Norwegian School of Economics

Bergen, fall 2015

# Bankruptcy Prediction

*Static Logit and Discrete Hazard Models incorporating Macoreconomic Dependencies and Industry Effects*

**Suleman Sheikh and Muhammad Yahya**

**Supervisor: Professor Liam Brunt**

Master thesis, MSc, Finance

NORWEGIAN SCHOOL OF ECONOMICS

[This page intentionally left blank]

# Abstract

In this thesis, we present firm default prediction models based on firm financial statements and macroeconomic variables. We seek to develop reliable models to forecast out-of-sample default probability, and we are particularly interested in exploring the impact of incorporating macroeconomic variables and industry effects. To the best of our knowledge, this is the first study to account for both macroeconomic dependencies and industry effects in one analysis. Additionally, we investigate the impact of the 2008 financial crisis on bankruptcies.

We develop five models, one static logit model and four hazard models, and compare the out-of-sample predictive performance of these models. To explore the impact of industry effects and the financial crisis, our study includes 562 U.S. public companies across all sectors (except financial) that filed for bankruptcy between 2003 and 2013. These were matched to a control group of non-bankrupt firms.

We find that the cash flow, profitability, leverage, liquidity, solvency, and firm size are all significant determinants of bankruptcy. The ratio of cash flow from operations to total liabilities, and total debt to total assets, are the most significant variables in the static logit model. In addition to these ratios, cash to total assets and net income to total assets are also among the most important covariates in the hazard models. Next, we find that the forecasting results are improved by incorporating macroeconomic variables. Finally, we find that the hazard model with macroeconomic variables and industry effects has the best out-of-sample accuracy.

**Keywords**: Bankruptcy prediction, static logit model, discrete hazard model, macroeconomic patterns, industry effects.

# Acknowledgements

This thesis was written in the fall of 2015 at Norwegian School of Economics, as a part of our Master of Science degree in Economics and Business Administration.

We hope this thesis will contribute to the interesting field of economics and that it sheds light on the dispute between static logit and discrete hazard rate models.

We would like to thank Norwegian School of Economics for making available the databases that build the foundation of this research. We also thank senior advisor Eivind Bernhardsen (FSA Norway) for providing helpful feedback and insights.

Most of all, we would like to express our sincere gratitude to our supervisor, Professor Liam Brunt, for his support and invaluable guidance. His time and valuable suggestions lead us to improve the quality of this research.

Finally, our special thanks goes out to our beloved families, for their wholehearted support and encouragement.

# Contents

# List of Tables

# List of Figures

[This page intentionally left blank]

# 1. Introduction

"Bankruptcy is more likely when the economy moves from boom to recession" (Lennox, 1999). Recession and bankruptcy are two terms of great importance for the economy and society, as events since 2008 have shown. Consequently, researchers have been studying bankruptcy prediction more frequently than ever (Jones & Hensher, 2008). The topic is of such importance that it concerns all stakeholders: from employees to managers, investors, and regulators. However, to fully understand the term bankruptcy, we have to define it first.

Most of the bankruptcy prediction studies define failure legalistically. This provides an objective criterion to easily classify the firms that are being examined. The legal definition of failure is also applied in this study. According to the Title 11 of the U.S. Code, the term "insolvent" is defined as: *"Financial condition such that the sum of such entity's debts is greater than all of such entity's property".*[1] In other words, the company is considered insolvent if the total liabilities of a firm are greater than its total assets.

Insolvency is one of the most significant threats for corporations today, despite their size and the nature of their operations. Substantial evidence shows that business failures have occurred at higher rates over the last three decades than at any time since the early 1930s (Charitou, Neophytou, & Charalambous, 2004). As an illustration, there were more than a thousand banks in the U.S. that failed during the period 1985 to 1992 - more than 100 every year (Cole & Wu, 2009).

Different factors lead to business failures. Many economists emphasize the roles of heavy debts, high interest rates and reduced profits. Furthermore, government regulations can

[1] U.S.C. Title 11, Chapter 1 § 101 (32) (A).

affect specific industries and contribute to financial distress.[2] Studies show that small, private, and newly founded companies are more vulnerable to financial distress, rather than large well-established public firms. However, the recent global financial crisis proves that even the larger corporations are vulnerable. It also reminds us how important a well-functioning banking system is for economic growth. The regulators, for instance, took over numerous banks and financial institutions during the financial crisis to keep them as going concerns and avoid a credit crunch (Cole & Wu, 2009). Frozen international credit markets generated a global recession and increased unemployment.

The consequences of the financial crisis emphasize the importance of credit risk management. Credit risk can be defined as "a borrower's failure to meet contractual obligation" (Jones & Hensher, 2008). This failure may arise whenever a borrower is expecting to use uncertain future cash flows to pay a current debt and may eventually lead to insolvency. Hence, predicting the probability of corporate default can be valuable for both creditors and investors. For banks, this can lead to improved lending practices as well as setting interest rates that reflects credit risk. Naturally, investors can also benefit from these predictions, as they can preclude investing in businesses with high probability of default. However, bankruptcy prediction affects more than just banks and investors. Default probabilities can also be used to assist managers, auditors and regulatory agencies. To emphasize the importance of this topic; note that auditors can risk potential lawsuits if they fail to provide early warning signals of failing firms (Lennox, 1999).

Predicting corporate bankruptcies is therefore an important and widely studied topic (Wilson & Sharda, 1994). Indeed, to predict the probability of default accurately, reliable

[2] Government agencies can set restrictions that lead to increased costs, comprised profits or even lawsuits. For instance, U.S. Patent and Trademark Office (USPTO) can impose heavy fines for patent and trademark violations; Food and Drug Administration (FDA) can withhold approvals for pharmaceutical companies; Environmental Protection Agency (EPA) can file lawsuits against firms for violating environmental rules.

empirical models are much needed. This allows the stakeholders to take either preventive or corrective action.

## 1.1 Motivation and Objective of the Study

The subject of bankruptcy prediction is both interesting and challenging, as it affects all stakeholders in the business world. In addition, the subject brings together economic and legal (institutional) issues. Moreover, researching this topic allows us to choose a sample where we can evaluate the impact of the recent financial crisis.

The main objective of our study is to develop a reliable default prediction model using recent data. We compare the accuracy of forecasting bankruptcy using a static logit model and four hazard rate models. In the static logit model, we use cross-sectional data, whereas in the hazard rate models we use time-varying data to better exploit the richness of our data. We also try to see if the predictive power of the hazard models can be improved by incorporating macroeconomic dependencies and industry effects simultaneously. To the best of our knowledge, this is the first research to apply both of these in one analysis. Several previous studies have incorporated either one, but not both together (Chava & Jarrow, 2004; Hill, Perry, & Andes, 2011; Nam, Kim, Park, & Lee, 2008). Further, we also want to test whether there is significant increase in the number of bankruptcies post-2008.

To answer these research questions, we based our research on U.S. listed companies. Conducting research on U.S. companies is a nice natural experiment, because legislation varies by state. However, the culture and the business structure are similar across the country – which makes it a cleaner comparison than cross-country research.

## 1.2 Limitations

Our data set consists of all U.S. firms that filed for bankruptcy during 2003-2013, and had available data. We compiled annual historical data from company financial statements. Employing annual data obscures the fact that the companies' financial position might be significantly different at the time of filing for bankruptcy. However, comparable monthly and quarterly data are unavailable so we cannot do better here, even though inclusion might improve the predictive power of the models (Baldwin & Glezen, 1992; Chava & Jarrow, 2004). We are also aware of the fact that the models could have been improved by adding market data, such as market capitalization, market to book ratio, firm age or number of employees (Campbell, Hilscher, & Szilagyi, 2008; Lennox, 1999; Shumway, 2001). However, the market data was also omitted due to unavailability – for most of the companies. Hence, the models only rely on financial ratios and macroeconomic dependencies. Moreover, the models do not account for the bankruptcy exit date of the companies. If these dates were available, then we could identify how many corporations that filed for Chapter 11 bankruptcies actually managed to reorganize successfully and exit bankruptcy. By contrast, a great strength of our study is a control group matched to the sample of defaulting firms.

## 1.3 Outline of the Thesis

The structure of the thesis is as followed. In the next section, we review the previous research on default prediction. Section 3 explains the applied methodology for the models, misspecification tests, and different approaches used for model evaluation. In section 4, we thoroughly describe the sample and the data collection process. Section 5 examines the variable selection. Section 6 presents and discusses our results. Section 7 evaluates the performance of our models. Section 8 concludes.

# 2. Literature Review

In this section, we summarize previous research. The literature review is divided in five subsections. In the first, second and third subsections we present the research on discriminant analysis, logit models and hazard models respectively. This is followed by the discussion on other research. The last subsection presents a comparison of logit and hazard models.

## 2.1 Research on Discriminant Analysis (DA)

Predicting firm default probability is a vastly researched field. Numerous researchers have attempted to build reliable bankruptcy prediction models. Altman (1968) used Beaver's (1966) pioneering work in this field to create the first statistical model. His data set included 66 failed and non-failed manufacturing companies over the period 1946 to 1965. He used Multivariate Discriminant Analysis (MDA) in order to construct a model that utilized financial ratios for predicting corporate defaults. The resulting model attained global prominence and is known as the *Z*-score. Altman found that a firm is more likely to fail if the firm is highly leveraged, unprofitable, and suffers cash flow difficulties (Lennox, 1999).

The MDA is criticized mainly for two assumptions: the multivariate normal distribution assumption that it imposes on explanatory variables; and the assumption of independent and identical distribution, for instance, that firms were selected randomly from the population of non-failed and failed firms (Jones & Hensher, 2008; Lennox, 1999).

## 2.2 Research using Logit and Probit Models

Due to shortcomings discussed in the preceding subsection, researchers tried models that relax these assumptions, which led to the application of logit and probit models. Ohlson's (1980) study was based on observations from 105 failed firms and 2058 non-failed firms employing data from 10-K financial statements. The model generates the *O*-score, which is similar to Altman's *Z*-score. He identified four factors as statistically significant for predicting the probability of default: I) size of company, II) a measure of performance, III) a measure of current liquidity, and IV) a measure of firm's financial structure. The major disadvantage of his model is that it takes no account of the market data of the firms. Zmijewski (1984) employed a probit model. His study consisted of 81 failed and 1600 non-failed firms between 1972 and 1978. His research indicated three variables as statistical significant in explaining the probability of default: I) return on assets; II) financial leverage; III) liquidity.[3]

Lennox (1999) re-evaluated the performance of probit, logit and DA. He employed these models on sample of 949 public companies in UK. His two most important findings were that: the leverage and cash flow of a firm has non-linear effects on probability of default; and probit and logit models are better than DA at predicting bankrupt firms.

Westgaard & Van der Wijst (2001) found that the logit model is able to predict defaults sufficiently well, using liquidity, financial coverage, size of the firm, solidity, cash flow to debt and age of the firm.

These models take into account only the cross-sectional data of the firms and thus ignore the fact that the characteristic of a firm changes over time. In other words, these

[3] Liquidity = Current assets/Current liabilities

models do not consider the time-varying covariates of the firm while predicting the probability of bankruptcy (Shumway, 2001).

## 2.3 Research on Hazard Models

Shumway (2001) suggested a hazard model that capture the changes in firm characteristics over time. His data set consisted of 300 failed firms over the period 1962 to 1992. The use of hazard models yielded better results for predicting probability of bankruptcy. Most of the financial ratios that were significant in static models became insignificant when employed in the hazard rate model. Moreover, he incorporated market variables in the hazard rate model, which proved to be significant in predicting default. He emphasized using firm age as a baseline to capture the common characteristics among firms. The variables included in his research are; the past stock returns, market size, and the idiosyncratic standard deviation of returns.

Chava & Jarrow (2004) further improved the already superior forecasting performance of Shumway's (2001) model by incorporating industry effects. They estimated the hazard rate model using both monthly and yearly data over the period 1962-1999. Further, they emphasize the importance of using monthly or quarterly data, as it markedly captures changes in firm's characteristics and thus improves forecasting. Additionally, they found that incorporating industry effects significantly changes both the intercept and slope coefficients.

Nam et al. (2008) compared a static logit model with two hazard models, with and without macroeconomic variables, as a baseline specification. The sample consisted of 367 Korean companies over the period 1991-2000. They used two macroeconomic variables; change in the interest rate and the volatility of foreign exchange. The results indicate that the dynamic models with time-varying covariates yield superior performance compared to

static logit models. The hazard model with macroeconomic variables was also more accurate in predicting probability of default. Bellotti & Crook (2009) concluded that including macroeconomic variables, such as interest and unemployment rates, significantly affects default probability and improves prediction accuracy.

Hill et al. (2011) conducted an event history analysis on financially distressed firms. Their paper considered the difference between financially distressed firms that survive and those that ultimately go bankrupt. They also incorporated two macroeconomic variables, the interest rate and the unemployment rate, to reflect changes in the overall economy. Their data set included 75 failed firms between 1977 and 1987. Both macroeconomic variables were found to be significant, and improved the overall performance of the model.

## 2.4 Other Research

In addition to the statistical models mentioned above there is another approach that has emerged over recent years – neural networks (NN). This approach is applied to different business areas including credit analysis and bankruptcy prediction. NN are computer systems that identify specific patterns, and use these patterns to solve given problems. Empirical evidence proves that the computer systems can provide at least as reliable results as the traditional statistical models (Charitou et al., 2004). Despite the fact that we cannot implement this method, we want to mention its existence in predicting default probability.

## 2.5 Comparison of Logit with Hazard Models

The common finding in the abovementioned research is that the predictive power of MDA is weaker than static logit models and the hazard rate models. The latter, with time-

varying and macroeconomic covariates, are better at providing forecasts for both in- and out-of-sample estimates.

Shumway (2001) provided a detailed comparison of hazard rate and static logit models for predicting bankruptcy. He argued that static models are inconsistent due to the nature of bankruptcy data. Due to infrequency in bankruptcy events, the researchers use data that spans over several years in order to obtain a suitable sample for analysis. However, the underlying characteristics of most firms change over time, which is not captured by static logit. Most researchers use the data for each firm in the year preceding bankruptcy, thus ignoring the data for the healthy firms that may eventually file for bankruptcy. This might result in selection bias in the estimates (Hillegeist, 2001).

Secondly, the hazard rate models are preferred over the static models due to its ability to incorporate all the available information in order to determine each firm's risk of default at each point in time. The dependent variable of a hazard rate model would be the time that a firm spent in the healthy group (Shumway, 2001).

Finally, due to the incorporation of time-varying data for each firm over several years, the out-of-sample forecasting ability of the hazard rate models would be more than the logit model. For instance, the hazard model can be seen as binary logit model that treats each firm year as a separate observation (Shumway, 2001). Furthermore, in this thesis, we have chosen data for each firm for the five preceding years until it files for bankruptcy in year $t$, so we have five times more data than the cross-sectional logit.

Nonetheless, researchers are still using single period logit to predict bankruptcy. There is empirical evidence showing that the out-of-sample predictive power of simple logit model is better, or at least comparable, to the more advanced models (Fantazzini & Figini, 2009; Galil & Sher, 2015; Halling & Hayden, 2006; Nam et al., 2008). This indicates that even

though the cross-sectional logit might not seem to be an accurate specification, it might still able to provide good or superior results for forecasting out-of-sample defaults.

To summarise, there is no academic consensus for favouring complex models over static logit. Therefore, it is necessary to examine the predictive ability of these models to resolve the controversy among them.

In this thesis, we will adapt the models implemented by Chava & Jarrow (2004), Hill et al. (2011), Nam et al. (2008), Ohlson (1980), and Shumway (2001). Additionally, we considered the global financial crisis, macroeconomic variables, and industry effects, simultaneously, to forecast their effect on predicting bankruptcy. To the best of our knowledge, previous studies had incorporated either macroeconomic variables (Nam et al., 2008; Shumway, 2001) or industry effects (Chava & Jarrow, 2004). Therefore, we want to measure the effect of implementing these together in one analysis. We expect superior out-of-sample predictive ability of this model compared to the above-mentioned models.

# 3. Methodology

In this chapter, we present the theoretical underpinnings of the econometric framework that forms the basis of our analysis. We first discuss logit models, then hazard rate models. We present our model specifications, and the frameworks employed to check for the presence of functional misspecification, omitted variables and heteroscedasticity. The section concludes with a description of several validation tests.

## 3.1 The Logit Model

Researchers have recently preferred logit models over discriminant analysis, because logit models do not impose any assumptions regarding the distribution of predictors. Also, logit models provide results in terms of probabilistic outcomes and therefore do not require any score to be converted into probabilities, which can be an additional source of error (Ohlson, 1980).

Logit models assume that, for a firm with a given set of predictors, there is a certain probability that the firm will default. The dichotomous dependent variable takes the value of 1 for a bankrupt firm or 0 for a healthy firm.

$$\boldsymbol{X} = \left(X_{ij}\right), \quad j = 1, \dots, n; \; i = 1, \dots, k. \tag{1}$$

Where X is the set of independent variables that contribute towards default and $\boldsymbol{\beta}$ is the vector of unknown parameters, $\boldsymbol{k}$ is the number of explanatory variables, and $\boldsymbol{n}$ is the number of firms. For instance, the data for $\boldsymbol{i}^{\text{th}}$ firm is given by $\boldsymbol{X_i}$. The logit model provides the probability of $Y_i = 1$, given $\boldsymbol{X_i}$, as the cumulative standard logistic distribution function. Given the estimates of parameter $\boldsymbol{\beta}$, the probability of default for firm $\boldsymbol{i}$ can be estimated using the following equation:

$$p_i = Pr(Y_i = 1|\boldsymbol{X_i}) = F(\beta_0 + \beta_i \boldsymbol{X_i}) = \frac{1}{1 + e^{-(\beta_0 + \beta_i \boldsymbol{X_i})}}, \quad i = 1, \dots, n \tag{2}$$

Where $F(\beta_0 + \beta_i \boldsymbol{X_i})$ is the cumulative logistic distribution (Greene, 2003) and:

$$Y_i = \begin{cases} 1, & if\ Y_i^* > 0, \\ 0, & otherwise, \end{cases}$$

To estimate the model parameters of $\boldsymbol{\beta}$ vector, the log-likelihood function of the following form is maximized (Baltagi, 2002):

$$ln(L) = \sum_{i=1}^{n} \left[ Y_i ln F(\beta_0 + \beta_i \boldsymbol{X_i}) + (1 - Y_i) ln\left(1 - F(\beta_0 + \beta_i \boldsymbol{X_i})\right) \right] \tag{3}$$

In this thesis, all the analysis is carried out using Stata.[4] In addition to the single-period (cross-sectional) logit, there are multi-period logit models that will be discussed in the following sub-section.

## 3.2 Hazard Models

This sub-section discusses discrete hazard models. Hazard models are classified as a type of survival models. The covariates in hazard models are related to the amount of time that passes before the occurrence of an event (i.e. bankruptcy filing). In other words, each firm has multiple observations for each predictor and its risk for moving from healthy to bankrupt changes over time depending on these covariates.

