A comparative study of different survival analysis models for bankruptcy prediction

Li, Ting

2008

https://doi.org/10.32657/10356/2181

A COMPARATIVE STUDY OF DIFFERENT SURVIVAL ANALYSIS MODELS FOR BANKRUPTCY PREDICTION

LI TING

NANYANG TECHNOLOGICAL UNIVERSITY

A COMPARATIVE STUDY OF DIFFERENT SURVIVAL ANALYSIS MODELS FOR BANKRUPTCY PREDICTION

LI TING

HUMANITIES AND SOCIAL SCIENCES SCHOOL

2008

A Comparative Study of Different Survival Analysis Models for Bankruptcy Prediction

Li Ting

Humanities and Social Sciences School

A thesis submitted to the Nanyang Technological University in fulfillment of the requirement for the degree of Doctor of Philosophy

2008

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

ACKNOWLEDGEMENTS

This thesis has been possible only with the assistance, mentor advice and support in various ways from many parties. Particularly I would like to express my sincere gratitude to:

My supervisors, Associate Professor Koh Hain Chye and Associate Professor Cao Yong, for their loving care and guidance, particularly when times were tough.

My internal examiners, Associate Professor Wu Yuan and Associate Professor Wang Qinan, the anonymous external examiners for their valuable comments.

Professor Ruth M. Ripley from University of Oxford, for her generosity in providing me the whole package of syntax for all the survival ANNs applied in this study, and her patience in helping me out in coding-related problems.

I take full responsibility for any errors that remain.

ABSTRACT

Survival analysis is one of the most advanced techniques in bankruptcy prediction. However, to date, only few nonlinear techniques in survival analysis have been implemented in financial applications. This study introduces four nonlinear survival analysis, namely, partial logistic artificial neural networks ("PLANNs") (Biganzoli et al., 1998), the Cox's survival artificial neural networks ("Cox's ANNs") (Faraggi, 1995), the Weibull parametric survival artificial neural networks ("Weibull ANNs") (Ripley, 1998) and the log-logistic parametric survival artificial neural networks ("log-logistic ANNs") (Ripley, 1998) into bankruptcy prediction.

Based on the data of about 1,000 US corporations in consumer goods/services industries, estimation and prediction results of linear regression and neural networks are presented. A comprehensive comparison among the outputs from different models is conducted. Relevant topics such as misclassification costs and the optimal structure of neural networks are also discussed. The results of this study show that survival artificial neural networks ("ANNs") are superior to linear survival approaches in terms of prediction performance.

CHAPTER 1 INTRODUCTION

1.1 Research Motivation

Statistic models on corporate bankruptcy prediction for small and medium enterprises ("SMEs") have been developed for many years and are widely applied in financial institutions. An accurate bankruptcy prediction model can effectively reduce the operating cost of credit management and mitigate the potential credit risk for financial institutions, and hence is highly desirable for financial industries as well as researchers.

To achieve an accurate prediction on bankruptcies, many models have been developed in the past decades, such as multiple discriminant analysis ("MDA") (Altman, 1968), logistic regression and Merton's asset value model (Merton, 1974). Among of these approaches, survival models are gaining more and more attentions due to their capability in providing rich information, which includes not only the probabilities of defaults for firms but also the timing of bankruptcies.

Substantial efforts have been made to explore the applicability of survival models on SME bankruptcies. Mainly three types of estimation methods have been derived until now, namely, non-parametric estimation models (Kaplan and Meier, 1958), parametric estimation models (Collett, 2003) and semi-parametric estimation models (Cox, 1972). Successful applications of these estimations methods on bankruptcy prediction were frequently presented in the literature, for instance, by Lane et al. (1986), Bandopadhyaya (1994), Wheelock and Wilson (1995), Lee and Urrutia (1996) and etc. However, in these studies the estimation and prediction performance of survival regressions are usually compared with the results of earlier credit risk assessment models, such as MDA and logistic regression. Few studies have been conducted on comparing the performance among different survival estimation methods. Consequently, the questions such as under what situation which survival regression performs better have not been adequately addressed in the literature before.

As is known, due to the high flexibility in modeling the covariates' forms, the non-linear models usually can obtain better performance in estimation and/or prediction than their linear counter parties. Hence it is no surprise to observe the invention of the survival artificial neural networks ("ANNs"), a combination of traditional linear survival analysis and neural networks, which can achieve higher prediction accuracy. Various survival ANNs have been developed and implemented in bio-medical research since the early of 1990s (Ravdin and Clark 1992; Faraggi 1995; Mani 1999; Ripley 1998). However, very few applications of survival ANNs in financial topics have been observed by now. Will the scarcity of applications indicate the inapplicability of survival ANNs in financial research? Or instead, it reveals a promising research area in bankruptcy prediction analysis?

The general objectives of this study are to investigate the applications of three linear survival regressions and three non-linear survival ANNs on bankruptcies prediction, to provide a comprehensive comparison among these models and to comment on the model superiority in terms of their prediction outputs based on current empirical outputs, which have not yet been conducted before. Furthermore, topics such as the misclassification cost in prediction results of survival linear regressions and the determination of the optimal structure of neural networks are also discussed.

1.2 Outline of the Dissertation

This thesis investigates model assumptions, estimation method and other characteristics of six linear and non-linear survival models, namely, Kaplan-Meier method (Kaplan and Meier, 1958), parametric hazard regressions (Collett, 2003), the Cox's proportional hazard regression (Cox, 1972), partial logistic artificial neural networks ("PLANN") (Biganzoli et al., 1998), the Cox's survival ANN (Faraggi, 1995) and parametric survival ANNs (Ripley, 1998). Prediction accuracies of these six models based on a same set of data are compared. During the comparison among these models, questions such as under what condition which model can perform better than the others, and will the non-linear survival models always be superior to their linear counterparties are addressed in details.

In the following chapter, previous applications of linear survival models and simple neural networks in bankruptcy analysis are reviewed; furthermore, the development and applications of the survival neural networks in bio-medical researches are also discussed. Methodology developments on the six target models and other relevant technical issues such as validation of proportionality assumptions are elaborated in Chapter 3 and 4. A five-step data preparation is described in Chapter 5. Chapter 6 demonstrates the estimation results from all the six models, with a focus on the significance tests of parameters, numerical signs of parameters and the shape parameter estimation (for parametric models only). Comparisons among models in terms of the estimation capability are also conducted. After incorporating misclassification costs and the neural networks' structure, a comprehensive comparison of model prediction accuracies is carried out in Chapter 7.

The conclusions, limitations of this study and suggestions to further develop survival ANNs are included in Chapter 8.

CHAPTER 2 LITERATURE REVIEW

In this chapter, the applications of linear survival models in bankruptcy prediction in prior research is reviewed first, and then research studies about applications of survival ANNs in biomedical literature are discussed. The literature review here serves as a foundation for an in-depth understanding of the research motivations presented earlier. Gaps in the existing research are highlighted.

2.1 Overview of Linear Bankruptcy Prediction Models

Researchers have long been concerned with predicting corporate bankruptcy. Numerous linear prediction methods have been developed, such as by Altman (1968), Dopuch et al. (1987), Bell and Tabor (1991), Laitinen (1994) and etc.

In linear regression analysis, three generations have been evolved during the past decades (Duffie and Wang, 2003). The first generation of empirical corporate failure analysis is pioneered by Altman (1968). He applies multiple discriminant analysis ("MDA") to conduct his "Z-score" model to predict bankruptcy. Altman's seminal work has inspired further researches in developing the MDA model. These have included changing financial ratios in MDA models to improve model performance (Mensah 1983, Keasey and Watson1986, Lawrence and Bear 1986, Gentry Newbold and Whiteford1987) and adding non-financial ratios into models (Peel Peel and Pope 1986, Whittred and Zimmer 1984, Booth 1983). At the same time, much criticism of MDA models arose because the assumptions required by these models are usually violated, e.g., normal distribution assumptions on financial factors (Karels and Prakash 1987, Richardson and Davidson 1983).

The second generation of linear regression models, namely logistic/probit models (Ohlsn 1980, Gentry Newbold and Whiteford 1987, Dopuch et al. 1987), extends the MDA models. Although both MDA and logistic/probit can achieve high accuracy, these techniques have several shortcomings (see Altman et al. 1981, Eisenbeis 1977, Joy and Tollefson 1975). As stated in Lane et al. (1986), one drawback in using MDA and logistic/probit models to predict bank failures is that they can't provide information on the timing of bankruptcy, or equivalently, the explicit time to bankruptcy. How ever, the explicit time to bankruptcy of a firm is a piece of information that is as crucial as its probability of bankruptcy for investors. The awareness of an impending bankruptcy in one year or a possible bankruptcy in five years may result in different investor behaviors and decisions. The outputs of MDA and logistic/probit models are primarily limited to posterior probabilities.

To provide timing information on bankruptcy, survival regression is used for corporate bankruptcy prediction. Thus, the latest generation of modeling is dominated by survival models. Numerous applications of using survival regressions on bankruptcy predictions have been explored (Lane et al.1986, Whalen 1991, Bandopadhyaya 1994, Henebry 1996, Shumway 2001, etc.). The following section will provide a more detailed literature review on these applications, in which adventures of survival regression over other linear models are discussed, and limitations of these survival regressions are also highlighted.

2.2 Applications of Linear Survival Models on Bankruptcy Analysis

Early financial applications of survival analysis are concentrated on bank failure prediction, with attentions on comparing survival regressions with other linear models. For example, in 1986 Lane et al. construct a bank failure prediction using the

Cox's proportional hazard regression (Cox, 1972) with time-independent covariates. Their results show the superiority of survival analysis to the traditional MDA models: the Cox model can provide additional information regarding the relationship between time before default and various explanatory variables without the loss of prediction capability. In addition, no significant difference exists in total prediction accuracy in both models while the Cox's regression possesses a lower type I error rate which is regarded as more important in bank failure prediction. Subsequent to Lane et al. (1986), Henebry (1996) demonstrates the using of the Cox's time-varying covariate ("Cox's TVC") survival regression in bank failure prediction by adding cash flow information in covariates which obtained higher prediction accuracy. Similar work has been done by Whalen (1991), Wheelock and Wilson (1995, 2000) and Molina (2002).

Due to the advantage of dynamically capturing changes of firm features from the firm's initiation to its bankruptcy, the Cox's TVC model (Cox, 1972) is gaining more and more attentions in recent years. Audretsch and Mahmood (1995) apply a Cox's TVC model in studying the survival of new firms with a dataset of more than 12,000 U.S. manufacturing. They find that both the structure of ownership and start-up size can substantially shape the likelihood of survival. Shumway (2001) compares the dynamic TVC hazard model with "static models" (i.e., the single-period classification models such as MDA and logistic) and summarizes the following advantages of Cox's TVC model over other linear regressions: First, the Cox's TVC model can automatically account for a firm's period at risk when sampling periods are long while the MDA and logistic models can not. Shumway argues that, usually firms have different periods at risk, i.e., some file for bankruptcy after many years of being at risk while other fail in their first year. The MDA and logistic models use same horizons for all firms (usually 1 year before default) which ignore the differences in the period at risk. Second, the Cox's TVC model can utilize all available data in multiple years but MDA and logistic model can not. As a consequence, the TVC model can produce more efficient out-of-sample forecasts by using time series data in multiple years. As is shown in Shumway (2001), the Cox's TVC models can identify more significant variables and have higher accuracy in out-of-sample forecasting. Third, the Cox's TVC can generate asymptotically consistent outputs but MDA and logistic models can not. Shumway shows that theoretically the Cox's TVC model is equivalent to a multi-period logistic model, but the outputs of Cox's TVC model is asymptotically consistent while outputs of static models are not.

Other than the Cox's proportional hazard, estimation methods such as Weibull and log-logistic hazards are also employed in corporate bankruptcy analysis and performance of these survival regressions is compared with other linear approaches as well. For example, Lee and Urrutia (1996) conduct a study on insolvency in the insurance industry with using the Weibull survival regression. They compare the prediction results of the Weibull survival regression and the logistic regression and find that the Weibull hazard model is superior to the logistic model in identifying more significant variables with attaining comparable forecasting accuracy.

With several estimation methods in different hazards distribution assumptions available, the following questions become interesting: among these survival regression estimation methods, which one is better? Under what circumstance, which survival estimation method should be adopted? Unfortunately little literature can be found to comprehensively address these questions.

Besides, there are two problems with the above linear approaches. First, linear models require that the set of variables used to distinguish bankrupt and non-bankrupt firms should be generally linear. Second, in linear regressions variables are treated as completely independent. In reality, however, a variable may signal bankruptcy both when it is higher than normal and when it is lower than normal, or a variable's value may be considered acceptable under some conditions, yet risky under others (Karels and Prakash, 1987). To overcome these limitations, non linear approaches, such as artificial neural networks ("ANNs") are adopted. The literature review on bankruptcy prediction with ANNs, especially survival ANNs, is provided in the following section.

2.3 Overview of Survival Neural Networks

ANN models were introduced into bankruptcy studies from early 1990s and become popular in this field in later years (Bell et al. 1990, Tam and Kiang 1990 & 1992, Coats and Fant 1993, Klersey and Dugan 1994, Koh and Tan 1999, Etheridge et al. 2000, etc.). The ANNs have been proven to be at least as successful as linear models in terms of overall accuracy (Kim and Scott, 1991; Altman, Marco and Varetto, 1994; Podding, 1994; Yang, Platt and Platt, 1999; Odom and Sharda, 1990; Webb and Lowe, 1990; and Utans and Moody, 1991). For example, in Tam and Kiang (1992), the bank failure prediction capability of ANNs is compared to that of other models such as MDA, logistic and Decision Trees. The empirical results show that in the out-of-sample prediction, ANNs achieve the highest overall accuracy and the lowest type I error rate among all these models compared.

The techniques of standard neural networks and linear survival analysis can be combined to create a new type of models that can conduct the nonlinear survival analysis. Historically, the ANNs for survival analysis were initially devised from early 1990s in medical research, especially in oncology studies. By now, four types of survival ANNs have been developed, namely, the Kaplan-Meier survival ANNs, the

Cox's survival ANNs, the parametric survival ANNs and the hierarchical survival ANNs.

- The Kaplan-Meier survival ANNs

The Kaplan-Meier survival ANNs which use the non-parametric method (Kaplan and Meier, 1958) in estimating the hazards is frequently applied in biomedical research (Burke1994, Ravdin and Clark 1992, De Laurentiis and Ravdin 1994, Lapuerta et al 1995). The Kaplan-Meier ANNs produce the survival probabilities and usually are equipped with complex output encoding settings. For example, in Street (1998), the network has ten output nodes which represent the survival probabilities of events happened in the first year, the second year and so on. The limitations of Street's approach are that his model cannot generate monotonically decreasing outputs, which means a non-monotone survival curve is possible. Furthermore, no extension is provided to deal with time-varying inputs under this framework.

A variation on the Street's method was developed by Mani(1999). He uses the hazard probability instead of survival probability as outputs. The survival probabilities may then be estimated by the ordinary relationship of hazard function and the survival function in a traditional Kaplan-Meier approach.

The generated survival curve in Mani's model thus is monotonically decreasing which simplifies the interpretation and increases robustness (Mani et al., 1998). However, the topic of time-varying inputs has been left unaddressed.

Analogous to Mani, Brown et al.(1997) suggests a single neural network with multiple outputs to predict hazard rate. In their network, the output of network is the hazard function. The authors suggest training the neural network to minimize the sum of squared error criterion. Brown's approach can have monotonic survival curves, but again, no extension is presented to deal with time-varying inputs.

- The Cox's survival ANNs

The original idea of the Cox's survival ANN was presented in Faraggi (1995). In his breast cancer research, by replacing the linear combination of variables and parameters in traditional Cox's proportional hazard regression with the output of a standard network, Faraggi successfully introduces the nonlinear transformation to the Cox's proportional hazard analysis. Faraggi's method allows preserving all the advantages of the classical proportional hazard model, and in addition, it can easily be extended to incorporate time-varying covariates. However, the proportionality assumption still must be held under this approach.

Faraggi's method in Cox's survival ANN has been further developed by Mariani et al. (1997) and Ripley (1998). Particularly, in Ripley's approach, the input nodes of the network can directly connect with not only the hidden nodes but also the output nodes, which offer higher flexibility of the model.

Biganzoli (1998) designed another ANN under the framework of the Cox's proportional hazard distribution, the partial logistic survival ANN ("PLAAN"), which can be implemented directly with standard software. Different from other Cox's survival ANNs, the outputs of PLAAN are conditional hazard probabilities instead of the conditional survival probabilities.

- The parametric survival ANNs

The parametric survival ANNs were firstly developed and implemented on breast cancer relapse prediction by Ripley in her 1998 paper. Similar to the Cox's survival ANN designed by Faraggi (1995), the parametric survival ANNs is obtained simply by replacing the linear combination of the covariates with the output of a standard neural network, with using same maximum likelihood mechanism as in the parametric survival regressions. Usually Weibull survival ANNs and log-logistic survival ANNs are representatives of parametric survival ANNs.

- The hierarchical survival ANNs

The hierarchical survival ANNs was introduced by Ohno-Machado (1995, 1997). Basically the hierarchical survival ANN is a system of standard neural networks which use the outputs from the previous network as the inputs for the next one. For example, in Ohno-Machado et al. (1995), the authors initially use the first network to classify observations into big groups, letting all targeted low frequency patters fall into one big group. Later, he uses the second ANN to further classify this big group into smaller groups. Repeatedly, the last ANN is used to classify the target patterns from others.

Although essentially identical to design of standard neural networks in each individual network in the hierarchical networks, Machado's models can provide not only the accumulated survival probability but also the conditional survival probability which differentiates Machado's models as a kind of survival ANNs. Besides, hierarchical models can easily include time-dependent inputs.

However, the hierarchical survival ANNs have two limitations. First, as Ohno-Machado et al. admitted, hierarchical models may result in non-monotonic survival curves. In addition, the necessity of using multiple neural networks and how to combine these networks represents an important scalability problem which makes the hierarchical model less suitable for handling large data set.

2.4 Applications of Survival Neural Networks on Bankruptcy Prediction

While survival neural networks have been used extensively in medical research, the only application of survival ANNs in financial analysis found by now is the working paper on personal loan analysis by Baesens et al. (2004).

In this study, a Kaplan-Meier survival ANN was employed and the prediction capability of the survival ANN is compared with logistic regression as well as the Cox's proportional hazard regression. They show that although the prediction accuracy in the survival ANN is lower than the logistic model, the performance of survival ANN is better than Cox's proportional hazard regression. They emphasized that survival ANNs are desirable approaches in business failure prediction for the reason that they can provide much higher prediction performance than traditional linear survival analysis.

However, since the Cox's proportional hazard regression and parametric methods have more advantages than the Kaplan-Meier model, it will be more meaningful to apply the Cox's survival ANN and the parametric survival ANN as the comparable representatives of nonlinear survival approaches when comparing with the Cox's proportional hazard regression. In addition, since the only difference between the Cox's/parametric survival ANNs and their linear counter parties is their nonlinear combination of covariates in their likelihood functions, the comparisons of outputs from the Cox's/parametric survival ANNs and the corresponding linear regressions can be more convincing in proving whether the non-linearity has superiority to linearity in combining covariates. Unfortunately, no relevant studies could be found in previous research and the comparison between survival regressions and survival ANNs in bankruptcy prediction still almost remains as a blank.

This thesis is targeted to fill this blank by comparing outputs from different linear survival regressions with outputs from their corresponding ANNs based on a same set of data.

CHAPTER 3 METHODOLOGY– SURVIVAL REGRESSIONS

To facilitate the descriptions, this chapter starts with a short description on parameter notations. After that the three estimation methods used in survival regressions, namely, non-parametric method (Kaplan and Meier, 1958), semi-parametric method (Cox, 1972) and parametric methods (Collett, 2003), are introduced one by one. For all of these three methods their assumptions are discussed, advantages are compared and limitations are highlighted.

3.1 Notations of Parameters

The survival function S(t) is defined as the probability of a firm surviving beyond a specified time period t or experiencing failure after time t. S(t) can be expressed as

$$S(t) = 1 - F(t) = P\{T \geq t\} \tag{3.1}$$

where $F(t) = P(T < t)$ which is the cumulative distribution function specifying the probability that an event (bankruptcy) has happened before time t.

The hazard function can be expressed as

$$h(t) = f(t)/S(t) \tag{3.2}$$

where $f(t)$ is the corresponding density function for F(t). Roughly, $h(t)$ is the rate at which events (bankruptcies) will happen immediately after t, given that firms survive until t. A more precise definition of $h(t)$ is

$$h(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T \leq t + \Delta t | T \geq t]}{\Delta t} \tag{3.3}$$

As is shown in Kiefer (1988), the relationship among $F(t)$, $f(t)$, $S(t)$, and $h(t)$ can be illustrated as follows: if the time interval from the year of a firm's initiation to the year it files for the bankruptcy is defined as time to bankruptcy, the probability that the time to bankruptcy is less than t years is $F(t)$; equivalently, the probability that time to bankruptcy is longer than t years is $S(t)$. The probability that time to bankruptcy ends between t and t+Δt years is $f(t)*\Delta t$, while the probability that time to bankruptcy ends between t and t+Δt conditional on the firm keeping survived until year t is $h(t)*\Delta t$.

Combining equations (3.1) and (3.2), it leads to

$$h = f / S = (dF / dt)/ S = (-dS / dt)/ S \qquad (3.4)$$

From which, the following relationship between $h(t)$ and $S(t)$ can be derived:

$$h(t) = -d \ln S(t)/ dt \qquad (3.5)$$

Besides $h(t)$ and $S(t)$, the integrated hazard $\Lambda(t) = \int_0^t h(u)du$ is also a useful function in practice. The relation of integrated hazard to the survival function is:

$$S(t) = \exp[-\Lambda(t)]. \qquad (3.6)$$

3.2 The Kaplan-Meier Non-Parametric Method

The Kaplan-Meier non-parametric method (Kaplan and Meier, 1958) is usually employed as a rough graphical analysis in survival studies.

In the Kaplan-Meier method, the estimator of the hazard function $h(t_j)$, which is the probability of a firm going bankrupt at year t_j conditional upon it having survived until t_j, can be written as:

$\hat{h}(t_j) = h_j / n_j$ where h_j is the number of bankruptcies at year t_j and n_j is the number of firms surviving until year t_j.

The corresponding estimator for the survival function is

$$\hat{S}(t_j) = \prod_{i=1}^{j} (n_i - h_i)/n_i = \prod_{i=1}^{j} \left(1 - \hat{h}_i\right) \tag{3.7}$$

Graphically, the Kaplan-Meier survival curve appears as a step function with a drop at each year. In plotting the hazard curve, the conditional probability of failure is an estimator of the probability that a firm will get bankrupt in the interval of one year, given that it makes its survival to the start of the year.

3.3 The Parametric Method

The parametric survival regressions (Collett, 2003) are also known as the accelerated failure time ("AFT") models (Allison, 1995). In the most general form, the AFT model describes a relationship between the survival functions of any two individuals, e.g., if $S_i(t)$ is the survival function for individual i, then for any other individual j, the AFT model holds that

$S_i(t) = S_j(\theta_{ij} t)$ for all t

Where θ_{ij} is a constant that is specific to the pair (i, j). This model says, in effect, that what makes one individual different from another is the rate at which they age. Assume the survival time t satisfies $y(t) = X\beta + \sigma W$, where W follows some given distribution and y is a given transformation. In the AFT model, $y = \log(t)$.

Let Ti be a random variable denoting the time to bankruptcy for the *ith* individual in sample, and let $x_{i1},...x_{ik}$ be the values of k covariates for that individual, the model is then

$$\log T_i = \beta_0 + \beta_1 x_{i1} + ...\beta_k x_{ik} + \sigma\varepsilon_i \qquad (3.8)$$

Where ε_i is a random disturbance term, and $\beta_0...\beta_k$, and σ are parameters to be estimated.

The likelihood function in parametric survival regression is:

$$L(\theta) = \sum_1^n d_i \ln f(t_i, \theta) + \sum_1^n (1 - d_i) \ln S(t_i, \theta) \qquad (3.9)$$

here, d_i =1 if *ith* individual is defaulted, d_i =0 if non-defaulted.

Here, the defaulted firms are contributing a density term $f(t_i, \theta)$ and non-defaulted firms are contributing a probability $S(t_i, \theta)$. So, in the parametric survival analysis all required to construct a likelihood function are the hazard function and the survival function.

According to the distribution of Ti, there are generally five types of hazard distributions, i.e., exponential, Weibull, gamma, log-normal and log-logistic.

- **The Exponential Model**

It is the simplest model in parametric survival regressions. It specifies that ε_i has a standard extreme-value distribution, and constrains σ =1.

Hazard function of exponential model is

$$h(t|x) = h\exp(x\beta) \qquad (3.10)$$

Survival function of exponential model is

$$S(t|x) = \exp[-h + \exp(xb)] = \exp(-ht)^{\exp(x\beta)} \qquad (3.11)$$

It is easy to find that for the exponential model, the hazard is constant over time.

- **The Weibull Model**

The hazard function of the Weibull model is

$$h(t|x) = \alpha\gamma t^{\gamma-1}\exp(x\beta) \qquad (3.12)$$

The survival function of Weibull model is

$$S(t|x) = \exp[-\alpha t^{\gamma}\exp(x\beta)] \qquad (3.13)$$

where $\gamma = 1/\sigma$

In the Weibull model, the assumption that ε_i has a standard extreme-value distribution is still held, but the assumption that σ =1 is relaxed. When σ >1, the hazard decreases with time; when 0.5< σ <1, the hazard is increasing at a decreasing rate; when 0< σ <0.5, the hazard is increasing at an increasing rate; and when σ =0.5, the hazard function is an increasing straight line with an origin at 0.

