# University of Cincinnati

**Date: 5/21/2012**

**I, Shaonan Tian , hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Business Administration.**

It is entitled:

**Essays on Corporate Default Prediction**

Student's name: **Shaonan Tian**

This work and its defense approved by:

**Committee chair: Yan Yu, PhD**

**Committee member: Hui Guo, PhD**

**Committee member: Martin Levy, PhD**

UNIVERSITY OF Cincinnati

2901

# Essays on Corporate Default Prediction

A dissertation submitted to the

Graduate School

of the University of Cincinnati

in partial fulfillment of the

requirements for the degree of

**Doctor of Philosophy**

in the Department of Operations and Business Analytics

of the College of Business

2012

by

**Shaonan Tian**

M.S., Business Analytics, University of Cincinnati (2009)

B.S., Software Engineering, Zhejiang University (2006)

Committee Chair: Yan Yu

## Abstract

Corporate bankruptcy prediction has received paramount interest in academic research, business practice and government regulation. The recent financial crisis, during which unexpected corporate insolvencies had caused severe damage to the aggregate economy, highlights the crucial importance of an accurate corporate default prediction. Consequently, accurate default probability prediction is extremely important. The purpose of this research is to offer a unique contribution to the extant literature. This dissertation consists of three essays.

In the first essay (Chapter 1), we propose to apply a discrete transformation family of survival models to corporate default risk predictions. A class of Box-Cox transformations and logarithmic transformations are naturally adopted. The proposed transformation model family is shown to include the popular Shumway's model and grouped relative risk model. We show that a transformation parameter different from those two models is needed for default prediction using the bankruptcy data. In addition, out-of-sample validation statistics show improved performance. The estimated default probability is further used to examine a popular asset pricing question whether the default risk has carried a premium. Due to some distinct features of bankruptcy application, the proposed class of discrete transformation survival models with time-varying covariates is different from the continuous survival models in the literature. Their links and differences are also discussed.

Essay 2 (Chapter 2) introduces a robust variable selection technique, the least absolute shrinkage and selection operator (LASSO), to investigate formally the relative importance of various bankruptcy predictors commonly used in the existing literature. Over the 1980 to 2009 period, LASSO admits the majority of Campbell, Hilscher, and Szilagyi's (2008) predictive

variables into the bankruptcy forecast model. Interestingly, the total debt to total assets ratio and the current liabilities to total assets ratio constructed from only accounting data also contain significant incremental information about future default risk. LASSO-selected variables have superior out-of-sample predictive power and outperform (1) those advocated by Campbell, Hilscher, and Szilagyi (2008) and (2) the distance to default from Merton's (1974) structural model. Furthermore, study on the international market reveals the uniform significance brought by the activity indicator, *sales / total assets*.

Essay 3 (Chapter 3) devotes special care to an important aspect of the bankruptcy prediction modeling: data sample selection issue. To investigate the effect of the different data selection methods, three models are adopted: logistic regression model, Neural Networks (NNET) and Support Vector Machines (SVM). A Monte Carlo simulation study and an empirical analysis on an updated bankruptcy database are conducted to explore the effect of different data sample selection methods. By comparing the out-of-sample predictive performances, we conclude that if forecasting the probability of bankruptcy is of interest, complete data sampling technique provides more accurate results. However, if a binary bankruptcy decision or classification is desired, choice based sampling technique may still be suitable.

**Keywords: Corporate Bankruptcy Prediction, Logistic Regression, Proportional Hazard, Survival Analysis, LASSO**

## Acknowledgements

I would like to express my sincere gratitude and appreciation to my advisor Dr. Yan Yu for her valuable guidance through my research work for these five years and for teaching me the aspiring and energetic attitude toward work which can benefit me lifetime. This work would not be possible without your kindness support and generosity. I also would like to thank my dissertation committee members, Dr. Martin Levy and Dr. Hui Guo for your generous advice during my study and taking the precious time to review my dissertation.

Special thanks to my co-author, Dr. A. Adam Ding. Thank you for being such a great and outstanding researcher and I truly treasure the collaboration and fruitful work we have accomplished. I would also like to extend my appreciation to all the QAOM faculty members for the generous help and scientific support during my study.

Finally, I would like to give my sincere, special and heartfelt thanks to my husband Dingchuan Xue and my parents for their continuous and consistent support and encouragement throughout all of my education and my life both in China and in the US. Appreciation also goes to my friends, my colleagues and classmates for their valuable advice and suggestions.

# Table of Contents

# List of Tables

# List of Figures

# 1 Chapter One

## A Class of Discrete Transformation Survival Models with Application to Default Probability Prediction

## 1.1 Introduction

Corporate bankruptcy has long been one of the most significant threats for many businesses. It not only increases the financial loss to its creditors but also has a negative impact on the society and the aggregate economy. More alarmingly, data released by the Administrative Office of the U.S. courts show that in the recent decades business failures have occurred at higher rates than at any time since the early 1930's. The default loss has also maintained at a startling level of trillions of dollars.

Accurate default probability prediction is of great interest to all academics, practitioners and regulators. Corporate default forecasting models are used by regulators to monitor the financial health of banks, funds, and other institutions. Practitioners use default probability forecasts in conjunction with models to price corporate debt and for internal rating based approach (Schönbucher 2003; Lando 2004). Academics use bankruptcy forecasts to test various conjectures such as the hypothesis that default risk is priced in stock return (Campbell, Hilscher and Szilagyi 2008). Given recent economic condition, the importance of accurate default predictive model validation is even more substantially promoted by the Basel Committee on Banking Supervision under the current framework of Basel II [1].

Despite vast literature on bankruptcy prediction (see Altman 1993 for a survey), most research prior to past decade is concentrated on static modeling using cross-sectional data (e.g. Altman 1968; Ohlson 1980; Zmijewski 1984). Though multi-period firm characteristics are observed, prior researchers only choose to use one period observation with a single-period logistic regression or discriminant analysis.

On the other hand, the event of default can be considered a terminal event for the company. This is mathematically equivalent to the death event in the survival analysis which has generated a huge body of literature. From the viewpoint of survival analysis, predicting the time to default based on the various measurable financial and market variables at current time naturally corresponds to analyze covariate effects on the survival time. In continuous time survival analysis, the proportional hazards model (Cox 1972), which can also be referred as continuous relative risk model, is the most popular covariate effect model. The partial

[1] Basel II is an international business standard that requires financial institutions to maintain enough cash reserves to cover risks incurred by operations.

likelihood estimator for the proportional hazards model is studied in Breslow (1974), Cox (1975) and Efron (1977). Alternatively, covariate effects can be modeled as proportional odds whose biological motivation is discussed in Bennett (1983). Mathematical properties of the covariate effect estimator in the proportional odds model are studied by Wu (1995), Murphy, Rossini, and van der Vart (1997). For time-invariant covariates, a generalized odds model family is proposed that includes both proportional hazards and proportional odds model as special cases (Harrington and Fleming 1982, Clayton and Cuzick 1986, Dabrowska and Doskum 1988a). Estimation of covariate effect parameter for the generalized odds model is studied in Clayton and Cuzick (1986), Dabrowska and Doskum (1988b), Cheng, Wei and Ying (1995, 1997), Scharfstein, Tsiatis, and Gilbert (1998). Zeng and Lin (2006, 2007) extend the generalized odds model family to a transformation family for time-varying covariates. In discrete time survival analysis, there are a few different extensions of the Cox proportional hazards model. They include the discrete logistic model proposed in Cox (1972), the grouped relative risk model studied in Kalbfleisch and Prentice (1973) and the discrete relative risk model. These discrete time models are summarized in an encompassing formulation in Kalbfleisch and Prentice (2002, Page 136).

While there is a natural correspondence between survival analysis and the default risk modeling, such a link has not been explored in the literature until Shumway (2001) proposed a discrete hazard model. The conditional default probability $\pi_{i,k}$ that the $i$th firm files for bankruptcy at time $t_k$ given it survives past time $t_{k-1}$ is modeled through a multi-period logistic regression by the $i$th firm's specific characteristics $\mathbf{Z}_i(t_k)$ at time $t_k$,

$$\pi_{i,k} = \frac{1}{1 + \exp(-\alpha - \boldsymbol{\beta}^\tau \mathbf{Z}_i(t_k))}.$$

Here the time-varying covariate values $\mathbf{Z}_i(t_k)$ are usually firm's financial ratios obtained from accounting statements and firm's market variables from public trading record; $\boldsymbol{\beta}$ is the covariate effect parameter and $\alpha$ is a scalar parameter. Not surprisingly, Shumway (2001) also shows that hazard, or survival modeling, is advantageous by coping with time-varying panel data, while static model ignores the fact that firms change through time and may produce biased and inconsistent bankruptcy probability estimates. This discrete hazard model quickly gains popularity in corporate bankruptcy prediction and is used in Chava

and Jarrow (2004); Bharath and Shumway (2008); Campbell et al. (2008). This popular Shumway's discrete hazard model is in fact the discrete logistic model of Cox (1972) for time-varying covariates. The Cox proportional hazards model has also been used for bankruptcy prediction in Duffie, Saita and Wang (2007); Duffie, Eckner, Horel, and Saita (2009).

In this paper, we propose a class of discrete-time transformation model family to bankruptcy probability prediction with time-varying covariates

$$\pi_{i,k} = \begin{cases} 1 - \frac{1}{[1+c\exp(\alpha_k+\boldsymbol{\beta}^\tau \mathbf{Z}_i(t_k))]^{1/c}}, & c > 0; \\ 1 - \exp[-\exp(\alpha_k + \boldsymbol{\beta}^\tau \mathbf{Z}_i(t_k))], & c = 0. \end{cases} \tag{1.1}$$

or

$$\pi_{i,k} = \begin{cases} 1 - \frac{1}{1+\frac{1}{\rho}\{\exp[\rho\exp(\alpha_k+\boldsymbol{\beta}^\tau \mathbf{Z}_i(t_k))]-1\}}, & \rho > 0; \\ \frac{1}{1+\exp(\alpha_k+\boldsymbol{\beta}^\tau \mathbf{Z}_i(t_k))}, & \rho = 0. \end{cases} \tag{1.2}$$

Here $c$ (or $\rho$) is a scalar transformation parameter. The proposed discrete-time transformation models are derived formally in Section 1.3 by applying a monotonic transformation on difference of the minus log survival functions. Inverse of a class of Box-Cox transformations and inverse of a class of logarithmic transformations are used. Essentially we are proposing to apply the transformation families in Zeng and Lin (2006) for continuous survival models to an equivalent formulation of the discrete-time survival models in Kalbfleisch and Prentice (2002). This class of transformation model family contains the Shumway (2001) model (or discrete logistic model, Cox 1972) when $c = 1$ or $\rho = 0$ and the grouped relative risk model (Kalbfleisch and Prentice 1973) when $c = 0$ or $\rho = 1$ as special cases.

The estimated covariate effect parameter $\boldsymbol{\beta}$ can be used to rank companies' default risk based on their covariate values $\mathbf{Z}_i(t_k)$: companies with higher $\boldsymbol{\beta}^\tau \mathbf{Z}_i(t_k)$ values would have higher default risk at given time $t_k$. Thus $\boldsymbol{\beta}^\tau \mathbf{Z}_i(t_k)$ can be considered as a credit score like those given out by the rating agencies such as Moody's and S&P. Abundant literature and ongoing research are dedicated to obtain good credit scores. However, actual default probability is needed to assess the portfolio risk for calculating banking reserves as in Basel II. Actual default probability is also essential to combine with the loss given default measure (Schuermann 2005).

Figure 1.1 shows the default probability curves for different transformation parameter values. We can see that same scores would correspond to different default probabilities

under different transformation parameters. In our empirical analysis, we apply the proposed discrete-time transformation model family on a comprehensive bankruptcy data set spanning from 1981 to 2006. Log-likelihood plots of the fit on quarterly and annually firm observations show that the optimal transformation parameter resides near the point $c = 10$ which is neither Shumway's model nor the grouped relative risk model.

[**Insert Figure 1.1**]

In addition, out-of-sample prediction with withholding period 2002-2006 shows improved accuracy ratio and model goodness-of-fit. We also investigate the asset pricing implication conjectured by Fama and French (1996) that investors require a positive return premium for holding stocks with high default probabilities. We sort stocks into portfolios by the predicted default probability using the proposed discrete transformation survival model. We find that stocks with higher default probabilities deliver anomalously lower returns, which challenges the original Fama and French's (1996) conjecture. Our findings, however, are consistent to those documented by some recent literature such as Campbell et al. (2008). Furthermore, a simulation study is conducted, showing promises of the proposed transformation model family.

The original motivation of our proposed class of Box-Cox and logarithmic transformations for discrete survival models comes from the continuous time generalized odds-rate model of Dabrowska and Doskum (1988) with time-invarying covariate and Zeng and Lin (2006, 2007) for time-varying covariate. We are proposing a similar extension of this class of transformation families to discrete failure time distribution. However, our proposed models are different from those continuous transformation models. In Section 1.3 we show in detail that the transformation is on the difference of the minus log survival functions, while transformation of Zeng and Lin (2006) depends on the entire history of covariate values. Simple discrete extension of Zeng and Lin (2006) is ill-defined, due to some unique features for bankruptcy prediction application. In particular: Firstly, actual calendar time needs to be used. This is because same firm specific characteristics at different calendar time may expose to different default probability due to different macroeconomic conditions. Secondly, differing from the classical survival model set-up, companies from the bankruptcy database

do not share a common starting point. This is due to the use of calendar time (see discussion in Section 1.3.5). Data for most firms start from the beginning of the sample observation period despite that they have prior accounting statements and trading activities. On the other hand, a number of firms enter in the middle of the sample period since they have just started public trading. Hence, the transformation model of Zeng and Lin (2006), depending on the entire history of covariates from the same starting point, is not well-defined. Thirdly, the proposed class of discrete transformation model family enjoys the same appealing "memoryless feature" as Shumway (2001) and common discrete survival models (Kalbfleisch and Prentice 2002). That is, the conditional default probability only depends on the last available observation, instead of the whole path of covariates as in Zeng and Lin (2006). The first two features of bankruptcy prediction imply that a practical model needs to be memoryless. Finally, by nature, the accounting and market information used in the default prediction study are collected only at discrete time period over a fixed time window. For example, accounting data from sample period 1981-2006 are obtained for our study from the quarter end balance sheet, income statement and cash flow report.

The rest of paper is organized as following. Section 1.2 describes the bankruptcy data we use in the study. Section 1.3 presents the class of discrete time transformation survival model family and links to continuous time survival analysis. Empirical results of the corporate bankruptcy application and a simulation study are given in Section 1.4. The Appendices give detailed mathematical derivations.

## 1.2 Bankruptcy Data

In our study, we develop a comprehensive bankruptcy database by merging the Center for Research in Security Prices (CRSP) with Compustat from Standard & Poor's (COMPUSTAT) database through Wharton Research Data Services (WRDS)[2]. The CRSP database provides a complete collection of security data including price, return, and volume data for the three major stock exchange markets: NYSE, AMEX, and NASDAQ. COMPUSTAT maintains quarterly accounting information for companies including reports of Income Statement, Balance Sheet, and Statement of Cash Flows etc. Our bankruptcy database includes

[2]website: http://wrds-web.wharton.upenn.edu/wrds/

all the publicly traded companies in the United States between 1981 and 2006. To measure the probability of default using our proposed transformation model, we need a set of exploratory variables and an event indicator of bankruptcy for default companies. In our empirical study, we define a firm as default if it files under either Chapter 7 or Chapter 11 bankruptcy code. Because it usually takes a long time to settle bankruptcy disputes, in some cases, the COMPUSTAT updates the default status with a substantial delay. The delay makes it difficult to identify accurately the corporate default in the most recent period. To address this issue, we end our sampling period in the year 2006 and restrict our sampling time window from 1981 to 2006. In addition, eight covariate measures are constructed for the exploratory variables: profitability, leverage, short-term liquidity, the market-to-book ratio (MB), volatility and excess return over the S&P 500 index return, as well as the firm's relative size to the S&P 500 index value and the price. The formation of this set of covariates follows Campbell et al. (2008) closely.

Note that a firm may exit the database at any time due to its financial health status, it may also enter the database in different time periods. For most healthy firms, their Initial Public Offering (IPO) dates were prior to the year of 1981. In this case, the firms enter our database coincidentally with the start of our sampling period. On the other hand, there may be firms with IPO dates after 1981. This case is particularly common during the "dot-com bubble" period in the late 1990s and early 2000. Under such situation, we record the first observation of the firm at the time of its first trading date. For example, the IPO date for the IT company "Microsoft Corporation" was March 1986. Therefore, the firm "Microsoft Corporation" enters our database in the year of 1986, instead of the starting time of our sampling period, 1981. Such property of having "different starting time" is unique for bankruptcy data, and therefore, needs special treatment when linking with survival models.

## 1.3 Model

### 1.3.1 A Class of Discrete Time Transformation Survival Models

Financial data are discrete in nature. For example, commonly used predictors such as firms' financial ratios are obtained through accounting statements quarterly. Hence, a discrete time

model is needed for corporate bankruptcy prediction.

Suppose there are $K$ fixed observation time $t = t_1, t_2, ..., t_K$ for the whole observation period. For example, these are the quarter end date for quarterly data. Total $i = 1, 2, ..., n$ public firms are in the data base during the sample period, each with observed data $(B_i, X_i, \Delta_i, \mathbf{Z}_{i,k})$, $k = 1, ..., K$. Here $B_i$ denotes the starting time – the first time the firm is publicly traded during the observation period. If a firm in the data base is traded prior to $t_1$, then $B_i = t_1$. Here $X_i$ denotes the last time the firm is observed during the observation period. It is subject to right censoring by the end of observational period. If a firm files for bankruptcy after $t_K$, then $X_i = t_K$. $\mathbf{Z}_{i,k} = \mathbf{Z}_i(t_k)$ is the $d$-dimensional covariate vectors for firm $i$ at time $t = t_k$. $\Delta_i$ is the so called censoring indicator in survival analysis: $\Delta_i = 1$ if the $i$th firm enters bankruptcy filing process at $t = X_i$; $\Delta_i = 0$ otherwise. A healthy firm may experience early exit from the data base, such as merger or acquisition. In those cases, $X_i < t_K$ but $\Delta_i = 0$.

Denote $T$ the calendar death time. For firms with common starting point $B = t_1$, without loss of generality if $t_1 = 0$, then $T$ is equivalent to the survival time in the literature. Let $S_{\mathbf{Z}}(t) = Pr(T > t|\mathbf{Z} = \mathbf{z})$ be the survival function given $\mathbf{Z}$. Here $\mathbf{Z}$ is the covariate process over the whole time period. Denote

$$\pi_{i,k} = Pr(T = t_k | X \geq t_k, \mathbf{Z}(t_k) = \mathbf{Z}_{i,k})$$

the conditional probability that the firm files for bankruptcy at time $t_k$ given it is at risk at time $t_k$ (survival past time $t_{k-1}$).

Formally, let

$$G[-\log \frac{S_{\mathbf{Z}}(t_k)}{S_{\mathbf{Z}}(t_{k-1})}] = exp[\boldsymbol{\beta}^\tau \mathbf{Z}(t_k)] G[-\log \frac{S_{\mathbf{0}}(t_k)}{S_{\mathbf{0}}(t_{k-1})}], \tag{1.3}$$

where $G$ is a strictly increasing transformation function with $G(0) = 0$ and $G(\infty) = \infty$; $\boldsymbol{\beta}$ is a $d$-dimensional covariate effect parameter; $S_{\mathbf{0}}(\cdot)$ is the (baseline) survival function when $\mathbf{Z} \equiv \mathbf{0}$.

Our class of discrete time transformation survival model family then takes the form in (1.1)

$$\pi_{i,k} = \begin{cases} 1 - \frac{1}{[1 + c \exp(\alpha_k + \boldsymbol{\beta}^\tau \mathbf{Z}_{i,k})]^{1/c}}, & c > 0; \\ 1 - \exp[-\exp(\alpha_k + \boldsymbol{\beta}^\tau \mathbf{Z}_{i,k})], & c = 0. \end{cases}$$

for transformation function $G_c$ that belongs to family

$$G_c(x) = \begin{cases} \frac{1}{c}[exp(cx) - 1], & c > 0; \\ x, & c = 0. \end{cases} \tag{1.4}$$

Or it takes the form in (1.2)

$$\pi_{i,k} = \begin{cases} 1 - \frac{1}{1+\frac{1}{\rho}\{\exp[\rho \exp(\alpha_k + \boldsymbol{\beta}^\tau \mathbf{z}_{i,k})] - 1\}}, & \rho > 0; \\ \frac{1}{1+\exp(\alpha_k + \boldsymbol{\beta}^\tau \mathbf{z}_{i,k})}, & \rho = 0. \end{cases}$$

for transformation function $G_\rho$ that belongs to family

$$G_\rho(x) = \begin{cases} \frac{1}{\rho} log[1 + \rho(exp(x) - 1)], & \rho > 0; \\ exp(x) - 1, & \rho = 0. \end{cases} \tag{1.5}$$

Appendix 1 gives detailed derivations of the discrete time transformation survival models (1.1) and (1.2) based on (1.3) and monotonic transformations (1.4) and (1.5). Here $G_c$ and $G_\rho$ are common monotonic transformations used in the continuous survival analysis literature. Inverse of $G_c$ is equivalent to a class of logarithmic transformations and inverse of $G_\rho$ is similar to the class of Box-Cox transformations considered in Chen et al. (2002) and Zeng and Lin (2006) etc. For firms with different starting point $B = t_{k^*} > t_1$, for $k \geq k^*$, let the conditional default probability $\pi_{i,k}$ take the values as in (1.1) and (1.2). This is feasible due to their memoryless features.

Note when $c = 0$ or $\rho = 1$, the proposed discrete time transformation survival model is equivalent to the classical grouped relative risk model. When $c = 1$ or $\rho = 0$, it is the popular so-called discrete hazard model proposed in Shumway (2001) and is then followed by most bankruptcy prediction literature such as Chava and Jarrow (2004), Campbell et al. (2008) etc. Our transformation model would estimate the transformation parameter $c$ (or $\rho$) in addition to parameters $\alpha_k$ and $\boldsymbol{\beta}$.

### 1.3.2 Estimation and Algorithm

The likelihood function for the proposed discrete transformation survival model is

$$\prod_{i=1}^{n} \prod_{k:\ B_i < t_k \leq X_i} {\pi_{i,k}}^{\Delta_{i,k}} (1 - \pi_{i,k})^{1-\Delta_{i,k}},$$

where $\Delta_{i,k} = \Delta_i I\{X_i = t_k\}$ and $\pi_{i,k} = Pr(T = t_k | X \geq t_k, \mathbf{Z}(t_k) = \mathbf{Z}_{i,k})$ is modeled by (1.1) and (1.2). The log-likelihood function is then

$$L = \sum_{i=1}^{n} L_i = \sum_{i=1}^{n} \sum_{k:\ B_i < t_k \leq X_i} \Delta_{i,k} log(\pi_{i,k}) + (1 - \Delta_{i,k}) log(1 - \pi_{i,k}). \tag{1.6}$$

Therefore, for a fixed $c$ (or $\rho$) value, the fitting of $\alpha_k$ and $\boldsymbol{\beta}$ can be implemented mathematically using logistic regression with a specified link function on independent $\Delta_{i,k}$'s even though $\Delta_{i,k}$'s are dependent in the data set. When assumption of the grouped relative risk model or the Shumway (2001) model holds, our estimates would give about the same results. Our transformation model would provide better fit when the transformation parameters are not close to 0 or 1.

For estimation, we could maximize (1.6) with (1.1) (or (1.2)) over the transformation parameter $c$ (or $\rho$), covariate effect parameter $\boldsymbol{\beta}$ and baseline parameter $\alpha_k$ simultaneously on the data set. For graphing purpose, we compute the parametric maximum likelihood estimator for the covariate effect parameter $\boldsymbol{\beta}$ and $\alpha_k$ over a grid window of the transformation parameter $c$ (or $\rho$). At each point value of $c$ (or $\rho$), log-likelihood given by (1.6) is maximized over the parameter $\boldsymbol{\beta}$ and $\alpha_k$. This is a standard nonlinear optimization procedure. Many scientific optimization package can conduct the maximization computation.