Due to the annual nature of our data set, bankruptcy can occur only at discrete points in time, $t = 1, 2, 3, \dots$ . Further, denote the failure time as $t_i$ for each firm. The dichotomous (dependent) variable, $y_i$, is equal to 1 if the firm defaults at $t_i$, and it takes the value of

[4] In Stata, we just need to define our dependent and independent variables and it provides us with the coefficients using a maximum likelihood estimator.

zero otherwise. The continuous random variable follows a probability mass function given by, $f(t_i, \boldsymbol{X_i};\ \beta)$, and has a cumulative density function given by the following expression, $F(t_i, \boldsymbol{X_i};\ \beta)$. The survivor function estimates the probability that a firm will survive up to time $t$ is given by $S(t_i, \boldsymbol{X_i};\ \beta)$ and it is defined as:

$$S(t_i, \boldsymbol{X_i};\ \beta) = 1 - \sum_{j<1} f(t_i, \boldsymbol{X_i};\ \beta) = 1 - F(t_i, \boldsymbol{X_i};\ \beta) \tag{4}$$

The hazard function provides the probability that the firm has filed for bankruptcy at $t$, which is conditional on surviving to $t$. The relationship between survivor function and the hazard rate can be expressed as:

$$h(t_i, \boldsymbol{X_i};\ \beta) = \frac{f(t_i, \boldsymbol{X_i};\ \beta)}{S(t_i, \boldsymbol{X_i};\ \beta)} \tag{5}$$

The $\beta$ parameters in the hazard rate models are estimated using the maximum likelihood function and it can be expressed as:

$$L = \prod_{i=1}^{n} h(t_i, \boldsymbol{X_i};\ \beta)^{y_i} S(t_i, \boldsymbol{X_i};\ \beta) \tag{6}$$

Where the parametric form of the hazard rate, $h(t_i, \boldsymbol{X_i};\ \beta)$, is often assumed. Hence, the model can incorporate the time-varying covariates by making $\boldsymbol{X}$ dependent on time (Shumway, 2001).

Hazard models are closely related to logit models, described in the preceding subsection. Shumway (2001) has proved that the likelihood estimator of a discrete-time hazard model with the hazard function, $h(t, \boldsymbol{X};\ \beta) = F(t, \boldsymbol{X};\ \beta)$, is equivalent to the multi-period logit model. The expression for logistic regression with time varying covariates $\boldsymbol{X_{it}}$ to estimate the parameters of $\beta$ for discrete hazard rate model are estimated using the following equation:

$$h(t, \boldsymbol{X};\ \beta) = F(t, \boldsymbol{X};\ \beta) = \frac{1}{1 + e^{-(\beta \boldsymbol{X}_{it})}} \tag{7}$$

Additionally, each bankrupt firm will only have one failure observation i.e. the dependent variables, $y_{it}$, will take the value of 1 for the bankruptcy year and value of 0 for the preceding years when the firm was healthy. To incorporate the time-varying covariates in a logit model, each annual financial ratio is used as a firm-year observation, which is similar to "event history analysis" by Hill et al. (2011). Event history analysis "considers the changes in independent variable over time" i.e. each firm has longitudinal data (panel data), and over time the state of the firm ($y_{i,t}$) changes from healthy to default (Hill et al., 2011).

To allow for the incorporation of baseline hazard rate, we can partition $\boldsymbol{\beta}$ into $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$. Following (Chava & Jarrow, 2004; Shumway, 2001), we get the following form of logistic regression with the parameters $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ for our hazard rate model with a baseline hazard rate:

$$h(t, \boldsymbol{X};\ \beta) = F(t, \boldsymbol{X};\ \beta) = \frac{1}{1 + e^{-(k_t \beta_1 + \boldsymbol{X}_{it} \beta_2)}} \tag{8}$$

From the above equation, we can see the hazard model consisting of $k_t$, which is a time-dependent variable, also called the baseline of hazard function. It expresses the hazard rate of a firm if the covariates $\boldsymbol{X}$ are absent. By contrast, $\boldsymbol{X}_{it}\boldsymbol{\beta_2}$ is a function of idiosyncratic characteristics of the firm represented by financial ratios. By incorporating the time varying covariates, we are accounting for the change in financial condition over time (Shumway, 2001).

The baseline hazard function is normally represented by some macroeconomic variable. Omitting the baseline from the model is analogous to estimating an exponential hazard model in which the probability of a firm's failure does not depend on a baseline. We use

the same approach as Hill et al. (2011) and Shumway (2001) for entering the macroeconomic variables as the baseline rate, i.e. by including them as covariates to take into account the temporal dependence in the hazard rate. Further, he used $ln(firm's\ age)$ as proxy for baseline. Other researchers used different baseline proxies, for instance, Hillegeist (2001) used the changes in interest rate and the rate of recent defaults, and Hill et al. (2011) used the prime rate and the unemployment rate.

## 3.3 Specifications of Models

We formulated five model specifications to investigate the performance improvement from using discrete hazard rate models over a static logit model. Additionally, we tested for the performance improvement in hazard rate models by adding macroeconomic variables and industry effects. Further, we tested for the improvement in predictive power of hazard rate models when both macroeconomic variables and industry effects are employed simultaneously in the model.

### 3.3.1 Model I

In the static logit model, we have just one observation per firm, hence, the covariates of the firm do not change over time. This can be formulated as follows:

$$Pr(y_i = 1) = \frac{1}{1 + e^{-(\beta \boldsymbol{X}_i)}} \tag{9}$$

Where $\boldsymbol{X}_i$ is the cross-sectional observation for each firm in the sample.

### 3.3.2 Model II

The hazard rate model with covariates that vary over time can be specified as the following logit form:

$$Pr(y_i = 1) = \frac{1}{1 + e^{-(\boldsymbol{X}_{i,t}\beta)}} \quad (10)$$

Where $\boldsymbol{X}_{i,t}$ are the changes in independent variables over time. This is the exponential hazard model where the firm's probability of failure does not depend on some baseline (Nam et al., 2008; Shumway, 2001).

*Hypothesis 1: Time-varying models have a better predictive power than cross-sectional logit model.*

To test this hypothesis, we compared the predictive performance of Model I and Model II "out-of-sample". This is measured using the classification matrix, where the overall predictive accuracy of the two models is compared.

### 3.3.3 Model III

In Model III, we added macroeconomic variables alongside the time-varying covariates in the hazard model:

$$Pr(y_i = 1) = \frac{1}{1 + e^{-(\boldsymbol{k}_t\beta + \boldsymbol{X}_{i,t}\beta)}} \quad (11)$$

Where $k_t$ is the baseline hazard model to capture changes in the macroeconomic environment. We tested the significance of the CPI, stock market returns, GDP, unemployment and the prime rate.

*Hypothesis 2: The predictive power of Model III will be better than both Model I and Model II, as a result of adding macroeconomic variables.*

### 3.3.4 Model IV

In Model IV, we included industry effects in Model II, alongside the time-varying covariates in the hazard model:

$$Pr(y_i = 1) = \frac{1}{1 + e^{-(\boldsymbol{X}_{i,t}\beta + \beta(Industry) + \beta(\boldsymbol{X}_{i,t} \times Industry))}} \quad (12)$$

Where $Industry$ is a dummy variable that takes value of 1 for a particular sector and 0 otherwise. Further, each industry dummy variable is interacted with the covariates in order to estimate change in slope coefficient between different industries. Using this criterion allowed us to test for changes in significance of the selected covariates from one industry to another and the change in intercept as well (Chava & Jarrow, 2004; Hill et al., 2011). Unlike Chava & Jarrow (2004), we did not group the sectors into three industries; instead, we treated each sector as segregate, only exception of "Energy" and "Consumer discretionary". These sectors are relatively different from other sectors and therefore, we group the interaction terms for these two sectors. Chava & Jarrow (2004) reported that using this approach is equivalent to estimating a separate hazard model for each industry.

*Hypothesis 3: Inclusion of industry effects does not improve the predictive power of the hazard rate model.*

*Hypothesis 4: Inclusion of industry effects would not be significant in the hazard rate model.*

To test hypothesis 4, we used the likelihood ratio test to gauge for significance. The model with industry dummies and interaction variables is used as the unconstrained model and this is tested against the model without industry dummies. Under the null hypothesis, there is no significant effect of inclusion of these industry variables in the model.

### 3.3.5 Model V

In the final specification, we investigate the forecasting accuracy of including time-varying covariates, macroeconomic variables, and the industry effects, simultaneously. The hazard model with inclusion of these variables can be formulated as the following functional form:

$$Pr(y_i = 1) = \frac{1}{1 + e^{-\left(\boldsymbol{k}_t\beta + \boldsymbol{X}_{i,t}\beta + \beta(Industry) + \beta(\boldsymbol{X}_{i,t} \times Industry)\right)}} \tag{13}$$

This specification, not only accounts for changes in the macroeconomic environment on bankruptcy, but also considers the effects on each industry.

*Hypothesis 5: The predictive power of Model V would be better than the previous models.*

This is due to the fact that we are accounting for both macroeconomic variables and industry effects in this model.

## 3.4 Tests for Misspecification

In this subsection, we discuss possible tests for functional form misspecification, omitted variables, and heteroscedasticity.

### 3.4.1 Test for Specification Error and Omitted Variables

Some variables are found to have non-linear effects on bankruptcy probability, such as leverage and cash flow (Lennox, 1999). This might lead to omitted variable bias due to exclusion of the respective quadratic terms. Hence, we need to test for functional form misspecification. We adopt the framework proposed by (Pregibon, 1979, 1980) to test for the inclusion of non-linear independent variables. Here, we generate the predicted values ($\hat{p}$) and the square of the predicted values ($\hat{p}^2$), and use these as independent variables, which are regressed on the binary dependent variable in the auxiliary regression, as shown below.

$$\hat{p} = \hat{\beta}\boldsymbol{X} \tag{14}$$

$$y = \beta_0 + \beta_1\hat{p} + \beta_2\hat{p}^2 \tag{15}$$

If the model is correctly specified, then squared term of the predicted values should not be significant or have much predictive power. If the square of predicted values are significant then this indicates that the quadratic terms of the independent variables need to be added or that we have omitted some relevant variable(s) from the model (Pregibon, 1979, 1980).[5]

### 3.4.2 Test for Heteroscedasticity

Heteroscedasticity can be a problem in the logit model, which can result in the parameter estimates being inconsistent (Davidson & MacKinnon, 1984; Verbeek, 2008). The problem of heteroscedasticity occurs when the variance of $\varepsilon_i$ depends on some exogenous variables $z_i$ and can be expressed as:

$$V\{\varepsilon_i\} = kh(z_i'a) \tag{16}$$

Where $k = {\pi^2}/{3}$ and $h$ is some function $h > 0$ with $h(0) = 1$, and $h'(0) \neq 0$.

We tested for heteroscedasticity by plotting the standardized residuals against the fitted values for visual inspection of heteroscedasticity, and we used the "White test". The square term of the standardized residuals is used as dependent variable and is regressed on predicted values and the square of predicted values from the first regression. If the coefficients are significant then there is problem of heteroscedasticity, indicating that the variance of the error term depends on exogenous variable.

To overcome the problem in cross-sectional logit, we used robust standard errors (Allison, 2012). However, for the hazard rate model, the observations are likely to be dependent within clusters.[6] Hence, we used cluster robust standard errors (Allison, 2012; Arminger, Sobel, & Clogg, 1995; Hilbe, 2009; Hosmer Jr, Lemeshow, & Sturdivant, 2013; Long & Freese, 2006).

[5] We used the Stata command "linktest" in order to detect the specification error after fitting our logit model.
[6] Firms are referred as clusters in this sense.

## 3.5 Approaches for Model Evaluation

In this sub-section, we present several tests to evaluate the out-of-sample predictive power of the models, thus being able to compare the models.

### 3.5.1 Classification Matrix for Model Accuracy

The coefficients of the fitted model are used to classify the out-of-sample outcomes (Hosmer Jr et al., 2013). The classification table does not only provide the accuracy of the model in predicting default, but also reflects the embedded uncertainties in the model. There are two ways the uncertainties are embedded in the model. First, the classification of a bankrupt firms as non-bankrupt (Type I error), and the classification of a non-bankrupt firms as bankrupt (Type II error). The costs associated with Type I error are; a lender might lose principal, interest, and potential lawsuits, and an investor might lose his investment. For the Type II error; an investor might lose the foregone profit he could have generated by undertaking the investment opportunity in those firms, and a lender might lose foregone interest and more importantly a potential customer (Bellovary, Giacomino, & Akers, 2007). The following table provides an overview of the two types of errors discussed:

**Table I. Classfication matrix**

| | Observed | |
|---|---|---|
| Classified | Bankrupt | Non-bankrupt |
| Bankrupt | Correctly predicted | Type II error |
| Non-bankrupt | Type I error | Correctly predicted |

Notes: The following table shows the two types of errors. Type I error is the classification of bankrupt firms as non-bankrupt. Type II error is the classification of non-bankrupt firms as bankrupt. Sensitivity is the correctly predicted bankrupt firms and specificity is correctly predicted non-bankrupt firms.

In order to estimate the classification table, we need a threshold point above which the model distinguishes between bankrupt and non-bankrupt firms. The intersection between sensitivity[7] and specificity[8] can be used as a threshold point (Hosmer Jr et al., 2013), because around this point, the Type I and Type II error are at the optimal level. We obtained the intersection point by plotting the sensitivity against the specificity.[9] Additionally, we have provided the sensitivity analysis by using different threshold for estimation of classification table and its impact on the overall accuracy of model.

### 3.5.2 Area under the Receiver Operating Characteristics Curve

Receiver Operating Characteristics (ROC) plots the probability of true default $(sensitivity)$ against the incorrectly predicted default $(1 - specificity)$. This measure evaluates the ability of the fitted model to assign higher probability when the outcome $(y = 1)$ than to the outcome $(y = 0)$. The Area Under the ROC (AUROC) curve can range from 0.5 to 1. An AUROC close to 1 indicate the better ability of the model in differentiating between the two outcomes (Hosmer Jr et al., 2013).[10]

### 3.5.3 Goodness-of-fit Deciles

Hosmer & Lemesbow (1980) proposed grouping the estimated out-of-sample probabilities into deciles. Further, Lemeshow & Hosmer (1982) suggested using a group size of 10, which would result in the first group containing the $n_1' = {}^{n}/_{10}$ firms having the smallest estimated probabilities, whereas the last group containing the $n_{10}' = {}^{n}/_{10}$ firms having the largest estimated probabilities (Hosmer Jr et al., 2013). Consistent with Chava

[7] Sensitivity is the correct classification of the actual bankrupted firms.
[8] Specificity is the correct classification of the actual non-bankrupted firms.
[9] We used the Stata command "lsens" to plot the sensitivity and specificity for the out-of-sample period.
[10] We used the Stata command "lroc" for the years 2011, 2012 and 2013 to get the ROC curve and the AUROC.

& Jarrow (2004) and Shumway (2001), we can compare the model based on its ability to allocate defaulted firms across different deciles. The model is considered good if it allocates higher percentage to the top decile.[11]

### 3.5.4 Brier Score

The Brier Score (BS) is a commonly used measure for evaluating probabilistic forecasts (Roulston, 2007). The BS measures the disagreement between the observed outcomes and the forecasted outcomes. The score lies between 0 to 1 and the lower score reflects the better probability forecast of the model. The following equation is used for the estimation of BS:

$$Brier\ Score = \sum_j \left(y_j - \hat{p}_j\right)^2 \Big/ N \tag{17}$$

Where $N$ reflects the number of observations, and $\hat{p}_j$ is the forecast default probabilities.

### 3.5.5 Likelihood Ratio Test

The likelihood ratio test (LRT) is used to compare the goodness-of-fit of a constrained model over the unconstrained model. For instance, the model with industry dummies and interaction variables is set as the unconstrained model and tested against the model without industry dummies. We used LRT to estimate the significance of industry dummies. Under the null hypothesis, the bankruptcy prediction is not affected by the industry effects i.e. $\beta_{IND} = 0$.

$$G = -2ln\left[\frac{(Likelihood\ without\ the\ variable(s))}{(Likelihood\ with\ the\ variable)}\right] \tag{18}$$

[11] We implemented this approach using Stata command "estat gof, group(10) table".

# 4. Data

In this section, we present the data for both bankruptcy[12] and financials. This is followed by the discussion of the requirements for data set inclusion, the control group, and data quality. The section concludes with a discussion of construction of data sets for analysis and the selection of period for in- and out-of-sample.

## 4.1 Sample Selection

The sample for this study consists of 562 U.S. companies that filed for bankruptcy between 2003 and 2013. The list of defaulted firms is obtained from the Bloomberg terminal (hereafter BB). The terminal possesses data for over 800,000 securities worldwide. It gathers the data from a combination of different sources; stock exchanges, the companies directly, public filings and global news. The bankruptcy data is gathered from court dockets, company filings, and press releases. We used a function in BB[13] that allowed us to systematically narrow down securities by different criterions. We selected our companies based on the following criterions:

I. Country of domicile: United States. (173,956)

II. Public companies. (115,824)

III. The bankruptcy filing took place between 01.01.2003 – 31.12.2013. (1,340)

IV. Total assets known. (1,013)

V. Bankruptcy defined under Chapter 7 and Chapter 11. (815)

VI. All sectors, except financial. (753)

[12] List of defaulted firms.

[13] The function is called "Equity Screening" in Bloomberg.

Firstly, we chose the United States as the country of domicile because many elements are thereby held constant (culture, currency, government), but we can still see if the difference in state legislation has any effect on bankruptcy. BB has data on over 170,000 securities in the U.S.; 3,290 of these entities filed for bankruptcy. Moreover, we included companies from all sectors except financial, to get as much variation in the data as possible. (Obviously, the financial sector has certain special characteristics such as capital structure, which makes it unlike other sectors. So it is standard procedure to drop that sector.)

Second, there was a tremendous amount of missing data on private companies that went bankrupt. Hence, we included only public companies in our sample because they naturally had more data available than private companies. Considering only public companies means that all companies have a similar basis for comparison. *Screening result: 115,824*.

Third, by choosing the time period between 2003 and 2013 we cover 11 years, five years prior to the financial crisis, and five years after it. Choosing this period will provide us a sufficient time frame to analyse the impact of the financial crisis. *Screening result: 1,340.*

Fourth, around 40% of the aforementioned companies had no data or significantly missing data. The preliminary sample is therefore narrowed down with respect to total assets. Despite this criterion, there were still companies that did not have data on total assets. *Screening result: 1,013.*

Fifth, there were 17 different types of bankruptcy filings on BB. However, there are only two types of corporate bankruptcies that are legally defined: Chapter 7 (liquidation) and Chapter 11 (reorganization). Hence, the sample consists of Chapter 7 and Chapter 11 filings, which is consistent with majority of previous research. *Screening result: 815.*

Finally, we excluded the financial sector because it could not be treated on equal terms with the other sectors. *Screening result: 753.*

However, in order to ensure that our company list contained all bankruptcy filings during the period 2003-2013; we enhanced our company list by collecting bankruptcy data from other sources as well:

In addition to screening, we searched for "bankruptcy" on BB. This resulted in a list of 541 companies filing for bankruptcy between 1995 and 2013. A notable commonality for the firms on this list is that the minimum total liabilities was 500 million U.S. dollars. We crosschecked these 541 companies with the 753 companies that we found by screening. Out of the 541 companies, there were 105 filings that were not included in our sample. We incorporated these companies in our list, which meant that we had 858 bankruptcy filings.

In addition, we also found a list of the 20 largest companies that filed for bankruptcy between 2003 and 2013 from Bankruptcydata.com[14], which totalled 220 companies over 11 years. Out of these companies, 80 firms were not included in our list. By adding these companies, we had a total of 938 bankruptcy filings in our list.