The Weibull distribution has long been the most popular parametric model in the bio-statistical literature for two reasons. First, it has a relatively simple survival function that is easy to handle mathematically; second, in addition to be an AFT model, the Weibull model is also a proportional hazards model. This means that its coefficients (when suitably transformed) can be interpreted as relative hazard ratios. In fact, the Weibull model (and its special case, the exponential model) is the only model that is simultaneously a member of both these classes.

In this study, the applications of the Weibull distribution both in parametric regression approach and the parametric survival ANN approach are exploited.

- **The Log-Normal Model**

The hazard function of log-normal model is:

$$h(t|x) = 2(\pi)^{1/2}\gamma t^{-1}\exp\left(\frac{-\gamma^2(\log(\lambda t))^2}{2}\right) \tag{3.14}$$

The survival function of log-normal model is:

$$S(t|x) = 1 - \Phi(\gamma\log(\lambda t)) \tag{3.15}$$

where $\lambda(x) = \exp(x\beta)$, Φ is the incomplete normal integral.

Unlike the Weibull model, the log-normal model has a non-monotonic hazard function. The hazard is 0 when t=0. It rises to a peak and then declines to toward 0 as t goes to infinity. The log-normal is not a proportional hazards model, and its hazard cannot be expressed in a closed form since it involves the c.d.f. Φ of a standard normal variable.

- **The Log-logistic Model**

The hazard function for log-logistic model is:

$$h(t) = \frac{\lambda\gamma(\lambda t)^{\gamma-1}}{\left[1+(\lambda t)^{\gamma}\right]^2} \tag{3.16}$$

The survival function for log-logistic Model is:

$$S(t) = \frac{1}{1+(\lambda t)^{\gamma}} \tag{3.17}$$

Where $\lambda(x) = \exp(x\beta)$, $\gamma = 1/\sigma$

The log-logistic model is another model that allows for an inverted U-shaped hazard, which assumes that ε_i has a logistic distribution with p.d.f.

$$f(\varepsilon) = \frac{e^{\varepsilon}}{(1+e^{\varepsilon})^2} \tag{3.18}$$

A little algebra shows that the survival function can be written as

$$\log\left[\frac{S(t)}{1-S(t)}\right]=\beta_0^*+\beta_1^*x_1+\ldots+\beta_k^*x_k-\gamma\log t \qquad (3.19)$$

Where $\beta_i^*=\beta_i/\sigma$, $\gamma=1/\sigma$ for i=1, .., k, which is nothing more than a logistic model.

Because $S(t)$ is the probability of surviving to time t, $\frac{S(t)}{1-S(t)}$ is the odds of surviving to time t. Let $S_i(t)$ and $S_j(t)$ be the survival functions for any two individual i and j. The log-logistic model says that

$$\frac{S_i(t)}{1-S_i(t)}=\vartheta_{ij}\left(\frac{S_j(t)}{1-S_j(t)}\right) \text{ for all t} \qquad (3.20)$$

where ϑ_{ij} is some constant that is specific to the part (i, j).

In this study, the log-logistic distribution is implemented in both survival regressions and survival ANNs.

When the shape parameter in log-logistic distribution is no less than 1, (σ>=1), the hazard curve of this log-logistic distribution decreases in a decreasing speed; when 0<σ<1, the hazard curve shows an inverted U shape, that is, the hazard initially increase with time, after hitting a peak it then declines to 0 gradually.

- **The Gamma Model**

The generalized gamma model has one more parameter than any of the above models; its hazard function can take on a wide variety of shapes. In particular, the exponential, Weibull, and log-normal models (but not the log-logistic) are all special cases of the generalized gamma model. Although the generalized gamma model has wide richness in its hazard model, because of its complicated formula for hazard model and considerable difficulty in computation, the gamma model is seldom used in practice.

3.4 The Cox's Proportional Hazard ("Cox's PH") regression

The Cox's proportional hazard model (Cox, 1972) has been widely used in economics and other disciplines. In this model the hazard function depending on a vector of explanatory variables x with unknown coefficients β and h_0 is factored as

$$h(t, x, \beta, h_0) = \Psi(x, \beta) h_0(t) \tag{3.21}$$

Where h_0 is a baseline hazard corresponding to $\Psi(\ )$=1. A specification of $\Psi(\ )$ usually used is $\Psi(x, \beta) = \exp(x'\beta)$

So, it is easily to get

$$h(t, x, \beta, h_0) = \exp(x'\beta) h_0(t) \tag{3.22}$$

Here, the covariates x's are assumed to be constant over time.

Because $\partial \ln h(t, x, \beta, h_0)/\partial x = \partial \ln \Psi(x, \beta)/\partial x$, the proportional effect of x on the conditional probability of having a bankruptcy does not depend on how long this firm's time to bankruptcy is.

In a important special case $\Psi(x, \beta) = \exp(x'\beta)$,

$$\partial \ln h(t, x, \beta, h_0)/\partial x = \beta \tag{3.23}$$

Hence the coefficient can be interpreted as the constant proportional effect of x on the conditional probability of bankruptcy.

The survival function for t is given by

$$S(t) = \exp[-\Lambda_0(t) \exp(x'\beta)] \tag{3.24}$$

where $\Lambda_0(t) = \int h_0(u) du$ is the integrated baseline hazard.

In the Cox's PH models, β's still are estimated by maximizing the likelihood function as in other survival regression models. But the Cox's PH models use the partial likelihood method which does not need to specify the form of the baseline hazard function h_0.

In this study, the applications of the Cox's PH method in both regression approach and survival ANN approach are exploited.

Under the framework of the Cox's PH method, partial likelihood function, treatments on ties of defaults and regression diagnostics are three essential parts which deserve more attentions. Sections 3.4.1, 3.4.2 and 3.4.3 are therefore dedicated to explain these issues in details.

3.4.1 Partial Likelihood Method

The conditional probability that individual firm *i* is the one that got bankrupt at t_i, given that the firms in the set R_i are at risk of bankruptcy, and given further that exactly one bankruptcy occurred at t_i, is

Prob{firm i got bankrupt at $t_i \mid R_i$ and one bankruptcy at t_i }=

$$\frac{\Pr ob\left\{ firm\ i\ got\ bankrupt\ at\ t_i \middle| R_i \right\}}{\Pr ob\left\{ one\ got\ bankrupt\ at\ t_i \middle| R_i \right\}}$$

Thus, a general expression for the partial likelihood for data with fixed covariates from a proportional hazard model is

$$PL = \prod_{i=1}^{n} \left[\frac{\exp(\beta x_i)}{\sum_{j=1}^{n} Y_{ij} \exp(\beta x_j)} \right]^{\delta_i} \qquad (3.25)$$

where Y_{ij} =1 if $t_j \geq t_i$ and Y_{ij} =0 if $t_j \prec t$; δ_i =0 if firm i is non-default and δ_i =1 if i is default.

The intuition here is that, in the absence of all information about the baseline hazard, only the order of the bankruptcies provides information about the unknown coefficients.

As usual, it is convenient to maximize the logarithm of the likelihood, which is

$$\log PL = \sum_{i=1}^{n} \delta_i \left[\beta x_i - \log\left(\sum_{j=1}^{n} Y_{ij} \exp(\beta x_j) \right) \right] \quad (3.26)$$

Most partial likelihood programs use some version of the Newton-Raphson algorithm to maximize this function with respect to β .

Compared with other estimation approaches in survival analysis, the Cox's PH function is the most popular one. There are some reasons for this enormous popularity (Allison 1995): first, unlike the parametric methods, the Cox's method does not require that some particular probability distribution is chosen to represent survival times. As a consequence, the Cox's method is considerably more robust; second, the Cox regression makes it relatively easy to incorporate time-dependent covariates, that is, covariates may change in value over the observation period; third, the Cox regression permits a kind of stratified analysis that is very effective in controlling for dummy variables.

3.4.2 Handling of Ties

Tied event times (same time to bankruptcy for numerous firms) are a common feature in corporation data since harsh external environment (i.e., economy crisis) may drag several firms into defaults at same time and tightly correlated business interest probably pull a firm into default when its business partners are default.

When there are tied bankruptcy times in the sample, the partial log likelihood function involves permutations so it can be time-consuming to compute. Breslow (1974) has derived a satisfactory approximate log likelihood function when the number of ties is not large. Farewell and Prentice (1980) show that the Breslow approximation deteriorates as the number of ties at a particular point in time becomes a large proportion of the number of cases at risk. Efron (1977) derives another approximation to the true likelihood that is significantly more accurate than the Breslow approximation and often yields estimates that are much closer to the exact results than Breslow's approximation. This improvement comes with only a trivial increase in computation time.

Efron's method is adopted in handling all tied situations in this study.

3.4.3 Regression Diagnostics

It is important to inspect the goodness of fit before the models are used for prediction. In survival analysis, there are various aspects and numerous methods to accomplish the inspection of how well the model is specified. In this study, only the test on the appropriate form of covariates and the test on the validation of the proportionality assumption are conducted.

- Martingale Residuals to Test for the Appropriate Form of Covariates

To define the martingale residual in the most general sense, suppose that, for the jth firm in the sample, it has a vector $X_j(t)$ of possible time-dependent covariates. Let $N_j(t)$ have a value 1 at time t if this firm got bankrupt and 0 if not.

Let $Y_j(t)$ be the indicator that firm j is under study at a time just prior to time t. Finally, let b be the vector of regression coefficients and $\hat{H}_0(t)$ the estimator of the cumulative baseline hazard rate. The martingale residual is defined as

$$\hat{M}_j = N_j(\infty) - \int_0^{\infty} Y_j(t)\exp\left[b^t X_j(t)\right] d\hat{H}_0(t),\ j = 1, \ldots, n \tag{3.27}$$

When all the covariates are fixed, the martingale residual reduces to

$$\hat{M}_j = \delta_j - \hat{H}_0(T_j)\exp\left(\sum_{k=1}^{p} X_{jk} b_k\right) = \delta_j - r_j,\ j = 1, \ldots, n \tag{3.28}$$

The residuals have the property of $\sum_{j=1}^{n} \hat{M}_j = 0$. Also, for large samples the $\hat{M}_j$ ’s are an uncorrelated sample from a population with a zero mean.

According to Klein and Moeschberger (1997), the martingale residuals can be interpreted as the difference over time of the observed bankruptcies minus the expected number of bankruptcies under the assumed Cox model, that is, the martingale residuals are an estimate of the excess number of bankruptcies seen in the data but not predicted by the model.

The rationale of using martingale residuals to examine the best functional form of covariates is: suppose that the covariate vector X is partitioned into a vector X^*, which is known as the proper functional form of the Cox model, and a single covariate X_1 whose proper functional form is unsure. Assume that X_1 is independent of X^*. Let $f(X_1)$ be the best function of X_1 to explain its effect on the Cox’s survival regression. The Cox model is, then,

$$H\left(t \middle| X^*, X_1\right) = H_0(t)\exp\left(\beta^* Z^*\right)\exp\left[f(X_1)\right] \tag{3.29}$$

To find f , the Cox model is fit to the data based on X^* and compute the martingale residuals, $\hat{M}_j$, j=1,…, n. these residuals are plotted against the value of X_1 for the *jth* observation . A smoothed fit of the scatter diagram is typically used. The smoothed-fitted curve gives an indication of the function f . If the plot is linear, for example, no transformation of X_1 is needed (Klein and Moeschberger, 1997; Therneau et al., 1999). However, if some martingale residuals plots are not linear, it actually implies that, the simple linear combination of covariates in the Cox's PH regression may be inappropriate.

- Validation Test on the Proportionality Assumption

In this thesis, two types of tests on the proportionality assumption of Cox's regression are employed.

The first test is relatively simple—the stratification graphic test on the proportionality. According to Allison(1995), The stratification can be used for testing the proportionality assumption in the Cox's PH model. Specifically, the graphs of log-log survival function for two values in stratified dummy variable can be used for this test. Take a dummy variable D as an example, the rationale of this test can be illustrated as: assume $h_1(t)$ is the hazard function for a firm whose D = 1 and $h_2(t)$ is the hazard function for another firm whose D = 0. If two hazard functions, $h_1(t)$ and $h_2(t)$, are proportional, then

$$h_1(t) = \tau h_2(t) \qquad (3.30)$$

where τ is the constant of proportionality.

Because $S(t) = \exp\left[-\int h(u)du\right]$ (see Equation 3.6) it is easily to show that $S_1(t) = [S_2(t)]^{\tau}$.

Taking the logarithm, multiplying by -1, and taking the logarithm a second time yields

$$\log[-\log S_1(t)] = \log\tau + \log[-\log S_2(t)] \tag{3.31}$$

Which says that the two log-log survival curves differ by a constant amount, $\log\tau$.

That is, in plotting the log-log survival function, when the hazards are proportional, the log-log survival functions should be parallel.

The second test used in examining the proportionality assumption is the tests of Schoenfeld residuals.

The Schoenfeld residual can be derived from (3.26). As demonstrated in Hosmer et al. (1999), the Schoenfeld residuals defined by Schoenfeld (1982) can be generalized as followed:

$$\frac{\partial PL}{\partial\beta_k} = \sum_{i=1}^{n}\delta_i\left\{x_{ik} - \frac{\sum_{j\in R(t_i)} x_{jk}e^{x_j'\beta}}{\sum_{j\in R(t_i)} e^{x_j'\beta}}\right\} = \sum_{i=1}^{n}\delta_i\left[x_{ik} - \overline{x_{w_ik}}\right] \tag{3.32}$$

where $\overline{x_{w_ik}} = \dfrac{\sum_{j\in R(t_i)} x_{jk}e^{x_j'\beta}}{\sum_{j\in R(t_i)} e^{x_j'\beta}}$

The estimator of the Schoenfel residual for the ith subject on the kth covariate is obtained from equation (3.32) by substituting the partial likelihood estimator of the coefficient, $\hat{\beta}$, $\hat{r}_{ik} = \delta_i(x_{ik} - \hat{\overline{x_{w_ik}}})$, Where $\overline{\hat{x}}_{w_i k} = \dfrac{\sum_{j \in R(t_i)} x_{jk} e^{x_j' \hat{\beta}}}{\sum_{j \in R(t_i)} e^{x_j' \hat{\beta}}}$ is the estimator of the risk set conditional mean of the covariate.

Notice that since $\hat{\beta}$ is the solution to the f.o.c. equal to zero, which is just equation (3.32), the sum of the Schoenfeld residuals is zero.

In assessing the proportional hazard assumption, if the proportional hazards assumption is hold, the Schoenfeld residuals should be a random walk. Conversely, assume that some variable has a large positive effect early but that the effect trails off, the proportional hazards do not hold. The fitted models will underestimate the true effect of this variable for small t, and overestimate it for large t, which can be reflected into an early positive trend followed by a late negative trend in the Schoenfeld residuals.

In further development, Harrell (1986) suggests using the correlation of rank (time) with this residual as a test for non-proportional hazards. Grambsch and Therneau (1994) propose that, a rescaled Schoenfeld residual can correct for correlation among the covariates and be more interpretable. Specifically, Grambsch and Therneau scale the Schoenfeld residuals by an estimator of its variance which yields a residual with greater diagnostic power, sometimes, this rescaled residuals are called Grambsch and Therneau Residuals:

$$r_i^* = \left[\widehat{Var(\hat{r}_i)} \right]^{-1} \hat{r}_i \qquad (3.33)$$

where $\hat{Var}(\hat{r}_i)_{kk} = \sum_{j \in R(t_i)} \hat{w}_{ij}(x_{jk} - \overline{\hat{x}}_{w_i k})^2$ which is the diagonal elements in the Variance-Covariance matrix;

$$\hat{Var}(\hat{r}_i)_{kl} = \sum_{j \in R(t_i)} \hat{w}_{ij}(x_{jk} - \overline{\hat{x}}_{w_i k})\left(x_{jl} - \overline{\hat{x}}_{w_i l}\right) \quad (3.34)$$

where $\hat{w}_{ij} = \dfrac{e^{x_j' \hat{\beta}}}{\sum_{l \in R(t_i)} e^{x_l' \hat{\beta}}}$

Grambsch and Therneau suggest using of an easily computed approximation for the scaled Schoenfeld residuals:

$$\left[\hat{Var}(\hat{r}_i)\right]^{-1} = m\hat{Var}\left(\hat{\beta}\right) \quad (3.35)$$

since $\hat{Var}(\hat{r}_i)$ tends to be fairly constant overtime.

In addition, Grambsch and Therneau (1994) propose both the statistical test and graph test for assessing the proportional hazard assumption. Assumed that, $\beta_j(t) = \beta_j + \gamma_j g_j(t)$, $g_j(t)$ is a specified function of time. Grambsch and Therneau show that the scaled Schoenfeld residuals have, for the jth covariate, a mean at time t of approximately

$$E\left(r_j^*(t)\right) \approx \gamma_j g_j(t) \quad (3.36)$$

this suggests that, a plot of the scaled Schoenfeld residuals over time may be used to visually assess whether the coefficient γ_j is equal to zero and, if not, what nature of $g_j(t)$ may be.

Both the graphic scaled Schoenfeld test and the statistical Schoenfel test are employed. In the statistical test, results of a chi-squared test of γ =0 for each covariate and an overall chi-squared test are shown. The plot method gives a plot for each covariate of the scaled Schoenfeld residuals against $g(t)$, with a spine smooth and point-wise confidence bands for the smooth provided.

By far, except the incorporation of time varying covariate, all the other key characteristics and relevant tests of traditional Cox's PH method have been covered. Nonetheless, as a important extension of the traditional Cox's PH model, the Cox's TVC model could never been neglected in relevant research. The subsequent section is therefore focused on the Cox's TVC model and its applicability in bankruptcy prediction.

3.5 The Cox's Time Varying Covariate ("Cox's TVC") Model

In general, covariates are divided into two categories: fixed and time-varying. A covariate is time-varying *if the difference between covariate values from two different subjects may be changing with time* (Hosmer and Lemeshow 1999).

Let T be the time to bankruptcy of interest, and let X be a set of possibly time-dependent covariates. X(t) is used to denote the value of X at time t, and $\overline{X(t)} = \{X(s): 0 \leq s \leq t\}$ to denote the history of the covariates up to time t. it is convenient to formulate the effects of covariates on the time to bankruptcy through the hazard function. The conditional-hazard function of T given $\overline{X(t)}$ is

$$h\left(t \middle| \overline{X}\right) = \Pr\left(T \in [t, t+dt) \middle| T \geq t, \overline{X}(t)\right) \tag{3.37}$$

where (t, t + *dt*) is a small interval from t to t + *dt*. The Cox TVC model specifies that

$$h(t|x(t)) = h_0(t)\exp(\beta x(t)) = h_0(t)\exp\left[\sum_{k=1}^{p}\beta_k x_k(t)\right] \qquad (3.38)$$

Where β 's are a set of unknown regression parameters and $h_0(t)$ is an unspecified baseline hazard function.

The idea of time-varying covariates is introduced by Cox (1972). When the covariates change with time, the covariates X(t) can be either stochastic or deterministic, and the ratio of the hazards of two firms with distinct values of X(t) is no longer a constant and hence the hazard rates of the two firms are no longer proportional.

Although the Cox's TVC model can use richer data and hence is more powerful in estimation, its dependency on the time-dependent covariates also results in the lost of its ability in prediction. Therneau and Foundation (1996) point out that when there are time dependent covariates, the predicted survival curve can present something of a dilemma. Focused on public healthy study, Fisher and Lin (1999) give two reasons for this dilemma: first, because the model depends on the value of a changing quantity (the time-dependent covariate), at a future time the future values are usually unknown until they are actually observed. Second, if the future values of some covariates (e.g., blood pressure) have already been known, the existence of the values implies that the subject has not reached a death endpoint.

However, it is still possible to obtain the projected survival probability for particular pattern of changes in the time dependent covariates. Therneau and Foundation (1996) demonstrate an application of using time-varying covariates for survival function prediction. The limitation of this approach is that, each time it can only predict one company's survival function for all of its survived years, so when data of large numbers of companies are given, this method becomes really time-consuming.

Based on all above considerations, in this study the Cox's TVC model is just used for estimation and no prediction application is conducted.

CHAPTER 4 METHODOLOGY - STANDARD NEURAL NETWORKS AND SURVIVAL NEURAL NETWORKS

In contrast to the linear approaches demonstrated in Chapter 3, Chapter 4 illustrates the non-linear approaches, such as neural network as well as its survival analysis applications for bankruptcy prediction. For each ANN described, the structure and likelihood function are introduced, advantages and limitations are briefly discussed.

4.1 Standard Artificial Neural Networks ("ANNs")

The most frequently used networks in bankruptcy prediction are the feed-foreword neural networks, and the most popular algorithm in training the feed-forward neural networks is the back-propagation learning algorithm. The back-propagation learning algorithm can be divided into two phases: forward-propagation and backward-propagation. Suppose there are s records of firms, each described by an input vector of accounting ratios $X_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ and an output vector of status of firms $D_i = (d_{i1}, d_{i2}, \ldots, d_{in}), 1 \le i \le s$. In forward-propagation, X_i is fed into the input layer, and an output $Y_i = (y_{i1}, y_{i2}, \ldots, y_{in})$ is generated on the basis of the current W. The value of Y_i is then compared with the actual (or desired) status D_i by calculating the squared error $(y_{ij} - d_{ij})^2, 1 \le i \le n$, at each output unit. Output difference are summed up to generate an error function E defined as

$$E = \sum_{i=1}^{s} \sum_{j=1}^{n} \frac{(y_{ij} - d_{if})^2}{2} \tag{4.1}$$

The objective is to minimize E by changing W so that all input vectors are correctly mapped to their corresponding output vectors. Thus the learning process can be cast as a minimization problem with objective function E defined in the space of W.

The second phase performs a gradient descent in the weight space to locate the optimal solution. The direction and magnitude change according to $\text{A}w_{ij} = -\frac{\partial E}{\partial w_{ij}}\varepsilon$, where $0<\varepsilon<1$ is a parameter (the learning rate) controlling the convergence rate of the algorithm, which in fact controls how efficiently the network sets its weights (Dhar and Stein 1997). If the network adjusts the weights too gradually(slightly), training the networks can take a long time, on the other hand, if the network adjusts the weights too fast, maybe it will miss the optimal weight sets.

A problem of over-fitting usually occurred in training the neural networks. Over-fitting means the network fits the training sample too well to fit the validation data, and thus the optimal set of internal parameters of the networks got from the training data may not be able to provide good prediction performance in validation data (Bishop 1995). To avoid over-fitting, first the training data set must be representative of the problem being tackled; second, the number of hidden nodes must be properly determined. When the number of hidden nodes is too few, the network may not be flexible enough to capture the actual features of the data, but if the number of hidden nodes is too many, the network becomes so flexible that it even starts to fit the noise in the data, which results in over-fitting.

The over-fitting also can be controlled by weight decay setting. As is known, the network structures with different degrees of complexity can be obtained by: (1) choice of the number of hidden nodes; (2) introduction of penalty term weight decay in the loss function (Callan, 1999). Weight decay penalizes large weight values by modifying the error function as:

$$E^* = E + \theta \sum \omega^2 \tag{4.2}$$

θ is exactly the weight decay setting. It is suggested that (Ripley 1994), θ shall range from 0.01 to 0.1 depending on the degree of fit expected. The use of penalty has the advantage of improving the convergence of optimization algorithms, in addition to avoid over-fitting.

When weight decay is used, it is a common practice to rescale covariates by multiplying them with appropriate factors so as to approximately span from 0 to 1, so as to be comparable with hidden unit outputs.

Although seem complicated apparently, ANNs have certain attributes that make them attractive as modeling techniques. As shown by Etheridge et al. (2000), ANNs do not require the data to be some form of distribution, such as multivariate normal distribution. ANNs also can accommodate numerous variables, without the detraction of multi-colinearity. Further more, because they are nonlinear procedures, ANNs are more versatile and robust than linear statistical techniques (Liang et al. 1992, Etheridge and Sriram 1997).

4.2 Survival ANNs

In this study, four survival ANNs, namely, partial logistic ANNs, the Cox's proportional hazard ANNs and two parametric survival ANNs (Weibull and log-logtistic) are studied and implemented with using a same set of US companies' data in bankruptcy prediction.

4.2.1 Partial Logistic Artificial Neural Networks ("PLANN")

The PLAAN is designed by Biganzoli et al.(1998, 2002). It is a flexible ANN approach in a discrete survival time context, which can provide smoothed hazard function estimation and allow for non-linear covariate effects.

The structure of this network model can be described as: the input layer is composed of J nodes plus one bias node, within the J nodes one is for time to bankruptcy while the others are for the covariates. The input nodes are fully connected with the H nodes of the hidden layer. A single output neuron estimates conditional failure probability values from the connections with the hidden and the bias units.

Since this network also uses the partial likelihood maximum method and use the logistic translation in the output part of each node, it is named as partial logistic ANN.

The total error function for PLANN is

$$E = -\sum_{i=1}^{n}\sum_{l=1}^{l_i}\{d_{il}\log h_l(x_i, a_l) + (1 - d_{il})\log[1 - h_l(x_i, a_l)]\} \qquad (4.3)$$

where x_i is the input node which represents one explanatory variable. (Remember that, one of the input nodes is the time to bankruptcy). The target variable is represented by the status dummy d_{il} .

The linkage between neural networks and the Cox's PH analysis in PLAAN is presented by the derivation of h_l in (4.3). The discrete hazard rate h_l in this error function actually can be derived from the Cox's PH model. Cox (1972) proposes the proportional odds model for grouped survival times as follows:

$$\frac{h_l(x_i)}{1-h_l(x_i)} = \frac{h_l(0)}{1-h_l(0)}\exp\left(\beta^T x_i\right) \tag{4.4}$$

where x_i is the covariate vector for subject i, β is the vector of regression coefficients and $h_l(0)$ is the baseline hazard for firms with x=0. Defining $\theta_l = \log\left[\frac{h_l(0)}{1-h_l(0)}\right]$, expression (4.4) is written as

$$h_l(x_i, a_l) = \frac{\exp\left(\theta_l + \beta^T x_i\right)}{1+\exp\left(\theta_l + \beta^T x_i\right)} \tag{4.5}$$

As is known, the logistic activation function employed in a standard neural network is

$$\phi_h(u) = \frac{\exp(u)}{1+\exp(u)}, \tag{4.6}$$

where u is the linear combination of input values and their parameters from the previous layer.

