Credit Modelling 50 years after Altman's Z-score

Master Thesis 2020

Luca Sorlini

Student number: 102652

M.Sc. in Economics and Business Administration

Cand. Merc, ASC

Accounting, Strategy and Control

Supervisor: Jens Dick-Nielsen

Hand-in date: 15/05/2020

Characters: 141,162

Total pages: 80

Contents

Abstract

This dissertation investigates the relationships between probability of default and firm-level covariates. Two hypotheses are tested by comparing different specifications. I find abundant evidence suggesting that some univariate relationships between firm-level covariates and probability of default and some ceteris paribus relationship between firm-level covariates and probability of default are not linear. The empirical results presented in this paper suggest also that, when a large sample is used, additive models outperform standard linear models. Finally, I also provide an explanation for any non-linearity that is presented.

1. Introduction

Corporate default represents an event with potentially severe consequences for real economies. Estimating the probability that a corporation defaults is therefore highly important for both identifying imminent threats, which then can determine the choice of policy to be adopted, and to correctly price financial instruments involving credit risk, whose market can be considered large by any standard.

This dissertation contributes to a wide literature focused on the estimation of probability of default by exploring the validity of a specific assumption that is often made to estimate credit risk models. In fact, the relationship between firm-level covariates and probability of default is usually assumed to be *linear* to enable the fitting of generalized linear models. This is the case for a number of highly influential articles where credit risk models are presented.

I explore the validity of this assumption by estimating, rather then assuming, the form of these relationships and find persuasive evidence toward the idea that some of these are non-linear. Functions thus estimated are also shown through the extensive use of plots, which help understanding how the ceteris paribus marginal effect of a variable can change across the interval of values that variable can assume.

Further, I present evidence on the relative superiority, in terms of predictive power, of additive models over generalized linear models. The latter are widely adopted in credit risk applications, so this finding is particularly interesting, as it appears that the former have been relatively ignored.

Finally, I discuss the functions that are presented, providing some economic explanations for their form and any non-linearity that is detected.

The remaining of this paper is structured as follows: The next section presents hypotheses H_1 and H_2, the testing of which has guided this research. Part 1 presents a literature review on the statistical modelling of probability of default. Part 2 presents the methodologies that have been adopted to test the hypothesis H_1. Particularly important in this part of the dissertation is section 4.5 “Quantitative Methods”, where the statistical models being compared are presented. Part 3 presents the empirical findings, together with comparisons of goodness of fit and predictive power of the models that are presented. A robustness check is also presented. Finally, Part 4 presents a discussion providing economic explanations for some empirical findings presented in Part 3, and a criticism section, where a critical perspective on the research is offered.

2. Problem hypotheses

This paper investigates the relationship between probability of default and microeconomic covariates. Partly, the purpose of the research that has been carried out is to verify empirically the following hypotheses:

H_1: *at least one of the univariate relationships between firm-level covariates and probability of default is non-linear*.

H_2: *at least one of the ceteris paribus relationships between probability of default and firm-level covariates is non-linear*.

Testing H_1 is relatively straightforward as it entails the comparisons of two univariate regressions, one with a linear model, the other with a non-linear model. The two fits can then be compared to see if any of the two is better than the other. As the reader will note, a visual inspection of the two fits suffices to test H_1. I find clear evidence towards the acceptance of H_1.

Testing H_2 is more complicated than testing H_1. The approach adopted in this research is to compare the goodness-of-fit and predictive power of statistical models that assume these relationships to be linear with statistical models that relax this assumption and estimate non-linear approximations of these ceteris paribus relationships. I find empirical evidence, in line with current research, towards the relative superiority of the models in the second group over the models in the first and ultimately towards the acceptance of H_2. In fact, a number of comparative statistics and visualization tools all suggest that some of the relationships are non-linear.

The second objective of this paper is to describe the non-linearities that have been estimated and offer a possible economic explanation for their form. In fact, whereas some patterns of the functions that were estimated can caused by endogeneity or overfitting, others are clearly in line with economic theory.

All in all, the true underlying objective of this research is to contribute to the advancement of our understanding of corporate failure and its microeconomic drivers. Testing H_1, H_2 and present explanations for some empirical findings here presented should therefore be considered as complementary tasks.

3. Part I: Literature review

The literature on credit risk modelling offers numerous and diverse approaches to compute probability of default. Given the importance that accurate estimation of credit risk plays first in the real economy and, more specifically, in the banking sector, the topic attracted the attention of many scholars and professionals from such sector. Further, major corporate defaults, such as the cases of Enron, World.com and Lehman Brothers, clearly demonstrated the threats that these phenomena pose to the well-functioning of societies and to the well-being of their members. As a result, the last 100 years saw a consistent and incremental development of our understanding of credit risk.

The next paragraphs present a literature review of corporate credit risk modelling, with a focus on models based on microeconomic, firm-level variables. It should be stressed the literature was reviewed with three clear objectives: finding an appropriate set of covariates to be included in the models; identifying a linear model that is both accurate and widely adopted; identifying methodologies to detect non-linearities and adjust estimates to take them into account.

3.1 Univariate analysis

In its early stages, credit analysis relied heavily on univariate analysis of financial ratios. This methodology compares single financial ratios of defaulted firms with those of firms which did not default to see if there is any systematic difference. During the embryonal phase of this kind of analysis (i.e. during the first years of the 20th century), the current ratio was the indicator mainly used to assess short-term risk of bankruptcy (reference). However, the set of financial ratios used in univariate analysis started expanding in the 1920s (reference), though most of the studies in this period (e.g. Justin, 1924) were lacking a "scientific" (i.e. statistical) methodological framework. Many authors (e.g. Foulke, 1931; Fitzpatrick, 1932; Rasmer and Foster, 1931; Marwin, 1942; Walter, 1957; Hickman, 1958; Saulnier, 1958; Moor and Atkinson, 1961) demonstrated the usefulness of specific ratios to predict corporate default and identify financial distress. Thus, this period saw the set of financial ratios that was used in credit analysis expanding significantly and culminated with the work of Beaver (1966). In his article "Financial ratios as predictors of failure", Beaver reviews thirty financial ratios and compares their power in predicting corporate default between one and five years from the availability of the data. Among the thirty, he finds six ratios that are useful for estimating probability of default. Among these are financial leverage, the current ratio and net income to total assets, which were selected to be included in the models that were compared to test H_2.

As it can be noted, for a considerable period of time credit research focused on financial ratio analysis. The clear advantages of such analysis are mainly two: first, it is easy to compute; second, it can be communicated to and understood by laymen. However, univariate analysis fails to measure the effect that variables have on probability of default when controlling for other covariates. That is, it fails to measure the "ceteris paribus" effect of the regressors. Hence, users need to adopt heuristics or scoring systems to assign weights to each ratio to then estimate overall corporate probability of default. This practice is often highly subjective (Petersen, Plenborg and Kinserdal, 2017).

Research in univariate analysis of the relationship between single accounting ratios and probability of default has never ended. An example relevant to this dissertation is Acharya, Davydenko and Strebulaev (2012). They explored the relationship between credit spreads and cash holdings and find that "conservative cash policies" can be associated with higher probability of default because of endogeneity. That is, firms with large cash holdings might have higher probability of default because the decision to hold large reserves of liquid assets can be adopted by firms with looming solvency issues. Their findings suggest that the relationship between cash and probability of default is not linear and non-monotonic. I find evidence against this claim, though the empirical findings do not entirely undermine the claim made in that article.

3.2 Altman's Z-score

To address the problems of univariate analysis, Altman (1968) wrote an influential paper that can be seen as the first example of statistical measurement of credit risk using financial ratios. In his work "Financial ratios, discriminant analysis and prediction of corporate bankruptcy", Altman expresses his desire to "*bridge the gap [...] between traditional ratio analysis and the more rigorous statistical techniques which have become popular among academicians in recent years*" (The Journal of Finance, 1968, pp.589, vol. XXIII, no. 4). The classification technique adopted in his work is Multivariate Discriminant Analysis (MDA). Using a sample of 30 bankrupted firms and 30 non-bankrupted firms, all from the manufacturing sector, and five independent variables (Working Capital / Total Assets; Retained Earnings / Total Assets; EBIT / Total Assets; Market Value of Equity / Book Value of Total Debt; Sales / Total Assets), he uses MDA to classify the two groups, bankrupt and non-bankrupt by separating the two classes as neatly as possible with a linear classifier. The result is the division of the space populated by all manufacturing corporations in three regions: one populated by healthy firms with low probability of bankruptcy; one populated by financially distressed firms with high probability of bankruptcy; one grey area populated by firms with medium probability of bankruptcy. In order to identify which of the three regions a firm belongs to, a Z-score was presented in Altman's paper. This can be computed using the formula shown in Appendix 1. A score below 1.81 is associated with a high probability of default; a score above 2.99 is associated with a low probability of default; a score between 1.81 and 2.99 is

associated with a medium probability of default. The classifier was then refined into the ZETA® model (Altman, Halderman and Narayanan, 1977) to improve accuracy and marketability (Agarwaal and Taffler, 2008).

MDA was initially considered as a potential linear model to use in the testing of H_2. However, as it is explained in some detail in a later paragraph, the method is not the most suitable for the purpose of this research.

Further, the formulation of Altman's model is not rigorous and is in sharp contrast with a central part of finance theory. In fact, since accounting data is a representation of a company's historical performance, no expectation about future performance is embedded in this kind of information. From a theoretical point of view, given the backward-looking nature of financial ratios, market information should have superior predictive power than these as it reflects market's expectations about companies' future. In addition, according to some scholars (e.g. Hillegeist, 2004), since financials are prepared on a going-concern basis, accounting information has little power in predicting bankruptcy. Further, due to the flexibility of accounting principles and standards, financials can be manipulated to distort firms' financial representations. This can be done for several reasons, such as providing basis for higher management's compensation or increasing the market's appetite for a company's stock (See chapter 16 of Petersen, Plenborg and Kinserdal for a wide range of detailed examples). Manipulation of this kind results in financials that do not reflect companies' underlying economics and should therefore carry little information about probability of default.

3.3 Merton's structural model

The issues raised by academics were addressed by Nobel laureate Robert C. Merton shortly after Altman's publication. Following the notable advances in option pricing theory made by Black and Scholes (1973), a pricing model for bonds, supported by a theoretically rigorous formulation, was presented in Merton (1974). In his revolutionary paper "On the pricing of corporate debt", Merton models the equity of a company as a call option on the firm's assets. He then applies the Black and Scholes formula to show that the risk premium of corporate debt (i.e. the corporate credit risk) depends exclusively on 1) the volatility of the firms' assets and 2) the ratio between the present value of the firm's debt and the market value of its total assets (Merton, 1974). Merton's corporate bond pricing model provided the foundations for Moody's Kealhofer Merton Vasicek (KMV) model: this computes a firm's Distance to Default, corresponding to the number of standard deviations that separate the market value of a company's equity from zero. Distance to default can then be used, in turn, to estimate the firm's probability of default, by compute how likely a swing that brings the value of the assets below the value of the liabilities is.

As mentioned above, according to proponents of Efficient Market Hypothesis (EMH), Merton's model has the theoretical rigour that Altman's Z-score cannot rely on. In fact, in an efficient market, the price of a company's equity reflects all available information, public and non-public, including, but not limited to, financial ratios. Moreover, structural models enjoy also the possibility of using information in continuous time, whereas accounting information is released, usually, every quarter and can become obsolete.

Yet, Merton's model relies on a number of assumptions. One of these is that firms issue only one zero coupon bond. In addition, Merton's model assumes that companies do not issue any security, equity or debt, between the issuance of the bond and maturity. This can be regarded as highly unlikely. Furthermore, the model requires the estimation of one unobservable variable, namely market value of total assets, forcing users to use proxies instead of it (Linderstrøm, 2013) or to derive it from a system of equations. Due to these limitations, the emergence of structural models did not coincide with a halt in the use and development of credit risk models using accounting information and the two became the most widely adopted approaches.

A large number of scholars investigated the predictive power and informational content of structural models (e.g. Hillegeist, 2004; Eom, Helwege and Huang, 2004; Gilchrist and Zakrajsek, 2012. A notable instance can be found in Barath and Shumway (2008). They investigate the predictive ability of a distance to default model estimated by solving the set of equation presented in Merton (1974) and compare it with a naïve alternative estimated by using proxy variables and other, more comprehensive specifications. They find that distance to default is not a sufficient statistic to estimate probability of default. In addition, they find that the naïve alternative has higher out-of-sample predictive power than the one computed by solving the system of equations set out in Merton's paper. The findings presented by the two authors suggest that Merton's distance to default model has some predictive power, though this stems mainly from its functional form. As a consequence, variables used to estimate distance to default were considered to estimate the models presented in this dissertation. More specifically, volatility was found to carry predictive power in predicting corporate default. In addition, given the findings presented in Barath and Shumway (2008) relative to the accuracy of the naïve alternative, which approximates the market value of total assets by summing the market valuye of the equity to the book value of total debt, the proxy there used was also adopted in this research to estimate market value of total assets. This resulted again in a gain in predictive power.

3.4 Structural vs Accounting-based models

Agarwaal and Taffler (2008) compare the predictive power of the UK-based Z-score model with two structural models: one found by solving a set of equations, following the application in Hillegeist et al. (2004); the other

estimated using the approximations adopted in the naïve model presented in Barath and Shumway (2004). They use a sample of public UK firms and find that the UK-based Z-score model significantly outperforms Hillegeist's model in terms of predictive power but that the difference in performance between the UK-based Z-score and the naïve model is not statistically significant. Further, they investigate the information content of the different approaches and find that both carry unique information content. Finally, they compare the performance of the models taking into account the difference in misclassification costs. In fact, in credit risk classification, where a firm is classified as either 0 = "no-default" or 1 = "default", type II errors are typically more costly than type I errors due to a higher Loss Given Default. They simulate a corporate loan market where only two banks, one using the naïve model and the other using the Z-score, are the sole credit providers and competitors. They find that the bank using the Z-score is significantly more profitable than the other. Findings presented in this article provide substantial evidence that both market and accounting information should be used in estimating a credit model, which is the case in this dissertation.

3.5 Reduced-form models

During the 1980s, some advances in the field of mathematical finance (e.g. Pliska, 1981; Ho and Lee, 1986; Johnson and Stultz, 1987) created the assumptions for credit research to progress. As a consequence, the last decade of the 20th century saw the emergence of a new kinds of credit risk models, known as reduced-form models, that would later become the dominating methodology to estimate credit risk. The first reduced-form model has been presented in Jarrow and Turnbull (1995) who identified a need for "a new theory for pricing and hedging derivative securities involving credit risk" (The journal of Finance, vol.L, no.1, March 1995, pp. 53). Jarrow and Turnbull took an approach that is different than the ones adopted by precursors. In fact, in a reduced-form model, default intensities are modelled, as opposed to the single corporate probability of default. That is, focus is on modelling default intensities instead of identifying what causes default (Linderstrøm, 2013). This class of models maintain the theoretical rigor of structural models but rely on less stringent and more likely assumptions. Further, all exogenous variables can be directly observed, contrary to structural models where the firm's market value of total assets has to be estimated. These two advantages lead to the wide adoption of these kinds of models. Jarrow and Turnbull (1995), Jarrow, Lando and Turnbull (1997), Lando (1998) and Duffie and Singleton (1999) represent the main works that laid the foundations for default intensity-based modelling (Linderstrøm, 2013). This approach was considered but not adopted in this paper as it does not fulfil the aim set out in section 2. More specifically, given that the objective of this research is not to estimate the most accurate model, but rather to explain what causes default and, most importantly, how it causes it, reduced-form models were not adopted in this research.

3.6 Probability of default and logistic regression

Fuelled by innovations in computer technology, research in credit risk, just like research in many other statistical areas, intensified in the last 20 years of the 20th century. One important instance of this phenomenon is Ohlson (1980), who presents one of the first logit models to estimate probability of default. Logistic regression is more computationally intensive than MDA and this is probably why MDA applications emerged earlier than logistic regression. Moved by the need to address "some well-known problems associated with multivariate discriminant analysis" (Journal of Accounting Research, vol. 18, no.1, 1980, pp.111), which concerned violations of assumptions of MDA that Ohlson detected in earlier applications (e.g. Altman, 1968), Ohlson developed a logit model using four, statistically significant variables: size, enterprise value, net income to total assets, working capital and liquid assets to total assets. The article resulted in Ohlson's O-score. Ohlson O-score is effectively the corporate's probability of default estimated by the model. However, due to data imbalances, Ohlson identified as cut-off point (that is the point above which a firm's class is predicted to be 1 = Default) the value of 3.8%. Ohlson's logit model is more rigorous from a statistical perspective than Altman's MDA model, as the latter relies on a number of assumptions which are generally violated. For instance, the variance/covariance matrices of the regressors should be the same for both defaulting and non-defaulting groups and the independent variables should be normally distributed (Ohlson, 1980). In addition, MDA results in a score, whereas logistic regression effectively results in the probability that the observations belong to one of two classes, which is more interpretative (Ohlson, 1980). Finally, nowadays there is no difference, in terms of computational time, between fitting linear discriminants or logit/probit models, as this happens almost instantaneously. All of these facts constitute a large set of reasons to prefer logistic regression over MDA. Logistic regression was therefore selected as the linear methodology to use to test H_2. In fact, it should be noted that logit models are not exempt from making stringent assumptions. Particularly, it is assumed that there is a linear relationship between the log-odds (see section 4.5 for a definition of log-odds) of belonging to a class and the regressors. Testing this assumption coincides with testing H_2 and thus represents the objective of this research.

Zmijewski (1984) builds a probit model, which slightly differs from a logit model because of the choice of cumulative distribution (see section 4.5 for further elucidations), to investigate the effect that under-sampling of non-defaulted firms have on estimated probability of default. He finds that this practice results in biased estimated probability of defaults and classification rates. He therefore suggests using a sample characterized by an average probability of default close to the one of the population. This suggestion was followed in this dissertation.

Shumway (2001) present several points in favour of the superiority of dynamic logit models over static models. That is, he shows that estimators in logit models that do not take into account the development of covariates

over time tend to be inconsistent and biased. The dynamic models presented in his paper have better out-of-sample forecasts than Altman's MDA and Zmijewski 's probit model. Further, the dynamic models estimated using market variables have the highest performance among all the ones presented in that article. Finally, the sample he uses excludes all firms with a SIC code between 6000 and 6999, which corresponds to financial firms (including holdings) and real estate companies. This is because companies belonging to these industries do not share the same economics of other companies. Following the example set out in Shumway (2001), the exclusion of financial companies, the inclusion of lagged variables and the use of market values when available are practices that were implemented also in this dissertation.

Industry effects were explored by Chava and Jarrow. They confirm Shumway's claim that dynamic logit models have better performance than static models and therefore provide additional evidence in favour of dynamic models. Further, they show that industry effects play an important role in determining both intercepts and coefficients in hazard models. Finally, they document that using monthly financial information, instead of yearly, improves the predictive power of a model. Contrary to this last finding, I use annual information instead of monthly. This was done to limit the sample to a manageable size. After all, the objective of this research is not to estimate the most accurate model, but to investigate relationships between covariates and probability of default. Industry effects are not explored in this paper. This is mainly because the set that was used is not rich in defaults and therefore sub-setting it according to industries would lead to a very limited and unreliable number of default cases per industry. It is however important to mention Chava and Jarrow's work as it represents an opportunity for further research to expand the findings of this investigation.

Campbell, Hilscher and Szilagyi (2008) present a dynamic logit model used to investigate the relationship between excess returns and financial distress, also known as the "distress puzzle" (see Friewald, Wagner and Zechner, 2014 for futher information on the topic). In that article, financial distress is measured as probability of default, hence the need of a model estimating probability of default.

Cambell, Hilshcher and Szilagyi's paper interests this dissertation in several ways. First, given the relatively high accuracy, covariates included in that model were all considered for inclusion in the models here presented. In addition to measures of leverage, return on assets (net of interest expense), liquid assets over total debt and volatility also the logarithm of the share price, excess returns age and size were tested for inclusion. Most of the variables resulted indeed to be statistically significant. Second, the choice of methodology, that is logistic regression, further supported the idea, already presented in Ohlson (1980), that logit models represent a valid linear benchmark to use to test H_2. Third, the authors present in that article a way to include lagged information as recommended in Shumway (2001). Although the methodology adopted in this paper is not exactly the same as the one applied there, the two are very similar. In addition, the authors used a very large sample that seemed to benefit the predictive power of their model. This decision was adopted also in this research.

3.7 Splines, GAMs and credit risk

The 1990's represented a period of particular statistical innovation. The emergence of the fields of artificial intelligence and statistical learning (also known as machine learning, predictive analytics, big data etc…) coincided with the development of highly sophisticated quantitative methods for both regression, classification and clustering. However, the high degree of automation, the use of feature spaces and the transformation of input variables that some of these methods require conceal the link between input variables and response. As a consequence, "black box" methods, such as neural networks and support vector machines with radial basis functions, were not considered to explore the questions set out in section 2 as their ability to shed light on the underlying relationships between probability of default and accounting/market data is of limited comprehensibility. If estimating a model with a high predictive power was the sole objective of this research, the choice of method could have taken into consideration black box methods as well. Yet, given the purpose of this research, the literature regarding credit risk modelling with such black box methods is not presented.

Nevertheless, some methods from the statistical learning literature have revealed to be both flexible and interpretable. This is the case of additive models, a class of models that leverages the flexibility of piecewise polynomial functions and basis expansions, that can now be estimated by common statistical software (Hastie, Tibshirani and Friedman, 2009). A thorough description of additive models can be found in section 4.5 (this borrows heavily from chapter 9 of Hastie, Tibshirani and Friedman, 2009 and Wood, 2017).

By using expansions of the original input space, additive models are more flexible than traditional linear regression as no assumption about the relationship between regressors and regressand is imposed. Despite this advantage, applications of this class of models in credit risk research are limited. This fact is probably due, to some extent, to the high risk of overfitting that these methods entail (see section 4.5 for more on this regard).

Berg (2007) is one the first examples of applications of additive models in a credit risk context. He estimates a Generalized Additive Model (GAM) on a sample of Norwegian private firms and compares its fit and out-of-sample performance with the ones of an MDA, a logit model and a neural network. He finds that the GAM out-performs all other models in goodness-of-fit and accuracy. He also documents non-linearities for some accounting variables. Berg (2007) provides some evidence that additive models have superior predictive power and are more suitable for measuring probability of default than most models. As a result, his work heavily influenced the choice of non-linear modelling methodology to test H_2.

Giordani, Jacobson, von Schedvin and Villani (2014) (herafter referred to as Giordani et al.) represents a remarkable study of the non-linear relationships between accounting variables and probability of default. They estimate a GAM on a very large sample of Swedish firms and use natural splines as basis expansion (see section 4.5 for a description of additive models). They find that their GAM has higher predictive power than a

linear logit model. The article also presents a wide array of plots to visualize both the non-linearities (an example of which can be find in Appendix 2) and the predictive accuracy of the model (Appendix 3). As it will be explained in section 4.5, the adoption of basis expansions, and in particular of natural splines, might create difficulties in communicating empirical results as the underlying functions between dependent and independent variables cannot be summarized by a number (due to the non-linearity that characterize these functions). The choice of tools to communicate empirical results has drawn extensively from Giordani et al. (2014). Further, this research was also structured in a similar vein to the one there presented, as it was deemed to effectively fulfil the purpose to explore non-linear relationships. Moreover, that article presents univariate logistic regression and compares it with the univariate spline regression by plotting one against the other. This methodology was adopted to test H_1.

Both Berg (2007) and Giordani et al. (2014) inspired this research. In fact, their recent contributions explored a relatively ignored area of research. In addition, there seems to be a potential gap in credit risk research as no GAM that estimates probability of default for public companies could be found in the literature by the author of this dissertation. This has been considered as an opportunity that motivated this investigation.

This concluded Part 1. To summarize, the literature review resulted in the selection of the covariates to be tested for inclusion. These are return on assets (Net income/ total assets), financial leverage, volatility, the current ratio, cash holdings over total debt, share price and excess returns. As in the case of Campbell, Hilscher and Szilagyi (2008) I also include lagged information of return on assets and excess returns on the company stock. Logistic regression was selected as linear model that is used to test H_2, whereas natural splines were selected to estimate non-linear alternatives, again used to test H_2.

Part 2 proceeds by presenting the methodologies that have been applied.

4. Part II: Methodologies

This research is highly quantitative in nature and falls under the category of statistical modelling (Veal and Darcy, 2014). As such, no conclusion has been drawn if not supported by objective evidence produced according to a framework that is widely accepted by the scientific community. The quantitative methods that have been adopted in this investigation are not exclusively implemented in credit risk applications, but in many different fields such as medical research, phenome recognition, software development and many others. As a consequence, a vast body of knowledge has been produced to demonstrate the validity of the underlying theories on which these methods rest.

