MANAGEMENT SCIENCE AND ENGINEERING
Vol. 4, No. 4, 2010, pp. 24-32
www.cscanada.org

ISSN 1913-0341 [Print]
ISSN 1913-035X [Online]
www.cscanada.net

Studies of Discriminant Analysis and Logistic Regression Model Application in Credit Risk for China's Listed Companies

ZHU Kong-lai[1]

LI Jing-jing[2]

Abstract: With the appearance of listed companies' credit issues and frequent credit crisis, investors are increasingly concerned about credit risk analysis for listed companies. In view of the current development methods of credit risk analysis and the importance of identifying corporate financial risk, this paper designed an effective indicator system and established the credit evaluation models of China's listed companies by taking advantage of their 2009 financial data. Combined with the reality of China's listed companies, we use the established models to discriminate and analyze. The result of empirical research on the credit risk analysis for listed companies is that Logistic regression model is superior to discriminant analysis model.

Key words: Credit Risk; Discriminant Analysis; Logistic Regression Model; Principal Component Analysis

1. INTRODUCTION

Listed companies are the cornerstone of the securities market. Their behavior and financial situation affect the securities markets and the interests of investors directly. Since Shanghai Stock Exchange was established on December 19, 1990, China's securities market expanded rapidly and has become an important part of our national economy. In recent years, with the expansion of the scale securities market, listed companies are exposed more and more problems. Unsound regulation of the market itself and immature supervision give some listed companies opportunities to release false information maliciously in order to achieve their own purposes, such as list reorganize. For various reasons, the amounts of listed companies dropping into financial crisis are rapidly increasing and the ability to resist risks generally reduced. How to objectively evaluate a company's financial operating conditions, reveal the potential risks and indicate the crisis before the enterprise has not yet developed to the worsening financial situation is particularly important. Because of this way, it's possible to give the operator advance warning and avoid severe financial crisis happening. In view of the current development methods of credit risk analysis and the importance of identifying corporate financial risk, this paper will use the discriminant analysis, Logistic regression model and principal component analysis which belongs to the multivariate statistical analysis to establish the credit evaluation models of China's listed companies, discriminate and analyze by taking

[1] University of Jinan, Jinan, 250022, Shandong, China. E-mail:zklai@sina.com.
[2] Shandong University of Finance, Jinan, 250014 Shandong, P.R. China.

*Recevied 13 July 2010; accepted 18 October 2010

advantage of their financial data in 2009. Then study the gap between calculation results and the actual situation. Finally the paper will compare discriminant accuracy and warning effect of the discriminant analysis and Logistic regression model.

2. LITERATURE REVIEW

Altman was the first to apply discriminant analysis to analyze the financial crisis, bankruptcy and default risk. He selected the most influential indicators in the light of the contribution to common factors that is extracted from twenty-two financial ratios by using principal components analysis, applied discriminant analysis to analyze 33 financially distressed companies and 33 nons and built the famous Z-Score model. The model is better for short-term (2 years) predictive ability. Later on he improved the Z-Score model, and then established the ZETA model that was superior to Z-Score Model. (Altman, 1983) Domestic scholars utilize foreign research methods to create forecasting model of China. Chen Jing made univariate analysis and two-groups linear discriminant analysis on the basis of financial statements data that derived from 27 ST companies and 27 non-ST companies from 1995 to 1997.As a result, the overall accuracy was 92.6%.(CHEN, 1999) Zhao Jian-mei et al (2003) used univariate analysis and Altman's Z-Score model to analyze 80 listed companies of China stock market and study the financial distress prediction empirically, then found that multivariate analysis can be more fully reflected the company's overall situation. (ZHAO & WANG, 2003)