It is easy to find that (4.5) exactly bears same appearance of the logistic activation function (4.6). Therefore, by using the traditional logistic activation function the error function (4.3) in PLAAN is minimized.

Besides, also notice that in traditional neural networks, for a binary classification problem, a likelihood function named "cross-entropy error" is usually employed:

$$E = -\sum\sum\left\{y_{ik}^0 \log \hat{y}_k(x_i, w) + \left(1-y_{ik}^0\right)\log\left[1-\hat{y}_k(x_i, w)\right]\right\} \tag{4.7}$$

where y_{ik}^0 is the binary target variables, $\hat{y}_k(x_i, w)$ is the predicted value.

It is easy to find that when substituting the binary target variables y_{ik}^0 in (4.7) with the status dummy d_{il} in (4.3), and let the predicted value $\hat{y}_k(x_i, w)$ is the output of the logistic activation function in output lay of the networks which actually is $h_l(x_i, a_l)$, Equation (4.7) actually is an identical expression of the error function (4.3).

As a result, the particular error function of PLAAN actually can be coding with the standard syntax for traditional neural networks by choosing the options "entropy" in the likelihood function.

In the error function (4.3) for PLAAN, the total error is summed both over the n firms and over time intervals $l = 1,2,...,l_i$ in which the firm i is observed. By using the error function (4.3) in a neural network model with no hidden nodes and the logistic activation function $\phi_h(u)$, a linear logistic regression model equivalent to (4.4) is obtained. Biganzoli et al. (1998) propose the generalization of the partial logistic regression model to a feed forward ANN by the addition of a hidden layer of nodes. The output of PLAAN model can be used to explore the shape of the hazard function depending on time and covariates.

The PLANN model is not constrained to proportionality assumptions since interaction of time and covariate effects will be modeled implicitly. Therefore an advantage of PLAAN is the possibility of straightforward use of time-dependent covariates.

4.2.2 The Cox's PH Survival ANNs

Besides of the above approach, another type of survival ANNs which apply the Cox's PH function is the survival ANNs designed by Faraggi and Simon (1995).

The Cox's PH survival ANN is a combination of the standard neural network and the Cox's PH regression. Assumed there are N firms and P input variables are used for each firm's survival prediction. Consider the *ith* firm, where $x_i = (x_{i0} = 1, x_{i1}, ...x_{ip})$ is a $1 \times (P+1)$ input vector (i=1…N). $\omega_h = (\omega_{0h}, ..., \omega_{Ph}) \times 1$ is a vector of parameters connecting the inputs to the hidden node h (h=1,.., H). The input to the h node in the hidden layer for the *ith* firm is the weighted sum

$$x_i\omega_h = \sum_{p=0}^{P} x_p \omega_{ph} = \omega_{0h} + \sum_{p=1}^{P} x_p \omega_{ph} \tag{4.8}$$

The output from any node in the hidden layer is its input transformed by an activation function, such as the traditional logistic function $f(z) = \{1 + \exp(-z)\}^{-1}$. Hence the output of the i th firm from the h node in the hidden layer is $f(x_i\omega_h)$. Finally the output of the network is a linear combination of the outputs from the nodes in the hidden layer with parameters $\alpha = (\alpha_0, \alpha_1, ..., \alpha_H)$. Particularly it is just a linear combination of outputs from all hidden nodes with their parameters, and there is no activation function in the output node. Consequently the output from a single hidden layer neural network with H hidden nodes for a given firm with input vector $x_i = (1, x_{i1}, ..., x_{iP})$ is

$$g(x_i, \theta) = \alpha_0 + \sum_{h=1}^{H} \alpha_h f(x_i\omega_h) = \alpha_0 + \sum_{h=1}^{H} \frac{\alpha_h}{1 + \exp(-x_i\omega_h)} \tag{4.9}$$

As is shown in Chapter 3, the hazard function in the Cox PH model is

$$h(t, x_i) = h_0(t)\exp(x_i\beta)$$

The $h_0(t)$ is the baseline hazard and $\beta = (\beta_0 = 0, \beta_1, ..., \beta_P)$ are estimated by maximizing the log of the partial likelihood:

$$\sum_i \left\{ x_{(i)}\beta - \log \sum_{j \in R_i} \exp(x_j b) \right\}$$

Faraggi and Simon (1995) design a neural network with the Cox's partial likelihood function by replacing the linear combination of the covariates $x_i\beta$ in the hazard function with the output of the neural network $g(x_i, \theta)$ in equation (4.9):

$$h(t, x_i) = h_0(t)\exp(g(x_i, \theta)) \tag{4.10}$$

and the likelihood function becomes:

$$y = \log \lambda(x) = \sum_i \left\{ g(x_{(i)}, \theta) - \log \sum_{j \in R_i} \exp[g(x_j, \theta)] \right\} \tag{4.11}$$

Maximum partial-likelihood estimates of weights can be obtained by using the Newton-Ralphson method which is suggested by Faraggi and Simon (1995), or quasi-Newton algorithm which is used by Mariani et al. (1997).

To find a suitable local maximum, multiple starting values should be used to initiate the optimization algorithm. Faraggi and Simon (1995) also adopt a penalized likelihood function approach based on the following formula:

$$L_c(w) + \eta \sum w^2 \tag{4.12}$$

where $L_c(w)$ is the likelihood function and η is a shrinkage parameter.

This approach has the advantage of controlling the effective model complexity, so as to avoid over-fitting on training data, and therefore improving the generalization ability of the final model.

The Cox's PH survival ANN features the following merits: first, it can be used for large scale data set; second, this method allows preserving all the advantages of the classical proportional hazards model; third, the standard approach still assumes that the hazards are proportional; forth, this approach also allows for time-varying inputs, but it is less flexible in modeling the baseline variation.

Faraggi and Simon's method shows a way which can effectively convert a traditional survival linear analysis into a non-linear survival model by replacing the linear function of the covariates with the output of a standard non-linear model, such as a standard neural network. This method has been further developed by Mariani et al. (1997) and Ripley (1998). This approach requires non-standard software since the function to be minimized (the log partial likelihood) is not a sum over all firms: for each bankrupt firm it requires a sum over all the other firms still in the risk set at the time of that firm's bankruptcy. In Mariani et al. (1997), they use same software designed by Faraggi and Simon with one small adjustment: they use weight decay at all stages while in Faraggi and Simon (1995) the weight decay is removed from the final stages of optimization. In Ripley (1998), she re-designs the framework but leaves theoretical designs remained the same as Faraggi and Simon (1995).

4.2.3 The Parametric Survival ANNs

The parametric survival ANNs implemented here are proposed by Ripley (1998).

As is shown in Chapter 3, the density function and survival function of the Weibull distribution are

$$h(t|x) = \gamma t^{\gamma-1}\lambda$$

$$S(t|x) = \exp(-t^{\gamma}\lambda)$$

For the log-logistic distribution, the density function and survival function are

$$h(t) = \frac{\lambda\gamma(\lambda t)^{\gamma-1}}{\left[1+(\lambda t)^{\gamma}\right]^2}$$

$$S(t) = \frac{1}{1+(\lambda t)^{\gamma}}$$

where γ is the shape parameter.

In both cases, $\log\lambda$ is modeled as a function of x by a standard neural network with a single linear output:

$$y = \log\lambda(x) = \sum_j \omega_{jo} x_j + \sum_h \omega_{ho} l\left(\sum_j \omega_{jh} x_j\right) \qquad (4.13)$$

Similar to the relationship between the Cox's PH survival ANN and the Cox's PH regression, the parameter survival ANN models substitute the linear combination of covariates by the output of a standard neural network in their likelihood function. Specifically, in parametric regression methods, λ equals to $\exp(x\beta)$, while in parametric survival ANNs λ equals to $\exp(y)$, where y is the single output of a neural network, as demonstrated in (4.13).

As shown in Chapter 3, the likelihood function in parametric survival regressions can be simply demonstrated as:

$$\prod_{\delta_i=1} f(t_i)\prod_{\delta_i=0} S(t_i)$$

Where δ_i is the indicator variable for observed bankruptcy. Thus the first product is over the defaulted firms, the second over the non-defaulted ones.

Substituting the $f(t_i)$ and $S(t_i)$ for the Weibull distribution into above likelihood function, with $\lambda = \exp(y)$, after taking log, the log likelihood for parametric survival ANN in Weibull distribution is:

$$\sum_{firms} \delta_i[\log\gamma + (\gamma - 1)\log t_i + y_i] - (1 - \delta_i)(t_i^{\gamma} \exp y_i) \tag{4.14}$$

Similarly, the log-likelihood function for parametric survival ANN in log-logistic distribution is:

$$\sum_{firms} \delta_i \{y_i + \log\gamma + (\gamma - 1)(y_i + \log t_i) - \log[1 + (e^{y_i} t_i)^{\gamma}]\} - \log[1 + (e^{y_i} t_i)^{\gamma}] \tag{4.15}$$

The solutions can be achieved by minimizing the log likelihood over the weights and the shape parameter γ.

CHAPTER 5 DATA PREPARATION

The theoretical descriptions on survival regressions and survival ANNs have been accomplished by Chapter 3 and 4. The following chapters are about to apply these survival regressions and survival ANNs based on a same set of US company data, empirically study the prediction powers of each model, and comprehensively compare their performance among each other.

Chapter 5 is focused on data preparation. First, the data and variable selections are described, and then the sampling methods are discussed. Further, data reduction procedure is presented and finally a preliminary finding based on data is shown.

5.1 Data and Variables

For this study, data are downloaded from Company Analysis™[1] ranged from 1985 to 2002. Using the Dow Jones Global Classification Standard, the research is restricted to the following four industries: cyclical consumer goods, non-cyclical consumer goods, cyclical consumer service and non-cyclical consumer services. The following industry types are excluded: basic, energy, financial, healthy care, technology, telecommunications and utility.

Cyclical consumer goods and services refer to the industries whose sales and operation are changing cyclically, such as automobiles, media and retails. Non-cyclical consumer goods and services refer to the industries whose sales and operation are relatively constant over time, such as foods and beverage production.

[1] Company Analysis™, provided by Thomson Financial, is a Windows based financial analysis system. It provides financial statement data for 18,000 companies from over 70 countries.

Initially 37 variables were collected including accounting ratios, stock market information and cash flow information which were all employed by previous bankruptcy analysis studies (Altman 1968, Sung, Chang and Lee 1999; Marais et al. 1984, Bell and Tabor 1991, Koh 1992, etc.). These variables can help in analyzing the four facets of a company's performance and condition (Dominiak and Louderback 1985), specifically, internal liquidity, operating performance, financial risk, growth potential.

In addition to above variables, another two dummy variables need to be included: "GS" (which indicates whether the firm is a goods provider or a service provider) and "Cyclical" (which indicates whether the goods/service belongs to cyclical item or non-cyclical item). Specifically, the "GS" of 0 is for the service providers and 1 for the goods providers; the "Cyclical" is "0" for the non-cyclical providers and 1 for cyclical providers. Consequently, the whole dataset can be divided into four groups: goods cyclical, goods non-cyclical, service cyclical and service non-cyclical.

There are two dependent variables for all survival regressions: the "status" and the "survival time". Specifically, the status variable is a dummy variable whose value is 1 for the bankrupt firms and 0 for the non-bankrupt firms. The survival time is the number of years from the firm's inception to its filling for bankruptcy. This setting of survival time can be found in several previous survival studies on company bankruptcies. For instance, Shumway (2001) sets his dependent variable about survival time as the number of calendar years a firm has been traded on the NYSE or AMEX. In Audrestch and Mahmood (1995), they set the time to hazard variable (survival time) as the life span of a company, from setup to the closure.

Only one dependent variable, the status, is required in survival ANNs.

5.2 Building Samples

In this study, all non-bankrupt firms are selected to match the bankrupt firms on the basis of industry. Company size is used as one covariate in our models and hence it is not used as a matching criterion. When the firm got bankrupt, its counterparty in the non-bankrupt group is denoted “censored” at the same year.

Nine models are exploited in this study. Except for the Cox’s TVC model which requires multiple-year records for each firm, all other eight models need only one-year records which are values of variables at 31 December of the year prior to the bankruptcy or the censoring year for all firms. The former data set is referred as TVC dataset since it is used exclusively for the Cox’s TVC model and the latter data set is referred as one- year dataset.

The TVC dataset is constructed in the same way as Philip Cheng (2002): The observation period (1985-2002) is divided into 17 at-risk intervals. Each at-risk interval is set as the time range from the end of the previous year to the end of current year. A record is created for each at-risk interval for each firm, that is, yearly covariates are obtained for each firm. Specifically, if a firm entered into the study in December 1994 and became bankrupt in December 2000, there were six records for this firm with each record pertaining to its yearly accounting ratios from 1994 to 2000. Within these six yearly records, only the record for year 2000 is marked as 1 in the status variable, which indicates that the firm became bankrupt in the at-risk interval December 1999 to December 2000. For the other five years’ records, the status variable is marked as 0 to indicate that the firm is a non-bankrupt one. No more records will be created for this firm after the at-risk interval in which it filed bankrupt.

If a firm that survives throughout the entire observation period, its status will equal to 0 for all the at-risk intervals.

The survival time variable gets a same setting in both the one-year-dataset and the TVC dataset. More specifically, for the bankrupt companies, the survival time is measured from the initiation of the firm to the last year before its bankruptcy; for the non-bankrupt firms, the survival time is the entire period it stayed in the study. Particularly, for the non-bankrupt firms, the survival time is referred as censored time. In both datasets, for instance, if a firm listed in 1994 and filed for bankruptcy in December 2000, its survival time equals to 6 years. If a firm listed in 2000 and remained non-bankrupt by the end of the observation period in the study (which is 2002), its censored time equals to 2.

For the TVC model, as the survival time is measured from the beginning of the observation period, the horizon of prediction is the length of the observation period. Further, there is a survival probability estimated for each firm at each of the at-risk intervals throughout the observation period.

To compare models based on their prediction performance, a training sample and a validation sample are constructed for each model. The training sample is used to train the model and estimate all the parameters while the validation sample is used to generate predictions with parameters estimated in the training phase. Usually more data should be assigned for model training compared with data assigned for prediction, because the correctness of prediction is mainly depended on the accuracy of model training. If too few data are allocated as training sample, it will result in large variances in model outputs. Besides, if the training sample can not cover all features of dataset, the model outputs will also become biased. However, leaving enough data for prediction is also important. If too few data are allocated to prediction sample, the assessments based on the model output may not be reliable and meaningful any more. After taking balance of the training sample and prediction sample, randomly 67% of total records are selected for the training sample and the remaining is for the validation sample.

5.3 Data Reduction

Currently 37 variables are all loaded into the datasets. However most of them bear similar information and therefore is redundant. To apply the dataset into the model, a variable reduction process is required. Besides, the datasets need to be further refined to remove missing values and outliers before been imported into the model engine.

Subsequently, Section 5.3.1 is to reduce the variable number while minimizing the information loss in this process. Section 5.3.2 is to conduct the data cleansing which removes the missing values and outliers.

5.3.1 Variable Reduction

Factor analysis is implemented to identify the variables eventually imported into the models. For the one-year dataset, the factor analysis results show that 37 variables can be allocated into 11 factors with averagely three or four variables within one factor. Traditionally, the variable that has the highest score within each factor in the rotated component matrix can convey the richest information among the variables within the same factor. Therefore, by selecting the variable with the highest score in each factor, the number of variables can be reduced from 37 to 11 without losing too much information. All these selected 11 variables are to be employed by the models. The rotated component matrix in one-year dataset are shown in table 5.1.

Table 5.1 Rotated Component Matrix in Factor Analysis for One-Year Dataset

Rotated Component Matrix											
	Factor										
	1	2	3	4	5	6	7	8	9	10	11
CA/TA	**0.93**	0.02	-0.02	-0.20	0.01	0.05	0.16	0.06	-0.03	0.10	0.00
Sales/FA	0.92	0.01	-0.01	-0.12	-0.01	0.07	0.19	0.06	-0.03	0.05	0.05
Sales/TA	0.74	0.17	-0.09	-0.01	0.00	0.09	0.07	-0.12	0.05	-0.02	-0.34
Fixed Asset/TA	-0.28	0.02	-0.03	-0.11	-0.01	0.01	0.04	-0.08	0.03	0.08	0.02
LOG(TA)	0.05	**1.00**	-0.01	-0.01	0.00	0.01	0.04	0.00	-0.01	0.01	0.00
LOG(SALES)	-0.01	-1.00	0.01	0.01	0.00	0.00	-0.03	0.00	0.01	0.00	0.00
Share price at period end	0.08	0.99	-0.01	0.00	0.00	0.02	0.05	-0.01	0.00	0.00	0.00
Debt/TA	0.05	0.03	**0.84**	0.02	0.03	-0.02	0.05	-0.18	-0.04	-0.08	-0.04
TL/TA	-0.16	-0.06	0.77	-0.26	0.06	-0.03	-0.23	0.25	-0.04	0.04	0.04
Long loan/TA	0.18	0.06	0.76	0.13	0.02	0.02	0.14	-0.30	-0.01	-0.09	-0.04
Cash/CL	-0.10	-0.01	-0.17	**0.90**	-0.06	0.01	-0.05	-0.02	-0.01	0.02	-0.02
CA/CL	-0.23	0.01	0.04	0.87	-0.06	0.02	0.10	-0.05	-0.01	-0.02	0.09
CA/short Loan	-0.04	-0.01	-0.11	0.81	0.01	0.00	-0.04	0.05	-0.02	0.02	-0.02
CA/sales	0.00	0.00	0.05	-0.03	**0.95**	0.18	0.01	-0.03	0.00	0.01	0.00
Inventory /Sales	0.03	0.00	0.08	0.01	0.87	-0.14	0.00	0.04	0.01	0.00	0.00
Sales growth	-0.02	0.00	-0.03	-0.07	0.78	0.18	-0.01	0.04	-0.01	0.02	0.00
CFO/long debt	0.11	0.02	0.05	0.00	0.08	**0.88**	-0.02	0.04	0.01	-0.01	-0.02
CFO/total debt	0.09	0.01	-0.01	0.04	-0.10	0.82	-0.03	0.09	0.01	-0.02	-0.01
Cashflow/total debt	0.07	0.01	0.06	-0.03	-0.02	-0.64	-0.10	0.33	0.01	-0.03	0.06
Interest cover	0.04	0.00	-0.01	-0.03	0.17	0.46	0.03	-0.04	-0.01	0.02	0.06
PBIT/sales	0.29	0.02	-0.22	0.08	0.01	0.05	**0.82**	-0.03	0.04	-0.12	0.01
Earning/TA	0.38	0.17	0.09	0.11	0.01	0.09	0.80	-0.18	0.04	-0.14	0.01
Trading profit/FA	0.02	-0.01	-0.06	-0.18	-0.02	-0.04	0.61	0.22	-0.04	0.26	0.06
Net debt increase	0.20	0.00	0.07	0.03	0.06	-0.12	-0.07	**0.78**	0.04	0.00	0.18
Total net flow/TA	-0.09	0.00	-0.02	-0.02	-0.01	0.03	0.24	0.57	-0.03	-0.11	-0.40
Earnings/Equity	0.08	-0.01	-0.06	-0.03	0.01	0.00	0.03	0.04	**0.83**	-0.01	0.06
Net debt/Equity	0.14	0.00	0.02	0.00	0.01	0.01	0.01	0.02	-0.83	-0.02	0.07
PBIT/Capital	0.06	0.00	0.01	0.00	0.00	0.00	0.00	-0.07	0.01	**0.73**	0.04
Net income/Total net flow	0.04	0.01	0.02	0.04	-0.01	0.00	-0.01	-0.05	0.01	0.03	0.12
Reported eps (gross/net)	0.32	-0.02	0.04	-0.30	-0.02	0.02	0.23	-0.04	-0.01	0.50	0.07
Earnings per share growth	-0.16	0.01	0.01	0.02	-0.01	0.02	0.09	0.06	-0.01	-0.10	**0.84**
Increase in reserves	0.03	0.01	0.07	0.11	-0.01	0.01	-0.13	0.04	0.01	0.02	0.18
Working capital movements	-0.04	0.02	0.01	0.10	0.04	0.01	-0.08	0.04	0.01	0.61	-0.14

- Extraction Method: Principal Component Analysis.
- Rotation Method: Varimax with Kaiser Normalization.
- A Rotation converged in 9 iterations.

For the TVC dataset, 12 factors are identified. To facilitate the model comparison between TVC model and others, the variable of the last factor loaded in the rotated component matrix is excluded from our analysis. This action won't loss too much information since the last factor loaded in the rotated component matrix is the one that has the least information richness and usually bears very limited explanation power. As a result, 11 variables are selected which eventually enter the Cox's TVC model. The rotated component matrix in TVC dataset is shown in table 5.2.

Table 5.2 Rotated Component Matrix in Factor Analysis for TVC dataset

Rotated Component Matrix												
	Factor											
	1	2	3	4	5	6	7	8	9	10	11	12
Earning/TA	**0.97**	0.00	0.00	-0.11	0.07	0.00	0.00	0.04	-0.05	-0.01	0.00	0.00
NI/FA	0.95	0.02	0.02	-0.05	0.05	0.00	0.03	0.07	-0.04	0.00	-0.01	0.00
PBIT/Total Asset	0.90	0.01	0.01	-0.17	0.31	0.00	0.03	0.09	-0.04	-0.01	0.01	0.00
Net Income/TA	0.84	0.00	0.00	-0.07	-0.43	0.00	-0.02	0.04	-0.03	-0.01	-0.02	-0.01
PBIT/Capital employed	0.34	0.00	0.00	0.05	0.15	0.00	0.02	-0.06	0.04	0.10	0.02	0.01
CFO/total debt	0.00	**0.95**	0.00	-0.01	0.00	0.16	0.01	0.00	0.00	0.00	-0.01	0.00
Interest cover	0.02	0.92	0.00	0.00	0.00	0.12	0.01	0.03	0.00	0.02	-0.01	0.00
CFO/long debt	0.01	0.78	0.00	-0.01	0.00	0.03	0.00	0.03	0.00	0.01	-0.02	0.01
PBIT/sales	0.02	0.00	**1.00**	0.00	0.00	0.00	0.01	0.04	0.00	0.00	-0.02	0.00
Trading profit/Sales	0.01	0.00	0.99	0.00	0.00	0.00	0.01	0.04	0.00	0.00	-0.02	0.00
CA/sales	0.00	0.00	-0.98	-0.01	0.00	0.00	-0.02	-0.03	0.00	0.00	0.02	0.00
Debt/TA	-0.02	-0.01	0.00	**0.90**	-0.01	-0.03	0.20	-0.01	0.01	-0.13	0.00	0.01
TL/TA	-0.15	-0.01	0.00	0.88	0.02	-0.03	-0.01	0.07	-0.01	-0.04	-0.04	0.02
Long loan/TA	0.03	-0.02	0.00	0.78	-0.02	0.01	0.21	0.02	0.03	-0.12	0.02	0.01
Reserve/TA	0.10	0.00	0.00	-0.71	-0.01	0.02	0.09	0.06	0.02	-0.06	0.06	0.00
Sales/TA	0.10	0.00	0.00	0.00	**0.99**	0.00	-0.01	0.00	-0.04	0.00	0.00	0.00
Sales/Fixed Asset	0.13	0.00	0.00	0.00	0.98	0.00	-0.06	0.00	-0.04	-0.01	0.00	0.00
CA/CL	0.00	0.16	0.00	-0.02	0.00	**0.88**	-0.06	-0.02	0.00	0.00	0.00	0.00
CA/short loan	0.00	0.06	0.00	-0.02	0.00	0.88	-0.03	0.01	0.00	0.00	-0.02	0.01
QA/short loan	-0.01	0.31	0.00	-0.03	0.00	0.84	0.00	-0.03	0.00	-0.01	0.05	0.00
FA/TA	0.02	0.02	0.02	0.12	-0.02	-0.06	**0.90**	0.16	0.01	-0.04	-0.08	0.03
CA/TA	-0.02	-0.01	-0.02	-0.14	0.02	0.08	-0.87	-0.24	-0.01	0.01	0.08	-0.05
IntangibleA/TA	0.01	-0.02	0.01	0.02	-0.02	0.03	0.69	-0.09	-0.01	0.02	0.05	-0.10
LOG(TA)	0.06	0.02	0.10	0.00	0.04	-0.03	0.07	**0.91**	0.00	0.18	0.01	-0.05
LOG(sales)	0.05	0.00	0.01	0.01	-0.05	-0.01	0.19	0.91	0.00	0.16	0.07	-0.07
Net debt/Equity	0.04	0.00	0.00	0.02	0.05	0.00	0.01	0.00	**0.98**	-0.01	0.00	0.00
Earning/Equity	0.27	0.00	0.00	0.01	0.56	0.00	0.00	0.00	-0.73	0.01	0.00	0.00
Reported eps (gross/net)	0.07	0.02	0.01	-0.15	0.00	-0.03	-0.03	0.06	-0.01	**0.77**	-0.05	-0.06
Share price at period end	0.02	-0.01	0.00	-0.03	0.00	0.02	0.00	0.25	-0.01	0.68	0.08	0.00
Total net flow/TA	0.02	0.06	-0.06	-0.09	0.00	0.03	-0.09	0.01	0.00	0.01	**0.75**	-0.04
Sales growth	0.01	-0.01	0.03	0.03	0.00	-0.01	0.12	-0.22	0.01	0.08	0.42	-0.23
Working capital movements	0.01	0.00	0.01	0.03	0.00	0.01	0.05	-0.19	0.01	0.18	0.05	**0.72**
Increase in reserves	0.01	0.00	0.00	-0.01	0.00	0.01	0.08	-0.05	0.01	0.27	0.00	-0.64
Earnings per share growth	0.01	0.00	0.00	0.02	0.00	0.00	0.05	-0.08	0.01	0.14	-0.04	0.08

- Extraction Method: Principal Component Analysis.
- Rotation Method: Varimax with Kaiser Normalization.
- A Rotation converged in 6 iterations.