However, we note that when the sampling time period extends or the sampling frequency increases (for example, from annual data to quarter data), the parameter space may expand significantly, which may potentially lead to some common numerical problems in large-scale nonlinear optimization. Here we adopt the profile-likelihood method (Murphy and van der Vaart, 2000). Our algorithm is as below. For a given $c$ (or $\rho$) over a fixed grid window,

- **Step 0**. Initialize parameter $\widehat{\boldsymbol{\beta}}^{(0)}$ and $\widehat{\alpha}_k^{(0)}$ for $k = 1, 2, ..., K$. Sensible initial values, such as a vector of 0 or 1 can be used. Alternatively, we may use the parameter estimated through Shumway's multi-period logistic regression as an initial estimate.

- **Step 1**. Given estimated $\widehat{\boldsymbol{\beta}}^{(j)}$, obtain $\widehat{\alpha}_k^{(j+1)}$ for $k = 1, 2, ..., K$. Each $\alpha_k$ is estimated only on the set of firm observations at time $t_k$ for $k = 1, 2, ..., K$. Hence, the likelihood given by equation (1.6) is maximized over one parameter for each time period. This can be done almost instantaneously.

- **Step 2**. Given estimated $\widehat{\alpha}_k^{(j+1)}$ for $k = 1, 2, ..., K$, estimate $\widehat{\boldsymbol{\beta}}^{(j+1)}$ on the entire data set. This step involves the entire dataset to conduct the optimization procedure. However, the dimension of covariate parameter $\boldsymbol{\beta}$ is relatively small so that the convergence is quite fast.

Then iterate step 1 and step 2 until convergence.

From our experience, the merit of the profile-likelihood algorithm is to expedite the computational process, especially over a relatively large parameter space. Alternatively, Newton-Raphson method when applying a special feature may also be applied. See Kalbfleisch and Prentice (2002, Page 139) for details.

Note that the optimization algorithm does not guarantee a global maximum value on likelihood. To avoid the local minimum situation, we experiment on different initial values of $\boldsymbol{\beta}$ and $\alpha_k$. Our experience shows that the sensible initial values, such as a vector of 0 or 1 may easily lead to a satisfactory performance.

### 1.3.3 Large Sample Properties

The proposed class of transformation models on the bankruptcy prediction has three parametric components: $d$-dimensional covariate effect parameter $\boldsymbol{\beta}$, $K$-dimensional baseline parameter $\alpha_k = \alpha(t_k)$ and the scalar transformation parameter $c$ (or $\rho$). This is a usual parametric problem set-up. Hence, under some mild conditions, large sample properties for estimates of our proposed class of discrete transformation models (1.1) or (1.2) can be easily established following the standard parametric maximum likelihood estimators with total fixed $d + K + 1$ parameters to estimate, similar to those in Lehmann and Casella (1998) and Kalbfleisch and Prentice (2002). Root-n consistency and asymptotic normality are readily available for further inference.

Formally, assume the firm's birth time $B$ is a random variable with $P(B = t_k) > 0$ for $k = 1, 2, ..., K$. Given $B = t_{k^*}$, the covariate process $\mathbf{Z} = (\mathbf{Z}(t_{k^*}), \mathbf{Z}(t_{k^*+1}), ..., \mathbf{Z}(t_K))$ is generated from a distribution with probability density function $f_{k^*}$. For the calendar death time $T$, the conditional probability of $T = t_k$ is given by model (1.1) or (1.2), censoring indicator $\Delta = I(T \leq t_K)$ and $X = min(T, t_K)$. The observed random variables for the company are

$\boldsymbol{\chi} = (B, X, \Delta, \mathbf{Z} = [\mathbf{Z}(B), ..., \mathbf{Z}(X)])$. We observe $n$ random replicates $(B_i, X_i, \Delta_i, \mathbf{Z}_i)$ from this probability distribution.

Assumptions below are needed to establish the following asymptotic theorem.

**Assumptions**

*(i) The true parameter* $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^T, \alpha_{0,1}, \cdots, \alpha_{0,K}, c_0)$ *or* $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^T, \alpha_{0,1}, \cdots, \alpha_{0,K}, \rho_0)$ *is an interior point of* $\omega$*, an open subset of the parameter space* $\Omega$*.*

*(ii)* $L_1(x, \boldsymbol{\theta})$ *admits third derivatives with respect to* $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_{d+K+1})$ *for all* $\boldsymbol{\theta} \in \omega$*. Furthermore, there exist functions* $M_{jkl}$ *such that*

$$\left| \frac{\partial^3}{\partial\theta_j \partial\theta_k \partial\theta_l} L_1(\mathbf{x}, \boldsymbol{\theta}) \right| \leq M_{jkl}(\mathbf{x}) \text{ for all } \boldsymbol{\theta} \in \omega,$$

*where*

$$m_{jkl} = E_{\boldsymbol{\theta}_0}[M_{jkl}(\boldsymbol{\chi})] < \infty \text{ for all } j, k, l.$$

*(iii) The Fisher Information* $\mathbf{I}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[\{\frac{\partial}{\partial\boldsymbol{\theta}} L_1(x, \boldsymbol{\theta})\}^{\otimes 2}]$ *exists. Here and below, for a vector* $\mathbf{v}$*,* $\mathbf{v}^{\otimes 2}$ *denotes* $\mathbf{v}\mathbf{v}^\tau$*. We assume that* $\mathbf{I}(\boldsymbol{\theta})$ *is positive definite for all* $\boldsymbol{\theta} \in \omega$*.*

**Theorem** *Under Assumptions (i) to (iii), the likelihood equation (1.6) has a consistent root* $\hat{\boldsymbol{\theta}}$ *and* $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ *is asymptotically normal with mean zero and variance* $[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}$*.*

For a fixed number of $K$ time periods, the parameters are $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \alpha_1, ..., \alpha_K, c)$ or $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \alpha_1, ..., \alpha_K, \rho)$. Our estimators are the parametric maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ for the discrete transformation models (1.1) or (1.2). Under the standard regularity conditions (e.g., Lehmann and Casella 1998, Chapter 6; Kalbfleisch and Prentice 2002, Chapter 3), for large samples, $\hat{\boldsymbol{\theta}}$ has the true value $\boldsymbol{\theta}_0$ for the parameter $\boldsymbol{\theta}$ as asymptotic mean and the inverse of Fisher Information, $\frac{1}{n}\mathbf{I}^{-1}(\boldsymbol{\theta}_0)$ as asymptotic variance. The Fisher Information can be estimated by

$$\hat{\mathbf{I}}(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} \Big\{ \sum_{k:\ B_i < t_k \leq X_i} (\frac{\Delta_{i,k}}{\pi_{i,k}} - \frac{1-\Delta_{i,k}}{1-\pi_{i,k}}) \frac{\partial}{\partial\boldsymbol{\theta}} \pi_{i,k} \Big\}^{\otimes 2} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

The variance and confidence intervals for the parameters can then be calculated using this information. Notice that one of the standard regularity condition is that the true parameter value $\boldsymbol{\theta}_0$ is an interior point of the parameter space. $c = 0$ (grouped relative risk model) and $\rho = 0$ (Shumway's model) are on the boundary of the parameter space. Therefore,

two families (1.1) and (1.2) instead of only one are considered in practice. For data from those two models, we would use the other transformation families. The boundary points $c = 0$ (grouped relative risk model) and $\rho = 0$ (Shumway's model) correspond to interior points $\rho = 1$ and $c = 1$, respectively, of the other family. The two families together include extensions of these two standard models from both directions.

### 1.3.4 Link to Continuous Time Survival Models

The proposed class of discrete transformation survival models is originally motivated from the continuous generalized odds-rate model of Dabrowska and Doskum (1988a) with time-invarying covariate and Zeng and Lin (2006) for time-varying covariate $\mathbf{Z}(t)$. In essence we are applying inverse of Box-Cox transformation and logarithmic transformation families in Zeng and Lin (2006) to an equivalent formulation of the discrete survival transformation models in Kalbfleisch and Prentice (2002).

Transformation models (1.3) are different from those in continuous survival analysis. The transformation $G$ is on $[-\log \frac{S_{\mathbf{Z}}(t_k)}{S_{\mathbf{Z}}(t_{k-1})}]$ or equivalently on the difference of minus log of survival functions. Equation (1.3) can be rewritten as

$$G[-\log S_{\mathbf{Z}}(t_k) - (-\log S_{\mathbf{Z}}(t_{k-1}))] = exp[\boldsymbol{\beta}^{\tau}\mathbf{Z}(t_k)]G[-\log S_{\mathbf{Z}}(t_k) - (-\log S_{\mathbf{Z}}(t_{k-1}))]. \quad (1.7)$$

Instead, the transformation of Zeng and Lin (2006) is on the cumulative hazard function $\Lambda_{\mathbf{Z}}(t) = -\log S_{\mathbf{Z}}(t)$

$$G(\Lambda_{\mathbf{Z}}(t)) = \int_0^t e^{\boldsymbol{\beta}^T \mathbf{Z}(s)} d\Lambda(s),$$

which depends on the entire history of covariate values $\{\mathbf{Z}(s)\}_{s=0}^{t}$. Here $\Lambda(.)$ is an unspecified increasing function.

After some calculation (see Appendix 2), Zeng and Lin (2006)'s model can be reexpressed by

$$\frac{d}{dt}G[\Lambda_{\mathbf{Z}}(t)] = exp[\boldsymbol{\beta}^{\tau}\mathbf{Z}(t)]\frac{d}{dt}G[\Lambda_0(t)]. \quad (1.8)$$

Equation (1.7) can be regarded as taking the difference of $-\log S_{\mathbf{Z}}(t_k)$ first and then take transformation $G$ on the difference. Equation (1.8) applies the transformation $G$ on $\Lambda_{\mathbf{Z}}(t) = -\log S_{\mathbf{Z}}(t_k)$ first before taking difference. Again note that (1.7) and (1.8) are fundamentally different transformation models.

Zeng and Lin (2006) can not be extended similarly to the bankruptcy prediction application. When the covariate $\mathbf{Z}(t)$ is time-varying, their likelihood function depends on the values of the covariate process $\mathbf{Z}(s)$ for all time $s$ between $s = 0$ and $s = X_i$. Therefore their maximum likelihood estimator can not be found unless we observe the whole covariate process $\mathbf{Z}(s)$ between $s = 0$ and $s = X_i$. For a new public company starting at time $B_i > 0$, certainly the covariate process $\mathbf{Z}(s)$ does not exist for the time interval between $s = 0$ and $s = B_i$. That is, $\Lambda_{\mathbf{Z}}(B_i)$ is unknown as long as there is covariate effect and the time-varying covariate is not known for $t < B_i$. Hence, simple discrete extension of Zeng and Lin (2006) is ill-defined for bankruptcy prediction application. By applying the transformation families $G_c(\cdot)$ and $G_\rho(\cdot)$ on the differences as in equation (1.7), this approach yields measures that only depend on covariate values at one last period but not on the whole past history of covariates.

Note when $\mathbf{Z}$ is time invariant covariates, the generalized odds-rate model of Dabrowska and Doskum (1988a) has similar transformation identity as equation (1.8) since

$$G[\Lambda_{\mathbf{Z}}(t)] = exp[\boldsymbol{\beta}^\tau \mathbf{Z}] G[\Lambda_{\mathbf{0}}(t)],$$

where $\Lambda_{\mathbf{Z}}(t)$ is their cumulative hazard function in continuous time.

Inverse of logarithmic transformation $G_c(\cdot)$ of (1.4) and inverse of Box-Cox Transformation $G_\rho(\cdot)$ of (1.5) are commonly used in continuous time transformation survival models (e.g. Dabrowska and Doskum 1988; Chen et al. 2002; Zeng and Lin 2006). The generalized odds-rate model family includes the classical Cox proportional hazard model ($c = 0$) and the proportional odds model ($c = 1$) as special cases. The Cox proportional hazards model corresponds to $c = 0$, a boundary point on the parameter space, for the family $G_c$. Transformation $G_\rho(\cdot)$ of (1.5) yields alternative families where the Cox proportional hazards model corresponds to an interior point of the parameter space with the Cox proportional hazards model ($\rho = 1$) and the proportional odds model ($\rho = 0$) as special cases.

Though the transformation (1.3) or (1.7) is very different from the transformation for continuous survival analysis, this transformation alone is not new for discrete survival models. In fact, Kalbfleisch and Prentice (2002) suggested an equivalent encompassing formulation

of transformation

$$h(\pi_{i,k}) = \boldsymbol{\beta}^\tau \mathbf{Z}_i(t_k) + h(\pi_{0,k}),$$

where $h(\cdot)$ is a monotone-increasing and twice-differentiable function mapping $[0, 1]$ to $(-\infty, \infty)$ with $h(0) = -\infty$. Three common examples are provided for the function $h(\cdot)$ that yields the grouped relative risk model, the discrete logistic model and the discrete relative risk model. In discrete failure time regression, conditional default probability is traditionally called the hazard at time $t_k$. Here $\pi_{i,k} = 1 - S_{\mathbf{Z}_i}(t_k)/S_{\mathbf{Z}_i}(t_{k-1})$ can also be considered as the differential increment of the cumulative hazard function $\Lambda_{\mathbf{Z}}(t_k)$ (see Kalbfleisch and Prentice 2002, Page 9). Denote the baseline hazard $\pi_{0,k} = 1 - S_0(t_k)/S_0(t_{k-1})$ when $\mathbf{Z} = \mathbf{0}$. Therefore, equations (1.3) and (1.7) can be rewritten as

$$G[-\log(1 - \pi_{i,k})] = exp[\boldsymbol{\beta}^\tau \mathbf{Z}_i(t_k)] G[-\log(1 - \pi_{0,k})].$$

That is, transformation families $G_c(\cdot)$ and $G_\rho(\cdot)$ are actually on the minus logarithm of one minus the hazard at time $t_k$. And equivalently,

$$h(x) = \log[G(-\log(1 - x)],$$

where in our paper $G(\cdot)$ takes the forms of $G_c(\cdot)$ of (1.4) (inverse of logarithmic transformation of Zeng and Lin 2006) and $G_\rho(\cdot)$ of (1.5) (inverse of Box-Cox transformation of Zeng and Lin 2006). Note that when $\mathbf{Z} = \mathbf{0}$, $h(\pi_{0,k}) = \log\{G(-\log(1 - \pi_{0,k})\} = \alpha_k$. Hence the parameter $\alpha_k$ in models (1.1) and (1.2) may be regarded as a reparametrization of the baseline hazard $\pi_{0,k}$.

In this paper, we incorporate a class of logarithmic transformations and Box-Cox transformations with additional transformation scalar parameters $c$ and $\rho$, which enables a flexible class of model family. Application to bankruptcy database and a simulation study in Section 1.4 show promises of the proposed class of transformation model family.

### 1.3.5 Discussion

Doksum and Gasko (1990) pointed out a correspondence between logistic regression models in binary regression analysis and the generalized odds-rate model in survival analysis with time invarying covariates. The binary regression is done for the indicator variable on survival

past a fixed time point. In Section 1.3.1, we similarly show that the discrete transformation (1.3) along with $G_c(\cdot)$ of (1.4), the inverse of logarithmic transformations, and $G_\rho(\cdot)$ of (1.5), the inverse of Box-Cox transformations, with time-varying covariates in survival analysis corresponds to logistic regressions (1.1) or (1.2) on indicators for survival in each discrete time period as if those indicators were independent. This new correspondence links the survival analysis model techniques to the application of bankruptcy probability modeling.

Literature on survival analysis is extensive in biomedical fields. Although it is natural to model the bankruptcy of a company as a survival event, the advanced survival analysis model results have rarely been applied to the bankruptcy prediction. There is a technical reason for this. In most survival analysis theory, the focus is on the survival probability curves from a common starting time $t = 0$, generally a clinic event such as the beginning of a medical treatment or diagnose of the disease. That is, the time used is not the actual calendar time. The literature on modeling of bankruptcy prediction, in contrast, generally adopts calendar time. It is important to understand the reason for the difference. In the biomedical applications, it is reasonable to expect that individuals with same covariate value history (physical characteristics, medical treatments, etc.) will follow the same biological process after, say, a surgical operation. Therefore, those patients should have same probability to survive after one year, regardless the operation was done in calendar year 1985 or year 1990. However, we would not expect companies with same covariate value history (say same financial measures for the last 5 years) to have same one year bankruptcy probability in calendar year 1985 as in calendar year 1990. This is due to the different macroeconomics environment in 1985 versus 1990. With the actual calendar time, we can no longer ensure a common starting time $t = 0$ for all individuals as companies are not all started at the same time. The application of survival analysis models to the bankruptcy prediction needs to take this difference into account. For example, the transformation model (1.8) is not well-defined unless covariate process $\mathbf{Z}(t)$ exists from the common starting time. So it can not be applied directly to the bankruptcy prediction problem.

### 1.3.6 Model Evaluation

Accurate default probability prediction is of great importance for practitioners and regulators to gauge the exposure to default risk, which, as we had learnt from the recent worldwide financial crisis, could have a devastating effect on personal investment and aggregate economy. In the past decade, academic researchers have been striving to develop sophisticated corporate bankruptcy models that provide improved out-of-sample forecasts. In the existing literature, the usefulness or goodness of a prediction model is judged mainly by its ability of accurately distinguishing companies with high default probability from companies with low default probability in the out-of-sample context. Along the release and implementation of Basel II, an accurate default probability becomes even crucial in practice due to its decisive role in determining capitals that banks need to put aside to protect against certain financial and operational risks.

In particular, for a company with covariate values $\mathbf{Z}(t) = \mathbf{Z}_{new}$, its default probability according to the model (1.1) is given by $\pi_{new} = \pi(\mathbf{Z}_{new}; \alpha_t, \boldsymbol{\beta}, c) = 1 - exp\{-G_c^{-1}[exp(\alpha_t + \boldsymbol{\beta}^\tau \mathbf{Z}_{new})]\}$. Hence with parameters estimated from data, the predicted default probability for this company at time $t$ is $\hat{\pi}_{new} = \pi(\mathbf{Z}_{new}; \hat{\alpha}_t, \hat{\boldsymbol{\beta}}, \hat{c})$. Denote $\hat{\pi}_{j,new}$ the predicted probabilities for company $j$ with covariate $\mathbf{Z}_j(t) = \mathbf{Z}_{j,new}$, $j = 1, ..., m$. In practice, we are interested in out-of-sample prediction of default probability at time period $t$ that is in the future beyond the fitted time periods.

Generally, it is recognized that there are two components of the performance of predicted probabilities of binary events (Hosmer and Lemeshow 2000; Sobehart, Keenan and Stein 2001; Wilks 2006; Cook 2008): (a) discrimination, that is, the ability to discriminate between those subjects experiencing the event of interest and those not; (b) calibration, that is, providing correct prediction probability for event occurrence.

The corporate default literature has often reported only measures of discrimination. For example, Shumway (2001) and Chava and Jarrow (2004) use a crude measure of decile ranking to compare the out-of-sample performance using different predictor variables. Alternatively, accuracy ratio is another commonly used gauge of default model prediction evaluation (Duffie, Saita and Wang 2007). Following the literature, the discriminant ability

of our model is evaluated through accuracy ratio, which is defined as twice the difference between the area under the ROC (Receiver Operating Characteristic) curve (AUC) and 0.5. Accuracy ratio of 0 corresponds to the random forecast, and accuracy ratio of 1 corresponds to the perfect forecast.

Note that without proper calibration, models above can be used to produce default scores that rank the risks of default for different companies, but may not provide accurate default probabilities for a portfolio needed in bank reservation level calculations required by BASEL II. In statistics literature, one way to check the calibration is through the Hosmer-Lemeshow test (Hosmer and Lemeshow 1980; Hosmer and Lemeshow 2000). Like the Chi-square goodness-of-fit test statistics, Hosmer-Lemeshow test is a popular way to evaluate the model deviation. A high Hosmer-Lemeshow statistics, or equivalently, a low p-value for the Hosmer-Lemeshow test indicates a poor calibration.

For our default prediction model, the discrimination is achieved through the scores $\hat{\boldsymbol{\beta}}^{\tau}\mathbf{Z}_{j,new}$. Since the default rate in practice is very low, the models with $c = 0$ (grouped relative risk model), $c = 1$ (Shumway's model) and $c = 10$, for example, are all very close for most companies. Thus simply fitting Shumway's model can result in scores that provide similar rankings of default probabilities to our scores. Hence we would expect the accuracy ratios of our model prediction to be close to those of Shumway's model (Chava and Jarrow 2004; Bharath and Shumway 2008; Campbell et al. 2008) or those making Cox proportional hazards assumption (Duffie, Saita and Wang 2007; Duffie et al. 2009). Some slight improvement in accuracy ratio may result from a better estimate of $\boldsymbol{\beta}$ with correct value specification on the transformation parameter $c$.

On the other hand, we do expect our model to improve the calibration over Shumway's model ($c = 1$, $\rho = 0$) or the grouped relative risks model ($c = 0$, $\rho = 1$) if the true transformation parameters do not take those special values. This should be reflected via a lower value of Hosmer-Lemeshow test statistics or a bigger p-value from the Hosmer-Lemeshow test.

In summary, there are two most important gauges for formal evaluation of default probability models: discrimination and calibration. In principal, we are searching for the model

that maximizes the discriminatory power subject to proper calibration. For the binary outcomes such as default, one possible statistical test for calibration is Hosmer-Lemeshow test. For models that are properly calibrated, those with higher accuracy ratio or equivalently higher AUC are more useful for practitioners. It is worth noting that both discriminatory power and calibration performance should be examined using not only the in-sample data but also out-of-sample validation data to prevent over-fitting. In fact, the standard for internal rating based approach to Basel II is to use one-year-ahead default probability prediction and out-of-sample backtesting for default probability model validation. This is implemented in our current application (e.g. Table 1.3).

Note that future studies are warranted in order to incorporate censoring. Under the assumption of independent censoring, for the discrete-time survival model here, the discriminatory power and calibration performance for the one-time-period-ahead default probabilities can be checked as above using uncensored cases. However, generally there may be bias for the tests for calibration (e.g. Hosmer-Lemeshow test) and measures of discriminatory power (e.g. accuracy ratio or AUC) for checking multiple-time-periods-ahead default probabilities. Inverse probability weighting approach (e.g. Gerds and Schumacher 2006; Uno et al. 2011) may be considered. For dependent censoring, a model of the dependence and adjustments to the calibration test and estimation of the discriminatory measures are needed, which certainly deserves future study.

## 1.4 Empirical Results

### 1.4.1 The Data

To measure the distress risk using our proposed transformation survival model, we need an event indicator of bankruptcy for the distressed firms and a set of exploratory variables. In our work, we classify a firm as in distress if it files under either Chapter 7 or Chapter 11 bankruptcy protection code. To obtain the list of distressed firms, we take the firms as default if its reported deletion reason is liquidation or bankruptcy by COMPUSTAT or it was delisted from CRSP due to the same reason. As such, our bankruptcy database includes $1,565$ firms that went bankrupted from January 1981 to December 2006. The

default indicator equals to one in the month that firm was delisted due to Chapter 7 or Chapter 11 bankruptcy filing. All the other exiting reasons such as merger or acquisition would set the default indicators to zero.

Table 1.1 reports the properties of firm defaults by year in more details. We note that the default rate exhibits substantial variation across time. The pattern reflects mainly the fact that firms have more difficulties in fulfilling their financial obligations during business recessions than during business expansions. Specifically, in our sample, the default rate peaks in the year 1991, when the economy was in a recession accompanied by a severe credit crunch due to monetary tightening (e.g., Bernanke, Lown and Friedman 1991). Similarly, the default rate is at an elevated level in the year 2001, when the economy fell into recession following the burst of the technology bubble. Such time-varying default rate highlights the importance of taking into account the distinct macro-economic effect featured by different calendar time on default probability, as in our proposed survival model.

[**Insert Table 1.1**]

To construct the exploratory variables for our model, we merge the daily and monthly equity data from CRSP with the quarterly and annually updated accounting data from COMPUSTAT. We adopt eight covariates as in Campbell et al. (2008), which has been considered to be the state-of-art model in the bankruptcy literature. Furthermore, in a separate working paper on dynamic variable selection, we also confirm that these variables remain significant in explaining default probability. Among these eight covariates, three are accounting ratio measures. Profitability (NIMTA) is calculated by dividing the net income by the market value of the total asset. Leverage (TLMTA) is calculated by dividing total liability by the market value of the total asset. Short-term liquidity (CASHMTA) is calculated by dividing the cash and short-term asset by the market value of the total asset. Here the market value of the total asset is the sum of the firm market equity and its book liability.