Since the 80s of the 20th century, Logistic regression analysis replaced the traditional discriminant analysis gradually. Martin (1977) selected eight financial ratio indicators from twenty-five financial indexes such as net profit rate of total assets and established Logistic regression model for twenty-three bankrupt banks during the period of 1975-1976. Compared the predict power among models that involves the Z-Score model, ZETA model and Logistic model, it proved that Logistic model was is superior to others. (Earky, 1977) Wu Shinong(2003) selected all seventy ST companies that were treated specially between 1998 and 2000 in China A-share market and the same amount non-ST companies as the matching sample, then used cross-section analysis, univariate analysis, linear probability model, Fisher two-groups linear discriminant analysis, Logit models and other statistical methods to predict the financial difficulties of the company. His paper pointed that prediction accuracy rate of Logit model was 93.53%. (WU, 2003) Liang Qi (2005), taking into account high-dimensional and high correlation of the financial data of listed companies, used principal component analysis to reduce the dimensions of data, and modified the Logistic regression equation by using the components as dependent variables. The results showed that the accuracy rate of classification and prediction for business failure through the modified model were higher than that through the simple Logistic regression analysis. (LIANG) Li Meng, School of Economics Nankai University, proved that Logit model had the quality of high-credible identification, prediction and generalization, and was an effective instrument for commercial bank's credit risk assessment. (LI, 2005). Pang Su-lin's study showed that, the discrimination accuracy rate of Logistci regression model in general has reached 99.06%, indicating that the discriminant accuracy of Logistic model was high. (PANG, 2006) In addition, it’s relatively common that the Logistic model was applied in financial risk prewarning and credit guarantee risk by domestic scholars.

3. RESEARCH METHODS AND SAMPLE DATA

3.1 Study sample select

In the traditional methods, the ST(Special Treatment) stocks are seen as defaulted companies, while the non-ST stocks are seen as normal companies. China Securities Regulatory Commission (CSRC) required the stock exchange give stocks of listed companies in abnormal financial condition special treatment(ST for short) on March 1998, so it’s logical to measure credit default risk by using ST companies’ financial data. In this paper the total sample includes failed-managed groups and normal-managed groups that were contained by 130 listed companies from Shanghai and Shenzhen stock exchange. Failed-managed groups were composed by 73 listed companies initially which were excuted “ST” by Shanghai and Shenzhen stock exchange in 2009, then considering the integrality and availability of data, the valid sample is 65 ultimately. While choosing the sample of the normal-managed group, we considered the factors such as industry and asset scale of the companies in the failed-managed group and defined 65 non-ST companies as the matching sample. Then, from 130 companies, 45 ST and 45 non-ST companies, 90 in total were chose to form the training sample set which was used to estimate the model and test the accuracy of classification , composed by 90 samples. The rest of the 40 companies were formed a test sample set which was mainly used to test the prediction accuracy.

From the view of the industies's classfication of the valid model, according to the CSRC's classification of listed companies, the failed-managed group covers 9 industry deparments, including 3 agriculture, forestry, husbandry and fishing deparments, 41 manufacturing industry, 1 electric power, gas and water production and supply industry, 1 transportation and ware-housing industry, 5 information technology industry, 3 wholesale and retail industry, 4 real estate industry, 2 social services industry and 5 comprehensive industry. The industries of samples cover most of the CSRC's classification of listed companies, so it is strongly representative and the results of the model have great practicability.

3.2 The forming of index system

Credit risk, called default risk, is the possibility of loss suffered by banks, investors or counterparty, which generated by borrowers, securities issuers, or the counterparty who are unwilling or unable to fulfill the contract conditions for a series of reasons and make a default. Enterprise's default risk can be measured by the possibility of default within a certain time, which is named the default probability. There are many influence factors of credit risk and the mainly are moral level, management ability, capital scale, the operating environment, guarantee and the continuity of management. Based on the comprehensive consideration the factors of the credit risk, the the evaluation index system executed by the Chinese financial government department and the existing related research indexes of the listed companies' performance evaluation from both the financial position and an integrated evaluation of operating results of listed companies, the index are designed according the following principles: (1) operational principle. The index should be comprehensible, applicable and simple and we should fully consider the quantifiable and easy accessed of the index. (2) scientific principle. The chosen index should try to meet the requirements of the comprehensive evaluation of the listed company and can reflect the production, operation and the future development of the whole situation, then covers the factors that may affect the companies' credit status and includes the various static and dynamic index. (3) available principle. The obtain of the index data should be stable and reliable. Considering the caliber and availability of the statistical information of the listed companies, we have to eliminate some index reasonale in theory. (4) flexible principle. The chosen index should have certain universality and flexibility in order to suitable for companies from different industries.

