble in the US, these five interrelated factors resulted in severe stresses on the financial system and a number of financial institution failures.

The second point in the above summary mentions about the problems from credit models. The Turner Review (2009) stated that the predominant assumption under the increasing scale and complexity of the securitised credit market was that increased complexity had been matched by the advanced mathematical techniques for measuring and managing the resulting risks. Value-at-Risk (VAR) models, the core of many of the techniques, make predictions about forward-looking risk from the observation of past patterns of price movement. However, there are fundamental questions about the validity of VAR as a measure of risk. (1) Short period of observations in the past. (2) Wrong assumption of normal distributions. (3) Systemic versus idiosyncratic risk. (4) Non-independence of future events.

Some empirical researches also reported model issues in the financial crisis. Murphy (2009) argued that the defaulting mortgages are only a component and symptom of deeper problems in this financial crisis, and the root cause of the crisis was the mispricing in the massive mortgage securitization and credit default swaps (CDS) market. Any investment in a debt requires compensation not only for the time value of money but also a premium for the credit risk of the debt. However, the credit risk premium was largely underestimated before the financial crisis. Rajun et al (2008) made an analysis of the very large forecasting errors that result from the application of sophisticated mathematical models which fit historical data extremely well but ignore human judgement of 'soft' information. Some investors trusted the credit ratings provided by a few rating agencies such as Moody's and Standard & Poors (S&P), which themselves evaluate credit largely using only mathematical models. Those models, which analyse the past relationships between debt defaults and a few variables purely based on statistics, can ignore very important factors and possibilities (Woellert and Kopecki 2008). Thus those models did not perform well over the financial crisis. Luo et al (2009) investigated CDO revaluation in the financial crisis, they analyzed the structural causes of CDO mispricing, and suggested that model mis-specification and data quality can have substantial effects on CDO valuation. They reported the models considering frailty factors are more predictive powerful and accurate than no-frailty models. Demyanyk and Hemert (2008) analysed the quality of subprime mortgage loans, and found the quality of loans deteriorated for six consecutive years before the crisis. The problems could have been detected by models before the crisis, but they were masked by high house price appreciation between 2003 and 2005.

1.3 Scope of the study

With the implementation of Basel II, banks which adhere to advanced Internal-Rating-Based (IRB) approaches need to produce their own models for estimation of Probability of Default (PD) and Loss Given Default (LGD). Therefore, modelling PD and LGD is becoming more important than before. This is the main reason for doing this research. The 2007-2009 financial crisis made a disaster in global financial market, and the huge financial losses led to economic recessions over the period. One of the peculiarities of this financial crisis is that the credit risk models did not work well because they could not respond to the macroeconomic changes. Thus, to avoid the financial crisis happening again in the future, one of the precautions is to build robust credit risk models which can respond to the macroeconomic changes well. This is the second reason for doing this research.

Probability of Default (PD) is the likelihood that a loan will not be repaid and will fall into default. For the corporate default probability estimation, structural approaches and reduced form approaches (the theories will be reviewed in chapter 2) are widely used. For small business and retail default probability estimation, logistic regression is the most common technique for estimating the drivers of default based on a historical data base of defaults. There is no academic literature on invoice discounting default probability estimation; in industries, banks use logistic regression model to make predictions, however, the model in one bank did not perform well during this financial crisis. We try to introduce macroeconomic variables in logistic regression models and also make segmentations to estimate the probability of default for invoice discounters. This will be discussed in Chapter 6 of this thesis.

Loss Given Default (LGD) is the final loss of an account as a percentage of the exposure, given that the account goes into default. Most LGD modelling research has concentrated on corporate lending where LGD was needed as part of the bond pricing formulae. On consumer side, modelling PD has been the objective of credit scoring systems for fifty years but modelling LGD is not something that had really been addressed until the advent of the New Basel regulations. Now with the financial crisis in mortgage lending, there is also a practical need for lenders to be able to forecast the losses in their defaulted loans, so as to set aside the appropriate level of provisions.

Modelling LGD appears to be more difficult than modelling PD for two reasons. Firstly, much of the data may be censored (debts still being paid) because of the long time scale of recovery. Second, debtors have different reasons for defaulting and these lead to different repayment patterns. This thesis makes a study of modelling LGD for unsecured personal loans. In Chapter 3, we use survival analysis to model LGD, because survival analysis can handle censored data, and segment the whole default population and build mixture distribution models for modelling LGD. Comparisons are made between survival analysis models and linear regression model, and between mixture distribution models and single distribution models. In Chapter 4, we use payment-patterns variables before and shortly after default and short-term recovery rate (RR) variables in LGD prediction models, and see whether these variables help estimate LGD. In Chapter 5, we consider how to bring macroeconomic variables into LGD prediction models, and examine the influence of macroeconomic environment on debt losses.

Chapter 2

Literature Review

In the first section of this chapter we will review the literature of LGD modelling in corporate sector. In the second section the literature of LGD modelling in consumer sector will be reviewed, although not much work has been done on LGD modelling for consumers. Section three will talk about the PD modelling approaches for corporate lending, and section four will review the classification techniques for modelling PD for consumer lending, which are usually described as credit scoring. In section five, we will review the application of survival analysis in credit scoring, and this approach will be used to model LGD for consumer loans in chapter 3. In section six, the literature on invoicing discounting and factoring will be reviewed. Not much modelling research has been done, but some empirical studies have been made in this area.

2.1 LGD for the corporate sector

This section will review the literature of LGD modelling in corporate sector. Firstly, the theoretical analysis will be briefly discussed. Secondly, some empirical studied will be reviewed. Lastly, two commercial LGD models will be described.

2.1.1 Theoretical Contributions to LGD

There are two main credit risk models: (i) structural approach models and (ii) reduced form models; they treat the recovery rate (RR) differently. In structural approaches (proposed by Merton 1974), the default process of a firm is driven by the value of the firm's assets, and the default occurs when the market value of a firm's assets is lower than that of its liabilities. Therefore, the creditors' payoff at the maturity of the debt is smaller of the two amounts: the face value of the debt or the market value of the firm's assets. Under these models, RR at default is a function of the structural characteristics of the firm: asset volatility and leverage. The payment to the debt holders is a function of the residual value of the defaulted firm's assets; therefore the RR is an endogenous variable. (See Altman et al 2005)

Reduced form models (Jarrow and Turnbull 1995, Duffie 1998, Jarrow et al 1997) do not condition default on the value of the firm, and PD and RR are modelled independently from the structural features of the firm, its asset volatility and leverage. They assume an exogenous RR that is independent of the PD, and both PD and RR vary stochastically through time. The RR is parameterised differently in Reduced from models. Jarrow and Turnbull (1995) assume that a bond at default would have a market value equal to an exogenously specified fraction of an otherwise equivalent default-free bond. Duffie (1998) assumes that bonds of the same issuer, seniority and face value have the same RR at default, regardless of the remaining maturity. Jarrow et al (1997) allow for different debt seniorities to translate into different RRs for a given firm.

2.1.2 LGD in Empirical Studies

Schuermann (2005) summarise 5 characteristics of LGD: (i) Recovery distributions are bimodal. RR is either high or quite low, but lower recoveries are more common than higher recoveries. (ii) Seniority and collateral matter. RR of loans is higher than bonds because loans are typically senior to bonds; recoveries for senior secured debts are higher than unsecured debts. (iii) Recoveries vary across the business cycle. There is strong evidence that recoveries in recessions are lower than during expansions. (iv) The impact of industry. For corporate bonds, Altman and Kishore (1996) find evidence that some industries such as utilities do better than others. (v) Size probably doesn't matter. Although size is an important determinant in models of default, it seems to have no strong effect on debts recoveries.

Bank loan RRs have been analysed by Asnarnow and Edwards (1995) and by Eales and Bosworth (1998). They use US data and focus on small business loans and larger consumer loans. Dermine and Neto de Carvalho (2003) analyse the determinants of LGD rates using Portuguese data of 371 defaulted loans to SMEs. For corporate bonds, Altman and Fanjul (2004) breakdown bond recoveries by original rating of different seniorities. Altman and Kishore (1996) and Verde (2003) concentrate on RRs across different industries. Renault and Scaillet (2004) and Friedman and Sandow (2003) try to estimate the entire RR probability distribution rather than focusing on its expected value.

2.1.3 Commercial LGD Models

Moody's KMV model, LossCalc

The most widely used model is the Moody's KMV model, LossCalc (Gupton 2005). It is developed based on 3026 global observations of LGD for defaulted loans, bonds and preferred stock occurring between January 1981 and December 2003. The dataset includes 1424 defaulted public and private firms in all industries. The RRs are based on secondary market pricing of default debt as quoted one month after the date of default. They find RRs are not normally distributed, and are well described by Beta distribution. Using a

Beta transformation, the RRs are converted to be Normally distributed. They consider using 9 predictive factors which are in 5 categories to explain RRs. These 5 categories are (i) Collateral and other support. (ii) Debt type and seniority classes. (iii) Firm-level information. (iv) Industry. (v) Macroeconomic and geographic. Then, the regression techniques are used to regress the transformed RRs on the factors mentioned above but without an intercept term. The final step is to apply inverse Beta transformations for predicted values and get the predicted RRs.

LossCalc is validated in out-of-time and out-of-sample tests. It includes time-varying factors (updated firm-level information, macroeconomics), which are a Basel requirement, and uses LGD histories that are longer than seven years that Basel requires in all of the countries / regions covered. Thus, LossCalc is viewed as a robust and validated global model of LGD.

Standard & Poor's Recovery Ratings

Another popular model, Recovery Ratings, is created by Standard & Poor's Ratings Services (Chew and Kerr 2005). Its analytical process is comprised of a few steps: review transaction structure and borrower's projections, establish simulated path to default, forecast borrower's free cashflow at default, determine valuation, identify priority debt claims and value, determine collateral value available to lenders, and finally, based on the analysis above, it classifies the loans into 6 classes which cover different recovery ranges. It remains early days for recovery ratings, and it is in the process of being further developed and improved.

2.2 LGD for the consumer sector

2.2.1 Approaches from Basel Accord

Approaches from corporate LGD modelling are not appropriate for consumer credit LGD modelling since there is no continuous pricing of the debt as is the case on the bond market. The Basel Accord (BCBS 2004 paragraph 465) suggests using implied historic LGD as one approach in determining LGD for retail portfolios. One difficulty with this approach is that it is accounting losses that are often recorded and not the actual economic losses. The alternative method suggested in the Basel Accord is to model the collections or work out process. Such data is used by Dermine and Neto de Carvalho (Dermine and Neto de Carvalho 2006) for bank loans to small and medium sized firms in Portugal, because small firms are considered as the retail portfolio by Basel. They make an empirical RR study based on univariate mortality analysis and use a multivariate approach to analyse the determinants of RRs and a log-log form of the regression to estimate LGD.

2.2.2 LGD modelling for secured loans

Calem and LaCour-Little (2004) look at estimating both default probability and recovery of mortgage loans. They estimate recovery by employing spline regression to accommodate the non-linear relationships that are observed between both loan-to-value ratios and recovery, which achieves an R-square of 0.25. Lucas (2006) suggests the idea of using the collection process to model LGD for mortgage loans. The collection process is split into whether the property was repossessed and the loss if there was repossession. So a scorecard is built to estimate the probability of repossession where Loan to Value is the key and then a model used to estimate the percentage of the estimated sale value of the house that was actually realised at sale time. Somers and Whittaker (2007) propose the use of quantile regression in the estimation of predicted discount (Haircut) in sale price observed in the case of repossessed properties. For mortgage loans, a one-stage model is built by Qi and Yang (2009). They model LGD directly using characteristics of defaulted loans, and find LTV (Loan to Value) is the key variable in the model and achieve an adjusted R square of 0.610, but only a value of 0.15 without including LTV. Leow et al (2009) add some other variables besides LTV, and find the model is significantly improved by adding other variables. They also compare a two-stage model with a one-stage model, and conclude the two-stage model is superior to the one-stage model.

2.2.3 LGD modelling for unsecured loans and credit cards

For unsecured consumer credit, the only approach is to model the collections process, and now there is no security to be repossessed. The difficulty in such modelling is that the Loss Given Default, or the equivalent Recovery Rate, depends both on the ability and the willingness of the borrower to repay, and on decisions by the lender on how vigorously to pursue the debt. This is identified at a macro level by Matuszyk et al (2010), who use a decision tree to model whether the lender will collect in house, use an agent on a percentage commission or sell off the debts, - each action putting different limits on the possible LGD. If one concentrates only on one mode of recovery, in house collection for example, it is still very difficult to get good estimates. Matuszyk et al (2010) look at a few types of regression models including Box-Cox transformation, OLS regression, Beta transformation, Log normal transformation, WOE approach, and find WOE approach achieves the highest R-square of 0.23. Bellotti and Crook (2008) also look at various versions of regression techniques and conclude the OLS regression achieves the lowest Mean Square Error (MSE) and Least Absolute Value regression model based on a fractional logit transformation of RR gives least Mean Absolute Error (MAE). Bellotti and Crook (2009) add economic variables to the OLS regression model and find unemployment rate and interest rate influence RR and models including these two factors are improved, but in all cases the results in terms of R-square are poor - between 0.05 and 0.2.

Querci (2005) investigates geographic location, loan type, workout process length and borrower characteristics for loans to small businesses and individuals from an Italian bank, but concludes none of them is able to explain LGD though borrower characteristics are the most effective.

2.3 PD Models for the Corporate Sector

In the corporate sector, there are generally two main approaches to the modelling of credit risk: structural approach models (also known as Merton models, Merton 1974) and reduced form models (Artzner and Delbaen 1995 and Jarrow and Turnbull1995). The structural approach, which is based on Black-Scholes option pricing (Black and Scholes 1973), models the economic process of default, whereas reduced-form models decompose risky debt prices in order to estimate the random intensity process underlying default. Besides, Accounting based models, which are based on the financial ratios from annual accounts, have been looked at by some researchers, and it is used in credit risk modelling for small and medium sized enterprises (SMEs) in recent years (Altman and Sabato 2007). There are also some other models, such as VAR approach models and insurance approach models, and they will be reviewed in this section.

2.3.1 Structural approach models

Structural approach was first proposed by Robert Merton (1974). He exploited and extended the options models of Black and Scholes (1973). Merton’s model of risky debt starts with a set of assumptions that allow the modeller to view equity as an option on the assets of the company. From this insight, the value of debt can be derived. The major work within the structural approach models is the modelling of the evolution of the firm’s value and of the firm’s capital structure.

For the case of a single bond of face value (D) maturing at the time (T), Merton's approach assumes default at time T in the event that $A_t \leq D$. This model treats the process A, the market value of the firm's assets, as a log-normal diffusion, which allows the firm's equity to be priced with the Black-Scholes formula as though it is a call option on the total asset value A of the firm, struck at the face value of debt. The value of the debt is then simply obtained by subtracting this equity option price from the initial asset value.

The associated model of the default probability is illustrated in Figure 2.1 (Rikkers 2006), where the total value of assets A is approximated as the sum of the market value of equity and the book value of liabilities. Looking forward from "now", the default probability is obtained from the probability distribution of asset values at the maturity date T.

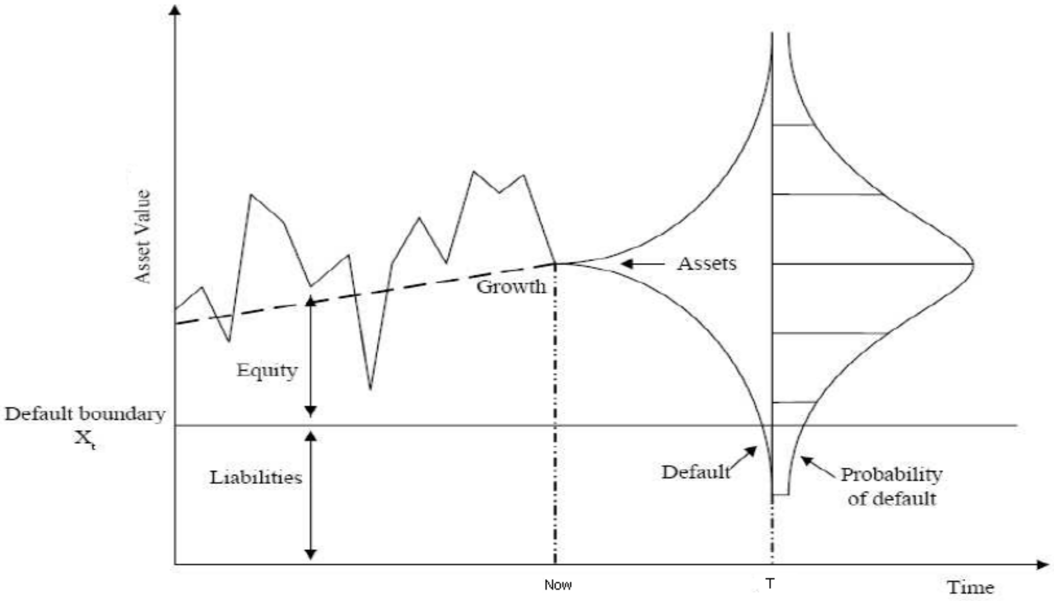

Figure 2.1 Explanation of Merton type model

The Moody's - KMV Approach

One implementation of the Merton approach is Moody's KMV model, which uses it to estimate Distance to Default (DD). This is then mapped onto the probability of default that is Expected Default Frequency (EDF). As outlined in Crosbie and Bohn (2002), the Moody's-KMV approach consists of four steps. (i) Estimate asset value and volatility. (ii) Calculate a "Default Boundary". (iii) Calculate the Distance to Default (DD). (iv) Map DD into Default Probability (PD). The correlations in default between the different loans in a portfolio are calculated by using Monte Carlo or multi-step simulations.

The primary advantage of structural models is that they utilize stock price data that is predictive and highly responsive to changes in the firm's financial condition. The disadvantage is their reliance on distributional assumptions (i.e., normality) that imply default probabilities that sometimes are not true. (Saunders and Allen 2002)

2.3.2 Reduced form approaches

Reduced form models go back to Artzner and Delbaen (1995) and Jarrow and Turnbull (1995). The dynamics of the intensity are specified under the market-implied probability. It is not interested in why the firm defaults but interested in when the firm defaults, the intensity model is calibrated from market prices.

The simplest version of intensity default models defines default as the first arrival time τ of a Poisson process with some constant mean arrival rate, called intensity, often denoted λ. With this:

The probability of survival is $p = e^{-\lambda}$, meaning that the time to default is exponentially distributed.

The expected time to default is $1/\lambda$.

The probability of default over a time period of length Δ, given survival to the beginning of this period, is approximately $\Delta\lambda$, for small Δ.

Once the default event actually occurs, the intensity of course drops to zero. When we speak of an intensity λ, we mean the intensity prior to default. It is normally implausible to assume that the default intensity λ is constant over time. If we use $\lambda(t)$ to describe the intensity at time t, the probability of survival for t years is

$$p(t) = e^{-\int_0^t \lambda(t)dt}.$$

Any given non-negative process λ can be used to parameterise the dynamics of default. No economic model of firm default is needed for this purpose any more. There are no formal commercial models exactly based on reduced form approaches, but there are two models often viewed as in this branch, and we define them as Markov Chain approach models.

2.3.3 Markov Chain approach models

Markov Chain approaches look at changes in bond prices to give view of underlying changes in PD. KPMG’s Loan Analysis System (LAS) is an extension of this approach. It uses a net present value (NPV) approach to credit risk pricing that evaluates the loan’s structure. A lattice or “tree” analysis is used to evaluate the impact of revaluations on credit risk pricing. The loan’s value is computed for all possible transitions through various states, from credit upgrades and prepayments, to restructurings, to default.

Using bond prices is a problem because it depends both on PD and LGD. One needs to separate them. KAMAKURA’s Risk Manager (KRM) does it by modelling both debt and equity prices, since PD and LGD appear in different ways in the two sets of prices. It decomposes credit spreads into PD and LGD by the use of both debt and equity prices in order to better separate the default intensity process from the loss recovery process. The default hazard rate is modelled as a function of stochastic default-free interest rates, liquidity factors, and lognormal risk factors, such as a stochastic process for the market index. (For details see Saunders and Allen 2002)

These two models are referred to as mark-to-market (MTM) models, which calculate value at risk (VAR) based on the change in the market value of loans. They do not concentrate on predicting default losses. Since markov chain approach models are purely empirical, they cannot be evaluated by interpreting their economic assumptions and implications. The primary advantages of markov chain models over structural models are their relative ease of computation and their better fit to observed credit spread data.

2.3.3 Accounting based models

In the case of the Accounting based models, the initial work uses a univariate model to predict business failures using a set of financial ratios (Beaver 1967). In his model, a dichotomous classification test to determine the error rates a potential creditor would experience if firms are classified on the basis of individual financial ratios as failed or non-failed. Six financial ratios from among original 30 ratios are selected as best indicators of performance. These are cash flow to total debt, net income to total assets, total debt to total asset, working capital to total assets, current ratio, and no-credit interval.

Altman (1968) uses a multiple discriminant analysis technique (MDA) to solve the inconsistency problem linked to Beaver’s univariate analysis and to assess a more complete financial profile of firms. Altman examines twenty-two financial ratios, eventually selecting five as providing in combination the best overall prediction of corporate bankruptcy, thus developed Z-Score model.

$$Z = 0.012 \cdot X1 + 0.014 \cdot X2 + 0.033 \cdot X3 + 0.006 \cdot X4 + 0.999 \cdot X5$$

where X1 = Working capital / Total assets

X2 = Retained Earnings / Total assets

X3 = Earnings before interest and taxes / Total assets

X4 = Market value equity / Book value of total debt

X5 = Sales / Total assets

These five financial ratios reflect five financial aspects of the firm, which are liquidity, profitability, leverage, solvency and capital turnover.

Ohlson(1980), for the first time, applies the logistic regression model to the default prediction's study. The practical benefits of the logit methodology are that it does not require the restrictive assumptions of MDA and is less sensitive to extreme values. He bases the analysis on nine predictors which reflect four characters of the firm; they are size of the company, measure of financial structure, measure of performance, and measure of current liquidity.

After the work of Ohlson (1980), lots of researchers use logit models to predict default. Casey and Bartczak (1985) investigate the use of operating cash flows as possible predictor of bankruptcy. Gentry et al (1985) use a cash-based funds flow model to classify bankruptcy and non-bankruptcy. Aziz et al (1988) also make a study of cash flow based models for bankruptcy prediction. Becchetti and Sierra (2002) find some non-balance sheet variables have some predictive power on the probability of firm failure. Keasey and Watson (1987) investigate whether a model utilising a number of 'non-financial' variables is able to predict small company failure more accurately than models based solely upon financial ratios. Mossman et al (1998) make a comparison of four types of bankruptcy prediction models, which are based on financial statement ratios, cash flows, stock returns, and return standard deviations. Shumway (2001) develops a hazard model for forecasting bankruptcy, where he finds several market-driven variables are strongly related to bankruptcy probability. He et al (2005) re-estimate Ohlson's model (1980) and Shumway's model (2001), and observe that

Shumway’s model performs marginally better than Ohlson’s model. Lin et al (2007a) uses logistic regression to predict default of small businesses using different definitions of financial distress. Lin et al (2007b) compare Merton models and logistic regression models on modelling default of small business under different circumstances. Altman and Sabato (2007) compare a set of credit risk models for small and medium sized enterprises (SMEs), and conclude the logistic regression models are better than the generic corporate model (known as Z’’-Score, developed by Altman (2005)) and MDA model. Altman et al (2009) find that some qualitative data make a significant contribution to increasing the default prediction power of risk models built specifically for SMEs.

2.3.4 VAR approach models

Value at Risk (VAR) models seek to measure the minimum loss of value on a given asset or liability over a given time period at a given confidence level. The typical model of VAR approach is CreditMetrics, which was first introduced in 1997 by J.P. Morgan and its co-sponsors. CreditMetrics seeks to answer the question: If next year is a bad year, how much will I lose on my loans and loan portfolio?

CreditMetrics tries to use available data on a borrower’s credit rating, the probability that rating will change over the next year, recovery rates on defaulted loans, and credit spreads and yields in the bond or loan market, to estimate the market value (P) and the volatility or standard deviation of that market value (σ), then the VAR can be directly calculated. (For details see Saunders and Allen 2002)

However, CreditMetrics VAR calculations assume that transition probabilities are stable across borrower types and across the business cycle. This assumption of stability is problematic. There is empirical evidence that default rates are sensitive to the state of the business cycle and rating transitions may depend on the state of the economy [see Wilson (1997a,b) and Nickell, Perraudin, and Varotto (2001)]. One way to build in business cycle effects and take a forward-looking view of VAR is to model macroeconomic effects on the probability of default and associated rating transitions. CreditPortfolio View Model, which was produced by McKinsey in 1997, uses macro simulation approach to overcome some of the biases resulting from assuming static or stationary probabilities period to period. (For details see Saunders and Allen 2002)

CreditMetrics involves a full valuation in which both an upgrade and a downgrade rating to loan values are considered, thus it is a MTM model which calculates VAR based on the change in the market value of loans. CreditPortfolio View can be used as either an MTM or a DM (default mode) model, because it can allow for credit upgrades and downgrades as well as defaults in calculating loan value losses and gains and hence capital reserves, and it also can consider only two states of the world: default and non-default.

2.3.5 Insurance Approach

Credit Suisse Financial Products (CSFP) developed a model, Credit Risk Plus, similar to the one a property insurer selling household fire insurance might use when assessing the risk of policy losses in setting premiums. Because of default rate uncertainty and severity of the losses uncertainty, Credit Risk Plus rounds and bands loss severities or loan exposures, and produces a distribution of losses for each exposure band. Summing these losses across exposure bands produces a distribution of losses for the portfolio of loans. (For details see Saunders and Allen 2002)

Credit Risk Plus is different from CreditMetrics in the objectives and the theoretical foundations. Credit Risk Plus only considers two states of the world – default and non-default – and the focus is on measuring expected and unexpected losses rather than expected and unexpected changes in value as under CreditMetrics. Thus, Credit Risk Plus is a default mode (DM) model and it can only work at portfolio level while other models can work at individual loan levels.

2.3.6 Summary of commercial models

	Moody's KMV	LAS/ KAMAKURA	Credit Metrics	Credit Portfolio View	Credit Risk Plus
Produced by	KMV Moody's	KPMG/ KRM	JP Morgan	McKinsey	Credit Suisse
Definition of risk	DM	MTM	MTM	MTM or DM	DM
Risk drivers	Asset values	Debt and equity prices	Asset values	Macroeconomic factor	Expected default rate
Risk Measured	Default Loss	Default Loss	Change in Market value	Change in Market value	Default Loss
Events modelled	Defaults	Defaults	Defaults + Migration	Defaults + Migration	Defaults
Numerical approach	Analytic and simulation	Econometric	Simulation	Simulation	Analytic

Table 2.1 Summary of commercial models

Table 2.1 (based on Saunders and Allen 2002) summarises the similarities and differences of commercial models based on different approaches which we have discussed earlier. In that discussion we concentrated on the risk events estimated and the approaches used rather than the risk drivers. No commercial model is based on accounting approach, thus this approach is not listed in Table 2.1.

2.4 PD models for the consumer sector

The initial credit scoring approach is Discriminant Analysis (DA) which was proposed by Fisher (1936), this approach could be viewed as a form of linear regression (Thomas, et al 2002), and was ever the most popular statistical method. Afterwards, logistic regression became the most common statistical method, because it needs less restrictive assumptions than DA. Classification tree is an alternative statistical approach for credit scoring. Also, there are some non-statistical approaches from artificial intelligence or operational research, such as neural networks, linear programming, genetic algorithms, nearest neighbours. Although they are not wildly used in practice, sometimes they have good performance in a specific task. The classification techniques in credit scoring were reviewed by Rosenberg and Gleit (1994), Hand and Henley (1997), Thomas (2000), and Hand (2001). This section briefly reviews these techniques.

2.4.1 Discriminant Analysis

Discriminant analysis (DA) was introduced by Fisher (1936) to differentiate between different types of irises. Basic idea is using some classification tool to minimise the distance between cases within a group, and maximise the differences between cases in different groups.

Let $Y = \omega_1 X_1 + \omega_2 X_2 + \ldots + \omega_p X_p$ be any linear combination of the characteristics. One measure of separation is how different are the mean values of Y for the two different groups of goods and bads in the sample. Thus one looks at the difference between $E(Y|G)$ and $E(Y|B)$ and choose the weights ω_i with $\sum_i \omega_i = 1$, which maximize this difference. (For details, see Thomas et al 2002)

There are two basic assumptions behind DA: one is the independent variables included in the model are multivariate normally distributed; another is the group variance-covariance matrices are equal across the good and bad groups. However, there are some arguments about these assumptions, some people thought they were critical (Eisenbeis 1977, 1978, Rosenberg and Gleit 1994), and some people thought they did not have much influence (Reichert et al 1983, Hand et al 1998).The first published work of using discriminant analysis to produce a scoring system seems to be that of Durand (1941) who uses the method to make predictions of credit repayment. Grablowsky and Talley (1981) compare linear discriminant analysis and probit analysis by using data from a large Midwestern retail chain in the USA. Other work of the use of discriminant analysis in credit scoring is given by Lane (1972), Apilado et al (1974) and Moses and Liao (1987).

2.4.2 Linear Regression

In credit scoring, or any instance where there is a binary outcome, linear regression is also referred to as linear probability modelling (Anderson 2007). The end result is an estimate of p (Good), the formula for which is

$$p(Good) = \beta_0 + \sum_j \beta_j x_j + e$$

The probability for each record is the sum of a constant and the products of a series of weights β_j and variable values x_j, where the variables take on different values for each record and weights differ for each variable j (the error term e is ignored).

There are also some assumptions behind Linear Regression. But in most cases, these assumptions do not hold. The most problematic are ‘normally distributed error terms’ and ‘homoscedasticity’, because the target variable only has two possible values, 0 and 1, also the predicted values often fall outside the 0 to 1 range. Orgler (1970) uses regression analysis in a model for commercial loans. Orgler (1971) uses regression analysis to construct a scorecard for evaluating outstanding loans. Other studies of using regression include Fitzpatrick (1976), Lucas (1992), and Henley (1995).

2.4.3 Logistic Regression

The most common method for building credit scorecards today is logistic regression (Thomas, et al 2002), and the first to publish credit scoring results from logistic regression model is Wiginton (1980). A logistic regression model is simply one where the explanatory variables time their coefficients are assumed to be linearly related to the natural log of the odds that default will happen. That is

$$Ln(\frac{p}{1-p}) = \beta_0 + \sum_j \beta_j x_j + e$$

Where, p is the probability that default will not occur and $\frac{p}{1-p}$ is the odds that default will not occur.

Logistic regression is designed specifically for the case where the dependent variable is binary. The logistic model prevents the predicted probabilities greater than 1 or less than 0 by working with the odds of the event happening instead of the probability. The assumptions behind the Logistic regression are less than other statistic methods and can be held in most of credit scoring cases. Logistic regression’s primary disadvantage is its computational intensiveness, but improvements in computers have made this less of an issue.

2.4.4 Classification Trees

Classification trees, sometimes called recursive partitioning algorithms (RPA), classify the consumers into groups, each group being homogeneous in its default risk and as different from the default risks of other groups as is possible (Thomas, L. C. 2009). The tree is started by choosing one variable and splitting the attributes of the variable into two subsets in order to maximize the difference in the default risk between the two subsets as large as possible according to the splitting rule. This process is repeated on each of the two subsets created in turn until the new subsets are too small or the difference between two subsets is not statistically significant based on the stopping rule. Almost always this results in a tree which overfits the data in that it performs far less well on other samples. Thus, it is usual to use a second sample to check whether the splits suggested are also significant on that set. If not, the original tree is pruned back until it has splits that meet the splitting criteria on both samples. Then according to the assignment rule, each end node is classified as good or bad. Figure 2.2 is an example of classification tree, there it uses three variables to classify the whole sample into 4 groups, two of which are good customers and two are bad customers.

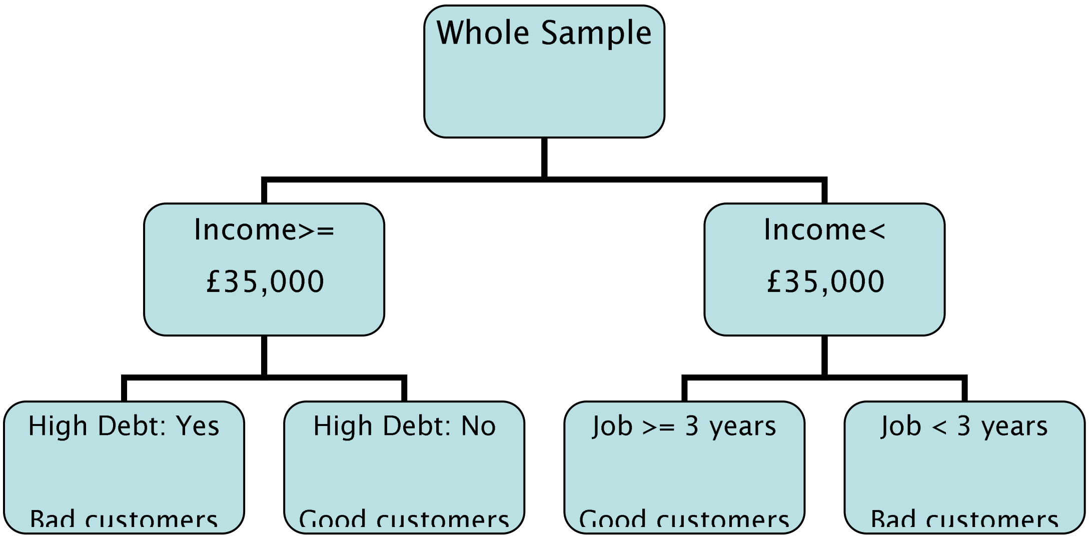

Figure 2.2 An example of classification tree

The tree approach was firstly developed in statistics by Breiman et al (1984). Makowski (1985) was one of the first to use classification trees in credit scoring. After that, the applications of such methods in credit scoring were looked at by Coffman (1986), Carter and Catlett (1987).

