3rd International Conference on Economic and Business Management (FEBM 2018)

Research on Financial Early Warning of Listed Companies Based on Lasso-logistic Model

Yutong Han[a], Qi Sun[b], Zhuoxi Yu[c]*
School of Management Science and Information Engineering
Jilin University of Finance and Economics
Jilin Province Key Laboratory of Fintech
Changchun, China
412167001@qq.com[a], 949667392@qq.com[b], yzx8170561@163.com[c*]

***Abstract*—The financial status is an important factor affecting the survival and development of enterprises. When we conduct financial early-warning model analysis, the selection of index variables and the estimation of model parameters directly affect the prediction accuracy of the early-warning model. Lasso is a variable selection method for shrinkage estimation. By constructing a penalty functions to realize variety selection and retaining the advantages of subset shrinkage; Lasso can be applied to time series, high-dimensional graphics discrimination and selection. In this paper, the Lasso method and the Logistic model are combined to construct an early-warning model reflecting the financial status within the enterprise. The experimental results show that the model can effectively select relatively important influencing factors and also has good predictive evokes.**

Keywords—Financial warning; Lasso algorithm; Logistic model; Variable selection

I. Introduction

The financial crisis of listed companies has certain signs in advance, and these signs are often reflected in some indicators related to corporate finance. We can build a financial early warning indicator system and establish an early warning model. There are many methods for constructing financial early warning models, such as solely use discriminant analysis and logistic regression. However, the common crisis early warning model can't get rid of the assumptions and limitations of traditional statistical methods on data, and can't self-learn and adjust, it has great limitations in data selection. Therefore, this paper starts from the statistical and intelligent modeling technology, comprehensively considers the timeliness of prediction, the number of samples and the application of statistical learning techniques, and uses the Lasso method combined with the Logistic model to construct an early warning model. A better early warning effect has been achieved, and the prediction results of the Lasso-Logisitic method are summarized to show the financial status of the listed company, which provides a basis for decision makers to make decisions based on relatively comprehensive considerations.

In the research of the financial early warning model, it is generally considered whether the enterprise has financial distress, that is, whether it is predicted by ST. The Logistic model was introduced into the financial risk assessment earlier. Deng Jing et al.'s research shows that the Logistic model has a good predictive effect [1], but when there are too many independent variables, the model will have problems such as multicollinearity and unrelated variable interference [2], so before modeling, we need to make the variable selection first. For the variable selection problem of regression models, the subset selection method and the ridge regression have some improvement compared with the least squares estimation, but their variable selection is still not stable enough and the interpretability is not strong [3]. The Lasso method was proposed by Robert Tibshiranni in 1996, it not only accurately selects the variables strongly related to the class label,but also overcomes the instability of the ridge regression, the calculation is simple and has become a research hotspot [4]. Since the financial early-warning problem is a two-category prediction problem, it is considered to combine the Lasso method with the logistic model and apply it to the analysis of the financial status of listed companies. Chinese scholars have applied the Lasso-logistic model into personal credit risk early-warning and supply chain financial credit risk evaluation, and all of them have achieved good prediction results. Moreover, by comparison, the Lasso-logistic model is more concise than the stepwise regression, and it can dig out the key influencing factors[5]. According to the multi-dimensionality of the listed company's financial data and the correlation between the variables, the Lasso-logistic model is applied to the financial early warning. The experimental results show that the Lasso-logistic model can simultaneously perform variable selection and parameter estimation without multiple repetitions of operation [6], and achieved a good early-warning effect.

II. Method introduction

A. Lasso Method

The Lasso method is an estimation method proposed by Tibshirani (1996) to achieve a reduced set from indicators[7]. By constructing a penalty functions to simplify the regression coefficients of the model;some coefficients are even compressed to zero to achieve model selection. The method has strong stability, and the objective function is as follows:

$Equation(1):$

$$\arg\min\left\{\sum_{i}^{n}\left(y_i-\sum_{j}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j}^{p}\left|\beta_j\right|\right\} \qquad (1)$$

$Equation(2):$

$$subject\,to\sum_{j}^{p}\left|\beta_j\right|\le t \quad , \qquad (2)$$

where $i=(1,2,...,n)$, $j=(1,2,3,...,p)$. x_{ij} is the j th component of the i th variable, y_i is the output variable corresponding to the i th variable, β_j is a regression coefficient for the j th variable. $\lambda\sum_{j}^{p}\left|\beta_j\right|$ is the penalty function, when λ become larger, the greater the punishment, leave fewer variables. In addition, automatic compression of the regression coefficients can be achieved by controlling the harmonic parameter t. When the value of t is small, some of the less relevant coefficients can be compressed to zero, and finally the variable selection is implemented. At present, the most commonly used algorithm for solving the lasso problem is the Least Angle Regression (LARS) algorithm with relatively fast calculation speed [8]. The solution process can be realized by the package Lars's package in the R software.

