International Journal of Security and Its Applications
Vol.8, No.4 (2014), pp.37-50
http://dx.doi.org/10.14257/ijsia.2014.8.4.04

Research on Distinguish the Accounting Information Distortion Based on the Principle Components-logistic Regression Model

Zhenjie Li

School of Economics and Management, Yantai University
ytmysky@163.com

Abstract

Nowadays, the accounting information distortion of listed company is generally common in the market. It has caused adverse effects to the enterprise itself and even the development of the securities market. In order to solve this problem, domestic and foreign scholars have done a series of researches from different perspectives and create a lot of detection model to identify financial reporting fraud more correctly. As far as we are concerned, these models' index selection, calculation, prediction and application are not so satisfying and few efficient recognition models can be applied generally. In this paper, we combine the method of principal component analysis with logistic regression method. Then we select variables from the financial data that reflect the profitability, turnover, the establishment of the enterprise and some other perspectives. This accounting information distortion detection model is created by improving the method and index selection which has a higher correct recognition rate. We have chosen the 2012 financial statements from 56 firms for sample and the forecasting accuracy of the model reached 92.86%. We can get that it has obvious advantages compared to the predicted results from simple logistic regression model.

Keywords: *Financial fraud, Principal component analysis, Logistic regression*

1. Introduction

Financial reporting fraud has been a common phenomenon in the securities market. For instance, in the international market, in 2001 American Enron event shocked the whole world; in 2002 the world communication and Lantian Limited by Share Ltd set bad examples again. In the domestic market, Qiong Minyuan event happened in 1998; Yinguangxia event happened in 2001; Tianyi Science and Technology was caught financial fraud in 2004; Unisplendour Guhan Group Corporation Limited was also got punished in 2008 and Wanfu Biotechnology event was exposed in 2012. Among these fraud cases, some are typical report fraud and some cases may finished by cooperating with the auditing department. Anyway the financial reporting fraud caused very bad influence to the securities market. Therefore, the financial reporting fraud detection is one of the most urgent problems to the auditing department and the securities regulatory authorities. It must be controlled strictly to promote our securities market and national economic development.

The financial reporting fraud can be divided into financial data fraud and non-financial data fraud. It can also be divided into annual report fraud, interim report fraud, (letting) prospectus fraud, not timely disclosure notice on important matters and other information misrepresentation. We can find that the financial data fraud occupies a larger proportion of financial fraud events from the occurrence of financial fraud. In addition, according to a research of 2009 [1], in the past ten years the financial reporting fraud happened 20 times annually which accounted for 49.21% of the total fraud cases. These companies falsify the

ISSN: 1738-9976 IJSIA
Copyright ⓒ 2014 SERSC

financial statements by making up or inflating income, concealing or not timely disclosing of important matters. In this paper, the study of financial reporting fraud points at the annual report fraud and the sample data is from annual report.

As the carrier of an enterprise to send its business operation in formation, financial statements are provided to reflect its financial situation, operating results and cash flow synthetically. The financial report of each issue should fairly reflect the financial condition, profitability, growth status and so forth. Many enterprises tamper their important data in the financial statements for concealing or overstating. Financial reporting fraud is the result of internal factors and external factors. For the enterprise itself, it is generally due to the information asymmetry, interest driven, corporate governance failure and the independence of audit deficiency. We can also catch the external factors like in adequate supervision, laws and regulations deficiency. Although China has a series of laws, they need to be reinforced compared to those in other countries. We usually pay too much attention to punish the enterprises while ignore the main persons in charge who only get some public condemnation or penalty. Financial reporting fraud has a continual damage effect on the company's development and also the securities market. Therefore, strengthening the internal and external audit as well as the correct identification of the fraud is an urgent task in the development of the securities market. In order to correctly identify the fraudulent financial statements, this paper will select all kinds of influencing factors of financial reporting fraud as indicator variables. We will create a recognition model by the combining principal component analysis and logistic regression method. The results will be compared with previous research achievements and the model will be found out the advantages and disadvantages.