Classification trees have some advantages compared to other techniques: this approach is non-parametric, and well suited to categorical analysis. Its main strength is its ability to identify patterns, find and exploit interactions. The results are very transparent and easy to implement. However, there are also some disadvantages: as trees become bushier, there are fewer cases in each node, bringing with it the potential for overfitting and unreliable results. Classification trees are best used for quick and dirty data exploration, gaining insight into data, identifying key predictive variables, or acting as a benchmark for other models (Anderson 2007).

2.4.5 Neural Networks

Neural networks (NNs) can be described as networks of computing elements that can respond to inputs, and learn to adapt to the environment (Anderson 2007). A neural network consists of a number of inputs (variables), each of which is multiplied by a weight, and the products are summed and transformed in a 'neuron' and the result becomes an input value for another neuron. The end result is similar to a decision tree, but the details are much finer and decision rules are very complex. Most credit applications of neural networks have been to scoring corporations (Altman et al 1994, Rosenberg and Gleit 1994, Tam and Kiang 1992), also this approach was used in scoring consumers (Desai et al 1996, 1997).

There are some advantages of NNs. It can process huge amounts of data, discover and track interactions in the data, deal with non-linear relationships within the data. There are also some disadvantages: they are computation-intensive, needing long time to train a model. They are expensive to implement and maintain. They are 'black-box' approach; the models are difficult to interpret. The NNs are not suitable for any practice where the decision logic must be interpreted. However, they are well suited where accurate and adaptive predictions are critical, and transparency is not so important (Anderson 2007).

2.4.6 Linear Programming

Linear programming (LP) was proposed for classification by Freed and Glover (1981), and gives rise to an additive scorecard. It is not based on the probabilities of belonging to any one of the classes being classified but simply on how wrong is the scorecard when it assigns a data point incorrectly. If an applicant is wrongly classified, the misclassification error is how far the wrong way from the cut-off value. Linear programming seeks to minimize the sum of such misclassification errors. Assume there are *n* applicants in the data set, the applicant *i* has characteristics ($x_{i1}, x_{i2}, \cdots, x_{ip}$). For ease of notation let us assume that the first n_G applicants are the goods and the remaining applicants from $n_G + 1$ to $n_G + n_B$ are the bads. The linear programming can be expressed as:

Minimize $a_1 + a_2 + \cdots + a_{n_G+n_B}$

Subject to $\omega_1 x_{i1} + \omega_2 x_{i2} + \cdots + \omega_p x_{ip} \geq c - a_i$, $1 \leq i \leq n_G$,

$\omega_1 x_{i1} + \omega_2 x_{i2} + \cdots + \omega_p x_{ip} \leq c + a_i$, $n_G + 1 \leq i \leq n$,

$a_i \geq 0$, $1 \leq i \leq n$,

$\omega_1, \omega_2, \cdots, \omega_p$ are weights for each characteristics x, and c is cut-off score. For the goods, the first constraint says that their score should be above the cut-off score, and there is only an error a_i if their score is below the cut-off. For bads, the second constraint says that the score should be below the cut-off score, and there is only an error a_i if the score is above the cut-off score. All the errors a_i are positive, and objective is to find the weights $\omega_1, \omega_2, \cdots, \omega_p$ that minimize the sum of the errors.

Joachimsthaler and Stam (1990) review the substantial literature on this approach for classification problems. Hardy and Adrian (1985) and Nath et al (1992) compare LP with other statistical approaches for classification problems.

There are two significant advantages of the linear programming approach. One is that it deals with large number of variables very well. The second is that if one wants to ensure certain relationships between the attribute scores, this is very easy to do by adding extra constraints in the linear programme which ensure these requirements between the coefficients are met (Thomas 2009). However, this approach is computationally intensive, and the statistical significance of the point allocations cannot be tested. While it is technically possible to use this technique for credit scoring, it is seldom used in practice (Anderson 2007).

2.4.7 Other approaches

Expert system is a kind of system which incorporates human experts' learning into a set of rules, some of which are trained using an inference engine from data input to the system. Some studies using expert systems for credit risk analysis were done by Zocco (1985), Davis (1987), Davis et al (1992), Leonard (1993a, 1993b). One of the attractive features of expert systems is their ability to explain their recommendations and decisions, thus quite applicable under the legal requirements for banks to give reasons for rejecting applicant.

Genetic algorithms are heuristic search algorithms, which try to find an optimal result within a search space through survival-of-fittest evolution (Fractal Analytics 2003). In the credit scoring context one has a number of scorecards which mutate and bend together according to their fitness at classification (Thomas 2000). Fogarty and Ireson (1993) and Albright (1994) were some of the first to apply this approach in credit scoring. Desai et al (1997) and Yobas et al (1997) compare it with other classification techniques.

Nearest neighbours technique is used to determine group membership by finding cases within a set of training data whose predictors are most similar to a new case for which group membership is not known. Chatterjee and Barcun (1970) were one of the first to suggest this approach to credit scoring. Henley and Hand (1996) make a detailed investigation of nearest neighbour methods applied to data from a large mail order company and suggest the system has the advantage that new data points can be added easily and the system be updated with no change to the underlying coding.

2.5 Survival Analysis

Survival analysis is a relatively new application that offers an advantage of predicting time to the event of interest and therefore, lays the foundation for estimating the applicant's profitability (Banasik et al, 1999; Stepanova and Thomas, 2001). The theoretical detail of survival analysis will be reviewed in Chapter 3. Narain (1992) is one of the first authors who investigate survival analysis to credit scoring. Narain applies one type of proportional hazard approach to loan data and shows that it gives a reasonable approximation to the time until default. Banasik et al (1999) compare performance of exponential, Weibull and Cox's nonparametric models with logistic regression and find that survival-analysis methods are competitive with, and sometimes superior to, the traditional logistic-regression approach. Kelly and Hand (1999) describe the use of survival analysis in building models which can allow for uncertainties in the definitions of what is meant by 'good' and 'bad'. Hand and Kelly (2001) use survival analysis models to predict default probability of a newly launched credit product, where no historical data are available for standard scorecard built. Stepanova and Thomas (2001) develop techniques based on Cox's proportional hazards model incorporating behavioural information to develop survival analysis approaches to behavioural scoring. Stepanova and Thomas (2002) look at three extensions of Cox's proportional hazards model applied to personal loan data. A number of diagnostic tests to check adequacy of the model fit show the models fit well. Sohn and Shin (2006) investigate a reject inference method based on the confidence interval of a median survival time to delayed repayment. Andreeva (2006) explores the application of survival analysis models to the data of revolving credit from three European countries. Survival analysis national and generic models are produced and their predictive quality is very close to logistic regression. Andreeva et al (2007) investigate the relationship between present value of net revenue from a revolving credit account and times to default and to second purchase. Bellotti and Crook (2009) use survival analysis models including macroeconomic variables to predict default of credit cards. The study shows that inclusion of macroeconomic variables improves model fit and affects PD yielding a modest improvement in predictions of default. Banasik and Crook (2010) consider the application of augmentation to profit scoring applicants by means of survival analysis.

2.6 Invoice Discounting and Factoring

Invoice discounting and factoring are two forms of short-term financing, often used to improve a company's working capital and cash flow position. Invoice discounting allows a business to draw money against its sales invoices before the customer has actually paid. To do this, the business borrows a percentage of the value of its sales ledger from a finance company, effectively using the unpaid sales invoices as collateral for the borrowing. Factoring is a financial transaction whereby a business sells its accounts receivable (i.e., invoices) to a third party (called a factor) at a discount in exchange for immediate money with which to finance continued business (Wikipedia). The main difference between these two forms is that factoring is the sale of receivables whereas invoice discounting is borrowing where the receivable is used as collateral, thus factoring involves three parties: invoice seller, factor and debtor; and invoice discounting involves only two parties: invoicing company and finance company.

There is very little literature on invoice discounting and factoring, especially on invoicing discounting, we can not find any academic papers. Some qualitative research was done on factoring. Mian and Smith (1992) theoretically argue that there is a potential to develop a robust theory regarding receivable management policy in business finance. Smith and

Schnucker (1994) examine organizational structure, where the economics of the factoring decision is evaluated. They claim that economies of scale have an impact on the decision to integrate. Summers and Wilson (2000) find evidence of a ‘financing demand’ explanation for the use of factoring, and they argue that the motivation to use factoring is more related to a demand for asset-based finance from small companies than to firm-level choice about organizational structure. Soufani (2000) makes an interview based survey to 21 factoring companies and uses the information to evaluate the profile of businesses using factoring and the extent to which the provision of invoice financing services is focused upon particular groups of firms as delineated by characteristics such as sector, size, age and type of ownership. Soufani (2002) makes a survey of 3805 SMEs and builds a logistic regression model to test the hypotheses he makes about the motivation for using factoring and the type of business choosing it in terms of their demographic characteristics, and also whether firms’ financial distress, relationship with their banks, size and value of the collateral, total value of the firms debt have an effect on their choice of using factoring. No credit risk models were reported in literature for invoice discounting and factoring.

Chapter 3

Comparisons of linear regression and survival analysis in modelling LGD[1]

3.1 Introduction

Modelling PD, the probability of default has been the objective of credit scoring systems for fifty years but modelling LGD is not something that had really been addressed in consumer credit until the advent of the new Basel regulations. Modelling LGD appears to be more difficult than modelling PD, because of two reasons. Firstly, much of the data may be censored (debts still being paid) because of the long time scale of recovery. Linear regression does not deal that well with censored data. Second, debtors have different reasons for defaulting and these lead to different repayment patterns. For example, some people do not want to repay; some people can not repay because of permanent changes in their situation, while for others the reason for non repayment is temporary. One distribution may find it hard to model the outcomes of these different reasons. Survival analysis though can handle censored data, and segmenting the whole default population is helpful to modelling LGD for defaulters with different reasons for defaulting.

[1] A paper based on this work has been accepted by *International Journal of Forecasting*, and will be published soon.

In this chapter, we use linear regression and survival analysis models to build predictive models for recovery rate and recovery amount, and hence LGD. Both single distribution and mixture distribution models are built to allow a comparison between them. This analysis will give an indication of how important it is to use models – survival analysis based ones – which cope with censored debts and also whether mixed distribution models give better predictions than single distribution model.

The comparison will be made based on a case study involving data from an in house collections process for personal loans. In section two, the data will be described. In section three we briefly review the theory of linear regression and survival analysis models, and build and compare single distribution models using linear regression and survival analysis based models. In section four we explain the idea of mixture distribution models as they are applied in this problem, and create mixture distribution models, so that comparisons can be made between single distribution approach models and mixture distribution approach models. In section five, the conclusions obtained will be summarised.

3.2 Data

The data in this research is data on defaulted personal loans from a major UK bank. The debts occurred between 1987 and 1999, and the repayment pattern was recorded until the end of 2003. In total 27278 debts were recorded in the data set, of which, 20.1% debts were paid off before the end of 2003, 14% debts were still being paid, and 65.9% debts were written off beforehand. The range of the debt amount was from £500 to £16,000; 78% of debts are less than or equal to £5,000 and only 3.6% of them are greater than £8,000. Loans for multiples of thousands of pound are most frequent, especially 1000, 2000, 3000 and 5000. Twenty one characteristics about the loan and the borrower were available in the data set such as the ratio of the loan to income, employment status, age, time with bank, loan purpose and term of loan.

The recovery amount is calculated as:

default amount – last outstanding balance (for non-write off loans)

OR default amount – write off amount (for write off loans)

The distribution of recovery amount is given in Figure 3.1, ignoring debts that are still being repaid but this graph could be misleading as it does not describe the original debt.

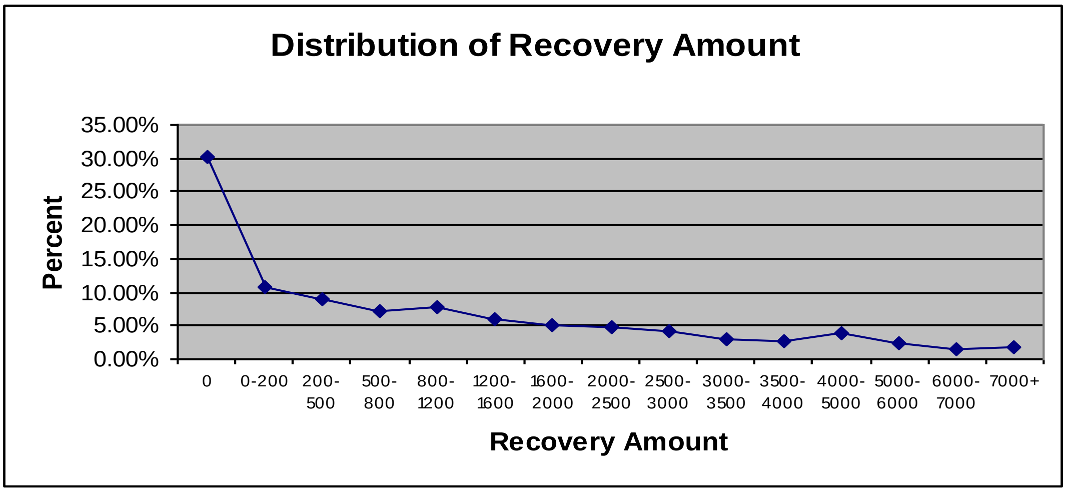

Figure 3.1 Distribution of recovery amount in the data set

$$\text{The recovery rate} \quad \frac{\text{Recovery Amount}}{\text{Default Amount}}$$

is more useful as it describes what percentage of the debt is recovered. The average recovery rate in this data set is 0.42 (not including debts still being paid). Some debts could have negative recovery rate, if the defaulted amounts generate interest and fees in the months after default, but the debtors did not pay anything, so the outstanding balance keeps increasing. Whether fees and interest are allowed to be added after default is determined by banking rules and the lender's accounting conventions. The vast majority of UK lenders do not add fees and so the amount owed is frozen at default and the recovery rate is the amount repaid as a percentage of this. We use this convention in this research and so recovery rates only increase with time. It also means we redefine all negative recovery rates to be zero.

If fees and interest are included it is possible for the recovered amount to exceed the amount at default. In this case should one allow RR>1 or redefine it to be 1. We choose the latter course of action, which is consistent with fees being a cost in the recovery process and not part of the debt which is repaid. This is what mortgage and car finance companies do in that the fees are taken out of the money received for selling the repossessed property before addressing whether the remainder is enough to cover the defaulted balance of the loan. For credit card and personal loan recoveries there is less uniformity but normally a collections department will not charge fees or add interest to the defaulted balance during the recovery process.

Without the truncations in the two ends, Figure 3.2 shows the distribution of recovery rate is a bimodal shape with two peaks at 0 and 1 recovery rates. With those conventions mentioned above, the distribution of recovery rate is a bathtub shape, see Figure 3.3. (This distribution excludes the debts still being paid.) 30.3% debts have 0 recovery rate, and 23.9% debts have 100% recovery rate, others are relatively evenly distributed between 0 and 1. The truncation at the 1 end is consistent with the Basel Accord requirement for conservatism. Redefining negative recovery rate as zero is not conservative, but negative recovery rates really correspond to no payments and the lender adding extra fees. Thus redefining these as zero recoveries does reflect the actual actions of the borrower.

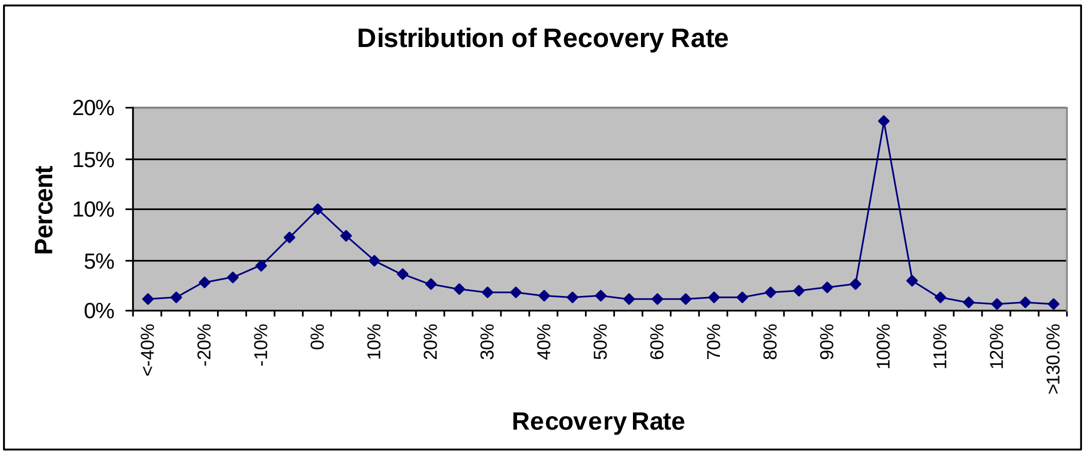

Figure 3.2 Distribution of recovery rate (without truncation)

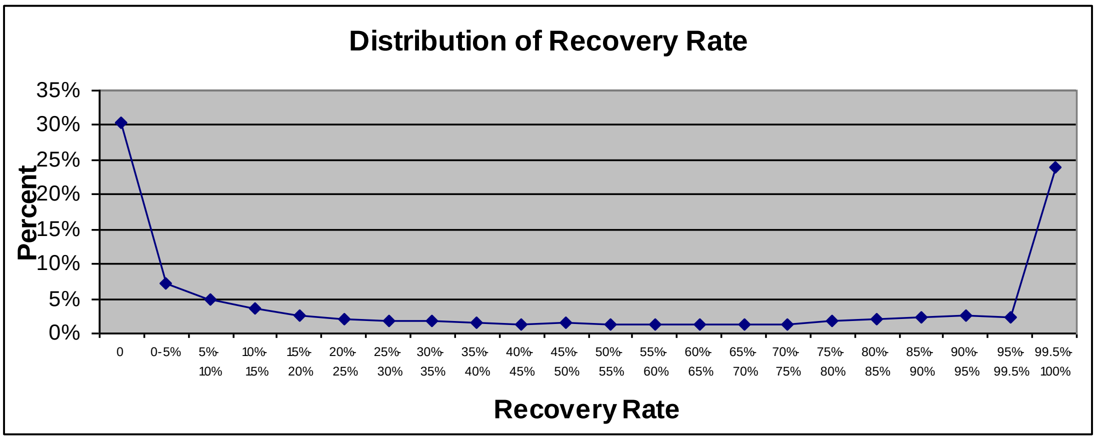

Figure 3.3 Distribution of recovery rate (with truncation)

The whole data is randomly split into 2 parts; the training sample contains 70% of observations for building models, and the test sample contains 30% of observations for testing and comparing models.

3.3 Single distribution approaches

3.3.1 Linear regression

Linear regression is the most obvious predictive model to use for recovery rate (RR) modelling, and it is also widely used in other financial area for prediction. Formally, linear regression model fits a response variable y to a function of regressor variables $x_1, x_2, ..., x_m$ and parameters. The general linear regression model has the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m + \varepsilon \qquad (3.1)$$

Where in this case

y is the recovery rate or recovery amount

$\beta_0, \beta_1, ... \beta_m$ are unknown parameters

$x_1, x_{2,} ..., x_m$ are independent variables which describe characteristics of the loan or the borrower

ε is a random error term.

In linear regression, one assumes that the mean of each error component (random variable ε) is zero and each error component follows an approximate normal distribution. However, the distribution of recovery rate tends to be bathtub shape, so the error component of linear regression model for predicting recovery rate does not satisfy these assumptions.

3.3.2 Survival analysis

Survival analysis concepts

Normally in survival analysis, one is dealing with the time that an event occurs and in some cases the event has not occurred and so the data is censored. In our recovery rate approach, the target variable is how much has been recovered before the collection's process stops, where again in some cases, collection is still under way, so the recovery rate is censored. The debts which were written off are uncensored events; the debts which are still being paid are censored events, because we don't know how much more money will be paid or could be paid. If the whole loan is paid off, we could treat this to be a censored observation, as in some cases, the recovery rate (RR) is greater than 1. If one assumes recovery rate must never exceed 1, then such observations are not censored. Since we redefine the cases where RR>1 so that RR=1, we will consider all recovery rates at 1 to be censored.

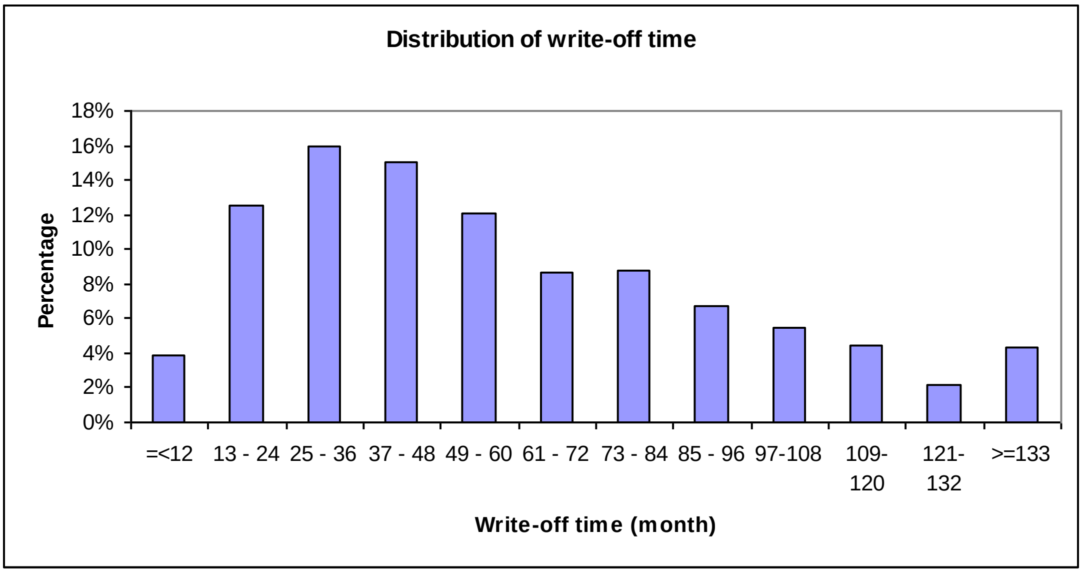

Figure 3.4 Distribution of write-off/paid off time

Since the recovery process takes so long, survival analysis has an advantage over the regression approaches, in that one can use the data for the cases in the recovery process, and not have to wait until they have either paid off completely or been written off. Figure 3.4 shows the distribution of time between default and being written off or paid off in full for the data set of this research. It shows the mean write-off/pay off time is 58 month, with a standard deviation of 34 months, and a longest time of 173 months. So in the regression approach one is using data on cases which on average are at least five years since default.

Suppose T is the random variable of the percentage of the debt recovered (defined as RR in this case) which has probability density function f. If an observed outcome, t of T, always lies in the interval [0, +∞), then T is a survival random variable. The cumulative density function F for this random variable is

$$F(t) = P(T \le t) = \int_0^t f(u)du \qquad (3.2)$$

The survival function is defined as:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(u)du \qquad (3.3)$$

Likewise, given S one can calculate the probability density function, f(u),

$$f(u) = -\frac{d}{du}S(u) \qquad (3.4)$$

The hazard function h(t) is an important concept in survival analysis because it models imminent risk. Here the hazard function is defined as the instantaneous rate of no further payment of the debt given that t percentage of the debt has been repaid,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t \mid T \ge t)}{\Delta t} \qquad (3.5)$$

The hazard function can be expressed in terms of the survival function,

$$h(t) = \frac{f(t)}{S(t)}, \qquad t > 0 \qquad (3.6)$$

Rearranging, we can also express the survival function in terms of the hazard,

$$S(t) = e^{-\int_0^t h(u)du} \qquad (3.7)$$

Finally, the cumulative hazard function, which relates to the hazard function, $h(t)$,

$$H(t) = \int_0^t h(u)du = -\ln S(t) \qquad (3.8)$$

is widely used.

It should be noted that f, F, S, h and H are related, and only one of the functions is needed to be able to calculate the other four.

There are two types of survival analysis models which connect the characteristics of the loan to the amount recovered – accelerated failure time models and Cox proportional hazards regression.

Accelerated failure time models

In an accelerated failure time model, the explanatory variables act multiplicatively on the survival function. They either speed up or slow down the rate of 'failure'. If g is a positive function of x and S_0 is the baseline survival function then an accelerated failure model can be expressed as

$$S_x(t) = S_0(t \cdot g(x)) \tag{3.9}$$

Where the failure rate is speeded up where $g(x) < 1$. By differentiating (3.9), the associated hazard function is

$$h_x(t) = h_0[tg(x)]g(x) \tag{3.10}$$

For survival data, accelerated failure models are generally expressed as a log-linear model, which occurs when $g(x) = e^{\beta^T x}$. In that case, one can show that the random variable T satisfies

$$\log_e T_x = \mu_0 + \beta^T x + \sigma Z \tag{3.11}$$

where Z is a random variable with zero mean and unit variance. The parameters, β, are then estimated through maximum likelihood methods. As a parametric model, Z is often specified as the Extreme Value distribution, which corresponds to T having an Exponential, Weibull, Log-logistic or other types of distribution. When building an accelerated failure model, the type of distribution of the dependent variable has to be specified.

Using accelerated failure time ideas to model recovery rates, leads to problems in that they do not allow the target variable to have a zero value nor can there be a value t* so that S(t*)=1 for all cases. Thus to use this approach one must allow RR>1 and not redefine such recovery rates to be 1; one also needs to use a logistic regression model to first classify which loans will have zero recovery rate, and use the accelerated failure approach on those which are predicted to have positive recovery rate.

Cox proportional hazards regression

Cox (1972) proposed the following model

$$h(t;x)=e^{(\beta^T x)}h_0(t) \tag{3.12}$$

Where β is a vector of unknown parameters, x is a vector of covariates and $h_0(t)$ is called the baseline hazard function.

The advantage of this model is that we do not need to know the parametric form of $h_0(t)$ to estimate β, and also the distribution type of dependent variable does not need to be specified. Cox (1972) showed that one can estimate β by using only the rank of the failure times to maximise the likelihood function.

3.3.3 Single distribution models results

Linear regression models

Two multiple linear regression models are built, one is for recovery rate as the target variable and one is for recovery amount as the target variable. In the former case, the predicted recovery rate could be multiplied by the default amount, and so the recovery amount could be predicted indirectly; in the latter case, a predicted recovery rate can be obtained by dividing the predicted recovery amount by the default amount.

The stepwise selection method was used for all regression models. Coarse classification was used on categorical variables so that attributes with similar average target variable values are put in the same class. The two continuous variables ‘default amount’ and ‘ratio of default amount to total loan’ were transformed into ordinal variables as well, and also their functions (square root, logarithm, and reciprocal) and their original form were included in the model building in order to find the best fit for the Recovery Rate.

The results are reported using a number of measures, R^2, the coefficient of determination is a common measure of goodness of fit for regression models, in that it measures how much of the square of the differences between the recovery rate of individual debtors and the mean recovery rate is explained by the RR model. Although R^2 of up to 0.8 are common in time series analysis, in real problems involving individual people, R^2 around 0.1 to 0.2 are not unusual. If one is only interested in how well the model is ranking the debtors, the Spearman coefficient is more appropriate. The Spearman rank correlation reflects how accurate the ranking of the predicted values is. It is a non-parametric measure of statistical dependence between two variables and assesses how well the relationship between two variables can be described using a monotonic function. If one is concerned about the error between the actual RR and the predicted RR for each individual then Mean absolute error (MAE) or Mean square error (MSE) would be the measure of importance. (MAE and MSE values for Recovery Amount will be much greater than those for Recovery Rate as the latter is always bounded between 0 and 1).

The R-squares for these models are small, (see Table 3.1, which gives the results on the training samples). This is consistent with previous authors (Bellotti and Crook 2009, Dermine and Neto de Carvalho 2006, Matuszyk et al 2010), but they are statistically significant. From the results, we can see modelling recovery rate directly is better than indirect modelling by first estimating the recovery amount. Surprisingly, better recovery amount results are also obtained by predicting recovery rate first and then calculating recovery amount rather than estimating the amount directly.

	R-square	Spearman	MAE	MSE
Recovery Rate from recovery rate model	0.1066	0.3183	0.3663	0.1650
Recovery Rate from recovery amount model	0.0354	0.2384	0.4046	0.2352
Recovery Amount from recovery amount model	0.1968	0.2882	1239.2	2774405.4
Recovery Amount from recovery rate model	0.2369	0.3307	1179.6	2637470.7

Table 3.1 Linear regression models results (from training sample)

The details of the recovery rate model whose results are given in the first row of Table 3.1 are given in Table 3.3. The most significant variable is 'the ratio of default amount to total loan', which has a negative relation with recovery rate. This gives some indication of how much of the loan was still owed before default occurs, and if a substantial portion of the loan was repaid before default then the Recovery Rate is also likely to be high. The second most significant variable is 'second applicant status', where loans with a second applicant have higher recovery rate than loans without a second applicant. Other significant variables, using p value as a measure, include: employment status, residential status, and default amount. The coefficient of the reciprocal of default amount looks very large but is only multiplying small values; so the overall impact although significant is not the largest effect.

The years of default were also allowed as independent variables since they represent the best one could hope to do if one used economic variables to represent the temporal changes in the credit environment. However, they were not very significant in the model, and only two were selected in. Table 3.2 lists the results of the model where 'default year' was left out, and we can see all the measuring criterion do not have large difference from the model results in Table 3.1. The fact they were not that significant means it was felt that adding in economic variables would have a minor impact in these models, thus we did not consider to put economic variables in the model at this stage, but will look at their impact in a later stage of Chapter 5.

	R-square	Spearman	MAE	MSE
Recovery Rate from recovery rate model	0.1057	0.3156	0.3665	0.1651
Recovery Rate from recovery amount model	0.0366	0.2377	0.4043	0.2308
Recovery Amount from recovery amount model	0.1947	0.2853	1240.4	2781574.5
Recovery Amount from recovery rate model	0.2370	0.3305	1179.9	2637215.4

Table 3.2 Linear regression models without variable 'default year' (results are from training sample)

In the recovery amount model, the variables which entered the model are very similar to recovery rate model. Because predicting recovery amount directly from the recovery amount model is worse than predicting it indirectly via the recovery rate model, the coefficient details of recovery amount model are not given.

Variable	Parameter Estimate	Standard Error	P-value
Intercept	0.682	0.029	<.0001
Employment status 1	0.098	0.013	<.0001
Employment status 2	0.144	0.015	<.0001
Mortgage	0.047	0.009	<.0001
Visa card	-0.036	0.010	0.0003
Insurance indicator 2	-0.053	0.009	<.0001
No. of dependant 2	0.027	0.010	0.0086
Personal loan account	0.024	0.008	0.0019
Residential status 1	-0.037	0.011	0.0005
Residential status 3	-0.041	0.017	0.0148
Residential status 4	-0.113	0.013	<.0001
Saving account	0.014	0.007	<.0351
Loan term1	-0.063	0.019	0.0007
Loan term2	-0.027	0.010	0.0080
Loan term4	0.042	0.011	0.0002
Second applicant status 1	-0.107	0.014	<.0001
Second applicant status 2	-0.051	0.017	0.0025
Second applicant status 3	-0.127	0.009	<.0001
Loan purpose 1	-0.069	0.016	<.0001
Loan purpose 2	-0.040	0.009	<.0001
Loan purpose 3	-0.051	0.012	<.0001
Loan purpose 4	-0.044	0.010	<.0001
Time at address 2	0.033	0.011	0.0029
Time at address 3	0.037	0.010	0.0003
Time at address 4	0.051	0.013	<.0001
Time at address 5	0.066	0.015	<.0001
Time at address 6	0.074	0.015	<.0001
Time at address 7	0.090	0.014	<.0001
Time with the bank 1	-0.030	0.015	0.0403
Time with the bank 5	0.032	0.010	0.0017
Time in occupation 1	0.029	0.013	0.0268
Time in occupation 2	0.039	0.013	0.0025
Time in occupation 3	0.044	0.015	0.0037
Time in occupation 4	0.047	0.015	0.0022
Time in occupation 5	0.090	0.016	<.0001
Monthly expenditure	0.036	0.016	0.0202
Monthly income 1	0.066	0.013	<.0001
Monthly income 2	0.060	0.013	<.0001
Default year 90	0.031	0.010	0.0021
Default year 96	0.029	0.011	0.0077
SQR default amount	-0.003	0.000	<.0001
REC default amount	-58.398	8.933	<.0001
Default rate	-0.012	0.001	<.0001

Table 3.3 Coefficients of variables in single distribution linear regression model for RR

Survival analysis models

There are two reasons why survival analysis may be a useful approach to Recovery Rate and LGD modelling. Firstly, debts still being repaid cannot be included in the standard linear regression approach. Survival analysis models can treat such repayments as censored, and include them easily in the model building. Secondly, the recovery rate is not normally distributed, so modelling it using linear regression makes the standard errors of parameters unstable, thus affects the significance tests. Bathtub shape distribution causes errors not normally distributed, which violates the assumptions of linear regression. Survival analysis models can handle this problem; different distributions can be set in accelerated failure time models, and Cox model's approach allows any empirical distribution.