The next sections will describe how this research was conducted from a methodological perspective.

4.1 Philosophy and approach

First of all, the methodological approach adopted to carry out this research can be defined as strictly post-positivistic. Positivism developed in the 19th century and requires any truth to be objectively supported by data (Veal and Darcy, 2014). As in the case of a natural scientist, the positivistic researcher typically formulates hypotheses that have to be either accepted or rejected, depending on what the relevant theories suggest. Hypotheses that are accepted constitute the objective truth, which does not depend on individuals' perspectives or interpretations. Instead, it represents a sort of "noumenon".

In this regard post-positivism differs, in that post-positivists do not claim to have discovered an objective truth. Rather, any hypothesis that is not rejected is simply "deemed to be not falsified" (Veal and Darcy, 2014, pp. 36). In the eyes of positivists, hypotheses that are accepted helped researcher formulating theories which can guide our understanding until proved wrong.

In line with this approach, I make use of a wide range of quantitative methods that are typically employed by the scientific community and social scientists. It is indeed remarkable what can be accomplished due to the advances in computer technology and statistical theory. Further, I do not deem any finding presented in this research to be conclusive, as other approaches, more brilliant and sophisticated than the one presented here, can always be adopted and result in evidence that contradicts the findings here presented. It would be unreasonable to expect findings of such a research to be conclusive, especially given the researches in economics cannot rely on controlled environments, such as laboratories.

4.2 Research Characteristics

This investigation is applied (as opposed to non-applied) as it tries to explore real world phenomena (Veal and Darcy, 2014). Although, potentially, its findings could be generalized to the wide class of all public companies, in practice it takes under examination a specific sample of American public companies and its findings can be deemed valid only for the population of American public companies it is representative of. Further applied research of this kind is required to provide ground for the generalizability of the findings that are presented in sections 5.4 - 5.7.

This dissertation is also inherently deductive. Deduction seeks to formulate hypotheses based on theory and empirics and to test whether these hypotheses should be accepted or refused. Nevertheless, as in many other cases, this research cannot be considered exclusively deductive. Despite the little room for inductive processes,

these will be used to address specific interesting issues. More specifically, as mentioned in section 1, some economic explanations of the relationships that I document are offered. Elaborating these explanations entailed the use of induction, as such general explanations are formulated by the observation of the particular cases being studied. The deductive approach had to be adopted also because of the highly exploratory nature of this research. In fact, the ultimate goal of this dissertation is to contribute to our general understating of the drivers of corporate bankruptcy by investigating a specific topic that received little attention. As a consequence, formulating specific research questions was deemed to be premature. Obviously, the aim of this research is not limited to testing H_1 and H_2, but also to document some counter-intuitive relations that are not usually grasped by common sense.

Given the methodological approach adopted, this research is also highly empirical, as opposed to theoretical, as it makes extensive use of data analysis (Veal and Darcy, 2014). In fact, empirical research aims at drawn conclusion based on what the data suggests, contrary to non-empirical research that leverage exclusively logical reasoning. The former is vastly employed by econometricians and researchers in credit risk (e.g. Shumway, 2001; Campbell, 2008, Giordani et al., 2014, etc...). An instance of a notable exception is Merton (1974), that clearly shows how both kinds of research are necessary to the development of theories and understanding of complex phenomena. However, no research can be deemed to be exclusively empirical, as all the methodologies that can be implemented to extrapolate meaningful metrics from data rely on some theory that was logically derived (Veal and Darcy, 2014). Thus, it would be more appropriate to say that this dissertation is "mostly" empirical.

Finally, this research is characterized by a high degree of objectivity. This is a direct consequence of the choice of philosophy, research methods and research topic. However, as for other methodological characteristics presented in this section, it must be stressed that absolute objectivity cannot be achieved by any researcher, as subjectivity always interferes to some degree: this must be true for some specific choices, such as choice of topic (Veal and Darcy, 2014). The author acknowledges that some decisions related to this dissertation were taken arbitrarily. To maintain a high level of transparency, when the direct consequences of those decisions are presented, some attention is paid to the subjectivity behind these decisions.

4.3 Validity and Reliability

At the core of high-quality research is the adoption and implementation of valid and reliable methodologies.

Validity is "*the extent to which the information presented truly reflect the phenomena which the researcher claims it reflects*" (Veal and Darcy, 2014, pp.49). It is therefore paramount to, first, ensure that the research is valid, and, second, to disclose any potential threats to the validity of a specific research.

Internal validity concerns the relationship between the characteristics of the phenomena under investigation and the information relative to those characteristics. The information used by the researcher should be representative of those characteristics for a research to be internally valid. In this dissertation, the financial nature of the phenomena under investigation minimizes problems concerning internal validity. In fact, the financial aspects of a corporation can be easily communicated and represented. The financial data used in this dissertation is either as disclosed by corporations, in the case of accounting variables, or as quoted by the stock exchanges, in the case of market information. In addition, the use of market information mitigates the risk of relying on data that does not reflect the underlying economics of firms, as market information cannot be manipulated and distorted by individual firms and could even capture attempts of misrepresentation (Brealey, Myers and Allen, 2014).

External validity concerns the generalizability of the findings of the research. In fact, if the researcher fails to obtain a sample that is representative of the population being studied, the findings apply only to that specific sample and the research would suffer from little external validity. Ideally, as mentioned in the previous section this dissertation attempts to obtain findings that can be generalized to the universe of all public corporations. However, whether the findings of this paper can be generalized, to some extent, to all public corporations, is probably not true. Due to differences in culture, institutions, accounting standards and principles, legal systems and other factors among states, further research is required to assess the generalizability of the conclusions drawn in this research. Nevertheless, because of the dimensions of the sample used, findings of this investigation can be safely regarded as being generalizable to the universe of American public corporations. Sections 4.6 and 4.7 offer a thorough description of the sample and ample evidence of why it should be representative of such a population.

The second dimension, reliability, concerns the inter-temporal generalizability of the results. Conclusions that highly depend on the period characterizing the sample and that do not stand the test of time should be considered as suffering from little reliability (Veal and Darcy, 2014). This is a particular issue in this dissertation as there are numerous factors that could affect the relationships being investigated in this dissertation. Changes in accounting standards, such as the enactment of the Sarbanes-Oxley act, or changes in culture, such as the emergence of the IPO culture during the '90s, all represents potential threats to the reliability of this study. To address this problem, the relationships being investigated are also estimated for different time periods, and the resulting differences are explored. This is done to detect any evolution or development of the relations being studied. I find that, although hypothesis H_2 seems to be valid across all the

different time periods that were considered, little inter-temporal generalizability characterizes the relations that have been documented. These in fact seem to evolve over time.

Further, because of the many factors that shape social realities, the author of this research would not expect the findings to be completely replicable, if another sample is used. To find perfectly replicable findings seems to be somewhat normal in natural sciences but tends not to be the case in social sciences and especially in economics due to the uncontrollable environment where data and information is produced (Veal and Darcy, 2014). However, results from such a study, that is an identical statistical modelling research with a different sample, would be expected to be similar to the results reported in this dissertation, in terms of acceptance of H_1. However, the relations I documents seem to depend on factors that differ from state to state.

4.4 Research scope

A researcher should identify and communicate the boundaries that divide what is being studied from what is not (Veal and Darcy, 2014). This grants a certain degree of specificity to the research and allows the reader to understand how it relates to previous works.

Firstly, this dissertation does not investigate models used to estimate the Expected Loss on a given bond. This requires the estimation of both probability of default and Loss Given Default (LGD), which require different methodologies. While focusing extensively on the former, the latter is almost completely ignored.

Secondly, only methodologies that can be fully understood by users are explored in this research. Thus, "black-box" models from the fields of machine learning and Artificial Intelligence (AI) were not taken into account. This can be regarded as a direct consequence of the hypothesis being tested and of the problem at hand. In fact, values fitted by these methods cannot be fully comprehended due to the transformations of the input space that they entail.

Finally, this research focuses mainly on relationships between probability of default and microeconomic (firm-level) covariates. As a result, macroeconomic variables enter some of the models presented in this paper only to control for their effects or to create proxy variables.

4.5 Quantitative Methods

Before describing the quantitative methods that have been applied, to fully understand the nature of the problem guiding this research, the concept of "linear" has to be defined. In mathematics, the function between

an independent variable x_1 and a dependent variable y is said to be linear if it can be represented by a straight line with formula:

$$y = \beta_0 + \beta_1 x_1 \quad (1)$$

Y is said to be linear in X and the coefficients β_i represent the marginal effect of x_1 on y. However, there is another notion of "linear", that is easily encountered in statistical applications. In fact, a model is said to belong to the class of Generalized Linear Models (GLMs) if its predictor is in the linear form $\beta_0 + \beta_1 x_1, \ldots, \beta_i x_i$. However, the effect of the predictors on y could be mediated by a link function, whose form depends on the probability distribution that is adopted. For instance, if the observations are thought to have a Poisson (i.e. exponential) distribution, the relationship between y and x can be represented as:

$$ln(\lambda_i) = \beta_0 + \beta_1 x_1 \quad where\ y_i = Poisson(\lambda_i). \quad (2)$$

ln( ) is the link function, as it links the linear predictor $\beta_0 + \beta_1 x_1$ to the parameter of the probability distribution. A model with functional form as in formula (2 is said to be linear in the β_i coefficients. Yet, depending on the probability distribution and the link function, the model could be non-linear in the x_i. Throughout the rest of this dissertation, when a model is described as linear, the second notion of linearity presented above is intended. Thus, in the remaining, a linear model is essentially a model belonging to the GLM class.

This section proceeds as follows: firstly, the linear functional form that was adopted to test hypothesis H_1 is presented, together with the methodologies to fit such a functional form and the assumption on which it rests. Secondly, the GLM class is extended by presenting additive models. This class differs in that it is not linear, as opposed to GLMs. In that section, the focus is on the different assumptions on which additive models rely. Finally, the concept of basis expansion is introduced in some detail, together with the notion of spline regression.

Binomial Generalized Linear Models

In statistics, a model belonging to the binomial family, which in turn belongs to the class of GLMs, estimates the probability that a certain observation belongs to one of two classes. If the classes are coded as 0 and 1, as they typically are, then the probability P that an observation belongs to class 1, conditional on an input vector **x**, is given by

$$P(y = 1 \mid \mathrm{X} = \mathbf{x}) = P(y = 1 \mid \mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_k) = G\,(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) = G(\beta_0 + \boldsymbol{\beta}\mathbf{x}) \quad (3)$$

where G is the link function. Since the sum of all probabilities must equal 1, and given that there are only two classes ,1 and 0, the probability is given by:

$$P(y = 0 \mid \mathrm{X} = \mathbf{x}) = 1 - P(y = 1 \mid \mathrm{X} = \mathbf{x}) \tag{4}$$

The main binomial models used in applied research are the logit, with link function:

$$\log \frac{P(Y=1|X)}{P(Y=0|X)} = G\,(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) = \frac{exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}{1 + exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)} \tag{5}$$

where $\log \frac{P(Y=1|X)}{P(Y=0|X)}$ are the log-odds, and the probit, with link function:

$$\begin{aligned} G\,(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) &= \Phi(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) \\ &\equiv \int_{-\infty}^{\mathbf{x\beta}} \phi\,(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) dv \end{aligned} \tag{6}$$

Where ϕ is the normal standard density function:

$$\phi = (2\pi)^{-1/2}\, e^{(-\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k/2)}. \tag{7}$$

Although economists seem to favour the probit link due to some differences in the assumptions regarding the distribution of the errors (Wooldridge, 2019), this paper mainly focuses on logit models, due to the wide adoption of these models in credit research (logit models are estimated in Shumway, 2001, Campbell, Hilscher and Szilagyi, 2008, and Giordani et al., 2014). Logit and probit models differ in the cumulative distribution used to compute the response probability. The logit uses the logistic cumulative distribution, whereas the probit relies on the normal cumulative distribution (these are compared in appendix 4).

In estimating probabilities, both logit and probit models are far superior than Linear Probability Models (LPMs) (Wooldridge, 2019). This is because the latter assumes that partial marginal effects are constant. Hence, it can fit probabilities below 0% and above 100%, which can hardly be interpreted. Moreover, decreasing/increasing marginal effects can reflect more appropriately some aspects of the DGP. For instance, and with reference to the context of this research, it is indeed highly unlikely that a one-unit increase in net earnings has the same marginal effect on probability of default for firms that are very profitable and for firms that are loss making. It is far more likely that the increase in earnings decreases substantially the probability of default of the loss-making firm and decreases only slightly the probability of default of the company with positive net income.

However, a model linear in **x,** as the LPM, has a clear advantage in that the parameters are highly informative, since they express the partial marginal effects of each x_i on y. To compute partial marginal effects of a regressor on the regressand, in a logit model, the first order derivative with respect to that variable has to be used. In this dissertation, however, another approach, similar to the one used in Giordani et al. (2014) will be used. This is presented in some detail in a later section.

Logit models rely on several assumptions:

1) The relationship between the log-odds and **x** is linear.
2) The observations are independent from each other.
3) Independent variables are not affected by multicollinearity (i.e. high covariance).
4) The sample used has to be large.

Assumption 1 is the object of this research. I present persuasive evidence supporting the claim that the relationship between log-odds and the input vector **x** is non-linear and that models that relax this assumption and adjust for eventual non-linear relationships are superior than logit models.

Assumption 2 is considered to be respected in this application, as the sample used to estimate the logit models presented in later sections is not a result of ensemble methods.

Assumption 3 was explored by estimating the correlation matrix. This is presented in more detail in section 5.3.

Assumption 4 can be safely considered as respected due to the size of the sample that has been used (number of observations = 111,788).

Finally, logit and probit models are estimated by Maximum Likelihood Estimation (MLE). The density function of a binary variable y, taking only values in the interval [0, 1], conditional on the input matrix **X**, is

$$f(y|\mathbf{X};\boldsymbol{\beta}) = [\mathbf{G}(\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{X})]^{y}[\mathbf{1}-\boldsymbol{G}(\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{X})]^{\mathbf{1}-\boldsymbol{y}} \quad (8)$$

.

In fact, when y = 0,

$$[\mathrm{G}(\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{X})]^{y} = 1 \rightarrow f(y|\mathbf{X};\boldsymbol{\beta}) = [\mathbf{1}-\boldsymbol{G}(\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{X})]^{\mathbf{1}-\boldsymbol{y}} \quad (9)$$

.

and when y = 1,

$$[1-G(\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{X})]^{\mathbf{1}-\boldsymbol{y}} = 1 \rightarrow f(y|\mathbf{X};\boldsymbol{\beta}) = [\mathbf{G}(\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{X})]^{y} \quad (10)$$

.

Defining the log-likelihood function as:

$$\ell_i(\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{X}) \equiv \boldsymbol{y_i log}[\boldsymbol{G}(\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{X})] + (\mathbf{1}-\boldsymbol{y_i})\boldsymbol{log}[\mathbf{1}-\boldsymbol{G}(\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{X})] \quad (11)$$

.

the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ maximizes the likelihood function

$$argmax\ \mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{\ell}_i(\boldsymbol{\beta}). \tag{12}$$

w.r.t. $\boldsymbol{\beta}$.

The likelihood function represents the relationship between the probability of drawing the sample at hand and different values of the parameter vector $\boldsymbol{\beta}$. Therefore, $\widehat{\boldsymbol{\beta}}$ represents the most "likely" parameters given the sample used to estimate them.

Additive models

To expand the GLM framework presented in the previous section, a new class of models has to be introduced. In the general case, additive models assume that the relationship between a dependent variable y and independent variables $x_1, x_2, \dots, x_n$ can be expressed as

$$y = f_1(x_1) + f_2(x_2) + \dots + f_n(x_n), \tag{13}$$

where f is a "smoothing" function. More specifically and with focus on this application, additive logistic regression assumes the relationship between log odds and independent variables to be in the form

$$\log \frac{P(Y=1|X)}{P(Y=0|X)} = log\left(\frac{\mu(X)}{1-\mu(X)}\right) = \alpha + f_1(x_1) + f_2(x_2) + \dots + f_n(x_n) \tag{14}$$

where $\mu(X) = P(Y=1|X)$, α is a constant and $log\left(\frac{\mu(X)}{1-\mu(X)}\right)$ is the logit link for an additive model. This can be compared with the case of a GLM with logit link:

$$\log \frac{P(Y=1|X)}{P(Y=0|X)} = log\left(\frac{\mu(X)}{1-\mu(X)}\right) = \alpha + \beta_1(x_1) + \beta_2(x_2) + \dots + \beta_n(x_n) \tag{15}$$

As it can be noticed, whereas linear models assume only a coefficient that multiply the regressor, a non-linear model uses a whole function, which can be non-linear and non-monotonic. It should be stressed that, by assuming a relationship as in formula (14) between independent and dependent variables, interaction terms between independent variables, such as $f_1(x_1, x_2)$, are assumed away. Although Wood (2017) notes that the

additivity assumption is a "fairly strong one" (Wood, 2017, pp. 131), additive logistic regression has higher flexibility than logistic regression. In fact, the only assumption that is relaxed between the two classes of models is exactly the one about the linear relationship between log-odds and regressors. This relationship is indeed linear for GLMs, whilst it is affected by the functions $f_i(x_i)$ in additive models. Once functions f_i are defined, additive models are fitted by standard MLE. This relative simplicity of fitting these models represents one of the reasons behind the selection of this method.

A problem that arises with the estimation of additive models concerns the choice of the smoothing functions f_i. These need to be prespecified before fitting additive models. The literature of statistical learning offers many possibilities which can usually be grouped under the umbrella term *"basis expansions"*. In this dissertation the choice of smoothing functions followed Giordani et al. (2014) and resulted in the selection of natural splines. This is due to the results that this methodology led to in the case of that journal article. Natural splines, splines and basis expansions are presented in the next section.

Basis expansions and splines

One way to estimate the underlying relationship between probability of default and firm-level covariates, with more flexibility than logit models can allow for, is to expand the feature space **X** and to fit the linear models in such an expanded space. Denoting by $h_m(X) : \mathbb{R}^p \longmapsto \mathbb{R}$ the mth transformation of X, a linear basis expansion would model

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X) \tag{16}$$

With m taking values between 1 and M (Hastie, Tibshirani and Friedman, 2009). As it can be noted from formula (16), the "trick" is to transform the input vectors **x** and fit linear models using the transformations thus obtained. In fact, each basis expansion $h_m(X)$ is multiplied by a coefficient β_m. This rather simple idea allows to build upon the class of GLMs by relaxing the assumptions about linearity and by fitting non-linear, rather than linear, functions. $f(X)$ can then be regarded as an estimate of the underlying function between response and input variables, conditional on all other regressors.

However, formula 16 is very general, suggesting that there are numerous ways to expand the input matrix X. The class of basis expansions that was adopted in this research to test hypothesis H_1 and H_2 is known as natural cubic splines. As mentioned in the Part 2, choosing this specific set of basis expansions followed the methodology adopted in Giordani et al. (2014) and, to a lesser extent, Berg (2007). In addition, splines usually

entail relatively simple transformations, which ease the interpretability and communicability of the results, as opposed to, for instance, kernel methods.

A cubic regression spline is a piecewise local polynomial used for regression, rather than interpolation. For piecewise, it is meant that the feature space X is divided in, say, n regions by knot points $\xi_1, \xi_2, \dots, \xi_{n-1}$ (the Greek letter ξ follows the notation used in Hastie, Tibshirani and Friedman, 2009). The term polynomial indicates that basis functions $h_1 = X,\ h_2 = X^2, \dots, h_k = X^k$ are used. For local it is meant that the polynomial is fitted in a specific interval of values between two knots.

A cubic spline is made of four basis expansions $h_1 = 1,\ h_2 = X,\ h_3 = X^2,\ h_4 = X^3$ to expand the feature space, where h_1 is a constant term and $h_2,\ h_3,\ h_4$ enable the fitting of a polynomial function of third degree. A cubic spline requires also constraints imposing continuity at the knot points. Introducing the superscript ξ_i^-, to indicate the negative side of the knot ξ_i, and ξ_i^+, to indicate the positive side of the knot ξ_i, a continuity constraint can be expressed as $f(\xi_i^-) = f(\xi_i^+)$. However, the same constraint can be expressed more conveniently in the form of a basis such as

$$h_k(X) = (X - \xi_i)_+ \tag{17}$$

where again the subscript indicates the positive part of the function. Essentially, formula 17 specifies that from knot point ξ_i onward, the basis h_k will co-determine values of $f(X)$ together with all other bases whose effect is not nihil in that specific range.

The fact that the local polynomials fitted are of third order and continuous proves to be useful in many applications, as the resulting function *f(x)* has continuous first- and second-order derivatives. Cubic splines have degrees of freedom equal to 4 (the number of basis expansions) plus number of knots. A practical comparison of piecewise polynomials and cubic splines is offered in Appendix 5 and Appendix 6.

Natural cubic splines, which is precisely the set of basis expansions adopted in Giordani et al. (2014), represent an attractive sub-group of cubic splines. In fact, this specific type of spline relies on an additional assumption, that is, the underlying relationship being estimated is linear beyond the boundary knots. At a first glance the benefits of such an assumption are difficult to grasp. Yet, given that the function is assumed to be linear beyond the extreme knots implies that continuity constraints in those knots can be dropped, and additional knots can be more profitably placed interiorly, leaving the degrees of freedom unchanged. In addition, assuming that the underlying function is linear beyond the boundary knots, where little data is given, is a reasonable assumption (Hastie, Tibshirani and Friedman, 2009).

By using the set of basis functions just presented, natural cubic splines can generally be implemented to estimate more flexible models. In fact, if the underlying relationship between independent and dependent

variables is assumed to be linear when indeed it is non-linear, the model would suffer from underfitting and result in poor in-sample fit and out-of-sample performance. However, the use of basis expansions could lead to the estimation of a model that is too specific to the sample at hand. This model would then be said to suffer from overfitting and have poor predictive power, despite having high in-sample fit.

There are several ways to reduce the probability of overfitting and all of them entail the control of two parameters: the degree of the polynomial and the number of knots. The first is generally held fixed, which is also the case in this dissertation. In fact, there are few applications that require a polynomial with order higher than three (Hastie, Tibshirani and Friedman, 2009). As for the number of knots, there are different approaches.

Before proceeding with the presentation of such approaches, it should be noted that, once the number of knots is chosen, these are then placed at either equidistant quantiles, thus placing more knots where more data is present, or equidistant values. Giordani et al. (2014) offer a third option, that is, the position of the knots can be determined through k-means clustering. In this dissertation, knots are placed at equidistant quantiles, as this seems to be common practice in applied research (Hastie, Tibshirani and Friedman, 2009; Wood, 2017; Perperoglu, 2019).

One way to avoid overfitting models with splines to the sample at hand is by stepwise backward deletion. The process requires the estimation of increasingly simple nested models until a specification that is parsimonious and that has a good in-sample fit is identified. If only (in-sample) fit was used as criterion, complex models would always be preferred over simple ones. Vice versa, if parsimony was the only criterion used for model selection, models would reduce to simple constants, as they represent the most basic functional form. Stepwise backward deletion starts with an overly complex model, which in the case of splines translates into a large number of knots. It then proceeds by estimating models with a smaller number of knots and comparing the two specifications using a criterion that measures in-sample fit whilst penalizing model complexity. The two most used bases for comparison are the Aikake Information Criterion (AIC)

$$AIC = 2k - log(\mathcal{L}) \tag{18}$$

where k is the number of parameters and $\mathcal{L}$ is the log-likelihood of the model, and the Bayesian Information Criterion (BIC)

$$BIC = k\,log(n) - 2\,log(\mathcal{L}) \tag{19}$$

where, again, k is the number of parameters, $\mathcal{L}$ is the log-likelihood of the model and n is the number of observations (Hastie, Tibshirani and Friedman, 2009). Both AIC and BIC increase as parameters are included and decrease with the likelihood, which, it is reminded, is a measure of in-sample fit. Therefore, models that have low AIC and BIC should be preferred over models with high AIC and BIC.

Stepwise backward deletion has been implemented in Giordani et al. (2014) to select the number of knots of natural splines and is also implemented here.

To summarize, in order to provide evidence towards the non-rejection or the rejection of H_2, two classes of models will be fitted and compared. The first, which has been defined as linear, is a binomial GLM model with logit link. The second, which has been defined as non-linear, is a binomial additive model with logit link function and natural splines as smoothing functions. If no significant difference, in terms of in-sample-fit and, most importantly, out-of-sample performance, is found or if the linear model results to be more accurate than the non-linear, then H_1 should be rejected, as no evidence that the underlying relationships are non-linear is presented. Vice versa, if non-linear specifications are found to be more accurate and to better estimate these underlying relations, then H_2 cannot be rejected. As it is shown in later sections, a large body of evidence has been produced towards the non-rejection of H_2.