We follow Daniel and Titman (2006) to determine the firm book equity in market-to-book ratio (MB) calculation. To avoid the outliers of the book equity, we adjust the book equity by adding 10% of the difference between the book equity and the market equity. Four

market variables based on the equity information are excess return (EXRET), firm's market capitalization or relative size (RSIZE), volatility (SIGMA) and stock price (PRICE). We use the standard deviation of the daily stock price over the previous three month to estimate the volatility. Both the excess return and the firm's relative size are evaluated over the S&P 500 index as the market value. Except for volatility, the other three market measures including the excess return, the firm's relative size, and the price enter the estimation model in its log scale.

To be consistent with previous literature, we further modify our data set in the following ways. Companies report their accounting data with a delay. To ensure we use the accounting information that is available at the time of the forecast, we lag all the annually updated accounting measures, including net income, total liability and the cash and short-term assets by four months and quarterly updated accounting data by two months. All the covariate measures are constructed after such lagging on the accounting data. Great care has also been taken in aligning the fiscal time with the calendar time.

Table 1.2 reports the summary statistics of the eight covariate measures. Panel A gives the summary statistics on the entire $1,812,730$ firm-month observations. Panel B summarizes the eight covariates' statistical property on the default group only, a total of $1,565$ observations. Comparison between Panel A and Panel B demonstrates strong distribution differences among the eight covariates between the entire data set and the default group.

[**Insert Table 1.2**]

### 1.4.2 Results

We apply the proposed class of discrete transformation survival models in (1.1) and (1.2) to the quarterly updated bankruptcy data from 1981 to 2006 using the two-step profile likelihood algorithm described in Section 1.3.2. Figure 1.2 gives the model maximum log-likelihood values over a grid of transformation parameters $c$ and $\rho$. From Figure 1.2, we observe that the log-likelihood decreases monotonically over $\rho$ so that the optimal model resides in the $G_c$ transformation. The log-likelihood increases sharply from $c = 0$ and stabilizes around the interval from $c = 10$ to $c = 12$ with little variation. The maximum is

achieved around the point of $c = 11.5$. Clearly neither Shumway's model ($c = 1$ or $\rho = 0$) nor the grouped relative risk ($c = 0$ or $\rho = 1$) is optimal on the bankruptcy probability prediction for this data set.

[**Insert Figure 1.2**]

Similarly, Figure 1.3 plots the observed values for the log-likelihood function of the fit on the annually updated data set under the transformation functions $G_c$ in (1.4) and $G_\rho$ in (1.5). Again, we observe that the log-likelihood value is maximized neither close to the $c = 1$ for Shumway's model nor $c = 0$ for the grouped relative risk model, but at around $c = 10$.

[**Insert Figure 1.3**]

We further withhold data from 2002 to 2006 for validation purpose and estimate the default probability via expanding window approach. To that end, we predict the probability of default for each validation year on an estimation window from the start of the sampling period up to the forecasting period to eliminate look-ahead bias. For the reason of computational efficiency, we use annually updated firm data. We investigate accuracy ratio and the Hosmer-Lemeshow goodness-of-fit test statistics on the validation data set by comparing three different models: the optimal transformation model with $c = 10$, Shumway's model $c = 1$ and the grouped relative risk model $c = 0$.

Table 1.3 displays our out-of-sample performance evaluation results. We find our optimal transformation model with $c = 10$ yields a slightly better accuracy ratio. The high p-value for the Hosmer-Lemeshow test attests the out-of-sample improvement of our selected model in the overall significance.

[**Insert Table 1.3**]

Proper calibration requires good estimation of the transformation parameter $c$ (or $\rho$) and baseline $\alpha_k = \alpha(t_k)$ in addition to good estimation of covariate parameter $\boldsymbol{\beta}$. When calculating the out-of-sample Hosmer-Lemeshow test, we ignored the prediction of $\alpha(t_k)$ here by using its data estimation. The Hosmer-Lemeshow tests show that the standard models of $c = 1$ or $c = 0$ are not properly calibrated for out-of-sample predictions while

the selected transformation model of $c = 10$ significantly improves the calibration. Using correct transformation parameter also leads to better estimation of $\boldsymbol{\beta}$ thus improving the discrimination as well, although the improvement in this component is small. The out-of-sample prediction of $\alpha(t_k)$ requires further modeling and will be a topic for future research.

### 1.4.3 Asset Pricing Implication

We further investigate Fama and French's (1996) conjecture that investors require a positive return premium for holding distressed stocks. As in Campbell et al. (2008), we measure the distress premium by sorting stocks according to their predicted default probabilities, estimated from the selected optimal discrete transformation survival model. In particular, at the beginning of each year from 1985 to 2006, we update the firm's default probability only using historically available data. We then form 10 portfolios according to their default risk distribution and hold each portfolio for one year. Detailed specification of the cut-off percentile points for such ten portfolios is listed in Table 1.4. Note that those percentile cut-off points are not equally spaced, but provide finer grids to the tail of the distribution. Table 1.4 summarizes our findings.

[**Insert Table 1.4**]

We first investigate the average of simple return[4] in excess of the S&P 500 index return for each portfolio. From Panel A, we observe stocks with high conditional default probabilities have substantially lower returns than do stocks with low conditional default probabilities. For example, the portfolio with the lowest default probability has an annualized average excess return of 3.29%, compared with the $-12.03\%$ for the portfolio with the highest default probability. Such a finding poses a significant challenge to Fama and French's (1996) conjecture that distressed stocks have a higher expected return.

We then examine the question "can stock return anomalies be explained by the three-factor model" (Fama and French 1996) by regressing each portfolio's value-weighted return on the standard Fama and French three factors[5]. Panel B shows portfolio with high distress

[4] equal-weighted return

[5] Data are available at Professor Kenneth R. French's website.

risk tends to have high loadings on the size factor (Small Minus Big or SMB) and value risk factor (High Minus Low or HML), yet an anomalously negative alpha. Alpha is the estimated intercept after fitting the value-weighted return on the three factors. A significant alpha suggests that the risk premium is not fully priced by the three factors. In the first row of Panel A, we observe that the reported alpha for the highest default risk portfolio in the 99th to the 100th percentile is $-18.96\%$ with a t-statistics of $-19.62$. The alpha deviates significantly from 0. Therefore, it shows that distress risk estimated from the selected optimal transformation model cannot be fully explained by the commonly used Fama and French's (1996) three-factor model. Additional risk factor may be needed in order to explain the anomalous stock return. These results confirm the findings in Campbell et al. (2008), where a discrete hazard model with $c = 1$ is used.

Panel C shows the average of the default probabilities of stocks ($\hat{p}$), market capitalization (rSize) and market-to-book equity ratio (MB) within each portfolio. We can see that as the average default probability increases, the portfolio tends to have a monotonically decreased market capitalization. By contrast, the market-to-book equity ratio first decreases and then increases when the default risk is elevated. In summary, our asset pricing implication demonstrated by Table 1.4 imposes challenges to the standard Fama and French's (1996) conjecture, but is consistent with Campbell et al. (2008)'s findings among others.

### 1.4.4 Simulation

We further conduct a simulation study that is designed to mimic the real bankruptcy data. For illustration purpose, suppose there are 26 fixed time periods. At the start of the sampling period, we first generate $N_1 = 4000$ firms. At each following time period, $N_k$ new companies are generated from a Poisson distribution with mean parameter estimated from the real data. For each company, eight time-varying covariates are independently generated at each time period during its lifetime following distributions similar to the real data. For example, one covariate mimicking profitability variable NIMTA is generated from a normal distribution with the same mean and standard deviation as those from the real data while a covariate mimicking PRICE is generated from a lognormal distribution. The discrete baseline values $\alpha_k$ are generated from a random walk process. The default probability follows a transformation

model family (1.1) with the true transformation parameter $c = 10$. The censoring indicators are then generated using a Bernoulli distribution. For evaluation purpose, we apply our proposed transformation survival models on the first 25 time periods, and withhold the last time period for out-of-sample comparison.

Table 1.5 reports some summary statistics for the covariate parameter estimates from the 400 Monte Carlo simulations, including the bias and standard error. We also show the 95% confidence interval coverage, a percentage that the true parameter lies within the 95% confidence interval determined by the asymptotic standard error. Here, the asymptotic standard error is derived by taking the square root of the inverse of the diagonal entries from the fisher information matrix as described in Section 1.3.3. Results from the grouped relative risk model ($c = 0$) are shown in Panel A, discrete logistic model ($c = 1$) in Panel B and model with $c = 10$ in Panel C. Panel D displays the results from our transformation model fit. We note that the estimation biases from discrete logistic model and grouped relative risk model are far worse than those from the model $c = 10$, and yet, our optimal model from the proposed transformation model class achieves the performance very close to the model with $c = 10$. In addition, our optimal model consistently provides much improved 95% confidence interval coverage comparing to the case of $c = 0$ and $c = 1$ for each covariate estimate. Such performance is close to the case as if the transformation parameter $c$ is known.

[**Insert Table 1.5**]

Table 1.6 gives the probabilistic measures. we calculate the mean absolute deviation (MAD) between the estimated and true default probability both on the estimation and validation sample. In addition, we examine accuracy ratio and Hosmer-Lemeshow test statistics across different models. Unlike directly reporting the Hosmer-Lemeshow test statistics as in the empirical study, we report the selection percentage over the grouped relative risk model ($c = 0$), discrete logistic model ($c = 1$) and the optimal model from the proposed transformation model class ($c = \textit{optimal}$) based on their Hosmer-Lemeshow test statistics. We find both of the discrete logistic model and the grouped relative risk model yield much larger MAD than the model with $c = 10$, whereas, our optimal model performs very close to the model with $c = 10$. For accuracy ratio, though our optimal model's performance

remains very close to the model with $c = 10$, the improvement over the two benchmark models with $c = 0$ and $c = 1$ is small. This indicates that a measure like accuracy ratio that evaluates discrimination power may not be very sensitive in this case. However, such findings are consistent with our empirical results. The selection percentage, presented in the last row, shows that according to the Hosmer-Lemeshow goodness-of-fit test statistics, about 99.75% of the simulation runs suggest that neither the classical grouped relative risk model nor the Shumway's discrete hazard model is correct for the default data. From our limited simulation study, Hosmer-Lemeshow test appears to be useful to select the correct model specification.

[**Insert Table 1.6**]

### 1.4.5 Conclusion

We applied our proposed class of discrete transformation survival model to the bankruptcy data from 1981 to 2006. An optimal model with $c$ around 10 was selected using two-step profile likelihood estimation. This model is recommended over Shumway's model ($c = 1$) or the grouped relative risk model ($c = 0$), when accurate default probability prediction is needed. Out-of-sample performance is examined through annual expanding window approach for withholding sample from 2002 to 2006 and is shown with improved accuracy ratio as well as model goodness-of-fit. Further asset pricing implication challenges the famous Fama and French's (1996) conjecture on investors demand risk premiums for distress companies. However, the findings are consistent with some recent literature such as Campbell et al. (2008). Results of a simulation study are consistent with the empirical findings. The proposed class of discrete transformation survival model may be potentially used in other applications with similar characteristics.

## 1.5 Appendix

**Appendix 1:** *Derivations of the discrete transformation models (1.1) or (1.2).*

For our discrete time model where $T$ can only take values at $t_1, ..., t_K$, the survival function $S(t) = P(T > t)$, $t = t_1, ..., t_K$.

$$G[-\log \frac{S_{\mathbf{Z}}(t_k)}{S_{\mathbf{Z}}(t_{k-1})}] = exp[\boldsymbol{\beta}^\tau \mathbf{Z}(t_k)] G[-\log \frac{S_{\mathbf{0}}(t_k)}{S_{\mathbf{0}}(t_{k-1})}]. \tag{A.1}$$

Assuming independent censoring,

$$\pi_{i,k} = Pr(T = t_k | X \geq t_k, \mathbf{Z}(t_k) = \mathbf{Z}_{i,k}) = Pr(T = t_k | T \geq t_k, \mathbf{Z}(t_k) = \mathbf{Z}_{i,k}).$$

Hence it can be represented as

$$\begin{aligned} \pi_{i,k} &= Pr(T = t_k | T \geq t_k, \mathbf{Z}(t_k) = \mathbf{Z}_{i,k}) \\ &= 1 - \tfrac{S_{\mathbf{Z}}(t_k)}{S_{\mathbf{Z}}(t_{k-1})} \\ &= 1 - \exp\{-G^{-1}[G(-\log \tfrac{S_{\mathbf{0}}(t_k)}{S_{\mathbf{0}}(t_{k-1})}) \exp(\boldsymbol{\beta}^\tau \mathbf{Z}_{i,k})]\} \\ &= 1 - \exp\{-G^{-1}[\exp(\alpha_k + \boldsymbol{\beta}^\tau \mathbf{Z}_{i,k})]\} \end{aligned}$$

where $\alpha_k = \log[G(-\log \frac{S_{\mathbf{0}}(t_k)}{S_{\mathbf{0}}(t_{k-1})})]$.

When $G$ belongs to the family (1.4)

$$G_c^{-1}(x) = \begin{cases} \frac{1}{c}\log(1+cx), & c > 0; \\ x, & c = 0. \end{cases}$$

Hence

$$\exp[-G_c^{-1}(x)] = \begin{cases} \frac{1}{(1+cx)^{1/c}}, & c > 0; \\ \exp(-x), & c = 0. \end{cases}$$

This gives

$$\pi_{i,k} = \begin{cases} 1 - \frac{1}{[1+c\exp(\alpha_k + \boldsymbol{\beta}^\tau \mathbf{Z}_{i,k})]^{1/c}}, & c > 0; \\ 1 - \exp[-\exp(\alpha_k + \boldsymbol{\beta}^\tau \mathbf{Z}_{i,k})], & c = 0. \end{cases} \tag{A.2}$$

When $c > 0$, note left hand side of equation (A.1) indicates

$$G[-\log \frac{S_{\mathbf{Z}}(t_k)}{S_{\mathbf{Z}}(t_{k-1})}] = \frac{1}{c}[exp(c(-\log \frac{S_{\mathbf{Z}}(t_k)}{S_{\mathbf{Z}}(t_{k-1})}))-1] = \frac{1}{c}[exp(\log \frac{S_{\mathbf{Z}}^c(t_{k-1})}{S_{\mathbf{Z}}^c(t_k)})-1] = \frac{1}{c}\frac{S_{\mathbf{Z}}^c(t_{k-1}) - S_{\mathbf{Z}}^c(t_k)}{S_{\mathbf{Z}}^c(t_k)}.$$

Right hand side of equation (A.1) indicates

$$exp[\boldsymbol{\beta}^\tau \mathbf{Z}(t_k)] G[-\log \frac{S_{\mathbf{0}}(t_k)}{S_{\mathbf{0}}(t_{k-1})}] = exp[\boldsymbol{\beta}^\tau \mathbf{Z}(t_k)] \frac{1}{c} \frac{S_{\mathbf{0}}^c(t_{k-1}) - S_{\mathbf{0}}^c(t_k)}{S_{\mathbf{0}}^c(t_k)}.$$

Equation (A.1) indicates

$$\frac{S_{\mathbf{Z}}^c(t_{k-1}) - S_{\mathbf{Z}}^c(t_k)}{S_{\mathbf{Z}}^c(t_k)} = exp[\boldsymbol{\beta}^\tau \mathbf{Z}(t_k)] \frac{S_{\mathbf{0}}^c(t_{k-1}) - S_{\mathbf{0}}^c(t_k)}{S_{\mathbf{0}}^c(t_k)}.$$

In particular, when $c = 1$ this leads to

$$\frac{S_{\mathbf{Z}}(t_{k-1}) - S_{\mathbf{Z}}(t_k)}{S_{\mathbf{Z}}(t_k)} = exp[\boldsymbol{\beta}^\tau \mathbf{Z}(t_k)] \frac{S_{\mathbf{0}}(t_{k-1}) - S_{\mathbf{0}}(t_k)}{S_{\mathbf{0}}(t_k)},$$

equivalently

$$\frac{\pi_{i,k}}{1 - \pi_{i,k}} = exp[\boldsymbol{\beta}^\tau \mathbf{Z}(t_k)] \frac{\pi_{i,0}}{1 - \pi_{i,0}}.$$

This is equivalent to binary regression model with logit link with panel data, which is the popular so-called discrete hazard model in bankruptcy literature following Shumway (2001).

When $c = 0$, Equation (A.1) indicates

$$-\log \frac{S_{\mathbf{Z}}(t_k)}{S_{\mathbf{Z}}(t_{k-1})} = exp[\boldsymbol{\beta}^\tau \mathbf{Z}(t_k)](-\log \frac{S_{\mathbf{0}}(t_k)}{S_{\mathbf{0}}(t_{k-1})}).$$

or

$$\frac{-\log S_{\mathbf{Z}}(t_k) - (-\log S_{\mathbf{Z}}(t_{k-1}))}{-\log S_{\mathbf{0}}(t_k) - (-\log S_{\mathbf{0}}(t_{k-1}))} = exp[\boldsymbol{\beta}^\tau \mathbf{Z}(t_k)].$$

This is equivalent to proportional hazards model in continuous case when regarding $-\log[S(t)]$, the cumulative hazard function in continuous case, as step functions with jumps at $t = t_1, ..., t_K$ and the difference of $-\log[S(t)]$ is equivalent to hazard rate function in continuous sense.

For $G$ in the family (1.5),

$$G_\rho^{-1}(x) = \begin{cases} \log\{1 + \frac{1}{\rho}[\exp(\rho x) - 1]\}, & \rho > 0; \\ \log(1 + x), & \rho = 0. \end{cases}$$

Hence

$$\exp[-G_\rho^{-1}(x)] = \begin{cases} \frac{1}{1 + \frac{1}{\rho}[\exp(\rho x) - 1]}, & \rho > 0; \\ \frac{1}{1+x}, & \rho = 0. \end{cases}$$

This gives

$$\pi_{i,k} = \begin{cases} 1 - \frac{1}{1 + \frac{1}{\rho}\{\exp[\rho \exp(\alpha_k + \boldsymbol{\beta}^\tau \mathbf{z}_{i,k})] - 1\}}, & \rho > 0; \\ \frac{1}{1 + \exp(\alpha_k + \boldsymbol{\beta}^\tau \mathbf{z}_{i,k})}, & \rho = 0. \end{cases}$$

**Appendix 2:** *Derivation of transformation equation (1.8).*

Zeng and Lin (2006) assumes that

$$\Lambda_{\mathbf{Z}}(t) = G^{-1}\{\int_0^t e^{\boldsymbol{\beta}^T \mathbf{z}(s)} d\Lambda(s)\},$$

where $\Lambda(.)$ is an unspecified increasing function. The baseline hazard is

$$\Lambda_{\mathbf{0}}(t) = G^{-1}\{\int_0^t 1 d\Lambda(s)\} = G^{-1}\{\Lambda(t)\}.$$

Notice that their $\Lambda(t)$ is not the baseline hazard, $G^{-1}\{\Lambda(t)\}$ is. So their $\Lambda(t) = G\{\Lambda_{\mathbf{0}}(t)\}$. We have $d\Lambda(t) = dG\{\Lambda_{\mathbf{0}}(t)\}$. Therefore, $dG\{\Lambda_{\mathbf{Z}}(t)\} = e^{\boldsymbol{\beta}^T \mathbf{z}(t)} d\Lambda(t) = e^{\boldsymbol{\beta}^T \mathbf{z}(t)} dG\{\Lambda_{\mathbf{0}}(t)\}$.

| **Year** | **Num of Bankruptcy** | **Number of Firms** | **Default Rate** (in %) |
|---|---|---|---|
| 1981 | 32 | 4085 | 0.78 |
| 1982 | 45 | 4345 | 1.04 |
| 1983 | 36 | 4474 | 0.81 |
| 1984 | 49 | 4785 | 1.02 |
| 1985 | 53 | 4944 | 1.07 |
| 1986 | 80 | 4971 | 1.61 |
| 1987 | 39 | 5203 | 0.75 |
| 1988 | 68 | 5444 | 1.25 |
| 1989 | 73 | 5404 | 1.35 |
| 1990 | 81 | 5362 | 1.51 |
| 1991 | 125 | 5345 | 2.34 |
| 1992 | 93 | 5357 | 1.74 |
| 1993 | 47 | 5615 | 0.84 |
| 1994 | 50 | 6508 | 0.77 |
| 1995 | 51 | 7003 | 0.73 |
| 1996 | 57 | 7226 | 0.79 |
| 1997 | 71 | 7587 | 0.94 |
| 1998 | 95 | 7554 | 1.26 |
| 1999 | 63 | 7188 | 0.88 |
| 2000 | 69 | 7008 | 0.99 |
| 2001 | 79 | 6738 | 1.17 |
| 2002 | 79 | 6310 | 1.25 |
| 2003 | 44 | 5881 | 0.75 |
| 2004 | 25 | 5634 | 0.44 |
| 2005 | 20 | 5572 | 0.36 |
| 2006 | 10 | 5523 | 0.18 |

Table 1.1: **Firm data**: This table lists the number of defaults and number of active firms each year in our sampling period. The number of active firm is calculated by averaging over the number of active firms across all months of the year.

| Variable | NIMTA | TLMTA | EXRET | RSIZE | SIGMA | MB | PRICE | CASHMTA |
|---|---|---|---|---|---|---|---|---|
| Panel A: Entire Data Set | | | | | | | | |
| Mean | −0.0073 | 0.4285 | −0.0131 | −10.6007 | 0.5996 | 2.1824 | 2.1981 | 0.1209 |
| Median | 0.0045 | 0.3970 | −0.0096 | −10.7276 | 0.4685 | 1.5392 | 2.4532 | 0.0460 |
| Std. Dev | 0.1070 | 0.2868 | 0.1701 | 2.0776 | 0.4952 | 1.7278 | 1.3281 | 0.8102 |
| observations: 1,823,241 | | | | | | | | |
| Panel B: Bankruptcy Group | | | | | | | | |
| Mean | −0.1137 | 0.7164 | −0.2373 | −13.4158 | 1.6761 | −0.3841 | −0.3678 | 0.1911 |
| Median | −0.0444 | 0.8386 | −0.1110 | −13.6183 | 1.4030 | 0.0171 | −0.4700 | 0.0292 |
| Std. Dev | 0.4151 | 0.2905 | 0.5180 | 1.7834 | 1.4362 | 1.5942 | 1.5809 | 1.6681 |
| observations: 1,562 | | | | | | | | |

Table 1.2: **Summary Statistics for quarterly updated firm-month observations**

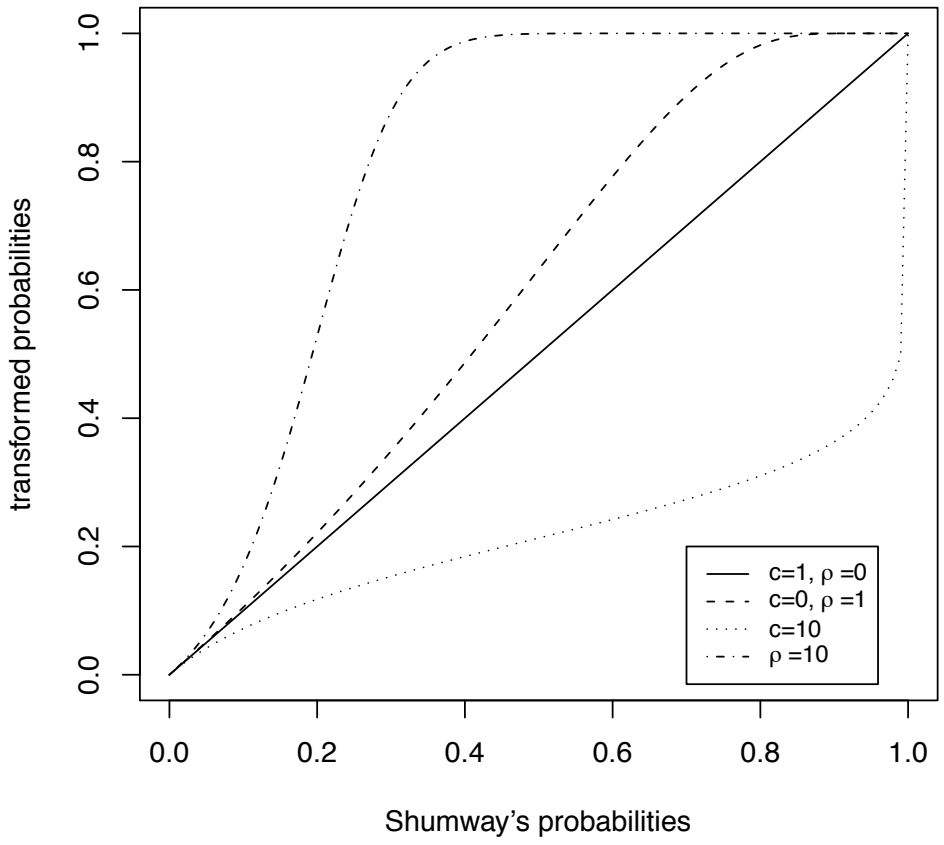



Figure 1.1: This plots how the same scores are translated into different probabilities under different transformation parameter $c$ or $\rho$ values. The solid line gives the probabilities converted according to Shumway's model $c = 1$ or $\rho = 0$. The optimal parameter value from the fit on the real bankruptcy data is around $c = 10$.

| **c** | **Accuracy Ratio** | **Hosmer Lemeshow Test** | |
|---|---|---|---|
| | | $\chi^2$ | **p-value** |
| $c = 0$ | 0.7956 | 22.8885 | 0.0035 |
| $c = 1$ | 0.8006 | 19.4592 | 0.0126 |
| $c = 10$ | 0.8095 | 8.7790 | 0.3613 |

Table 1.3: Out-of-Sample Accuracy Ratio and Hosmer Lemeshow Test statistics and its p-value for $c = 0$ (grouped relative risk model), $c = 1$ (Shumway's model) and $c = 10$ (the optimal selected transformation model) under the transformation function $G_c$ on annually updated bankruptcy data.