However, many of these companies were dropped due to incomplete financial statements (explained under 4.2 financial data). Our final sample included 562 companies that filed for bankruptcy between 2003 and 2013. The following table illustrates the distribution of filings.

[14] The industry's largest collection of corporate bankruptcy information – except for financial data.

Table II. Bankruptcy filings by year

| Filing year | Total failed | Total active | Failed to active firms | Failed in sample | Percentage in sample |
|---|---|---|---|---|---|
| 2003 | 146 | 9,856 | 1.48% | 91 | 16.2% |
| 2004 | 76 | 10,785 | 0.70% | 54 | 9.6% |
| 2005 | 63 | 11,719 | 0.54% | 42 | 7.5% |
| 2006 | 49 | 12,227 | 0.40% | 39 | 6.9% |
| 2007 | 65 | 12,726 | 0.51% | 44 | 7.8% |
| 2008 | 119 | 13,285 | 0.90% | 85 | 15.1% |
| 2009 | 186 | 13,924 | 1.34% | 74 | 13.2% |
| 2010 | 77 | 14,369 | 0.54% | 39 | 6.9% |
| 2011 | 78 | 15,049 | 0.52% | 37 | 6.6% |
| 2012 | 87 | 15,704 | 0.55% | 38 | 6.8% |
| 2013 | 67 | 16,087 | 0.42% | 19 | 3.4% |
| Total | 1,013 | 16,087 | 6.30% | 562 | 100.0% |

Notes: The table shows the number of bankruptcy filings during the time period in our sample (2003-2013). Total failed and active firms are the total number of filings registered on Bloomberg. Failed to active firms shows the percentage in each year. Failed in sample is the number of firms in our sample. The number of firms differs from the total, as we have only included the firms with available data. Percentage in sample shows how many percent of the filings were each year.

In our sample, most bankruptcy filings took place during 2003; then there is a conspicuous number of bankruptcies in 2008 and 2009. It is also noteworthy that the exclusion of the financial sector results in fewer bankruptcies during the financial crisis, as opposed to 2003. Except for the financial sector, our sample includes companies from all indices, sectors and states. Unlike previous research, the companies were grouped in sectors by "Bloomberg Industry Classification System" (BICS). However, the "Standard Industrial Classification" (SIC) code can also be used to classify the companies according to sectors. The following table illustrates the number of bankruptcy filings by sector.

**Table III. Bankruptcy frequency by sector**

| Sector | Total failed | Percentage |
|---|---|---|
| Communications | 62 | 11.0% |
| Consumer discretionary | 155 | 27.6% |
| Consumer staples | 22 | 3.9% |
| Energy | 51 | 9.1% |
| Health care | 68 | 12.1% |
| Industrials | 69 | 12.3% |
| Materials | 39 | 6.9% |
| Technology | 85 | 15.1% |
| Utilities | 11 | 2.0% |
| Total | 562 | 100.0% |

Notes: The table presents the bankruptcy filings during the time period 2003-2013 among the different sectors in our sample. Consumer discretionary has most bankruptcy filings, whereas, Utilities is the smallest sector.

Finally, we could assign the state of domicile and the state of incorporation to each of our companies. The data could not be incorporated intuitively in our models, but we wanted to check if there were any distinct features in the data.

**Table IV. Bankruptcy filings by state**

| State of domicile | | | State of incorporation | | |
|---|---|---|---|---|---|
| State | N | Percentage | State | N | Percentage |
| CA | 85 | 15.1% | DE | 391 | 69.6% |
| TX | 63 | 11.2% | NV | 43 | 7.7% |
| FL | 54 | 9.6% | FL | 13 | 2.3% |
| NY | 45 | 8.0% | TX | 13 | 2.3% |
| NJ | 26 | 4.6% | CO | 11 | 2.0% |
| MA | 25 | 4.4% | NY | 9 | 1.6% |
| IL | 22 | 3.9% | CA | 7 | 1.2% |
| MI | 21 | 3.7% | MN | 7 | 1.2% |
| OH | 18 | 3.2% | OH | 7 | 1.2% |
| CO | 17 | 3.0% | VA | 6 | 1.1% |
| Total | 376 | 66.9% | Total | 507 | 90.2% |

Notes: N is the number of bankruptcy filings. Only the ten states with most bankruptcies are included in this table. We can observe that total bankruptcies are 376 and 506, whereas our total sample size consists 562 filings.

As we can observe from Table IV, state of domicile does not have any specific pattern, compared to state of incorporation; almost 70% of the companies are incorporated in Delaware. Further investigation revealed that the state of Delaware is very favourable for firm incorporation (Black, 1999).[15]

## 4.2 Financial Data

Compustat on Wharton Research Data Services (WRDS) was primarily used to obtain the financial statements for each company. However, some of the companies on Compustat had missing data. Therefore, we also obtained financial data from BB. Although some companies had data, the data for years prior to bankruptcy was missing. Hence, we set some requirements for a firm to qualify in order to be included in our data set.

### 4.2.1 Requirements for Inclusion

First, companies need to have data for at least four consecutive years prior to the filing year because less than four years' data might cause misleading results in hazard rate model (Chava & Jarrow, 2004). Second, at most we included data for five years. This cut-off point was set because around 60-70% of the firms did not have data prior to the fifth year before bankruptcy. Also, too old data would likely have a negligible effect on the event of bankruptcy. Third, when the same company filed multiple bankruptcies, only one filing was included. In almost 5% of the cases, the same company had filed for bankruptcy more than once. We could include both bankruptcy filings in our data set by treating each filing as

[15] Reasons include: Delaware General Corporation Law – advanced and flexible corporation statutes; Court of Chancery – Delaware's court for corporations; Secretary of State's Office – thinks and acts like a corporation, rather than a government bureaucracy.

separate "bankruptcy" observation, or only include one filing for each company. The following table illustrates two cases of multiple bankruptcies:

**Table V. Multiple bankruptcies**

| | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Company A | | | | | FY | | FY | | | | |
| Data | A | A | A | A | #N/A | A | #N/A | | | | |
| Company B | | | | | FY | | | | | | FY |
| Data | | A | A | A | #N/A | | A | A | A | A | #N/A |

Notes: FY: Filing year; A: Available; #N/A: Not available.

Company A files for bankruptcy during 2004, and then again in 2006. In this case, we only include the filing in 2004 as the filing in 2006 violates the first criteria. Similarly, in the case of Company B, we include only the filing in 2010. In none of the cases with multiple bankruptcies were the financial statements coherent. Hence, we could not include more than one filing for each company, to stay consistent with requirement 1.

### 4.2.2 Control Group

Importantly BB was also used to obtain a sample of financially healthy companies.[16] The control group was matched with a 1:1 ratio, via the nearest neighbour method. Each healthy company was matched with a failed company in terms of sector (BICS-code) and total asset size. A failed company's total assets four years prior to the bankruptcy year were used to match with the total assets of the healthy company.[17] The fourth year asset size is used to match because at that time both companies can be considered "healthy". Matching companies with this procedure avoids over-fitting with failed companies, which could lead to biased results (Lennox, 1999).

[16] Healthy company: no filing for bankruptcy protection law during the timeframe.

[17] There is a maximum deviation of 8% between the total assets of a failed and healthy company.

### 4.2.3 Quality of the Data

To get valid and reliable results, it is important to have a clean data set. To increase the quality of the data set, we have thoroughly reviewed all collected data. Crosschecking the bankruptcy data revealed that some companies were listed twice under different tickers. Consequently, the "duplicate" companies were dropped. Before dropping companies with respect to financial data, we checked the financial statements for each company individually. Every ticker and company name was searched on both databases (BB and Compustat), before concluding that there was no available data.[18] Several previous researchers had set floors and ceilings in order for a company to qualify in the data set. Using such arbitrary criteria could result in biased estimates. Therefore, in this thesis, we have chosen to include all the companies to maintain variation in our data set. Additionally, there is no reason to exclude companies with lower values as size is one of the control variables in our models. Furthermore, the companies in the control sample were also crosschecked for not being repeated. Hence, we are confident in the quality of our data.

## 4.3 The Data Sets

Every company had its financial statements in its own Excel file. To be able to conduct the analysis in Stata, we had to construct a data set including all the financial statements in one file. Moreover, we used the financial statements to compute the necessary financial ratios (explained in section 5.1). The macroeconomic dependencies were also incorporated in the data set (explained in section 5.2). Since the methodology is different for the different models, we have constructed two different data sets.

[18] We tried to determine whether there was a pattern for the companies that had no historical data available. The noteworthy similarity of these firms was that approximately 80% of these companies were comparatively small. By small companies we mean companies that had total assets below 1 million U.S. dollars.

The first data set consists cross-sectional data - one observation for each firm. The included observation was the data one year prior to the filing year. This is necessary because there are typically no financial filings for the year in which a firm goes bankrupt. For instance, if a company went bankrupt in 2010, the data set included data for 2009. This data set consists of 562 failed firms and 562 non-failed firms, hence 1,124 firm-year observations. By contrast, the second data set is a panel containing all observations for each firm. Since we have either four or five years of data, the data set includes 5,524 firm-year observations.

## 4.4 In- and Out-of-Sample

The final data sets were subsequently split into two sub-samples. Both sub-samples included failed and non-failed companies. The first sub-sample included failed companies between 2003 and 2010, and is used as training data set to fit the model for out-of-sample prediction. The second sub-sample included failed companies between 2011 and 2013, and is used to evaluate the fitted model. To choose the latest observation period as out-of-sample prediction is consistent with the majority of previous research (Chava & Jarrow, 2004; Nam et al., 2008, Shumway, 2001).

**Table VI. Sample distribution**

| Data sample for Model I | Training data set | Evaluation data set | Total |
|---|---|---|---|
| Defaulted companies | 468 | 94 | 562 |
| Non-defaulted companies | 468 | 94 | 562 |
| Total | 936 (83%) | 188 (17%) | 1,124 |

| Data sample for Model II-V | Training data set | Evaluation data set | Total |
|---|---|---|---|
| Defaulted firm years | 468 | 94 | 562 |
| Non-defaulted firm years | 4,125 | 837 | 4,962 |
| Total | 4,593 (83%) | 931 (17%) | 5,524 |

Notes: Training data set: 2003-2010. Evaluation data set: 2011-2013.

# 5. Variable Selection

This section focuses on the selection of variables that are employed in our models. First, we have discussed the selection of financial ratios used. Second, we present the selection of various macroeconomic variables to capture the change in overall business surroundings.

## 5.1 Financial Ratios

We review the financial statements of each company to determine their financial structure. Previous research suggests that companies are more likely to fail if they are unprofitable, have high leverage and have cash flow difficulties. Hence, we identified all the financial ratios that have been examined in logit and hazard models in the mainstream literature.[19] Many of these ratios were an obvious transformation of other ratios. Hence, we performed a correlation analysis, to determine which ratios were highly correlated.[20] In cases with high correlation, the ratios with weak performance in the previous literature were dropped. Further, we estimated the variance inflation factor (VIF) of these ratios by running a linear regression. The employed ratios are used as independent variables in order to estimate the inflation of a coefficient that is caused due to linear dependence on other predictors. We remove those variables from the model that were causing an increase in the VIF. A VIF of 10 is considered as high inflation factor in this research, which is consistent with previous research (O'brien, 2007). Table VII presents the VIF of the employed ratios and Table VIII presents the correlation matrix of these ratios. An alternative approach would be to use Principal Component Analysis (PCA) to select variables with highest

[19] A table of all tested ratios in previous research is reproduced in Table A-II in the appendix. We also identified all the market variables tested in previous research but as the market data were not available, we could not test these in our models.

[20] Variables are considered highly correlated if the correlation coefficient is larger than 0.6.

explanatory power in an efficient way. We did not have time to implement this here, but may use it in future research.

**Table VII. Variance inflation factor of employed ratios**

| Variable | VIF | 1/VIF |
|---|---|---|
| R5 | 8.93 | 0.1120 |
| R5sq | 8.11 | 0.1232 |
| R11sq | 5.94 | 0.1683 |
| R11 | 5.92 | 0.1690 |
| R20 | 4.97 | 0.2013 |
| R20sq | 4.52 | 0.2214 |
| R18 | 2.36 | 0.4242 |
| Lag Unemployment rate | 1.71 | 0.5864 |
| R17 | 1.51 | 0.6608 |
| Lag Interest rate | 1.44 | 0.6962 |
| d2008_1 | 1.29 | 0.7751 |
| R1sq | 1.29 | 0.7781 |
| R25 | 1.15 | 0.8693 |
| R10 | 1.1 | 0.9080 |
| Mean VIF | 3.59 | |

Notes: R1: Cash flow from operations/Total liabilities; R5: Net income/Total assets; R10: Current liabilities/Total assets; R11: Total debt/Total assets; R17: Cash/Total assets; R18: Working capital/Total assets; R20: Current assets/Current liabilities; R25: ln (Total assets). The table provides the variance inflation factor (VIF) when employing different ratios using the linear regression model. The ratios are used as independent variables in order to estimate the inflation of a coefficient due to linear dependence on other predictors. We removed those variables from the model that were causing an increase in the VIF. A VIF of 10 is considered as a high inflation factor in this thesis, consistent with previous research.

The explanatory variables considered for our models are shown in Table IX, and divided into seven categories. The table also includes the macroeconomic variables, which will be explained in section 5.2.

**Table VIII. Correlation analysis of the employed ratios**

| | R1 | R5 | R10 | R11 | R17 | R18 | R20 | R25 | R1sq | R5sq | R11sq | R20sq | D2008_1 | lagUnemp | lagInterest |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| R1 | 1 | | | | | | | | | | | | | | |
| R5 | 0.31 | 1 | | | | | | | | | | | | | |
| R10 | 0.01 | -0.15 | 1 | | | | | | | | | | | | |
| R11 | 0.07 | -0.33 | 0.09 | 1 | | | | | | | | | | | |
| R17 | -0.41 | -0.17 | 0.08 | -0.12 | 1 | | | | | | | | | | |
| R18 | -0.03 | 0.64 | -0.23 | -0.50 | 0.08 | 1 | | | | | | | | | |
| R20 | -0.16 | 0.05 | -0.01 | -0.10 | 0.26 | 0.14 | 1 | | | | | | | | |
| R25 | 0.16 | 0.04 | 0.16 | 0.11 | -0.16 | -0.14 | -0.05 | 1 | | | | | | | |
| R1sq | -0.85 | -0.17 | -0.01 | -0.12 | 0.36 | 0.09 | 0.21 | -0.12 | 1 | | | | | | |
| R5sq | -0.16 | -0.93 | 0.18 | 0.30 | 0.11 | -0.64 | -0.04 | 0.05 | 0.09 | 1 | | | | | |
| R11sq | 0.01 | -0.38 | 0.12 | 0.90 | 0.02 | -0.52 | -0.06 | 0.05 | -0.04 | 0.36 | 1 | | | | |
| R20sq | -0.07 | -0.02 | 0.00 | -0.01 | 0.06 | 0.02 | 0.85 | 0.02 | 0.08 | 0.00 | -0.01 | 1 | | | |
| d2008_1 | 0.01 | -0.05 | 0.00 | 0.11 | -0.02 | -0.09 | 0.00 | 0.04 | -0.01 | 0.04 | 0.09 | 0.00 | 1 | | |
| lagUnemp | 0.03 | -0.03 | 0.01 | 0.06 | -0.02 | -0.05 | -0.01 | 0.02 | -0.04 | 0.02 | 0.04 | 0.01 | 0.40 | 1 | |
| lagInterest | 0.03 | 0.01 | 0.01 | 0.00 | -0.04 | 0.03 | -0.01 | 0.01 | -0.04 | -0.02 | -0.01 | 0.00 | 0.00 | 0.50 | 1 |

Notes: R1: Cash flow from operations/Total liabilities; R5: Net income/Total assets; R10: Current liabilities/Total assets; R11: Total debt/Total assets; R17: Cash/Total assets; R18: Working capital/Total assets; R20: Current assets/Current liabilities; R25: ln (Total assets). The table provide the correlation matrix of the employed ratios. As can be seen, all the ratios have correlation of less than 0.6, except for R5 and R18, which is 0.64. However, using the VIF, we can see that this is not causing any problem in the estimations.

**Table IX. Ratios tested in default prediction**

| Notation | Exp. signs | Variable definition | Origin |
|---|---|---|---|
| | | *Cash Flow* | |
| CFOTL | - | Cash flow from operations/Total liabilities | (Lennox, 1999) |
| CFOFE | - | Cash flow from operations/FE | (Zeitun & Tian, 2007) |
| | | *Profitability* | |
| NISALES | - | Net income/Sales | (Park & Han, 2002) |
| NITA | - | Net income/Total assets | (Zmijewski, 1984) |
| NITE | - | Net income/Total equity | (Park & Han, 2002) |
| NITL | - | Net income/Total liabilities | (Park & Han, 2002) |
| EBITTA | - | Earnings before interest & tax/Total assets | (Altman, 1968) |
| RETA | -/+ | Retained earnings/Total assets | (Altman, 1968) |
| | | *Leverage* | |
| CLTA | + | Current liabilities/Total assets | (Zmijewski, 1984) |
| EBITIE | - | Earnings before interest & taxes/Interest | Own |
| TDTA | + | Total debt/Total assets | (Zmijewski, 1984) |
| TDTE | + | Total debt/Total equity | (Zeitun & Tian, 2007) |
| TETA | -/+ | Total equity/Total assets | Own |
| | | *Size* | |
| TA | - | Total assets | (Park & Han, 2002) |
| Ln(TA) | - | Log of total assets | (Ohlson, 1980) |
| | | *Liquidity* | |
| CASHTA | - | Cash/Total assets | (Nam et al., 2008) |
| WCTA | - | Working capital/Total assets | (Altman, 1968) |
| | | *Solvency* | |
| QATA | - | Quick assets/Total assets | (Zmijewski, 1984) |
| CACL | - | Current assets/Current liabilities | (Zmijewski, 1984) |
| | | *Macroeconomic* | |
| INTEREST | + | Interest rate | (Hill et al., 2011) |
| UNEMLP | + | Unemployment rate | (Hill et al., 2011) |
| GDP | - | Gross domestic product | (Simons & Rowles, 2009) |
| CPI | - | Consumer price index | Own |
| SMR | -/+ | Stock market return | Own |

Notes: The notation will be used to identify the ratios in the outputs from Stata. Cash flow from operations = NI + Depreciation ± Change in WC; Working capital (WC) = Current assets – Current liabilities; Interest = Interest expenditure; Financial expenditures (FE) = Interest expenditure + Short-term debt; Quick asset = (Current assets – Inventories)/Current liabilities;. The macroeconomic variables are explained in section 5.2.

Cash flow: The net amount of cash and cash-equivalents moving in and out of a business is called the cash flow. Positive cash flow can be reinvested, used to pay debt, expenses, dividends to shareholders, or simply stored as a buffer for the future. A negative cash flow, however, implies that the liquid assets are decreasing (Casey & Bartczak, 1985).

Profitability: It is essential for a company to generate sufficient margin on its operations on a long-term basis, otherwise there is a high probability of the company facing financial difficulties (Pompe & Bilderbeek, 2005). Unprofitable companies also have lower going-concern value than profitable companies. Hence, they should be more likely to default (Myers, 1977). Stable profitability is not only vital to service the debt but also necessary to maintain the ability to obtain external finance. Hence, profitability can be considered as the driving factor for both liquidity and solidity. Persistent negative profits will reduce solidity in the long run, and liquidity in the short run.

