

Application of Business Risk Prediction Model: Based on the Logistic Regression Model

Jing Tang^{1,2} & Shen-Le-ping¹

¹ School of Business, Administration, South China University of Technology, China

² Guangdong University of Foreign Studies, City Guangzhou 510420, China

Correspondence: Jing Tang, School of Economics and Trade, Guangdong University of Foreign Studies, Da Xue Chen Campus, 510006, Guangzhou, China. E-mail: 200411285@oamail.gdufs.edu.cn

Received: April 4, 2014

Accepted: May 15, 2014

Online Published: June 22, 2014

doi:10.5539/ijbm.v9n7p139

URL: <http://dx.doi.org/10.5539/ijbm.v9n7p139>

Abstract

Credit risk is one of the three components making up financial risk. Under the New Basel Capital Accord, default risk has been listed as the most important factor for credit risk among all elements that affect risk of credit. Banks in China currently leave large quantities of cash idle due to difficulty in loan recovery. This essay first analyzes the distributional features of variables' cross-section data concerning the default rate. Based on credible data, this research then undertakes the choice of an appropriate default prediction model. The Binary Logistic Regression Model is adopted here to build the default rate model of business credit risk and to analyze the risk information generated, in hopes of helping banks find the correct loaning strategies.

Keywords: credit risk, default risk, logistic, regression

1. Introduction

Business risk is always one of the dominant problems of co-relational research in corporate governance, and it is also the focus of social groups of different interests and the key point to make a revolutionary breakthrough. Business risk, covered by the compensation package of senior executives, always skips out of the general public's eyesight as it is much more covert and inconspicuous than other compensation. However, in our country, especially in state-owned enterprises, monetary rewards will always exceed monetary compensation more than a great many times and the monetary remuneration can even be ignored compared with the large amount of co-related consumption. Thus, a great number of researchers keep focusing on figuring out and analysing the factors that affect the level of business risk and trying to work out how to restrict the existing self interest of senior executives.

As is shown in the existing researches, business risk is not only the problem deriving from its internal governance. Under the macro background of our country's economic transition, it may be too arbitrary to regard business risk as a sort of agency cost or compensation incentive from a sole aspect of corporate internal governance. In the climate of economic transition, the cost of maintaining social relations, the transaction cost due to the lack of formal system and other expenditures are also included in the business risk. In different types of enterprises, the proportion that each cost takes varies in business risk due to the differences in an enterprise's construction of social relations and the political connection and situation it is faced with and in terms of that, business risk appears in different characters in general and the differences of these characters decide whether business risk generates desirable or undesirable influences on corporate performance.

Therefore, starting off from the macro background under which our country is in an economic transition period, this article is going to discuss the causes, characters and its economic effects of business risk from the overall perspective of how an enterprise build up the political connections with the government and in what kind of political and policy environment it develops. In this way, the research in factors that affect business risk of enterprise is broadened. Moreover, as for the fact that whether business risk of enterprise exerts positive impact or the negative one on enterprise value has aroused heated debate in academia circle, this article will give a brand-new analysis about the controversy in the conclusion to tickle the problem above by classifying the characters of enterprise business risk.

2. Literature Review

America ranks among the top in the world in researching and practicing corporate internal governance theory. Only a few companies disclose the management's internal governance in their periodic reports. Therefore, there are no more achievements of empirical research about the internal governance than that from voluntary disclosure studied by some researchers such as Raghunandan, Rama and Hermanson. Most of their researches focus on the economic consequences generated by the defect in the internal governance and likewise, domestic experts place the same emphasis on their empirical researches though from different angles. There are only a small number of experts studying the quality of the internal governance of Chinese listed companies. Business risk is always one of the dominant problems of co-relational research in corporate governance, and it is also the focus of social groups of different interests and the key point to make a revolutionary breakthrough. As is shown in the existing researches, business risk is not only the problem derived from its internal governance. Under the macro background of our country's economic transition, it may be too arbitrary to regard business risk as a sort of agency cost or compensation incentive from a sole aspect of corporate internal governance. In the climate of economic transition, the cost of maintaining social relations, the transaction cost due to the lack of formal system and other expenditures are also included in the business risk. In different types of enterprises, the proportion that each cost takes varies in business risk due to the differences in an enterprise's construction of social relations and the political connections and situation it is faced with and in terms of that, business risk appears in different characters in general and the differences of these characters decide whether business risks generate desirable or undesirable influences on corporate performance. Traditional banks have tried to work under the override control of passing down expertise and resulted in many flaws. In order to eliminate the defects, it is necessary to establish an objective model to quantify the risk in favour of the banks. The Binary Logistic Regression Model, which is highly accurate in risk prediction and remains stable, is currently the main approach taken by financial industries to exercise financial credit ratings for different businesses.