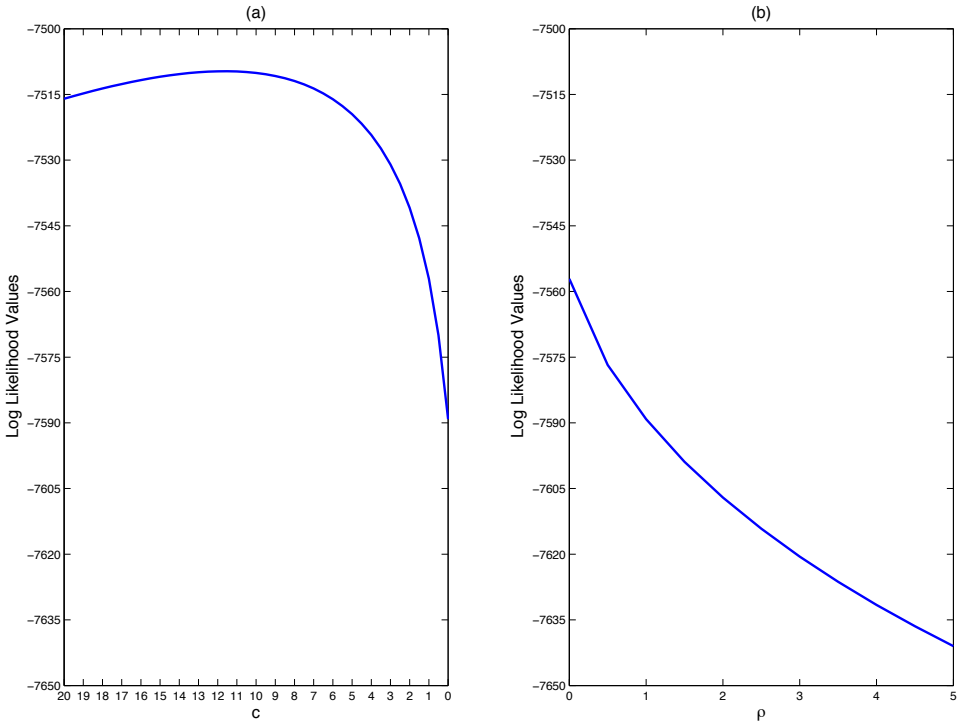



Figure 1.2: This plots the observed values of the log-likelihood functions for the quarterly bankruptcy data from 1981 to 2006: (a) pertains the transformation function $G_c$; (b) pertains the transformation function $G_\rho$.

| **Portfolios** | 0005 | 0510 | 1020 | 2040 | 4060 | 6080 | 8090 | 9095 | 9599 | 9900 |
|---|---|---|---|---|---|---|---|---|---|---|
| Panel A: Portfolio Return | | | | | | | | | | |
| Return(%) | 3.2875 | 2.1037 | 1.3872 | 0.5437 | −0.3416 | −1.3635 | −2.8139 | −4.5521 | −6.6981 | −12.0232 |
| Panel B: Fama-French Three-Factor Regression Coefficient | | | | | | | | | | |
| Alpha(%) | 0.7998 | −0.3106 | −0.6905 | −1.2457 | −2.4045 | −4.2112 | −6.7159 | −9.4513 | −12.1528 | −18.9572 |
| | (11.21)** | (−3.37)** | (−8.04)** | (−13.25)** | (−13.52)** | (−18.48)** | (−19.34)** | (−22.62)** | (−21.78)** | (−19.62)** |
| RM | −0.1295 | −0.0119 | 0.0767 | 0.1239 | 0.1839 | 0.3275 | 0.4276 | 0.6669 | 0.6641 | 0.6746 |
| | (−7.11)** | (−0.50) | (3.50)** | (5.16)** | (4.05)** | (5.63)** | (4.82)** | (6.25)** | (4.66)** | (2.74)** |
| HML | −0.2178 | 0.0476 | 0.2256 | 0.3332 | 0.1819 | 0.2373 | 0.1442 | 0.4827 | 0.5235 | 0.7920 |
| | (−7.95)** | (1.34) | (6.84)** | (9.23)** | (2.66)** | (2.71)** | (1.08) | (3.01)** | (2.44) | (2.13)* |
| SMB | 0.0289 | 0.1767 | 0.1905 | 0.3195 | 0.5177 | 0.8115 | 1.0051 | 1.3279 | 1.2795 | 1.5402 |
| | (1.30) | (6.16)** | (7.12)** | (10.91)** | (9.34)** | (11.43)** | (9.29)** | (10.20)** | (7.36)** | (5.12)** |
| Panel C: Portfolio Characteristics | | | | | | | | | | |
| $\hat{p}$(%) | 0.0270 | 0.0448 | 0.0671 | 0.1205 | 0.2321 | 0.4599 | 0.8844 | 1.5032 | 2.7754 | 6.5600 |
| rSize | −5.1747 | −6.2812 | −6.6933 | −7.0707 | −7.9217 | −9.2124 | −10.1094 | −10.3917 | −10.9139 | −11.1155 |
| MB | 3.6068 | 2.9953 | 2.5172 | 2.0419 | 1.9462 | 1.9039 | 1.6422 | 2.8665 | 0.7333 | 1.1215 |

Table 1.4: **Asset Pricing Results**: We sort all the stocks according to their predicted default probabilities, obtained from the selected optimal discrete transformation survival model. Ten portfolios are constructed based on percentile cut-off. For example, first column "0005" refers to the portfolio of stocks with the default probability below the 5th percentile and the second column displays portfolio of stocks in 5th to 10th percentile. Panel A reports the average of simple excess return for each portfolio. Panel B summarizes the results of estimate on constant alpha and three factor loadings with their corresponding t-statistics obtained by regressing the value-weighted return on the Fama-French three (RM, HML,SMB) factors. Panel C reports the average fitted default probability ($\hat{p}$), market capitalization (rSize) and market-to-book equity ratio (MB). * denotes significant at 5%, ** denotes significant at 1%.

| **Coefficient** | **SIGMA** | **NIMTA** | **LTMTA** | **CASHMTA** | **RSIZE** | **MBE** | **PRICE** | **EXRET** |
|---|---|---|---|---|---|---|---|---|
| Panel A: $c = 0$ | | | | | | | | |
| Bias | −0.0825 | 0.0416 | −1.0259 | 0.0816 | 0.0408 | −0.0003 | 0.2371 | 0.3052 |
| Standard Error | 0.0294 | 0.0526 | 0.1195 | 0.0634 | 0.0078 | 0.0073 | 0.0155 | 0.0298 |
| Coverage of 95% C. I. | 0.1275 | 0.7825 | 0.0000 | 0.6100 | 0.0000 | 0.9150 | 0.0000 | 0.0000 |
| Panel B: $c = 1$ | | | | | | | | |
| Bias | −0.0627 | 0.0319 | −0.8499 | 0.0616 | 0.0307 | −0.0002 | 0.1779 | 0.2292 |
| Standard Error | 0.0304 | 0.0580 | 0.1015 | 0.0688 | 0.0086 | 0.0078 | 0.0191 | 0.0342 |
| Coverage of 95% C. I. | 0.3650 | 0.8700 | 0.0000 | 0.7950 | 0.0750 | 0.9525 | 0.0000 | 0.0000 |
| Panel C: $c = 10$ | | | | | | | | |
| Bias | −0.0027 | 0.0050 | −0.0003 | 0.0008 | −0.0001 | 0.0000 | −0.0015 | −0.0021 |
| Standard Error | 0.0402 | 0.0762 | 0.0253 | 0.0919 | 0.0110 | 0.0104 | 0.0169 | 0.0374 |
| Coverage of 95% C. I. | 0.9625 | 0.9450 | 0.9525 | 0.9325 | 0.9650 | 0.9675 | 0.9425 | 0.9550 |
| Panel D: $c = optimal$ | | | | | | | | |
| Bias | −0.0027 | 0.0050 | −0.0014 | 0.0009 | 0.0000 | 0.0000 | −0.0014 | −0.0019 |
| Standard Error | 0.0402 | 0.0762 | 0.0523 | 0.0917 | 0.0110 | 0.0104 | 0.0190 | 0.0392 |
| Coverage of 95% C. I. | 0.9675 | 0.9425 | 0.9450 | 0.9325 | 0.9575 | 0.9675 | 0.9375 | 0.9550 |

Table 1.5: Summary of Covariate Parameter Estimates from 400 Simulations, including biases and standard errors. Coverage of 95% Confidence Interval is calculated through asymptotic results using estimated Fisher information.

| **Metrics** | $c = 0$ | $c = 1$ | $c = 10$ | $c = optimal$ |
|---|---|---|---|---|
| Panel A: In-Sample Evaluation | | | | |
| MAD | 0.0090(0.0034) | 0.0069(0.0025) | 0.0013(0.0004) | 0.0014(0.0005) |
| Panel B: Out-of-Sample Evaluation | | | | |
| MAD | 0.0151(0.0073) | 0.0119(0.0057) | 0.0017(0.0016) | 0.0018(0.0016) |
| Accuracy Ratio | 0.7244(0.1175) | 0.7251(0.1167) | 0.7268(0.1160) | 0.7268(0.1160) |
| Selection Percentage (%) | 0.00 | 0.25 | $--$ | 99.75 |

Table 1.6: In-Sample and Out-of-Sample Evaluation of default probability estimates from 400 Simulations. Shown are Mean Absolute Deviation (MAD), Accuracy ratio and Hosmer-Lemeshow statistics based Selection Percentage.

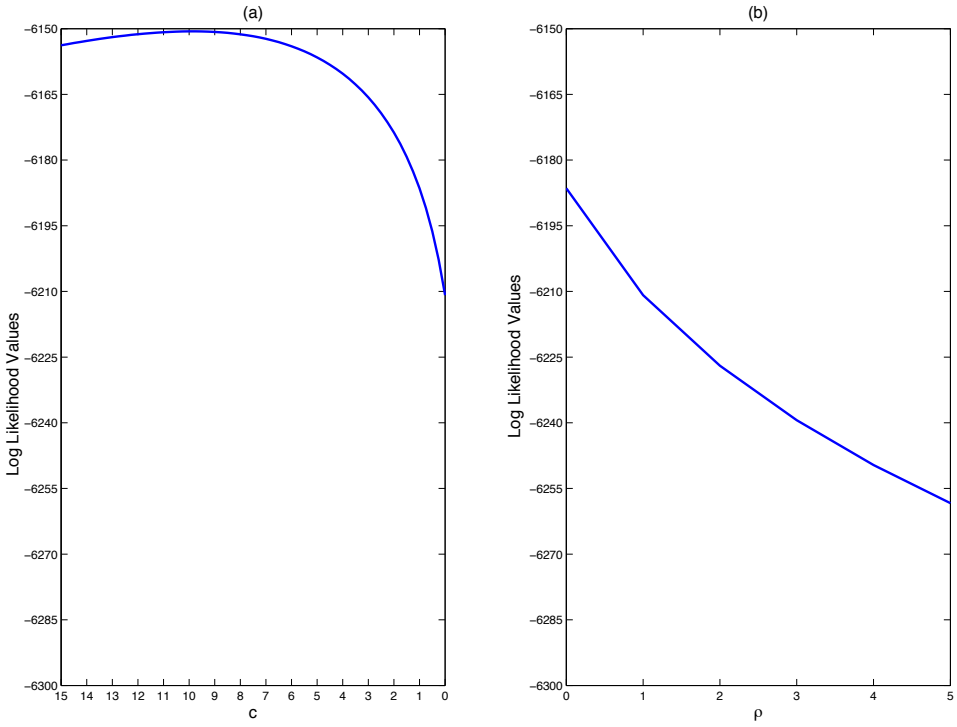



Figure 1.3: This plots the observed values of the log-likelihood functions for the annual bankruptcy data from 1981 to 2006: (a) pertains the transformation function $G_c$; (b) pertains the transformation function $G_\rho$.

# 2 Chapter Two

## Variable Selection and Corporate Bankruptcy Forecasts

## 2.1 Introduction

Accountants and financial economists have considered various predictive variables in the reduced-form corporate bankruptcy forecast model. Earlier studies, e.g., Beaver (1966), Altman (1968), Ohlson (1980), and Zmijewski (1984), have routinely used accounting ratios, i.e., financial ratios constructed from only accounting data, as a gauge of default risk. In an attempt to improve the empirical performance of the reduced-form model, Shumway (2001) advocates for incorporating market variables in the bankruptcy forecast, in addition to two accounting ratios.[1] In a similar vein, Campbell, Hilscher, and Szilagyi (2008; CHS thereafter) introduce new market variables and financial ratios; and they also propose a modification of the accounting ratios adopted in Shumway (2001) by using the market value of assets rather than the book value. While Shumway (2001) and CHS (2008) have shown their models exhibit noticeable improvement over the models proposed in previous studies, none of existing studies has provided a formal analysis on the relative importance for a comprehensive set of bankruptcy predictors. In this paper, we try to fill the gap by introducing a robust variable selection technique proposed by Tibshirani (1996)—the least absolute shrinkage and selection operator (LASSO).

Statisticians develop variable-selection methods to achieve two main objectives—(1) identifying relevant predictive variables and (2) improving prediction accuracy (see, e.g., Fan and Li (2001)). A formal variable-selection analysis thus allows us to shed new light on the corporate bankruptcy forecast literature in two important ways. First, it enables us to identify from an exhaustive set of bankruptcy predictors proposed in existing studies a parsimonious subset of the most relevant ones. Such identification has important implications for testing bankruptcy theories, designing regulations in credit markets, and conducting credit risk analysis.

[1] Accounting researchers, e.g., Ohlson (1980), have noted that market data are potentially important in bankruptcy forecasts. These authors, however, do not pursue this investigation because their main research interest is the informativeness of accounting data for bankruptcy rather than the search for a good bankruptcy forecast model.

Second, as we confirm in this paper, the selected reduced-form model has superior in-sample and out-of-sample predictive power and outperforms the prominent models in the existing literature.

LASSO penalizes regression coefficients through a shrinkage method and thus provides a sparse variable-set solution. LASSO has been widely used in variable-selection studies (see, e.g., Efron, Hastie, Johnstone, and Tibshirani (2004)) and is a state-of-the-art variable selection tool. LASSO enjoys the easy interpretability as the traditional subset variable selection does but has added benefits of the stability of model selection and prediction accuracy. Compared with other commonly used variable selection methods such as the subset or stepwise selection, LASSO has several desirable statistical properties that suit particularly for the main empirical issues that we try to address in this paper. First, given the rareness of default events, robustness is a necessary requirement of variable section techniques used for bankruptcy forecasts. LASSO is quite robust to small perturbations of data changes. Second, the shrinkage method improves prediction accuracy. Third, LASSO naturally produces an entire variable selection path to provide the relative importance of the selected variables. Finally, LASSO is computationally efficient, especially when there are a large number of candidate predictors.

We adopt LASSO variable selection using time-varying covariates for the panel data. Variable selection and modeling that use full information of the panel data are obviously appealing. Shumway (2001) shows that the discrete hazard model using time-varying panel data has important advantages compared with static models using cross-sectional data (e.g., Altman (1968), Ohlson (1980) and Zmijewski (1984)). Specifically, the latter ignore the fact that firms change over time and thus may produce biased and inconsistent bankruptcy probability estimates.

In the empirical analysis, we construct a comprehensive bankruptcy database by merging daily and monthly equity data from the Center for Research in Security Prices (CRSP) with annual financial information from COMPUSTAT. A company is in default if it files for either Chapter 7 (liquidation) or Chapter 11 (reorganization) bankruptcy protection. We include an exhaustive list of 39 financial ratios and market variables that have been used in the bankruptcy literature as candidate default-risk predictors. As in Shumway (2001), Chava and Jarrow (2004), CHS (2008), and others, we model the bankruptcy risk using the discrete hazard model.

Over the full sample spanning the 1980 to 2009 period, LASSO selects seven predictive variables into the reduced-form bankruptcy forecast model. We find strong support for Shumway's (2001) argument of including market variables in the bankruptcy forecast. Two market variables advocated by Shumway (2001) and CHS (2008), i.e., stock return volatility and the excess stock return, and one newly-proposed market variable by CHS (2008), i. e. stock price, enter into the LASSO-selected reduced-form model. Shumway (2001) shows that (1) the net income to total assets ratio and (2) the total liabilities to total assets ratio constructed using accounting information are significant predictors even when controlling for market variables in the bankruptcy forecast. CHS (2008), however, suggest that we should modify these two variables using the market value of assets instead of the book value. Our formal variable selection analysis allows us to shed light on this debate: LASSO selects CHS's modified financial ratios but not Shumway's (2001) original variables.

Of CHS's (2008) eight predictive variables, five variables are selected into our reduced-form bankruptcy forecast model. These results indicate that CHS have done a reasonably good job in selecting the bankruptcy predictors.[2] Nevertheless, the formal variable selection analysis

[2] Of 5 predictive variables proposed by Shumway (2001), 4 variables enter into our LASSO-selected reduced-form model either directly or in a modified form.

improves CHS's model in two important ways. First, LASSO identifies two new predictive variables—(1) the current liabilities to total assets ratio and (2) the total debts to total assets ratio. This result confirms the important role of accounting ratios in the bankruptcy forecast. Second, three of CHS's predictive variables, i.e., the market capitalization, the market to book ratio, and the ratio of cash and short-term assets to the market value of assets, are not included in the LASSO-selected reduced-form model.

Note that variables that are not selected do not necessarily correspond to their statistical insignificance in in-sample estimation. For example, using our data, we confirm CHS's (2008) finding that market capitalization, the market to book ratio, and the ratio of cash and short-term assets to the market value of assets have statistically significant in-sample predictive power for the default risk, although these variables are not selected. The variable selection results are strikingly stable across various subsample periods; specifically, we select the identical set of predictive variables over the 1980 to 2000, 1980 to 2002, 1980 to 2005, and 1996 to 2009 periods. These results also highlight a potentially important advantage of LASSO—it allows us to identify a stable parsimonious set of most relevant explanatory variables that may have superior out-of-sample predictive power.[3] Indeed, the LASSO-selected bankruptcy model outperforms CHS's model in the out-of-sample forecast.

The distance-to-default (DD) constructed from Merton's (1974) structural model is a popular bankruptcy risk measure for practitioners. CHS (2008) and Bharath and Shumway (2008), however, argue that DD provides little additional information about future bankruptcy beyond the variables used in their reduced-from models. The formal variable selection analysis

[3] In a similar vein, Pesaran and Timmermann (1995) advocate for selecting a parsimonious set of stock market return predictors on their performance in the historical sample, and find that the data-determined model has significant out-of-sample predictive power for market returns. However, unlike the bankruptcy forecast, Pesaran and Timmermann (1995) find that the set of selected stock market return predictors change substantially across time.

allows us to test this conjecture directly. When we add DD as a candidate predictor along with the other 39 predictive variables, DD is not selected by LASSO and the set of selected predictors is identical to that without DD as a candidate. In the out-of-sample forecast, the performance of DD alone model is similar to, or slightly better than, that of CHS's reduced-form model. In contrast, the LASSO-selected reduced-form model performs noticeably better than DD alone model.

While CHS (2008) have advocated for constructing financial ratios using the market value of assets, accounting researchers, e.g., Beaver, McNichols, and Rhie (2005), have reiterated the relevance of accounting ratios in the bankruptcy forecast by showing that their predictive power is strikingly stable across time. We provide support for both arguments. Specifically, LASSO selects the market value of assets for the net income to total assets ratio and the total liabilities to total assets ratio but chooses the book value for the current liabilities to total assets ratio and the total debts to total assets ratio. These results possibly reflect the fact that, while the former is a measure of a company's ability to pay off debts, the latter is arguably an indicator of a company's tolerance toward bankruptcy risk. Specifically, George and Hwang (2010) argue that the target or book leverage depends on a company's bankruptcy costs—companies with high (low) bankruptcy costs tend to have low (high) target leverage.[4]

The remainder of the paper proceeds as follows. We discuss briefly the discrete hazard model and the LASSO variable selection method in Section 2.2. Data description is presented in Section 2.3. Section 2.4 reports the empirical findings. In Section 2.5, we present a comparison study on distance-to-default and expected-default-frequency from Merton's structural model under discrete hazard model framework. We offer some concluding remarks in Section 2.6.

## 2.2 Model and LASSO Variable Selection

[4] See also Johnson, Chebonenko, Cunha, D'Almeida, and Spencer (2011).

### 2.2.1 Discrete Hazard Model

To investigate each firm's default risk at any specific time, we use the discrete hazard model to forecast bankruptcy over the next time period (Shumway (2001)). The discrete hazard model implies a logistic regression between the dependent variable and time-varying covariates. For one-year-ahead or equivalently twelve-month-ahead default risk prediction, the model can be expressed in the formula below

$$\mathrm{P}\left(\mathrm{Y}_{i,t+12}=1\middle|\mathrm{Y}_{i,t+12-1}=0,\mathrm{X}_{i,t}\right)=\frac{e^{\beta_0+\beta^*\mathrm{X}_{i,t}}}{1+e^{\beta_0+\beta^*\mathrm{X}_{i,t}}} \quad , \tag{2.1}$$

where $\mathrm{X}_{i,t}$ is the covariate vector of time-varying firm-specific explanatory variables. In particular, $\mathrm{X}_{i,t}$ may be a vector of observed financial ratios or market variables at time *t* for firm *i*. Coefficient $\beta$ is a covariate effect parameter vector and $\beta_0$ is a scalar parameter. *t* can be any observation time period. In this paper, *t* represents month-end for monthly data. $\mathrm{Y}_{i,t+12}$ is the associated default indicator after 12 months or one year. $\mathrm{Y}_{i,t+12}$ is set to 1 if the company *i* files for bankruptcy protection code within 12 months from time *t* and zero otherwise.

### 2.2.2 LASSO Variable Selection

Many accounting ratios and market variables have been introduced to improve the prediction accuracy in the bankruptcy research literature (e.g. Beaver (1966), Altman (1968), Beaver et al. (2005), Shumway (2001), Chava and Jarrow (2004), and CHS (2008)). However, there is no consensus on which variables should be included in the reduced-form bankruptcy model of equation (2.1). Moreover, explanatory determinants are often established subjectively based on expert or field judgment prior to the analysis. To formally identify relevant variables from a comprehensive variable set considered in the literature, we introduce a state-of-the-art LASSO variable selection using the available panel data.

Variable selection has long been an important topic in the statistics literature. Variable selection is essential to identify relevant predictive variables and to improve prediction accuracy. Recent development in variable selection literature shows promising evidence from the emerging of new penalized shrinkage approaches. In his seminal work, Tibshirani (1996) proposes a regularization approach LASSO, for simultaneous parameter estimation and variable selection that yield nice statistical properties. In contrast to subset regression that either zeros a coefficient or inflates it, shrinkage method tries to zero some coefficients, shrink others and thus more stable (Breiman (1995); Fan and Li (2001)). This state-of-the-art variable selection method is not only computationally efficient, but also possesses superior performance in terms of stability and prediction accuracy. This robust feature is particularly important for the bankruptcy forecasting because corporate bankruptcy has been a rather rare event, even if all the historical information of panel data is considered.