The domestic and foreign researches on the financial fraud were mainly focused on the motivation, influencing factors and identification methods. Financial fraud detection can be classified into financial fraud signal judgment and recognition model creation. Khanh Nguyen [2] presented the identification of financial reporting fraud, such as providing specific information on the financial report analysis, financial data comparative analysis for previous years, some index ratio of relatively large impact analysis. Kinney [3] thought the financial reporting fraud can be more likely to appear in those financial distressed companies because these enterprise managers have stronger motivation to cover up their temporary financial difficulties. Albrecht [4] thought that problems about interest burden and cash flow turnover will increase the possibility of fraud, such as the issue of income needs, descent of earning quality and high levels of debt. Different scholars have different financial judgments so we don’t have a centralized, perfect system about these conjectures. We cannot apply all these methods when judge the financial fraud and that is not functional. Then the detection models appeared. Green and Choi [5] used artificial neural network technology to construct the financial reporting fraud detection model which was built based on the original financial data. It was founded to be very effective when applied in the sample. Benish [6] proposed to detect financial fraud report by searching whether there is any false accounting data. 74 companies from 1957 to 1993 in USA by CSRC for accounting fraud sample and other listed company for normal samples, he used 8 financial indicators to establish the Probitregression prediction model of which the accurate prediction rate is 75% and was applied actually. Summers and Sweeny [7] found that usually higher inventory turnover, rapid sales growth and rate of return on total assets occurred in the previous year. Efstathins Kirkosa [8] chose 38paired healthy enterprises and unhealthy ones from manufacturing industry of Greek and used 10financial indicators as input variables to establish a decision tree, neural network and Bias belief network model. Two of the overall mean variance hypothesis testing and Logistic regression equations are used by Lou Quan [9] to study the characteristics of financial reporting fraud companies. Chen Guoxin [10] uses China's listed company data to establish the 4 indicators

 Copyright ⓒ 2014 SERSC

of fraud. Cheng Liang [11] established the one-way ANOVA recognition model based on fraudulent financial information fraud detection and the overall accuracy was 86%.MaoDaowei, Zhu Min [12]selected 136 samples and used analytical review methods to establish a one-way ANOVA model, multivariate, discriminate model, linear probability model and Logistic regression model. They used the models to forecast and the Logistic regression model prediction accuracy was up to 80%. Wang Ya and Yuan Quan [13] selected 35 samples and 35 control samples, chose five indicator variables to establish a Logistic regression model to quantify the financial reporting fraud identification, the model's accuracy is up to 84%.

Based on principal component analysis and logistic regression methods, we select the appropriate indicators from each respect of the enterprise such as the financial situation and the basic aspects to build recognition model which has a higher accuracy compared with ordinary single-factor model and logistic regression model. In this paper, the second and third part will explain the basic idea of Logistic regression method and the establishment of ideological principal component logistic regression model in detail. A simple logistic regression model and principal component-logistic regression model will be created in the fourth part. We will use the two models to predict and compare the effect and make a conclusion.

2. The Basic Idea of Logistic Regression Method

Logistic regression model, a multiple-variable analysis category, was proposed by J. Berkson in 1944 [14]. It is a common method of sociology, biostatistics, clinical, quantity psychology, econometrics, marketing and other statistical empirical analysis. It is most widely used in medicine, and then gradually expands the scope of application. logistic regression is mainly used to predict the dependent variable and the relationship between a set of discrete explanatory variables. The most commonly used is binary logistic regression and the value of the variable contains only two categories, for example: good or bad; yes or no. Financial reporting fraud judgment of this study is also binary type, that "fraud" or "non-corrupt." X usually represents the explanatory variable and $P(Y=1 \mid X)$ means probability of $Y=1$ under condition X . Logistic regression's mathematical expression is: $\log(p/(1-p)) = A + BX = L$, $p/(1-p)$ is called ODDS ratio which means occurrence to nonoccurrence ratio. We can get $P(Y=1 \mid X) = 1/(1+e^{-L})$ from the former formula and create a model like follows:

$$p_i = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n}}$$

Among it, x_i represents number i index variable, β_i is the coefficient of i -th index variable, p_i represents the probability of occurrence in number i observation, $1-p_i$ represents the probability of nonoccurrence. According to the sample data model parameters can be calculated in the maximum likelihood method and then we can predict. When Logistic regression model is applied to financial reporting fraud detection, the closer is p to 1, the lower is the possibility of financial reporting fraud and the closer is it to 0, the higher. The model contains the following assumptions:

①Data should be derived from a random sample;

②The dependent variable y is assumed to be a function of n independent variables x_i and there is non-linear relationship between the dependent and independent variables. $(i = 1, 2, \ldots n)$

Copyright ⓒ 2014 SERSC

③The dependent variable should be dichotomous variables; take financial reporting fraud for an example, “yes” is 1 and “no” is 0.

④ There is no multi collinearity between the independent variables.