Table 5.3 and 5.4 summarize the 11 variables selected in one-year dataset and TVC dataset after factor analysis.

Table 5.3 11 Variables Selected for One-Year Dataset

Log (TA)	Debt/TA	CA/TA	Cash/CL	PBIT/Sales
CFO/Long debt	CA/Sales	Net Debt Increase	Earnings/Equity	EPS Growth
PBIT/Capital				

Table 5.4 11 Variables Selected for TVC Dataset

Earning/TA	CFO/Total Debt	PBIT/Sales	Debt/TA	Sales/TA
CA/CL	FA/TA	Log (TA)	Net Debt/Equity	EPS
Total Flow/TA				

The definitions of each selected 11 variables for one-year dataset are presented in table 5.5. The relationships between the variables and the company's survival can indicate the expected signs of the variables in survival regressions. The expected signs and the relationship between the variables and company's survival are also shown in table 5.5. Relevant information for the TVC dataset is in table 5.6.

As a result, all the models studied in this thesis have 13 covariates: the 11 variables selected from factor analysis plus the two dummies ("GS" and "Cyclical").

Table 5.5 Variable Definitions, Relationship with Survival and Expected Signs for One-Year Dataset

	Variable	Variable	Relationship with Survival	Expected Sign
1	Cash/CL	Cash / Current Liability	The higher the ratio is, the better internal liquidity for that company. A favorable factor for a company's survival.	positive in the parametric models negative in the Cox's model
2	CFO/Longdebt	Cash flow from Operations / Long-term Debt	The higher the ratio is, the lower the financial risk. A favorable factor for a company's survival.	positive in the parametric models negative in the Cox's model
3	NetdebtIncre	Net Debt Increased Per Year	The higher the net debt increased, the higher the financial risk. An unfavorable factor for a company's survival.	negative in the parametric models positive in the Cox's model
4	Debt/TA	Long-term Debt / Total Asset	The higher the ratio is, the higher the financial risk. An unfavorable factor for a company's survival	negative in the parametric models positive in the Cox's model
5	PBIT/Sales	Profit before Interest and Tax / Sales	The higher the ratio is, the higher the operating profitability. A favorable factor for a company's survival	positive in the parametric models negative in the Cox's model
6	PBITCapital	Profit before Interest and Tax / Capital	The higher the ratio is, the higher the operating profitability is. A favorable factor for a company's survival.	positive in the parametric models negative in the Cox's model
7	CA/Sales	Current Asset / Sales	The higher the ratio is, the lower the operating efficiency. An unfavorable factor for a company's survival.	negative in the parametric models positive in the Cox's model
8	Earnings/Equity	Earnings / Equity	The higher the ratio is, the higher the operating efficiency is. A favorable factor for a company's survival.	positive in the parametric models negative in the Cox's model
9	CA/TA	Current Asset / Total Asset	The higher the ratio, the higher the capital liquidity is. A favorable factor for a company's survival.	positive in the parametric models negative in the Cox's model
10	Log(TA)	Log (Total Asset)	The larger this number is, the large the company is. Usually, larger companies can survive longer than smaller ones.	positive in the parametric models negative in the Cox's model
11	EPSgrowth	Growth in Earnings Per Share	The higher the number is, the faster the company grows. A favorable factor for a company's survival.	positive in the parametric models negative in the Cox's model

Table 5.6 Variable Definitions, Relationship with Survival and Expected Signs for TVC Dataset

	Variable	Variable	Relationship with Survival	Expected Sign
1	CA/CL	Current Asset / Current liability	The higher the ratio is, the better internal liquidity for that company. A favorable factor for a company's survival.	positive in the parametric models negative in the Cox's model
2	Debt/TA	Long-term Debt / Total Asset	The higher the ratio is, the higher the financial risk. An unfavorable factor for a company's survival	negative in the parametric models positive in the Cox's model
3	Netdebt/Equity	Net Debt / Equity	The higher the ratio is, the higher the financial risk. An unfavorable factor for a company's survival.	negative in the parametric models positive in the Cox's model
4	PBIT/Sales	Profit before Interest and Tax / Sales	The higher the ratio is, the higher the operating profitability. A favorable factor for a company's survival	positive in the parametric models negative in the Cox's model
5	Eps	Earnings per Share	The higher the number is, the higher the profitability shown. A favorable factor for a company's survival.	positive in the parametric models negative in the Cox's model
6	Sales/TA	Sales / Total Asset	The higher the ratio is, the higher the operating efficiency is. A favorable factor for a company's survival.	positive in the parametric models negative in the Cox's model
7	Earning/TA	Earnings / Total Asseet	The higher the ratio is, the higher the operating efficiency is. A favorable factor for a company's survival.	positive in the parametric models negative in the Cox's model
8	Log(TA)	Log (Total Asset)	The larger this number is, the large the company is. Usually, larger companies can survive longer than smaller ones.	positive in the parametric models negative in the Cox's model
9	Totalflow/TA	Total Cash Flow / TA	The higher the ratios, the better the status of company is. A favorable factor for a company's survival.	positive in the parametric models negative in the Cox's model
10	FA/TA	Fixed Asset / Total Asset	The higher the ratio, the lower the capital liquidity is. An unfavorable factor for a company's survival.	negative in the parametric models positive in the Cox's model
11	CFO/TotalDebt	Cashflow from Operation / Total Debt	The higher the ratios, the better the status of company is. A favorable factor for a company's survival.	negative in the parametric models positive in the Cox's model

5.3.2 Exclusion of Outliers and Missing Values

In the one-year dataset, if the values go beyond three times of the standard deviation of the variables, these numbers are excluded as outliers. However, if a variable gets too many values identified as outliers (i.e. CA/Sales) according to this rule, the outlier exclusion of this variable will base on subjective judgments instead.

After excluding outliers, a total of 1563 records are remained, within which 512 are bankrupt and 1051 are non-bankrupt.

When handling the missing values, if a record has one missing value in any of the13 covariates, the whole record is deleted from the data set.

After going through all these procedures above, finally the one-year dataset has 804 records with 281 bankrupts and 523 non-bankrupts.

No missing value exclusion is processed in the TVC data since the Cox's TVC model can handle the missing values automatically.

5.4 Data Rescaling

After the data reduction, both the one-year dataset and TVC dataset are ready for the survival regressions. However, these two datasets require one more step on data preparation, data rescaling which rescale all values into range of 0 to 1, to facilitate the running of ANNs.

A two-step rescaling method is used in this study.

Step 1: Standardized Normalization

$$y' = \frac{y - \bar{y}}{S_y}$$

where y is original value, $\bar{y}$ and S_y are mean and standard deviation respectively, and y' is the normalized value.

Step 2: Min-Max Normalization: $y'' = \frac{y' - y'_{min}}{y'_{max} - y'_{min}}$

where y'_{min} and y'_{max} are minimum and maximum values of the y' generated in Step 1.

After these two steps, all values in both datasets are rescaled into the range of 0 to 1.

5.5 A Preliminary Finding

Before applying any computation, a particular pattern of the data is observed in both one-year dataset and TVC dataset: the younger firms go bankrupt more frequently than the older ones. Specifically, for both one-year dataset and TVC dataset, for firms whose survival time are less than 3 years, about 95% of them default. However, for firms with survival time longer than 15 years, only 3% of them eventually end up with bankruptcy. Therefore, there is a strong negative connection between the survival time and the bankruptcy of firms.

It is straightforward to explain this pattern since a long established company probably has better internal efficiency (i.e., experienced staffs and sophisticated management) and competitive advantages (i.e., well-known branding and larger market share) compared with those newly incepted ones.

CHAPTER 6 ESTIMATION RESULTS OF SURVIVAL REGRESSIONS AND MODEL TRAINING OF NEURAL NETWORKS

This chapter is mainly to address the estimation outputs of survival regressions and neural networks designs for survival ANNs.

For survival regressions, estimation results, i.e., coefficient of each covariate, sighs of coefficients, number of significant covariates and goodness of fit for each regression, are discussed.

For survival ANNs, no estimation output can be generated. However, since estimation belongs to the scope of model training for regressions, correspondingly, the neural network's model training issues, such as number of hidden node, are discussed here.

In this chapter, only model comparisons among survival regressions are conducted.

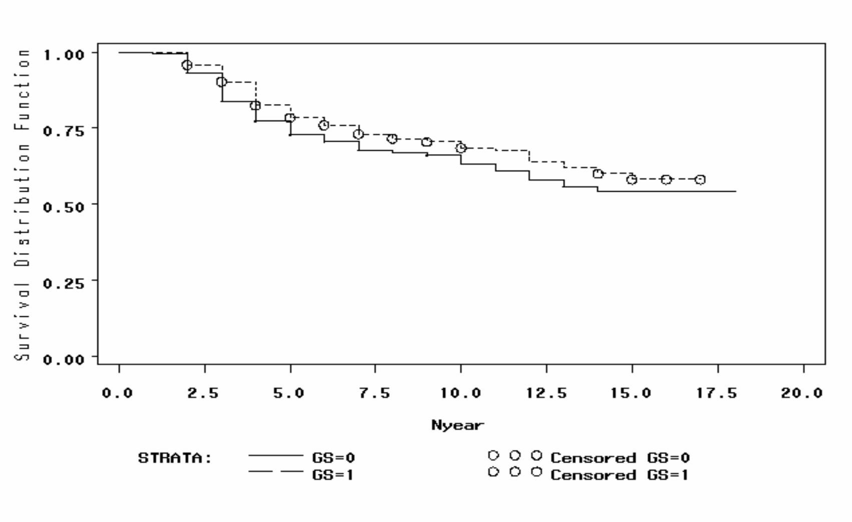
6.1 Estimation Results of Linear Survival Analysis

6.1.1 Kaplan-Meier Estimation of the Survival Curve and the Hazard Curve

The Kaplan-Meier non-parametric approach is the simplest one among all survival estimation methods. Although as simple as it is, it still can shed some lights on the companies' survival experience.

Based on one-year dataset, figure 6.1 shows the survival curves generated by Kaplan-Meier's method for consumer goods providers (GS=1) and the consumer service providers (GS=0). As shown, in the first two years, both service providers and goods providers have similar survival probability. After three years, however, consumer service providers have a lower survival curve than consumer goods providers. This reflects a higher business risk for consumer service providers compared with that of consumer goods providers.

Figure 6.1 Kaplan-Meier Survival Curve for Consumer Goods Providers and Consumer Service Providers



Similarly, figure 6.2 shows the difference on the survival curves between the cyclical goods/service providers (Cyclical=1) and the non-cyclical goods/service providers (Cyclical=0). Obviously, except for the beginning 2 – 5 years, non-cyclical goods/service providers experience a systematically lower survival probability compared with cyclical providers, in addition, this discrepancy deteriorates as time goes by.

Figure 6.2 Kaplan-Meier Survival Curves for Cyclical Consumer Goods/Service Providers and Non-cyclical Consumer Goods/Service Providers

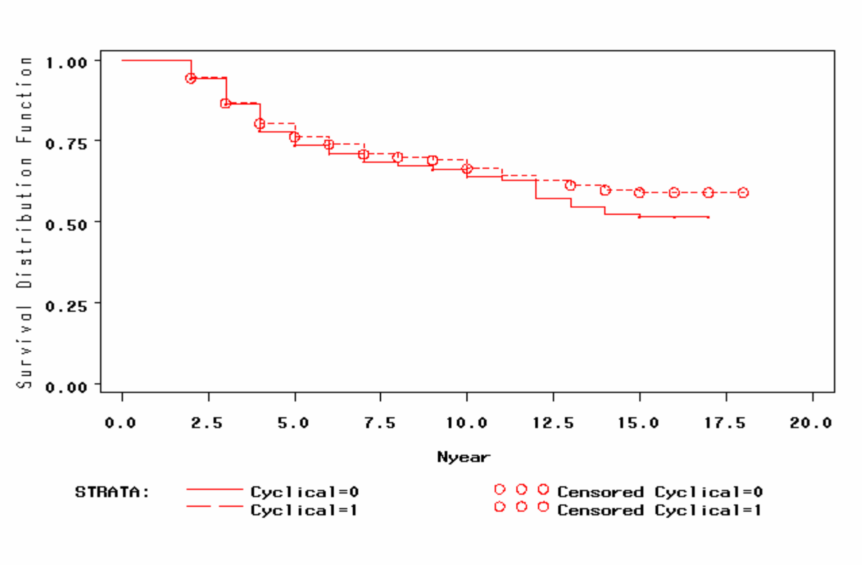
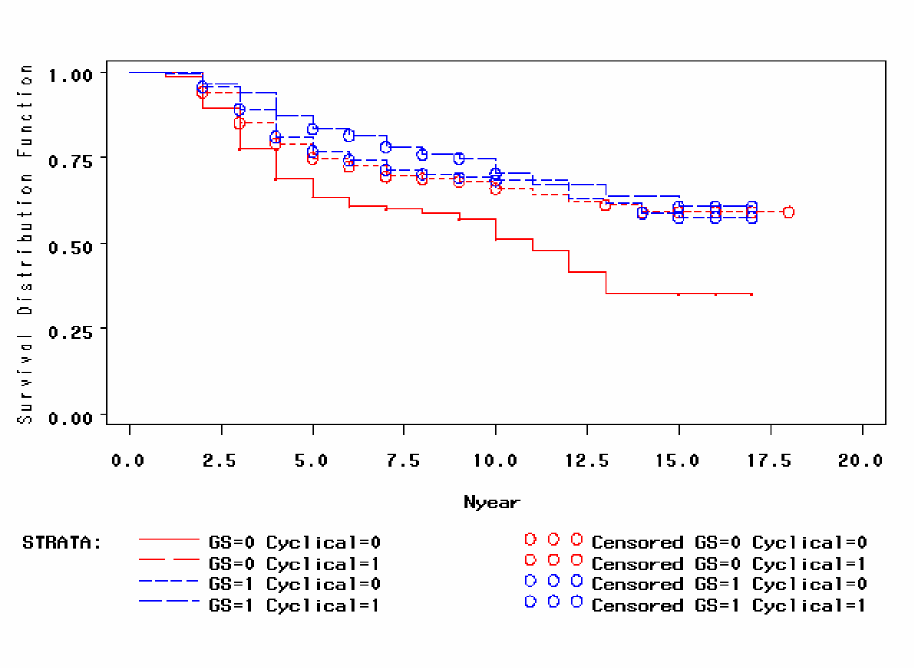


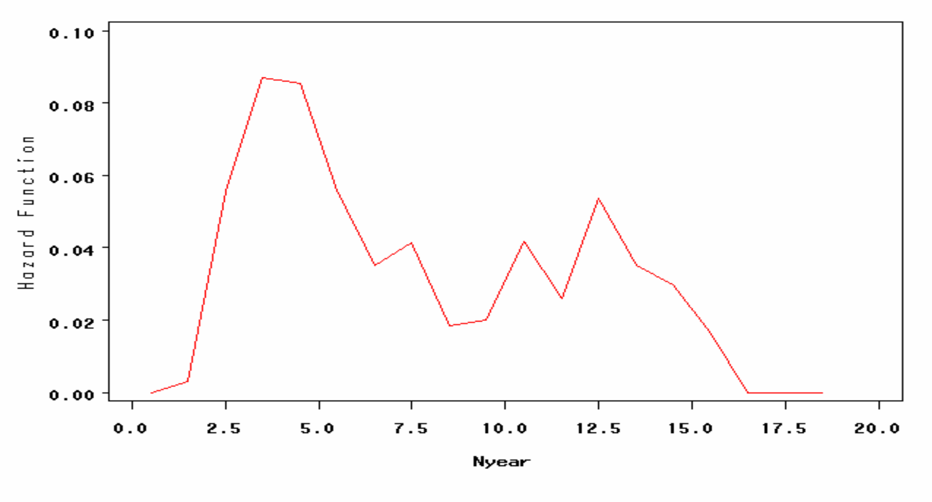
Figure 6.3 exams the survival status of four groups within one-year dataset: cyclical goods providers, cyclical service providers, non-cyclical goods providers and non-cyclical service providers. As expected, the non-cyclical service providers exhibit the worst survival status. Their survival probability decreases much faster than the other three groups except for first two years. For the other three groups, they generally follow a similar survival pattern within the research time horizon.

Figure 6.3 Kaplan-Meier Survival Curves for Cyclical Goods Providers, Cyclical Service Providers, Non-Cyclical Goods Providers and Non-Cyclical Service Providers



Besides of the survival curves, hazard curve can also be generated. Figure 6.4 shows a general hazard curve for all records in one-year dataset. This empirical hazard curve shows that hazard rate for the one-year dataset is changing with time.

Figure 6.4 Kaplan-Meier Hazard Curve



6.1.2 Estimation Results of Parametric Survival Regressions

6.1.2.1 P-Value, Sign and Estimated Effect of Covariates on Survival Time

The estimation results of parametric survival regressions using the Weibull distribution and log-logistic distribution are shown in tables 6.1 and 6.2. There are many similarities between the outputs of these two models: first, only "GS", "Cyclical" and "LOG (TA)" are significant at a 5% level; second, the estimated sign of all covariates in the two model outputs are identical.

Table 6.1 Estimation Results of Parametric Survival Regression - the Weibull Distribution

Name of Distribution	Weibull
Log Likelihood	-557.6

Analysis of Maximum Likelihood Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		Chi- Square	Pr > ChiSq
Intercept	0.27	0.37	-0.51	0.99	0.49	0.52
GS	0.42	0.13	0.17	0.67	11.27	0.00
Cyclical	0.48	0.12	0.24	0.72	14.86	0.00
PBIT/Sales	-0.01	0.01	-0.04	0.02	0.46	0.50
CA/Sales	0.00	0.01	-0.01	0.02	0.19	0.66
CA/TA	0.06	0.22	-0.37	0.49	0.07	0.80
Debt/TA	-0.08	0.16	-0.38	0.22	0.26	0.61
CFO/Long Debt	0.00	0.00	0.00	0.00	0.20	0.66
Log (TA)	0.36	0.06	0.25	0.47	41.99	<.0001
Cash/CL	0.00	0.00	0.00	0.00	0.44	0.51
Net Debt Increase	0.00	0.00	0.00	0.00	1.78	0.18
Earnings/Equity	0.00	0.00	0.00	0.00	0.10	0.75
EPS Growth	0.00	0.00	0.00	0.00	2.15	0.14
PBIT/Capital	0.00	0.00	0.00	0.00	0.83	0.36
Scale	0.73	0.04	0.65	0.82		
Weibull Shape	1.36	0.08	1.21	1.53		

Table 6.2 Estimation Results of Parametric Survival Regression - the Log Logistic Distribution

Name of Distribution	log-logistic
Log Likelihood	-543.76

Analysis of Maximum Likelihood Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		Chi- Square	Pr > ChiSq
Intercept	-0.51	0.45	-1.38	0.37	1.3	0.25
GS	0.50	0.13	0.25	0.75	15.33	<.0001
Cyclical	0.54	0.13	0.29	0.79	17.76	<.0001
PBIT/Sales	-0.01	0.02	-0.04	0.02	0.86	0.35
CA/Sales	0.01	0.01	-0.01	0.02	0.42	0.52
CA/TA	0.07	0.22	-0.37	0.50	0.09	0.77
Debt/TA	-0.16	0.19	-0.54	0.22	0.67	0.41
CFO/Long Debt	0.00	0.00	0.00	0.00	0.44	0.51
LOG (TA)	0.43	0.07	0.30	0.56	39.31	<.0001
Cash/CL	0.00	0.00	0.00	0.00	0.59	0.44
Net Debt Increase	0.00	0.00	0.00	0.00	2.59	0.11
Earnings/Equity	0.00	0.00	0.00	0.00	0.07	0.79
EPS Growth	0.00	0.00	0.00	0.00	2.84	0.09
PBIT/Capital	0.00	0.00	0.00	0.00	1.26	0.26
Scale	0.60	0.04	0.54	0.68		

Table 6.3 and 6.4 provide the comparisons between the expected signs of all covariate and their actual estimated signs for the Weibull model and the log-logistic model.

Table 6.3 Estimated Effects on Covariate's Survival Time and Signs - the Weibull Distribution

WEIBULL Pamrametric Regression

Covariates	Parameter	Effect on Survival Time	P-Value	Expected Sign	Sign Indicator *
GS	0.42	1.52	0.00	†	√
Cyclical	0.48	1.62	0.00	†	√
PBIT/Sales	-0.01	0.99	0.50	†	x
CA/Sales	0.00	1.00	0.66	—	x
CA/TA	0.06	1.06	0.80	†	√
Debt/TA	-0.08	0.92	0.61	—	√
CFO/Long Debt	0.00	1.00	0.66	†	x
Log (TA)	0.36	1.43	<0.0001	†	√
Cash/CL	0.00	1.00	0.51	†	√
Net Debt Increase	0.00	1.00	0.18	—	x
Earnings/Equity	0.00	1.00	0.75	†	√
EPS Growth	0.00	1.00	0.14	†	x
PBIT/Capital	0.00	1.00	0.36	†	√

* "√" indicates that the estimated sign is consistent with the expected sign;
"x" indicates that the estimated sign is opposite to the espected sign.

Table 6.4 Estimated Effects on Covariate's Survival Time and Signs - the Log-logistic Distribution

Log-Logistic Parametric Regression

Covariates	Parameter	Effect on Survival Time	P-Value	Expected Sign	Sign Indicator *
GS	0.50	1.65	<0.0001	†	√
Cyclical	0.54	1.72	<0.0001	†	√
PBIT/Sales	-0.01	0.99	0.35	†	x
CA/Sales	0.01	1.01	0.52	—	x
CA/TA	0.07	1.07	0.77	†	√
Debt/TA	-0.16	0.85	0.41	—	√
CFO/Long Debt	0.00	1.00	0.51	†	x
LOG (TA)	0.43	1.54	<0.0001	†	√
Cash/CL	0.00	1.00	0.44	†	√
Net Debt Increase	0.00	1.00	0.11	—	x
Earnings/Equity	0.00	1.00	0.79	†	√
EPS Growth	0.00	1.00	0.09	†	x
PBIT/Capital	0.00	1.00	0.26	†	√

* "√" indicates that the estimated sign is consistent with the expected sign;
"x" indicates that the estimated sign is opposite to the espected sign.

It is noted that most covariates got consistent signs with our expectations. Among the 13 covariates used, only five covariates got the opposite signs compared with the expectation, namely, “CFO/Long Debt”, “PBIT/Sales”, “Net debt Increase”, “EPS Growth” and “CA/Sales”.

Statistically, a negative sign implies that one percent increase in the covariate will decrease the survival time by Exp (estimated parameter) percent (See Section 3.3). For instance, in the Weibull model if the estimated parameter for DebtTA equals to -0.08, it means that when debt over total asset of a company increase by 1 percent, the number of survival years will be shortened by 0.92 percent (exp(-0.08)). The effects on the change of survival time based on one percent change in the covariate for each of the 13 covariates in both models are also shown in table 6.3 and 6.4.

6.1.2.2 The Log-logistic model fits data better than the Weibull model

A scale parameter (σ) is estimated for both the Weibull model and the log-logistic models. This parameter controls the shape of hazard curve (see Chapter 3). The estimated σ is 0.73 in the Weibull model, which represents a monotonically increasing (although increase in a decrease speed) hazard. Similarly, the estimated of σ =0.60 for the log-logistic model represents a “hump” shape hazard. Recall that the hazard curve under the non-parametric (Kaplan-Meier) method, which is regarded as a simple reflection of characteristics of true data, demonstrates a fluctuated curve, the hazard curve assumed in the log-logistic model probably fit the data better than that the one assumed in the Weibull model.

The estimation results also provide the log-likelihood value for each model. The larger it is, the better the data fit the model. However, since log-likelihood values are negative, larger log-likelihood values have smaller absolute values. As shown in

Table 6.1 and 6.2, the log-likelihood for the Weibull and the Log-logistic models are -557.61 and -543.76 respectively, therefore, the loglogistic model fits the data better than the Weibull model.

6.1.3 Estimation Results of the Cox's PH Regression

6.1.3.1 Global Test for Model Significance

In this study the companies' survival time is measured as number of years instead of months or days and hence the survival time of over 500 bankrupt companies ranges in a small interval, from 1 to 17. When two firms have a same survival time, a tie is occurred. Obviously the data have many ties. Ties will incur a rather long computation time for the Cox's PH model if they are not properly handle. As discussed in Section 3.4.2, Efron's method is selected to speed up the computation.

Table 6.5 Estimation Results of the Cox's Proportional Hazard Regression

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	38.14	13	0.00
Score	40.71	13	0.00
Wald	41.03	13	<.0001

Analysis of Maximum Likelihood Estimates

Covariate	Parameter	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
GS	-0.54	0.17	10.46	0.00	0.58
Cyclical	-0.64	0.17	14.92	0.00	0.53
PBIT/Sales	0.01	0.02	0.51	0.48	1.01
CA/Sales	-0.01	0.01	0.29	0.59	0.99
CA/TA	-0.03	0.30	0.01	0.91	0.97
Debt/TA	0.10	0.22	0.20	0.65	1.10
CFO/Long Debt	0.00	0.00	0.32	0.57	1.00
LOG (TA)	-0.34	0.08	17.03	<.0001	0.71
Cash/CL	0.00	0.00	0.86	0.35	1.00
Net Debt Increase	0.00	0.00	1.72	0.19	1.00
Earnings/Equity	0.00	0.00	0.05	0.82	1.00
EPS Growth	0.00	0.00	2.50	0.11	1.00
PBIT/Capital	0.00	0.00	0.60	0.44	1.00

Table 6.5 shows a test on global null hypothesis, which are analogous to the F-test in linear regression models, is provided. Within this test, three alternative chi-square statistics are calculated: likelihood-ratio statistic, score statistic, and Wald statistic. All these three statistic tests are asymptotically equivalent.