Survival analysis models can be built for modelling both recovery rate and recovery amount. The event of interest is the percentage recovered when the debt is written off, so written-off debts are treated as uncensored; debts which were paid off or were still being paid are treated as censored. All the independent variables which are used in the linear regression model building are used here as well, and they are coarse classified again and dummy variables used to represent the various classes created. Continuous variables were firstly split into 10 to 15 bins to become 10 to 15 dummy variables, and these were put in a proportional hazard model without any other characteristics. Observing the coefficients from the model output, bins with similar coefficients were combined. The same method was used for nominal variables. This follows the approaches first proposed by Stepanova and Thomas (2002). Two continuous variables 'default amount' and 'ratio of default amount to total loan' were included in the models both in their original form and as coarse classified versions. The variables of 'default year' were kept in the model, and we did not build a model where they were left out, because the previous research showed they did not make big impact on model performance.

Because accelerated failure time models can not handle 0's existing in target variable, observations with recovery rate 0 should be removed off from the training sample before building the accelerated failure time models. This is also something that could be done for proportional hazards model, so that one is estimating the spike at RR=0, separately from the rest of the distribution. This leads to a new task: a classification model is needed to classify recovery 0's and non-0's (recovery rate greater than 0). Therefore, a logistic regression model is built based on the training sample before building the accelerated failure time models. In the logistic regression model, the variables 'month until default' and 'loan term' are very significant, though they were not so important in the linear regression models before. The other variables selected in the model are similar to those in the previous regression models. The Gini coefficient is 0.32 and 57.8% 0's were predicted as non-0's and 21.5% non-0's were predicted as 0's by logistic regression model. Cox regression models allow 0's to exist in the target variable; so two variants of the Cox model were built – one where one first separated out those with RR=0 by building a logistic regression model, and a one stage model where all the data was used to build the Cox model.

For the accelerated failure life models, the type of distribution of survival time needs to be chosen. After some simple distribution tests, Weibull, Log-logistic and Gamma distributions were chosen for the recovery rate models; and Weibull and Log-logistic distributions were chosen for the recovery amount models.

Unlike linear regression, survival analysis models generate a predicted distribution of the recovery values for each debt, rather than a precise value. Thus, to give a precise value, the quantile or mean of the distribution needs to be chosen. In all the survival models, the mean and median values are not good predictors, because they are too big and generate large MAE and MSE compared with predictions from some other quantiles. The optimal predicting quantile points are chosen based on minimising the MAE and/or MSE. The lowest MAE and MSE are found with quantile levels lower than median, and the results from the training sample models are listed in Table 3.4 and Table 3.5. The optimal quantiles are obtained empirically but it would be interesting to see whether there is any theoretical justification for them, which would be useful in using quantile regression in LGD modelling (Whittaker et al 2005). The model details of Cox-with 0 recoveries are found in Table 3.6, while the baseline hazard function for the model excluding the RR=0 values is given in Figure 3.4. In Figure 3.4, we can see that the hazard is higher in the two ends of recovery rate range, and lower in the middle range.

Recovery Rate	Optimal quantile	Spearman	MAE	MSE
Accelerated (Weibull)	34%	0.24731	0.3552	0.1996
Accelerated (log-logistic)	34%	0.25454	0.3532	0.2015
Accelerated (gamma)	36%	0.16303	0.3597	0.1968
Cox-with 0 recoveries	46%	0.24773	0.3631	0.2092
Cox-without 0 recoveries	30%	0.24584	0.3604	0.2100

Table 3.4 Survival analysis models results for recovery rate (training sample)

Recovery Amount	Optimal quantile	Spearman	MAE	MSE
Accelerated (Weibull)	34%	0.30768	1129.7	3096952
Accelerated (log-logistic)	34%	0.31582	1117.0	3113782
Cox-with 0 recoveries	46%	0.29001	1174.5	3145133
Cox-without 0 recoveries	30%	0.30747	1140.25	3112821

Table 3.5 Survival analysis models results for recovery amount (training sample)

Variable	Parameter Estimate	Standard Error	P-value
Mortgage	-0.142	0.024	<.0001
Visa card	0.106	0.027	0.0001
Personal loan account	-0.087	0.021	<.0001
Employment status 1	-0.079	0.040	0.0497
Employment status 2	0.064	0.033	0.0498
Employment status 3	0.328	0.045	<.0001
Insurance indicator 2	0.099	0.030	0.0009
Insurance indicator 3	0.115	0.032	0.0003
Marital status	0.090	0.031	0.0038
No. of dependant	-0.064	0.021	0.0026
Residential status 1	0.092	0.029	0.0015
Residential status 3	0.265	0.029	<.0001
Second applicant status 1	-0.225	0.025	<.0001
Second applicant status 2	-0.145	0.046	0.0015
Loan purpose 1	0.146	0.022	<.0001
Loan purpose 2	0.130	0.026	<.0001
Age of applicant	-0.051	0.024	0.0325
Time at address	-0.163	0.023	<.0001
Time in occupation	-0.147	0.024	<.0001
Time with the bank 1	-0.060	0.023	0.0108
Time with the bank 2	-0.115	0.030	0.0001
Time with the bank 3	-0.215	0.031	<.0001
Affordability	0.170	0.031	<.0001
Default rate 1	0.090	0.027	0.0007
Default rate 2	0.183	0.028	<.0001
Default rate 3	0.324	0.039	<.0001
Default rate 4	0.340	0.050	<.0001
Default rate 5	0.439	0.052	<.0001
Default amount 1	0.112	0.044	0.0104
Default amount 3	-0.068	0.027	0.0107
Default amount 4	0.059	0.027	0.0289
Default amount 5	0.183	0.040	<.0001
Default amount 6	0.210	0.044	<.0001
Month until default 1	0.120	0.039	0.0020
Month until default 2	0.067	0.027	0.0128
Default year 91	0.101	0.027	0.0002
Default year 92	0.082	0.038	0.0324
Default year 93	0.116	0.045	0.0103
Default year 97	-0.190	0.046	<.0001
Default year 98	-0.216	0.046	<.0001
Default year 99	-0.165	0.064	0.0097

Table 3.6 Coefficients of variables in single distribution Cox regression model (including 0 recoveries) for recovery rate

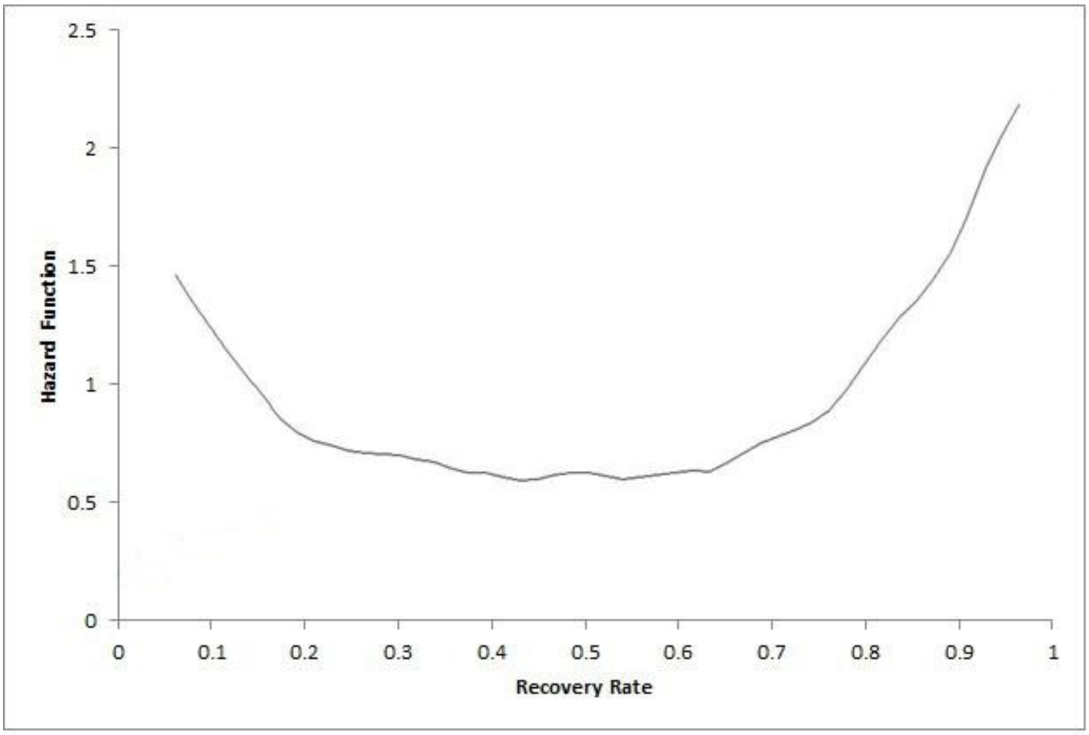

Figure 3.5 Baseline hazard function obtained from Cox model excluding RR=0

Using a quantile value has some advantages in this case and quantile regression has been applied in credit scoring research. Whittaker et al (2005) use quantile regression to analyse collection actions, and Somers and Whittaker (2007) use quantile regression for modelling distributions of profit and loss. Benoit and Van den Poel (2009) apply quantile regression to analyse customer life value. Using quantile values to make prediction avoids outlier influences. In particular when using survival analysis, the mean value of a distribution is affected by the amount of censored observations in the data set, so use a quantile value is a good idea when making predictions using it.

If the Spearman rank correlation test is the criterion to judge the model, we can see, from the above results tables (Table 3.4 and Table 3.5), the accelerated failure time model with log-logistic distribution is the best one among several survival analysis models. We can also see the optimal quantile point is almost the same regardless of the distribution in accelerated failure time models. The number of censored observations in the training sample does influence what the optimal quantile point is. If some of the censored observations are deleted from the training sample, the optimal quantile points move towards the median. This investigation was done by deleting half of the censored observations, the optimal quantile point moved from 34% to 44% in accelerated failure models, from 46% to 62% in Cox model including 0's, and from 30% to 46% in Cox model excluding 0's.

3.3.4 Comparisons of single distribution models

The comparison of the models is based on the results using the test sample. For debts still being paid, the final recovery amount and recovery rate are not known, and they can't be measured properly, thus these observations are removed from the test sample. This is unfortunate since it means one is comparing the methods only using debts which have been completely written off or paid off. Yet one of the advantages of survival analysis is that it can deal with loans which are still paying. The results from all the single distribution models when applied to the test sample are listed in Tables 3.7 and 3.8.

Recovery Rate	R-square	Spearman	MAE	MSE
(1) Linear Regression	0.0904	0.29593	0.3682	0.1675
(2) A – Weibull	0.0598	0.25306	0.3586	0.2042
(3) A – log-logistic	0.0638	0.25990	0.3560	0.2060
(4) A – gamma	0.0527	0.23496	0.3635	0.2015
(5) Cox – including 0's	0.0673	0.27261	0.3546	0.2006
(6) Cox – excluding 0's	0.0609	0.25506	0.3564	0.2072
(7) Linear Regression*	0.0292	0.22837	0.4077	0.2432
(8) A – weibull*	0.0544	0.24410	0.3606	0.2070
(9) A – log-logistic*	0.0591	0.25315	0.3575	0.2077
(10) Cox – including 0's*	0.0425	0.22646	0.3693	0.2216
(11) Cox – excluding 0's*	0.0504	0.23269	0.3624	0.2108

*: results from recovery amount models

Table 3.7 Comparison of recovery rate predictions from single distribution models (test sample)

In the test sample, R-square is calculated by linear regression approach, where the real RR is regressed on the predicted RR, and the R-square is reported by regression output. Only one independent variable is used in this approach, thus the R-square reported equals the square of the Pearson correlation coefficient between actual RR and predicted RR. From the recovery rate predictions in Table 3.7, if R-square and Spearman ranking test are the criterion to judge a model, we can see (1) Linear Regression is the best one, and (5) Cox-including 0's is the second best model. In the training sample, accelerated failure time model with log-logistic distribution outperforms the Cox models, but for the test sample, the Cox model including 0's is more robust than the accelerated failure models. In terms of MSE, linear regression always achieves the lowest MSE as one would expect as it is minimising that criterion. All the survival models have similar results. For MAE, the results are very consistent, except the linear regression models are poor. Modelling recovery rate directly (rows 1 to 6 in Table 3.7) gives better results than modelling it indirectly via recovery amount, whose results are in rows 7 to 11 of Table 3.7. Almost all the R-square and Spearman test from recovery amount models are lower than these from recovery rate models.

Recovery Amount	R-square	Spearman	MAE	MSE
(1) Linear Regression	0.1807	0.28930	1212.1	2634270
(2) A – weibull	0.1341	0.30594	1123.5	3026908
(3) A – log-logistic	0.1318	0.31178	1111.7	3047317
(4) Cox – including 0's	0.1572	0.31788	1138.9	2887499
(5) Cox – excluding 0's	0.1400	0.30437	1125.3	3017661
(6) Linear Regression*	0.2068	0.32522	1162.4	2549591
(7) A – weibull*	0.1424	0.31149	1116.1	2982477
(8) A – log-logistic*	0.1396	0.31697	1105.9	3014320
(9) A – gamma*	0.1413	0.30139	1141.5	2972807
(10) Cox – including 0's*	0.1628	0.34619	1101.9	2906821
(11) Cox – excluding 0's*	0.1377	0.31246	1107.4	3028183

*: results from recovery rate models

Table 3.8 Comparison of recovery amount predictions from single distribution models (test sample)

From the recovery amount results in Table 3.8, we see that modelling recovery amount directly (rows 1 to 5) is not as good as estimating recovery rate first (rows 6 to 11). The (6) Linear Regression* model achieves the highest R-square while (10) Cox-including 0's* model achieves the highest Spearman ranking coefficient. Both of them are recovery rate models and the predicted recovery amount is calculated by multiplying predicted recovery rate by the default amount. Regression models and Cox-including 0's models outweigh the accelerated failure time models. In the test sample, Cox-including 0's model beats the other survival models. The reason is that the logistic regression model which is used before the other models to classify 0 recoveries and non-0 recoveries generates more errors in the test sample, but Cox-including 0's model is not affected by this model.

3.4 Mixture distribution approaches

Models may be improved by segmenting population and building different models for each segment, because some subgroups maybe have different features and distributions. For example, small and large loans have different recovery rates, long established customers have higher recovery rate than relatively new customers (the latter may have high fraudulent elements which lead to low RR), and recovery rate of house owners is higher than that of tenants (because the former has more assets which may be realisable). Segmenting on recovery rate is a way of splitting who will not pay or permanently cannot pay from those who temporarily cannot pay. One could develop more sophisticated segments but using the RR values is an obvious first approach to a mixture model.

The development of finite mixture (FM) models dates back to the nineteenth century. In recent decades, as result of advances in computing, FM models proved to offer powerful tools for the analysis of a wide range of research questions, especially in social science and management (Dias, 2004). A natural interpretation of FM models is that observations collected from a sample of subjects arise from two or more unobserved/unknown subpopulations. The purpose is to unmix the sample and to identify the underlying subpopulations or groups. Therefore, the FM model can be seen as a model-based clustering or segmentation technique (McLachlan and Basford, 1998; Wedel and Kamakura, 2000).

In order to investigate different features and distributions in subgroups, we model the recovery rate by segmenting first. A classification tree model is built to generate segments with different features. Then, linear regression and survival models are built for each segment, so that mixture distribution models can be created.

Mixture distribution models have the potential to improve prediction accuracy and they have been investigated by other researchers for modelling RR. Matuszyk et al (2010) suggested to separate LGD=0 and LGD>=0 for unsecured personal loans, and then modelling LGD by using different models in each segment. Bellotti and Crook (2009) suggested to separate RR=0, 0<RR<1, and RR=1 for credit cards, and then for the group 0<RR<1, use Ordinary Least Squares regression or Least Absolute Value regression to model RR and achieved R-square 0.077. One possible reason for modelling RR by mixture distribution is people’s different views about repayment. Some debtors want to pay back, but they have financial troubles and can’t pay back; but some debtors deliberately do not want to pay.

For these reasons, we build a mixture model where the segments aim to have different recovery rate ranges. There are other ways of segmenting – age and size of loan, percentage of loan already paid off - which may also separate out the won’t pays from the can’t permanently pays and can’t temporarily pays, but using Recovery Rate to segment has the advantage of building on the work of others and of the inherent view that RR=0 must contain the won’t pays. The default years were not considered as variables to segment on because they did not appear significant in the single distributions, but it might be worth exploring this further in due course. We describe two approaches to achieving appropriate segments.

3.4.1 Method 1

The recovery rate is treated as a continuous variable and also the target variable, and a classification tree model is built to split the whole population into a few subgroups, in order to maximise the difference of average recovery rate between the subgroups.

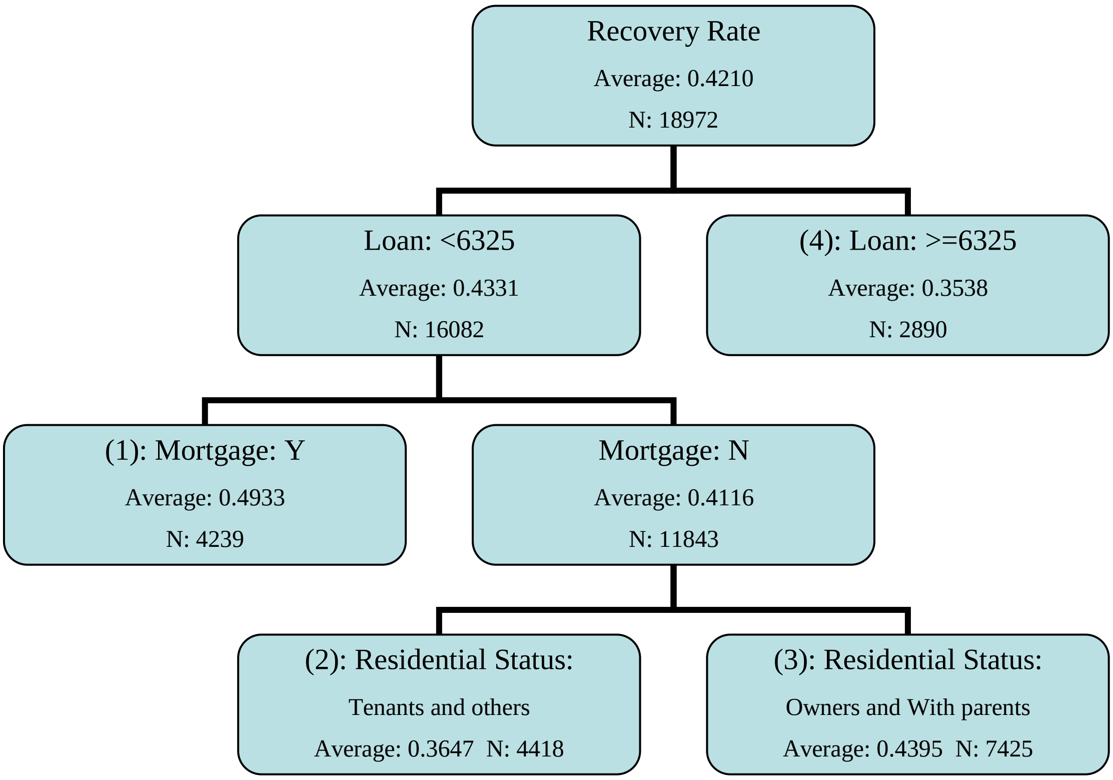

Figure 3.6 Method 1: Classification tree for recovery rate as continuous variable

SAS Enterprise Miner was used to produce a classification tree and the option ‘Tree Depth’ was set 4, because we did not want too many segments. The tree went down to the 4th level and the whole population was split to 14 end groups. However, we did not need so many groups to build mixture distribution models and also some end nodes had too few observations, thus the tree was pruned upward to make sure each end node contains at least 15% population. As is seen from the tree in Figure 3.6, the whole population was eventually split into 4 segments. Generally, large amount loans have lower recovery rate than small amount loans; if the debtors have a mortgage with this bank, then their loans have higher recovery rate than those without a mortgage with the bank; house owners or living with parents have higher recovery rates than people of tenants or those with ‘other’ residential status.

Linear regression model and survival models are built on each of the segments. The previous research shows that better predicted recovery amount results are obtained from predicting recovery rate first and then multiplying by the default amount, so only recovery rate models are built here. The models are built based on training samples and tested on test samples.

Recovery Rate	R-square	Spearman	MAE	MSE
Regression	0.0840	0.28544	0.3693	0.1688
Accelerated	0.0660	0.26625	0.3549	0.2055
Cox-including 0’s	0.0752	0.28581	0.3518	0.1967
Cox-excluding 0’s	0.0636	0.26236	0.3549	0.2067

Table 3.9 Recovery rate from mixture distribution models of method 1 (test sample)

In all four segments, linear regression is always the best modelling technique, as it has the highest R-square and Spearman coefficient; so after piecing together the 4 segments, linear regression model still has the highest R-square. Among the accelerated failure time models, the best fit in the first three segments are achieved with the log-logistic distribution models, and the best fit in the last segment is with Weibull distribution model. So the test results for the accelerated failure time models are made up of three log-logistic distribution models and one Weibull distribution model. In the Cox-regression modelling, the Cox model including 0’s (without logistic regression to predict 0 or non-0 recoveries) performs better than Cox model excluding 0’s (with logistic regression first) in all four subgroups. This means it is not better to predict 0 recoveries by logistic regression first. The results of the four approaches are given for the recovery rate in Table 3.9 and for the recovery amount in Table 3.10.

Recovery Amount	R-square	Spearman	MAE	MSE
Regression	0.1942	0.31824	1166.7	2593870
Accelerated	0.1346	0.31820	1102.3	3030185
Cox-including 0’s	0.1574	0.35314	1100.5	2976283
Cox-excluding 0’s	0.1357	0.31564	1105.8	3068188

Table 3.10 Recovery amount from mixture distribution models of method 1 (test sample)

In terms of R-square, among mixture distribution models, the linear regression models are the best; but in terms of Spearman ranking test, the Cox model-including 0’s outperforms the linear regression model, especially for predicting recovery amount.

Compared with the analysis from single distribution models, the results from mixture distribution models are disappointing and are somewhat worse than the results from the single distribution models. In terms of R-square, the best mixture distribution model is linear regression, but its R-square is still lower than that from the single distribution linear regression model. In terms of Spearman ranking coefficient, the best mixture distribution model is the Cox model-including 0’s. The Spearman ranking coefficient for the recovery rate is a little bit lower than 0.29593 which is the best one in the single distribution models; the Spearman ranking coefficient for the recovery amount is higher than 0.34619 which is the highest in the single distribution models. Thus, it seems mixture distribution models only improve the Spearman rank coefficient in the case of recovery amount predictions.

3.4.2 Method 2

Another way to separate the whole population is to split the target variable into three groups: the first group RR<0.05 (almost no recoveries), the second group 0.05<RR<0.95 (partial recoveries), and the third group RR>0.95 (full recoveries). These splits correspond to essentially no, partial or full recovery.

Recovery rate can be treated as an ordinal variable, with three classes - recovery rate less than 0.05 is set to 0, recovery rate between 0.05 and 0.95 is set 1, and recovery rate greater than 0.95 is set 2. A classification tree with the three classes as the target variable was tried, but the results were disappointing because each end node had similar distribution over the three classes. As an alternative a classification tree was first built to separate 0's and non-0's, so the whole data is split into two groups. Figure 3.7 is this tree, two end nodes with gray shade are classified as '0' group and other end nodes are classified as '1 or 2' group. Then a second classification tree was built for the non-0's group, in order to separate them into 1's and 2's. Figure 3.8 is this tree, the end node with gray shade is classified as '2' group. So eventually the population was split into 3 subgroups and this gave slightly better results. The distribution of the 3 groups is shown in Figure 3.9. The population in the first segment (most zero repayments) have the following attributes: no mortgage and loan term less than or equal to 12 months, OR no mortgage, time at address less than 78 months and have a current account. The population in the third segment (highest full repayment rate) have attributes: loan less than £4320 and insurance accepted. The rest of the population are allocated to the second segment as is shown in Figure 3.9.

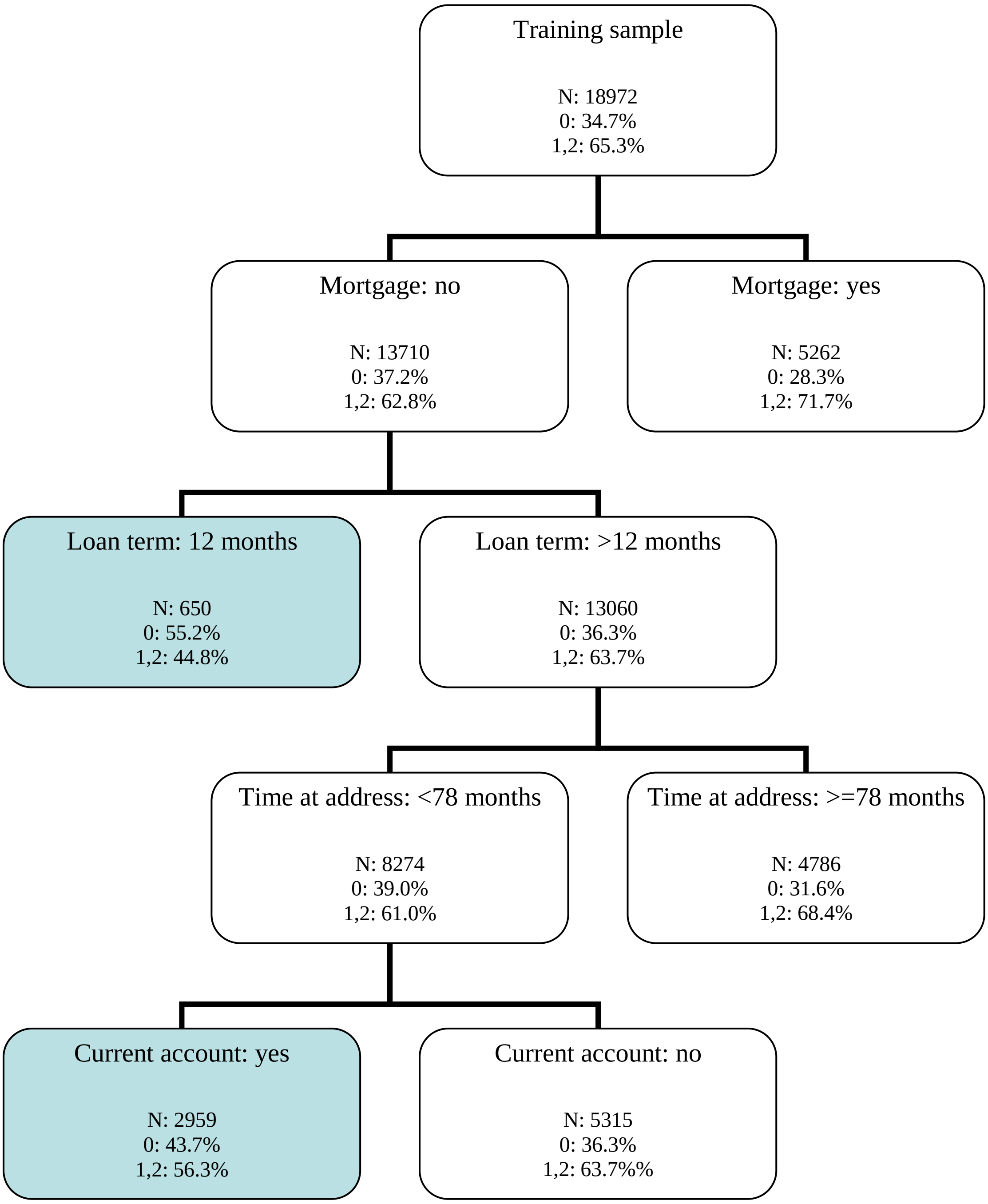

Figure 3.7 Method 2: First tree to separate ‘0’ and non ‘0’ groups

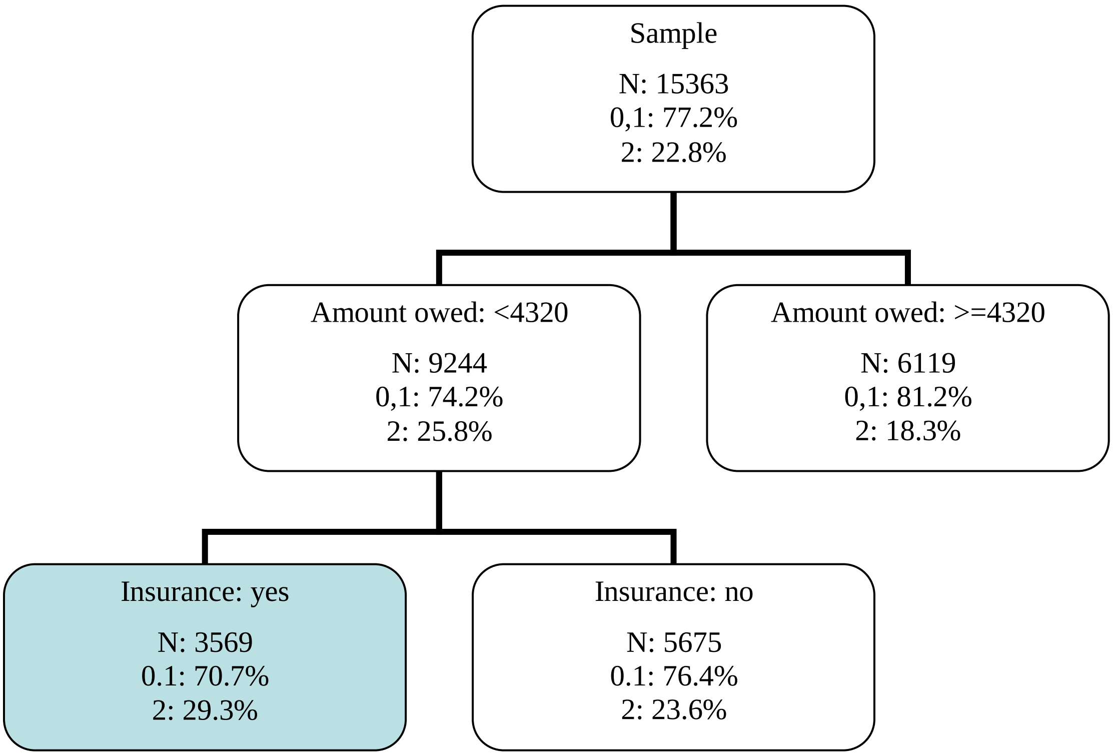

Figure 3.8 Method 2: Second tree to separate '2' and non '2' groups

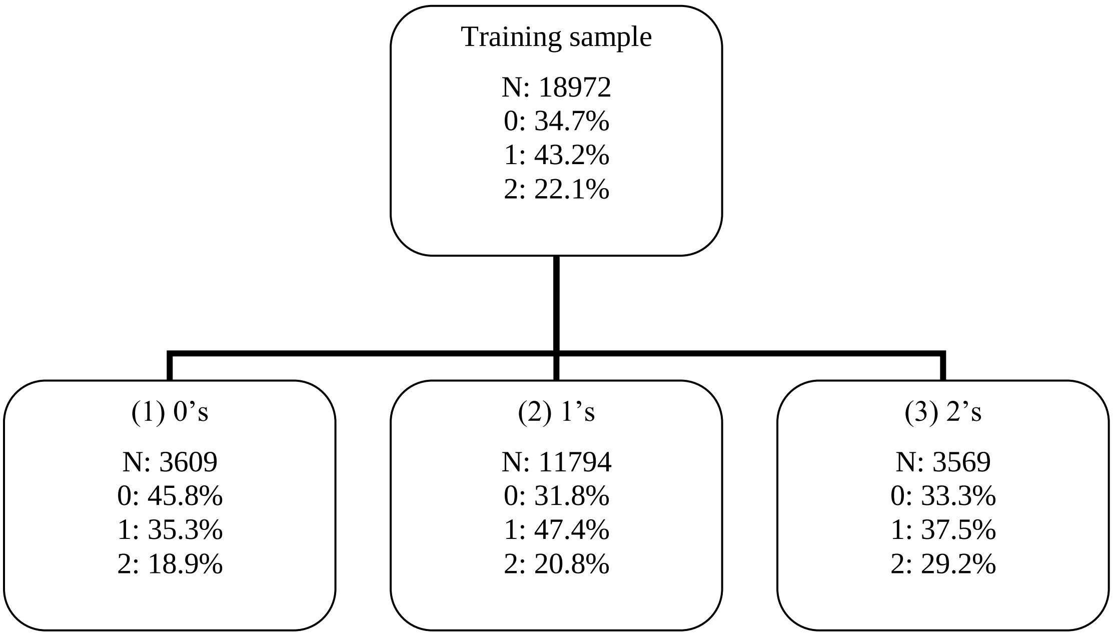

Figure 3.9 Method 2: Classification result for recovery rate as ordinal variable

This classification is very coarse. Group (1) aims at debts with recovery rate less than 0.05, but only 45.8% debts actually belong to this group; group (2) is for the debts with recovery rate between 0.05 and 0.95, but only 47.4% debts are in this range; group (3) is for the debts with recovery rate greater than 0.95, but, only 29.2% debts in this group have recovery rate greater than 0.95.