We generally use the maximum likelihood method or iterative method in Logistic regression model parameter estimation. The simulation effects of the model will be judged by the overall goodness of fit of the model, whether the explanatory variables can explain the role and the extent of the explanation.

Logistic regression method is a typical method of discrete choice models and it has obvious advantages compared with the linear regression when analyzing correlations between variables. On the one hand, according to domestic and foreign scholars’ research, applicability and accuracy of Logistic regression model is significantly higher than linear model in financial accounting due to the nature and number of influencing factors. On the other hand, Logistic regression model has overcome the disadvantages of multiple linear analyses because it doesn’t restrict the distribution of the independent variables and has no strict assumptions about the distribution of types, or the covariance matrix. This method is currently occupies the mainstream status in the research field of discriminant analysis. However, Logistic regression model still has some short comings in the identification of financial reporting fraud. Logistic regression model is difficult to use because there’re too many variables which make the model complicated. Many variables which have small little contribution to the model affect the sound effects of other variables and the accuracy of model predictions. Additionally it’s also difficult to fully meet all the hypotheses. When the degree of correlation between variables increases, standard errors of coefficients estimation increase dramatically. Meanwhile, coefficients are very sensitive to the sample and model settings. Small changes in model setting like adding or deleting cases in the overall sample will cause great changes in coefficient estimation. Financial ratios are calculated by the financial statements of mutual relations and the degree of correlation between similar indicators are very large. There will be serious multi collinearity interference if these variables are fit into the model without any treat. Regrettably, this problem has not been realized by most domestic and foreign scholars. So the stability and accuracy of those models could not bode well.

Ohlson [15] was the first one who applied the logistic regression method in financial crisis warning. The sample of 105 bankrupt companies and 2058 non-bankrupt companies are chosen from 1970 to 1976. The author analyzed their distribution in the ruin probability interval as well as the relationship between the two types of errors and the dividing point. Then he found that the company size, capital structure, performance and financing ability make the prediction accuracy reach 96.12%. This approach enabled financial warning greatly and overcame many traditional problems, including assumptions like variables belong to normal distribution, bankrupt and non-bankrupt firms have the same covariance matrix. Then logistic regression method was applied in the financial field universally. Jiang Guohua and Wang Hansheng [16] proposed a prediction model of "ST" listed companies’ bases on logistic regression whose accuracy reached 64%andshowedits good predictive capacity. Wang Ya also used Logistic regression model to identify and judge the financial reporting fraud, get more accurate results than other models. Some other Logistic methods later derived from the Original Logistic regression model.

3. The Principle Components-logistic Regression Method

The first step to create the principle components-logistic regression model is principal component analysis. We should use several principal components unrelated to each other

 Copyright ⓒ 2014 SERSC

instead of numerous original variables to avoid overlapping information. Then it comes to the logistic regression analysis based on the principal component analysis to create the principle components-logistic regression model and grasp the possibility of listed companies' financial failure from the perspective of probability. Principal component analysis and logistic regression method are both applied in the financial field a lot but their combination is not often seen. Yang Shengyuan [17] has combined the principal component analysis with the logistic regression in our corporate financial distress prediction. Geng Kehong and Li Zhansheng [18] established the principle components-logistic regression model when creating listed companies' financial failure prediction model. Li Shaoxuan and Zhang Ruili [19] have also applied the principle components-logistic regression method in the study of internal factors of information disclosure of listed companies controlling. By comparison, the principle components-logistic regression model has a higher accuracy than other models like simple logistic model or the principle component model.

Zhang Aimin [20] use the principal component analysis for the first time in the study of listed companies' financial failure prediction and the model created base on this method is called principal component analysis model. In financial studies, we often encounter multiple indicators and the independent and dependent variables can both be categorical variables or numeric variables. Due to the big number of variables, they often have some relevance that makes the observed data reflects overlapping information. When there're many variables, it is difficult to study the distribution of the sample in a high-dimensional space. The simplest and most straightforward solution to these problems is to reduce the number of variables which on the other hand will inevitably cause some loss and incompleteness of information. To solve this problem better, people want to explore a more effective solution which will greatly reduce the number of variables involved in data modeling and won't cause too much loss of information. The principle component analysis is such a widely used method that can effectively reduce the dimension of variable.

In terms of minimal information loss, principal component analysis replaces those original variables with several integrated indicators which have the following characteristics:

① The number of principal components is far less than the original variables.
After the original variables are replaced by a few factors which will take part in data modeling, the calculation workload process will be greatly reduced.