According to the above principles, we choose 5 five financial factors: the profit ability, the repayment of debt ability, the operation ability, the growth ability and the capital structure. After the analysis of samples, finally the evaluation index system of the credit risk is as follows:

Profitability indicators include: return rate on total assets(X_1), rate of return on net worth (X_2), net earnings per share(X_3).

Repayment of debt indicators include: flow rate(X_4), quick ratio(X_5), debt to assets ratio(X_6), capital adequacy ratio(X_7).

Operating capacity indicators include: accounts receivable turnover ratio(X_8), inventory turnover ratio(X_9), asset turnover ratio(X_{10}), liquid assets turnover ratio(X_{11}).

Growth capacity indicators include: the net asset growth rate(X_{12}), the total asset growth rate(X_{13}), the operating income growth rate (X_{14}).

3.3 Data preprocessing

The high-correlation and high-dimensionality of the credit data of listed companies affect the process and results of analysis in the process of analyzing discriminant analysis and Logistic regression analysis to predict and study enterprise default risk. It is necessary to consider the information provided by all indicators as much as possible for the purpose of improving the accuracy of identification and prediction for enterprise operation status. Because of the uncorrelated feature among principal components, the importance of any one of principal components can be reflected by the proportion explained by this principal component of the variance explained by all. If the pre-k ($k < p$) principal components can explain most of the variance of the original data, we can say k-dimensional principal component preserve the original information of the p-dimensional space maximumly. On the basis of existing research results both at home and abroad, this paper introduced principal component analysis in order to achieve dimensionality reduction and minimize the loss of the information contained in the original data. The principal components replacing the original data are unrelated to each other.

This paper used SPSS16.0 to do principal component analysis of data being standardized. Due to the more indexes introduced, it's likely to lose useful information if the principal component extraction is not enough. By way of avoiding the occurrence mentioned above, while extracting principal components, we chose the cumulative contribution rate as reference point and emploied the largest variance orthogonal rotation method (Varimax) to change the distribution of amount of information from different principal components. After the treatment, it's easy to explain the economic implication of the principal components. Based on the standard that cumulative contribution rate should be up to 80%, we normalized the raw data and then extracted 6 principal components from 14, which extracted 85.401% from the original variabl. Let $Y_1, Y_2, Y_3, Y_4, Y_5, Y_6$ respectively equal to the first, second, third, fourth, fifth, sixth principal component. According to the component matrix and component score coefficient matrix, we can understand the correlationship between the principal components and the original data. In the light of rotated component matrix, we defined $Y_1, Y_2, Y_3, Y_4, Y_5, Y_6$ respectively as the growth capacity indicator, the long-term solvency indicator, the short-term solvency indicator, the operating capacity indicator, the profitability indicator, the financial capacity indicator.

4. EMPIRICAL ANALYSIS

4.1 Discriminant analysis model structure and empirical analysis

4.1.1 Discriminant analysis model structure

According to a certain amount of listed companies' given information of a grouping variable and the corresponding explanatory variables, we introduce linear fitting approach to building the best discriminant equation in order to predit business circumstances of listed companies belonging to different groups. The referential statistical decision rule is that variance is minimum within group and the largest between groups. The assumptions of discriminant analysis are as follows: hypothesis 1 is that there is no linear relationship among discriminant variables; hypothesis 2, is that the covariance matrix of variables in each group is equal; hypothesis 3 is that discriminant variables are multi-normal.