In the previous analysis, the linear regression model and Cox-including 0’s model are the two best models, so here only the linear regression model and the Cox-including 0’s regression model are built for each of the three segments. The models results from the combined test sample are listed in Tables 3.11 and 3.12.

Recovery Rate	R-square	Spearman	MAE	MSE
Regression	0.0734	0.26453	0.3695	0.1688
Cox including 0’s	0.0570	0.25869	0.3588	0.2051

Table 3.11 Recovery rate from mixture distribution models of method 2 (test sample)

Recovery Amount	R-square	Spearman	MAE	MSE
Regression	0.2054	0.31356	1169.4	2564149
Cox including 0’s	0.1669	0.33888	1125.7	2930725

Table 3.12 Recovery amount from mixture distribution models of method 2 (test sample)

From Tables 3.11 and 3.12, we can see that, for recovery rate, the linear regression model is still better than the Cox regression model in terms of R-square and Spearman coefficient; for recovery amount, the R-square of the linear regression model is higher than that of the Cox regression model, but the Spearman coefficient of linear regression is lower than that of the Cox model. Compared with the results from single distribution models, these mixture models do not improve the R-square or the Spearman ranking coefficient.

3.5 Conclusions

Estimating Recovery Rate and Recovery Amount has become much more important both because of the new Basel Accord regulation and because of the increase in the number of defaulters due to the 2007-2009 recession.

This chapter makes a comparison between single distribution and mixture distribution models of predicting recovery rate and recovery amount for unsecured consumer loans. Linear regression and survival analysis are the two main techniques used in this research where survival analysis can cope with censored data better than linear regression. For survival analysis models we investigated the use of proportional hazard models and accelerated failure time models though the latter have certain problems that need to be addressed – they do not allow 0’s to exist in the target variable and the recovery rate cannot be bounded above. This can be overcome by not defining RR>1 to be censored at 1 and by first using a logistic regression model to classify which loans have zero and which have non zero recovery rates. Cox’s proportional hazard regression models can deal with 0’s in the target variable and can deal with the requirement that RR≤1 for all loans. So that approach was tried both with logistic regression used first to split off the zero recoveries and without using logistic regression first. In all cases we used the approaches to model both recovery rate and recovery amount, and for all the models it turns out it is better to model recovery rate and then use the estimate to calculate the recovery amount rather than modelling the recovery amount directly.

		R square	Spearmen Rank Coefficient	MAE	MSE
Single distribution model	Linear regression	0.0904	0.29593	0.3682	0.1675
	Accelerated – log-logistic	0.0638 (<.0001)	0.25990 (0.0009)	0.3560 (0.0003)	0.2060 (<.0001)
	Cox– excluding 0's	0.0609 (<.0001)	0.25506 (0.0002)	0.3564 (0.0004)	0.2072 (<.0001)
	Cox- including 0	0.0673 (0.0002)	0.27261 (0.0323)	0.3546 (<.0001)	0.2006 (<.0001)
Mixture distribution model Method 1	Linear regression	0.0840 (0.3146)	0.28544 (0.3332)	0.3693 (0.9836)	0.1688 0.4237
	Cox - including 0	0.0752 (0.0152)	0.28581 (0.3506)	0.3518 (<.0001)	0.1967 (<.0001)
Mixture distribution model Method 2	Linear regression	0.0734 (0.0138)	0.26453 (0.0104)	0.3695 (0.9814)	0.1688 (0.4456)
	Cox - including 0	0.0570 (<.0001)	0.25869 (0.0007)	0.3588 (0.0043)	0.2051 (<.0001)

Table 3.13 Model comparisons for Recovery Rate (test sample)

		R square	Spearmen Rank Coefficient	MAE	MSE
Single distribution model	Linear regression	0.2068	0.32522	1162.4	2549591
	Accelerated – log-logistic	0.1396 (<.0001)	0.31697 (0.4368)	1105.9 (0.0003)	3014320 (<.0001)
	Cox– excluding 0's	0.1377 (<.0001)	0.31246 (0.2304)	1107.4 (0.0006)	3028183 (<.0001)
	Cox- including 0	0.1628 (<.0001)	0.34619 (0.0466)	1101.9 (<.0001)	2906821 (0.0002)
Mixture distribution model Method 1	Linear regression	0.1942 (0.1372)	0.31824 (0.5104)	1166.7 (0.7467)	2593870 (0.5614)
	Cox - including 0	0.1574 (<.0001)	0.35314 (0.0079)	1100.5 (<.0001)	2976283 (<.0001)
Mixture distribution model Method 2	Linear regression	0.2054 (0.8845)	0.31356 (0.2937)	1169.4 (0.7281)	2564149 (0.6541)
	Cox - including 0	0.1669 (<.0001)	0.33888 (0.1963)	1125.7 (0.0165)	2930725 (<.0001)

Table 3.14 Model comparisons for Recovery Amount (test sample)

Table 3.13 and Table 3.14 are model comparisons for Recovery Rate prediction and Recovery Amount prediction. The single distribution linear regression models in both tables are regarded as benchmark models. Other models are compared with these two models. The numbers in brackets are p-value of the significance tests of comparing each model with the benchmark linear regression model. Since these turn out to be the best models overall, the p-values are measuring whether the other models are statistically significantly different from the best model. R square is calculated by regressing the actual RR or recovery amount on predicted RR or recovery amount, and it equals to the square of Pearson correlation coefficients, thus the R square significance test is based on Pearson correlation test. Pearson correlation and Spearman ranking coefficients are tested by Fisher's *z* transformation using SAS *proc corr* procedure. MAE and MSE are tested by student's *t* test using SAS *proc univariate* procedure.

In the comparison of the single distribution models, the research result shows that linear regression is better than survival analysis models in most situations. For recovery rate modelling, see upper half of Table 3.13, linear regression achieves significantly higher R-square and Spearman rank coefficient than survival analysis models. The same situation happens to recovery amount modelling, see upper half of Table 3.14. The Cox model without logistic regression first is the best model among all the survival analysis models. This is surprising given the flexibility of distribution that the Cox approach allows. Of course one would expect MSE to be minimised using linear regression on the training sample because that is what linear regression tries to do. However, the superiority of linear regression holds for the other measures both on the training and the test set. One reason may be the need to split off the zero recovery rate cases in the accelerated failure time approach. This is obviously difficult to do and the errors from this first stage result in a poorer model in the second stage.

Another reason for the survival analysis approach not doing so well is that to make comparisons we used test sets where the recovery rate was known for all the debtors. That is they all had either paid off or been written off. So there was no opportunity to test the models predictions on those who were still paying, which is of course the type of data that is used by the survival analysis models but not by the regression based models. Finally in the survival analysis approach, there is the question of whether loans with RR=1 are really censored or not. Assuming they are not censored would lead to model lower estimate of RR, which might be more appropriate for the conservative philosophy of the Basel Accord.

The mixture models do not give a real improvement. Seeing Table 3.13 and 3.14, linear regression model and survival analysis models in mixture distribution approach are not better than their counterparts in single distribution approach in terms of R-square, Spearman Ranking coefficient, and MSE, except the improvement in Spearman coefficient for recovery amount modelling. (See Table 3.14, Cox regression model in mixture distribution model method 1, the Spearman ranking coefficient is significantly higher (p-value 0.0079) than that in single distribution linear regression model.) It is because finding suitable segments is difficult and the resultant subgroups are not as homogeneous as one would wish. In segmentation 1, four segments have different average RR's, but in each segment the individual RR still varies between 0 and 1. In segmentation 2, we tried to split the whole population into 3 segments of 'no recovery', 'partial recovery' and 'full recovery', but in each of the 3 segments, all 3 recovery statuses are mixed. This leads to the mixture models do no give a real improvement.

Chapter 4

Payment Patterns and Short term RR

4.1 Introduction

This chapter will look at arrear patterns before and after debtors' default and use arrear information to model RR. The results of Chapter 3 suggest that linear regression is the best method to model RR. Thus we use it as the main modelling method in this chapter. Firstly, we use arrear-pattern variables of before default and repayment-pattern variables of after default in RR prediction models, then try to build two-stage models. The first stage is to predict the number of payments in the first 12 months or 24 months after default, and the second stage is to use predicted information from the first stage to model overall RR. Secondly, we try to use linear regression to model both short term RR (12-month RR and 24-month RR) and overall RR with short term RR as independent variables, and also look at the relationship between short term RR and overall RR and build two-stage models. The first stage is to model short term RR and the second stage is to use predicted short term RR from the first stage to model overall RR. In the end, we measure the RR predictions in another way.

4.2 Data description

The data set in this research is the same data set as in chapter 3. It contains yearly payment amount information, and also monthly payment pattern records. Because of some missing data and errors in payment pattern records, some observations with missing data or errors are left out, and the total population for model building is less than before. It is thought the payment patterns could reveal something about final loss or recovery for each default debt. Thus we consider the information on payment patterns before default to predict short term and overall RR at the time of a debtor's default. The information of payment patterns in the first 1 or 2 years after default can be used to update the RR predictions at the time of 1 or 2 years after default.

4.3 Using pre-default arrear patterns to model RR

The arrear patterns of the 12 months before default are summarised by two kinds of variables: how many missed payments in 12 months just before default, and how they went to default.

Most debtors have 3 missed payments before default, which is the definition of default. However, not every case has only 3 missed payments in 12 months before default. Some observations have one or two missed payments a long time before default, which is beyond the 12 months before default, and the payments are always behind the normal schedule. So in the period of 12 months before default, they have only two or one missed payments. Some observations paid a large amount of money in advance – for example, the amount equal to five payments were paid in one month – then no payments in the following months, so it is seven months after that big payment before the debt went into default. So in this case, there are at least 7 missed payments in 12 months before default.

Another piece of payment information we can derive is default behaviour, which is how the debts go to default. Some observations go to default by missing 3 consecutive payments, but in some cases, their 3 missed payments are scattered. Some observations which have 3 consecutive missed payments also have missed payment beforehand, but they cured themselves after that. For example, if a debtor missed one month payment, but two payments were paid in the next month, we say this debtor was cured; in some cases there were no missed payment before the 3 consecutive missed payments, so no cure happens. All these types of different payment behaviour can be summarised by a few variables and can be used as predictive variables in RR prediction models.

The relationship between number of missed payments before default and the final RR is summarised in Table 4.1.

Missed payments	No. of observations	Percentage of observations	Mean of final RR
1	40	0.17%	0.5037
2	496	2.13%	0.4425
3	11625	49.98%	0.3967
4	6309	27.12%	0.4412
5	3196	13.74%	0.4698
6	1150	4.94%	0.4665
7	288	1.24%	0.4379
8	62	0.27%	0.4786
9	21	0.09%	0.5016
10	10	0.04%	0.4489
11	12	0.05%	0.6965
12	50	0.21%	0.3614

Table 4.1 Missed payments in 12 months before default and final RR

Table 4.1 is the table for the average RR of debts with different number of missed payments before default. About half of the whole population have 3 missed payments in 12 months before default, and the RR of these debts is lower than other debts (except debts with 12 missed payments). Some debts have less than 3 missed payments, because they had some missed payments earlier than 12 months before default. Some debts have 4 or 5 missed payments, due to some of them being cured after one or two missed payments, and then subsequently going to default. Some debts have a large number of missed payments, the reason is debtors paid large payments at one time, and after that they didn't pay anything. So after the large amount of payments were offset in the following few months, the debts went to default and they have more than 3 missed payments. We can not see any trend of the average RR in different missed payments groups. It seems there is no obvious relationship between the number of missed payments before default and the final RR, and the correlation coefficient between them is very small (0.01).

	No. of observations	Percentage of observations	Mean of final RR
(1) 3 consecutive missed payments, no missed payment before	7049	30.3%	0.3698
(2) 3 consecutive missed payments, have missed payments before	7607	32.7%	0.4378
(3) Not 3 consecutive missed payments	8603	37.0%	0.4572

Table 4.2 Default behaviour and final RR

Table 4.2 lists 3 types of default behaviour and the corresponding average RR. We can see debtors who go into default with 3 consecutive missed payments and no missed payment beforehand have the lowest RR; debtors who go into default but not in 3 consecutive missed payments have the highest RR; and the average RR of debtors who go into default in 3 consecutive missed payments and also have missed payments before hand is in the middle of the other two. This shows default behaviour may be a good variable to predict the RR of debts with different default behaviour.

The regression model of RR prediction without arrear information has a R square 0.1055. We try to add the arrear information into the RR prediction models and see whether the model will be improved. Three dummy variables are created for the number of missed payments before default. Missed payments less than or equal to 3 are taken as the reference variable; 4 missed payments is a dummy variable ‘missed_pay_4’; 5 missed payments is a dummy variable ‘missed_pay_5’; and 6 or more missed payments is a dummy variable ‘missed_pay_6’. Two dummy variables are created for default behaviour. In Table 4.2, the first behaviour is a dummy variable ‘behaviour_1’; the second behaviour is a dummy variable ‘behaviour_2’; and the third behaviour is the reference variable. Putting these 5 dummy variables into the regression model, the R square is increased to 0.1076. Two default behaviour variables are statistically significant with P value less than 0.0001, but the number of missed payments before default is not as important as default behaviour. Only one of the three dummy variables is selected by stepwise selection procedure and the P value is not really low. (See Table 4.3)

Variables	Parameter Estimate	Standard Error	t Value	P Value
Missed_pay_5	0.0187	0.0100	1.87	0.0609
Behaviour_1	-0.0521	0.0083	-6.32	<0.0001
Behaviour_2	-0.0385	0.0079	-4.86	<0.0001

Table 4.3 Arrear variables in RR prediction models

4.4 Using early default payment patterns to estimate final RR

It is also found that the number of payments in the first 12 or 24 months after default has some relationship with the final RR of each debt. Generally speaking, the more payments in the early default period, the higher RR is.

Figure 4.1 and 4.2 show the number of payments in 12 and 24 months after default and their corresponding average RR.

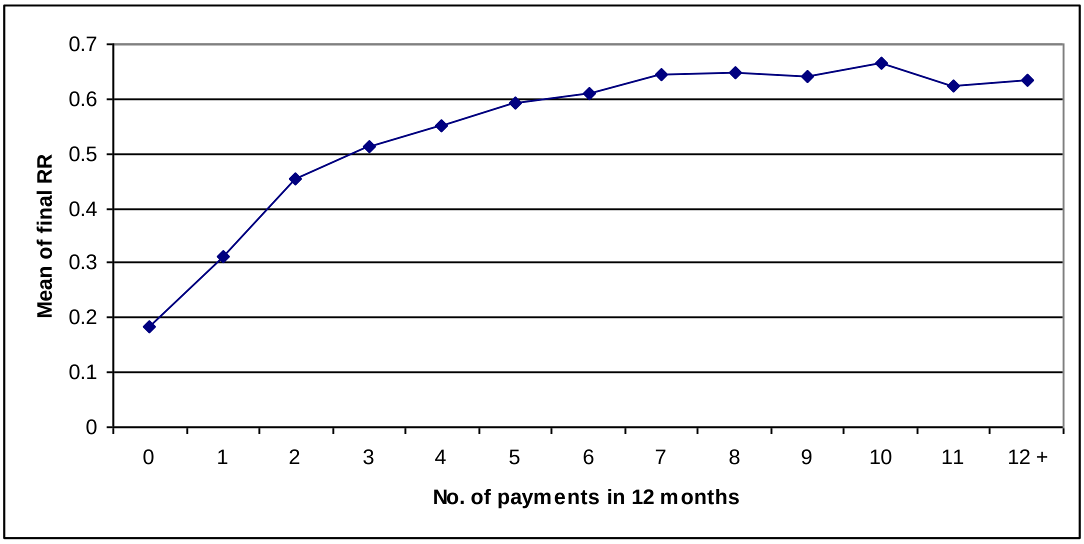

Figure 4.1 No. of payments in 12 months after default and final RR

In Figure 4.1, 30 percent of defaulters do not make any payment in the first 12 months after default, and they have the lowest average RR. The average RR of each group increases along with number of payments increasing until up to 7 payments, after that the average RR fluctuates around 0.64. We can conclude that the more payments in the first 12 months after default, the higher the final RR, but after the number of payments grows to 7, the number of payments has no influence on the final RR any more.

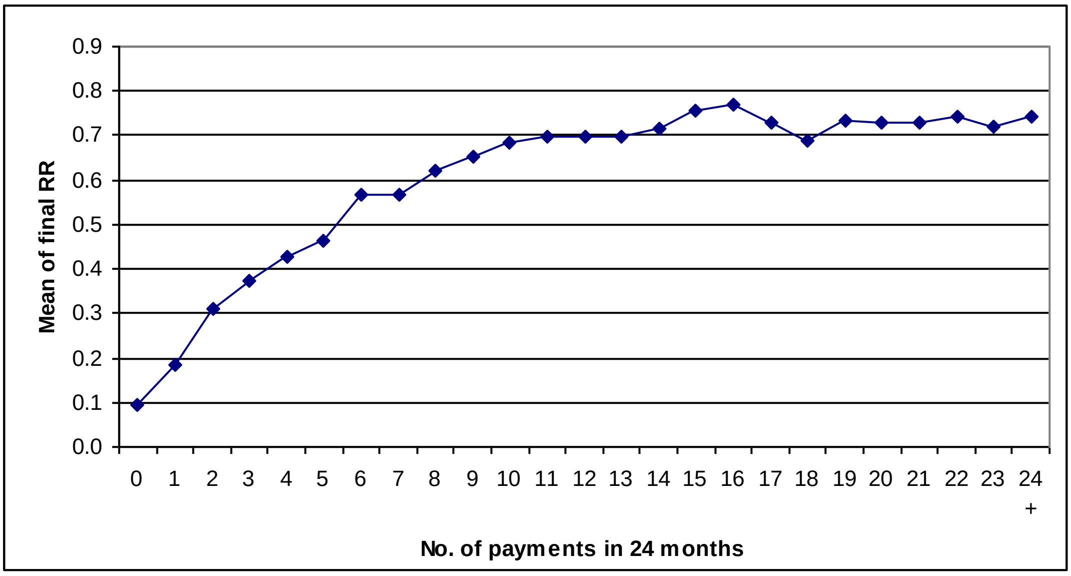

Figure 4.2 No. of payments in 24 months after default and final RR

The same situation happens in the first 24 months after default. We can see from Figure 4.2 that debtors who do not make any payments in the first 24 months have the lowest average RR and average RR grows up with number of payments growing until 11 payments, after that the average RR fluctuates around 0.70.

We can create some variables from the information about 12/24 months payments and put these variables in the prediction models to predict final RR. Three different ways to deal with the information of 12/24 months payments are adopted. One way is to put the number of payments directly in the model as one variable. We find that there are a large number of observations without any payment in 12/24 months, so another way to capture payment information is to create a dummy variable to indicate whether a debtor paid or not in the first 12/24 months after default. We also notice the average RR increases with the increase of number of payments in the early stage, but after a certain number of payments, the average RR stops increasing, which means the relationship between number of payments and final RR is not linear in the late stage. So, a few dummy variables could be created by combining adjacent groups with similar average RR. For 12 months payments, 4 dummy variables are created: the no payment group is taken as the reference group, a one payment group, a two or three payments group, a four to six payments group, and a seven payments or more group. For 24 months payments, 5 dummy variables are created: the no payment group is still taken as a reference group; a one payment group, a two or three payments group; a four or five payments group, a six to ten payments group; and an eleven payments or more group. These variables could be put in the prediction models with the other application variables and also the variables of payment patterns before default. The model results are listed in Table 4.6.

	With payment pattern before default	No. of payments in 12 months	Whether paid or not in 12 months	Four dummy variables for 12 months	No. of payments in 24 months	Whether paid or not in 24 months	Five dummy variables for 24 months
R square	0.1076	0.2232	0.2009	0.2531	0.3344	0.2365	0.3850
No. of variables in the model	42	38	41	37	42	43	42

Table 4.4 Model measurements of linear regression models including early payment information (training sample)

Table 4.4 lists R squares of, and the number of variables in, the linear regression models including early payment information. These 7 models are built using the whole population, and no test samples are used. The reason is that we just want to examine whether the payment-pattern variables are helpful in modelling RR, and there is no intention of predicting RR at this stage. In Table 4.4, the first column is the model excluding the variables of payment pattern after default, it is a benchmark model here, and the models in other columns are the ones including different format variables of payment pattern after default. From Table 4.4, we can see that models including number of payments in early default give greatly improved R square and yet the number of variables entering the model is not increased. Thus the R square improvement is not due to the increase of independent variables. All the newly created payment-pattern variables are statistically significant with

P-value less than 0.0001 in each model. Models including information of payments in 24 months are better than those including information of payments in 12 months, probably this is due to we get more repayment information in 24 months time than in 12 months time. To transform the payment number into dummy variables is better than to put payment numbers directly in the models, because the R square of the models with a few dummy variables (the fourth and seventh columns in Table 4.4) is significantly higher that that in other models. The results here suggest if banks record the payment patterns of each debt, in 12 or 24 months time they can update their RR prediction to get much more accurate prediction results.

4.5 Two-stage models with predicted number of payments in early default

If we know how many payments will be paid in the first 12 or 24 months after default, we can get much more accurate RR predictions at the time of defaults occurring, which has been proved in the last section, but the problem is we don't know them in advance until it happens. However, if we can make predictions for the number of payments in the first 12/24 months after default at the time when default occurs, then we can put the predicted number of payments in the RR prediction models, and if the predictions of number of payments are good, the final RR predictions will be improved. In this section, prediction models will be built to predict the number of payments in the first 12/24 months after default, and then two-stage models will be built to predict final RR based on the predicted early payments.

A simple way is to build linear regression models to predict the number of payments in early default period directly. The number of payments is a dependent variable, and the application variables and the variables of payment patterns before default are used as independent variables. The R square of linear regression models is 0.1111 and 0.1049 in test samples for

12 months payments prediction and 24 months payments prediction respectively. This result is, as we expected, disappointing.

Logistic regression models can also be tried to predict whether a debtor will pay or not in 12 months and 24 months. Two logistic regression models are built, and the Gini coefficient is 0.36 and 0.38 for test data sample for 12 months predictions and 24 months predictions respectively.

Cumulative logistic regression models could also be considered to predict payment information in 12 months and 24 months after default. Before building cumulative logistic regression models, we need to create an ordinal variable as the target variable. For 12 months prediction, an ordinal variable is created with 5 values: no payment is 0, one payment is 1, two or three payments is 2, four to six payments is 3 and seven or more payments is 4. For 24 months prediction, an ordinal variable is created with 6 values: no payment is 0, one payment is 1, two or three payments is 2, four or five payments is 3, six to ten payments is 4 and eleven or more payments is 5. This is following the approach of creating dummy variables before. Two cumulative logistic regression models are built to estimate the two target variables. Cumulative logistic regression model gives each value of the target variable a probability, which is the probability of the target value being that value, the sum of the probabilities for one observation is 1. The value with the highest probability is set as the predicted target result. However, the predictions made by cumulative logistic regression are not exciting. In 12 month payments predictions, only 0 and 4 exist in the predicted results. In 24 month payments predictions, only 0, 4 and 5 exist in the predicted results. If putting these predictions in the RR models, lots of payments information would be lost. An expedient is to put the predicted probabilities of each value of the target variable in the RR prediction models which include the dummy variables for payment numbers. This way can best use the predicted information from cumulative logistic regression models.

The results of final RR prediction models which include the predicted payments information are listed in Table 4.5. All the models are built based on training sample, where the actual payment information is used, the results in Table 4.5 are based on test data sample where the predicted payment information are used. The same independent variables are used in both stage-one and stage-two models, thus the collinearity may be existing in two-stage models in theory. In model building, the variance inflation factor (VIF) is tested for diagnosing collinearity in each model, and only the variables whose VIF value is less than 5 are left in the models. Thus, though collinearity within each stage is prevented, there is a possibility of collinearity between variables in different stages. This problem always arises in multi-stage scoring systems and provided the VIF values between the score of the previous stage and the new variables introduced is low, does not lead to difficulties.

Table 4.5 is the results of RR prediction models including predicted payments information after default. Model (1) doesn't include the predicted payments information and is a one-stage model, and it is in this table as a benchmark model. The numbers in brackets are the p-values of significance tests. The measures from two-stage models are tested whether they are significantly different from those of one-stage benchmark model (Model 1). Model (2) to Model (5) include the predicted payments information of 12 months after default. Model (2) includes the number of payments predicted from the linear regression model. Model (3) includes the binary variable of whether the debtors pay or not predicted from the logistic regression model. Model (4) includes the probabilities of each early payment band predicted from the cumulative logistic regression model; the probabilities are put into the dummy variables for the number of payments in RR prediction model. For model (5), the predicted number of payments predicted from the linear regression model in Model (2) is ranked in ascending order, and then binned into five groups to form an ordinal variable (with 5 values defined in Model 4). The ordinal variable is transformed into four dummy variables and the dummy variables are put in the final RR prediction model which includes dummy variables of payment information. Model (6) to Model (9) are the equivalent models based on 24 months payments predictions.

Linear regression models for RR predictions	R square	Spearman ranking coefficient	MAE	MSE
(1) One-stage Model (benchmark)	0.0975	0.29841	0.3637	0.1642
(2) No. of payments in 12 months (from linear regression)	0.0875 (0.1326)	0.28104 (0.1136)	0.3664 (0.2060)	0.1661 (0.2571)
(3) Whether paid or not in 12 months (from logistic regression)	0.0725 (<.0001)	0.27189 (0.0160)	0.3616 (0.4071)	0.1768 (<.0001)
(4) Four dummy variables for 12 months (from cumulative logistic regression)	0.0914 (0.3646)	0.28768 (0.3272)	0.3675 (0.0672)	0.1653 (0.4861)
(5) Four dummy variables for 12 months (from linear regression)	0.0577 (<.0001)	0.22921 (<.0001)	0.3656 (0.4938)	0.1844 (<.0001)
(6) No. of payments in 24 months (from linear regression)	0.0859 (0.0781)	0.27922 (0.0806)	0.3673 (0.1931)	0.1669 (0.1880)
(7) Whether paid or not in 24 months (from logistic regression)	0.0661 (<.0001)	0.27407 (0.0270)	0.3665 (0.2777)	0.1819 (<.0001)
(8) Five dummy variables for 24 months (from cumulative logistic regression)	0.0854 (0.0657)	0.28053 (0.1034)	0.3691 (0.0092)	0.1664 (0.1611)
(9) Five dummy variables for 24 months (from linear regression)	0.0479 (<.0001)	0.21679 (<.0001)	0.3700 (0.0019)	0.2001 (.0001)

Table 4.5 Final RR prediction models including predicted payment information in early default period (test sample)

In terms of R square and Spearman ranking coefficient, the two-stage models including payments predictions are not better than the one-stage model without payments predictions. The MAE and MSE performance measures are compatible with R square and Spearman coefficient; Model (1) has the lowest MSE and second lowest MAE. Although the lowest MAE appears in Model (3), the other 3 measures of Model (3) are worse than majority of other models. The results from the previous sections show that models including the real information of number of payments are largely improved from the models without payments information. The models in this section are trained from the training data sample where the real number of payments in early 12/24 months are used. Thus the variables of ‘number of payment in 12/24 months’ are very good and significant variables and have big influences on the model. However, in the test data sample, the number of payments in 12/24 months is predicted from the stage-one models. These predicted values are not very accurate and so they ruin the subsequent predictions from the two-stage models.

4.6 Predicting final Recovery Rate from early Recovery Rate

In this section, we will look at the relationship between short term RR and final RR, and see whether early RR can also help model final RR. Recovery rate at 12-month and 24-month after default can be calculated. These two variables can be used as independent variables in modelling final recovery rate. The relationship between 12/24-month RR and final RR is shown in Table 4.6 and 4.7.

Categories	No. of observations	range of 12 month RR	mean of 12 month RR	mean of final RR
0	11969	< or = 0	0	0.2436
1	1129	0 - 0.016	0.0076	0.3508
2	1125	0.016 - 0.034	0.0246	0.4310
3	1131	0.034 - 0.0556	0.0446	0.5143
4	1129	0.0556 - 0.0818	0.0682	0.5252
5	1124	0.0818 - 0.112	0.0966	0.5993
6	1130	0.112 - 0.151	0.1309	0.6564
7	1128	0.151 - 0.2013	0.1739	0.6734
8	1129	0.2013 - 0.2886	0.2398	0.7244
9	1124	0.2886 - 0.492	0.3678	0.7796
10	1132	0.492 +	0.8390	0.9043

Table 4.6 12-month RR and final RR

According to ascending ranking of 12-month RR, the whole population is split into 11 categories. Category 0 contains observations with 0 recovery rate in 12 month, and it contains more than half of the total population. Other categories are set up by nearly equal size of the observations. The range of 12-month RR of each category is in the third column, and the fourth and fifth column are the mean of 12-month RR and the mean of final RR for each category. We can see there is a strong positive linear relationship between 12-month RR and final RR. The relationship between 24-month RR and final RR is reflected in Table 4.7, the positive linear relationship is also very clear.

Categories	No. of observations	range of 24 month RR	mean of 24 month RR	mean of final RR
0	9592	< or = 0	0	0.1599
1	1360	0 - 0.023	0.0104	0.2239
2	1372	0.023 - 0.0529	0.0380	0.3163
3	1366	0.0529 - 0.0914	0.0710	0.3908
4	1364	0.0914 - 0.14	0.1145	0.5070
5	1360	0.14 - 0.20	0.1680	0.5925
6	1373	0.20 - 0.278	0.2379	0.6728
7	1368	0.278 - 0.38	0.3255	0.7391
8	1360	0.38 - 0.547	0.4533	0.8041
9	1369	0.547 - 0.905	0.7060	0.8623
10	1366	0.905 +	0.9817	0.9894

Table 4.7 24-month RR and final RR

Including 12-month RR in the model as an independent variable to predict final RR, the regression model for final RR prediction has an R square of 0.2434. If using an ordinal variable for 12-month RR (values from 0 to 10, the same as categories in Table 4.8) in stead of the real value in the model, the R square can be increased to 0.2938. If including the value of 24 month RR in the model, the R square of linear regression model to predict final RR is 0.4179. If an ordinal variable of 24 month RR is included in the model as an independent variable, the R square can be increased to 0.4729. Therefore, we can say 12/24-month RR are very good variables to predict final RR.

The R square is improved, but this is a little bit of cheating to include 12/24-month RR to predict final RR, because 12/24-month RR is part of final RR. It is like a riddle game, we have already known part of the answer, so it is easier to get the whole correct answer. But it is still very useful for banks to upgrade the LGD predictions in 12 months and 24 months time after default.

Another idea is to use the information of 12/24-month RR to predict the RR of the remaining default amount. It is like we know part of overall RR, and to predict the other part of overall RR. Table 4.8 and 4.9 list the mean of remaining RR for each category of 12-month RR and 24-month RR.

Categories	No. of observations	Range of 12 month RR	Mean of 12 month RR	Mean of remaining RR	Percentage of Repaid off finally
0	11631	< or = 0	0	0.2658	13.4%
1	1097	0 - 0.016	0.0076	0.3667	18.0%
2	1088	0.016 - 0.034	0.0247	0.4456	23.1%
3	1113	0.034 - 0.0556	0.0446	0.5158	27.0%
4	1099	0.0556 - 0.0818	0.0681	0.5213	27.6%
5	1105	0.0818 - 0.112	0.0966	0.5884	32.2%
6	1125	0.112 - 0.151	0.1309	0.6321	37.2%
7	1127	0.151 - 0.2013	0.1739	0.6299	36.4%
8	1122	0.2013 - 0.2886	0.2398	0.6677	40.6%
9	1118	0.2886 - 0.492	0.3677	0.6793	42.0%
10	552	0.492 +	0.6931	0.5027	34.2%

Table 4.8 12-month RR and remaining RR after 12 months

Cases which have been paid off or been written off within 12 or 24 months after default are left out. In Table 4.8, we can see in Category 10 the number of observations (the second column) drops a lot, because about half of the debts in this category have been paid off within 12 months. The far right column is the percentage of observations which paid off finally in each category on the base of the current numbers in the second column. It increases until category 9 and drops in category 10. The fifth column in Table 4.8 is the average RR of the remaining default amount after 12 months since default occurs. We can see the mean of remaining RR is increasing from Category 0 until Category 9. The remaining RR of Category 10 has an obvious drop; this is because cases which paid off within 12 months have been left out from this group, and among other cases only 34.2% of them paid off in the end, which is less than Category 9, thus the remaining RR for category 10 is lower than category 9.

Categories	No. of observations	Range of 24 month RR	Mean of 24 month RR	Mean of remaining RR	Percentage of Repaid off finally
0	7791	< or = 0	0	0.211	10.0%
1	1083	0 - 0.023	0.0106	0.2804	13.0%
2	1155	0.023 - 0.0529	0.0380	0.3552	16.8%
3	1206	0.0529 - 0.0914	0.0712	0.4037	21.4%
4	1295	0.0914 - 0.14	0.1149	0.4827	24.3%
5	1320	0.14 - 0.20	0.1682	0.5502	29.8%
6	1343	0.20 - 0.278	0.2379	0.6116	36.6%
7	1350	0.278 - 0.38	0.3255	0.6515	39.0%
8	1324	0.38 - 0.547	0.4528	0.6935	44.9%
9	1102	0.547 - 0.905	0.6896	0.6679	47.8%
10	177	0.905 +	0.9502	0.5915	31.1%

Table 4.9 24-month RR and remaining RR after 24 months

Table 4.9 reflects the story of remaining RR after 24 months default, which is very similar to what happened to remaining RR after 12 months default.