In our discrete hazard model framework, the LASSO estimate is obtained by minimizing the negative log-likelihood function and a roughness penalty on the sum of the absolute value of the covariate parameter, the so-called "$l_1$ penalty" or equivalently with certain "$l_1$ constraint". LASSO uniformly penalizes the coefficients. Such penalization or constraints placed on parameter estimate enjoys nice properties theoretically and computationally. LASSO also provides the nice feature of continuous shrinkage in coefficient estimate, with some coefficients to exact zeros.

Best-subset and stepwise (backward/forward) variable selection are commonly-used classical variable selection approaches. When the total number of variables M in consideration is moderate, best-subset selection involves selection of the best model from $2^{M-1}$ different combination according to some criterion such as Akaike information criterion (AIC) (Akaike

(1974)). The computation becomes forbiddingly intensive as M increases. For example, when a comprehensive set of 39 accounting ratios and market variables are considered in this paper, a combination of $2^{39-1}$ or about 275 billion models need to be built for an exhaustive best-subset search. In practice, stepwise-subset selection is usually adopted as a surrogate for best-subset by sequentially deleting or adding one variable at a time based on some significant tests. However, due to the nature of stepwise selection, algorithm may yield a local optimal solution rather than the global optimal solution. Through comprehensive simulations studies, Breiman (1995) shows that subset regression is instable even with small changes of data. For example, removing one sample data point can result in drastically different selection of significant variables. Such feature is certainly not favorable in terms of model robustness, which is a crucial feature desired by the bankruptcy prediction. Furthermore, the subset variable selection procedures ignore the stochastic errors in the variable selection stage (Fan and Li (2001)).

### 2.2.3 Model Evaluation

Corporate bankruptcy modeling involves binary events of default versus non-default. To evaluate the overall in-sample performance of the discrete hazard model, formal model information criterion based on the negative of log-likelihood and a complexity penalty can be used. For example, AIC is a popular goodness-of-fit measurement for likelihood-based model selection using two times the number of parameters as a penalty. A model with the smaller AIC is desirable. Generally, a good model attempts to balance its accuracy and complexity, which are often termed as the tradeoff between bias and variance by the statisticians. For instance, a bankruptcy prediction model with increasing number of explanatory variables will always yield better in-sample likelihood, however, not necessarily better AIC, and most importantly possibly worse out-of-sample prediction due to overfitting or data-snooping. The primary goals to

introduce LASSO variable selection process in this paper are indeed two folds: to identify relevant predictive variables and to achieve high prediction accuracy in a parsimonious model.

The area under the ROC (Receiver Operating Characteristic) curve (AUC) (Hosmer and Lemeshow (2000)) is a very popular measure to evaluate the model discriminatory power, that is, the ability to discriminate between the binary events, bankruptcy and non-bankruptcy, using the predicted bankruptcy probabilities. Equivalently, accuracy ratio is calculated as double of the difference between AUC and 0.5. It is another commonly used gauge for corporate bankruptcy model evaluation (Duffie, Saita and Wang (2007)). In particular, accuracy ratio of 0 or AUC of 0.5 corresponds to a random forecast, and accuracy ratio or AUC of 1corresponds to a perfect forecast.

Accurate out-of-sample bankruptcy prediction is essential in addition to good in-sample fit. In fact, Basel II standard for internal rating based approach is to use one-year-ahead default probability prediction and out-of-sample back-testing for default model validation. We compare the out-of-sample predictive performance and implement a similar strategy used in Shumway (2001) to predict the bankruptcy probability in the recent withheld test sample. We first build our discrete hazard rate model using the bankruptcy data over the training period. With the variables selected by LASSO and the coefficient estimates from the discrete hazard model fitting, we predict the probabilities of bankruptcy for the firms over the testing period and report the out-of-sample accuracy ratio and AUC.

We also report the out-of-sample decile-rankings, which are commonly used in the bankruptcy literature (Shumway (2001); Chava and Jarrow (2004)). Specifically, for each year in the testing period, we rank the predicted probability of bankruptcy in deciles. The companies with highest probabilities of default are ranked in the first decile. With the smallest probabilities

of bankruptcy, companies are ranked in the last decile. Year by year, we aggregate the total number of bankruptcy filings. Meanwhile, we count the number of firms with its predicted probability of bankruptcy ranked in the first decile for each year. The smaller discrepancy between the two counts implies a better predictive model in the out-of-sample performance.

## 2.3 Data

We obtain bankruptcy information and market and financial variables by merging the daily and monthly equity data from CRSP with annually updated accounting data from COMPUSTAT over the 1980 to 2009 period. Note that companies report their accounting data with a delay. To ensure we use the accounting information that is available at the time of the bankruptcy forecast, all the annually updated accounting measures are lagged by four months. For each predictor variable, we truncate at the lowest and highest percentiles to alleviate the effect of outliers.

To estimate the discrete hazard model, we need to construct an event indicator for bankruptcy. A company is in default if it files for bankruptcy under either Chapter 7 (liquidation) or Chapter 11(reorganization) bankruptcy protection. The bankruptcy indicator of a company equals 1 if the company exits the database due to bankruptcy filing and equals zero otherwise. Specifically, we assign the bankruptcy indicator a value of zero for healthy firms and firms that exited from our database due to other reasons such as merger and acquisition. In Figure 2.1, we plot the total number of firms that file for bankruptcy in each year over the 1980 to 2009 period. Consistent with findings in earlier studies, Figure 2.1 shows that bankruptcy filings exhibit strong countercyclical patterns with peaks following the 1981-82, 1990-91, 2001, and 2007-09 business recessions.

We consider an exhaustive list of 39 financial and market variables as candidate

bankruptcy predictors; and Table 2.1 provides a brief description for each variable.[5] Those predictive variables are drawn from previous studies in bankruptcy literature, including Beaver (1966), Altman (1968), Ohlson (1980), Zmijewski (1984), Shumway (2001), Chava and Jarrow (2004), Dwyer, Kocagil, and Stein (2004), Beaver, McNichols, and Rhie (2005), Härdle, Lee, Schäfer, and Yeh (2009), Bharath and Shumway (2008), CHS (2008), Ding, Tian, Yu, and Guo (2012), and many others. Earlier studies, e.g., Beaver (1966), Altman (1968), Ohlson (1980), Zmijewski (1984), use various accounting ratios in the bankruptcy forecast, and Altman's (1968) Z-score and Ohlson's (1980) O-score have been the standard distress risk measures for both practitioners and academic researchers. Accounting researchers, e.g., Ohlson (1980), have conjectured that including market variables may improve substantially the bankruptcy forecast, and Shumway (2001) first provides empirical support for this conjecture by applying the discrete hazard model to panel data. Specifically, Shumway (2001) shows that three market variables—the relative market capitalization (RSIZE), the stock return in excess to the market return (EXCESS RETURN), and stock return volatility (SIGMA)—have significant predictive power for bankruptcy risk. Shumway (2001) also finds that two accounting ratios, the net income to total assets ratio (NIAT) and the total liabilities to total assets ratio (LTAT), are also significant bankruptcy predictors. Overall, Shumway (2001)'s reduced-form model performs substantially better than those proposed in earlier accounting studies, e.g., Altman (1968) and Zmijewski (1984).

Shumway's (2001) market-variable-augmented reduced-form model has become popular in the bankruptcy forecast literature, and CHS (2008) try to improve its empirical performance in

[5] The variables used in this study are available only for publicly traded companies. Dwyer, Kocagil, and Stein (2004) propose some alternative predict variables in the forecast of credit risk for privately held companies. As a robustness check, we also include these variables as candidate predictors and find that none of them is selected by LASSO. For brevity, these results are not reported but are available on request.

three ways. First, CHS add a new market variable, the stock price, as a bankruptcy predictor. Second, CHS advocate for constructing financial ratios using the market value of assets rather than the book value. That is, CHS replace NIAT and LTAT by the net income to the market value of total assets ratio (NIMTA) and the total liabilities to the market value of total assets ratio (LTMTA), respectively. Last, CHS include two new financial ratios as bankruptcy predictors: The market-to-book equity ratio (MB) and the ratio of cash and short-term investment to the market value of total asset (CASHMTA).[6] CHS find that their model has a better in-sample fit than does Shumway's (2001) model. Nevertheless, neither Shumway (2001) nor CHS choose the variables in their reduced-form models via a formal variable selection analysis, and we try to fill the gap in this paper.

In Merton's (1974) bond pricing model, the likelihood of default or the distance-to-default (DD) depends on the difference between the face value of the firm's debts and the market value of its assets divided by the volatility of the firm's asset value. The distance-to-default is a leading alternative bankruptcy risk measure, and there is an ongoing debate about the relative performance of the structural versus the reduced-form bankruptcy forecast model. Hillegeist, Keating, Cram and Lundstedt (2004) find that the default probability derived from the structural model performs substantially better than the Z-score or O-score in the bankruptcy forecast. CHS (2008) and Bharath and Shumway (2008), however, find that the distance-to-default provides no additional information about future default risk beyond the market variables and financial ratios employed in their reduced-form model. To address this issue, we follow Vassalou and Xing (2004) and construct the distance-to-default using CRSP and COMPUSTAT data.[7]

[6] MB correlates negatively with the cross-section of stock returns, and Fama and French (1996) suggest that this relation possibly reflects the fact that MB correlates negatively with distress risk.

[7] Practitioners, e.g., Moody's KMV, adopt the empirical distribution of the distance-to-default estimated from a large database, which may potentially yield better fitting and prediction results than our simple distance-to-default

## 2.4 Empirical Analysis

### 2.4.1 LASSO Variable Selection Results

In Figure 2.2, we report the LASSO variable selection results for the full sample spanning the 1980 to 2009 period.[8] The upper panel illustrates the evolution of estimated coefficients on all candidate predictive variables listed in Table 2.1 over the LASSO variable selection process. The horizontal axis indicates the constraint—the maximum value for the sum of absolute coefficients; for each constraint, the vertical axis reports the respective coefficient estimates of all candidate predictive variables. LASSO estimates are close to zero for restrictive constraints, and variables are sequentially selected into the predictive regression as their LASSO estimates increase in magnitude and become nonzero when the constraint is relaxed. Variables with stronger predictive power will enter the process earlier, showing their relatively higher importance. The lower panel of Figure 2.2 illustrates the evolution of in-sample AIC corrected with a factor of finite sample size (AICC) as the constraint becomes less restrictive. Similar to AIC, AICC provides a goodness-of-fit measurement with regard to information loss (see, e.g., Hurvich and Tsai (1989)); a smaller value of AICC indicates a better model fitting.

Shumway (2001) and CHS (2008) advocate for incorporating market information in bankruptcy forecast because compared with accounting information, market information has three advantages. First, the stock price is a forward looking variable that incorporates all available information. Second, the stock volatility is a direct determinant of the default probability in Merton's (1974) structural model. Third, the market value is a more accurate measure of the true value of assets than the book value. Consistent with Shumway (2001) and

---

measure (see Hamilton, Sun and Ding (2011)). However, the information of the empirical distribution is proprietary and is unavailable to us for comparison.

[8] In this paper, we implement empirical analysis using SAS software. As a robustness check, we find the same LASSO variable selection results using R provided by Efron, Hastie, Johnstone, and Tibshirani (2004).

CHS's conjecture, Figure 2.2 shows that the stock price (PRICE) and the stock return volatility (SIGMA) are the first two predictive variables that are selected by LASSO. Similarly, the excess stock return (EXCESS RETURN) is also selected by LASSO albeit at a relatively late stage. We also find support for CHS argument for using the market value of assets instead of the book value. Specifically, Shumway (2001) use the book value of equity in the construction of the net income to total assets (NIAT) and the total liabilities to total assets ratio (LTAT). CHS use the market value of equity for the net income to total assets (NIMTA) and the total liability to total assets (LTMTA). We find that both NIMTA and LTMTA are selected in the bankruptcy forecast, while NIAT and LTAT are not. Overall, our variable selection results are strikingly consistent with the variables advocated by Shumway (2001) and CHS (2008). Specifically, except for the market capitalization, all the variables proposed by Shumway (2001) enter into the LASSO variable selection either directly or in a modified form. Similarly, LASSO chooses five out of eight variables used in CHS; two other CHS variables, corporate cash holdings (CASHMTA) and the market-to-book equity ratio (MB), are not selected by LASSO, however.

Figure 2.2 also offers new insight on the bankruptcy predictors. Two financial ratios constructed using only accounting data, the current liabilities to total book assets ratio (LCTAT) and the total debt to total book assets ratio (FAT), enter the LASSO-selected bankruptcy forecast model. The accounting ratios LCTAT and FAT are common risk measurements in evaluating a company's ability to pay off its debts. LCTAT and FAT provide incremental information about future default risk possibly because the leverage measure constructed using the book value reveals a company's tolerance toward the default risk. Specifically, LCTAT and FAT reveal a company's target leverage level. This interpretation is consistent with the endogenous leverage hypothesis advanced by George and Hwang (2010), who argue that target leverage level depends

on the company's bankruptcy costs. Intuitively, if a company has high bankruptcy costs, it is optimal for the company to maintain a low target leverage level to reduce the default risk. Therefore, LCTAT and FAT forecast bankruptcy risk because they are proxies for risk tolerance; for example, ceteris paribus, companies that more tolerant with bankruptcy risk are more likely to run into bankruptcy in future. To the best of our knowledge, this link of accounting variables with future bankruptcy risk is novel. In the next subsection, we show the accounting ratios are statistically significant in in-sample estimate and improve the out-of-sample performance of the reduced-form model as well.

As a robustness check, we repeat the LASSO analysis using various subsample periods, including the 1980 to 2000, 1980 to 2002, 1980 to 2005 and 1996 to 2009 periods. Interestingly, we find the set of LASSO-selected variables are strikingly stable across time.

### 2.4.2 In-Sample Estimation and Out-of-Sample Forecast

In Table 2.2, we present the discrete hazard model estimation results over the entire bankruptcy database, spanning from 1980 to 2009. Column 1 reports the results for the reduced-form model with LASSO-selected variables. The predictive variables are all statistically significant at the 1% level with expected signs. For comparison, in column 2, we also report the results for CHS's (2008) model, which are similar to those reported in CHS, although we use an updated sample period. The LASSO-selected model has a lower AIC value than the CHS model, indicating that the former has less information loss and thus provides a better fit for the bankruptcy data. Similarly, with a higher AUC value, the LASSO-selected model has better discriminatory power than the CHS model. As a robustness check, we re-estimate the LASSO-selected model and the CHS model using the sample spanning the 1980 to 2002 period, and entries in column 3 and column 4 show that results are qualitatively similar to those reported in

column 1 and column 2 of Table 2.2. To summarize, as expected, the LASSO-selected model provides a better in-sample explanation for the bankruptcy data than the CHS model.

We then evaluate our model's out-of-sample predictive ability by splitting the bankruptcy data into a training sample ending in 2002 and a testing sample over the 2003 to 2009 period. Specifically, we employ the discrete hazard model on the bankruptcy records over the training period, and evaluate its out-of-sample predictive performance on the testing sample. Over the holdout sample period 2003 to 2009, we sort stocks equally into ten portfolios, with the expected default probability decreasing from decile 1 to decile 10. In Table 2.3, we report the percentage of correctly-identified bankruptcy filings in each decile and the out-of-sample accuracy ratio. Consistent with the in-sample estimation results reported in Table 2.2, we find that LASSO-selected model exhibits a better discriminatory power than the CHS (2008) model in the out-of-sample tests. The LASSO-selected model delivers an almost 80 percent correct prediction rate in the two top deciles (column 1), comparing with 66 percent for the CHS model (column 2). The CHS model yields an out-of-sample accuracy ratio of 0.636 (equivalent to AUC of 0.818), which is lower than that of 0.682 (equivalent to AUC of 0.841) for the LASSO-selected model. To summarize, the variables selected via the LASSO method display superior out-of-sample predictive performance.

## 2.5 A Comparison with Distance-to-Default

In this section, we investigate whether the distance-to-default provides additional predictive power when included in the reduced-form model. Specifically, we add the distance-to-default to the candidate predictor set and then apply the LASSO variable selection method to determine the most important forecasting variables over the full sample spanning the 1980 to 2009 period. Interestingly, Figure 2.3 shows that LASSO variable selection coefficient path is

almost identical to that reported in Figure 2.2, which is obtained without the distance-to-default as a candidate predictor. The distance-to-default is not selected by LASSO, and including the distance-to-default as a candidate variable does not affect our results in any qualitatively manner.

In Table 2.3, we summarize the predictive power of the distance-to-default in out-of-sample tests. We find that the distance-to-default only model (column 4) exhibits an out-of-sample AUC value of 0.824 (or equivalently, an accuracy ratio of 0.648). Comparing with CHS (2008) model (column 2), the distance-to-default only model shows improved overall discriminatory ability, but its predictive ability is still weaker than the LASSO-selected reduced-form model (column 1).

As a robustness check, we augment the LASSO-selected reduced model by the distance-to-default. Though the in-sample coefficient estimate of the distance-to-default remains statistically significant in the augmented LASSO-selected reduced-form model, we do not observe any significant improvement when adding the distance-to-default to the LASSO-selected reduced-form model in the out-of-sample context (column 3).

Therefore, in the discrete hazard model setup, the distance-to-default provides little supplementary information about future bankruptcy risk beyond the market variables and accounting variables used in the previous reduced-form models. Our finding differs from those reported by Hillegeist et al. (2004) because these authors compare the default probability constructed from the structural model with only accounting ratios. In contrast, our results provide a formal statistical confirmation for the conjecture by Bharath and Shumway (2008) and CHS (2008), who emphasize the importance of market information in the bankruptcy forecast.

## 2.6 Conclusion

We introduce a state-of-the-art variable selection method, LASSO, to the discrete hazard

model of corporate bankruptcy and document three important findings. First, we find that accounting ratios provide significant supplemental information about future default beyond market variables and financial ratios constructed using the market value of assets. Second, the reduced model selected via the LASSO method performs better in out-of-sample prediction than the models adopted in the previous studies, including the CHS (2008) model. Last, the distance-to-default does not provide additional predictive power in the reduced-form models.

**Table 2.1: Variable Description**

| Variable | Description | Variable | Description |
|---|---|---|---|
| NIAT | Net Income / Total Asset | APSALE | Accounts Payable / Sales |
| NISALE | Net Income / Sales | LOG(AT) | log(Total Assets) |
| OIADPAT | Operating Income / Total Asset | INVCHINVT | Growth of Inventories / Inventories |
| OIADPSALE | Operating Income / Sales | FFOLT | Funds from Operations / Total Liabilities |
| EBITAT | Earnings before Interest and Tax / Total Asset | MVEF | Market Equity (Yearly) / Total Debit |
| (EBIT+DP)/AT | (Earnings before Interest and Tax + Amortization and Depreciation) / Total Asset | LT/(LT+MKET) | Total Liabilities / (Total Liabilities + Market Equity) |
| EBITSALE | Earnings before Interest and Tax / Sales | RELCT | Retained Earnings / Current Liabilities |
| SEQAT | Equity / Total Asset | CASHAT | Cash and Short-term Investment / Total Assets |
| REAT | Retained Earnings / Total Asset | LCTSALE | Current Liabilities / Sales |
| (LCT-CH)/AT | (Current Liabilities – Cash) / Total Asset | FAT | Total Debt / Total Assets |
| LTAT | Total Liabilities / Total Assets | LCTAT | Current Liabilities / Total Asset |
| LOG(SALE) | log(Sale) | NIMTA | Net Income /(Market Equity + Total Liabilities) |
| CHAT | Cash / Total Assets | LTMTA | Total Liabilities /(Market Equity + Total Liabilities) |
| CHLCT | Cash / Current Liabilities | CASHMTA | Cash and Short-term Investment /(Market Equity + Total Liabilities) |
| QALCT | Quick Assets / Current Liabilities | RSIZE | Log(Market Capitalization) |
| ACTLCT | Current Assets/ Current Liabilities | PRICE | Log(Price) |
| WCAPAT | Working Capital / Total Assets | MB | Market-to-Book Ratio |
| LCTLT | Current Liabilities / Total Liabilities | SIGMA | Stock Volatility |
| INVTSALE | Inventories / Sales | EXCESS RETURN | Excess Return Over S&P 500 Index |
| SALEAT | Sales / Total Assets | | |

Note: This table provides description for 39 bankruptcy predictors that we consider in the variable selection analysis.

**Table 2.2: Discrete Hazard Model Estimations**

| | LASSO | CHS | LASSO (1980-2002) | CHS (1980-2002) |
|---|---|---|---|---|
| Panel A: Parameter Estimations | | | | |
| LCTAT | 0. 5641<br>(3.30)** | | 0. 6557<br>(3.63)** | |
| FAT | 0. 0013<br>(5.57)** | | 0. 0013<br>(5.41)** | |
| NIMTA | -1.0104<br>(5.74)** | -1.1949<br>(6.63)** | -1.1475<br>(6.20)** | -1.3940<br>(7.38)** |
| LTMTA | 1.3582<br>(10.26)** | 1.7785<br>(13.22)** | 1.1910<br>(8.45)** | 1.6707<br>(11.74)** |
| CASHMTA | | -0.7096<br>(3.07) ** | | -0.9904<br>(3.81)** |
| RSIZE | | -0.0939<br>(3.81)** | | -0.1130<br>(4.24)** |
| PRICE | -0.5644<br>(17.05)** | -0.5330<br>(13.63)** | -0.5630<br>(16.53)** | -0.5142<br>(12.58)** |
| MB | | 0.0693<br>(3.92)** | | 0.0810<br>(4.48)** |
| SIGMA | 0.5491<br>(7.92)** | 0.5367<br>(7.76)** | 0.4472<br>(6.22)** | 0.4293<br>(5.99)** |
| EXCESS RETURN | -0.8803<br>(5.22)** | -0.8769<br>(5.18)** | -0.8320<br>(4.71)** | -0.8332<br>(4.69)** |
| INTERCEPT | -7.8232<br>(63.23)** | -8.8070<br>(26.74)** | -7.6472<br>(59.83)** | -8.8584<br>(25.03)** |
| Panel B: Goodness-of-Fit Statistics | | | | |
| AIC | 14683 | 14712 | 13035 | 13053 |
| AUC | 0.711 | 0.710 | 0.720 | 0.717 |

Notes: The table reports the estimation results of the discrete hazard model. Unless otherwise indicated, we use the annual sample spanning the 1980 to 2009 period. Column "LASSO" is the LASSO-selected reduced-model. Column "CHS" is the CHS (2008) model. Column "LASSO (1980-2002)" is the LASSO-selected reduced-model for the 1980 to 2002 period. Column "CHS" is the CHS model (2008) for the 1980 to 2002 period. In-sample AIC and AUC (the area under the ROC curve) is shown in the third row. The absolute z-statistics are reported in the parenthesis, and ** denotes significance at the 1% level.

**Table 2.3: Out-of-Sample Performance over the year 2003 to 2009**

| | LASSO | CHS | LASSO+DD | DD |
|---|---|---|---|---|
| Accuracy Ratio | 0.682 | 0.636 | 0.682 | 0.648 |
| AUC | 0.841 | 0.818 | 0.841 | 0.824 |
| Percentage of Bankruptcy Filings | | | | |
| 1 | 59.62 | 58.65 | 59.62 | 55.77 |
| 2 | 19.23 | 7.69 | 18.27 | 20.19 |
| 3 | 5.77 | 12.5 | 7.69 | 7.69 |
| 4 | 5.77 | 7.69 | 2.88 | 3.84 |
| 5 | 0.96 | 5.77 | 3.85 | 1.92 |
| 6-10 | 8.65 | 7.69 | 7.69 | 10.57 |

Note: The table report three out-of-sample performance measures, including the out-of-sample accuracy ratio and the out-of-sample AUC (area under the ROC curve), and decile ranking. Column "LASSO" is the LASSO-selected reduced-model. Column "CHS" is the CHS (2008) model. Column "LASSO+DD" is the LASSO-selected model augmented by the distance-to-default. Column "DD" is the reduced-form model with the distance-to-default as the only predictive variable.

Figure 2.1: Number of Bankruptcy Filings in Each Year: 1980 to 2009.