Testing H_1 instead, requires only the visual comparison of the fit of univariate logistic regressions against the fit of univariate logistic regression with natural splines. This process is indeed simpler than testing H_2.

Part 2 proceeds by describing the data that were used to estimate the models, together with the data handling techniques that were implemented.

4.6 Data Sources

Annual firm data was imported from Standard and Poor's COMPUSTAT database for all available American, publicly listed firms on each year spanning the period 1970-2016. The choice of period was arbitrary and followed the example set out in Campbell, Hilscher and Szilagy (2008). Among the many variables that were considered, the accounting variables that were selected to build the data set that has been used in this research are Net Income, Current Assets, Cash and Cash Equivalents, Total Liabilities, Current Liabilities, whereas other firm data sourced from COMPUSTAT are industry code, CUSIP identification code (CUSIP), financial year (FYR) and date of release of accounting information (DATE). The calculation of Book Value of Equity follows the indications set out in Fama and French (1993).

Market Data for the same group of companies was imported from the Center for Research in Security Prices (CRSP), that is affiliated with the University of Chicago Booth School of Business. The market variables that were selected to be part of the dataset are market capitalization of companies' equity, equity annual volatility (calculated using daily returns across the previous year), 1-year-trailing returns on the companies' equity and stock price. Again, the calculations of companies' market capitalization of companies' equity follows Fama and French (1993).

In line with Campbell, Hilscher and Szilagyi (2008), taking into consideration Shumway's findings supporting the superiority of market-based variables over accounting-based variables (Shumway, 2001), and given the evidence regarding the high predictive power of proxy variables presented in Barath and Shumway (2008), accounting and market data has been combined to create an approximation of market-based unobservable variables. Net Income to Market Value of Total Assets (NIMTA) was calculated as the ratio between companies' Net Income over the sum of the market capitalization of companies' equity and the book value of their total debt. The rationale behind this calculation is that book values of companies' total debt represent good approximations of market values of their debt. Thus, summing book values of companies' debt and their market cap would result in a good approximation of the market values of companies' assets. Applying the same logic, Total Liabilities over Market value of Total Assets (TLMTA) was computed as the ratio of the book value of total liabilities to the sum of companies' equity market capitalization and book value of total debt. Following Campbell, Hilscher and Szilagyi (2008), the natural logarithm of companies' stock prices (LPRICE) was taken. In addition, in line with the same article, the natural logarithm of the ratio between companies' market cap and the S&P 500 index level was also taken to measure companies' size. To do so, daily levels of the Standard and Poor 500 index at close were sourced from Yahoo Finance and matched to each observation DATE. The process resulted in the computation of the variable SIZE. Further, firms' age (AGE) was calculated based on the number of previous year-firms sharing the same CUSIP identification code. In other words, AGE represents the number of years a firm appears in the sample. As a consequence, it should be stressed that AGE is a rough approximation of actual firms' ages.

Contrary to Campbell, Hilscher and Szilagyi (2008), I also include the current ratio (CR), defined as current assets over current liabilities, due its wide adoption throughout the early history of credit analysis. Further, I also include the ratio of cash and cash equivalents to the book value of total debt. In fact, the findings presented in Acharya, Davydenko and Strebulaev (2012) and in Giordani et al., (2014) suggest that the relationship between this particular financial ratio and companies' credit risk is non-linear and non-monotonic.

Bankruptcy dates, defined as the date a company files for either chapter 7 or chapter 11, were sourced from Georgia Tech Ernest Scheller Jr. College of Business' bankruptcy database. A dummy variable (DEF1Y), where 1 indicates a default in the next financial year and 0 indicates otherwise, was then constructed.

Firms sharing the same financial year were sorted in quintiles, first, according to the market capitalization of their equity and, then, according to their book-to-market ratio (i.e. the book value of companies' equity divided by their market value). Thus, for each financial year, 25 portfolios (five market capitalization portfolios times five book-to-market portfolios) were created and each firm-year was then assigned to one of these. Data regarding average returns of portfolios created in this way was sourced from the Fama and French library and matched to the relevant portfolios. Portfolio returns were then subtracted from firm-year's returns to compute excess returns (EXR).

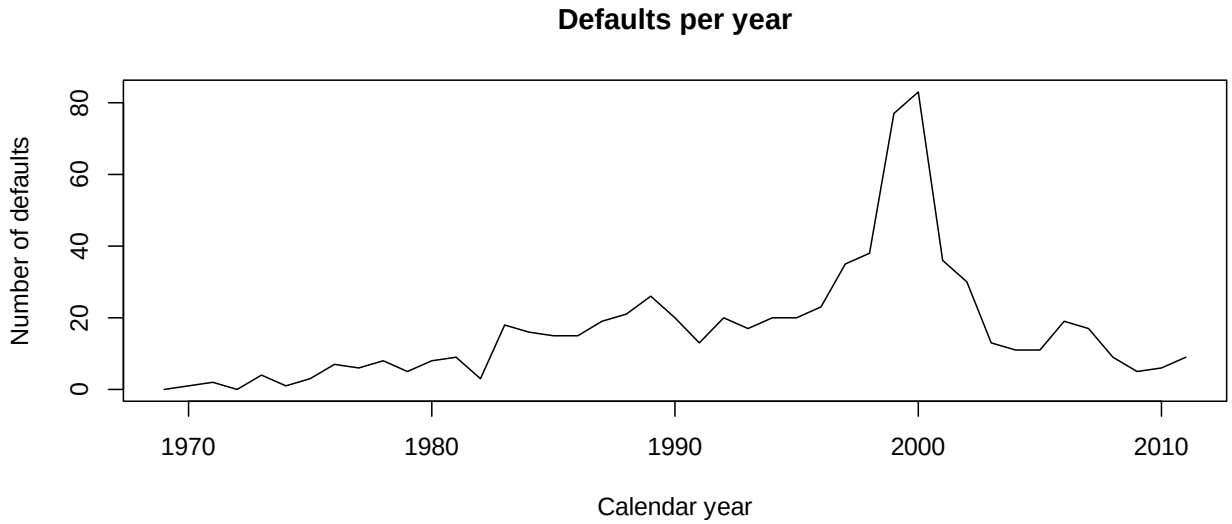
Following Shumway (2001) and Campbell, Hilscher and Szilagyi (2008), I also include NIMTA and EXR 1-year-lagged information (NIMTA_lag1 and EXR_lag1). When lagged information is not available, that value is substituted with NIMTA or EXR (i.e. when the lag is not available, I use the information from the same year).

4.7 Data Handling

The "raw" dataset resulting from the process described in the previous paragraph comprises 193,252 year-firms. However, the log-transformations (used to calculate LPRICE and SIZE) and the use of ratios resulted in Not Available (NA), Not A Number (NaN) and infinite values which were removed to allow the fitting of statistical models. In addition, as in the case of Shumway (2001), firms with SIC code between 6000 and 6500, characterizing companies that provide financial services, and between 6700 and 6800, characterizing holding companies, were removed. The final dataset included American, non-financial, publicly-listed companies from 1970 to 2011 and counted 111,788 year firms.

Plot 1 below shows number of defaults per year in the final dataset.

Plot 1*: Number of defaults per year as in the total sample that has been used.*



As it can be easily noted, the highest numbers of defaults in the sample coincide with the burst of the dot.com bubble in calendar years 1999 and 2000. As in the case of Campbell, Hislcher and Szilagyi (2008), the sample used in this dissertation sees the number of defaults increase in the 1980s and 1990s. A second peak is then

reached in 2008, though, since financial firms and private firms are not part of the sample, it is not as pronounced as the peak in 1999-2000.

A training set and a testing set were created by separating observations with DEF1Y=1 from observations with DEF1Y=0 (to ensure that the proportions of defaulting firms a non-defaulting firms is the same for both training and testing set). Given the high risk of overfitting incurred by using models with splines and the related need for ample out-of-sample performance diagnostics, a training set was constructed using two thirds of the defaulting observations and two thirds of the non-defaulting observations, all of which were selected at random. The testing set was then constructed using the remaining observations, one third of the defaulting group and one third of the non-defaulting group.

Extreme values, defined as the values belonging to the top and bottom 0.3% quantile, of NIMTA, CR and CD were removed. The resulting training set counts 73,603 year-firms, of which 429 are defaulting, whereas the resulting testing set counts 36,890 firm-years, of which 237 are defaulting.

Summary statistics for training set, testing set and the combined set (training + testing) are offered in Appendices 7, 8 and 9). Interestingly the average of EXR is negative and far from 0 (-12.7%). This should not be the case. However, in estimating the models that are presented in Part 3, EXR proved to be statistically significant for all of them, suggesting that the difference is probably due to a negative bias, rather than a miscalculation.

5. Part II: Econometric Models

5.1 Choice of covariates

The literature review presented in Part I offers extensive material supporting the choice of specific covariates. The contributions there presented were used to form expectations about the predictive power of specific firm characteristics and to select which ones of these characteristics should be considered for inclusions in the econometric models that are presented. In line with the empirical approach of this research, no variable was taken into consideration purely based on common sense or without any empirical evidence supporting its correlation with probability of default.

Following the results of earlier researches, probability of default is expected to be:

- Positively correlated with firms' asset volatility (measured by VOL).
- Positively correlated with leverage (measured by TLMTA).

- Negatively correlated with the current ratio (measured by CR).
- Negatively correlated with return on assets (measured by NIMTA).
- Negatively correlated with size (measured by SIZE).
- Correlated with age (measured by AGE).
- Correlated with positive excess returns on the companies' equity (measured by EXR).
- Correlated with the share of total assets that is liquid (measured by CD).

The presence of non-linearities is supported by the findings of Berg (2007), Acharya, Davydenko and Strebulaev (2012) and Giordani et al. (2014). Given the contributions of these scholars and the empirical nature of the problem, no assumptions regarding the sign of these correlations is made in this paper.

5.2 Economics of credit risk modelling

Before presenting the methodologies that have been adopted to estimate the econometric models used in this research, it is important to explain what a good credit risk model is and what makes a specification better than alternatives.

As for almost all statistical models, a good credit risk model has high predictive power. That is, it is characterized by a tendency to accurately identify defaulting firms before a credit event (default or bankruptcy) takes place, with minor errors. From an informational point of view, good statistical models capture some underlying aspects and characteristics of a DGP that allows user to draw intelligent conclusions and to forecast confidently. Thus, a credit risk model is considered to be better than suitable alternatives if it has a lower prediction error or higher accuracy.

However, from a practical point of view, in classifying corporations as either characterized by a high risk of default or by a low risk of default, users might find more useful a model that neatly separates the two groups. For instance, there is a range of values for which Altman's Z-score is associated with a grey area or medium probability of default. Models that fail to clearly separate firms with high probability of default from firms with low probability of default could result in confusion about the classification problem at hand. Should a corporation with a Z-score within the grey area range be classified as defaulting or non-defaulting?

To test hypothesis H_2, comparisons of model functional forms and specifications are presented. Consequently, measures of the first basis of comparison, that is predictive power, and of the second basis of comparison, that is the ability of neatly separating classes, were adopted. Most of the techniques used in this paper to compare the difference in predictive powers among the specifications being compared are discussed in the works of

Stein (2007) and/or adopted by several scholars such as Agarwaal and Tuffler (2008), Campbell et al. (2008) and Giordani et al. (2014). As for the second basis, the author of this paper finds in Shumway (2001) the most appropriate methodology to compare model specifications according to their ability to clearly separate the two classes. To keep this reading fluent, such methodologies are introduced in some detail right before being applied in Part 3.

The reader should note that, given that estimating the expected loss on a corporate loan implies the calculation of LGD, in addition to probability of default, no asset prices will be used to compare model specifications. Further, it should also be noted that Type I, false positives, and Type II errors, false negatives, in a practical context, do not have the same economic consequences for lenders. False negatives, that is, borrowers classified as non-defaulting when they default (false positives are the opposite), imply the loss of part of the principal being lent. As a result, Type II errors have much higher costs than Type I errors. This is also observed and stressed in Agarwaal and Taffler (2008). As in that article, one should take into account this asymmetry when comparing credit models. Finally, a credit analysis of any kind should identify high probability of default in a timely manner (Petersen, Plenborg, Kinserdal, 2017). As stated in Campbell, Hilscher and Szilagyi (2008), predicting corporate default a month before the event is not useful *"just as it would not be useful to predict a heart attack by observing a person dropping to the floor clutching his chest"* (Journal of Finance, Vol. LXIII, no. 6, pp. 2900). As a consequence, the input variables used in this research were publicly available one financial year before corporate bankruptcy takes place, which seems to be the common practice in academia (e.g. Ohlson, 1980; Zmijewski, 1984; Shumway, 2001; Giordani et al., 2014).

5.3 Model Specifications and Functional Forms

All the linear specifications that have been estimated as part of this research rely on the assumptions discussed in section 4.5. As mentioned there, testing the assumption on linearity is the objective of this research, whereas the assumption on independency and the assumption on the size of the sample are considered to be respected. However, whether the assumption on the absence of multicollinearity is also considered to be respected requires further analysis. Table 1 below shows the correlation matrix for the set of variables used to estimate the three models.

Table 1: *correlation matrix of training set*

	NIMTA	TLMTA	VOL	LPRICE	EXR	CR	CD	DEF1Y	AGE	SIZE
NIMTA	100,0000%	-2,5835%	-30,2141%	37,5859%	-0,6529%	-5,4964%	-12,5879%	-8,1953%	9,5022%	25,1231%
TLMTA	-2,5835%	100,0000%	-8,8801%	-18,7075%	1,8600%	-44,2375%	-43,1594%	9,7035%	9,6685%	-8,4103%
VOL	-30,2141%	-8,8801%	100,0000%	-42,7535%	12,1041%	12,5962%	17,9555%	5,7408%	-25,0911%	-38,3443%
LPRICE	37,5859%	-18,7075%	-42,7535%	100,0000%	2,8327%	-10,6697%	-12,3645%	-9,6754%	26,9403%	79,3064%
EXR	-0,6529%	1,8600%	12,1041%	2,8327%	100,0000%	-2,7567%	-2,2397%	-2,3570%	0,6365%	4,3015%
CR	-5,4964%	-44,2375%	12,5962%	-10,6697%	-2,7567%	100,0000%	82,6863%	-3,2375%	-14,0983%	-16,3996%
CD	-12,5879%	-43,1594%	17,9555%	-12,3645%	-2,2397%	82,6863%	100,0000%	-2,2812%	-15,3645%	-14,0612%
DEF1Y	-8,1953%	9,7035%	5,7408%	-9,6754%	-2,3570%	-3,2375%	-2,2812%	100,0000%	-1,7077%	-6,6073%
AGE	9,5022%	9,6685%	-25,0911%	26,9403%	0,6365%	-14,0983%	-15,3645%	-1,7077%	100,0000%	25,7403%
SIZE	25,1231%	-8,4103%	-38,3443%	79,3064%	4,3015%	-16,3996%	-14,0612%	-6,6073%	25,7403%	100,0000%

There is no clear correlation threshold above which two independent variables are said to be colinear and almost all variables are correlated to some extent (Wooldridge, 2019). In this dissertation, an arbitrary value of 0.5 was chosen as threshold. The decision of setting 0.5 correlation as the multicollinearity threshold rules out the joint inclusion of the pair of variables CR/CD, whose correlation is approximately 0.827. This poses some problems in terms of variable selection, as both variables have been extensively used in applied research. In fact, the current ratio is one of the main metrics used in univariate analysis, whereas the correlation between liquid assets (as a share of total assets) and probability of default has been documented in Campbell, Hilscher and Szilagyi (2008), Acharya, Davydenko and Strebulaev (2012) and Giordani et al. (2014). The solution adopted to solve this problem entailed the estimation of two model specifications per functional form: one including CR and excluding CD, the other including CD and excluding CR.

Two functional forms were adopted to estimate the models presented in this dissertation. The first is a dynamic (i.e. including lagged information) GLM, with binomial family and logit link, which represents the linear alternative. The second functional form is a dynamic additive model, with binomial family and logit link, that uses natural splines to augment the feature space X. This methodology is similar to the one applied in Giordani et al. (2014) and represent one of the two non-linear alternatives. Number of knots is selected by stepwise backward deletion and these are then placed at equidistant quantiles.

Given that three different functional forms are used to estimate models on two different sets of covariates, a total of four models was estimated.

5.4 Empirical results

Univariate spline regression

Before proceeding with the presentation of results obtained from the estimation of the four models introduced in Part 2, a preliminary univariate regression with natural splines is used to test H_1. In this section, univariate linear and non-linear models are compared to see if there is any notable difference in fit.

Therefore, in analytical form this could be expressed as

$$\log \frac{P(DEF1Y = 1|X)}{P(DEF1Y = 0|X)} = log\left(\frac{\mu(X)}{1 - \mu(X)}\right) = \alpha + f_i(x_i) \qquad (20)$$

where x_i is one among NIMTA, TLMTA, VOL, EXR, LPRICE, SIZE, CR and CD and f_i is a natural spline with 6 degrees of freedom (as it is shown below the relative superiority in fit displayed by natural splines is so apparent in some cases that there is no need to select number of knots by stepwise backward deletion).

The plots below offer a visual representation of the empirical results of univariate spline regression between probability of default and individual firm-level covariates, for each percentile of a sample that comprehend training and testing set. In other words, after ordering observations according to the magnitude of a specific independent variable, percentiles are formed and the average probability of default, as well as the average value of the specific covariate are calculated for each percentile. The resulting 100 observations are plotted on graphs where y is the average probability of default and x is the covariate in question. In every plot, the blue line represents a logistic regression with natural spline of 6^{th} order (degrees of freedom = 6), whereas the red line represents the function fitted by linear logistic regression. Plots were constructed using the R-package "ggplot" (Wickam, 2016), which allows to fit univariate GLM estimates, while the natural splines were fitted using both "ggplot" and "splines" (Venables and Bates).

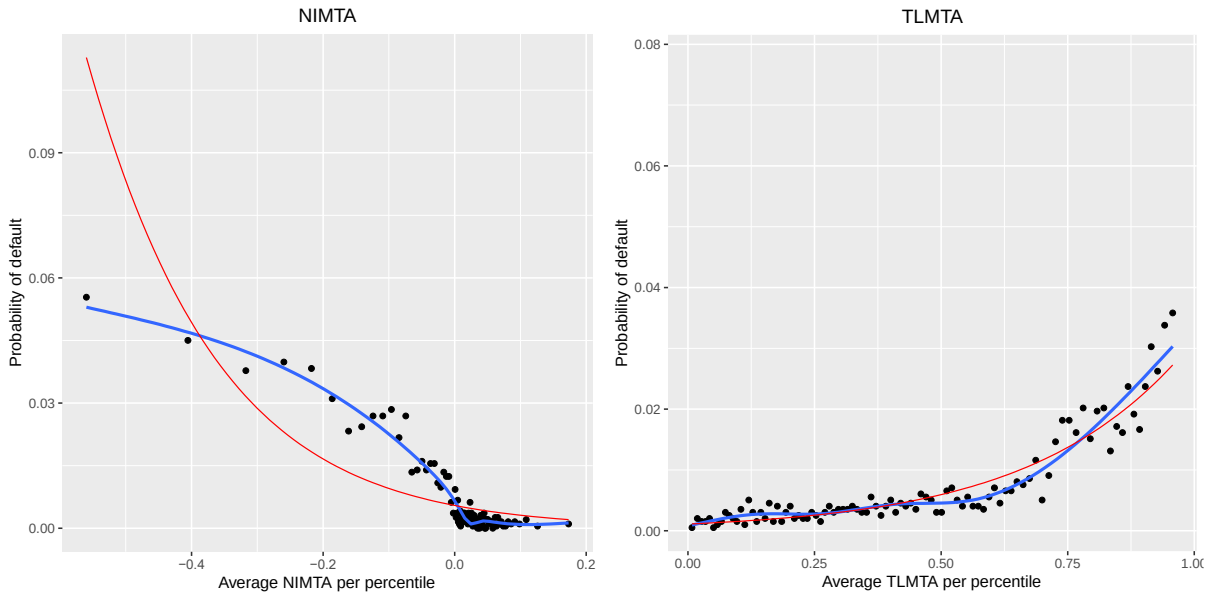
It should be stressed that these are univariate splines and do not represent the ceteris paribus relations between probability of default and independent variables. As a result, the empirical findings presented in this section can only be used to test H_1. However, if some of the univariate relations are non-linear, as they seem to be from the plots, it is likely that also some of the ceteris paribus relations are non-linear.

Plot 2 shows the univariate relation between average NIMTA and average probability of default. The differences between the function fitted using a logistic regression with natural spline and linear logistic regression is significant. In fact, the linear model exhibits a poor fit for almost the whole range of values that has been considered. For very low values of NIMTA (below -0.4 or -40%) the linear model vastly overestimates probability of default, while for values between -0.4 and 0 it underestimates it. Further, it seems to overestimate it also for positive values. The natural spline displays a much better fit, although it is probably

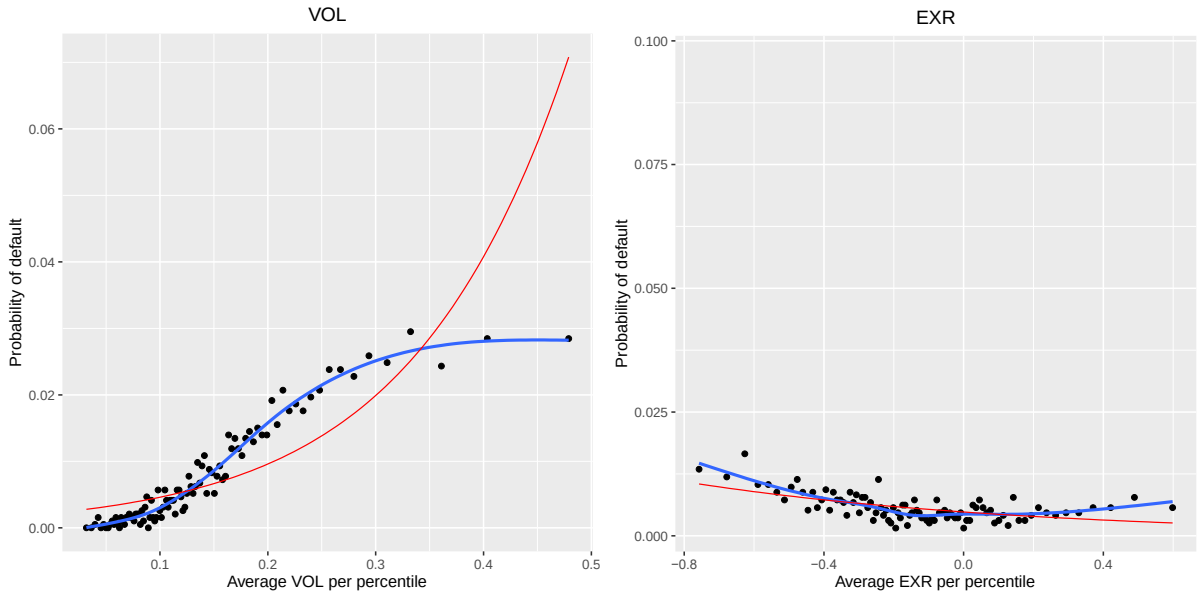
overfitted, as probability of default seems to increase for positive NIMTA values in the range 0.03 - 0.05. The natural spline captures also what seems to be a non-monotonic pattern, as probability of default remain stably close to 0 for all positive values.

Plot 3 shows the univariate relationship between probability of default and TLMTA. The difference between the two fits is not as striking as in the case of NIMTA. However, the natural spline seems to capture a non-monotonic pattern better than the linear model. In fact, the natural spline fitted in Plot 3 is relatively stable for low values of TLMTA, where probability of default and TLMTA do not appear to be correlated but increases definitely faster beyond TLMTA = 0.5 (leverage = 0.5/0.5 = 1). This is not the case for the linear model, which fits constantly increasing marginal effects. The linear model seems also affected by an upward bias approximately in the region 0.5 – 0.7, where it estimates a higher probability of default.