The information about 12/24-month RR can be used to predict the remaining RR. An ordinal variable is used for 12/24-month RR, also two dummy variables are adopted for the Category 9 and Category 10, because these two categories do not follow the general trend. Using only the variables of the 12/24-month RR, without any other variables, to build linear regression models to predict remaining RR, the models' R squares are 0.1262 and 0.1603 respectively for 12 month and 24 month. If other application variables are added in the model, R squares can be increased to 0.1710 and 0.1899 for 12-month and 24-month remaining RR prediction. The increase of R square is not very significant, so we can say the information of 12-month RR and 24-month RR is very important to predict the remaining RR after 12 months and 24 months since default.

4.7 Two-stage models with predicted early RR

12-month RR and 24-month RR are very important variables in modelling final RR and remaining RR. If we know these values at debtors' default time, we can get more accurate predictions of final RR at the time when debtors default. So, prediction models can be built to predict 12-month and 24-month RR by using payment information before default and the application variables, then put predicted short term RR values into final RR prediction models. If the predicted values of 12/24-month RR are good, the final RR prediction from the two-stage models will be better than to predict it directly just using payment information before default and application variables.

Two linear regression models are built to predict 12/24-month RR by using payment information before default and application variables. The R squares are 0.0961 and 0.0995 respectively for 12-month RR and 24-month RR. The prediction results are not satisfactory in terms of R squares, which suggests the predictions of 12-month RR and 24-month RR are not very accurate.

Two-stage models can be built on the base of 12/24-month RR prediction models. The training data sample is used to build models, where the real values of 12/24-month RR are put in. For the test data sample, the predicted 12/24-month RR values from the 12/24-month RR prediction models are put into the models trained from training data sample to predict the final RR. However, the prediction results from two-stage models are very disappointing; the model measurement results of the two-stage models for the testing data are listed in Table 4.10.

	R square	Spearman Ranking coefficient	MAE	MSE
(1) Without 12/24 RR predictions	0.0975	0.29841	0.3637	0.1642
(2) 12 month RR predictions	0.0951 (0.7240)	0.29501 (0.7546)	0.3650 (0.5270)	0.1645 (0.8506)
(3) Binned 12 month RR predictions	0.0639 (<.0001)	0.24510 (<.0001)	0.3613 (0.3954)	0.1875 (<.0001)
(4) 12 month RR group predictions	0.0720 (<.0001)	0.26827 (<.0001)	0.3578 (0.0053)	0.1856 (<.0001)
(5) 24 month RR predictions	0.0911 (0.3352)	0.28845 (0.3626)	0.3660 (0.2735)	0.1652 (0.5339)
(6) Binned 24 month RR predictions	0.0590 (<.0001)	0.23628 (<.0001)	0.3633 (0.9001)	0.2068 (<.0001)
(7) 24 month RR group predictions	0.0716 (<.0001)	0.26767 (<.0001)	0.3553 (0.0009)	0.2014 (<.0001)

Table 4.10 Measurement results of two-stage models including predicted 12/24-month RR (test sample)

In Table 4.10, Model (1) ‘Without 12/24 RR prediction’ in the second row is a one-stage model where no 12/24-month RR information is involved in the model building; it is listed in the table as benchmark just for model comparisons. Model (2) and Model (5) ‘12/24 month RR prediction’ are results of two-stage models where predicted values of 12/24-month RR from stage one (linear regression models) are directly put into final RR prediction models.

Another idea of building two-stage models is to use ordinal variables of 12/24-month RR rather than actual values. The ordinal variables are transformed from the continuous variables of 12/24-month RR. For the test sample, the predicted values of 12/24-month RR from stage one models are ranked in ascending order, and then transformed to be ordinal variables with

11 values (0-10 in Table 4.6 and 4.7) according to the proportions of each categorical group of ordinal RR in the training sample. Model (3) and Model (6) ‘Binned 12/24 month RR prediction’ use the transformed ordinal variables of 12/24-month RR. In the training samples, we create the ordinal variable based on the real values of 12/24-month RR; in the test samples, the ordinal variables are transformed from the predicted values of 12/24-month RR from stage one models. Predictions from these two-stage models are worse, because the R square and Spearman ranking coefficient are much lower than those of Model (2) and Model (5).

We can also predict the 12/24-month RR categories directly. The 12/24-month RR in the training sample were binned into 11 bins and transformed into ordinal variables as set in Table 4.6 and Table 4.7. These two ordinal variables can be used as dependent variables to build regression models to predict the categories of 12/24-month RR. The predicted values for the categories are ranked in ascending order firstly, and then are binned into 11 groups to form ordinal variables with value from 0 to 10 according to the proportion of each category in the training sample. After that, put the predicted grouped 12/24-month RR categories into the test data sample, and use the model trained from ordinal variables of 12/24-month RR to predict final RR. Model (4) and Model (7) (‘12/24 month RR group prediction’) are built in this way, and the measuring results are still disappointing.

From Table 4.10, we can see the two-stage models including the predicted 12/24-month RR do not make improvements compared with one-stage models. In the test sample, the R squares and Spearman ranking coefficients are still lower than those from the model without 12/24 month RR information. MSE is compatible with R square; Model (1) has the lowest MSE. This suggests the predictions of 12/24-month RR are not good enough to help predict final RR.

4.8 Measuring results in another way

In the above sections, we use R-square and Spearman Ranking Coefficient to measure models' performance. The R square is low and the results are poor and disappointing. But there is another measure which relates to what the industry wants from these models in terms of Basel Accord. In that one segments the portfolio and estimates LGD for a segment not for an individual loan. According to the value of predicted RR, we can rank the whole test population in ascending order, and then the whole test population can be split into 10 groups with equal size in each decile. The average value of real RR and the average value of predicted RR can be calculated in each group, then give the average value of predicted RR to each observation as their predicted RR in every group and then the measurement can be made between the average value of predicted RR and average value of real RR for the whole test observations.

Model	(1) No payment prediction		(2) 12 month payment prediction		(3) 24 month payment prediction		(4) 12 month RR prediction		(5) 24 month RR prediction	
Deciles	Real Ave RR	Predicted Ave RR	Real Ave RR	Predicted Ave RR	Real Ave RR	Predicted Ave RR	Real Ave RR	Predicted Ave RR	Real Ave RR	Predicted Ave RR
10%	0.1761	0.1739	0.2068	0.1798	0.2065	0.1765	0.1731	0.1854	0.1884	0.1904
20%	0.2780	0.2767	0.2652	0.2804	0.2772	0.2791	0.2954	0.2798	0.2699	0.2800
30%	0.3655	0.3321	0.3342	0.3303	0.3426	0.3300	0.3478	0.3305	0.3626	0.3262
40%	0.3450	0.3733	0.3901	0.3728	0.3892	0.3724	0.3552	0.3705	0.3508	0.3652
50%	0.3949	0.4104	0.3951	0.4094	0.4015	0.4088	0.3874	0.4058	0.3891	0.4014
60%	0.4180	0.4468	0.4263	0.4440	0.4199	0.4443	0.4337	0.4395	0.4499	0.4340
70%	0.4801	0.4808	0.4763	0.4784	0.4546	0.4789	0.4765	0.4756	0.4646	0.4684
80%	0.5166	0.5201	0.4856	0.5195	0.4952	0.5179	0.5197	0.5142	0.5260	0.5066
90%	0.5470	0.5742	0.5635	0.5702	0.5606	0.5674	0.5534	0.5650	0.5454	0.5553
100%	0.6649	0.6751	0.6436	0.6636	0.6395	0.6595	0.6438	0.6598	0.6393	0.6459
Correlation		0.9922		0.9942		0.9908		0.9957		0.9922
MAE		0.0151		0.0158		0.0197		0.0119		0.0131
MSE		0.0004		0.0003		0.0006		0.0002		0.0003

Table 4.11 Measuring the results in another way (test sample)

The 5 models with relatively good performance are selected from the previous sections to be measured in this new way, and the measuring results are listed in Table 4.11. Model (1) is a one-stage model, and does not include payment information. Model (2) to Model (5) are two-stage models. Model (2) and Model (3) predict number of payments in the first 12 and 24 months by linear regression first, and then put the predicted number of payments in the final RR prediction models. Model (4) and Model (5) predict the 12-month RR and 24-month RR by linear regression first, and then put the predicted early RR in the final RR prediction models. In all the 5 models, the real average RR and the predicted average RR are very close in each decile, this is fantastic results. Model (1) splits the real average RR to the maximum limit (from 0.1761 to 0.6649) among all 5 models. But a flaw in Model (1) is that the real average RR in 30% and 40% deciles are in the wrong order. This also happens in Model (5). We use Correlation, MAE and MSE to measure models' performance, as mentioned before. The real average RR and the predicted average RR are given to observations in each group, and Correlation, MAE and MSE are calculated between the real average RR and the predicted average RR; and the measuring results are very good. We can see all the 5 correlation coefficients are greater than 0.99, the MAE are less than 0.02 and the MSE are less than 0.001. In terms of these three criteria, the best model should be Model (4), but other models are only slightly inferior.

These results are surprisingly good. The reason may be that linear regression models are good at estimating the sample means. Thus the mean of each decile is estimated well and the systematic errors are caught. However, linear regression models are less successful at catching idiosyncratic errors. Thus measuring the prediction in individual loans are poor. These results are meaningful in practice. Banks can use these models to predict RR for individual loans. They can then split the portfolio into a few groups according to predicted RR in ascending order and calculate the average predicted RR for each group. The average predicted RR is very close to the actual average RR in each group. But, the drawback is, in each group, the actual RR of each observation still varies.

4.9 Conclusions

Payment-pattern variables before default are useful in modelling final RR, but the models including them seem are not significantly improved. Including the payment information of early 12 or 24 months after default is good to predict final RR, because models' R square is largely improved. However, these payment information can only be known after 12 or 24 months time since default, if we want to use this information in the time when debtors default, we have to make predictions. The prediction models of 12/24-month payments and 12/24-month RR were built in linear regression and logistic regression, the prediction results are disappointing. Two-stage models including the predicted 12/24-month payment information are worse than one-stage models without payment information. It proves that the simple models are always better than complex ones. Two-stage model is not a better choice for model builders who want to set more accurate LGD predictions for defaulters in the early stage, but still worthwhile for these who need to update the LGD predictions in the late stages after the early payment information is known. However, in all cases, in another measuring way, the results of the models are much better when predicting the means of segments of the loan portfolio than when predicting RR for individual loans.

Chapter 5

Macroeconomic Variables in RR prediction models

5.1 Introduction

Macroeconomic conditions have great influence on people's life, for example, in recessions lots of people loose jobs and are concerned for their livelihood, so they do not want to spend much money; but in prosperity people have positive feelings and would like to make purchase. Therefore, we want to examine whether macroeconomic environment influences debtors' repayment conditions. In this section, five macro-economic variables are considered to be included in the short term recoveries and final recovery rate prediction models; and their influences on the debts' payment behaviour will be examined.

5.2 Choice of macroeconomic variables

Not much research has been done on the influences of macroeconomic factors on LGD. Grippa et al (2005) made a survey on 250 Italian banks about Recovery Rates of bank loans, and also made a multivariate analysis. They reported time dummies were significant variables, reflecting the different effect of the economic cycle on the final recovery rate. But they did not test any specific economic variables. Dermine and de Carvalho (2005) estimated LGD for loans to small and medium sized firms in Portugal. They did a mortality analysis and included annual GDP growth as an independent variable. However, they found that GDP growth was not significant. They suggested that this may be due to no serious economic recessions during the period of analysis, 1995 – 2005. Bellotti and Crook (2009) included interest rates and unemployment rates in LGD models for credit cards in the UK. They reported these two economic variables were significant, and improved the prediction accuracy in the hold out sample both in individual loan level and portfolio level. But they only predicted one-year recovery rate, and did not consider final recovery rate. In this chapter, we consider to investigate the influences of 5 macroeconomic variables on recovery rate.

The five macro-economic variables we considered are:

GDP Growth: Quarter on quarter previous year change, seasonally adjusted
RPI Growth: Retail Prices Index, percentage change over 12 months
Unemployment Rate: All aged 16 and over, percentage, seasonally adjusted
Interest Rate: Bank of England interest rate
House Price Index Growth: Quarter on quarter previous year change, Halifax

‘GDP Growth’ is the main measure of a country's overall official economic output, and reflects the whole economic conditions. High GDP growth reflects economy growing fast and social prosperity. Low or negative GDP growth reflects economic shrinkage. Each resident is a member of a society, and their daily life and behaviour is influenced, more or less, by macroeconomic environment. ‘RPI Growth’ reflects the whole retail price fluctuations. High RPI means price increase and inflation in the whole market. On the one hand, this makes consumers feel they are becoming poor, and can’t afford to the high consumptions, thus they would not like to spend more; on the other hand, people would feel the value of their property and assets is increasing, and they are becoming rich, so this feeling makes people excessively consume easily. ‘Unemployment Rate’ indicates the employment conditions of a whole country. It influences the retail credit risk but only on the portfolio level. On the individual level, if a person looses the job, his or her potential losses would be high; if a person does not loose the job, no influence to them at all. 'Interest Rate' is usually used as an instrument to deal with inflation by government. It has some influence on individual consumptions. Under low interest rate, people would like to borrow money from banks; but under high interest rate, people are reluctant to do so because they know they would pay large sums of interest to banks. 'House Price Index Growth' reflects the price fluctuations in properties. High growth makes house owners feel they have more valuable properties and becoming rich, and low or negative growth makes them feel they are becoming poor. This feeling would influence their borrowing and repaying behaviours.

The Macroeconomic variables are correlated, because the same economic conditions usually generate similar economic outcomes. For example, good economic environment is usually with high GDP Growth, and low Unemployment Rate, and downturn economic condition is often with low GDP Growth and high Unemployment Rate. The Pearson correlation coefficients between economic variables from the period between 1987 and 2000 are shown in Table 5.1.

	GDP Growth	RPI Growth	Unemp. Rate	Interest Rate	House Price
GDP Growth	1				
RPI Growth	-0.41	1			
Unemp. Rate	-0.14	-0.13	1		
Interest Rate	-0.46	0.94	0.01	1	
House Price	0.55	0.20	-0.22	0.22	1

Table 5.1 Pearson correlation coefficients of macroeconomic variables (1987-2000)

From table 5.1, we can see ‘Interest Rate’ and ‘RPI Growth’ have strong positive correlation, so it may be confusing to have them both in the model. We leave ‘Interest Rate’ out, the reason of picking this one rather than ‘RPI Growth’ is that interest rate is always manually set and is controlled by government policy, but ‘RPI Growth’ is more objective. It reflects the real economic condition changes and is not controlled by government policy. The correlation between other variables is not very strong, although ‘GDP Growth’ has some negative correlation with ‘RPI Growth’ and has positive correlation with ‘House Price Index’. So, except for ‘Interest rate’, the other four economic variables are put in the prediction models, and their performance will be examined.

5.3 Economic variables and final Recovery Rate

The relationship between the economic variables and final RR can be seen from Table 5.2. Table 5.2 lists Pearson correlation coefficients of 5 economic variables from different time lags before or after default time with final RR. These time lags are 6 months before, 3 months before, and 1 month before default, the month of default, 1 month after, 3 months after, 6 months after and 12 months after default. (GDP Growth is a quarterly growth value. For the ‘1 month before’ variables, we use GDP Growth value from the last quarter, so it is the same value as ‘3 month before’.)

	Final Recovery Rate						
	6 month before	3 month before	1 month before	default month	3 month after	6 month after	12 month after
GDP Growth	-0.05 (<.0001)	-0.04 (<.0001)	-0.04 (<.0001)	-0.03 (<.0001)	-0.02 (<.0001)	-0.02 (<.0001)	-0.01 (0.0088)
RPI Growth	0.01 (0.0723)	0.01 (0.2798)	0.002 (0.7521)	0.001 (0.7862)	-0.001 (0.6749)	-0.001 (0.9636)	-0.01 (0.4881)
Unemp Rate	0.07 (<.0001)	0.08 (<.0001)	0.08 (<.0001)	0.08 (<.0001)	0.08 (<.0001)	0.08 (<.0001)	0.07 (<.0001)
House Price	-0.04 (<.0001)	-0.05 (<.0001)	-0.05 (<.0001)	-0.06 (<.0001)	-0.06 (<.0001)	-0.07 (<.0001)	-0.05 (<.0001)

Table 5.2 Pearson correlation coefficients between each of 5 economic variables and final recovery rate

In Table 5.2, all the correlation coefficients are less than 0.1, and we cannot see any strong correlation between any economic variables and the final RR. The P-values in the brackets show the significance of correlation coefficients. Although the correlation coefficients are small, the correlations between ‘GDP Growth’, ‘Unemployment Rate’, ‘House Price Index’, and final RR are significantly different from 0. Putting the 7 groups of economic variables from different time lags individually in the one-stage linear regression RR prediction model, in some cases, none of 4 economic variables are selected in; in other cases, only ‘unemployment rate’ is selected into the model by stepwise selection process, but its t-value is not high enough and its p-value is not less than 0.01, which shows it is not a significant variable. Thus, the regression model reinforces the findings from Table 5.2: the final RR does not have strong relationship with economic variables. So, for this data set: macroeconomic conditions appear to have little influence on Recovery Rate.

5.4 Economic variables and number of payments

The previous section suggests the economic variables do not very much influence final RR. In this section, we want to examine whether economic variables influence the short term payment behaviours. If they have relationship with short term payment behaviours, we want economic variables to help us predict number of payments in the first 12 and 24 months after default with the intention to improve the prediction accuracy of two-stage recovery rate models. First of all, the Pearson correlation between economic variables and number of payments is checked like before, and the correlation coefficients are listed in Table 5.3 and Table 5.4.

	Number of Payments in 12 months after default						
	6 month before	3 month before	1 month before	default month	3 month after	6 month after	12 month after
GDP Growth	-0.09 (<.0001)	-0.11 (<.0001)	-0.11 (<.0001)	-0.12 (<.0001)	-0.14 (<.0001)	-0.16 (<.0001)	-0.23 (<.0001)
RPI Growth	0.17 (<.0001)	0.17 (<.0001)	0.19 (<.0001)	0.20 (<.0001)	0.25 (<.0001)	0.29 (<.0001)	0.30 (<.0001)
Unemp Rate	0.01 (0.0152)	0.03 (<.0001)	0.04 (<.0001)	0.04 (<.0001)	0.06 (<.0001)	0.07 (<.0001)	0.13 (<.0001)
House Price	0.18 (<.0001)	0.11 (<.0001)	0.05 (<.0001)	0.03 (<.0001)	-0.05 (<.0001)	-0.11 (<.0001)	-0.17 (<.0001)

Table 5.3 Pearson correlation coefficients between economic variables and number of payments in 12 months after default

	Number of Payments in 24 months after default						
	6 month before	3 month before	1 month before	default month	6 month after	12 month after	18 month after
GDP Growth	-0.06 (<.0001)	-0.08 (<.0001)	-0.08 (<.0001)	-0.09 (<.0001)	-0.12 (<.0001)	-0.17 (<.0001)	-0.21 (<.0001)
RPI Growth	0.13 (<.0001)	0.14 (<.0001)	0.15 (<.0001)	0.16 (<.0001)	0.23 (<.0001)	0.24 (<.0001)	0.19 (<.0001)
Unemp Rate	0.001 0.7055	0.01 0.2260	0.01 0.0226	0.02 0.0056	0.04 (<.0001)	0.08 (<.0001)	0.12 (<.0001)
House Price	0.16 (<.0001)	0.12 (<.0001)	0.07 (<.0001)	0.05 (<.0001)	-0.07 (<.0001)	-0.12 (<.0001)	-0.13 (<.0001)

Table 5.4 Pearson correlation coefficients between economic variables and number of payments in 24 months after default

Table 5.3 shows the Pearson correlation coefficients between economic variables of different time lags before or after default and number of payments in 12 months after default. Table 5.4 shows the Pearson correlation coefficients between economic variables and number of payments in 24 months after default. From the two tables above, we can see the number of payments has a negative relationship with ‘GDP growth’ and a positive relationship with ‘RPI growth’. It has a very weak positive relationship with ‘unemployment rate’; and it has a positive relationship with ‘house price index’ before default and a negative relationship with ‘house price index’ after default. We can also notice that the correlation becomes stronger with the time going towards the point at which it is measured, especially in ‘GDP growth’ and ‘RPI growth’, which suggests people’s repaying behaviour is more related to the current economic conditions rather than economic conditions before default. The P-values in the table show most of correlation coefficients are significantly different from 0, only except ‘unemployment rate’ from some time lags.

We want to predict the number of payments in the early periods after default happens, so the economic variables from the time period after default can not be used in prediction models. The economic variables which are considered to put in the models are those from three time lags: 6 months before, 3 months before and 1 month before default, and the three groups of economic variables are put in the prediction models for number of payments separately. After careful observation and comparison, 3 economic variables are chosen. They are ‘GDP Growth’ 1 month before, ‘RPI Growth’ 1 month before, and ‘House Price’ 6 months before; these 3 variables have relatively larger correlation coefficients with number of payments in 12 and 24 months after default and their t-values in the regression models output are larger than those of the same variables from other time lags. Different combinations of these 3 variables are put in the models in order to find the model with best economic variables for the number of payments predictions. In order to evaluate whether these economic variables are helpful to predict final RR, we put the predicted number of payments in the two-stage RR models which include payment patterns and see their R square, Spearman ranking coefficients and mean square error (MSE).

	No. of Payments Prediction in 12 months after default			Final Recovery Rate		
Economic Variables	Spearman	R square	MSE	Spearman	R square	MSE
(1) No Economic Variables	0.33529	0.1136	13.233	0.28104	0.0875	0.1661
(2) GDP, RPI, House Price	0.36115 (0.0082)	0.1597 (<.0001)	12.545 (0.0002)	0.27427 (<.0001)	0.0824 (0.3886)	0.1672 (0.4896)
(3) GDP, RPI	0.34307 (0.3938)	0.1274 (0.0411)	13.027 (0.2694)	0.28215 (0.9115)	0.0875 (0.9919)	0.1661 (0.9881)
(4) RPI	0.34557 (0.2964)	0.1236 (0.1388)	13.085 (0.4300)	0.28174 (0.9594)	0.0879 (0.9860)	0.1660 (0.9596)
(5) RPI, House Price	0.36232 (0.0057)	0.1603 (<.0001)	12.536 (0.0002)	0.27495 (0.5031)	0.0826 (0.4051)	0.1671 (0.5054)
(6) House Price	0.35927 (0.0143)	0.1565 (<.0001)	12.592 (0.0007)	0.27432 (0.5047)	0.0823 (0.3845)	0.1672 (0.4816)

Table 5.5 Model results for 12-month number of payments predictions and final recovery rate predictions with economic variables (test sample)

Table 5.5 lists the measuring results from test sample of regression models (including different combinations of economic variables) which predict the number of payments in 12 months after default and also the final recovery rate. Model (1) does not include any economic variables, and is as benchmark model for model comparison. Measures from other models are tested to check whether they are significantly different from those of Model (1). We can see, with all 3 economic variables in the Model (2) (the second row in Table 5.5), the prediction outcomes of number of payments are improved significantly, because the Spearman ranking coefficient and R square are raised and the MSE is lowered. All have significant p-values, (0.0082, <.0001, 0.0002) compared with the Model (1) without economic variables (the first row in Table 5.5). The most significant improvements in the left hand side of Table 5.5 happen in the fifth row for the Model (5) including 'RPI' and 'House Price Index'. This achieves the highest Spearman ranking coefficient and R square and the lowest MSE among all the models in Table 5.5, looking at the results for predicting number of payments in first 12 months.

We put the predicted number of payments from the models in the left hand side of Table 5.5 into the recovery rate prediction model (two-stage models including payment-pattern variables, which are in the right hand side of Table 5.5). However, Model (2) with all 3 economic variables and the Model (5) with 'RPI' and 'House Price Index' now have worse performance than the Model (1) without economic variables, because Spearman ranking coefficients and R squares are decreased and MSE is increased (seeing the right hand side of Table 5.5). However, the p-values suggest that apart from the Spearman ranking coefficient of Model 2, the other model results cannot be considered statistically different from those of Model 1. This is also the case for Model 4, though it beats Model 1, its p-values of 0.9594, 0.9860, and 0.9596, say this improvement is not significant.

	No. of Payments Prediction in 24 months after default			Final Recovery Rate		
Economic Variables	Spearman	R square	MSE	Spearman	R square	MSE
(1) No Economic Variables	0.32599	0.1019	39.372	0.27922	0.0859	0.1664
(2) GDP, RPI, House Price	0.34489 (0.0557)	0.1347 (<.0001)	37.937 (0.0261)	0.26952 (0.2129)	0.0786 (0.1311)	0.1680 (0.2221)
(3) GDP, RPI	0.33078 (0.6303)	0.1103 (0.1944)	39.002 (0.5708)	0.28051 (0.7870)	0.0854 (0.6533)	0.1665 (0.7463)
(4) RPI	0.33151 (0.5791)	0.1071 (0.4258)	39.145 (0.7299)	0.28025 (0.7730)	0.0864 (0.7563)	0.1663 (0.8331)
(5) RPI, House Price	0.34516 (0.0523)	0.1346 (<.0001)	37.941 (0.0266)	0.27006 (0.2309)	0.0787 (0.1342)	0.1680 (0.2258)
(6) House Price	0.34244 (0.0960)	0.1330 (<.0001)	38.011 (0.0354)	0.26957 (0.2148)	0.0784 (0.1229)	0.1681 (0.2130)

Table 5.6 Models results for 24-month number of payments predictions and final recovery rate predictions with economic variables (test sample)

The same story happens to the prediction of number of payments in 24 months. From Table 5.6, we can see the Model (2) with all 3 economic variables (the second row in Table 5.6) has the best performance in the prediction of number of payments in 24 months. However, in the two-stage RR prediction models, the best results are from the Model (4) with only ‘RPI Growth’ in (the fourth row in Table (6)); the improvements are also very tiny and insignificant.

5.5 Cumulative logistic regression models with economic variables

In Chapter 4, cumulative logistic regression models were used to predict the probability of each payment band, and then put the predicted probability into RR prediction models including payment bands as independent variables. Unfortunately, the two-stage models with the predicted probabilities are worse than the one-stage model. Here, we put the economic variables into the cumulative logistic regression models to predict the probability of each payment band again, and then put the predicted probabilities of payment bands in the two-stage final RR prediction models. The results of RR predictions from the two-stage models with cumulative logistic regression models as stage-one model are shown in Table 5.7.

	RR model including 12 months payments bands predictions from cumulative logistic regression			RR model including 24 months payments bands predictions from cumulative logistic regression		
Economic Variables	Spearman	R square	MSE	Spearman	R square	MSE
(1) No Economic Variables	0.28768	0.0914	0.1653	0.28053	0.0854	0.1664
(2) GDP, RPI	0.28792 (0.9814)	0.0915 (0.9981)	0.1653 (0.9997)	0.28064 (0.9915)	0.0852 (0.9816)	0.1665 (0.9915)
(3) RPI	0.28804 (0.7561)	0.0917 (0.9543)	0.1653 (0.9994)	0.28068 (0.9902)	0.0854 (0.9971)	0.1664 (0.9991)
(4) RPI, House Price	0.28142 (0.4259)	0.0875 (0.5627)	0.1661 (0.9153)	0.27389 (0.3138)	0.0808 (0.8743)	0.1674 (0.9243)

Table 5.7 RR prediction results of two-stage models from cumulative logistic regression models with economic variables (test sample)

In Table 5.7, the left hand side presents the results of the two-stage RR prediction model including predicted probabilities of payment bands in 12 months, and the right hand side presents the results of the two-stage RR prediction model including predicted probabilities of payment bands in 24 months. It is noticed that model (3) which includes the economic variable ‘RPI Growth’ is the best one in both sides, compared with the model without economic variables, but it only improves the Spearman ranking coefficient and the R square a little (not significant due to high p-value) and there is no improvements in MSE. The improvements are so small that they can be neglected.

5.6 Economic variables and 12/24-month Recovery Rate

The Pearson correlation coefficients between economic variables and 12/24-month recovery rate are listed in Table 5.8 and Table 5.10. The time lag effects are also considered, so the economic variables from a certain time period before or after default are used to calculate correlation coefficients.

	12 months Recovery Rate						
	6 month before	3 month before	1 month before	default month	3 month after	6 month after	12 month after
GDP Growth	0.08 (<.0001)	0.08 (<.0001)	0.08 (<.0001)	0.07 (<.0001)	0.07 (<.0001)	0.06 (<.0001)	0.02 (0.0003)
RPI Growth	-0.04 (<.0001)	-0.02 (0.0014)	-0.01 (0.3060)	0.001 (<.8834)	0.03 (<.0001)	0.05 (<.0001)	0.07 (<.0001)
Unemp Rate	-0.02 (0.0017)	-0.04 (<.0001)	-0.05 (<.0001)	-0.05 (<.0001)	-0.07 (<.0001)	-0.07 (<.0001)	-0.07 (<.0001)
House Price	0.11 (<.0001)	0.09 (<.0001)	0.08 (<.0001)	0.08 (<.0001)	0.06 (<.0001)	0.05 (<.0001)	0.05 (<.0001)

Table 5.8 Pearson correlation coefficients between economic variables and 12-month RR

From Table 5.8, we can not see any strong correlation between economic variables and 12-month RR due to small values, although most of them are significantly different from 0. All 4 economic variables from each period are put into the 12-month RR prediction regression model (including the application variables); only 'Unemployment Rate' is selected by the stepwise selection process in most cases. Although 'GDP Growth' has a positive correlation with 12-month RR, it is never selected into the model in any time lag periods.

Putting economic variables from different time periods into the regression model separately, and comparing economic variables' performance in terms of t-value and P-value and models' performance based on models' R squares and Spearman ranking coefficients, the economic variables from 1 month before default are chosen and put into the model to predict 12-month RR. Using stepwise regression, only the 'Unemployment Rate' is left in the model with P-value less than 0.001. The predicted 12-month RR from this model is put into the final RR prediction model with 12-month RR as an independent variable (two-stage model). The two-stage model results are presented in Table 5.9. We can see, in the left side of Table 5.9 for 12-month RR prediction, the R square is improved from 0.0978 to 0.1005 (insignificant due to high p-value), but the Spearman ranking coefficient is decreased in the model with 'Unemployment Rate', and almost no change happens to MSE. In the right side of Table 5.9 for final RR prediction, both Spearman ranking coefficient and R square are decreased and MSE has a slight increase. The model comparison results in Table 5.9 show that economic variables make a small contribution to 12-month RR predictions but they are not helpful in two-stage model to predict final RR.

	Prediction of 12 months RR			Final Recovery Rate		
Economic Variables	Spearman	R square	MSE	Spearman	R square	MSE
No Economic Variables	0.29236	0.0978	0.0431	0.29501	0.0951	0.1645
Unemp Rate	0.29013 (0.8377)	0.1005 (0.6950)	0.0430 (0.9859)	0.29391 (0.9189)	0.0947 (0.9523)	0.1646 (0.9874)

Table 5.9 Model results of 12-month RR prediction and final RR prediction with economic variables (test sample)

	24 months Recovery Rate						
	6 month before	3 month before	1 month before	default month	6 month after	12 month after	18 month after
GDP Growth	0.11 (<.0001)	0.11 (<.0001)	0.11 (<.0001)	0.12 (<.0001)	0.12 (<.0001)	0.10 (<.0001)	0.07 (<.0001)
RPI Growth	-0.10 (<.0001)	-0.09 (<.0001)	-0.08 (<.0001)	-0.08 (<.0001)	-0.03 (<.0001)	0.002 (0.4348)	0.02 (0.0044)
Unemp Rate	0.002 (<.4742)	-0.03 (<.0001)	-0.05 (<.0001)	-0.06 (<.0001)	-0.09 (<.0001)	-0.11 (<.0001)	-0.11 (<.0001)
House Price	0.08 (<.0001)	0.09 (<.0001)	0.09 (<.0001)	0.09 (<.0001)	0.09 (<.0001)	0.10 (<.0001)	0.11 (<.0001)

Table 5.10 Pearson correlation coefficients between economic variables and 24-month RR

Table 5.10 is the Pearson correlation coefficients between economic variables from different time periods and 24-month RR. We can see that ‘GDP Growth’ and ‘House Price Index’ have relatively strong positive correlation with 24-month RR. With time period moving towards the right end, the sign of correlation coefficient of ‘RPI Growth’ changes from negative to positive, and ‘Unemployment Rate’ changes from 0 to negative. Putting the economic variables from 6 months before, 3 months before and 1 month before default into the 24-month RR prediction models separately, it is found that the model with economic variables of 6 months before default has the highest R square and also the economic variables which are selected by stepwise mechanism (‘GDP Growth’, ‘RPI Growth’, and ‘House Price’) have the correct signs (compatible with the signs of correlation coefficients in Table 5.10). ‘RPI Growth’ in other two models (models with economic variables of 3 months before and 1 month before default) has wrong signs (positive signs, which are against correlation coefficients in Table 5.10). ‘Unemployment Rate’ is not selected in any of 3 models. Therefore, we decide to put economic variables from 6 months before default into 24-month RR prediction model to predict 24-month RR, and then put the predictions into two-stage model to predict final RR. The model results are listed in Table 5.11.