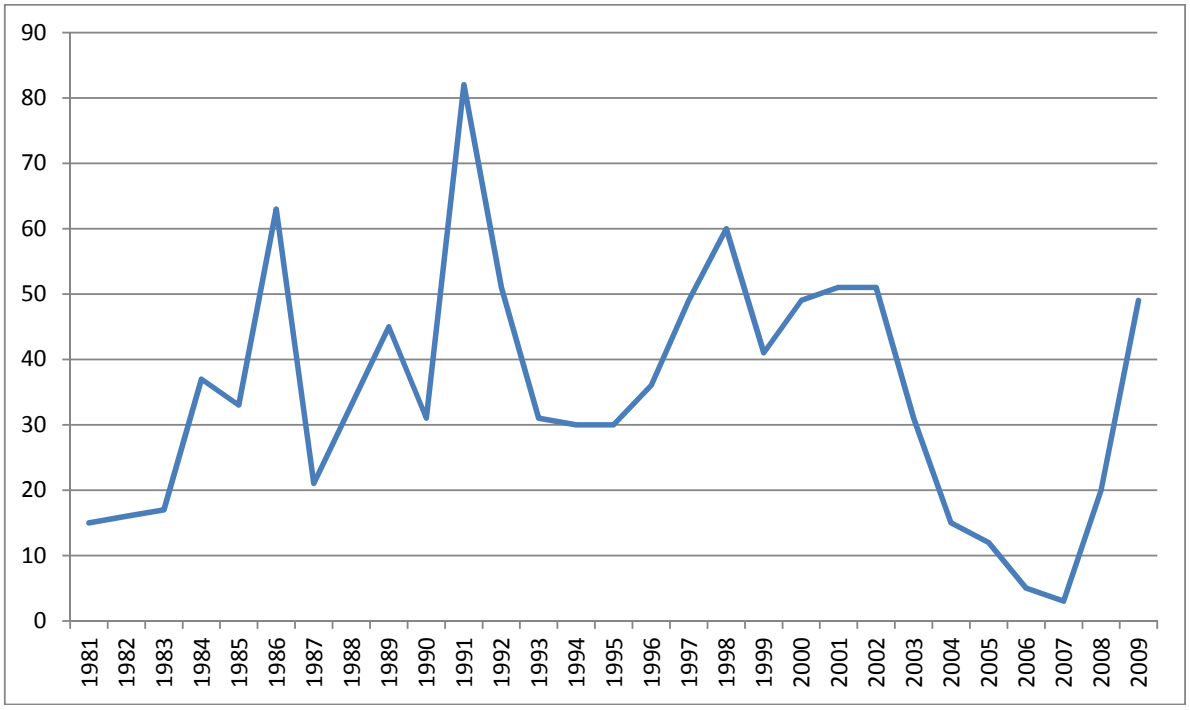


Figure 2.2: Coefficient Paths using LASSO Variable Selection with 39 Explanatory Variables

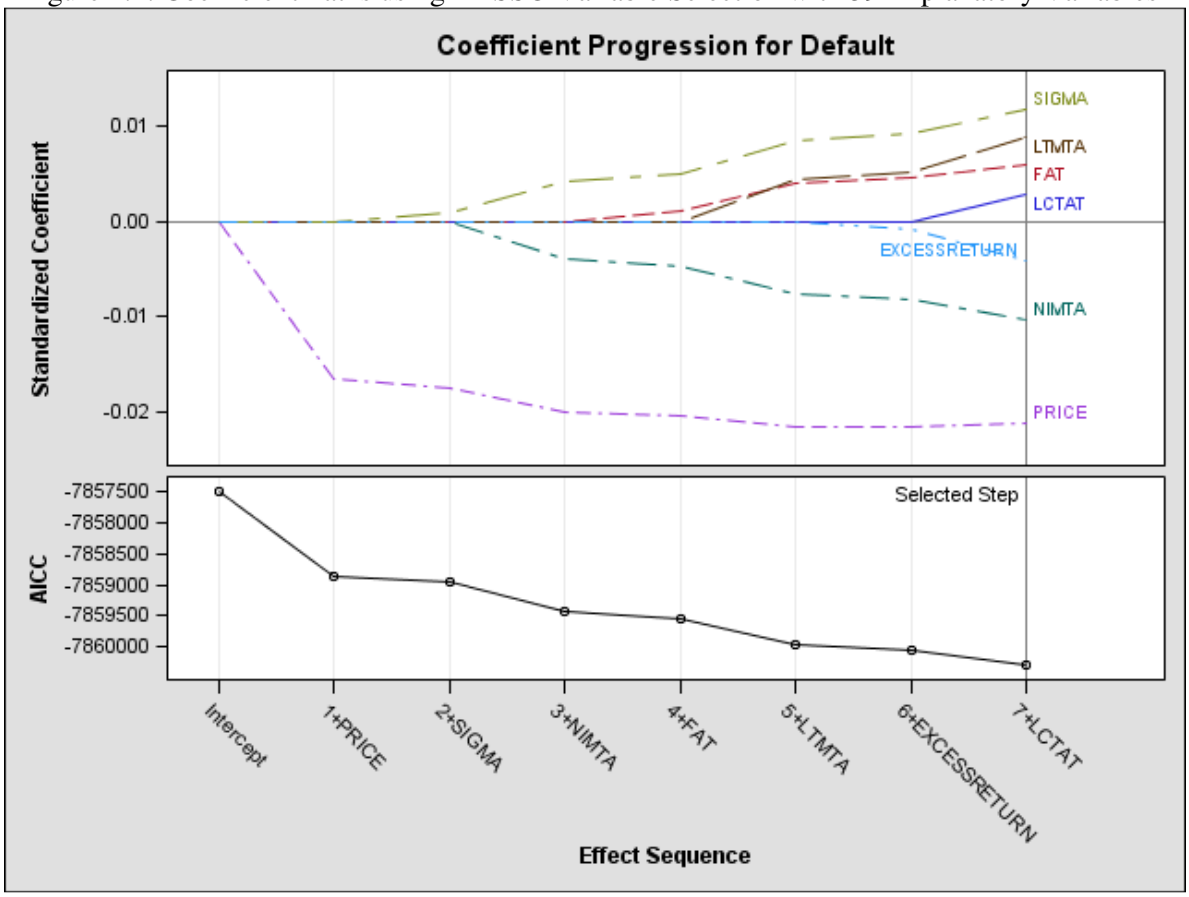
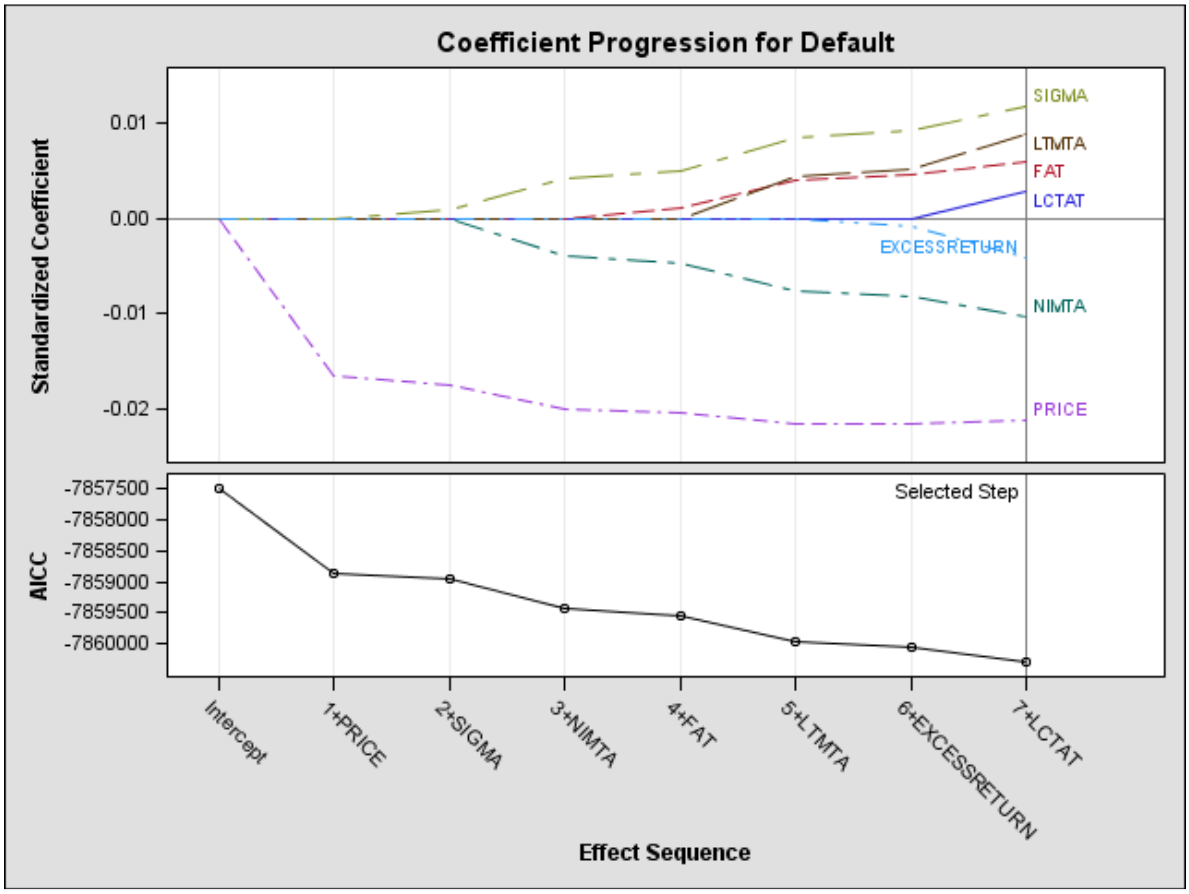


Figure 2.3: Coefficient Paths using LASSO Variable Selection with 40 Explanatory Variables and Distance to Default



# 3 Chapter Three

# Data Sample Selection Issues for Bankruptcy Prediction

## 3.1 Introduction

There is continuing interest in developing and refining corporate bankruptcy prediction studies, especially given today's tumultuous market environment. Corporate bankruptcy not only incurs serious financial loss to its creditors, but also has a high cost to the society and the country's economy (Warner, 1977). In the United States, there are 245,695 reported bankruptcies filings made in the first quarter of year 2008 alone. While in 2002, the reported default loss reached 100 millions. This loss soared to an extremely high level of a trillion dollars today. Consequently, bankruptcy prediction studies that aid financial and investment decision-making have become very important to all those involved: owners, shareholders, lenders, suppliers, and government (Dimitras *et al.*, 1996).

Within the corporate bankruptcy arena, researchers have intensively devoted themselves to develop bankruptcy prediction models, for example, Altman (1968), Ohlson (1980), Odom and Sharda(1990) and Chava and Jarrow (2004). However, the issue of data sample selection, an indispensible and crucial step necessary for testing any bankruptcy prediction model, has received considerable less attention.

Methodological issues on data selection were first investigated by Zmijewski (1984). In general, there are two major sampling techniques on data sample selection: choice based sampling technique and complete data sampling technique (Zmijewski, 1984). Choice based sampling technique implies keeping all of the available bankrupted company records in the data sample, while at the same time, keeping only part of the non-bankrupted companies to match with the bankrupted ones. Such matching is conducted either by random selection or by some criteria, like industry code, size of the company, etc. This approach is widely used in studies such as Beaver (1966), Altman (1968), Norton and Smith (1979), where models are estimated on data samples obtained by combining all distressed firms with the exact same number of the matched non-distressed firms. In addition, with the extensive number of artificial intelligent studies used in the corporate bankruptcy forecasting, choice based sampling technique has been applied as one of the most common data selection methods in neural networks (Odom and Sharda, 1990; Tam and Kiang, 1992; Latcher *et al*., 1995),

support vector machines (Härdle *et al*., 2005; Kim and Sohn, 2010) and etc. Choice based sampling technique successfully remedies the potential problem of extremely low frequency rate of bankruptcy events in the population. However, due to the inconsistent bankruptcy rates between the data sample and the population, choice based sampling technique might introduce biased parameter estimates and probability estimates. On the other hand, the complete data sampling technique carries all of the available companies' records subject to a "complete data criterion" in population into the data sample. Data samples used in Ohlson (1980), Vassalou and Xing (2004), Chava and Jarrow (2004), Bharath and Shumway (2008) are constructed by maintaining their entire "known" corporate records. This data sample selection approach effectively eliminates the previous estimation bias but it might require more computational efforts.

Despite the existence of different data sample selection methods, there have been few studies conducted to explore data sample selection issues in the corporate bankruptcy research. This paper uses both a corporate bankruptcy data set and simulation examples to examine the data selection issues. For the corporate data set, we constructed an updated bankruptcy database consisting of firms traded either on "NYSE", "AMEX", or "NASDAQ" from the "Compustat" database.

For the simulation examples, we conducted a full simulation study using the Monte Carlo method. Simulation studies are important tools for model validation in the statistics and engineering literature, they are not popular in the finance literature. One of the nice features of simulation is the fact that the underlying true patterns are known. Such feature is of particular interest under the framework of Basel II. Basel II significantly promotes the importance of validating the bankruptcy prediction model. However, corporate bankruptcy probabilities can only be obtained from estimation. Alternatively, with simulation, one can gain better insight into the effect of data sample selection methods on the bankruptcy predictions, and derive more solid and convincing methodology. Monte Carlo simulation studies can effectively eliminate random sample errors through multiple replications, resulting in reliable simulation conclusions. Hence, Monte Carlo simulations are applied in

order to investigate the effect of data sample selection methods on different bankruptcy prediction studies more thoroughly.

Bankruptcy classification is important to help the creditors make financial decisions. To measure the goodness of fit for data selection methods, several common criteria are used in this work. In the context of forecasting binary bankruptcy decisions, predicted classifications could result in two types of misclassification errors. One is referred as Type I error, the error of misclassifying a bankrupted firm as a non-bankrupted one. Type II error is considered as misclassifying a non-bankrupted firm as a bankrupted one. As in most existing bankruptcy prediction studies, we assess the overall misclassification rate, i.e. we count the total misclassification errors regardless of the error type. Such an approach is equivalent to assigning a symmetric cost to these two misclassification errors (Hsieh, 1993). However, it is worth noting that in the bank decision context, the costs associated with these two types of misclassification errors might differ. From the empirical study of *commercial bank loan* in Altman *et al*. (1977), it is demonstrated that the cost of misclassifying a bankrupt firm as a non-bankruptcy is approximate 35 times the cost of the other misclassification error. In Weiss and Capkun (2005), Type I error can be compared to the cost of lending to a defaulted firm. The loss includes both of the lending principle and interest. The Type II error cost is considered as the opportunity cost for not lending to a healthy firm. Thus, compared to Type I error, cost of Type II error is much less or even negligible. To address such asymmetric misclassification cost issues, we incorporate a higher error cost of misclassifying a bankrupted firm as a non-bankruptcy, and compare the weighted misclassification rate.

On the other hand, knowing the probability of bankruptcy is of great interest, like the credit risk and bond pricing studies (Merton, 1974). For example, as one of the four key parameters used in the internal rating based (IRB) approach (Schuermann and Hanson, 2004), the probability of default assessment has a crucial influence on credit loss estimation. The product of the probability of default and the loss given default yields the so-called *expected loss*. This resulting expected loss estimate is the key part of the New Accord Capital Basel. New Accord Capital Basel is currently used by the major banks on a daily basis to report the

regulatory capital (Crouchy *et al*., 2000). To evaluate the accuracy of the predicted probability, the simulation study compares the differences between the underlying true probability and the estimated probability of bankruptcy.

It is worth noting that all the goodness-of-fit measurements stated above are evaluated on the same hold-out prediction data sample. This design provides a consistent benchmark of the out-of-sample predictive performance for different data sample selection methods.

In our study for both simulation and empirical data, we discovered that for binary bankruptcy classification problems, we concurred with Zmijewski (1984) findings in terms of negligible choice based sample bias in the overall classification rate for the financial distress models, if the model is assessed under symmetric misclassification cost. Within the scenario of an asymmetric misclassification error cost, our results indicate that applying 0.5, a fixed cut-off probability, for different data sampling methods leads to an increased number of Type I errors. The implication of this is not favored because the high loss cost is often associated with Type I errors. A further study suggests if "expensive" Type I errors are assumed, the "in-sample" bankruptcy rate is a recommended cut-off probability. Both our updated corporate bankruptcy data and simulation results suggest this cut-off probability choice provides stable Type I errors. In particular, this conclusion could be justified within the logistic regression model setting. Thus, we argue that, within the logistic regression model context, the choice based sampling technique is sufficient to provide the equivalent predictive ability as the complete data sampling technique if the in-sample bankruptcy rate is applied as the cut-off probability.

Under the circumstance that an explicit probability of undergoing financial distress may be desired, our simulation results recommend applying the complete data sampling technique. The complete data sampling technique presents negligible deviance of the estimated probability from the true probability of bankruptcy. The choice based sampling technique yields a biased probability estimate. In logistic regression, the main coefficient estimation bias lies in the intercept term. Therefore, correcting the intercept could eliminate the probability bias significantly.

On the corporate bankruptcy data set, our study further investigates two non-linear classification methods, Neural Network (NNET) and Support Vector Machines (SVM), to examine the prediction effect of data sample selection methods. NNET and SVM have been widely used in the field of artificial intelligence. But recently, those methods have also gained significant popularity in bankruptcy classification studies. In our limited empirical study, both NNET and SVM present higher number of correct predictions of the bankruptcy records if choice based sampling technique is used. In particular, when the training sample is consisted of equal number of bankrupted companies and the non-bankrupted companies, NNET and SVM outperform logistic regression in classification prediction. However, failing to adjust the cut-off probability would produce devastating predictions. Applying 0.5 as the cut-off probability might possibly result in no successful bankruptcy prediction. Given the superior importance of predictive accuracy in the bankruptcy records, the choice based sampling technique is recommended for both NNET and SVM.

Our paper is organized as follows. Section 3.2 presents our approach to compare different data sample selection methods. Section 3.3 describes the simulation designs used in this study and briefly discusses the simulation results. This is then followed by the real-life bankruptcy database description of Section 3.4 Section 3.5 describes an empirical study analysis. Section 3.6 concludes the paper.

**3.2 Model**

Our main focus of this paper is to study the prediction effect of data sample selection methods. For validation purpose, we adopt the logistic regression model, the most extensively used statistical model in bankruptcy prediction studies, to illustrate and compare the different data sample selection methods. Other models could also be used.

First, let $n$ be the sample size, and $\mathrm{Y} = (\mathrm{Y}_1, \mathrm{Y}_2, \cdots, \mathrm{Y}_n)^{\mathrm{T}}$ be the observed cross-sectional binary outcome vector, indicating a firm's bankruptcy status: $\mathrm{Y}_i = 1,$ if the $\mathrm{i}^{\mathrm{th}}$ company filed bankruptcy during the sample period and $\mathrm{Y}_i = 0,$ otherwise.

The logistic model is:

$$Logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \mathrm{X}_i * \beta \quad \text{for } i = 1,2,\cdots,n, \tag{3.1}$$

where the predictor $\mathrm{X}_i$ is an independent covariate vector for the i[th] observation, $\beta$ is the regression parameter vector, and $p_i$ is the probability of the i[th] firm going bankruptcy. That is,

$$p_i = \mathrm{P}(\mathrm{Y}_i = 1 | \mathrm{X}_i). \tag{3.2}$$

Let $\hat{\mathrm{Y}}_i$ be the predicated bankruptcy status. $\hat{\mathrm{Y}}_i$ is predicted to be 1 if the estimated probability of a firm going bankrupt, $\hat{p}_i$, is greater than a certain threshold, and 0 otherwise. Such a threshold is the so-called *cut-off probability,* denoted by $p_{cut}$. Determining an optimum cut-off probability in the bankruptcy classification prediction study is also important. We will elaborate more on this part in our simulation analysis (Section 3.3).

There has been much research developed for variable selection methods to form a good predictor $\mathrm{X}_i$ in order to achieve improved predictive performance. For illustration purpose, our study applies Return on Assets (ROA), Financial Leverage (FINL) and Liquidity (LIQ) as the covariate $\mathrm{X}_i$, the same set of financial ratios as in Zmijewski (1984). Alternative covariate vectors consisting of cash-flow ratios and turn-over ratios could also be used.

Several criteria are employed to measure the goodness of fit. If a binary bankruptcy classification prediction is of interest, we establish a two by two misclassification matrix, as in Table 3.1. In this misclassification matrix, let entry A report the total number of correct bankruptcy predictions, i.e. when $\hat{\mathrm{Y}}_i = \mathrm{Y}_i = 1$ for $i = 1,2,\cdots,n,$ and D report the total number of correct non-bankruptcy predictions, where $\hat{\mathrm{Y}}_i = \mathrm{Y}_i = 0$. Entry B reports the total number of Type I errors, i.e. when $\hat{\mathrm{Y}}_i = 0$ but $\mathrm{Y}_i = 1$. C reports total number of Type II errors, i.e. $\hat{\mathrm{Y}}_i = 1$ but $\mathrm{Y}_i = 0$. In addition, we report the misclassification rates under two different scenarios. One is for the symmetric misclassification cost case, and the other is for

the asymmetric misclassification cost case. The symmetric misclassification rate (MR) is derived by taking the quotient of the total number of misclassified cases over the total number of observations. Applying our matrix notation, MR can be calculated as

$$\text{MR} = \text{Misclassification Rate} = \frac{\text{B} + \text{C}}{(\text{A} + \text{B} + \text{C} + \text{D})} * 100\% . \tag{3.3}$$

In contrast, the asymmetric misclassification rate, denoted by the weighted misclassification rate (*w*MR), incorporates the asymmetric misclassification cost ratio parameter $\rho$ (Nanda and Pendharkar, 2001) to (3.3). Hence, using matrix notation:

$$wMR = \text{weighted Misclassification Rate} = \frac{\text{B} * \rho + \text{C}}{(\rho + 1)(\text{A} + \text{B} + \text{C} + \text{D})} * 100\%, \tag{3.4}$$

where $\rho$ is the cost ratio of Type I errors over Type II errors. Note that because the cost of making Type I errors and/or Type II errors are intangible and immeasurable (Koh, 1992), and it may also vary within different industries, it is not easy to obtain a good estimate for the value of $\rho$. Nevertheless within our banking decision context, it remains safe to assume that $\rho \geq 1$.

On the other hand, given that the true underlying probability of bankruptcy $p_i$ is known in the simulation study, the model goodness-of-fit can be measured by the 1-norm distance (also known as Manhattan distance) and 2-norm distance (also known as Euclidean distance) between the true probability $p_i$, and the estimated probability $\hat{p}_i$. That is,

$$\text{1-norm distance between } p \text{ and } \hat{p}: \; D_{1,p} = \sum_{i=1}^{n} \frac{1}{n} \left| \hat{p}_i - p_i \right|, \tag{3.5}$$

$$\text{2-norm distance between } p \text{ and } \hat{p}: \; D_{2,p} = \sqrt{\sum_{i=1}^{n} \frac{1}{n} (\hat{p}_i - p_i)^2} . \tag{3.6}$$

Here $p$ and $\hat{p}$ without the subscript "*i*" denote the corresponding probability vectors for the *n* observations, where *n* is the sample size.

The 1-norm distance and 2-norm distance between the true regression parameter vector $\beta$ and the estimated regression parameter $\hat{\beta}$ are also reported, denoting by,

1-norm distance between $\beta$ and $\hat{\beta}$: $D_{1,\beta} = \sum_{j=1}^{k} \frac{1}{k} \left| \hat{\beta}_j - \beta_j \right|$, (3.7)

2-norm distance between $\beta$ and $\hat{\beta}$: $D_{2,\beta} = \sqrt{\sum_{j=1}^{k} \frac{1}{k} \left( \hat{\beta}_j - \beta_j \right)^2}$ . (3.8)

Here k is the number of regression parameters. In our case, k = 4.

**3.3 Simulation Analysis**

To address the data sample selection issues, we conducted a Monte Carlo simulation study. In this simulation study, the multivariate vector, $\mathrm{X}_i$, is designed to follow a multivariate normal distribution, using the notation,

$$\mathrm{X}_i \sim N( \ , \Sigma) \qquad i = 1,2,\cdots,n ,$$

where the vector is obtained by concatenating the respective expectations of the three independent financial ratio variables: ROA, FINL, and LIQ; $\Sigma$ is the corresponding covariance variance matrix, and n is the sample size. In the subsequent simulation analysis, we present the results obtained at a sample size $n$ = 10000. Data sets with different sample sizes have also been simulated and we reached similar conclusions.

To mimic the real data situation, we choose appropriate values for and $\Sigma$. Since our simulation study adopts the unbiased parameter estimator $\beta = (-4.8, -3.599, -5.406, -0.1)$ from Zmijewski (1984) as the regression parameter vector, the values of and $\Sigma$ are therefore assigned as the reported sample statistics for ROA, FINL, and LIQ respectively, in that sample period 1972 to 1978. Also $p_i$, the underlying true probabilities of bankruptcy, is computed by fitting the logistic regression model in (3.1).

In order to relate the binary classification variable $\mathrm{Y}_i$ with its underlying probability, $p_i$, a Bernoulli distribution is applied to generate the variable $\mathrm{Y}_i$ and represented by $\mathrm{Y}_i \sim Bernoulli(p_i)$.