Leverage: Companies often borrow capital to finance their investments and operations. Leverage increases risk and highly leveraged firms have a higher probability of default (Altman, 1968; Lennox, 1999; Ohlson, 1980; Shumway, 2001; Zeitun & Tian, 2007; Zmijewski, 1984)

Size: We have captured the size by log of total assets (Ohlson, 1980). However, one could also measure the size of a company by the number of employees (Lennox, 1999). With our limited resources, the latter variable was not available to us.

Liquidity: Liquidity is a measure of how quickly an asset or security can be sold without significant reduction in value, and cash is considered as the liquid asset. Companies usually get drained of their liquid assets prior to bankruptcy. Hence, they issue more short-term debt to fulfil their obligations. However, banks may tighten lending practices for financially distressed firms. Intuitively, we could assume that the more liquid a company, the more

likely it is to survive Liquidity ratios are found to be useful in default prediction (Altman, 1968; Nam et al., 2008).

Solvency: Solvency is a company's ability to meet its long-term obligations. If a company is not able to fulfil its long-term obligations, then it is considered insolvent. Furthermore, when a company is insolvent, it must enter bankruptcy. By definition, lower solvency implies a higher risk of default (Zmijewski, 1984).

## 5.2 Macroeconomic Variables

Theoretically, all macroeconomic variables might have an impact directly or indirectly on the hazard rate of a firm (Nam et al., 2008). Chen (2010) suggests that in times of recession, a firm expects its cash flows to have lower growth, be more volatile, and more correlated with the market. Besides there is higher risk at such times, which lowers the continuation value for shareholders and consequently increases the probability of default in recessions (Chen, 2010).

Unemployment rate: The unemployment rate is a leading macroeconomic indicator. Generally, the unemployment rate is higher in recession periods compared to non-recession periods. The figure illustrates the development of the unemployment and bankruptcy rates over time.

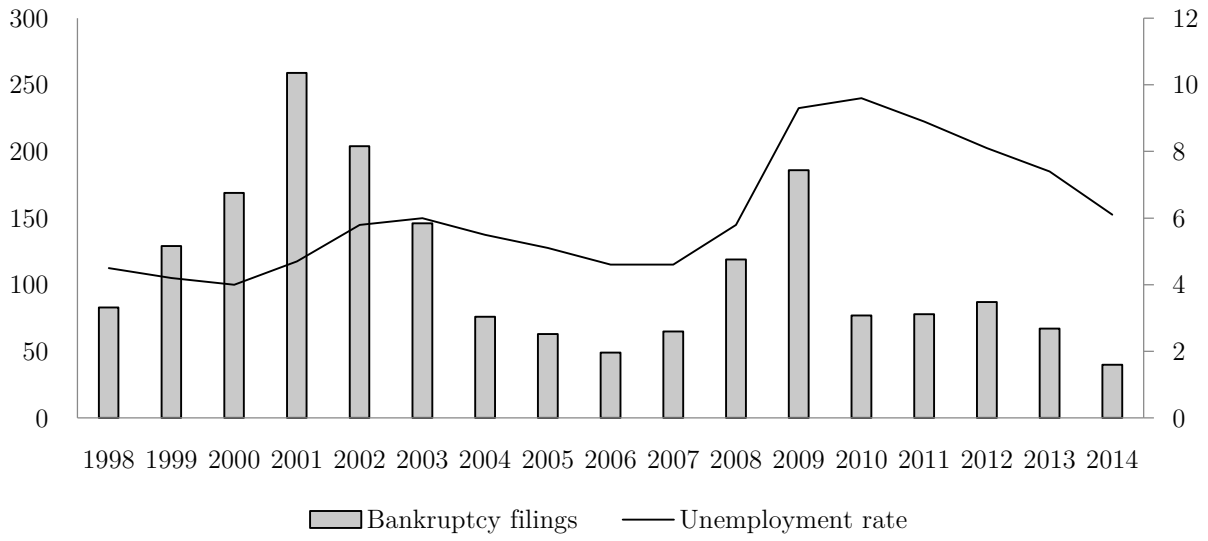


**Figure 1. Unemployment rate to bankruptcy filings by year.**
The figure shows total bankruptcy filings each year differ from our sample because the figure includes all bankruptcy filings from Bloomberg (not only the companies with available data). The financial sector is also taken into account.

The unemployment rate is slowly decreasing from 2003 till 2006. We then experienced an exogenous shock in 2008. The shock caused an unexpected increase in the unemployment rate from 4.6 (2007) to 5.8 (2008). We observe an even larger aftermath from 2008 to 2009, the unemployment rate increasing swiftly to 9.3 (2009). This can be regarded as a direct consequence of the financial crisis. The graph emphasizes that total bankruptcy filings and the unemployment rate each year are not necessarily accumulating equally. Nonetheless, we can observe a relationship; the unemployment rate is greater in or after years with many bankruptcy filings. Hence, the unemployment rate will be included with a one-year lag (Hill et al., 2011).

Interest rate: The interest rate will naturally affect all companies with debt. As the interest rate increases, the interest expenditures will also increase. High interest rates will affect the firm's borrowing ability, future cash flows, and its overall health. The interest rate a particular company gets is also determined by its default risk. We use the prime rate

as suggested by (Hill et al., 2011). The prime rate is the interest rate that commercial banks charge their most credit-worthy customers. We are aware of that most of the bankrupted companies will not be qualified to borrow at the prime rate. However, we use the prime rate as it directly reflects other lending rates and other aspects of the macro economy (such as liquidity).

Gross domestic product (GDP): Aggregated demand is reflected in GDP, and the sales of firms are related to aggregate demand. Higher GDP growth implies higher growth in firm revenues, whereas low GDP growth suggests lower growth in revenues. Hence, low GDP growth will make it harder for firms to generate income and the probability of default increases if firms struggle to generate sufficient income to fulfil their obligations (Simons & Rolwes, 2009).

Other variables: To the best of our knowledge, no other macroeconomic variables in previous research have been proved significant in estimating the probability of default. Nevertheless, we wanted to ensure that we have not omitted any variables that might have an impact. Hence, we used inflation and stock market returns as well to test for their impact in our models. Inflation is measured by the consumer price index (CPI). Whereas, stock market returns must be measured by a broad index. A common approach is to use the S&P 500. Since we have companies of all sizes, states and sectors, we gathered data for not only the S&P 500, but also NASDAQ and DOW JONES. We also computed the average value of the three indices to test for its significance in predicting default.

# 6. Empirical Results

In this section, we present and discuss the empirical findings from our analysis. Table X and Table XI provides the coefficients estimated using maximum likelihood from the static logit and the hazard rate models. The coefficients are estimated by fitting the model on in-sample observations between 2003 and 2010. The estimated coefficients are then used to predict the out-of-sample bankruptcies between 2011 and 2013.

We tested for the functional misspecification of the model using the procedure described in the methodology section 3.4.1. The idea of the test is that the quadratic terms of the predicted values will be insignificant, if the model is correctly specified. We also tested for the presence of heteroscedasticity by plotting standardized residuals against the fitted values and by using the White test procedure presented under methodology section 3.4.2.

Table A-III in the appendix reports the maximum likelihood estimates from the auxiliary logistic regression for functional misspecification. The quadratic term in the auxiliary logit regression for Model IA is significant, indicating that we have omitted relevant quadratic terms of the variables from the model. Therefore, we have added the relevant quadratic terms of covariates in the models. Further testing indicates that the square term of the auxiliary logistic regression is no longer significant. Similarly, we have tested for the functional misspecification for the hazard rate models after adding the quadratic terms of the covariates. Except for Model IA, the quadratic terms in the auxiliary logistic regression turn out to be insignificant.

The visual inspection of the scatter plots of standardized residuals against the fitted values is presented in Figure A-1 in the appendix, reflecting that the models might be exposed to heteroscedasticity. Therefore, we tested for the presence of heteroscedasticity using the White test procedure presented in section 3.4.2. Table A-IV in the appendix presents the results from the White test, reflecting that the models are exposed to

heteroscedasticity. Therefore, we used robust standard in the case of cross-sectional logit to correct for the heteroscedasticity. In addition, for the hazard rate models we have used the cluster robust standard errors, as it relaxes the assumption that the observations are not necessarily independent within clusters. The use of robust standard errors corrects the heteroscedasticity without changing the signs or magnitude of the coefficients. The results from the first three model specifications are presented in the following table:

Table X. Maximum likelihood estimates for Models I-III

| | Coefficients | | | |
|---|---|---|---|---|
| Variables | Model IA | Model IB | Model II | Model III |
| Cash flow from operation/Total liabilities | -0.383*** | -1.031*** | -0.631*** | -0.655*** |
| | (0.0872) | (0.239) | (0.149) | (0.151) |
| Net Income/Total assets | 0.110 | -0.680** | -0.408** | -0.425*** |
| | (0.0857) | (0.277) | (0.160) | (0.163) |
| Current liabilities/Total assets | | | 0.00102*** | 0.000935*** |
| | | | (0.000293) | (0.000315) |
| Total debt/Total assets | 0.473*** | 2.205*** | 1.128*** | 1.180*** |
| | (0.150) | (0.369) | (0.207) | (0.215) |
| Cash/Total assets | | | -1.222*** | -1.323*** |
| | | | (0.368) | (0.384) |
| Working capital/Total assets | -0.0903 | -0.313* | -0.130* | -0.156* |
| | (0.104) | (0.170) | (0.0740) | (0.0806) |
| Current assets/Current liabilities | -0.382*** | -0.265** | -0.198** | -0.186** |
| | (0.0545) | (0.117) | (0.0910) | (0.0896) |
| ln (Total assets) | -0.0203*** | -0.0161** | -0.0121*** | -0.0115*** |
| | (0.00599) | (0.00633) | (0.00401) | (0.00402) |
| (CFO/TL)^2 | | -0.148*** | -0.0884*** | -0.0887*** |
| | | (0.0528) | (0.0310) | (0.0310) |
| (Net Income/Total assets)^2 | | -0.116*** | -0.0703*** | -0.0752*** |
| | | (0.0413) | (0.0269) | (0.0268) |
| (Total debt/Total assets)^2 | | -0.581*** | -0.263*** | -0.285*** |
| | | (0.115) | (0.0648) | (0.0693) |
| (Current assets/Current liabilities)^2 | | 0.00188** | 0.000389** | 0.000364** |
| | | (0.000775) | (0.000180) | (0.000177) |
| d2008_1 | | | 1.891*** | 1.744*** |
| | | | (0.125) | (0.135) |
| Lag Unemployment | | | | 0.626*** |
| | | | | (0.0338) |
| Lag Interest | | | | 0.00672 |
| | | | | (0.0262) |
| Constant | 0.904** | -0.126 | -2.263*** | -5.261*** |
| | (0.439) | (0.504) | (0.177) | (0.275) |

*(Continued)*

**Table X. Maximum likelihood estimates for Models I-III** (*Continued*)

| | Coefficients | | | |
|---|---|---|---|---|
| Variables | Model IA | Model IB | Model II | Model III |
| Model Fit | 166.16 | 109.04 | 442.54 | 719.97 |
| Pseudo $R^2$ | 0.1281 | 0.2226 | 0.1492 | 0.2127 |
| Observations | 936 | 936 | 4,593 | 4,593 |
| Functional misspecification | Yes | No | No | No |
| Heteroscedasticity test | Yes | Used RSE | Used CRSE | Used CRSE |
| Macro-Variables | Yes | Yes | No | Yes |
| Non-linear forms | No | Yes | Yes | Yes |

Notes: For Model IB, the robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1. For Model II and Model III, the standard errors are clustered by firm and cluster robust standard errors (CRSE) are presented in parentheses. The static logit model is estimated using one observation for each firm one year prior to bankruptcy in the years 2003-2010. The hazard rate models are estimated using multiple observations for each firm prior to bankruptcy. Further, we incorporated the macroeconomic variables in the hazard rate models to test their significance in predicting default. The chi-square is presented in the model fit row. CFO/TL = Cash flow from operation/Total liabilities

## 6.1 Model I

The cross-sectional logit model is estimated using the observation of each firm one year prior to bankruptcy. Model IA is estimated without the inclusion of quadratic forms of the independent variables. The variables are identified by moving from a general model to a parsimonious model. To elaborate this point, we first included all the relevant independent variables in the model. Then, we stepwise removed the least significant variables from the model. The parsimonious model includes only six independent variables, which were found to be significant. One key point to note here, although the variables net income to total assets (NITA) and working capital to total assets (WCTA) are insignificant, these variables are still included in Model IA, because these are considered as important determinants in predicting default. NITA captures the capability of managers to efficiently utilize assets to

generate earnings. WCTA indicates the surplus of current assets over current liabilities as a proportion of total assets, reflecting the liquidity of the firm. However, further analysis of functional misspecification and heteroscedasticity indicates that Model IA is prone to omitted variable bias due to functional form misspecification and heteroscedasticity. Therefore, we have re-estimated Model IA by including the required quadratic terms of the independent variables, and by using robust standard errors. We have used the same procedure as explained above by approaching from a general model to a parsimonious model. However, prior to removing a variable, we tested for its joint significance with the quadratic term. For instance, if a variable, its quadratic term, or both are insignificant then we tested for their joint significance in the model. The re-estimation of Model IA generates Model IB. Further tests show that, after including the quadratic terms of independent variables, the model is no longer exposed to omitted variables arising from functional misspecification.[21]

The results from Model IB illustrate that the firm is more likely to go bankrupt if it is highly levered, as evidenced by the positive coefficient on total debt to total assets (TDTA). The positive coefficient on TDTA and the negative coefficient on its quadratic term means an increase in firm leverage increases the probability of default, but this effect diminishes as leverage increases. Although the existence of turning point is contrary to the conventional wisdom, few companies in the sample have a TDTA ratio of more than 1.9. This simply indicates that we cannot forecast the effect of TDTA for values more than 1.9 and an increase in TDTA always increases the probability of default in our data. For the lower values of TDTA, an increase in the leverage has a large positive effect on bankruptcy.

In contrast, the firm's probability of default decreases if the firm's cash flow to total liabilities and net income to total assets increases. The cash flow from operations to total liabilities (CFOTL) reflects the ability of a company to cover its short- and long-term

[21] Table A-III in the appendix reports the output for functional misspecification test.

liabilities (such as servicing its debt) by utilizing its cash flows from operations. By contrast, NITA provides an idea of efficiency of management in utilizing assets to generate earnings. The negative coefficients on both the linear and the quadratic term of these ratios indicates an exponential decline in probability of default.

The ratio of current assets to current liabilities (CACL) provides an overview of short-term solvency of a firm. The negative coefficient on the linear term indicates that the firm's probability of default decreases if it holds enough short-term assets to cover its short-term liabilities. The positive coefficient on its quadratic term leads us to estimate a turning point at 70, although only two firms in the sample have values greater than 70. This again indicates that we cannot really forecast the effect of CACL for the values more than 70 and an increase in CACL will always decrease the probability of default. WCTA provides insight about the liquidity of a company. The negative coefficient on WCTA reflects that the probability of firm going bankrupt decreases as it holds a higher proportion of WCTA.[22]

In addition, we have controlled for the effect of firm size. The coefficient on natural log of total assets (ln (TA)) shows that as firm size increases the probability of default falls. This reflects the fact that smaller firms are more prone to bankrupt than bigger firms.

## 6.2 Model II and Model III

We estimated Model II and Model III using a hazard rate model by taking into account multiple observations for each firm over time. Model II is estimated without inclusion of macroeconomic variables and Model III is estimated by including the macroeconomic variables in the hazard model. We get to a parsimonious model by undertaking the same approach as that of Model IB. The signs of the coefficients are the same as that of cross-

[22] Working capital = Current assets – Current liabilities

sectional logit and hence the interpretation of these coefficients is the same. In addition, the hazard model identified two more covariates as statistically significant in explaining default; current liabilities to total assets (CLTA) and cash to total assets (CASHTA). Further, we tested for the increased likelihood of bankruptcy in post-2008 by adding a dummy variable for all the firm-year observations after 2008. In contrast with the static logit model, the hazard model has identified post-2008 as statistically significant. The positive sign on the coefficient indicates that a firm is more likely to go bankrupt post-2008.

The output from the hazard models are reported in Table X. Most of the covariates in Model II and Model III are statistically significant at the 5% level except for WCTA, which is significant at 10% level. The positive sign on current liabilities to total assets (CLTA) indicates that the firm is more likely to default if current liabilities increases as a proportion of total assets. However, its economic impact is negligible. In addition, the coefficient on cash to total assets (CASHTA) indicates that a firm is less likely to go bankrupt if it holds a significant portion of cash and cash-equivalents proportional to total assets. Moreover, in economic terms, this coefficient is found to have the most significant impact on predicting default in our model. The coefficient on CFOTL is negative and significant indicating that a firm is less likely to bankrupt if it generates enough cash flow to cover its liabilities. Similarly, the coefficients on NITA, CASHTA, and WCTA indicate that the probability of default decreases if a firm efficiently employs its assets, holds enough cash as a portion of total assets, and has enough working capital relative to total assets. The coefficient on ln (TA) captures the fact that as size increases the probability of bankruptcy decreases, reflecting that smaller firms are more prone to default.

Additionally, we tested the significance of adding macroeconomic variables in the hazard rate model as a baseline. Unlike the static logit model, the lagged unemployment rate is statistically significant at 1% in the hazard rate model. Moreover, the magnitude of

this coefficient is economically significant improving our ability to predict default. This is consistent with the findings of Hill et al. (2011) and Shumway (2001), highlighting the weakness that the static logit model does not take into account the effect of macroeconomic variables that are same for all the firms.

## 6.3 Model IV and Model V

To investigate the effect of industry on bankruptcy prediction, we estimated Model IV and Model V with slope and intercept dummy variables for different sectors in our sample. The industry effects capture the impact of structural characteristics of the industry on the performance of a firm. Using this estimation is equivalent to estimating a separate hazard rate model for each industry in the sample (Allison, 2012; Chava & Jarrow, 2004). However, our approach offers an efficient way to "test" these 8 models in one overarching model.[23] If our model includes eight financial ratios then we need to generate 64 interaction terms[24] to capture the changes in slope coefficients. For this reason, most previous researchers have chosen to estimate the probability of bankruptcy in only one industry.

To incorporate the industry effects, we have generated the dummy variables for each sector included in our sample. For a particular sector, the dummy variable takes the value of 1 for firms in that sector and 0 otherwise. We use these dummies to estimate the change in intercept for the eight sectors in our sample. Furthermore, the slope coefficients for each industry are estimated by multiplying the industry dummy with each of the ratios. This generates eight slope variables for each of the ratios included in the model.

[23] "Energy" and "Consumer discretionary" are different from other sectors and therefore we group the interaction terms for these two sectors. ENECD is 1 if sector is Energy or Consumer discretionary.

[24] Eight ratios are multiplied by each sector; eight ratios × eight sectors = 64 interaction terms.