	Prediction of 24-month RR			Final Recovery Rate		
Economic Variables	Spearman	R square	MSE	Spearman	R square	MSE
No Economic Variables	0.29533	0.1056	0.0879	0.28845	0.0911	0.1652
'GDP Growth' 'RPI Growth' 'House Price'	0.29752 (0.8431)	0.1086 (0.6760)	0.0877 (0.9427)	0.28687 (0.8847)	0.0896 (0.8185)	0.1655 (0.9391)

Table 5.11 Model results of 24-month RR prediction and final RR prediction with economic variables (test sample)

Table 5.11 shows the model results of 24-month RR prediction and final RR prediction. There are very similar stories with 12-month RR predictions. We can see in the left side of Table 5.11 for prediction of 24-month RR, the model with economic variables achieves a little insignificant improvement, but in the right side for final RR, the model with economic variables has an even worse performance compared with the model without economic variables. Therefore, the same conclusion can be made: economic variables help short RR predictions, but do not help predict final RR in the two-stage model.

5.7 Economic variables and remaining Recovery Rate after 12 and 24 months

In the previous chapter, it was shown that 12/24-month RR can help to predict the remaining RR after 12/24 months since default. In this section, a similar investigation will be done on whether economic variables help to predict the remaining RR.

For the remaining RR after 12 months, the economic variables from 6 months, 9 months, and 12 months after default are introduced into the model. The correlation coefficients between these economic variables and the remaining RR are calculated, but the results are disappointing. Only ‘Unemployment Rate’ has a relatively strong positive correlation (coefficient is 0.07 with p-value <.0001). The coefficients of other economic variables are less than 0.02. Also, the average values of these economic variables from 3 periods are worked out, and the correlation with remaining RR is still not obvious. After comparison and trying each set of them in the remaining RR model, economic variables from 6 months after default are chosen to put into remaining RR prediction model. Only ‘Unemployment Rate’ is selected by the stepwise regression model. The model results are in Table 5.12.

	Prediction of remaining RR after 12 months since default		
Economic Variables	Spearman	R square	MSE
No Economic Variables	0.40815	0.1609	0.1645
Unemp Rate	0.40982 (0.8720)	0.1613 (0.9661)	0.1644 (0.9947)

Table 5.12 Model results of predicting remaining RR after 12 months with economic variables (test sample)

From Table 5.12, we can see the model with ‘Unemployment Rate’ of 6 months after default has a little improvement. The Spearman ranking coefficient and the R square are increased a little, and MSE is lowed a little. However, compared with the model without economic variables, the p-values of three measures are high, which suggests this improvement is too small to be impressive.

For the remaining RR after 24 months since default, the economic variables from 18 months, 21 months, 24 months after default and their average values of this 6 months period are considered. The correlation coefficients between these economic variables and remaining RR are calculated. The same situation as with remaining RR after 12 months happens, except ‘Unemployment Rate’ has some positive correlation (0.09 with p-value <.0001). Other variables have almost no correlation with remaining RR after 24 months. The ‘Unemployment Rate’ from 18 months after default is better than that from other time periods. Thus it was chosen to put into the prediction model for remaining RR after 24 months. The models results are in Table 5.13. From Table 5.13, we can draw the same conclusion as remaining RR after 12 months: the improvement from economic variables is not significant.

	Prediction of remaining RR after 24 months since default		
Economic Variables	Spearman	R square	MSE
No Economic Variables	0.42729	0.1736	0.1824
Unemp Rate	0.43007 (0.7976)	0.1743 (0.9423)	0.1822 (0.9958)

Table 5.13 Model results of predicting remaining RR after 24 month with economic variables (test sample)

5.8 Conclusions

On this data set, economic variables appear not to be useful to predict final Recovery Rate. The research shows that there is no apparent correlation between economic variables and final RR, and the economic variables do not enter linear regression models of final RR prediction as important independent variables. The reason probably is that the recovery period spans a long time, usually 3 to 5 years and in some cases 8 or 9 years. Debtors' repayment behaviour is influenced by the current economic conditions, and it is not just related to the economic conditions at a certain time before or after default.

For the short term payments predictions, the economic variables have some correlation with the number of payments in 12 or 24 months after default. 'GDP Growth', 'RPI Growth', and 'House Price Index' are important variables in the prediction models for number of payments in 12 or 24 months. Thus the prediction results are improved in terms of Spearman ranking coefficient, R square, and MSE. However, better prediction of number of payments does not bring better prediction of final Recovery Rate. Putting the predicted number of payments in the two-stage models for RR, the prediction results do not have obvious improvements. Some two-stage models with economic variables are even worse than the ones without economic variables.

For the 12 or 24-month RR prediction, economic variables help to improve model results. In the 12-month RR prediction, ‘Unemployment Rate’ enters the model. In the 24-month RR prediction, ‘GDP Growth’, ‘RPI Growth’, and ‘House Price Index’ enter the model. They improve prediction accuracy to a little degree. Putting the predicted 12 or 24-month RR from models including economic variables into the two-stage models for final RR prediction, the prediction results for final RR from two-stage models with economic variables are not improved, compared with two-stage models without economic variables.

In the prediction of remaining RR after 12 or 24 months since default, ‘Unemployment Rate’ from 6 months after default and 18 months after default enter the prediction models of 12 months remaining RR and 24 months remaining RR respectively. The prediction results are slightly improved by the models with this economic variable.

So, economic variables are useful to predict short term number of payments and RR, but for final RR, they do not help, even when using two-stage models.

Chapter 6

Modelling Probability of Default for Invoice Discounting

6.1 Introduction

Invoice discounting is a form of short-term borrowing often used to improve a company's working capital and cash flow position. It allows a business to draw money against its sales invoices before the customer has actually paid. To do this, the business borrows a percentage of the value of its sales ledger from a finance company, effectively using the unpaid sales invoices as collateral for the borrowing. Invoice discounting has some difference from debt factoring. Invoice discounting only involves two parties: invoicing company and finance company. In debt factoring, the invoice sellers sell the invoices as receivable to factors (finance company), and the factors collect money from the invoice payers (debtors), so debt factoring involves three parties.

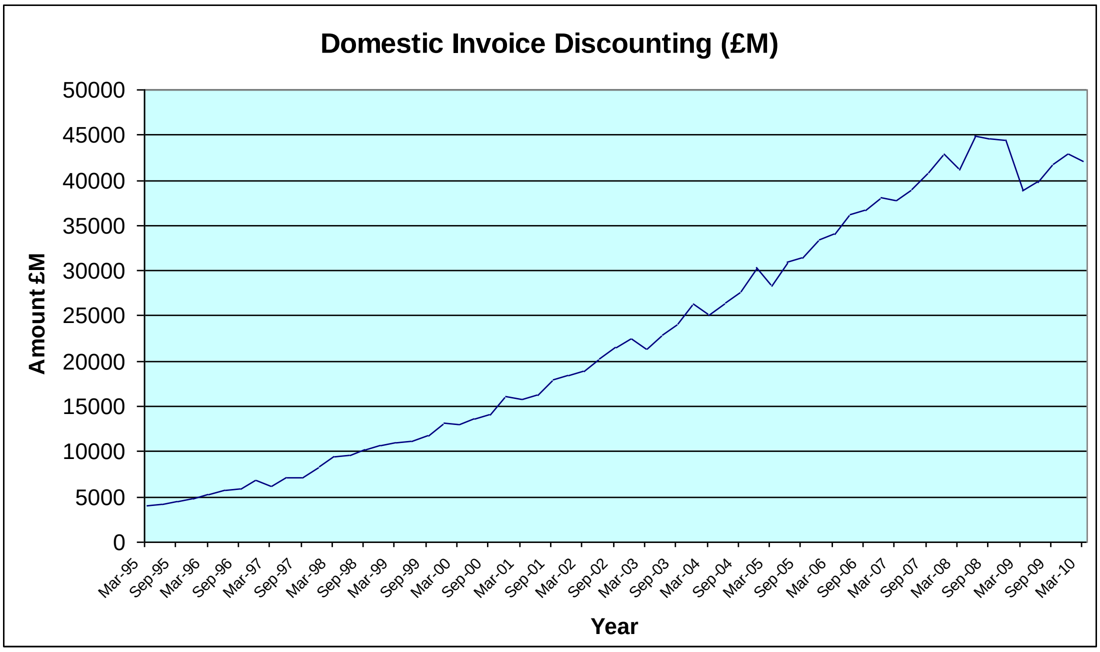

Figure 6.1 Amount of domestic invoice discounting (www.abfa.org.uk)

Figure 6.1 reflects the increasing trend of domestic invoice discounting from 1995 to 2010 in the UK. It is based on quarterly submissions of all members of Asset Based Finance Association (ABFA). We can see an obvious increasing trend from 1995 to early 2008. However, a small drop occurs in 2009 due to the financial crisis. The amount in Figure 6.1 probably does not represent the total amount of invoice discounting of the whole country (because it is only based on the members of ABFA), but the trend reflects the real increasing trend of invoice discounting in the UK. With the amount of borrowing increasing, the control of default risk of invoice discounting becomes more important.

In invoice discounting, default means the invoicing company defaults, at which point the bank cannot collect on the invoices. However unlike other corporate lending, the bank has very up to date information on the state of the firm, by seeing the value of the invoice being issued, and by observing financial statement submitted to the bank.

Invoice discounters do use scoring to assess the likelihood of default of businesses. Unlike consumer behaviour scoring, finance companies can close down accounts or seek further collateral from businesses, so they need to get the probability of default of the borrower not just its ranking. So, the Hosmer-Lemeshow test and Expected versus Actual number of defaulters are important, not just KS and GINI. The Hosmer–Lemeshow test is a statistical test for goodness of fit for logistic regression models. The test assesses whether or not the observed event rates match expected event rates in subgroups of the model population and it specifically identifies subgroups as the deciles of fitted risk values. With changes in economic environment in 2008 onwards, lenders found default predictions poor even though their scorecards continued to discriminate well. Can the predictions be improved? The strategy in this research is a) build a scorecard using data a finance company also used, in fact our results are slightly better than that of finance company; b) add economic variables to scorecard to improve predictions; c) add interactive economic-invoice behaviour variables to improve predictions; d) include multi-period variables to improve predictions; e) look at segmentation and build scorecard for each segment.

6.2 Data

6.2.1 Data description

The invoice discounting data set used in this project is provided by a major UK bank. It is panel data recording monthly information about invoicing and the company being invoiced. The records start from July 2003 and end at March 2009, and have the invoicing information on 5826 firms, among which 1184 default. The dependent variable is a binary variable, which describes whether the firm defaults within next 12 months (‘0’ means non-default, and ‘1’ means default). There are 75 independent variables, some of which are about the firms’ basic information, and some of which are about firms’ invoicing condition and sales ledgers.

Some good firms disappear from the data set, and no reasons are recorded. The yearly leaving rate is between 15% - 20%, so, when we make predictions for the number of defaulters, we should consider this factor. The default firms usually have 12 consecutive ‘1’s in dependent variable, because ‘1’ means default happens within the following 12 months. For example, if a firm defaults at August 2005, the dependent variable appears a ‘1’ from July 2004 to August 2005. Some firms have less than 12 ‘1’s; there are two reasons for this. One is errors where data is wrongly recorded; another reason is the data records end at March 2009, some firms default in the months after March 2009, therefore less than 12 ‘1’ are recorded. For example, a firm defaults in October 2009; the dependent variable starts ‘1’ from November 2008 and ends at March 2009, so only 5 ‘1’s are recorded. Some firms have more than 12 ‘1’s, the reason is that after they default, they still submit financial report to the bank, so the bank keeps recording the data. A small number of bad firms are cured in the end, which is found by some ‘0’s appearing after 12 ‘1’s. This research does not consider curing issue, so these observations are left out from the data set.

The quarterly default rate is reflected in Figure 6.2, where only the first ‘1’ of 12 ‘1’s in the default firms is counted, and other ‘1’s are ignored, so the report default rate leads the actual default rate by 12 months. The third quarter of 2003 is the first quarter in this data set, some ‘1’s in July 2003 do not exactly say default happens in June 2004, and some defaults which happened before June 2004 are recorded as ‘1’s in July 2003, but we can not tell this, so all the ‘1’s in July 2003 are counted, which make default rate in this quarter higher than average. We can see the default rate goes up from the fourth quarter of 2007, because the serious financial crisis happened in the second half year of 2008. The default rate suddenly goes down in the first quarter of 2009, which is counter intuitive. The reason is that this data set was released in February 2010, so the first quarter of 2009 does not include all the default information happening in the first quarter of 2010. This makes the first quarter of 2009 odd, and the data in this quarter is left out in model building.

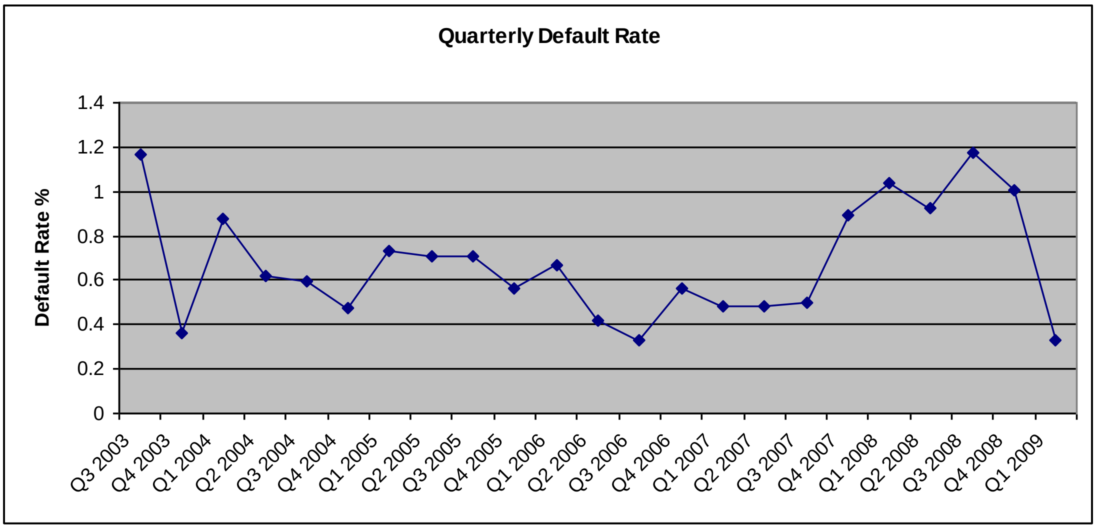

Figure 6.2 Quarterly default rate of invoicing firms

6.2.2 Data cleaning

The total number of observations in the data set is 173,542. Deleting the 'cured' observations and observations with lots of missing data, the remaining population is 157,883. Checking each variable and then deleting observations with outliers (extremely small or large values), cuts the sample to 137,271. These are the cases used to build the model.

6.2.3 Sampling

We use the data from July 2003 to December 2008 to build and test model, and we try to split the whole population into 3 parts: training sample used in model building, and in-time test sample and out-of-time test sample used in model test. Observations from July 2008 to December 2008 are kept untouched as out-of-time test population. All firms from July 2003 to June

2008 are randomly split into two parts; one part includes about 2/3 firms as training population, and remaining part includes about 1/3 firms as in-time test population. In both the training population and the in-time test population, the bad observations (‘1’s) only are 6 percent, and the ratio of good observations (‘0’s) to bad observations (‘1’s) is about 15:1. Samples with such a low bad rate are not that robust, so we decide to drop off some good observations and keep all the bad observations and try to make the ratio of good to bad is 3:1. Thus, in training population, all 1’s observations (4666) are kept, and we randomly select some of 0’s observations (14479, about 20% of the whole 0’s in training population) and the ratio of ‘0’ to ‘1’ is 3:1. These observations are the training sample and are used in model building. In in-time test population, we do the same thing; keep all 1’s observations (2247), and randomly select some of 0’s observations (7002, about 20% of the whole 0’s in in-time test population) to make the ratio of ‘0’ to ‘1’ be 3:1. For the out-of-time test sample, all observations from July 2008 to December 2008 are used, 12502 0’s (89.87%) and 1409 1’s (10.13%), and no sampling is made. These observations are used as out-of-time model test in order to test the model’s robustness for forecasting. Observations after December 2008 are left out in model building and test.

6.2.4 Variable transformations

Firstly, a univariate analysis is undertaken for each candidate variable. For categorical variables, the percentage of ‘1’s in each category of a variable is observed, and then the categories with similar percentage of ‘1’s are combined together to form a dummy variable. For the continuous variables, firstly they are coarsely classified into 15-20 groups according to their values in ascending order, and then we observe the bad rates in each group. For most grouped variables, we do not see the obvious monotone trend of bad rates, thus we consider to transform them to be binary variables as well. In the grouped variables, the adjacent groups are combined if they have similar bad rates and most continuous variables are transformed to between 3-6 binary variables. Two continuous variables, ‘UtilisationAverage90’ and ‘DebtTurnClient’, are used in ordinal format, which uses 15-20 numbers (1 to 15 or 20, same as coarsely classified groups) in one variable, because in their coarsely classified groups, there have strong monotone trends.

Some independent variables are strongly correlated, because they reflect the same features of a firm, and so the correlations between independent variables are tested. There are strong correlations between 17 pairs of variables, the highest correlation coefficient gets 0.95. For these correlated variables, the correlations of their transformed variables are tested again, and still obvious correlations exist. The correlated variables are put into a logistic regression model one by one to see their Wald Chi-square and P-value in order to choose the best one. After carefully comparing them, only one variable from each correlated variable group is selected for model building.

6.3 Model and model results

6.3.1 The model

Logistic regression is the most popular approach in predicting default, especially in financial industries. In this research, we also use logistic regression approach as the main method to predict default of invoicing companies. The form of logistic regression is

$$\log\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \qquad (1)$$

where, p is the probability of default

$x_1, x_{2,} ..., x_k$ are independent variables which describe characteristics of the invoicing account or the invoicing firm

$\beta_0, \beta_1, ... \beta_k$ are unknown parameters

ε is a random error term.

From equation (1), we can derive p, the probability of default

$$p = \frac{1}{1 + \exp - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)} \tag{2}$$

However, sampling was done for the training sample and the in-time test sample, but not done for the out-of-time test population, thus the equation (1) and (2) should be adjusted to predict in out-of-time test sample. For instance, if we only keep α percent of good observations in model training, when the model is used to make predictions in the out-of-time test sample, the formula is adjusted so that

$$\log\left[\frac{p}{\alpha(1-p)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \tag{3}$$

$$p = \frac{1}{1 + \frac{1}{\alpha} \exp - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)} \tag{4}$$

In order to realise this adjustment in the out-of-time test sample, if the model trained from the sampled training sample gives probability of default p, the actual probability of default p' in out-of-time test sample is

$$p' = \frac{1}{1 + \frac{1}{\alpha}(\frac{1}{p} - 1)} \tag{5}$$

6.3.2 Model results

The independent variables used in building the model are the transformed variables created before, and they are transformed from the variables already given in the data set, we do not create new variables or add external variables at this moment. ‘Stepwise’ is set as variable selection mechanism, and this mechanism selects 38 variables whose P-values are less than 0.05. According to the bank’s suggestions, only the most significant 20 variables are kept in the model. The variables coefficients of the model (Model 1) are in Table 6.1.

All the 20 variables selected in the model have P-values less than 0.0001, which indicates all of them are very significant. We define default as ‘1’ in the target variable, and this model predicts ‘1’s, so the positive coefficients mean more likely to default and negative coefficients mean safer. From Table 6.1, we can see that if the account has longer history, it is less likely to default, because ‘Account_Duration_3’ has a larger negative coefficient than ‘Account_Duration_2’. ‘AdditionalABLLine_1’ describes a supplementary product, it is a binary variable and some firms bought it (set as 1) and some firms did not (set as 0). It has a positive coefficient, which means the firms which bought it are more likely to default. ‘ATTrend3m’ is ‘Annualised_Turnover_Trend_3months’, which is calculated by dividing the average value of estimated annual sales turnover last 3 months by the estimated value of last month. So the bigger ratio means the turnover goes down. From the coefficients for ‘ATTrend3m’, we can see the bigger ratio, the more likely to default, which indicates the estimated annual turnover going down leads to default more easily. ‘BandName’ is a binary variable, and it describes two types of band name: ‘Corporate’ or ‘Global’. The coefficient reflects that ‘Global’ firms are less likely to default than ‘Corporate’ firms.

```
                                          Standard   Wald
Parameter                   Estimate        Error    Chi-Square    Pr > ChiSq

Intercept                    -4.6399       0.0869     2854.0756        <.0001

Account_Duration_2           -0.3433       0.0520       43.5183        <.0001
(1235-2081 days)
Account_Duration_3           -0.5092       0.0508      100.2635        <.0001
(>2081 days)
AdditionalABLLine_1           0.5352       0.0639       70.0789        <.0001
(A supplementary product. Yes)
ATTrend3m_2                   0.2117       0.0524       16.3572        <.0001
(AnnualisedTurnoverTrend3months, 1.01-1.04)
ATTrend3m_3                   0.4154       0.0624       44.2631        <.0001
(AnnualisedTurnoverTrend3months, 1.05-1.08)
ATTrend3m_4                   0.6215       0.0886       49.1614        <.0001
(AnnualisedTurnoverTrend3months, 1.09-1.11)
ATTrend3m_5                   0.7471       0.0611      149.3005        <.0001
(AnnualisedTurnoverTrend3months, >1.11)
BandName_1                   -0.4099       0.0827       24.5987        <.0001
(‘Corporate’ or ‘Global’; ‘Global’)
Bank_1                        1.2858       0.0497      668.2168        <.0001
(Banking with this bank or others; ‘Others’)
DisapprovalsAge_01            0.2109       0.0405       27.1083        <.0001
(Disapprovals due to ‘age’; >5.61)
EntitleAvgTr3m_01             0.2531       0.0573       19.4894        <.0001
(Entitlement average trend 3 months; >1.13)
FinancialRating_2             0.2855       0.0455       39.3415        <.0001
(‘6’)
FinancialRating_3             0.5249       0.0509      106.5264        <.0001
(‘7’,’8’,’9’)
LedgerDiffer_01               0.2295       0.0416       30.4531        <.0001
(Ledger Difference; <=0.97)
PayColleRatio_1               0.4121       0.0844       23.8346        <.0001
(Payment Collection Ratio; <=0.53)
Top5PcDebtors_1              -0.5818       0.0817       50.6893        <.0001
(Top 5 biggest debtors’ percentage of the current Sales Ledger; <=22.22)
Top5PcDebtors_2              -0.2743       0.0623       19.4128        <.0001
(Top 5 biggest debtors’ percentage of the current Sales Ledger; 22.23-34.96)
Top5PcDebtors_4               0.4061       0.0442       84.2220        <.0001
(Top 5 biggest debtors’ percentage of the current Sales Ledger; >=62.65)
UtilisationAve90_0            0.1499      0.00460     1064.0180        <.0001
(Utilisation average in last 3 months; ordinal variable, 1-18)
DebtTurnClient_0              0.0341      0.00502       46.1980        <.0001
(Ordinal variable, 1-16)
```

Table 6.1 Model (1) Logistic regression model variables and coefficients

‘Bank’ is also a binary variable; the firms banking with other banks rather than this bank are more likely to default. ‘DisapprovalsAge’ describes the amount of disapproved invoices because of the reason ‘Age’. We can see that the firms with larger disapproved amount are more likely to default. ‘EntitleAvgTr3m’ is the ratio of average entitlement value last 3 month to the entitlement value last month. We can see that the larger this ratio, the more likely to default, which says firms whose entitlement value going down are easier to default. ‘FinancialRating’ is a measure of the client financial strength based on their accounts. It has 9 bands, ‘1’ is the least risk, and ‘9’ is the most risk. We can see ‘FinancialRating_3’, which stands for band 7, 8 and 9, has a high positive coefficient and this is consistent with the logic. ‘LedgerDifference’ is a measure of reconciliation errors, we can see that firms with this value less than or equal to 0.97 have higher risk. ‘PaymentCollectionRatio’ is calculated by dividing the sum of payments (the money bank paid to firms) in last month by the sum of cash (the money collected from firms) in last month. The ratio less than 0.53 has higher risk than others. ‘Top5PcDebtors’ is the percentage of the current SalesLedger that is represented by the top 5 biggest debtors; high concentration to a few large debtors is considered to be high risk. The coefficients for this variable confirm to this logic. ‘Top5PcDebtors_1’ which describes the percentage less than 22.22 has a big negative coefficient, and is the least likely to default. ‘Top5PcDebtors_5’ which describes the percentage greater than 62.5 has a big positive coefficient, and is the most likely to default. ‘UtilisationAve90’ is calculated by dividing the average amount of overdraft in last 3 months by the average amount of entitlement in last 3 months. This variable is an ordinal variable, and has 18 values from 1 to 18 to represent 18 bands. The positive coefficient indicates the higher the band is, the more likely to default. ‘DebtTurnClient’ describes how many days it takes assignments to cover the closing sales ledger. It is also an ordinal variable and has 16 values from 1 to 16 to represent 16 bands. The positive coefficient indicates the higher the band, the more likely to default. All the signs and values of these 20 variables are checked with the univariate analysis, and they do not contradict with each other. The extremely significant variables are ‘UtilisationAvg90’ and ‘Bank’, they have large Wald Chi-Square value.

	Training	In time Test	Out of time Test
Gini	0.62	0.63	0.60
KS	46.46	48.83	46.34
Hosmer-Lemeshow test (Chi square and p-value)	43.16 (<.0001)	26.93 (<.001)	1470.94 (<.0001)
Actual number of defaults	4666	2247	1409
Expected number of defaults	4666	2201	605

Table 6.2 Model (1) Logistic regression model measurement results

Table 6.2 lists the measurements such as Gini, KS, and Hosmer-Lemeshow test. From Gini and KS, we can see the model performance is quite good in all of the training, in-time test, and out-of-time test samples. This tells us the model has no obvious overfitting. The only flaw is that Gini is a little bit lower in out-of-time test sample than that in training sample and in-time test sample. However, from Hosmer-Lemeshow test, we can see that the Chi-square in the out-of-time test sample is much higher (1470.94) than that in the training sample and the in-time test sample. This indicates that the predicted number of defaulters in each 10 decile does not match the actual number of defaulters in the 10 deciles well. The problem is the expected number of defaulters (the sum of the predicted probability of default of every observation) given by model predictions is much smaller than the actual number of defaulters in the out-of-time test sample. We can see that in the out-of-time test sample, the expected number of defaulters are 605, which is much lower than 1409 of actual number of defaulters. In the training and the in-time test samples, the number of expected defaulters is the same as or very close to the number of actual defaulters. This is the reason why the Chi-squares of Hosmer-Lemeshow tests are not that high in the training sample and the in-time test sample. The expected number of defaulters in the out-of-time test sample is the adjusted number, where the number predicted from the model is multiplied by 0.85, as we found using historical data that on average of 15% companies would stop using invoice discounting within one year. Thus multiplying by a staying rate coefficient is necessary. With this calculation, the expected number of defaulters is 605. To calculate a confusion matrix, we categorise the 605 borrowers with the highest probability of default as defaulters, and the remaining 13306 as non-defaulters. Checking who actually defaulted, leads to the confusion matrix in Table 6.3, where the accuracy is 88.4%, however, only 19.66% of the defaulters are predicted as defaulters.

Frequency Percent Row Percent Column Percent		Predicted		
		0	1	Total
Actual	0	12148 87.40 97.17 91.30	354 2.54 2.83 58.51	12502 89.87
	1	1158 8.32 82.19 8.70	251 1.80 17.81 41.49	1409 10.13
	Total	13306 95.65	605 4.35	13911 100.00

Table 6.3 Confusion Matrix for out-of-time sample predictions from Model 1

6.4 Incorporating macroeconomic conditions

6.4.1 Adding macroeconomic variables

Because the expected number of defaulters predicted by Model (1) in the out-of-time sample is much fewer than the actual number of defaulters, we consider introducing some economic variables into the model in order to deal with this issue. Macroeconomic conditions indeed affect default risk considerably not only in company lending, but also in consumer lending. In this research, we want to see how the macroeconomic conditions influence the default risk of invoice discounting.

The economic variables we considered are:

GDP Growth: Quarter on quarter previous year change, seasonally adjusted

RPI Growth: Retail Prices Index, percentage change over 12 months

Unemployment Rate: All aged 16 and over, percentage, seasonally adjusted

Interest Rate: Bank of England interest rate

Production Index: Production Index for manufacturing industries, seasonally adjusted

Business Confidence Index: measuring overall finance professionals' sentiments toward the short-term future economic situation; from the Institute of Chartered Accountants

FTSE All-share: The monthly highest prices of FTSE All-Share

These economic variables are easily obtained and some of them have been used in academic research. 'GDP', 'RPI', 'Unemployment Rate', and 'FTSE' are suggested as important in consumer finance by Tang et al (2007), Liu and Xu (2003). 'GDP', 'RPI', 'Unemployment Rate', 'Interest Rate', and 'FTSE' are investigated by Figlewski et al (2007) in corporate credit risk, and they find they have influences on corporate credit risk. We consider additional another two economic variables, 'Production Index', and 'Business Confidence Index', and think they might be related to manufacturing firms and small-and-medium-sized enterprises, because these firms make up a large proportion of all invoicing companies in the data set. 'Production Index' gives an insight into how the manufacturing sector is performing and this may be relevant to other service forms which are invoicing such manufactures. 'Business Confidence Index' gives a country wide view of what managers are thinking may be the future for their organisations. Among these economic variables, 'GDP' and 'Business Confidence Index' are quarterly data, other variables are monthly data. 'FTSE All-share' is the monthly highest price. It is standardised when adding in the models due to its large values. There are certain correlations between these economic variables. Table 6.3 lists the Pearson correlation coefficients between each economic variable (From 2003 to 2009).

	Unemp rate	*RPI*	*GDP*	*Interest*	*Produc Index*	*Business Confidence Index*	*FTSE*
Unemp rate	1						
RPI	-0.71	1					
GDP	-0.93	0.76	1				
Interest	-0.80	0.94	0.85	1			
Produc Index	-0.84	0.88	0.92	0.95	1		
Confidence Index	-0.83	0.51	0.89	0.61	0.74	1	
FTSE	-0.08	0.53	0.29	0.60	0.55	0.04	1

Table 6.4 Pearson correlation coefficients between economic variables (2003-2009)

From Table 6.4, we can see that except ‘FTSE’, all the other economic variables have strong correlation between each other, which suggests the economic variables consistently reflect the same macroeconomic conditions. Because of this strong correlation, we can not put all the economic variables in the model at the same time.

To choose the best economic variables, first of all, each economic variable is put in the model one by one together with the previous important 20 variables. We do not consider time lag effects at this stage, thus for each observation, the 7 economic variables are from the month the observation is in. ‘GDP’ and ‘Business Confidence Index’ have negative signs, with high Wald Chi-Square (greater than 100) and extremely small p-value (smaller than 0.0001). The negative sign is reasonable and accords with the logic, which says good economic environment with high GDP growth and high Business Confidence usually leads to low default rate. ‘Production Index’ and ‘FTSE’ have negative signs but the Wald Chi-Squares are not that high (around 20) and P-Values are not smaller than 0.0001. However, ‘Unemployment Rate’, and ‘Interest Rate’ are not selected by stepwise system with 0.05 as entry and removal significance level. ‘Retail Price Index’ is selected, but with a positive sign. This is counter intuitive, as we would expect increases in RPI lead to a lowing in default rate and here a highing is being suggested.