As is shown in table 6.5, all three statistics present very small p-values, which indicate that the null hypothesis of all 13 coefficients equaling to 0 should be rejected.

6.1.3.2 Risk Ratio, P-Value and Sign for each Covariate

Table 6.5 also shows the estimation results of the Cox's PH regression. As observed, there is no intercept which is a typical characteristic of partial likelihood estimation. Different from the other regressions, a hazard ratio is provided for each covariate. For dummy variables, the hazard ratio can be interpreted as the ratio of the estimated hazard for those with a value of 1 to those with a value of 0 (controlling for other covariates). For example, if the estimated hazard ratio for “GS” is 0.58 it means that the hazard of bankruptcy for consumer goods firms (GS=1) is only about 58 percent of the hazard for consumer service (GS=0) firms. Similarly, if the hazard ratio for “Cyclical” dummy is 0.53, the hazard of getting bankruptcy for cyclical goods or service firms (Cyclical=1) is about 53% for non-cyclical goods or service firms (Cyclical=0).

For quantitative covariates, according to Allison (1995), a more helpful statistic, the effect on hazard, is obtained by subtracting 1 from the hazard ratio and multiplying the result by 100. This gives the estimated percentage change in the hazard for each one-unit increase in the covariate. For instance, for variable LOG (TA), when the hazard ratio is 0.71, the effect on hazard equals 100(0.71-1) = -29. Therefore, for one percent increase in LOGTA, the hazard of bankruptcy goes down for 29 percent.

Table 6.6 demonstrated the effect on hazard (in percentage) of all covariate. These values can be used as indicators for the importance of the covariates.

Table 6.6 Estimated Effects on Covariate's Survival Time and Signs - the Cox's PH model

Covariate	Parameter	Hazard Ratio	Effect on Hazard	P-Value	Epected Sign	Sign Indicator*
GS	-0.54	0.58	-41.90	0.00	—	√
Cyclical	-0.64	0.53	-47.40	0.00	—	√
PBIT/Sales	0.01	1.01	1.40	0.48	—	x
CA/Sales	-0.01	0.99	-0.60	0.59	†	x
CA/TA	-0.03	0.97	-3.30	0.91	—	√
Debt/TA	0.10	1.10	10.40	0.65	†	√
CFO/Long Debt	0.00	1.00	0.10	0.57	—	x
LOG (TA)	-0.34	0.71	-28.60	<0.0001	—	√
Cash/CL	0.00	1.00	-0.10	0.35	—	√
Net Debt Increase	0.00	1.00	0.00	0.19	†	√
Earnings/Equity	0.00	1.00	0.00	0.82	—	√
EPS Growth	0.00	1.00	0.00	0.11	—	x
PBIT/Capital	0.00	1.00	-0.10	0.44	—	√

* "√" indicates that the estimated sign is consistent with the expected sign;
"x" indicates that the estimated sign is opposite to the expected sign.

There are three covariates significant at a 5% level in the Cox's PH model output: "GS", "Cyclical" and "LOGTA". It is noted that the three significant covariates in the Cox's PH model are same as the ones in the parametric models. Beside, the coefficients as well as the associated p values estimated in Cox's model are also similar to their counter parties in parametric models.

The signs of parameters in the Cox's PH model are opposite to those in parametric models. This is not surprising since in parametric model the estimation is based on a log-survival format, while in the Cox's PH model the estimation is based on a log-hazard format. According to Table 6.6, in estimating the signs of all 13 covariates, the Cox's PH regression correctly estimates 9 covariates' signs and only does wrong in 4 covariates' signs. Compared with parametric survival regressions, the Cox's PH regression makes an improvement with correctly estimating the sign of variable “Net debt Increase”, whose sign is mistakenly estimated by both parametric regressions.

Before reporting the estimation output of the Cox’s TVC model, two more tests still need to be conducted for the Cox’s PH model: the function form test and the proportionality test. This is because the output of the function form test may help to justify the introduction of nonlinear model, and the output of the proportionality test will explain the necessity of introducing the Cox’s TVC model.

6.1.3.3 Assessing Functional Form

As discussed in Chapter 3, martingale residuals can be employed to examine the best functional form. This target can be accomplished by plotting the martingale residuals from a model with the variable of interest removed, versus the variable of interest. To help to estimate the relationship, a smooth line can be added to the plot of martingale residuals. In this study, all 11 covariates with continuous value are examined with martingale residuals test. If most variables investigated show non-linear property, the linearity assumption of models should be denied.

When the smoothed line which represents the relationship between a covariate and the associated the martingale residuals is reasonably linear and around 0, the correct form of this covariate is considered as linear. However, as displayed in Figure 6.5, for most covariates their smoothed lines about martingale residuals are obviously non-liner, which reveals the fact that, a linear model is not appropriate. To attain more accurate prediction results, maybe some none-linear approaches, such as artificial neural networks (ANNs), shall be employed.

Figure 6.5 Martingale Residuals for 11 Continuous Covariates

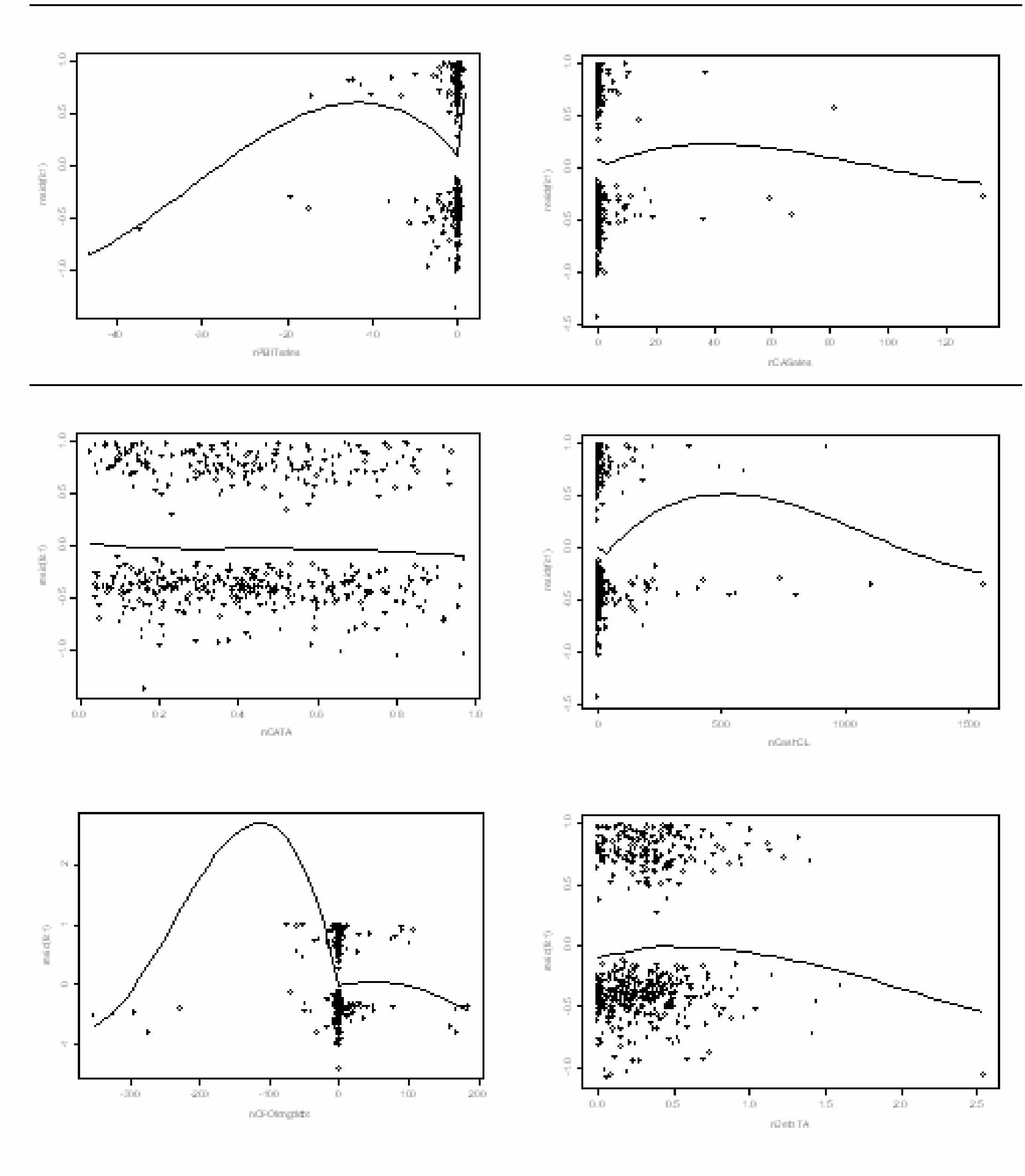
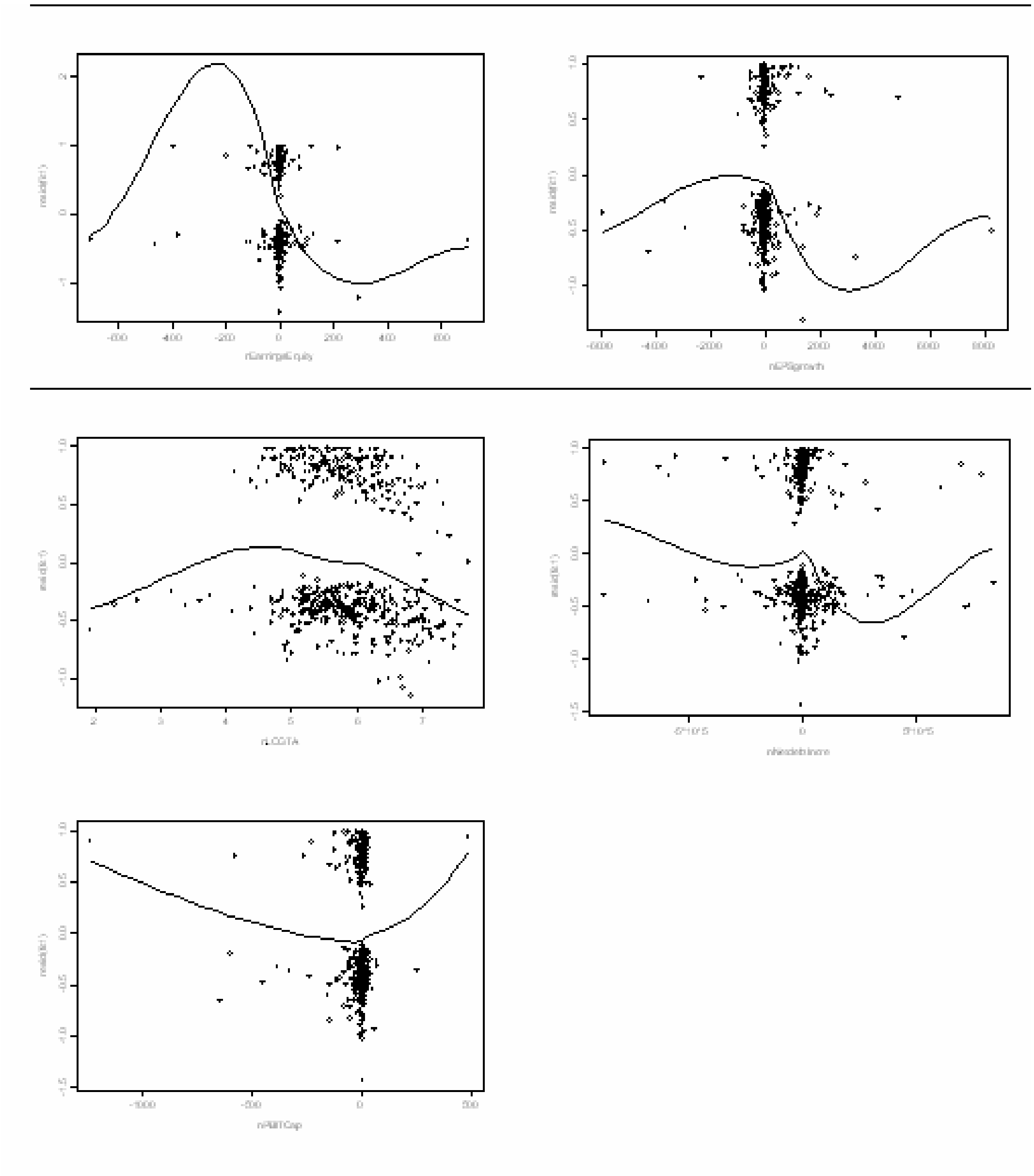


Figure 6.5 Martingale Residuals for 11 Continuous Covariates (Continued)



6.1.3.4 Test of Proportionality

The validity of the Cox's PH regression analysis relies much on the assumption of proportionality.

One simple method for testing the proportionality of the Cox's PH model is to add covariates by log-time interactions to the model and assess their significance. The rationale of this test comes as follows: As is shown by (Frank and Harrell, 2001), in the Cox's semi-parametric model, the hazard function is assumed as the following form:

$$h(t, x, \beta, h_0) = \exp(x\beta) h_0(t)$$

thus, $\log h(t|x) = \log h_0(t) + x\beta$.

The effect of the predictors is assumed to be the constant at all values of t, since $\log h(t)$ can be separated from $x\beta$.

However this method is practically hard to implement. First, it is hard to determine which covariates should be chosen to form time interaction terms. Second, the exact form of time-interaction with the covariates is unknown. Many time-interaction forms of the 13 covariates have been tested but none of the trials could achieve significant level better than 5%. Therefore the proportionality assumption should not be denied based on the above test.

Another test on the proportionality is the scaled Schoenfeld residuals statistic test. As observed in table 6.7, for most covariates the none-hypothesis of proportionality assumption can not be rejected and the general test of covariates again confirms this conclusion. Only two covariates demonstrate some none-proportional characteristics: test on "GS" reject the none-hypothesis of proportionality at a 5% significant level, and test on "Log (TA)" is significant at a 10% level.

Table 6.7 Statistical Test of Proportionality - Scaled Schoendfeld's Residuals

Covariate	Rho	Chi Sqared	P
GS	0.14	4.59	0.0322**
Cyclical	0.06	0.78	0.38
PBIT/Sales	-0.02	0.03	0.87
CA/Sales	0.04	0.40	0.53
CA/TA	0.01	0.01	0.93
Debt/TA	0.09	1.44	0.23
CFO/Long Debt	0.01	0.01	0.92
LOG (TA)	0.15	3.15	0.0761*
Cash/CL	-0.04	0.38	0.54
Net Debt Increase	0.04	0.46	0.50
Earnings/Equity	0.03	0.06	0.81
EPS Growth	-0.02	0.08	0.78
PBIT/Capital	-0.04	0.58	0.45
GLOBAL	NA	10.30	0.67

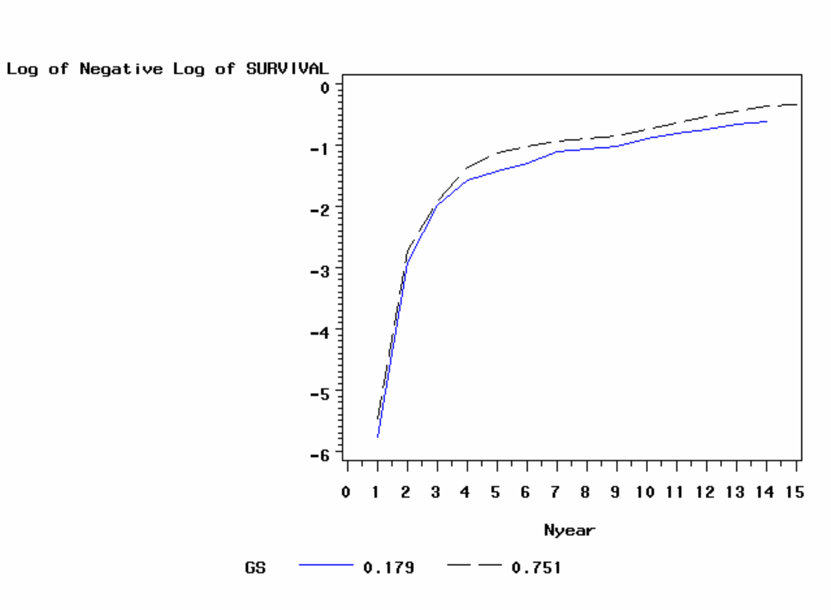
** Significant at 5% level
* Significant at 10% level

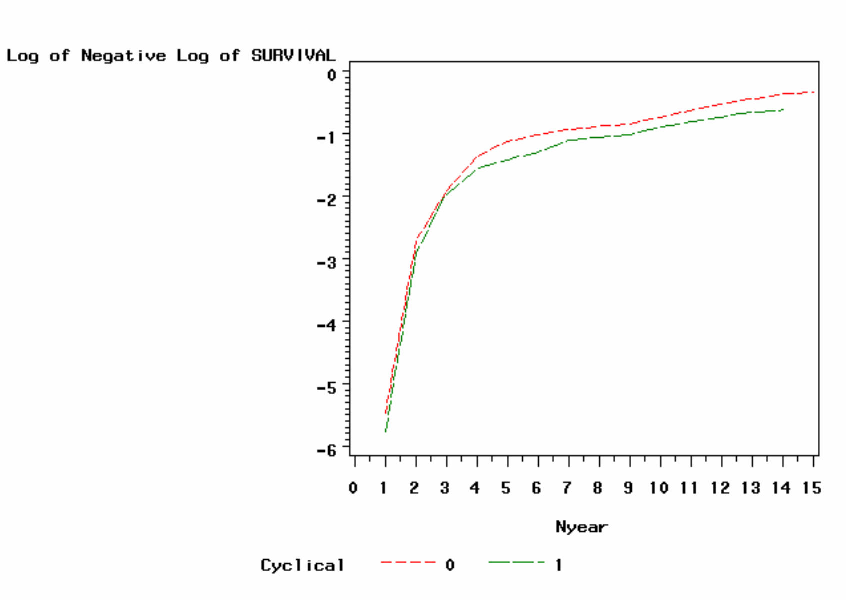
In additional to statistical tests, graphic tests are also frequently used for proportionality test in the Cox's PH model. Here, two graphic tests are presented.

The first approach uses the log-log survival functions for testing of the proportional assumption. As illustrated in Chapter 3, when the graphs of $\log[-\log S_1(t)]$ and $\log[-\log S_2(t)]$ are parallel, the proportionality assumption is valid.

Figure 6.6 shows that in "GS" dummy the log-log survival curve in consumer goods group is overlapping with the log-log survival curve in consumer services group for the first three or four years. Later, these log-log survival curve of these two groups diverse, and roughly keep parallel to the end of time. Similar pattern appeared on "Cyclical". This implies that the proportionality does not exist in the whole examination time range, but such proportional hypothesis can be held in the later years. However, this graphic test is only applicable for dummy variables.

Figure 6.6 Graphic Proportionality Test 1 - Log-log Survival Plots for "GS" and "Cyclical"





The second graphic test, the scaled Schoenfeld residuals test, can be used to test the proportionality assumption in covariates other than dummies.

As described in Chapter 3, when the smoothed scaled Schoenfeld residuals curve is flat and laying around 0, the proportionality hypothesis can be regarded as true. As is shown in figure 6.7, most covariates get roughly flat curves around 0, except for “Cyclical”, “GS” and “Log(TA)”, which is consistent with our findings in statistical test.

Figure 6.7 Graphic Proportionality Test 2 - Plots of Scaled Shoenfeld's Residuals for All Covariates

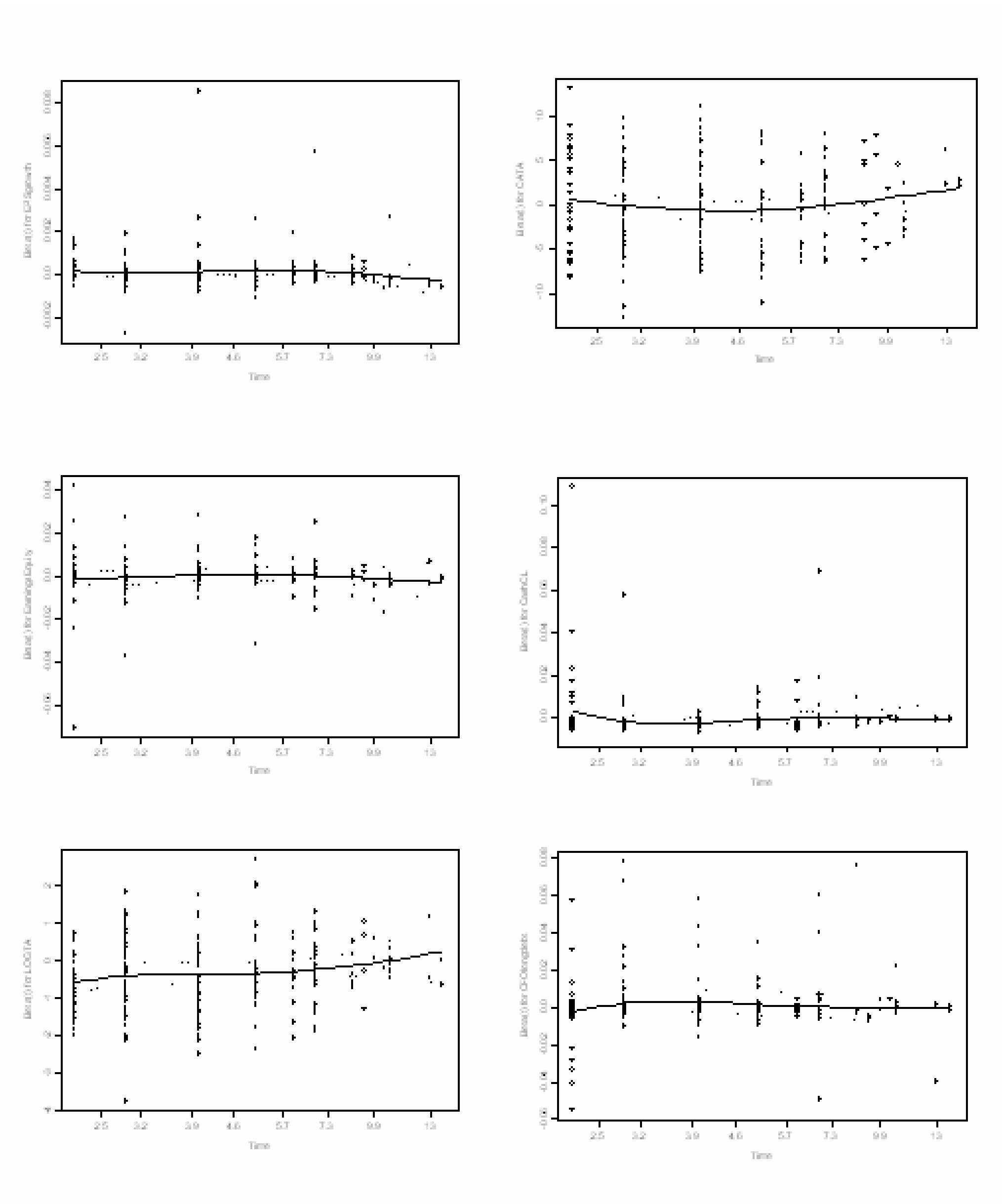
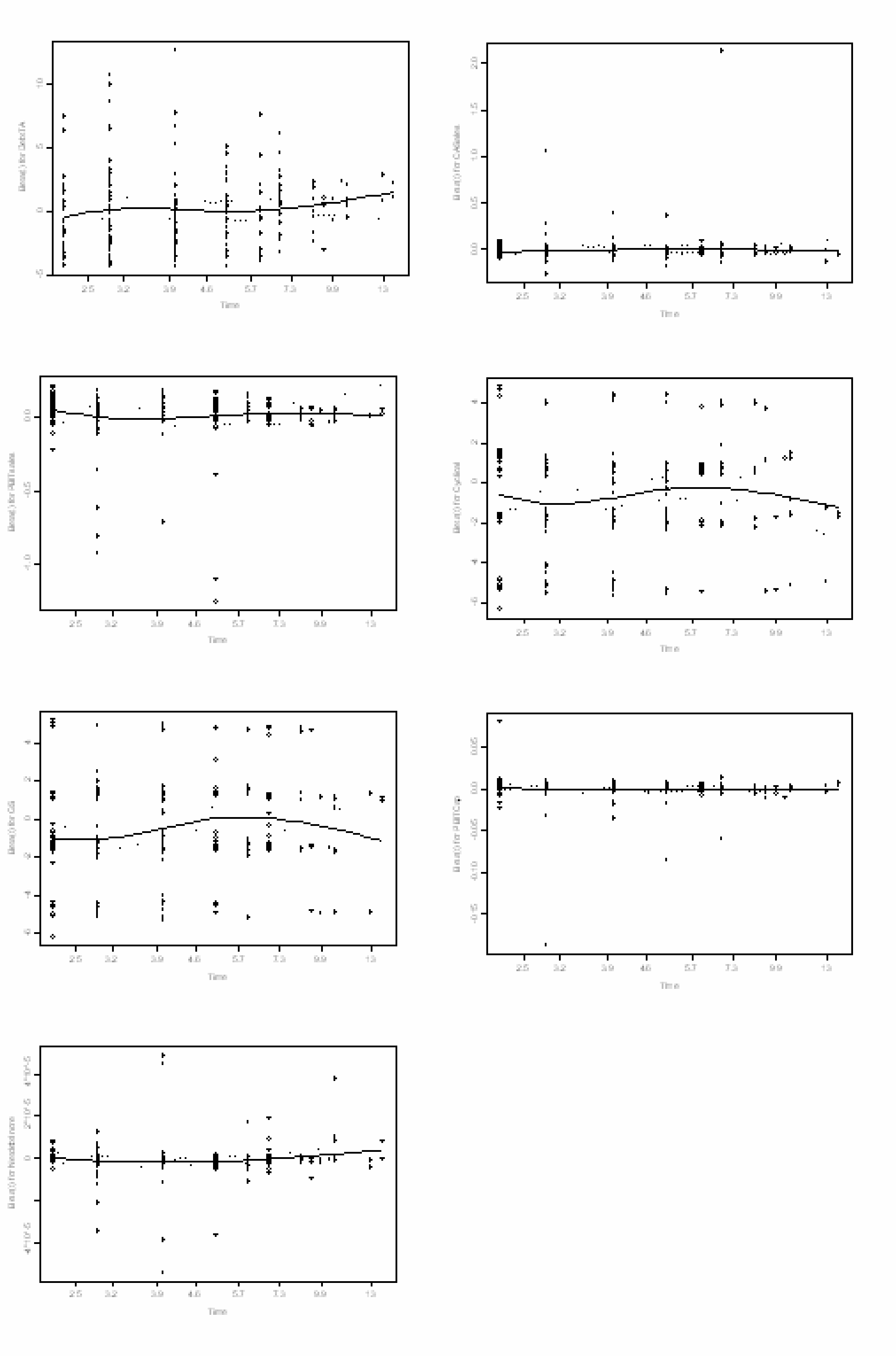


Figure 6.7 Graphic Proportionality Test 2 - Plots of Scaled Shoenfeld's Residuals for All Covariates (Continued)



After considering about all the results from above tests on the proportionality assumption, we conclude that, the proportionality assumption is valid for most of covariates but can not be held in the whole covariates set.