To estimate models with industry effects, we followed the same approach as previously to find a parsimonious model. For Model IV, we incorporated the industry-specific dummies and the interaction terms in Model II. The main idea is to test whether incorporating industry effects improves the predictive power of the hazard rate model. Further, we stepwise removed the least significant interaction variables from the model. The base group used in estimation of these variables is "Communication". Therefore, the explanation of these coefficients would be relative to Communication. The parsimonious model containing 13 interaction dummies found to be statistically significant and is presented in Table XI. In addition, the intercepts of ENECD, Consumer staples, Industrials, Materials, and Technology are positive and statistically significant.[25] This indicates that, relative to Communication, these sectors have more chance of going bankrupt and are relatively riskier. This is also consistent with the number of bankruptcy filings in Table III, as Consumer discretionary and Energy (ENECD) has the largest number of filings, followed by Technology and Industrials. The interaction dummies similarly indicate the sensitivities to the ratios over the reference group. For instance, the interaction dummy of Industrials with CACL indicates an increase in CACL would decrease the probability of bankruptcy over Communication, reflected by the value of -0.84 (-0.688 - 0.152). In addition, we used the likelihood ratio test to determine the significance of having industry variables in the model using the procedure discussed in section 3.5.5. The model with the intercept industry dummies is set as the unconstrained model, and the model without industry intercepts is set as the constrained model. We tested the null hypothesis that the coefficients on industry dummies are equal to zero. The entire set of industry variables is statistically significant using the LRT. This is evidence of needing industry effects in the model, which is consistent

[25] The full models with all interaction terms are in Table A-V in the appendix.

with the findings of Chava and Jarrow (2004). Moreover, the signs and magnitudes of coefficients are consistent with their findings.

In Model V, we incorporated both the macroeconomic variables and the industry effects in the hazard rate model to test for the improvement in the predictive power. By incorporating both of these factors, we are taking into account the changes in structural characteristics of the industry in which the firm operate and overall changes in the macroeconomic environment.

To estimate Model V, we incorporated the intercept and slope effects of industry dummies in Model III.[26] As in Model IV, we stepwise removed the least significant interaction variables from the model. The base group used for estimation of Model V is again "Communication". The parsimonious model containing 12 significant interaction dummies is reported in Table XI. The intercepts of ENECD, Consumer staples, Industrials, Materials, and Technology are found to be positive and significant. This indicates that these industries are riskier and have more chances of going bankrupt relative to Communication, which is also consistent with the number of bankruptcy filings in Table III. The statistically significant interaction dummies reflect the sensitivity to the ratios over the reference group. For instance, the interaction of ENECD with TDTA reflects the fact that an increase in TDTA would result in an increase in the default probability over Communication, reflected by the value of 0.998 (1.498 - 0.500).

Additionally, we incorporated two macroeconomic variables; lagged of unemployment and the lagged interest rate. To the best of our knowledge, this is the first research in the field to incorporate both the industry effects and the macroeconomic variables simultaneously. Lagged of unemployment is significant and positive at the 1% level, revealing that an increase in unemployment increases the probability of default for all firms.

[26] The full model with all industry dummies and interaction terms are attached in Table A-V in the appendix.

In addition, we tested for the significance of industry intercept dummies using the likelihood ratio test, following the procedure in section 3.5.5. The entire set of dummies is found to be statistically significant. The results for the LRT are presented in Table XII. In addition, we present the ceteris paribus effect of changes in covariates on the default probability. Table A-VI and Table A-VII in the appendix shows the marginal effects of change in the covariates from the mean on the default probability.

**Table XI. Maximum likelihood estimates for Models IV and Model V**

| | Coefficients | |
|---|---|---|
| Variables | Model IV | Model V |
| Cash flow from operation/Total liabilities | -0.685*** | -0.720*** |
| | (0.146) | (0.151) |
| Net income/Total assets | -0.332** | -0.358** |
| | (0.168) | (0.172) |
| Current liabilities/Total assets | 0.000757** | 0.000960*** |
| | (0.000315) | (0.000289) |
| Total debt/Total assets | 1.464*** | 1.498*** |
| | (0.235) | (0.244) |
| Cash/Total assets | -0.819** | -0.921** |
| | (0.404) | (0.439) |
| Working capital/Total assets | -0.116 | -0.213*** |
| | (0.0757) | (0.0773) |
| Current assets/Current liabilities | -0.152* | -0.140 |
| | (0.0889) | (0.0895) |
| Ln (Total assets) | -0.0141*** | -0.0121*** |
| | (0.00442) | (0.00427) |
| ENECD | 0.685*** | 0.821*** |
| | (0.219) | (0.238) |
| Consumer staples | 1.176** | 1.364** |
| | (0.515) | (0.563) |
| Health care | 0.130 | 0.200 |
| | (0.199) | (0.218) |
| Industrials | 1.265*** | 1.405*** |
| | (0.375) | (0.373) |
| Materials | 1.394*** | 1.299*** |
| | (0.452) | (0.460) |
| Technology | 0.597*** | 0.698*** |
| | (0.215) | (0.233) |
| Utilities | -0.430 | -0.358 |
| | (0.605) | (0.594) |
| Current liabilities/Total assets × ENECD | 0.0223** | 0.0144* |
| | (0.00895) | (0.00869) |
| Total debt/Total assets × ENECD | -0.432** | -0.500** |
| | (0.183) | (0.197) |
| Cash/Total assets × ENECD | -2.555*** | -2.700*** |

(*Continued*)

Table XI. Maximum likelihood estimates for Models IV and Model V (*Continued*)

| | Coefficients | |
|---|---|---|
| Variables | Model IV | Model V |
| | (0.921) | (0.963) |
| Current liabilities/Total assets × Consumer staples | -2.367* | -2.546* |
| | (1.221) | (1.309) |
| Working capital/Total assets × Consumer staples | -2.503** | -2.533* |
| | (1.193) | (1.297) |
| Working capital/Total assets × Industrials | 0.389* | 0.468** |
| | (0.208) | (0.187) |
| Current assets/Current liabilities × Industrials | -0.688*** | -0.679*** |
| | (0.232) | (0.219) |
| Cash flow from operations/Total liabilities × Materials | 0.749** | 0.813*** |
| | (0.335) | (0.290) |
| Total debt/Total assets × Materials | -0.560** | -0.602** |
| | (0.270) | (0.245) |
| Current assets/current liabilities × Materials | -0.527** | -0.442* |
| | (0.246) | (0.231) |
| Total debt/Total assets × Technology | -0.944*** | -0.727*** |
| | (0.265) | (0.232) |
| Working capital/Total assets × Technology | -0.277** | |
| | (0.122) | |
| Current liabilities/Total assets × Utilities | 3.827** | 4.502*** |
| | (1.786) | (1.719) |
| (Cash flow from operation/Total liabilities)^2 | -0.0859*** | -0.0881*** |
| | (0.0264) | (0.0266) |
| (Net income/Total assets)^2 | -0.0641** | -0.0717** |
| | (0.0279) | (0.0285) |
| (Total debt/Total assets)^2 | -0.265*** | -0.280*** |
| | (0.0660) | (0.0694) |
| (Current assets/Current liabilities)^2 | 0.000303* | 0.000278 |
| | (0.000176) | (0.000177) |
| Lag Unemployment | | 0.642*** |
| | | (0.0462) |
| d2008_1 | 1.928*** | 1.786*** |
| | (0.132) | (0.141) |
| Constant | -2.760*** | -5.908*** |
| | (0.245) | (0.359) |

(*Continued*)

**Table XI. Maximum likelihood estimates for Models IV and Model V (*Continued*)**

| | Coefficients | |
|---|---|---|
| Variables | Model IV | Model V |
| Model Fit | 490.15 | 583.91 |
| Pseudo $R^2$ | 0.1686 | 0.2313 |
| Observations | 4,593 | 4,593 |
| Functional misspecification | No | No |
| Heteroscedasticity test | Used CRSE | Used CRSE |
| Macroeconomic Variables | No | Yes |
| Non-linear forms | Yes | Yes |
| Industry effect | Yes | Yes |

Notes: Cluster robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1. The hazard rate models are estimated using the data from 2003 to 2010 for fitting the model. Model IV (industry effects) and Model V (industry effects and macroeconomic variables). ENECD is 1 if sector is Energy or Consumer discretionary.

**Table XII. Likelihood ratio test**

| | Model IV | Model V |
|---|---|---|
| Unrestricted model - 2LOG(LF) | 490.15 | 583.91 |
| Restricted model - 2LOG(LF) | 481.31 | 567.75 |
| Chi-Square | 24.46 | 26.96 |
| *P*-Value | 0.0009 | 0.0003 |

Notes: To test if industry dummies are significant, a likelihood ratio test is conducted. The unrestricted model includes the industry effects, whereas the restricted model does not include the industry effects. Under null hypothesis there is no significance of industry effects.

Table XII presents the likelihood ratio test to gauge the significance of having industry dummies in the models. The model with industry effects is set as unconstrained model and it is tested against the model without industry effects.

## 6.4 Summary of Results

To estimate the probability of default, we have analysed five models and compared their out-of-sample predictive ability. The static logit model is fitted using the cross-sectional observations for each firm, one year prior to bankruptcy. The remaining models are estimated using time-varying covariates of each firm in hazard rate models. We have further refined these models, by adding macroeconomic variables and industry effects, first separately and then together.

The cross-sectional logit suggests that six variables are significant in predicting default probability. The signs and magnitudes of estimated coefficients of these variables are consistent with previous research. The negative signs on CFOTL, NITA, WCTA, CACL, and ln (TA) shows that an increase in these variables would lead to decrease in probability of bankruptcy. The significance of quadratic terms in the model indicates that the path to bankruptcy is non-linear, which is consistent with the findings of Lennox (1999).

In Model II, eight covariates are found significant in predicting probability of default, consistent with research by Chava & Jarrow (2004) and Shumway (2001). In addition to the coefficients estimated in static logit model, the hazard rate model identified two other coefficients that are statistically significant, the CLTA and CASHTA. The coefficient CASHTA indicates that the probability of bankruptcy decreases as the firm has more cash & cash equivalents relative to total assets. In addition, the hazard rate model identifies that post-2008 the probability of bankruptcy is higher.

In Model III, we added variables to capture changes in the macroeconomic environment. The lag of unemployment is found both statistically and economically significant, consistent with Hill et al. (2011).

In Model IV, we incorporated industry effects on both the slope and intercept in the hazard rate model. In total, 13 interaction terms are statistically significant mostly for

Energy and Consumer discretionary sectors (ENECD). We used the LRT to gauge the significance of these variables. The model containing the industry dummies is set as unconstrained model and it is tested against the model without inclusion of industry dummies. Under the null hypothesis, there is no significant effect of adding these to the model. As can be seen from Table XII, the industry dummies as a whole are statistically significant.

In Model V, we incorporated both the macroeconomic variables and the industry effects in the model to test whether this significantly outperforms the other models in predicting default out-of-sample. The significance of industry intercept dummies is estimated using the same approach as discussed in Model IV by using the LRT. The industry intercept dummies are found as jointly significant. Moreover, the lag of unemployment is also significant in this model. This indicates that Model V would outperform other models as it captures idiosyncratic variations, industry specific variation and the changes in the macroeconomic conditions.

# 7. Model Evaluation

In this section, we evaluate the out-of-sample predictive power of the five models for the validation period 2011 to 2013, using the measures discussed in section 3.5. The coefficients of the fitted models are used to predict defaults for these years. Basically, these are validation tests to evaluate how accurately the models are able to differentiate between "true" and "false" defaults in practice. Several previous researchers have highlighted weak out-of-sample performance of models in predicting default. Hence, we conducted different tests to evaluate the performance of our models.

## 7.1 Classification Matrix

The classification matrix provides the accuracy of the model, at a given threshold level, in predicting out-of-sample defaults along with sensitivity, specificity and the two types of embedded uncertainties in out-of-sample prediction. Recall that sensitivity is the true classification of default firm when the firm actually defaulted; and specificity is the true classification of non-defaulted firms as non-defaulted. The two types of uncertainty are Type I and Type II errors. The Type I error is the classification of a bankrupt firm as non-bankrupt, and the Type II error is the classification of non-bankrupt firm as bankrupt. The optimal threshold point is where we have the highest sensitivity and specificity. However, an investor or lender can decide the optimal level based on his risk preference.

Table XIII reports sensitivity analysis for the classification matrix using three threshold points around the intersection of sensitivity and specificity, which is considered the optimal threshold level. We found the intersection by plotting sensitivity against specificity as shown in Figure A-2 in the appendix. The graph plots the sensitivity and specificity for all possible cut-off probabilities, it can be seen that specificity decreases as sensitivity increases.

From Table XIII, the static logit (Model IB) correctly classifies 74.47% of the out-of-sample bankruptcies at the threshold level of 0.4, which is also its maximum percentage of correct classifications. The threshold level of 0.4 indicates that the model is good at differentiating between the defaulted and non-defaulted firms. However, the sensitivity and specificity of the model changes significantly upon changing the threshold. The sensitivity decreases from 90.43% to 59.57% when the threshold increases from 0.4 to 0.6, which increases the Type I error from 10% to 40%. By contrast, the specificity increases at a relatively similar percentage rate, indicating a reduction in Type II errors.

Model II and Model IV are inferior to the static logit model in terms of sensitivity. The optimal threshold level for these models is around 0.3. However, the correct classification of the models at this level is only 57.57% and 58.54%, respectively. In addition, the sensitivity of these models decreases at a significantly higher rate than Model IB and thus increases Type I error. In terms of specificity, Model II and Model IV are able to classify around the same percentage of non-bankrupted firms as Model IB. This indicates that both of these models are worse in terms of predicting the true defaults but are comparable to cross-sectional logit in terms of predicting true non-default.

Model III and Model V, with inclusion of macroeconomic variables, are the best performing models. The optimal threshold level for these models are around 0.87 and 0.88, respectively. The higher threshold level reflects the fact that these models are superior in discriminating between the defaulted and non-defaulted firms. Further, at the optimal threshold level the correct classification for these two models are 83.57% and 82.81%. In addition, Model III (which includes only macroeconomic variables) has sensitivity of 92.55% and specificity of 82.56% indicating that this model has the lowest percentage of Type I and Type II errors.

The overall results of the classification matrix indicate that Model III and Model V are the best performing models. These models outperform the rest of the models estimated in terms of correct classification, sensitivity, and specificity. In addition, the Type I error of Model III at the optimal threshold is lower than Nam et al. (2008) reported (13.89%); however, the Type I error of Model V is slightly higher. Both Model III and Model V outperform Nam et al. (2008) in terms of the overall percentage of correct classifications. They reported an overall classification of 80.55% for the model with a macroeconomic baseline. In addition, the overall classification of the static logit model is also higher than they reported, which is 72.22%. However, they estimated the Type I error for their cross-sectional logit model at 8.33%, which is slightly lower than our estimate of 9.57%.

# Table XIII. Classification matrix

**Model IB (Static Logit Model)**

| | | Observed | | |
|---|---|---|---|---|
| | | Bankrupt | Non-bankrupt | Total |
| Classified | Bankrupt | 85 | 39 | 124 |
| | Non-bankrupt | 9 | 55 | 64 |
| | Total | 94 | 94 | 188 |

| | | |
|---|---|---|
| Correclty classified default if predicted Pr(D) >= 0.4 | | 74.47% |
| Sensitivity | Pr( +\| D) | 90.43% |
| Specificity | Pr( -\|~D) | 58.51% |
| False - rate for true D (Type I error) | Pr( -\| D) | 9.57% |
| False + rate for true ~D (Type II error) | Pr( +\|~D) | 41.49% |

| | | Observed | | |
|---|---|---|---|---|
| | | Bankrupt | Non-bankrupt | Total |
| Classified | Bankrupt | 56 | 13 | 69 |
| | Non-bankrupt | 38 | 81 | 119 |
| | Total | 94 | 94 | 188 |

| | | |
|---|---|---|
| Classified default if predicted Pr(D) >= 0.6 | | 72.87% |
| Sensitivity | Pr( +\| D) | 59.57% |
| Specificity | Pr( -\|~D) | 86.17% |
| False - rate for true D (Type I error) | Pr( -\| D) | 40.43% |
| False + rate for true ~D (Type II error) | Pr( +\|~D) | 13.83% |

| | | Observed | | |
|---|---|---|---|---|
| | | Bankrupt | Non-bankrupt | Total |
| Classified | Bankrupt | 26 | 6 | 32 |
| | Non-bankrupt | 68 | 88 | 156 |
| | Total | 94 | 94 | 188 |

| | | |
|---|---|---|
| Correctly classified default if predicted Pr(D) >= 0.8 | | 60.64% |
| Sensitivity | Pr( +\| D) | 27.66% |
| Specificity | Pr( -\|~D) | 93.62% |
| False - rate for true D (Type I error) | Pr( -\| D) | 72.34% |
| False + rate for true ~D (Type II error) | Pr( +\|~D) | 6.38% |

**Model II (Hazard Model)**

| | | Observed | | |
|---|---|---|---|---|
| | | Bankrupt | Non-bankrupt | Total |
| Classified | Bankrupt | 88 | 389 | 477 |
| | Non-bankrupt | 6 | 448 | 454 |
| | Total | 94 | 837 | 931 |

| | | |
|---|---|---|
| Correctly classified default if predicted Pr(D) >= 0.3 | | 57.57% |
| Sensitivity | Pr( +\| D) | 93.62% |
| Specificity | Pr( -\|~D) | 53.52% |
| False - rate for true D (Type I error) | Pr( -\| D) | 6.38% |
| False + rate for true ~D (Type II error) | Pr( +\|~D) | 46.48% |

| | | Observed | | |
|---|---|---|---|---|
| | | Bankrupt | Non-bankrupt | Total |
| Classified | Bankrupt | 40 | 112 | 152 |
| | Non-bankrupt | 54 | 725 | 779 |
| | Total | 94 | 837 | 931 |

| | | |
|---|---|---|
| Correctly classified default if predicted Pr(D) >= 0.5 | | 82.17% |
| Sensitivity | Pr( +\| D) | 42.55% |
| Specificity | Pr( -\|~D) | 86.62% |
| False - rate for true D (Type I error) | Pr( -\| D) | 57.45% |
| False + rate for true ~D (Type II error) | Pr( +\|~D) | 13.38% |

| | | Observed | | |
|---|---|---|---|---|
| | | Bankrupt | Non-bankrupt | Total |
| Classified | Bankrupt | 7 | 23 | 30 |
| | Non-bankrupt | 87 | 814 | 901 |
| | Total | 94 | 837 | 931 |

| | | |
|---|---|---|
| Correctly classified default if predicted Pr(D) >= 0.7 | | 88.18% |
| Sensitivity | Pr( +\| D) | 7.45% |
| Specificity | Pr( -\|~D) | 97.25% |
| False - rate for true D (Type I error) | Pr( -\| D) | 92.55% |
| False + rate for true ~D (Type II error) | Pr( +\|~D) | 2.75% |