The above designed procedure produces the entire simulation data set. All of the

following validation steps are then applied to this data set. In particular, we start by partitioning this simulation data set into two subsamples, an "estimate data subsample" (training data sample population), and a "prediction data subsample" (testing data sample). With different data sampling techniques, training data samples are selected from this training data sample population. The testing data sample is used to assess the goodness of fit metrics across all the training data samples. Such out-of-sample testing design not only provides a good benchmark for model comparison and validation, it also effectively eliminates the "in-sample over-fitting" effect (Clark, 2004). As stated previously, consistent simulation conclusions could be reached by running the Monte Carlo simulation for 100 replications. For illustration purposes, one simulation with training data sample population consisted of 403 "1"s and 4598 "0"s and testing data sample with 402 "1"s and 4597 "0"s is presented.

As a subsequent step, three choice based data samples are selected using the choice based sampling technique. Every choice based training data sample includes all 403 "1"s records in training data sample population. However, the number of randomly selected "0"s varies from 403, 806 to 1612. As a result, the in-sample bankruptcy rates are 0.5, 0.3333 and 0.2 respectively. One more data sample is selected from the training data sample population using the complete data sampling technique. Such selection method creates a complete replication of the original training data sample population. Therefore, a total of four different training data samples, three from the choice based sampling technique and one from the complete data sampling technique, are available to be modeled and compared by their goodness of fit measures evaluated on the same testing data sample. Here we present the results with the cost ratio parameter $\rho = 35$ as suggested in Altman *et al.* (1977) to derive the weighted misclassification rate (wMR) as in equation (3.4) for the purpose of illustration. Other choices of the cost ratio capture similar trends.

Table 3.2 reports the resulting misclassification matrices for the model developed from the four training data samples. Clearly, the upper block of Table 3.2 demonstrates that as the composition of the training data sample approaches the training data sample population, the total misclassification rate decreases if applying 0.5 as the cut-off probability. Such fact

concurs with Zmijewski (1984) findings. However, it is also worth noting that the weighted misclassification rate displays an increasing trend.

Next, let us focus on the case of the asymmetric cost: in extreme, assume that the Type I error cost is infinity while the Type II error cost is 0. Then, the comparison on the weighted misclassification rate reduces to comparing the reported entry B for each training data sample. It is then obvious that the number of Type I errors is almost 5 times larger across these four training data samples. Such significant increase implies that it is not appropriate to apply a fixed cut-off probability across different training data samples, given that the cost ratio is asymmetric. Therefore, as the data sample approaches the complete sample, we conclude that using a fixed cut-off probability would incur increased Type I errors, but decreased total misclassification errors.

The lower part of Table 3.2 describe the resulting misclassification matrices if applying a cut-off probability equal to the corresponding in-sample bankruptcy rate in the "choice based training data sample". Such in-sample bankruptcy rate is computed by dividing the number of bankrupted companies by the total number of companies in the estimate data sample. It is shown that every reported number in entry A through D does not vary much. As a direct result, a series of stable misclassification rates are captured for both the unweighted and weighted misclassification rates. Such results confirm the statement that under the logistic regression model, the choice based sampling technique produces a predictive classification ability equivalent to the complete data sampling technique when the cut-off probability applied on the hold-out testing sample is set to the respective training sample's bankruptcy rate.

Furthermore, it is also shown in Table 3.2 that with one training data sample, choice of cut-off probabilities has a significant effect on the weighted misclassification rates. This is due to the dramatic change in the reported number of Type I errors, which dominates the weighted misclassification rate computation. Therefore, in different choice-based data samples, an appropriate choice of cut-off probability depends on the specification of the cost ratio. Also, if more attention is needed for Type I errors, the choice based sampling technique

with a cut-off probability equal to the training sample's in-sample bankruptcy rate may be recommended.

The first two rows of Table 3.3 report the resulting 1-norm distance $D_{1,\mathrm{p}}$ and 2-norm distance $D_{2,\mathrm{p}}$ between the true probability, $p$, and the estimated probability of bankruptcy, $\hat{p}$. One can see that the probability difference decreases significantly in the first three samples, and becomes negligible in the complete data sample. Figure 3.1 depicts the graphical comparisons between the true probability and the predicted probability of bankruptcy on the same predication data sample for the four simulation samples. An up-lifting effect is observed in the first plot, which is obtained from the training sample with equal number of bankruptcy records and non-bankruptcy records. But the trend fades out gradually, and a straight line is plotted for the complete data sample. This is an evidence of the fact that oversampling the bankrupted records would produce an inevitable biased probability estimate effect. Based on these simulation facts, we propose to apply complete data sampling method to reduce the deviance of the predicted probability from the true probability. Table 3.3 and Figure 3.1 yield consistent results in this sense. Therefore, we conclude that if the probability of bankruptcy is of interest, the complete data sample selection is a recommended sampling technique.

The 1-norm distance $D_{1,\beta}$ and 2-norm distance $D_{2,\beta}$ between $\beta$ and $\hat{\beta}$ are reported in the last two rows of Table 3.3. Similar to the previous comments, the difference between the true regression parameter and the estimated regression parameter is smallest in the complete data sample.

We also tried Monte Carlo simulation with 100 replications, and get similar results. Thus, we only report the above results in the interest of parsimony. We see that for binary bankruptcy classification predictions, the choice based sample technique may still be appropriate and displays satisfactory misclassification rates. Especially in logistic regression model setting, it is verified that with some moderate cut-off probability justification, choice based sampling technique yields the same predictive effect as the complete data sampling

technique. However, necessary attention should be paid to determine a suitable cut-off probability based on different specifications of the cost ratio "$\rho$". Regarding the assignment of a higher cost associated with Type I error, our simulation study demonstrates that applying the cut-off probability equal to the training data sample's in-sample bankruptcy rate presents smaller weighted misclassification rates and thus outperforms the fixed cut-off probability. Alternatively, if a probability of bankruptcy is desired, the complete sampling technique is recommended. The complete sampling technique provides the negligible difference between the estimated probability and the true probability of bankruptcy, whereas, the choice-based sampling technique would result in a biased probability estimate.

**3.4 Data**

In section 3.3, we mainly discuss the data sample selection issues from the perspective of simulation. In the subsequent empirical analysis, we employ a similar procedure and further verify the proposed recommendations.

As part of the empirical study, we construct a bankruptcy database by including all publicly traded companies available in the "Compustat North America" with Standard Industrial Classification (SIC) code less than 6000. We further screened the corporate data sample by excluding firms other than those traded on either "NYSE", "AMEX", or "NASDAQ" stock exchange during 1972 to 2000. The "Compustat" database contains the corporate financial data until 2009. However, after 2000, there is a notable dropping in the number of the reported bankrupted firms. Only a few bankruptcy filings were recorded after that. Such trend is not in line with the reports from other major research works. Hence, we limit our analysis only up to the year 2000.

In our study, a firm is defined as bankrupted if it makes either a Chapter 7 or Chapter 11 filing between January 1972 and December 2000. Applying all these filter conditions resultes in a database consisted of a total of 172 bankrupted companies and 7218 non-bankrupted companies. This is a more up-to-date database compared to Zmijewski (1984)'s work, which was built on a time horizon from 1972 to 1978.

For our empirical study, we apply different data sampling techniques on a cross sectional data set. For the bankrupted companies, only the most recently reported financial record, prior to the bankruptcy filing, will be collected. for the non-bankrupted company, we randomly select a fiscal year from the sample period. Only the selected year's annual accounting data entry for a non-bankrupted company will be used for the following empirical analysis.

Figure 3.2 reports the total number of bankruptcies over the sample period. It is shown that the number of bankruptcies increases over the entire period, and rise dramatically in the middle of 1980s and early 1990s.

Then, a partition strategy, similar to the one we used on our simulation data set is applied to this data set. It results in an "estimate data subsample" of 86 bankrupted companies and 3609 non-bankrupted, and a "prediction data subsample" of 86 bankrupted companies and 3609 non-bankrupted companies. Again, for model comparison and validation purpose, the prediction data subsample is designed to capture the out-of-sample predictive performance across different data sampling techniques.

**3.5 Empirical Application**

To investigate the recommendations derived from the Monte Carlo simulation study on this corporate bankruptcy data set, we apply the same validation procedure to the training data sample population, and report our goodness-of-fit measures on the testing data sample. In particular, three training data samples are selected using the choice based sampling technique. Every choice based data sample includes all of the 86 bankruptcy records, as defined. But we manipulate the composition ratio "r", for each estimate data sample, by randomly selecting different numbers of non-bankruptcy records. And "r" is the ratio of the number of bankruptcies over the number of non-bankruptcies. As a result, the three training data sample contains 86 (i.e. $r = 1:1$), 258 (i.e. $r = 1:3$) to 860 (i.e. $r = 1:10$) non-bankrupted companies with the in-sample bankruptcy rates 0.5, 0.25 and 0.0909 respectively. Furthermore, the complete data sampling technique creates one more data sample of the 86 bankrupted companies, and 3609 non-bankrupted companies. Therefore, we

report the misclassification matrices and the two types of misclassification rates for these four training data samples as the goodness of fit in the regime of the binary bankruptcy forecasting study.

In addition to the application of the logistic regression validation model, we further investigate two popular classification methods, neural network (NNET) technique and Support Vector Machines (SVM) approach, to study the prediction effect of the different data sample selection methods on this corporate bankruptcy data set. The misclassification matrices and the two types of misclassification rates are reported.

Note that the power of this empirical study is limited to provide the probability deviance comparisons for the estimated probabilities of bankruptcy from the true probability, as reported in the simulation analysis. This is because the validation of such comparison is violated due to the unavailability of the true probability of bankruptcy. Again, in order to provide a convincing benchmark conclusion, the same hold-out test sample is used across models for validation comparisons.

### 3.5.1 Logistic Regression

Table 3.4 reports the empirical results of using the logistic regression model for our bankruptcy data set. It is shown that using 0.5 as the cut-off probability, the total misclassification rate has decreased significantly, whereas the weighted misclassification rate increases as the sample bankruptcy rate approaches the population bankruptcy rate. However, if the cut-off probability is chosen to be the in-sample bankruptcy rate, similar "stable misclassification rate and weighted misclassification rates" phenomenon is observed here as in the simulation data set.

Figure 3.3 compares the weighted misclassification rates associated with different cost ratios, "$\rho$", for different data sampling techniques. To obtain this plot, we calculate the weighted misclassification rates by iterating the cost ratio "$\rho$", from 1 to 200, for three samples: the first two choice based data samples and the last one, the complete data sample. The black dot and the red diamond depict the choice based sampling technique for the data sample with "$r$" as 1:1 and 1:3. The complete data sample is plotted with blue triangle

elements. Each resulting misclassification matrix is computed by applying the in-sample bankruptcy rate as the cut-off probability. Note that the weighted misclassification rates generated by the three samples are very close to each other. Also, the three lines converge to one thicker line when $\rho$ grows bigger. This plot visually verifies the stable misclassification entries reported in Table 3.4 when the cut-off probability is set to the bankruptcy rate in sample.

Therefore, we reach consistent conclusion for our empirical study and the simulation study. With the logistic regression model, choice based sampling technique and the complete data sampling technique yield similar predictive effect if the cut-off probability is set to the bankruptcy rate in the estimate sample. In an asymmetric cost case, the strategy of choice based sampling technique with a cut-off probability equal to the in-sample bankruptcy rate is recommended.

### 3.5.2 Neural Networks

Both the Neural Network (NNET) and SVM are important non-linear classification methods with much attention in recent development. In this section, we further examine the data sample selection issues for bankruptcy prediction with the neural network approach through an empirical study.

Neural network is a nonparametric learning system. It is constructed with inter-connected processing units, organized in layers. In general, there are three layers in the neural network: input layer, output layer, and hidden layer. Pioneer works, such as Odom and Sharda (1990), Tam and Kiang (1992), and Zhang *et al*. (1999), have shown the superior bankruptcy prediction results achieved by the neural network approach. Particularly, in the case of severe loss incurred by the Type I errors, neural network is suggested as a "more robust approach" (Latcher *et al*., 1995). With the virtues stated above, we conduct the empirical work with two simple neural network configurations: one without hidden layers and the other has a 5-unit hidden layer. The results of the corresponding classification matrices are presented in the upper panel and lower panel, respectively, of Table 3.5.

We observe the followings. First, with one training data sample, applying the bankruptcy rate in sample as the cut-off probability always yields higher misclassification rate but lower weighted misclassification rate than the case of using a fixed 0.5 as the cut-off probability. In particular, if applying in sample bankruptcy rate as the cut-off probability, less bankruptcy predictions are correctly captured, when the data sample approaches to the population composition. Second, for both Two-layer NNET and three-layer NNET, choice based sampling technique managed to provide higher rate of correct prediction of bankruptcy records than the complete data sampling technique. In contrast, the predictive performance hit the bottom if complete data sampling technique is adopted combining with the cut-off probability set to 0.5. In our case, for the two-layer neural network, the number of correct bankruptcy predictions decreased to zero drastically from 54 by changing from choice based sampling technique to complete data sampling technique using 0.5 as the cut-off probability. And in the three-layer neural network configuration, the number of the correct bankruptcy predictions displays the same dropping trend, but in a slower rate. This is possibly due to the "rare event" phenomenon of the bankruptcy data in nature. Third, it is also worth noting that both the misclassification rates and the weighted misclassification rates reported in the lower part are much lower than those in the upper part. This is directly due to the additional layer, which is designed to generate a better learning model.

With the neural network approach, choice based sampling technique demonstrates higher accuracy in capturing the bankruptcy records and setting cut-off probability as the bankruptcy rate in sample provides relatively stably lower weighted misclassification rate. Thus, for this empirical study, if "expensive" Type I errors are assumed, choice based sampling technique with cut-off probability equal to the bankruptcy rate in sample is recommended.

### 3.5.3 Support Vector Machines

In addition, we examine the data sample selection issues using the Support Vector Machines (SVM) for this empirical data. The results of misclassification matrices using the SVM approach are shown in Table 3.6. Support Vector Machine, developed by Vapnik

(2000), is a non-linear classification method developed from statistical learning theory. It has been widely applied in the expert system, text recognition, just to name a few. The appealing properties of SVM are its transparency and accuracy in the predictive study. Thus, under the Basel II framework, which greatly promotes the model's accuracy, SVM becomes an attractive candidate for the corporate bankruptcy prediction study. Many researchers including Härdle *et al.* (2005) and Kim and Sohn (2010) have reported the significant improvement in bankruptcy prediction accuracy using SVM technique.

Table 3.6 demonstrates the classification ability of SVM controlled at different level of capacity C -- the regularization term in the Lagrange formulation. The capacity C is related to the generalization ability of the SVM. In Table 3.6, the classification results with capacity C = 1 is presented in the upper panel, and the capacity reported in the lower panel is set to 100.

Similar trends shown in the neural network setting are also observed here. For instance, applying the bankruptcy rate in sample as the cut-off probability outperforms the fixed cut-off probability in the weighted misclassification rate, but underperforms it in the symmetric cost case.

Besides, we also notice that, especially in the complete data sample when C = 1, SVM totally missed the bankruptcy records when a fixed 0.5 cut off probability is used. Such "zero entry" for the correct count of bankruptcy is the least desirable. The very low frequency rate of the bankruptcy events of the data in nature might be the reason for such outcome. And one possible remedy is to adjust the cut-off probability, as illustrated here, from 0.5 to the bankruptcy rate in sample. Such a cut-off probability adjustment increases the number of correct bankruptcy prediction by 10 when C = 1, and by 2 when C = 100.

In this empirical study, SVM also managed to provide higher bankruptcy classification accuracy than the logistic regression in the data sample of equal number of the bankruptcy records and the non-bankruptcy records. In particular, SVM with C = 1 provides 65 correct "1" predictions, whereas the logistic provides only 49 correct bankruptcy predictions. Given the above limited empirical study, complete data sampling technique may not be recommended under the SVM model setting. Choice based sampling technique with 0.5 cut

off probability is capable of providing superior prediction results. Thus in SVM, if "expensive" Type I errors are the major concerns, choice based sampling technique is recommended.

In summary, the empirical results obtained by fitting the logistic regression model are consistent with the simulation results. Hence, under logistic regression setting, we recommend special care to be devoted in choosing a suitable cut-off probability for the binary bankruptcy classification problem. The choice for the cut-off probability depends on the cost ratios between the Type I errors and Type II errors. In a symmetric cost case, complete data sampling technique is always recommend, whereas in an asymmetric cost case, choice based sampling technique with cut-off probability equal to the in sample bankruptcy rate would provide lower weighted misclassification rate. On the other hand, extrapolating from the simulation results, if a precise probability of bankruptcy is of interest, the complete data sampling technique is recommended.

NNET and SVM have also been able to produce accurate classification results. In the asymmetric cost case with higher loss associated with Type I error, NNET and SVM outperform the logistic regression by presenting more correct bankruptcy predictions. The suggested strategy is to choose equal number of records from both groups: bankruptcy and non-bankruptcy, and the cut-off probability is set to 0.5.

To conclude for this empirical study, for both NNET model and SVM setting, if Type I errors are assumed to incur a significant asymmetric cost, choice based sampling technique with 0.5 as the cut-off probability is recommended.

**3.6 Conclusion**

Our study developed a framework to investigate the data sample selection issues for bankruptcy prediction studies and provided guidelines for data sample selection process for both academics and practitioners. With the presence of different data sample selection methods, we conclude that the method used to select the data sample depends on the objective of a given study. In the context of binary bankruptcy prediction studies, the choice based sampling technique is a suitable and easy-to-implement data selection method. However, if one takes into account the different misclassification costs incurred by the two

different types of misclassifications, it is essential to make an appropriate choice of cut-off probability. It is shown in our study that if within the decision role similar to a bank lender, i. e. misclassifying a bankruptcy as a non-bankruptcy incurs a much higher cost than misclassifying a non-bankrupted firm as a bankruptcy, applying a fixed or predefined cut-off probability to classify bankruptcy across different data samples implies increased unfavorable costly errors, whereas, selecting a cut-off probability equivalent to the estimate data sample bankruptcy rate leads to a considerably more stable out-of-sample predictive ability. On the other hand, if a precise probability of bankruptcy is of interest, the complete data sampling technique is recommended because estimation obtained via this sampling technique results in negligible deviance from the underlying true patterns.

**Table 3.1: Misclassification Matrix**

*Entry A reports the total number of correct bankruptcy predictions, and entry D reports the total number of correct non-bankruptcy predictions. Entry B reports the total number of misclassifying a bankruptcy as a non-bankruptcy while entry C reports the total number of misclassifying a non-bankruptcy as a bankruptcy. The last two appended rows report the total misclassification rate (MR), and weighted total misclassification rate (wMR).*

| | $Y_i = 1$ | $Y_i = 0$ |
|---|---|---|
| $\hat{Y}_i = 1$ | A | C |
| $\hat{Y}_i = 0$ | B | D |
| *Misclassification Rate (MR)* | | $\frac{B + C}{A + B + C + D}$ |
| *weighted Misclassification Rate (wMR)* | | $\frac{B * \rho + C}{(\rho + 1)(A + B + C + D)}$ |

**Table 3.2: Simulated data sample results for bankruptcy classification study**

*Table 3.2 reports the resulting misclassification matrices of the three different choice based data samples and one complete data sample for a binary classification bankruptcy prediction study along with the total misclassification rate. "r" is the composition ratio of the estimate data sample, reported as the number of bankruptcies over the number of non-bankruptcies in the training data sample. Two types of cut-off probability are used, one is fixed at 0.5 and the other is varied according to the bankruptcy rate in each estimation sample.*

| | **Choice Based r = 1:1** | | **Choice Based r = 1:2** | | **Choice Based r = 1:4** | | **Complete Sampling** | |
|---|---|---|---|---|---|---|---|---|
| $p_{cut} = 0.5$ | 356 | 570 | 330 | 370 | 290 | 201 | 200 | 61 |
| | 46 | 4027 | 72 | 4227 | 112 | 4398 | 201 | 4536 |
| | *M.R.* | *12.32%* | *M.R.* | *8.84%* | *M.R.* | *6.26%* | *M.R.* | *5.26%* |
| | *w.M.R.* | *1.21%* | *w.M.R.* | *1.61%* | *w.M.R.* | *2.29%* | *w.M.R.* | *3.94%* |
| $p_{cut}$ = Bankruptcyrate in sample* | 356 | 570 | 358 | 595 | 363 | 610 | 368 | 629 |
| | 46 | 4027 | 44 | 4002 | 39 | 3978 | 34 | 3968 |
| | *M.R.* | *12.32%* | *M.R.* | *12.78%* | *M.R.* | *13.50%* | *M.R.* | *13.27%* |
| | *w. M.R* | *1.21%* | *w. M.R* | *1.19%* | *w. M.R* | *1.10%* | *w. M.R* | *1.01%* |

**Bankruptcy rate is calculated as dividing the total number of records in sample by the number of the bankruptcy records.*

**Table 3.3: Probability distance result from simulated data samples**

*The first two rows of Table 3.3 report the 1-norm and 2-norm distance, calculated as in (3.5) and (3.6) respectively, between the true probability and the estimated probability of bankruptcy on the "predicted data subsample" across these four data samples. The third and fourth row of Table 4 report 1-norm and 2-norm distance, calculated as in (3.7) and (3.8) respectively, between the true regression parameter and the estimated regression parameter on the "predicted data subsample" across these four data samples.*

| | **Choice Based r = 1:1** | **Choice Based r = 1:2** | **Choice Based r = 1:4** | **Complete Sampling** |
|---|---|---|---|---|
| $D_{1,p}$ ***(*0.01)*** | 12.447 | 8.323 | 4.595 | 0.409 |
| $D_{2,p}$ ***(*0.001)*** | 48.501 | 23.490 | 8.306 | 0.098 |
| $D_{1,\beta}$ | 3.209 | 3.104 | 2.867 | 2.846 |
| $D_{2,\beta}$ | 36.246 | 32.440 | 31.013 | 30.216 |

**Table 3.4: Empirical data sample analysis using logistic regression**

*Table 3.4 reports the resulting misclassification matrices along with the total misclassification rate of these four data samples for a binary classification bankruptcy prediction study. "r" is the composition ratio of the estimate data sample, reported as the number of bankruptcies over the number of non-bankruptcies in the training data sample. Two types of cut-off probability are used, one is fixed at 0.5 and the other is varied according to the bankruptcy rate in each estimation sample.*

| | **Choice Based r = 1:1** | | **Choice Based r = 1:3** | | **Choice Based r = 1:10** | | **Complete Sampling** | |
|---|---|---|---|---|---|---|---|---|
| $p_{cut} = 0.5$ | 49 | 1579 | 2 | 40 | 1 | 16 | 0 | 2 |
| | 37 | 2030 | 84 | 3569 | 85 | 3593 | 86 | 3607 |
| | *M.R.* | *43.73%* | *M.R.* | *3.36%* | *M.R.* | *2.73%* | *M.R.* | *2.38%* |
| | *w. M.R* | *2.16%* | *w. M.R* | *2.24%* | *w. M.R* | *2.25%* | *w. M.R* | *2.26%* |
| $p_{cut}$ = Bankruptcyrate in sample* | 49 | 1579 | 51 | 1590 | 47 | 1477 | 51 | 1665 |
| | 37 | 2030 | 35 | 2019 | 39 | 2131 | 35 | 1994 |
| | *M.R.* | *43.73%* | *M.R.* | *43.98%* | *M.R.* | *41.03%* | *M.R.* | *46.01%* |
| | *w. M.R* | *2.16%* | *w. M.R* | *2.12%* | *w. M.R* | *2.14%* | *w. M.R* | *2.17%* |

**Bankruptcy rate is calculated as dividing the total number of records in sample by the number of the bankruptcy records.*