Plot 2 (on the left) and plot 3 (on the right): *Univariate logistic regression vs univariate logistic regression with natural splines. The blue line is fitted using a natural spline whereas the red line is fitted using a linear model.*



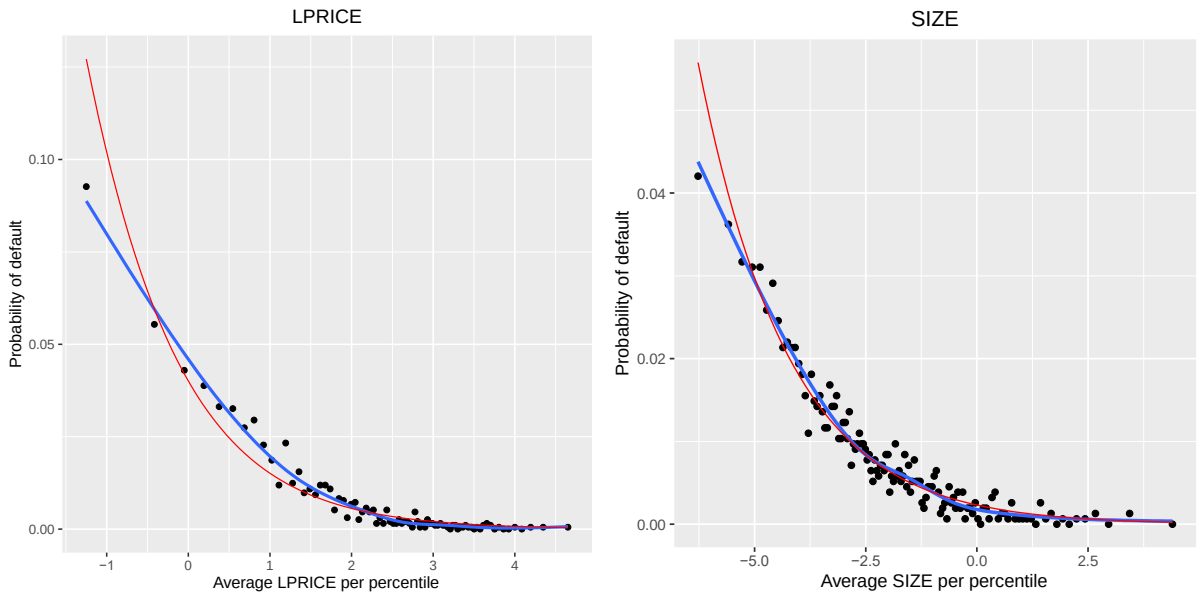
Plot 4 (on the left) and plot 5 (on the right): *Univariate logistic regression vs univariate logistic regression with natural splines. The blue line is fitted using a natural spline whereas the red line is fitted using a linear model.*



Another remarkable difference in terms of fit between the linear univariate logit regression and the one with natural spline can be observed in Plot 4, which displays the univariate relationship between probability of default and volatility. The linear logit model displays a very poor fit in this case and misestimates probability of default for almost the entire range of VOL values. It is interesting that, as in the case of NIMTA, the two curves, red and blue, have opposite second order derivatives, except for low values of VOL or high values of NIMTA. In addition, note that the linear model estimates an exponential trend where the non-linear estimates a logarithmic one (VOL > 0.15), which represents a substantial difference in fit.

This is not the case for EXR, shown in Plot 5. There the two fits are similar, though the natural spline seems to better estimate probability of default at extreme EXR values. However, it should be noted that, again, the spline captures a non-monotonic pattern that the linear model fails to account for. In fact, probability of default appears to increase at very positive values of EXR (above 0.3).

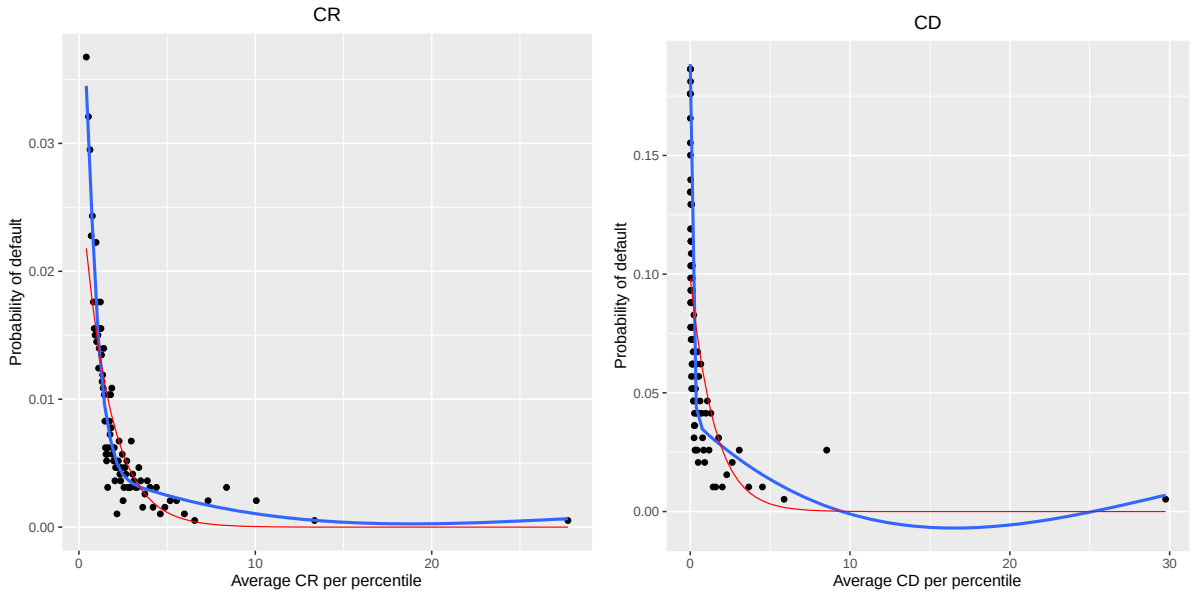
Plot 6 (on the left) and plot 7 (on the right): *Univariate logistic regression vs univariate logistic regression with natural splines. The blue line is fitted using a natural spline whereas the red line is fitted using a linear model.*



Plots 6 and 7 show, respectively, the univariate relationship between probability of default and LPRICE and the univariate relationship between probability of default and SIZE. In the case of LPRICE, the natural spline displays a trend that is highly similar to the of the linear function, except for the range [0 – 1.5], where the spline displays better fit. In fact, there is almost no difference in fit for LPRICE values above 1.5 and below 0. In the case of plot 7, differences between the two univariate relationships are difficult to spot. In fact, except for extremely low values of SIZE, the two functions are essentially the same.

Finally, Plot 8 and Plot 9 shows the univariate relationships between probability of default and CR and probability of default and CD, respectively. There is a clear similarity among the functions estimated in the two plots, which is in line with expectations. However, Plot 9 clearly suffers from a lack of sparseness in the covariate space as many observations share the same values both in terms of probability of default and CD. In addition, the natural spline is undoubtedly overfitted, since probability of default is negative for some specific values of CD. Plot 9 might indeed represent the only plot where the linear fit should be preferred over the non-linear. In the case of Plot 8, the natural spline displays a very good fit, especially for very low values of CR. However, it is not significantly different from the one of the linear model.

Plot 8 (on the left) and plot 9 (on the right): *Univariate logistic regression vs univariate logistic regression with natural splines. The blue line is fitted using a natural spline whereas the red line is fitted using a linear model.*



The plots just described provide evidence supporting the non-rejection of hypothesis H_1. In fact, it seems as if some univariate relationships between probability of default and firm-level covariates, most notably the ones shown in Plot 2 and 4, are non-linear. Further, they also offer some evidence concerning the risk of overfitting and how non-linear models can be misused. Finally, given that some of the univariate relationships between probability of default and firm-level covariate are non-linear, and even non-monotonic in some particular cases, plots 2 – 9 suggest that also the ceteris paribus relationships between firm-level covariates and probability of default ight be non-linear.

Econometric models

In this section, empirical results obtained from the estimation of models are presented and described in some detail. These are then compared to test H_2.

The linear model

As introduced in section 4.5, two binomial GLM models with logit link were estimated. Both specifications were fitted using the function “glm” from the R-package “stats”, which is provided with the basic installation of R software version 3.6.1 (RDocumentation.org).

An unrestricted logit model, Logit 1U was fitted to the training set, selecting as independent variables NIMTA, TLMTA, VOL, CR, LPRICE, EXR, NIMTA_lag1, EXR_lag1, SIZE, AGE and as dependent variable DEF1Y. The model can then be represented by the formula

$$P(DEF1Y = 1|\,\mathbf{X}) = G\,(\boldsymbol{\beta}^{T}\mathbf{X}) = \frac{exp(\boldsymbol{\beta}^{T}\mathbf{X})}{1 + exp(\boldsymbol{\beta}^{T}\mathbf{X})} \quad where \quad \boldsymbol{\beta}^{T}\mathbf{X} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1\, NIMTA \\ \hat{\beta}_2\, TLMTA \\ \hat{\beta}_3 VOL \\ \hat{\beta}_4\, CR \\ \hat{\beta}_5\, LPRICE \\ \hat{\beta}_6\, EXR \\ \hat{\beta}_7\, NIMTA\_lag1 \\ \hat{\beta}_8\, EXR\_lag1 \\ \hat{\beta}_9\, SIZE \\ \hat{\beta}_{10}\, AGE \end{bmatrix} \tag{21}$$

The second unrestricted logit model (Logit 2U) was fit to the same training set used to fit Logit 1U, with the only exception that the independent variable CR was changed with CD. Thus, formula for Logit 2U is

$$P(DEF1Y = 1|\,\mathbf{X}) = G\,(\boldsymbol{\beta}^{T}\mathbf{X}) = \frac{exp(\boldsymbol{\beta}^{T}\mathbf{X})}{1 + exp(\boldsymbol{\beta}^{T}\mathbf{X})} \quad where \quad \boldsymbol{\beta}^{T}\mathbf{X} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1\, NIMTA \\ \hat{\beta}_2\, TLMTA \\ \hat{\beta}_3 VOL \\ \hat{\beta}_4\, CD \\ \hat{\beta}_5\, LPRICE \\ \hat{\beta}_6\, EXR \\ \hat{\beta}_7\, NIMTA\_lag1 \\ \hat{\beta}_8\, EXR\_lag1 \\ \hat{\beta}_9\, SIZE \\ \hat{\beta}_{10}\, AGE \end{bmatrix} \tag{22}$$

Table 2 below shows summary statistics, coefficients estimate, standard error, z-score and p-values for Logit 1U and Logit 2U.

In model Logit 1U almost all variables have signs as presented in Campbell, Hilscher and Szilagyi (2008). The only exception is SIZE whose sign is positive, implying that firms' probability of default increases with firms' size.

Table 2: *Summary statistics for models Logit 1U and Logit 2U*

	Logit 1U				Logit 2U			
	Estimate	Std. Error	Z value	P-value	Estimate	Std. Error	Z value	P-value
Intercept	-6,384774	0,318191	-20,066	0.0000000000000002	-6,62807	0,30871	-21,471	0,0000000000000002
NIMTA	-0,225664	0,097961	-2,304	0.02124	-0,23949	0,09846	-2,432	0,015
TLMTA	4,290933	0,268254	15,996	0.0000000000000002	4,39916	0,28012	15,705	0,0000000000000002
VOL	1,409959	0,323821	4,354	0.000013359647658189	1,45854	0,32095	4,544	0,0000055090
CR	-0,133063	0,048112	-2,766	0.00568				
CD					-0,13485	0,09607	-1,404	0,16
LPRICE	-0,556992	0,068565	-8,124	0.000000000000000453	-0,57732	0,06911	-8,354	0,0000000000000002
EXR	-0,763647	0,143862	-5,308	0.000000110709913525	-0,77146	0,14368	-5,369	0,000000079
NIMTA_lag1	-0,037534	0,123335	-0,304	0.76088	-0,04428	0,12341	-0,359	0,72
EXR_lag1	-0,376023	0,149521	-2,515	0.01191	-0,38148	0,14996	-2,544	0,011
SIZE	0,011577	0,042626	0,272	0.78593	0,02258	0,04276	0,528	0,597
AGE	-0,007279	0,006704	-1,086	0.27756	-0,007059	0,006718	-1,051	0,293
AIC =	4226,40				AIC =	4233,1		

However, the estimate of the coefficient of SIZE is not statistically significant at conventional levels. This is also the case of NIMTA_lag1 and AGE. The likelihood ratio test (Formula in Appendix 10) was carried out to test the joint exclusion of NIMTA_lag1, SIZE and AGE. This test compares the log-likelihood of two nested models to see if there is any statistically significant difference in fit (Wooldridge, 2019). Thus, the null hypothesis H_0 is

$$H_0: \boldsymbol{\beta}_{Logit\ 1U} = \boldsymbol{\beta}_{Logit\ 1} \tag{23}$$

where Logit 1 is the logit model estimated excluding the three variables. In other words, the null hypothesis states that there is no significant difference between the parameters estimated in the unrestricted model and the parameters estimated in the unrestricted one. The likelihood ratio test resulted in an insignificant difference (value of 0.71), implying that we fail to reject hypothesis H_0. The three variables were therefore dropped. This decision was further supported by the AIC, which is presented for unrestricted and restricted models in the respective summary statistics in table 2 and 3 (summary statistics for the likelihood ratio test can be found instead in appendix 11).

Size, Age and Nimta_lag1 were also found to be statistically non-significant, at conventional levels, in model Logit 2U. Surprisingly, also CD resulted to have a p-value of 0.16. This result is contrary to the findings presented in Campbell, Hilscher and Szilagyi (2008). The likelihood ratio test was carried out (summary in Appendix 12) to test the joint exclusion of CD, NIMTA_lag1, SIZE and AGE from the model specification

and resulted again in an insignificant difference (value of 0.41), leading to the exclusion of the four variables from Logit 2. Again, this decision was supported by the AIC.

Given that Logit 2 and Logit 1 are nested, with Logit 2 being a more parsimonious specification of Logit 1, since the latter includes CR in addition to the variables in the former, Logit 2 was ultimately dropped in whole. Summary statistics for model Logit 1 can be found in the table 3 below.

Table 3: *Summary statistics for model Logit 1 and Logit 2 (which was ultimately abandoned).*

	Logit 1				Logit 2			
	Estimate	Std. Error	Z value	P-value	Estimate	Std. Error	Z value	P-value
Intercept	-6,484650	0,270520	-23,971	0.0000000000000002	-6,95158	0,22167	-31,360	0,0000000000000002
NIMTA	-0,246730	0,087400	-3,056	0.00224	-0,25009	0,08016	-3,120	0,00181
TLMTA	4,275970	0,268226	15,940	0.0000000000000002	4,59844	0,24536	18,742	0,0000000000000002
VOL	1,430150	0,321140	4,453	0.0000084550	1,45891	0,31793	4,589	0,0000044586
CR	-0,132720	0,047950	-2,768	0.00564				
LPRICE	-0,552730	0,044950	-12,297	0.0000000000000002	-0,54818	0,04524	-12,116	0,0000000000000002
EXR	-0,764340	0,143160	-2,339	0.0000000934	-0,76658	0,14273	-5,371	0,0000000784
EXR_lag1	-0,374310	0,149450	-2,505	0.01191	-0,37661	0,14976	-2,515	0,01191
AIC =	4221,80				AIC =	4229		

Additive logistic regression with natural splines

The second pair of models, Logit 1N and Logit 2N (the N stands for Natural splines), was estimated on the same training set used for Logit 1 and Logit 2, respectively. Natural splines were adopted to estimate non-linear functions between probability of default and each variable. So that the two models could be represented with the formula:

$$NS\,1 = G(\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{HX}) = \frac{exp(\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{HX})}{1 + exp(\boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{HX})} \quad where \quad \boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{HX} = \begin{bmatrix} \hat{\beta}_0 \\ \widehat{\boldsymbol{\beta}}_{\mathbf{1}}\boldsymbol{h}_{\mathbf{1}}\,NIMTA \\ \widehat{\boldsymbol{\beta}}_{\mathbf{2}}\boldsymbol{h}_{\mathbf{2}}\,TLMTA \\ \widehat{\boldsymbol{\beta}}_{\mathbf{3}}\boldsymbol{h}_{\mathbf{3}}\,VOL \\ \widehat{\boldsymbol{\beta}}_{\mathbf{4}}\boldsymbol{h}_{\mathbf{4}}\,CR/CD \\ \widehat{\boldsymbol{\beta}}_{\mathbf{5}}\boldsymbol{h}_{\mathbf{5}}\,LPRICE \\ \widehat{\boldsymbol{\beta}}_{\mathbf{6}}\boldsymbol{h}_{\mathbf{6}}\,EXR \\ \widehat{\boldsymbol{\beta}}_{\mathbf{7}}\boldsymbol{h}_{\mathbf{7}}\,NIMTA\_lag1 \\ \widehat{\boldsymbol{\beta}}_{\mathbf{8}}\boldsymbol{h}_{\mathbf{8}}\,EXR\_lag1 \\ \widehat{\boldsymbol{\beta}}_{\mathbf{9}}\boldsymbol{h}_{\mathbf{9}}\,SIZE \\ \widehat{\boldsymbol{\beta}}_{\mathbf{10}}\boldsymbol{h}_{\mathbf{10}}\,AGE \end{bmatrix} \tag{24}$$

H = $(\boldsymbol{h_1}, \boldsymbol{h_2}, \ldots, \boldsymbol{h_{10}})$ is a vector of natural splines and each $\boldsymbol{h_i}$ (i = 1, 2, …, 10) is in turn a vector of basis functions $h_{i1}, h_{i2}, \ldots, h_{ij}$, where j depends on the degrees of freedom of the natural cubic spline. Thus, **H** is a matrix of basis expansions. In addition, each $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ (i =1, 2, … , 10) is a vector of coefficients $\hat{\beta}_{i1}, \hat{\beta}_{i2}, \ldots, \hat{\beta}_{ij}$, each multiplying a specific basis function h_{ij}. To avoid repetitions, I used the notation CR/CD to indicate that CR was included in Logit 1N, and CD in Logit 2N. The R-package "gam" (Hastie, 2019) was used to fit Logit 1N and 2N.

To select model specifications, I adopted the same approach set out in Giordani et al. (2014). That is, in estimating Logit 1N and 2N, the degrees of freedom were held the same for each natural spline and the resulting AIC statistics were compared to select the functional form. Once that the functional form is selected, each variable was removed, and the resulting AIC statistics were again compared to identify the best model specification.

Stepwise backward deletion resulted in the selection of natural splines with 5 degrees of freedom for each variable in both additive models, implying that the number of basis expansions, as well as the number of coefficients, equals five for each spline. SIZE and AGE were excluded in model Logit 1N, whereas SIZE, AGE and Nimta_lag1 were excluded from Logit 2N. The beta coefficients and AIC of Logit 1N and Logit 2N are presented in tables 4 and 5 below.

Marginal effects

As mentioned in section 4.5, in a logit model the beta coefficients do not directly represent the partial marginal effects that each independent variable has on the dependent. Therefore, partial marginal effects need to be estimated using the first-order partial derivative with respect to each variable, at specific values of **X**.

Table 4: *coefficients and AIC of model Logit 1N*

	Logit 1 N									
	NIMTA					TLMTA				
basis	1	2	3	4	5	1	2	3	4	5
estimate	34,31408	33,23093	23,27787	-30819262	-749,46358	1,64172	1,42051	1,6435	4,59167	3,60518
	VOL					CR				
basis	1	2	3	4	5	1	2	3	4	5
estimate	2,53339	22,68274	5,33632	-24,65305	-59,32317	-0,12913	-0,31358	-172,51754	-91,7463	-11,61702
	LPRICE					EXR				
basis	1	2	3	4	5	1	2	3	4	5
estimate	1,55558	0,41515	-1,69186	-0,39231	-7,27305	0,44573	0,2775	-2,87326	7,9968	15,66418
	EXR_lag1					NIMTA_lag1				
basis	1	2	3	4	5	1	2	3	4	5
estimate	1,85034	1,22081	1,04382	1,61947	9,62347	-0,58787	-0,55091	-5,26181	-7,08157	-6,00501
					AIC =	3638,376				
					Intercept =	-46,817				

Table 5: *coefficient and AIC statistics for model Logit 2N*.

	Logit 2 N									
	NIMTA					TLMTA				
basis	1	2	3	4	5	1	2	3	4	5
estimate	33,26188	32,20235	19,94397	-261,30227	-651,72445	1,01329	0,63391	0,93903	3,32656	2,98835
	VOL					CD				
basis	1	2	3	4	5	1	2	3	4	5
estimate	2,50487	2,69458	5,37363	-24,58616	-59,2931	-0,68523	-1,31082	-116,16471	-70,27283	-23,33359
	LPRICE					EXR				
basis	1	2	3	4	5	1	2	3	4	5
estimate	1,63281	0,4386	-1,65399	-0,32748	-7,29807	0,41586	0,25409	-2,93351	7,92321	15,64388
	EXR_lag1									
basis	1	2	3	4	5		AIC =	3612,84		
estimate	1,71345	1,04856	-0,58904	-0,03273	-3,70128		Intercept =	-45,0023		

Deciding to use mean values for each variable in **X** coincides with the estimation of Average Marginal Effects (AME) (Wooldridge, 2019). However, given that the average probability of default of the sample at hand is already a very low value, computing partial marginal effects could result in little information regarding what drives corporate bankruptcy. This is due to the concept of marginal effect, that is, the change in independent variable due to a one-unit-increase change in the independent. As an example that could be used to clarify what is just said, one can imagine the very limited extent to which an increase in earnings of 1 USD affects probability of default. Obviously, this example is extreme and does not fully apply to the case of this research, since returns on assets, and not earnings, are used. Yet, it gives the idea.

Moreover, given that, in a logit model, partial marginal effects are not constant, AME provide information regarding only a specific level of **X**. Since the main objective of this research is to explore if and how partial effects change across an interval of values, AME do not represent an attractive way to convey information regarding what drives corporate default.

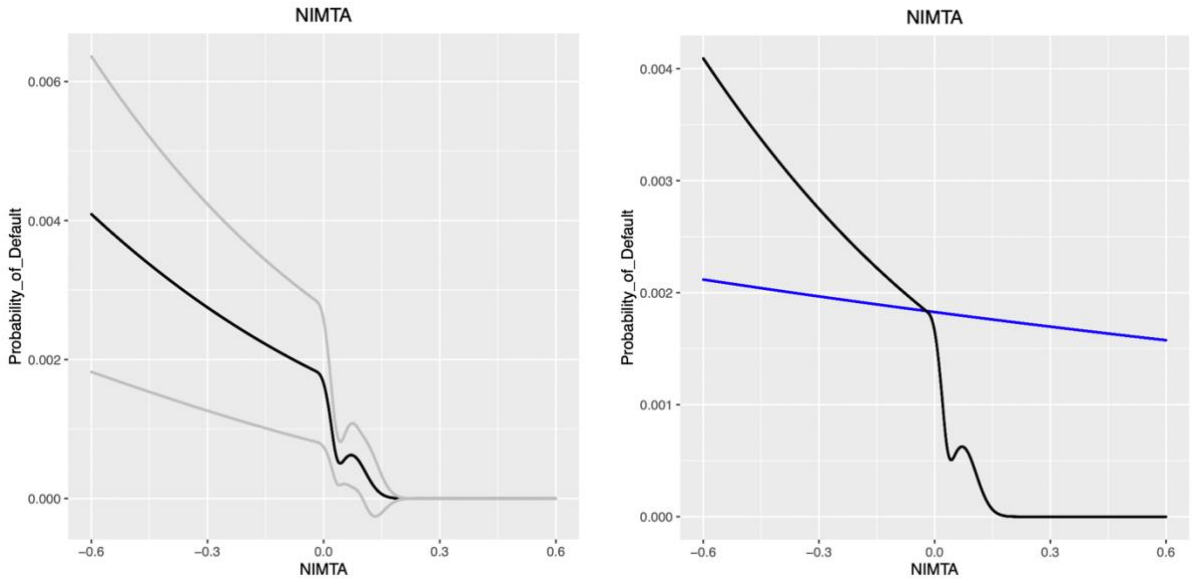
As an alternative, I adopt a methodology presented in Giordani et al. (2014), which allows to plot the whole function of partial effects, across a large interval of values. The logic is to visualize the conditional functions for each independent variable. Given that these functions represent the partial (ceteris paribus) effects of a specific independent variable on probability of default, one way to estimate them is to hold constant all independent variables except one, which varies across a range of values of interest.