B. Lasso-Logistic Model

Assumed the observations $\left(X^i,y_i\right)$, $i=1,2,3,...,n$ are independent and identically distributed, $X^i=\left(x_{i1},x_{i2},...,x_{ip}\right)$ is the ith observation of the model independent variable, y_i is the dependent variable for the model. In the financial early warning analysis of the company, the dependent variable is whether the company is judged as a financial crisis company, which is a binary discrete variable with a value of 0 or 1. Therefore, the conditional probability of the Logistic model is:

$Equation(3):$

$$\log\left\{\frac{p\left(y_i=1\middle|X^i\right)}{1-p\left(y_i=1\middle|X^i\right)}\right\}=\eta_\beta\left(X^i\right) \quad . \qquad (3)$$

The coefficient estimator $\hat{\beta}$ in the Lasso-logistic regression model can be written as:

$Equation(4):$

$$\hat{\beta}=\arg\min_{\beta}\sum_{i=1}^{n}\left\{y_i\eta_\beta\left(X^i\right)-\log\left\{1+\exp\left[\eta_\beta\left(X^i\right)\right]\right\}\right\}+\lambda\sum_{j=1}^{p}\left|\beta_j\right| \qquad (4)$$

where $\eta_\beta\left(X^i\right)=\beta_0+\sum_{j=1}^{p}x_{ij}\beta_j$, the establishment of the above formula applied the minimization theorem of the convex function and the determination of the log-likelihood function [9].

By controlling the harmonic parameter λ, the automatic compression of the regression coefficient can be conducted to achieve the purpose of variable selection. At present, there are three methods for determining the λ value: the cross-validation method, the generalized cross-validation method and the unbiased estimation of the analysis [10]. The first two are mainly applicable to the case where the distribution of the observed samples is unknown, which is consistent with the data characteristics of this paper. Therefore, the generalized cross-validation method is used to determine the value of the optimal λ, which can be realized by the function cv.glmnet() in the R software. Its specific expression is:

$Equation(5):$

$$\hat{\lambda}=\arg\min\frac{\left\|y-X\beta(\lambda)\right\|^2}{n\left\{1-p(\lambda)/n\right\}^2} \qquad (5)$$

Where

$Equation(6):$

$$p(\lambda)=tr\left\{X\left\{X^TX+\lambda\left(DLAG\left(\left|\hat{\beta}_1\right|,...,\left|\hat{\beta}_p\right|\right)\right)^{-1}\right\}^{-1}X^T\right\} \qquad (6)$$

we refer to $\frac{\left\|y-X\beta(\lambda)\right\|^2}{n\left\{1-p(\lambda)/n\right\}^2}$ as the GCV statistic and the verification value of the generalized cross model.

III. Experimental simulation and results analysis

A. Index System Construction

If a domestic listed company loses money for two consecutive years, its stock will be treated specially (Special Treatment, ST), and in serious cases, it will face the risk of delisting. ST can reflect the abnormality of the company's monetary status or other conditions. Considering the dynamic nature of the financial crisis, this paper selects the annual financial report data of the T-3th year from 2013 to 2016 during the observation period [11], according to the ratio of ST company to non-ST company 1:1, 80 companies are taken as the study sample, 1200 sample data are taken as a training set to build a model. Descriptive statistical analysis of the original data was performed in the SPSS software, and a small number of missing values were processed using the complement method. The sample data in this paper is mainly from the RESSET database.

The selection of financial risk fore-warning indicators can directly influence the prediction effect of the crisis early warning model, and domestic and foreign scholars have made different attempts to select the indicators. These attempts are not unintentionally randomly to select the indicators; they are tended to choose indicators or related ratios that affect the business conditions of the enterprises. Although the variable screening methods are not uniform, the selection principles are basically similar. The most important considerations in the selection of indicators are those that have a greater correlation with the research issues and tend to select the stable nummary indicators. Based on the chosen experience summarized in the literature, the selection of the financial indicators is based on five aspects.15 index variables are selected as the explanatory variables. These include the liquidity ratio and the quick ratio,etc., which reflect the solvency of the enterprises, the accounts receivable turnover rate and the inventory turnover rate reflecting the operation capability of the enterprises, the net asset yield and the rate of pay on assets reflecting the profitability of the enterprises, the net profit growth rate and the net asset growth rate,etc., which reflect the development ability of the enterprises, the cash recovery rate of the total assets ,etc., which reflect the cash flow level of the enterprises[12]; the specific delimitings are defined in Table I.