3. Methods of Credit Risk Measurement

3.1 Expert Judgment

When the expert judgment is taken for main reference, what matter most are the expertise, subjective judgment of experts and their balancing of certain key factors. [1] Under this approach, we use the key-element analysis of C5 most of the time, which is based on borrower's moral character, capability to pay, capital store, collateral and economic cycle, known as C5. Experts in financial institutions inspect these five crucial aspects and then make a comprehensive qualitative analysis, so that borrower's willingness to pay as well as ability to pay can be fairly presented. Some banks generalize this method as SW factors, namely the borrower(who), loan term (when), collateral (what) and way of payment (how). Other banks call it the P5 factors, namely personal information, purpose of borrowing, payment, protection and prospect.

3.2 Credit Rating Methods

The most renowned method is the OCC Loans Rating approach developed by the US Office of Currency Control, which is used by supervisors and bankers in America and abroad to assess if there are enough reserves for the possible loss of loans. Under this method, current loan portfolios are divided into 5 categories (4 low-quality levels and 1 high-quality level), each of which is provided with the percentage of reserves banks should hold for imminent loan losses (For high-quality loans, the default rate is assumed as zero). Then the aggregate numbers are acquired to assess the sufficiency of reserves prepared for possible loan losses.

3.3 Credit Scoring Models

One of the famous models is the Z Score Model established by NYU business professor Edward Altman in 1968. [3] This model is a kind of discrimination model, which uses the discriminate analysis techniques in mathematical statistics to carry out a statistical analysis of banks' past loaning cases, and choose a couple of indicators that can best reflect the financial status of borrowers, have most influence on loan quality, and that are most valuable for assessing or prediction. Then a discrimination function that can best distinguish risk of different loans will be built to assess borrowers' credit risk.

3.4 KMV Model

The KMV Model is based on the Black-Scholes Option Pricing Theory. It assumes that the probability of a bank going bankrupt is largely dependent on the relations between the firm's values of assets and liabilities, as well as fluctuation rates of the firm's market price. When the market value of a company declines and goes below a certain level, the company is prone to breach its contracts of liabilities. In 1994, the KMV Company in America developed the index EDF (Expected Default Frequency) to measure risk of loans and predict the possibility of a

borrowing firm unable to pay back its principal and interest in due time over the next year. The computation of EDF is under the prerequisite that the market value relations of liabilities and stocks greatly rely on the stability of the firm’s assets value. Despite the fact that changes in assets are difficult to detect, we can nevertheless infer these changes from the firm’s equity shares changes through observation of its stock prices.

3.5 Credimerties Model

The Credimerties Model is brought up by JP Morgan and several other financial institutions in April of 1997, hoping to improve the market transparency by becoming an industrial standard. In this model, the change in firms’ credit level is considered to be the direct source of credit risk. It assesses the credit risk of assets by adopting the VRA value of the assets at a certain confidence level, which is calculated by acquiring the expected value and standard deviation of the assets.

4. Introduction of the Logistic Regression Model

Belgian mathematician Verhulst asserts that the original form of S Curve is:

$$p = \frac{e^{f(x)}}{1 + e^{f(x)}}$$

$$f(x) = a + \beta_1x + \beta_2x^2 + \dots + \beta_kx^k$$

In the function above, *a* denotes the intercept or constant, which can be replaced by β ; *f* is a multinomial of *X*, which is an expected value.