Thus, the plots below were created on artificial datasets, constructed in the following way: for a specific interval of values, which differs depending on the variable, 10,000 equidistant observations are created; the resulting ten 1 X 10,000 vectors (one for each variable) are then used to create ten artificial datasets, where the values of one specific variable at the time corresponds to one of the 1X10,000 vectors created, while all other independent variables have

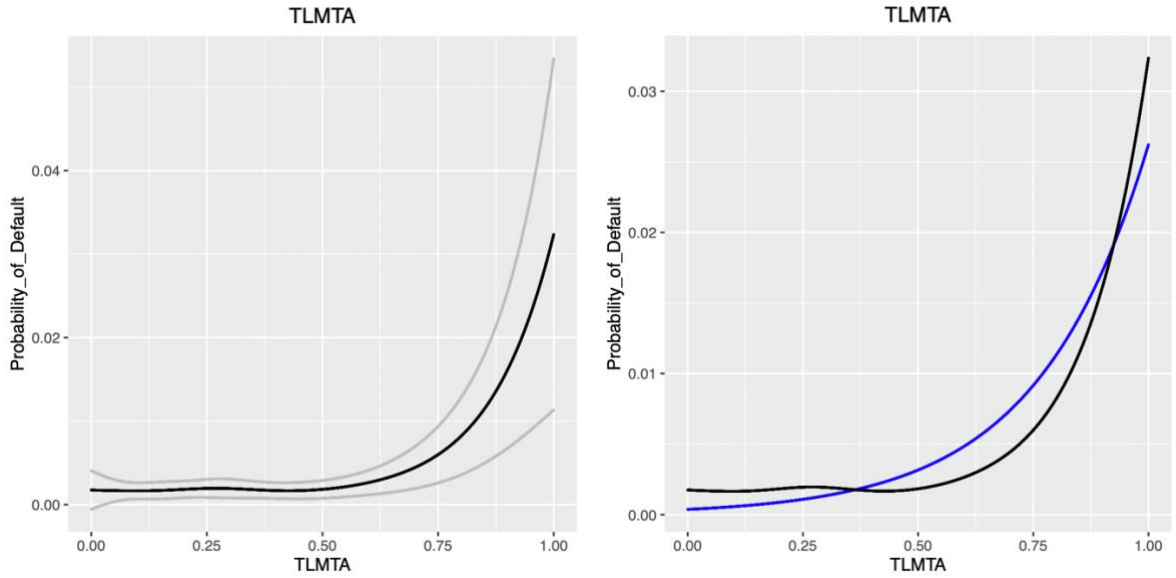
a value corresponding to the arithmetic mean of that variable. Models Logit 1N and Logit 2N are then used to estimate probability of default using the ten 10 X 10,000 datasets. The resulting estimates are finally plotted. The conditional functions fitted by both the linear and non-linear model are presented. 95% confidence intervals are also presented for the conditional functions fitted by the non-linear models. These were computed by adding 1,96*the standard error estimated by the function "gam" as part of the fitting process. Unless otherwise specified, the plots below were created using model Logit 1N. Plots created with Logit 2N are in Appendix 13, except for the plot of the conditional mean function of CD. As it can be observed, in some cases the conditional mean function estimated with natural splines differ substantially from the one fitted by the linear model. The case of NIMTA (Plots 10 and 11) particularly exemplifies this. In fact, the NIMTA conditional mean function fitted by the Logit 1N not only is highly non-linear, with clear changes in trends at NIMTA = 0 and NIMTA = 0.16, but also non-monotonic, which represent a puzzling finding further discussed in section 6.1. Naturally, this does not necessarily mean that also the underlying function displays this form, since the black line in Plots 10 and 11 is only an estimate. The conditional mean function of NIMTA, together with TLMTA and, to a lesser extent, VOL are highly similar to the univariate natural spline logistic regression fitted for the same variables in Plots 2, 3 and 4.

TLMTA seems to be, potentially, the major driver of corporate default. This can be observed both from the highest value of probability of default estimated by Logit 1N, which is 3%, or more than four times the average

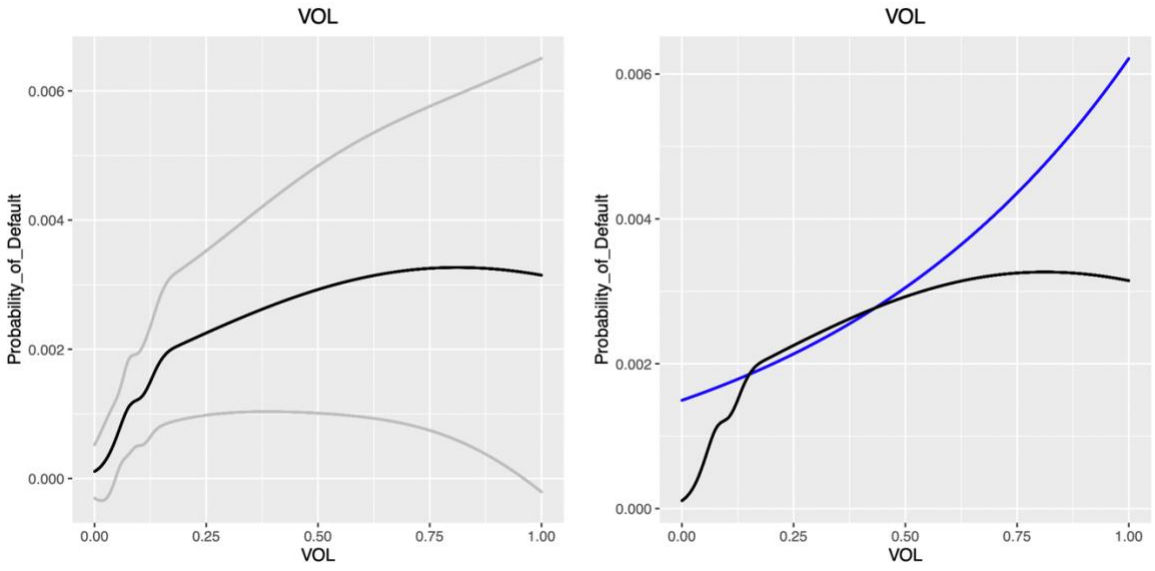
Plot 10 (on the left) and 11 (on the right): *The black line is the conditional mean function of NIMTA fitted by Logit 1N. The grey lines represent upper and lower boundaries of the 95% confidence interval. The blue line is the conditional mean function fitted by Logit 1.*



Plot 12 (on the left) and 13 (on the right): *The black line is the conditional mean function of TLMTA fitted by Logit 1N. The grey lines represent upper and lower boundaries of the 95% confidence interval. The blue line is the conditional mean function fitted by Logit 1.*



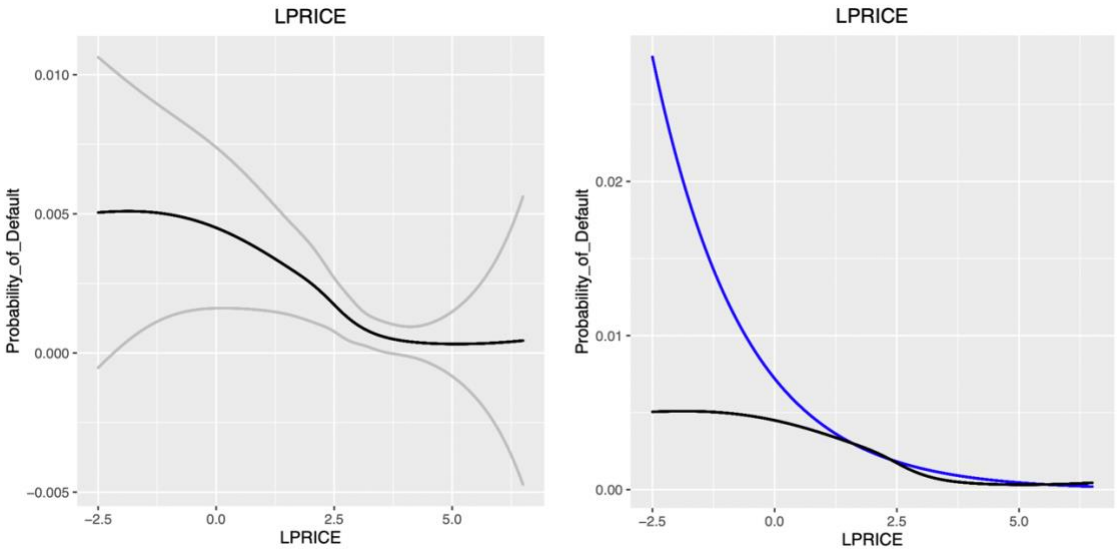
Plot 14 (on the left) and 15 (on the right): *The black line is the conditional mean function of VOL fitted by Logit 1N. The grey lines represent upper and lower boundaries of the 95% confidence interval. The blue line is the conditional mean function fitted by Logit 1.*



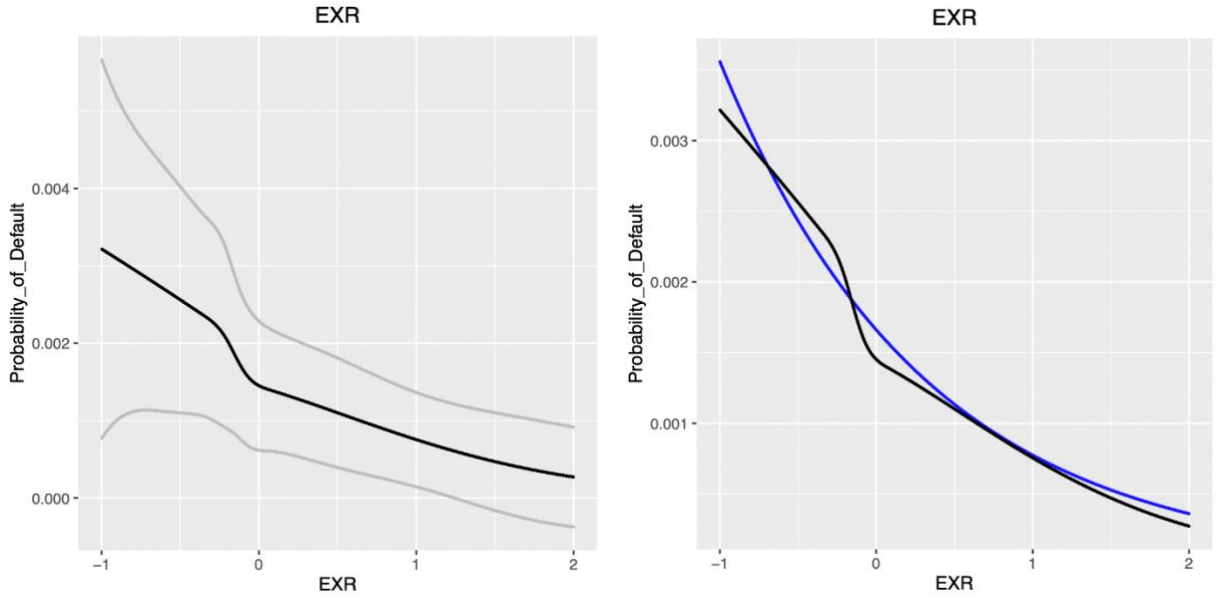
probability of bankruptcy. This is especially striking if one takes into consideration that, despite a TLMTA value of approximately 1 associated with that probability of default, all the values of the other variables are set at their arithmetic means. This finding is in line with the findings presented in Giordani et al. (2014). As in the case of the univariate plot, the natural spline seems to capture a relatively stable pattern where leverage is below 1 (TLMTA = 0.5), where the linear model fits a constantly increasing trend. In addition, the confidence interval is also remarkably narrow in that range.

It is interesting that the partial marginal effects on probability of default, as fitted by model Logit 1N, seem to turn negative for high values of VOL. However, it should be noted that the 95% confidence interval widens significantly in that region, indicating a low quality of estimates. Further, there is a clear difference between the constant term fitted by the linear model and the one fitted by the non-linear model. In fact, for VOL = 0, the natural spline seems to imply a very low probability of default, whereas the linear model fits a probability of approximately 0.015%. Moreover, both in the cases of VOL and LPRICE, the linear model fits an exponential trend where the non-linear model fits a seemingly reverting one. This happens for low values of LPRICE and, as just mentioned, for high values of VOL. However, also in the case of LPRICE the trend is stable where the confidence interval widens. Hence, these patterns should not be relied too heavily on. The partial effect of LPRICE on probability of default stabilizes also beyond a value of approximately 3.75. This trend is captured by both the linear and the non-linear models.

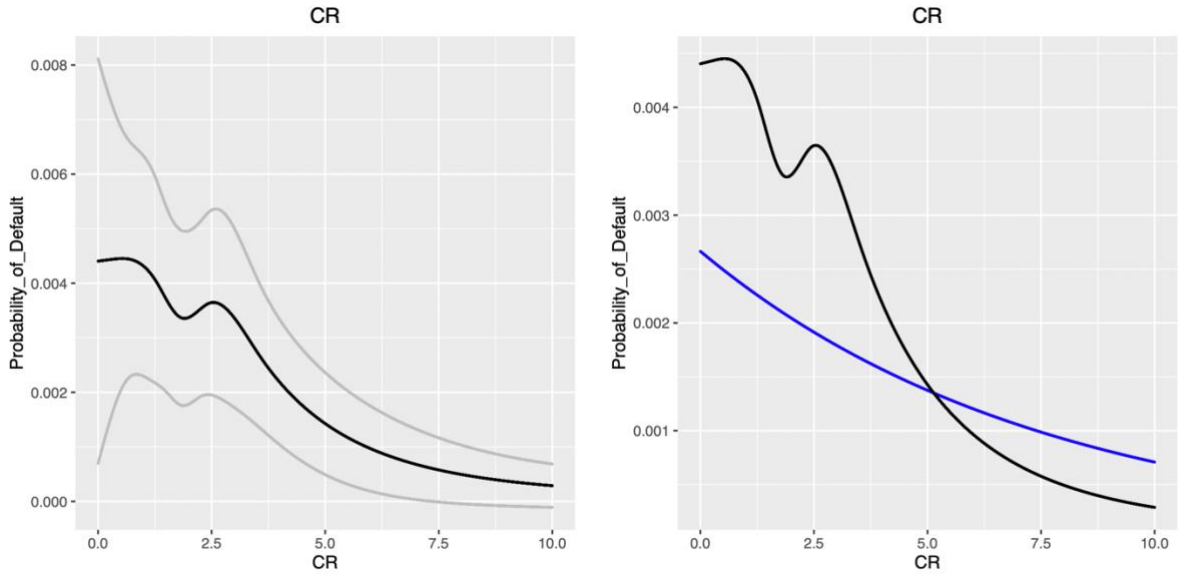
Plot 16 (on the left) and 17 (on the right): *The black line is the conditional mean function of LPRICE fitted by Logit 1N. The grey lines represent upper and lower boundaries of a confidence interval. The blue line is the conditional mean function fitted by Logit 1.*



Plot 18 (on the left) and 19 (on the right): *The black line is the conditional mean function of EXR fitted by Logit 1N. The grey lines represent upper and lower boundaries of the 95% confidence interval. The blue line is the conditional mean function fitted by Logit 1.*



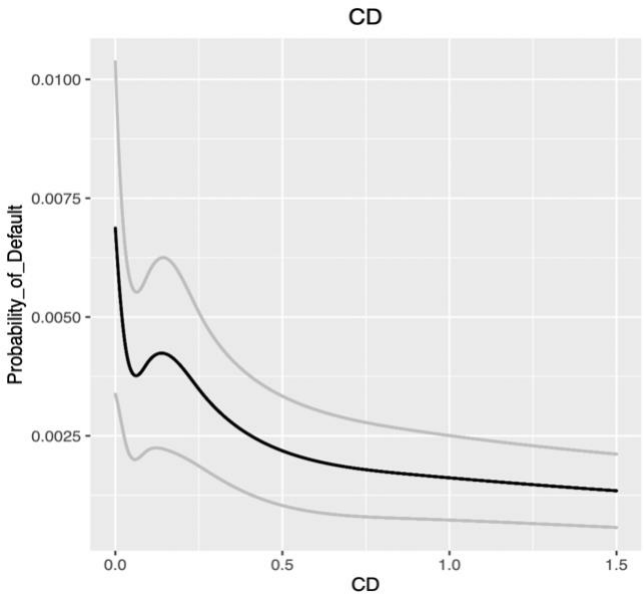
Plot 20 (on the left) and 21 (on the right): *The black line is the conditional mean function of EXR fitted by Logit 1N. The grey lines represent upper and lower boundaries of the 95% confidence interval. The blue line is the conditional mean function fitted by Logit 1.*



The difference in conditional mean functions for the variable EXR is not pronounced, though the non-linear model estimates a partial effect whose pattern rapidly changes, featuring two kinks at EXR =-0.2 and EXR = 0.

Finally, the non-linear estimates of the ceteris paribus relationship of CR and CD on probability of default share some features. Particularly, they both display non-monotonic trends, with estimated marginal partial effects changing sign at mid-to-low values (CR= 1.9 and CD=0.05). This appears to be counter intuitive and probably due to overfitting. However, whereas for extremely low values of CR probability of default stabilizes, it grows almost asymptotically at extremely low levels of CD. Once again, the difference in fit between the conditional mean function of CR fitted by the linear model and the one fitted by the non-linear model is significant. This comparison could not be offered for CD, as this variable was not included in the linear model (note that the function was estimated using model Logit 2N).

Plot 22: *The black line is the conditional mean function of CD fitted by Logit 2N. The grey lines represent upper and lower boundaries of a confidence interval.*



5.5 Model comparison: in-sample fit

Having presented the four model specifications that were estimated and selected, this section compares their goodness of fit using multiple approaches and different measures.

In-sample contingency tables and accuracy

Using Stern's terminology (Stern, 2007), a contingency table (also known as confusion matrix) is a 2 X 2 table showing number of true positives, true negatives, false positives and false negatives. The accuracy of a model can be defined as the number of correctly classified observations as a percentage of total number observations. Confusion matrices were constructed for each of the models being compared using the sample on which these were estimated. In doing so a cutoff point, such as the one introduced in Ohlson (1980), had to be established. As mentioned in section 3.6, binary classification requires to select a value above which an observation is classified as 1, and below which an observation is classified as 0. This was selected to be 0,7%, which is approximately the average probability of default in the sample that has been used. The resulting tables can be observed in Appendix 14.

The accuracy of each model was then computed as the sum of the values in the upper left and lower right cell, divided by the sum of values in all cells. The process just described resulted in the following ranking:

- Logit 2N (accuracy = 84.626%)
- Logit 1N (accuracy = 84.137%)
- Logit 1 (accuracy = 82.586%)

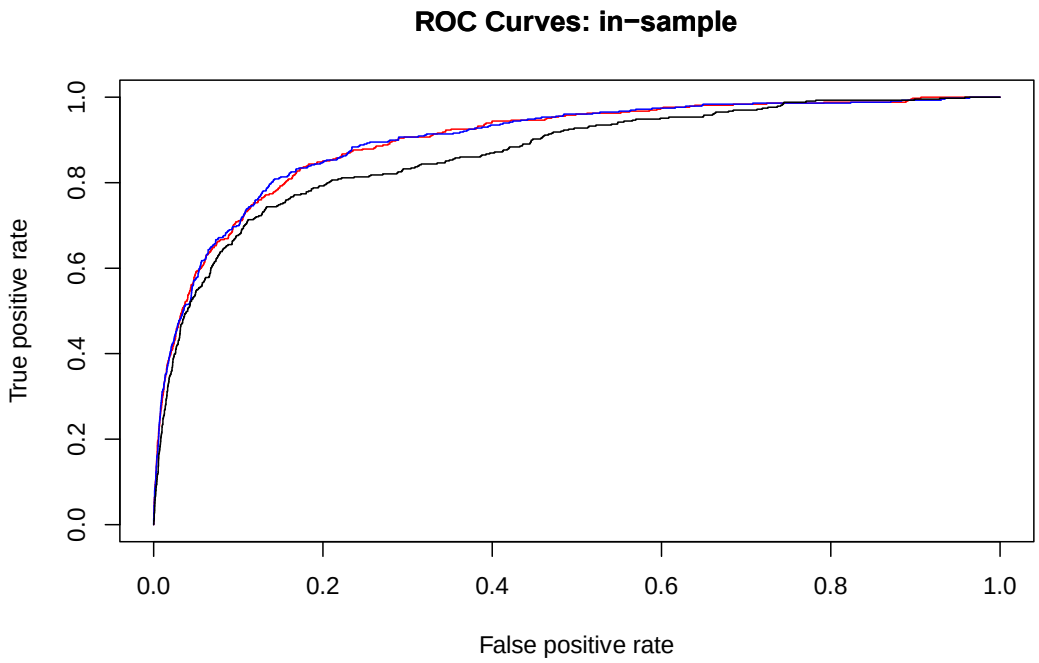
Model accuracy seem to further support hypothesis H_2, as the linear model ranks last. Logit 2N has the highest accuracy, though Logit 1N has accuracy very close to the one of the non-linear alternative.

ROC curves

Relative (or Receiver) Operating Characteristic (ROC) curves generalize the framework underlying confusion matrices by showing the performance of a model at any cutoff point. A ROC curve represents the percentage of true positives, on the Y-axis, as a function of the percentage of false positives (on the x-axis) (Stein, 2007). The relationship between ROC curves and contingency tables can be observed in Appendix 15. A binary basic model has an error rate of 50%, since it is no better than a random guess, and will therefore have a ROC curve equivalent to a straight line with slope = 1. On the other hand, a perfect model has a ROC curve identical to two orthogonal straight segments intersecting at point (0; 1). The Area Under the Curve (AUC) can be used to compare two models. This is usually calculated as a fraction of the area under the curve of a perfect model, which has an AUC if 1. If the curve of a model lies regularly above the curve of another model, then the former can be safely regarded as more accurate than the latter.

ROC curves for the four model specifications are presented in plot 23 below.

Plot 23*:in-sample ROC curves for Logit 1 (in black), Logit 1N (in red) and Logit 2N (in blue)*



Further, comparing the model specifications by AUC resulted in the following ranking:

- Logit 2N (AUC = 89.970%)
- Logit 1N (AUC = 89.874%)
- Logit 1 (AUC = 86.772%)

Both the comparison based on AUC and the plot showing the ROC curves provide substantial evidence of the better in-sample fit of the non-linear specifications.

Given the ROC curves of the non-linear specifications, all lie above the black one, representing the ROC curve of the linear specification, and that all the comparisons of measures of in-sample-fit lead to the same result, it can be concluded that the non-linear specifications have better in-sample fit. However, this does not necessarily mean that the non-linear specifications have higher predictive power as well. This is tested in the following section.

Accuracy vs parsimony: AIC and BIC

As mentioned in section 4.5, the AIC and BIC measure the in-sample fit by comparing log-likelihood, whilst penalizing the inclusion of additional terms. Despite the fact that both statistics do not directly test the out-of-sample performance of a model, it is legitimate to expect a model with lower AIC or BIC to outperform alternatives built on the same sample. As a consequence, preliminary comparisons of the AIC and BIC of the models are presented before more direct measures and tests of out-of-sample performance.

The four models ranked by AIC see Logit 2N as most accurate (AIC = 3612,84), Logit 1N as second (AIC = 3638,376) and Logit 1 as third (AIC = 4047.446). This is also the case if the BIC was used instead (Appendix 16). Notice that the difference in AIC between the Logit 2N and 1N are of different magnitude than the difference in AIC between Logit 1N and 1.

Such a ranking support the claim that at least one of the ceteris paribus relationships between probability of default and firm level covariates is non-linear. In fact, all non-linear models have a lower AIC than the linear one. In addition, the AIC and BIC also suggest that specification Logit 2N has higher predictive power than the non-linear alternative.

5.6 Out-of-sample performance

Several measures of out-of-sample performance were employed to compare the three models. The literature on credit risk modelling offers many alternatives and this relatively high richness in methods was exploited to test H_1.

Confusion Matrices

As in the case of section 5.5, confusion matrices are the first tool that is presented to compare the predictive power of the four models. However, table 6 through 8 were created using the testing set, as opposed to the training set. More specifically, models were used to estimate the probability of default of observations which were not used to estimate the models. Guesses thus generated were then used to classify observations as either defaulting or non-defaulting and results were compared to actual values (i.e. DEF1Y). A cut-off point of 0.007 was used for classification purposes. The resulting tables are shown below.

Table 6, 7 and 8 (from left to right): *out-of-sample contingency tables.*

Logit 1		
	Actual	
Prediction	0	1
0	30148	50
1	6505	187

Logit 1N		
	Actual	
Prediction	0	1
0	30555	41
1	6098	196

Logit 2N		
	Actual	
Prediction	0	1
0	30698	42
1	5955	195

The non-linear specifications seem to have better out-of-sample performance. This is supported by the fact that both Type I errors and Type II errors occurred more frequently in testing model Logit 1 than in testing Logit 1N and 2N. Comparing model out-of-sample accuracy, which was computed as explained in the previous section, resulted in the following ranking:

- Logit 2N (83.744%)
- Logit 1N (83.359%)
- Logit 1 (82.223%)

which, again, supports the hypothesis that non-linear models are superior in this application. Indeed, both non-linear specifications resulted in higher accuracy. Further, Logit 2N appears to be marginally more accurate than Logit 1N, suggesting that a measure of cash holding of a company as a fraction of total debt is more informative than the current ratio.