**Model III (Hazard Model with macroeconomic variables)**

| | | Observed | | |
|---|---|---|---|---|
| | | Bankrupt | Non-bankrupt | Total |
| Classified | Bankrupt | 94 | 231 | 325 |
| | Non-bankrupt | 0 | 606 | 606 |
| | Total | 94 | 837 | 931 |

| | | |
|---|---|---|
| Correctly classified default if predicted Pr(D) >= 0.7 | | 75.19% |
| Sensitivity | Pr( +\| D) | 100.00% |
| Specificity | Pr( -\|~D) | 72.40% |
| False - rate for true D (Type I error) | Pr( -\| D) | 0.00% |
| False + rate for true ~D (Type II error) | Pr( +\|~D) | 27.60% |

| | | Observed | | |
|---|---|---|---|---|
| | | Bankrupt | Non-bankrupt | Total |
| Classified | Bankrupt | 93 | 194 | 287 |
| | Non-bankrupt | 1 | 643 | 644 |
| | Total | 94 | 837 | 931 |

| | | |
|---|---|---|
| Correctly classified default if predicted Pr(D) >= 0.8 | | 79.05% |
| Sensitivity | Pr( +\| D) | 98.94% |
| Specificity | Pr( -\|~D) | 76.82% |
| False - rate for true D (Type I error) | Pr( -\| D) | 1.06% |
| False + rate for true ~D (Type II error) | Pr( +\|~D) | 23.18% |

| | | Observed | | |
|---|---|---|---|---|
| | | Bankrupt | Non-bankrupt | Total |
| Classified | Bankrupt | 87 | 146 | 30 |
| | Non-bankrupt | 7 | 691 | 901 |
| | Total | 94 | 837 | 931 |

| | | |
|---|---|---|
| Correctly classified default if predicted Pr(D) >= 0.87 | | 83.57% |
| Sensitivity | Pr( +\| D) | 92.55% |
| Specificity | Pr( -\|~D) | 82.56% |
| False - rate for true D (Type I error) | Pr( -\| D) | 7.45% |
| False + rate for true ~D (Type II error) | Pr( +\|~D) | 17.44% |

(*Continued*)

## Table XIII. Classification matrix *(Continued)*

**Model IV (Hazard Model with Industry effects)**

| | | Observed | | |
|---|---|---|---|---|
| | | Bankrupt | Non-bankrupt | Total |
| Classified | Bankrupt | 81 | 373 | 454 |
| | Non-bankrupt | 13 | 464 | 477 |
| | Total | 94 | 837 | 931 |

| | | |
|---|---|---|
| Correctly classified default if predicted Pr(D) >= 0.3 | | 58.54% |
| Sensitivity | Pr( +\| D) | 86.17% |
| Specificity | Pr( -\|~D) | 55.44% |
| False - rate for true D (Type I error) | Pr( -\| D) | 13.83% |
| False + rate for true ~D (Type II error) | Pr( +\|~D) | 44.56% |

| | | Observed | | |
|---|---|---|---|---|
| | | Bankrupt | Non-bankrupt | Total |
| Classified | Bankrupt | 62 | 203 | 265 |
| | Non-bankrupt | 32 | 634 | 666 |
| | Total | 94 | 837 | 931 |

| | | |
|---|---|---|
| Correctly classified default if predicted Pr(D) >= 0.4 | | 74.76% |
| Sensitivity | Pr( +\| D) | 65.96% |
| Specificity | Pr( -\|~D) | 75.75% |
| False - rate for true D (Type I error) | Pr( -\| D) | 34.04% |
| False + rate for true ~D (Type II error) | Pr( +\|~D) | 24.25% |

| | | Observed | | |
|---|---|---|---|---|
| | | Bankrupt | Non-bankrupt | Total |
| Classified | Bankrupt | 45 | 112 | 157 |
| | Non-bankrupt | 49 | 725 | 774 |
| | Total | 94 | 837 | 931 |

| | | |
|---|---|---|
| Correctly classified default if predicted Pr(D) >= 0.5 | | 82.71% |
| Sensitivity | Pr( +\| D) | 47.87% |
| Specificity | Pr( -\|~D) | 86.62% |
| False - rate for true D (Type I error) | Pr( -\| D) | 52.13% |
| False + rate for true ~D (Type II error) | Pr( +\|~D) | 13.38% |

**Model V (Hazard Model with macroeconomic variables and Industry effects)**

| | | Observed | | |
|---|---|---|---|---|
| | | Bankrupt | Non-bankrupt | Total |
| Classified | Bankrupt | 90 | 240 | 330 |
| | Non-bankrupt | 4 | 597 | 601 |
| | Total | 94 | 837 | 931 |

| | | |
|---|---|---|
| Correctly classified default if predicted Pr(D) >= 0.7 | | 73.79% |
| Sensitivity | Pr( +\| D) | 95.74% |
| Specificity | Pr( -\|~D) | 71.33% |
| False - rate for true D (Type I error) | Pr( -\| D) | 4.26% |
| False + rate for true ~D (Type II error) | Pr( +\|~D) | 28.67% |

| | | Observed | | |
|---|---|---|---|---|
| | | Bankrupt | Non-bankrupt | Total |
| Classified | Bankrupt | 89 | 200 | 289 |
| | Non-bankrupt | 5 | 637 | 642 |
| | Total | 94 | 837 | 931 |

| | | |
|---|---|---|
| Correctly classified default if predicted Pr(D) >= 0.8 | | 77.98% |
| Sensitivity | Pr( +\| D) | 94.68% |
| Specificity | Pr( -\|~D) | 76.11% |
| False - rate for true D (Type I error) | Pr( -\| D) | 5.32% |
| False + rate for true ~D (Type II error) | Pr( +\|~D) | 23.89% |

| | | Observed | | |
|---|---|---|---|---|
| | | Bankrupt | Non-bankrupt | Total |
| Classified | Bankrupt | 76 | 142 | 218 |
| | Non-bankrupt | 18 | 695 | 713 |
| | Total | 94 | 837 | 931 |

| | | |
|---|---|---|
| Correctly classified default if predicted Pr(D) >= 0.88 | | 82.81% |
| Sensitivity | Pr( +\| D) | 80.85% |
| Specificity | Pr( -\|~D) | 83.03% |
| False - rate for true D (Type I error) | Pr( -\| D) | 19.15% |
| False + rate for true ~D (Type II error) | Pr( +\|~D) | 16.97% |

Notes: The classification matrix provides the accuracy of the model, at a given threshold level, in predicting out-of-sample defaults along with sensitivity, specificity and the two types of embedded uncertainties in out-of-sample prediction. We used three different threshold levels for each model in order to test the change in overall classification, Type I, and Type II errors. Sensitivity is the true classification of actual defaulted firms and specificity is the classification of true non-defaulted firms.

## 7.2 Goodness-of-fit Deciles

In order to estimate the out-of-sample performance of the models we also conducted bankruptcy prediction test using deciles. This test is same as the test employed by Chava & Jarrow (2004), Nam et al. (2008) and Shumway (2001) in order to validate out-of-sample prediction. The main idea behind this test is to use the coefficients estimated from the testing sample to predict bankruptcies for the out-of-sample period. As mentioned earlier, the models are estimated using the data from 2003 to 2010 in order to forecast the out-of-sample bankruptcies between 2011-2013. The probabilities are estimated for each year and then the companies are grouped into different deciles based on probabilities of default. Further, we aggregated the number of bankruptcies in each decile for each of the three years, as reported in Table XIV.

Based on forecasting ability, Model V (with inclusion of macroeconomic and industry specific variables along with the other covariates) is superior as it correctly identified around 50% of the defaults in the first decile. This is followed by Model III (46%), Model IV (31%), and Model II (25%). The static logit model is inferior to all the hazard rate models as it allocates only 15% of correctly identified bankruptcies in the top decile. Nam et al. (2008) accumulated the number of bankruptcies in the top two deciles. Employing the same procedure, Model III (with inclusion of only macroeconomic variables along with the other covariates) is the superior performer, with around 78% of bankruptcies in the first two deciles. This is followed by Model V (76%), Model IV (53%), and Model II (47%). The forecasting estimates of Model III and Model V are superior in comparison with estimates provided by Nam et al. (2008). They reported in the top 2 deciles a total of 70% of bankruptcies being correctly specified. Moreover, our estimates are close to the estimates provided by Chava & Jarrow (2004). They estimated out-of-sample forecasts of around 84% in top 2 deciles.

Table XIV. Out-of-sample forecasting accuracy

| Decile | Static Logit | Hazard rate models | | | |
|---|---|---|---|---|---|
| | Model IB | Model II | Model III | Model IV | Model V |
| 1 | 14 | 24 | 44 | 29 | 46 |
| 2 | 17 | 20 | 29 | 20 | 25 |
| 3 | 17 | 17 | 20 | 16 | 18 |
| 4 | 13 | 14 | 1 | 10 | 2 |
| 5 | 10 | 11 | 0 | 6 | 2 |
| 6 | 8 | 6 | 0 | 6 | 0 |
| 7 | 9 | 2 | 0 | 5 | 0 |
| 8 | 5 | 0 | 0 | 1 | 1 |
| 9 | 1 | 0 | 0 | 1 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 |
| Total | 94 | 94 | 94 | 94 | 94 |

Notes: The table presents the forecasting accuracy over 2011-2013 by using the fitted model over the period 2003-2010.

## 7.3 Area under ROC Curve

Area under ROC (AUROC) compares the sensitivity (true default) of the model to 1 – specificity (false default). AUROC is considered a more complete description of accuracy (Hosmer Jr et al., 2013). The AUROC ranges from 0.5 to 1 and an AUROC close to one is considered optimal; an AUROC close to 0.5 suggests that a model has no discrimination ability between two outcomes. Table XV reports the results of AUROC and Figure A-3 in the appendix reports the ROC curves for all five estimated models. Model III (with just the macroeconomic variables) is the best performing model and is followed by Model V. The AUROC close to 1 for these models indicates their superior ability. These estimated results are similar to the findings of Chava & Jarrow (2004). The AUROC for their best performing model is 0.9449. Although the industry effects are found to be statistically significant in-sample, it appears that they do not significantly increase out-of-sample accuracy.

**Table XV. AUROC for all five models**

| Model | AUROC |
|---|---|
| Model IB | 0.8324 |
| Model II | 0.7896 |
| Model III | 0.9188 |
| Model IV | 0.7854 |
| Model V | 0.9051 |

Notes: The table present the Area under ROC (AUROC) for each model. The AUROC close to 1 for the models indicates superior ability in classification.

## 7.4 Brier Score

The Brier Score (BS) is a commonly used measure for evaluating probabilistic forecasts (Roulston, 2007). It is an aggregate measure of the disagreement between the predicted and observed outcomes. The BS tests for the calibration of the model as well as the discrimination ability of the model between the two outcomes. Lower scores reflect superior performance of the model. Table XVI presents the results obtained from the Brier Score for the five models.

**Table XVI. Brier Score for all five models**

| Models | Brier Score |
|---|---|
| Model IB | 0.1720 |
| Model II | 0.0802 |
| Model III | 0.0770 |
| Model IV | 0.0786 |
| Model V | 0.0749 |

Notes: Lower score reflects superior performance in terms of both discrimination ability and the calibration of the models. Model V with macro variables and industry effects performs best among all models.

Model V (with both industry effects and macroeconomic variables) performs best, followed by Model III (with just macroeconomic variables). The static logit model has the highest score among all the models, reflecting the inferiority of the model. Although ranked higher than Model II and Model IV when testing for discrimination ability, the model performs worse than all the employed hazard rate models when accounting for discrimination and calibration. This is due to the fact that hazard rate models assign probabilities to the outcome close to zero if $y_i = 0$ and close to one for the outcomes $y_i = 1$. Alternatively, the Type II error of the static logit model on the basis of BS is significantly higher than the hazard rate models which drives down the BS close to zero for the hazard rate models.

# 8. Conclusion

Higher rates of business failure over recent years emphasizes the importance of credit risk management. Therefore, the main aim of this research has been to develop reliable default prediction models by testing the out-of-sample forecasting accuracy of a static logit model and four hazard rate models. In addition, we validated the out-of-sample forecasting accuracy of the hazard models with macroeconomic variables and industry effects over the static logit model using these recent data.

We studied macroeconomic, industry-specific, and idiosyncratic determinants of corporate failures in a sample of 562 bankrupt firms, and 562 non-bankrupt firms, over the period 2003 to 2013. The sub-sample of failed firms between 2011 and 2013 is used for out-of-sample evaluation. Both investors and lenders can benefit from the findings of this research. The investors can avoid investing in firms with high probability of default, whereas the lenders can ensure that their lending practices conform to the credit risk.

First, we find that hazard rate models (without macroeconomic covariates as a baseline) perform as well as the static logit model in terms of the classification matrix and AUROC. However, all of the hazard rate models significantly outperform the static logit model in terms of allocation of bankrupt firms in top decile and Brier Score. Second, we demonstrate the performance improvement of hazard rate model by employing industry-specific and macroeconomic variables. Consistent with the findings of Hill et al. (2011), Nam et al. (2008) and Shumway (2001), the hazard rate models with macroeconomic variables significantly outperform other models. In terms of overall discrimination ability and calibration, Model V (including both macroeconomic and industry effects) is the best performing model, consistent with the findings of Chava & Jarrow (2004).

The main reason for the lacklustre performance of the static logit model is its inability to account for the change in firm characteristics over time and in

macroeconomic dependencies. Our results indicate that ignoring changes in firm characteristics over time reduces the out-of-sample performance of the static logit model, consistent with the findings of Chava and Jarrow (2004), Hill et al. (2011), Nam et al. (2008), and Shumway (2001). Within the hazard rate models, we find that the macroeconomic variables significantly improve forecasting accuracy. This is evident from the out-of-sample validation tests of the classification matrix, AUROC, goodness-of-fit deciles, and Brier Score. Although, industry effects are statistically significant, they do not substantially improve the forecasting accuracy of the hazard rate model.

Our models are based on annual financial data, which obscures the fact that a company's financial position might be significantly different at the time of filing for bankruptcy. Employing monthly or quarterly data might improve the predictive power of the models. This can be a prospective area for future research. Second, the models could have been improved by adding market data. Further research may reveal the performance improvements by employing these data in a hazard rate model. Future research can also include exit dates from bankruptcies, to identify how many corporations that filed for Chapter 11 bankruptcies actually managed to reorganize successfully and exit bankruptcy. Further, we have selected variables based on a correlation matrix and the Variance Inflation Factor. However, future research can employ Principal Component Analysis (PCA) to select variables with highest explanatory power. It would also be interesting to predict bankruptcies based on the management structure. One could identify changes in director holdings over time, and test if this has any predictive power for bankruptcies (since management has more information about the financial position of the firm). This could also be an avenue to attempt to identify fraud.

# References


Allison, P. D. (2012). *Logistic regression using SAS: Theory and application*: SAS Institute.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance, 23*(4), 589-609.

Arminger, G., Sobel, M. E., & Clogg, C. C. (1995). *Handbook of statistical modeling for the social and behavioral sciences*. New York: Plenum Press.

Baldwin, J., & Glezen, G. W. (1992). Bankruptcy prediction using quarterly financial statement data. *Journal of Accounting, Auditing & Finance, 7*(3), 269-285.

Baltagi, B. H. (2002). *Econometrics* (3rd ed.). Berlin: Springer.

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of accounting research*, 71-111.

Bellotti, T., & Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society, 60*(12), 1699-1707.

Bellovary, J. L., Giacomino, D. E., & Akers, M. D. (2007). A review of bankruptcy prediction studies: 1930 to present. *Journal of Financial education*, 1-42.

Black, L. S. (1999). *Why corporations choose Delaware*. United States Corporation Company, 1-14.

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *The Journal of Finance, 63*(6), 2899-2939.

Casey, C., & Bartczak, N. (1985). Using operating cash flow data to predict financial distress: some extensions. *Journal of Accounting Research*, 384-401.

Charitou, A., Neophytou, E., & Charalambous, C. (2004). Predicting corporate failure: empirical evidence for the UK. *European Accounting Review, 13*(3), 465-497.

Chava, S., & Jarrow, R. A. (2004). Bankruptcy prediction with industry effects. *Review of Finance, 8*(4), 537-569.

Chen, H. (2010). Macroeconomic conditions and the puzzles of credit spreads and capital structure. *The Journal of Finance, 65*(6), 2171-2212.

Cole, R. A., & Wu, Q. (2009). Is hazard or probit more accurate in predicting financial distress? Evidence from US bank failures. *Munich Personal RePEc Archive,* paper number (29182)*,* 1-47.

Davidson, R., & MacKinnon, J. G. (1984). Convenient specification tests for logit and probit models. *Journal of Econometrics, 25*(3), 241-262.

Fantazzini, D., & Figini, S. (2009). Random survival forests models for SME credit risk measurement. *Methodology and Computing in Applied Probability, 11*(1), 29-45.

Galil, K., & Sher, N. (2015). Predicting default more accurately: to proxy or not to proxy for default? *Available at SSRN 2618190*.

Greene, W. H. (2003). Econometric analysis, 5th. *Ed.. Upper Saddle River, NJ*.

Halling, M., & Hayden, E. (2006). Bank failure prediction: a two-step survival time approach. *Available at SSRN 904255*.

Hilbe, J. M. (2009). *Logistic regression models*: CRC Press.

Hill, N. T., Perry, S. E., & Andes, S. (2011). Evaluating firms in financial distress: An event history analysis. *Journal of Applied Business Research (JABR), 12*(3), 60-71.

Hillegeist, S. A. (2001). *Corporate Bankruptcy: Do Debt Covenant and Disclosure Quality Measures Provide Information Beyond Options and Other Market Variables?*, School of Management at Northwestern University.* Corresponding author: Kellogg Graduate School of Management, Northwestern University, Evanston, IL.

Hosmer, D. W., & Lemesbow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics-Theory and Methods, 9*(10), 1043-1069.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398): John Wiley & Sons.

Jones, S., & Hensher, D. A. (2008). Advances in credit risk modelling and corporate bankruptcy prediction.

Lemeshow, S., & Hosmer, D. W. (1982). A review of goodness of fit statistics for use in the development of logistic regression models. *American journal of epidemiology, 115*(1), 92-106.

Lennox, C. (1999). Identifying failing companies: a re-evaluation of the logit, probit and DA approaches. *Journal of Economics and Business, 51*(4), 347-364.

Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata*: Stata press.

Myers, S. C. (1977). Determinants of corporate borrowing. *Journal of financial economics, 5*(2), 147-175.

Nam, C. W., Kim, T. S., Park, N. J., & Lee, H. K. (2008). Bankruptcy prediction using a discrete-time duration model incorporating temporal and macroeconomic dependencies. *Journal of Forecasting, 27*(6), 493-506.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, 109-131.

O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity, 41*(5), 673-690.

Park, C.-S., & Han, I. (2002). A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction. *Expert Systems with Applications, 23*(3), 255-264.

Pompe, P. P., & Bilderbeek, J. (2005). The prediction of bankruptcy of small-and medium-sized industrial firms. *Journal of Business Venturing, 20*(6), 847-868.

Pregibon, D. (1979). *Data analytic methods for generalized linear models.*

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Applied statistics*, 15-14.

Roulston, M. S. (2007). Performance targets and the Brier score. *Meteorological Applications, 14*(2), 185-194.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model*. *The Journal of Business, 74*(1), 101-124.

Simons, D., & Rolwes, F. (2009). Macroeconomic default modeling and stress testing. *International Journal of Central Banking, 5*(3), 177-204.

Verbeek, M. (2008). *A guide to modern econometrics*: John Wiley & Sons.

Westgaard, S., & Van der Wijst, N. (2001). Default probabilities in a corporate bank portfolio: A logistic model approach. *European Journal of operational research, 135*(2), 338-349.

Wilson, R. L., & Sharda, R. (1994). Bankruptcy prediction using neural networks. *Decision support systems, 11*(5), 545-557.

