Corporate Default Prediction:

Models, Drivers and Measurements

Submitted by Yangzhengxuan Wang to the University of Exeter as a Thesis for the Degree of Doctor of Philosophy in Finance,

December 2011.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature: .....

Abstract

This thesis identifies the optimal set of corporate default drivers and examines the prediction performance of corporate default measurement tools, using a sample of companies in the United States from 1970 to 2009.

In the discussion of optimal default drivers, feature selection techniques including the t-test and stepwise methods are used to filter relevant default information collected from previous empirical studies. The optimal default driver information set consists of quantitative parameters from accounting ratios, market indices, macroeconomic indicators, default history, and firm age. While both accounting ratios and market information dominate the explanatory ability, followed by default history, macroeconomic indicators contribute additional explanation for default risk. Moreover, industry effects show significance across alternative models, with the retail industry presenting as the sector with highest risk. The results are robust in both traditional and advanced random models.

In investigating the optimal prediction method, two newly developed random models, mixed logit and frailty model, are tested for their theoretical superiority in capturing default clusters and unobservable information for default risk. The prediction ability of both models has been improved upon using the extended optimal set of default drivers. While the mixed logit model provides better prediction accuracy and shows stability in robustness checks, the frailty model benefits from computational efficiency and explains default clusters more thoroughly.

This thesis further compares the prediction performance of large dimensional models across five categories based on the default probabilities transferred from alternative results in different models. Besides the traditional assessment criteria - covering the

receiver operating characteristic curve, accuracy ratios, and classification error rates – this thesis thoroughly evaluates forecasting performance using innovative proxies including model stability under financial crisis, profitability and misclassification costs for creditors using alternative risk measurements. The practical superiority of the two advanced random models has been verified further in the comparative study.

Acknowledgements

First and foremost, I offer my sincerest gratitude to my supervisor, Richard D. F. Harris, who with his invaluable guidance has supported me throughout my thesis patiently and tolerantly, from preliminary data collection to the conclusions. This thesis would not have been possible without his effective input. One simply could not wish for a better, or kinder, supervisor.

My great thanks goes to my dearest friends, Peter, Debora, Diaz, William, Wenjing, Fiona, Yibo, Ian and Jane, who have helped me in every corner of my life in the UK. They have shared the highs and lows, and welcomed me as a family member.

Thanks to the University of Exeter Business School, which has provided the support and the friendly environment, thus enable me produce and complete my thesis. Thanks to the HOST UK, which helps international students, like myself, to make friends in this country, understand local culture, and gain insights into Britain. Throughout my studies I have benefited greatly from this charity, and it helped make my life easier and happier.

Thanks to AXA has generously funded my studies. Without which I could not have concentrated effectively on my work.

Finally, my appreciation goes to my parents, my grandmother, and my beloved Junxiong, for their understanding and their unwavering support throughout my studies.

Table of Contents

List of Tables

CHAPTER TWO

CHAPTER THREE

CHAPTER FOUR

CHAPTER FIVE

List of Figures

CHAPTER TWO

CHAPTER THREE

CHAPTER FOUR

CHAPTER FIVE

Abbreviations

ACM	Advance Choice Models
AR	Accuracy ratios
AUROC	Are Under the Receiver Operating Characteristic Curve
CPM	Conditional Probability Models
CPV	Credit Portfolio View
CRSP	Centre for Research in Security Prices
DA	Discriminant Analysis
FASB	Financial Accounting Standards Board
FRED	Federal Reserve Bank of St. Louis
HIM	Hazard Intensity Models
IIA	Independence of Irrelevant Alternatives
IID	Independent and Identically Distributed
IT	Information Techniques
MBS	Market-Based Structure models
MDA	Multivariate Discriminant Analysis
NN	Neural Network
O-score	Olson's (1980) measure of the probability of bankruptcy
OL	Opportunity Loss
OLOA	Opportunity Loss On Assets
OV	Original Variables
R&D	Research and Development
ROA	Return On Assets
RS	Rough Set technique
S.D.	Standard Deviation
S&P	Standard & Poor's
UDA	Univariant Discriminant Analysis
WRDS	Wharton Research Data Services
Z-score	Altman's (1968) measure of the probability of bankruptcy

Chapter 1: Introduction

1.1 Objectives and Motivations

This thesis has three objectives. First, it investigates the optimal information required for corporate default prediction by using a comprehensive dataset. Second, the thesis introduces and tests empirically two advanced random models: mixed logit model and Cox frailty survival model, to extract significant relationships and explain the corporate default behaviour more accurately. Third, the thesis compares and evaluates the performance of various large dimension corporate default prediction models using both statistical and economic criteria.

1.1.1 The Concept of Corporate Default

The definition of corporate default (corporate failure) is a crucial concern in this field of research. Misspecification of the concept of corporate default violates the assumption of dichotomous dependent variables, which classifies firms into default and non-default populations, and thus deteriorates the prediction accuracy.

Early studies (Altman, 1968; Ohlson, 1980) narrowly restricted the concept of corporate default to bankruptcy. According to Altman (1968), "bankruptcy refers to those firms that are legally bankrupt and either placed in receivership or have been granted the right to reorganize under the provisions of the National Bankruptcy Act" (p.589). However, the legal definition of bankruptcy varies between countries, and is heavily influenced by banks, creditors, governments and other parties. Most importantly, there are differences between bankruptcy and corporate default. On the one hand, as Balcaen & Ooghe (2006) suggest, "the moment of legal failure often does not reflect the 'real' failure event; it is possible that a firm, showing many of the characteristics of a failing company, does not show a change

in its legal situation. Or, instead of showing a 'failing' legal situation, troubled companies may also merge with another firm or reorganize" (p.19). On the other hand, "firms file the legal status of bankruptcy for strategic reasons, such as eliminate rising debts. Or, companies may file for bankruptcy due to 'acts of God' and may be forced into bankruptcy even though their previous financial results were excellent" (Muller, Steyn-Bruwer & Hamman, 2009, p22). All these cases would pollute the dependent variable and thus lead to poor prediction.

As for the development of corporate default databases, extending the concept from bankruptcy to general corporate default is more common in empirical studies (e.g. Chava & Jarrow, 2004; Duffie, Saita & Wang, 2007; Campbell, Hilscher & Szilagyi, 2008). Bonfim (2009) defines the default risk as occurring when "credit and interest have become overdue within the last 3-6 months" (p.283). For Lennox (1999), corporation failure occurs "if a firm entered liquidation, receivership or administration" (p.350). Dahiya, Saunders & Srinivasan (1994) define financial distress as occurring when "a firm has insufficient cash flows to meet the payments on its debt, which includes: (1) the default on a firm's public debt, and (2) the filing by a firm for bankruptcy protection under Chapter 11" (p.379).

For the purpose of this research, considering the data available, the definition of corporate default follows the classification of Moody's default risk database, by which a firm is defined to have defaulted in the following conditions:

"A missed or delayed disbursement of interest and/or principal, including delayed payments made within a grace period; bankruptcy, administration, legal receivership, or other legal blocks (perhaps by regulators) to the timely payment of interest and/or principal; or a distressed exchange occurs where: (i) the issuer offers debt holders a new security or package of securities that amount to a diminished financial obligation (such as preferred or common stock, or debt with a lower coupon or par amount, lower seniority,

or longer maturity); or (ii) the exchange had the apparent purpose of helping the borrower avoid default." (Moody, 2007)

1.1.2 The Importance of Corporate Default Prediction

Researchers have studied the forecasting of corporate default risk for almost four decades. Default risk is also a risk for society, which is crucial for many parties (Balcaen & Ooghe, 2006), including creditors, bankers, regulators, managers, auditors, governments and shareholders. Understanding the evaluation methods and driving factors of credit risk could help creditors to maximise their profits. It could also help investors and assets managers to reduce consequential losses on their portfolios, because an unsatisfactory financial condition would deteriorate a firm's performance (Opler & Titman, 1994). Forecasting corporate default rates accurately is a significant issue for the assessment of financial stability. Therefore, regulators and policymakers could benefit from accurate prediction models. For bankers, a well-preformed default risk prediction model could help to avoid profit missing due to suboptimal capital allocation. Governments require an accurate prediction model to mitigate the effects of ill performing companies in terms of short-term operational and fundamental features. Shareholder return is driven by firm performance (share price), capital structure and dividend strategy. Lastly, auditors could be empowered to run a more adequate assessment of the firm's health and provide early warning signals through strengthened default prediction processes.

Both the direct and indirect costs of default events are enormous (Opler & Titman, 1994; Charalambous, Charitou & Neophytou, 2000; Charitou, Evineophytou & Charalambous, 2004; Balcaen & Ooghe, 2006). In recent years, large defaults of big firms such as Enron, WorldCom, Kmart World, Lehman Brothers, Washington Mutual and General Motors negatively impacted the interests of their employees, shareholders, creditors, clients and suppliers. In severe cases, corporate default events contribute to a global financial crisis and economic recession fuelling speculation on sovereign default. Accordingly, accurate

prediction of corporate default is vital for both "the individual" and "the society" (Balcaen & Ooghe, 2006). On the other hand, misspecification of a healthy firm as being in financial distress can cause not only opportunity loss for creditors, but also market value reduction for investors and shareholders.

A sound corporate default prediction model produces results that are directly applicable to bond rating, debt pricing, and loss provisioning for investors and creditors. The results also help to determine the capital requirement for banks, thus facilitating process of capital allocation. Researchers are also interested in the abnormality of stock returns, and the volatility of financial distress in firms (Vassalou & Xing, 2004; Campbell, Hilscher, & Szilagyi, 2008; Chava & Purnanandam, 2010). Meanwhile Eisdorfer (2008) has investigated risk-shifting behaviour in financially distressed firms. Other studies (Franzen, Rodgers & Simin, 2007; Liu & Tonks, 2008) have investigated R&D spending, and the Universities Superannuation Scheme, based on the results of default predictions.

1.1.3 Opportunity for Corporate Default Prediction

The renewed interest in exploring default prediction models is attributed to several factors: changes in the economic and financial environment; developments of firm procedure and international regulation systems; the availability of default data; and evolution of new quantitative methods. These factors arouse the interests in investigation of new prediction models as well as testing the validation of traditional methods under new environments.

Over the last forty years, economic indicators have fluctuated unexpectedly. Globalisation and information have changed corporate operation procedures tremendously. Competition between firms and financial institutions becomes more intensive. In the loan market, attention has been paid to higher risk loans. The default rate in many countries has risen spectacularly (Balcaen & Ooghe, 2006). Global recessions in the early 90s, the IT bubble in the late 20th century, and the current national debt crisis have made it painfully

clear that the financial system needs quantitative methods to assess credit risk and flag early warnings more accurately. New financial products have appeared in investment banks and other financial institutions. The explosive growth of the credit derivative market has also given a new importance to credit risk and, accordingly, the estimating of corporate default risk has intensified since 2008 under a global revision of financial regulation driven by financial crises and contagion.

Under the uncertainties and changes of the financial market, financial regulators and governments have updated their policies to secure a stable financial system. The recent interest in investigating default models traces back to the policy of Basel II, which allowed banks to use internal credit risk models to determine capital charges according to risk level. This encouraged banks to develop default evaluation tools to maximise profit. During the recent financial crisis triggered by credit risk, Basel III created a requirement to use long-term data horizons to estimate probabilities of default. Also, Basel III addressed the need for creditors to conduct stress tests and back testing in recessionary scenarios. These new policies increase the interests of retesting and investigation of default prediction models.

Early studies suffered from sample problems such as the limitation of the default dataset, and missing data, which seriously restricted the degree of prediction accuracy (Zmijewski, 1984; Chava & Jarrow, 2004; Hodges, Clusker & Lin, 2005). With the evolution in availability of the default dataset, as well as the publication of private information, the possibility of predicting default risk more accurately has grown. The development of the Internet has made the business of data collection more profitable, and so some financial agencies such as Fitch, Moody's, S&P, and Bankruptcy.com collect information for research and business purposes.

Finally, the availability of numerical methods and the development of computer techniques also increase the possibility of developing advanced models (Balcaen & Ooghe,

2006). New methods in discrete choice models, structure models and comparison proxies broaden the techniques in default risk assessment.

1.2 Principal Contributions

As defaults are extremely rare events, the shortage of data is a main concern for this kind of research. How to use the data more efficiently is a key question. This thesis constructs the most updated and comprehensive panel dataset, containing 2,123 defaults in 639,573 observations in the United States, from 1970 to 2009.

To ensure the efficiency of sample usage, the thesis adopts dynamic observations following studies (Shumway, 2001; Chava & Jarrow, 2004; Campbell, Hilscher, & Szilagyi, 2008) instead of the static sample in the other studies (Jones & Hensher, 2007; Beaver, McNichols, & Rhie, 2005). The dynamic observations method applies each firm's time series data with time-varying dependent variables, while the static sample method uses an observation of each firm only one period before a default event. In other words, the dynamic method uses multi-period observations for the same firm, while the static method associates with single period observation for each firm. Moreover, the thesis uses quarterly observations, instead of yearly as in most previous studies, following the suggestions from Chava and Jarrow (2004) that smaller intervals of the sample will achieve better prediction accuracy.

1.2.1 Default Drivers

This first unanswered question in predicting corporate default is: what are the most important components that trigger default? Among the extensive default drivers used in previous studies, the optimum information set for predicting default remains ambiguous. While Altman (2000) recommends accounting ratios as default drivers, Campbell, Hilscher & Szilagyi (2008) prefer firm-level market information. Further investigation on the debate over the importance of accounting ratios vs. market information in default

prediction is required. Moreover, the importance of macroeconomic indicators is suggested in previous studies (e.g. Duffie, Saita & Wang, 2007), which use macroeconomic variables in survival models to improve prediction accuracy. However, the application of macroeconomic variables in other prediction models (such as logit model, mixed logit model, etc.) remains unclear. Furthermore, although industry effects are suggested by Chava & Jarrow (2004), Campbell, Hilscher & Szilagyi (2008) argue that industry effects are insignificant. Finally, some information related to corporate default may not be numerically measurable (Zavgrenc, 1985), such as unmeasured qualities of assets, the creative ability of management, random events, and the decisions of regulators and the courts of law. The existence of these unobservable variables raises particular interests.

This thesis is a systematic investigation of the impact of alternative factors in an alternative framework provided by both traditional and random models. Unlike most previous studies (e.g. Abdullah, 2008) that adopt variables randomly from other papers, this thesis picks default drivers empirically using the t-test and stepwise methods. The findings contribute to the debate on the relative importance of accounting default and market default drivers. The thesis also bridges the two strands of literature on default risk prediction that tends to focus separately on the micro determinants, at firm level, with wider macro determinants. The thesis therefore assesses the relative explanatory power of both firm level information and macroeconomic information and investigates the link between business and credit cycles. Furthermore, the thesis verifies the significance of industry effects under alternative methods. Most importantly, the thesis suggests two advanced random methods to capture the unobservable default drivers that are conditioned on the optimal variable set.

1.2.2 Advanced Random Models

The second important issue in predicting default risk is the shortage of advanced models that could capture default clusters and correlations. Traditional models fail to represent

default clusters, and idealistically assume the independence between explanatory variables. While the discriminant analysis method is judged for the assumption of normality variables, the logit model suffers the limitation of Independence of Irrelevant Alternatives[1] (IIA) (Train, 2003) (Train K. , 2003). On the other hand, the innovative studies using advanced models (Duffie, Saita & Wang, 2007; Hensher & Jones, 2007; Hensher, Jones & Greene, 2007) are based on a narrow range of explanatory variables that lead to poorer predicting power. The advanced discrete choice models limit the default drivers within accounting ratios, while survival analysis and frailty models contain only market and macroeconomic information.

This thesis introduces two advanced random models, the mixed logit and frailty models, which could efficiently capture default clusters and unobservable information, thus improving prediction accuracy. It bridges the variables used in traditional models with the theoretical advantage of advanced random models, which allows the existence of the unobservable variables to be explored beyond the well-specified optimal variable set. This thesis extends the existing random methods of default prediction in three ways: first, by investigating whether for US firms the market variables and macroeconomic indicators add additional power to the prediction accuracy with the mixed logit model, as argued in Kalotay (2007); second, by exploring whether the usage of accounting ratios, explored in traditional methods, has affected prediction ability in frailty structure methods; and third, by exploring the potential default cluster factors with the frailty model. To this author's knowledge, this is the first empirical application of the mixed logit model using a sample taken from the United States. In particular, the thesis represents an advance in the comparison of advanced models from different categories.

[1] "For any two alternatives i and k, the ratio of the logit probabilities P_{ni}/P_{nk} does not depend on any alternatives other than i and k. Since the ratio is independent from alternatives other than i and k, it is said to be independent from irrelevant alternatives, or IIA." (p 45-46,Train, 2003)

1.2.3 Optimal Predicting Models

A further uncertain topic in predicting corporate default is the optimal method in theory and practice. Existing empirical studies present fragmentary and complex suggestions for the optimal default predicting models. Comparison studies concentrate mostly on traditional models, such as Z-score, logit model and the Merton's structure model (Lennox, 1999; Chava & Jarrow, 2004), with a constrained comparison criterion, the classification rate, which largely depends on the selection of cut-off points. With the development of new predicting methods, evaluation and comparison of traditional methods and new methods - such as the survival models and advance random models - are essential, especially using multidimensional objective practical comparison criteria.

This thesis collects large dimension default risk measurements from five categories, and assesses their predicting performances under the well-specified optimal information set. It bridges models from different categories with the default probability, thus allowing performance comparisons among the traditional discriminant analysis, conditional probability analysis, discrete choice model, parametric and semi-parametric survival analysis, and recommends the optimal default risk measurement. To this author's knowledge, the survival analysis models, the frailty models in particular, have not been assessed with models from other categories. Also, the logit model is the only reference model accompanying the mixed logit method; however the performance of the mixed logit model over other methods remains unclear. Moreover, less attention has been paid to the performance of the probit method. Besides the dimensions and number of models included in the comparison study, this thesis departs from previous model comparison research further with its model assessment criteria. After applying some commonly used criteria, such as the area under the receiver curve (AUROC), the accuracy ratio (AR), and Type I and Type II error comparisons, this thesis looks in depth at three more assessment benchmarks: the model stability during financial crisis, profits for creditors using alternative methods and two types of losses (capital loss and opportunity loss) associated

with misclassification.

1.3 Structure of the Thesis

This thesis contains six chapters: this introduction, a literature review, three empirical study chapters covering related topics and a conclusion chapter.

The literature review presented in Chapter 2 focuses on two main areas: a summary of default prediction models and a summary of previously used default drivers. Most of the methods mentioned so far are documented with their advantages and disadvantages. Default drivers are reviewed and recorded by sector. This thesis further summarizes the frequencies in previous studies of both the varying default models and alternative default variables. Other issues are also briefly reviewed, including the samples of previous empirical studies, model assessment methods, and variable selection and combination techniques.

Based on the literature review, Chapter 3 investigates the optimal default drivers with the dynamic logit model. It begins by describing the sample and data sources of empirical studies, followed by a number of investigations of optimal default drivers. Specifically, this thesis uses both the t-test and stepwise methods to eliminate variables from extensive sources. Robustness of the optimal variable set is verified with alternative sample periods and prediction horizons. The chapter ends with a discussion of sample prediction accuracy.

Chapter 4 starts with a discussion of the methodologies of the two random models, the mixed logit model and the frailty survival model. To compare the prediction power of these advanced models, two classic models are used for baseline comparison. For the mixed logit model, the first empirical result is a comparison of the standard logit model with the advanced mixed logit model, followed by the mixed logit model tests with alternative variable combinations, different simulation times and coefficient assumptions. The second part of the results compares the frailty model over the traditional hazard

model following a specification of variable set test with standard Cox survival model. The chapter further investigates the frailty factors, and then ends with a comparison of these two random models, theoretically and empirically.

The third empirical component is presented in Chapter 5, which compares twelve models using alternative criteria. To assess overall prediction accuracy, the chapter presents the results of the receiver operating characteristic curve, the accuracy ratios, the model stabilities, and the classification rate over different cut-off points. To analyse the economic value for creditors using the alternative model, the chapter compares indicators such as return over assets, assets loss under default, and opportunity loss, using the competitive credit market system.

After summarising the conclusions and contributions of the empirical findings, the final chapter also provides suggestions for further research.

Chapter 2: Literature Review

2.1 Introduction and Overview

The earliest recorded bankruptcy prediction research dates back to FitzPatrick (1932). Over the 20th century, exhaustive effort has been given to investigating the behaviour of corporate and bank default. Default prediction has developed with a wide range of methodologies and a broad collection of covariates. The empirical tests cover the U.S. and other countries. This chapter summarises and discusses research work in default prediction mainly from 1966 to 2010.

A few papers have reviewed this topic from varying perspectives and research models. Scott (1981) listed early stage empirical models and addressed the overlap problem of theoretical models including the single period model, the gambler's ruin model, and models with either perfect or imperfect access to external capital. Altman (1984; 1996) documented the empirical results of Z-score models internationally. Dimitras, Zanakis & Zopounidis (1996) confirmed the international interests of business default prediction by reviewing 47 papers from 1966 to 1993. They ranked these papers according to different countries, methods, and accounting ratios. They concluded that discriminant analysis is the most commonly used method, and working capital over total asset ratio is the most commonly used ratio. Summary and comparison of commercial models for portfolio credit risk is presented by several studies (e.g. Crouhy, Galai & Mark, 2000; Gordy, 2000). Moreover, Charitou, Neophytou & Charalambous (2004) documented failure prediction studies in the UK. Balcaen & Ooghe (2006) summarised the main features of classic models and their specific properties including assumptions, advantages and disadvantages. Alternative models and their properties are discussed in Balcaen & Ooghe (2004).

However, advanced random methods are excluded in their discussions, thus it remains unclear that if advanced random methods have a better performance than other classical methods. Kumar & Ravi (2007) reviewed papers using primarily artificial intelligent techniques. They concluded that researchers have employed nearly all the artificial intelligent techniques to predict corporate default, and the Neural Network method is the most commonly used. Aziz & Humayon (2006) compared the prediction accuracy of varying methods from different studies and stated that artificially intelligent expert system models perform only marginally better than statistical or theoretical models.

The aim of this review chapter is to present and summarise the development of default prediction methods, and address new trends in default prediction measurements in the 21st century. Particularly, the chapter discusses the advantages and disadvantages of both the prediction methods and independent variable categories. The papers summarised in this chapter are predominantly from journals in the fields of Finance, Accounting, Management and Operational Science. This review departs from previous reviews in the following respects: (1) it categorises papers into both methodologies and variables; (2) it is by far the most up-to-date and comprehensive review of new developments in the last decade (including the advanced discrete model and frailty model); (3) the variables categorised in this chapter include not only firm level information but also macro level indicators; (4) model comparison technologies and feature selection methods are also presented and summarised.

The review is conducted primarily in two dimensions: methodologies and variables. In the methodology dimension, the methods summarised comprise discriminate analyses (DA), conditional probability models (CPM), market-based structure models (MBS), Hazard intensity models (HIM), Advance Choice Models (ACM) and Intelligent Techniques (IT). Within the dimension of variables, the review categorises the observable variables suggested in previous studies into five groups: accounting ratios, market variables, macroeconomic variables, industry effect and other significant variables. Furthermore,

this chapter is written across other dimensions such as: sample chosen problem, feature selection technique and models comparison technology.

The rest of this chapter is organized as follows. Section 2 focuses on the default prediction methodologies, and summarises the various models. Section 3 reviews the variables used in previous studies. Other issues including sample chosen, feature selection and model comparison technique are presented in Section 4. Section 5 draws together conclusions.

2.2 Prediction Models

Researchers disagree as to the best default forecasting models. Figure 2.1 (see page 57) presents the summarised frequency of each method used in 165 previous studies from 1960 to 2010. Multivariate discriminant analysis is the most commonly used method, followed by the logit model and neural network.

The current section summarises the most commonly used methods and key conclusions regarding various default models. Advantages and disadvantages of each model are analysed. Particular attention is given to the way new models have developed in recent years, such as the mixed logit, frailty and nested logit models.

[Figure 2.1, p.67 about here]

2.2.1 Discriminant Analysis

The discriminant analysis method was the dominant model in predicting defaults in the 1960s and 1970s (Bellovary, Giacomino & Akers, 2007). The discriminant analysis group consists of the simplest Univariate Discriminant analysis (UDA) and Multiple Discriminant analysis (MDA) methods, generating Z-scores and Zeta scores. It has been shown that MDA has better prediction capability, with a higher accuracy rate than UDA (Bhargava, Dubelaar & Scott, 1998). Given the higher accuracy of MDA (Bellovary, Giacomino & Akers, 2007), it is generally regarded as a benchmark for other methods (Daubie & Meskends,

2002; Balcaen & Ooghe, 2006) and is thus the most commonly applied (Dimitras, Zanakis & Zopounidis, 1996; Balcaen & Ooghe, 2006; Aziz & Dar, 2006). MDA is also popular as an indicator for default risk in other financial phenomenan and analyses, such as abnormal returns for distressed firms (Wood & Piesse, 1987) and bank behaviour (Tamari, 1984). Several studies have shown that MDA has a higher success rate in recognising failure in firms than both the logit model and Neural Network method (Muller, Steyn-Bruwer & Hamman, 2009).

2.2.1.1 Univariate Discriminant Analysis

The application of discriminant analysis in default prediction dates back to Beaver (1966), who introduced single ratio discriminant analysis, whereby the default rate is evaluated by one ratio at each discrete time. UDA is easy to apply and favoured by studies investigating the significance of single variables, such as cash flow and return on assets (Bhargava, Dubelaar & Scott, 1998), and total liability over total assets (Miller, 2009). However, inconsistency and assumption of linearity are criticisms highlighted by other studies (Keasey & Watson, 1991; Balcaen & Ooghe, 2006). Also, the sufficiency of the single accounting ratio in explaining default behaviour is questionable, making UDA susceptible to problems of endogeneity and omitted variable bias since the correlations between ratios are neglected.

2.2.1.2 Multiple Discriminant Analysis

Pioneering MDA studies include Altman (1968), Deakin (1972), and Blum (1974), which sought to score a firm's credit risk using accounting ratios. All firms are classified into two groups, bankruptcy and non-bankruptcy, according to a chosen cut-off point in the score range. The high stability of MDA over other models has been shown in several studies (e.g. Bhargava, Dubelaar & Scott, 1998; Miller, 2009). The extension format of this method includes non-linear MDA (Aziz, Emanuel & Lawson, 1988). Moreover, this method has

been applied in different countries (Taffler, 1984; Izan, 1984; Takahashi & Kurokawa, 1984; Wang & Campbell, 2010; Lifschutz & Jacobi, 2010). Merrill Lynch has also explored the Z-Score as a default predictor.

The practical superiority of MDA, compared to the market-based model, is documented by Agarwal & Taffler (2008), who show that a bank using the Z-score approach would achieve a higher profit than a bank employing a market-based model such as Black-Scholes-Merton. Moreover, the Z-score model is empirically more accurate than market-based models, although neither model is statistically sufficient for failure prediction (Agarwal & Taffler, 2007; 2008). Compared to the logit model, the DA estimator is asymptotically efficient under certain assumptions (Lo, 1986). It is also noted that MDA saves the computational space for the analyst by using cut-off points (Altman, 1984).

However, although it is a popular baseline position, the traditional Z-score model is criticized for several reasons: First, the accuracy of the assumptions of MDA is contested in numerous studies (Lennox, 1999; Balcaen & Ooghe, 2006; Lin & Piesse, 2004). The basic premise of MDA is that all variables used to determine failure should be assumed to follow the normal distribution; but this is criticized in practice (Lennox, 1999; Lin & Piesse, 2004; Balcaen & Ooghe, 2006). Actually, most financial ratios are non-normal (Ezzamel & Mar-Molinero, 1990). Both non-univariate normality and multivariate normality have been discussed at length in the literature (Lin & Piesse, 2004). Once the normality assumption is violated, the prediction accuracy is damaged, since DA is not consistent (Lo, 1986). Also, the assumption of equal variance-covariance matrices across the bankruptcy and the non-bankruptcy parties needs to be tested before using linear MDA (Ohlson, 1980; Zavgrenc, 1985; Luoma & Laitinen, 1991; Lennox, 1999; Balcaen & Ooghe, 2006). If the dispersion matrix is not equal for two groups, then the linear classification procedure should not apply. Some studies (Taffler, 1984; Izan, 1984; Takahashi & Kurokawa, 1984) applied a more suitable quadratic discriminant method for non-equal variance-covariance matrices; however, Altman's original linear model has been shown to perform better than the

quadratic discriminant method (Altman, Haldeman & Narayanan, 1977; Neophytou & Molinero, 2001; Altman, 2000). These assumptions further limit DA in capturing default clusters and correlations between variables.

Second, MDA is a static model, which could only predict corporate default one step ahead. It does not explicitly incorporate time (Looney, Wansley & Lane, 1989). This makes manipulation of information inefficient (Shumway, 2001), and results in a failure to capture the risk trend over time for the same firm (Robertson & Mills, 1988). The stability of MDA has also been debated (Barnes, 1990).

Third, the interoperable ability of dependent and independent variables is poor. The relative importance of each explanatory variable in MDA is arbitrary (Ohlson, 1980; Balcaen & Ooghe, 2006). The matching principle of MDA does not produce an interpretable coefficient, since the least squares estimation does not apply to the binary dependent variable (Balcaen & Ooghe, 2006). This leads to confusion when reading the coefficients, and increases the difficulty for practitioners (Zavgrenc, 1985; Robertson & Mills, 1988). Furthermore, it is questioned whether the classification method for the Z-score should be determined as ex post-based rather than ex ante (Piesse & Wood, 1992). On the other hand, the Z-score could not be interpreted as default probability, and classification of the Z-score is largely dependent on prior probability, hence the MDA system is argued to be a false and biased prediction model (Ohlson, 1980; Balcaen & Ooghe, 2006). Moreover, Agarwal & Taffler (2007) have argued that the negative Z-score cannot explain default behaviour.

2.2.2 Conditional Probability Model

The conditional probability model was popular in the 1980s and 1990s (Bellovary, Giacomino & Akers, 2007). This model predicts default probability with a flexible set of factors using the maximum likelihood estimator. Based on different probability assumptions, the conditional probability model includes the linear probability model,

probit model and logit model. The linear probability model assumes that the probability of default has a linear distribution, while the probit model and logit model assume that it has a normal distribution and a logistic distribution respectively.

2.2.2.1 Linear Probability Model

Linear probability is applied by limited studies (e.g. Platt, 1989) since under the linear assumption, the default probability varies by the same increment in response to equal change in explanatory variables (Platt & Platt, 1990). In practice, the assumptions of linear probability are difficult to meet when the error term is not normally distributed and is heteroscedastic. Moreover, the dependent variable is normally arbitrarily distributed in this area of research, which gives less predicting power (Aziz & Dar, 2006).

2.2.2.2 Probit Model

The studies using the probit model and its extensions are also generally limited, but it is favoured in a few recent empirical tests. The superiority of a well-specified non-linear probit model over a traditional DA model is suggested by Lennox (1999), after testing misspecification and heteroscedasticity. Zmijewski (1984) used the probit model to examine the sample collection problem in default prediction. Amato & Furfine (2004) applied the ordered probit model to investigate credit rating and credit cycles. Bonfim (2009) tested macroeconomic variable effects using a random probit model. Given their theoretical similarities, the advantages and disadvantages of the probit model are analysed in the logit model subsection.

2.2.2.3 Logit Model

The conditional logit model is the most popular conditional probability model (Balcaen & Ooghe, 2006). The first generation of the logit model, also called O-score, is a static model dating back to Ohlson (1980). This model shows strong prediction ability with updated

data (Begley, Ming & Watts, 1996). Other researches (Zavgrenc, 1985; Keasey & McGuinness, 1990; Johnsen & Melicher, 1994) have extended the methodology with wider independent variables, longer prediction periods, and multiple default states. There have also been studies internationally such as in New Zealand (Peursem & Pratt, 2002), Taiwan (Wu & Kuo, 2004), the UK (Lin & Piesse, 2004), Malaysia (Abdullah, 2008), Turkey (Vuran, 2009), Norway (Westgaard & Wijst, 2001), and the Asia-Pacific region (Chi & Tang, 2006). The extended form of logit model includes the cumulated logit model, which could be used in multi-state default studies (Jones & Hensher, 2004; 2007) and rating transition (Kim & Sohn, 2008). However, this single period logit model uses only one, non-randomly selected observation per firm, which causes sample selection bias. Time-varying changes in default risk are omitted, which leads to cross-sectional dependence in the sample (Hillegest, Keatin, Cram & Lundstedt, 2004). These problems have been shown by Shumway (2001) to result in biased, inefficient and generally inconsistent coefficient estimations.

The second generation of the logit model, called the multi-period logit or dynamic logit, was first introduced by Shumway (2001) and has been widely used (e.g. Chava & Jarrow, 2004; Abdullah, 2008; Campbell, Hilscher & Szilagyi, 2008; Nam, Kim, Park & Lee, 2008). Theoretically, it equates to a discrete hazard model with the logit assumption. The superior efficiency of the multi-period logit model over the single period logit model is attributed to: (1) incorporating time-varying variables into failure prediction; (2) utilising more data; (3) distinguishing the corporate risk exposure period (Shumway, 2001). Empirically, the prediction power of the dynamic logit model over the single period logit model has also been demonstrated (Shumway, 2001; Abdullah, 2008; Nam, Kim, Park & Lee, 2008). In practice, the dynamic logit model has also been used for both the Kamakura Risk Information Service and McKinney's Credit Portfolio View.

Compared to the MDA model, the conditional logit model shows some theoretical advantages. First, the logit model relaxes the assumptions concerning normal distribution of the explanatory variables and the prior probabilities of default (Johnsen & Melicher,

1994). Secondly, the regression results from the logit model show a score from zero to one, which could explain default probability exactly. Moreover, the coefficient of explanatory variables could indicate the significance of each variable in predicting defaults. Furthermore, quantitative variables could be used as explanatory variables in the logit model, allowing more information to be investigated in relation to defaults, such as default history and industry effects. Finally, the non-linear shape of the logit function is appealing (Platt & Platt, 1990; Balcaen & Ooghe, 2006), which in practice indicates that it is hard to change the state of an extreme healthy (or distressed) firm.

The prediction ability and performance reported in previous studies for the logit model is mixed. The superiority of a well-specified single period logit model over the MDA model is confirmed by Lennox (1999) after testing misspecification rates and heteroscedasticity. In comparison studies (e.g. Zavgrenc, 1985; Begley, Ming & Watts, 1996; Chava & Jarrow, 2004), the logit model has shown better performance on overall predictive accuracy than alternative techniques including the MDA model and decision tree conditioning (Muller, Steyn-Bruwer & Hamman, 2009), and also some other intelligent techniques including the Neural Network, and recursive partitioning methods in the retail industry (Hu & Ansell, 2009). Additionally, the logit is more robust than DA in parameter estimation (Lo, 1986). Chava & Jarrow (2004) have also stressed the advantages of the multi-period logit model over the single-period logit model. On the other hand, Franzen, Rodgers & Simin (2007) argue that the logit method is less accurate with R&D intensity increasing over time. In a long-term empirical test, Campbell, Hilscher & Szilagyi (2008) showed that the multi-period logit model could fit the general time pattern quite well, but under-predicted the default frequency of failure in the 1980s and over-predicted it in the 1990s. Also, prediction ability decreases as the time horizon increases (Campbell, Hilscher & Szilagyi, 2008).

However, the assumption of the logit model, which is related to Independent and Identically Distributed (IID) errors and Independence of Irrelevant Alternatives (IIA), is

argued by previous studies (e.g. Train, 2003; Wu & Kuo, 2004; Hensher, Jones & Greene, 2007; Jones & Hensher, 2007). Moreover, Train (2003) has argued that these assumptions weaken the prediction power of the logit model, especially in a multi-state default model (Jones & Hensher, 2007). Besides, the logit model is argued to be sensitive to the multi-collinearity problem, missing values, and extreme non-normality (Balcaen & Ooghe, 2006). Hillegest, Keatin, Cram & Lundstedt (2004) complain that the logit model is a sample-based model, and as such its estimation accuracy is largely dependent on the sample selection process. This exposes any predictions to a risk of sample selection bias. Most importantly, the default cluster and default correlation between variables are not captured by the logit model. Finally, the suitability of the function for variable combination is questioned by Hwang, Cheng & Lee (2007). Extending the linear format of the logit model with an unspecified function may help to mitigate the parameter assumption.

2.2.3 Advanced Choice Models

Recently, advanced choice models including the latent class multinomial logit, error component logit, nested logit, and mixed logit, have been introduced as default risk predictors. The pioneering application of the choice method in default risk is Jones & Hensher (2004) who use the mixed logit method. An extended form of the mixed logit model, namely error component logit analysis, is also discussed in Hensher, Jones & Greene (2007). Among these methods, the mixed logit model is highly recommended in terms of underlying behavioural realism, desirable econometric properties, and predictive performance (Hensher & Jones, 2007). However existing empirical tests are limited to Australian datasets.

2.2.3.1 Mixed Logit Model

The mixed logit model is an open form discrete choice method with random parameters. Although the mixed logit method is not new, and has been applied in political analysis and

marketing research, its application in corporate default is in its infancy. Theoretically, as an open form discrete choice model, the mixed logit model fully relaxes the IID condition and IIA assumption from the conditional probability model, such as the logit model (Jones & Hensher, 2004). The key advantage of the mixed logit model in predicting default is its ability to capture additional information from the randomly weighted coefficients. This enables the model to explain corporate default clusters more precisely. Also, unobservable variables can be incorporated into the prediction, largely avoiding sample shortage. Practically, Jones & Hensher (2004) have confirmed the superior prediction power of the mixed logit model over the multinomial logit model regarding out-of-sample performance. Sampling problems and the distributions of random coefficients are discussed further in Hensher & Jones (2007).

However, as the mixed logit model has seldom been applied, it remains unclear whether this method would be efficient in other countries, such as the United States or the United Kingdom. Moreover, the predictive power of the mixed logit model compared with less sophisticated methods, such as the DA, is unknown. Kalotay (2007) has pointed out several problems in Jones & Hensher (2004), for example, market variables and macroeconomic indicators were ignored, and evaluation of the mixed logit model using the area under the receiver operating curve and the accuracy ratio is unknown.

2.2.3.2 Nested Logit Model

An alternative advanced model, the multinomial nested logit model, has also been applied to predict corporate default by Jones & Hensher (2007). In an empirical study, they compared the strengths and weaknesses of the multinomial nested logit model and the standard logit model. A two-level nested logit model is found to be the best model for a four-state default sample in Australia. Both the out-of-sample test and the model-fit parameters show the superiority of the multinomial nested logit model over the multinomial logit model (Jones & Hensher, 2007).

Since the nested logit model has a closed form solution, the calculation process is simpler and more convenient than that of mixed logit model (Jones & Hensher, 2007). It incorporated firm-specific unobserved heterogeneity to some extent (Jones & Hensher, 2007). However, this nested logit model fails to capture the correlation across nests, while the mixed logit model could seize the correlation between variables efficiently. Moreover, comparing with the mixed logit model, nested logit model only partially corrects the IID and IIA condition. Finally, judgement is required in determining which alternatives can be appropriately partitioned into nests (Jones & Hensher, 2007).

2.2.4 Market-Based Structure Model

The first market-based structure model (MBS) to analyse corporate default risk is constructed in Merton (1974), which modelled corporate default risk assuming that default is triggered when total assets (TA) are lower than the total liability (TL). This model views equity as a call option on the assets of the firm and assumes the strike price of the option equal to the face value of the liabilities. In practice, this model is also used by Moody's KMV, a crucial institution for valuing credit risk.

This theoretical model is preferred since it is identical with the efficient market hypothesis, and reliable even when accounting policy changes (Agarwal & Taffler, 2008). Also, this model is not sample dependent (Agarwal & Taffler, 2008). Hillegest, Keatin, Cram & Lundstedt (2004) recommend that the market-based structure model with market information performs better than classical models such as MDA and logit models with only accounting information. Miller (2009) confirms that the market-based structure model outperforms MDA and UDA using the TLTA ratio, in terms of both accuracy ratio and the I error. This superior performance is attributed to the use of significant additional information, namely assets volatility, provided by the MBS model. Moreover, Miller (2009) provides evidence that the MBS model generates more durable credit ratings than both MDA and UDA models. Furthermore, the BSM method provides an independent estimation

for every publicly traded firm, and releases the bias in sample selection procedures.

However, the assumptions required in the Merton model - for instance the normality hypothesis with the stock return and the single coupon loan constraint on each firm - are questioned in several papers (Hillegest, Keatin, Cram & Lundstedt, 2004; Duffie, Saita & Wang, 2007; Agarwal & Taffler, 2008; Bharath & Shumway, 2008). These idealistic assumptions introduce errors and biases into the process of predicting the probability of default. In addition, the variables required in model estimation, such as asset value and volatility and the expected return on assets, is not practically observable, and so prediction accuracy relies on the proxy accuracy of asset value and volatility (Altman & Saunders, 1998; Crouhy, Galai & Mark, 2000). However, as Zhou, Xie & Yuan (2008) argue, asset value can change with a changing financial environment, and may cause a non-stationary process. Moreover, this model only includes information from the financial market; its prediction ability would lessen if the market efficiency assumption were violated. Bharath & Shumway (2008) assessed the Merton model and concluded that it is not sufficient to explain default, as it limits default drivers in market scope. Additionally, the economic benefit acquired by financial institutions using this market-based model is inferior compared to some accounting-based models, such as the Z-scores (Agarwal & Taffler, 2008). Finally, this model could not be used for non-publicly traded firms, but of course such firms also suffer from default risk (Altman & Saunders, 1998).

2.2.5 Hazard Intensity Survival Model

Instead of determining whether a firm would default on its credit, another family of models, called duration models or hazard models, has been introduced to capture the survival time of a credit event. Besides predicting the possibility of default, these models can also estimate the time of default, and the life cycle of a firm (Whalen, 1991). There are various choices of intensities within both the continuous and discrete models. Continuous models include parametric models, semi-parametric models and non- parametric models.

Examples of the parametric models include the Weibull model, the Gompertz model, the exponential model, the log-normal model, the generalized gamma model and the log-logistic model. The Cox proportion model is an example of the semi-parametric model. The non-parametric models include Kaplan-Meier and Nelson-Aalen.

The hazard intensity model is preferred for relaxing the assumption about the distribution of independent variables (Luoma & Laitinen, 1991; Whalen, 1991). The Cox proportion model is the most popular Hazard model in default prediction, and was first applied to analyse credit risk for banks (Looney, Wansley & Lane, 1989; Whalen, 1991; Bonfim, 2009). Early studies (Looney, Wansley & Lane, 1989; Luoma & Laitinen, 1991; Whalen, 1991) claimed superior prediction ability for the static hazard model - which assumes that the covariates are constant over time - over the MDA and logit models. However, the static assumption limits the predicting potential of the hazard model. Recently, time-varying covariates have been applied in both the discrete hazard model and continuous hazard model. The time-varying covariates enable the hazard model to predict for several future periods, which enables efficient manipulation of data. Bonfim (2009) suggests that both Weibull and log-logistic distributions have a better prediction power over other parametric hazard models using the time-varying covariates. The time dimension in the hazard model may enhance the prediction power when incorporating macroeconomic factors (Bonfim, 2009).

The discrete multi-period hazard model was first employed by Shumway (2001) and subsequently in several other studies (Chava & Jarrow, 2004; Beaver, McNicholes & Rhie, 2005; Campbell, Hilscher & Szilagyi, 2008). This dynamic discrete model was also validated as a superior forecasting model to static models using expanded databases (Chava & Jarrow, 2004). Meanwhile, Carling, Jacobson, Linde & Roszbach (2007) applied the Cox proportion model with multi-period discrete covariates, and others (Das, Duffie, Kapadia & Saita, 2007; Duffie, Saita & Wang, 2007) have explored continuous time-varying explanatory variables with the Cox hazard model.

However, the hazard model may suffer from left censoring problems when analysing loan defaults over a short period (Bonfim, 2009). Also, the prediction of default time would be affected by the information release period (Luoma & Laitinen, 1991). Moreover, Koopman, Kraussl, Lucas & Monteiro (2009) argue that any model using only observant variables may underestimate default risk. Das, Duffie, Kapadia & Saita (2007) test the doubly stochastic assumption of Duffie, Saita & Wang's (2007) model, and suggest that there were missing variables in their empirical test. Most importantly, both the static hazard model and the dynamic hazard model fail to capture corporate default correlations and clustering.

Interest in capturing the default cluster and missing variables has led to recent investigations into the frailty hazard model. Zhou, Xie & Yuan (2008) recommend the theory of the frailty Cox model, which is supported by the empirical study by Duffie, Eckner, Horel & Saita (2009), which examines corporate default correlations. The authors introduce a time-varying latent frailty factor to the hazard model, and give evidence that there are unobserved variables that influence the prediction of corporation default. The frailty model is employed to investigate other credit risk related topics, including credit rating changes (Koopman, Lucas & Monteiro, 2008) and the relationship between credit cycles and macro variables (Koopman, Kraussl, Lucas & Monteiro, 2009). The significance of the unobservable common factors has been shown in these studies, amongst others.

The random intensity model has also been employed to identify the relationship between credit cycle valued by a change of credit rating and macro variables (Koopman, Kraussl, Lucas & Monteiro, 2009). Koopman, Lucas & Schwaab (2010) demonstrate the significance of frailty factors for different industries under large sets of macroeconomic variables. Koopman, Lucas, & Monteiro (2008) conclude that the impact of the random factors is higher for downgrade than for upgrade. They also conclude that the frailty factors contribute more before and during financial crisis.

2.2.6 Artificial Intelligent Techniques

This section reviews the main intelligent techniques used in previous studies. Kumar & Ravi (2007) conclude that, within the intelligent group, neural networks are the most popular methods, followed by Rough sets, Case-based reasoning, Operations research, Evolutionary approaches, and other techniques subsuming Fuzzy logic.

2.2.6.1 Neural Network

Neural network (NN) methods transfer information from explanatory variables via layers and nodes to predict default probability, similar to the vast network of neurons in the human brain. Daubie & Meskends (2002) and Bellovary, Giacomino, & Akers (2007) point out that neural networks dominated the research of the 1990s. Neural network methods are not subject to statistic assumptions such as linear relation and/or multivariate normality (Charalambous, Charitou & Neophytou, 2000). They also hold the advantage of dealing with nonlinear relations and accommodating missing information (Charalambous, Charitou & Neophytou, 2000).

There is a considerable literature using NN to predict bankruptcy, because of its high level of accuracy, for example Atiya (2001). There are also studies comparing the performance of NN with other, methods including logit (Barniv, Agarwal & Leach, 1997; Charalambous, Charitou & Neophytou, 2000; Lin & McClean, 2001), MDA (Wilson & Sharda, 1994; Jo & Han, 1997; Barniv, Agarwal & Leach, 1997; Charalambous, Charitou & Neophytou, 2000; Lin & McClean, 2001; Muller, Steyn-Bruwer & Hamman, 2009; Min & Lee, 2005), case-based reasoning (Jo & Han, 1997), and decision tree (Muller, Steyn-Bruwer & Hamman, 2009). A general conclusion derivable from these studies is that the stability, robustness, and prediction power of NN is superior to other methods.