ROC Curves

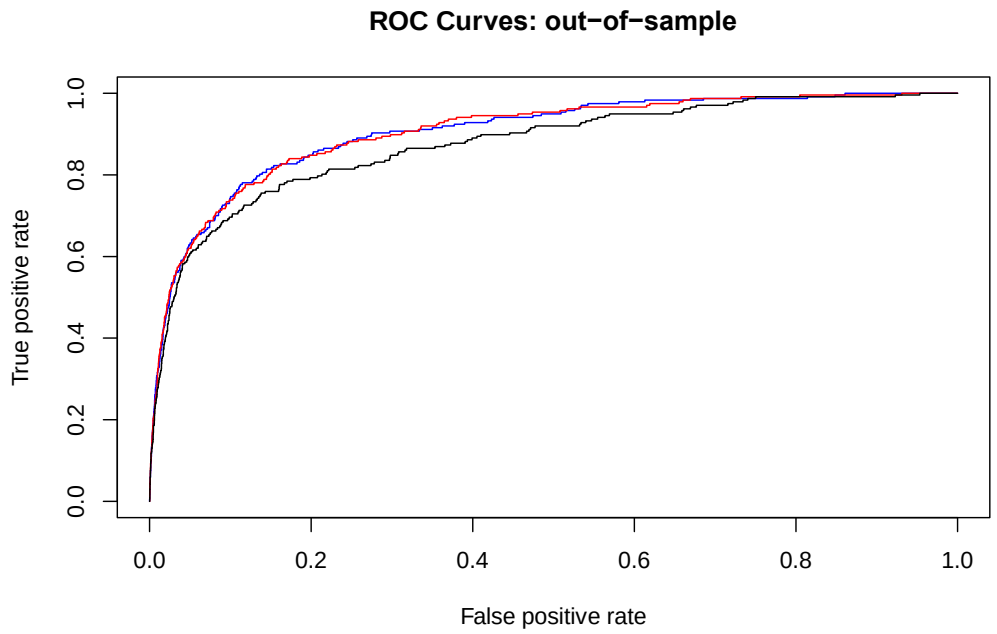
ROC curves were also constructed, on the testing set, to measure out-of-sample performance. As in the case of Plot 23, Plot 24 (below) suggests that non-linear models have higher predictive power, since the ROC curve of the linear model lies unequivocally under the ones of the two non-linear specifications.

As a matter of fact, ranking models by AUC results in the following order:

- Logit 2N (AUC = 0.9044)
- Logit 1N (AUC = 0. 90434)
- Logit 1 (AUC = 0.8674).

The relative superiority of Logit 2N and Logit 1N is supported once more. Further, Logit 2N appears to be again marginally more accurate than the non-linear alternative.

Plot 24: *out-of-sample ROC curves for Logit 1 (in black), Logit 1N (in red) and Logit 2N (in blue)*



Probability deciles

Following the example of Shumway (2001), the observations forming the testing set were sorted according to the associated probability of default estimated by the three models, from lowest to highest. Deciles were then formed to explore how many actual defaults took place in each group. This is done to see how clearly the model signals risk of default. In fact, models that assign a high probability of default to defaulting firms provide a clearer signal to users. A table created with this method for a model that perfectly discerns defaulting from non-defaulting firms would see all defaulting firms in the last decile (since average probability of default is less than 10%). Table 9 shows the results of this analysis for the three models.

Table 9: *probability deciles and number of defaults.*

	Logit 1		Logit 1N		Logit 2N	
Decile	Defaulting firms	Defaulting firms (as % of total)	Defaulting firms	Defaulting firms (as % of total)	Defaulting firms	Defaulting firms (as % of total)
1	2	0,84%	0	0,00%	1	0,42%
2	5	2,11%	3	1,27%	2	0,84%
3	1	0,42%	3	1,27%	5	2,11%
4	4	1,69%	2	0,84%	1	0,42%
5	5	2,11%	3	1,27%	1	0,42%
6	7	2,95%	3	1,27%	1	0,42%
7	15	6,33%	7	2,95%	13	5,49%
8	17	7,17%	17	7,17%	16	6,75%
9	19	8,02%	33	13,92%	32	13,50%
10	162	68,35%	166	70,04%	165	69,62%

Though the differences in number of firms classified in the last decile across the three models are not as material as the one between dynamic and static models documented in Shumway (2001), they suggest that the non-linear models signal risk of default more clearly, nonetheless. This finding is further supported by the differences in number of defaulting firms classified in the bottom 20%, which are 181 (76,37%) for Logit 1, 199 (83,96%) for Logit 1N and 197 (83,12%) for Logit 2N (approximately 10% more defaulting firms classified in the bottom 20%). Moreover, Logit 1N ranks the highest number of defaulting firms in the last decile, whereas Logit 2N ranks the highest number of defaulting firms in the last 2 deciles. This finding supports the idea that there is little difference between the two non-linear specifications.

Probability of default: estimated vs actual

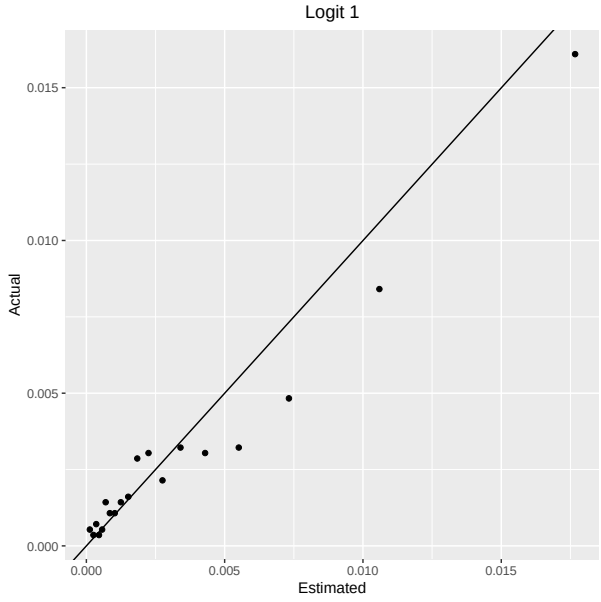
The last approach implemented to compare the predictive power of the three models follows the example presented in Giordani et al. (2014), where the estimated probability of default is plotted against the actual. Since probability of default is a latent variable that cannot be directly observed, it has to be approximated. In that article, the authors sort firms by estimated probability of default and group observations by percentile. The actual probability of default for each percentile is then approximated by taking the number of defaults for each percentile and dividing it by the number of firms in that percentile. A 2-dimensional scatterplot is then used to plot points having as x-coordinate the probability of default estimated by the model and as y-coordinate the approximation of actual probability of default. A perfect model would see all points lying on a straight line with slope = 1 and passing through the origin.

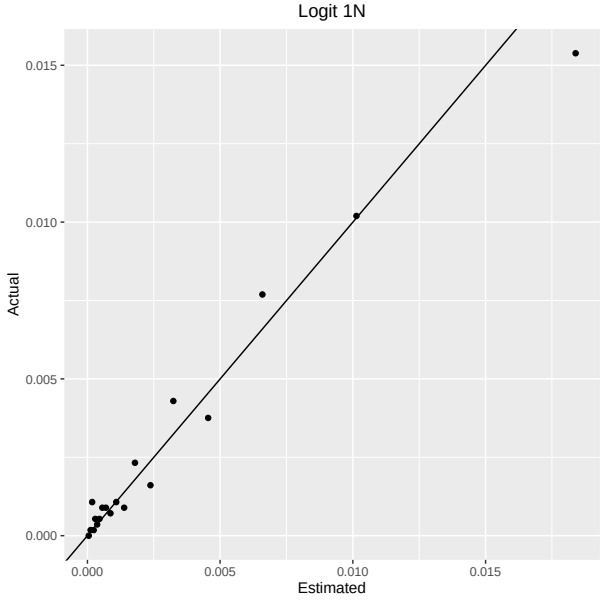
Although the same approach was adopted in this dissertation, some remarks have to be made. In Giordani et al. (2014), a sample of more than 4 million year-firms is used. Consequently, each percentile counts 40,000 firms, leading to substantial sparseness in the number of defaults per percentile. Given that roughly one fortieth of the observations was used for this research, dividing groups according to percentiles would have led to little sparseness, which in turn would have impaired significantly the quality of the plots. Thus, instead of creating 100 groups, each corresponding to a percentile, I created 20, each counting 5% of the sample. Actual probability of default is then computed and compared with the estimated for each of the 20 groups (or portfolios). The results are then plotted. It should be stressed that the whole sample, a combination of training and testing set, was used. This is in line with the method used in Giordani et al. (2014).

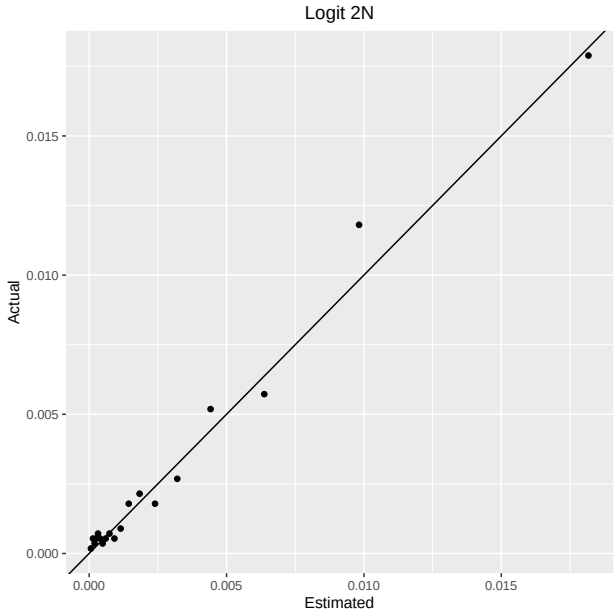
A visual inspection of plot 25 suggest that the probabilities fitted by the linear model are downwardly biased for portfolios with little probability of default and upwardly biased for portfolios with high probability of default. This “s-shaped” trend is also documented in Giordani et al. (2014), although it is much more distinct there due to the large number of points that are plotted. However, it should be noted that in a practical setup,

given the high costs associated with false negatives, the opposite kind of bias would not be preferred. In other words, it is preferable to underestimate probability of default for safe firms and overestimate it for risky ones rather than the opposite. Bias of this kind are not present in the case of Logit 1N and 2N, which seem to estimate probabilities of default which are closer to the actual values. It is particularly interesting that the probability of default estimated by Logit 2N for the riskiest group (i.e. the one further on the right) almost lie on the straight line, though this does not necessarily mean that the model is accurate to this extent in estimating probability of default for risky observations in general. Further, model Logit 2N seemingly fits more accurate probabilities also for the safest groups than the other specifications do, which, once more, supports the claim that CD is a better predictor than CR.

Plot 25 (on the left), 26 (on the right) and 27 (below): *actual versus estimated probability of default for each of twenty group sorted by estimated probability of default.*







Out-of-sample performance: conclusions

This section has compared the out-of-sample performance of the three specifications Logit 1, 1N and 2N taking full advantage of the wide range of tools and methodologies that are available for tasks of this kind. The comparison unequivocally showed that the two non-linear specifications have higher predictive power than the linear one. Logit 1N and 2N have indeed higher accuracy, AUC and result in better estimation of average probability of default across a portfolio, as well as fewer Type I and Type II errors. In addition, the non-linear specifications discern defaulters from non-defaulters more clearly, as they tend to assign higher probabilities of default to risky observations. Therefore, the empirical evidence unilaterally supports the claim that the ceteris paribus relationships between probability of default and firm-level covariates are non-linear.

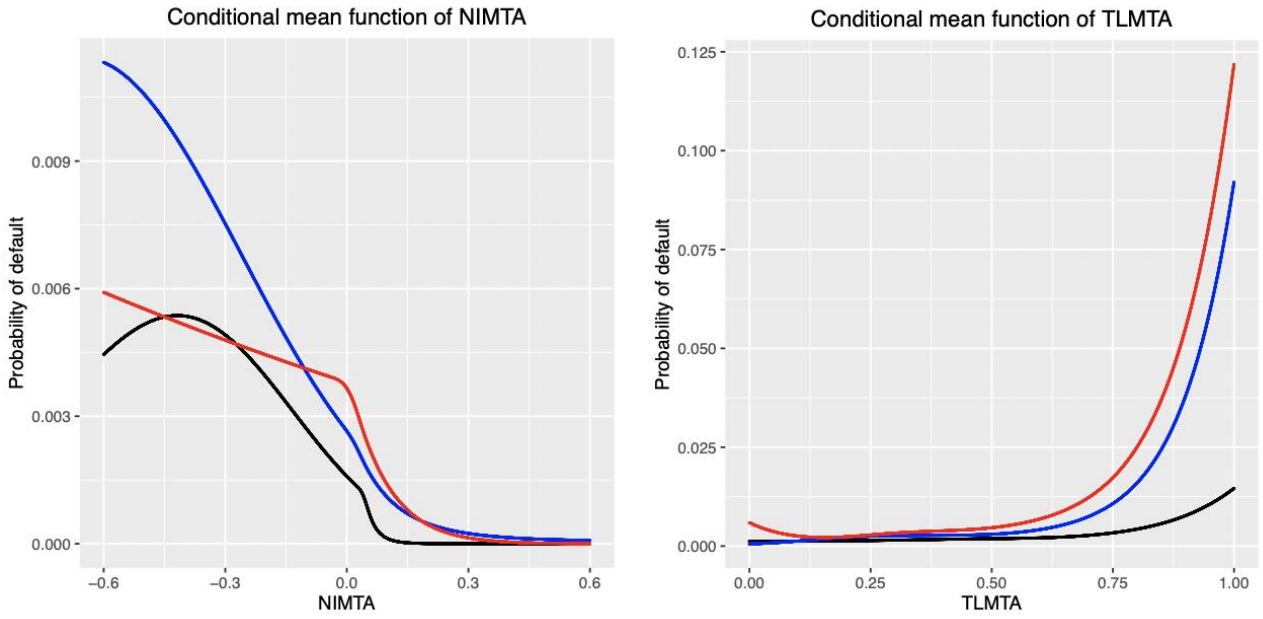
5.7 Robustness check: stability of non-linear relations

The fact that the non-linear specifications have better in-sample fit, and out-of-sample performance, provide some evidence towards the non-rejection of H_2. However, before drawing conclusions, it is important to show that the non-linearities are consistently present across time and do not originate just in specific time intervals.

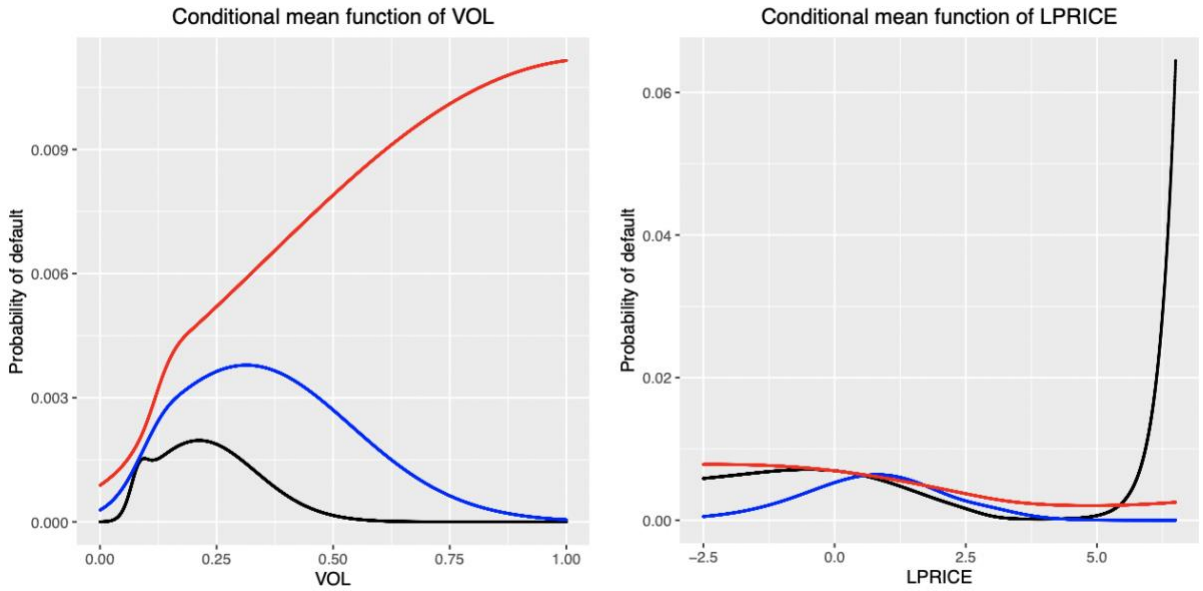
Taking inspiration from Giordani et al. (2014), specifications identical to the model Logit 1N and 2N were re-estimated on three different data sets, each comprising firms in different time intervals. The first subset comprises all firms with DATE up to 1989; the second subset comprises all observations with DATE between 1990 and 2001, the third subset comprises all year-firms with DATE beyond 2001. The periods were chosen so that each sub-sample comprises approximately the same number of year-firms. This resulted in a sub-sample spanning the period 1970-1989, one spanning the period 1989-2000 and a third spanning the period 2000-2011. Conditional mean functions were then computed as in section 5.4 and plotted to compare the results.

In the case of NIMTA (Plot 28) the shape of the conditional mean function between the period 1970 – 1989 and 2000 – 2011 tend to be very similar. In fact, both functions feature a kink approximately at NIMTA=0 and the different levels of probability of default associated with extreme negative values are merely due to difference in default rate across the two periods (circa 0.4% in 1970-1989 and 1.05% in 2000-2011). Moreover, the change in trend at NIMTA = -0.4 for the black line suggests that the function was probably overfitted in that range.

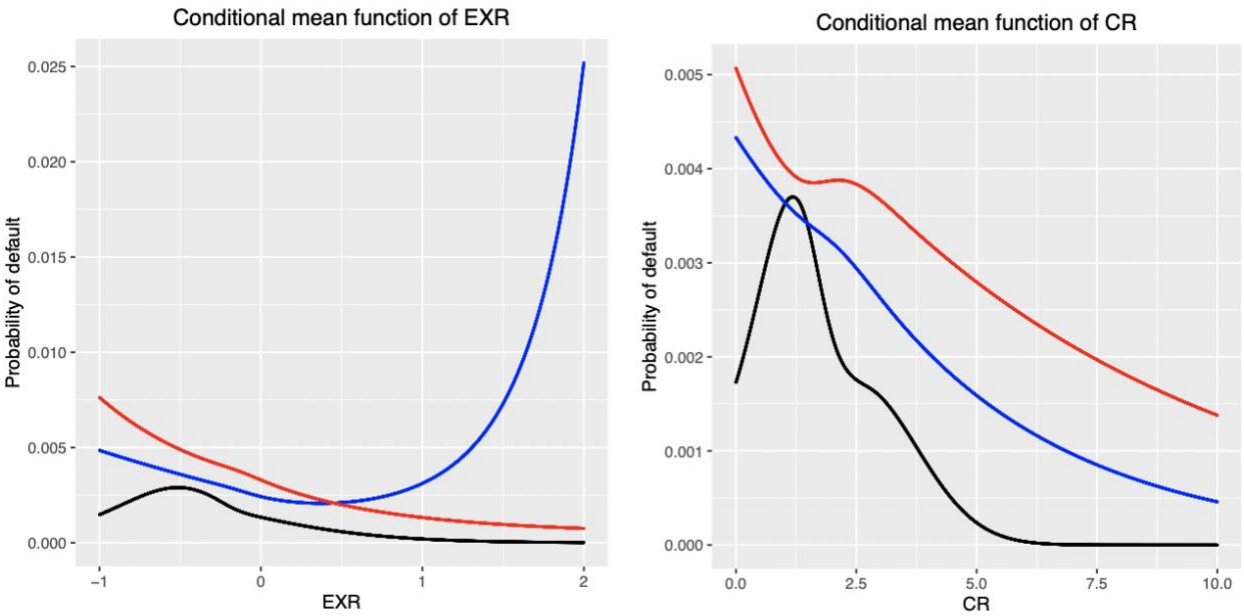
Plot 28 (on the left) and 29 (on the right): *The black line is the conditional mean function estimated in the period 1970-1989, the blue line is the mean function estimated in the period 1990-2000 and red function the conditional mean function estimated in the period 2001-2011.*



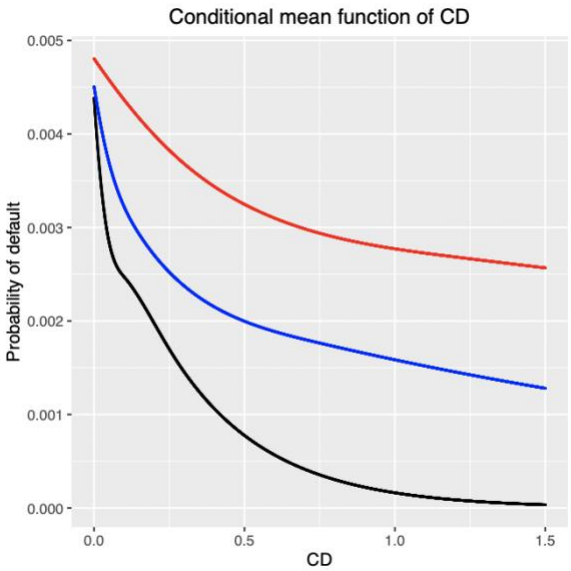
Plot 30 (on the left) and 31 (on the right): *The black line is the conditional mean function estimated in the period 1970-1989, the blue line is the mean function estimated in the period 1990-2000 and red function the conditional mean function estimated in the period 2001-2011.*



Plot 32 (on the left) and 33 (on the right): *The black line is the conditional mean function estimated in the period 1970-1989, the blue line is the mean function estimated in the period 1990-2000 and red function the conditional mean function estimated in the period 2001-2011.*



Plot 34: *The black line is the conditional mean function estimated in the period 1970-1989, the blue line is the mean function estimated in the period 1990-2000 and red function the conditional mean function estimated in the period 2001-2011.*



However, the conditional mean function changes in the period 1989-2000, as in this period it does not feature a kink, rather a smooth decline. In addition, the different level of probability of default for extreme negative values of NIMTA cannot be explained by a higher average probability of default, as this was 0.68%, which is very close to the average across the whole sample. As such, this relationship might have changed in the period 1989-2000. The "IPO culture" and the "irrational exuberance" of market participants might have played a role in redefining the features of the ceteris paribus relationship between probability of default and return on assets. However, the causes of this change are not clear.

TLMTA is remarkably stable over time, which suggests that financial leverage has always been highly correlated with probability of default, at least for non-financial firms. The different levels at extreme positive values of TLMTA can be again explained by the different average probabilities of default across the three periods.

It is interesting that the conditional mean function of VOL has a non-monotonic trend in the period 1970-2000. This is puzzling, since excess returns are controlled for and can be probably due to overfitting. It should also be noted that, as it is shown in section 5.4, probably the confidence interval widens substantially in the part of the scatterplot to the right of VOL = 0.5. This is further discussed in Part 4.

The conditional mean function of the variable LPRICE is also probably overfitted for extreme positive values of that variable. For the remaining it seems relatively stable over time and display a trend which is highly

similar to the shown in section 5.3, though the blue curve is more bell-shaped than the other two curves. Probably another symptom of overfitting.

Plot 32 showing the conditional mean function for the variable EXR offers an interesting example of how culture and market frictions can cause these relationships to variate. In fact, in the periods 1970-1989 and 2000-2011 the functions are highly alike, except for a difference in trend at very low values of EXR. Yet, in the period 1989-2000, the functions assume a totally different curve, suggesting that, beyond a certain level of excess returns, corporate risk of bankruptcy increases. This is not in line with the efficient market hypothesis and, as we have seen, does not generally apply. However, the form of the function is probably due to the IPO culture that later translated into the Dot.com bubble and into a wave of corporate defaults. Indeed, in the '90s, a large number of online-based companies listed their shares on stock markets, despite negative earnings. Further, due to the confusion about the role that the internet could play in economies, many companies' valuations increased by hundreds of percentage points (Shiller, 2001). The result was a very volatile market and the implementation of several "bump-and-dump" schemes (Schiller, 2001), and this is why EXR is positively correlated with probability of default at very high values of this variable. This represents a good example of how a statistical model can capture some aspects of a DGP.

The functions of CR and CD are very similar, except for some overfitting characterizing the black line in Plot 33. Again, the difference in level of probability of default associated with very positive value of both variables are due to the different average probabilities of the samples.

5.8 Empirical results: conclusions

Sections 5.5 and 5.6 provided a thorough comparisons of in-sample-fit and out-of-sample performance of the models estimated. Such as an extensive comparison resulted in substantial material towards the non-rejection of H_2. In addition, section 5.7 presented a robustness check of the non-linear models presented, and showed how some of these non-linearities, though depending on factors such as culture and market frictions, are persistently present across all time periods analyzed. I therefore fail to reject hypothesis H_2 and conclude that, most likely, one of the ceteris paribus relationships between probability of default and firm-level covariates is non-linear. This claim is in line with current research, in particular with the works of Giordani et al. (2014) and Berg (2007). The non-rejection of H_2 does not imply that the underlying relations between probability of default and firm level covariate are as documented. Rather, it implies that assuming these relationships to be linear, though constituting a necessary assumption in some cases (e.g. when the sample is small), affects the predictive power of the model being estimated, as this assumption is violated. Thus, when large samples are

available, it is suggested to estimate, rather than assume, the form of the relationships between firm level covariates and probability of default. This suggestion applies both when the model is used to explore some characteristics of the DGP in credit risk applications, and when the model is used to estimate probability of default.