② The principal components can reflect most information of the original variables.
Factors are not a simple trade-off of the original variables, but the result of the reorganization of the original variables. Therefore they will not cause much information loss and can represent most original variables.

③ These principal components should be uncorrelated.
We can get some new comprehensive indexes (principle components) from the principle component analysis and they can effectively solve the problems like overlapping information and multi collinearity.

④ Principal component has named interpretative.

In conclusion, the principal component analysis is a multivariate statistical analysis method which can condense original variables into a few factors and make them have named interpretative. The basic idea is trying to replace a number of original indicators with some relevance like $X_1, X_2, \ldots, X_n$ with a less number of unrelated comprehensive indexes like F_m. Then how to extract the composite indicators to reflect the greatest degree of information the original variable X_n represents and simultaneously ensure the new indicators remain independent of each other.

Copyright ⓒ 2014 SERSC

Assume F_1 represents the principle component which means the first linear combination of the original indicators:

$$F_1 = a_{11}X_1 + a_{21}X_2 + \ldots + a_{n1}X_n$$

As we know, each main component of the extracted amount of information can be measure by its variance $Var(F_1)$ and greater variance means more information F_1 contains. We often want to get the largest amount of information from the first principal component and therefore F_1 should have the greatest variance among all those linear combinations of $X_1, X_2, \ldots, X_n$. That's why F_1 is called the first principle component.

If the first principal component is insufficient on behalf of the original n indicators and then we can select the second principal component F_2. In order to effectively reflect the original information, F_1 and F_2 don't reflect duplicate information which means they're independent of each other. Expressed in mathematical language: their covariance $Cov(F_1, F_2) = 0$, so F_2 has the greatest variance except F_1 and is called the second principle component.

$$\begin{cases} F_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1n}X_n \\ F_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{mn}X_n \\ \quad \ldots\ldots \\ F_m = a_{m1}X_1 + a_{m2}X_2 + \ldots + a_{mn}X_n \end{cases}$$

According to the analysis above:

(1) F_1 and F_2 are irrelevant, $Cov(F_i, F_i) = 0$ and $Var(F_i) = a_i \sum a_i$, $\sum a_i$ is the Covariance matrix of X.

(2) F_1 has the greatest variance among all those linear combinations of $X_1, X_2, \ldots, X_n$ and F_m has the greatest variance except $F_1, F_2, \ldots, F_{m-1}$.

(3) $F_1, F_2, \ldots, F_m$ are the 1st, 2nd ,…, and m th principle component and they are new indicators.

The analysis above shows that there're two main tasks of principal component analysis:

(1)To determine the expression of each principal component F_i $(i = 1, 2, \ldots n)$ on the original variables $X_j (j = 1, 2, \ldots, n)$ or we can say the coefficient $a_{ij} (i = 1, 2, \ldots, m; j = 1, 2, \ldots, n)$. It can be proved mathematically that, the characteristic root of the covariance matrix of the original variable is the main ingredients' variance, so the top m characteristic roots represent the top m variances of principle components. The corresponding feature vectors of top m characteristic roots are the coefficients recorded as a_i of principle components' expressions (we select λ_i in this way to ensure the variance are arranged in descending order). For restrictions, a_i use the corresponding unitized feature vector of λ_i what means that $a_i' a_i = 1$.

(2)Calculate the principal component loads which a reflection of the degree of correlation between the main components and the original variables.

$$P(Z_k, x_i) = \sqrt{\lambda_k}\, a_{ki} (i, = 1, 2, \cdots, p; k = 1, 2, \cdots, m)$$

The general calculation steps of the principal component analysis:

 Copyright ⓒ 2014 SERSC

Step1 : To standardize the original data, gather P dimensional random vector $X = (X_1, X_2, \ldots, X_n)^T$, the sample size is N and $x_i = (x_{i1,} x_{i2,\ldots,} x_{in})^T, i = 1, 2, \ldots, N, N > n$. Then construct a sample array and make a standardized transformation to the sample array as follows:

$$Z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, i = 1, 2, \ldots, N; j = i = 1, 2, \ldots, n$$

$\bar{x}_j = \frac{\sum_{i=1}^{N} x_{ij}}{N}, s_j^2 = \frac{\sum_{i=1}^{N} (x_{ij} - \bar{x}_j)^2}{N-1}$, then we can get the standard array Z .