The major criticism of NN is that its input and processing is a black box (Altman & Saunders, 1998; Peursem & Pratt, 2002; Kumar & Ravi, 2007). The contribution of each

variable to defaults could not be shown directly from the neural network. Thus, this method could not be used to investigate the importance of each factor, which largely limits the ability of auditors and managers to identify default sources. Second, the prediction ability of the NN method largely depends on its physical architecture, such as the number of layers and the number of elements in each layer (Charalambous, Charitou & Neophytou, 2000). Prediction ability varies with different algorithms (Pendharkar & Rodger, 2004). There have been criticisms of over-parameterization in NN, and also that the weight in NN is not directly interpretable (Charalambous, Charitou & Neophytou, 2000). Finally, the long process time and over fitting is argued by other studies (e.g. Aziz & Dar, 2006).

2.2.6.2 Rough Set

Rough set (RS) is an approximation using the lower and upper approximations of the original set. These do not require particular functional form, and are free of the adoption of restrictive assumptions concerning the distributions of both model variables and errors (Beynon & Peel, 2001). Kumar & Ravi (2007) point out that RS can sometimes be impractical to apply as it may lead to an empty set. The accuracy of the RS model largely depends on a specific dataset. The RS model also suffers problems such as over-sensitivity to noise, multimodality, and lack of performance-oriented fitting to task requirements (Aziz & Dar, 2006). The prediction ability found in previous studies using RS is reasonable, but lower than that of the MDA and logit models (Beynon & Peel, 2001; McKee & Lensberg, 2002). Ahn, Cho & Kim (2000) suggest that combining RS and NN could generate better prediction ability.

2.2.6.3 Decision Tree

Decision Tree (DT) forecasts default risk by a recursive partitioning technique, which partitions data into sub-classes and then recursively replaces each of the subsets with a DT node until the final nodes of the tree contain unique risk outcomes: default or non-

default. DT is a non-parametric method, which is not subject to assumptions such as multivariate normality distribution, equal covariance, and distribution assumptions.

Frydman, Altman & Kao (1985) show that, in most cases, DT outperforms MDA in terms of prediction accuracy, in cross-validated and bootstrapped results. Lin & McClean (2001) also confirm that DT performs better than both MDA and logit. However, in an empirical test by Beynon & Peel (2001), the out-of-sample prediction accuracy is poorer for DT than MDA, logit or RS.

One of the disadvantages of the DT is that, as a forward selection method, it may lose some important explanatory variables. Also, over-fitting using DT is cautioned by Kumar & Ravi (2007) and Aziz & Dar (2006). Most importantly, DT is only a classification technique; it is not possible to compare different types of default risk between firms, which largely limits its potential ability in risk management.

2.2.6.4 Case-Based Reasoning Model

Case-based reasoning (CBR) classifies a new firm based on previously predicted firms in the same domain of explanatory variables. Park & Han (2002) conclude that employing the CBR analytic hierarchy process, which uses k nearest neighbour weighting performs better than pure k nearest neighbour CBR. The prediction power of CBR is poorer than that of the logit model (Bryant, 1997). CBR is seen as lacking a convincing methodology, namely interviewing human experts and collecting cases. The index selection process is also questionable. Jo & Han (1997) also conclude that CBR is not appropriate for default prediction due to the low correlation between dependent and independent variables.

Other models used without preference in previous studies in default prediction include the Markov model, Genetic Algorithms, Balance Sheet Decomposition, Multidimensional scaling approach, Sequential minimum optimization, Linear gambler's ruin score etc.

2.3 Variables

This section summarises the default drivers used in previous studies from alternative sources and categorises the significant observable variables into five groups: accounting ratios, market variables, macroeconomic variables, industry indicators and other variables mentioned in previous studies. Table 2.1 presents 76 empirical studies, which have been summarized in this section. For each category, this section discusses the definition, advantage, disadvantage and key conclusions from previous literatures.

[Table 2.1, p 61 about here]

2.3.1 Accounting Ratios

Public financial reports contain considerable financial accounting information on firm performance, which is why early attempts at bankruptcy prediction were all based on accounting ratios (e.g. FitzPatrick, 1932; Beaver, 1966; Altman, 1968; Ohlson, 1980). Corporate default is not a sudden event; it depends on long-term underperformance, and this could be revealed by accounting information. Moreover, accounting ratios could be used for both public and private firms. Also, Agarwal & Taffler (2008) suggested that the assessment of a loan is normally based on accounting information, which makes accounting ratios more likely to reflect firm credit risk.

[Figure 2.2 p.68 about here]

[Table 2.2, p 63 about here]

In previous studies there are more than 185 different accounting ratios with significance in predicting corporate defaults. Figure 2.2 displays accounting ratios that have shown to be significant in at least four applications published from 1960 to 2010. The definition of these variables is listed in Table 2.2. The Return Over Asset (ROA) ratio (Net Income/Total Assets, or NITA) is by far the most important predictor of default risk, followed by

Working Capital ratio, Earnings before Interest, Tax over Total Assets, Sales ratio, Retained Earnings ratio, Total Debt ratio, Current ratio, and Total Liability ratio. The extended form of accounting ratios, such as the Depreciation form or the quadric form is used in some studies (Lennox, 1999).

The key ratios and variables are summarized in five categories: profitability, liquidity, operation efficiency, capital structure and firm size. In summary, firms with low profit, low liquidity, high leverage, high leverage and small total asset size are more likely to default.

The profitability ratios reveal the ability of a firm to generate profit. They illustrate how well management is employing the firm's assets to make profit. A higher profit associates with a more efficient management in utilizing the asset, thus decrease the default risk. Previous studies (McKee & Lensberg, 2002; Park & Han, 2002; Ohlson, 1980; Zmijewski, 1984; Lo, 1986; Lennox, 1999; Shumway, 2001; Peursem & Pratt, 2002; Wu & Kuo, 2004; Vuran, 2009 ;Atiya, 2001; Back, Laitinen, & Sere, 1996; etc) reveal that the amount and stability of corporate profit influence the probability of corporate default. There are five ratios have shown to be significant for more than four times in previous studies, the net income ratio (NITA), earning before interests over total assets ratio (EBITA), earning before interests over total liability ratio (EBIATL), binary ratio indicates the profit before default (INTWO). NITA is the most commonly used accounting variable in previous default prediction research, as shown in Figure2.2. This ratio reveals how profitable a company is relative to its total assets. Both EBITA and NITA reflect firms' return on their investments. EBIATL reveals the percentage of the earnings over total liability. INTWO shows whether the firm made a profit or loss during the last two quarters. As presented in Table 2.2, a firm with high NITA, EBITA, EBIATL and low INTWO ratio associated with low default risk. (Beaver W. H., 1966; Ohlson, 1980; Taffler R. J., 1984; Lo, 1986; Campbell, Hilscher, & Szilagyi, 2008; Chava & Jarrow, 2004)

Operational efficiency reflects managerial ability and how efficiently the firm is using its current and fixed assets. Efficient management leads to optimal allocation of resources,

and integration to achieve corporate goals effectively. In such a situation, operational income increases, and raising capital for future investment opportunities becomes easier, resulting in a potential increase in profit (Lin & Piesse, 2004). Seven ratios that reflect operational efficiency showed significance for more than 4 times in previous studies (Frydman, Altman, & Kao, 1985; Bryant, 1997; Ohlson, 1980; Altman & Narayanan, 1996; Park & Han, 2002; Lattinen, 1993; Johnsen & Melicher, 1994; Muller, Steyn-Bruwer, & Hamman, 2009; Back, Laitinen, & Sere, 1996; etc) are: Assets Turnover ratio (SATA), Sales Growth rate (SALEG), Retained Earnings ratio (RETA), Cash Flow over Sales (CFSA), Cash Flow over Debt (CFTD), Cash Flow over Total Liability (CFTL) and Net Operation Cash Flow over Total Assets (NOCFTA). SATA measures a firm's efficiency in managing its assets in relation to the revenue generated. RETA indicates the proportion of net earnings reinvested in the firm itself. CFSA provides information of a firm's ability to turn sales into cash. CFTD and CFTL indicate a firm's ability to cover total debt with its quarterly cash flow from operations. NOCFTA reveals a firm's ability to generate cash in relation to the asset size. As presented in Table 2.2, a firm with high SATA, SALEG, RETA, CFSA, CFTD, CFTL and NOCFTA suggested efficient operation, such associated with lower default risk (Altman E. I., 1968).

Liquidity ratios reveal the ability of firms to meet their debt obligation. Whether a firm could pay back its debt and interest directly decides the default risk (Lin & Piesse, 2004). In previous studies (Chava & Jarrow, 2004; Wu & Kuo, 2004; Johnsen & Melicher, 1994; Lennox, 1999; Chi & Tang, 2006; Deakin, 1972; Barnes, 1990; Back, Laitinen, & Sere, 1996; Beynon & Peel, 2001; Hillegest, Keatin, Cram, & Lundstedt, 2004; Jones & Hensher, 2004; etc), the Current ratio (CACL), Quick ratio (QACL), Working Capital ratio (WCTA), Cash ratio (CASHCL), Quick Assets over Total Assets (QATA), Current Assets over Total Assets (CATA), Current Liability over Total Assets (CLTA), and Cash over Total Assets (CASHTA), are all surrogates for solvency for more than 4 times. WCTA, QATA, CLTA, CATA indicate a firm's ability to cover its short term financial obligations comparing with the total asset

size. CACL, QACL, CASHCL reveal whether a firm's short-term assets are readily available to pay off its short-term liabilities. A high or increasing of these ratios indicate the liquidity of the firm is improving. The choice based of cash-based, current assets-based, quick ratio-based or working capital-based liquidity ratio is not conclusive. As presented in Table 2.2, a firm with high CACL, QACL, WCTA, CASHCL, QATA, CATA, CASHTA ratio and low CLTA ratio is expecting lower probability of the default risk (Beaver W. H., 1966; Altman E. I., 1968; Ohlson, 1980; Taffler R. J., 1984; Lo, 1986; Campbell, Hilscher, & Szilagyi, 2008).

Capital structure reveals the composition of a firm's liabilities. Previously suggested (Jones & Hensher, 2004; Hensher, Jones, & Greene, 2007; Abdullah, 2008; Vuran, 2009; Miller, 2009; Wilson & Sharda, 1994; Bonfim, 2009; Campbell, Hilscher, & Szilagyi, 2008; Beaver, McNicholes, & Rhie, 2005; Begley, Ming, & Watts, 1996; etc) capital structure ratios include capital structure ratios useful for default prediction include: Total Liability ratio (TLTA), Total Debt over Total Equity (TDTE), Total Debt over Total Assets (TDTA), Market Equity over Total Liability (METL), Market Equity over Total Debt (METD) and Total Liability and Assets Difference Indictor (OENEG). The debt ratio (TDTE, TLTA) reflect the proportion of a firm's assets that are provided via debt. OENEG summarizes the total liability and total assets as a binary variable. Ratios TDTE, METL and METD indicate the proportion of equity and debt a firm is using to finance its assets. Modigliani and Miller's famous proposition shows that debt leverage lowers tax payments, whereas equity does not present the same benefit as dividend payments are not tax deductible. A higher debt rate therefore potentially increases the value of the firm. Debt, however, has its disadvantage if it leads to costly financial distress. Thus, the optimal capital structure exists where the benefits of debt financing trades off these potential costs as the increased borrowing lead to an increase in the risk of default risk (Lin & Piesse, 2004). As presented in Table 2.2, a firm with high TLTA and low TDTE, TDTA associate with low default risk (Beaver W. H., 1966; Ohlson, 1980; Lo, 1986; Campbell, Hilscher, & Szilagyi, 2008).

Finally, firm size measured using the total asset has shown significant in some studies

(Altman, Haldeman, & Narayanan, 1977; Ohlson, 1980; Johnsen & Melicher, 1994; Peursem & Pratt, 2002; etc). LOGTAGNP indicates the logarithm GNP based total assets firm size and LOGTA is the logarithm of the total assets. Ohlson (1980) shows that smaller firms are associated with high default risk, as summarized in Table 2.2. However, (Hol, 2007) argues that larger firms that have more liquidity and solidity are less likely to go bankrupt.

Using only accounting ratios as prediction factors is argued by several studies (Gilbert, Menon, & Schwartz, 1990; Robertson & Mills, 1988; Shumway, 2001; Hillegest, Keatin, Cram & Lundstedt, 2004). Market-value Total Assets ratios have greater prediction ability than book-value Total Assets ratios (Campbell, Hilscher & Szilagyi, 2008). Chava & Jarrow (2004) also argue that accounting variables add little predictive power when market variables are already included in the bankruptcy model.

There are seven reasons to question the effectiveness of accounting ratios as a measurement of default risk. (1) The release time of accounting information is arbitrary and available, at best, quarterly, which can deteriorate prediction accuracy (Luoma & Laitinen, 1991) and cause an infrequently updated valuation (Bystrom & Kwon, 2007). (2) Accounting information is designed to measure prior information rather than future performance (Hillegest, Keatin, Cram & Lundstedt, 2004; Vassalou & Xing, 2004; Bystrom & Kwon, 2007; Agarwal & Taffler, 2008). (3) Financial statements are designed under the going-concern principle, assuming that firm would not default (Hillegest, Keatin, Cram & Lundstedt, 2004). (4) Accounting information includes only asset value, not asset volatility (Hillegest, Keatin, Cram & Lundstedt, 2004; Vassalou & Xing, 2004). Asset value could be underestimated relative to market value due to the conservatism principle. The default risk could be different even in firms with the same leverage ratios if the degree of asset volatility differs. Higher asset volatility is normally associated with a higher default risk. (5) The reliability of accounting ratios is therefore questionable, especially in terms of comparability across sectors, countries, accounting rules and other categories of

difference. For example, Hodges, Cluskey & Lin (2005) and Platt & Platt (1990) have demonstrated that total assets tend to increase at a higher rate than the balance of the cash account, thus deteriorating the prediction power of cash over the total assets ratio. Inflation may also influence the predicting power of accounting variables. As Platt, Platt, & Pedersen (1994) have demonstrated that accounting ratios also have the effect of deflation (6) Accounting information is at risk of manipulation (Agarwal & Taffler, 2008). Changes to accounting systems, and the development of accounting standards, can influence predicting ability considerably. Moreover, Beaver, McNicholes & Rhie (2005) conclude that the prediction power of accounting ratios is lessened due to increased managerial discretion under general accepted accounting principles, or the increase in intangible assets not offset by improvements from additional FASB standards to ensure they are adequately represented in financial statements.

2.3.2 Market Variables

Incorporating market variables into default prediction is suggested in several studies (e.g. Keasey & Watson, 1991; Chava & Jarrow, 2004). However, studies using market information are limited. The most commonly used market variables are stock returns and stock volatilities, at both firm and macro-market level. Figure 2.3 lists all the significant market variables used in previous studies. Table 2.3 presents the definition of all the market variables used in previous studies and the following chapters. Following the recent trend, this thesis will investigate all these market variables in the following chapters. These market variables could groups into six categories: stock return (RETURENFQ and EXCESS), stock volatility (SDF, PRINCETREND and QPRID), market capital size (LOGMC and FMVTMV), stock price (PRICE and LOGPRICE), market to book ratio (MBR) and the earnings per share ratio (EPS).

The relationship between firm stock return and the default risk has been studied by several literatures (e.g.Campbell, Hilscher, & Szilagyi, 2008). RETURNFQ is the most

commonly used variable to indicate the stock return, which has been shown as significant by Taffler (1984), Beaver, McNicholes & Rhie (2005), Duffie, Saita & Wang (2007), Campbell, Hilscher & Szilagyi (2008), Chava & Jarrow (2004), and Hwang, Cheng & Lee (2007). Campbell (2008) and (Chava & Jarrow, 2004) suggest an alternative form to measure the stock return, which is the difference between the firm return and the whole market return (EXCESS). This is the second most popular market variable. Campbell, Hilscher, & Szilagyi (2008) sugguest that an decrease in EXCESS reduces the probability of default by 28% of its initial value. It furthermore shows that stocks with high default risk associates with anomalously low average returns. Vassalou & Xing (2004), on the other hand, address that higher stock return is associated with higher default risk only when the firm size is small and the book to market ratio is high.

The importance of the firm level stock volatility has been addressed by recently literatures (e.g. Shumway, 2001; Chava & Jarrow, 2004; Campbell, Hilscher, & Szilagyi, 2008; Beaver, McNichols, & Rhie, 2005). In this thesis, the equity risk is measured by the standard deviation of the firm's monthly stock return (SDF), the price trend (PRICETREND), and the price gap (QPRID). As presented in Table 2.3, high default risk associates with high stock volatility (Campbell, Hilscher, & Szilagyi, 2008; Chava & Jarrow, 2004).

Campbell (2008) has also suggested the stock price (PRICE) in the logarithm form LOGPRICE as default predictor. An increase in equity prices leads to a decrease in firm leverage, and thus indirectly pushes down default probabilities. The significance of trends in market share price has been shown by Wu & Kuo (2004). Campbell, Hilscher, & Szilagyi (2008) sugguest that an decrease in price per share reduces the probability of default by 56% of its initial value.

Market to book ratio (MBR) has also been shown to be significant by several studies (e.g. Altman E. I., 1968; Back, Laitinen & Sere, 1996; Lin & Piesse, 2004; Campbell, Hilscher & Szilagyi, 2008). Altman E. I. (1968) sugguests that this measrue reveals the degree that a

firm's assets can decline in value before the liabilities exceed the assets and the firm becomes financial distressed. Campbell, Hilscher, & Szilagyi (2008) sugguest that an increase in MBR reduces the probability of default by 9% of its initial value. Other notable variables include the firm size (FMVTMV) as measured by the log-ratio of firm market capitalisation over total market capitalisation and the natural logarithm of the firm market capacity (LOGMC), which has been shown to be significant by Frydman, Altman & Kao (1985). Campbell, Hilscher, & Szilagyi (2008) sugguest that an increase in maket capitalization reduces the probability of default by 17% of its initial value. Moreover, Foreman (2003) suggests the Earnings per share (EPS) as default driver and a higher EPS associates with a low default risk for the telecommunication industry (Foreman, 2003).

[Figure 2.3 p. 69 about here]

[Table 2.3, p 64 about here]

The benefits of using market information as default drivers are as follows. 1) Market information contains investor expectations of future performance rather than purely past information (as a financial statement does) (Vassalou & Xing, 2004). 2) The volatility of the stock return could be well represented by market variables (Beaver, McNicholes & Rhie, 2005). 3). While accounting information is available at best on a quarterly basis, market-based data is more current, for example stock price changes are available daily (Beaver, McNicholes & Rhie, 2005) or at an even higher frequency. 4) Market variables are not subject to the influence of accounting regulation or manipulation from management. In an efficient market, market variables would provide more information than accounting variables (Agarwal & Taffler, 2008). However, market price potential does not reveal all information from financial statements if the market is not efficient (Hillegest, Keatin, Cram & Lundstedt, 2004). Moreover, market ratios are not available for private companies.

2.3.3 Macroeconomic Variables

Using macroeconomic variables as an extension of classical models is suggested by Keasey & Watson (1991). The importance of systematic factors as indicators of credit risk has been emphasised by others (Couderc & Renault, 2005; Carling, Jacobson, Linde & Roszbach, 2007; Bonfim, 2009; Koopman, Kraussl, Lucas & Monteiro, 2009). Few studies use macroeconomic variables as a default predictor. Of the four large credit portfolio models used in financial risk management - CreditMetrics, KMV, CreditRisk and Credit Portfolio View (CPV) - CPV is the only one that uses macroeconomic variables to predict the probability of default. Previous studies investigating macro effects on default are limited to loan defaults instead of corporate defaults (Carling, Jacobson, Linde & Roszbach, 2007; Bonfim, 2009).

Macro factors have the advantage that they are not susceptible to managerial manipulation, and easy to incorporate into time-varying models. They can be used for both public and private firms, domestically and internationally. Carling, Jacobson, Linde & Roszbach (2007), Hol (2007), and Bonfim (2009) have shown that although firm-level explanatory variables are documented as central factors leading to corporate default, macroeconomic variables could contribute further to establishing credit risk.

[Figure 2.4, p.70 about here]

[Table 2.4, p 65 about here]

Figure 2.4 presents the significant macro variables in credit risk studies from 1960 to 2010. Market index return, GDP growth, short- and long-term interest rates, and default spreads, are the most popular default predictors. Table 2.4 distinguishes four blocks of original macroeconomic variables: economic cycle, bank lending and investment condition, bond yield and interest rate and stock market information.

The stock market information block includes return on S&P 500 (SPRETURN) and the

standard deviation of S&P 500 return (SDM). SPRETURN measures both short- and medium-term economic performance, and negative impact on default probability is expected. The volatility of the S&P 500 is often used as a proxy for the volatility of indexed firms' assets, which is an important driver of the Merton default model (Couderc & Renault, 2005). The monthly standard deviation over the last quarter is used as the estimation of total market risk. If the stock market were efficient, this block would reveal all the macroeconomic level information from other blocks.

Investigating the link between default risk and economic information helps to understand the relationship between the credit cycle and economic cycle. The general economic information contains quarterly observations of Consumer Price Index (CPIAUCSL), Producer Price Index (PPIACO), Real Gross National Product (GNPC96), Gross National Product (GNP), Real Gross Domestic Product (GDPC96), Gross Domestic Product (GDP), Civilian Unemployment Rate (UNRATE), Unemployment rate, Real Imports of Goods and Services (IMPGSC1), Real Export of Goods and Services (EXPGSC1), National Income (NICUR), and Industrial Production Index (INDPRO).

Besides the information of the general economic indicators and the stock index, this thesis expects that indicators about bond yield and interest rate add valuable information to the default probability. The original analysis contains 9 original variables of bond yields and interest rates: AAA and BAA rating corporate bond yield (AAA, BAA), 3 month and 6 month treasury bills (DTB3, DTB6), 1 year and 10 year treasury constant maturity rates (DGS1, DGS10), 1 month certificate of deposit rates (CD1M), effective federal funds rates (DFF), and bank prime loan rates (DPRIME). In addition, this chapter uses d the following indicators: structure of different corporate bonds (BAA-AAA), long-term and short-term treasury rates (DGS10-DGS1) and BAA bond rates and 10-year treasury rates (DGS10-BAA). A higher bond yield attracts more investors with increased risk appetite, but exposes a firm to paying back a higher amount of interest, which may decrease the total net income.

The bank lending and investment block includes information about loans, investment, money supply, and mortgages. The four variables used to describe the loans are: commercial and industrial loans (BUSLOANS), individual loans (CONSUMER), real estate loans (REALLN) and total loan and lease in all commercial banks (LOANS). This chapter also includes the investment at all commercial banks (INVEST) and Federal Government Debt (GFDEBTN). This chapter measures the aggregate money supply by the economic with the Non-M1[2] Components of M2[3] (NOM1M2) since the data of M2 money stocks (M2) and M1 money stocks (M1) is not sufficient for the 40-year test. These policy indicators of the Federal Reserve Bank may influence default probability indirectly.

Conclusions of the relationship between macroeconomic variables and default risk include the following. 1). Kim & Sohn (2008) claim that discount rate, unemployment rate, and GDP growth rate have high correlation with the proportion of risk rating downgrades to upgrades. 2) GDP growth, the short-term interest rate, default spreads, and stock market volatilities remain relatively significant after adding unobservable variables into the duration model (Koopman, Kraussl, Lucas & Monteiro, 2009). 3) The volatility of foreign exchange rates is an explanatory variable in assessing credit risk (Nam, Kim, Park & Lee, 2008). 4) The aggregate U.S. failure rate is negatively related to the rate of economic growth and credit availability, and positively related to the formation rate of new businesse s (Platt, 1989).

Some studies have investigated the link between business cycles and credit cycles. Conclusions include the following: Vassalou & Xing (2004) and Amato & Furfine (2004) conclude that default risk is a systematic risk that varies with business cycles. Bankruptcy is more likely when the economy moves from boom to recession, while it is less likely to occur if the economy is currently in recession (Lennox, 1999). Macro-variables tend to

[2] M1 is the total of all physical currency part of bank reserves + the amount in demand accounts ("checking" or "current" accounts).
[3] M2 is M1 + most savings accounts, money market accounts, retail money market mutual funds, and small denomination time deposits (certificates of deposit of under $100,000).

explain the default rate more when in the low default regime than the high default regime (Banachewicz, Lucas & Vaart, 2008). Macro-factors can be relatively strong explanatory variables when the default risks are increasing (Koopman, Kraussl, Lucas & Monteiro, 2009). Credit cycles and business cycles are interdependent, but the correlation is far from perfect (Banachewicz, Lucas & Vaart, 2008).

The main criticism of using macro variables as a default predictor is that all firms are exposed to the same macroeconomic conditions, thus a model based only on macro factors may predict only market default intensity instead of default probability for a single firm. Koopman, Lucas & Schwaab (2010) conclude that although systematic factors accounts for 30% of the total default risk, any model based only on macroeconomic variables would either under- or over-estimate the default risk. It is also important to point out that macro variables may lose their significance when adding other and micro-variables (Koopman, Kraussl, Lucas & Monteiro, 2009).

2.3.4 Industry Effects

The potential importance of industry effects as an explanatory variable to analyse credit risk is addressed in several studies (Chava & Jarrow, 2004; Couderc & Renault, 2005; Carling, Jacobson, Linde & Roszbach, 2007; Banachewicz, Lucas & Vaart, 2008; Bonfim, 2009; Koopman, Kraussl, Lucas & Monteiro, 2009; Koopman, Lucas & Schwaab, 2010). Industry effects are posited as a potential explanation for the credit contagion phenomenon. Chava & Jarrow (2004) have emphasised that different industries face various levels of competition and different accounting conventions, which may contribute to differences in default probability. The importance of industry-related variables is shown by Platt & Platt (1990). Using industry-related variables would help in dealing with the data instability problem, and could translate all firms into the same scale (Platt & Platt, 1990; Barnes, 1990).

However, the conclusions about sector effects on default risk are relatively mixed. Banachewicz, Lucas & Vaart (2008) have indicated that credit risk varies considerably across industries. Also, several studies have shown the significance of industry dummy variables (Platt, 1989; Lattinen, 1993; Lennox, 1999; Westgaard & Wijst, 2001; Chava & Jarrow, 2004; Jones & Hensher, 2004; Hol, 2007; Bonfim, 2009). In the mixed logit model experiment, Jones and Hensher (2004) concluded that the new economy sector, including mainly biotechnology firms, Internet firms and high technology firms, hold higher risks of corporate default (Jones & Hensher, 2004; 2007). Platt (1989) has indicated that (i) the construction and financial services sectors are more prone to bankruptcy (ii) the industries of paper, chemicals, metals, and transportation equipment were responsible for over half of inter-industry failures and (iii) the industries of mining, leather, stone, wholesale, and commercial services had no significant effect on failures in other industries. Also, Platt (1989) has suggested that the retail industry was most vulnerable to failures in other industries. Other studies focus their predictions on one industry, such as manufacturing (Altman, 1968; Becchetti & Sierra, 2003), finance (Looney, Wansley & Lane, 1989), construction (Abidali & Harris, 1995), oil and gas (Platt, Platt & Pedersen, 1994), motor (Piesse & Wood, 1992), telecommunications (Foreman, 2003), and retail (Hu & Ansell, 2009). Couderc & Renault (2005) argue that some industries lead the default cycle while others default during the economic recovery period. However, Bhargava, Dubelaar & Scott (1998) suggest there is no difference in the prediction accuracy rate of the retail and manufacturing industries, although their sample parameter is variable. Prediction power is reduced in financial firms when investigating the prediction of corporate bankruptcy with industry effects (Chava & Jarrow, 2004). Campbell, Hilscher & Szilagyi (2008) also claim insignificance of industry effects on explaining corporate default.

Due to the limited number of relevant databases previously available, some papers limit their industry effect investigation to bankruptcy, a subsample of corporate default (e.g. Platt, 1989; Lattinen, 1993; Lennox, 1999; Westgaard & Wijst, 2001; Chava & Jarrow, 2004;

Hol, 2007). Other papers test the industry effect by dividing companies into a limited number of industry groupings (Chava & Jarrow, 2004; Jones & Hensher, 2004; 2007; Banachewicz, Lucas & Vaart, 2008). These limitations affect the identification of industry effect. There are at least two more ways to investigate industry effect. First, it is important to know if the industry effect is significant for other types of default. Second, a smaller classification of industry groups could explain industry effects more accurately. If those industries in the larger groupings (Chava & Jarrow, 2004; Jones & Hensher, 2004; 2007) hold opposite default rates, then the industry effect could vanish inside the group.

2.3.5 Other Significant Variables

[Figure 2.5, p.71 about here]

Besides the industry effect, four variables are suggested to be significant in more than two studies (Figure 2.5). (1) The calendar effect (measured by year) dummy proves a significant variable in Blum (1974), Hol (2007), Amato & Furfine (2004), and Bonfim (2009). Hol (2007) shows the calendar effect is stronger during the late 1990th than that of middle 1990th. (2) Firm age is also considered as a key predictor by Chi & Tang (2006) and Westgaard & Wijst (2001). Chi & Tang (2006) show positive significant relation between firm age and default risk. (3) Firm size measured by the number of employees is also suggested. Smaller firms suffer higher default risk (Lennox 1999; Amato & Furfine, 2004). Smaller banks suffer higher default risk (Looney, Wansley & Lane, 1989). Dichev (1998) has shown that the size effect was strong during the 1960s and 1970s but disappeared after 1980. (4) Although Merton's model itself is not sufficient to explain corporate default, Merton's distance to default has been indicated as a significant predictor by Duffie, Saita, & Wang (2007) and Bharath & Shumway (2008). However, Campbell, Hilscher & Szilagyi (2008) argue that corporate default risk cannot be adequately summarized by a measure of distance to default inspired by Merton's (1974) pioneering structural model.

Moreover, the importance of management characteristics is stressed by Abidali & Harris

(1995). Gilson (1989) has suggested that senior management change has a relationship with default action. Lussier & Halabi (2010) suggest that total years of education, difficulty of staffing, whether they have specific plan, and professional advice are key variables in measuring small business success. Park & Han (2002) recommend management variables including personnel and staff hiring policies, technology development, pricing competitive advantage, quality of management, working conditions, relationship between labour and capital, industry reputation, and past payment records. Management information in annual reports is tested for small Finnish firms by Lattinen (1993) and some are suggested to be significant such as: tidiness of the report, length of text explaining the past, changes of directors, changes in the number of employees and changes in fixed assets. A separate A-score model is constructed by Abidali & Harris (1995) to investigate the importance of management variables. Fraudulent activity and resignation of management staff are shown to be significant by Barniv, Agarwal & Leach (1997). Foreman (2003) demonstrates that firms with uncertain legal and regulatory outcomes have higher default risks. However, it is difficult to assess the public data for management variables; previous studies using survey variables (e.g. Abidali & Harris, 1995; Barniv, Agarwal & Leach, 1997; Park & Han, 2002) may suffer data mining problems and sampling issues.

2.4 Samples Used by Empirical Studies

Shortage of the sample size is one of the major concerns in corporate default prediction research. Most previous studies (e.g. Altman, 1968; Altman & Narayanan, 1996; Charalambous, Charitou, & Neophytou, 2000; Deakin, 1972; Foreman, 2003; Hwang, Cheng, & Lee, 2007; Wu & Kuo, 2004; Wilson & Sharda, 1994) limit their default sample to 100 default observations. This could deteriorate the forecasting power, and the reliability of the empirical tests.

Another problem in previous studies that has been argued is that the sample used is not representative of the population (Piesse & Wood, 1992; Lin & Piesse, 2004). Although the

defaulting firms are carefully selected, the matching control sample of healthy firms violates the random assumption for sample selection (Lennox, 1999; Lin & Piesse, 2004). Some studies selected healthy firms with the same properties as failed firms (such as size and industry), while others overestimate default frequency by selecting a sample with an equal number of defaulting and healthy firms. Under this assumption, the default rate suddenly rises to 50%, which is significantly larger than the average rate in the total population. This results in an overestimate of type II errors and underestimates type I errors. Simultaneously, classification accuracy is overestimated since the sample is not representative. Sueyoshi & Goto (2009) suggest putting more weight on the default group, which may help to solve the imbalance problem.

2.5 Model Assessment Methods

Assessing classification rates is the most commonly used method to validate default prediction, and is applied by numerous studies (e.g. Altman, 1968; Ohlson, 1980; Zavgrenc, 1985; Luoma & Laitinen, 1991; Johnsen & Melicher, 1994; Begley, Ming & Watts, 1996; Lennox, 1999; Agarwal & Taffler, 2008; Miller, 2009). Misclassification rates with different cut-off values are used to benchmark different models (Lennox, 1999). The optimal cut-off point requires information about both the misclassification costs for Type I and Type II errors, and the prior probability of default rate. Some studies address the difference between Type I error costs and Type II costs. Looney, Wansley & Lane (1989) and Whalen (1991) claim that Type I errors are more important than Type II errors in practice. Muller, Steyn-Bruwer & Hamman (2009) conclude that Type I error costs are 20 to 38 times greater than Type II error costs. However, there is a shortage of empirical evidence on the misclassifications costs for both errors together.

The Receiver Operating Characteristics (ROC) curve is also a popular technique to compare models, and is used by Begley, Ming & Watts (1996), Chava & Jarrow (2004), and Agarwal & Taffler (2008). Other criteria used as comparison techniques include: (1)

accuracy ratios (Vassalou & Xing, 2004; Miller, 2009); (2) information content (Zavgrenc, 1985; Agarwal & Taffler, 2007; Agarwal & Taffler, 2008); (3) stability (Miller, 2009); (4) durability (Miller, 2009); (5) Log-likelihood ratio (Zmijewski, 1984) and (6) Pseudo R square (Zmijewski, 1984).

2.6 Variable Combination and Selection

There are only a few papers combining variables from different categories. On the one hand, using a wide range of variables would improve prediction power. Hillegest, Keatin, Cram & Lundstedt (2004) argue that neither accounting information or market information are sufficient to explain default risk and probabilities. Furthermore, Beaver, McNicholes & Rhie (2005) suggest that market variables could offset information from financial statements in predictive ability. A further suggestion (Shumway, 2001) is to combine market driven variables with accounting ratios as a preferred explanatory set for corporate bankruptcy prediction. The superior prediction power of Campbell, Hilscher & Szilagyi (2008) compared to Chava & Jarrow (2004) comes from their usage of a wider set of explanatory variables across market-based and accounting information. The improvement of combining explanatory sets with financial ratios and market variables has also been shown using the NN model (Atiya, 2001) and the logit model (Platt, Platt & Pedersen, 1994). While some accounting ratios such as the NITA ratio lose their significance when adding market variables, overall predicting ability extracted from the dataset increases (Shumway, 2001). In the credit derivative market, Das, Hanouna & Sarin (2009) suggest that interacting accounting information with market variable gives better explanatory ability for credit default swaps. Finally, Nam, Kim, Park & Lee (2008) show that macroeconomic variables add further explanatory power to a dynamic logit model.

On the other hand, using many variables would lead to high correlation between variables, and increase computation difficulty and processing time. Main methods used in selecting variables include the t-test (Jo & Han, 1997; Abdullah, 2008), correlation matrix (Atiya,

2001; Abdullah, 2008), stepwise regression (Gilbert, Menon & Schwartz, 1990; Jo & Han, 1997), univariate method (Atiya, 2001; Abdullah, 2008), Principle Component Analysis (PCA) (Couderc & Renault, 2005; Hu & Ansell, 2009) and Factor Analysis (FA) (Barnes, 1990). Although prediction accuracy varies greatly with the methods used on variable selection (Back, Laitinen & Sere, 1996; Jo & Han, 1997), studies that compare different feature selection methods on default risk are limited. Tsai (2009) concludes that the t-test is the best selection technique in terms of overall prediction accuracy and Type I errors, while the stepwise method extracts the most features, and has higher prediction accuracy.

2.7 Conclusion

Although the topic of default prediction has been studied for decades, further investigation regarding both the methodology and default drivers is essential. Various prediction models have different advantages and disadvantages. MDA is the most commonly used model, followed by logit and NN. However, there is no solid conclusion of the best prediction model in assessing corporate default risk. Model comparison studies are limited by the number of underlying models, assessment criteria, and the size of the default databases. New methods including mixed logit and frailty hazard model could theoretically capture default clusters and unobservable variables, but further investigation using a broader covariant set is essential. Moreover, assessment of traditional models under newly developed econometric research methods is also needed. All categories of explanatory variables provide unique information with limitations; the optimal information set in default prediction remains unclear. Combining predictors from different categories, with well-selected variables using feature selection techniques, is suggested for following chapters.

Table 2.1

Lists of Previous Empirical Tests for Default Drivers Summary

This table presents the empirical studies used to summarize the default drivers. It also presents the region of the empirical tests.

No.	Author and year	Country
1	Abdullah, 2008	Malaysia
2	Abidali & Harris, 1995	UK
3	Agarwal & Taffler, 2007	UK
4	Agarwal &Taffler, 2008	UK
5	Altman E. I., 1968	US
6	Altman, Haldeman, & Narayanan, 1977	US
7	Amato & Furfine, 2004	US
8	Atiya, 2001	US
9	Back, Laitinen, & Sere, 1996	Finland
10	Banachewicz, Lucas, & Vaart, 2008	US
11	Barnes, 1990	UK
12	Barniv, Agarwal, & Leach, 1997	US
13	Barniv, Agarwal, & Leach, 2000	US
14	Beaver, Mcnichols & Rhie 2005	US
15	Beaver, 1966	US
16	Becchetti & Sierra, 2003	Italy
17	Begley, Ming, & Watts, 1996	US
18	Betts & Belhoul, 1987	UK
19	Beynon & Peel 2001	UK
20	Bhargava, Dubelaar, & Scott, 1998	US
21	Blum, 1974	US
22	Bonfim, 2009	Portugal
23	Steyn-Bruwer & Hamman 2006	South Africa
24	Bryant, 1997	US
25	Campbell, Hilscher, & Szilagyi, 2008	US
26	Charalambous, Charitou, & Neophytou, 2000	UK
27	Chava & Jarrow 2004	US
28	Chi & Tang, 2006	Asia Pacific
29	Couderc & Renault, 2005	66% US
30	Das, Duffie, Kapadia, & Saita, 2007	US
31	Deakin 1972	US
32	Duffie, Eckner, Horel, & Saita, 2009	US
33	Duffie, Saita, & Wang, 2007	US
34	El Hennawy & Morris, 1983	UK
35	Foreman, 2003	US
36	Frydman, Altman, & Kao, 1985	US
37	Gilber, Menon & Schwartz, 1990	US
38	Hensher & Jones, 2007	Australia
39	Hensher, Jones, & Greene, 2007	Australia
40	Hillegest, Keatin, Cram, & Lundstedt, 2004	US
41	Hol 2007	Norway
42	Hwang, Cheng, & Lee, 2007	US
43	Izan, 1984	Australia

44	Johnsen & Melicher, 1994	US
45	Jones & Hensher, 2004	Australia
46	Jones & Hensher, 2007	Australia
47	Laitinen 1993	Finland
48	Lennox, 1999	UK
49	Lin & Piesse, 2004	UK
50	Lo, 1986	US
51	Luoma & Laitinen, 1991	Finland
52	Lussier & Halabi, 2010	Chile.US and Croatia
53	McKee & Lensberg, 2002	US
54	Miller, 2009	US
55	Muller, Steyn-Bruwer, & Hamman, 2009	South Africa
56	Nam, Kim, Park, & Lee, 2008	Korea
57	Ohlson, 1980	US
58	Park & Han, 2002	Korea
59	Peursem & Pratt, 2002	US
60	Platt & Platt, 1990	US
61	Platt, Platt, & Pedersen, 1994	US
62	Platt, 1989	US
63	Psillaki & Margaritis, 2010	France
64	Shumway, 2001	US
65	Sueyoshi & Goto, 2009	US
66	Taffler, 1984	UK
67	Taffler, 1983	UK
68	Taffler & Tisshaw, 1977	UK
69	Vuran, 2009	Turkey
70	Wang & Campbell, 2010	China
71	Watase, 1984	Japan
72	Westgaard & Wijst, 2001	Norway
73	Wilson & Sharda, 1994	US
74	Wu & Kuo, 2004	Taiwan
75	Zavgrenc, 1985	US
76	Zmijewski, 1984	US

Table 2.2

Accounting Variables Definitions

Table 2.2 describes the definitions and the abbreviations of all accounting variables that have been shown as significant for default prediction in at least four previous empirical studies. All the ratios are grouped into five categories: profitability, liquidity, operational efficiency, capital structure and firm size.

Variables	Definition	Expected Signs
Profitability:		
NITA	Net income/ total asset	Negative
EBITA	Earnings before interest and tax/total assets	Negative
EBIATL	Earnings before interest and tax/total liabilities	Negative
INTWO	Equals to one if net income<0 of last two quarters	Positive
Operation Efficiency:		
SATA	Assets Turnover Ratio =Sales/total assets	Negative
SALEG	Sales growth=(Sales-prior period sales)/prior period sales	Negative
CFSA	Operating Income After Depreciation/sales	Negative
CFTD	Operating Income After Depreciation /(short term debt + long term debt)	Negative
CFTL	Operating Income After Depreciation/total liabilities	Negative
NOCFTA	(Revenue-Operating Expense) /total assets	Negative
RETA	Retained earnings/total assets	Negative
Liquidity:		
CACL	Current ratio=Current assets/current liability	Negative
QACL	Quick ratio =Quick assets/current liabilities	Negative
WCTA	Working capital ratio=Working capital/total assets	Negative
CASHCL	Cash/current liability	Negative
QATA	Quick assets/total assets	Negative
CATA	Current assets/total assets	Negative
CLTA	Current liability/total assets	Positive
CASHTA	Cash /total assets	Negative
Capital structure:		
TLTA	Total liability /total assets	Positive
TDTE	(Short term debt + long term debt)/total equity	Positive
OENEG	Equals to one if total liability>total assets, 0 otherwise	Positive
TDTA	(Short term debt + long term debt)/ total assets	Positive
METL	(Common share outstanding + preferred stock)*stock price/total liability	Negative
METD	(Common share outstanding + preferred stock)*stock price /(short term debt + long term debt)	Negative
Firm size:		
LOGTAGNP	Log (total assets/GNP price level index)	Negative
LOGTA	Natural logarithm of the firm's total assets	Negative

Table 2.3

Market Variables Definition

Table 2.3 presents the definitions and the abbreviations for market variables. Log is the abbreviation for Natural logarithm.

Variable	Definition	Expected Signs
PRICE	Quarterly firm share price	Negative
LOGPRICE	Log(firm share price)	Negative
RETURNFQ	Log quarterly stock return based on firm monthly return	Negative
EPS	Earnings per share=net income/common stock outstanding	Negative
QPRID	Quarterly firm price gap=quarterly highest price-lowest price	Positive
PRICETREND	$\frac{H_t - H_{(t-1)} + L_t - L_{(t-1)}}{H_t + H_{(t-1)} + L_t + L_{(t-1)}}$ H_t (L_t) highest (lowest) price in quarter t	Positive
LOGMC	Log(Firm market capitalize)	Negative
MBR	$\text{market to book ratio} = \frac{\text{Stock price}}{\text{stockhoders book equity/number of shares outstanding}}$	Negative
SDF	Standard deviation of firm monthly stock return over quarter	Positive
FMVTMV	Log(firm market value/total S&P 500 value)	Negative
EXCESS	Log(1+RETURNFQ)-log(1+SPRETURN)	Negative

Table 2.4

Macroeconomic Variables Definition

Table 2.4 describes the definitions of the original macroeconomic variables and the abbreviations used for the Natural logarithm and Growth rates of the original variables.

Original Variables	Definition	Natural logarithm variables	Growth rate variables
Stock market information			
SPRETURN	Log (quarterly S&P 500 index return)		
SDM	S&P 500 standard deviation of monthly stock return over quarter		
Economic cycle			
CPIAUCSL	Consumer Price Index for All Urban Consumers: All Items	CPIAUCSLL	CPIAUCSL
PPIACO	Producer Price Index: All Commodities	PPIACOL	PPIACOG
GNPC96	Real Gross National Product	GNPC96L	GNPC96G
GNP	Gross National Product	GNPL	GNPG
GDPC96	Real Gross Domestic Product, 3 Decimal	GDPC96L	GDPC96G
GDP	Gross Domestic Product, 1 Decimal	GDPL	GDPG
UNRATE	Civilian Unemployment Rate	UNRATEL	UNRATEG
UNEMPLOY	Unemployed	UNEMPLOYL	UNEMPLOYG
IMPGSC1	Real Imports of Goods & Services, 1 Decimal	IMPGSC1L	IMPGSC1G
EXPGSC1	Real Exports of Goods & Services, 1 Decimal	EXPGSC1L	EXPGSC1G
NICUR	National Income	NICURL	NICURG
INDPRO	Industrial Production Index	INDPROL	INDPROG
Bond yield and interest rate			
AAA	Moody's Seasoned Aaa Corporate Bond Yield	AAAL	AAAG
BAA	Moody's Seasoned Baa Corporate Bond Yield	BAAL	BAAG
DTB3	3-Month Treasury Bill: Secondary Market Rate	DTB3L	DTB3G
DTB6	6-Month Treasury Bill: Secondary Market Rate	DTB6L	DTB6G
DGS1	1-Year Treasury Constant Maturity Rate	DGS1L	DGS1G
DGS10	10-Year Treasury Constant Maturity Rate	DGS10L	DGS10G
CD1M	1-Month Certificate of Deposit: Secondary Market Rate	CD1ML	CD1MG
DFF	Effective Federal Funds Rate	DFFL	DFFG
DPRIME	Bank Prime Loan Rate	DPRIMEL	DPRIMEG
BAA-AAA	Difference between the Aaa and Baa Bond Yield		
DGS10-DGS1	Difference between 10year and 1 year treasury		
DGS10-BAA	Difference between DGS10 and BAA		

Bank lending and investment			
BUSLOANS	Commercial and Industrial Loans at All Commercial Banks	BUSLOANSL	BUSLOANSG
CONSUMER	Consumer (Individual) Loans at All Commercial Banks	CONSUMERL	CONSUMERG
REALLN	Real Estate Loans at All Commercial Banks	REALLNL	REALLNG
LOANS	Total Loans and Leases at Commercial Banks	LOANSL	LOANSG
INVEST	Total Investments at All Commercial Banks	INVESTL	INVESTG
GFDEBTN	Federal Government Debt: Total Public Debt	GFDEBTNL	GFDEBTNG
NOM1M2	Non-M1 Components of M2	NOM1M2L	NOM1M2G
M2	M2 Money Stock		
M1	M1 Money Stock		

Figure 2.1

Summaries of Default Prediction Models

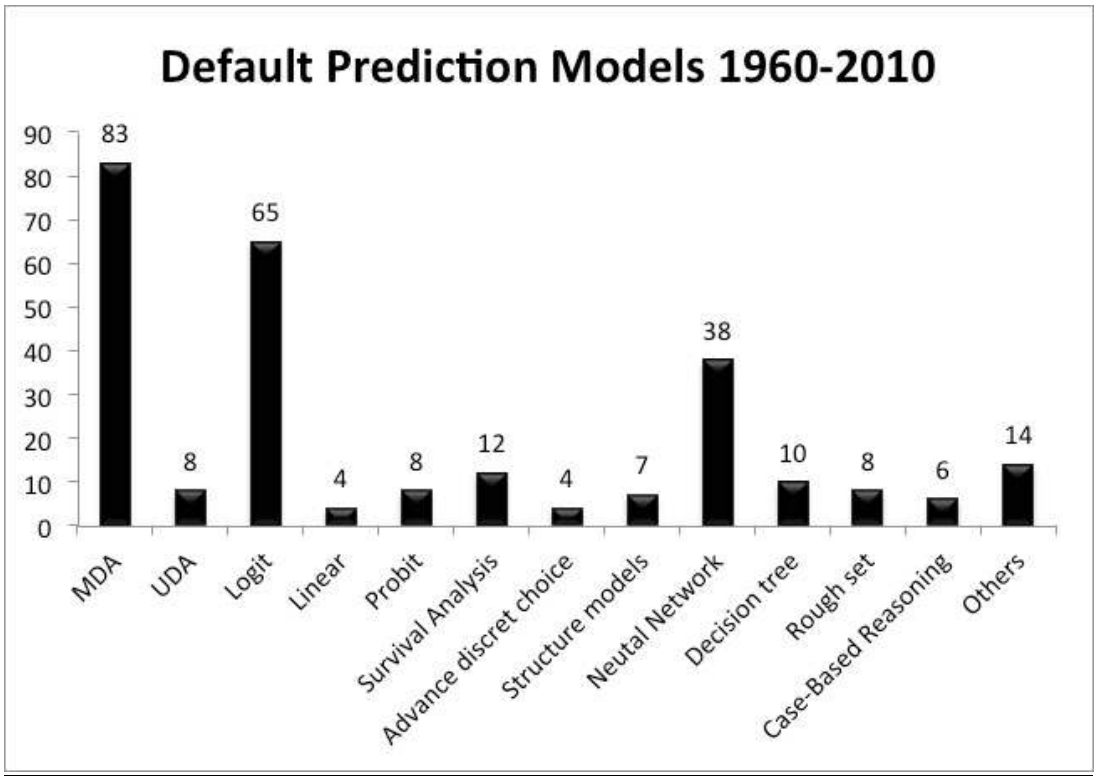


Figure 2.1 summarises the historical frequency of default prediction models from 1960 to 2010. Other methods include: the Markov model, Genetic Algorithms, Balance Sheet Decomposition, Multi-dimensional scaling Approach, Sequential Minimum Optimization, and the Linear Gambler's Ruin score.