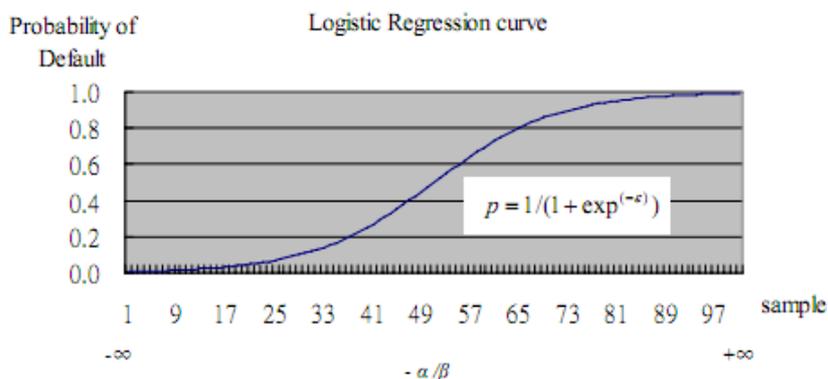


Figure 1. The logistic regression curve

The logistic regression is used to describe a dependent variable *y_i* and its relationship with several independent variables *x_i*. The basic form of this model is very similar to that of traditional linear regression. However, it does not obey the linear regression assumptions that the independent variables should be normally distributed, and that the dependent variable should be continuous.

The S Curve in Logistic Regression Model is non-linear distributed. The Cumulative Distribution Function, which is similar to random variable, is often analyzed through the Logistic Regression Model. In estimation, the Maximum Likelihood Estimation approach is adopted to estimate unknown parameters, so that the problems unsolvable through linear probability model can be settled.

The Logistic Regression Model is based on the assumption below:

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

Y_t is the dependent variable observed. When the event of *Y_t > 0* happens, it represents the situation of business default, which is denoted as *Y (t=1)*; In other situations where the event does not happen, it is denoted as *Y (t=0)* to show that default does not happen.

In the Logistic Regression, the non-linear relationship between the conditional probability of an event happening *p (Y_t = 1|X_t)* and *X* is usually monotonic, which means *P (Y_t = 1|X_t)* increases as *X* increases and decreases as *X* decreases.

The Logistic Regression function is denoted as below:

$$y_i^* = a + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i$$

y_i^* denotes the default probability and scores of liability quality of sample firm i , $i = 1, 2, \dots, n$;

x_{ij} denotes the explanatory variable j of the sample firm i ;

a denotes the intercept or constant in the Logistic Regression Model;

β_j denotes the regression parameter j , calculated under the largest estimation level;

ε_i denotes the possible error in the Logistic Regression.

Assuming that the theoretically existent continuous dependent variable Y_t represents the possibility that a default happens, the range of Y_t is infinitely small to infinitely great. When the value of Y_t steps over a critical value C (for example, $c = 0$), the event of default happens.

$$p(y_i = 1 | x_i) = \frac{1}{1 + e^{-\infty}} = 0$$

$$p(y_i = 1 | x_i) = \frac{1}{1 + e^{-\infty}} = 1$$

From the functions above, we know that the probability of $Y_t=1$ can be denoted as:

$$\begin{aligned} p(y_i = 1 | x_i) &= p[(a + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i) > 0] \\ &= p[\varepsilon_i > -(a + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i)] \\ &= p[\varepsilon_i < -(a + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i)] \end{aligned}$$

We already know that ε_i is logistically distributed, so the cumulative distribution function should be

$$p(y_i = 1 | x_i) = \frac{1}{1 + e^{-\varepsilon_i}} = 1$$

ε_i is defined as a series of linear functions influencing the occurrence of default, whose rules can be applied to unitary and multiple regression. If we assume that the probability of i sample firms defaulting $p(y_i = 1 | x_i) = p_i$, p_i is a non-linear function made up of variable x_i , this non-linear function can also be converted into a linear function.

When estimating unknown population parameters in linear regression, we mainly adopt the Least Square approach, the principle of which is to choose estimated values of parameters according to the linear regression model, leaving the deviation squared between predicted dependent variable and estimated value from the model at the smallest value. However, when the dependent variable happens to be a categorical variable instead of a continuous one, the linear regression rules cannot be applied, and the maximum likelihood estimation approach can be used for non-linear regression.