Secondly, all the 7 economic variables are put in the model at the same time together with the other 20 variables, ‘FTSE’ and ‘Business Confidence’ are very significant with negative signs and extremely small P-Value; ‘Production Index’ and ‘GDP’ are also selected in, but their signs are counter intuitive (positive signs) and are not highly significant. Considering all the factors above, we decide to use two versions of economic variables putting in the logistic regression model; one version includes ‘Business Confidence Index’ and ‘FTSE’, and the other version has ‘GDP’ only.

```
                                            Standard     Wald
             Parameter          Estimate       Error     Chi-Square     Pr > ChiSq

Version 1:   Confidence         -0.0258      0.00236      119.3080        <.0001
(Model 2)     FTSE_st           -0.1509       0.0251       36.1971        <.0001

Version 2:       GDP            -0.2899       0.0384       56.9679        <.0001
(Model 3)
```

Table 6.5 Coefficients of economic variables in the model (Model2 and Model 3)

Version 1	Training	In time Test	Out of time Test
Gini	0.63	0.63	0.59
KS	47.12	49.00	48.80
Hosmer-Lemeshow test (Chi square and p-value)	32.51 (<.0001)	29.95 (<.001)	63.81 (<.0001)
Actual number of defaults	4666	2247	1409
Expected number of defaults	4666	2202	1306

Table 6.6 Model 2 results with ‘Business Confidence Index’ and ‘FTSE’ in the model

Version 2	Training	In time Test	Out of time Test
Gini	0.63	0.63	0.59
KS	46.69	49.06	44.33
Hosmer-Lemeshow test (Chi square and p-value)	36.38 (<.0001)	35.32 (<.0001)	66.44 (<.0001)
Actual number of defaults	4666	2247	1409
Expected number of defaults	4666	2207	1321

Table 6.7 Model 3 results with 'GDP' in the model

Table 6.5 presents coefficients of economic variables of two versions. All the 20 variables selected in Model (1) (excluding economic variables) are also selected in the Model (2) (including 'Confidence index' and 'FTSE') and Model (3) (including 'GDP growth'), and their coefficients are very similar to those in Model (1), so they are not listed here again. 'GDP', 'Business Confidence Index' are their original values; 'FTSE' is the standardised form (original value minus the average value and divided by the standard deviation), because their original values are too large. All three economic variables have negative signs, which indicate higher 'GDP', 'Business Confidence Index', and 'FTSE' lead to lower default risk; and this is consistent with logic. These two new models achieve very similar performance (see Table 6.6 and Table 6.7). The Gini, KS and Chi-square of Hosmer-Lemeshow test are very similar between the two models. Compared with Model (1) (Table 6.2), KS coefficients in Model (2) and (3) are almost the same as those in Model (1); Gini in the training sample has a little increase, but in the out-of-time test sample, it drops a little. In terms of Hosmer-Lemeshow test, the Chi-square value in the out-of-time test sample has been improved considerably, as it drops from 1470 to 63 and 66. The largest improvements are seen from the expected number of number of defaulters in the out-of-time test sample. Model (1) without economic variables only predicts 605 defaulters, but Model (2) and Model (3) with economic variables predict more than 1300 defaulters, which is very close to the actual number of defaulters. Thus including economic variables has improved the predicted default rate considerably without losing much on discrimination. The confusion matrixes (Table 6.8 and Table 6.9) are produced for out-of-time sample predictions from Model 2 and Model3. The percent of actual defaulters correctly predicted as defaulters (sensitivity) is 31.51% and 31.94%, which are improved from Model 1.

Frequency Percent Row Percent Column Percent		Predicted		
		0	1	Total
Actual	0	11476 83.67 93.11 92.34	862 6.20 6.89 66.00	12502 89.87
	1	965 6.94 68.49 7.66	444 3.19 31.51 34.00	1409 10.13
	Total	12605 90.61	1306 9.39	13911 100.00

Table 6.8 Confusion Matrix for out-of-time sample predictions from Model 2

Frequency Percent Row Percent Column Percent		Predicted		
		0	1	Total
Actual	0	11631 83.61 93.03 92.38	871 6.26 6.97 65.93	12502 89.87
	1	959 6.89 68.06 7.62	450 3.23 31.94 34.07	1409 10.13
	Total	12590 90.50	1321 9.50	13911 100.00

Table 6.9 Confusion Matrix for out-of-time sample predictions from Model 3

6.4.2 Interactions between economic variables and other variables

In the last section, we saw the macroeconomic conditions indeed affect the default risk of invoicing companies. In this section, we investigate that whether there are interactions between economic variables and other variables, and so the economic conditions affect firms with different characteristics in different ways. It is worthwhile to see whether interactions exist between economic variables and firm characteristics and how the interactions influence models' performance.

Interaction variables are created by multiplying the dummy variables made before with economic variables. For example, 'Account Duration' is cut into 4 ranges and is expressed by 3 dummy variables and one reference variable. The interaction variables of 'Account Duration' with 'GDP' are created by multiplying the 3 dummy variables with 'GDP', and multiplying reference variable (when 3 dummies equal to 0) with 'GDP'. Thus, 4 interaction variables between 'Account Duration' and 'GDP' are constructed. A similar approach is undertaken for other transformed dummy variables and the economic variables ('Business Confidence Index', 'FTSE', and 'GDP'). This leads to 55 interaction variables.

Using as a base Model (2) (including 'Business Confidence Index' and 'FTSE-all share'), the interaction variables of 'Business Confidence Index' and 'FTSE-all share' with other variables are added, and the stepwise approach is used to select the most important ones. 4 interaction variables are selected; the details of model (Model (4)) can be seen in Table 6.10.

In Table 6.10 for Model (4), we can see the 20 important variables and the 2 economic variables which were selected before are still in the model, and their signs are unchanged and values are just slightly adjusted. 4 interaction variables (the last 4 variables in Table 6.10, the explanations for each variable are given) are selected, but they are not extremely significant in terms of Wald Chi-square. From Table 6.11, we can see that in the out-of-time test sample, the Gini goes down to 0.57 compared with 0.59 in Model (2), which suggests the ranking is a little worse when using interaction variables. The good thing is the expected number of defaulters is increased to 1383, which is even closer to the actual number.

Using as a base Model (3) (including ‘GDP’), the interaction variables of ‘GDP’ with other variables are added, and the stepwise approach selects the most important ones. 4 interaction variables are selected; the details of the model (Model (5)) can be seen in Table 6.12.

```
                                              Standard        Wald
Parameter              DF     Estimate         Error    Chi-Square    Pr > ChiSq

Intercept               1      -4.6270        0.0897     2660.2453        <.0001
Account_Duration_2      1      -0.2911        0.0538       29.2601        <.0001
Account_Duration_3      1      -0.4616        0.0523       77.9699        <.0001
AdditionalABLLine_1     1       0.5014        0.0648       59.9159        <.0001
ATTrend3m_2             1       0.1984        0.0527       14.1771        0.0002
ATTrend3m_3             1       0.4025        0.0629       40.9477        <.0001
ATTrend3m_4             1       0.5958        0.0893       44.5324        <.0001
ATTrend3m_5             1       0.7215        0.0619      135.9351        <.0001
BandName_1              1      -0.4233        0.0831       25.9798        <.0001
Bank_1                  1       1.3327        0.0513      675.7101        <.0001
DisapprovalsAge_01      1       0.2268        0.0408       30.9586        <.0001
EntitleAvgTr3m_01       1       0.2556        0.0578       19.5745        <.0001
FinancialRating_2       1       0.2925        0.0461       40.2734        <.0001
FinancialRating_3       1       0.5443        0.0513      112.5567        <.0001
LedgerDiffer_01         1       0.2139        0.0419       26.0984        <.0001
PayColleRatio_1         1       0.4251        0.0849       25.0563        <.0001
Top5PcDebtors_1         1      -0.5595        0.0821       46.4677        <.0001
Top5PcDebtors_2         1      -0.2908        0.0627       21.5286        <.0001
Top5PcDebtors_4         1       0.4049        0.0446       82.4601        <.0001
UtilisationAve90_0      1       0.1545       0.00477     1049.4924        <.0001
DebtTurnClient_0        1       0.0340       0.00505       45.3418        <.0001
Confidence              1      -0.0353       0.00302      136.0526        <.0001
FTSE_st                 1      -0.1308        0.0257       25.9569        <.0001
AD_Conf                 1       0.0173       0.00361       23.0917        <.0001
(Account Duration =<829; dummy*Confidence)
LD6_Conf                1      -0.0329       0.00896       13.4535        0.0002
(LedgerDifference 0.98; dummy*Confidence)
FR1_FTSE                1      -0.2828        0.0825       11.7578        0.0006
(FinancialRatingCode 1,2; dummy*FTSE)
UA2_Conf                1       0.0213       0.00532       16.0582        <.0001
(UtilisationAve90; dummy*Confidence)
```

Table 6.10 Coefficient details of Model (4) with interaction variables of 'Confidence' and 'FTSE'

	Training	In time Test	Out of time Test
Gini	0.63	0.63	0.57
KS	47.03	49.27	44.17
Hosmer-Lemeshow test (Chi square)	31.75 (<.001)	21.49 (<.01)	81.69 (<.0001)
Actual number of defaults	4666	2247	1409
Expected number of defaults	4666	2204	1383

Table 6.11 Measurement results of Model (4) with interaction variables of 'Confidence' and 'FTSE'

```
                                          Standard        Wald
Parameter              DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept               1     -4.1495      0.1026     1634.6326        <.0001
Account_Duration_2      1     -0.3352      0.0523       41.0621        <.0001
Account_Duration_3      1     -0.4868      0.0510       91.1876        <.0001
AdditionalABLLine_1     1      0.5274      0.0642       67.5592        <.0001
ATTrend3m_3             1      0.2359      0.0681       12.0196        0.0005
ATTrend3m_4             1      0.4343      0.0926       22.0028        <.0001
ATTrend3m_5             1      0.5356      0.0677       62.6701        <.0001
BandName_1              1     -0.4509      0.0830       29.5010        <.0001
Bank_1                  1      1.3274      0.0502      697.9842        <.0001
DisapprovalsAge_01      1      1.0059      0.1488       45.6907        <.0001
FinancialRating_2       1      0.3043      0.0458       44.1984        <.0001
FinancialRating_3       1      0.5328      0.0510      108.9626        <.0001
LedgerDiffer_01         1      0.2180      0.0418       27.2059        <.0001
PayColleRatio_1         1      0.3719      0.0850       19.1393        <.0001
Top5PcDebtors_1         1     -0.5412      0.0819       43.6524        <.0001
Top5PcDebtors_2         1     -0.2719      0.0625       18.9302        <.0001
Top5PcDebtors_4         1      0.3925      0.0444       78.1872        <.0001
UtilisationAve90_0      1      0.1528     0.00462     1094.8247        <.0001
DebtTurnClient_0        1      0.0187     0.00586       10.1372        0.0015
ATT_GDP                 1     -0.0779      0.0186       17.5308        <.0001
(AnnualisedTurnoverTrend3months 0.91-1, dummy*GDP)
DTC5_GDP                1      0.1359      0.0271       25.2113        <.0001
(DebtTurnClient >106, dummy*GDP)
EAT_GDP                 1     -0.1053      0.0206       26.1081        <.0001
(EntitlementAvgTrend3months =<1.13, dummy*GDP)
DA01_GDP                1     -0.3086      0.0551       31.4203        <.0001
(DisapprovalAge >5.61, dummy*GDP)
```

Table 6.12 Coefficient details of Model (5) with interaction variables of 'GDP'

	Training	In time Test	Out of time Test
Gini	0.63	0.63	0.54
KS	46.78	48.65	39.43
Hosmer-Lemeshow test (Chi square)	43.64 (<.0001)	24.27 (<.005)	146.0 (<.0001)
Actual number of defaults	4666	2247	1409
Expected number of defaults	4666	2212	1309

Table 6.13 Measurement results of Model (5) with interaction variables of 'GDP'

In Table 6.12 for Model (5), we can see that the basic 20 variables are still in the model, but ‘GDP’ is not selected in; 4 interaction variables (the last 4 in Table 6.12) are selected in the model and all of them have p-value smaller than 0.0001, but one of them with a positive sign, which is counter intuitive, maybe because there are interactions between the interaction variables. From Table 6.13, we can see that the Gini and KS in training and in-time test sample still have similar performance as before, but in the out-of-time test sample Gini goes down further to 0.54 and KS drops to 39.43. This suggests interaction variables make the discrimination poorer. They also do not improve the predicted number of defaults. Thus the confusion matrix is not produced for the two models.

The interaction variables created before in this section do not effectively improve model’s performance, thus we consider to create some other interaction variables. The trend variables in the data set reflect the firms’ dynamic economic situation, and the macroeconomic trend variables reflect the dynamic macro economic environment. So, it is a good idea to create some dummy variables to describe the changes of both firms’ and macro economic situations.

In the previous sections, ‘GDP growth’ added in the model is a yearly growth rate, which means comparing GDP of this quarter with that of 4 quarters ago. It is worthwhile to compare GDP of this quarter with that of the last quarter, and we call it ‘GDP quarterly growth’. ‘GDP quarterly growth’ is always increasing from 1989 until the 2008 economic recession. From the second quarter of 2008, ‘GDP quarterly growth’ became negative. Thus, if we use ‘GDP quarterly growth’ to create new interaction variables, this leads to a problem: in the training and in-time test sample, ‘GDP quarterly growth’ is positive, except in the last quarter, but in the out-of-time test sample, ‘GDP quarterly growth’ is negative. So, the interaction variables trained from training sample do not work properly in out-of-time test sample. Therefore, ‘GDP quarterly growth’ was not used to create this kind of interaction variables. ‘Confidence Index’ quarterly change does not have this problem. It has negative values even in non-recession period. We use ‘Confidence Index’ quarterly change to create some interaction variables with other trend variables which are already available in the data set or can be easily generated from original variables in the data set. Taking ‘Confidence Index’ quarterly change and ‘CashAverage’ 3 months trend as an example, 4 interaction variables are created and the logic is as following:

If CashAvg30 – CashAvg90 >=0 and Confidence Index quarterly change >=0, then take this situation as reference variable;
If CashAvg30 – CashAvg90 <0 and Confidence Index quarterly change >=0, then create a dummy CashAvg_Conf1=1; else Cash_avg_Conf1=0;
If CashAvg30 – CashAvg90 >=0 and Confidence Index quarterly change <0, then create a dummy CashAvg_Conf2=1; else Cash_avg_Conf2=0;
If CashAvg30 – CashAvg90 <0 and Confidence Index quarterly change <0, then create a dummy CashAvg_Conf3=1; else Cash_avg_Conf3=0;

36 new interaction variables from 12 interacting variables in the data set with ‘Confidence Index’ are created and they are put in Model (2) (inclulding ‘Confidence Index’ and ‘FTSE’). However, the results are disappointing. Only 3 new interaction variables are selected in the model, and none of them have large Wald Chi-square value and extremely small p-values. It is not expected that Gini, KS, and expected number of defaulters will change much from Model (2), so these measurements are not calculated. Since this kind of interaction variables of ‘Confidence Index’ are not helpful, interaction variables of other economic variables are not created; we don’t expect they will be better than interaction variables of ‘Confidence Index’.

6.4.3 Using economic variables one quarter/month before

In the models built before, the values of economic variables are the ones which occur in the quarter or month of the data extract. But usually, in reality, economic variables are published a couple of months later, so in practical work, we do not know the values of current economic variables, thus it is difficult to use them. For this reason, we consider using the economic variables in the previous quarter or month before the data extract, and to see the model performance in this case.

‘GDP’ and ‘Business Confidence Index’ are quarterly data, so we use their last quarter values. Other economic variables are monthly data, and their last month values are used. All 7 economic variables from last quarter or month are carefully investigated, and ‘Business Confidence Index’, ‘FTSE-all share’ and ‘GDP’ are still the most significant variables. Thus, we do the same thing as before: put both ‘Confidence’ last quarter and ‘FTSE’ last month in the model as one version of economic variables, and put ‘GDP’ last quarter in the model as another version of economic variables, and still keep the basic most important 20 variables.

```
                                              Standard         Wald
              Parameter          Estimate        Error   Chi-Square    Pr > ChiSq

Version 1     Confidence1         -0.0244      0.00314      60.6717        <.0001
(Model 6)     FTSE1_st            -0.1155       0.0263      19.3284        <.0001


Version 2      GDP1               -0.1147       0.0454       6.3867        0.0115
(Model 7)
```

Table 6.14 Coefficients of economic variables of last quarter or month in the models

Model 6	Training	In time Test	Out of time Test
Gini	0.63	0.63	0.60
KS	46.63	49.20	45.51
Hosmer-Lemeshow test (Chi square and p-value)	35.57 (<.0001)	40.72 (<.0001)	195.96 (<.0001)
Actual number of defaults	4666	2247	1409
Expected number of defaults	4666	2200	1077

Table 6.15 Measurement results of Model (6) with ‘Business Confidence Index’ from last quarter and ‘FTSE’ from last month in the model

Model 7	Training	In time Test	Out of time Test
Gini	0.62	0.63	0.60
KS	46.60	48.94	45.79
Hosmer-Lemeshow test (Chi square and p-value)	45.77 (<.0001)	31.91 (<.001)	948.17 (<.0001)
Actual number of defaults	4666	2247	1409
Expected number of defaults	4666	2202	714

Table 6.16 Measurement results of Model (7) with 'GDP' from last quarter in the model

From Table 6.14 we can see the 3 economic variables still have negative sign, which accords with logic. 'Confidence' and 'FTSE' are still very significant with extremely small P-values, but 'GDP' is not so significant any more, the P-value is 0.0115. Other 20 basic important variables are still in the models, they are still important, and their signs and coefficients are almost the same as before, so they are not listed here again. Table 6.15 and Table 6.16 are the models measurement results. We can see that in the training sample and the in-time test sample Gini and KS are as good as before. Even in the out-of-time test sample these two coefficients are still good, and no obvious drop compared with Model (4) and Model (5). However, the expected number of defaulters drops a lot in the out-of-time test samples of both models, especially in Model (7) with 'GDP' last quarter, it drops to 714. This tells us that the economic variables from last quarter or month do not properly describe current economic situation. In terms of expected number of defaulters, Model (6) with 'Confidence' and 'FTSE' is better than Model (7) with 'GDP, which suggests that, 'Confidence' and 'FTSE' are more forward looking estimats than 'GDP', and 'GDP' only reflects what has happened and can't tells us what will happen next.

6.4.4 New trend variables

Based on the variables already given in the data set, we can create some new variables which can better reflect the change trend of firms' business conditions. There have already been some trend variables in the data set, so we consider generating something different. We pick up 9 variables to create new variables, and they are:

'EntitlementAvg30'
'Exposure'
'FinancialRatingCode'
'LedgerDifference'
'LTVALL'
'NetAssignmentsAvg30'
'PaymentCollectionRatio'
'Top1PcDebtors'
'Top5PcDebtors'

3 types of trend variable are created for each selected variable: the ratio of the value of this month to the value of last month; the ratio of the value of this month to the average value of last 3 months; the ratio of the value of this month to the average value of last 5 months. 'EntitlementAvg30' and 'NetAssignmentsAvg30' already had trend variables in the data set, but the new ones created here are different. For the 'FinancialRatingCode', the new trend variables are created by recording whether the rating codes upgrade or downgrade in the last 3 or 5 months, so 2 new variables are created. Thus, totally 28 new trend variables are created.

Correlations between 28 new trend variables are checked and it is found that there are correlations between some variables, especially those of the 3 month trend and 5 month trend of the same variable. Thus, a univariate analysis is made, in order to find and keep the best one from the correlated variables. In the univariate analysis, all the 28 new trend variables are transformed to ordinal variables with 15-20 groups for each variable. The percentage of ‘good/bad’ observations in each group is calculated, and the adjacent groups are combined if they have similar percentages so as to form dummy variables; it is also noticed that the trend variables of ‘Top1PcDebtors’ and ‘Top5PcDebtors’ have no variation between groups when classifying ‘good/bad’ observations, so they are left out. In the end, after leaving out correlated and ineffective variables, 35 dummy variables are created. In order to choose the most important variables from the 35 dummy variables, we put only these 35 variables in a logistic regression model. 24 variables enter the model, and some of them are very significant with extremely small P-value. The 12 strongest dummy variables from these 24 variables are selected according to their p-value and the order of entering the model, and then we put these 12 variables together with the other 20 basic variables selected before in the logistic regression model. All the 20 basic variables are still very significant, however, only 5 variables from the new 12 variables are selected in the model (see Model (8) in Table 6.17), and none of them are significant at the 0.0001 level of P-value.

```
                            Standard        Wald
Parameter                  Estimate       Error    Chi-Square    Pr > ChiSq

Intercept                   -4.7222      0.0960     2420.4746        <.0001
Account_Duration_2          -0.2522      0.0555       20.6699        <.0001
Account_Duration_3          -0.5116      0.0547       87.5648        <.0001
AdditionalABLLine_1          0.7396      0.0694      113.4728        <.0001
ATTrend3m_2                  0.2527      0.0564       20.0757        <.0001
ATTrend3m_3                  0.5265      0.0687       58.7661        <.0001
ATTrend3m_4                  0.5560      0.0982       32.0345        <.0001
ATTrend3m_5                  0.6994      0.0712       96.6081        <.0001
BandName_1                  -0.4110      0.0887       21.4790        <.0001
Bank_1                       1.3135      0.0535      601.8227        <.0001
DisapprovalsAge_01           0.2305      0.0443       27.0405        <.0001
EntitleAvgTr3m_01            0.2235      0.0617       13.1414        0.0003
FinancialRating_2            0.2933      0.0496       34.9761        <.0001
FinancialRating_3            0.5202      0.0572       82.8319        <.0001
LedgerDiffer_01              0.2467      0.0455       29.3757        <.0001
Top5PcDebtors_1             -0.5958      0.0895       44.2708        <.0001
Top5PcDebtors_2             -0.3065      0.0676       20.5748        <.0001
Top5PcDebtors_4              0.3812      0.0484       62.0100        <.0001
UtilisationAve90_0           0.1543     0.00504      935.5134        <.0001
DebtTurnClient_0             0.0331     0.00556       35.5337        <.0001
PCRTrend5_1                  0.2877      0.0772       13.8873        0.0002
RatingChange3_2              0.2685      0.0914        8.6264        0.0033
EntitTrend5_3                0.1178      0.0515        5.2239        0.0223
LTVTrend3_2                 -0.1192      0.0605        3.8808        0.0488
 NetAssTrend5_1               0.2012      0.0797        6.3791        0.0115
```

Table 6.17 Coefficient details of Model (8) with new trend variables

Model 8	Training	In-time Test	Out-of-time Test
Gini	0.64	0.64	0.60
KS	47.67	50.55	46.04
Hosmer-Lemeshow test (Chi square)	46.90 (<.0001)	17.14 (<.05)	1246.56 (<.0001)
Actual number of defaults	4047	1979	1330
Expected number of defaults	4047	1949	601

Table 6.18 Measurement results of Model (8) with new trend variables

Table 6.17 is the coefficient details of Model (8) with new trend variables, the last 5 variables are newly created variables, which are ‘PaymentCollectionRatio’ 5 month trend, ‘FinancialRatingCode’ downgrade in last 3 month, ‘EntitlementAvg30’ 5 month trend, ‘LoanToValue’ 3 month trend, ‘NetAssignmentAvg30’ 5 month trend. We can see that none of them have significant Wald Chi-square and p-value. The 20 basic variables are still important, but only one of them, ‘EntitlementAvgTrend3month’, p-value becomes less significant. This is because the 5 newly created trend variables contain a variable of ‘EntitlementAvg30’ 5 month trend, these two reflect the similar trend story, thus the new one makes the old one (‘EntitlementAvgTrend3month’) unimportant. Table 6.18 is the measurement results of Model (8), comparing it with Model (1) in Table 6.2, we can see Gini and KS have a little improvements in training and in-time test samples, but no any improvement in out-of-time test sample. Also, the expected number of defaulter is quite low, no any improvement compared with Model (1).

6.5 Segmentation

In this section, we try to split the population into segments, build scorecards for each segment, and see whether there will be prediction improvements from segmentation. In practice, most lenders have several scorecards for different segments. There are a number of reasons why segmentation is needed. (1) Practical reasons. Lenders use behaviour scores in scorecards, but new clients have no behaviour scores; therefore a different scorecard is needed for new applicants. (2) Strategic reasons. Segmentation is able to deal with some groups different, such as VIP group, student group. (3) Statistic reasons. Segmentation is one of ways of dealing with variable interactions. Here, we want to see whether segmentation helps the statistical prediction. For each segment, we still hold training, in-time test, and out-of-time test samples. The out-of-time test samples will be the same observations as before, so that comparisons of models performance can be made objectively. In the training and in-time test samples, sampling is done again to maintain the good/bad ratio to be 3:1 in each segment, thus the observations are slightly different from before.

Two ways of segmentation are tried. One way is to split the firms into 4 segments in terms of firm type. (1) Manufacturing (38% of firms). (2) Wholesale and retail trade, and repairing service (24% of firms). (3) Financial intermediation, and real estate, renting and business activities (22% of firms). (4) Others (16% of firms). Another way is to split the whole population into 3 segments in terms of firms' estimated annualised turnover. (1) Small firms (turnover less than 1 million, 22% of population). (2) Medium firms (turnover between 1 million and 4 million, 45% of population). (3) Big firms (turnover greater than 4 million, 33% population). The second way is based on monthly estimated annualised turnover; due to fluctuation of this value, some firms will be in small firms group this month and in medium firms group next month. These two ways are the most obvious and easiest ways to segment; also they are suggested by practitioners in the industry.

In the training samples of each segment, univariate analysis is performed for each independent variable and they are regrouped to be dummy variables according to the good/bad rates in each band. Thus, in different training samples, the independent dummy variables are different. Two economic variables, 'Business Confidence Index', 'FTSE all shares', are put in the models for each segment, because these two outperform other economic variables in the previous analysis. Prediction models are built for each segment, and they are used to make predictions for the in-time test and out-of-time test samples in each segment. Then training samples, in-time test samples, and out-of-time test samples are combined together to form the whole training, in-time test, out-of-time test samples in two segmentation ways. The Gini, KS, and Hosmer-Lemeshow tests are calculated for the combined training, in-time test, out-of-time test samples.

Segmentation 1	Training	In time Test	Out of time Test
Gini	0.68	0.57	0.56
KS	52.80	43.94	42.43
Hosmer-Lemeshow test (Chi square and p-value)	31.33 (<.0001)	30.96 (<.0001)	187.82 (<.0001)
Actual number of defaults	4666	2247	1409
Expected number of defaults	4666	2172	1382

Table 6.19 Measurement results of the segmentation (1) – firm types

Segmentation 2	Training	In time Test	Out of time Test
Gini	0.63	0.60	0.52
KS	46.48	45.06	39.36
Hosmer-Lemeshow test (Chi square and p-value)	40.47 (<.0001)	179.73 (<.0001)	129.94 (<.0001)
Actual number of defaults	4666	2247	1409
Expected number of defaults	4666	2140	1392

Table 6.20 Measurement results of the segmentation (2) - turnover

From Table 6.19 and Table 6.20 we can see that segmentation overfits. Gini and KS in both in-time and out-of-time test samples are much lower than those in training samples in both segmentations. The only improvement occurs in the training sample of segmentation (1) (Table 6.19), the Gini and KS are higher than those in any other models built before (Such as Model (2) in Table 6.6), but they deteriorate a lot in the in-time test and out-of-time test samples. The overfitting probably results from too few observations in some segments. There are 4 segments in segmentation (1), and 3 segments in segmentation (2), thus, on average, the number of observations in each segment of segmentation (1) is less than that of segmentation (2), so the overfitting is more serious in segmentation (1) than in segmentation (2). The number of expected defaulters in both segmentations is very close to the number of actual defaulters, which is better than Model (2) in Table 6.6, but the high Chi square of Hosmer-Lemeshow test suggests the distribution of expected number of defaulters in the ten bands is not as good as before. Thus, we conclude segmentation does not improve predictions.

The independent variables in the segmentation models are very similar with the basic 20 variables in the models built before. ‘Account_Duration’, ‘Bank’, ‘Top5PcDebtor’, and ‘UtilisationAve90’ are still the very significant variables. In small segments, there are fewer independent variables entering the models than in the big segment. For example, in the fourth segment (‘Others’, 16% of firms) of segmentation (1), there are only 12 variables with significant p-value entering the model, but in the first segment (‘Manufacturing’, 38% firms) segmentation 1, there are 26 variables with significant P-value in the model. The signs for each variable in the models are checked, and two variables whose signs are different from the signs in the overall model built before are left out of the models. The two economic variables ‘Business Confidence Index’ and ‘FTSE-all shares’ still have negative signs, their values in different segments are slightly different, see Table 6.21 and Table6.22.

Segmentation (1)	Economic variables	Coefficient	Standard Error	Wald Chi-Square	P-Value
(1) Manufacturing	Confidence	-0.0303	0.0037	67.91	<0.0001
	FTSE	-0.1825	0.0386	22.40	<0.0001
(2) Wholesale	Confidence	-0.0205	0.0053	15.25	<0.0001
	FTSE	-0.1872	0.0546	11.77	0.0006
(3) Business activities	Confidence	-0.0256	0.0062	17.14	<0.0001
	FTSE	-0.1541	0.0686	5.05	0.0246
(4) Others	Confidence	-0.0427	0.0055	61.07	<0.0001
	FTSE	-	-	-	-

Table 6.21 Economic variables in the models of segmentation (1)

Segmentation (2)	Economic variables	Coefficient	Standard Error	Wald Chi-Square	P-Value
(1) Small firms	Confidence	-0.0179	0.0041	19.49	<0.0001
	FTSE	-0.1008	0.0420	5.76	0.0164
(2) Medium firms	Confidence	-0.0288	0.0035	68.03	<0.0001
	FTSE	-0.1479	0.0374	15.61	0.0006
(3) Big firms	Confidence	-0.0289	0.0056	26.33	<0.0001
	FTSE	-0.2839	0.0602	22.25	<0.0001

Table 6.22 Economic variables in the models of segmentation (2)

Table 6.21 and Table 6.22 describe the coefficients of economic variables in two types of segmentation models. ‘Confidence Index’ and ‘FTSE’ are still important in all segments (except in ‘Others’ of segmentation 1, where ‘FTSE’ does not enter the model), as well as when they were put into overall scorecard. ‘FTSE’ has the same effect on ‘Manufacturing’ and ‘Wholesale’, and a slightly less effect on ‘Business activities’, but no effect on ‘Others’. However, in the other segmentation, ‘FTSE’ has much larger effect on ‘Big firms’ than ‘Medium firms’, and larger on ‘Medium firms’ than ‘Small firms’. This makes sense, since big firms are closer in their size and scale of operations to FTSE All-share firms.

Less difference is between segments for ‘Confidence Index’. Changes in confidence have a lower impact on wholesale firms and on small firms. The result for small firms is a little surprising. It might be expected confidence to affect them more, but it may be that they have shared a small client base, the impact is very individual. So confidence measured by a national survey may be less effective at picking up the idiosyncratic effects of the customers of an individual firm.

6.6 Conclusions

In this research, invoice discounting firms are viewed as consumers, and logistic regression models are built to measure their default risk. Totally ten models have been built, the measures of model performance based on out-of-time test sample are listed in Table 6.23.

There are 75 independent variables already available in the data set. Based on them, more than 200 dummy or ordinal variables are created and used in model building. The most significant 20 variables are selected and kept in the basic model. In the out-of-time test sample, the Gini is 0.60, and KS is 46.34. (see Model 1 in Table 6.23) However, the number of expected defaulters is 605, which is much lower than the actual number of defaulters 1409.

To deal with this problem, we consider putting macroeconomic variables in the model. Two versions of economic variables are tried. One version is ‘Business Confidence Index’ and ‘FTSE all shares’ (Model 2); another version is ‘GDP growth’ (Model 3). The two models with economic variables effectively solve the problem above, and the expected number of defaulters is increased to 1306 and 1321, which are close to the actual number of defaulters, although the Gini decreased a little bit.

Models (Out of time test samples)	Gini	KS	Hosmer-Lemeshow test (Chi square)	Expected number of defaults
Model (1)	0.60	46.34	1470.94	605
Model (2)	0.59	48.80	63.81	1306
Model (3)	0.59	44.33	66.44	1321
Model (4)	0.57	44.17	81.69	1383
Model (5)	0.54	39.43	146.0	1309
Model (6)	0.60	45.51	195.96	1077
Model (7)	0.60	45.79	948.17	714
Model (8)	0.60	46.04	1246.56	601
Model (9)	0.56	42.43	187.82	1382
Model (10)	0.52	39.36	129.94	1392

Table 6.23 Model comparisons based on out-of-time test sample

We consider there might be interactions between the given variables and the macroeconomic variables, so some interaction variables are created. But only a few are entered in the model and they are not very significant, and do not improve models' Gini, KS, and expected number of defaulters (Model 4 and Model 5).

Considering the values of economic variables are usually published a couple of month after the target month, thus we consider using economic variables' value of one quarter or month before the target month. 'Confidence Index' and 'FTSE' are still important (Model 6); but 'GDP growth' becomes less important than before, and its p-value is not smaller than 0.01 (Model 7). Although Gini and KS are still as good as before, the number of expected defaulters drops a lot to 1077 and 714 in the two models with two versions of lagged economic variables respectively. The model with lagged 'GDP growth' is worse than the model with lagged 'Confidence Index' and 'FTSE', we can say that the pre-dictability of 'Confidence Index' and 'FTSE' is better than that of 'GDP growth'.

We try to create some new trend variables in order to compare the value of this month with the value of last month, the average value of last 3 months and the average value of last 5 months. However, few of the new trend variables enter in the model, and do not improve the models predictive performance (Model 8).

Finally, we try to split the whole population into segments in terms of firm types (Model 9) and turnover volumes (Model 10) and build models for each segment. However, the segmentation makes models overfit. The Gini and KS in the in-time test and out-of-time test samples are much lower than those in the training samples in both segmentations. We conclude the segmentation does not improve the models predictive performance.

Chapter 7

Conclusion

With the implementation of New Basel Accord, banks who adhere to advanced Internal-Rating-Based (IRB) approaches have to develop their own empirical models based on historical data for probability of default (PD), loss given default (LGD) and exposure at default (EAD). The 2007-2009 financial crisis led to disastrous consequences in global financial market. Part of the cause is that the risk models in banks, rating agencies, and other financial institutes seemed not to work well. They did not respond to macroeconomic changes and underestimated the credit risks. Therefore, building sound credit risk models is becoming more and more important, so as to maintain a healthy and stable international financial market. This thesis looks at modelling Loss Given Default (through modelling Recovery Rate (RR, RR=1-LGD)) for unsecured personal loans and modelling Probability of Default for invoice discounting. This chapter summarizes the findings and the contributions of this research and discusses possible future research directions.