Figure 2.2

Summaries of Significant Accounting Variables

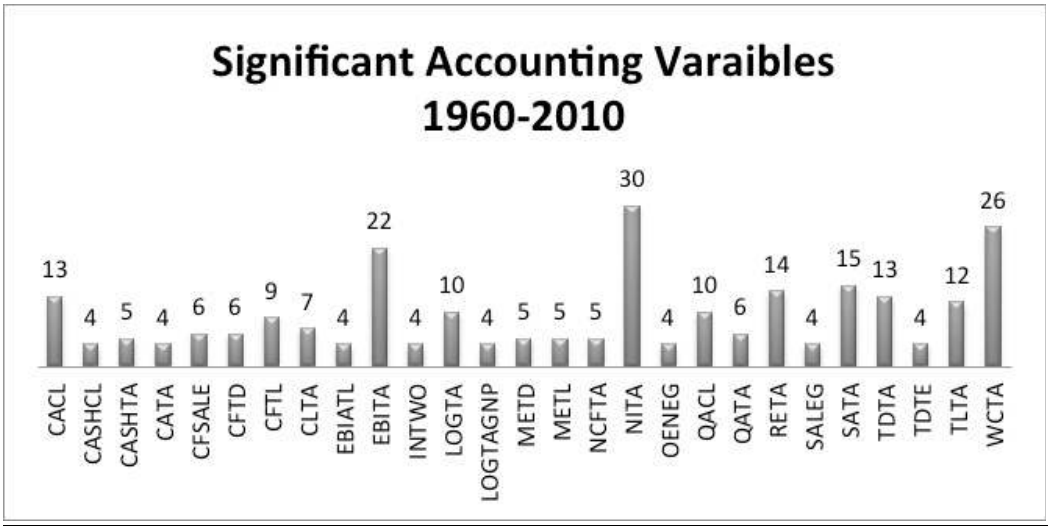


Figure 2.2 summarises the most significant accounting variables in previous empirical studies from 1960 to 2010. The abbreviations of variables are presented in Table 2.2.

Figure 2.3

Summaries of Significant Market Variables from Previous Studies

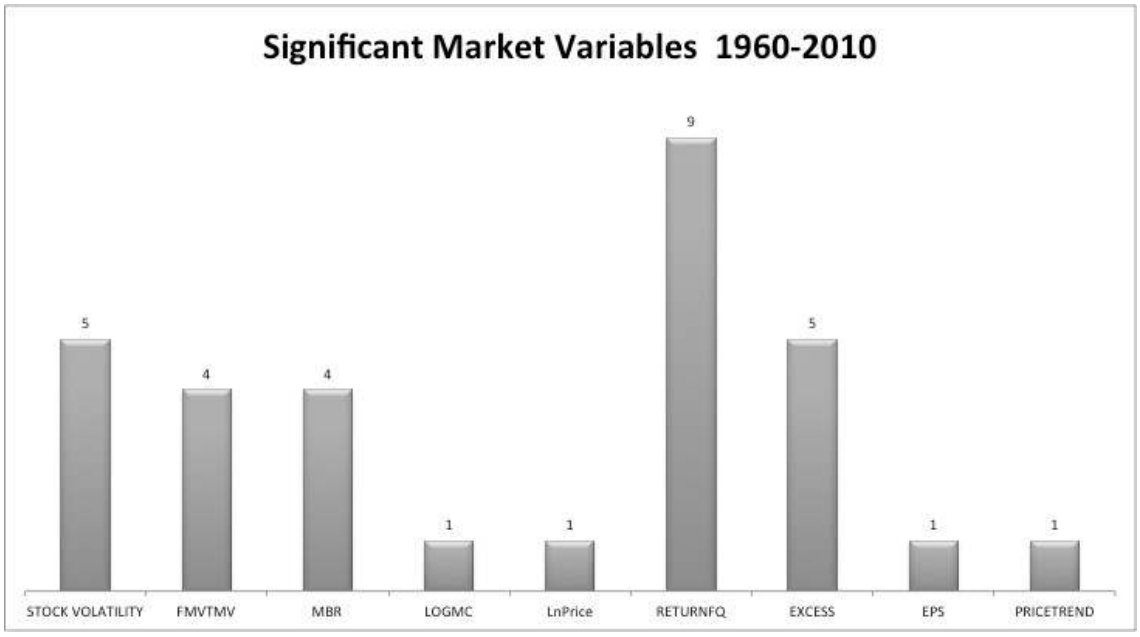


Figure 2.3 summarises the significant market variables in empirical studies from 1960 to 2010. The Abbreviations of variables are presented in Table 2.3.

Figure 2.4

Summaries of Significant Macro Variables from Previous Studies

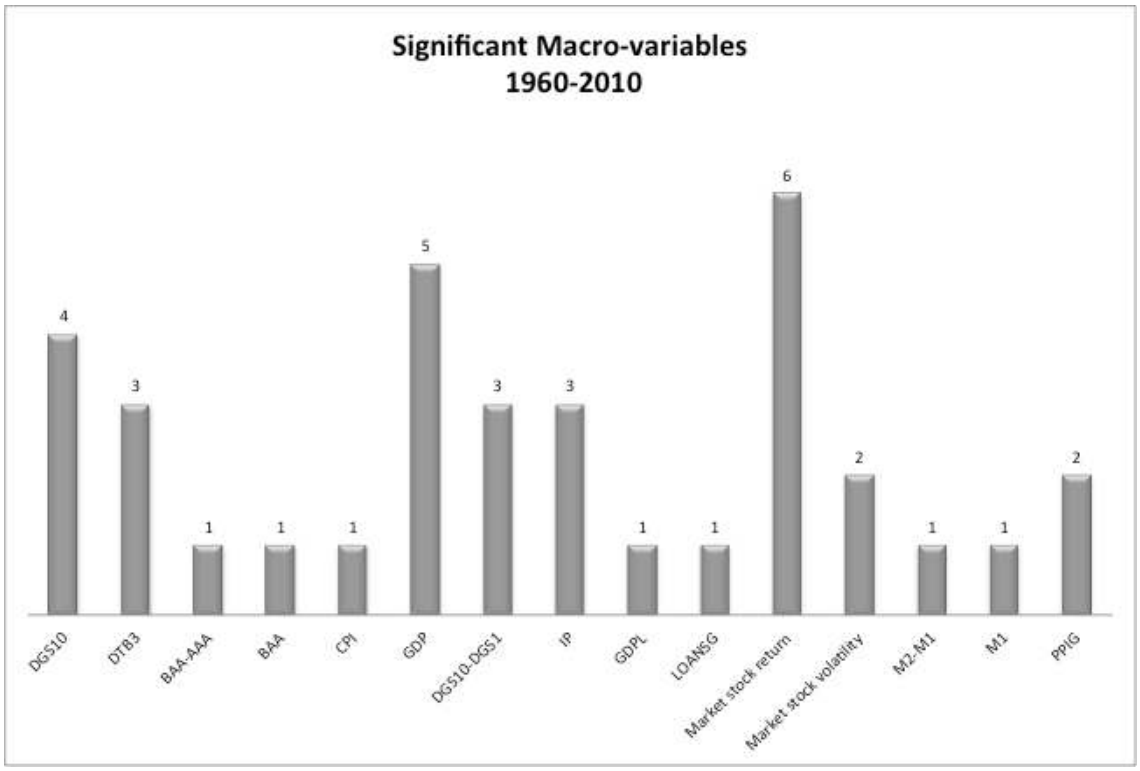


Figure 2.4 summarises the significant variables in empirical studies from 1960 to 2010. The abbreviations of variables are presented in Table 2.4.

Figure 2.5

Industry Effects and Other Significant Variables

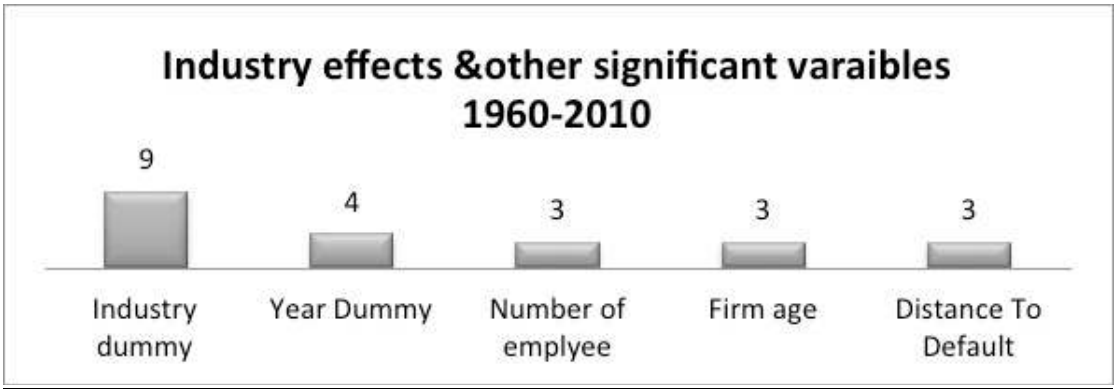


Figure 2.5 summarises the significant variables in empirical studies from 1960 to 2010.

Chapter 3: Variable Selection for Default Prediction

3.1 Introduction

What information contributes to predicting default events? Accounting-based information has the benefit of broad applications in both private and public firms, and has been widely used to predict corporate default since the 1960s (Beaver, 1966; Altman, 1968). With the development of the market-based model, market variables have been recommended as default drivers rooted in market efficiency theory. Recently, the potential advantage of combining accounting ratios and market information as explanatory default drivers has been claimed by several studies (Chava & Jarrow, 2004; Beaver, McNicholes & Rhie, 2005; Campbell, Hilscher & Szilagyi, 2008). Since the latest 2008 financial crisis, the correlation between credit and economic cycles has aroused new interests. Bonfim (2009) investigates the macro effects based on Portugal's bank default, while Koopman, Kraussl, Lucas & Monteiro (2009) confirm the stronger explanatory power of macroeconomic variables under higher default risk pressure. Furthermore, the importance of industry effects for corporate default has been addressed extensively (Chava & Jarrow, 2004; Jones & Hensher, 2004; Carling, Jacobson, Linde & Roszbach, 2007; Bonfim, 2009; Koopman, Kraussl, Lucas & Monteiro, 2009). If all this information counts towards default risk, a further question arises: what is the optimal information set to be used for corporate default prediction?

Using the most comprehensive and up-to-date dataset, the first objective of this chapter is to identify the role of variables from different categories. Although the importance of the explanatory variables from alternative categories has been addressed in previous studies,

a few questions remain unanswered. First, it remains open to debate whether market variables could replace the role of accounting variables in default prediction. While the market information-based model is favoured by some studies (e.g. Chava & Jarrow, 2004; Hillegest, Keatin, Cram & Lundstedt, 2004), Agarwal & Taffler (2008) argue that the accounting information-based model has practical preference for creditors in terms of differential error costs for different classification errors. This chapter is a discussion of the preference for using accounting information and market information for default prediction, and also a reference for testing the market efficiency theory. The second unanswered question is whether macroeconomic indicators can contribute to default prediction on top of firm-level information. Although studies combining firm variables with macroeconomic indicators have recently been tested in a bank default study (Bonfim, 2009), empirical study in corporate default prediction in general remain limited. This chapter contributes to the literature on the role of macroeconomic information in default prediction and, further, to the debate on the relationship between business cycle and credit risk cycle conditioning on firm-level information. The third unanswered question is the explanatory powers of industry effect on default prediction, which remains undetermined in previous studies. While Chava & Jarrow (2004) and Hensher & Jones (2007) detect the significance of the industry effects for default prediction, Campbell, Hilscher & Szilagyi (2008) claim them to be insignificant. This chapter represents a further effort to test the industry influence on default prediction.

A further objective of this chapter is to investigate the optimal variable set for corporate default prediction. A comprehensive set of information is used to explain default behaviour accurately, thus improving overall prediction ability. However, repeated information contains a high correlation between variables, which diminishes the accuracy of the prediction. Also, superfluous and redundant variables increase the computational time-cost, which would not be preferred in practice. It is well known that parsimonious models generate much more accurate out-of-sample forecasts because of their efficiency.

Furthermore, identifying the optimal variable set helps to isolate the roles of individual variables.

[Figure 3.1, p.123 about here]

To balance the number of variables and select the optimal variable set, the investigation was divided into two stages. Figure 3.1 presents an overview of the procedure and associated methodologies. For the first stage, a wide range of information was collected from different categories. Instead of borrowing the variable set from one or two studies, as in most previous studies, this thesis selected key variables originating from more than 80 empirical studies. Particular attention was paid to macroeconomic information. Several new macroeconomic variables (such as certificates of deposit and investments at all commercial banks) were collected and proved to be significant in default predictions. For the second stage, variables were selected using both t-test and stepwise feature selection techniques, based on the dynamic logit model to solve the overlapping problem of the original variable set. This stage drops irrelevant information and keeps the correlation between variables at a tolerably low level, whilst maintaining comparatively higher prediction accuracy according to indicators covering the likelihood ratios, AUROC and Pseudo R^2.

Using the optimal variable set, obtained by the analyses of these two stages, this chapter further investigates the roles of variables from alternative categories. To begin with, the chapter explores prediction abilities with various regressions using either sole or combined variables from different categories. Then, the chapter investigates deeply the characteristics of variables from different categories, at various time horizons and alternative sub-periods. Industry effects are tested for both private and public firms with alternative horizons. Finally, the chapter verifies the prediction ability with both in-sample and out-of-sample classification and related AUROC.

The contribution of this chapter, therefore, is fourfold. The first contribution is to improve

the default identity ability of the dynamic logit model by applying comprehensive multidimensional information. In tandem, the contributions of explanatory variables from different categories are identified. The original information set contains 125 variables from accounting ratios, financial market indicators, macroeconomic factors and other additional factors. The appropriateness of the dynamic logit model is tested, relying on information from a single sector. The prediction accuracies from each regression show that both accounting ratios and market variables capture significant information, while additional information such as default history and company age also significantly improves the explanatory ability of the predictive model. Several macroeconomic variables show significance in predicting default risk, although the single macroeconomic variable set has limited ability to capture the probability of default. In summary, combining all information greatly increases the overall corporate default prediction power of the model.

Secondly, extensive empirical analysis is conducted to investigate the optimal default drivers using t-test and stepwise selection methods. The significant tiers figure is drawn, and the optimal variable set for default prediction demonstrated, which combines a small number of default drivers with relatively high prediction ability. The optimal variable set contains 26 variables from all four categories, keeping 20% of the original variables, with 99.25% prediction ability for the original 125 variable information set. Default history (DH) is shown to rank as the most important variable, followed by Return on Assets and financial ratios on market capacity. The selected optimal variable set shows considerably better performance in comparison with previous studies.

Thirdly, the prediction ability of each variable is investigated at a longer horizon and several sub-periods. As expected, prediction ability deteriorates as the horizon increases. Two variables, Earnings per Share and Stock Price, become more important in a longer-horizon prediction, while most accounting variables become less informative as the prediction stretches further into the future. The macroeconomic variables are more

significant in long period predictions than sub-period tests.

Finally, this chapter verifies that default risk varies with industry. The amount of information available from industries improves the ability to predict default risk, especially for private firms, which have a shortage of market information. The industry effect is also shown to play a significant explanatory role in longer-horizon predictions. The most risky industries are shown to be agriculture, forestry, fishing and retail trade, while public administration and the finance, insurance and real estate sectors are shown to have the lowest default risk.

The layout of this chapter is organized as follows. Section 2 outlines the methods of both the dynamic logit model and feature selection techniques. Section 3 provides details of all the data and variables, divided by categories. Section 4 compares the results of the regressions. Section 5 concludes the research method and findings.

3.2 Methodology

This section briefly describes the selection techniques and the default prediction method applied in this chapter. It begins by describing the preliminary selection procedure with both the t-test and stepwise techniques. Attention is then paid to the ''dynamic'' logit model or ''hazard'' logit model (Shumway, 2001; Chava & Jarrow, 2004; Campbell, Hilscher & Szilagyi, 2008). This model is selected in this chapter for variable selection investigation because it has the advantages of utilising data efficiently, incorporating time-varying elements naturally, computing coefficients economically, and most importantly, allowing the results to be traced and interpreted directly.

3.2.1 Variable Selection Method

Feature selection is the process of selecting from the original variables those features that are significant and relevant for a regression. From the 1960s, massive selections of variables (more than 180 accounting ratios, more than 10 market variables, and more

than 30 macroeconomic variables) from different categories have been available and used to estimate corporate default. However, variables in the same sector could be highly correlated. In addition, using a huge number of variables would increase the costs of computing time and thus reduce the efficiency of the prediction. The feature selection procedure addresses a shortage of theoretical foundation in selecting the optimal independent variable set for default prediction. Also, the use of a stepwise procedure helps to reduce multi-collinearity problems, which may decrease the prediction accuracy. Tsai (2009) suggested the t-test and stepwise methods are the preferred feature selection techniques because they can extract information effectively while keeping most of the prediction ability from the original data set. Other methods such as principal component analysis and factor analysis are also used as variable filters in previous research. However, instead of selecting variables from the original sets, these methods transfer the original variables into a new variable set. The original meaning of those variables is lost in the transformation. In other words, it is not possible to assign an interpretation to the transferred variables set.

3.2.2 The T-Test

The t-test is used to determine whether there is a significant difference between the means of both default and healthy firms. The technique helps to decide whether the mean values of the two groups are from the same population, and whether they appear different because of a chance error or a significant difference. The means are more likely to be significantly different between groups if the sample size is larger.

The null hypothesis of the t-test is that the mean of the default firms $\overline{X}_1$ and the healthy firms $\overline{X}_2$ are equal. The test for $\overline{X}_1 = \overline{X}_2$ when the variances of each group is unequal is given by:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{(s_{x_1}^2 / n_{x_1} + s_{x_2}^2 / n_{x_2})^{1/2}} \quad (1)$$

Where n_{x_1} n_{x_2} are the sample size, $s_{x_1}^2$ $s_{x_2}^2$ are the sample variance.

3.2.3 Stepwise Regression

There are four main stepwise approaches: forward selection, backward elimination, forward-backward selection and backward-forward selection. Forward selection is a bottom-up procedure. The procedure starts with no variables; new features are added to the variable set if they are statistically significant. On the other hand, backward elimination is a top-down procedure. The regression starts with the complete variable set. They are tested one by one for statistical significance and variables are deleted based on their contribution to the likelihood ratio statistics. Therefore, variables that do not contribute significantly to the statistics are not included in the final variable set. The forward-backward and the backward-forward selections combine both bottom-up selections and top-down eliminations. They test at each stage for potential default indicators to be included or excluded.

3.2.4 Dynamic Logit Model

This section briefly describes the dynamic logit model. In a two-state corporate default situation, this thesis models the default event of firm *i* in the period *t* as a random variable y_{it} , which is called the explained variable or dependent variable, taking on the values zero or one, which indicate whether or not a company has defaulted, as showed below:

$$y_{it} = \begin{cases} 1 & \text{if firm } i \text{ default at time } t \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The main purpose here is to calculate the probability of a default event occurring for a certain company, which is formulated as:

$$p(X) = P(y_{it} = 1|X) = P(y_{it} = 1|x_1, x_2 \dots x_k) \tag{3}$$

$X \equiv (x_1, x_{2,}, \dots, x_k)$ is defined as the vector of explanatory variables, or independent variables, which stands for the set of factors influencing the default behaviour of a company.

The binary logit model is the easiest and most widely used to predict the probability of corporate default. Its popularity comes from the fact that the formula for the default probability is interpretive and has a closed form solution (Train, 2009), which makes the calculation time-efficient. To set up the logit model, this chapter explores a mapping logistic function $G(x)$, which is used to link the explanatory variable X with the response probability y when default has happened.

$$G(z) = \frac{1}{1+\exp(-z)} = \frac{\exp(z)}{1+\exp(z)} \qquad -\infty < z < +\infty \tag{4}$$

This link function $G(z)$ is strictly between 0 and 1, for all real numbers. This property makes the logistic model the first choice when a probability is estimated, because a probability would never be above 1 or under 0. Another reason for its popularity is the elongated S-shape of the logistic function $G(z)$, as shown in Figure 3.2. The S-shape of the logistic function applies to corporate default prediction if the variable z is viewed as representing an index that combines contributions of several risky factors including firm-level information and macroeconomic information, and $G(z)$ as representing the total risk of corporate default for a given value of z. Then, the S-shape of $G(z)$ indicates that the effect of z on an individual firm's risk is minimal for a low value of z until some threshold is reached. The risk then rises rapidly over a certain range of intermediate z values, and then remains extremely high around 1, once z gets large enough. This threshold idea is

thought, by credit risk analysts to apply to a variety of company conditions, since corporate default is an extreme high-risk event.

[Figure 3.2 p. 124 about here]

The representative utility is usually specified to be linear $z = \alpha_0 + \beta X$, where X, is a $1\times K$ vector as the collection of observed independent variables x_1, x_2, and so on up to x_k on a group of subjects, such as accounting, market, or macroeconomic. The purpose of using this information is to describe the probability that the firm will default during a defined study period, say t to $t+1$. The $K \times 1$ vector β_i is the coefficients for explanatory variables, which would be estimated by regression below, given this information of explanatory variables at the time t and the firm status at time $t+1$:

$$p(X) = P(y_{i,t+1} = 1 \left| x_{i,t}, \varepsilon_i \right.) = \frac{\exp(\alpha_0 + \beta_i x_{i,t})}{1 + \exp(\alpha_0 + \beta_i x_{i,t})} \tag{5}$$

ε_i is the error part. After a logit transformation, denoted as logit $p(X)$, where $p(X)$ denotes the logistic model as previously defined, it could simplify to the linear sum found in the denominator of the formula for $p(\mathbf{X})$:

$$\text{logit}\, p(\mathbf{x}) = \ln_e[\frac{p(\mathbf{x})}{1-p(\mathbf{x})}] = \alpha + \sum \beta_i x_i \tag{6}$$

This is the logit form of the logistic model, which is more convenient to use. The logit form also gives an expression for the log-odds of processing the default for a certain company with a specific explanatory set of X's. α performs as baseline odds or the background effect. The coefficient β_i represents the change in the log-odds that would result from one unit change in the variable x_i, when all other x_i are fixed (Kleinbaum, Klein & Pryor, 2002).

Since the logistic model is a nonlinear model, the maximum likelihood estimation is the preferred method to estimate the coefficient compared with the least squares estimation.

Moreover, maximum likelihood estimation requires no restrictions of any kind on the characteristics of the independent variables, which means the independent variables can be nominal, ordinal, or interval (Kleinbaum, Klein & Pryor, 2002). The density of y_i given X can be expressed as:

$$f(y \mid X;\beta) = [G(X\beta)]^{y}[1-G(X\beta)]^{1-y}, y = 0,1 \tag{7}$$

The log-likelihood for observation i is a function of the K×1 coefficient and the data set ($x_{i,t}, y_{i,t+1}$):

$$l_i(\beta) = y_i \log[G(\mathbf{x}_i\beta)] + (1-y_i)\log[1-G(\mathbf{x}_i\beta)] \tag{8}$$

Using this model to investigate the factors that influence the default of a firm, the sign of the effect of each explanatory variable could be decided by the sign of the estimated coefficient β_i (Wooldridge, 2002).

3.3 Data Description

This section summarises all the default indicators and explanatory variables used in the rest of the thesis. The empirical test is based on U.S. firms from January 1st, 1970 to December 31st, 2009. Chava & Jarrow (2004) have suggested that the smaller the observation intervals, the stronger the predication power. The prediction in this thesis is therefore quarterly based because this is the minimum available interval for standardised accounting variables.

3.3.1 The Corporate Default Database

The financial distress indicator used to describe the status of a firm is collected from four sources: (1) the Moody's corporate default risk service database, which contains over 20,000 corporate and sovereign entities from 1970 to present and is updated monthly; (2)

BankruptcyData.com, which is a searchable online database and provided U.S. business bankruptcy filings from federal bankruptcy districts (this data source is also used by Jorion & Zhang, 2009); (3) COMPUSTAT North America quarterly and yearly databases; (4) the Fitch-rated Defaults.

The original default sample contains is 6,296 default events during the period from 1970 to 2009. However, 3184 default events do not exist in either COMPUSTAT or CRSP database. Presumably, they are most private firms. As a result, only 3112 default events are kept at this stage. At the second stage, 989 default firms have been drop because either relevant accounting or market information is missing for the whole observation period in the COMPUSTAT or CRSP database. It is worth noticing that to avoid the accounting information release lag, this thesis replaces the missing variable with previous period observation up to two years for default firms. Finally, those observations with sufficient accounting and market information in both COMPUSTAT and CRSP databases are kept. The final default sample consists of 2,123 default events for one-quarter ahead of prediction. The details of the process for selecting the final default events sample are listed in Figure 3.3 and Table 3.1.

[Table 3.1, p.103 about here]

[Figure 3.3, p.125 about here]

During the peak default periods (early 1990th, early 2000th and late 2000th), large amount private firms suffer high default risk, which results in the gaps between the original sample and the final sample. The difference between the original default sample and the final default sample influent the accuracy of the default prediction, however, compared with existing studies (Altman E. I., 1968; Ohlson, 1980; Chava & Jarrow, ,2004; Campbell, Hilscher, & Szilagyi, 2008), this is the most comprehensive and updated corporate default sample from the U.S. It captures the most recent default information particularly from the recent financial crisis since 2007. Also, as the default data are collected from four different

sources, it avoids bias arising from the use of a single data source. The yearly default number is presented in Figure 3.4. In these default firms, 1,920 firms have sufficient information for testing half-year ahead prediction, 1,882 for one-year ahead prediction, and 1,686 for two years ahead prediction.

[Figure 3.4, p.126 about here]

It is noteworthy that 1,979 firms defaulted only once, 128 firms defaulted twice, 15 firms defaulted three times, and one firm defaulted a fourth time. The average yearly default rate for all firms over the sample period is 0.32%, which is derived from 0.27% bankruptcy and 0.05% default for other reasons, such as failure to pay interest rates, distressed exchange rates.

Figure 3.4 provides the yearly default number of the final sample used in the research from 1970 to 2009. The default number is relatively small before 1983, and then increases to almost 100 in the early 1990s. This matches with the early 1990s recession. As expected, the highest quarterly default number appears in 2001, as 204 firms. Many firms default during the early 2000s because of the technology bubble. The default number falls down to 30 in 2006, and hits another peak in 2009 because of the current financial crisis.

Non-default firms consist of all firms with available explanatory information in both the COMPUSTAT and CRSP databases. The final combined sample of default and non-default firms consists of 639,573 firm-quarter observations representing 17,663 individual firms. This larger sample should better capture the actual information of publicly traded firms from 1970 to 2009, thereby improving the accuracy of the coefficient estimates and increasing the power of my model compared with previous studies.

[Figure 3.5 p.127 about here]

Figure 3.5 presents default intensity, which is the default number over the number of active firms every quarter. The average quarterly default rate over 40 years is 0.33%.

Compared with Figure 3.4, the default rate has the same peaks, but this default intensity is not stable before 1980. The default rate is quite high in 1971 since the total active firms before 1976 is limited in the COMPUSTAT database. Even the number of defaults is small. The number of firms quarterly in the sample ranges from a minimum of 25 in the last quarter of 1970 to a maximum of 6,339 in the last quarter of 1997.

It is also important to point out that the highest default intensity is in the fourth quarter of 2001, but the rate is almost the same in 2001 and the end of 2008. Nevertheless, default is still an extremely rare event among public firms. The data indicates that the quarterly default rate generally reflects the overall health of the economy, with relatively high rates during the recessions of the early 1990s, early 2000s and late 2000s, and low rates during the expansion years of the mid-1990s and mid-2000s.

[Figure 3.6 p.128 about here]

The default rate also varies considerably by industry. The partitioning of corporate defaults by industry classification is reported in Figure 3.5. There are ten main industry sectors distinguished by the SIC code in the COMPUSTAT database. The name of each industry is presented in Panel C of Figure 3.6.

The default numbers of each sector are presented in Panel A of Figure 3.6. The four sectors with industry code as 4 (Manufacturing), 5 (Transportation, Communications and Utilities), 7 (Retail Trade) and 9 (Services) have relatively high default rates, and represent over 80% of the total default sample.

Panel B in Figure 3.6 presents the default rate within certain industries measured as the default number over the number of active firms in each sector. The average default rate for all industries is 0.33% while the default rate for individual sectors differs. The manufacturing sector (industry code 4) experienced the highest default number (874) with the lowest default rate (0.26%). The sectors with the codes 1 (Agriculture, Forestry, And Fishing), 3 (Construction), and 10 (Public Administration), on the other hand, have

higher sector default rates while taking lower sector default numbers. Sector 7 (retail trade) has the highest default rate (0.59%) and the third highest default number (285).

3.3.2 Explanatory Variables

In order to evaluate the relation between default rates and potential influential independent variables, a large set of variables were gathered covering not only accounting variables, but market-based and macro variables as well. The definition of each variable has been presented in Chapter 2. This section describes the data sources, and the summarised statistics for explanatory variables.

The accounting data is collected from the COMPUSTAT North America quarterly and yearly, and the market information is collected from CRSP. The definitions and details of each variable are presented in Table 2.2. Both accounting and market data are lagged by a quarter so that they are available to the market at the time of estimation. In cases where accounting or market data are missing, the previous available observations are substituted.

To eliminate the outlier effect, all variables from COMPUSTAT and CRSP higher than the 99 percentile of each variable are set to that value, and all values lower than 1 percentile of each variable are replaced using the same method. This method is used in previous studies (e.g. Shumway, 2001), while Campbell, Hilscher & Szilagyi (2008) take the 95th percentile as the outlier boundary.

3.3.2.1 Accounting Variables

Accounting variables have been used to estimate corporate bankruptcy since 1960. More than 180 accounting variables have been shown to be significant in existing papers. Those variables that have been shown as empirically significant in more than four papers have been selected as the initial accounting variables set. The definition of each variable is listed in Table 2.2. The summary statistics for each accounting variable are presented in Table 3.2. The ratios and variables are summarized in five categories: profitability,

liquidity, operation efficiency, capital structure and firm size.

[Table 3.2, p.104 about here]

A comparison of the default firms with non-default firms reveals that the default firms have apparent difference from the rest of the sample. Firstly, the quarter before the default event, the default firms are typically less profitable (the mean of NITA, EBITA, EBIATL and RETA of default firms are all smaller than non-default firms; the mean of INTWO of the default firms indicates that most default firms generate a net loss before default). Another feature from profitability ratios is that although the non-default group has more observations, the standard deviations of NITA, EBITA and EBIATL are lower than that of the default group. This reveals the instability of income from default firms. Secondly, the default firms operate inefficiently: the sales decline 2% in the quarter before the defaults; the operating income ratios (CFSA, CFTD, CFTL, and NOCFTA) are all smaller than that of the healthy firms. Thirdly, the default group also suffers liquidity problems as the Cash ratio, Current ratio and Working Capital ratio of default groups are largely below the sample and non-default group average means. Fourthly, the total liability ratio of default firms is almost doubles that of the healthy firms (0.945 for default firms and 0.575 for healthy firms). Default firms are more likely to have higher debt assets, as the TDTA is higher and the TDTE lower in the default group. Finally, smaller firms are found to be more likely to default since failing firms have a slightly smaller LOGTA and LOGTAGNP value (a measurement of firm size).

3.3.2.2 Firm-level Financial Market Variables

Compared with accounting ratios, research interest in using market variables as default prediction factors is more recent. If the market efficiency assumption holds, then all the information would be included in the stock price, thus market indicators could replace all the other information in default prediction. The market variables included in the initial empirical test would be those shown to be significant in previous studies. Table 2.3

presents the definition of all the market variables used in this chapter.

Understanding the descriptive statistics of all these market variables reported in Table 3.2, the default groups have relatively lower stock price, lower stock return, lower EPS, smaller market capitalisation and higher market risk. On aggregate the stock price of healthy firms ($14.58) is almost seven times that of default firms ($2.74). The LOGPRICE for default firm is -0.14, while it is 1.93 for healthy firms. The default firms lost 15.9% on average one quarter before default, almost seven times that of healthy firms as 2%. Not surprisingly, the SDF as 26.95 is twice that of the healthy firms as 13.55, showing that the stock price of default firms is more volatile and suffers higher market risk. EPS of default firms is -0.28, while it is 0.18 for healthy firms. Comparing the market capitalization size, default firms are associated with smaller LOGMC as 2.25, while that of the healthy firm is 4.41.

3.3.2.3 Other Firm-level Variables

This chapter measures default history as a binomial variable (DH). It sets the value to one if a firm has defaulted before, and zero otherwise. The firm age (AGEQ) is defined as the number of quarters that a firm appears in the database. Although this variable suffers the left censoring problem, there are only 65 firms set up on or before 1970 in our sample. The description statistic of these variables is presented in Table 3.2. The mean of DH in the default group is more than 20 times higher that of the non-default firms.

3.3.2.4 Macro-level Variables

The macroeconomic variables were extracted from the Federal Reserve Bank of St. Louis (FRED) Website (following Couderc & Renault, 2005), and CRSP. Table 2.4 presents all the macroeconomic level variables selected in the initial test. Obviously, some pairs of the above variables such as (CPI & PPI, GNP & GDP, BUSLOAN & REALLN) are highly correlated, which would lessen statistical significance on the full set of variables. However,

the main purpose at this stage is to identify the contributions of variables from different categories towards default probabilities. A complete variable set allows the possibility of accurate prediction. This chapter uses a feature selection technique to extract the most core factors of default at the next stage. It is worth mentioning that the natural logarithm and the growth rate of most of the original variables were also included (see Table 2.4 and Table 3.3). This expands the range of original macroeconomic indicators and provides the best chance to identify the most efficient explanatory variables. Table 3.3 presents the summary statistics of all the macroeconomic variables divided in three groups: the original variables, the natural logarithm form of the matching original variables and the growth rate of the matching original variables.

[Table 3.3, p.105 about here]

Within the Stock market information category in Table 3.3, the default firms associate with lower market return and lower standard deviation. The average SPRETURN for the non-default firms is 0.02, while it is 0.014 with the high default risk firms. The average SDM for healthy firm is 0.039, while it is 0.037 with the default firms.

In the economic cycle category in Table 3.3, although the default groups associate with higher CPIAUCSL, PPIACO, IMPGSC1, EXPGSC1, GNPC96, GNP, GDPC96, GDP, NICUR, INDPRO and the matching natural logarithm, the growth rates of these variables are lower than that of the healthy firms. The UNRATE, LUNTAE and GUNRATE of the default firms are all lower than those of the non-default firms.

The summary statistics in Table 3.3 with the bond yield and interest rate category (AAA, BAA, DTB3, DTB6, DGS1, DGS10, CD1M, DFF, DPRIME, BAA_AAA, DGS10_BAA, DGS10_DGS1), describe the default firms exhibiting lower bond yields and interest rates, lower natural logarithms of the bond yield and interest rate, and also lower growth rates of the bond yield and interest rate, compared with healthy firms.

In the bank lending and investment category, Table 3.3 shows that during the period of

default, banks hold relatively higher amounts of loans (BUSLOANS, CONSUMER, REALLN, LOANS) and investment (INVEST) and that the GFDEBTN and NOM1M2 are also higher than the non-default period. However, it is difficult to attract new loans during the default period, as the growth rates of BUSLOANS, CONSUMER, REALLN and LOANS are relatively low. The growth rate of INVEST, on the other hand, increases in the default period.

3.4 Results

3.4.1 Variable Selection

This section presents the results of the selection of features as default indicators. It presents the feature selection process and results following both the t -test and stepwise methods. The objective is to use the lowest amount of variables to capture the most information from different categories. In the meantime, the extracted sample should keep most of the prediction accuracy and efficiency.

3.4.1.1 Mean Difference T-Test

A 0.95 confidence interval was used in the t-test to assess if the means of the default and the healthy groups are equal. The value of the t-test for firm-level variables is shown in Table 3.2 and the t values of the mean test results for macro-level variables are presented in Table 3.3**.** In sum, firm-level variables have higher t values compared with macroeconomic variables. At firm level, all variables except SATA show significant difference of the means between the default firms and non-default firms at the 99% level. The stock price (PRICE) distinguishes the two groups most significantly.

At market level, the probability that BAA_AAA and DGS10G show difference of the means between the default and non-default firms is less than 90%. The variable GFDEBTNG has a 1% chance of having the same mean values between groups. If the probability of difference were set at 0.1%, five more variables (AGEQ, SPRETURN, BAAG, AAAG and

NOM1M2G) would be excluded, in that there are no differences in these variables between the default and non-default group. The variable UNEMPLOYL also has a 0.01% chance of having the same mean between groups.

3.4.1.2 Stepwise Tests

For the stepwise test, this chapter sets the first feature selection stage with a significance level of 10%. Figure 3.8 compares the performance of stepwise with alternative approaches. The backward induction method eliminates variables with a significance value over 10%. The forward-backward method includes variables with a significance value under 10.1% and drops variables with a significance value over 10% in the new set of data for regression. The backward-forward method starts with a full model with all the original variables, eliminating variables with significance values over 10% and including dropped variables at each step if the significance value is under 10.1% in the new data set.

[Table 3.4, p.106 about here]

In the four stepwise approaches, backward and backward-forward approaches generate similar results, while the forward and forward-backward methods present similar conclusions. The backward and backward-forward methods achieve the best prediction in terms of the likelihood ratios and the Pseudo-R^2. However, the backward approach is more efficient than the backward-forward approach, because the backward-forward approach takes twice the computation time than that of the backward approach. It is worth noting that all the approaches select all market variables, the default history, and firm ages. Most accounting variables selected are similar except QATA, CATA, SATA, RERA and CFTD. Only two macroeconomic variables are selected by all the four approaches, which are the SPRETURN and INVESTL. The backward approach keeps the most number of variables (58), while the forward-backward method keeps the least number of variables, held at 40. On account of its prediction accuracy and computational efficiency, and the similarity between different stepwise approaches, the rest of this chapter will therefore

use the backward stepwise method for further stepwise tests.

3.4.1.3 Feature Selection Methods Comparison

Table 3.4 compares the performance of two different feature selection methods, the t-test and the backward stepwise method, with alternative significance levels. It is notable that five variables (DGS10_BAA, BAA_AAA, DGS10_DGS1, GDPL and GDPC96L) are deleted at the first stage because of the problem of collinearity. The total number of variables entered in the original regression is 125. The regression with the original variable set has the log-likelihood ratio as -9098.02, the Pseudo-R2 as 0.361, and AUROC as 0.9433. Overall, at the same significance level, the t-test shows higher prediction accuracy because it results in a higher Log-Likelihood ratio, a higher Pseudo-R^2, and a higher AUROC rate. On the other hand, the stepwise method extracts the initial variables effectively, while under the same probability rate the stepwise method keeps a lower number of explanatory variables. This quality is an effect of delimiting the correlations between variables. Also, comparing these methods with the same number of variables, for example when the number equals 34, the stepwise method provides a better prediction in terms of the AUROC, Pseudo R^2 and Log-likelihood ratio. Furthermore, although the means show significant difference between groups, the t-test cannot select the best-combined variable set. For instance, in the 29-variable logit regressions of similar means between groups, five insignificant variables (DFFL, CD1ML, CFTD, CACL and TDTA) were selected and included in the regression. Therefore, this chapter employs the backward stepwise as the feature selection technique to address the feature selection issue.

[Table 3.5 p.108 about here]

3.4.2 Variable Contributions

[Figure 3.7 p.129 about here]

Figure 3.7 ranks the significance and power of explanatory variables taken from the initial

pool of variables using the backward selection method. All variables presented in Figure 3.7 are significant for predicting corporate default, and are extracted by the filter of the backward stepwise technique.

Figure 3.7 indicates that the most important variables for default probability are Default History (DH) and Net Income (NI) ratio, and this chapter marks these as level one variable. The second most significant level of variables includes Firm Size, measured in terms of Market Capacity (FMVTMV) and Total Assets (LOGTAGNP). The third level of significance includes Cash ratio (CASHTA), Firm-level Stock Return (RETURN), and Stock Risk (SDF). At the fourth level, the first macroeconomic variable, Market Index Return (SPRETURN), is selected with five more firm level variables (EBITA, SALEG, QPRID, EXCESS, and PRICE). Five accounting ratios (EBIATL, CFTL, QACL, WCTA and METL) are added at the fifth significant level, while six further macroeconomic variables and the firm age (AGEQ) are significant at the sixth level.

Figure 3.7 further shows the optimal variable set for prediction, at all significance levels, combining information from different categories. Firm-level information – including accounting ratios, market variables and default history – contributes most to explaining default probability, while macroeconomic variables add additional information for predicting corporate default from level four.

[Table 3.6 p.109 about here]

The detailed regression results of the backward logit selections with alternative significance levels are presented in Table 3.6. The signs of the variable coefficients remain stable across different variable sets.

In the accounting ratios section, all the listed variables indicate that firm profit and firm size are selected at the 10% significance level. The signs of NITA, EBITA and INTWO in the profit section show that default risk is associated with low profit. The positive sign of EBIATL is unexpected; presumably this results from correlation between variables. When

cutting the number of variables to 14 or fewer by changing the significance level, the remaining indicators of profit (NITA, EBIATA) clearly confirm that profit is negatively related to the probability of default. The results in the operation and leverage sections are all as expected. Firms are more likely to default if the operation is less efficient, with a lower sales growth rate and limited operating income. High leverage firms suffer greater credit risk. However, the leverage variables are least significant in accounting ratios because half the indicators in this section have been dropped as they show statistical insignificance, and all indicators are excluded when the number of variables is reduced to 14. Seven variables are kept at the 10% significance level to reveal the relation between default risk and liquidity. It is difficult to draw conclusions at this level since the sign of the variables is controversial due to high correlations. When the stepwise regression reduces the variable number to 46, with four liquidity indicators remaining, the signs of these indicators, Working Capital ratio, Cash over Total Assets (CASHTA) and Current Liability over Total Assets, suggest that a less liquid firm is more likely to default. This result is confirmed in the 14-variable results, when the Quick Ratio is excluded and the coefficient of CLTA is increasing. Finally, a large firm by asset size is more likely to default, as shown by the sign of LOGTAGNP after dropping the correlated variable LOGTA from the 26-variable regression.

For market variables at firm level, we can conclude that a higher default risk is negatively correlated with firm stock price (PRICE, LOGPRICE), Earnings per Share (EPS) and Market to Book ratio (MBR), while it is positively correlated with the market volatility measure as the standard deviation (SDF), Price Gap (QPRID) and Price Trend (PRICETREND). Two pairs of more highly correlated variables (RETURNQF&EXCESS; LOGMC&FMVTMV) have conflicting signs for coefficients from column 2 to column 4. However, when the LOGMC is excluded at the 0.01% significance level, the coefficient of FMVTMV increases accordingly and the sign of FMVTMV suggests that a smaller market capacity is associated with a higher default risk. The sign of RETURNFQ in the 8-variable regression, after dropping the

EXCESS, indicates that firm stock return is negatively correlated to default probability.

The coefficient of default history and firm age stays stable across all significant levels, suggesting these two variables reflect distinct information about default behaviour. A young firm with a default history is more likely to default.

In the macroeconomic variable section, the S&P 500 index return shows the strongest explanatory power for default risk. A higher market index return (S&P 500) is associated with a higher default probability. Other macroeconomic variables including CD1M, GNP, INVEST, GNPL and INVESTL enter the explanatory set when the number of variables increases to 26 with the p value as 0.0001.

Among the nine default indicator levels, Level six with 26 variables is recommended as the optimal variable set, for three reasons. The first reason is that it maintains most of the prediction power of the original variable set (99.25% of the AUROC curve). Second, it keeps only 20% of the variables from the original variable set, which largely decreases the correlation between variables and extracts repeated information. Figure 3.8 presents correlations between variables under the process of variable selection. The number of pairs with correlations over 0.9 decreases from over 600 in the original variable set to fewer than 100 in the optimal variable set. The number of pairs with correlations over 0.5 drops from over 1600 in the original variable set to under 200. Although the 19- variable (or fewer) set eliminates more correlations, the 26-variable set includes more macroeconomic information, which is one of the main objectives of this thesis. Thirdly, the significance of the variables in the 26-variable set remains relatively stable when the combination of variables changes (see Table 3.8). Stability is a critical standard for selecting independent variables. If it sets the significance lever to 0.1% (level seven), more variables lose their significance in a single sector prediction model.

In order to compare the best variable set with previous studies, this chapter tests alternative variable sets from previous papers, using the example of investigating U.S.

bankruptcy by the dynamic logit method. The comparison tests take the best predicting variable sets in each paper. The coefficients, Z-statistic and general prediction indicators are presented in Table 3.10. The sample period runs from 1970 to 2009. In summary, the optimal variable set presented in this thesis out-performs all previous studies in terms of the likelihood ratios, Pseudo R^2 and the AUROC. The new model shows increased prediction power even when using 5 variables from significance level 1 and level 2 (9th column in Table 3.6). When including both accounting and market indicators in variable sets following previous studies (Chava & Jarrow, 2004; Beaver, McNicholes & Rhie, 2005; Campbell, Hilscher & Szilagyi, 2008), the result has shown better performance than prediction based on a single accounting ratio (as per Altman, 1968; Ohlson, 1980). Most variables in these prior tests show significance in the updated sample.