5. Application of the Logistic Regression Model

5.1 Logistic Regression Model

Assuming that there is a population made up of N companies, choosing n companies as samples randomly from $Y_1, Y_2, Y_3, \dots, Y_N$, and we get $Y_1, Y_2, Y_3, \dots, Y_N$. If we let $p_i = p(y_i = 1 | x_i)$ be the conditional probability of given x_i when $y_i = 1$; then under the same condition, we can get the conditional probability $p(y_i = 0 | x_i) = 1 - p_i$ when $y_i = 0$. So we can also acquire the probability of an estimated value:

$$p(y_i) = p_i^{y_i} (1 - p_i)^{1 - y_i}$$

While under the conditional probability from the samples $p_i = p(y_i = 1 | x_i)$, we denote the likelihood function as:

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

Since the logistic distribution is symmetric, and maximizing likelihood function $L(\theta)$ is a difficult task, we first calculate the maximum value of $\ln[L(\theta)]$ which is the natural logarithm of likelihood function, rather than calculate the likelihood function $L(\theta)$ directly.

Take the sample information into $L(\theta)$, and then take the logarithm of $L(\theta)$, we will see that the likelihood logarithm of Logistic Regression Model is as follows:

$$\ln[L(\theta)] = \sum_{i=1}^n \left[y_i \ln \left(\frac{p_i}{1-p_i} \right) + \ln(1-p_i) \right]$$

In order to get the values of population parameters α and β when the likelihood function is at its largest value, we take the partial derivative of $\alpha^{-\beta_k}$ and give it the value of zero. Then we will get n+ 1 non-linear formula. Since we need to figure out the value of $\alpha^{-\beta_k}$, first convert these non-linear formulas into polynomials using the Taylor's Series. Supposing the value of x is already known, while Δx remains unknown, so we will get:

$$f(x + \Delta x) = f(x) + \frac{f'(x)}{1!}(\Delta x) + \frac{f''(x)}{2!}(\Delta x)^2 + \wedge$$

If variable $\Delta x \rightarrow 0$, then only take the first two terms for use, leaving out the rest. This is the Newton's Method.

$$g = \mathcal{N} = \begin{bmatrix} \frac{\partial f}{\partial \alpha} \\ \frac{\partial f}{\partial \beta_1} \\ \frac{\partial f}{\partial \beta_k} \end{bmatrix}, \beta = \begin{bmatrix} \alpha \\ \beta_1 \\ m \\ \beta_k \end{bmatrix}, H = \begin{bmatrix} \frac{\partial^2 f}{\partial \alpha^2} & \frac{\partial^2 f}{\partial \alpha \partial \beta_1} & \wedge & \frac{\partial^2 f}{\partial \alpha \partial \beta_k} \\ \frac{\partial^2 f}{\partial \alpha \partial \beta_1} & \frac{\partial^2 f}{\partial \beta_1^2} & \wedge & \\ M & & & \\ \frac{\partial^2 f}{\partial \alpha \partial \beta_k} & \frac{\partial^2 f}{\partial \beta_1 \partial \beta_k} & \wedge & \frac{\partial^2 f}{\partial \beta_k^2} \end{bmatrix}$$

Then:

$$\Delta f = g^T * \Delta \beta + \frac{1}{2}(\Delta \beta^T * H * \Delta \beta)$$

And

$$g + H * \Delta \beta = 0$$

$$\Delta \beta = -H^{-1} * g$$

So we can get the value of $p(y_i)$, which is the probability of default occurrence.

5.2 Security Analysis

To summarize the theoretical and experimental analyzing results, we can see that the use of multiple chaotic systems in generating chaotic pseudo random sequences does have reached desired results, and the sequences are provided with chaotic system's long periodicity, as well as the extreme sensitivity to initial conditions and control parameters. The generation of sequences only need control parameters and initial values. It is also fast and convenient to generate chaotic pseudo random sequences of any length that have good self-correlation and cross-correlation. Therefore, inside a key set, there are not many weak keys (The correlation between the key and output is more than what a good key should have) or equivalent keys (From one set of key and output can another set of key and outpour be deduced). This should be a desirable pseudo random number generator.