7.1 Summary

7.1.1 Linear Regression or Survival Analysis for Modelling LGD

Chapter 3 builds linear regression and survival analysis models for estimating recovery amount and recovery rate. The research shows that modelling recovery amount directly is worse than modelling RR first, and then multiplying predicted RR by default amount to get predicted recovery amount indirectly. In all cases of modelling recovery amount and modelling RR, linear regression is better than survival analysis models, based on a few measures such as R-square, Spearman ranking coefficient, MAE and MSE. Only one exception is that one survival analysis model achieves a higher Spearman ranking coefficient than linear regression model for recovery amount modelling. Among all the survival analysis models, Cox proportional hazard model is always better than accelerated failure time models. The reason might be Cox models do not depend on the distributions of the target variable.

However, this conclusion is derived from the measures based on test sample, where the censored observations (debts still being paid) are left out, because we do not know what their final RR is, thus can not measure their predictions. This is unfair to survival analysis models, which are good at dealing with censored data and include the censored observations in the training sample.

7.1.2 Single Distribution or Mixture Distribution Models for LGD

Chapter 3 also builds mixture distribution models for estimating RR. The first stage is to segment the population. Two ways of segmenting are tried. One way is to attempt to maximize the distance of average RR between segments, and another way is to split no-recoveries, partial-recoveries, and full-recoveries. Then the second stage is to build linear regression and survival analysis models for each segment, and then combine the test sample from each segment into the whole test sample. Comparisons are made between single distribution models and mixture distribution models. However, mixture distribution models do not outperform single distribution models in term of R-square, Spearman ranking coefficient, MAE and MSE.

7.1.3 Payment Patterns and Short Term RR

Chapter 4 looks at the payment patterns before default and immediately after default and their relationship with final RR. The variables of payment patterns before default are useful, but the model inclulding these variables is not significantly improved. The variables of payment patterns after default are very useful in modelling RR, but they can be observed only after a period of default. If we want to use these variables at the time of default, we have to predict them then. Thus, two-stage models are built. Stage one is to predict the payment patterns in early default period, and stage two uses the predicted information from stage one to build RR prediction models. Unfortunately, two-stage models do not improve the prediction accuracy.

Chapter 4 also investigates the relationship between short term (12 months or 24 months) RR and final RR. Including the variables of short term RR, the model is very much improved. Thus, two-stage models are built again. The same outcome as before occurs in that two-stage models are not better than one-stage models.

7.1.4 Influence of Macroeconomic Variables on RR

Chapter 5 adds macroeconomic variables into RR prediction models. For final RR prediction, economic variables seem not to have any influence.

Maybe this is because the recovery time is long, and it is very hard to find macroeconomic variables from a specific time to influence final RR. However, adding economic variables into payment patterns prediction models and short term RR prediction models, the models performance is improved. Thus, it is suggested that macroeconomic variables do have influence on short term recoveries. But two-stage models including economic variables are still worse than one-stage models.

7.1.5 Modelling PD for Invoicing Discounting

Chapter 6 leaves LGD modelling and turns to PD modelling. In chapter 6, a logistic regression model is built for invoicing discounting. The data set is from a major UK bank, whose model did not work well, and predicted too few defaulters in this financial crisis. After adding economic variables, the model works well and the bank’s problem is successfully solved. We conclude that the default probability of invoice discounting firms is influenced by macroeconomic conditions. Segmentation is applied to invoice discounting firms; however, the segmentation does not improve the scorecard’s performance. This might be due to too few observations in some segments, which leads to model overfit.

7.2 Contribution to Literature

7.2.1 LGD Modelling

Survival analysis approach is used to model LGD in this thesis, and this is new in the academic literature and in industries. This approach overcomes a difficulty in recovery modelling, that is a large number of censored observations exist in recovery data. Mixture distribution models are also built for modelling LGD for the first time in literature. The aim of this is to build models for different debtor groups who have different reasons to default and have various capacity and willingness to repay.

This thesis looks at the payment patterns before and immediately after default, and their relationship with RR. The payment-pattern variables are used in modelling LGD for the first time in the literature. This thesis also looks at the relationship between short term RR and final RR, and two-stage models are built to model LGD.

7.2.2 PD Modelling

A PD model for invoice discounting is built for the first time in academic literature. Successfully including macroeconomic variables, the model can effectively respond to macroeconomic changes. This overcomes the problem the data-issuing bank met in this financial crisis.

7.3 Research Limitation

The LGD dataset in this research includes debtors' payment patterns – how many number of payments in each month. So, we can identify whether a debtor made a payment or not, and one payment or a few number of payments. But we do not know how much money was paid in each payment. If we know it, some further research could be done. For example, whether paying 50 pounds per month is better than paying 30 pounds per month; or whether paying equal amount in each month is better than paying varied amount in each month.

7.4 Suggestions for Future Research

7.4.1 Censored or uncensored for paid-off debts

In Chapter 3, the paid-off observations are set as censored data. We assume that they are stopped paying by the bank or by themselves when they have paid off the debts, but they still have financial capacity to continue to pay. It is analogous to that in medical research some patients do not die of the disease, but die from other external causes, thus these cases are usually set as censored observations. Is this reasonable for paid-off debts in modelling LGD? An extra research was done, where paid-off observations were set as uncensored observations, and Cox regression models (including 0’s and excluding 0’s) were built for modelling final RR. Table 7.1 is a comparison table, which lists the model results of linear regression, Cox models built before where paid-off observations are set as censored, and new Cox models where paid-off observations are set as uncensored. From Table 7.1 we can see that when paid-offs are set as uncensored, the optimal cut off point moves towards the medium, R squares are improved a little, but MSE of Cox with 0’s model is worse, compared with Cox models where paid-offs are set as censored. From the 4 measures (R-square, Spearman ranking coefficient, MAE, MSE), we can not conclude which way is better to set paid-off observations censored or uncensored, because most measures have similar performance, some of them are improved but some of them are worsened. Thus, some further research needs to be done to find out which approach is better and more practical. However, whether paid-off observations are considered censored or uncensored, all the Cox models are worse than linear regression model in terms of the 4 measures.

		paid-off as censored		paid-off as uncensored	
	Linear Regression	Cox with 0	Cox without 0	Cox with 0	Cox without 0
cut off point		46%	30%	48%	32%
R-square	0.0904	0.0673	0.0609	0.0727	0.0657
Spearman	0.2959	0.2726	0.2551	0.2674	0.2573
MAE	0.3682	0.3546	0.3564	0.3819	0.3618
MSE	0.1675	0.2006	0.2072	0.2674	0.2035

Table 7.1 Comparison for Cox models setting paid-off debts uncensored

7.4.2 How to make segmentations

In order to build mixture distribution models, segmentation are done in chapter 3 based on RR values. However, mixture distribution models do not improve prediction accuracy, and the reason might be that the segmentation is poor. Is it a good way to make segmentation only according to RR values? We try to split the whole population into people who won't pay and people who can't pay. It is very difficult to do so. In each segment, RR varies from 0 to 1, thus this poor segmentation affect the performance of mixture distribution models. A further research could be done to involve a few other variables together with RR to split the population, and cluster analysis could be a segmentation tool.

7.4.3 How to measure the predictions for censored observations properly

In the test sample of LGD modelling, censored observations (debts still be paid) are removed, because we don't know their actual RR, so can't measure the predictions. Then we use R-square, Spearman ranking coefficient, MAE, MSE to measure the prediction accuracy in the test sample. However, this is partial to linear regression. Firstly, the reason we consider to apply survival analysis in LGD modelling is that it can deal well with censored observations. The censored observations are kept in the training sample to build survival analysis models, but are left out when building linear regression models. However, in the test sample the censored observations are removed, this is unfair for survival analysis. Secondly, the core mechanism of linear regression is to maximise R-square and minimise MSE, thus linear regression always achieves the highest R-square and lowest MSE. Thus, in terms of these two measures, linear regression is always better than survival analysis. If an out-of-time test sample is set, we can have an estimate of every loan recovery rate, so we should be able to compare that with an "actual" recovery rate. For loans which are paid off or are written off during the out-of-time sample, we use the actual recovery rate at the time of write off or completion of payment. For those which are still paying at the end of the sample, we can use their recovery rate to that date. Clearly, that is a slight underestimate, but it is like the lender deciding to write off all loans at that point. The out-of-time test sample was not set in this research due to the low default rate and write-off rate in the last 3 years.

7.4.4 Continuing to look for key variables

Although this research suggests linear regression is better than survival analysis in modelling LGD, the R square of linear regression is still very low (around 0.1). Thus, if there are the magic variables we have not found? Payment-pattern variables in early default period and short-term RR variables are very helpful in modelling final RR, but to predict them at the time of default is very difficult based on the variables currently used. If we can find some variables which can be known at an early stage and have the similar power as payment-pattern variables and short-term RR variables, LGD prediction accuracy could be significantly improved. Macroeconomic variables have some influence on payment patterns predictions and short-term RR predictions, but have little influence on final RR predictions. It is a further task to look for the more effective macroeconomic variables to predict final RR.

7.4.5 Write-off policy

LGD or RR depends not only on debtors’ willingness or capacity to repay, but also on lenders’ collection strategy and write off policy. It would be useful if the ‘write off policy’ could be separated out from the borrowers’ performance. However, it is very hard to do so in this data set. Payment pattern models would allow one to find what the optimal write off policies are. For example, how long a time after a borrower stops paying is it optimal to write off a debt. Moreover, the cost generated during the recovery process should be considered, and so optimality would mean that the amount recovered in the future exceeds these costs.

References

Adair, A., Berry, J., Haran, M., Lloyd, G. & McGreal, S. (2009) The Global Financial Crisis: Impact on Property Markets in the UK and Ireland. Report by the University of Ulster Real Estate Initiative Research Team. http://news.ulster.ac.uk/podcasts/ReiGlobalCrisis.pdf

Albright, H. T. (1994) Construction of a polynomial classifier for consumer loan applications using genetic algorithms, Department of Systems Engineering, University of Virginia, Working Paper.

Altman, E. I. (1968) Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, *Journal of Finance*, Vol.23, No.4

Altman, E. I. (2005) An Emerging Market Credit Scoring System for Corporate Bonds, *Emerging Markets Review*, Vol. 6, No. 3

Altman, E. I. & Eberhart, A. (1994) Do Seniority Provisions protect bondholders investments, *Journal of Portfolio Management*, Summer, pp67-75

Altman, E. I. & Fanjul, G. (2004) Defaults and Returns in the High Yield Bond Market: Analysis Through 2003, NYU Salomon Centre Working paper, January.

Altman, E. I., Haldeman, R. & Narayanan, P. (1977) ZETA Analysis: A new model to identify bankruptcy risk of corporations, *Journal of Banking and Finance* 1, pp29-54

Altman, E. I. & Kishore, V. M. (1996) Almost Everything You Wanted to Know About Recoveries on Defaulted Bonds, *Financial Analysts Journal*, November/December.

Altman, E. I., Marco, G., & Varetto, F. (1994) Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking and Finance* 18, 505–529

Altman, E. I., Resti, A. & Sironi, A. (2001) Analyzing and Explaining Default Recovery Rates, A Report Submitted to The International Swaps & Derivatives Association, December 2001

Altman, E. I., Resti A. & Sironi A. (2005) Loss Given Default; a review of the literature in Recovery Risk, *Recovery Risk*, ed by Altman, E.I., Resti, A, Sironi, A. Risk Books, London, pp 41-59

Altman, E. I. & Sabato, G. (2007) Modelling Credit Risk for SMEs: Evidence from the U.S. Market, *Abacus*, 43(3), 332-357

Altman, E. I., Sabato, G. & Wilson, N. (2009) The Value of Non-financial Information in SEM Risk Management, Proceedings of Credit Scoring & Credit Control XI.

Altman, R. (2009) 'Altman - The Great Crash', *Foreign Affairs*. http://www.foreignaffairs.org/20090101faessay88101/roger-c-altman/the-great-crash-2008.html. Retrieved 2009-02-27

Anderson, R. (2007) *The Credit Scoring Toolkit*, Oxford University Press, Oxford.

Andreeva, G. (2006) European generic scoring models using survival analysis. *Journal of Operational Research Society* 57 (10), 1180–1187

Andreeva, G., Ansell, J. & Crook, J. (2007) Modelling profitability using survival combination scores. *European Journal of Operational Research* 183: 1537–1549

Apilado, V. P., Warner, D. C. & Dauten, J. J. (1974) Evaluative techniques in consumer finance. *Journal of Financial Quantitative Analysis*, March, 275-283

Artzner, P. & Delbaen, F. (1995) Default Risk Insurance and Incomplete Markets, *Mathematical Finance*, 5, pp. 187-95.

Asarnow, E. & Edwards, D. (1995) Measuring Loss on Defaulted Bank Loans: A 24-Year Study, *Journal of Commercial Lending*, 77(7), 11-23

Aziz, A., Emanuel, D.C. & Lawson, G.H. (1988) Bankruptcy Prediction—an Investigation of Cash Flow Based Models, *Journal of Management Studies*, Vol. 25, No. 5

Banasik, J., Crook, J.N. & Thomas, L.C. (1999) Not if but when will borrowers default. *Journal of Operational Research Society* 50 (12), 1185–1190

Banasik, J. & Crook, J. (2010) Reject inference in survival analysis by augmentation. *Journal of the Operational Research Society* 61, 473-485

Basel Committee on Banking Supervision (BCBS), (2004, updated 2005), International Convergence of Capital Measurement and Capital standards: a revised framework, Bank of International Settlement, Basel.

Beaver, W. (1967) Financial Ratios as Predictors of Failure, *Journal of Accounting Research*, Vol. 4, (Supplement).

Becchetti, L. & Sierra, J. (2002) Bankruptcy Risk and Productive Efficiency in Manufacturing Firms, *Journal of Banking and Finance*, Vol. 26, No. 8

Benoit, D. F. & Van den Poel, D. (2009) Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: An application in financial services, *Expert Systems with Applications* 36 10475-10484

Bellotti, T. & Crook, J. (2008) Modelling and Estimating Loss Given Default for Credit Cards, CRC working paper 08-1, University of Edinburgh.

Bellotti, T. & Crook, J. (2009) Loss Given Default models for UK retail Credit Cards, CRC working paper 09-1, University of Edinburgh.

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

Bernanke, B. S. (2009) ‘Four Questions About the Financial Crisis’,
Federalreserve.gov. 2009-04-14

http://www.federalreserve.gov/newsevents/speech/bernanke20090414a.htm. Retrieved 2010-05-01

Black, F. and Scholes, M. (1973) The pricing of options and corporate liabilities, *Journal of Political Economy*, 81, 637-659

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984) Classification and Regression Trees, Wadsworth, Belmont, Califonia.

Buckley, J. J. & James, I. R. (1979) Linear regression with censored data, *Biometrika* 66, 429-436

Calem, P. S. & LaCour-Little, M. (2004) Risk-based capital requirements for mortgage loans. *Journal of Banking & Finance*, 28 (3), 647- 672

Carter, C. & Catlett, J. (1987) Assessing credit card applications using machine learning. *IEEE Expert*, fall, 71-79

Casey, C. & Bartczak, N. (1985) Using operating Cash Flow Data to Predict Financial Distress: Some Extensions, *Journal of Accounting Research* Vol. 23 No.1 (Spring 1985):384-401

Chatterjee, S. & Barcun, S. (1970) A nonparametric approach to credit screening, *Journal of American Statistical Association* 65, 150–154

Chew, W. H. & Kerr, S. S. (2005) Recovery Ratings: Fundamental Approach to Estimation Recovery Risk, *Recovery Risk*, ed by Altman E.I., Resti A, Sironi A. Risk books, London, p87-97

Chorafas, D.N. (1990) Risk Management in Financial Institutions. Butterworth & Co.: London.

Clementi, D. (2000) ‘Crisis Prevention and Resolution – Two Aspects of Financial Stability’, Deputy Reserve Bank Governor’s inaugural lecture at the South Bank University Centre for Monetary and Financial Economics, 6 September. Bank of England: London.
http://www.bankofengland.co.uk/publications/speeches/2000/speech97.htm

Coffman, J. Y. (1986) The proper role of tree analysis in forecasting the risk behaviour of borrowers. MDS Reports 3, 4, 7 and 9. Management Decision Systems, Atlanta.

Cox, D.R. (1972) Regression Models and Life Tables (with discussion), Journal of the Royal Statistical Society, B34, 187-220

Crosbie, P. & Bohn, J. (2002) 'Modeling Default Risk', KMV, 2002

Davis, D. B. (1987) Artificial intelligence goes to work. *High Technol*, Apr., 16-17

Davis, R. H., Edelman, D. B. & Gammerman, A. J. (1992) Machine-learning algorithms for credit- card applications. *IMA Journal of Mathematics Applied in Business and Industry*, 4, 43-51

De Servigny, A. & Oliver, R. (2004) Measuring and managing Credit Risk, McGraw Hill, Boston

Demyanyk, D. & Van Hemert, O. (2008) Understanding the Subprime Mortage Crisis, Working paper series.
Available at SSRN: http://ssrn.com/abstract=1020396

Dermine, J. & Neto de Carvalho, D. (2003) Bank Loan Losses-given-default – Empirical Evidence, October, mimeo.

Dermine, J. & Neto de Carvalho, C. (2005) Bank loan losses given default: A case study, *Journal of banking and Finance* 30, 1219-1243

Desai, V. S., Crook, J. N., & Overstreet, G. A. (1996) A comparison of neural networks and linear scoring models in the credit environment. *European Journal of Operational Research* 95, 24–37

Desai,V. S., Convay, D. G., Crook, J. N., & Overstreet, G. A. (1997) Credit scoring models in the credit union environment using neural networks and genetic algorithms. *IMA Journal of Mathematics Applied in Business and Industry* 8, 323–346

Dias Jose G. (2004) Finite Mixture Models, Rijksuniversiteit Groningen.

Duffie, D. (1998) Defaultable Term Structure Models with Fractional Recovery of Par, Graduate School of Business, Stanford University.

Durand, D. (1941) Risk Elements in Consumer Instalment Financing. New York: National Bureau of Economic Research

Eales, R. & Bosworth, E. (1998) Severity of Loss in the Event of Default in Small Business and Larger Consumer Loans, *Journal of Lending & Credit Risk Management,* May, 58-65

Eisenbeis, R. A. (1977) Pitfalls in the application of discriminant analysis in business, finance, and economics. Journal of Finance., 32, 875-900

Eisenbeis, R. A. (1978) Problems in applying discriminant analysis in credit scoring models. Journal of Banking and Finance, 2, 205-219

Engelmann, B. & Rauhmeier, R. (2006) The Basel II Risk Parameters, Springer, Heidelberg

Figlewski, S., Frydman, H. & Liang, W. (2007) Modelling the Effect of Macroeconomic Factors on Corporate Default and Credit Rating Transitions. Working Paper No. FIN-06-007, NYU Stern School of Business.

Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems, Ann. Eugenics, 7, 179-188

Fitzpatrick, D. B. (1976) An analysis of bank credit card profit. *Journal of Bank Research*, 7, 199-205

Fogarty, T. C. & Ireson, N. S. (1993) Evolving Bayesian classifiers for credit control — a comparison with other machine learning methods. *IMA Journal of Mathematics Applied in Business and Industry* 5, 63–76

Fractal Analytics (2003) Comparative Analysis of Classification Techniques: A Fractal Whitepaper. Fractal Analytics: Mumbai, India.

Freed, N. & Glover, F. (1981) A linear programming approach to the discriminant problem, *Decision Sciences* 12, 68-74

Friedman, C. & Sandow, S. (2003) Ultimate Recoveries, Risk, August, pp.69-73

Gentry, J. A., Newbold, P. and Whitford, D. T. (1985) Classifying Bankrupt Firms With Funds Flow Components, *Journal of Accounting Research*, Vol. 23, No. 1

Grablowsky, B. J. & Talley, W. K. (1981) Probit and discriminant functions for classifying credit applicants: a comparison. *Journal of Economics and Business*, 33, 254-261

Grippa, P. S., Iannotti, F. & Leandri, F. (2005) Recovery rates in the banking industry: stylised facts emerging from the Italian experience. In Altman E, Resti A and Sirona A (eds) *Recovery Risk*, Risk Books, London.

Gup, B. E. (2004) The New Basel Capital Accord, Thomson Edition, 2004

Gupton, G. (2005) Estimation Recovery Risk by means of a Quantitative Model: LossCalc, *Recovery Risk*, ed by Altman E.I., Resti A, Sironi A. Risk books, London, p61-86

Hand, D. J. (2001) Modelling Consumer Credit Risk, *IMA Journal of Management Mathematics*, 12, 139-155

Hand, D. J. & Henley, W. E. (1997) Statistical classification methods in consumer credit. *Journal of the Royal Statistical Society*, Series A 160, 523–541

Hand, D.J. & Kelly, M.G. (2001) Lookahead scorecards for new fixed term credit. *Journal of Operational Research Society* 52, 989–996

Hand, D. J., Oliver, J. J. & Lunn, A. D. (1998) Discriminant analysis when the classes arise from a continuum. *Pattern Recognition* 31: 641-650

Hardy, W. E. & Adrian, J. L. (1985) A linear programming alternative to discriminant analysis in credit scoring. *Abribus* 1, 285–292

He, Y., Kamath, R. & Meier, H.H. (2005) An Empirical Evaluation of Bankruptcy Prediction Models for Small Firms: An Over-The-Counter (OTC) Market Experience, *Academy of Accounting and Financial Studies Journal*, Volume 9, Number 1

Henley, W. E. (1995) Statistical aspects of credit scoring. PhD Thesis. The Open University, Milton Keynes.

Henley,W. E. & Hand, D. J. (1996) A k-NN classifier for assessing consumer credit risk. *The Statistician* 65, 77– 95

Hosmer D.W. & Lemeshow S. (1989) Applied Logistic Regression, Wiley, New York.

Jarrow, R. A., Lando, D. & Turnbull, S. M. (1997) A Markov Model for the Term Structure of Credit Risk Spreads, *Review of Financial Studies*, 10, pp. 481-523

Jarrow, R. A. & Turnbull, S. M. (1995) Pricing Derivatives on Financial Securities Subject to Credit Risk, *Journal of Finance*, 50(1), pp.53-86

Joachimsthaler, E. A. & Stam, A. (1990) Mathematical programming approaches for the classification problem in two-group discriminant analysis. *Multivariate Behavioural Research* 25, 427–454

Keasey, K. & Watson, R. (1987) Non-Financial Symptoms and the Prediction of Small Company Failure: A Test of Argenti’s Hypotheses, *Journal of Business Finance and Accounting*, Vol. 14, No. 3

Kelly, M.G. & Hand, D.J. (1999) Credit scoring with uncertain class definitions. *IMA Journal of Mathematics Applied in Business and Industry* 10: 331-345

Krugman, P. (2009) ‘Revenge of the Glut’. nytimes.com (New York Times). http://www.nytimes.com/2009/03/02/opinion/02krugman.html?pagewanted=print.

Lahart, J. (2007) ‘Egg Cracks Differ In Housing, Finance Shells’. WSJ.com (Wall Street Journal).
http://online.wsj.com/article/SB119845906460548071.html?mod=googlenews_wsj.
Retrieved 2008-07-13.

Lane, S. (1972) Submarginal credit risk classification. Journal of Financial and Quantitative Analysis., 7, 1379-1385

Leonard, K. J. (1993a) Empirical Bayes analysis of the commercial loan evaluation process. *Statistics & Probability Letters*, 18, 289-296

Leonard, K. J. (1993b) Detecting credit card fraud using expert systems. *Computers & Industrial Engineering*, 25, 103-106

Leow, M., Mues, C. & Thomas, L. C. (2009) Modelling Loss Given Default for Mortgage Loans, Presentation in Conference of Credit Scoring and Credit Control XI, Edinburgh, August 2009.

Lin, S.M., Ansell, J. & Andreeva, G. (2007a) Predicting default of a small business using different definitions of financial distress. Proceedings of Credit Scoring & Credit Control X.

Lin, S.M., Ansell, J. & Andreeva, G. (2007b) Merton models or credit scoring: modelling default of a small business. Working Paper, Credit Research Centre.

Liu, J. & Xu, X.E., (2003) The predictive power of economic indicators in consumer credit risk management, *RMA Journal* , September.

Lucas, A. (1992) Updating scorecards: removing the mystique. In Credit Scoring and Credit Control (eds L. C. Thomas, J. N. Crook and D. B. Edelman), pp. 180-197. Oxford: Clarendon.

Lucas, A. (2006) Basel II Problem Solving; QFRMC Workshop and conference on Basel II & Credit Risk Modelling in Consumer Lending, Southampton 2006.

Luo, D., Tang, D. Y. & Wang, A. (2009) A Little Knowledge Is A Dangerous Thing: Model Specification, Data History, and CDO (Mis)Pricing, Working paper series, University of Hong Kong. Available at
http://www.finance.nsysu.edu.tw/SFM/17thSFM/program/FullPaper/110-1365508837.pdf

Makowski, P. (1985) Credit scoring branches out. *Credit World*, 75, 30-37

Matuszyk, A., Mues, C. & Thomas, L.C. (2010) Modelling LGD for unsecured personal loans: Decision tree approach, *Journal of Operational Research Society*, 61, 393-398

McLachlan, G. J. & Basford, K. E. (1998), Mixture Models: Inference and Applications to Clustering, New York: Marcel Dekker.

Merton, R. (1974) On the pricing of corporate debt: The risk structure of interest rates, *Journal of Finance*, 29, 449-470

Mian, S. & Smith, C. W. (1992) Accounts receivable management policy: theory and evidence. *The Journal of Finance*, 47, 169–201

Moses, D. & Liao, S. S. (1987) On developing models for failure prediction. *Journal of Commercial Bank Lend*, 69, 27-38

Mossman, C. E., Bell, G. G., Swartz, L. M. & Turtle, H. (1998) An Empirical Comparison of Bankruptcy Models, *The Financial Review*, Vol. 33, No. 2, 1998

Murphy, A. (2008) An Analysis of the Financial Crisis of 2008: Causes and Solutions, Working paper series, Oakland University.
Available at SSRN: http://ssrn.com/abstract=1295344

Narain, B. (1992) Survival analysis and the credit granting decision. In: Thomas LC, Crook JN and Edelman (eds). *Credit Scoring and Credit Control*. OUP: Oxford, UK, pp 109–121

Nath, R., Jackson, W. M. & Jones, T. W. (1992) A comparison of the classical and the linear programming approaches to the classification problem in discriminant analysis. *Journal of Statistical Computation and Simulation,* 41, 73–93

Nickell, P., Perraudin, W. & Varotto, S. (2001) Stability of Rating Transitions, *Journal of Banking and Finance* 24, no. 1/2 (2001): 203–228

Ohlson, J. (1980) Financial Ratios and the Probabilistic Prediction of Bankruptcy, *Journal of Accounting Research*, Vol. 18, No.1, 1980

Qi, M. & Yang, X. (2009) Loss given default of high loan-to-value residential mortgages, *Journal of Banking & Finance*, 33, 788-799

Orgler, Y. E. (1970) A credit scoring model for commercial loans. *Journal of Money, Credit and Banking*., Nov., 435-445

Orgler, Y. E. (1971) Evaluation of bank consumer loans with credit scoring models. *Journal of Bank Research*, spring, 31-37

Querci, F. (2005) Loss Given Default on a medium-sized Italian bank's loans: an empirical exercise, The European Financial Management Association, Genoa, Genoa University. http://www.efmaefm.org/efma2005/papers/206-querci_paper.pdf

Quinlan, J. R. (1993) C4.5: Programs for Machine Learning, Morgan Kaufman, San Mateo, California.

Rajun, U., Serun, A. & Vig, V. (2008) The Failure of Models that Predict Default: Distance, Incentives, and Default, Chicago GSB Research Paper No. 08-19

Reichert, A. K., Cho, C. C. & Wagner, G. M. (1983) An examination of the conceptual issues involved in developing credit-scoring models. *Journal of Business & Economic Statistics*, 1, 101-114

Renault, O. & Scaillet, O. (2004) On the way to Recovery: A Nonparametric Bias Free Estimation of Recovery Rate Densities, *Journal of Banking and Finance*, 28(12), pp. 2915-31

Reuters.com (2009) "Bloomberg-U.S. European Bank Writedowns & Losses-November 5, 2009". Reuters.com.
http://www.reuters.com/article/marketsNews/idCNL554155620091105?rpc=44.
Retrieved 2010-05-01

Rikkers, F. (2006) A Structural Form PD Model for SMEs, Evidence from the Dutch Market. The 2006 PREBEM Conference

Rosenberg, E. & Gleit, A. (1994) Quantitative methods in credit management: a survey. *Operations Research* 42, 589–613

Ross, V. (2009) “Lessons from the Financial Crisis” Speech by Verena Ross, Chatham House conference on Global Financial Regulation
http://www.fsa.gov.uk/pages/Library/Communication/Speeches/2009/0324_vr.shtml

Safavian, R. & Landgrebe, D. (1991), A survey of decision tree classifier methodology, *IEEE Transactions on Systems*, Man, and Cybernetics 21, 660-674

Saunders, A. & Allen, L. (2002) Credit risk measurement: new approaches to value at risk and other paradigms, New York, [Great Britain]: John Wiley, c2002.

Schuermann, T. (2005) What Do We Know About Loss Given Default? *Recovery Risk*, ed by Altman E.I., Resti A, Sironi A. Risk books, London, p3-24

Shumway, T. (2001) Forecasting Bankruptcy More Accurately: A Simple Hazard Model, *Journal of Business*, 74, 101-124, 2001

Smith, J. K. & Schnucker, C. (1994). An empirical examination of organizational structure: the economics of the factoring decision. *Journal of Corporate Finance*, 1, 119–138

Sohn, S. Y. & Shin, H. W. (2006) Reject inference in credit operations based on survival analysis. *Expert Systems with Applications* 31: 26–29

Somers, M. & Whittaker, J. (2007) Quantile regression for modelling distributions of profit and loss, *European Journal of Operational Research* 183 (2007) 1477-1487

Soufani, K. (2000) The role of factoring in financing UK SMEs: a supply side analysis. *Journal of Small Business and Enterprise Development*, Volume 8, number 1

Soufani, K. (2002) On the determinants of factoring as a financing choice: evidence from the UK. *Journal of Economics and Business* 54 (2002) 239-252

Stepanova, M. & Thomas, L. C. (2001) PHAB scores: Proportional hazards analysis behavioural scores. *Journal of Operational Research Society* 52, 1007–1016

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

Summers, B. & Wilson, N. (2000) Trade Credit Management and the Decision to Use Factoring: An Empirical Study, J*ournal of Business Finance & Accounting*, 27(1) & (2), January/March 2000

Tam, K. Y. & Kiang, M. Y. (1992) Managerial applications of neural networks: the case of bank failure predictions. *Management Science* 38, 926–947

Tang, L., Thomas, L. C., Thomas, S. and Bozzetto, J-F. (2007) It's the economy stupid: Modelling financial product purchases. *International Journal of Bank Marketing*. 25: 22-38

The Turner Review (2009) A regulatory response to the global banking crisis, Financial Services Authority, UK
http://www.fsa.gov.uk/pubs/other/turner_review.pdf

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.

Thomas, L. C. (2009) Consumer Credit Models: Pricing, Profit, and Portfolios, Oxford University Press, Oxford.

Thomas, L. C., Edelman, D. B. & Crook, J.N. (2002) Credit Scoring and its Applications, SIAM, Philadelphia.

Van Gestel, T. & Baesens, B. (2009) Credit Risk Management, Basic conceptes: financial risk components, rating analysis, models, economic and regulatory capital. Oxford University Press.

Verde, M. (2003) Recovery Rates Return to Historic Norms, Fitch Ratings, September.

Wedel, M. & Kamakura, W. A. (2000) Market Segmentation. Conceptual and Methodological Foundations (2nd ed.), International Series in Quantitative Marketing, Boston: Kluwer Academic Publishers.

Whittaker, J., Whitehead, C. & Somers M. (2005) The neglog transformation and quantile regression for the analysis of a large credit scoring database. *Applied Statistics-Journal of the Royal Statistical Society Series C* 54, 863–878

Wiginton, J. C. (1980) A note on the comparison of logit and discriminant models of consumer credit behaviour, *Journal of financial Quantitative Analysis*, 15,757-770

Wikipedia, Factoring (Finance), Invoice Discounting
http://en.wikipedia.org/wiki/Factoring(finance)
http://en.wikipedia.org/wiki/Invoice_discounting

Wilson, T. (1997a) Credit Risk Modeling: A New Approach. New York: McKinsey & Co., 1997a (mimeo).

Wilson, T. (1997b) Portfolio Credit Risk (Parts I and II). Risk Magazine (September and October 1997b).

Woellert, L. & Kopecki, D. (2008) “Moody’s, S&P Employees Doubted Ratings, E-Mails Say.” Bloomberg online (October 22, 2008),
http://www.bloomberg.com
/apps/news?pid=20601087&sid=a2EMlP5s7iM0&refer=home

Yobas, M. B., Crook, J. N. & Ross, P. (1997) Credit scoring using neural and evolutionary techniques, Credit Research Centre, University of Edinburgh, Working Paper 97/2.

Zocco, D. P. (1985) A framework for expert systems in bank loan management. *Journal of Commercial Bank Lending* 67, 47–54