B. Experimental Results and Analysis

In this paper, the above 15 variables are used as the explanatory variables, and explained variable Y is a binary discrete variable. The abnormality of the financial status of the enterprise is recorded as 1, and the normal enterprise is recorded as 0. The R software was used to analyze and select the variables, and the Lasso-logistic model was constructed, and the relative accuracy was tested.

In order to determine the value of λ, we use the Gelmnet Package in the R software to perform a 10-fold cross-validation on the sample data, and obtain theopposite trend graph with the number of the explanatory variables shown in Fig. 1. The abscissa is the value of λ, and the ordinate is the different variation of the model error corresponding to the value of λ. For each value of λ, a confidence interval for the target parameter numbers can be obtained around the mean of the target parameter numbers indicated by the red dot. The two dashed lines give two special values of λ. After the value of λ reaches a certain value, continue to increase the number of model explanatory variables, that is, to reduce the value of λ, does not significantly improve the performance of the model [13]. The lambda.1se function in the program can give a model with good performance but the least number of the explanatory variables.

Fig. 2 is a solution path diagram of the Lasso coefficient. Each curve represents the trajectory of the coefficient of an independent variable. The ordinate is the value of the coefficient, and the abscissa is the number of non-zero coefficients under different log (Lambda) values. The coefficient of each variable becomes larger as the value of λ becomes smaller. Combining the two graphs and according to Tibshirani's value experience [14], the most suitable λ value is chosen to be 0.028926. At this time, the Lasso-logistic model parameter estimation results are obtained, and the seven variables with non-zero coefficients are selected. The estimated values are shown in Table II. These variables pass the significance test at 5% level.

TABLE I. FINANCIAL EARLY WARNING INDICATOR SYSTEM

Indicator code	Indicator name	Indicator definition
x_1	Net asset yield	Net profit / Average net assets
x_2	Rate of return on assets	Profit before interest and taxes / Total assets
x_3	Operating profit rate	Operating profit / Total business income
x_4	Liquidity ratio	Current assets / Current liabilities
x_5	Quick ratio	Quick-moving assets/Current liabilities
x_6	Operating income growth rate	(Current operating income - Previous operating income) / Previous operating income
x_7	Net profit growth rate	(Current net profit - Last period net profit) / Last period net profit
x_8	Net asset growth rate	(End-of-year net assets - Net assets at the beginning of the period) / Net assets at the beginning of the period
x_9	Inventory turnover rate	Cost of goods sold / Average inventory balance
x_{10}	Accounts receivable turnover rate	Net sales revenue / Average balance of accounts receivable
x_{11}	Current asset turnover rate	Main business income / Average current assets
x_{12}	Total asset turnover rate	Sales revenue / Total assets
x_{13}	Sales cash ratio rate	Net cash flow from operating activities / Main business income
x_{14}	Total asset cash recovery rate	Net operating cash / Average total assets
x_{15}	Assets and liabilities rate	(Current liabilities + Long-term liabilities) / Total assets

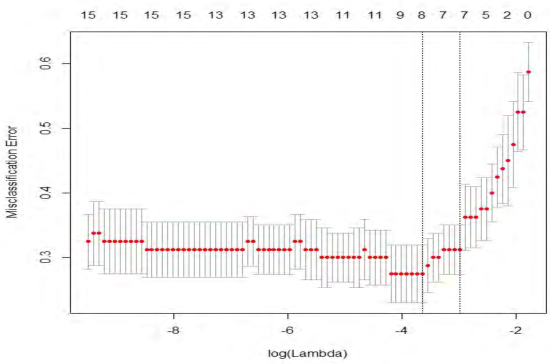

Fig. 1. The value of parameter λ correspond to the number of variables

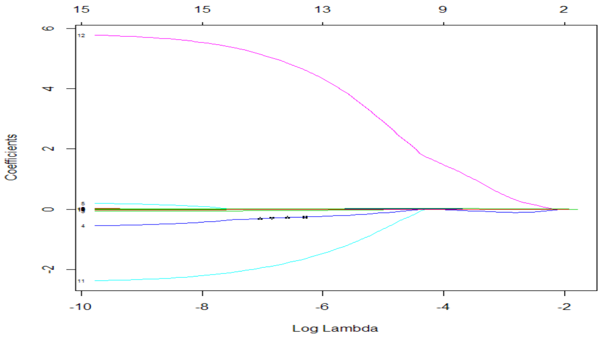

Fig. 2. Lasso Coefficient Solution Path Diagram

It can be seen from Table II, the liquidity ratio, the total asset turnover rate and the asset-liability ratio have a great impact on the financial situation of the enterprise. The total asset turnover rate has the greatest impact. The management and investors should focus on it when they inspect the development level of the enterprise. Secondly, the most important factors are the net assets yield, the operating profit rate, the operating income growth rate and the accounts receivable turnover rate.