Step2: solve the correlation coefficients to the standard array Z : $R = \left[r_{ij} \right]_n xn = \frac{Z^T Z}{N-1}$ and $r_{ij} = \frac{\sum z_{kj} \square z_{kj}}{n-1}, i, j = 1, 2, \ldots, n$.

Step3: Solve the characteristic equation of sample correlation matrix R : $\left| R - \lambda I_n \right| = 0$ and get n characteristic roots. Ensure the principle components and the value of m by $\frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{n} \lambda_j} \geq 0.85$ to make the utilization of information be more than 85%.For each $\lambda_j (j = 1, 2, \ldots, m)$, solve the equations $Rb = \lambda_j b$ and get the unit eigenvector b_j^0 .

Step4: Transform the standardized variables to principle components $F_{ij} = z_i^T b_j^0, j = 1, 2, \ldots, m$. $F_1, F_2, \ldots, F_m$ are the 1st, 2nd ,…, and m th principle component.

Step5: To evaluate the m principal components comprehensively and calculate their weighted sum to obtain the final evaluation value, the weights for each is the variance contribution rate of the principle component.

In the financial report, we must select n proper and unrelated variables first according to the object of study. The index system also requires structured. If targets are too complicated, it will take a lot of effort during screening. After finishing selecting variables we need to standardize the sample data and calculate the characteristic value as well as the variance contribution rate of each variable. As illustrated in Table 1, we sort the variables in descending order of its contribution rate and calculate the cumulative variance contribution. Generally speaking, if the cumulative contribution rate reaches 85% or higher at X_i , it means that the i variables reflect most information.

Table 1. Principal Components Selection

variables	X_1	X_2	X_3	……	X_n
Characteristic value	λ_1	λ_2	λ_3	……	λ_n
The contribution rate	θ_1%	θ_2%	θ_3%	……	θ_n%
The cumulative contribution rate(%)	θ_1%	θ_1% + θ_2%	θ_1% + θ_2% + θ_3%	……	100%
The 1st principle component F_1	a_{11}	a_{12}	a_{13}	……	a_{1n}
The 2nd principle component F_2	a_{21}	a_{22}	a_{23}	……	a_{2n}
……	……	……	……	……	……
The m-th principle component F_m	a_{m1}	a_{m2}	a_{m3}	……	a_{mn}

After getting the result, we use the principal components to replace the original

variables and create a logistic regression model with only i variables:

$$p_i = \frac{e^{\beta_0 + \beta_1 F_1 + \beta_2 F_2 + \cdots + \beta_m F_m}}{1 + e^{\beta_0 + \beta_1 F_1 + \beta_2 F_2 + \cdots + \beta_m F_m}}$$

We can use the model finished to predict and have 0.5 as the cut-off point. Take this paper for example, if $p_i > 0.5$ we will judge the company for financial reporting fraud and if $p_i < 0.5$ the company is regular. We cannot make the estimation when $p = 0.5$.

4. The Empirical Analysis

In this paper, we select 26 listed companies during 2009 to 2011 whose annual report's audit opinion was" a disclaimer of opinion" or "no opinion" as the sample of financial reporting fraud companies and other 30 companies whose audit opinion was "unqualified opinion" as the comparison sample.

The first step to take advantage of each piece of information to correctly identify the financial reporting fraud is grasping the enterprise's internal motivation. The vast majority of financial fraud is driven by profit and it can be divided into pressure-driven and greedy nature. The pressure may be obtaining the listing qualifications, raising the stock price, ease the investors, urging creditors to ease policy or avoiding getting special treatment. These pressures are all likely to carry out the financial reporting fraud. The greed may be pursuit of profit, tax evasion or further business expansion. Some companies are not managed properly and have some internal problems. Based on the analysis above, we will analyze the factors from the company's size, profitability, liquidity, growth prospects, credit condition and shareholding structure.

Table 2. Index System of Financial Reporting Fraud Research

First grade	Second grade	Third grade
Financial reporting fraud of listed company	Assets and Liabilities	Assets、Asset-liability ratio
	Profitability	Earnings per share、Return on equity、Operating profit cash guarantee rate
	Liquidity	Assets turnover ratio、Accounts receivable turnover rate
	Growth prospects	Operating profit growth rate、Sales growth rate
	Credit condition	Has the company cheated before?
	Shareholding structure	Whether it's state-controlled?

① Assets and liabilities: A company's assets and liabilities are directly linked with its information disclosure. Generally speaking, an asset-rich company will not be in a bad financial condition except it's in urgent need of circulating capital. So we select

 Copyright ⓒ 2014 SERSC

total assets and asset-liability ratio and record them as TA (X_1) and ALR (X_2). T Are presents the scale of the enterprise and ALR represents the asset structure. It will be about appropriate if ALR's value is between 40% and 60%.