Zeitun, R., & Tian, G. G. (2007). Capital structure and corporate performance: evidence from Jordan. *Australasian Accounting, Business and Finance Journal, Forthcoming, 1*(4), 40-61.

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting research*, 59-82.

# Appendix

## A. Descriptive Statistics

**Table A-I. Descriptive statistics for the whole sample**

| | All firms | | Non-failed | | Failed | |
|---|---|---|---|---|---|---|
| Variable | Mean | Std. Dev | Mean | Std. Dev | Mean | Std. Dev |
| *Cash flow* | | | | | | |
| CFOTL | -0.22 | 1.19 | -0.11 | 1.3 | -0.32 | 1.06 |
| CFOFE | -1.95 | 150.28 | 12.94 | 166.74 | -16.89 | 130.03 |
| *Profitability* | | | | | | |
| NISALES | -9.28 | 89.26 | -8.82 | 96.54 | -9.74 | 81.32 |
| NITA | -0.42 | 1.22 | -0.25 | 0.99 | -0.59 | 1.39 |
| EBITTA | -0.3 | 1.02 | -0.18 | 0.88 | -0.41 | 1.13 |
| NITE | -0.52 | 34.75 | -0.43 | 20.06 | -0.61 | 44.9 |
| NITL | -0.13 | 26.21 | 0.86 | 30.19 | -1.13 | 21.46 |
| RETA | -25.24 | 530.03 | -21.64 | 423.82 | -28.85 | 618.61 |
| *Leverage* | | | | | | |
| CLTA | 4.16 | 102.04 | 3.78 | 80.5 | 4.55 | 119.83 |
| TDTA | 0.38 | 0.58 | 0.29 | 0.53 | 0.46 | 0.6 |
| TDTE | 0.61 | 72.74 | 1.11 | 33.74 | 0.1 | 97.26 |
| EBITIE | -14.03 | 121.66 | -3.93 | 124.7 | -24.16 | 117.68 |
| TETA | -3.58 | 102.05 | -3.08 | 80.54 | -4.08 | 119.84 |
| *Size* | | | | | | |
| TA | 1411.58 | 13147.73 | 1442.96 | 12222.76 | 1380.1 | 14016.98 |
| *Liquidity* | | | | | | |
| CASHTA | 0.19 | 0.24 | 0.22 | 0.26 | 0.16 | 0.22 |
| WCFOTA | -0.05 | 1.18 | 0.09 | 1.06 | -0.19 | 1.27 |
| *Solvency* | | | | | | |
| QATA | 0.21 | 6.59 | 0.09 | 9.31 | 0.34 | 0.24 |
| CACL | 2.76 | 8.56 | 3.3 | 5.85 | 2.21 | 10.57 |
| *Activity* | | | | | | |
| TASALES | 14.79 | 200.02 | 19.05 | 258.5 | 10.52 | 114.33 |

Notes: The table provides summary statistics of the explanatory variables implemented before the parsimonious models. Each firm year is considered as a separate observation. The total sample includes 5,524 firm years (2,767 and 2,757 firm years, respectively for non-failed and failed sample). The interpretation of the signs is an increase in a variable with the negative coefficient decreases the probability of a firm going bankrupt, and a positive sign increases the probability. CFO: cash flow from operations; TL: total liabilities; NI: net income; TA: total assets; CL: current liabilities; TD: total debt; WC: working capital; CA: current assets.

## B. All Financial Ratios

**Table A-II. All financial ratios tested in previous research**

| Category | Variable name | Variable definition |
|---|---|---|
| *Leverage* | RETA | Retained earnings/Total assets |
| | TETA | Total equity/Total assets |
| | TETD | Total equity/Total debt |
| | TETL | Total equity/Total liabilities |
| | TLTA | Total liabilities/Total assets |
| | TLTE | Total liabilities/Total equity |
| | TDTE | Total debt/Total equity |
| | TDTA | Total debt/Total assets |
| *Cash flow* | CFOTA | Cash flow from operations/Total assets |
| | CFOCL | Cash flow from operations/Current liabilities |
| | CFOTE | Cash flow from operations/Total equity |
| | CFOSALES | Cash flow from operations/Sales |
| | CFOTL | Cash flow from operations/Total liabilities |
| | CFOFE | Cash flow from operations/FE |
| *Liquidity* | CATA | Current assets/Total assets |
| | CACL | Current assets/Current liabilities |
| | CLCA | Current liabilities/Current assets |
| | CLTA | Current liabilities/Total assets |
| | CLTE | Current liabilities/Total equity |
| | QATA | Quick assets/Total assets |
| | QACL | Quick assets/Current liabilities |
| | CASHTA | Cash/Total assets |
| *Profitability* | EBITTA | Earnings before interest & taxes/Total assets |
| | EBITCL | Earnings before interest & taxes/Current liabilities |
| | EBITFA | Earnings before interest & taxes/Fixed assets |
| | EBITTE | Earnings before interest & taxes/Total equity |
| | EBITTL | Earnings before interest & taxes/Total liabilities |
| | EBITIE | Earnings before interest & taxes/Interest |
| | NIFA | Net Income/Fixed assets |
| | NISALES | Net Income/Sales |
| | NITL | Net Income/Total liabilities |
| | NITA | Net Income/Total assets |
| | NITE | Net Income/Total equity |
| | WCTA | Working capital from operations/Total assets |
| | WCTE | Working capital from operations/Total equity |
| | WCSALES | Working capital from operations/Sales |
| *Activity* | CASALES | Current Assets/Sales |
| | TESALES | Total equity/Sales |

*(Continued)*

**Table A-II. All financial ratios tested in previous research** (*Continued*)

| Category | Variable name | Variable definition |
|---|---|---|
| | TASALES | Total assets/Sales |
| | QASALES | Quick assets/Sales |
| | SALESCA | Sales/Current assets |
| | SALESTA | Sales/Total assets |
| | SALESFA | Sales/Fixed assets |
| *Size* | TA | Total assets |
| | Ln(TA) | Log of total assets |

Notes: Most of the ratios are gathered from the paper of Charitou et al. (2004). They summarize a substantial number of ratios that have been tested in previous research. This table also includes ratios they did not account for, whereas the market ratios are not included (as they were not available, and therefore not tested). Working capital (WC) = Current assets – Current liabilities; Cash flow from operations = NI + Depreciation ± Change in WC; Financial expenditures (FE) = Interest expenditure + Short-term debt; Quick asset = (Current assets – Inventories)/Current liabilities; Interest = Interest expenditure.

## C. Misspecification Tests

**Table A-III. Functional misspecification test**

| Variables | Model IA | Model IB | Model II | Model III | Model IV | Model V |
|---|---|---|---|---|---|---|
| Predicted | 1.0311*** | 1.0194*** | 1.027*** | 0.8443*** | 0.910*** | 0.9049*** |
| | (0.1042) | (0.1053) | (0.2075) | (0.1363) | (0.2463) | (0.1724) |
| Predictedsq | 0.0174*** | 0.0542 | 0.0082 | -0.04717 | -0.0289 | -0.0305 |
| | (0.0024) | (0.0449) | (0.0665) | (0.0438) | (0.088) | (0.0744) |
| Constant | -0.0105 | -0.0554 | 0.01033 | -0.0752 | -0.0356 | -0.0358 |
| | (0.0713) | (0.0795) | (0.0132) | (0.1068) | (0.1029) | (0.0931) |

Notes: Predicted: Predicted values; Predictedsq: Square of predicted values. To test for functional form misspecification, we ran an auxiliary logistic regression where the predicted values and the squares of the predicted values are regressed on the dichotomous dependent variable. Under the null hypothesis, if the model is correctly specified then the square of the predicted values would not be significant. As can be seen from the table, the square of the predicted values are significant for Model IA reflecting that we have omitted relevant variables due to functional form misspecification. After adding quadratic terms of the required independent variables, the square of the predicted value is no longer significant.

## D. Heteroscedasticity Plots

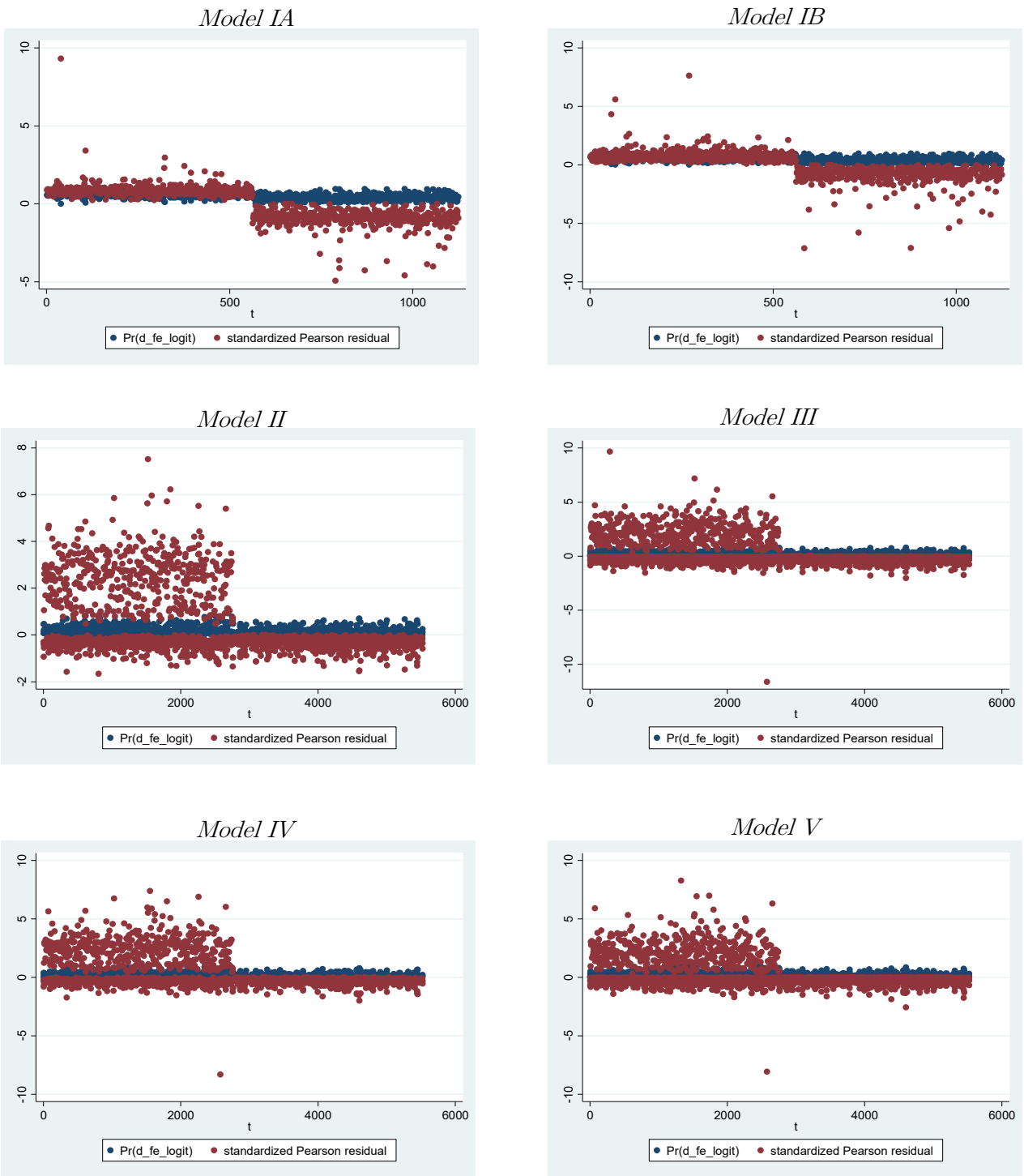


**Figure A-1. Heteroscedasticity plots.** Pr(d_fe_logit) is the predicted values. The graph portrays the plot of standard Pearson residuals against the predicted values for the visual inspection of heteroscedasticity. As it can be seen from the graph, the standard errors are not homoscedastic.

## E. Heteroscedasticity Test

**Table A-IV. Test for heteroscedasticity**

| Variables | Model IA | Model IB | Model II | Model III | Model IV | Model V |
|---|---|---|---|---|---|---|
| Predicted values | -4,484*** | -48.36*** | -26.95** | -3.728 | -2.558 | 3.660** |
| | (1,152) | (8.020) | (12.64) | (5.569) | (3.397) | (1.601) |
| Predicted square | 3,975*** | 43.94*** | 41.39* | 5.420 | 3.425 | -4.822* |
| | (1,204) | (7.999) | (23.11) | (10.19) | (5.657) | (2.676) |
| Constant | 1,158*** | 12.45*** | 3.529*** | 1.423*** | 1.294*** | 0.612*** |
| | (268.1) | (1.785) | (0.977) | (0.444) | (0.276) | (0.134) |
| Observations | 936 | 936 | 4,593 | 4,593 | 4,593 | 4,593 |
| R-squared | 0.018 | 0.039 | 0.001 | 0.000 | 0.000 | 0.001 |

Notes: Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1. To test whether our models are exposed to heteroscedasticity, we use the procedure of White, where the squared standardized residuals are regressed on the predicted values and the square of the predicted values. Further, we evaluated the model by post-multiplying the number of observations by the $R^2$. Under the null hypothesis, if the estimated value is greater than the critical value of chi-square, we can reject the null hypothesis of homoscedasticity. This suggests that our models might be exposed to heteroscedasticity. Therefore, we used standard measures to cope with this issue by using the robust standard errors (cross-sectional data) and cluster-robust standard errors (longitudinal data).

## F. Optimal Cut-off Points

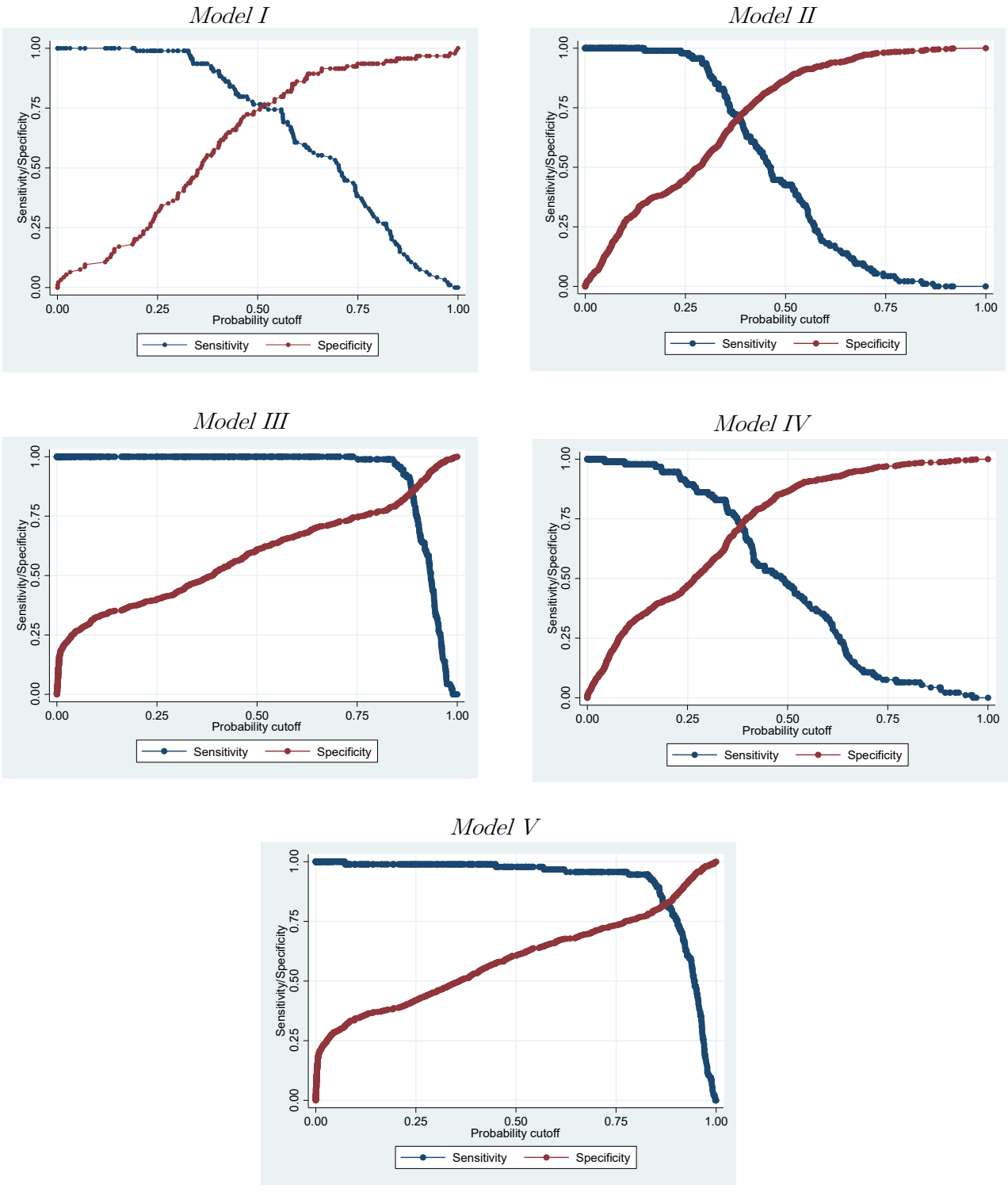


**Figure A-2. Optimal cut-off points.** The figure shows the relationship between sensitivity and specificity. Sensitivity is the correct classification of true default and specificity is referred as the correct classification of true non-default. As can be seen from the graphs, except for Model III and Model V, as the sensitivity increases the specificity decreases.