[Figure 3.8, p.130 about here]

[Table 3.7, p.111 about here]

3.4.3 Regressions with Variables from Different Categories

In order to investigate the role of variables from different categories, Table 3.8 presents the regression results of the dynamic logit model. The tests are based on different combinations of explanatory variables. Each column presents the coefficients, Z-values and significance levels of the different tests. In the lower part of Table 3.8, some additional information on the estimations are displayed including the number of variables, log-likelihood index, the Pseudo R^2 value and AUROC.

[Table 3.8, p.112 about here]

The strongest fit in Table 3.8 is model 1, with the largest Pseudo R^2 value of 0.3524, Log-likelihood of -9219.57, and AUROC of 0.9391. This result suggests that corporate default can be attributed not only to accounting ratios, but also to market information, macroeconomic indicators and other segments of information include DH and AGEQ. A

combination of firm-level information and macroeconomic information would be a preferred explanatory variable set. In model 1, all variables show significance at the 1% level.

Considering the prediction ability, there is little to choose between the single accounting variable set-based model (Model 2) and the market variables-based model (Model 3), although Model 3 has a slightly larger Pseudo R^2 and log-likelihood index. The results show that the stock market alone is not efficient enough to include all categories of information for default events. When market information was combined with accounting information in Model 6, the prediction ability improved considerably, yielding a Pseudo R^2 of 0.2687 and AUROC of 0.9147.

Model 4 gives the weakest prediction, with a Pseudo R^2 of 0.0152 demonstrating that the macro variables alone are not sufficient to explain corporate default behaviour, even though all the macro variables present as significant. The improvement in the AUROC and Pseudo R^2 when comparing Model 1 to Model 7 shows that the macro variables can add extra explanatory power and supplement the forecasting ability.

The result in Table 3.8 also indicates that other firm-level information, such as the default history (DH), plays an important role in default prediction. While the prediction ability reveals limited power for default prediction in Model 5, the forecasting ability is improved from Model 6 to Model 7 by adding the default history and firm age as explanatory variables. The positive coefficient of DH in Table 3.8 suggests that a firm is more likely to default if it has previous defaulted. The negative coefficient of AGEQ reveals that new firms are more likely to default.

3.4.4 Regressions of Different Horizons

Table 3.8 summarises the results of a one-quarter ahead of the prediction time period. This section explores the significance of variables from different categories conditioned on

a longer time horizon. In Table 3.9, the conditional probability of default is estimated in 1 quarter, 2 quarters, 3 quarters, 4 quarters, 5 quarters, and 8 quarters ahead respectively.

[Table 3.9, p.114 about here]

As the horizon increases in Table 3.9, the overall prediction ability declines, as one would expect. This pattern is not just found in the optimal variables set, as similar results are found if the regression is run using more or fewer variables than suggested in Figure 3.7. It is observed that, generally speaking, accounting variables become less important compared with market variables as the time horizon increases. Two independent variables are particularly important for long-term prediction: the z-statistic on the Earnings per Share (EPS) ratio increases with the horizon; and the z-statistic on the stock price (PRICE) also increases slightly. The firm stock return and firm excess return have z-statistics that decay particularly rapidly with the horizon, indicating that they are primarily short-term signals of credit risk. Firm age, Return on Assets, Earnings over Total Liability, Working Capital ratio, log GNP and log Investment at bank are intermediate, with effects that decay more slowly. The coefficient on the Quick Ratio QACL switches to the expected sign on a longer horizon, presumably as a result of the decrease of the significance of other variables from the liquidity section (WCTA, CLTA), which are more short-term in nature. It is notable that most macroeconomic variables switch signs in a longer horizon prediction.

The number of observations and the number of default firms varies with the time horizon in Table 3.9. This is because increasing the horizon forces observations to be dropped at both the beginning and the end periods of the sample. The number of observations for each regression is reported in the bottom row. Following the study by Campbell, Hilscher & Szilagyi (2008), the sections below run the subsample regressions for different horizons. The similar results from this subset test tell us that the slight difference in sample size in Table 3.9 is not responsible for the results.

3.4.5 Different Sub-periods

Table 3.10 reports the estimation and prediction results for alternative sub-periods. The total period is divided into 4 sub-periods and each sub-period contains ten years of observations. This section of the thesis seeks to investigate the prediction power of explanatory variables from different categories during different decades. Generally speaking, the optimal variable set performs stably, and the prediction ability is strong for all four decades. All the AUROC are above 0.9, and sub-period prediction of the 1970s, 1990s and 2000s are even better than the whole 40-year prediction.

[Table 3.10, p. 116 about here]

Most macroeconomic variables lose significance in the sub-sample period prediction. This indicates that macroeconomic variables are more likely to be long-term default drivers.

However, the S&P 500 return (SPRETURN) shows significance in most periods at 1% significant level. The 1-month certificate of deposit (CD1m), the amount of the bank investment (INVEST), and its natural logarithm (INVESTL), show significance in the last decade model. This may reveal that macroeconomic information plays a more important role in the new financial environment.

The significance levels of the market variables (FMVTMV, RETURNFQ, PRICE, EPS and SDF) increase steadily from the 1970s to the 21st century. One explanation is that the market has become more efficient with the development of modern techniques. On the other hand, the effect of firm age (AGEQ) has become weaker recently. Some accounting and market variables became insignificant in the earliest decade (1970 to 1979), presumably as a result of the shortage of observations in early years. Also, the sign differences between the period 1970 to 1980 and other periods might reveal either the data shortage of the COMPUSTAT or the survivorship problem of the sample. It is worth noticing that only 54279 observations in the period of 1970 to 1980.

3.4.6 Industry Effects

This section investigates the importance of industry effects in predicting corporate default using a dynamic logit model. As presented in Figure 3.6, all firms are divided by ten industries using the SIC code. Following previous studies (Chava & Jarrow, 2004; Hensher & Jones, 2007), ten classifications are combined into four sub-groups for estimating purposes. The combination criterion is the default rate of each industry calculated in Figure 3.6. The four groups for investigating industry effect are: i) agriculture, forestry, fishing and retail trade (default rate above 0.45%); ii) mining, construction, wholesale trade and service (default rate between 0.35% and 0.38%); iii) transportation, communications, utilities and manufacturing (default rate between 0.29% and 0.3%); iv) public administration, finance, insurance and real estate (default rate below 25%).

This section investigates industry effects for both public and private firms using the optimal variable set with the one-quarter horizon. It also tests industry effects over longer horizons of 2, 4, and 8 quarters ahead. The results are presented in Table 3.11. Compared with non-industry effect prediction, tests with additional industry dummy variables have increased prediction ability. The AUROC improves with industry effect regressions compared with the matching horizon predictions in Table 3.10.

[Table 3.11, p. 118 about here]

The industry effect is statistically significant for both public and private firms. Private firms are more likely to be influenced by industry sector. This, presumably, is the result of the shortage of market information for private firms. The negative sign of industry 2 to industry 4 (RIND2, RIND3 and RIND4) indicates that industry 1 (agriculture, forestry, fishing and retail trade) has the highest unconditional default probability, as expected. Industry 4, on the other hand, is the least likely to default, consistent with Figure 5. Industry 2 has a slightly smaller coefficient than industry 3 in all tests. Another important conclusion from Table 3.10 is that industry effect is significant for both short and long

time horizon prediction. The industry effect is more significant in a longer horizon prediction, such as half-year and one-year.

I have also tested the interactive industry variables as previous studies suggest: the product of industry dummy and NITA, the product of industry dummy and TLTA and the product of industry dummy and FMVTMV. The results do not show significance consistently.

3.4.7 Classification and Prediction Ability

[Figure 3.9, p. 131 about here]

To evaluate the prediction power of the optimal variable set with the logit model, this chapter compares the actual default rate and the predicted default rate in Figure 3.8. The predicted default rate is calculated as a fraction of the companies over time. The model captures most of the default risk trends and variations. It gives a strong signal for all high-defaulted periods including the early 1990s, early 2000s, and the current financial crisis. In particular, the prediction model describes the current financial crisis almost perfectly. However, the model somewhat under predicts in the early 1980s.

3.4.7.1 Classification Rates

As a classification default model, the estimated error can be measured in two ways. First, the model under Type I error can indicate low default risk when the default risk is actually high. Second, the model suggests high credit risk when, in fact, the default risk is low. This is referred to as Type II error. The accuracy and error scenarios are described in Table 3.12. It is notable that the costs associated with Type I and Type II errors in default risk classification are different. Chapter 5 illustrates the details of classification costs. This chapter refers to the rate of the correctly classified default firms as sensitivity, and the rate of correctly classified non-default firm as specificity. The sensitivity and specificity rates vary with the cut-off points for the Type I and II errors. The increase in sensitivity is,

unfortunately, associated with a decrease in specificity.

[Table 3.12, p. 120 about here]

Table 3.13 presents the forecasting classification rates for each group and the whole model from 2000 to 2009. The out-of-sample prediction uses the sample from 1970 to 1999 to estimate default probability from 2000 to 2009. The accuracy rate of both in-sample and out-of-sample prediction is over 98% when cut-off points are chosen at and above 0.05. The prediction ability is slightly deteriorated in out-of-sample performance. Considering the high cost of the Type I error, the optimal cut-off point in Table 3.13 is 0.005. With this cut-off point, the overall prediction accuracy is 86.84% for the out-of-sample performance and 89.03% for in-sample performance. The Type I error for the in-sample test is 8.98%, while the Type I error of the out-of-sample test is 11.12%. The Type II error in this case is also considerably lower, with 13.16% for the out-of-sample test and 10.98% for the in-sample test.

[Table 3.13, p. 121 about here]

3.4.7.2 AUROC

The AUROC rate has been referenced to compare models throughout the analyses in this thesis. In this section, the AUROC is drawn upon to investigate the validation of the models for both in-sample and out-of-sample tests. The AUROC provides all cut-off points that might be chosen. This gives a clear and straightforward parameter to conduct a model comparison: the higher the AUROC, the better the prediction. In this section, 1000 bootstrap simulations are conducted to calculate the AUROC area. Table 3.14 reports details of the AUROC estimation. The rolling out-of-sample test has also been investigated, as suggested by Sobehart, Keenan & Stein (2001). The AUROC plot in Figure 3.10 confirms the prediction ability of the optimal variables set, and that both the in-sample and out-of-sample tests achieve considerably high AUROC rates.

[Table 3.14, p.122 about here]

[Figure 3.10, p.132 about here]

3.5 Conclusion

This chapter analysed default probability with the most comprehensive and up-to-date database on corporate default from the United States. The dynamic logit stepwise selection procedure and t-tests enabled the significance of default drivers to be marked by different levels.

More importantly, this chapter investigated information not only from financial statements, but other previously overlooked sources which has formed a model with improved prediction ability, namely market variables, macroeconomic variables, default records, firm age effect and industry effect. Figure 3.7 clearly indicates that the default history and the Return on Assets are the most significant default drivers, followed by firm size and current liability over total assets. The optimal variable set contains information from both firm level and macroeconomic level. This optimal variable set out-performs the best variable sets of previous studies and captures default intensity considerably efficiently and consistently over different decades. In the optimal variable set, the variables PRICE and EPS are more significant for longer horizon prediction, while the macroeconomic variables show more significance at a long period prediction. Moreover, the industry effects are detected under alternative horizons with both private and public firms.

Table 3.1

Default Database Constructions

Table 3.1 describes the procedure of the default data selection for the empirical test. The total default number decreases from 6,296 to 2,123 after including matching quarterly level explanatory information.

Inclusion criteria in steps	Default number
Initial default sample from Moody's, COMPUSTAT, BankruptcyData.com and Fitch Rating	6296
Firms exist in COMPUSTAT and CRSP database (Public firms)	3112
Firms for which both the financial statement and stock price are available in COMPUSTAT and CRSP database	2123

Table 3.2

Descriptive Statistics of Firm-level Variables

Table 3.2 presents the statistical summary of firm-level information from 1970 to 2009. The summary is recorded in quarterly units. The summary statistics information are shown for all firms, default firms, and non-default firms. It also includes the result of the t-test in the last column. The total number of observations is 639,573, with 2,123 observations in the default group. The abbreviations of variables are presented in Table 2.2 and Table 2.3. Y dummy stands for the year dummy variables, AGEQ is the firm's age, and DH is default history equal to 1 if a firm has defaulted before, 0 otherwise.

Variable	Min	Max	Mean	S.D.	Mean	S.D.	Mean	S.D.	
Observation Number	All firms: N=639573				Default Firms N=2123		Non- Default Firms: N=637450		T mean test
NITA	-0.20	0.05	-0.01	0.06	-0.09	0.08	-0.01	0.06	43.37
EBITA	-0.16	0.08	0.01	0.06	-0.05	0.07	0.01	0.06	39.40
EBIATL	-0.31	0.23	0.02	0.12	-0.06	0.11	0.02	0.12	35.70
RETA	-4.43	0.56	-0.34	1.22	-1.39	1.62	-0.34	1.21	29.89
INTWO	0.00	1.00	0.50	0.50	0.93	0.26	0.49	0.50	-78.46
SATA	0.04	0.78	0.31	0.20	0.30	0.22	0.31	0.20	0.06
SALEG	-0.39	0.67	0.05	0.24	-0.02	0.21	0.05	0.24	15.88
CFSA	-1.91	0.31	-0.09	0.50	-0.40	0.64	-0.09	0.50	22.65
CFTD	-1.54	1.66	0.07	0.60	-0.23	0.49	0.07	0.60	27.68
CFTL	-0.32	0.21	0.01	0.12	-0.06	0.10	0.01	0.12	34.08
NOCFTA	-1.12	0.14	-0.02	0.17	-0.09	0.21	-0.02	0.17	16.03
CACL	0.07	13.08	2.25	2.02	1.14	1.48	2.25	2.02	34.64
QACL	0.07	13.08	2.24	2.02	1.14	1.48	2.24	2.02	34.30
WCTA	-0.34	0.63	0.19	0.25	-0.08	0.27	0.19	0.25	46.29
CASHCL	0.01	3.55	0.61	0.93	0.23	0.55	0.62	0.93	31.76
QATA	0.03	0.96	0.49	0.25	0.46	0.26	0.49	0.25	5.78
CATA	0.03	0.96	0.49	0.25	0.46	0.26	0.49	0.25	6.03
CLTA	0.07	0.82	0.30	0.20	0.55	0.27	0.29	0.19	-44.31
CASHTA	0.00	0.52	0.12	0.14	0.07	0.10	0.12	0.14	20.42
TLTA	0.18	1.21	0.58	0.26	0.95	0.27	0.58	0.26	-62.27
TDTE	-1.06	3.83	0.76	1.08	0.59	1.79	0.76	1.07	4.20
OENEG	0.00	1.00	0.07	0.26	0.49	0.50	0.07	0.26	-38.18
TDTA	0.01	0.79	0.30	0.22	0.50	0.26	0.30	0.22	-35.13
METL	0.15	16.36	3.26	4.18	0.84	1.98	3.27	4.18	56.16
METD	0.25	217.54	22.93	52.45	5.50	22.98	22.99	52.51	34.77
LOGTAGNP	-9.75	1.40	-4.02	2.40	-4.92	2.10	-4.02	2.40	19.68
LOGTA	0.91	8.90	4.77	2.24	4.10	2.02	4.77	2.24	15.18
PRICE	0.23	49.63	14.54	14.28	2.74	6.61	14.58	14.28	81.87
LOGPRICE	-1.47	3.90	1.93	1.50	-0.14	1.33	1.93	1.50	71.52
RETURNFQ	-0.62	0.47	-0.02	0.26	-0.16	0.30	-0.02	0.26	20.56
EPS	-0.67	1.19	0.18	0.44	-0.28	0.40	0.18	0.44	52.91
QPRID	0.13	14.59	4.22	3.98	1.76	2.47	4.23	3.98	45.82
PRICETREND	-0.25	0.18	-0.01	0.11	-0.07	0.12	-0.01	0.11	22.88
LOGMC	0.48	8.59	4.41	2.28	2.25	1.62	4.41	2.28	61.45
MBR	-0.74	9.15	2.23	2.35	0.84	2.24	2.23	2.35	28.58
SDF	1.75	42.59	13.60	10.89	26.95	13.84	13.55	10.85	-44.57
FMVTMV	-17.48	-9.32	-13.36	2.27	-15.77	1.67	-13.35	2.27	66.34
EXCESS	-1.03	0.42	-0.09	0.30	-0.26	0.43	-0.09	0.30	19.04
Y dummy	1970	2009	1994	9.26	1996	8.01	1994	9.26	-16.40
AGEQ	0.00	157.01	35.65	31.67	37.87	31.35	35.64	31.67	-3.28
DH	0.00	1.00	0.02	0.15	0.47	0.50	0.02	0.14	-41.52

Table 3.3

Macroeconomic Variables Descriptive Statistics and T-test

Table 3.3 presents the summary statistics and the results of the t-test for macroeconomic variables from 1970 to 2009. It reports information from the original variables, the natural logarithm of the original variables, and the growth rate of the original variables. The total number of observations is 639,573, with 2,123 observations in the default group. The abbreviations of variables are presented in Table 2.4.

Variable name	Original variables (OV)			Log (OV)			Growth rate of OV		
	Means		T-test	Means		T-test	Means		T-test
	Non-default firms	Default firms		Non-default firms	Default firms		Non-default firms	Default firms	
SPRETURN	0.020	0.014	2.83						
SDM	0.039	0.037	-5.518						
CPIAUCSL	144.449	157.458	-16.60	2.136	2.184	-19.19	0.009	0.007	10.18
PPIACO	121.357	128.260	-13.35	2.071	2.101	-16.47	0.007	0.004	5.80
GNPC96	9228.835	9961.422	-15.12	3.948	3.986	-16.83	0.007	0.006	6.47
GNP	7586.428	8615.441	-14.22	3.819	3.896	-18.18	0.015	0.013	13.04
GDPC96	9173.415	9903.871	-15.20	3.946	3.984	-16.91	0.007	0.006	6.99
GDP	7541.653	8564.968	-14.28	3.817	3.894	-18.22	0.015	0.012	13.63
UNRATE	6.080	5.942	4.34	0.772	0.762	4.63	0.002	0.015	-10.08
UNEMPLOY	7794.562	7988.984	-4.39	3.883	3.891	-3.79	0.005	0.018	-10.06
IMPGSC1	1072.525	1227.846	-12.73	2.952	3.033	-16.00	0.016	0.011	7.75
EXPGSC1	817.275	924.063	-13.17	2.849	2.922	-15.94	0.014	0.010	7.95
NICUR	6659.582	7559.648	-14.20	3.762	3.839	-18.09	0.015	0.012	14.22
INDPRO	72.028	76.840	-14.15	1.844	1.876	-15.38	0.006	0.003	8.52
AAA	8.115	7.646	11.23	0.894	0.871	10.36	-0.004	-0.007	2.76
BAA	9.179	8.727	10.29	0.949	0.931	9.13	-0.004	-0.007	2.76
DTB3	5.235	4.411	14.30	0.619	0.501	11.98	-0.009	-0.038	5.99
DTB6	5.368	4.511	14.91	0.643	0.533	12.90	-0.007	-0.038	7.12
DGS1	5.839	4.901	15.18	0.687	0.584	13.48	-0.003	-0.027	5.60
DGS10	7.000	6.300	13.78	0.816	0.771	13.27	-0.003	-0.003	0.12
CD1M	5.888	4.996	14.04	0.684	0.581	12.29	-0.011	-0.060	12.38
DFF	6.154	5.114	14.46	0.679	0.550	12.20	0.004	-0.033	6.25
DPRIME	8.281	7.517	13.27	0.892	0.848	12.70	-0.004	-0.024	9.71
BAA_AAA	1.064	1.081	-1.599						
DGS10_BAA	-2.179	-2.427	12.671						
DGS10_DGS1	1.161	1.399	-9.893						
BUSLOANS	739.520	849.217	-16.27	2.819	2.896	-20.06	0.014	0.006	15.05
CONSUMER	436.525	489.697	-14.00	2.589	2.659	-18.38	0.015	0.012	8.36
REALLN	1281.145	1519.237	-10.93	2.960	3.079	-17.32	0.024	0.023	4.06
LOANS	2852.787	3313.228	-12.51	3.369	3.460	-17.55	0.018	0.015	11.24
INVEST	958.823	1104.660	-12.08	2.893	2.981	-16.42	0.018	0.021	-5.31
GFDEBTN	4442253	5162263	-12.95	6.538	6.645	-18.15	0.022	0.021	1.76
NOM1M2	3021.295	3485.011	-14.02	3.419	3.499	-18.55	0.016	0.016	2.73

Table 3.4

Stepwise Regressions with Alternative Approaches

Table 3.4 compares the stepwise performance using different approaches. The total number of observations is 639,573, with 2,123 observations in the default group. For each variable, the table reports the value of the estimated coefficient. The backward method eliminates variables with a significance value over 0.1. The forward method includes variables with a significance value under 0.1. The forward-backward method includes variables with a significance value under 0.1, and drops variables with a significance value over 0.1 in the new set regression. The backward-forward method starts with a full model with all the original variables, eliminates variables with a significance value over 0.1, and includes dropped variables at each step if their significance value is under 0.1 in the new set. The abbreviations of variables are presented in Table 2.2, Table 2.3 and Table 2.4. Y dummy stands for the year dummy variables, AGEQ is the firm's age, and DH is default history equal to 1 if a firm defaulted before, 0 otherwise. ∗denotes significant at 10%, ∗∗denotes significant at 5%, *** denotes significant at 1%.

	Backward	Backward forward	Forward	Forward backward
Var. No.	58	52	41	40
NITA	-3.935***	-4.108***	-3.937***	-4***
EBITA	-5.933***	-5.921***	-5.81***	-5.799***
EBIATL	5.883***	5.908***	5.757***	5.775***
INTWO	0.237**	0.226**	0.224**	0.22**
SALEG	-0.729***	-0.76***	-0.737***	-0.734***
CFTL	-5.817***	-5.368***	-5.763***	-5.349***
CACL	-3.468*	-3.548*	-3.379*	-3.462*
QACL	3.570*	3.647**	3.482*	3.561*
WCTA	-1.083***	-1.085***	-1.07***	-1.065***
CLTA	1.411***	1.413***	1.424***	1.426***
CATA	0.328**	0.392**		
CASHTA	-3.154***	-3.151***	-3.12***	-3.166***
CASHCL	0.194**	0.188**	0.187**	0.192**
TLTA	0.450***	0.537***	0.446***	0.446***
TDTE	0.088***	0.085***	0.089***	0.09***
METL	-0.063***	-0.063***	-0.063***	-0.063***
LOGTAGNP	0.677***	0.656***	0.671***	0.675***
QATA			0.319*	0.321**
LOGTA	-0.397***	-0.395***	-0.391***	-0.391***
SATA		-0.218*		
RETA		0.044*		
CFTD		-0.136*		-0.125*
FMVTMV	-0.600***	-0.596***	-0.574***	-0.565***
RETURNFQ	-5.335***	-5.357***	-5.338***	-5.301***
SDF	0.034***	0.034***	0.034***	0.034***
LOGPRICE	-0.127***	-0.136***	-0.127***	-0.125***
QPRID	0.105***	0.105***	0.104***	0.103***
PRICETREND	0.897***	0.915***	0.887***	0.925***
EXCESS	3.931***	3.948***	3.947***	3.914***
LOGMC	0.347***	0.353***	0.322***	0.312***
PRICE	-0.042***	-0.041***	-0.042***	-0.042***
EPS	-0.468***	-0.454***	-0.478***	-0.46***
MBR	-0.036**	-0.035**	-0.037***	-0.038***

DH	2.859***	2.862***	2.852***	2.842***
AGEQ	-0.004***	-0.004***	-0.004***	-0.004***
SPRETURN	3.722***	3.849***	3.997***	4.061***
CD1M	-0.492***	-0.398***		
DFF	0.147***	0.082**		
UNRATE	-0.942***			
DFFL	-2.753***	-1.412***		
DTB3L	3.290***	2.309***		
UNRATEL	13.858***			
CONSUMERL	-10.876***			
LOANSL	23.472***			
GDPC96G	16.478***	9.261*		
GNP	-0.002***	-0.002***		
INDPROG	-10.232***	-5.278*		
DTB3G	-0.842***	-0.408***		
DPRIMEG	1.361**			
INVEST	0.007***	0.004**		
BAA	0.834**			
GNPL	35.478***	28.319***		
IMPGSC1	-0.003**	-0.001**	-5.112***	
INVESTL	-24.423***	-11.729***	-5.272***	-3.202***
DTB3	0.266**			
DGS10	-0.545**			
BAAL	-13.170**			
BUSLOANS	0.011***	0.008***		
IMPGSC1L	7.260**			
BUSLOANSL	-26.313***	-17.724***		
DGS10L	5.938*			
UNRATEG	-2.074**	-1.468*		
CD1MG			-0.791***	-0.522***
CD1ML			-0.386***	
GNPC96L			20.037***	
DPRIMEG			1.175**	
INDPROL			-2.413***	7.512**
LOANSG			-0.25***	
NOM1M2L		0.001***	1.736*	
DFFL				-0.372***
INDPRO				-0.05***
CPIAUCSLG				15.526***
GDPC96		0.001***		
DPRIME		0.249***		
NOM1M2				3.089***
Constant	-71.726***	-42.136***	-62.415***	-23.692***
Log-Likelihood	-9120.61	-9121.21	-9152.26	-9153.93
Pseudo-R2	0.3594	0.3594	0.3572	0.3571

Table 3.5

Comparisons of the T-test and the Stepwise Selection

Table 3.5 compares the feature selection ability of the t-test and the stepwise technique under the original variable set containing 125 explanatory variables. The Log-likelihood, Pseudo-R^2, AUROC and the variable numbers are reported as the assessment criteria. The Level is given as a reference for Figure 3.7.

	Log-Likelihood	Pseudo-R^2	AUROC	Var. NO.	Log-Likelihood	Pseudo-R^2	AUROC	Var. NO.	Level
All variables	-9098.02	0.361	0.9433	125					
Significance level	**T-test**				**Backward stepwise**				
0.1	-9099.30	0.3609	0.9432	123	-9120.61	0.3594	0.9425	58	9
0.01	-9100.09	0.3608	0.9432	122	-9137.78	0.3582	0.9418	46	8
0.001	-9132.33	0.3586	0.9430	117	-9169.31	0.3560	0.9409	34	7
0.0001	-9132.72	0.3585	0.9430	116	-9219.57	0.3524	0.9391	26	6
1.00E-09	-9149.05	0.3574	0.9424	104	-9292.99	0.3473	0.9362	19	5
1.00E-12	-9149.35	0.3574	0.9424	100	-9363.70	0.3423	0.9352	14	4
1.00E-24	-9158.76	0.3567	0.9420	92	-9553.95	0.3290	0.9302	8	3
1.00E-86	-9403.46	0.3395	0.9353	34	-9910.40	0.3039	0.9235	5	2
1.00E-137	-9459.57	0.3356	0.9347	29	-10779.10	0.2429	0.8837	2	1

Table 3.6

Regressions with Different Significance Tiers

Table 3.6 reports the regression results of various best-combined variable sets selected by the stepwise method in terms of different significance levels. The sample period is 1970 to 2009. For each variable, the table reports the value of the estimated coefficient. Total number of observations is 639,573. *denotes significant at 10%, **denotes significant at 5%, *** denotes significant at 1%.

Var. No. / Var. Name	58	46	34	26	19	14	8	5	2
NITA	-3.935***	-3.812***	-3.815***	-4.014***	-5.133***	-5.397***	-8.746***	-9.472***	-12.150***
EBITA	-5.933***	-6.068***	-6.172***	-6.334***	-5.880***	-5.437***			
EBIATL	5.883***	5.908***	5.958***	5.938***	5.553***				
INTWO	0.237**								
SALEG	-0.729***	-0.726***	-0.722***	-0.720***	-0.735***	-0.752***			
CFTL	-5.817***	-5.914***	-5.891***	-5.821***	-5.785***				
CACL	-3.468*								
QACL	3.570*	0.148***	0.151***	0.152***	0.160***				
WCTA	-1.083***	-0.876***	-0.904***	-1.010***	-0.987***				
CLTA	1.411***	1.613***	1.600***	1.703***	1.797***	2.130***	2.215***	2.492***	
CATA	0.328**								
CASHTA	-3.154***	-2.553***	-2.536***	-2.588***	-2.753***	-2.908***	-2.724***		
CASHCL	0.194**								
TLTA	0.450***	0.488***	0.474***						
TDTE	0.088***	0.091***	0.063***						
METL	-0.063***	-0.061***	-0.071***	-0.089***	-0.101***				
LOGTAGNP	0.677***	0.679***	0.643***	0.346***	0.363***	0.402***	0.324***	0.368***	
LOGTA	-0.397***	-0.403***	-0.345***						
FMVTMV	-0.600***	-0.606***	-0.601***	-0.324***	-0.284***	-0.331***	-0.344***	-0.458***	
RETURNFQ	-5.335***	-5.310***	-5.282***	-5.122***	-5.223***	-5.288***	-0.375***		
SDF	0.034***	0.034***	0.034***	0.035***	0.035***	0.035***	0.038***		
LOGPRICE	-0.127***	-0.139***	-0.149***						
QPRID	0.105***	0.109***	0.108***	0.110***	0.110***	0.105***			
PRICETREND	0.897***	0.905***	0.907***						
EXCESS	3.931***	3.924***	3.927***	3.965***	4.034***	4.057***			
LOGMC	0.347***	0.354***	0.333***						
PRICE	-0.042***	-0.045***	-0.046***	-0.062***	-0.078***	-0.080***			
EPS	-0.468***	-0.538***	-0.518***	-0.451***					
MBR	-0.036**	-0.037***							
DH	2.859***	2.849***	2.829***	2.833***	2.740***	2.747***	2.819***	2.862***	3.744***
AGEQ	-0.004***	-0.004***	-0.004***	-0.004***					
SPRETURN	3.722***	3.918***	4.043***	3.991***	4.403***	4.409***			
CD1M	-0.492***	-0.328***	-0.093***	-0.092***					
DFF	0.147***	0.166***							
UNRATE	-0.942***	-0.536***							
DFFL	-2.753***	-2.794***							
DTB3L	3.290***	3.318***							
UNRATEL	13.858***	7.458***							
CONSUMERL	-10.876***	-8.539***							
LOANSL	23.472***	13.690***							
GDPC96G	16.478***	15.858***							
GNP	-0.002***	-0.002***	-0.001***	-0.000***					
INDPROG	-10.232***	-8.322***							
DTB3G	-0.842***	-0.615***							
DPRIMEG	1.361**	1.522***							
INVEST	0.007***	0.008***	0.005***	0.003***					
BAA	0.834**								
GNPL	35.478***	43.412***	29.626***	14.790***					
IMPGSC1	-0.003**								
INVESTL	-24.423***	-26.031***	-18.927***	-14.162***					
DTB3	0.266**								
DGS10	-0.545**								
BAAL	-13.170**								
BUSLOANS	0.011***	0.009***	0.005***						

IMPGSC1L	7.260**								
BUSLOANSL	-26.313***	-19.736***	-10.252***						
DGS10L	5.938*								
UNRATEG	-2.074**								
Constant	-71.726***	-70.098***	-41.415***	-25.048***	-9.785***	-10.344***	-11.470***	-12.547***	-6.864***
AUROC	0.9425	0.9418	0.9409	0.9391	0.9362	0.9352	0.9302	0.9235	0.8837
Log-Likelihood	-9120.61	-9137.78	-9169.31	-9219.57	-9292.99	-9363.70	-9620.17	-9910.40	-10779.05
Pseudo-R^2	0.3594	0.3582	0.356	0.3524	0.3473	0.3423	0.3243	0.3039	0.2429

Table 3.7

Regressions with the Best Variable Set in Previous Studies

Table 3.7 compares dynamic logit models using different variable sets from precious studies using the updated dataset. For each variable, the table reports the value of the estimated coefficient. The sample period is from 1970 to2009. Total number of observation is 639,573. The abbreviations of variables are presented in Table 2.2 and Table 2.3. CABCASHMTA is stock of cash and short-term investments over the market value of total assets. CABTLMTA is total liabilities over market value of total assets. NITMAAVG is the geometrically declining weighted net income over market value of total assets. EXCESSAVG is the geometrically declining weighted log of gross excess return over value-weighted S&P 500 return. CJIND2 and CJIND3 are industry effects. CJIND2TLTA and CJIND3TLTA are the industry effects plus the TLTA. CJIND2NITA and CJIND3NITA are the industry effects plus the NITA. *denotes significant at 10%, **denotes significant at 5%, *** denotes significant at 1%.

Papers / Variables	ALTMAN 1968	OLSON 1980	BEAVER 2005	CHAVA 2004	CAMPBELL 2008
NITA		-6.190***	-3.196***	-6.533***	
LOGTAGNP		0.203***			
EBITA	-9.792***			1.308**	
EBIATL			0.175		
CFTL		-0.133			
WCTA	-3.238***			-0.960***	
METL				-0.334***	
CACL				0.094***	
INTWO		1.718***			
TLTA		3.724***	2.674***	1.083***	
SATA	0.532***			-0.363***	
CABTLMTA					4.536***
CABCASHMTA					-3.610***
NITMAAVG					-49.063***
NITL		-1.188***			
OENEG		-0.208**			
CHIN		0.304***			
CLCA		0.020***			
RETA	0.100***				
METD	-0.019***				
EXCESSAVG					-0.848***
FMVTMV				-0.125***	0.124***
RETURNFQ			-0.810***		
LOGPRICE					-0.217***
EXCESS				-0.330***	
LOGMC			-0.170***		
MBR					0.014
SDF			0.045***	0.043***	0.037***
CJIND3				-1.266***	
CJIND2				-0.679***	
CJIND2TLTA				0.960***	
CJIND3TLTA				1.562***	
CJIND2NITA				1.847***	
CJIND3NITA				1.838*	
Constant	-5.590***	-9.285***	-8.198***	-8.903***	-7.760***
Pseudo-R^2	0.1178	0.1734	0.1856	0.2059	0.2554
Log-Likelihood	-12560.4	-11769.4	-11594.9	-11306.1	-10602
AUROC	0.8247	0.8700	0.8958	0.8982	0.9114

Table 3.8

Regressions with Different Variable Categories

Table 3.8 presents the dynamic logit prediction with different variable combinations. The sample period is from 1970 to2009. Total number of observations is 639,573. For each variable, the first row reports the value of the estimated coefficient and the second row reports the value of the z- statistic. Model 1 is based on the optimal variable set. Model 2 is based on single accounting information. Model 3 is based on single market information. Model 4 is based on single macroeconomic information. Model 5 is based on the default history and the firm ages. Model 6 is based on both the accounting and market information. Model 7 is for private firms. The abbreviations of variables are presented in Table 2.2, Table 2.3 and Table 2.4. Y dummy stands for the year dummy variables, AGEQ is the firm's age, and DH is default history equal to 1 if a firm has defaulted before, 0 otherwise. *denotes significant at 10%, **denotes significant at 5%, *** denotes significant at 1%.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
DH	2.833***				4.045***		2.831***
	-48.30				-83.97		-49.10
AGEQ	-0.004***				-0.010***		-0.005***
	-4.57				-14.12		-6.04
NITA	-4.014***	-8.448***				-3.765***	-3.602***
	-7.22	-16.94				-7.20	-6.54
CASHTA	-2.588***	-1.490***				-1.913***	-2.747***
	-9.90	-6.37				-7.76	-10.67
CLTA	1.703***	2.392***				1.714***	1.765***
	-11.85	-16.76				-12.61	-12.35
LOGTAGNP	0.346***	0.176***				0.601***	0.314***
	-17.39	-15.34				-33.97	-16.00
SALEG	-0.720***	-0.798***				-0.699***	-0.731***
	-7.73	-9.44				-7.87	-7.88
EBITA	-6.334***	-0.39				-7.449***	-6.244***
	-7.07	-0.48				-8.62	-7.01
QPRID	0.110***		-0.044***			0.037***	0.064***
	-8.34		-3.21			-2.95	-5.05
EBIATL	5.938***	2.577***				6.669***	5.806***
	-6.63	-2.96				-7.91	-6.52
CFTL	-5.821***	-4.754***				-5.476***	-5.796***
	-7.15	-5.97				-7.22	-7.16
QACL	0.152***	0.186***				0.127***	0.159***
	-7.10	-10.77				-6.39	-7.53
WCTA	-1.010***	-1.654***				-1.098***	-0.942***
	-6.72	-12.11				-7.73	-6.35
METL	-0.089***	-0.460***				-0.072***	-0.106***

	-5.70	-19.26				-4.60	-6.69
FMVTMV	-0.324***		-0.233***			-0.493***	-0.253***
	-13.47		-11.94			-23.88	-10.96
RETURNFQ	-5.122***		-1.870***			-1.579***	-2.048***
	-13.17		-8.24			-6.91	-8.47
EXCESS	3.965***		1.082***			1.336***	1.610***
	-13.27		-5.97			-7.37	-8.47
PRICE	-0.062***		-0.046***			-0.070***	-0.054***
	-9.31		-5.95			-10.81	-8.14
EPS	-0.451***		-2.814***			-0.625***	-0.528***
	-5.40		-37.12			-7.32	-6.3
SDF	0.035***		0.048***			0.036***	0.032***
	-18.15		-26.18			-19.44	-16.93
SPRETURN	3.991***			-0.453*			
	-10.06			-1.77			
CD1M	-0.092***			-0.206***			
	-5.00			-11.58			
GNP	-0.000***			-0.000***			
	-4.82			-3.41			
INVEST	0.003***			0.002***			
	-6.67			-4.42			
GNPL	14.790***			21.329***			
	-6.99			-10.43			
INVESTL	-14.162***			-18.107***			
	-7.55			-10.25			
Constant	-25.048***	-5.711***	-9.985***	-33.325***	-5.999***	-11.249***	-9.381***
	-7.88	-71.78	-31.71	-10.74	-167.58	-37.91	-29.40
Variable No.	26	11	7	6	2	18	20
Log-Likelihood	-9219.57	-11660.7	-11581.8	-14020.4	-11709.3	-10414.7	-9320.56
Pseudo-R^2	0.3524	0.1810	0.1865	0.0152	0.1776	0.2685	0.3454
AUROC	0.9391	0.8784	0.8750	0.6232	0.7295	0.9147	0.9360

Table 3.9

Regressions with Different Horizons

Table 3.9 summarizes the regression using the 26-optimal variable set for lags of 1 quarter, 2 quarters, 3 quarters, 4 quarters, 5 quarters and 8 quarters. The sample period is from 1970 to 2009. For each variable, the first row reports the value of the estimated coefficient and the second row reports the value of the z- statistic. The abbreviations of variables are presented in Table 2.2, Table 2.3 and Table 2.4. Y dummy stands for the year dummy variables, AGEQ is the firm's age, and DH is default history equal to 1 if a firm has defaulted before, 0 otherwise. *denotes significant at 10%, **denotes significant at 5%, *** denotes significant at 1%.

	1Q	2Q	3Q	4Q	5Q	8Q
DH	2.833***	-1.001***	-0.704***	-0.483***	-0.277**	0.253**
	-48.3	-9.45	-6.62	-4.44	-2.52	-2.20
AGEQ	-0.004***	0.004***	0.004***	0.003***	0.001	-0.002*
	-4.57	-5.02	-4.15	-2.9	-1.36	-1.81
NITA	-4.014***	-3.685***	-1.622***	-1.276**	-0.203	0.747
	-7.22	-6.68	-2.78	-2.06	-0.31	-0.96
CASHTA	-2.588***	-1.268***	-1.229***	-1.340***	-1.369***	-1.336***
	-9.90	-4.73	-4.74	-5.26	-5.37	-5.20
CLTA	1.703***	1.870***	1.247***	0.828***	0.599***	0.371**
	-11.85	-13.14	-9.09	-6.03	-4.33	-2.45
LOGTAGNP	0.346***	0.631***	0.529***	0.501***	0.414***	0.284***
	-17.39	-30.34	-24.4	-22.25	-17.57	-10.80
SALEG	-0.720***	-0.685***	-0.651***	-0.591***	-0.264***	0.006
	-7.73	-7.34	-6.99	-6.35	-2.84	-0.06
EBITA	-6.334***	-4.597***	-5.274***	-6.433***	-3.379***	-1.022
	-7.07	-4.89	-5.27	-6.04	-2.96	-0.78
QPRID	0.110***	0.049***	0.024*	0.033**	0.070***	0.095***
	-8.34	-3.51	-1.82	-2.53	-5.80	-8.14
EBIATL	5.938***	5.267***	5.695***	2.598**	1.514	0.615
	-6.63	-5.58	-6.47	-2.54	-1.44	-0.55
CFTL	-5.821***	-5.319***	-5.634***	-2.949***	-2.256**	-2.086**
	-7.15	-6.32	-7.28	-3.22	-2.39	-2.04
QACL	0.152***	0.04	-0.024	-0.083***	-0.073***	-0.024
	-7.10	-1.46	-0.84	-2.79	-2.69	-1.08
WCTA	-1.010***	-1.366***	-0.780***	-0.163	-0.086	0.171
	-6.72	-8.65	-5.04	-1.05	-0.57	-1.13
METL	-0.089***	-0.037**	-0.029**	-0.029**	-0.022*	-0.032***
	-5.70	-2.28	-2.08	-2.24	-1.88	-2.95
FMVTMV	-0.324***	-0.475***	-0.354***	-0.292***	-0.233***	-0.165***
	-13.47	-20.25	-14.24	-11.48	-8.73	-5.77

RETURNFQ	-5.122***	-1.957***	0.602	0.085	0.197	-0.26
	-13.17	-4.93	-1.53	-0.21	-0.50	-0.61
EXCESS	3.965***	1.407***	-0.904***	-0.365	-0.371	0.311
	-13.27	-4.65	-3.02	-1.19	-1.20	-0.89
PRICE	-0.062***	-0.090***	-0.086***	-0.089***	-0.093***	-0.058***
	-9.31	-12.9	-12.91	-13.99	-15.15	-11.55
EPS	-0.451***	-0.915***	-0.993***	-0.819***	-0.928***	-0.852***
	-5.4	-9.94	-10.86	-9.23	-10.46	-9.95
SDF	0.035***	0.012***	0.005***	0.004*	0.005**	0.006**
	-18.15	-5.92	-2.64	-1.82	-2.27	-2.38
SPRETURN	3.991***	1.054***	-0.692*	-0.11	-0.491	1.501***
	-10.06	-2.67	-1.77	-0.28	-1.20	-3.27
CD1M	-0.092***	-0.042**	-0.030*	0.005	0.030*	0.101***
	-5.00	-2.36	-1.73	-0.29	-1.80	-6.14
GNP	-0.000***	-0.000***	-0.000*	0.000	0.000	0.000**
	-4.82	-4.33	-1.89	-0.23	-0.13	-2.44
INVEST	0.003***	0.002***	0.001**	0.000	0.000	-0.002**
	-6.67	-4.24	-2.11	-0.51	-0.26	-2.53
GNPL	14.790***	12.176***	10.713***	8.035***	6.866***	-1.952
	-6.99	-5.56	-4.97	-3.70	-3.10	-0.89
INVESTL	-14.162***	-9.898***	-9.395***	-7.140***	-5.994***	2.35
	-7.55	-5.12	-4.85	-3.63	-2.97	-1.14
Constant	-25.048***	-26.939***	-21.854***	-17.638***	-16.155***	-7.048**
	-7.88	-8.22	-6.95	-5.68	-5.21	-2.37
Observations	639573	631851	627852	623854	619704	607140
Log-Likelihood	-9219.57	-9763.56	-10407.2	-10762.8	-11088.6	-10910.4
Pseudo-R^2	0.3524	0.2473	0.1926	0.1543	0.119	0.0601
AUROC	0.9391	0.9100	0.8854	0.8631	0.8333	0.7537

Table 3.10

Regressions with Different Periods

Table 3.10 summarizes the regression result in different decades using the optimal variable set. The abbreviations of variables are presented in Table 2.2, Table 2.3 and Table 2.4. For each variable, the first row reports the value of the estimated coefficient and the second row reports the value of the z- statistic. Y dummy stands for the year dummy variables, AGEQ is the firm's age, and DH is default history equal to 1 if a firm defaulted before, 0 otherwise. *denotes significant at 10%, **denotes significant at 5%, *** denotes significant at 1%.

	1970-2009	1970-1980	1980-1989	1990-1999	2000-2009
DH	2.833***	3.827***	3.130***	2.784***	2.845***
	-48.3	-8.21	-20.23	-27.19	-32.8
AGEQ	-0.004***	0.272***	0.004	-0.007***	-0.003***
	-4.57	-8.12	-1.07	-4.19	-2.84
NITA	-4.014***	-8.466**	-2.059	-3.948***	-4.227***
	-7.22	-2.10	-1.51	-3.88	-5.21
CASHTA	-2.588***	1.572	-1.757***	-5.024***	-1.769***
	-9.90	-0.90	-2.74	-8.78	-4.95
CLTA	1.703***	2.497**	1.660***	1.797***	1.585***
	-11.85	-2.31	-5.22	-7.50	-6.72
LOGTAGNP	0.346***	-0.489**	0.124**	0.366***	0.371***
	-17.39	-2.33	-2.27	-9.73	-12.48
SALEG	-0.720***	0.874	-0.683***	-0.932***	-0.656***
	-7.73	-1.48	-3.53	-5.80	-4.31
EBITA	-6.334***	-4.865	-8.821***	-6.794***	-5.856***
	-7.07	-0.54	-4.17	-4.21	-4.43
QPRID	0.110***	-0.124	0.110***	0.138***	0.096***
	-8.34	-1.34	-3.36	-5.79	-4.84
EBIATL	5.938***	7.021	4.527**	5.993***	6.727***
	-6.63	-1.12	-2.31	-3.39	-5.24
CFTL	-5.821***	-6.193	-1.775	-6.020***	-7.513***
	-7.15	-1.05	-0.99	-3.74	-6.46
QACL	0.152***	0.238*	0.154***	0.145***	0.164***
	-7.10	-1.70	-3.36	-3.80	-4.90
WCTA	-1.010***	-1.860**	-1.441***	-0.975***	-1.018***
	-6.72	-2.09	-4.46	-3.97	-3.95
METL	-0.089***	0.099	-0.04	-0.092***	-0.145***
	-5.70	-1.06	-1.33	-3.1	-5.27
FMVTMV	-0.324***	-0.179	-0.243***	-0.359***	-0.283***
	-13.47	-0.82	-4.25	-8.39	-7.38
RETURNFQ	-5.122***	9.421***	-1.763*	-3.422***	-8.311***
	-13.17	-3.35	-1.72	-4.94	-14.71
EXCESS	3.965***	-7.910***	0.46	2.493***	6.826***
	-13.27	-3.31	-0.61	-4.71	-15.56
PRICE	-0.062***	0.056**	-0.032**	-0.073***	-0.085***
	-9.31	-2.17	-2.37	-5.86	-6.76
EPS	-0.451***	0.309	-0.423**	-0.445***	-0.513***
	-5.40	-0.65	-2.34	-2.94	-3.73
SDF	0.035***	-0.018	0.019***	0.029***	0.046***
	-18.15	-0.93	-4.34	-8.76	-15.04

SPRETURN	3.991***	-8.456**	1.369	3.025***	6.784***
	-10.06	-2.45	-1.35	-3.97	-11.08
CD1M	-0.092***	-0.133	-0.033	-0.136	-0.091***
	-5.00	-0.65	-0.81	-1.36	-3.18
GNP	-0.000***	0.007	0.002	0.000	0.000
	-4.82	-0.50	-0.72	-0.05	-0.22
INVEST	0.003***	-0.179	-0.016	0.009	0.007***
	-6.67	-0.68	-0.75	-0.71	-3.11
GNPL	14.790***	-48.649	-3.56	-1.493	-9.115
	-6.99	-0.75	-0.13	-0.04	-0.28
INVESTL	-14.162***	88.585	1.756	-22.069	-25.349***
	-7.55	-0.72	-0.08	-0.80	-2.85
Constant	-25.048***	-38.658	-2.462	51.589	94.664
	-7.88	-0.37	-0.06	-0.69	-0.94
Observations	639573	54279	168691	221076	195527
Log-Likelihood	-9219.57	-280.24	-2045.71	-3126.16	-3496.07
Pseudo-R^2	0.3524	0.3545	0.2891	0.3567	0.4103
AUROC	0.9391	0.9683	0.9079	0.9526	0.9590

.