The form of discriminant function is as follow:

$$Z_i = b_0 + \sum_{k=1}^{m} b_k X_{ik} (i = 1, 2, \cdots, n; k = 1, 2, \cdots, m) \tag{1}$$

Where n is the amount of listed companies selected in sample, $X = (X_1, X_2, \cdots, X_m)'$ is a m-dimensional random vector, $X_k (k = 1, 2, \cdots, m)$ is a influence variable for the credit risk assessment,

b_0 is constant, b_k is a discriminant coefficient, Z_i is the cut value of i listed company's operating condition.

Linear discriminant analysis is a method using linear discriminant rule to classify and predict business conditions of listed companies, that is, the principle assigning i listed company to the failure group g' is that the posterior probability that i belongs to g' is greater than that i belongs to g , i.e.

$$P(g'|X_i) > P(g|X_i)(g' \neq g) \tag{2}$$

Assuming variables selected in the credit risk assessment satisfy the three basic assumptions of discriminant analysis, so $\hat{P}(g'|X_i) = \frac{q_{g'} \cdot \exp(-D_{ig'}^2/2)}{\sum_{g=1}^{l} q_g \cdot \exp(-D_{ig}^2/2)}$ (3)

Where, l is the number for the model group, $q_{g'}$ and q_g are the prior probability that i belongs to g' or g , $D_{ig'}^2$ and D_{ig}^2 are the distance between the score of i projection vector and the score of projection indexes means of g' or g . Solving formula (2) is equal to seeking a count making molecule of formula (3) or its natural logarithm maxmize, that is,

$$L_{ig'} = \ln q_{g'} - (D_{ig'}^2/2) = [\bar{X}'_{g'} S^{-1}]X_i + [-\bar{X}'_{g'} S^{-1} \bar{X}_{g'}/2 + \ln q_{g'}] \tag{4}$$

4.1.2 Empirical analysis

Run SPSS 16.0 software for analysis. Tests of equality of group means showed that the average of short-term solvency and operating capacity was not equal at 0.05 significance level. The test whether each group covariance matrix was equal based on the Box's test, the null hypothesis (equal population covariance matrices) was rejected at the 0.05 significance. This means that covariance matrix of each group are not equal. It violates the original hypothesis of the discriminant analysis. The significance tests of discrimination function pronounced that the function was significant at the 0.05 level by the Wilks' Lambda test.

Table 1: Standardized Canonical Discriminant Function Coefficients

	Function
	1
the short-term solvency indicator	.849
the operating capacity indicator	.446

Judging from the last key indicators, short-term solvency indicators and operating capacity index were selected for the final discrimination model construction; standardized discrimination function, expressed as $z = 0.849y_3^* + 0.446y_4^*$, here ， y_3^* , y_4^* are the standardized variables of y_3, y_4 ; non-standardized discrimination function, that was $z = 0.896y_3 + 0.537y_4$, according to the discrimination function calculated the discrimination score for each observation. According to the results, discrimination function in the normal operation group which center of the gravity was 0.603, as the two groups of the same size, critical partition points to 0, under the discrimination score of each observation classified the observations.

Table 5 illustrates the classification differences between predicted by the discrimination function and the original in the training sample and test sample. The average accuracy for the training sample is 74.4% and the test samples is 70.0%. On the one hand, judging from the model's discrimination ability, predictive validity is better. On the other hand, whether in the training samples or test samples, the probability of the ST company was judged as a normal company (training sample: 17.8%; test samples: 30%) was lower than the probability of the normal company was judged as a ST company(training sample: 33.3%; test samples: 30%), While the misjudgment cost of ST company (Type II Error) was higher than the normal company which was judged as a ST firm (Type I Error). (MA, 2008) Therefore, the result predicted by this model was relatively satisfactory.

Table 2: Classification Results[a,b]

				Predicted Group Membership		
			Classification	ST stock	Non-ST stock	Total
Cases Selected	Original	Count	ST stock	37	8	45
			Non-ST stock	15	30	45
		%	ST stock	82.2	17.8	100.0
			Non-ST stock	33.3	66.7	100.0
Cases Not Selected	Original	Count	ST stock	14	6	20
			Non-ST stock	6	14	20
		%	ST stock	70.0	30.0	100.0
			Non-ST stock	30.0	70.0	100.0

a. 74.4% of selected original grouped cases correctly classified.
b. 70.0% of unselected original grouped cases correctly classified.