## G. ROC Curves

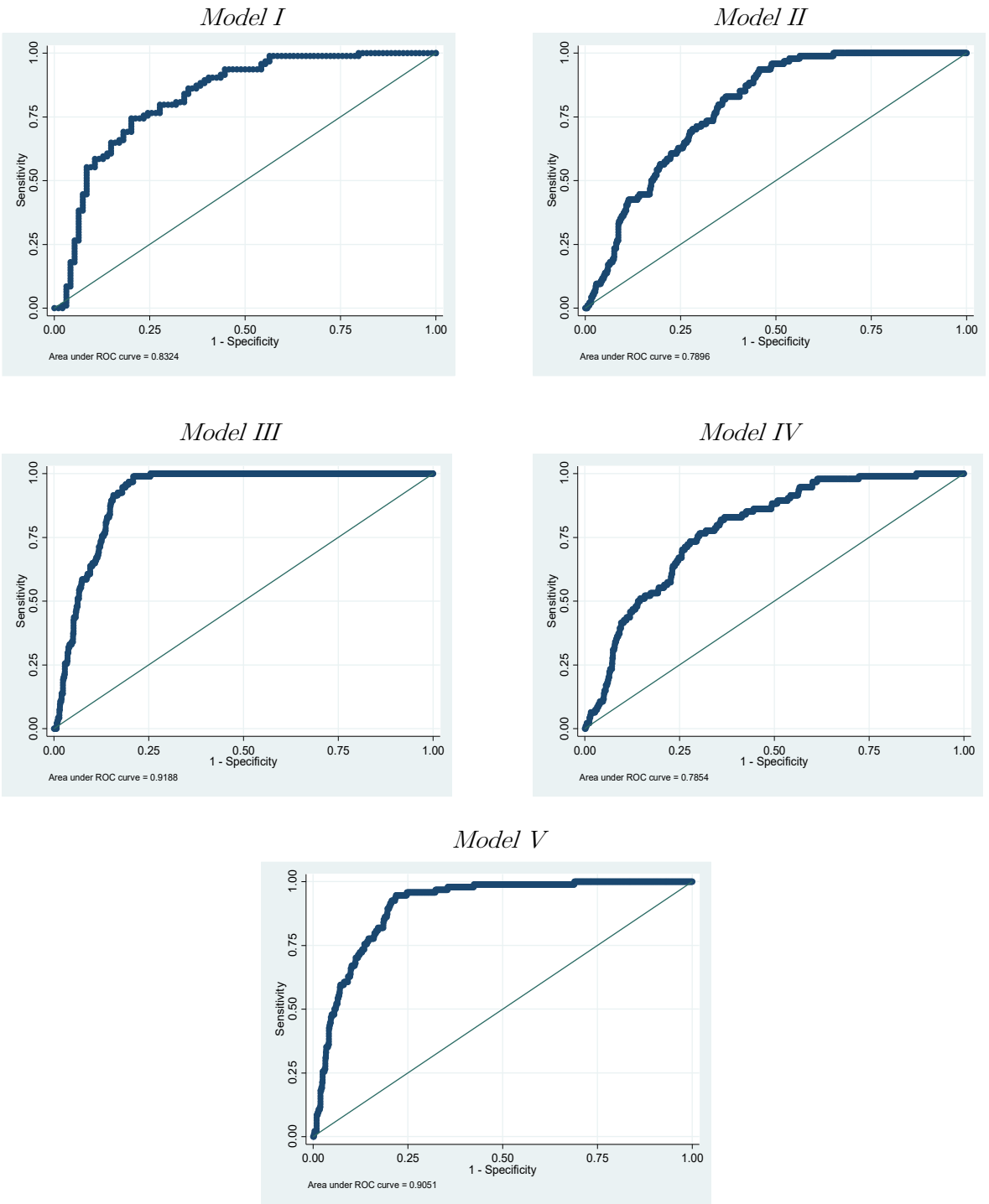


**Figure A-3. ROC Curves.** Receiver Operating Characteristics (ROC) plots the probability of true default (sensitivity) against the incorrectly predicted default (1 – specificity). The AUROC close to 1 indicates superior ability of the model in discriminating between the two outcomes. As can be seen from the graphs, Model III and Model V have an AUROC very close to 1 and thus are outperforming the other models in terms of discrimination ability.

## H. Results with all Interaction Terms

**Table A-V. Model IV and Model V with inclusion of all interaction terms**

| | Coefficients | |
|---|---|---|
| Variables | Model IV | Model V |
| Cash flow from operations/Total liabilities | -0.665** | -0.818** |
| | (0.271) | (0.334) |
| Net income/Total assets | -0.346* | -0.368* |
| | (0.198) | (0.207) |
| Current liabilities/Total assets | -0.0456 | -0.0997 |
| | (0.0923) | (0.181) |
| Total debt/Total assets | 1.521*** | 1.501*** |
| | (0.316) | (0.329) |
| Cash/Total assets | -1.622 | -1.902* |
| | (1.115) | (1.131) |
| Working capital/Total assets | -0.177 | -0.354 |
| | (0.176) | (0.309) |
| Current assets/Current liabilities | -0.00338 | 0.0110 |
| | (0.104) | (0.0816) |
| Ln (Total assets) | -0.0369** | -0.0270 |
| | (0.0165) | (0.0168) |
| ENECD | 0.771* | 0.964** |
| | (0.404) | (0.442) |
| Consumer staples | 0.582 | 0.589 |
| | (0.881) | (0.980) |
| Health care | -0.0449 | 0.165 |
| | (0.458) | (0.502) |
| Industrials | 1.322** | 1.500*** |
| | (0.518) | (0.545) |
| Materials | 1.327** | 1.317** |
| | (0.564) | (0.593) |
| Technology | 0.715 | 0.876* |
| | (0.491) | (0.524) |
| Utilities | -5.035 | -4.233 |
| | (5.650) | (4.164) |
| Cash flow from operations/Total liabilities × ENECD | -0.184 | -0.0660 |
| | (0.302) | (0.363) |
| Net income/Total assets × ENECD | 0.0798 | 0.0449 |
| | (0.169) | (0.168) |
| Current liabilities/Total assets × ENECD | 0.0788 | 0.138 |
| | (0.0928) | (0.183) |

*(Continued)*

**Table A-V. Model IV and Model V with inclusion of all interaction terms** *(Continued)*

| Variables | Coefficients | |
|---|---|---|
| | Model IV | Model V |
| Total debt/Total assets× ENECD | -0.473* | -0.445 |
| | (0.281) | (0.283) |
| Cash/Total assets × ENECD | -1.481 | -1.491 |
| | (1.449) | (1.464) |
| Working capital/Total assets × ENECD | 0.196 | 0.403 |
| | (0.245) | (0.353) |
| Current assets/Current liabilities × ENECD | -0.309** | -0.322** |
| | (0.148) | (0.130) |
| Ln (Total assets) × ENECD | 0.0243 | 0.0144 |
| | (0.0174) | (0.0178) |
| Cash flow from operations/Total liabilities × Consumer staples | -0.397 | -0.405 |
| | (0.601) | (0.662) |
| Cash flow from operations/Total liabilities × Health care | -0.168 | 0.0575 |
| | (0.309) | (0.380) |
| Cash flow from operations/Total liabilities × Industrials | 0.347 | 0.489 |
| | (0.402) | (0.445) |
| Cash flow from operations/Total liabilities × Materials | 0.751** | 0.951** |
| | (0.356) | (0.392) |
| Cash flow from operations/Total liabilities × Technology | 0.0972 | 0.256 |
| | (0.260) | (0.334) |
| Cash flow from operations/Total liabilities × Utilities | -1.528 | -2.214 |
| | (3.707) | (3.274) |
| Net income/Total assets × Consumer staples | -0.243 | -0.462 |
| | (0.769) | (0.918) |
| Net income/Total assets × Health care | 0.0446 | 0.123 |
| | (0.302) | (0.364) |
| Current liabilities/Total assets × Consumer staples | -2.588 | -2.633 |
| | (1.661) | (1.835) |
| Current liabilities/Total assets × Health care | -0.0104 | -0.0261 |
| | (0.122) | (0.214) |
| Current liabilities/Total assets × Industrials | 0.0343 | 0.0683 |
| | (0.0961) | (0.186) |
| Current liabilities/Total assets × Materials | 0.0625 | 0.114 |
| | (0.0928) | (0.181) |
| Current liabilities/Total assets × Technology | 0.0465 | 0.101 |
| | (0.0923) | (0.181) |

*(Continued)*

Table A-V. Model IV and Model V with inclusion of all interaction terms *(Continued)*

| Variables | Coefficients | |
|---|---|---|
| | Model IV | Model V |
| Current liabilities/Total assets × Utilities | 4.387 | 3.993 |
| | (9.091) | (7.911) |
| Total debt/Total assets × Consumer staples | 0.414 | 0.812 |
| | (0.747) | (0.739) |
| Total debt/Total assets × Health care | -0.0396 | 0.0247 |
| | (0.368) | (0.406) |
| Total debt/Total assets × Industrials | -0.279 | -0.265 |
| | (0.336) | (0.364) |
| Total debt/Total assets × Materials | -0.737 | -0.710* |
| | (0.485) | (0.428) |
| Total debt/Total assets × Technology | -1.081*** | -0.981*** |
| | (0.324) | (0.339) |
| Total debt/Total assets × Utilities | 7.319 | 6.387 |
| | (9.651) | (7.691) |
| Cash/Total assets × Consumer staples | 2.243 | 1.891 |
| | (2.435) | (2.410) |
| Cash/Total assets × Health care | 0.456 | 1.119 |
| | (1.320) | (1.395) |
| Cash/Total assets × Industrials | 1.540 | 2.074 |
| | (1.495) | (1.668) |
| Cash/Total assets × Materials | 1.601 | 2.148 |
| | (1.554) | (1.664) |
| Cash/Total assets × Technology | 0.759 | 0.882 |
| | (1.278) | (1.299) |
| Cash/Total assets × Utilities | 6.878 | 9.266 |
| | (8.251) | (9.140) |
| Working capital/Total assets × Consumer staples | -2.373 | -1.962 |
| | (2.247) | (2.648) |
| Working capital/Total assets × Health care | -0.112 | -0.162 |
| | (0.311) | (0.424) |
| Working capital/Total assets × Industrials | 0.358 | 0.461 |
| | (0.285) | (0.397) |
| Working capital/Total assets × Materials | 0.243 | 0.335 |
| | (0.374) | (0.379) |
| Working capital/Total assets × Technology | -0.200 | 0.0289 |
| | (0.230) | (0.346) |
| Working capital/Total assets × Utilities | -0.504 | -5.378 |
| | (14.76) | (15.38) |

*(Continued)*

Table A-V. Model IV and Model V with inclusion of all interaction terms *(Continued)*

| | Coefficients | |
|---|---|---|
| Variables | Model IV | Model V |
| Current assets/Current liabilities × Consumer staples | -0.254 | -0.223 |
| | (0.369) | (0.391) |
| Current assets/Current liabilities × Health care | -0.119 | -0.145 |
| | (0.158) | (0.160) |
| Current assets/Current liabilities × Industrials | -0.867*** | -0.877*** |
| | (0.253) | (0.234) |
| Current assets/Current liabilities × Materials | -0.758*** | -0.699*** |
| | (0.288) | (0.266) |
| Current assets/Current liabilities × Technology | -0.200 | -0.195 |
| | (0.253) | (0.244) |
| Current assets/Current liabilities × Utilities | -0.00767 | -0.0374 |
| | (0.343) | (0.404) |
| Ln (Total assets) × Consumer staples | 0.0542* | 0.0411 |
| | (0.0283) | (0.0295) |
| Ln (Total assets) × Health care | 0.0195 | 0.0113 |
| | (0.0238) | (0.0238) |
| Ln (Total assets) × Industrials | 0.0152 | 0.0102 |
| | (0.0262) | (0.0274) |
| Ln (Total assets) × Materials | 0.0288 | 0.0214 |
| | (0.0210) | (0.0215) |
| Ln (Total assets) × Technology | 0.0155 | 0.00824 |
| | (0.0213) | (0.0212) |
| Ln (Total assets) × Utilities | 0.0377 | 0.0251 |
| | (0.0241) | (0.0253) |
| (Cash flow from operations/Total liabilities)^2 | -0.0943*** | -0.0852*** |
| | (0.0315) | (0.0317) |
| (Net income/Total assets)^2 | -0.0622** | -0.0681** |
| | (0.0299) | (0.0309) |
| (Total debt/Total assets)^2 | -0.253*** | -0.263*** |
| | (0.0704) | (0.0712) |
| (Current assets/Current liabilities)^2 | 0.000620*** | 0.000614*** |
| | (0.000211) | (0.000212) |
| Lag Unemployment | | 0.661*** |
| | | (0.0399) |
| Lag Interest | | 0.0101 |
| | | (0.0273) |

(*Continued*)

Table A-V. Model IV and Model V with inclusion of all interaction terms *(Continued)*

| | Coefficients | |
|---|---|---|
| Variables | Model IV | Model V |
| d2008_1 | 1.934*** | 1.785*** |
| | (0.135) | (0.145) |
| Constant | -2.660*** | -6.007*** |
| | (0.369) | (0.483) |
| Model Fit | 686.74 | 913.93 |
| Pseudo $R^2$ | 0.1774 | 0.2415 |
| Observations | 4,593 | 4,593 |
| Functional misspecification | No | No |
| Heteroscedasticity test | Used CRSE | Used CRSE |
| Macroeconomic Variables | No | Yes |
| Non-linear forms | Yes | Yes |
| Industry effect | Yes | Yes |

Notes: Cluster robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1. The table provides the results from Model IV and Model V with all the interactions terms. Model IV is estimated without the macroeconomic variables. Whereas, Model V is estimated using both macroeconomic variables and the industry effects. The Model fit row reports the chi-square of the models. ENECD is 1 if sector is Energy or Consumer discretionary.

# I. Marginal Effects (Model IA-III)

**Table A-VI. Marginal effects (Model IA-III)**

| | Marginal effects | | | |
|---|---|---|---|---|
| Variables | Model IA | Model IB | Model II | Model III |
| CFO/TL | -0.104*** | -0.257*** | -0.0388*** | -0.0268*** |
| | (0.0214) | (0.0591) | (0.00873) | (0.00589) |
| NI/TA | 0.0211 | -0.170** | -0.0250** | -0.0174** |
| | (0.0210) | (0.0692) | (0.0105) | (0.00713) |
| CL/TA | | | 6.25e-05*** | 3.82e-05*** |
| | | | (1.89e-05) | (1.34e-05) |
| TD/TA | 0.113*** | 0.550*** | 0.0693*** | 0.0482*** |
| | (0.0362) | (0.0931) | (0.0144) | (0.0100) |
| Cash/TA | | | -0.0751*** | -0.0540*** |
| | | | (0.0253) | (0.0176) |
| WC/TA | -0.0106 | -0.0781* | -0.00799* | -0.00637* |
| | (0.0250) | (0.0425) | (0.00482) | (0.00349) |
| CA/CL | -0.0951*** | -0.0662** | -0.0121** | -0.00760** |
| | (0.0131) | (0.0289) | (0.00483) | (0.00321) |
| ln (Total assets) | | -0.00401** | -0.000741*** | -0.000468*** |
| | | (0.00158) | (0.000251) | (0.000169) |
| (CFO/TL)^2 | | -0.0370*** | -0.00543*** | -0.00362*** |
| | | (0.0131) | (0.00181) | (0.00121) |
| (NI/TA)^2 | | -0.0289*** | -0.00432** | -0.00307*** |
| | | (0.0104) | (0.00176) | (0.00117) |
| (TD/TA)^2 | | -0.145*** | -0.0162*** | -0.0117*** |
| | | (0.0290) | (0.00430) | (0.00307) |
| (CA/CL)^2 | | 0.000468** | 2.39e-05** | 1.49e-05** |
| | | (0.000192) | (9.57e-06) | (6.36e-06) |
| d2008_1 | -0.0199 | -0.0297 | 0.116*** | 0.0712*** |
| | (0.0524) | (0.0541) | (0.0116) | (0.00792) |
| Lag Unemployment | -0.00160 | 0.00193 | | 0.0256*** |
| | (0.0186) | (0.0208) | | (0.00228) |
| Lag Interest | | | | 0.000275 |
| | | | | (0.00107) |

Notes: Delta-method standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1. The marginal effects measure the instantaneous rates of change. This table shows the marginal effects of change in covariates from the mean on the default probability. For instance, in Model II, a one percentage increase in total debt as a portion of total assets from its mean increases the probability of default by 6.93% indicative by the positive sign on its coefficient. CFO: cash flow from operations; TL: total liabilities; NI: net income; TA: total assets; CL: current liabilities; TD: total debt; WC: working capital; CA: current assets.

## J. Marginal Effects (Model IV-V)

**Table A-VII. Marginal effects (Model IV-V)**

| | Marginal effects | |
|---|---|---|
| Variables | Model IV | Model V |
| Cash flow from operation/Total liabilities | -0.0361*** | -0.0250*** |
| | (0.00724) | (0.00508) |
| Net income/Total assets | -0.0175* | -0.0125** |
| | (0.00936) | (0.00632) |
| Current liabilities/Total assets | 3.99e-05** | 3.34e-05*** |
| | (1.70e-05) | (1.07e-05) |
| Total debt/Total assets | 0.0771*** | 0.0521*** |
| | (0.0137) | (0.00962) |
| Cash/Total assets | -0.0431* | -0.0320** |
| | (0.0224) | (0.0162) |
| Working capital/Total assets | -0.00611 | -0.00743*** |
| | (0.00411) | (0.00287) |
| Current assets/Current liabilities | -0.00801* | -0.00488* |
| | (0.00434) | (0.00292) |
| Ln (Total assets) | -0.000741*** | -0.000420*** |
| | (0.000238) | (0.000155) |
| ENECD | 0.0361*** | 0.0286*** |
| | (0.0116) | (0.00831) |
| Consumer staples | 0.0619** | 0.0475** |
| | (0.0274) | (0.0198) |
| Health care | 0.00685 | 0.00695 |
| | (0.0104) | (0.00757) |
| Industrials | 0.0666*** | 0.0489*** |
| | (0.0194) | (0.0130) |
| Materials | 0.0734*** | 0.0452*** |
| | (0.0235) | (0.0158) |
| Technology | 0.0314*** | 0.0243*** |
| | (0.0114) | (0.00823) |
| Utilities | -0.0226 | -0.0125 |
| | (0.0319) | (0.0208) |
| Current liabilities/Total assets × ENECD | 0.00118** | 0.000502 |
| | (0.000476) | (0.000306) |
| Total debt/Total assets × ENECD | -0.0228** | -0.0174** |
| | (0.00968) | (0.00689) |

*(Continued)*

**Table A-VII. Marginal effects (Model IV-V)** *(Continued)*

| | Marginal effects | |
|---|---|---|
| Variables | Model IV | Model V |
| Cash/Total assets × ENECD | -0.135*** | -0.0940*** |
| | (0.0477) | (0.0330) |
| Current liabilities/Total assets × Consumer staples | -0.125* | -0.0886* |
| | (0.0646) | (0.0458) |
| Working capital/Total assets × Consumer staples | -0.132** | -0.0882* |
| | (0.0632) | (0.0454) |
| Working capital/Total assets × Industrials | 0.0205* | 0.0163** |
| | (0.0110) | (0.00652) |
| Current assets/Current liabilities × Industrials | -0.0362*** | -0.0236*** |
| | (0.0119) | (0.00760) |
| CFO/TL × Materials | 0.0394** | 0.0283*** |
| | (0.0171) | (0.00989) |
| Total debt/Total assets × Materials | -0.0295** | -0.0209** |
| | (0.0144) | (0.00865) |
| Current assets/current liabilities × Materials | -0.0278** | -0.0154* |
| | (0.0128) | (0.00797) |
| Total debt/Total assets × Technology | -0.0497*** | -0.0253*** |
| | (0.0143) | (0.00827) |
| Working capital/Total assets × Technology | -0.0146** | |
| | (0.00651) | |
| Current liabilities/Total assets × Utilities | 0.202** | 0.157** |
| | (0.0960) | (0.0621) |
| (Cash flow from operation/Total liabilities)^2 | -0.00452*** | -0.00307*** |
| | (0.00135) | (0.000905) |
| (Net income/Total assets)^2 | -0.00338** | -0.00250** |
| | (0.00155) | (0.00105) |
| (Total debt/Total assets)^2 | -0.0139*** | -0.00975*** |
| | (0.00367) | (0.00260) |
| (Current assets/Current liabilities)^2 | 1.60e-05* | 9.66e-06* |
| | (8.56e-06) | (5.76e-06) |
| lag Unemployment | | 0.0224*** |
| | | (0.00194) |
| d2008_1 | 0.102*** | 0.0621*** |
| | (0.0101) | (0.00716) |

Notes: Delta-method standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1. The marginal effects are instantaneous rates of change in covariates from the mean on the default probability. For instance, a one percentage increase in cash flow from operations as a portion of total liabilities from mean decreases the probability of default by 3.61% indicative by the negative sign on coefficient. CFO: cash flow from operations; TL: total liabilities.