②Profitability: The profitability is one of the core indicators of an enterprise's financial condition, which can be divided into quantity and quality of earnings. We can select listed company's earnings per share and return on equity as the quantitative indexes. Then record them as EPS and ROE. Operating profit cash guarantee rate can be selected as qualitative index and recorded as OPR. OPR= (net cash flow generated from operating activities +items should be added– items should be eliminated) /operating profit. Items should be added refer to the difference between income tax paid and income tax refunds, which is usually ensued by "income tax" in the income statement after refund deduction; items also should be eliminated refer to impairment of assets, fixed assets depreciation, amortization of intangible assets, long-term expenses amortization, accrued expenses and financial expenses belonging to the investing and financing activities. The index is used to represent the proportion of cash recovery in operating profit, which is usually between 0 and 1. 1 or higher means good quality of operation profit. 0 or lower means no substantial cash flow in operation profit which means bad quality and the lower, the worse.

② Liquidity: We selected asset turnover ratio and accounts receivable turnover rate to reflect the liquidity and record them as ATR (X_6) and ARR (X_7). ATR is used as a comprehensive evaluation of the enterprise management quality and efficiency of all assets and a high ATR can reflect a rapid turnover of total assets and a strong sales ability. The accounts receivable turnover plays a key role in the enterprise capital turnover. Many companies have large-scale as set but often have problems in liquidity, which relate with the turnover of accounts receivable and inventory. Accumulation of bad debt scan cut off an enterprise's life line. In addition, accounts receivable is also a common project used in financial reporting fraud to fabricate income. Therefore, these two indexes are well suited to judge the financial situation.

③ Growth prospects: We can use the operating profit growth rate and sales growth rate to reflect the development potential of an enterprise. Record them as OGR (X_8) and SGR (X_9) and the higher, the better.

④ Credit condition: We choose historical financial reporting fraud as an index to reflect the credit condition. It's a categorical variable recorded as RFH (X_{10}) so we should assign to it. If the listed corporation has got audit opinions like "a disclaimer of opinion" or "no opinion", we assign 1 to it; if its opinions were all "unqualified opinion", "unqualified opinion with explanatory notes" or "qualified opinion", we assign 0 to it.

⑤ Shareholding structure: Listed Corporation can be divided into state- controlled and non state-controlled. State-controlled group generally have stronger financial ability and the possibility of financial reporting fraud is relatively smaller. We record it as OS (X_{11}) and state-controlled enterprises get 1, non state-controlled ones get 0.

First, we will use all the sample data to establish a simple logistic regression model and have a look at the predictive effect. Then we get the coefficient of each indicator and significant test results as Table 3 shows:

Copyright ⓒ 2014 SERSC

Table 3. Results of Sample Data

Index	coefficient	Sig
TA (X_1)	- 1.367	0.453
ALR (X_2)	6.074	0.026
EPS (X_3)	- 16.619	0.000
ROE (X_4)	- 3.432	0.013
OPR (X_5)	- 4.307	0.047
ATR (X_6)	- 11.815	0.391
ARR (X_7)	- 13.730	0.067
OGR (X_8)	- 2.476	0.034
SGR (X_9)	- 1.304	0.128
RFH (X_{10})	1.241	0.021
OS (X_{11})	- 0.754	0.073
Constant	0.416	0.010

According to the result, TA (X_1) , ATR (X_6) , ARR (X_7) , SGR (X_9) and OS (X_{11}) are not significant so we eliminate the 5 indexes and keep the other 6. Then we get a logistic regression model as following:

$$p_i = \frac{e^{0.416+6.074RAL-16.619EPS-3.432ROE-4.307OPR-2.476OGR+1.241RFH}}{1+e^{0.416+6.074RAL-16.619EPS-3.432ROE-4.307OPR-2.476OGR+1.241RFH}}$$

According to the regression coefficients' significant test, EPS (X_3) and ROE (X_4) are found to be most significant. The value of the model's significance is 0.000, which means that it's significant at the level of 5%. Use the model to predict the 56 enterprises about the fairness of their 2012 annual financial reports. If we get that $p_i > 0.5$, the financial reporting fraud will be judged to be unfair. As Table 4 shows, the prediction accuracy rate of fraud is 73.08% and of non-fraud is 73.33%. The overall prediction accuracy of the model is 73.21%.