This section concludes part 3. The remaining part of the paper explores the non-linearities that have been documented and provides some economic explanations for their form.

6. Part IV: Discussion

The conditional mean functions presented in section 5.4 represent estimates of the underlying ceteris paribus effect that firm-level covariates have on probability of default. As such, they carry some information that can be interpreted to shed some light on the underlying drivers of corporate default. Interpretation of these are given in the following sections.

6.1 Return on Assets and Probability of Default

Since NIMTA is a measure of Return On Assets (ROA), net of interest expense, its conditional mean function carries some information regarding the relationship between ROA and probability of default. Generally, this relationship has negative correlation, which is in line with expectations. In fact, everything else being equal, firms that earn more should be able to generate more earnings to finance their operations and meet payments of obligations. The conditional mean function of NIMTA features a kink at NIMTA = 0, implying that the difference in probability of default between firms that are loss making and firms with positive earnings, is substantial. In this regard, it is interesting to note that a 60% increase in NIMTA from - 0.6 to 0 and a change in NIMTA from 0 to 0.15 are expected to have the same partial incremental effect on probability of default. Further, although the kink disappears during the ‘90s, the fact that it is present both in the periods 1970-1989 and 2000-2011 suggest that this is a structural feature of this relationship. An interpretation that can be given to this finding is that loss making firms tend to have unsustainable business models and are therefore more likely to default.

Maintaining the focus on the non-linearities of the function, it should be stressed that the non-monotonic pattern associated with low positive values of the variable is due to overfitting. Such a pattern is not present in the conditional mean function estimated across the three periods that have been compared in the previous section.

As a last note on this relationship, it is interesting to see that probability of default is 0 for values of NIMTA beyond 0.2. This further strengthens the argument that earnings represent a form of financing that is vital to the survival and success of firms.

6.2 Leverage and probability of default

As it was shown in the literature review, the correlation between leverage and short-term risk of default has been documented by many. It is intuitive that, as the indebtedness of a firm increases, also the probability that it will not be able to meet its debt obligations increases. Yet, maintaining the focus on the non-linear aspects of the conditional function estimated for TLMTA, it is possible to relate these findings to the wider literature on corporate finance.

More specifically, for a large interval of values, probability of default is not correlated with leverage and remains stable at a relatively low level. However, when leverage surpasses the value of 1 (TLMTA = 0.5), the function assumes an exponential trend. This is because, up to certain level of indebtedness, meet payment obligations by refinancing or generating earnings is very likely, and few dramatic events can cause the firm to fail in servicing its debt. However, beyond a certain level, more probable events can cause the firm to incur a default. In fact, using Merton's model, since the volatility of the assets does not depend on how those assets are financed, a higher leverage implies a higher likelihood that the value of the assets drops below the value of the liabilities, causing a default.

This finding is consistent with the trade-off theory on optimal capital structure. According to this theory, in finding the optimal debt-to-equity mix, financial managers take into consideration the present value of future tax shields and the bankruptcy costs, both of which increases with firm leverage (Braeley, Myears and Allen, 2014). Bankruptcy costs comprise all the value destruction associated with financial distress, which is in turn intimately related to probability of default (Campbell, Hilscher and Szilagyi, 2008). The theory states that, while the relationship between the present value of tax-shields and leverage is approximately linear, the one between bankruptcy costs and leverage is not linear and tends to be steeper for high values of leverage. Therefore, maximizing the difference between bankruptcy costs and the present value of future tax shields would result in an optimum which should then dictate the target capital structure of the firm.

Since probability of default can be seen as a measure of financial distress (Campbell, Hilscher and Szilagyi, 2008), it is possible to reinterpret the conditional mean function of TLMTA and think it as the relationship between financial distress and leverage. From this perspective, it is clear how the finding supports the trade-off theory.

6.3 Probability of default and asset volatility

The conditional mean function of VOL represents the estimated ceteris paribus relationship between probability of default and asset volatility. The seemingly non-monotonic trend is caused by two factors. First, the estimates of probability of default for values of VOL above 0.5 have a very wide 95% confidence interval, implying that they are not reliable. Second, outliers whose volatility is very high are more likely to have experienced positive excessive returns rather than negative excessive returns. This is due to the fact that there is no limit to the positive returns that shares can deliver, while negative returns are capped at -100%. Following these two observations, estimates of probability of default beyond VOL = 0.8 should be disregarded.

Focusing on low values of VOL, we note that very low values of asset volatility are generally associated with very low probabilities of default. More specifically, the non-linear conditional function of VOL appears to pass through the origin, as opposed to the one fitted by the linear model, which has a constant. Since the probability that a firm with certain cash flows defaults should be nihil, the former is in line with theory, whereas the latter is in sharp contrast with it. This is especially true if one recalls that the conditional mean function expresses the probability of default as a function of VOL, conditional on all other variables being at their mean value. Since the mean value of TLMTA is not 1 (suggesting that the average market capitalization of the companies' equity is positive), the average equity value is positive. Thus, if we were to imagine a company on the black line in Plot 14, with VOL = 0, implying that this company's cash flows are certain, the company's probability of default should also be 0, given that its equity is worth something. This example highlights how, in the case of VOL, the probability fitted as in the conditional function are theoretically more correct.

6.4 Probability of default and share price

The fact that firms with a low share price tend to have higher credit risk is not intuitive, as in a frictionless market such a phenomenon would not take place. However, such a relationship has long been documented (e.g. Dewing, 1934) and is intimately related to the broader "share price puzzle". In fact, there seems to be an interval of share prices which is optimum (Dyl and Elliott, 2006). The relationship between share price and probability of default is caused by the requirements that stock exchanges impose on companies that desire to issue equity on these channels. For instance, a minimum number of outstanding shares is imposed on the NYSE, NASDAQ and AMEX. As a consequence, companies cannot implement reverse stock-splits indefinitely and firms whose shares have a very low price might not have access to any method to raise it. This relationship is therefore caused by market frictions, contrary to other relationships that have been documented in this dissertation.

By inspecting Plot 16, and recalling from section 5.7 that the non-monotonic trend should be ignored for values of LPRICE below 0, the trend of the conditional mean function there plotted is steeper at low values of LPRICE and tends to flatten as this covariate increases. This feature of the fitted curve suggests that probability of default is high for firms with very low share prices, but very low beyond a certain price, which is in line with the share price puzzle as documented by Dyl and Elliott (2006).

6.5 Probability of default and excess returns

Observing Plot 18, the ceteris paribus relationship between excess returns on companies' equity, measured by the variable EXR, and probability of default can be seen. The trend is strictly negative for the whole range of values, which is in line with expectations. An excess return can be achieved when the market has better expectations regarding the future of a particular asset compared to alternative investments with a similar risk profile (Braeley, Myers and Allen, 2014). The fact that the relationship turned out to be significant and negative in all models estimated strengthens the arguments that asset prices embed information regarding the future. This is especially true if one recalls again that the relationship being estimated is conditional on the mean return on assets, which could be interpreted as "accounting returns". Hence, the fact that excess returns could be due to better accounting returns is controlled for, leaving only market's expectations. This finding supports the idea that market information should be used when available, a claim already made in other influential papers in the area of credit risk (e.g. Shumway, 2001).

6.6 Current ratio, cash holdings and probability of default.

As mentioned in section 5.4, the conditional mean functions of the current ratio and cash holdings as a fraction of total debt share many features and carry a similar informational content, as the inclusion of one variable in the non-linear model resulted in the exclusion of the other. In fact, both covariates measure, in different ways, the solvency of a firm. Yet, in the case of the current ratio, short-term solvency risk is measured, whereas, in the case of CD, general solvency risk is measured (Petersen, Plenborg and Kinserdal, 2014). Both conditional mean functions have a downward trend, which is rather steep in the region close to 0 and flattens at mid-values of the covariates. This is in line with expectations, since, as an improvement in solvency, is not expected to have the same effect on probability of default for firms that are already highly solvent and for firms that have looming solvency issues.

The non-monotonic trend in both conditional mean functions is probably due to overfitting. In fact, the same pattern does not feature the same functions computed in the periods 1970-1989, 1989-2000 and 2000-2011. As a result, the trend should be interpreted as strictly downward.

With focus on cash holdings over total debt and with reference to the literature review in section 2, recall the Acharya, Davydenko and Strebulaev (2012) found that cash holdings and credit spreads can be positively correlated. The findings here presented seem to contradict the ones presented in that article, as probability of default is negatively correlated to cash holdings over total debt. However, it should be noted that in Acharya, Davydenko and Strebulaev (2012), cash holdings are measured against total assets, not long-term debt. As a consequence, despite the findings being contradicting to some extent, the conditional mean function estimated for the variable CD does not necessarily undermine the findings presented in that article.

Finally, the fact that model Logit 2N proved to be marginally more accurate than the non-linear alternative can be explained by the relatively high importance of the amount of cash and liquid assets that firm should have to cope with unexpected events. In fact, given that the conversion of receivables and inventory into cash might be threatened by sudden, uncontrollable factors, cash assets carry more information regarding firms' solvency in the short term.

6.8 Criticism

This section offers a critical perspective on the methodologies that have been implemented to test H_1 and H_2.

The first thing to note is that, to conclude with certainty that some of the relationships between probability of default and firm-level covariates are non-linear requires the estimation of the true underlying function. Although it is true that no statistical model can achieve that, I made a number of simplifying assumptions. One of these concerns directly the estimation of the conditional mean functions that have been presented. In fact, interaction terms among regressors were assumed away to make the estimation of the additive logit model more practical and viable. It is maybe redundant to mention it again, but it should be stressed that this was done because the inclusion of such terms would have caused the number of knots parameter to be excessively sparse in the covariate space due to the "curse of dimensionality". This is indeed the approach adopted in Giordani et al. (2014), which represented one of the very few references available for this topic. Yet, it could be that, although the inclusion of all terms would have led to the problem just described, the inclusion of a few of these terms, the most relevant, would not have caused this problem, and could have potentially resulted in a more accurate model. This was decided to be left to further research, as in exploring an unknown area, it is probably better to take one step at the time.

Further, I have used yearly data. This again followed the example of Giordani et al. (2014) but does not represent the most accurate approach recommended by the literature, which is to use monthly data. Using monthly data would have probably resulted in a dataset of very large dimensions, the handling of which would have proved to be difficult. Since the objective of this research was not to estimate the most accurate model, but to explore some of the relationships between probability of default and firm-level covariates, this was also left to further research. I recognize that, especially for market values, whose frequency is very high, some of the relationships that I estimated could be different if I used monthly data instead of annual. Nevertheless, I do not expect that using annual data invalidates the testing of H_1 and H_2.

Finally, plotting the conditional mean functions of firm level covariates does not fully express all the feature of these relationships. Some conditional functions could change dramatically according to the level of the other variables. However, also this critique does not impair the testing of H_2, rather warrants caution about what Plots 10 throughout 22 really represent.

7. Conclusion

This research has explored the drivers of corporate default by testing whether one of the univariate relationship between probability of default and firm-level covariates is non-linear, as well as testing whether one of the ceris paribus relationships between probability of default and firm-level covariates is non-linear.

Testing the first hypothesis revealed to be relatively simple and followed the example set out in Giordani et al. 2014. By visualizing these relationships, it appears as the univariate relationship between probability of default and return on assets and probability of default and volatility are non-linear. Further, the latter relationship reveiled also to be non-monotonic. This evidence led to the non-rejection of hypothesis H_1.

To test the second hypothesis, two class of models have been compared: a GLM and an additive model. An extensive literature review provided guidance on the choice of model specification to be tested, on the sample to be used to estimate the models and on the choice of functional form. Particularly, the research followed Campbell, Hislcher and Szilagyi (2008) to choose covariates and sample size. The model presented in that article was indeed deemed to be accurate and to be a valid representative of the linear class of models. To relax the assumption on linearity of GLMs, inspiration was taken from the work of Giordani et al. (2014), where natural splines are used.

The dissertation proceeded by estimating three specifications: a binomial model with logit link and two additive binomial models with logit link and estimated using natural splines.

Comparisons of the in-sample fit and out-of-sample performance of the two models resulted in a relative superiority of the non-linear specifications that, although very similar between them, revealed to have higher predictive power. A robustness check confirmed that the non-linearities detected are not simply incidental but are also present in sub-periods. Following the empirical results, H_2 cannot be rejected, implying that it is likely that at least of the ceteris paribus relationships is non-linear.

Finally, the form of these relationships was discussed. Economic explanations were provided for certain aspects of the estimates of these relationships, whereas others were proved to be the result of overfitting.

8. Bibliography

Articles

Acharya, V., Davydenko, S., & Strebulaev, I. (2012). Cash Holdings and Credit Risk. *The Review of Financial Studies*, *25*(12), 3572–3609. doi: 10.3386/w16995

Agarwal, V., & Taffler, R. J. (2007). Comparing the Performance of Market-Based and Accounting-Based Bankruptcy Prediction Models. *Journal of Banking & Finance*, 1541–1551. doi: 10.1016/j.jbankfin.2007.07.014

Almamy, J., Aston, J., & Ngwa, L. N. (2016). An evaluation of Altmans Z-score using cash flow ratio to predict corporate failure amid the recent financial crisis: Evidence from the UK. *Journal of Corporate Finance*, *36*, 278–285. doi: 10.1016/j.jcorpfin.2015.12.009

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and The Prediction Of Corporate Bankruptcy. *The Journal of Finance*, *23*(4), 589–609. doi: 10.1111/j.1540-6261.1968.tb00843.x

Altman, E. I. (2018). A fifty-year retrospective on credit risk models, the Altman Z -score family of models and their applications to financial markets and managerial strategies. *The Journal of Credit Risk*, *14*(4), 1–34. doi: 10.21314/jcr.2018.243

Bharath, S. T., & Shumway, T. (2008). Forecasting Default with the Merton Distance to Default Model. *Review of Financial Studies*, *21*(3), 1339–1369. doi: 10.1093/rfs/hhn044

Campbell, J., Hilscher, J., & Szilagyi, J. (2008). In Search of Distress Risk. *The Journal of Finance* , *LXIII*(6), 2899–2939. doi: 10.3386/w12362

Friewald, N., Wagner, C., & Zechner, J. (2014). The Cross-Section of Credit Risk Premia and Equity Returns. *The Journal of Finance*, *69*(6), 2419–2469. doi: 10.1111/jofi.12143

Gilchrist, S., & Zakrajšek, E. (2012). Credit Spreads and Business Cycle Fluctuations. *American Economic Review*, *102*(4), 1692–1720. doi: 10.1257/aer.102.4.1692

Giordani, P., Jacobson, T., Schedvin, E. V., & Villani, M. (2014). Taking the Twists into Account: Predicting Firm Bankruptcy Risk with Splines of Financial Ratios. *Journal of Financial and Quantitative Analysis*, *49*(4), 1071–1099. doi: 10.1017/s0022109014000623

Jarrow, R. A., Lando, D., & Turnbull, S. M. (1997). A Markov Model for the Term Structure of Credit Risk Spreads. *The Review of Financial Studies*, *10*(2), 481–523. doi: 10.1093/rfs/10.2.481

Jarrow, R. A., & Turnbull, S. M. (1995). Pricing Derivatives on Financial Securities Subject to Credit Risk. *The Journal of Finance*, *50*(1), 53–85. doi: 10.1142/9789812819222_0017

Kealhofer, S. (2003). Quantifying Credit Risk I: Default Prediction. *Financial Analysts Journal*, *59*(1), 30–44. doi: 10.2469/faj.v59.n1.2501

Lando, D., & Nielsen, M. S. (2008). Correlation in Corporate Defaults: Contagion or Conditional Independence? *SSRN Electronic Journal*, 1–40. doi: 10.2139/ssrn.1108649

Merton, R. C. (1974). On The Pricing Of Corporate Debt: The Risk Structure Of Interest Rates*. *The Journal of Finance*, *29*(2), 449–470. doi: 10.1111/j.1540-6261.1974.tb03058.x

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, *18*(1), 109–131. doi: 10.2307/2490395

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., & Schmid, M. (2019). A review of spline function procedures in R. *BMC Medical Research Methodology*, *19*(1). doi: 10.1186/s12874-019-0666-3

Schaefer, S., & Strebulaev, I. (2008). Structural models of credit risk are useful: Evidence from hedge ratios on corporate bonds☆. *Journal of Financial Economics*, *90*(1), 1–19. doi: 10.1016/j.jfineco.2007.10.006

Shumway, T. (2001). Forecasting Bankruptcy More Accurately: A Simple Hazard Model. *The Journal of Business*, *74*(1), 101–124. doi: 10.1086/209665

Stein, R. (2007). Benchmarking default prediction models: pitfalls and remedies in model validation. *The Journal of Risk Model Validation*, *1*(1), 77–113. doi: 10.21314/jrmv.2007.002

Zmijewski, M. E. (1984). Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research*, *22*, 59–82. doi: 10.2307/2490859

Books

Brealey, R. A., Myers, S. C., & Allen, F. (2014). *Principles of corporate finance*. New-York: McGraw-Hill Education. ISBN: 978-0-771-5507-0

Hastie, T., Friedman, J., & Tisbshirani, R. (2017). *The Elements of statistical learning: data mining, inference, and prediction*. New York: Springer. ISBN: 978-0-387-84857-0

Petersen, C. V., Plenborg, T., & Kinserdal, F. (2017). *Financial statement analysis: valuation, credit analysis, performance evaluation*. Harlow, England: Financial Times/Prentice Hall. ISBN: 978-82-450-2102-8

Veal, A. J., & Darcy, S. (2014). *Research methods in sport studies and sport management: a practical guide*. Londres: Routledge. ISBN: 978-0-273-73669-1

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. Boca Raton: CRC Press. ISBN: 978-1-498-72833-1

Wooldridge, J. M. (2020). *Introductory econometrics: a modern approach*. Boston, MA: Cengage. ISBN: 978-1-337-55886-0

9. Appendix

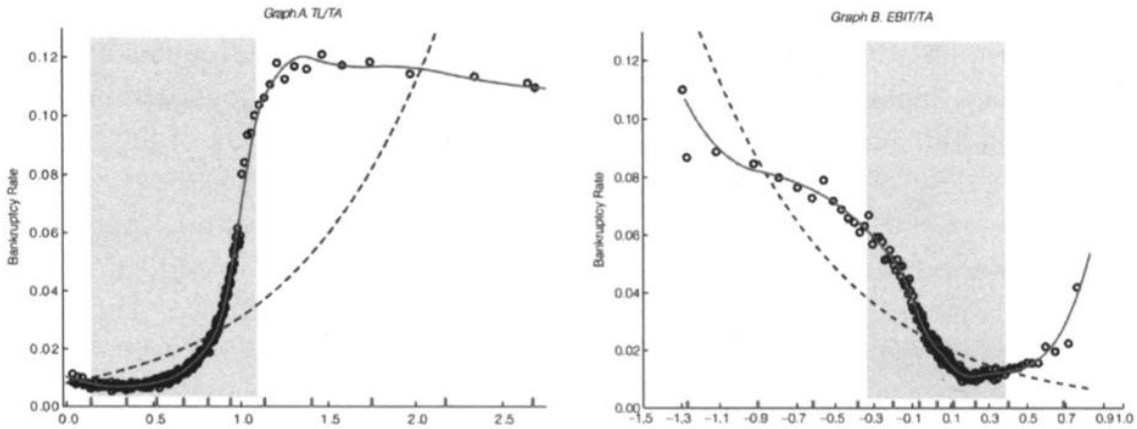
Appendix 1: Altman's Z-score formula

$$Z-\text{score} = 1.2\frac{\text{Working capital}}{\text{Total assets}} + 1.4\frac{\text{Retained earnings}}{\text{Total assets}} + 3.3\frac{\text{EBIT}}{\text{Total assets}} + 0.6\frac{\text{Market value of equity}}{\text{Book value of liabilities}} + 1.0\frac{\text{Sales}}{\text{Total assets}}$$

Appendix 2: An example of plot to visualize univariate relationship between probability of default and firm-level covariate. The dotted line represents the fit of univariate logistic regression, whereas the continuous line represents the fit of logistic regression with splines. (source: Giordani, Jacobson, von Schedvin and Villani, 2014)

Realized Bankruptcy Frequencies and Univariate Estimates

Figure 1 illustrates the realized bankruptcy frequencies (circles) and estimated bankruptcy probabilities, obtained from univariate logistic (dashed line) and univariate logistic spline models (solid line), for the five firm-specific variables over the full sample period 1991–2008. For each variable, the data have been sorted and grouped into 300 equal-sized groups. For each group, we calculate the realized bankruptcy frequency as the share of bankrupt firms over all firms, and then an average of the observations of the firm-specific variable at hand. The 300 group-data points are then plotted against each other to yield the circles. For each variable, the reported logistic spline fit is calculated based on a univariate spline model incorporating 11 knots, and likewise, the logistic fit is based on a univariate logistic model. The shaded areas in the graphs mark out regions containing 90% of the observations. The thicker tick marks on the horizontal axes indicate the location of the spline knots. SIZE is log of total sales. AGE measures log of firm age (+1 year) in number of years since first registered as a corporate firm.

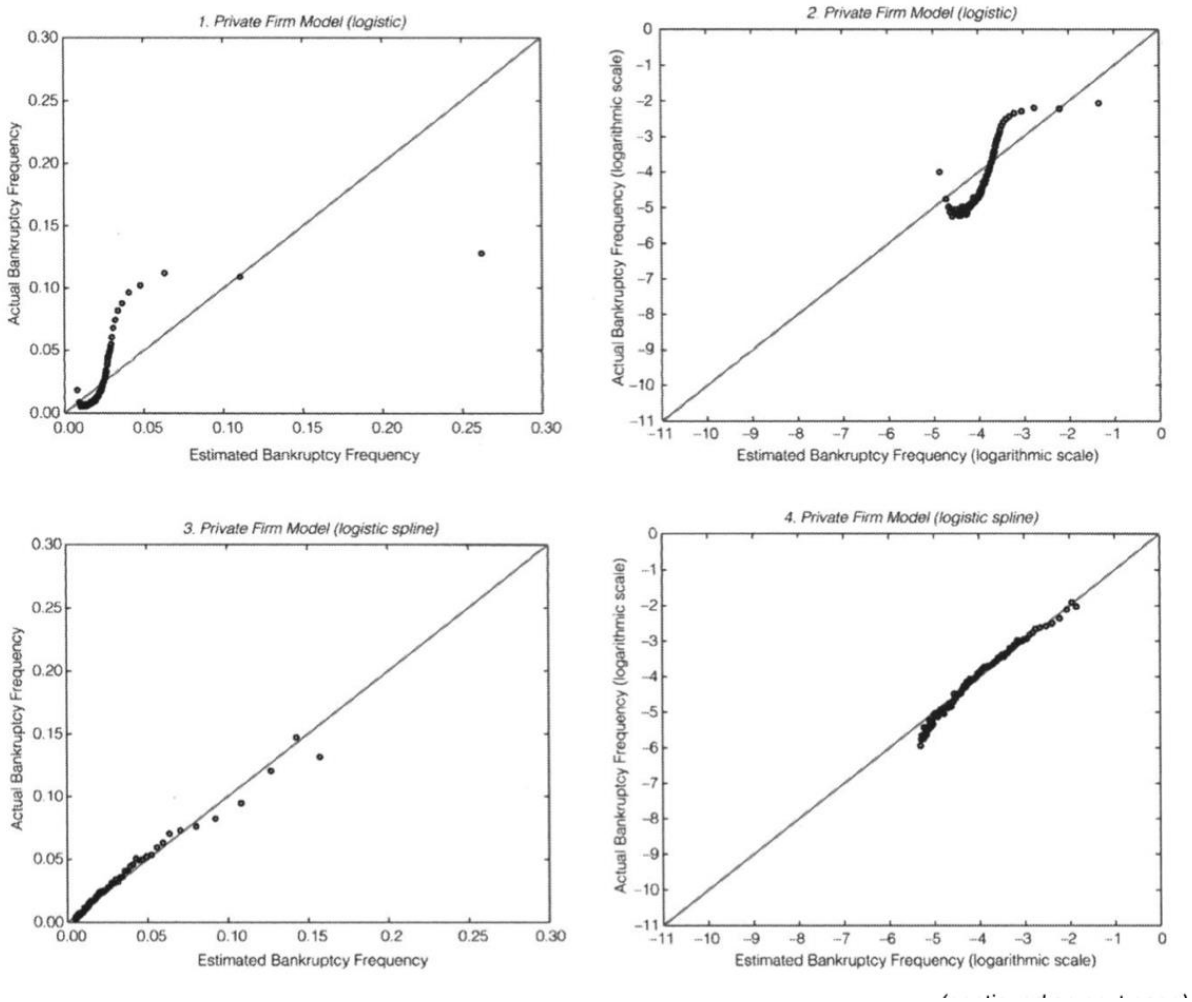


Appendix 3: A plot used to compare model goodness of fit (source: Giordani, Jacobson, von Schedvin and Villani, 2014)

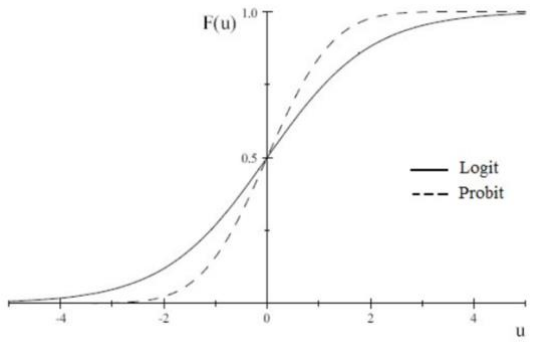
Predicted and Realized Bankruptcy Frequencies, In-Sample 1991–2008

Figure 2 illustrates in-sample estimated bankruptcy probabilities versus realized bankruptcy frequencies for the period 1990–2008. Graphs A and B correspond to the Private Firm Model and the Extended Private Firm Model, respectively, in Table 2. For each model, we sort all firm-year observations with respect to the size of their estimated bankruptcy probability and divide them into percentiles. We then calculate the average probability of bankruptcy and the share of realized bankruptcies within each percentile. The circles correspond to the pairs of estimated bankruptcy probabilities versus realized bankruptcy shares, and the 45-degree line illustrates a perfect fit. We have graphed the relationships using a probability scale (left-hand side) and a logarithmic scale (right-hand side).