The chaotic system is very sensitive to initial values and parameters, which is thereby able to provide a large set of keys, perfectly meeting the encryption requirements. By using key to put forward a new method, we can distribute the keys evenly in the key set. Compared with the pseudo random sequences generated by single chaotic system, the multiple chaotic system sequence encryption has higher security level. When adopting the floating point calculations of double precision, $x1(0)$, $x2(0)$ each has 63 care bits (Since both $x1(0)$ and $x2(0)$ have to be positive numbers, their sign bits remain zero). So the complexity of enumerable attacks is at least 2126. With complexity like this, it is fairly sufficient for most application. Using the binary sequences generated by the pseudo random number generator to cover information, which has reduced the correlation of encryption and single chaotic orbit, will render the challenge of relevant analysis even harder. Chaotic encryption belongs to sequence encryption; it is ineffective to attack with block encryption. In the meantime, choosing the information/encryption attacking schemes is barely feasible, because of the unipolarity of chaos and the iteration of chaotic signals.

5.3 Cost Analysis

The cost of calculation method consists of time cost and space cost. There are two kinds of time cost, namely time of preparation and time of encryption. Chaotic encryption is in the range of flow encryption, whose time of preparation is very little. During the encryption process, XOR manipulation is only conducted on units of data, therefore time is mainly spent on the generating operation of key flow. Compared to the block encryption calculation, which is currently very popular, the time cost is rather low.

Space cost is divided into the motionless space and running space. The motionless space refers to the space taken by the method itself after it turns into programs; usually it is shown by the length of executable code. Running space means the temporary space needed for the calculation to work during the encryption. There is no S-box

space for chaotic encryption calculation. Because the chaotic mappings we adopt are all uni-dimensional linear mapping, whose temporary variables are very few. Also, it generates key flows through cycling. The variables needed to be stored are few during cycling. Therefore, it is easier to realize and takes up very little space.

5.4 Realization Features

The encryption and decryption process of sequence code can be reused, which largely reduces the occupied space and showing better performance in both software and hardware. We have realized this method by the C# language.

5.5 Performance Analysis and Comparison

There exist non-enumerable stable non-periodic orbits;

There exists at least one unstable non-periodic orbit.

Supposing that $f_1(x_1, \mu_1)$, $f_2(x_2, \mu_2)$, $f_3(x_3, p_3)$ are three different uni-dimensional chaotic mappings $x_1(i+1)=f_1(x_1(i), \mu_1)$, $x_2(i+1)=f_2(x_2(i), \mu_2)$, $x_3(i+1)=f_3(x_3(i), p_3)$. μ_1, μ_2, p_3 are control parameters, $x_1(0), x_2(0), x_3(0)$ are the initial conditions, $\{x_1(i)\}$, $\{x_2(i)\}$, $\{x_3(i)\}$ are their chaotic orbits.

So we can define the multiple chaotic pseudo random number generator as (1):

In the formulas, $f_3(x_3, p_3)$ have to possess the attributes below:

- It should be surjection well defined at the interval $I=[a, b]$.
- It should possess ergodicity at the interval $I=[a, b]$.
- $\{x_3(i)\}$ should gradually possess independence when $i \rightarrow \infty$.
- Non-linear iteration system $x_3(i+1)=f_3(x_3(i), p_3)$ should have only one invariant distribution density function $\rho(x)$.

From above we know that the output of $f_1(x_1, \mu_1)$, $f_2(x_2, \mu_2)$ is the input of $f_3(x_3, p_3)$. If $f_1(x_1, \mu_1)$ or $f_2(x_2, \mu_2)$ remains constant, as $c \in [a, b]$, the threshold parameters we have are the chaotic sequences in dynamic changes, which means that the binarization of one chaotic orbit is completed with the help of another chaotic orbit. Every digit in the binary sequence is determined by three chaotic mapping orbits. We can know from chaotic system's extreme sensitivity to initial conditions that the security level has been greatly raised.