Table 3.11

Regressions with Industry Effects

Table 3.11 shows the regression results of industry effects for public and private companies with different horizons. The sample period is from 1970 to 2009. The abbreviations of variables are presented in Table 2.2, Table 2.3 and Table 2.4. For each variable, the first row reports the value of the estimated coefficient and the second row reports the value of the z- statistic. Y dummy stands for the year dummy variables, AGEQ is the firm's age, and DH is default history equal to 1 if a firm has defaulted before, 0 otherwise. *denotes significant at 10%, **denotes significant at 5%, *** denotes significant at 1%.

	Public 1Q	Private 1Q	Public 2Q	Public 4Q	Public 8Q
DH	2.804***	3.347***	-1.051***	-0.544***	0.186
	-47.61	-62.65	-9.90	-4.99	-1.61
AGEQ	-0.004***	-0.006***	0.004***	0.003***	-0.002
	-4.56	-7.82	-5.05	-2.96	-1.60
CD1M	-0.092***	-0.126***	-0.042**	0.004	0.100***
	-4.96	-6.93	-2.34	-0.26	-6.10
GNP	-0.000***	-0.000***	-0.000***	0.000	0.000**
	-4.92	-3.01	-4.39	-0.38	-2.31
INVEST	0.004***	0.002***	0.002***	0.000	-0.002**
	-6.76	-4.48	-4.33	-0.69	-2.36
GNPL	14.768***	15.503***	11.988***	8.134***	-1.767
	-6.98	-7.33	-5.47	-3.74	-0.80
EPS	-0.428***		-0.860***	-0.785***	-0.832***
	-5.13		-9.34	-8.85	-9.70
INVESTL	-14.153***	-14.337***	-9.824***	-7.262***	2.14
	-7.54	-7.73	-5.08	-3.69	-1.04
SPRETURN	3.952***	0.053	1.027***	-0.159	1.452***
	-9.95	-0.20	-2.60	-0.40	-3.16
NITA	-4.038***	-6.405***	-3.769***	-1.354**	0.638
	-7.26	-12.33	-6.81	-2.18	-0.82
FMVTMV	-0.325***		-0.481***	-0.295***	-0.167***
	-13.48		-20.29	-11.54	-5.78
RETURNFQ	-5.098***		-1.938***	0.121	-0.229
	-13.09		-4.88	-0.30	-0.54
CASHTA	-2.350***	-2.448***	-0.980***	-1.121***	-1.176***
	-8.88	-9.46	-3.61	-4.35	-4.52
SDF	0.035***		0.012***	0.004*	0.006**
	-18.09		-5.77	-1.75	-2.43
CLTA	1.655***	2.151***	1.805***	0.766***	0.312**
	-11.35	-14.59	-12.49	-5.50	-2.04
LOGTAGNP	0.342***	0.108***	0.634***	0.500***	0.280***
	-17.11	-8.16	-30.07	-21.98	-10.55
SALEG	-0.723***	-0.812***	-0.690***	-0.603***	-0.013
	-7.76	-8.91	-7.37	-6.47	-0.13
EBITA	-6.129***	-2.919***	-4.427***	-6.319***	-0.939
	-6.82	-3.36	-4.68	-5.91	-0.71
QPRID	0.111***		0.052***	0.034***	0.095***

	-8.43		-3.73	-2.64	-8.15
EXCESS	3.952***		1.402***	-0.388	0.29
	-13.22		-4.63	-1.26	-0.83
PRICE	-0.063***		-0.092***	-0.090***	-0.058***
	-9.45		-13.11	-14.11	-11.54
EBIATL	5.838***	4.013***	5.174***	2.535**	0.568
	-6.51	-4.32	-5.46	-2.46	-0.5
CFTL	-5.831***	-6.018***	-5.369***	-2.974***	-2.079**
	-7.15	-7.08	-6.35	-3.23	-2.02
QACL	0.161***	0.215***	0.050*	-0.077**	-0.017
	-7.37	-10.51	-1.77	-2.55	-0.75
WCTA	-1.182***	-1.302***	-1.561***	-0.315**	0.048
	-7.63	-8.56	-9.63	-1.98	-0.31
METL	-0.091***	-0.279***	-0.035**	-0.027**	-0.032***
	-5.76	-14.37	-2.16	-2.11	-2.88
RIND2	-0.373***	-0.392***	-0.679***	-0.599***	-0.539***
	-4.70	-5.05	-8.8	-7.80	-6.71
RIND3	-0.241***	-0.308***	-0.545***	-0.515***	-0.522***
	-3.30	-4.32	-7.68	-7.32	-7.15
RIND4	-0.743***	-0.882***	-1.129***	-0.990***	-0.890***
	-5.66	-6.78	-8.65	-7.51	-6.36
Constant	-24.697***	-23.082***	-25.958***	-17.175***	-6.681**
	-7.77	-7.29	-7.93	-5.52	-2.24
AUROC	0.9394	0.9199	0.9114	0.8657	0.7592
Observations	639573	639573	631851	623854	607140
Log-Likelihood	-9199.15	-9805	-9711.56	-10722.4	-10879
Pseudo-R^2	0.3539	0.3113	0.2513	0.1574	0.0628

Table 3.12

Accuracy and Error Scenarios for Default Classifications

Table 3.12 shows the default classification scenarios.

		Actual	
		Default firm	Healthy firm
Predicted	High default risk	Correct prediction (Sensitivity)	Type II error
	Low default risk	Type I error	Correct prediction (Specificity)

Table 3.13

In-sample and Out-of-sample Classification Rates

Table 3.13 presents the classification rate of in-sample and out-of-sample tests. The in-sample test is based on data from 2000 to 2009. The out-of-sample test using the data from 1970 to 2009 predicts the performance of 2000 to 2009. Sensitivity is the rate of the correctly classified default firms. Specificity is the rate of correctly classified non-default firm as.

	Out-of-sample					In-sample				
Cut off points	Sensitivity (%)	Specificity (%)	Type II error (%)	Type I error (%)	Accuracy rate (%)	Sensitivity (%)	Specificity (%)	Type II error (%)	Type I error (%)	Accuracy rate (%)
0.0001	99.89	11.34	88.66	0.11	11.76	99.68	28.60	71.40	0.32	28.94
0.0005	98.93	46.46	53.54	1.07	46.71	98.40	56.33	43.67	1.60	56.53
0.001	97.75	62.90	37.10	2.25	63.06	97.65	68.60	31.40	2.35	68.74
0.005	88.88	86.84	13.16	11.12	86.84	91.02	89.02	10.98	8.98	89.03
0.01	81.18	92.09	7.91	18.82	92.03	84.71	93.76	6.24	15.29	93.72
0.05	53.80	98.48	1.52	46.20	98.27	58.07	98.65	1.35	41.93	98.46
0.1	42.57	99.27	0.73	57.43	99.00	43.32	99.36	0.64	56.68	99.09
0.2	31.02	99.68	0.32	68.98	99.36	30.59	99.72	0.28	69.41	99.39
0.3	22.46	99.83	0.17	77.54	99.46	22.25	99.85	0.15	77.75	99.47
0.4	14.22	99.91	0.09	85.78	99.50	15.08	99.91	0.09	84.92	99.50
0.5	6.84	99.96	0.04	93.16	99.52	10.05	99.95	0.05	89.95	99.52
0.6	3.42	99.98	0.02	96.58	99.52	6.10	99.98	0.02	93.90	99.53
0.7	1.28	99.99	0.01	98.72	99.52	3.10	99.99	0.01	96.90	99.53
0.8	0.11	100.00	0.00	99.89	99.52	0.96	100.00	0.00	99.04	99.52
0.9	0.00	100.00	0.00	100.00	99.52	0.11	100.00	0.00	99.89	99.52

Table 3.14

AUROC Estimation of In-sample and Out-of-sample Tests

Table 3.14 illustrates details of the in-sample and out-of-sample AUROC. The out-of-sample is based on the sample of 1970 to 1999.

	Observation	Area	Std. Err.	[95% Conf. Interval]	
In-sample (1970-2009)	639573	0.939	0.0031	0.93309	0.94516
In-sample (2000-2009)	195527	0.959	0.0031	0.95287	0.96510
Out-of-sample (2000-2009)	195527	0.949	0.0034	0.94208	0.95523
Rolling out-of-sample (2000-2009)	195527	0.952	0.0032	0.94595	0.95864

Figure 3.1

Feature Selection Procedure

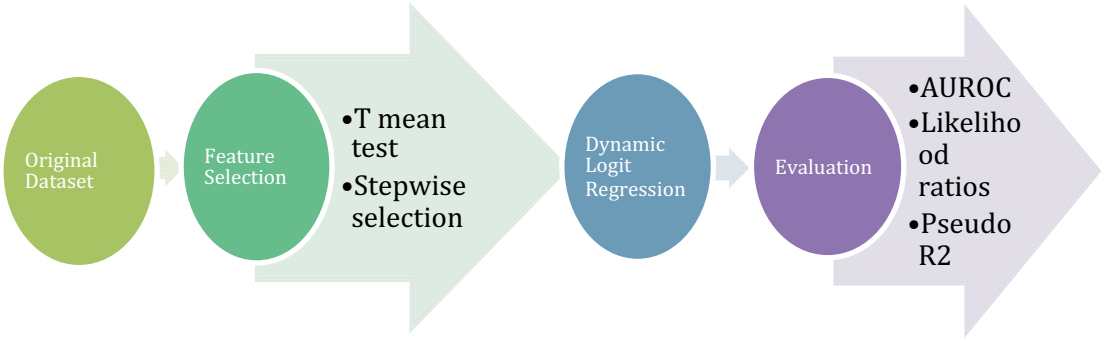


Figure 3.1 illustrates the feature selection procedure. The original dataset is delimited using both t-test and stepwise regression based on the dynamic logit model. The optimal variable set is decided with the evaluation indicators including AUROC, likelihood ratios and Pseudo R^2.

Figure 3.2

Shape of Logistic Function

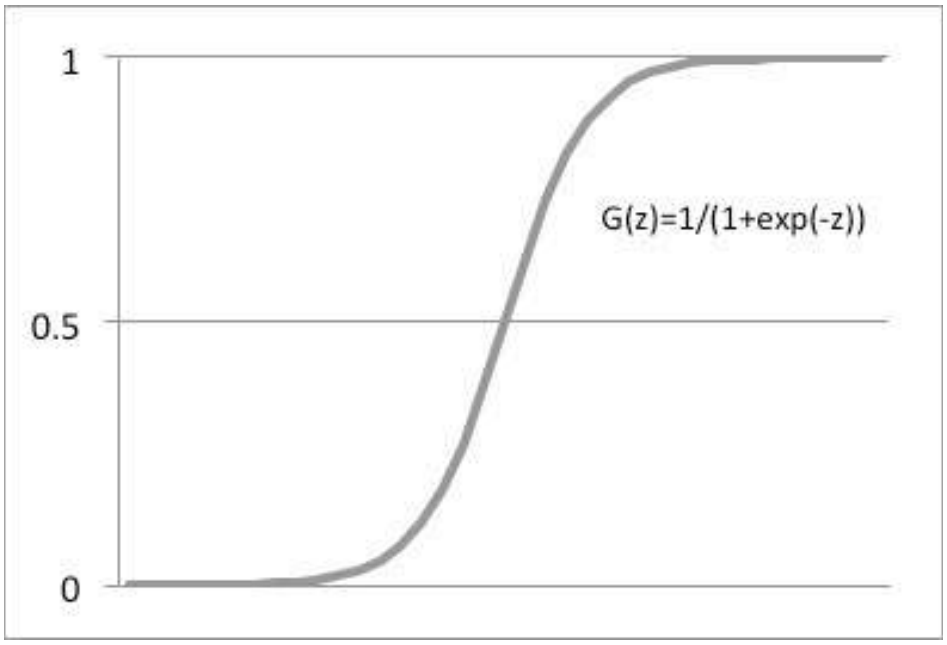


Figure 3.2 describes the S shape of the logistic function.

Figure 3.3

Default Number Dropping Process

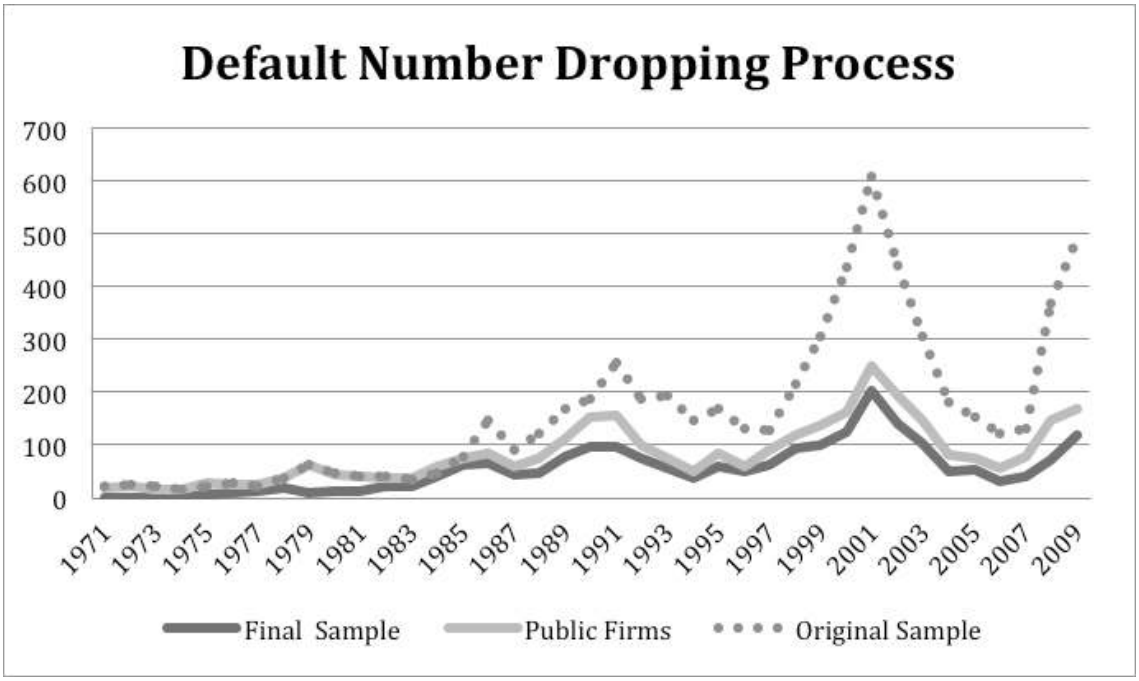


Figure 3.3 presents changing of the total number of defaults in the United States due to the shortage of firm-level information.

Figure 3.4

Yearly Default Numbers

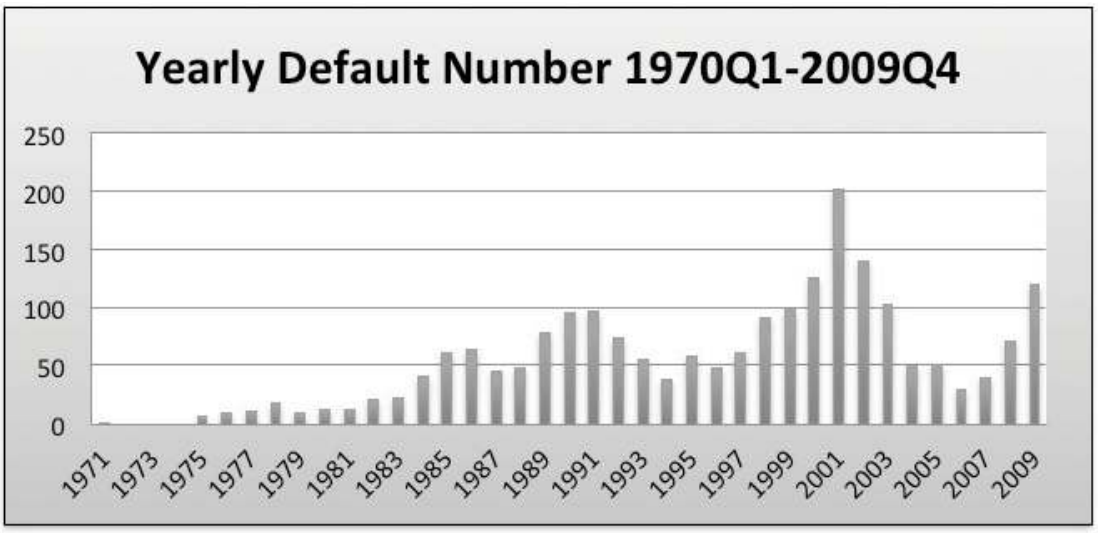


Figure 3.4 presents the total number of defaults in the United States each year. The peaks of the default numbers appear in the early 90s, early 00s and late 00s.

Figure 3.5

Quarterly Default Rates

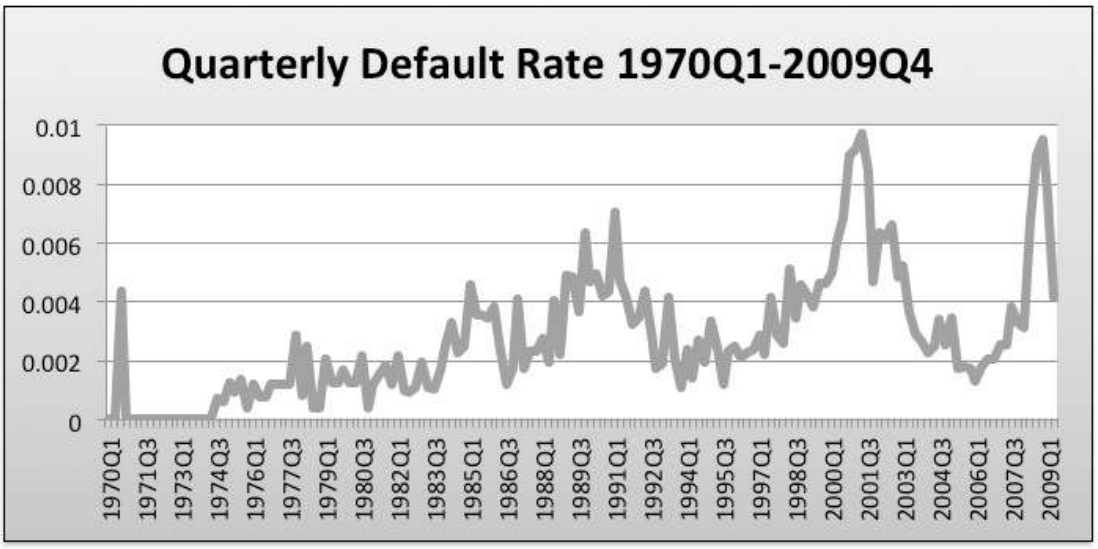
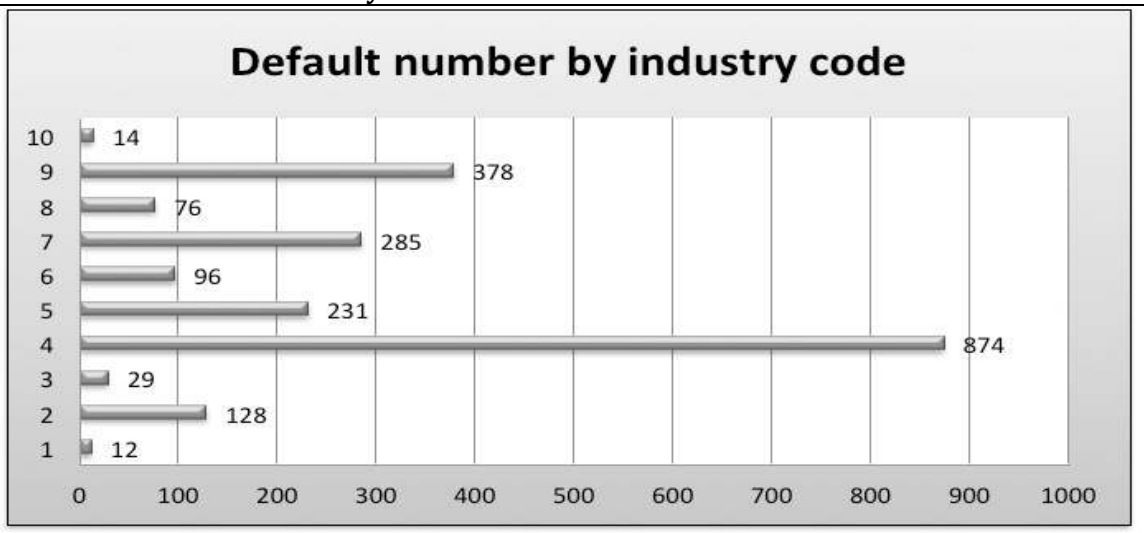


Figure 3.5 describes the quarterly default rates from 1970 to 2009.

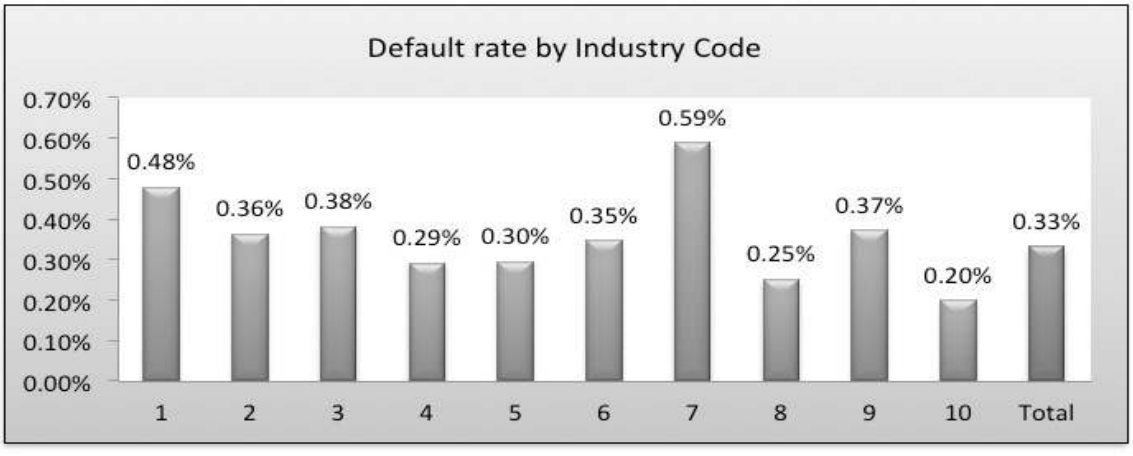
Figure 3.6

Default Information by Industry

PANEL A: Default Number by Sectors 1970 - 2009



PANEL B: Default Rate of Active Firms by Sectors 1970-2009



PANEL C: Industry Codes, Default Numbers and Active Firms

Industry Code	SIC Code	Industry Name	Default Number	Active Firms
1	<1000	Agriculture, Forestry, And Fishing	12	2,509
2	1000 to1499	Mining	128	35,179
3	1500 to 1799	Construction	29	7,616
4	2000 to 3999	Manufacturing	874	301,213
5	4000 to 4999	Transportation, Communications, and Utilities	231	78,272
6	5000 to 5199	Wholesale Trade	96	27,722
7	5200 to 5999	Retail Trade	285	48,348
8	6000 to 6799	Finance, Insurance, And Real Estate	76	30,196
9	7000 to 8999	Services	378	101,442
10	>=9100	Public Administration	14	7,076
TOTAL			2,123	639,573

Figure 3.6 presents the default numbers and default rates by industry from 1970 to 2009. Panel A describes the default numbers by industry codes. Panel B presents the default rates by industry codes. Panel C gives the references of the industry codes, default numbers and active firms in each industry.

Figure 3.7

Significance Levels of Selected Independent Variables

Figure 3.7 gives the significance tier of the default drivers, selected by the backward stepwise method. The abbreviations of variables are presented in Table 2.2, Table 2.3 and Table 2.4. The most significant default drivers are the default history and the net income ratio. Variables in dark grey are included in the optimal variable set.

Figure 3.8

Correlation Pairs Between Variables

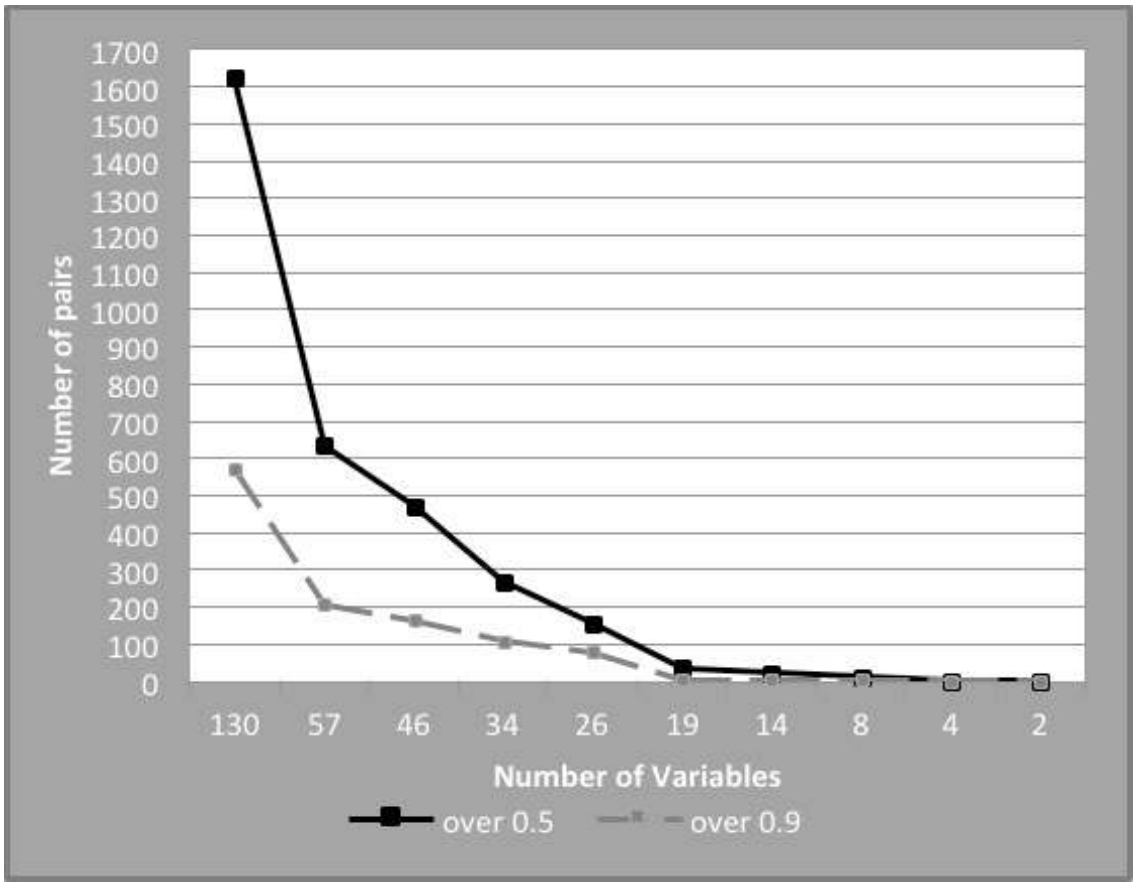


Figure 3.8 presents the change in the number of pairs of highly correlated variables under the feature selection process using the stepwise method. The black line is the change of the numbers of pairs with correlation over 0.5. The grey line shows the change of the numbers of pairs with correlation over 0.9.

Figure 3.9

Predicted vs. Actual Default Rates

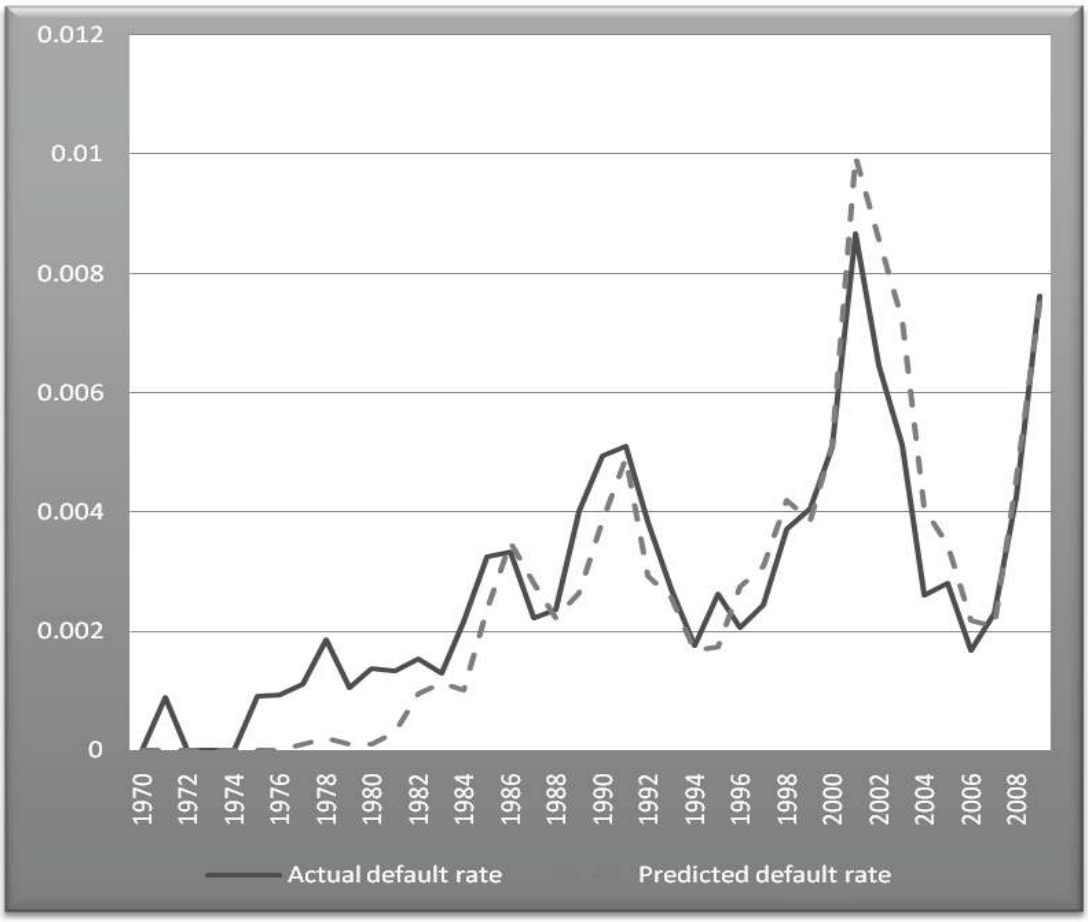


Figure 3.9 shows both the predicted and actual default rates for all firms from 1970 to 2009. Predicted defaults are calculated using fitted values of the optimal variable set (26 variables) with the logit model.

Figure 3.10

AUROC of In-sample and Out-of-sample Tests

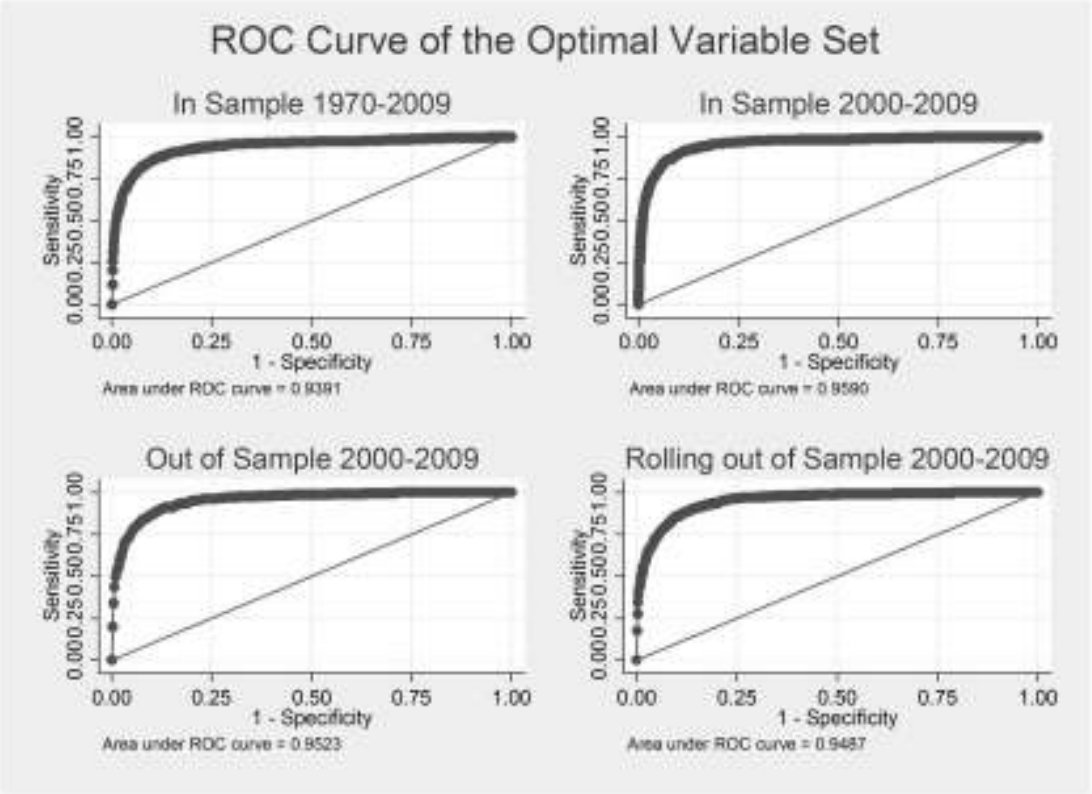


Figure 3.10 charts the area under the receiver operation curves. The AUROC of in-sample test from 1970 to 2009 is 0.9391. The AUROC of in-sample test from 2000 to 2009 is 0.959. The AUROC of out-of-sample test from 2000 to 2009 using the sample of 1970 to 1999 is 0.9523. The AUROC of rolling out-of-sample test from 2000 to 2009 is 0.9487.

Chapter 4: Corporate Default Prediction with Random Effects

4.1 Introduction and Background

This chapter investigates the default cluster and unobservable default factors using advanced random econometric models. Historically, corporate defaults are known to cluster in time (Figure 3.4 and Figure 3.5). The default rate was high in the early 1990s, early 2000s and late 2000s. Also, the historical default rates differ in industries (Figure 3.6). As presented in Chapter 3, the retail industry has the highest default rate, while default credit risk is found to be low in the public administration sector. Furthermore, Couderc and Renault (2005), Hol (2007) and Koopman, Kraussl, Lucas & Monteiro (2009) all address the relationship between credit cycle and business cycle. Das, Hanouna & Sarin (2009) summarise three explanations for default clusters: common risk factors, direct default contagion and generated default correlation. Although the phenomena and potential explanation of default clusters are addressed, the theoretical methods and empirical studies of correlated corporate defaults are quite recent and limited.

On the one hand, most previous studies using traditional models do not incorporate the cluster factors in the default predictions. On the other hand, existing random models are based on limited observable factors. Two classes of models have been suggested recently to capture default clusters: the advanced discrete model (Jones & Hensher, 2004; Hensher & Jones, 2007; Jones & Hensher, 2007) and frailty models (Das, Duffie, Kapadia & Saita, 2007; Duffie, Saita & Wang, 2007; Duffie, Eckner, Horel & Saita, 2009). Although the significance of the frailty effects and the unobservable factors are shown in both models, there are two gaps in these studies in terms of optimal selection of explanatory variables.

First, in the empirical tests of advanced discrete models (Jones & Hensher, 2004; Hensher & Jones, 2007; Jones & Hensher, 2007), market variables are missing in explaining the default clusters. A recent study (Chava & Jarrow, 2004) suggests that market information has greater explanatory ability for default behaviour. The results in Chapter 3 of this thesis also confirm that the combination of information from different categories predicts default risk more accurately and efficiently. Therefore, it is worth exploring the significance of random factors when adding market variables into the mixed logit model.

Secondly, a flaw in existing literature is that accounting variables are absent in previous frailty empirical tests (Das, Duffie, Kapadia & Saita, 2007; Duffie, Saita & Wang, 2007 Duffie, Eckner, Horel & Saita, 2009). In other words, previous frailty tests limit the observable variables within the macroeconomic and stock market categories. In an inefficient financial market, these variables would not necessarily reveal all the available predictive information. Random components have been shown to be significant in those studies, which could suggest that previous studies may have missed some variables with relevant the explanatory information. The omitted explanatory variables, however, could be composed of observable accounting information instead of unobservable information, as Koopman, Lucas & Schwaab (2010) suggest. Whether the accounting ratios contribute to the explanation of default probability in the survival analyses remains unanswered. Most importantly, it is unknown whether frailty factors remain significant when accounting variables are added. Furthermore, previous studies of both methods fail to incorporate default history indicators. In summary, advance models with limited observable variables are vulnerable to omitted variable bias and under-performance predicative power.

The primary objective of this chapter is to check the significance of the dynamic random factors in both the mixed logit models and the frailty models, conditioning on the optimal information set established in Chapter 3, which contains a wider range of observable covariates including both firm-level and macroeconomic-level information. The

significance of the random factors will explain the existence of unobservable variables beyond the optimal information set. The additional objective of this chapter is to compare the prediction performance between the two advance models, the mixed logit model and the frailty model, and to evaluate their abilities to capture the unobservable default information.

To test these two models, this chapter applies the quarterly sample reported in Chapter 3, containing 2,123 default events in the United States from 1970 to 2009. This chapter first examines the performance of the mixed logit models across different variable sets; in particular, it compares the prediction performance using the mixed logit model for both public and private firms. The chapter then looks more deeply at the optimal variables set with different simulation times, and also with the different coefficient assumptions. The following set of tests in this chapter examines the default behaviours with the proportional Cox survival models. It looks carefully for different observable information sets, starting with non-frailty survival models first. Most importantly, this chapter adds the common accounting ratios from traditional predictive methods, such as Z-score and O-score. Conditioning on the optional variable set, this chapter explores the default clusters in alternative dimensions: the industry-related cluster, the macroeconomic-related cluster and the calendar-related cluster. Finally, the chapter compares the prediction efficiency and accuracy for the random coefficient mixed logit model and the shared frailty model.

The main contributions of this chapter are fourfold. The first contribution is to show the significance of the random components in both the mixed logit model and the frailty model conditioning on the optimal information set. A significant result indicates that there are missing observable variables in previous empirical studies with the advanced random models. The results also suggest that there are unobserved components beyond the suggested optimal information set. For the mixed logit model, this chapter has shown the significance of the market and macroeconomic variables, which are missing as observable variables in previous studies. This evidence is consistent with the argument from Kalotay

(2007).

For the frailty model, this chapter has found that the accounting variables ignored in previous tests add additional explanatory power for default prediction. The results give further evidence of market inefficiency in default prediction, since the prediction ability for both advanced random models significantly improves with a wider variable set than only market variables. Beyond the optimal variable set, the significance of the random factors has been found in both models. These random parts capture additional information efficiently and thus improve prediction ability for default forecasting models. Those random variables presumably contain additional information from other accounting, market, or macroeconomic variables that are not included in the regression for correlation reasons. Also, the random factors model information that is difficult to collect and quantify, but is presumably important for default prediction. As suggested by Nwogugu (2007), these unobservable information could include: the reactions of the company's creditors and customers, changes in inventory valuation method, pension liabilities accounting, research and development expenditures; auditors' qualified opinions, the type and characteristics of creditors; the company's operating cash cycle, the existence of union labour dispute and major environmental liability and industry-related variables. Among these potential unobservable factors, three dimensions of clusters are verified particularly using the frailty models: industry difference, calendar period and macroeconomic conditions.

The second main contribution of this chapter is to provide the first empirical test to investigate the application of the mixed logit model in the United States. The mixed logit model shows considerably greater prediction power in capturing default risk over the standard logit model with the optimal variable set. Moreover, the tests with variables from different categories verify the preference of the optimal variable set developed in Chapter 3. The present chapter also verifies the stability and robustness of the mixed logit model with alternative simulations associated with the random replications, as well as both the

independent coefficient assumption and the correlated coefficient assumption. Industry effects are also revealed to be significant in the mixed logit model test.

The third contribution of this chapter is to improve the prediction ability of the Cox survival models by incorporating the optimal information set developed in Chapter 3. It demonstrates the significance of the accounting information and default history in default prediction with both the Cox survival analysis and the frailty models. Moreover, this chapter confirms the significance of industry effects using the Cox survival analysis.

Finally, this chapter provides the first empirical and theoretical comparison of the advanced random models from different families in default predictions. While the mixed logit model offers greater prediction accuracy in terms of the AUROC and classification rates, the frailty model suggests the potential unobservable dimensions and computes the default risk with greater efficiency.

The chapter is arranged as follows. Section 2 describes the numerical methods for both the mixed logit model and the frailty model. Section 3 presents the regression results: firstly from alternative mixed logit models in terms of variable combination, simulation times, and the assumptions of the coefficients; and secondly from the Cox survival models, with a comparison of frailty and non-frailty methods and investigation on frailty factors. Section 4 compares the two random models. Section 5 draws conclusions.

4.2 Methodology

Two approaches to modelling cluster defaults are described in this section. In the first approach, the mixed logit model, default cluster and unobservable information is modelled by random coefficients. In the second approach, the frailty model, the association between failure times is explicitly modelled with a random-effect term, normally called the “frailty”. These frailties stand for unobserved default variables shared by all firms of the same group. This section describes these two methods as used in the mentioned tests.

4.2.1 Mixed Logit Model

Mixed or random parameter ordered logit method is among a new breed of econometric models developed out of discrete choice theory (Train, 2003). It is considered the most promising discrete choice model (Hensher & Green, 2001). The development of the simulation method has led to the application of the mixed logit model and its variants, supplanting simpler models in many areas of economics, marketing, management, transportation, health, housing, energy research, and environmental science (Train, 2003).

The mixed logit model was first introduced to predict default risk by Jones & Hensher (2004) using a sample dating from 1996 to 2000 of firms in Australia. While the standard logit model presented in Chapter 3 limits the estimation to observed variables, the mixed logit model incorporates random effects more appropriately, and relaxes the assumption of IIA. The mixed logit model also accounts for correlation in unobserved factors over repeated choices, by taking useful information contained in the error part in the standard logit model (Train, 2003).

The mixed logit probability is a weighted average of the logit formula, evaluated at a different value of the coefficients β (Train, 2003). The most notable difference between the mixed logit model and the standard logit model is that the coefficients of the mixed logit model vary in population, while they are fixed in the standard logit model. This feature helps to capture unobserved factors that could influence default risk. The usual form of the mixed logit model is an integral of standard logit probabilities over a density of parameters:

$$P_{ni} = \int \left(\frac{e^{\beta' x_{ni}}}{e^{\beta' x_{nj}}}\right) f(\beta|b, W) d\beta \tag{9}$$

where $f(\beta|b,W)$ is the density distribution of random parameters, b is the mean of the

varying coefficients β_n, and W is the covariance of β_n. The density of β is assumed as normally distributed, following Jones & Hensher (2004).

The trigger for the wider application of the mixed logit model has been the development of simulation methods, which enable this open-formed model to be solved, and the random parameters to be estimated. In this thesis, the simulated maximum likelihood method is used to estimate the parameters. To estimate the probability, the first step is to draw the coefficients β from the normal distribution $f(\beta|\theta)$ using a Halton sequence, which is a method to generate points using a prime number as its base. For example, the Halton sequence for the prime number two is (1/2,1/4,3/4,1/8,3/8,5/8,7/8,1/16,3/16…) and the Halton sequence for the prime number three is (1/3,2/3,1/9,2/9,4/9,5/9,7/9,1/27…). The number of the Halton sequence is decided by the number of random coefficients. The length of each sequence is the product of the number of observations and the numbers of draws that the regression uses. For the random coefficient assuming as normal distribution, the coefficients are the inverse cumulative normal of each element of each sequence. Using the Halton sequence method could save considerable computational time and generate fewer errors compared to random draws (Train, 1999). The number of total observations and the replication times determine the length of the sequence. The second step is to calculate a logit probability conditional on β_n, which is:

$$L_{ni}(\beta_n) = \frac{e^{\beta X_{ni}}}{\sum_j e^{\beta X_{nj}}} \tag{10}$$

Repeating steps one and two many times, and averaging the results, obtains the unbiased simulated approximation to the actual probability satisfying $P_{ni} = E(\widetilde{P_{ni}})$ that:

$$\widetilde{P_{ni}} = \frac{1}{R}\sum L_{ni}(\beta^r) \qquad \gamma = 1,2 \dots R \tag{11}$$

where β^r is the γth draw from $f(\beta|\theta)$. The log-likelihood of the model is given by $LL(\theta) = \sum_{n=1}^{N} \ln P_n(\theta)$, and then the simulated log-likelihood is given by

$$SLL = \sum_{n=1}^{N} \ln \left(\frac{1}{R} \sum_{r=1}^{R} L_{ni}\right) \tag{12}$$

where R is the number of replications. The estimated parameters θ are those that could maximize the value of SLL. It is noteworthy that, although the simulated probability $\widetilde{P_{ni}}$ is an unbiased estimation for P_{ni}, the $Ln\widetilde{P_{ni}}$ is a biased estimation, since log operation is not a linear transformation. One method to diminish the bias in $Ln\widetilde{P_{ni}}$ in order to make the simulated Maximum Likelihood Estimation more accurate is to use more draws in the simulation (Train, 2009).

4.2.2 Frailty Model

This section presents the hazard models used for the subsequent empirical tests. For a random time t to default T, the probability density function of T is defined as $f(t)$ and the cumulative distribution function, or failure function, as $F(t) = P(T < t)$. In the survival analysis, this chapter also denotes the survivor function $S(t) = P(T > t) = 1 - F(t)$, and the hazard function $h(t) = f(t)/S(t)$, which can be interpreted as the instantaneous rate of failure, given survival up until time t. The relationship between these four functions is presented below:

$$h(t) = \frac{f(t)}{1-F(t)} = \frac{f(t)}{S(t)} = \frac{\partial H(t)}{\partial t} = \frac{\partial\left\{-\ln\left[S(t)\right]\right\}}{\partial t} \tag{13}$$

where $H(t)$ donates the cumulated hazard function.