4.2 Logistic regression model structure and empirical analysis

4.2.1 Logistic regression model structure

As a mainstreaming approach to quantitative credit risk, logistic regression model is flexible and simple, and moreover many of its assumptions are in line with economic realities and the distribution of financial data. The results are objective relatively. The model's greatest strength is that it solves the problem for discontinuous variables, especially for classification variable. The logistic model is based on the cumulative logistic probability function. In this function the probability that $y_i = 1$, as denoted by p_i, is given by the equation:

$$p = \frac{e^s}{1+e^s} \qquad (5)$$

$$s = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m$$

Where, $X_k (k = 1, 2, \cdots, m)$ is a impact variable in credit risk assessment, $\beta_j (j = 0, 1, \cdots, m)$ is a technical coefficient obtained by maximum likelihood estimation, $p \in (0,1)$ is the result of credit risk analysis.

If p_i for a certain listed company $i(i = 1, 2, \cdots, n)$ approximates to 0, it will be recognized as business failure; and vice versa.

$$P_i(y_i) = p_i^{y_i}(1-p_i)^{1-y_i}, i = 1, 2, \cdots, n \qquad (6)$$

Where $y_i = 0$ or $y_i = 1$, $y_i = \begin{cases} 0, & \text{if ST company} \\ 1, & \text{if non-ST company} \end{cases}$

Therefore , the likelihood function of n samples' joint PDF can be expressed as

$$L = \prod_{i=1}^{n} P_i = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \qquad (7)$$

Taking the natural logarithm on both sides above equation, it is a log likelihood function. In order to estimate the parameter $\beta_j (j = 0, 1, \cdots, m)$, the right way is to minimize the value of the log likelihood function. Take the partial derivative of each parameter, let them equal to 0, then obtain $m+1$ likelihood equations, solve each parameter $\overset{\wedge}{\beta_j}$ that is estimated value.

The logistic function is commonly used because it closely approximates the cumulative normal function and is simple to use. The slope of the cumulative logistic probability function is steepest in the region where

p_i = 0.50. In this region, changes in the independent variables have their greatest impact. The small slopes in the tails of the distribution imply that only small changes in probabilities will occur, given the same unit changes in the independent variables.

4.2.2 Empirical analysis

Run SPSS 16.0 software for analysis, the results are as follows:

Table 3: Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Y1	-.242	.551	.193	1	.660	.785
Y2	63.068	23.158	7.417	1	.006	2.456E27
Y3	10.530	2.898	13.202	1	.000	3.742E4
Y4	2.879	1.109	6.740	1	.009	17.793
Y5	-.364	.482	.572	1	.449	.695
Y6	1.843	.733	6.326	1	.012	6.318
Constant	-4.437	2.153	4.248	1	.039	.012

After ten times iteration in the process of estimating coefficient, the log likelihood value is minimum. The final result of cofficient estimation is

$$p = \frac{\exp(-4.437 + -0.242Y_1 + 63.068Y_2 + 10.53Y_3 + 2.879Y_4 - 0.364Y_5 + 1.843Y_6)}{1 + \exp(-4.437 + -0.242Y_1 + 63.068Y_2 + 10.53Y_3 + 2.879Y_4 - 0.364Y_5 + 1.843Y_6)}$$

Substitute the data of listed companies into the equation above and predict the credit risk. In this paper the classification node is 0.5 which is the default value of SPSS. If the result computed is smaller than 0.5, the company is judged into the failure group;the other hand, it belongs to the normal group.

Table 4: Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	55.885	6	.000
	Block	55.885	6	.000
	Model	55.885	6	.000

Omnibus Tests of Model Coefficients show that the significant level of P values in the last column are 0 almost and far less than 0.05. We concluded that the parameters of estimation model are significant. Therefore, the model is reasonable in the statistical sense. In the model, the overall explanatory power of six indexes is enough, but that of individual index are different. That the significance level of long-term solvency, operating capacity and financial capacity are less than 0.05 indicates their explanatory power is strong. However, the explanatory power of growth capacity and profitability indicators are weak.