By exercising encryptions on the same 1.37M BMP picture, a comparative analysis concerning the time performance and effect of encryption and decryption is carried out for DES, AES, RSA and the flow code that we design. The results are as follows:

Table 1. Properties analyses and compares of encryption algorithms

	DES	AES	RSA	Chaotic encryption	sequence
Time of encryption (μs)	215826	625830	158752	126820	
Encryption performance	Medium	Fair	Fair	Good	

From the table above, we can see the comparison for these algorithms. The encryption algorithm based on chaotic sequence takes little time and receives the best results, the reason for which is that it's easy to generate the key and XOR, as the encryption method, is very convenient in operation, so it costs less time; and the chaotic sequence it generated is pseudo random sequence, so the encryption performance is better. When we use initial key to decode, even if the key is changed by the minor difference of 10^{-15} , the effect results are completely different from the original picture, which has fully proved the sensitive property that chaotic sequences have towards initial values, showing the high security level of chaotic sequence algorithm.

The encryptions of DES algorithm and AES algorithm are mainly repeated replacement, so the time consumed will be longer relative to chaotic algorithm. When it comes to encryption effect, DES method is a little inferior, especially for a small amount of text. Also, its 56-bit key seems to be inadequate now, which is the reason why this method is gradually being taken over by DES and AES.

The RSA algorithm is fairly simple, thereby easy to understand and operate, and the encryption performance is so good that no one is able to decode. The algorithm of encryption and decryption is based on the mathematical factoring of large integers. If the length of key is N , $O(N)$ is the time needed for addition and subtraction, $O(N^2)$ is the time needed for multiplication and division. When it comes to power calculations, time as much as $O(N^3)$

will be occupied. The value of N is between 512 and 1024, so the calculation volume is alarming. So it appears much worse than the other three algorithms in terms of time performance, which might explain why RSR method is usually used for encryption of small amount of data.

6. Conclusion

Sufficient researches have been done to study on the economic consequences conducted by business risk in different types of enterprises. This study first discusses the risk problem. When the logistic model is realized on the computer, due to limited precision and limited conditions, the dynamic properties of digital chaotic system is a far cry from ideal chaotic system (property degeneration). Grounded on this problem, this essay designs a new pseudo random number generator based on the multiple chaotic system, and has conducted theoretical and experimental emulations, the result of which indicate that, adopting the multiple chaotic system to generate pseudo random sequence has reached expected performance. Therefore, it is a very ideal pseudo random sequence generator. This research is aimed at finding significant variables to effectively explain a model and establishing an accurate prediction model. When it comes to finding significant variables, by using the binary Logistic forward regression, a proper prediction model is built, which is able to predict the correct default probabilities quickly and objectively, help banks' decisions concerning risk, as well as setting appropriate interest rate and improving the credit risk management. As an organization gets larger, the organizing system becomes more complicated. Under the complex system, the existing managerial tools are able to handle superficial problems, however, it cannot detect many fundamental reasons underlying the problem, and the Entropy Theory in systematic science has managed to develop a complete theoretical system in terms of organizational complexity.

References

- Argenti, J. (1976). *Corporate Collapse: The Cause and Symptoms*. McGraw-Hill.
- DeMaris, A. (1992). *Logit Modeling*. Newbury Park, CA: Sage Publications.
- Dietsch, M., & Petey, J. (2004). Should SME Exposures be treated as retail or as corporate exposures? a comparative analysis of default probabilities and asset correlation in French and German SMEs. *Journal of Financial and Quantitative Analysis*, 1477–1493.
- Eliason, S. R. (1993). *Maximum Likelihood Estimation: Logic and Preactive*. Newbury Park, CA: Sage Publications.
- Elizabeth, M. (2001). *Handbook of Credit Scoring: the Basics of Scorecard Development and Validation*. Lessons Professional Publishing.
- Feinberg, S. (1985). *The Analysis of Crossclassified Categorical Data* (2nd ed.). Cambridge, MA: MIT Press.
- Huyghebaert Gaeremynck and Roodhooft. (2000). New Firm Survival: The Effect of Start-up Characteristics. *Journal of Business Finance and Accounting*, 27, 627–651.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).