To investigate the correlation between variables, this chapter explores the frailty model that the subjects occur in groups $j = 1, 2, \ldots J$. Conditional on the unobserved individual frailty v_j, the hazard intensity is given by

$$h_{ij}(t \mid v_j) = v_j h_{ij}(t \mid x_{ij}) = h_0(t) \exp(\beta_x x_{ij} + v_j) \tag{14}$$

where x_{ij} can be a mixture of firm-specific time-independent variables (accounting and stock market factors, default history and sector dummies) and macroeconomic time-dependent variables. The accounting ratios and stock market variables are unique for every firm, while the macroeconomic variables are common for all firms at a central time point. $h_0(t)$ is the unspecified baseline hazard function common to all firms. The term $\exp(\beta_x x_{ij} + \nu_j)$ allows the expected time to default to vary across firms, according to their observable covariates x_{ij} with the estimated coefficients $\hat{\beta}_x$ and unobservable covariates ν_j. The random effect (frailty) ν_j accounts for the within-subject correlation due to some unobserved common covariate information. If ν_j is greater than 1, then the specified firm fails faster than an average subject during that period. On the other hand, the firm is less likely to default if ν_j is less than 1. Each group has different values of random effect, and the variability in the frailty term reflects the heterogeneity of default risk between subjects. The frailty part is assumed to be independent and identically distributed and to follow a given statistical distribution. This chapter uses a gamma distribution with mean equal to 1 and unknown variance θ for computational convenience and convergence following (Chuang, Cai, Douglass, Wei & Dodson, 2005). When the frailties follow a gamma distribution, the likelihood of the ith group can be expressed compactly as

$$L_i = \left[\prod_{j=1}^{n_i} \left\{ h_{ij}\left(t_{ij}\right) \right\}^{d_{ij}} \right] \frac{\Gamma\left(1/\theta + D_i\right)}{\Gamma\left(1/\theta\right)} \theta^{D_i} \left\{ 1 - \theta \sum_{j=1}^{n_i} \ln \frac{S_{ij}(t_{ij})}{S_{ij}(t_{0ij})} \right\}^{-1/\theta - D_i} \qquad (15)$$

The survival times are measured in quarters.

4.3 Results

This section reports the results from the random models based on the sample described in Chapter 3. The standard logit model is also presented as a reference for the mixed logit models, and the standard Cox model is given as the benchmark for the frailty models.

4.3.1 Comparison between Logit and Mixed Logit Models

[Table 4.1, p. 154 about here]

Table 4.1 presents the estimated coefficients and z values for the mixed logit models (Models 2,3,4,5) given the standard dynamic logit model (Model 1) as a reference. The optimal variables set presented in Chapter 3 with industry effects is used to compare the overall prediction ability between the logit model (Model 1) and mixed logit model (Model 2). The mixed logit model shows a very strong overall prediction ability. Comparing the normal logit model in Model 1, the prediction ability of Model 2 is a considerable improvement, as the log-likelihood ratio increases from -9199.15 to -5566.84. One explanation of this improvement is that the mixed logit model captures more information. In the mixed logit results presented in Table 4.1, some variables have a single fixed coefficient while other variables have two parameters (mean and standard deviation) representing their coefficients.

The estimated mean value of the mixed logit model (Model 2) is broadly similar to that obtained with the standard logit model (Model 1). Generally speaking, the default history (DH) remains a core position in predicting default in the mixed logit model, as DH associates with the biggest z value in the mean parameter column. Firms are more likely to default if they have had previous financial distress. Industry effects have shown significance at the 1% level in Model 2. Macroeconomic variables have also overall shown statistical significance in the mixed logit test. On the other hand, two obvious differences between the estimated coefficients of the normal logit model and the estimated mean parameters of the mixed logit model are: the Quick ratio (QACL) which lost its explanatory power in the mixed logit model; and the coefficient of firm age (AGEQ) which changes sign from positive to negative, which suggests that an old firm is more likely to default.

Besides QACL, nine variables show as statistically significant in the unobservable

heterogeneity represented by standard deviations. This result shows the existence of unobservable variables conditioning on the optimal information set. A larger random coefficient value (S.D. value) is associated with a higher possibility of influence of the random parameter. The Cash ratio (CASHTA) has the largest random coefficient, followed by CLTA, WCTA and EBITA. The formula used to interpret the coefficient distribution for random variables is as follows:

$$\text{Estimated mean} + \text{Estimated standard deviation(S.D.)}*\text{N}(0,1)$$

$\text{N}(0,1)$ stands for the standard normal distribution. For example, with model 4 in Table 4.1, the default probability could be calculated as follow:

$$\begin{aligned} P = {} & [-3.205 + 3.721 * N(0,1)] * EPS - [0.391 - 0.864 * N(0,1)] * FMVTMV - \\ & [1.904 - 0.067 * N(0,1)] * RETURNQ + [0.04 + 0.033 * N(0,1)] * SDF + \\ & [0.05 + 0.084 * N(0,1)] * QPRID + [1.762 + 0.114 * N(0,1)] * EXCESS - \\ & [0.612 - 0.408 * N(0,1)] * PRICE \end{aligned} \tag{16}$$

It requires random draws from the standard normal distribution $\text{N}(0,1)$. It is worth noticing that, the average results of the random draws, with a large number of replication, would converge to the mean values of the coefficients of each variable.

Compared with the empirical test in Jones & Hensher (2004), this chapter tests the mixed logit method in Model 2 with a wider range of observable variables, which include not only the accounting information and industry effects mentioned in (Jones & Hensher, 2004), but also market variables, macroeconomic factors and other variables such as DH and AGEQ.

Further sector analyses are conducted for the mixed logit models. Model 3 contains only information from the financial statements as explanatory variables. Model 4 is based on the variables from the stock market. Comparing the log-likelihood ratios of Model 3 and Model 4, the market variables-based mixed logit model slightly outperforms the

accounting variables-based model. However, the significance of the accounting variables shows that market information does not cover all the available information which was interpreted as evidence of market inefficiency. This result is consistent with the findings of Chava & Jarrow (2004), but this chapter departs from their study by using the mixed logit model.

It is important to notice that some random parameters of accounting and market variables, including NITA, SDF, EXCESS, QPRID and CFTL, show significance in the single sector analyses (Models 3 & 4) but insignificance in the whole variable set (Model 2). This is an important finding that the observable random variables revealed in Jones & Hensher's study (2004) could either be macroeconomic variables or market variables. With a limited observable variables set in a mixed logit test, as carried out by Jones & Hensher (2004), the result could be less convincing. Comparing Model 2 with Models 3 and 4, this section concludes that the optimal variable information set largely improves the prediction ability for the mixed logit model and could be recommended for further application.

Finally, Model 5 in Table 4.1 contains the mixed logit model regression result for private firms. According to the log-likelihood ratios (-6126.6), the mixed logit model performs well for private firms, although it does not outperform public firms with the optimal information set. The S&P 500 index return (SPRETURN) and LOGTAGNP lost significance in the private firm regression, while the QACL ratio shows significance at 10% level. Comparing Model 2 and Model 5, the values of the standard deviations in CLTA, EBITA, CFTL and METL increase in the private firm test, while standard deviations of other variables stay relatively stable. Without market information for private firms, the random components have stronger effects than those in the public firms.

4.3.2 Mixed Logit Models with Model Specific Tests

Further considerations about the robustness of the mixed logit model are presented in this subsection. As a simulated random model, its stability is important for applications. This

section includes the results based on alternative replication times of the Halton draws and the results from alternative assumptions of the estimated coefficients.

4.3.2.1 Halton Draws

[Table 4.2, p. 156 about here]

This part investigates the variation of fixed and random coefficients as the number of simulated Halton draws increases. The results presented in Table 4.2 are the mixed logit models with the simulated number range from 50 to 500. The tests are based on nine random variables, which exhibit significance in standard deviation in Model 2 (Table 4.1.) The significance of the estimated standard deviation of the EBITA increases when the number of random variables is limited to nine. As expected, a smaller drawing number improves computational efficiency while a higher drawing number improves prediction accuracy. The computational time increases from fourteen hours to five days as the Halton replication times increases from 50 to 500, while the likelihood ratios increase from -5552.77 to -5470.37. The values of random components (S.D.) of variables (FMVTMV, CLTA, LOGTAGNP, EBITA and WCTA) increase as the number of replications increase. This result contradicts the conclusion made by Train (2000) that a smaller number of Halton draws would generate a smaller degree of estimation variance. Although the log-likelihood ratios improve steadily as the replications increase, the significance of the coefficients in both the fixed variables and random variables remains stable. The significance of the net income ratio (NITA) decreases and the QACL remains insignificant. The similarity of the coefficients suggests the stability of the mixed logit model. It, furthermore, suggests that the bias is negligible between mixed logit models using alternative times of Halton replications.

4.3.2.2 Correlated Coefficients

[Table 4.3, p. 158 about here]

In the previous tests using mixed logit model, the estimated coefficients of those random variables are assumed to be independent. However, the variables are from certain categories (such as accounting ratios), which presumably are correlated within the same sector. This leads to another interesting aspect of the mixed logit regression based on the assumption of correlated random coefficients. Table 4.3 compares the prediction abilities of correlated and uncorrelated random coefficient tests using 50 Halton draws from 1970 to 2009. Following Table 4.2, the regression contains nine random variables.

The prediction ability is slightly improved with the correlated coefficient assumption as the log-likelihood increases from -5552.77 to -5333.49. The improvement indicates that, with the correlated coefficient assumption, the model could capture more default information. Especially, the results show the unique ability of the mixed logit model, that it is able to capture the clusters between variables. Although the observable variables are selected carefully in Chapter 3, the correlations between observable variables in the same sector are ineluctable; however, traditional methods such as the Z-score and the logit model could not capture the correlations theoretically. As the additional information is provided by the covariant matrix, the significance of some variables slightly decreases, such as DH, CD1M, GNP, INVESTL and EPS. It is worth noting that the signs of the estimated means and standard deviations of all the variables do not change from the independent assumption to the correlated assumption. This further indicates the stability of the mixed logit model.

[Table 4.4, p. 160 about here]

The covariance matrix of the nine random variables is presented in Table 4.4. It is clear that the coefficients of most random variables are significant at the 1% level, while a few are significant at the 5% or 10% levels. Only six out of forty-six pairs of the random variables show insignificant covariance. However, it is noteworthy that the computation time for the uncorrelated coefficient test is fourteen hours, while it takes eight days for a

single test with the correlated coefficient assumption.

4.3.3 Results from Cox Frailty and Non-frailty Models

To initialise the survival model, this chapter assumes 1st January 1970 as the start date and 31st December 2009 as the end date. There are 17,663 active firms during this time period. The entry time of any firm is counted as when there are available data of the firm level information for that firm. Very few firms survived more than 25 years. The average default age is 9.47 years, as shown in Figure 4.1. Firms exit the observation for different reasons, including: acquisition or merger, bankruptcy, liquidation, becoming a private firm and missing data from WRDS. The empirical test of Bonfim (2009) using survival models faces a serious left censoring problem since the observation duration is limited to 6 years. The sample in this thesis does not suffer such a serious left censoring problem since there are only 65 firms created on or before 1970.

[Figure 4.1, p. 166 about here]

This chapter further investigates the use of non- parametric survival analyses by different industries. Figure 4.2 reports the survival estimation over different industries using the non-parametric Kaplan-Meier method, which allows estimation of survival over time, even when firms drop out or studied for different lengths of time. The survival probabilities decrease with time for all industries, with different speeds. Firms in the retail trade (industry 7) fail much more quickly than firms in other industries. The financial industry (industry 8) and the public sector (industry 10) are the safest industries; their firms have relatively long lives. These results are identical with the average default rates with alternative industries presented in Chapter 3, and suggest potential industry effects.

[Figure 4.2, p. 167 about here]

The following sections present results from the Cox survival model, with and without the frailty structures. It first investigates the sector effect with the standard Cox survival

model and then tests the frailty effects on the optimal information set.

4.3.3.1 Non-frailty Cox Model

Table 4.5 contains the results of standard Cox proportional hazard models with different specifications based on quarterly data ranging from 1970 to 2009.

[Table 4.5, p. 161 about here]

To compare the models in terms of explanatory power and overall prediction ability, two measures of model-fit have been calculated: the log-likelihood ratio and the AUROC. The AUROC reports the classification performance of models over all possible cut-off points. The best possible prediction model would yield an AUROC at 1, which stands for 100% corrected classified with both non-default firms and default firms.

Model 1 and Model 2 in Table 4.5 tests the non-frailty Cox proportional model with the optimal variables set concluded from Chapter 3. The variable firm age (AGEQ) suggested in Chapter 3 has not been included in the regressions due to the property of the survival model that the age of each firm is an underlying variable. Model 1 conditions on selected information from financial statements, stock markets, macroeconomic indicators and default history, while the industry effects are added in Model 2. Model 3 contains information only from accounting ratios, while Model 4 tests the prediction ability of market-based information. The prediction result of private firms is presented in Model 5.

A comparison of Models 3 and 4 in Table 4.5 shows that a market information based model has slightly stronger prediction power than a model based solely on accounting ratios according to the decrease in value of AUROC from 0.8508 to 0.828 from Model 4 to Model 3. It is important to note that both market information and accounting information show significance in Models 1 and 2. This demonstrates that previous frailty studies (Das, Duffie, Kapadia & Saita, 2007; Duffie, Eckner, Horel & Saita, 2009) suffer from a shortage of observable variables. In particular, accounting information and the default record have

not been explored using survival models in previous studies. This undermines conclusions that the frailty effects are significant because one of the explanations of the unobserved variables could either be accounting information or default record. Model 2 outperforms Model 1 in terms of the likelihood ratio and the AUROC. This shows that the industry effect dummy variables additionally contribute to default rate. This is consistent with the results shown in Figure 4.2 and the conclusion from Chapter 3.

It is interesting to note that Earnings per Share (EPS) and stock volatility (SD) contribute the most prediction power in the market information model. The default history has a better explanatory ability in private firms in Model 5 than it has in public firms as demonstrated in Models 1 or 2.

In a Cox proportion model, a positive coefficient is associated with a higher hazard rate, and therefore reduces the expected corporate survival duration. It is interesting to note that the sign of the coefficients of all variables in the Cox proportional model test in this chapter is identical with that of the logit model estimated in Chapter 3.

4.3.3.2 Frailty Cox Model

The mixed logit models in the previous section suggest the existence of unobservable factors in the default cluster. However, this does not indicate potential reasons of the default clusters. This subsection investigates the potential cluster factors using the frailty models.

[Table 4.6, p.163 about here]

Table 4.6 summarises the results from the frailty tests. Model 1 is the non-frailty Cox model presented as a benchmark for frailty studies. All the frailty models outperform the standard non-frailty Cox model in terms of the log-likelihood ratio and the AUROC. Model 2 contains results from the Cox proportional model with a shared industry frailty. The industry groups are divided by the industry code, as summarised in Panel C of Figure 3.6

in Chapter 3. In Model 2, the estimated industry theta for the industry frailty is 0.02; the assumption that the industry theta equals to zero is rejected at the 1% significance level. The estimated industry frailty effect for each group is plotted in Figure 4.3. Industries 1, 3, 4, 5 and 7 have a higher hazard rate, which associates with a higher default probability. On the other hand, industries 2, 8 and 10 present with a lower default risk compared with the average default probability. The retail industry (7) is the most risky industry, while public administration (10) is exposed to relatively lower default risk. An unreported model that combines industry frailty elements with all the variables in Model 2 of Table 4.6 shows that the industry frailty loses significance when added to industry dummy variables. It is interesting to note that the prediction ability of Model 2 in Table 4.6 is slightly better than that of Model 2 in Table 4.5.

[Figure 4.3, p. 168 about here]

It is well known that defaults cluster in time. Model 3 in Table 4.6 presents results of the Cox model with shared yearly frailty factors. The time period for the regression is 40 years (1970 to 2009), which leads to the total shared group for yearly frailty being 40. The estimated calendar theta for the calendar effect is 0.02, and the null hypothesis that the yearly theta equals to zero is rejected at the 1% significant level. The estimated yearly frailty effect for each group is plotted in Figure 4.6. The frailty effects capture the default cluster effectively, showing that during the intensity default periods - the early 1990s, early 2000s and late 2000s - the calendar frailty effect has a big value. The default risk is relatively low during the periods 1985 to 1988 and 1996 to 1998.

Model 4 investigates shared macroeconomic frailty based on the GDP growth rate. There are 160 macroeconomic frailty groups, since macroeconomic data are available on a quarterly basis during the forty-year period. This chapter has also tested macroeconomic frailty based on other macroeconomic indicators such as GNP, INVEST, GNPL, etc. Interestingly, the regressions report very similar results for these alternative

macroeconomic indicators: the macroeconomic thetas are all equal to 0.02 and significant at the 1% level. The estimated macroeconomic frailty for each group is plotted in Figure 4.3. In the figure, it is not straightforward to discern a clear relation between GDP growth and macroeconomic frailty. This result indicates that although there are significant links between business cycle and credit cycle, the relationship is not linear, and requires further investigation.

Finally, a combination Cox model of the estimated industry, calendar and macroeconomic frailties was tested in Model 5 (Table 4.6). The estimated frailties from Models 2, 3 and 4 were imported into Model 5 with a fixed coefficient equal to 1. As expected, Model 5 with all the frailties shows the best performance in terms of model fits: Model 5 has the highest AUROC and a likelihood ratio at -13908.5. Interestingly, all the models in Table 4.6 have similar coefficients and z statistics.

4.4. Method Comparison

This section compares the out-of-sample performance of two random models: the mixed logit model and the three-factor frailty model. It uses the sample from 1970 to 1999 to predict the default behaviour in the most recent decade (2000-2009), and compares these two methods with three dimensions: computational efficiency, information content and prediction accuracy.

In terms of computation efficiency, the frailty model is preferable. The mixed logit model requires 24 hours for a 50 times Halton draw test, and the computational time increases to seven days as the replication times of Halton draws reach 500. On the other hand, the frailty model takes only 2 hours to identify the frailty factors. Most importantly, it takes only seconds with the fixed frailties, such as Model 5 in Table 4.6.

The mixed logit model captures the unobservable variables with the random coefficients, where the unobservable variables are expressed as a whole. On the other hand, the frailty

model explains the individual scopes of the unobservable variables. It captures details of the default clusters factors and divides the unobservable information into three sections: industry, macroeconomic and calendar effects.

[Table 4.7, p.165 about here]

Previous studies fail to compare the model performance of survival analysis method with other methods, due to the fact that the output of the survival model is a hazard rate. This chapter converts the hazard rate back to the default probability in order to compare the prediction abilities of survival analysis with other methods using classification rates with different cut-off points and the AUROC. Table 4.7 reports the details of classification rates for both models according to different cut-off points. Generally speaking, the mixed logit model out-performs the survival frailty model in terms of overall accuracy rate and the AUROC ratio. The AUROC of the mixed logit model in the out-of-sample test is 0.963, higher than that of the three-factor frailty model, which is 0.9379. It seems that the frailty model has a stronger ability to identify the default firms when comparing two models with the same cut-off point. However, when the sensitivity rates for the two models are compared, such as a cut-off point as 0.1 for the mixed logit model and 0.6 for the frailty model, the mixed logit model out-performs the frailty model for all sensitivity, specificity, and the overall accuracy rate. This shows that the mixed logit model generates better predictions than the frailty model, and could capture unobservable information. A further comparison of these two models in terms of the economic costs and benefits for creditors will be discussed in Chapter 5.

4.5 Conclusion

This chapter investigated unobservable default information and default clusters using mixed logit models and frailty models, conditioning on the optimal variables set developed in Chapter 3. Both methods are able to capture additional information beyond the optimal observable variable set. This feature improved prediction accuracy for corporate default

rates. This chapter presented evidence that previous studies did not incorporate market information for the mixed logit test and accounting information for the frailty models, and that their conclusions are therefore vulnerable and incomplete. Prediction results improve with the optimal variable set developed for this thesis, which includes accounting information, market information, macroeconomic information, and the default history.

The market information-based model outperforms the accounting-based model in both the mixed logit model and the Cox proportional survival test. However, information from both sections plays a unique role that adds to the prediction ability in the optimal variable set. For these two methods, public firms have better predictability than private firms since more information is available.

In this chapter, the stability of the mixed logit model with alternative simulation times and different coefficient assumptions was demonstrated. While a higher number of Halton draw times is associated with stronger prediction, the computation time increases tremendously. Also, the mixed logit model captures correlations between observable variables through the correlation coefficient assumption. The model with correlated random coefficient has better predictability than the model with independent random coefficient. However, the computational efficiency decreases in accordance with the correlated coefficient assumption.

The mixed logit model and the frailty model have been compared with three scopes: information content, prediction accuracy and computation efficiency. In terms of the information content, the mixed logit model captured unobservable variables as a black box, while the frailty model divided the default clusters into three sections: macroeconomic frailty, industry frailty and calendar frailty. While the mixed logit models have superior prediction accuracy, the frailty model offers better computation efficiency.

Table 4.1

Standard Logit and Mixed Logit Results

Table 4.1 presents mixed logit model tests with 50 Halton draws from 1970 to 2009. For each variable, the first row reports the value of the estimated coefficient and the second row reports the value of the z- statistic. Model 1 shows the logit model as a reference with the optimal variable set with industry effects. The mixed logit methods have been tested based on different variables sets: optimal variable set with industry effects (Model 2), accounting variables set (Model 3), market variables set (Model 4) and private company information set (Model 5). Both the mean and standard deviation (S.D.) are reported for each mixed logit test. The total number of observations is 639,573. The abbreviations of variables are presented in Table 2.2, Table 2.3 and Table 2.4. RIND2 is industry dummy for mining, construction, wholesales trade and service. RIND3 is industry dummy for transportation, communications, utilities and manufacturing. RIND4 is industry dummy for public administration, finance, insurance and real estate. *** represents significance at the 1% level, ** represents significance at the 5% level, and * indicates significance at the 10% level.

	Logit	Mixed Logit							
	Model 1	Model 2		Model 3		Model 4		Model 5	
	ALL	ALL		ACCOUNTING		MARKET		PRIVATE	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
DH	2.804***	5.319***						5.404***	
	-47.61	-34.48						-41.62	
AGEQ	-0.004***	0.031***						0.021***	
	-4.56	-8.11						-6.34	
CD1M	-0.092***	-0.075***						-0.100***	
	-4.96	-2.84						-4.140	
GNP	-0.000***	-0.001***						-0.001***	
	-4.92	-6.92						-4.85	
INVEST	0.004***	0.006***						0.004***	
	-6.76	-6.82						-5.00	
GNPL	14.768***	22.042***						18.34***	
	-6.98	-6.47						-6.20	
INVESTL	-14.153***	-19.750***						-16.35***	
	-7.54	-6.55						-6.22	
SPRETURN	3.952***	6.194***						0.278	
	-9.95	-11.23						-0.86	
EPS	-0.428***	-0.869***	1.454***			-3.205***	3.721***		
	-5.13	-5.18	-4.95			-23.50	-22.05		
NITA	-4.038***	-2.573***	1.253	-5.21***	3.481**			-4.780***	0.257
	-7.26	-3.10	-0.56	-9.02	-2.49			-7.03	-0.082
FMVTMV	-0.325***	-0.686***	0.84***			-0.391***	0.864***		
	-13.48	-9.97	-11.78			-9.56	-15.11		
RETURNFQ	-5.098***	-7.554***	0.037			-1.904***	0.067		
	-13.09	-13.81	-0.18			-7.62	-0.48		

CASHTA	-2.350***	-6.787***	8.574***	-3.062***	4.821***			-6.547***	8.572***
	-8.88	-8.13	-8.24	-5.36	-5.21			-9.12	-11.30
SDF	0.035***	0.038***	0.013			0.04***	0.033***		
	-18.09	-13.61	-1.31			-16.42	-5.16		
CLTA	1.655***	2.922***	5.636***	2.539***	4.594***			3.185***	6.144***
	-11.35	-8.03	-9.63	-9.63	-12.04			-9.97	-13.61
LOGTAGNP	0.342***	0.493***	0.655***	0.138***	0.202***			-0.0242	0.487***
	-17.11	-10.58	-9.39	-6.88	-3.80			-0.84	-10.68
SALEG	-0.723***	-0.763***	0.074	-0.749***	0.044			-0.762***	0.056
	-7.76	-6.26	-0.39	-7.63	-0.29			-6.64	-0.33
EBITA	-6.129***	-7.643***	4.641**	-4.551***	3.48**			-4.776***	5.347***
	-6.82	-5.58	-2.13	-4.28	-2.21			-3.75	-2.62
QPRID	0.111***	0.139***	0.013			0.05**	0.084**		
	-8.43	-6.16	-0.39			-2.18	-2.12		
EXCESS	3.952***	6.563***	0.227			1.762***	0.114***		
	-13.22	-15.13	-1.30			-8.83	-0.69		
PRICE	-0.063***	-0.27***	0.197***			-0.612***	0.408***		
	-9.45	-5.31	-4.07			-15.23	-14.32		
EBIATL	5.838***	7.567***	0.119	4.924***	0.683			5.420***	1.546
	-6.51	-6.14	-0.06	-4.68	-1.03			-4.36	-1.42
CFTL	-5.831***	-6.238***	0.705	-6.037***	2.376***			-6.657***	2.713***
	-7.15	-5.68	-0.67	-6.30	-3.00			-5.90	-2.67
QACL	0.161***	0.021	0.106*	0.092**	0.084			0.0787*	0.0927*
	-7.37	-0.43	-1.75	-2.42	-1.06			-1.72	-1.77
WCTA	-1.182***	-1.719***	4.754***	-2.179***	3.343***			-1.812***	4.170***
	-7.63	-4.83	-8.32	-8.22	-8.93			-5.80	-10.53
METL	-0.091***	-0.162***	0.164***	-2.54***	1.749***			-1.283***	0.910***
	-5.76	-4.33	-4.61	-17.68	-16.61			-13.61	-14.4
RIND2	-0.373***	-1.107***						-0.964***	
	-4.7	-6.16						-6.05	
RIND3	-0.241***	-0.914***						-0.888***	
	-3.3	-5.28						-5.90	
RIND4	-0.743***	-1.549***						-1.480***	
	-5.66	-5.88						-6.22	
Log-Likelihood	-9199.15	-5566.84		-7889.94		-7809.30		-6126.61	

Table 4.2

Mixed Logit Tests with Different Halton Replications

Table 4.2 presents nine random variables mixed logit tests with Halton replication from 50 times to 500 times. The tests are based on the period from 1970 to 2009. The total number of observations is 639,573. For each variable, the first row reports the value of the estimated coefficient and the second row reports the value of the z- statistic. The abbreviations of variables are presented in Table 2.2, Table 2.3 and Table 2.4. RIND2 is industry dummy for mining, construction, wholesales trade and service. RIND3 is industry dummy for transportation, communications, utilities and manufacturing. RIND4 is industry dummy for public administration, finance, insurance and real estate. *** represents significance at the 1% level, ** represents significance at the 5% level, and * indicates significance at the 10% level.

	Halton 50		**Halton 100**		**Halton 200**		**Halton 500**	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
DH	5.243***		5.412***		5.612***		5.853***	
	-35.00		-33.61		-32.13		-30.77	
AGEQ	0.030***		0.033***		0.033***		0.036***	
	-7.91		-8.37		-7.75		-8.14	
CD1M	-0.068***		-0.0627**		-0.0721***		-0.064**	
	-2.61		-2.34		-2.60		-2.25	
GNP	-0.001***		-0.001***		-0.001***		-0.001***	
	-6.60		-6.97		-6.76		-6.81	
INVEST	0.006***		0.006***		0.006***		0.006***	
	-6.40		-6.50		-6.34		-6.35	
GNPL	20.783***		21.400***		22.090***		22.150***	
	-6.36		-6.31		-6.20		-6.05	
INVESTL	-18.410***		-18.570***		-19.020***		-19.080***	
	-6.33		-6.18		-6.04		-5.91	
SP 500 RETURN	6.216***		6.333***		6.753***		6.807***	
	-11.48		-11.36		-11.71		-11.55	
RIND2	-1.095***		-1.224***		-1.343***		-1.354***	
	-6.13		-6.54		-6.77		-6.59	
RIND3	-0.942***		-1.103***		-1.152***		-1.193***	
	-5.47		-6.16		-6.05		-6.03	
RIND4	-1.555***		-1.656***		-1.815***		-1.927***	
	-5.83		-6.06		-6.28		-6.38	
NITA	-2.258***		-2.430***		-2.056**		-2.085**	
	-2.81		-2.95		-2.39		-2.38	
RETURNFQ	-7.469***		-7.572***		-8.004***		-8.070***	
	-13.97		-13.70		-13.97		-13.78	
SDF	0.036***		0.036***		0.037***		0.037***	
	-13.66		-13.19		-13.10		-13.00	
SALEG	-0.741***		-0.729***		-0.741***		-0.728***	
	-6.10		-5.87		-5.75		-5.53	
EXCESS	6.507***		6.644***		7.062***		7.132***	
	-15.39		-15.23		-15.50		-15.28	
EBIATL	7.579***		7.477***		8.019***		7.653***	
	-6.19		-5.81		-6.03		-5.54	

CFTL	-6.320***		-6.413***		-6.902***		-6.504***	
	-5.71		-5.49		-5.71		-5.23	
QPRID	0.149***		0.154***		0.166***		0.164***	
	-6.81		-6.91		-7.02		-6.79	
QACL	0.046		0.0369		0.0325		-0.00126	
	-1.04		-0.81		-0.66		-0.02	
EPS	-0.901***	1.609***	-0.796***	1.154***	-1.038***	1.765***	-0.959***	1.493***
	-5.87	-7.67	-4.97	-3.93	-6.10	-7.54	-5.53	-5.69
FMVTMV	-0.601***	0.774***	-0.674***	0.912***	-0.598***	0.930***	-0.639***	0.950***
	-10.01	-8.92	-10.10	-11.79	-8.82	-11.18	-8.68	-10.55
CASHTA	-5.801***	7.499***	-6.480***	8.274***	-7.795***	10.39***	-8.050***	9.892***
	-7.55	-7.86	-7.88	-8.24	-8.68	-9.82	-8.56	-9.44
CLTA	2.927***	5.679***	3.060***	6.017***	2.960***	6.197***	3.269***	7.084***
	-8.75	-12.63	-8.55	-11.39	-7.79	-11.31	-7.91	-11.91
LOGTAGNP	0.451***	0.654***	0.480***	0.751***	0.439***	0.776***	0.504***	0.847***
	-8.96	-9.79	-9.60	-12.72	-8.21	-10.76	-8.73	-12.00
EBITA	-7.714***	5.563***	-7.515***	6.410***	-7.767***	6.971***	-8.181***	10.37***
	-2.54	-5.75	-5.42	-2.80	-5.30	-2.88	-5.37	-5.48
PRICE	-0.328***	0.227***	-0.350***	0.234***	-0.440***	0.303***	-0.432***	0.292***
	-11.05	-11.96	-10.00	-10.08	-11.20	-12.27	-11.15	-12.21
WCTA	-1.659***	4.469***	-2.018***	5.034***	-2.430***	6.078***	-2.423***	6.389***
	-4.84	-8.32	-5.48	-9.99	-5.90	-11.35	-5.74	-11.68
METL	-0.241***	0.245***	-0.270***	0.270***	-0.422***	0.409***	-0.366***	0.342***
	-3.73	-4.57	-4.90	-5.87	-6.12	-8.45	-4.90	-5.98
Observations	639573		639573		639573		639573	
Log-Likelihood	-5552.77		-5523.51		-5485.8		-5470.37	

Table 4.3

Mixed Logit Models with Correlated Coefficients

Table 4.3 presents mixed logit tests using both uncorrelated and correlated coefficients assumptions with nine random coefficients from 1970 to 2009. The simulation is based on 50 times Halton draws. The total number of observations is 639,573. The abbreviations of variables are presented in Table 2.2, Table 2.3 and Table 2.4. For each variable, the first row reports the value of the estimated coefficient and the second row reports the value of the z- statistic. RIND2 is industry dummy for mining, construction, wholesale trade and service industries. RIND3 is industry dummy for transportation, communications, utilities and manufacturing. RIND4 is industry dummy for public administration, finance, insurance and real estate. *** represents significance at the 1% level, ** represents significance at the 5% level, and * indicates significance at the 10% level.

	Uncorrelated coefficients		**Correlated coefficients**	
	Mean	S.D.	Mean	S.D.
DH	5.243***		5.799***	
	-35.00		-33.82	
AGEQ	0.030***		0.038***	
	-7.91		-8.87	
CD1M	-0.068***		-0.066**	
	-2.61		-2.33	
GNP	-0.001***		-0.001***	
	-6.60		-6.32	
INVEST	0.006***		0.005***	
	-6.40		-5.61	
GNPL	20.783***		20.027***	
	-6.36		-5.62	
INVESTL	-18.410***		-17.258***	
	-6.33		-5.45	
SP 500 RETURN	6.216***		7.049***	
	-11.48		-12	
RIND2	-1.095***		-1.287***	
	-6.13		-6.47	
RIND3	-0.942***		-1.046***	
	-5.47		-5.46	
RIND4	-1.555***		-1.705***	
	-5.83		-5.92	
NITA	-2.258***		-2.981***	
	-2.81		-3.46	
RETURNFQ	-7.469***		-8.256***	
	-13.97		-14.23	
SDF	0.036***		0.038***	
	-13.66		-13.31	
SALEG	-0.741***		-0.768***	
	-6.10		-5.95	
EXCESS	6.507***		7.295***	
	-15.39		-15.77	
EBIATL	7.579***		7.381***	
	-6.19		-5.35	

CFTL	-6.320***		-6.627***	
	-5.71		-5.41	
QPRID	0.149***		0.158***	
	-6.81		-6.67	
QACL	0.046		0.05	
	-1.04		-0.98	
EPS	-0.901***	1.609***	-0.714***	1.270***
	-5.87	-7.67	-4.44	5.44
FMVTMV	-0.601***	0.774***	-0.607***	1.540***
	-10.01	-8.92	-8.17	15.25
CASHTA	-5.801***	7.499***	-6.125***	6.870***
	-7.55	-7.86	-7.89	7.19
CLTA	2.927***	5.679***	2.777***	7.577***
	-8.75	-12.63	-6.82	12.97
LOGTAGNP	0.451***	0.654***	0.517***	1.348***
	-8.96	-9.79	-8.52	14.57
EBITA	-7.714***	5.563***	-7.441***	4.963***
	-2.54	-5.75	-5.06	3.77
PRICE	-0.328***	0.227***	-0.363***	0.263***
	-11.05	-11.96	-11.60	13.25
WCTA	-1.659***	4.469***	-2.640***	8.264***
	-4.84	-8.32	-5.91	14.58
METL	-0.241***	0.245***	-0.467***	0.462***
	-3.73	-4.57	-6.07	9.25
Observations	639573		639573	
Log-Likelihood	-5552.77		-5333.49	

Table 4.4

Covariance Matrix of Random Variables

Table 4.4 presents the covariance between the nine random variables in Table 4.3. The abbreviations of variables are presented in Table 2.2 and Table 2.3. *** represents significance at the 1% level, ** represents significance at the 5% level, and * indicates significance at the 10% level.

	EPS	FMVTMV	CASHTA	CLTA	LOGTAGNP	EBITA	PRICE	WCTA	METL
EPS	1.614***								
FMVTMV	-0.259*	2.371***							
CASHTA	8.594***	0.030	47.189***						
CLTA	-3.486***	2.359***	-16.507***	57.406***					
LOGTAGNP	-0.244**	-1.464***	-2.676***	-0.767	1.818***				
EBITA	3.100**	-3.222**	14.005*	-7.686	-0.388	24.630***			
PRICE	0.122***	-0.074***	0.695***	-0.771***	-0.116***	0.592***	0.069***		
WCTA	-4.296***	3.117***	-17.487***	34.757***	-1.852***	-24.192**	-0.202	68.293***	
METL	0.235***	-0.222***	0.925***	-2.065***	0.344***	0.478	0.023*	-1.689***	0.214***

Table 4.5

Cox Proportion Models on Different Information

Table 4.5 presents standard Cox proportional hazard models with different variable combinations based on the quarterly data from 1970 to 2009. For each variable, the first row reports the value of the estimated coefficient and the second row reports the value of the z- statistic. Model 1 presents results with the optimal variable set. Model 2 shows results with the optimal variable set with industry effects. Model 3 tests the Cox proportion model with accounting information. Model 4 reports results based on market information. Model 5 shows the results from private firms. The total number of observations is 639,573. The abbreviations of variables are presented in Table 2.2, Table 2.3 and Table 2.4. RIND2 is industry dummy for mining, construction, wholesales trade and service. RIND3 is industry dummy for transportation, communications, utilities and manufacturing. RIND4 is industry dummy for public administration, finance, insurance and real estate. *** represents significance at the 1% level, ** represents significance at the 5% level, and * indicates significance at the 10% level.

	Model 1	Model 2	Model 3	Model 4	Model 5
DH	2.499***	2.468***			3.109***
	-44.89	-44.12			-62.71
CD1M	-0.075***	-0.074***			-0.110***
	-4.24	-4.21			-6.25
GNP	-0.000***	-0.000***			-0.000**
	-3.95	-4.07			-2.15
INVEST	0.003***	0.003***			0.002***
	-5.58	-5.69			-3.21
GNPL	13.119***	13.194***			13.803***
	-6.49	-6.52			-6.74
EPS	-0.435***	-0.410***		-2.745***	
	-5.37	-5.09		-36.66	
INVESTL	-12.603***	-12.698***			-12.614***
	-7.03	-7.07			-7.02
SP 500 RETURN	3.790***	3.762***			0.235
	-10.38	-10.30			-0.95
NITA	-3.547***	-3.565***	-7.937***		-5.719***
	-6.97	-7.01	-16.20		-11.92
FMVTMV	-0.291***	-0.293***		-0.205***	
	-12.63	-12.73		-10.80	
RETURNFQ	-4.816***	-4.780***		-2.276***	
	-13.22	-13.11		-10.37	
CASHTA	-2.456***	-2.194***	-1.871***		-2.594***
	-9.98	-8.77	-8.06		-10.71
SDF	0.031***	0.031***		0.046***	
	-16.82	-16.70		-25.69	

CLTA	1.460***	1.392***	2.364***		1.941***
	-10.69	-10.04	-16.53		-13.74
LOGTAGNP	0.304***	0.303***	0.170***		0.093***
	-16.42	-16.29	-15.12		-7.66
SALEG	-0.643***	-0.641***	-0.739***		-0.730***
	-7.29	-7.27	-9.06		-8.53
EBITA	-4.843***	-4.668***	-0.571		-1.913**
	-6.04	-5.81	-0.72		-2.44
QPRID	0.105***	0.107***		-0.025*	
	-8.76	-8.85		-1.82	
EXCESS	3.825***	3.802***		1.564***	
	-13.81	-13.72		-8.92	
PRICE	-0.066***	-0.068***		-0.056***	
	-10.07	-10.24		-7.13	
EBIATL	4.976***	4.886***	2.594***		3.172***
	-5.87	-5.77	-3.16		-3.59
CFTL	-5.256***	-5.290***	-4.686***		-5.275***
	-6.82	-6.87	-6.24		-6.49
QACL	0.145***	0.154***	0.189***		0.185***
	-7.11	-7.40	-11.16		-9.70
WCTA	-1.024***	-1.216***	-1.625***		-1.104***
	-7.02	-8.07	-11.93		-7.68
METL	-0.099***	-0.099***	-0.444***		-0.275***
	-6.35	-6.37	-18.85		-14.31
RIND2		-0.361***			
		-4.94			
RIND3		-0.198***			
		-2.94			
RIND4		-0.688***			
		-5.52			
Observations	639573	639573	639573	639573	639573
Log-Likelihood	-13990.6	-13968.9	-16267.76	-16179.66	-14584.09
AUROC	0.9148	0.9157	0.8280	0.8508	0.8796

Table 4.6

Frailty Cox Survival Model Results

Table 4.6 presents results of the frailty models based on quarterly data from 1970 to 2009. For each variable, the first row reports the value of the estimated coefficient and the second row reports the value of the z-statistic. Model 1 presents a non-frailty model as a benchmark. Model 2 summarises the Cox survival model with industry frailty. Model 3 reports the survival result with calendar frailty. Model 4 reports the result with macroeconomic frailty. Model 5 combines industry, calendar and macroeconomic frailties. The total number of observations is 639,573. The abbreviations of variables are presented in Table 2.2, Table 2.3 and Table 2.4. *** represents significance at the 1% level, ** represents significance at the 5% level, and * indicates significance at the 10% level.

	Model 1	Model 2	Model 3	Model 4	Model 5
DH	2.499***	2.481***	2.511***	2.500***	2.493***
	-44.89	-44.28	-45.01	-44.81	-44.82
CD1M	-0.075***	-0.075***	-0.080***	-0.073***	-0.078***
	-4.24	-4.26	-3.50	-3.58	-4.41
GNP	-0.000***	-0.000***	-0.000***	-0.000***	-0.000***
	-3.95	-4.23	-2.94	-3.61	-4.37
INVEST	0.003***	0.003***	0.003***	0.003***	0.003***
	-5.58	-5.89	-4.03	-4.94	-6.07
GNPL	13.119***	13.552***	13.416***	13.011***	13.631***
	-6.49	-6.69	-5.23	-5.78	-6.84
EPS	-0.435***	-0.413***	-0.448***	-0.441***	-0.434***
	-5.37	-5.13	-5.52	-5.43	-5.38
INVESTL	-12.603***	-13.019***	-12.951***	-12.432***	-13.113***
	-7.03	-7.23	-5.50	-6.09	-7.39
SP 500 RETURN	3.790***	3.758***	3.745***	3.802***	3.734***
	-10.38	-10.29	-9.89	-9.22	-10.12
NITA	-3.547***	-3.528***	-3.491***	-3.577***	-3.501***
	-6.97	-6.94	-6.85	-7.01	-6.89
FMVTMV	-0.291***	-0.291***	-0.286***	-0.289***	-0.285***
	-12.63	-12.64	-12.34	-12.48	-12.36
RETURNFQ	-4.816***	-4.787***	-4.810***	-4.861***	-4.826***
	-13.22	-13.12	-13.16	-13.26	-13.21
CASHTA	-2.456***	-2.255***	-2.454***	-2.460***	-2.261***
	-9.98	-8.97	-9.96	-9.98	-9.17
SDF	0.031***	0.031***	0.031***	0.031***	0.030***
	-16.82	-16.69	-16.68	-16.78	-16.61
CLTA	1.460***	1.351***	1.461***	1.451***	1.343***
	-10.69	-9.51	-10.69	-10.62	-9.72
LOGTAGNP	0.304***	0.303***	0.298***	0.302***	0.295***
	-16.42	-16.18	-15.98	-16.20	-15.87
SALEG	-0.643***	-0.644***	-0.647***	-0.648***	-0.653***
	-7.29	-7.29	-7.32	-7.32	-7.38
EBITA	-4.843***	-4.706***	-4.995***	-4.904***	-4.921***
	-6.04	-5.85	-6.21	-6.10	-6.11
QPRID	0.105***	0.106***	0.104***	0.106***	0.105***
	-8.76	-8.74	-8.60	-8.73	-8.67
EXCESS	3.825***	3.805***	3.810***	3.841***	3.807***
	-13.81	-13.73	-13.73	-13.81	-13.74

PRICE	-0.066***	-0.068***	-0.065***	-0.066***	-0.066***
	-10.07	-10.22	-9.90	-9.98	-10.04
EBIATL	4.976***	4.944***	5.064***	5.033***	5.089***
	-5.87	-5.84	-5.96	-5.91	-5.97
CFTL	-5.256***	-5.366***	-5.272***	-5.268***	-5.398***
	-6.82	-6.97	-6.82	-6.80	-6.96
QACL	0.145***	0.151***	0.145***	0.144***	0.149***
	-7.11	-7.24	-7.05	-7.03	-7.15
WCTA	-1.024***	-1.244***	-1.021***	-1.019***	-1.236***
	-7.02	-8.04	-6.99	-6.98	-8.38
METL	-0.099***	-0.099***	-0.098***	-0.099***	-0.098***
	-6.35	-6.33	-6.33	-6.34	-6.31
Observations	639573	639573	639573	639573	639573
Log-Likelihood	-13990.6	-13977.1	-13984.6	-13987.2	-13908.5
Groups No.		10	40	160	
Industry theta		0.061***			
Calendar theta			0.02***		
Macro theta				0.02***	
AUROC	0.9148	0.916	0.9155	0.9157	0.9172

Table 4.7

Out-of-sample Comparison for Advanced Random Models

Table 4.7 compares the prediction accuracy of the mixed logit model and the three-factor frailty model. The ten-year prediction results are based on the sample from 1970 to 1999. Panel 1 presents results from the mixed logit model. Panel 2 shows results from the frailty model.

Panel 1:Mixed Logit classification 2000-2009 (AUROC 0.9630)					
Cut off point	Sensitivity	Specificity	Type I	Type II	Overall accuracy rate
0.9	5%	100%	95%	0%	99.5%
0.8	12%	100%	88%	0%	99.6%
0.7	18%	100%	82%	0%	99.6%
0.6	24%	100%	76%	0%	99.6%
0.5	29%	100%	71%	0%	99.6%
0.4	35%	100%	65%	0%	99.6%
0.3	42%	100%	58%	0%	99.6%
0.2	52%	100%	48%	0%	99.6%
0.1	66%	99%	34%	1%	99.2%
0.05	75%	99%	25%	1%	98.4%
0.01	87%	93%	13%	7%	93.2%
0.005	91%	89%	9%	11%	88.8%
0.001	96%	72%	4%	28%	71.9%
0.0005	97%	62%	3%	38%	62.0%

Panel 2:Three-Factor Frailty model classification 2000-2009 (AUROC 0.9379)					
Cut off point	Sensitivity	Specificity	Type I	Type II	Overall accuracy rate
0.9	46%	99%	54%	1%	98.6%
0.8	53%	98%	47%	2%	98.2%
0.7	58%	98%	42%	2%	97.8%
0.6	62%	97%	38%	3%	97.2%
0.5	68%	97%	32%	3%	96.5%
0.4	72%	96%	28%	4%	95.4%
0.3	77%	94%	23%	6%	93.7%
0.2	83%	91%	17%	9%	90.5%
0.1	90%	83%	10%	17%	82.6%
0.05	94%	71%	6%	29%	71.1%
0.01	99%	39%	1%	61%	39.0%
0.005	99%	27%	1%	73%	27.3%
0.001	100%	10%	0%	90%	10.7%
0.0005	100%	7%	0%	93%	7.2%

Figure 4.1

Ages at Defaults

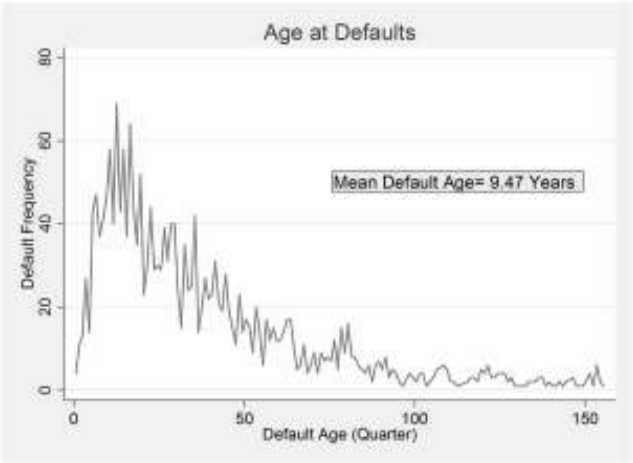


Figure 4.1 presents the age by quarter at defaults for 17,663 firms in the United States from 1970 to 2009. The average default age is 9.47 years.