Table 5: Classification Table[c]

			Predicted					
			Selected Cases[a]			Unselected Cases[b]		
Observed			**Classification**			**Classification**		
			ST stock	Non-ST stock	Percentage Correct	ST stock	Non-ST stock	Percentage Correct
Step 1	**Classification**	ST stock	39	6	86.7	14	6	70.0
		Non-ST stock	5	40	88.9	4	16	80.0
	Overall Percentage				87.8			75.0

The above table is the classification matrix which reflects the model identification capability. Upon the test, the average accuracy of Logistic regression model for the training sample is 87.8% and the test samples is 75.0%. Therefore the probability varied slghtly between committing Type I error and Type II. The two misjudgment costs are near.

4.3 Comparative analysis

Through the analysis above, considering both the average accuracy and misclassification cost, between the two statistical categories for credit evaluation model, the conclusion is that Logistic regression model is superior to discriminant analysis.

5. CONCLUSIONS

By structuring the discriminant analysis model and Logistic regression model which is suitable for credit risk, and taking empirical testing on China's financial ratios of 130 listed companies in 2009, conclusions are summarized below: (1) The principal component analysis is effective to avoid the impact on account of high-correlation and high-dimensionality of the listed companies' credit data. Principal components epitomize categories character without central information loss. It enhances the persuasiveness of discriminant model. This paper selected 6 principal components from 14 financial ratios, and the cumulative contribution rate reaches 85.401%. Meanwhile, according to the correlationship between the principal components and the original data, they are given different economic implications. (2) The discriminant analysis and Logistic regression model for credit risk assessment, based on the data of financial ratios of listed companies in China, is good for investors to determine the credict risk of the investment subject and reduce risk. Meanwhile, it could also provide valuable reference information to the internal management of listed company, when they plan to strengthen internal management and get rid of its financial difficulties.

In the process of exerting discriminant analysis model and Logistic regression model to evaluate credit risk of listed companies, there are some limitations as follows: （1）The premise of financial failure prediction is financial data released by enterprise must be authentic. As the accounting information distortion still exist currently, some financial data of enterprises do not accurately reflect the financial situation. (2) In theory, the sample introduced in discriminant analysis model must satisfy three basic assumptions, in this paper, it does not meet the assumptions due to limited sample size. So there was a certain deviation in the accuracy. (3) From the industry perspective, there are some differences in the financial ratios of different industries, this study does not consider the industry differences among listed companies, so some error is inevitable.

REFERENCES

CHEN Jing. (1999). Analysis of Forecast Financial Deterioration in Listed Companies. *Accounting Research,*(4),31-38.

E.I.Altman. (1983). *Corporate Financial Distress*. New York: John Wiley & Sons.

LIANG Qi. *Research on Credit Risk Measurement of Commercial Bank.* Beijing: China Financial Publishing House.

LI Meng. (2005). Application of Logit Model in Credit Risk Assessment of Commercial Bank. *Management Science,* (2), 33-38.

MA Ruo-wei. (2008). *Research on Financial Distress Prediction Model in Listed Companies*. Beijing: Intellectual Property Publishing House, 131-145.

M.D.Earky. (1977). Warning of Blank Failure: A Logit Regression Approach. *Journal of Banking and Finance,* 249-276.

PANG Su-ling. (2006). Application of Logistic Regression Model in Credit Risk Analysis. *Mathematics in Practice and Theory,* (9), 129-137.

WU Shinong. (2003). *Research on the Share Market Risk in China Share Market.* Beijing: China Renmin University Press.

ZHAO Jian-mei,WANG Chun-li. (2003). Empirical Study about Financial Crisis Forecast of Listed Companies in China.*Quantitative & Technical Economics Research*.

ZHANG Ling, ZHANG Gui-lin. (2000). Credit Risk Assessment in Development. *Forecasting,* (4), 72-75.