6.1.4 Estimation Results of the Cox's TVC Model

The first reason for employing TVC model comes from the results of proportionality test in previous part. Since when one covariate in a Cox's PH model is proven to be time-dependent, the proportionality assumption in the Cox's PH model becomes invalid.

Another reason for using TVC model relates with the economic meaning of the proportionality assumption. As is known, the covariates in the Cox's PH model are supposed to be time-independent over the entire interval of survival time. Since the survival time ranged from 1 year to 17 years in this study, the proportionality assumption requires all the accounting ratio covariates keeping constant over years, which is hard to be true. However, as stated in Shumway (2001), the Cox's TVC model is the only model that can incorporate multiple-year information of a firm, and hence theoretically it can provide preferable output to other models which just rely on one-year data.

Table 6.8 Estimation Results of the Cox's TVC Regression

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > Chi Sq
Likelihood Ratio	38.53	13	0.00
Score	54.88	13	<.0001
Wald	45.70	13	<.0001

Analysis of Maximum Likelihood Estimates

Covariate	Parameter	Standard Error	Chi - Square	Pr > ChiSq	Hazard Ratio
GS	-0.47	0.15	9.55	0.00	0.62
Cyclical	-0.28	0.15	3.27	0.07	0.76
Earning/TA	0.00	0.01	0.09	0.76	1.00
PBIT/Sales	-0.01	0.00	4.93	0.03	0.99
Debt/TA	0.00	0.00	3.56	0.06	1.00
Sales/TA	-0.14	0.08	3.40	0.07	0.87
CA/CL	0.00	0.00	4.37	0.04	1.00
FA/TA	0.00	0.00	15.14	<.0001	1.00
LOG (TA)	-0.19	0.08	6.25	0.01	0.83
Net Debt/Equity	0.00	0.00	0.11	0.74	1.00
Earning/Equity	0.00	0.00	2.29	0.13	1.00
Earnings per Share	0.00	0.00	0.07	0.80	1.00
Totalflow/TA	0.36	0.42	0.72	0.40	1.43

Table 6.8 shows the estimation results of the Cox's TVC model. Similar to the Cox's PH regression, the TVC model are significant in all three global non-hypothesis tests. But when concerned about significance in individual covariate, the TVC model exhibits stronger estimation capability: 8 out of 13 covariates employed are significant at a 10% level, within which 5 are significant at a 5% level. However, it is also noted that within total 13 covariates 6 of them have no effect on the hazard function estimation as their estimated hazard ratio is 1.

Table 6.9 shows effect on survival time of each covariate and the sign indicator which indicates whether the sign of covariate is correctly estimated or not.

As is shown, the TVC model correctly estimate 10 covariates' signs, while the Cox's PH model only gets 9 covariates' signs right.

Table 6.9 Estimated Effects on Covariate's Survival Time and Signs - the Cox's TVC model

Covariate	Parameter	Hazard Ratio	Effect on Hazard	P-Value	Epected Sign	Sign Indicator*
GS	-0.47	0.62	-37.80	0.00	—	√
Cyclical	-0.28	0.76	-24.30	0.07	—	√
Earning/TA	0.00	1.00	0.30	0.76	—	x
PBIT/Sales	-0.01	0.99	-0.80	0.03	—	√
Debt/TA	0.00	1.00	0.00	0.06	—	√
Sales/TA	-0.14	0.87	-13.20	0.07	—	√
CA/CL	0.00	1.00	0.00	0.04	—	√
FA/TA	0.00	1.00	0.00	<0.0001	—	√
LOG (TA)	-0.19	0.86	-13.80	0.01	—	√
Net Debt/Equity	0.00	1.00	0.00	0.74	†	x
Earning/Equity	0.00	1.00	0.00	0.13	—	√
Earnings per Share	0.00	1.00	0.00	0.80	—	√
Totalflow/TA	0.36	1.43	42.80	0.40	—	x

* "√" indicates that the estimated sign is consistent with the expected sign;
"x" indicates that the estimated sign is opposite to the espected sign.

6.1.5 Summary of Comparisons among Estimation Results from Survival Regressions

Table 6.10 provides a summery on the comparison among survival regressions employed in terms of significant covariates number and the number of correctly estimated signs. Clearly, the Cox's TVC model has the best estimation capability, since it can obtain the highest number of significant covariates and the highest number of correctly estimated signs of covariates. This result is conceivable since the Cox's TVC model uses the TVC dataset while other regressions only use the one-year dataset which contains much less information. Other than the Cox's TVC model, both log-logistic parametric regression and the Cox's PH show better estimation capability compared with Weibull parametric regression.

Table 6.10 Comparison among Four Survival Regressions: the Number of Significant Covariates and the Number of Correctly Estimated Signs

Regression Model	number of significant covariate (5%)	number of significant covariate (10%)	number of correctly estimated signs
Weibull parametric	3	0	8
log-logistic parametric	3	1	8
the Cox's PH	3	0	9
the Cox's TVC	5	3	10

6.2 Model Training of Neural Networks

6.2.1 Training the Standard ANNs

The standard ANN refers to the neural network that has the standard structure, the standard likelihood function and is easily to be implemented with the standard software. In this study, the input neurons are the same 13 covariates as in the Cox's PH model, the number of hidden neurons is ranged from 0 to 10, and the output neuron is the "status" dummy in data.

Although ANNs usually can provide much higher prediction accuracy than linear models， however, as stated in Saunders and Allen (2002), ANNs do nothing to illuminate the process or the relative importance of the variables which is usually referred as a black box problem. More over, because the internal structure of the ANN is hidden， it may not be easy to duplicate even using the same data inputs, which leads to a lack of accountability because the intermediate steps of the system cannot be checked (Altman et al., 1994).

Because of the black box problem in neural networks, there is little to compare between regression models and the neural networks about their estimation results. The remaining part of this chapter just briefly goes through the estimations and estimation results (if applicable) in survival ANNs. The model comparisons between regressions and neural networks are therefore focused on the comparison among prediction results which are intensively discussed in Chapter 7.

6.2.2 Training the Survival ANNs

Two important issues need to be carefully considered in training the survival ANNs, namely, the optimal structure and the local maxima.

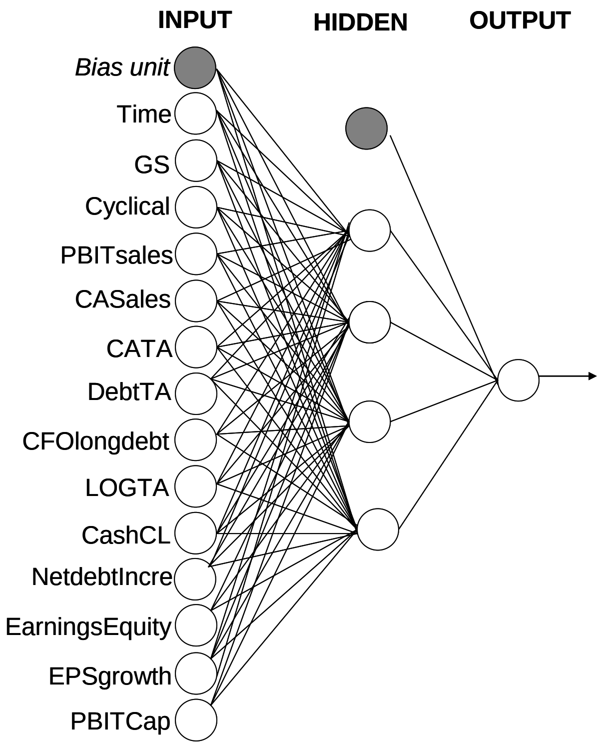
As stated in Chapter 4, the structure of networks is determined by both the number of hidden neurons as well as the parameter of weight decay. However, to determined the number of hidden neurons is not easy. If the number of hidden neurons is too small, the risk of under-fitting is high and the outputs maybe are statistically biased. But if the number of hidden neurons is too large, the risk of over-fitting increases fast and the outputs probably are no more statistically significant. Usually, a rule of thumb is that it should never be more than twice as large as the input layer (Berry and Linoff, 1997). Some other solutions are also available, e.g., Blum (1992) states that the number of hidden neurons should be between the number of input neurons and number of output neurons, while Wanas et al. (1998) say that, after considering the balance of the optimal performance and the computation cost, the optimal number of hidden neurons should be equal to log (number of training samples). But all of these studies agree that, training the neural networks with different numbers of hidden neurons and calculate the general accuracy ratio is necessary steps in determining the optimal number of hidden neurons. Therefore, in training the survival ANNs, the optimal structure of the ANNs is determined by seeking the best prediction result while changing the combination of number of hidden neurons and parameter of weight decay. In this study, same design is employed as Ripley (1998). Specifically, in each model of survival ANNs, the number of hidden neuron is varying with the following numbers, i.e., 0, 1, 2, 3, 5, and 10, combining with different weight decay parameters, with the larger weight decay parameter used for the models with more hidden nodes. The optimal structure of an ANN is the combination that generated the highest prediction results. The details will be discussed in tables 7.6 - 7.10.

The local maxima issue is related with the non-linear characteristic of ANNs. If the function of parameters is linear, the log likelihoods are concave and a unique maximum will exist. However, if hidden nodes are included, the function of parameters becomes nonlinear and the log likelihoods are no more simply concave. Consequently the unique maximum is no longer applicable and local maxima will be identified (Ripley, 1998). The solution is to train a same network for several times with different random starting weights and average these results. This is an approximation to integrate over the weight space in a Bayesian framework, and much less computer intensive than the full Bayesian approach. As demonstrated by Mathieson (1998), this technique is a good way to deal with local maxima when fitting neural networks.

6.2.2.1 Partial Logistic Artificial Neural Network (PLAAN)

The first type of survival ANNs employed in this study is PLAAN designed by Biganzoli et al.(1998). In PLAAN, 15 input nodes are used which include all 13 covariates in one year dataset, one time interval covariate, and one bias node. The time interval covariate is identical to the survival time in the Cox's PH regression. A structure graph of a PLAAN is demonstrated in figure 6.8.

Figure 6.8 Structure of PLAAN



PLAAN shares many common characteristics with the standard ANN. First, both of them have only one output node and this output node is represented by the event indicator: the "status" dummy. Second, both of them have same input variables: besides of the same 13 covariates used in survival regressions, the variable survival time and a random variable are used in both PLAAN and the standard neural networks. Third, these two models employ same activation function. Both of them use the traditional logistic function as the activation for both the hidden nodes and the single output node. The only difference of these two models is their likelihood functions. Specifically, in standard ANN, the likelihood function is

$$E = \sum_{i=1}^{s}\sum_{j=1}^{n}\frac{(y_{ij} - d_{if})^2}{2} \qquad (6.1)$$

However, for PLAAN model, the likelihood function is the cross-entropy error:

$$E = -\sum_{i=1}^{n}\sum_{l=1}^{l_i}\{d_{il}\log h_l(x_i, a_l) + (1 - d_{il})\log[1 - h_l(x_i, a_l)]\} \qquad (6.2)$$

(6.1) and (6.2) are identical as (4.1) and (4.3) correspondingly.

Therefore, as discussed in Chapter 4, for the standard ANN, the output of the network is just an "approximator" of the actual value, but for PLAAN its output can be viewed as the conditional bankruptcy probability.

6.2.2.2 Parametric Survival ANNs

Two types of hazard distributions, the Weibull distribution and the log-logistic distribution, are implemented in parametric survival ANNs.

The Weibull survival ANN is a network which assumes a monotonic hazard curve in its model specification. When training the Weibull survival ANN to find the optimal structure, the shape parameters "alpha" reported in all of the training networks range from 0.3 to 0.5, which implies that a monotonically increasing hazard is assumed in the Weibull survival network (see Section 3.3). This result is consistent with the result in the Weibull parametric regression, however, it also implies that the Weibull survival ANN suffers by the same limitation as the Weibull parametric regression does, since a monotonically increasing hazard curve contradicts to the features of the actual data. Thus, the Weibull survival ANN is vulnerable and probably only makes a weak fitness.

The log-logistic distribution, instead, can imply a non-monotonic hazard curve as well as a monotonic decreasing one. The shape parameters generated from different structures of the log-logistic survival ANNs range from 0.4 to 0.7 which reveals an inverted U-shaped curve (see 3.3), which is consistent with the results got from log-logistic parametric regression. Thus the log-logistic survival ANN suggests a better fitness than the Weibull survival ANN since the hazard curve assumed in log-logistic survival ANN is closer to the hazard curve plotted by Kaplan-Meier method in Section 3.2.

6.2.2.3 The Cox's Proportional Hazard (PH) Survival ANN

The Cox's PH survival ANN employed here is designed by Ripley (1998). Generally speaking, this survival ANN is almost identical to one designed by Faraggi and Simon (1995) except that the input nodes are allowed to link with the output nodes directly. Thus, if $\eta(x)$ is modeled as the output from a neural network with one linear output node, $\eta(x)$ can be expressed as:

$$\eta(x) = \sum_{j \neq 0} w_{jo} x_j + \sum_h w_{ho} l\left(\sum_j w_{jh} x_j \right) \tag{6.3}$$

where the x_j is the input covariate; w_{jh} s are the weight parameters connecting the input variables and the hidden nodes; w_{ho} s are the parameters between the hidden nodes and the output node; w_{jo} s are the parameters directly connecting the input nodes and the output node.

Similar like survival ANN designed by Faraggi and Simon (1995), Ripley substitutes the liner form βx_i with the nonlinear combination from a network to generate the non-linear log-likelihood function.

CHAPTER 7 PREDICTION RESULTS AND COMPARISON AMONG MODELS

After comparing different survival models based on parameter estimations in previous chapter, Chapter 7 shows the prediction outputs of all models and compares them based on their prediction performance. Specifically, this chapter starts with an introduction of four comparison criteria, namely, misclassification cost, sensitivity, specificity and accuracy. Then, the prediction outputs of each model, from linear models to all ANNs, are demonstrated one by one. Finally, the superiority of all models in terms of their prediction power are summarized at the end.

7.1 Comparison Criteria

Four values, namely, misclassification cost, sensitivity, specificity and accuracy are used as comparison criteria when comparing prediction outputs of different models. Specifically, misclassification cost is used for comparison among survival regressions, sensitivity, specificity and accuracy are applied for comparisons when survival ANNs is involved.

Misclassification cost is incurred when an individual in one group is mistakenly classified into another group. In bankruptcy prediction, if a bankrupt firm is mistakenly classified as a non-bankrupt firm, type I misclassification cost occurs; on the contrary, when a non-bankrupt firm is misclassified as a bankrupt one, type II misclassification cost occurs. Type I misclassification cost is much higher than the type II misclassification cost. Because when a bankrupt firm is predicted as a non-bankrupt firm, the investors may lose their entire investment which can be very substantial, on the other hand, when a non-bankrupt firm is predicted as a bankrupt one, the investors may not invest in that firm and consequently just lose the dividend and the capital gain that may be obtained otherwise. Therefore, type I error should be minimized or well under control in prediction. misclassification cost can be used as a simple metric which can effectively measure the prediction power of models.

Since all the survival regression and survival ANNs can only generate the predicted survival probability or hazard probability, a cut-off point is needed to classify the firms as the bankrupt ones or the non-bankrupt ones for the misclassification cost calculation. And since the choice of the cut-off point will directly impact the calculation of type I error, type II error as well as the misclassification cost, a optimal cut-off point which can minimize type I and II as well as misclassification cost should to be carefully chosen (Koh, 1992; Dopuch, Holthausen and Leftwich, 1987).

A solution for optimal cut-off point proposed by Afifi and Clark (1996) is adopted in this study for survival regressions.

According to Afifi and Clark (1996), the optimal cut-off point C is:

$$C = \frac{\bar{Z}_I + \bar{Z}_{II}}{2} + K$$

And

$$K = \ln \frac{q_{II} \cos t(I \ given \ II)}{q_I \cos t(II \ given \ I)}$$

Where

q_I is the prior probability of bankruptcy, which indicates the portion of bankrupt firms in the model training sample and equals to 0.008 in the one-year dataset;

q_{II} is the prior probability of non-bankruptcy, which indicates the portion of non-bankrupt firms in the training sample and is 0.992 (one minus the prior probability of bankruptcy);

cost (II given I) is the cost misclassifying a bankrupt firm as non-bankrupt;

cost (I given II) is the cost misclassifying a non-bankrupt firm as bankrupt;

$\bar{Z}_I$ is defined as the estimated survival probability of an "average" bankrupt firm, the covariates of which equal to the average of the bankrupt firms in the training sample;

$\bar{Z}_{II}$, is defined as the estimated survival probability of an "average" non-bankrupt firm, the covariates of which equal to the average of the non-bankrupt firms in the training sample.

Since it is impossible to estimate accurately the type II and type I misclassification cost in bankruptcy prediction (Altman, 1984), the type I and type II misclassification cost are analyzed through a range of relative values. In this study, this relative value $\frac{\text{cost (II given I)}}{\text{cost (I given II)}}$ ranges from 1:1 to 1:500 (Koh, 1992). Within this range, the cut-off point values vary from -2.0 to +2.0. Since only positive cutoff values can make sense, negative cutoff values are deleted.

Accuracy equals to $\frac{\textit{records correctly predicted}}{\textit{total number of records}}$ which measures the overall performance of the prediction. Specificity is the ratio of $\frac{\textit{survived firms correctly predicted}}{\textit{total number of survived firms}}$ which actually equals to 1 minus Type II error. Sensitivity is the ratio of $\frac{\textit{bankrupt firms correctly predicted}}{\textit{total number of bankrupt firms}}$ which equals to 1 minus type I error. When given a cut-off value and a set of predicted survival probabilities, a trade-off appear between the specificity and the sensitivity.

After working out the specificity and the sensitivity, it is easily to compute the total cost of misclassification. The total cost of misclassification can be defined as:

$$\begin{aligned}&total\ \cos t\ of\ \ misclassification = \\ &q_I Porb(II\ \ given\ I)\cos t(II\ \ given\ I) + q_{II}\Pr ob(I\ \ given\ II)\cos t(I\ \ given\ II)\end{aligned}$$

Where

Porb(II given I) = 1- sensitivity;

Porb(I given II) = 1- specificity.

Some studies (Koh, 1992; Chen and Church, 1992) suggest that the cut-off point C can be chosen by minimizing the total cost of misclassification. In this study, when the relative cost ratio of type II error over type I error changes from 1:1 to 1:500, the cut-off value that achieves the lowest misclassification cost under each relative ratio is obtained, and then the optimal cut-off value is the one that achieves the smallest of these lowest misclassification costs.

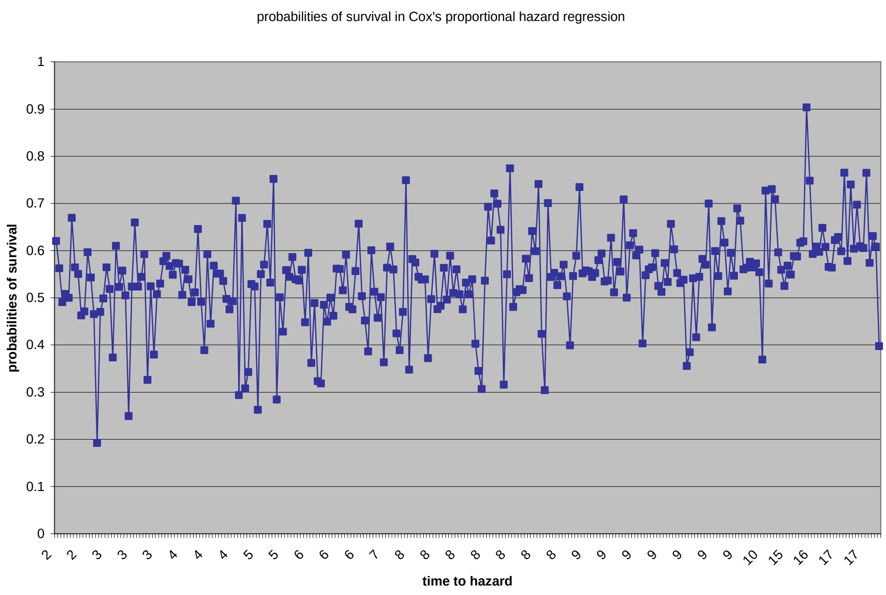
7.2 Prediction Performance and Model Comparison among Survival Regressions

Figure 7.1 to 7.4 demonstrate the plots of predicted survival probabilities against the hazard time generated by the Cox's PH regression, the Cox's TVC regression, the Weibull parametric survival regression and the log-logistic parameter survival regression.

The predicted survival probability here refers to the probability that a firm can remain non-bankrupt at its survival time. As defined in Chapter 5, the hazard time of a firm is the number of years that a firm remains survived, which is considered surrogate of firm age. As shown in figure 7.1 to 7.4, the curves of predicted survival probabilities from the four survival regressions demonstrate sort of upward slopes, which imply the probabilities of survival are higher for the older firms than the younger ones. This is compatible with the fundamental characteristic of the data found in Chapter 5: younger firms go bankrupt more frequently than the older ones.

Figure 7.1 shows that, the predicted survival probabilities in the Cox's PH regression ranged from 0.2 to 0.9, with most of them concentrated in a small interval from 0.5 to 0.6. Besides, the upward slope can hardly be identified. In this case, a cut-off value is very difficult to determine since bankrupt and non-bankrupt firms got similar predicted survival probabilities. Based on these observations, this model probably could not perform well.

Figure 7.1 The Probabilities of Survival Generated by the Cox's PH Regression



Similarly, in figure 7.2 the predicted survival probabilities are around 0.6 and no obvious upward slope is observed. Therefore, the Cox's TVC model probably also suffers a poor prediction performance.

Figure 7.2 The Probabilities of Survival Generated by the Cox's TVC Regression

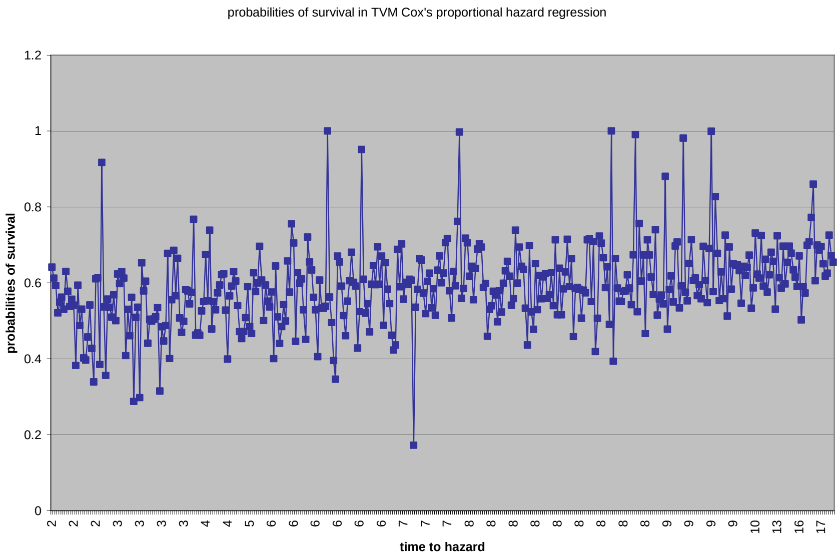


Figure 7.3 and 7.4 show the predicted survival probabilities in Weibull and log-logistic parametric regressions. The intervals of predicted survival probabilities in both models range from 0 to 1, and significant upward slopes are observed. Therefore, a cut-off value can be chosen which can classify all points above as non-bankrupt firms and points below as bankrupt ones, with most of older firms being classified as non-bankrupt and younger firms being classified as bankrupt ones.