 Copyright ⓒ 2014 SERSC

Table 4. Prediction of Simple Logistic Model

prediction / actual	Fraud	Non-fraud	Total	Prediction accuracy
Fraud	19	7	26	73.08%
Non-fraud	6	24	30	80%
Total	25	31	56	76.79%

Below we will establish a logistic regression model based on principal component analysis. Order the indexes in descending order of their contribution rate of information, as shown in Table 5:

Table 5. Variance Contribution Rate

Index	contribution rate	Accumulate contribution rate	The 1st principle component	The 2nd principle component	The 3rd principle component	The 4th principle component	The 5th principle component
EPS (X_3)	27.56%	27.56%	0.881	- 0.235	- 0.041	- 0.079	- 0.301
ARR (X_7)	22.43%	49.99%	- 0.037	- 0.029	0.836	- 0.057	- 0.432
OPR (X_5)	17.21%	67.2%	- 0.904	- 0.860	0.649	0.344	0.017
ROE (X_4)	14.07%	81.27%	- 0.259	0.255	- 0.135	0.671	0.099
OGR (X_8)	7.03%	88.3%	- 0.130	0.711	- 0.259	- 0.018	- 0.160
ALR (X_2)	4.25%	92.55%	0.447	- 0.089	0.046	0.578	- 0.174
RFH (X_{10})	2.14%	94.69%	0.093	- 0.191	0.261	0.319	0.007
OS (X_{11})	2.01%	96.7%	0.506	- 0.276	- 0.702	- 0.040	0.861
SGR (X_9)	1.74%	98.44%	- 0.340	0.663	- 0.103	- 0.034	- 0.249
ATR (X_6)	1.21%	99.65%	- 0.715	- 0.504	0.794	- 0.016	- 0.203
TA (X_1)	0.35%	100%	- 0.163	0.358	- 0.277	0.209	- 0.522

Copyright ⓒ 2014 SERSC

According to the Table, the accumulate information contribution of EPS (X_3), ARR (X_7), OPR (X_5), ROE (X_4) and OGR (X_8) reaches 88.3% which means that they contain the main information of all indexes. We have the five principle components as F_1、F_2、F_3、F_4、F_5 and use them to create a new logistic regression model as follows:

$$p_i = \frac{e^{0.409-1.217F_1-0.359F_2-0.406F_3+1.162F_4-0.793F_5}}{1+e^{0.409-1.217F_1-0.359F_2-0.406F_3+1.162F_4-0.793F_5}}$$

We judge the 2012 financial report to be fair when $p_i > 0.5$ and to be unfair when $P \le 0.5$. Test the model and get the value of significance is 0.000 which also means it's significant at the level of 5%. Similar to the simple logistic model, we can get that profitability plays a very important role in the model among all the indexes. Use the new model to predict the samples about their 2012 financial reports and get the results as follows:

Table 6. Prediction of Principle Components-Logistic Regression Model

prediction / actual	Fraud	Non-fraud	Total	Prediction accuracy
Fraud	25	1	26	96.15%
Non-fraud	3	27	30	90%
Total	28	38	56	92.86%

For the principle components-logistic regression model, prediction accuracy of fraud is 96.15% and of non-fraud is 90%. The overall prediction accuracy is 73.21% and 16 percentage points higher than the simple logistic model.

5. Conclusions and Recommendations

Based on the empirical analysis above, the model created by combining principal component analysis with logistic regression method has a good predictive effect. From a series of data calculated, we can draw the following conclusions: first of all, among the 11 indicators in the initial selection, the information contributions of EPS, ROE, OPR, ARR and OGR are higher, which means that the profitability, liquidity and growth prospects have a dominant influence on financial reporting fraud; the second point, several key indicators in the model are inverse indexes, greater profitability, higher turnover rate and better growth prospects mean a less chance of fraud; the last point, apart from these indicators, asset structure and shareholding structure also has a very important reference value in actual prediction, a firm with too much debt and concentrated owner ship is more prone to cheat. In addition, as an extension and expansion of simple logistic regression model, the principle component-logistic regression model has obvious advantages in listed companies' financial reporting fraud detection compared to the simple model and is more applicable.

This study has certain limitations and disadvantages; the prediction model has some prerequisites such as the sample data should be true, reliable and random. But due to the practical constraints, we select the sample data from what we can obtain and the data may not be reliable due to inadequate supervision and the imperfect legal system. These indicators will have an impact on the accuracy of the model. With the development of the securities market, strengthened supervision and the constant perfection of the laws,

 Copyright ⓒ 2014 SERSC

the model can be continuously improved and the accuracy will be further improved.