Figure 4.2

Kaplan-Meier Survival Estimation

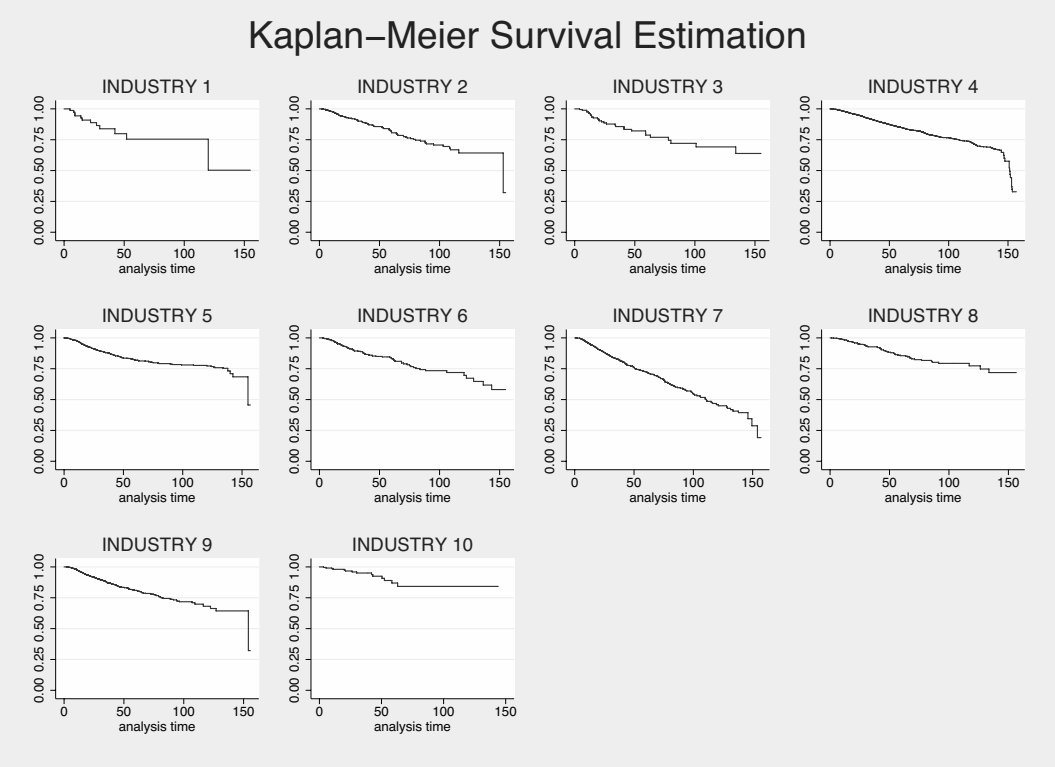


Figure 4.2 describes the Kaplan-Meier survival estimations for different industries. Analysis time is in quarters. Industry 1 is Agriculture, Forestry, and Fishing. Industry 2 is the Mining sector. Industry 3 is Construction. Industry 4 is the sector of Manufacturing. Industry 5 is Transportation, Communications, and Utilities. Industry 6 is Wholesale Trade. Industry 7 is Retail Trade. Industry 8 is Finance, Insurance, and Real Estate. Industry 9 is Services. Industry 10 is Public Administration.

Figure 4.3

Frailty by Different Categories

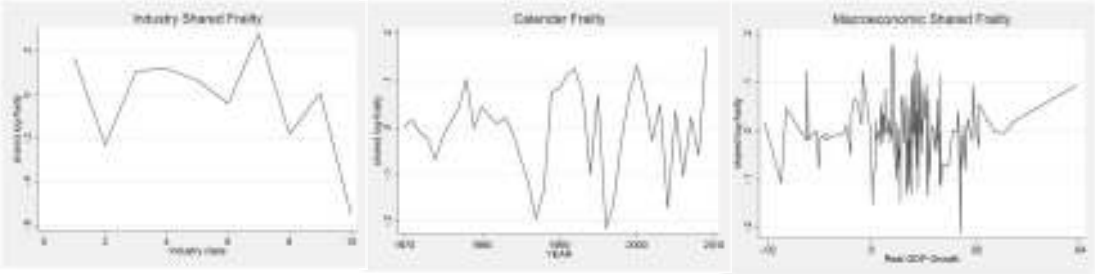


Figure 4.3 presents the frailty factors by groups. The figure on the left shows industry frailty. The middle figure shows yearly frailty. The figure on the right presents the macroeconomic frailty by GDP growth rate.

Chapter 5: Model Comparison

5.1 Introduction

Along with the development of default prediction models and sophisticated financial instruments, such as credit derivatives, there has been a renewed interest in credit risk model assessment and comparison. The existing empirical evidence, however, presents a fragmentary and complex picture that excludes the possibility of recommending a preferred method among the extensive range of default prediction models.

The traditional Z-score method is preferred by Agarwal & Taffler (2007, 2008), but is challenged by a number of other authors (Lennox, 1999; Chava & Jarrow, 2004; Hillegest, Keatin, Cram & Lundstedt, 2004). The superiority of the conditional probability models is proposed by Lennox (1999) and Chava & Jarrow (2004), but Jones & Hensher (2004) show that the mixed logit model out-performs the logit model. The frailty models are newly developed, but the performance of this model compared with other methods remains unclear. Bonfim (2009) documented that, in the parametric survival family, the Weibull and log-logistic distribution models provide more accurate predictions. However, it remains to be seen whether the parametric survival models out-perform other methods.

This chapter selects a range of statistical default prediction models in order to conduct a coherent and exhaustive assessment on default risk models. More importantly, this study contributes to the existing literature by comparing the newly developed advanced random models, including the mixed logit model and the frailty model, with less sophisticated measurements. Additionally, this is the first comparison between survival analysis methods and models from other branches of default risk studies. In order to make these methods comparable, this chapter transforms the outcomes of different models into the same scale with two steps. First, it rescales the results from discriminant analysis into the

probability range. Second, it transforms the hazard rate from survival analyses to the default probabilities. Excluding the univariate model, the comparison is within the same information set, the optimal variables group, suggested in Chapter 3. The information set contains details at both firm and macroeconomic levels. Models under consideration include the univariate model based on the total asset ratio, the multivariate discriminant model, the logit and probit model, the mixed logit model, parametric survival models with five different distribution assumptions, the semi-parametric Cox model, and the frailty model developed in Chapter 4.

This chapter measures the predicting ability of the twelve models under two main scopes: overall prediction accuracy and economic value for creditors. To assess the overall prediction accuracy, three criteria are examined: the area under the receiver operating characteristic curve (AUROC), the accuracy ratios (AR) and the models' stability in a period of financial crisis. The range of these evaluations includes in-sample tests, out-of-sample tests, and rolling out-of-sample tests - as suggested by Sobehart, Keenan & Stein (2001). Although the indicators of overall prediction accuracy conclude a simple method to measure prediction performance, the difference between two types of misclassification costs is instrumental for efficient default prediction.

To assess the economic costs, this chapter first lists and compares two types of error over varying cut-off points. Then, it uses the framework proposed by Blöchlinger & Leippold (2006), and the application from Agarwal & Taffler (2008), in order to compare market share, revenue, profitability, and Return over Assets for creditors using alternative default risk measurements. Importantly, this chapter extends the framework by comparing the opportunity loss of creditors, by employing different models. This chapter furthermore calculates Type I and Type II error derived loss, respectively, under the mixed regime framework of Blöchlinger & Leippold (2006). The Return over Assets ratio is used to indicate the effect of Type I error costs, and the opportunity loss over assets indicator is introduced to show the effect of Type II error cost. Finally, this chapter explores the

dynamic process of the competitive credit market that it excludes creditors gradually from the market if they suffer negative expected return associated with a default measurement.

The chapter reaches five main conclusions. 1) The mixed logit model generates the best prediction accuracy according to the AUROC and the accuracy rate, and the method is also the second most stable model. 2) Creditors using advanced random models covering the mixed logit model and the frailty Cox model would realise considerably higher revenues, profit, Return over Assets, and a lower opportunity loss over assets, than lenders using other methods. 3) There is little to choose between the traditional multivariable discriminant analysis (MDA) and the conditional probability models. MDA is the most stable model during the financial crisis and offers practical benefits for lenders in terms of the economic value it generates, while the conditional probability models have a slightly higher accuracy ratio and AUROC. 4) Among the conditional probability methods, the probit model somewhat surprisingly out-performs the popular logit model under the optimal variable set. 5) In the survival analysis family, the semi-parametric Cox survival model out-performs all the parametric survival models in terms of both the overall accuracy and the profit contribution to a financial institution.

The remainder of the chapter is presented as follows. Section 2 briefly describes the data sample and the methodologies of the twelve default risk measurements applied in the subsequent sections. Section 3 explains the assessment criteria for measuring the default risk. The comparison results with alternative benchmarks are reported in Section 4. Section 5 concludes the chapter.

5.2 Data and Methods

This section first describes the sample for the empirical tests, and then presents alternative models evaluated in the following sections.

5.2.1 Model Sample

To evaluate the performance of different models in the new economic environment, this chapter concentrates the default prediction on the most recent decade, from the year 2000 to 2009. This ten-year sample consists of 9,656 listed US companies with 935 default events. Table 5.1 describes the details of the sample by different subsample periods. This chapter compares models in both in-sample and out-of-sample test. In the in-sample test, the ten-year sample is used to compare the default performance, while, in the out-of-sample analysis; the first five-year subsample (2000-2004) is used to predict the default behaviour for the following subsample (2005-2009). In particular, this chapter assesses prediction ability with the expanding rolling out-of-sample analysis that is suggested by Sobehart, Keenan & Stein (2001). For example, it uses the subsample of 2000-2004 to forecast the default probability of 2005, and then uses the sample of 2000-2005 to forecast the performance of 2006.

[Table 5.1, p. 186 about here]

5.2.2 Default Prediction Methods

This subsection describes the models evaluated in what follows. It assesses the prediction ability of twelve models across five categories, including discriminant analysis, the conditional probability models, the discrete choice model and parametric and semi-parametric survival analyses. The models that have been chosen for this thesis include both the most popular benchmark measurements, such as the multiple discriminant model and the logit model, and models less commonly used such as the probit and parametric survival models. Most importantly, two random models have been included in comparison: the mixed logit model and the Cox frailty model.

According to the model outputs, these models group into three matching categories: default score models, default probability models and default time probability models.

The outcomes of both the univariate and multiple discriminant models are the default score. A higher Z-score associates with higher probability of default. Among the optimal independent variables and extensive accounting ratios, this chapter chooses the liability ratio that is the total liability over total assets for the univariate test, because: first, previous literature uses this ratio as a reference to compare other methods (Miller, 2009); and second, this liability ratio differentiates the default and non-default groups significantly in the t-test in Chapter 3. For the multiple discriminant analysis, this chapter assumes the Z-score is a linear combination of the optimal variable set following most previous studies (e.g. Altman E. I., 1968; Agarwal & Taffler, 2008)[4]. The optimal variable set is used to compare the MDA with other models under the same variable set containing firm level and macroeconomic level information. Moreover, our optimal MDA model based on unequal sample classification keeping the original default rate, which avoid any unrealistic assumption that the proportion of the defaulting and non-defaulting firms are similar. Furthermore, our optimal MDA outperforms the traditional five variables (Altman E. I., 1968) multiple discriminant analysis. The ten year in sample MDA Z-score is calculated as follow:

$$z = 0.77 * DH - 0.08 * \text{AGEQ} - 0.09 * \text{cd1m} - 1.11 * \text{gnp} + 1.04 * \text{invest} + 1.47 * \text{gnpl} - 0.10 * \text{EPS} - 1.51 * investl + 0.88 * SPRETURN - 0.28 * NITA - 0.66 * FMVTMV - 0.38 * RETURNFQ - 0.07 * CASHTA + 0.22 * SDF + 0.24 * CLTA + 0.76 * LOGTAGNP - 0.07 * SALEG - 0.25 * EBITA - 0.02 * QPRID + 0.32 * EXCESS + 0.10 * PRICE + 0.46 * EBIATL - 0.10 * CFTL + 0.14 * QACL - 0.13 * WCTA + 0.14 * METL \quad (17)$$

The benefit of the conditional probability model and the discrete choice model is that the outcomes of the models are interpreted as the default probabilities directly. A higher value of probability is associated with a higher default risk. The difference between the probit model and the logit model is simply the assumption of the residuals. Both the probit and the logit models have been selected in this chapter, since few studies compare the

[4] Other combinations such as the quadratic combination and logistic combination have been tested; the classification results are quite similar, thus have not reported for the purpose of this chapter.

performance of the probit model with methods from other families.

The outcomes of the survival models are the default time probability transferred from the survival probability results from the regressions. The default time probability indicates the probability that a firm will default before the current time. Parametric survival models with different distribution assumptions are tested in the chapter, including the exponential, Weibull, Gompertz, lognormal and log-logistic. The Cox proportional model and the frailty model described in Chapter 4 are also included. The frailty elements in the frailty model consist of industry frailty, macroeconomic frailty and yearly frailty.

[Table 5.2, p.187 about here]

Table 5.2 presents the detailed mathematical formulas behind each method. Generally speaking, β_j is the estimated coefficients for the default indicator j, and X_j is the j default indicator. N is the total number of default indicators. In the probit model and the lognormal model, Φ stands for the standard cumulative normal with the standard deviation σ. In the mixed logit model, specify the random distribution for the estimated coefficient vector β is specified as the normal distribution with mean b and the covariance W. In the parametric survival models, p, σ, γ are the ancillary parameters that need to be estimated during the regression. In the survival analysis, T is the length of time that each firm has existed and t is the current time.

5.3 Model Comparison Approaches

To compare the prediction accuracy of these twelve models, this chapter employs three methods: the AUROC evaluation of the overall prediction performance, the economic cost for both Type I and type two errors and model stability over different periods. To assess the practical performance of these models, this chapter compares the economic benefits and costs for lenders using each model.

5.3.1 AUROC Comparison

Originally used in medicine, engineering, and psychology, the area under the Receiver Operating Characteristic Curve (AUROC) is one of the most frequently applied methods for assessing the performance of pattern recognition. Recently, studies such as Chava & Jarrow (2004) have used AUROC to compare default predicting ability. The AUROC summarises both the Type I and Type II errors for every cut-off point; therefore, this method is applied to compare the global performance of different default prediction models. It makes the comparison visually and qualitatively assessable. Moody's refers to this curve as a cumulative accuracy profile, which is the equivalent tool. AUROC is a diagram of the sensitivity versus the absolute difference between one and the specificity in the diagnostic test. The sensitivity stands for the fraction of default cases that are properly classified, whereas the specificity is the proportion of non-default cases that are properly classified; therefore, the greater the value under the AUROC, the stronger the prediction ability.

At the end of each quarter, all the observations in the sample are ranked based on the default risk estimated as default score, default probability or default time probability, from highest risk to lowest risk. For every integer (x) between 0 and 1000, this chapter calculates the number of firms to default within one quarter within the first x% of the highest default risk firm observation.

An extension of the classification accuracy is the accuracy ratio (AR), as used by authors including Agarwal & Taffler (2008) and Sobehart, Keenan & Stein (2001). While the AUROC plots are a convenient way to visualise model performance, a single statistical measure is required to evaluate for both Type I and Type II errors. It is simply a linear transformation of the AUROC curve:

$$AR = 2 * (AUROC - 0.5) \tag{18}$$

Another measure that helps to determine prediction ability is credit stability. A stable model ensures better performance of the suggested models in practice. In the current

financial crisis, Standard & Poor's (2008) recognises that credit stability is a significant explicit default factor. The stability comparison also meets the stress test requirements from the recent Basel III policy. This chapter applies the standard deviation of the AUROC to measure the prediction stability. Mathematically, it is formulated as:

$$\sqrt{\frac{\sum(AUROC - \overline{AUROC})^2}{(N-1)}} \tag{19}$$

where $\overline{AUROC}$ is the sample mean of AUROC and N is the sample size.

5.3.2 Economic Benefits and Costs

Although the AUROC and the AR measure the general prediction performance, it is important to present details of Type I and Type II errors across different cut-off points, since the economic cost of Type I and Type II errors for default prediction are different. The default prediction model could err in two ways. First, a Type I error refers to cases where firms with high default risk are classified as low risk firms. Second, models indicating a high default probability when the default risk is actually low: experience Type II error. In a credit market, misclassification of a subsequently defaulting firm may drive the creditor's loss up to the whole loan amount, while a misclassification of a healthy firm costs mainly the interest income. The economic costs of different error types are presented in Table 5.3.

[Table 5.3, p. 188 about here]

Although studies such as Sobehart, Keenan & Stein (2001) have mentioned the importance of differentiating the economic cost of classification errors in a credit risk model, the choice of quantitative method to measure the expected total loss for the model error is limited. Blöchlinger & Leippold (2006) developed a price scheme to assess performance for creditors using different prediction models.

This chapter divides investors in a credit market into twelve creditor groups, according to

the default prediction methods, and assumes that the total market size is $100 billion. The creditors rank the default risk of each observation, and will offer customers different interest rates according to the level of the default risk. The higher the default risk, the more expensive the loans; creditors would refuse those customers with the highest risk. One can assume that all creditor groups reject customers with a default score in the top 5%. On the other hand, a company chooses a creditor group with the lowest credit spread in the market. If this creditor group has rejected the company, the company has to increase the budget until it obtains an offer. It is worth noting that the lowest credit quality companies may end up with no offer. If the credit spreads are equal for several credit groups, firms choose one randomly.

The credit spread created by Blöchlinger & Leippold (2006), assuming that creditors would only invest loans and bonds with positive net present value, is:

$$R = \frac{P\{Y = 1|S = s\}}{P\{Y = 0|S = s\}} LR + k \tag{20}$$

where $P\{Y = 1|S = s\}$ is the conditional default probability given the default score as s, and $P\{Y = 0|S = s\}$ is the conditional non-default probability given the default score s. LR is the loss rate. $LR = 1$ stands for a loss of 100%, the principle given default. k is the credit spread as the risk free loans. Similarly to Agarwal & Taffler (2008), this chapter groups the default score into twenty groups. A higher default score is associated with a higher default risk. It is also assumed that the LR is exogenous and the same for all creditor groups. Finally, this chapter works with the credit spread to seven decimal places.

To assess the economic value of applying different methods under this regime, the return on assets (ROA) is calculated following Agarwal & Taffler (2008), as:

$$ROA = \frac{Profit}{Assets\ lent} \tag{21}$$

This indicator associates with Type I errors, as the profit derives from the gap between the

revenue and the type I loss, which is the misclassification loss when creditor has accepted an actually high default risk firm.

To assess the potential profit loss associated with the Type II error, the opportunity loss on assets (OLOA) is constructed as:

$$OLOA = \frac{Opportunity\ Loss}{Assets\ lent} \tag{22}$$

If creditors reject a good quality customer, they calculate the potential profit as the opportunity loss (OL) with this misclassification. Although the opportunity loss could be similar for different creditors, the deterioration varies with the assets size.

5.4 Results

5.4.1 Statistical Summary

Table 5.4 summarises the prediction results for all twelve methods for both in-sample and out-of-sample prediction. The sample of 2000-2004 is used to predict the default risk of 2005-2009 as the out-of-sample test. All results are estimated using the optimal information set developed in Chapter 3 (except the TLTA model). This secures the model performance comparison as it uses the same information content, a crucial analysis issue that was addressed by Sobehart, Keenan & Stein (2001) and Agarwal & Taffler (2008). Because the results of the TLTA and the MDA methods are scores, they have been transformed as probabilities of default equivalent to other methods using

$$\frac{\text{Z-Min(Z)}}{\text{[Max(Z)-Min (Z)]}} \tag{23}$$

where Z stands for the original score. In the mixed logit test, seven uncorrelated random coefficients (EPS, FMVTMV, LOGTAGNP, EBITA, PRICE, WCTA, CASHTA) are used with 100 times Halton draws simulation. For both in-sample and out-of-sample results, the mean probabilities of default firms are significantly higher than those of non-default firms for all

twelve methods. Comparing the t-test, the frailty Cox model differentiates the default group most significantly in both in-sample and out-of-sample tests.

The true default probability rate for the in-sample test is 0.478%, while the true default probability rate for the out-of-sample test is 0.363%. It is of great interest to note that the estimated results of the mixed logit model are identical to this true value for both tests. The probit and logit model are also similar to the true value in the in-sample test, but lower in the out-of-sample tests. The predicted average probability of default from the survival models is 6-15 times higher than the true value in the in-sample test, and 9-17 times higher in the out-of-sample tests. The transformed TLTA and MDA methods are more than 50 times higher than the true average default probability. However, studies by Hillegest, Keatin, Cram & Lundstedt (2004) and Agarwal & Taffler (2008) suggest that the poor calibration is not relevant to the prediction ability and does not necessarily indicate that these models will lack information about the actual default probability.

[Table 5.4, p. 189 about here]

5.4.2 Overall Prediction Accuracy

Figures 5.1 and 5.2 present the area under the ROC curve comparison for in-sample and out-of-sample respectively. The 45° solid reference line in both figures indicates a default model with no predictive ability. The following can be observed from the figures: 1) Each of the twelve models does a better job of predicting default than the reference random model. 2) The mixed logit model has the best prediction ability for both in-sample and out-of-sample tests. 3) The probit model, the logit model MDA, the Cox survival model and the frailty Cox model show second best performance. 4) There is little to choose between the five survival parametric models in the in-sample test, while the prediction abilities are considerably reduced for models with log-normal and log-log distribution in the out-of-sample test. 5) The area under ROC is almost identical for the Cox survival model and the frailty survival model. 6) The univariate model (TLTA) shows the poorest performance in

Figure 5.1, while the survival analyses with log-normal and log-log distribution are less favourable in Figure 5.2.

[Figure 5.1, p.194 about here]

[Figure 5.2, p.195 about here]

Summary values of the area AUROC and the accuracy values for all the models are presented in Table 5.5. For each method, the table presents the in-sample prediction, out-of-sample prediction, and the rolling out-of-sample prediction for five years. All predictions show excellent (over 0.9) AUROC in the rolling out-of-sample prediction, except the univariate test. The random coefficient mixed logit model shows the best performance in all tests except the year out-of-sample test in 2007 using the sample from the year 2000-2006. The conditional probability models and the traditional multivariate discriminant analysis also show a strong performance. The semi-parametric survival model performs better than all the parametric survival models. Especially, the frailty survival model shows the best performance in the survival family. However, all the survival models show less accuracy than the conditional probability models, MDA, and the random discrete mixed logit model, in classifying default behaviour.

[Table 5.5, p.190 about here]

Comparing the accuracy ratios with previous studies, most of the models presented here have a higher accuracy ratio than the value of the Z-score model (0.79) in Agarwal & Taffler (2008) and the value of Moody's model (0.76) in Sobehart, Keenan & Stein (2001). This benefit in performance can be attributed to the optimal variable set and up-to-date comprehensive sample.

[Figure 5.3, p.196 about here]

Figure 5.3 compares the stability of the prediction ability according to the standard deviation of the AUROC over five years' rolling prediction. Although the sample size to

assess the credit stability is small, it includes the models' performance over the current financial crisis. As Standard & Poor's (2008) suggest, a period of moderate stress such as economic recession or financial crisis is the preferred period for evaluating credit stability.

The univariate (TLTA) method is the least stable model, since it does not contain market and economic information. During a moderate stress period, models with the optimal information set are less volatile and more likely to capture credit deterioration. The MDA is the most stable model, followed by the mixed logit model. The semi- parametric survival models are less stable than the parametric survival models and the conditional probability models. The probit model is slightly more stable than the logit model.

5.4.3 Misclassification Costs

This section employs the out-of-sample prediction to value the misclassification cost for different error types, because of the similarity between the out-of-sample and in-sample tests. Moreover, the out-of-sample test is closer to a practical application.

5.4.3.1 Classification Errors

Although the most efficient prediction model seeks to minimise both types of error, reducing one type of error is associated with increasing the other type. As previously discussed, the Type I and Type II rates depend on the selection of the cut-off points. For example, if the cut-off point is equal to 0.9, a company is classified as default if the expected default probability is over 90%, whereas a company is classified as non-default if the expected probability is less than 90%. Table 5.6 presents Type I, Type II and the overall classification accuracy over varying cut-off points for all the twelve models.

As the recovery rate and the interest rate for corporate bond and loans are varying across cases, it is hard to conclude which is the best method. However, one could discuss this in three scenarios. The first scenario is where Type I costs are higher than Type II costs, which is most likely and is suggested by previous studies (e.g. Lennox, 1999). In this case,

creditor loss is most of the principle but the return from the investment is much lower than the loss given default. For this investment, we would choose a low Type I error model with a bearable Type II error. Comparing the Type I error, MDA has the lowest classification error when the cut-off point is over 0.7, while the TLTA is least likely to class a healthy firm as a higher credit risk firm if the cut-off point is below 0.7. The frailty model and the Cox survival model also have a smaller Type I error than the survival parametric models and the conditional probability model. The second scenario is where Type I costs are lower than Type II costs, for example a bond with high credit spread, long term terminal value, and high recovery rate. In this case, we prefer a low Type II error with a bearable type I error. In terms of Type II errors, the probit model generates the least misclassifications. The logit and the mixed logit model also show lower classification errors than the survival models. Among the survival models, the log-logistic parametric model shows the least Type II classification errors. The traditional MDA method would have a similarly good performance with the conditional models if the cut-off point were under 0.01. The univariate TLTA model is most liable to misclassify a default firm as a low default risk firm. The third scenario is where the costs of the Type I and Type II are equal. In this case, the investment prefers a relatively low error for both types and a high overall prediction rate. If we restrict both errors to be smaller than 15% and the overall accuracy rate to be over 90%, only the MDA and the mixed logit model with cut-off point 0.01 satisfy the criteria.

[Table 5.5, p. 190 about here]

5.4.3.2 Economic Costs of Different Errors

The economic performance for financial institutions is compared in this thesis using different default prediction models. It assumes the loss given default to be 45% and the risk premium for a high quality investment (k) to be 0.3%, following Agarwal & Taffler (2008). The prior probability of failure is taken to be the same as the ex-post failure rate of

0.36% during the sample period.

[Table 5.6, p. 191 about here]

Table 5.6 presents the revenue, profit, Type I costs, Type II costs and other statistics under the competitive loan market described earlier. The creditors using the two random models (the mixed logit model and the frailty model) have the biggest market share at 27.79% and 27.47% respectively. This is because these two methods have a relatively good offer with a 0.32 average spread for the mixed logit model and 0.34 for the frailty model. Creditors using the MDA model have 10.39% market share. All creditors reject 0.29% of the total amount of bonds, which includes 4.47% of the total defaulted firms. Creditors using TLTA, the survival model with log-normal distribution and the survival model with log-log distribution accept 70% of default firms. This leads to a considerable number of Type I losses and eventually generates a negative return on the whole portfolio. A creditor using the TLTA model would have lost $33.82 million for the investment, while creditors using survival models, using log-normal distribution and log-log distribution would have lost $40.48 and $1.7 million respectively. The creditors using the mixed logit model and the frailty model are the most profitable with $82.92 and $88.07 million profit respectively. However, the creditors with the highest Return over Assets rate (0.37%) are those using the exponential distributed survival model.

[Table 5.7, p. 192 about here]

A creditor when using the survival model with Gompertz distribution would reject 0.96% of loans. This associates with 3.91 million potential profit losses, which amounts to 0.09% of its current investment assets. Those customers rejected by this creditor are reallocated to other creditors using other credit models. 0.27% of loans are rejected for all the other eleven methods. As the expected spread rate is different for various models, the opportunity loss for the same share loss is also varying. The lognormal and log-logistic distribution survival models have the second highest Type II costs, with 1.18 million. After

the OLOA of the Gompertz survival analysis (0.09%), the lognormal survival analysis, the TLTA method, the logit and the exponential survival methods also have substantial opportunity loss rates (0.042%, 0.041%, 0.041% respectively) compared with their assets size. The mixed logit model and the frailty model have the smallest opportunity loss rates. The MDA has the second smallest OLOA rate.

In a dynamic competitive credit market, creditors with a negative Return over Assets (ROA) would not survive. If a credit business starts with twelve creditor groups divided by the method they use to predict default risk, at each round, those creditors with negative ROA would exit the market or change their strategy for evaluating the default risk. Table 5.8 presents five rounds' economic costs of both Type I and II errors, after the existing credit groups redistribute the market. In the fifth round, the six credit groups all have a positive ROA. All creditors reject 1.53% of loans including 64.54% of defaulters. Considering both types of loss, the two random models are both excellent. Creditors using the frailty model achieve the highest ROA (0.236%), while creditors using the mixed logit bear the lowest opportunity loss rate (0.011%). Additional information in these random models largely supports creditors to achieve superior profit. It is a surprise that the traditional MDA model performs better than both logit and probit models according to both ROA and OLOA.

[Table 5.8, p. 193 about here]

5.6 Conclusion

This chapter compared the prediction performance of twelve models across five categories in the current financial situation. The results indicate that the models capable of capturing unobservable information have better prediction ability than the traditional methods. The mixed logit model is the most efficient and complete model in terms of overall prediction accuracy. Creditors using the frailty model and the mixed logit model achieve the best profits in the competitive credit market scenario. While lenders using the

frailty model generate the highest ROA with the second lowest potential profit loss rate, lenders avoid most opportunity loss with the second highest ROA under the mixed logit scheme. The mixed logit model is also relatively stable across the duration of a financial crisis. The results provide supplementary support for Chapter 4 and relevant studies including Hensher & Jones (2004, 2007). Moreover, the findings indicate that parametric survival models are not recommended for default prediction, in terms of either prediction stability or ability to generate profit for lenders. Although they show prediction abilities in default, they fail to create profit for lenders. This conclusion explains the selection preference of the Cox proportional hazard model in previous studies (Luoma & Laitinen, 1991; Duffie, Saita & Wang, 2007; Carling, Jacobson, Linde & Roszbach, 2007; Bharath & Shumway, 2008). Finally, although the traditional Z-score model is criticised by extensive studies, this chapter demonstrates its practical benefit in the competitive credit market structure. It is the most stable model, during a period of financial crisis and creditors who use this method will achieve a considerable profit, which is just slightly less than those using random models. This result is consistent with the conclusions of Agrawal & Taffler (2007), but provides a counterview to Chava & Jarrow (2004), who claim that the dynamic logit model is better than the traditional Z-score.

In practice, it is worth noticing that although the advanced random models (mixed logit model and the frailty model) outperform traditional models according to the criteria above, these models are less efficiency according to the calculation time in estimating the initial coefficients, as addressed in Chapter 4. However, once the coefficients have been estimated initially, the default risk of a new firm could be assessed within seconds for both the traditional and the advanced models.

Table 5.1

Subsample Details

Table 5.1 displays the sample structure by different sample periods. All observations are presented in quarterly bases.

Sample Period	Non-default Observation	Default Observation	Total Observation	Total Firm No.
2000	24,674	126	24,800	7,053
2001	23,201	203	23,404	6,663
2002	21,559	140	21,699	6,131
2003	20,021	103	20,124	5,665
2004	19,206	50	19,256	5,399
2005	18,526	52	18,578	5,212
2006	17,860	30	17,890	5,018
2007	17,303	40	17,343	4,886
2008	16,643	71	16,714	4,679
2009	15,599	120	15,719	4,336
2000-2004	108,661	622	109,283	8,504
2005-2009	85,931	313	86,244	6,456
2000-2009	194,592	935	195527	9,656

Table 5.2

Mathematical Methods for Comparison

Table 5.2 shows details of all default prediction models. $\boldsymbol{Z}$ is the default score. $\boldsymbol{P(Y_j = 1|X_j)}$ shows the default probability given information $\boldsymbol{X_j}$. $\boldsymbol{P(T < t|X_j)}$ represents the default time probability given information $\boldsymbol{X_j}$.

Methods		Mathematical formula
Discriminant analysis	Univariate	$Z = \frac{Total\ liability}{Total\ assets}$
	Multiple	$Z = \sum_{j=1}^{N} \beta_j X_j$
Conditional probability models	Probit	$P(Y_j = 1\|X_j) = \Phi\ (\sum_{j=1}^{N} \beta_j X_j)$
	Logit	$P(Y_j = 1\|X_j) = \frac{\exp\ (\sum_{j=1}^{N} \beta_j X_j)}{1 + \exp\ (\sum_{j=1}^{N} \beta_j X_j)}$
Discrete choice model	Mixed logit	$P(Y_j = 1\|X_j) = \int \frac{\exp(\sum_{j=1}^{N} \beta_j X_j)}{\sum \exp(\sum_{j=1}^{N} \beta_j X_j)} f(\beta\|b, W) d\beta$
Parametric Survival Analyses	Exponential	$P(T < t\|X_j) = 1 - S(t) = 1 - \exp\ (-\exp\ (\sum_{j=1}^{N} \beta_j X_j) t)$
	Weibull	$P(T < t\|X_j) = 1 - S(t) = 1 - \exp\ (-\exp\ (\sum_{j=1}^{N} \beta_j X_j) t^p)$
	Gompertz	$P(T < t\|X_j) = 1 - S(t) = 1 - \exp\ (-\exp\ (\sum_{j=1}^{N} \beta_j X_j) \gamma^{-1})$
	Lognormal	$P(T < t\|X_j) = 1 - S(t) = \Phi\ \left\{\frac{\log(t) - \sum_{j=1}^{N} \beta_j X_j}{\sigma}\right\}$
	Log-logistic	$P(T < t\|X_j) = 1 - S(t) = 1 - \left\{1 + \left(\exp\ (\sum_{j=1}^{N} \beta_j X_j) t\right)^{1/\gamma}\right\}^{-1}$
Semi-parametric survival analyses	Proportional Cox model	$P(T < t\|X_j) = 1 - S(t) = 1 - exp\left(-h_0(t) exp(\sum_{j=1}^{N} \beta_j X_j)\right)$
	Frailty-Cox Model	$P(T < t\|X_j) = 1 - S(t) = 1 - exp\left(-h_0(t) exp(\sum_{j=1}^{N} \beta_j X_j + \nu)\right)$

Table 5.3

Economic Costs for Different Error Types

Table 5.3 describes the different types of loss from misclassifications. The Type I misclassification relates to the loss of principle and market values. The Type II misclassification relates to the opportunity cost.

		Actual	
		Default firm	Healthy firm
Predicted	High default risk	Correct predicted	Type II misclassification loss: Opportunity costs (interest income loss)
	Low default risk	Type I misclassification loss: whole principle and potential loss in market value	Correct predicted

Table 5.4

Statistical Summaries of Default Predictions

Table 5.4 summarizes the mean of the default indicators from in-sample tests and out-of-sample tests. The abbreviations of methods are presented in Table 5.2.

	In sample 2000-2009				Out-of-sample 2005-2009			
Methods	ALL	Default	Non-default	T-test	ALL	Default	Non-default	T-test
TLTA	0.40822	0.76274	0.40651	-40.57	0.40912	0.79349	0.40772	-27.06
MDA	0.24788	0.59332	0.24622	-49.82	0.21367	0.60100	0.21226	-29.42
LOGIT	0.00478	0.17042	0.00399	-24.68	0.00235	0.10483	0.00198	-12.85
PROBIT	0.00477	0.15258	0.00406	-27.69	0.00179	0.08010	0.00151	-14.15
MIXLOGIT	0.00478	0.33724	0.00318	-32.54	0.00363	0.29838	0.00256	-17.61
S_EXOINENTIAL	0.03589	0.22085	0.03501	-25.28	0.04156	0.18877	0.04103	-13.42
S_GOMPERTZ	0.03585	0.22101	0.03496	-25.30	0.04154	0.18823	0.04100	-13.40
S_WEIBULL	0.03590	0.22077	0.03501	-25.27	0.04155	0.18896	0.04101	-13.39
S_LNORMAL	0.03626	0.20015	0.03547	-27.55	0.03536	0.12283	0.03504	-8.30
S_LOGLOG	0.03177	0.15831	0.03117	-26.46	0.03287	0.11986	0.03255	-9.04
S_COX	0.07351	0.67610	0.07061	-52.20	0.04285	0.58854	0.04086	-26.26
S_COXFRILTY	0.07338	0.67739	0.07048	-52.30	0.06299	0.67890	0.06075	-30.80
Observation	195527	935	194592	195527	86244	313	85931	86244

Table 5.5

AUROC and AR Comparison over Different Periods

Table 5.5 displays the overall accuracy comparison for in-sample, out-of-sample and rolling out-of-sample tests. The abbreviations of methods are presented in Table 5.2. Panel A provides results of the area under the receiver operating characteristic curve (AUROC). Panel B reports the results of the accuracy ratios.

Panel A: AUROC							
Methods	2000-2009	2005-2009	2005	2006	2007	2008	2009
TLTA	0.8126	0.8357	0.8445	0.7818	0.8606	0.8552	0.8250
MDA	0.9501	0.9550	0.9666	0.9534	0.9606	0.9488	0.9501
LOGIT	0.9590	0.9543	0.9691	0.9371	0.9814	0.9501	0.9367
PROBIT	0.9618	0.9565	0.9727	0.9404	0.9816	0.9606	0.9473
MIXLOGIT	0.9751	0.9757	0.9768	0.9487	0.9748	0.9751	0.9706
S_EXOINENTIAL	0.8942	0.8800	0.9520	0.9359	0.9745	0.9376	0.9230
S_GOMPERTZ	0.8944	0.8806	0.9522	0.9356	0.9739	0.9376	0.9234
S_WEIBULL	0.8940	0.8795	0.9519	0.9360	0.9749	0.9378	0.9445
S_LNORMAL	0.9000	0.7859	0.9445	0.9348	0.9722	0.9618	0.9373
S_LOGLOG	0.8873	0.7964	0.9434	0.9334	0.9729	0.9581	0.9235
S_COX	0.9384	0.9407	0.9571	0.9392	0.9644	0.9344	0.9081
S_COXFRILTY	0.9391	0.9425	0.9585	0.9395	0.9653	0.9352	0.9119
Panel B: Accuracy Ratio							
	2000-2009	2005-2009	2005	2006	2007	2008	2009
TLTA	0.6252	0.6714	0.6890	0.5636	0.7212	0.7104	0.6500
MDA	0.9002	0.9100	0.9332	0.9068	0.9212	0.8976	0.9002
LOGIT	0.918	0.9086	0.9382	0.8742	0.9628	0.9002	0.8734
PROBIT	0.9236	0.9130	0.9454	0.8808	0.9632	0.9212	0.8946
MIXLOGIT	0.9502	0.9514	0.9536	0.8974	0.9496	0.9502	0.9412
S_EXOINENTIAL	0.7884	0.7600	0.9040	0.8718	0.9490	0.8752	0.8460
S_GOMPERTZ	0.7888	0.7612	0.9044	0.8712	0.9478	0.8752	0.8468
S_WEIBULL	0.7880	0.7590	0.9038	0.8720	0.9498	0.8756	0.8890
S_LNORMAL	0.8000	0.5718	0.8890	0.8696	0.9444	0.9236	0.8746
S_LOGLOG	0.7746	0.5928	0.8868	0.8668	0.9458	0.9162	0.8470
S_COX	0.8768	0.8814	0.9142	0.8784	0.9288	0.8688	0.8162
S_COXFRILTY	0.8782	0.9170	0.8790	0.9306	0.8704	0.8238	0.8238

Table 5.6

Type I and Type II Errors

Table 5.6 presents Type I and Type II errors, and the overall classification accuracy over varying cut-off points for all the twelve models from 2005 to 2009 using the sample from 2000 to 2004. Panel A reports the Type I errors. Panel B displays the Type II errors. Panel C displays the overall prediction accuracy. The abbreviations of methods are presented in Table 5.2.

Panel A: Type I error

Cut off points	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.01	0.001	1.00E-04
TLTA (%)	100.00	51.44	38.98	26.84	19.81	12.78	9.58	4.15	0.64	0.00	0.00	0.00
MDA (%)	36.17	40.26	39.62	39.30	39.62	38.02	36.74	34.50	28.75	13.74	4.79	1.28
LOGIT (%)	100.00	100.00	99.68	99.04	97.12	93.93	88.82	82.11	68.37	22.36	5.43	2.24
PROBIT (%)	100.00	100.00	100.00	100.00	100.00	99.36	94.25	86.26	71.25	25.88	7.03	3.51
MIXLOGIT (%)	96.49	91.05	83.71	80.19	73.80	66.77	61.66	51.44	37.38	13.10	4.15	0.32
S_EXOINENTIAL (%)	98.08	96.81	95.85	93.93	92.33	88.82	82.75	70.29	39.62	3.83	0.32	0.00
S_GOMPERTZ (%)	97.76	97.12	95.85	94.57	92.01	89.46	82.75	69.01	38.98	3.83	0.32	0.00
S_WEIBULL (%)	97.76	97.44	95.85	94.57	92.01	89.46	83.07	69.65	39.30	3.83	0.64	0.00
S_LNORMAL (%)	98.72	96.49	95.53	94.89	94.25	93.29	91.05	83.07	64.86	24.92	10.86	5.75
S_LOGLOG (%)	100.00	98.08	97.12	95.85	94.89	93.29	90.42	81.47	65.18	24.60	10.22	3.51
S_COX (%)	65.18	58.47	52.72	46.65	42.81	35.78	30.03	23.64	16.29	3.83	0.64	0.00
S_COXFRILTY (%)	52.72	47.28	42.17	36.42	31.63	27.80	22.04	16.61	10.22	2.88	0.32	0.00

Panel B: Type II error

Cut off points	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.01	0.001	1.00E-04
TLTA (%)	0.00	9.44	11.86	15.83	22.54	33.69	49.26	68.95	90.70	100.00	100.00	100.00
MDA (%)	3.38	3.40	3.41	3.43	3.47	3.49	3.56	3.67	4.02	7.49	19.13	58.78
LOGIT (%)	0.00	0.00	0.00	0.00	0.01	0.02	0.06	0.10	0.27	3.06	20.55	56.77
PROBIT (%)	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.08	0.24	2.50	12.58	33.72
MIXLOGIT (%)	0.00	0.00	0.00	0.01	0.02	0.03	0.07	0.15	0.43	4.33	17.73	42.67
S_EXOINENTIAL (%)	0.18	0.50	0.73	0.95	1.27	1.62	2.17	3.61	8.90	53.60	85.71	97.07
S_GOMPERTZ (%)	0.23	0.52	0.71	0.98	1.30	1.64	2.16	3.63	8.93	53.23	85.24	96.93
S_WEIBULL (%)	0.23	0.52	0.71	0.98	1.30	1.64	2.16	3.61	8.94	53.12	85.08	96.86
S_LNORMAL (%)	0.11	0.50	0.80	1.12	1.48	2.04	2.64	4.09	8.62	32.45	51.64	64.61
S_LOGLOG (%)	0.01	0.14	0.52	0.77	1.15	1.71	2.57	4.38	8.67	29.82	53.52	70.80
S_COX (%)	0.50	0.68	0.89	1.15	1.52	2.06	2.97	4.78	9.54	41.66	72.41	91.62
S_COXFRILTY (%)	0.83	1.15	1.54	2.01	2.67	3.61	5.20	8.12	14.81	49.41	77.07	93.32

Panel C: Overall prediction accuracy

Cut off points	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.01	0.001	1.00E-04
TLTA (%)	99.64	90.40	88.04	84.13	77.47	66.39	47.41	31.29	9.63	9.63	0.36	0.36
MDA (%)	96.49	96.47	96.46	96.44	96.40	96.39	96.32	96.22	95.89	92.49	80.92	41.43
LOGIT (%)	99.64	99.64	99.64	99.64	99.64	99.64	99.61	99.60	99.48	96.87	79.51	43.43
PROBIT (%)	99.64	99.64	99.64	99.64	99.64	99.63	99.62	99.60	99.50	97.42	87.44	66.39
MIXLOGIT (%)	99.65	99.67	99.69	99.70	99.72	99.72	99.71	99.67	99.44	95.64	82.32	57.48
S_EXOINENTIAL (%)	99.47	99.15	98.92	98.71	98.40	98.06	97.54	96.15	90.99	46.58	14.60	3.28
S_GOMPERTZ (%)	99.42	99.13	98.94	98.68	98.37	98.05	97.55	96.14	90.96	46.95	15.07	3.43
S_WEIBULL (%)	99.42	99.13	98.94	98.68	98.37	98.04	97.54	96.15	90.95	47.06	15.23	3.49
S_LNORMAL (%)	99.53	99.15	98.85	98.54	98.18	97.63	97.04	95.62	91.17	67.58	48.51	35.60
S_LOGLOG (%)	99.63	99.50	99.13	98.89	98.51	97.96	97.11	95.34	91.12	70.20	46.64	29.45
S_COX (%)	99.26	99.11	98.92	98.68	98.33	97.81	96.94	95.15	90.43	58.48	27.85	8.71
S_COXFRILTY (%)	98.98	98.68	98.31	97.87	97.23	96.30	94.74	91.85	85.21	50.76	23.21	7.02

Table 5.7

Comparative Economic Costs of Different Default Models

Table 5.7 presents the economic values of different creditor groups using different default models from 2005 to 2009. Market share is the total number of issues as a percentage of total number of firm quarters. Share of defaulter is the number of the defaulters to whom a loan is granted over the total number of defaulters. Market size is assumed as $100 billion. Revenue is equal to market size ∗market share∗average credit spread. Loss is market size ∗ prior probability of default ∗ share of defaulter∗ loss given default. Profit is Revenue-Loss. Return on assets (ROA) is **profit**/(**market size** ∗ **market share**). Opportunity loss (OL) is market size ∗share loss∗average credit spread. The opportunity loss over assets (OLOA) refers to the opportunity loss over assets. Reject refers to issuers who have been rejected by all creditor groups. The abbreviations of methods are presented in Table 5.2.

	Market share %	Share of defaulter %	Average spread %	Revenue ($m)	Loss ($m)	Profit ($m)	ROA %	Share loss %	OL ($m)	OLOA %
Rejected	0.29	4.47								
TLTA	2.74	28.12	0.43	11.73	45.55	-33.82	-1.23	0.27	1.15	0.042
MDA	10.39	3.19	0.34	35.19	5.17	30.02	0.29	0.27	0.91	0.009
LOGIT	2.15	2.88	0.33	7.10	4.67	2.44	0.11	0.27	0.89	0.041
PROBIT	4.45	2.88	0.33	14.65	4.67	9.98	0.22	0.27	0.89	0.020
MIXLOGIT	27.79	4.15	0.32	89.65	6.72	82.92	0.30	0.27	0.87	0.003
S_EXOINENTIAL	2.69	0.64	0.41	10.94	1.04	9.90	0.37	0.27	1.09	0.041
S_GOMPERTZ	4.35	2.56	0.41	17.74	4.15	13.59	0.31	0.96	3.91	0.090
S_WEIBULL	5.96	3.19	0.41	24.23	5.17	19.06	0.32	0.27	1.09	0.018
S_LNORMAL	1.75	29.71	0.44	7.65	48.13	-40.48	-2.31	0.27	1.18	0.067
S_LOGLOG	4.21	12.46	0.44	18.49	20.19	-1.70	-0.04	0.27	1.18	0.028
S_COX	5.80	2.56	0.34	19.76	4.15	15.62	0.27	0.27	0.92	0.016
S_COXFRILTY	27.45	3.19	0.34	93.24	5.17	88.07	0.32	0.27	0.91	0.003

Table 5.8

Economic Costs in a Dynamic Competitive Credit Market

Table 5.8 presents the dynamic process in the competitive credit market. ROA refers to return over assets and OLOA refers to opportunity loss over assets. Models generating negative ROA would not appear in the following round of competition. The abbreviations of methods are presented in Table 5.2.