Figure 7.3 The Probabilities of Survival Generated by the Parametric Weibull Regression

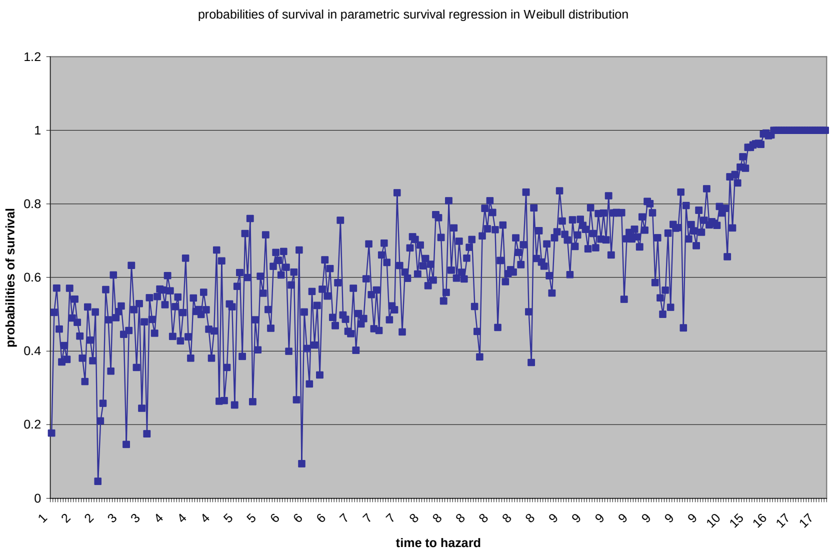
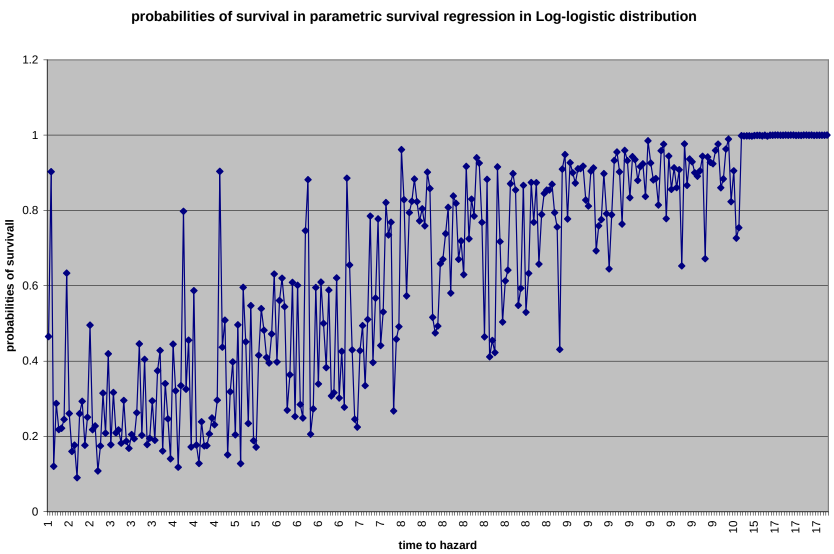


Figure 7.4 The Probabilities of Survival Generated by the Parametric Log-logistic Regression



Tables 7.1 to 7.4 show the prediction performance of all survival regressions and the cut-off value decided by minimizing the lowest misclassification cost at each

relative cost ratio. The row bolded indicates the optimal cut-off point which generates the global minimum misclassification cost within the range of relative cost ratio.

Table 7.1 Prediction Performance of the Cox's PH Regression

q1	q2	Cost (I given	Cost (II given	Cut-off Value	Prob (II given	Prob (I given	Accuracy	Specificit y	Sensitivit y	Misclassification Cost
0.008	0.992	1.000	110.000	0.402	0.806	0.157	0.609	0.843	0.194	0.865
0.008	0.992	1.000	115.000	0.447	0.694	0.230	0.602	0.770	0.306	0.867
0.008	0.992	1.000	120.000	0.489	0.565	0.309	0.599	0.691	0.435	0.849
0.008	0.992	1.000	125.000	0.530	0.398	0.435	0.579	0.565	0.602	0.829
0.008	**0.992**	**1.000**	**130.000**	**0.569**	**0.213**	**0.586**	**0.539**	**0.414**	**0.787**	**0.803**
0.008	0.992	1.000	135.000	0.607	0.148	0.728	0.482	0.272	0.852	0.882

Table 7.2 Prediction Performance of the Cox's TVC Regression

q1	q2	Cost	Cost	Cut-off Value	Prob (II given	Prob (I given	Accuracy	Specificit y	Sensitivit y	Misclassificatio n Cost
0.008	**0.992**	**1.000**	**105.000**	**0.386**	**0.958**	**0.012**	**0.645**	**0.988**	**0.042**	**0.816**
0.008	0.992	1.000	110.000	0.432	0.909	0.036	0.648	0.964	0.092	0.835
0.008	0.992	1.000	115.000	0.476	0.838	0.088	0.640	0.912	0.162	0.858
0.008	0.992	1.000	120.000	0.519	0.718	0.164	0.635	0.836	0.282	0.852
0.008	0.992	1.000	125.000	0.560	0.535	0.336	0.592	0.664	0.465	0.869
0.008	0.992	1.000	130.000	0.599	0.338	0.512	0.551	0.488	0.662	0.860
0.008	0.992	1.000	135.000	0.637	0.183	0.672	0.505	0.328	0.817	0.864
0.008	0.992	1.000	140.000	0.673	0.099	0.776	0.469	0.224	0.901	0.880

Table 7.3 Prediction Performance of the Parametric Weibull Survival Regression

q1	q2	Cost	Cost	Cut-off Value	Prob (II given	Prob (I given	Accuracy	Specificit y	Sensitivit y	Misclassification Cost
0.008	0.992	1.000	95.000	0.333	0.902	0.034	0.704	0.967	0.098	0.719
0.008	0.992	1.000	100.000	0.384	0.814	0.048	0.724	0.952	0.186	0.698
0.008	0.992	1.000	105.000	0.433	0.784	0.077	0.714	0.923	0.216	0.735
0.008	0.992	1.000	110.000	0.480	0.657	0.120	0.728	0.880	0.343	0.697
0.008	0.992	1.000	115.000	0.524	0.461	0.182	0.751	0.818	0.539	0.604
0.008	0.992	1.000	120.000	0.567	0.333	0.225	0.764	0.775	0.667	0.543
0.008	**0.992**	**1.000**	**125.000**	**0.607**	**0.245**	**0.292**	**0.748**	**0.708**	**0.755**	**0.535**
0.008	0.992	1.000	130.000	0.647	0.186	0.373	0.711	0.627	0.814	0.564
0.008	0.992	1.000	135.000	0.684	0.157	0.455	0.665	0.546	0.843	0.620
0.008	0.992	1.000	140.000	0.721	0.137	0.579	0.585	0.421	0.863	0.728
0.008	0.992	1.000	145.000	0.756	0.137	0.699	0.502	0.301	0.863	0.852
0.008	0.992	1.000	150.000	0.790	0.128	0.790	0.442	0.211	0.873	0.936
0.008	0.992	1.000	155.000	0.822	0.128	0.823	0.419	0.177	0.873	0.974

Table 7.4 Prediction Performance of the Parametric Log-logistic Survival Regression

q1	q2	Cost	Cost	Cut-off Value	Prob (II given	Prob (I given	Accuracy	Specificity	Sensitivity	Misclassification Cost
0.008	0.992	1.000	95.000	0.331	0.912	0.034	0.701	0.967	0.088	0.726
0.008	0.992	1.000	100.000	0.382	0.873	0.043	0.708	0.957	0.128	0.741
0.008	0.992	1.000	105.000	0.431	0.814	0.077	0.704	0.923	0.186	0.760
0.008	0.992	1.000	110.000	0.477	0.667	0.124	0.721	0.876	0.333	0.710
0.008	0.992	1.000	115.000	0.522	0.451	0.196	0.744	0.804	0.549	0.610
0.008	**0.992**	**1.000**	**120.000**	**0.564**	**0.314**	**0.239**	**0.761**	**0.761**	**0.686**	**0.534**
0.008	0.992	1.000	125.000	0.605	0.235	0.340	0.721	0.660	0.765	0.572
0.008	0.992	1.000	130.000	0.644	0.177	0.397	0.698	0.603	0.824	0.578
0.008	0.992	1.000	135.000	0.682	0.157	0.488	0.641	0.512	0.843	0.654
0.008	0.992	1.000	140.000	0.719	0.137	0.617	0.558	0.383	0.863	0.766
0.008	0.992	1.000	145.000	0.754	0.137	0.703	0.498	0.297	0.863	0.857
0.008	0.992	1.000	150.000	0.788	0.128	0.785	0.445	0.215	0.873	0.931
0.008	0.992	1.000	155.000	0.820	0.128	0.818	0.422	0.182	0.873	0.970

As is displayed in Tables 7.1 to 7.4, parametric regressions show better prediction performance than the Cox's approaches. This is supported by the following facts: first, with similar specificities, the accuracies and the sensitivities in parametric regressions are higher than those in the Cox's regressions. For example, given that the specificity is about 0.8, in parametric regressions the accuracy is about 0.72 (Table 7.3 and 7.4) and only about 0.62 in two Cox's regressions (Table 7.1 and 7.2). Similarly, when the specificity is around 0.8, the sensitivity in parametric regressions ranged from 0.33 to 0.54 (Table 7.3 and 7.4) while only 0.19 to 0.28 in two Cox's regression models (Table 7.1 and 7.2). Second, the parametric models achieve much lower misclassification cost than the Cox models do. As is shown from Tables 7.1 to 7.4, the total misclassification cost in the Weibull and the Loglogistic parameter regressions equal to 0.535 and 0.534 respectively, but these numbers hike up to 0.803 and 0.816 in the Cox's PH regression and the Cox's TVC model correspondingly. Table 7.5 summarizes the comparison among prediction performances under the optimal cut-off point of each model.

Table 7.5 The Comparison of the Prediction Results under the Optimal Cut-off Point among Survival Regressions

Model	Cut-off Value	Accuracy	Specificity	Sensitivity	Prob (II given	Prob (I given	Misclassification Cost
the Cox's PH Regression	0.57	0.54	0.41	0.79	0.21	0.59	0.80
the Cox's TVC Regression	0.39	0.65	0.99	0.04	0.96	0.01	0.82
the Weibull Regression	0.61	0.75	0.71	0.75	0.25	0.29	0.53
the Log-logistic Regression	0.56	0.76	0.76	0.69	0.31	0.24	0.53

Although lots of literature indicated that the Cox's model is superior to other survival models (i.e., Lane et al. 1986), in this study the parametric models defeat the Cox's models in terms of the prediction capability. This is because severely tied bankruptcy events are appeared in the data. For example, large amount of bankruptcies happened at a same time when their survival time ranges from 3 years to 5 years. Meyer (1995) shows that when many failures occur at the same time the likelihood of the Cox's PH model becomes intractable, which may result in poor and unreliable estimation and prediction results. Even though the Efron's method in adjusting ties of survival times has been implemented, the negative impact of the tied bankrupt data on the prediction performance probably couldn't be completely removed.

7.3 Prediction Performance and Model Comparison among all Neural Networks

7.3.1 Subjectively Determined Cut-off Value in Networks

The structure of a neural network is determined by the number of the hidden nodes and the value of the decay parameter. The optimal network structure is usually obtained by minimizing the misclassification cost (Hopwood et al., 1989; Koh 1992; Coats and Fant, 1993; Etheridge et al., 2000). However, this technique still can not be incorporated into any of survival neural network presented by now. The lack of the capability handling the misclassification cost is a limitation for survival neural networks approaches (Ripley 1998). Since the optimal cutoff point is also based on the minimum misclassification cost, the lack of misclassification cost function in survival ANNs results in the subjective determination of the optimal network structure as well as the optimal cutoff points.

Given that a cut-off value is chosen, three prediction measurement ratios, i.e., accuracy, specificity, and sensitivity, can be calculated. As discussed before, a high sensitivity is much more attractive than a high specificity for the investors, and a higher sensitivity can be achieved at the expense of the specificity by lowering the cut-off value. Therefore, in determining the optimal cutoff value a balance of sensitivity and specificity is properly considered with allocating a larger weight on sensitivity.

After the accuracy, specificity and sensitivity values are obtained, the optimal structure of the network is the one which achieves the highest sensitivity value. In this study, the number of hidden nodes change from 1 to 10 and the decay parameters change from 0.01 to 0.5, with the higher hidden node number corresponding to the higher decay parameter. To enhance the stability of the model outputs, averagely 7 to 10 different structures are tested for each model and the one which obtains the highest sensitivity is chosen as the optimal structure. Then the outputs based on this structure are used as the representative outputs and are compared with the representative outputs of the other models.

7.3.2 Comparison between Standard ANN and PLAAN

As is shown in Chapter 6, the standard ANN and PLAAN share many similarities in their structures. The similarities in structures explain the similarities in their prediction performance. As shown in Tables 7.6 and 7.7, these two networks have close accuracies, specificities and sensitivities.

Table 7.6 Prediction Performance of the Standard ANN

Hidden Nodes	Weight Decay	Accuracy	Specificity	Sensitivity
1	**0.01**	**0.92**	**0.95**	**0.87**
2	0.01	0.92	0.94	0.86
3	0.01	0.91	0.94	0.84
3	0.1	0.93	0.98	0.83
5	0.01	0.90	0.94	0.83
5	0.3	0.93	0.98	0.82
10	0.5	0.93	0.98	0.82
average		0.92	0.96	0.84

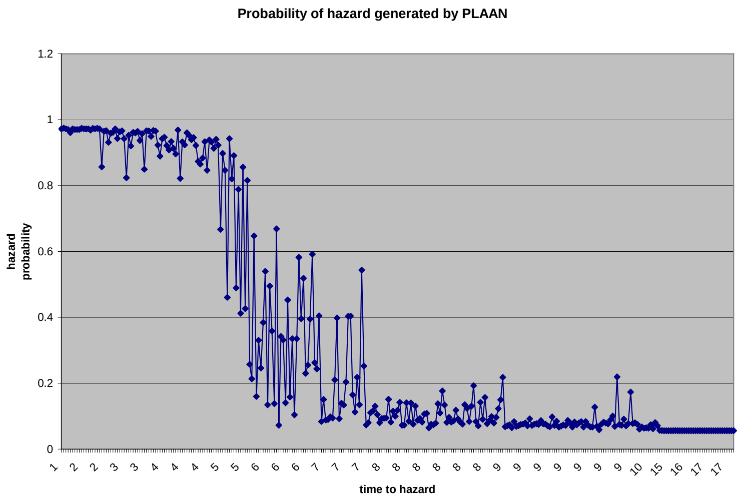
Table 7.7 Prediction Performance of PLANN

Hidden Nodes	Weight Decay	Accuracy	Specificity	Sensitivity
1	**0.075**	**0.93**	**0.96**	**0.85**
2	0.075	0.93	0.98	0.83
3	0.075	0.93	0.97	0.83
3	0.1	0.93	0.97	0.84
5	0.1	0.92	0.96	0.83
5	0.3	0.93	0.97	0.83
10	0.5	0.93	0.98	0.83
average		0.93	0.96	0.84

The major difference between PLAAN and standard ANN is their likelihood functions. As discussed in Chapter 4, the standard ANN achieves its optimal solution by minimizing the sum of squares of difference between actual values and predicted values, and during this procedure, gradient descent techniques like back-propagation is employed. PLAAN instead sets an entropy error function as its likelihood function and minimize it by using quasi-Newton algorithm to attain the optimal solution. Therefore, the standard neural network is just a non-linear generalization of logistic regression whose output is only a proxy of the actual status dummy, while PLAAN can produce the conditional bankruptcy function which demonstrate a successful combined analysis of survival analysis and neural networks.

Figure 7.5 shows the predicted conditional bankruptcy probabilities generated in PLAAN. Different from all other models, the probabilities produced by PLAAN is the probability of bankrupcy for a firm at its age. The negative slope here indicates that the younger firms take much higher risk of bankruptcy than those older ones, which is consistent with the finding in Chapter 5.

Figure 7.5 The Probabilities of Bankruptcy Generated by PLAAN



Since the outputs of standard ANN is just the “approximators” of the status dummy, no plot is generated for this ANN.

7.3.3 Comparison among the Cox’s Survival ANN and the Parametric Survival ANNs

The outputs generated from the Cox’s survival ANN and the parametric survival ANNs are not the probabilities since most of them are larger than 1.

Particularly, the outputs of the Cox's proportional survival ANN are called Cox's scores. Conversion functions are needed to convert these output values into predicted survival probabilities. In this study, the conversion function is provided by Prof. Ripley (1998). For the parametric survival ANNs, the parameter λ in $f(t_i)$ and $S(t_i)$ can be replaced by $\exp(y)$ since y equals to $\log\lambda(x)$ (refer to sector 4.2.3).

The predicted survival probabilities in the Cox's survival ANN and the two parametric survival ANNs are presented in figures 7.6 to 7.8. All three curves exhibit positive slope against survival time, however, figure 7.6 and 7.7 show very small interval (from about 0.9 to 1). The tight intervals implies that probably the prediction performances of the Cox's and the Weibull survival ANNs may not be reliable, since it will be difficult to find a good cutoff value within a very tight interval, with most points above correctly being classified as non-bankrupt firms and points below correctly being classified as bankrupt ones. Besides, it is also problematic that all bankrupt firms still could achieve high (above 0.9) survival probabilities at the last year before bankruptcies.

Figure 7.6 The Probabilities of Survival Generated by the Cox's Survival ANN

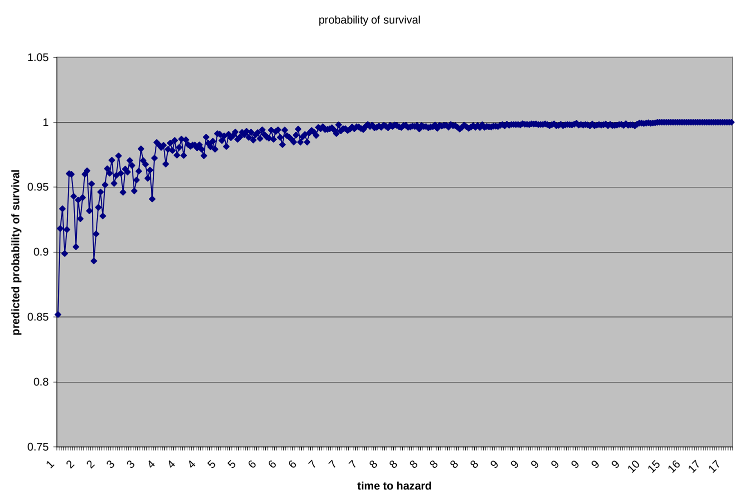


Figure 7.7 The Probabilities of Survival Generated by the Parametric Weibull Survival ANN

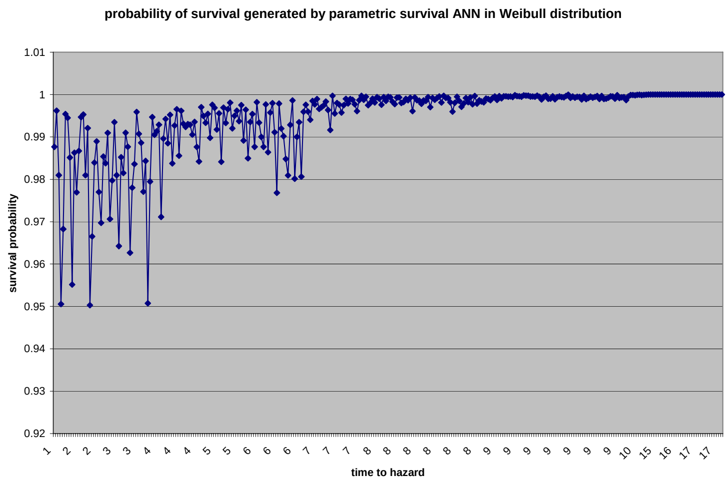
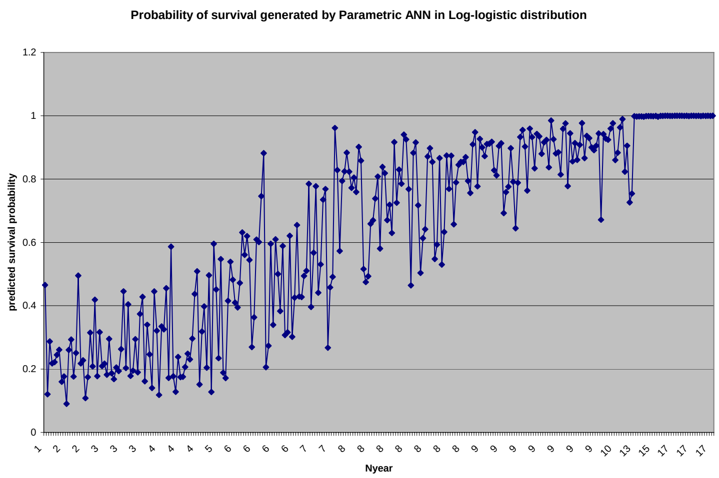


Figure 7.8 The Probabilities of Survival Generated by the Parametric Log-logistic Survival ANN



On the opposite, the survival probability curve produced by the log-logistic parametric survival ANN signals a good prediction performance: the survival probabilities range widely from 0.1 to 1 with older firms getting higher values and younger firms catching lower values. Consequently, it is relatively easy to find a cutoff value to realize a good classification. However, it is also noted that the volatility of the predicted survival probabilities becomes larger since the band of points along the upward slope becomes wider in this model.

Table 7.8 Prediction Performance of the Cox's Survival ANN

Hidden Nodes	Weight Decay	Accuracy	Specificity	Sensitivity
1	0.01	0.87	0.96	0.61
1	0.075	0.87	0.95	0.63
2	0.01	0.87	0.96	0.60
2	0.075	0.89	0.96	0.66
2	0.2	0.87	0.89	0.74
3	0.1	0.88	0.98	0.60
3	0.2	0.87	0.89	0.74
3	0.3	0.89	0.98	0.63
5	**0.2**	**0.87**	**0.90**	**0.74**
5	0.3	0.89	0.98	0.63
5	0.5	0.53	0.59	0.34
10	0.3	0.89	0.98	0.63
10	0.5	0.60	0.74	0.27
average		0.83	0.91	0.60

Table 7.9 Prediction Performance of the Parametric Weibull Survival ANN

Hidden Nodes	Weight Decay	Accuracy	Specificity	Sensitivity
1	**0.01**	**0.77**	**0.73**	**0.77**
2	0.01	0.77	0.73	0.76
2	0.075	0.61	0.83	0.12
2	0.2	0.75	0.96	0.25
3	0.1	0.46	0.32	0.70
3	0.2	0.81	0.81	0.72
3	0.3	0.81	0.88	0.59
5	0.2	0.81	0.81	0.72
5	0.3	0.81	0.88	0.59
5	0.5	0.78	0.93	0.39
10	0.3	0.81	0.88	0.59
10	0.5	0.82	0.88	0.64
average		0.75	0.81	0.57

Table 7.10 Prediction Performance of the Parametric Log-logistic Survival ANN

Hidden Nodes	Weight Decay	Accuracy	Specificity	Sensitivity
1	0.01	0.82	0.82	0.74
2	0.01	0.87	0.93	0.67
2	0.075	0.82	0.85	0.68
2	0.1	0.69	0.78	0.39
3	0.1	0.80	0.88	0.56
3	0.2	0.81	0.90	0.55
3	0.3	0.81	0.80	0.75
5	0.2	0.81	0.80	0.75
5	0.3	0.82	0.90	0.65
5	**0.5**	**0.81**	**0.78**	**0.79**
10	0.3	0.82	0.87	0.65
10	0.5	0.81	0.78	0.79
average		0.81	0.84	0.67

Tables 7.8 to 7.10 show the prediction performance of the Cox survival ANN, the Weibull survival ANN and the log-logistic survival ANN. Among these three models, the Weibull survival ANN is the worst of them since its accuracy, specificity and sensitivity are all lower than the corresponding values of the other two models. However, the comparison between the Cox's survival ANN and the log-logistic survival ANN is not that straightforward because the Cox's ANN got the higher accuracy and specificity while the log-logistic ANN got the higher sensitivity. Given the facts that the sensitivity is the most important one among the three ratios, and that

the survival probabilities generated by the Cox's ANN are concentrated within an extremely tight interval(from about 0.9 to 1), the output from log-logistic survival ANN probably should be considered better than that of the Cox's survival ANN.

7.3.4 Comparison among the Networks

Table 7.11 summarizes the prediction performances under the optimal structure of each model. Among all the ANNs, the standard ANN and PLAAN demonstrated best prediction performance. All three ratios of standard ANN and PLAAN got higher values than those of the other models. When the two parametric survival ANNs are compared with each other, the log-logistic model outperform the Weibull model in all three ratios. As discussed in Chapter 6, the log-logistic model fits the data better than the Weibull model in estimation, which may help to explain the better performance of log-logistic survival ANN. As discussed in previous Section (7.3.3), the results of comparing the Cox's ANN and the log-logistic ANN are mixed, and the log-logistic ANN should be considered as better than the Cox's ANN.

Table 7.11 Comparisons among Neural Networks

Model	Accuracy	Specificity	Sensitivity
Standard Neural Network	0.92	0.95	0.87
PLAAN	0.93	0.96	0.85
the Cox's survival ANN	0.87	0.90	0.74
the Weibull survival ANN	0.77	0.73	0.77
the Loglogistic survival ANN	0.81	0.78	0.79

Although the survival ANNs (except the PLAAN) generate lower prediction performance than the standard ANN, this inferiority can be compensated by the richer information the survival ANNs provided. As discussed before, theoretically the survival ANNs can generate the conditional survival probabilities of firms at any time during their life interval, which means, the survival ANNs technically can provide the survival probabilities in all the years of a firm's life when given enough information. This information is quite precious for investors and the researchers, especially under highly volatile markets. However, the standard ANN can only provide an "approximator" of survival at one point during the firm's life. Therefore, it is worthwhile to employ these survival networks, as long as their inferior prediction performance is justified by their richer information provided.