References

[1] Y. Dakai and H. Ying, "An Empirical Study on Fraudulent Financial Reports of Listing Corporation", Modern Accounting, no. 8, **(2009)**, pp. 4-9.
[2] K. Nguyen, "Financial Statement Fraud: Motives, Methods, Cases and Detection, **(2008),** pp. 40-49.
[3] W. Kinney, L. McDaniel, "Characteristics of firms correcting previously reported quarterly earnings", Journal of Accounting and Economics, no. 11, **(1989)**, pp. 71-93.
[4] W. S. Albrecht, G. W. Wemz and T. L. Williams, "Fraud Bring the Light to the Dark Side of Business, New York: IRWIN, Inc, **(l995),** pp. 15-52.
[5] B. P. Grccn and J. H. Choi, "Assessing the Risk of Management Fraud through Neural Network Technology Auditing", Spring, no. 16, **(1997),** pp. 14-28.
[6] D. M. Beneish, "Incentives and penalties related to earnings overstatements that violate GAAP", The Accounting Review 174, **(1999),** pp. 425-457.
[7] S. Summers and J. Sweeney, "Fraudulently misstated Financial Statements and Insider reading", The Accounting Review, vol. 73, **(1998)** January, pp. 131-146.
[8] E. Kirkosa, C. Spathisb and Y. Manolopoulosc, "Data Mining techniques for the detection of fraudulent financial statements", Expert Systems with Applications, 200, vol. 32, no. 4, pp. 995-1003.
[9] L. Quan, "Empirical Research on Financial Reporting Fraud of Listing Corporations in China", Securities Market Herald, no. 10, **(2003)**, pp. 35-36.
[10] C. Guoxin, L. Zhanjia and H. Feng, "An Empirical Research on Detection of Fraudulent Financial Reports -Based on Data of Chinese Listed Company", Auditing Research, no. 3, **(2007)**, pp. 88-92.
[11] C. Liang and W. Xuan, "Model of Analysis and Recognition of Accounting Frauds", Securities Market Herald, no. 8, **(2003)**, pp. 52-56.
[12] M. Daowei and Z. Min, "The Financial Judgment Method and Model in Corporations' Credit Status: Based on the Analysis of Listed Corporations' Fraud Financial Reports", Journal of Sichuan University (Social Science Edition), no. 3, **(2006)**, pp. 51-57.
[13] W. Ya and Y. Quan, "An Empirical Study on Listed Companies' Financial Reporting Fraud Identification", Economic Research Guide, no. 29, **(2012)**, pp. 139-140.
[14] J. Berkson, "Application of the Logistic Function to Bio-Assay," Journal of the American Statistical Association, vol. 39, **(1944)**, pp. 357-365.
[15] J. A. Ohlson, "Financial Ratios and the Probabilistic Prediction of Bankruptcy", Journal of Accounting Research, vol. 18, no. 1, **(1980),** pp. 109-1311.
[16] J. Guohua and W. Hansheng, "Research on financial statement analysis and listing Corporation ST prediction", Audit Research, no. 6, **(2004),** pp. 60-63.
[17] Y. Shengyuan, "Study on financial predicament forecast of listed company in China", Journal of Kunming University, no. 5, **(2009)**, pp. 94-99.
[18] L. Shaoxuan and Z. Ruili, "Study on influence factors of internal control information disclosure of listing Corporation: empirical analysis based on data of Shanghai and Shenzhen", Communication of Finance and Accounting, no. 9, **(2009)**, pp. 27-31.
[19] G. Kehong and L. Zhansheng, "Empirical Comparison among Different Predicting Models of Financial Failure for China's Listed Corporations", Journal of Shanxi Finance and Economics University, no. 5, **(2006)**, pp. 129-133.
[20] Z. Aiming, Z. Chunshan and X. Danjian, "Principle Components Prediction Model and the Empirical Study of Financial Failure of Public Company", Journal of Finance, no. 3, **(2001),** pp. 11-25.

Authors

Zhenjie Li, male, born on August, 1976 in Linyi, Shandong. Now, he is a lecturer in School of Economics and Management, Yantai University Lecturer, and his research directions are financial management, supply chain management, Logistics management and enterprise information.

Copyright ⓒ 2014 SERSC

Copyright ⓒ 2014 SERSC