	1		2		3		4		5	
	ROA %	OLOA %	ROA %	OLOA %	ROA %	OLOA %	ROA %	OLOA %	ROA %	OLOA %
TLTA	-1.234	0.042								
MDA	0.289	0.009	0.277	0.030	0.279	0.029	0.277	0.373	0.254	0.034
LOGIT	0.113	0.041	0.134	0.133	0.138	0.131	0.154	0.363	0.152	0.135
PROBIT	0.224	0.020	0.222	0.065	0.227	0.062	0.234	0.362	0.236	0.064
MIXLOGIT	0.298	0.003	0.291	0.010	0.293	0.010	0.293	0.355	0.296	0.011
S_EXOINENTIAL	0.368	0.041	0.311	0.120	0.031	0.081	-0.151	0.447		
S_GOMPERTZ	0.312	0.090	0.177	0.076	-0.079	0.044				
S_WEIBULL	0.320	0.018	-0.155	0.047						
S_LNORMAL	-2.313	0.067								
S_LOGLOG	-0.040	0.028								
S_COX	0.269	0.016	0.208	0.326	0.210	0.052	0.205	0.375	0.163	0.063
S_COXFRILTY	0.321	0.003	0.301	0.325	0.301	0.012	0.302	0.374	0.299	0.015
Total rejected %	0.29		1.08		1.08		1.10		1.53	
Default rejected %	4.47		34.19		34.19		35.78		64.54	

Figure 5.1

The AUROC of Twelve Models for In-sample Tests

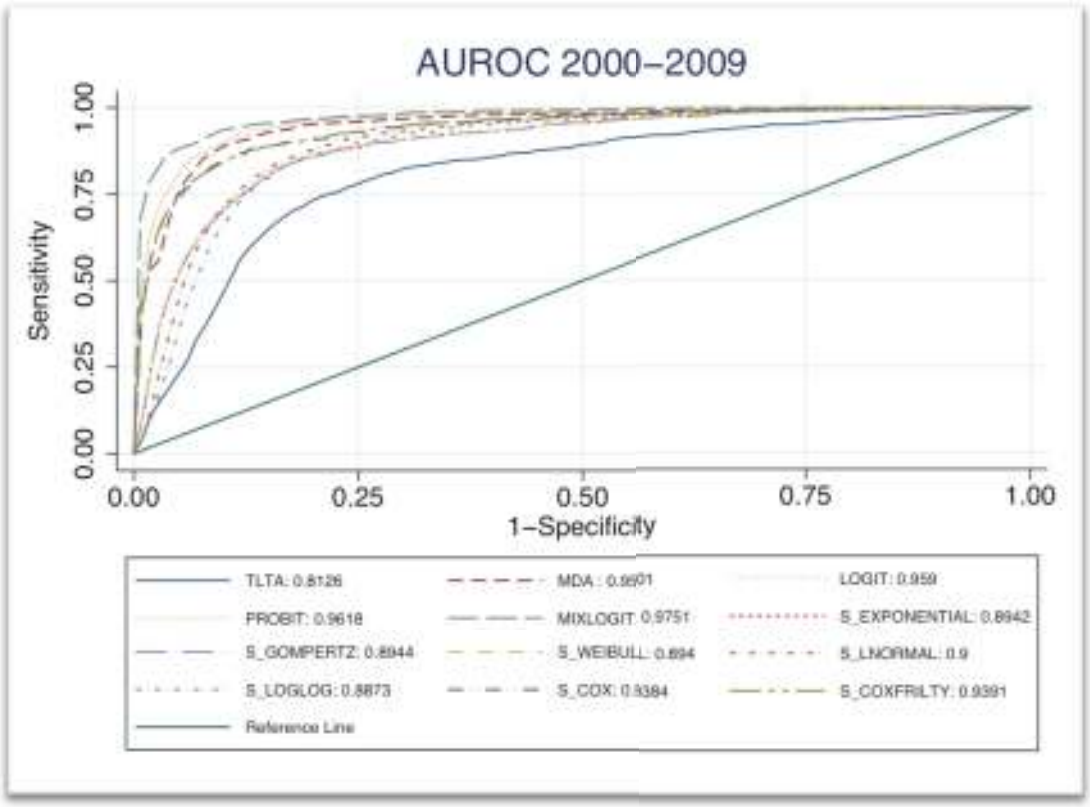


Figure 5.1 describes the AUROC for alternative models using the data from 2000 to 2009. The reference line indicates random model with no prediction ability.

Figure 5.2

The AUROC of Twelve Models for Out-of-sample Tests

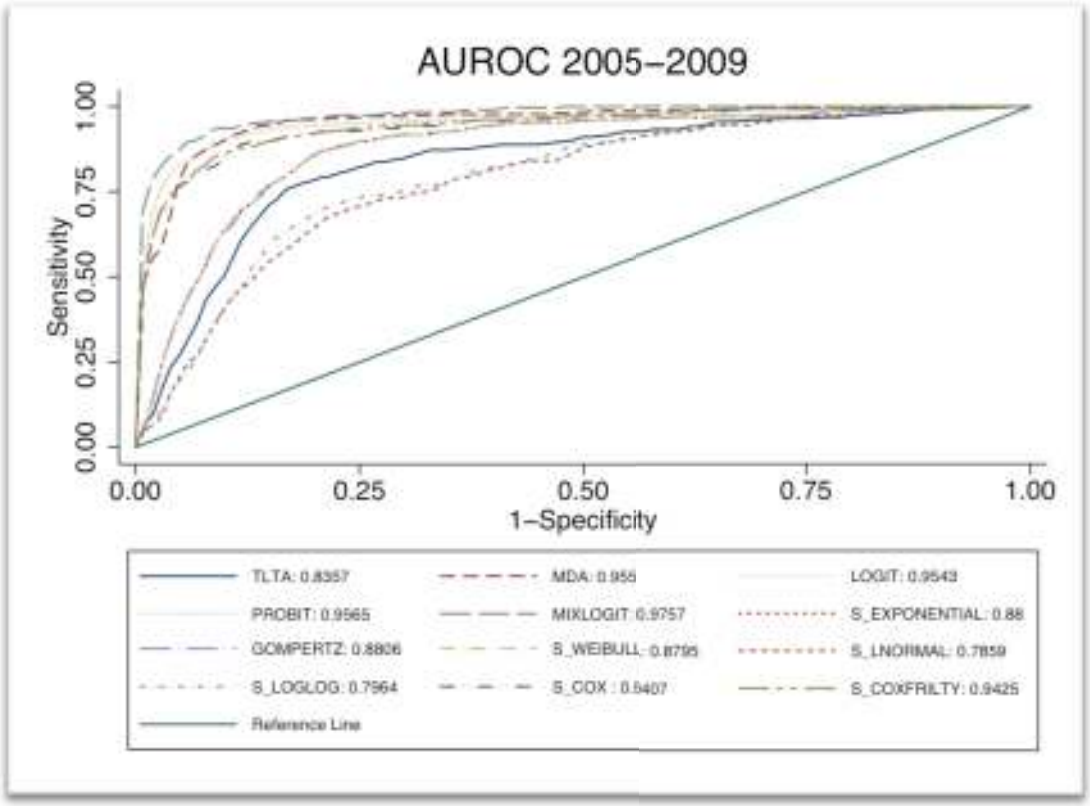
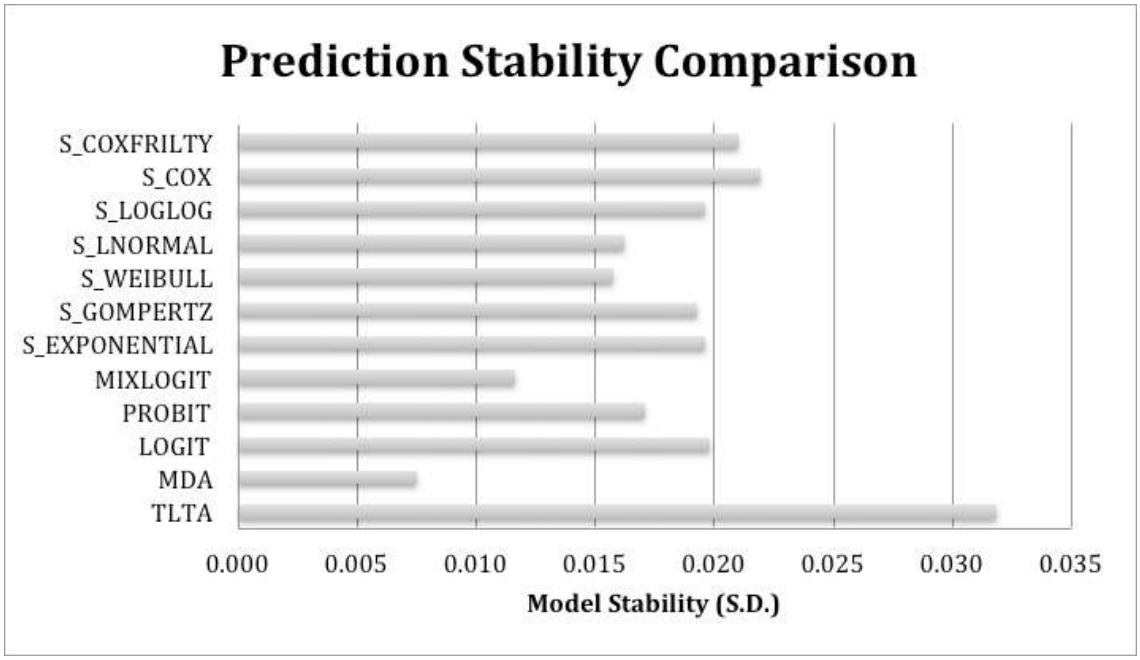


Figure 5.2 describes the AUROC for alternative models from 2005 to 2009 using the data from 2000 to 2004. The reference line indicates random model with no prediction ability.

Figure 5.3

Predicting Stability with Rolling Out-of-sample AUROC



This table compares the model stability under financial crisis. The TLTA model is the least stable model while the MDA is the most stable model.

Chapter 6: Conclusion and Recommendations

6.1 Summary of Research

This thesis empirically investigates prediction methods to calculate corporate default probability for firms in the United States. It contributes to the study of financial distress prediction methods for capturing default probabilities. In the four main chapters (2-5), the thesis has examined the optimal default information set and the preferred prediction model from different angles. Chapter 2 gave a comprehensive picture of previous corporate default studies, on both the default information and prediction methods. Chapter 3 discussed the optimal information set for default prediction under the dynamic logit model. Chapter 4 investigated two new random models under alternative sets of default information. These random models capture default cluster and unobservable information and thus improve the prediction accuracy. Chapter 5 conducted a comprehensive comparison of large dimension methods under the optimal information set, and gave suggestions for an optimal prediction model for future application. The empirical test was based on the most comprehensive and up-to-date default information collected from four sources. The whole sample contained 639,573 observations with 2,123 defaults across forty years.

In exploring the default drivers, this thesis began with a summary of significant variables from previous empirical studies of financial distress, and then graded these factors by their significance levels (see Figure 3.7) using feature selection techniques. These two procedures allowed an optimal default information set to be generated from numerous observable default drivers. Apart from the superior prediction accuracy with applying the optimal variable set in alternative models, the contributions of the optimal variable set

extend into two more directions. Firstly, it recognises the equally important but unique role of market information and accounting information in predicting default, through both traditional models and new random models. This provides evidence of market inefficiency in explaining the default behaviour. Secondly, it determined the relationship between the macroeconomic cycle and default risk cycle. The significance of macroeconomic variables was shown in alternative models. Most importantly, through the frailty survival model, it was found that default probability forms clusters in the macroeconomic cycle. Based on the optimal variable set, the thesis develops current literature with two additional results. First, this thesis verifies that industry effects contribute to default risk. The retail industry shows the highest default risk, while the financial sector and public sector keep a relatively low default risk profile. Second, the thesis confirms the existence of unobservable variables, which could be captured using advanced random models.

By tackling the question of the optimal default prediction models, this thesis adds to the existing research literature in two areas. First, the thesis improves the prediction performance of two newly developed random models, namely the mixed logit model and the frailty model. Missing observable variables are detected in both models. The significances of random effects are determined under the optimal information set for both models. For the mixed logit model, the thesis verifies the results with alternative assumptions. For the frailty model, the thesis explored default clusters in three categories: industry frailty, macroeconomic frailty and calendar frailty. It further contributes to the literature by comparing these two random models. Although the frailty model has greatest computational efficiency, the mixed logit model predicts default risk more accurately. The thesis then compares and explores the prediction performances of large dimension measurements in the recent financial environment, under alternative criteria. Twelve models in five categories were compared under the same information set according to prediction accuracy, prediction stability, economy benefit, and the misclassification cost of both the Type I and Type II errors. The practical advantage of the random model showed

that creditors using the frailty model and the mixed logit model achieve the best profits in competitive credit markets. The traditional Z-score model showed its benefit of stability during financial crisis, and a considerable profitability for creditors. Furthermore, the parametric survival models and univariate model were extracted from the competitive credit market for poor economic performance; thus they are not recommended for future application.

All the details of these conclusions and contributions are presented in section 2 below. Section 3 then summarises the limitations of the thesis, and also gives suggestions for further research.

6.2 Conclusions

6.2.1 Optimal Information Set

This thesis determined the optimal information set using the following steps. Firstly, a multidimensional extensive variable set was constructed according to the frequencies of existent significance shown from previous empirical tests. This thesis then extended the variable set with additional variables such as default history, firm age and alternative macroeconomic variables. This procedure secures the maximum integrity and reflects the range of available quantitative information. Second, this thesis extracted core variables from the original information set using the t-test and stepwise method based on the dynamic logit model. This procedure secures the computational efficiency of the prediction models by removing repeat and highly correlated information from the original set. Thirdly, this thesis validated the superiority of the optimal variables set over the information set from previous studies using the dynamic logit models. Fourthly, this thesis conducted robustness checks of the optimal information set with alternative subsample periods and different prediction horizons. This thesis furthermore verified the preference of the optimal variables set over single information sets with alternative models, including

the mixed logit model and the survival analysis models.

This thesis suggested a nine-tier significance figure of default information. The optimal variable set established in this thesis includes information of accounting ratios, stock market indicators, macroeconomic indicators, firm age and default history. It contains 26 out of the original 125 variables but includes 99.25% of their prediction ability. The core accounting ratios in the optimal variable set contain: profitability indicators (NITA, EBITA, EBIATL), operation efficiency indicators (SALEG, CFTL), liquidity indicators (QACL, WCTA, CASHTA, CLTA), capital structure indicators (METL) and firm size indicators (LOGTAGNP). The core market information in the optimal variable set includes firm market size, stock return, stock volatility, quarterly share price gap, Earnings per Share, share price, and excess return. The macroeconomic variables in the optimal information set include the whole stock market indicator (SPRETURN), the economic cycle indicators (GNP, GNPL), the interest rate indicator (CD1M) and the bank lending and investment indicators (INVEST, INVESTL). The CD1M, INVEST and INVESTL are the new default indicators developed in this thesis. Moreover, industry effects show significance over the optimal variable set in the traditional models and the random models. Retail is the most risky sector, while the public and financial sectors exhibits the lowest risk of default. Preference for the optimal information set over previous benchmark empirical studies was shown using a dynamic logit model. It also out-performs single variable information sets with not only the dynamic logit model, but also the mixed logit and survival models. The results are robust in the in-sample, out-of-sample, and rolling out-of-sample tests. Beyond the optimal information set, unobservable information was detected using the random models.

The results show that default history and the Net Income ratio are the most important factors in predicting default. Firms are more likely to default if they suffered financial distress before. Also, firms suffer more default risk if they are less profitable. Firm size and Liability ratios are second in significance. Although accounting ratios and market information play a fundamental role in default prediction, macroeconomic variables also

add moderate explanatory power. Firms are more likely to default if the operation is less efficient, with a lower sales growth rate and limited operating income. High leverage firms suffer more credit risk. A big firm (in asset size) is more likely to default, while a smaller market capacity is associated with a higher default risk. Firm stock return is negatively related to default probability. It is interesting to point out that the results show some that market variables, such as Earnings per Share and Stock Price, become more important in a longer-horizon prediction, while most accounting variables become less so.

The findings also contribute to the debate on the market efficiency theory. The market efficiency hypothesis has been tested with very few exceptions. Consistency is found across alternative markets (Jensen, 1978). I acknowledge that although the power of the market information is addressed with the highest frequency relative to the other theory-related patterns, it cannot be regarded as the sole predominant corporate default driver. This thesis shows that market information does not explain all default risk. Other information, especially accounting ratios and default history, are equally vital. While market information explains moderate default risk, the combined information set improves prediction accuracy considerably. This conclusion indirectly reveals that share price does not reflect all publicly available information thereby violating the principle of market efficiency.

In addition, part of the findings are consistent with Hol (2007), Couderc & Renault (2005), and Carling, Jacobson, Linde & Roszbach (2007), in that they identify that macroeconomic indicators increase the prediction accuracy of default risk. In the findings, this result is robust in alternative models. In particular, macroeconomic variables show stronger influence on default risk in a relatively long period test. This thesis also found the significance of macroeconomic frailty in the Cox survival analysis. Although the macroeconomic frailty is significant, there is no clear pattern showing the relation between the business cycle and credit cycle. However, the calendar frailty factor clearly shows that, during the economic crisis, adjusting the default risk to a higher probability of

default is necessary.

The findings furthermore support the existence of industry effects on default prediction. This result is consistent with Chava & Jarrow (2004) and Jones & Hensher (2004), but contrast with Campbell, Hilscher & Szilagyi (2008). The result is robust in longer-horizon predictions. In addition, the industry effect is more apparent for private firms. This thesis further detected industry effects with the frailty model, which concluded that the retail and agriculture, forestry, fishing industry sector suffers the highest default risk, while the public sector and financial sector are of the lowest default risk.

6.2.2 Optimal Prediction Methods

This thesis investigated the preferred prediction models under two main scopes: first, this thesis investigated and compared two newly developed random models; second, this thesis compared the performance of these two random models with other prediction methods through various measurements.

In the first scope, this thesis investigated two random models separately, followed by a comparison between them. In exploring the mixed logit model, this thesis first investigated the performance of the model across different variable sets of both public and private firms. Most importantly, this thesis added market variables into the observable information set. The comparison between logit model and mixed logit model was conducted using the optimal information set. The research further tested the robustness of the model with alternative coefficient assumptions and simulation times. In investigating the frailty model, this thesis first tested the normal Cox survival model with different information sets. In particular, departing from previous studies, this thesis added accounting information into the analyses. Then, it went further to explore potential frailty factors with alternative shared observable variables. Finally, this thesis compares the performance of these two random models, both theoretically and empirically.

Theoretically, the mixed logit model relaxes the restrictive IID and IIA assumptions, and incorporates random coefficients to capture observable and unobservable factors between and within firms. Also, the frailty model incorporates random structures as an additional explanatory variable. Both models capture the default cluster and unobservable information. Empirically, this thesis demonstrated that the mixed logit model out-performs the logit model, and the frailty survival model out-performs the normal survival models. Most importantly, this thesis improved the prediction accuracy of the random models by using the optimal information set. Market information improves the prediction power of the mixed logit model, while accounting information adds prediction accuracy to the frailty model. The maintained significance of the random coefficients of the mixed logit model shows robustness across alternative assumptions, including different simulation times and alternative coefficient assumptions. Furthermore, the frailty model detects three potential frailties of the default cluster: the industry frailty, the macroeconomic frailty and the calendar frailty. In the out-of-sample test, this thesis concluded that the mixed logit model has a better prediction accuracy, while the frailty model is more computationally efficient.

This thesis further compared the prediction performance of twelve models across five categories in the current financial situation. The comparison was based on not only the in-sample test, but also out-of-sample and rolling out-of-sample tests. To establish an identical scale, this thesis transferred the default score of the discriminant models and the hazard function of the survival models into default probability. The comparison methods of prior research mainly focus on the Type I and Type II errors over alternative cut-off points (Lennox, 1999; Lifschutz, 2010; Lifschutz & Jacobi, 2010). In this thesis, this thesis applied other comparison standards, such as AUROC, and prediction stability during financial crisis. Most importantly, this thesis addressed the economic costs for both Type I one and Type II errors and the profitability of creditors using alternative default measurements under the comparative credit markets.

The results showed the practical superiority of the advance random models. Creditors using the mixed logit model or the frailty model achieve the highest ROA, while bearing the lowest opportunity losses. Moreover, the mixed logit model is the most accurate predictor. Among the conditional probability models, the probit model slightly outperforms the logit model in terms of prediction accuracy, prediction stability, and economic benefit and costs. The traditional Z-score model is the most stable during financial crisis, and creditors using this model achieve relatively high profit. In the survival models, the semi-parametric model out-performs parametric models. Although the univariate model and all parametric survival models capture default information, it is not recommended for further application due to its poor performance in the competitive credit model.

What are the implications of these findings? The results and their conclusions are of importance to creditors, investors, regulators and other financial institutions and agencies, to assess the financial health of firms. This thesis will help banks and other creditors to choose the best default probability measurement method to profit maximise with the lowest opportunity losses and mitigate risk. The suggested optimal information set and optimal model could enhance the assessment accuracy of rating agencies. The thesis also provides references for regulators and governments to develop an effective financial regulatory system. Moreover, the framework of both the mixed logit model and the frailty model could extend to other research, including corporate event studies such as mergers and acquisitions, IPO and SEO, and other credit risk studies such as personal credit assessment, and government default risk assessment. I thus see my thesis as a stepping-stone for further research to look more closely into default risk assessment, as well as the application of advanced random models in other financial areas.

6.3 Research Limitations and Further Research Suggestions

While this thesis provides the most comprehensive evidence on corporate default prediction to date, several topics are worth pursuing in future research. The present study can be extended internationally by using samples from other countries, such as the United Kingdom, EU countries, and emerging economies. Although several studies conduct comparative default predictions between countries (e.g. Beynon & Peel, 2001; Park & Han, 2002; Becchetti & Sierra, 2003;Lin & Piesse, 2004; Agarwal & Taffler, 2007; Abdullah, 2008; Vuran, 2009; Wang & Campbell, 2010), there are gaps in the use of feature selection techniques, advance random models, and large dimension models. It would be worth investigating the performance of the optimal information set in other countries. It would be also interesting to investigate cross-country and cross-culture variations of corporate default probabilities.

Moreover, for the significance of industry effects, it would be worthwhile to test and compare default risk measurements within each industry, especially those with high default risk such as retail. Further important questions include whether default is contagious between industries, which would involve investigating the links and chains of default clusters between industries. In other words, it would be interesting to investigate if the default cluster of one industry could trigger the default cluster of another industry. Equally it would be informative to investigate cross-industry variations.

In attempting to select the optimal information set, this thesis used the backward stepwise method after a comparison with other approaches. However, one design problem is that this method does not contain all the possible variable combinations. There might be other combinations that perform better. However, with such an extensive number of variables, a full test of all possible sets would be expensive in time. For the feature selection process, this thesis used the dynamic logit model employed by Shumway (2001), Chava & Jarrow (2004), and Campbell, Hilscher & Szilagyi (2008), for its popularity and simplicity.

Although the optimal information set shows significance in alternative models, it would be interesting to expand and combine the stepwise feature selection method with other advance models, such as the suggested mixed logit model or the frailty model.

Leading on from Chapter 4, further research could extend the details of the newly established models. For example, for the mixed logit model, this thesis assumed the coefficients as normally distributed for simplicity. It would be worth examining whether different distribution assumptions of the coefficients in the mixed logit model could cause prediction differences. For the frailty model, this thesis detected three shared frailties. There might be other frailties within the same firm. Thus, it would be worth investigating other forms of frailties.

In the comparison section of Chapter 5, this thesis specified the mixed logit model as 100 Halton draws with the independent coefficients assumption, because of its computational efficiency. However, as shown in Chapter 4, the mixed logit model with correlated assumptions and higher simulation times is more likely to be accurate. Moreover, as it summarised in Chapter 2, there are other methods in predicting corporate default, such as the neutral network, decision tree etc. This thesis explored and compared most statistical models with alternative standards. It would be interesting to extend the comparative methods used in Chapter 5 with other predicting models including intelligent techniques, market-based structure model, and other advance choice models.

Bibliography

Abdullah, N. A. (2008). Predicting corporate failure of Malaysia's listed companies: comparing multiple discriminant analysis, logistic regression and the hazard model. *International Research Journal of Finance and Economics* 15, 201-217.

Abidali, A. F., & Harris, F. (1995). A methodology for predicting company failure in the construction industry. *Construction Management and Economics* 13, 189-196.

Agarwal, V., & Taffler, R. J. (2008). Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking & Finance* 32, 1541–1551.

Agarwal, V., & Taffler, R. J. (2007). Twenty-five years of the Taffler z-score model: does it really have predictive ability? *Accounting and Business Research* 37, 285-300.

Ahn, B. S., Cho, S. S., & Kim, C. Y. (2000). The integrated methodology of rough set theory and artificial neural network for business failure prediction. *Expert Systems with Applications* 18, 65-74.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance* 23, 589–609.

Altman, E. I. (2000). Predicting Financial Distress of Companies: Revisiting the Z-Score and Zeta® Models. Working paper, New York University.

Altman, E. I., & Narayanan, P. (1996). *Business failure classification models: An international survey.* Working paper, New York University Stern School of Business Finance Department.

Altman, E. I. (1984). The successes of business failure: An International Survey. *Journal of Banking and Finance* 8, 171-198.

Altman, E. I., & Saunders, A. (1998). Credit risk measurement: Developments over the last 20 years. *Journal of Banking & Finance* 21, 1721-1742.

Altman, E. I., Haldeman, R. G., & Narayanan, P. (1977). Zeta Analysis: A new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance* 1, 29-54.

Amato, J. D., & Furfine, C. H. (2004). Are credit ratings procyclical? *Journal of Banking and Finance* 28, 2641-2677.

Atiya, A. F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks* 12, 929-935.

Aziz, A., Emanuel, D. C., & Lawson, G. H. (1988). Bankruptcy prediction: An investigation of cash flow based models. *Journal of Management Studies* 25, 419-437.

Aziz, M. A., & Dar, H. A. (2006). Predicting corporate bankruptcy: Where do we stand? *Corporate Governance* 6, 18-33.

Back, B., Laitinen, T., & Sere, K. (1996). Neural Networks and Genetic Algorithms for bankruptcy predictions. *Expert Systems with Application* 11, 407-413.

Balcaen, S., & Ooghe, H. (2006). 35 years of studies on business failure: An overview of the classical statistic methodologies and their repeated problems. *British Accounting Review* 38, 63-93.

Balcaen, S., & Ooghe, H. (2004). Alternative methodologies in studies on business failure: Do they produce better results than the classic statistical methods? Working Papers of Faculty of Economics and Business Administration, Ghent University, Belgium 04/249.

Banachewicz, K., Lucas, A., & Vaart, A. V. (2008). Modelling portfolio defaults using hidden Markov models with covariates. *Econometrics Journal* 11, 155-171.

Barnes, P. (1990). The prediction of takeover targets in the UK by means of multiple discriminant analysis. *Journal of Business Finance & Accounting* 17, 73-84.

Barniv, R., Agarwal, A., & Leach, R. (1997). Predicting the outcome following bankruptcy filing: A three-state classification using neural networks. *Intelligent Systems in Accounting Finance and Management* 6, 177-194.

Bamiv, R., Agarwal, A., & Leach, R. L. (2000). Predicting Bankruptcy Resolution. *Journal of Business Finance & Accounting* , 497-520.

Beaver, W. H. (1966). Financial Ratios as Predictors of Failure. *Journal of Accounting Research* 4, 71-111.

Beaver, W. H., McNicholes, M. F., & Rhie, J. (2005). Have financial statements become less informative? Evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting Studies* 10, 93-122.

Becchetti, L., & Sierra, J. (2003). Bankruptcy risk and productive efficiency in manufacturing firms. *Journal of Banking & Finance* 27, 2099-2120.

Begley, J., Ming, J., & Watts, S. (1996). Bankruptcy Classification Errors in the 1980s: An Empirical Analysis of Altman's and Ohlson's Models. *Review of Accounting Studies* 1, 267-284.

Bellovary, J. L., Giacomino, D. E., & Akers, M. D. (2007). A Review of Bankruptcy Prediction Studies: 1930 to Present. *Journal of Financial Education* 33, 1-43.

Betts, J., & D.Belhoul. (1987). The Effectiveness of Incorporating Stability Measures In Company Failure Models. *Journal of Business Finance & Accounting* 14, 323-334.

Beynon, M. J., & Peel, M. J. (2001). Variable precision rough set theory and data discretisation: An application to corporate failure prediction. *Omega: The International Journal of Management Science* 29, 561-576.

Bharath, S. T., & Shumway, T. (2008). Forecasting default with the Merton distance to default model. *The Review of Financial Studies* 21, 1339-1369.

Bhargava, M., Dubelaar, C., & Scott, T. (1998). Predicting bankruptcy in the retail sector: An examination of the validity of key measures of performance. *Journal of Retailing and Consumer Services* 5, 105-117.

Blöchlinger, A., & Leippold, M. (2006). Economic benefit of powerful credit scoring. *Journal of Banking and Finance* 30, 851–873.

Blum, M. (1974). Failing Company Discriminant Analysis. *Journal of Accounting Research* 12, 1-25.

Bonfim, D. (2009). Credit risk drivers: Evaluating the contribution of firm level information and of macroeconomic dynamics. *Journal of Banking and Finance* 33, 281-299.

Bryant, S. M. (1997). A Case-based reasoning approach to bankruptcy prediction modelling. *Intelligent Systems in Accounting, Finance and Management* 6, 195-214.

Bystrom, H., & Kwon, O. (2007). A simple continuous measure of credit risk. *International Review of Financial Analysis* 16, 508-523.

Campbell, Y. J., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *Journal of Finance* 63, 2899-2939.

Carling, K., Jacobson, T., Linde, J., & Roszbach, K. (2007). Corporate credit risk modeling and the macroeconomy. *Journal of Banking and Finance* 31, 845-868.

Charalambous, C., Charitou, A., & Neophytou, E. (2000). Predicting corporate failure: Empirical evidence for the UK. Pre-print working paper. http://eprints.soton.ac.uk/36125/1/01-173.pdf

Charitou, A., Neophytou, E., & Charalambous, C. (2004). Predicting Corporate Failure: Empirical evidence for the UK. *European Accounting Review* 13, 465-497.

Chava, S., & Jarrow, R. A. (2004). Bankruptcy prediction with industry effects. *Review of Finance* 8, 537-569.

Chava, S., & Purnanandam, A. (2010). Is default risk negatively related to stock returns? *The Review of Financial Studies* 23, 2524-2559.

Chi, L., & Tang, T. (2006). Bankruptcy prediction: Application of logit analysis in export credit risks. *Australian Journal of Management* 31, 17-27.

Chuang, S., Cai, T., Douglass, C., Wei, L., & Dodson, T. (2005). Frailty approach for the analysis of clustered failure time observations in dental research. *Journal of Dental Research* 84, 54-58.

Clarke, J., Ferris, S. P., Jayaraman, N., & Lee, J. (2006). Are analyst recommendations biased? Evidence from corporate bankruptcies. *Journal of Financial And Quantitative Analysis* 41, 169-196.

Couderc, F., & Renault, O. (2005). Time-to-Default: Life cycle, global and industry cycle impacts. *Working paper, University of Geneva and FAME*.

Crouhy, M., Galai, D., & Mark, R. (2000). A comparative analysis of current credit risk models. *Journal of Banking & Finance* 24, 59-117.

Dahiya, S., Saunders, A., & Srinivasan, A. (1994). Financial distress and bank lending relationships. *The Journal of Finance* 58, 375-399.

Das, S. R., Duffie, D., Kapadia, N. & Saita, L. (2007). Common failings: How corporate defaults are correlated. *The Journal of Finance* 8, 537–569.

Das, S. R., Hanouna, P., & Sarin, A. (2009). Accounting-based versus market-based cross-sectional models of CDS spreads. *Journal of Banking & Finance* 33, 719-730.

Daubie, M., & Meskends, N. (2002). Business Failure Prediction: A review and analysis of the literature. In C. Zopounidis (ed.), *New Trends in Banking Management* (pp. 71-86). New York: Physica-Verlag.

Deakin, E. B. (1972). A discriminant analysis of predictors of business failure. *Journal of*

Accounting Research 10, 167-179.

Dichev, I. D. (1998). Is the Risk of Bankruptcy a Systematic Risk? *The Journal of Finance* 53, 1131-1147.

Dimitras, A. D., Zanakis, S., & Zopounidis, C. (1996). A survey of business failure with an emphasis on prediction method and industrial applications. *European Journal of Operational Research* 3, 487–513.

Duffie, D., Eckner, A., Horel, G., & Saita, L. (2009). Frailty correlated default. *Journal of Finance* 64, 2089-2123.

Duffie, D., Saita, L., & Wang, K. (2007). Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics* 83, 635-665.

EI Hennawy, R., & R. C. Morris. (1983). The significance of base year in developing failure prediction models. *Journal of Business Finance & Accounting* 10, 209-223.

Eisdorfer, A. (2008). Empirical evidence of risk shifting in financially distressed firms. *The Journal of Finance* 2, 609-637.

Ezzamel, M., & Mar-Molinero, C. (1990). The distributional properties of financial ratios in UK manufacturing companies. *Journal of Business Finance & Accounting* 17, 1-29.

FitzPatrick, P. J. (1932). *A Comparison of the Ratios of Successful Industrial Enterprises with Those of Failed Companies*. Washington: The Accountants Publishing Co.

Foreman, R. (2003). A logistic analysis of bankruptcy within the US local telecommunications industry. *Journal of Economics and Business* 55, 135–16.

Franzen, L., Rodgers, K. J., & Simin, T. T. (2007). Measuring Distress Risk: The Effect of R&D Intensity. *Journal of Finance* 62, 2931-2967.

Frydman, H., Altman, E. I., & Kao, D. (1985). Introducing recursive partitioning for financial

classification: the case of financial distress. *The Journal of Finance* XL, 269-291.

Gilbert, L. R., Menon, K., & Schwartz, K. B. (1990). Predicting bankruptcy for firms in financial distress. *Journal of Business Finance & Accounting* 17, 161-171.

Gilson, S. C. (1989). Management turnover and financial distress. *Journal of Financial Economics* 25, 241-262.

Gordy, M. B. (2000). A comparative anatomy of credit risk models. *Journal of Banking & Finance* 24, 119-149.

Hensher, D. A., & Jones, S. (2007). Forecasting corporate bankruptcy: Optimizing the performance of the mixed logit model. *Abacus* 43, 241-264.

Hensher, D. A., Jones, S., & Greene, W. H. (2007). An error component logit analysis of corporate bankruptcy and insolvency risk in Australia. *The Economic Record* 83, 86-103.

Hillegest, S. A., Keatin, E. K., Cram, D. P., & Lundstedt, K. G. (2004). Assessing the probability of bankruptcy. *Review of Accounting Studies* 9, 5-34.

Hodges, C., Clusker, G. R., & Lin, B. (2005). Analyzing bankruptcy predictors using time series data. *Journal of Accounting and Finance Research* 13, 159-168.

Hol, S. (2007). The influence of the business cycle on bankruptcy probability. *International Transactions in Operational Research* 14, 75-90.

Hu, Y., & Ansell, J. (2009). Retail default prediction by using sequential minimal optimization technique. *Journal of Forecasting* 28, 651-666.

Hwang, R., Cheng, K. F., & Lee, J. C. (2007). A semiparametric method for predicting bankruptcy. *Journal of Forecasting* 26, 317-342.

Izan, H. Y. (1984). Corporate distress in Australia. *Journal of Banking and Finance* 8, 303-320.

Jensen, M. C. (1978). Some anomalous evidence regarding market efficiency. *Journal of Financial Economics* 6, 95-101.

Jo, H., & Han, I. (1997). Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. *Expert Systems with Applications* 13, 97-108.

Johnsen, T., & Melicher, R. W. (1994). Predicting corporate bankruptcy and financial distress: Information value added by multinomial logit models. *Journal of Economics and Business* 46, 269-286.

Jones, S., & Hensher, D. A. (2004). Predicting Firm Financial Distress: A mixed logit model. *The Accounting Review* 79, 1011-1038.

Jones, S., & Hensher, D. A. (2007). Modelling corporate failure: A multinomial nested logit analysis for unordered outcomes. *British Accounting Review* 39, 89-107.

Jorion, P., & Zhang, G. (2009). Credit contagion from counterparty risk. *Journal of Finance* 64, 2053–2087.

Kalotay, E. (2007). Discussion of Hensher and Jones. *ABACUS* 43, 265-270.

Keasey, K., & McGuinness, P. (1990). The failure of UK industrial firms for the period 1976-1984: Logistic analysis and entropy measures. *Journal of Business Finance & Accounting* 17, 119-135.

Keasey, K., & Watson, R. (1991). Financial distress models: A review of their usefulness. *British Journal of Management* 2, 89-102.

Kim, Y., & Sohn, S. Y. (2008). Random effects model for credit rating transitions. *European Journal of Operational Research* 184, 561-573.

Kleinbaum, D. G., Klein, M., & Pryor, E. R. (2002). *Logistic Regression: A Self-Learning Text* (2nd ed.). New York: Springer.

Koopman, S. J., Kraussl, R., Lucas, A., & Monteiro, A. B. (2009). Credit cycles and macro fundamentals. *Journal of Empirical Finance* 16, 42-54.

Koopman, S. J., Lucas, A., & Monteiro, A. (2008). The multi-state latent factor intensity model for credit rating transitions. *Journal of Econometrics* 142, 399-424.

Koopman, S. J., Lucas, A., & Schwaab, B. (2010). Macro, industry and frailty effects in defaults: The 2008 credit crisis in perspective. Inbergen Institute Discussion Paper 10-004/2.

Kumar, P. R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques: A review. *European Journal of Operational Research* 80, 1–28.

Lattinen, E. K. (1993). The use of information contained in annual reports and prediction of small business failures. *International Review of Financial Analysis* 2, 155-176.

Lennox, C. S. (1999). Identifying failing companies: A re-evaluation of logit, probit and DA approaches. *Journal of Economics and Business* 51, 347-364.

Lifschutz, S., & Jacobi, A. (2010). Predicting bankruptcy: Evidence from Israel. *International Journal of Business and Management* 5, 133-141.

Lin, F., & McClean, S. (2001). A data mining approach to the prediction of corporate failure. *Knowledge-based Systems* 14, 189-195.

Lin, L., & Piesse, J. (2004). Identification of corporate distress in UK industrials: A conditional probability analysis approach. *Applied Financial Economics* 14, 73-82.

Liu, W., & Tonks, I. (2008). *Alternative risk-based levies in the pension protection fund for multi-employee schemes.* XFi Working Paper No. 08/01.

Lo, A. W. (1986). Logit versus Discriminant Analysis: A specification test and application to corporate bankruptcies. *Journal of Econometrics* 31, 151-178.

Looney, S. W., Wansley, J. W., & Lane, W. R. (1989). An examination of misclassifications with bank failure prediction models. *Journal of Economics and Business* 41, 327-336.

Luoma, M., & Laitinen, E. K. (1991). Survival analysis as a tool for company failure prediction. *Omega* 6, 673-678.

Lussier, R. N., & Halabi, C. E. (2010). A three-country comparison of the business success versus failure prediction model. *Journal of Small Business Management* 48, 360-377.

McKee, T. E., & Lensberg, T. (2002). Genetic programming and rough sets: A hybrid approach to bankruptcy classification. *European Journal of Operational Research* 138, 436-451.

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

Miller, W. (2009). Comparing models of corporate bankruptcy prediction: distance to default vs. Z-score. Working paper, Morningstar, Inc.

Min, J. H., & Lee, Y.-C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications* 28, 603-614.

Moody's (2007). FAQs of corporate default service. Retrieved 24/10/2011, from http://v2.moodys.com/moodys/cust/content/content.ashx?source=StaticContent/Free%20Pages/DRS/Corp_DRS_FAQ.pdf

Muller, G. H., Steyn-Bruwer, B. W., & Hamman, W. D. (2009). Predicting financial distress of companies listed on the JSE: A comparison of techniques. *South African Journal of Business Management* 40, 21-32.

Nam, C. W., Kim, T. S., Park, N. J., & Lee, H. K. (2008). Bankruptcy prediction using a discrete-time duration model incorporating temporal and macroeconomic dependencies.

Journal of Forecasting 27, 493-506.

Neophytou, E., & Molinero, C. M. (2001). Predicting corporate failure in the UK: A multidimensional scaling approach. Working Paper 01-172, University of Southampton Department of Accounting and Management Science.

Nwogugu, M. (2007). Decision-making, risk and corporate governance: A critique of bankruptcy/recovery prediction models. *Applied Mathematics & Computation* 185, 178-196.

Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research* 19, 109–131.

Opler, T. C., & Titman, S. (1994). Financial distress and corporate performance. *Journal of Finance* XLIX, 1015-1040.

Park, C.-S., & Han, I. (2002). A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction. *Expert systems with applications* 23, 255-264.

Pendharkar, P. C., & Rodger, J. A. (2004). An empirical study of impact of crossover operators on the performance of non-binary genetic algorithm based neural approaches for classification. *Computer & Operations Research* 31, 481-498.

Peursem, K. A., & Pratt, M. J. (2002). A New Zealand failure prediction model: Development and international implications. *Advances in International Accounting* 15, 229-247.

Piesse, J., & Wood, D. (1992). Issues in assessing MDA models of corporate failure: A research note. *British Accounting Review* 24, 33-42.

Platt, H. D. (1989). The determinants of inter-industry failure. *Journal of Economics and Business* 41, 107-126.

Platt, H. D., & Platt, M. B. (1990). Development of a class of stable predictive variables: The

case of bankruptcy prediction. *Journal of Business Finance & Accounting* 17, 31-51.

Platt, H. D., Platt, M. B., & Pedersen, J. G. (1994). Bankruptcy discrimination with real variables. *Journal of Business Finance & Accounting* 21, 491–510.

Psillaki, M., & Margaritis, L. E. (2010). Evaluation of credit risk based on firm performance. *European Journal of Operational Research* 201, 873-881.

Robertson, J., & Mills, R. W. (1988). Company failure or company health? Techniques for measuring company health. *Long Range Planning* 21, 70-77.

Scott, J. (1981). The probability of bankruptcy: A comparison of empirical predictions and theoretical models. *Journal of Banking and Finance* 5,317-344.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *Journal of Business* 74, 101–124.

Sobehart, J., Keenan, S., & Stein, R. (2001). Benchmarking Quantitative Default Risk Models: A Validation Methodology. *Algo Reserch Quarterly* 4, 57-72.

Standard & Poor's (2008). Standard & Poor's To Explicitly Recognize Credit Stability As An Important Rating Factor. Retrieved August 26, 2011, from www.standardandpoors.com/ratingsdirect.

Steyn-Bruwer, B., & Hamman, W. (2006). Company failure in South Africa: Classification and prediction by means of recursive partitioning. *African Journal of Business Management* 37, 7-18.

Sueyoshi, T., & Goto, M. (2009). Methodological comparison between DEA (data Envelopment Analysis) and DEA-DA (Discriminant Analysis) from the perspective of bankruptcy assessment. . *European Journal of Operational Research* 199, 561-575.

Taffler, R. J. (1984). Empirical models for the monitoring of UK corporations. *Journal of Banking and Finance* 8, 199-227.

Taffler, R. (1983). The assessment of company solvency and performance using a statistical model. *Accounting and Business Research* 15, 295-308.

Taffler, R., & Tisshaw, H. (1977). Going, going, gone: Four factors which predict. *Accountancy* 88, 50-54.

Takahashi, K., & Kurokawa, Y. (1984). Corporate bankruptcy prediction in Japan. *Journal of Banking and Finance* 8, 229-247.

Tamari, M. (1984). The use of a bankruptcy forecasting model to analyze corporate behavior in Israel. *Journal of Banking and Finance* 8, 293-302.

Train, K. (2000). Halton Sequences for Mixed logit. UC Berkeley: Department of Economics, UCB. Retrieved from: http://escholarship.org/uc/item/6zs694tp.

Train, K. (2003). *Discrete Choice Methods with Simulation.* Cambridge: Cambridge University Press.

Train, K. E. (2009). *Discrete Choice Methods* (2nd ed.). Cambridge: Cambridge University Press.

Train, K., & Sonnier, G. (2004). Mixed logit with bounded distributions of partworths. *Working paper, Department of Economics, University of California at Berkeley, CA*.

Tsai, C.-F. (2009). Feature selection in bankruptcy prediction. *Knowledge-Based Systems* 22, 120-127.

Vassalou, M., & Xing, Y. (2004). Default Risk in Equity Returns. *The Journal of Finance* 59, 831–868.

Vuran, B. (2009). Prediction of business failure: A comparison of discriminant and logistic regression analyses. *Istanbul University Journal of the School of Business Administration* 1, 47-65.

Wang, Y., & Campbell, M. (2010). Business failure prediction for publicly listed companies in China. *Journal of Business and Management* 16, 75-88.

Westgaard, S., & Wijst, N. V. (2001). Default probabilities in a corporate bank portfolio model approach. *European Journal of Operational Research* 135, 338-349.

Whalen, G. (1991). A proportional hazards model of bank failure: An examination of its usefulness as an early warning tool. *Federal Reserve Bank of Cleveland Economic Review* Q1, 21-31.

Wilson, R. L., & Sharda, R. (1994). Bankruptcy prediction using neural networks. *Decision Support Systems* 11, 545-557.

Wong, J. M., & NG, S. T. (2010). Company failure in the construction industry: A critical review and a future research agenda. Paper presented at FIG Int. Congress 2010, Sydney, Australia.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA: MIT Press.

Wood, D., & Piesse, J. (1987). The Information Value of MDA Based Financial Indictors. *Journal of Business Finance & Accounting* 14, 27-38.

Wu, C., & Kuo, C. (2004). Using non-financial information to predict bankruptcy: A study of public companies in Taiwan. *International Journal of Management* 21, 194-201.

Zavgrenc, C. V. (1985). Assessing the vulnerability to failure of American industrial firms: A logistic analysis. *Journal of Business Finance & Accounting* 12, 19–45.

Zhou, Y., Xie, S., & Yuan, Y. (2008). Statistical inference on the default probability in credit risk models. *Journal of System Engineering and Practice* 28, 206-214.

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*22 (Supplement), 59-86.