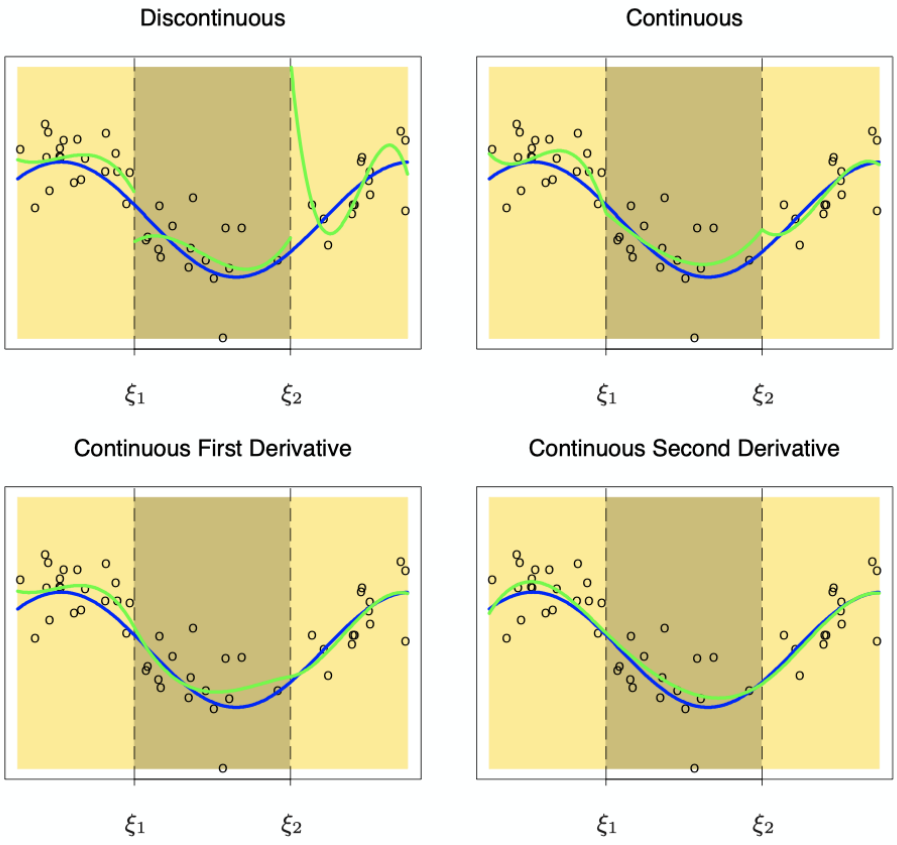
Graph A. Private Firm Models



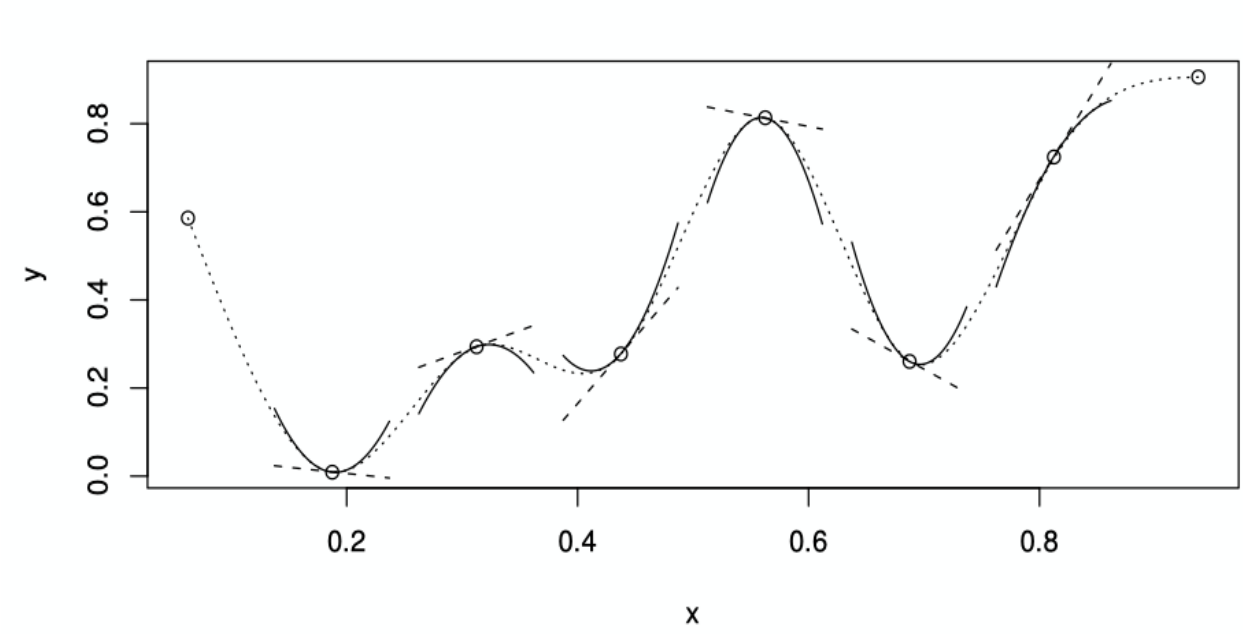
Appendix 4: logistic (logit) and normal (probit) cumulative functions.



Appendix 5: A series of piecewise-cubic polynomials (source: Hastie, Tibshirani and Friedman, 2009)



Appendix 6: A cubic spline as the combination of several piecewise polynomials (source: Wood, 2017).



Appendix 7: Summary statistics of the Training set

```
     CUSIP            DATE                FYR            NIMTA
68066520:   36   Min.   :1970-02-27   Min.   :1969   Min.   :-9.024107
77571110:   36   1st Qu.:1985-09-30   1st Qu.:1984   1st Qu.:-0.003812
88579Y10:   36   Median :1995-07-31   Median :1994   Median : 0.029703
20172310:   35   Mean   :1993-12-02   Mean   :1993   Mean   :-0.007979
21683110:   35   3rd Qu.:2002-04-30   3rd Qu.:2001   3rd Qu.: 0.049498
30161N10:   35   Max.   :2012-09-28   Max.   :2011   Max.   : 3.282979
(Other) :74685
    TLMTA               VOL              LPRICE            EXR
Min.   :0.0001092   Min.   :0.002131   Min.   :-3.474   Min.   :-1.55679
1st Qu.:0.1673975   1st Qu.:0.083244   1st Qu.: 1.792   1st Qu.:-0.31283
Median :0.3460233   Median :0.119927   Median : 2.639   Median :-0.14321
Mean   :0.3710929   Mean   :0.142645   Mean   : 2.476   Mean   :-0.12787
3rd Qu.:0.5541788   3rd Qu.:0.172669   3rd Qu.: 3.277   3rd Qu.: 0.03336
Max.   :0.9952697   Max.   :2.933677   Max.   : 7.821   Max.   : 5.56865

      CR                CD               DEF1Y              AGE
Min.   :   0.000   Min.   :  -0.0384   Min.   :0.000000   Min.   : 1.00
1st Qu.:   1.424   1st Qu.:   0.0445   1st Qu.:0.000000   1st Qu.: 4.00
Median :   2.109   Median :   0.1557   Median :0.000000   Median : 9.00
Mean   :   3.044   Mean   :   0.9746   Mean   :0.006435   Mean   :11.12
3rd Qu.:   3.215   3rd Qu.:   0.6461   3rd Qu.:0.000000   3rd Qu.:16.00
Max.   :1584.098   Max.   :1580.3007   Max.   :1.000000   Max.   :43.00

     SIZE           NIMTA_lag1          EXR_lag1
Min.   :-9.5379   Min.   :-15.363247   Min.   :-1.54547
1st Qu.:-2.4662   1st Qu.:  0.001087   1st Qu.:-0.31311
Median :-1.1792   Median :  0.030426   Median :-0.14267
Mean   :-1.0695   Mean   :  0.000974   Mean   :-0.12633
3rd Qu.: 0.2638   3rd Qu.:  0.050192   3rd Qu.: 0.03537
Max.   : 6.1099   Max.   :  2.459815   Max.   : 4.55351
```

Appendix 8: Summary statistics for the Testing set

```
     CUSIP            DATE                FYR           NIMTA
38238810:   23   Min.   :1970-02-27   Min.   :1969   Min.   :-15.363247
75511150:   22   1st Qu.:1985-10-31   1st Qu.:1985   1st Qu.: -0.003632
46121H10:   21   Median :1995-05-31   Median :1994   Median :  0.029301
45950610:   20   Mean   :1993-12-05   Mean   :1993   Mean   : -0.008037
87254010:   20   3rd Qu.:2002-04-30   3rd Qu.:2001   3rd Qu.:  0.049272
07181310:   19   Max.   :2012-09-28   Max.   :2011   Max.   :  1.658166
(Other) :36765
    TLMTA                VOL               LPRICE            EXR
Min.   :0.0003196   Min.   :0.0005292   Min.   :-3.466   Min.   :-1.63012
1st Qu.:0.1671781   1st Qu.:0.0831261   1st Qu.: 1.794   1st Qu.:-0.31008
Median :0.3481304   Median :0.1198279   Median : 2.639   Median :-0.14019
Mean   :0.3721215   Mean   :0.1423518   Mean   : 2.479   Mean   :-0.12347
3rd Qu.:0.5567685   3rd Qu.:0.1735422   3rd Qu.: 3.287   3rd Qu.: 0.03583
Max.   :0.9941356   Max.   :2.7805498   Max.   : 7.427   Max.   : 3.98399

      CR                CD               DEF1Y              AGE
Min.   :   0.001   Min.   : -0.0090   Min.   :0.000000   Min.   : 1.00
1st Qu.:   1.419   1st Qu.:  0.0440   1st Qu.:0.000000   1st Qu.: 4.00
Median :   2.111   Median :  0.1533   Median :0.000000   Median : 8.00
Mean   :   3.112   Mean   :  0.9678   Mean   :0.006424   Mean   :11.11
3rd Qu.:   3.196   3rd Qu.:  0.6448   3rd Qu.:0.000000   3rd Qu.:16.00
Max.   :1719.250   Max.   :654.6667   Max.   :1.000000   Max.   :43.00

     SIZE           NIMTA_lag1           EXR_lag1
Min.   :-7.8549   Min.   :-15.363247   Min.   :-1.60259
1st Qu.:-2.4576   1st Qu.:  0.001705   1st Qu.:-0.31301
Median :-1.1623   Median :  0.030503   Median :-0.14546
Mean   :-1.0671   Mean   :  0.000627   Mean   :-0.12755
3rd Qu.: 0.2589   3rd Qu.:  0.049782   3rd Qu.: 0.03501
Max.   : 6.0661   Max.   :  2.226938   Max.   : 4.50694
```

Appendix 9: Summary statistics for the whole set (Testing + Training)

```
     CUSIP             DATE                FYR           NIMTA
00095710:    43   Min.   :1970-02-27   Min.   :1969   Min.   :-15.363247
00282410:    43   1st Qu.:1985-09-30   1st Qu.:1984   1st Qu.: -0.003745
03741110:    43   Median :1995-06-30   Median :1994   Median :  0.029572
03965L10:    43   Mean   :1993-12-03   Mean   :1993   Mean   : -0.007998
05361110:    43   3rd Qu.:2002-04-30   3rd Qu.:2001   3rd Qu.:  0.049417
06738310:    43   Max.   :2012-09-28   Max.   :2011   Max.   :  3.282979
(Other) :111530
    TLMTA               VOL               LPRICE            EXR
Min.   :0.0001092   Min.   :0.0005292   Min.   :-3.474   Min.   :-1.6301
1st Qu.:0.1673546   1st Qu.:0.0832026   1st Qu.: 1.792   1st Qu.:-0.3120
Median :0.3466720   Median :0.1198750   Median : 2.639   Median :-0.1421
Mean   :0.3714324   Mean   :0.1425482   Mean   : 2.477   Mean   :-0.1264
3rd Qu.:0.5549527   3rd Qu.:0.1729435   3rd Qu.: 3.282   3rd Qu.: 0.0343
Max.   :0.9952697   Max.   :2.9336769   Max.   : 7.821   Max.   : 5.5687

      CR                CD                DEF1Y              AGE
Min.   :   0.000   Min.   :  -0.0384   Min.   :0.000000   Min.   : 1.00
1st Qu.:   1.423   1st Qu.:   0.0444   1st Qu.:0.000000   1st Qu.: 4.00
Median :   2.110   Median :   0.1550   Median :0.000000   Median : 9.00
Mean   :   3.067   Mean   :   0.9724   Mean   :0.006432   Mean   :11.12
3rd Qu.:   3.209   3rd Qu.:   0.6459   3rd Qu.:0.000000   3rd Qu.:16.00
Max.   :1719.250   Max.   :1580.3007   Max.   :1.000000   Max.   :43.00

     SIZE           NIMTA_lag1           EXR_lag1
Min.   :-9.5379   Min.   :-15.363247   Min.   :-1.60259
1st Qu.:-2.4634   1st Qu.:  0.001323   1st Qu.:-0.31310
Median :-1.1729   Median :  0.030457   Median :-0.14370
Mean   :-1.0687   Mean   :  0.000859   Mean   :-0.12673
3rd Qu.: 0.2621   3rd Qu.:  0.050072   3rd Qu.: 0.03529
Max.   : 6.1099   Max.   :  2.459815   Max.   : 4.55351
```

Appendix 10: likelihood ratio test formula:

$$LR = 2(\mathcal{L}_u - \mathcal{L}_r)$$

Where $\mathcal{L}$ is the log likelihood, and u and r indicate the unrestricted and restricted model, respectively.

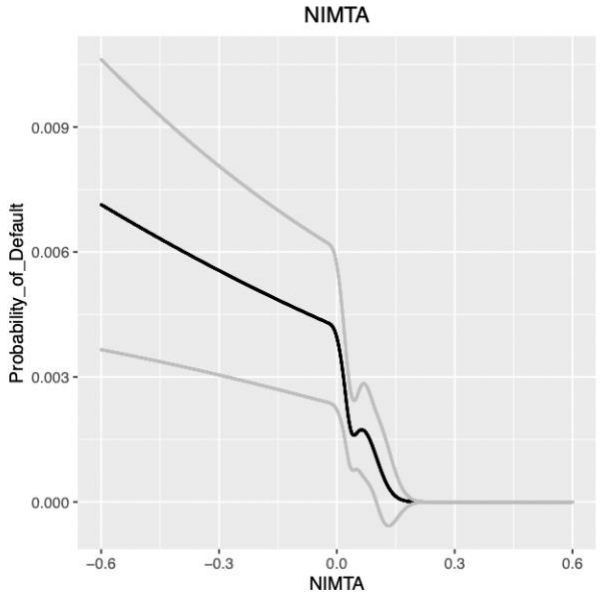
Appendix 11: summary of likelihood ratio test for restriction of model Logit 1U

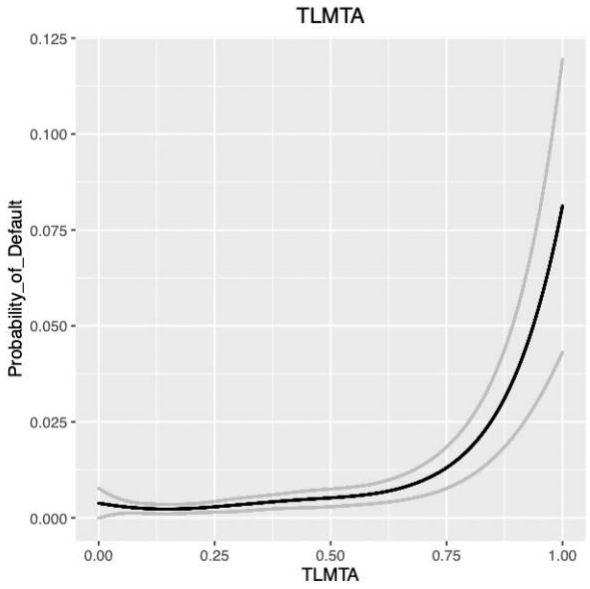
```
> lrtest(logit_1, logit_1U)
Likelihood ratio test

Model 1: DEF1Y ~ NIMTA + TLMTA + VOL + CR + LPRICE + EXR + EXR_lag1
Model 2: DEF1Y ~ NIMTA + TLMTA + VOL + CR + LPRICE + EXR + NIMTA_lag1 +
    EXR_lag1 + SIZE + AGE
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   8 -2102.9
2  11 -2102.2  3 1.3798     0.7103
```

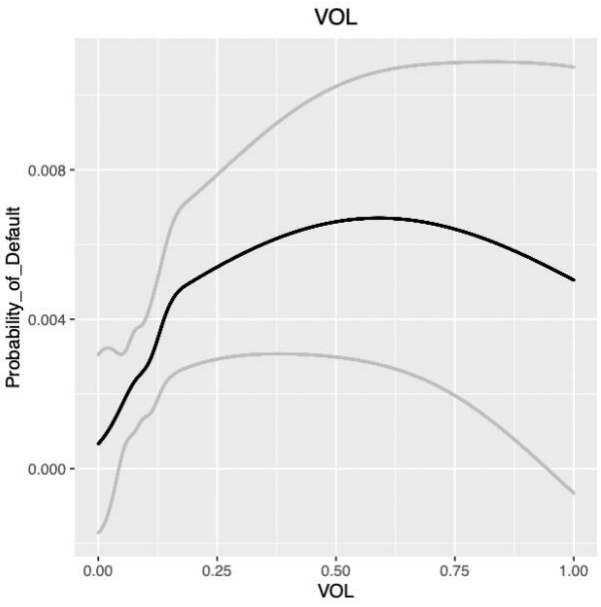
Appendix 12: summary of likelihood ratio test for restriction of model Logit 2U

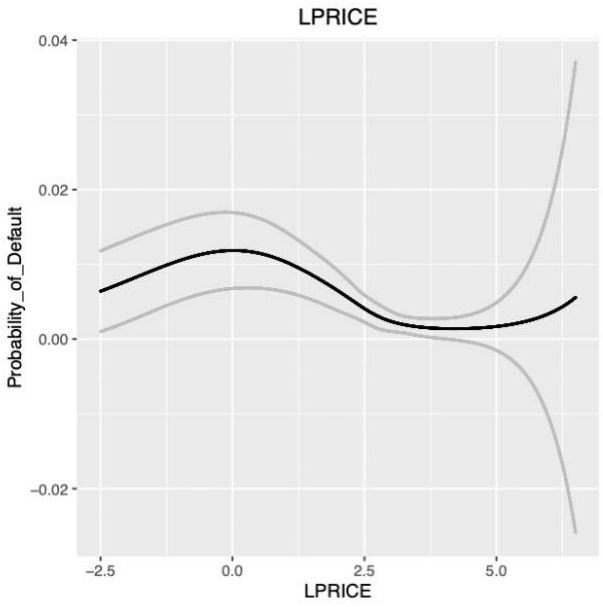
```
> lrtest(logit_2, logit_2U)
Likelihood ratio test

Model 1: DEF1Y ~ NIMTA + TLMTA + VOL + LPRICE + EXR + EXR_lag1
Model 2: DEF1Y ~ NIMTA + TLMTA + VOL + CD + LPRICE + EXR + NIMTA_lag1 +
    EXR_lag1 + SIZE + AGE
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   7 -2107.5
2  11 -2105.5  4 3.9661     0.4106
```

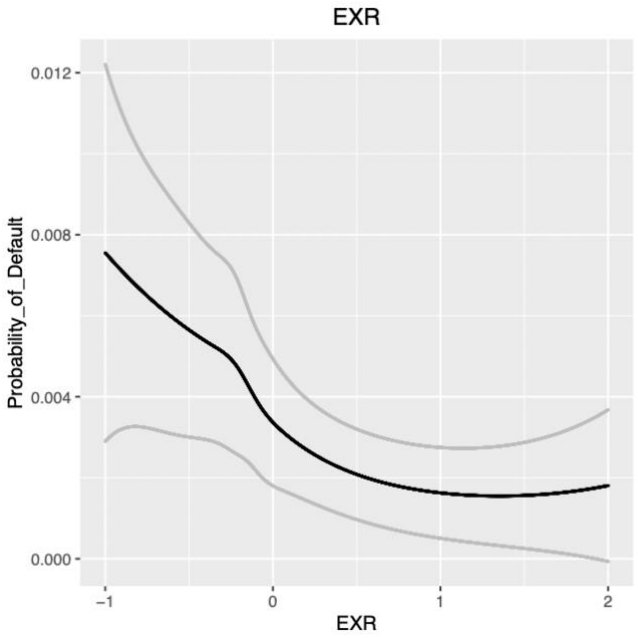
Appendix 13: plots of model Logit 1N











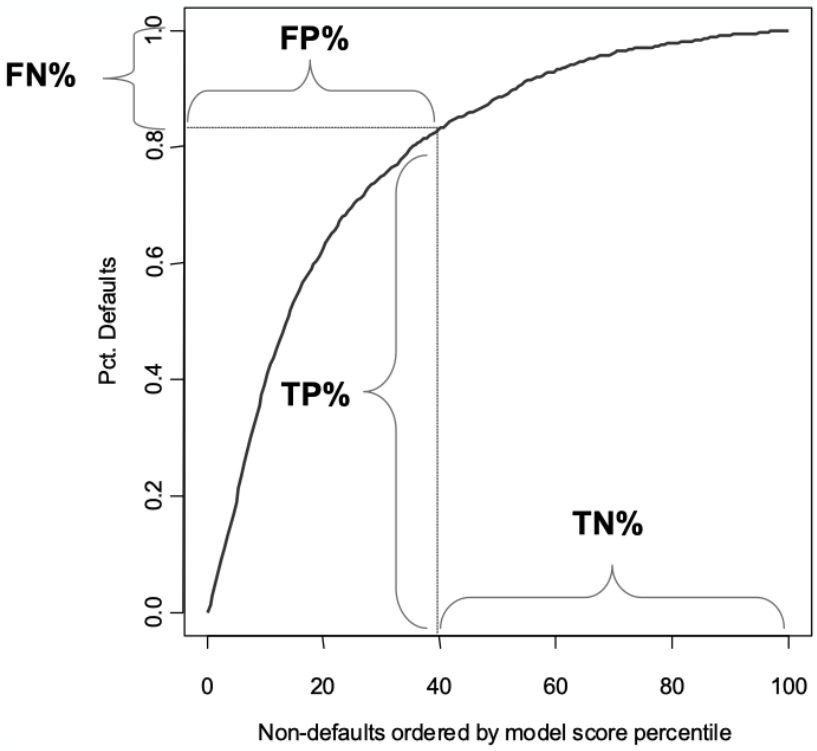
Appendix 14: in-sample contingency tables

Logit 2N		
	Actual	
Prediction	0	1
0	61938	80
1	11236	349

Logit 1N		
	Actual	
Prediction	0	1
0	61581	83
1	11593	346

Logit 1		
	Actual	
Prediction	0	1
0	60454	97
1	12720	332

Appendix 15: The relationship between ROC curves and contingency tables: TP = True Positives; FP = False Positives; TN = True Negatives; FN = False Negatives (Source: Stein, 2007).



Appendix 16: BIC for models Logit 1, 1N and 2N.

BIC Logit 1: 4296,468

BIC Logit 1N: 4015, 84

BIC Logit 2N: 3944,246