7.3.5 Comparison among the linear regressions and the networks

Table 7.12 shows the comparison among six models discussed in this study, namely, the Cox's PH regression, the Cox's survival ANN, the parametric Weibull regression, the Weibull survival ANN, the parametric log-logistic regression and the log-logistic survival ANN. The comparison among these six models is especially meaningful since each two of them are generated from a same survival analysis method, with the only difference that the regression uses the linear combination of covariates while the network counterparty uses the nonlinear forms.

Table 7.12 Comparisons among Regressions and Neural Networks

Model	Accuracy	Specificity	Sensitivity
the Cox's regression	0.54	0.41	0.79
the Cox's survival ANN	0.87	0.90	0.74
the Weibull regression	0.75	0.71	0.75
the Weibull survival ANN	0.77	0.73	0.77
the Loglogistic regression	0.76	0.76	0.69
the Loglogistic survival ANN	0.81	0.78	0.79

As is shown in table 7.12, all the networks outperform their linear counterparties in terms of prediction performance. Specifically, under the Cox's proportional hazard framework, the network gets a much higher accuracy (0.87) than the regression (0.54); under the Weibull and log-logistic parametric hazard framework, the networks did not get such a distinct superiority in their performance, but they still outperform the regressions in all three ratios.

As discussed in Chapter 4, the superiority of neural network to the linear regression comes from its universal approximation capability in data, which even do not require discovering covariates' exact function form. Besides, in networks, the obstacles like multi-colinarity are no more a problem, which instead harms the linear regression a lot. As observed from the estimation results in linear survival regressions, most of covariates are not significant individually but the whole set of covariates is highly significant in general tests. This suggests a serious multi-colinarity in data. This property of data actually penalizes the linear regression which instead helps to emphasize the comparative advantage of networks.

CHAPTER 8 CONCLUSIONS AND LIMITATIONS

This chapter summarizes the contributions as well as the limitations of this study. The potential developments of future research are also discussed.

8.1 Contributions

This study has three contributions to the development of survival analysis in corporate bankruptcy prediction.

First, four survival estimation regressions are applied and their estimation and prediction results from compared. Rather than comparing one survival regression with a MDA or Logit model which has been done a lot in the previous literature, the Cox's proportional hazard regression, the Cox's TVC model, the Weibull parametric survival regressions and the log-logistic survival regression are compared with each other based on their estimation and prediction results. In addition, the diagnostic tests on the proportionality assumption and the linear covariates assumption are conducted for the Cox's PH regressions. For the parametric survival regressions, the hazard shapes are compared with the hazard curve generated directly with the data by non-parametric method. Such comprehensive comparisons among various survival regressions are seldom observed in previous studies which make the first contribution of this study.

Second, four survival ANNs, namely, PLAAN, the Cox's survival ANN, the Weibull survival ANN and the log-logistic survival ANN, are firstly introduced into the bankruptcy prediction. Although both the survival linear regressions and the standard ANNs have won lots of success in bankruptcy prediction, to date, little work has been done in combining these two tools together to provide better bankruptcy prediction. Based on the development of survival ANNs in bio-medical science research, this study successfully implemented four survival ANNs in bankruptcy prediction. According to prediction output of various survival ANNs, curves of survival probabilities and conditional hazard probabilities are plotted. The optimal structure in each type of survival ANN is determined based on the model's prediction performance valued by the total accuracy, specificity and sensitivity.

Third, in the comparisons of prediction results between linear and nonlinear survival models based on the prediction outputs using same set of data, several interesting findings are captured: first of all, although the Cox's proportional hazard model is the most popular model employed in the literature, it does not certainly provide the best prediction results all the time. Especially when there are too many ties in the data, the prediction performance of parametric method will greatly outperform the Cox's model. Besides, linear restriction in model specification probably seriously weaken the model's prediction capability, which is proven by the remarkable increase in prediction performances of the survival ANNs when compared with their counter parties in linear approaches. The last but not the least, compared with the standard ANN, the survival ANNs incline to improve the information richness at the expense of prediction accuracy. As is shown in the Chapter 7, although survival ANNs can provide the additional timing information of bankruptcies, they are usually inferior to the standard ANN in terms of prediction capability. However, all above findings are just indicative; more researches need to be conducted to get more thorough understandings about the application of survival ANNs in bankruptcy prediction.

8.2 Limitations

8.2.1 Limitations of the Survival ANNs

Although the survival ANNs have shown stronger prediction capability than their linear counterparties, they are subject to limitations. As neural networks, survival ANNs bear typical limitations of neural networks, such as over-fitting and lack of the transparence in interpreting the effects of covariates. Besides, they are also associated with the two disadvantages particular to survival ANNs.

First, the survival ANNs can not integrate misclassification cost into their modeling procedure, which leaves the choosing of the optimal structure of network as a subjective decision. As is shown, for each network in survival ANNs, changes in the number of hidden nodes or in the weight decay parameter will both result in different output values. Even under the same combination of hidden nodes and weight decay, the network still may generate different output values because of the random initial values used when maximizing the likelihood function. If the misclassification cost is calculated for each set of outputs as in the linear models, the whole procedure of determining the optimal structure will become extremely time-consuming. Therefore, in this study subjective judgment is applied in deciding the optimal structure for each survival ANN.

Second, almost all survival ANNs (except PLAAN) suffer by lower prediction performance ratios compared with the standard ANN, especially for sensitivity value. Ripley (1998) admits that the Cox's and parametric survival ANNs usually produce poor sensitivity, since these networks are more successful in identifying survived records rather than the non-survived records. This behavior is undesirable since a high sensitivity is very critical in bankruptcy prediction. Unfortunately, as commented by Ripley (1998), this limitation is "difficult to eradicate".

In addition, researchers have pointed out that bankruptcy prediction using MDA, logistic, survival regressions all hold a common problem: data selection bias. Generally, characteristics of firms change from year to year and the process of bankruptcy varies from period to period. However, all the above-mentioned models consider only one set of explanatory variables. Researchers have to select when to observe each firm's characteristics. Usually, only the data in the year before bankruptcy for each firm are chosen. This introduces selection bias into their estimates (Shumway, 2001). Obviously same limitation is shared by all survival ANNs as well.

8.2.2 Limitations of this Study

In addition to above limitations, there are some other issues should be considered to avoid a misuse of conclusions or findings of this study.

First, all conclusions and findings are just based on current dataset. That said, when another set of data are employed, different conclusions may appear especially in terms of the performance ranking orders within groups of linear models and nonlinear models.

Second, in this study the final dataset after all data cleansing work contains about 800 records, which maybe is relatively insufficient to generate highly robust outputs for models with 13 covariates. Given the limited data resource, the issue of "the curse of dimensionality" is a drawback that is hardly to avoid. However, according to Donoho (2000), there are also "blessings" associated with dimensionality, one of which is related to a mathematical term "concentration of measure". Donoho comments that it could say that in many cases, there are really "few things that matter" and that the function will be constant on most of the space. Therefore, it is still possible to do statistics in a meaningful way under the circumstance of high number of dimensions.

Third, this study didn't really analyze the data features from the angle of the business analysis, and hence could not provide more profound and specific reasons on the superiority of the survival ANNs in bankruptcy analysis.

Due to all above limitations specific to this research as well as the common limitations of survival ANNs, the finding that survival ANNs can outperform survival regressions is just an indicative one. More work has to be conducted to validate the significance of the above finding.

8.3 Future Developments

As a pioneer work of applying non-linear survival analysis in bankruptcy prediction, this study is open to many further developments.

One development is to apply different data to verify whether the survival ANNs can consistently outperform their linear counterparties. Both real data from different industries and countries as well as simulated data with different distribution assumptions should be used. Different sets of covariates also should be applied. Tremendous efforts are required in this approach, but only by doing so, the question of how significant that the survival ANNs are superior to survival regressions can be well addressed.

Besides of that, other developments can be implemented to improve the performance of survival ANN. For example, to incorporate the misclassification cost into the likelihood function of survival ANNs. If a survival ANN can optimize the network's structure and calculate the misclassification cost simultaneously during its log-likelihood function maximization, subjective decisions in determining the optimal structure and minimum misclassification cost can be ultimately excluded from modeling of survival ANNs. Here Tam and Kiang (1992)'s solution in incorporating the misclassification cost into the likelihood function of a standard neural network points a good direction where the further study can proceed.

REFERENCE

Afifi, A. and V. Clark, 1996, Computer-aided Multivariate Analysis, Chapman & Hall;

Allison, P.D., 1995, "Survival Analysis Using SAS, A Practical Guild", SAS Publishing;

Altman, E. I., 1968, "Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy", *Journal of Finance* 23, 589-609;

Altman, E.I., R.B.AVERY, R.A.Eisenbeis and J.F.Sinkey, Jr., 1981, Application of classification techniques in business, banking, and finance, Contemporary studies in economic and financial analysis, Vol.3, (JAI press, Greenwich, CT);

Altman, E. I. 1984, "a Further Empirical Investigation of the Bankruptcy Cost Question", *The Journal of Finance*, 4, 1067-1089;

Altman, E.I., M. Giancarlo and F. Varetto, 1994, "Corporate Distress Diagnosis: Comparisons Using Linear Discriminant Analysis and Neural Networks (the Italian Experience), *Journal of Banking and Finance*, 18, 505-529;

Audretsch, D.B. and T. Mahmood, 1995, "New Firm Survival: New Results Using a Hazard Function", *The Review of Economics and Statistics*, Vol.77, 97-103;

Baesens, B., Gestel, T.V., Stepanova, M. and J.Vanthienen, 2004, “Neural Network Survival Analysis for Personal Loan Data”, working paper;

Bandopadhyaya A., 1994, “An Estimation of the Hazard rate of Firms under Chapter 11 Protection”, *The Review of Economics and Statistics*, Vol.76, 2, 346-350;

Bell, T.B., and R.H. Tabor, 1991, “Empirical Analysis of Audit Uncertainty Qualifications”, *Journal of Accounting Research*, 29, 350-370;

Berry, M.J.A., and Linoff, G. 1997, Data Mining Techniques, NY: John Wiley & Sons;

Biganzoli E., P. Boracchi, L.Mariani and E. Marubini, 1998, “Feed Forward Neural networks for the Analysis of Censored Survival Data: A Partial Logistic Regression Approach”, *Statistic in Medicine*, 17, 1169-1186;

Biganzoli E., P. Boracchi and E. Marubini, 2002, “A General Framework for Neural Network Models on Censored Survival Data”, *Neural Networks*, 15, 209-218;

Bishop, C. 1995, Neural Networks for Pattern Recognition. Oxford: University Press.

Blum, A. 1992, Neural Networks in C++, NY: Wiley.

Boger, Z., and Guterman, H. 1997, "Knowledge extraction from artificial neural network models," IEEE Systems, Man, and Cybernetics Conference, Orlando, FL.

Booth, P.J., 1983, "Decomposition Measures and the Prediction of Financial Failure", *Journal of Business Finance & Accounting*, 67-82;

Breslow, N.E., 1974, "Covariance Analysis of Censored Survival Data", *Biometrics*, 30:89-99;

Broomhead, D.S., D. Lowe, 1988, "Multivariate Functional Interpolation and Adaptive Networks", *Complex Systems*, 2:321-355;

Brown, S.F., A. Branford and W. Moran, "On the Use of Artificial Neural Networks for the Analysis of Survival Data", *IEEE Transactions on Neural Networks*, 8:1070-1077;

Burke, H.B., 1994, "Artificial Neural Networks for Cancer Research: Outcome Prediction", *Seminars in Surgical Oncology*, 10, 73-79;

Callan E., 1999, The Essence of Neural Networks, Prentice Hall Europe;

Cantor. Alan.B., 2003, Survival Analysis Techniques for Medical Research, SAS Publishing;

Chen, K.C.W. and B. K. Church, 1992, "Default on Debt Obligations and the Issuance of Going-Concern Opinions", Auditing: A Journal of Practice and Theory, 11(2), 30-49;

Cheng P.Y.K, 2002, "Predicting Bank Failures: A Comparison of Cox Proportional Hazards Model and the Time Varying Covariates Model", PhD dissertation;

Coats, P.K., and L.F. Fant, 1993, "Recognizing Financial Distress Patterns Using a Neural Network Tool", *Financial Management*, 22(3), 143-154;

Collett, D., 2003, *Modelling survival data in medical research*, Chapman & Hall;

Cox, D.R. and E.J. Snell, 1968, "A General Definition of Residuals (with Discussion)", *Journal of the Royal Statistical Society* B, 30: 248-275; -

Cox, D.R., 1972, "Regression models and life-tables (with discussion)", *Journal of the Royal Statistical Society* B, 34, 187-220;

De Laurentiis, M., P.M. Ravdin, 1994, "A Technique for Using Neural Network Analysis to Perform Survival Analysis of Censored Data", *Cancer Letters*, 77, 127-138;

Dhar V. and R. Stein, 1997, "Seven Methods for Transforming Corporate Data into Business Intelligence", Prentice-Hall, Inc.;

Dominiak, G. F., J. G. Louderback, 1985, Managerial Accounting, fourth edition, Kent Publishing, U.S.A;

Donoho, L.D., "High Dimensional Data Analysis: the Curses and Blessings of Dimensionality", 2000;

Dopuch, N., R. W. Holthausen and R.W. Leftwith, 1987, "Predicting Audit Qualifications with Financial and Market Variables", *The Accounting Review*, July 1987, 431-454;

Duffie, D. and K. Wang, 2003, "Multi-period Corporate Failure Prediction with Stochastic Covariates", working paper;

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

Eisenbeis, R.A., 1977, "Pitfalls in the Application of Discriminant Analysis in Business, Finance, and Economics", *Journal of Finance*, June, 875-900;

Etheridge H.L., Sriram, R.S., 1997, "A Comparison of the relative costs of Financial Distress Models: Artificial Neural Networks, Logit and Multivariate Discriminant Analysis", *Intelligent Systems in Accounting, Finance and Management*, 6, 235-248;

Etheridge H.L., Sriram, R.S. and Kathy Hsu H.Y., 2000, "A Comparison of Selected Artificial Neural Networks that Help Auditors Evaluate Client Financial Viability", *Decision Sciences*, 31:531-550;

Faraggi, D. and R. Simon, 1995, "A Neural Network Model for Survival Data", *Statistics in Medicine*, 14:73-82;

Farewell, V. and Prentice, R., 1980, "The Approximation of Partial Likelihood with Emphasis on Case-control Studies", Biometrika, 67:273-278;

Fisher L.D., Lin D.Y., 1999, "Time-Dependent Covariates in the Cox Proportional-Hazards Regression Model", *Annual Review of Public Health*.20:145-57;

Frank E. Harrell, 2001, "Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis", Springer;

Geman, S., Bienenstock, E. and Doursat, R. 1992, "Neural Networks and the Bias/Variance Dilemma", *Neural Computation*, 4, 1-58.

Gentry, J.A., Newbold, P. and D.T.Whiteford, 1987, "Funds Flow Components, Financial Ratios, and Bankruptcy", *Journal of Business Finance & Accounting*, 595-606;

Grambsch, P., Therneau, T.M., 1994, "Proportional Hazards Tests and Diagnostics Based on Weighted Residuals", *Biometrika*, 81:515:526;

Green, William H., 2003, *Econometric Analysis, fifth edition*, Prentice Hall;

Harhoff, D., K. Stahl and M. Woywode, 1998, "Legal Form Growth and Exit of West-German Firms: Empirical Results for Manufacturing, Construction, Trade and Service Industries", *Journal of Industrial Economics* 46, 453–488;

Harrell, F., 1986, "The PHGLM Procedure. SAS Supplemental Library User's Guide", Version 5, Cary, NC: SAS Institute, Inc.;

Henebry, K.L., 1996, "Do Cash Flow Variables Improve the Predictive Accuracy of Cox Proportional Hazards Model for Bank Failure?", *The Quarterly Review of Economics and Finance*, Vol.36, 3, 39-409;

Hillegeist, S.A., Keating, E.K., Cram, D.P., and K.G.Lundstedt, 2004, "Assessing the Probability of Bankruptcy", *Review of Accounting Studies*, 9, 5-34;

Hopwood, W., McKeown, J., and J. Mutchler, 1989, "A Test of the Incremental Explanatory Power of Opinions Qualified for Consistency and Uncertainty", *The Accounting Review*, 64(1), 28-48;

Hosmer, D.W., S.Lemeshow, 1999, Applied Survived Analysis—Regression Modeling of Time to Event Data, Springer;

Joy, M.O. and J.O.Tollefson, 1975, "On the financial applications of Discriminant Analysis", *Journal of Financial and Quantitative Analysis*, Dec., 723-739;

Kalbfleisch , J.D., R.L. Prentice,1973, "Marginal Likelihood Based on Cox's Regression and Life Model", *Biometrika*, 60:267-278;

Kaplan, E.L. and Paul. Meier, 1958, "Nonparametric estimation from incomplete observations", *Journal of the American Statistical Association*, 53, 457-481;

Karels, G.V. and A.J. Prakash, 1987, "Multivariate Normality and Forecasting of Business Bankruptcy", 573-592;

Keasey, K. and R.Watson, 1986, "Current Cost Accounting and the Prediction of Small Company Performance", *Journal of Business Finance & Accounting*, 51-70;

Kennedy R.L., Y.Lee, B.V.Roy, C..D.Reed and D.R.P.Lippman, 1998, Solving Data Mining Problems--Through Pattern Recognition, Prentice-Hall, Inc.

Klein, J.P., M.L., Moeschberger, 1997, "Survival Analysis, Techniques for Censored and Truncated Data", Springer;

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

Koh, H.C. and S.S.Tan, 1999, "A Neural Network Approach to the Prediction of Going Concern Status", *Accounting and Business Research*, Vol.29, 3, 211-216;

Laitinen, E.K., 1994, “Traditional Versus Operating Cash Flow in Failure Prediction”, *Journal of Business Finance and Accounting*, Vol.21, 2, 195-217;

Lancaster, T., 1990, The Econometric Analysis of Transition Data, New York: Cambridge University Press;

Lane, W.R., S.W. Looney, and J.W.Wansley, 1986, An Application of the Cox Proportional Hazards Model to Bank Failure, *Journal of Banking and Finance* 10, 511-531;

Lapuerta, P., Azen, S.P. and L.LaBree, 1995, “Use of Neural Networks in Predicting the Risk of Coronary Artery Disease”, Computers and Biomedical Research, Vol.28, 38-52;

Lawrene, E.C. and R.M. Bear, 1986, “Corporate Bankruptcy Prediction and the Impact of Leases”, *Journal of Business Finance and Accounting*, 571-585;

Lee, S.H. and J.L.Urrutia, 1996, Analysis and Prediction of Insolvency in the Property-Liability Insurance Industry: A comparison of Logit and Hazard Models. *The Journal of Risk and Insurance* 63, 121-130;

Liang, T.P., Chandler, J.S., Jan, I., & Roan, J., 1992, “An empirical Investigation of Some Data Efects on the Classification Accuracy of Probit, ID3, and Neural Networks”, *Contemporary Accounting Research*, 9(1), 306-328;

Mani, D.R., J. Drew, A.Betz and P. Datta, 1999, "Statistics and Data Mining Techniques for Life-time Value Modeling", In Proceedings of the Fifth ACM SIGKDD *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 94-103, San Diego, CA, U.S.;

Mariani, L., Coradini, D., Biganzoli, E., Marubini, E., Pilotti, S., Salvadori, B., Silvestrini, R., Veronesi, U., Zucali, R., and F. Rilke, 1997, "Prognostic Factors for Metachronous Contralateral Breast Cancer: A Comparison of the Linear Cox Regression Model and its Artificial Neural Network Extension", *Breast Cancer Research and Treatment*, 44, 167-178;

Marais, M.L., J.M. Patell and M.A. Wolfson, 1984, "The experimental Design of Classification Models: An Application of Recursive Partitioning and Bootstrapping to Commercial Bank Loan Classifications", *Journal of Accounting Research*, 22, 87-114;

Mathieson, M.J., 1998, "Ordinal Models and Predictive Methods in Pattern Recognition", Ph.D. thesis, University of Oxford;

Mensah, Y.M., 1983, "The Differential Bankruptcy Predictive Ability of Specific Price Level Adjustments: Some Empirical Evidence", *The Accounting Review*, 228-246;

Merton, R., 1974, "On the Pricing of Corporate Debt: the Risk Structure of Interest Rates", *The Journal of Finance*, 29, 499-470;

Meyer B.D., 1995, "Semi-parametric Estimation of Hazard Models", NBER working paper;

Molina, C.A., 2002, "Predicting Bank Failures Using a Hazard Model: the Venezuelan Banking Crisis", *Emerging Markets Review*, 3, 31-50;

Moody, J., C. Darken, 1989, "Fast Learning in Networks of Locally-Tuned Processing Units", *Neural Computation*, Vol 1, 2:281-294;

Odom, M.D. and R. Sharda, 1990, "A Neural Network Model for Bankruptcy Prediction", *in Proceedings of the International Joint Conference on Neural Networks*;

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

Ohno-Machado,L, Walker, M.G. and M.A. Musen, 1995, "Hierarchical Neural Networks for Survival Analysis", *Medinformation 95 Proceedings*, 828-832;

Ohno-Machado, L., 1996, "Sequential Use of Nerual Networks for Survival Prediction in Aids", *Journal of the American Medical Informatics Association*, 3:170-174;

Ohno-Machado, L, and M.A. Musen, 1997, "Modular Neural Networks for Medical Prognosis: Quantifying the Benefits of Combining Neural Networks for Survival Prediction", *Connection Science*, 9, 1, 71-86;

Ravdin, P.M., G.M.Clark, 1992, "A Practical Application of Neural Network Analysis for Predicting Outcome of Individual Breast Cancer Patients", *Breast Cancer Research and Treatment*, 22, 285-293;

Peel, M.J., D.A.Peel and PLF.Pope, 1986, "Predicting Corporate Failure—some Results for the UK Corporate Sector", *Omega International Journal of Management Science*, 5-12;

Richardson, F.M. and L.F.Davidson, 1983, "An exploration into Bankruptcy Discriminant Model Sensitivity", *Journal of Business Finance & Accounting*, 195-208;

Ripley, B.D., 1994, "Neural Networks and Flexible Regression and Discrimination", in Mardia, K.V.(eds), *Statistics and Images 2. Advances in Applied Statistics, 2*, Carfax, Abingdon, pp.39-57;l

Ripley, B.D., 1996, Pattern Recognition and Neural Networks, Cambridge: Cambridge University Press., 14, 34, 36, 51;

Ripley, B.D and R.M. Ripley, 2002, "Neural Networks as Statistical Methods in Survival Analysis", Artificial Neural Networks: Prospects for Medicine, Landes Biosciences Publishers;

Ripley, R.M., 1998, "Neural Network Models for Breast Cancer Prognosis", Ph.D. thesis, University of Oxford;

Ripley, R.M., 2001, "Survnnet: Neural Network Survival Models";

Rumelhart, D.E., G. Hinton and R. Williams, 1986, "Learning Representation by Back-Propagating Errors", *Nature*, 323, 9:533-536;

SAS-Institute Inc., SAS/STAT User's Guide, Version 8, SAS-Institute Inc: Cary NC;

Saunders A., L. Allen, 2002, "Credit Risk Measurement: New Approaches to Value at Risk and Other Paradigms", Second Edition, Wiley Finance;

Schoenfeld D, 1982, "Partial Residuals for the Proportional Hazards Regression Model", *Biometrika*, 69:239-241

Shumway, T. 2001, "Forecasting Bankruptcy More Accurately: A simple Hazard Model", *Journal of Business*, vol.74, no.1;

Street, W.N., 1998, "a Neural Network Model for Prognostic Prediction", *in Proceedings of he Fifteenth International Conference on Machine Learning (ICML)*, 540-546;

Sung T.K., N. Chang and G. Lee, 1999, "Dynamics of Modeling in Data Mining: Interpretive Approach to Bankruptcy Prediction", *Journal of Management Information Systems*, 16, 63-84;

Tam K.Y., M.Y. Kiang, 1990, "Predicting Bank Failures: A Neural Network Approach", *Appl. Artificial Intelligence*, Vol.4, 4, 265-282;

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

Therneau T.M., Foundation M., 1996 , "A package for Survival Analysis in S";

Therneau, T.M., Grambsch, P.M., and Fleming, T.R., 1990, Martingale-based residuals for survival models. *Biometrika* 77:147-160;

Utans, J. and J. Moody, 1991, "Selecting Neural Network Architechtures Via the Prediction Risk: Application to Corporate Bond Rating Prediction", *IEEE*, 35-41;

Venables, W.N., B.D. Ripley, 2002, "Modern Applied Statistics with S", fourth edition, Springer;

Wanas, N., Auda, G., Kamel, M.S., and Karray, F., 1998, "On the optimal number of hidden nodes in a neural network", *Electrical and Computer Engineering, IEEE Canadian Conference*, Volume 2, Issue 24-28, 918-921;

Webb, A.R. and D. Lowe, 1990, "The Optimized Internal Representation of Multilayer Classifier Networks Performs Nonlinear Discriminant Analysis", *Neural Networks*, Vol.3, 367-375;

Whalen, G., 1991, "A Proportional Hazards Model of Bank Failure: An Examination of its Usefulness as an Early Warning Tool", Federal Reserve Bnkd of Cleveland, *Economic Review*, 27, 1, 21-31;

Wheelock D.C. and P.W.Wilson, 1995, "Explaining Bank Failures: Deposit Insurance, Regulation, and Efficiency", *The Review of Economics and Statistics*, 689-700;

Whittred, G. and I. Zimmer, 1984, "Timeliness of Financial Reporting and Financial Distress", *The Accounting Review*, 287-295;