**Table 3.5: Empirical data sample analysis using Neural Network**

*Table 3.5 reports the resulted misclassification matrices along with the total misclassification rate of these four data samples. The model used for the upper part is the Neural Network without hidden layer, i.e. a two-layer structural Neural Network, and the lower part is reported for a Neural Network with a 5-units hidden layer. In each case, two types of cut-off probability are further used, one is fixed at 0.5 and the other is varied according to the bankruptcy rate in each estimation sample*

| ***NNET*** | | **Choice Based r = 1:1** | | **Choice Based r = 1:3** | | **Choice Based r = 1:10** | | **Complete sampling** | |
|---|---|---|---|---|---|---|---|---|---|
| ***Two-Layer*** | $p_{cut} = 0.5$ | 54 | 1548 | 1 | 17 | 0 | 2 | 0 | 3 |
| | | 32 | 2061 | 85 | 3592 | 86 | 3607 | 86 | 3606 |
| | | *M.R.* | *42.76%* | *M.R.* | *2.76%* | *M.R.* | *2.38%* | *M.R.* | *2.41%* |
| | | *w. M.R* | *2.00%* | *w. M.R* | *2.25%* | *w. M.R* | *2.26%* | *w. M.R* | *2.27%* |
| | $p_{cut}$ = Bankruptcy rate in sample* | 54 | 1548 | 64 | 1943 | 61 | 1849 | 26 | 1023 |
| | | 32 | 2061 | 22 | 1666 | 25 | 1760 | 60 | 2586 |
| | | *M.R.* | *42.76%* | *M.R.* | *53.18%* | *M.R.* | *50.72%* | *M.R.* | *29.31%* |
| | | *w. M.R* | *2.00%* | *w. M.R* | *2.04%* | *w. M.R* | *2.05%* | *w. M.R* | *2.35%* |
| ***Three-Layer*** | $p_{cut} = 0.5$ | 52 | 874 | 35 | 362 | 13 | 44 | 9 | 0 |
| | | 34 | 2735 | 51 | 3247 | 73 | 3565 | 77 | 3609 |
| | | *M.R.* | *24.57%* | *M.R.* | *11.18%* | *M.R.* | *3.17%* | *M.R.* | *2.08%* |
| | | *w. M.R* | *1.55%* | *w. M.R* | *1.61%* | *w. M.R* | *1.95%* | *w. M.R* | *2.03%* |
| | $p_{cut}$ = Bankruptcy rate in sample* | 52 | 874 | 50 | 808 | 40 | 649 | 47 | 533 |
| | | 34 | 2735 | 36 | 2801 | 46 | 2960 | 39 | 3076 |
| | | *M.R.* | *24.57%* | *M.R.* | *22.84%* | *M.R.* | *18.81%* | *M.R.* | *15.48%* |
| | | *w. M.R* | *1.55%* | *w. M.R* | *1.55%* | *w. M.R* | *1.70%* | *w. M.R* | *1.43%* |

**Bankruptcy rate is calculated as dividing the total number of records in sample by the number of the bankruptcy records.*

**Table 3.6: Empirical data sample analysis with SVM**

*Table 3.6 reports the resulted misclassification matrices along with the total misclassification rate of these four data samples. The model used is the Support Vector Machine with "Radial basis" kernel type, and the capacity parameter is realized as C = 1 and C = 100 respectively. In each case, two types of cut-off probability are further used, one is fixed at 0.5 and the other is varied according to the bankruptcy rate in each estimation sample.*

| ***SVM*** | | **Choice Based r = 1:1** | | **Choice Based r = 1:3** | | **Choice Based r = 1:10** | | **Complete sampling** | |
|---|---|---|---|---|---|---|---|---|---|
| | $p_{cut} = 0.5$ | 65 | 1408 | 18 | 126 | 0 | 1 | 0 | 0 |
| | | 21 | 2201 | 68 | 3483 | 86 | 3608 | 86 | 3609 |
| | | *M.R.* | *38.6%* | *M.R.* | *5.25%* | *M.R.* | *2.35%* | *M.R.* | *2.33%* |
| *C=1* | | *w.M.R.* | *1.61%* | *w.M.R.* | *1.88%* | *w.M.R.* | *2.26%* | *w.M.R.* | *2.26%* |
| | $p_{cut}$ = Bankruptcy rate in sample* | 65 | 1408 | 27 | 264 | 14 | 18 | 10 | 4 |
| | | 21 | 2201 | 59 | 3345 | 72 | 3591 | 76 | 3605 |
| | | *M.R.* | *38.6%* | *M.R.* | *8.74%* | *M.R.* | *2.71%* | *M.R.* | *2.17%* |
| | | *w.M.R.* | *1.61%* | *w.M.R.* | *1.75%* | *w.M.R.* | *1.91%* | *w.M.R.* | *2.00%* |
| | $p_{cut} = 0.5$ | 55 | 853 | 36 | 245 | 14 | 15 | 10 | 3 |
| | | 31 | 2756 | 50 | 3364 | 72 | 3594 | 76 | 3606 |
| | | *M.R.* | *23.9%* | *M.R.* | *7.98%* | *M.R.* | *2.35%* | *M.R.* | *2.14%* |
| ***C=10*** | | *w.M.R.* | *1.46%* | *w.M.R.* | *1.50%* | *w.M.R.* | *1.91%* | *w.M.R.* | *2.00%* |
| ***0*** | $p_{cut}$ = Bankruptcy rate in sample* | 55 | 853 | 45 | 418 | 17 | 57 | 12 | 42 |
| | | 31 | 2756 | 41 | 3191 | 69 | 3552 | 74 | 3567 |
| | | *M.R.* | *23.9%* | *M.R.* | *12.4%* | *M.R.* | *3.41%* | *M.R.* | *3.14%* |
| | | *w.M.R.* | *1.46%* | *w.M.R.* | *1.39%* | *w.M.R.* | *1.86%* | *w.M.R.* | *1.98%* |

**Bankruptcy rate is calculated as dividing the total number of records in sample by the number of the bankruptcy records.*

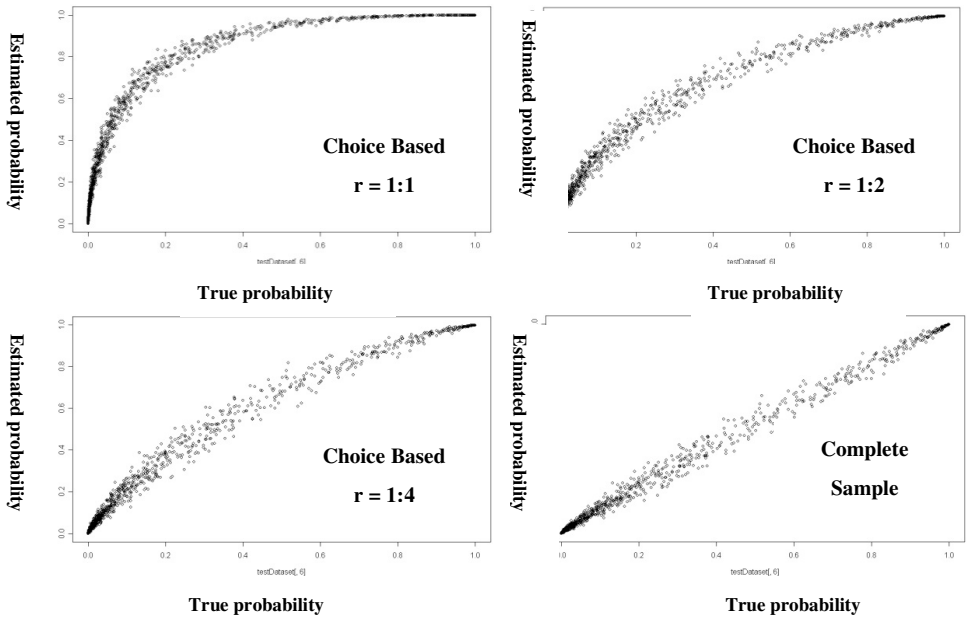



***Figure 3.1****: Comparisons between the true probability and the predicted probability of bankruptcy for the simulation samples.*

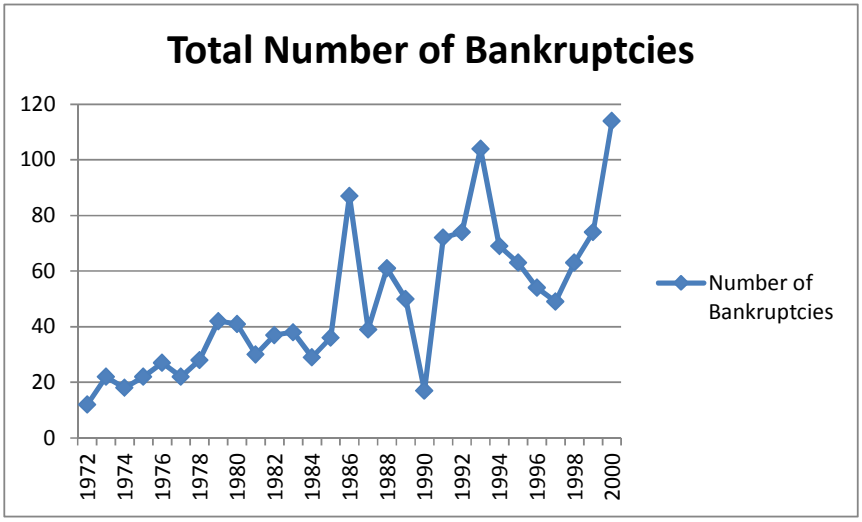



***Figure 3.2**: total number of bankruptcies by year.*

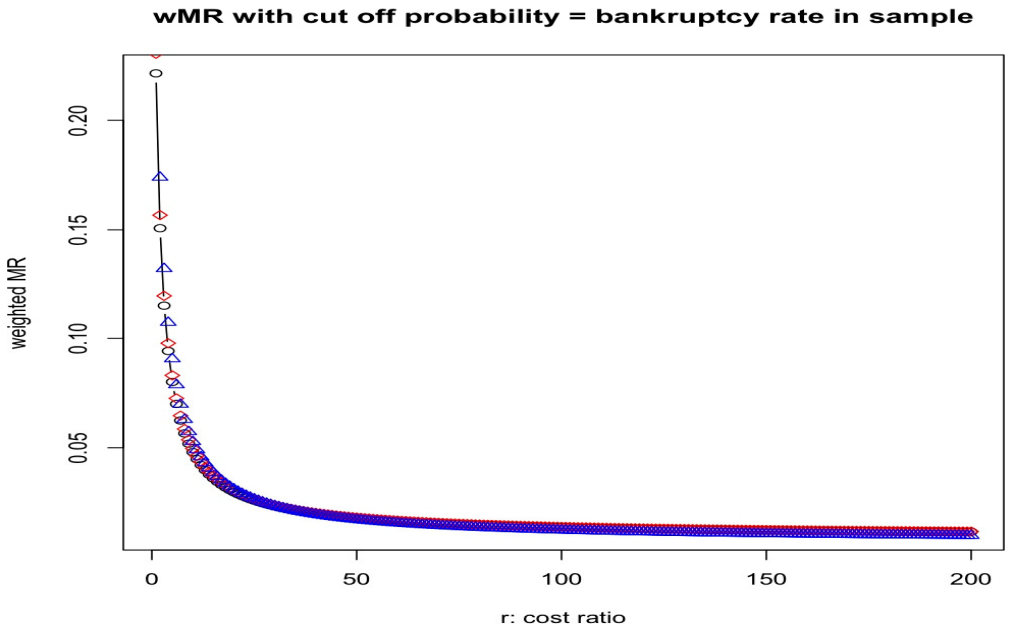



***Figure 3.3****: In logistic regression model, this figure displays the weighted misclassification rate with different cost ratio for* $\rho$ *= 1, 2, …, 200 using the in-sample bankruptcy rate as the cut-off probability. Black dot represents results for the choice based sampling technique with r = 1:1, Red diamond corresponding to the r = 1:3 while the blue rectangle for the complete data sampling technique, where "r" is the composition ratio of the estimate data sample, reported as the number of bankruptcies over the number of non-bankruptcies in the training data sample. Three curves virtually overlay each other.*

**Bibliography**


Akaike, H. (1974), "A New Look at the Statistical Model Identification", *IEEE Transactions, Automatic Control*, 19(6), 716-723.

Altman, E. I. (1968), Financial Ratios, Discriminant Analysis and Prediction of Corporate Bankruptcy, *Journal of Finance*, 23, 589-610.

Altman, E. I. (1993), *Corporate Financial Distress and Bankruptcy: A Complete Guide to Predicting and Avoiding Distress and Profiting from Bankruptcy*, New York: Wiley.

Altman, E. I., Haldeman, R. G., and Narayanan P. (1977), ZETA Analysis: a New Model to Identify Bankruptcy Risk of Corporations, *Journal of Banking and Finance*, 1, 29–54.

Beaver, W. H., (1966), Financial Ratios as Predictors of Failures, *Journal of Accounting Research* (Supplement 1966), 71-102.

Beaver, W. H., McNichols, M. F. and 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.

Bennett, S. (1983), Analysis of Survival Data by the Proportional Odds Model, *Statistics in Medicine*, 2, 273-277.

Bernanke, B., Lown, C., and Friedman, B. (1991), the Credit Crunch, *Brookings Papers on Economic Activity*, 205-247.

Bharath, S., and Shumway, T. (2008), Forecasting Default with the Merton Distance to Default Model, *Review of Financial Studies*, 21(3), 1339-1369.

Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garotte," *Technometrics*, 37, 373-384.

Breslow, N. E. (1974), Covariance Analysis of Censored Survival Data, *Biometrics*, 30, 89-99.

Campbell, J. Y., Hilscher, J., and Szilagyi, J. (2008), In Search of Distress Risk, *Journal of Finance*, LXIII, 6, 2899-2939.

Chava, S., and Jarrow, R. A. (2004), Bankruptcy Prediction with Industry Effects, *Review of Finance*, 8, 537-569.

Chen, K., Jin, Z., and Ying, Z. (2002), Semiparametric Analysis of Transformation Models with Censored Data, *Biometrika*, 89, 659-668.

Cheng, S. C., Wei, L. J., and Ying, Z. (1995), Analysis of Transformation Models with Censored Data, *Biometrika*, 82, 835-845.

Cheng, S. C., Wei, L. J., and Ying, Z. (1997), Prediction of Survival Probabilities with Semi-parametric Transformation Models, *Journal of the American Statistical Association*, 92, 227-235.

Clark, T. E., (2004), Can Out-of-Sample Forecast Comparisons Help Prevent Over Fitting? *Journal of Forecasting*, 23(2), 115 – 139.

Clayton, D., and Cuzick, J. (1986), The Semi-parametric Pareto Model for Regression Analysis of Survival Times, *Proceedings of International Statistical Institute, Amsterdam.*

Cook, N.R. (2008), Statistical Evaluation of Prognostic versus Diagnostic Models: Beyond the ROC Curve, *Clinical Chemistry*, 54, 17-23.

Cox, D. R. (1972), Regression Models and Life-tables, *Journal of the Royal Statistical Society, Series B*, 34, 187-220.

Cox, D. R. (1975), Partial likelihood, *Biometrika*, 62, 269-276.

Crouhy, M., Galai, D., and Mark, R., (2000), A Comparative Analysis of Current Credit Risk Models, *Journal of Banking & Finance*, 24, 59-117.

Dabrowska, D. M., and Doskum, K. A. (1988a), Estimation and Testing in a Two-Sample Generalized Odds-rate Model, *Journal of the American Statistical Association*, 83, 744-749.

Dabrowska, D. M., and Doskum, K. A. (1988b), Partial Likelihood in Transformation Models with Censored Data, *Scandinavian Journal of Statistics*, 15, 1-23.

Daniel, K., and Titman, S. (2006), Market Reactions to Tangible and Intangible Information, *Journal of Finance*, LXI, 4, 1605-1643.

Dimitras, A. I., Zanakis, S. H., and Zopounidis, C., (1996), A Survey of Business Failures with an Emphasis on Prediction Methods and Industrial Applications, *European Journal of Operational Research*, 90, 487-513.

Ding, A. A., Tian, S., Yu, Y., and Guo, H. (2012), "A Class of Discrete Transformation Survival Models with Application to Default Probability Prediction", *Journal of the American Statistical Association*, to appear.

Doksum, K. A., and Gasko, M. (1990), On a Correspondence between Models in Binary Regression Analysis and in Survival Analysis, *International Statistical Review*, 58, 3, pp. 243-252.

Duffie, D., Saita, L., and Wang, K. (2007), Multi-period Corporate Default Prediction with Stochastic Covariates, *Journal of Financial Economics*, 83, 635-665.

Duffie, D., Eckner, A., Horel, G., and Saita, L. (2009), Frailty Correlated Default, *Journal of Finance*, LXIV, 5, 2089-2123.

Dwyer, D. W., Kocagil, A. E., and Stein, R. M. (2004), "MOODY'S KMV RISKCALC v3.1 MODEL", Moody's KMV.

Efron, B. (1977), The Efficiency of Cox's Likelihood Function for Censored Data, *Journal of the American Statistical Association*, 72, 557-565.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression", *Annals of Statistics*, 32, 407-499.

Fama, E. F., and French, K. R. (1996), Multifactor Explanations of Asset Pricing Anomalies, *Journal of Finance*, 51, 55-84.

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348-1360.

Fawcett, T. (2006), "An introduction to ROC analysis," *Pattern Recognition Letters*, 27, 861-874.

George, T. J., and Hwang, C. (2010), "A resolution of the distress risk and leverage puzzles in the cross section of stock returns," *Journal of Financial Economics*, 96, 56-79.

Gerds, T.A. and Schumacher, M. (2006), Consistent Estimation of the Expected Brier Score in General Survival Models with Right-censored Event Times, *Biometrical Journal*, 48, 1029-1040.

Hamilton, D. T., Sun, Z., and Ding, M. (2011), "Through-the-Cycle EDF Credit Measures", Moody's KMV.

Härdle, W., Lee, Y. J., Schäfer, D., and Yeh, Y. R. (2009), "Variable Selection and Oversampling in the Use of Smooth Support Vector Machines for predicting the default risk of Companies", *Journal of Forecasting*, 28, 512-534.

Härdle, W., Moro, R. A., and Schäfer, D., (2005), Predicting Bankruptcy with Support Vector Machines, *SFB 649 Discussion Paper*, Economic Risk, Berlin.

Harrington, D. P., and Fleming, T. R. (1982), A Class of Rank Test Procedures for Censored Survival Data, *Biometrika*, 69, 553-566.

Hillegeist, S., Keating, E., Cram, D., and Lundstedt, K. (2004), :Assessing the Probability of Bankruptcy,", Review of Accounting Studies, 9, 5-34.

Hosmer, D. W., and Lemeshow, S. (1980), A Goodness-of-fit Test for the Multiple Logistic Regression Model, *Communication in Statistics A*, 10, 1043-1069.

Hosmer, D. W., and Lemeshow, S. (2000), *Applied Logistic Regression*, 2nd Ed., Wiley-Interscience.

Hsieh, S., (1993) A Note on the Optimal Cutoff Point in Bankruptcy Prediction Models, *Journal of Business Finance & Accounting*, 20(3), 457-464.

Hurvich, C. M., and Tsai, C. L. (1989), "Regression and time series model selection in small samples", *Biometrika*, 76, 297-307.

Johnson, T.C., Chebonenko, T., Cunha, I., D'Almeida, F., and Spencer, X. (2011), "Endogenous Leverage and Expected Stock Returns", *Finance Research Letters*, 8(3), 132-145.

Kalbfleisch, J. D., and Prentice, R. L. (1973), Marginal Likelihoods based on Cox's regression and life model, *Biometrika*, 64, 47-50.

Kalbfleisch, J. D., and Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data*, 2nd Ed., Wiley, Hoboken, NJ.

Kim, H.S., Sohn, S.Y., (2010), Support Vector Machines for Default Prediction of SMEs Based on Technology Credit, *European Journal of Operational Research*, 201, 838–846.

Koh, H. C., (1992), The Sensitivity of Optimal Cutoff Points to Misclassification Costs of Type I and Type II Errors in the Going-Concern Prediction Context, *Journal of Business Finance & Accounting*, 19 (2), 187-197.

Koh, K., Kim, S. J., and Boyd, S. (2007), "An Interior-Point Method for Large-Scale L1-Regularized Logistic Regression". *Journal of Machine Learning Research*, 8, 1519-1555.

Lacher, R. C., Coats, P.K., Sharma, S.C., and Fant, L.F., (1995), A Neural Network for Classifying the Financial Health of a Firm, *European Journal of Operations Research*, 85, 53-65.

Lando, D. (2004), *Credit Risk Modeling: Theory and Applications*, Princeton University Press.

Lehmann, E. L., and Casella, G. (1998), *Theory of Point Estimation*, Springer.

Meinshausen, N., and Buhlmann, P. (2006), "High-Dimensional Graphs and Variable Selection with the LASSO". *Annals of Statistics*, 34, 1436-1462.

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

Murphy, S. A., Rossini, A. J., and van der Vart, A. W. (1997), Maximal Likelihood Estimation in the Proportional Odds Model, *Journal of the American Statistical Association*, 92, 968-976.

Murphy, S. A., and van der Vart, A. W. (2000), On Profile Likelihood, *Journal of the American Statistical Association*, 95, 449-465.

Nanda, S., and Pendharkar, P., (2001), Linear Models for Minimizing Misclassification Costs in Bankruptcy Prediction, *International Journal of Intelligent Systems in Accounting, Finance & Management 10*, 155-168.

Norton, C. L., and Smith, R. E., (1979), A Comparison of General Price Level and Historical Cost Financial Statements in the Prediction of Bankruptcy, *the Accounting Review* 54, 72-87.

Ohlson, J. S. (1980), Financial Ratios and the Probabilistic Prediction of Bankruptcy, *Journal of Accounting Research*, 19, 109-131.

Odom M. D., and Sharda R., (1990), A Neural Network Model for Bankruptcy Prediction, *Proceedings of the IEEE International Conference on Neural Networks*, II, 1990, 163-168.

Pesaran, M. H., and Timmermann, A. (1995), "Predictability of Stock Returns: Robustness and Economic Significance", *Journal of Finance*, 50, 1201-1228.

Scharfstein, D. O., Tsiatis, A. A., and Gilbert, P. B. (1998), Semiparametric Efficient Estimation in the Generalized Odds-rate Class of Regression Models for Right-Censored Time-To-Event Data, *Lifetime Data Analysis*, 4, 355-391.

Schönbucher, P. J. (2003), *Credit Derivatives Pricing Models: Models, Pricing and Implementation*, Wiley.

Schuermann, T. (2005), What Do We Know about Loss Given Default? In: Altman, E., Resti, A., Sironi, A. (Eds.), *Recovery Risk: The Next Challenge in Credit Risk Management*, Risk Books, London.

Schuermann, T., and Hanson, S., (2004), Estimating Probabilities of Default, *Federal Reserve Bank of New York*, Staff Reports No. 190, July 2004.

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

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

Tam, K. Y., and Kiang, M. Y., (1992), Managerial Applications of Neural Networks: The Case of Bank Failure Predictions, *Management Science*, 38(7), 926-947.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso", *Journal of the Royal Statistical Society*, *Series B*, 58, 267-288.

Uno, H., Cai, T., Pencina, M.J., D'Agostino, R., Wei, L.J. (2011), On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data, *Statistics in Medicine*, 30, 1105-1117.

Vapnik, V.N., (2000), *The Nature of Statistical Learning Theory*, Springer, New York.

Vassalou, M., and Xing, Y., (2004), Default Risk in Equity Returns, *Journal of Finance*, 59(2), 831-868.

Warner, J. B., (1977), Bankruptcy Costs: Some evidence, *Journal of Finance*, 32, 337-347.

Weiss, L. A., and Capkun, V., (2005), The Impact of Incorporating the Cost of Errors into Bankruptcy Prediction Models, Working Paper.

Wieand, S., Gail, M.H., James B.R. and James K.L. (1989). A Family of Nonparametric Statistics for Comparing Diagnostic Markers with Paired or Unpaired %Data. {\sl Biometrika}, 76, 585-592.

Wilks, D. S. (2006), *Statistical Methods in the Atmospheric Sciences*, 2nd ed. Academic Press.

Wu, C. O. (1995), Estimating the Real Parameter in a Two-Sample Proportional Odds Model, *Annals of Statistics*, 23, 376-395.

Zeng, D., and Lin, D. Y. (2006), Efficient Estimation of Semiparametric Transformation Models for Counting Processes, *Biometrika*, 93, pp. 627-640.

Zeng, D., and Lin, D. Y. (2007), Maximum Likelihood Estimation in Semiparametric Regression Models with Censored Data, *Journal of the Royal Statistical Society, Series B*,69, pp. 507-564.

Zhang, G., Hu, M. Y., Patuwo, B. E., and Indro, D. C., (1999), Artificial Neural Networks in Bankruptcy Prediction: General Framework and Cross-Validation Analysis, *European Journal of Operational Research*, 116, 16-32.

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of the LASSO", *Journal of Machine Learning Research*, 7, 2541-2563

Zmijewski, M. (1984), Methodological Issues Related to the Estimation of Financial Distress Prediction Models, *Journal of Accounting Research*, 22, 59-82.