TABLE II. PARAMETER ESTIMATES FOR THE LASSO-LOGISTIC MODEL

Variable	Parameter estimate	Variable	Parameter estimate
Intercept term	-2.0184**	X_8	0
X_1	0.0015**	X_9	0
X_2	0	X_{10}	0.0025**
X_3	-0.0034**	X_{11}	0
X_4	-0.00529**	X_{12}	1.0467*
X_5	0	X_{13}	0
X_6	0.0011**	X_{14}	0
X_7	0	X_{15}	0.0232**

[a] Note: 1."*" indicates the variable is significant at 5% level. 2. Coefficient is 0, represents that the variable was removed.

Model fitting effect is an important criterion for judging the quality of the model. The established model is back-calculated with the original data, and the data of the test set is used for predictive analysis [15]. Table III shows the results of the judgment of the training sample data. The results show that the probability of the model misclassification the nummulary normative company as a financial crisis company is 17.5%, that is, the predictive accuracy rate for non-ST companies is 82.5%; The probability of a financial crisis company being judged as a nummulary groovy company is 25%, that is, the forecast accuracy rate for ST companies is 75%. The overall accuracy rate on the training set was 78.8%.

TABLE III. TRAINING SAMPLE JUDGMENT RESULTS

	Predicted		
Observed	0	1	**Accuracy**
0	33	7	82.5%
1	10	30	75%
Total percentage			78.8%

Table IV shows the prediction results of the test sample set. The results show that the accuracy rate for non-ST companies is 100%, the predictive accuracy for ST companies is 80%, and the overall accuracy rate for training sets is 90%. The model predicts better results.

TABLE IV. TEST SAMPLE PREDICTION RESULTS

	Predicted		
Observed	0	1	**Accuracy**
0	10	0	100%
1	2	8	80%
Total percentage			90%

IV. CONCLUSION

In view of the diversity of financial indicators, this paper uses the combination of the Lasso method and the Logistic model to come true predictive modeling while achieving streamlined monetary indicators. The obtained model has better prediction result, and the Lasso method can effectively prevent over-fitting, make the selection of variables simpler and more credible.The financial early-warning model of listed companies established in this paper can enable managers and investors to obtain more accurate and reliable calculates, and also can analyze and forecast the financial status of enterprises earlier. The control of risks has a good effect on the healthy development of the social economy and the steady operation of the national economy. In this paper, only a set of cross-section data is analyzed. If we want to improve the accuracy of prediction, we can consider the use of longitudinal data [16], and more scientifically select the evaluation index system. Under the condition of sufficient sample size, we also can consider combining the characteristics of the industry and discuss separately [17].

ACKNOWLEDGMENT

The research was supported by the “Thirteenth Five-Year Plan” Science and Technology Research Project of the Education of Jilin Province under grant No. 2016103.

REFERENCES

[1] Y.L. Zhang, X.L. You, W. Qiang, Y.Q. Zheng, and Y.T. Li, “Enterprise Financial Crisis Warning and Key Indicator Selection Based on LASSO,” Journal of Henan Normal University (Natural Science Edition), vol. 44(03), pp. 160-165, 2016.

[2] W.F. Shi, X.G. Hu, and K. Yu, “A Method for Equalizing Lasso Feature Selection for High Dimensional data,” Computer Engineering and Applications, vol. 48(01), pp. 157-161, 2012.

[3] Y. Li, J.X. Li, and M. Shuangge, “Research on Enterprise Financial Early Warning Model of Unbalanced Data [J/OL],” Mathematical Statistics and Management, vol. 35(05), pp. 893-906, 2016.

[4] P.W. Cheng and F. Yu, “Application of LASSO and A-LASSO Methods in the Selection of Fnancial Early Warning Model Variables,” China Securities and Futures, vol. (03), pp. 110-111, 2013.

[5] Y.Y. Gu, Financial Early Warning Analysis of Listed SMEs Based on Lasso and Cox Model, Lanzhou University, 2016.

[6] Y.D. Lu and Y.L. Jin, “Empirical Research on Supply Chain Financial Credit Risk Based on Lasso-logistic Model,” Management Modernization, vol. 36(02), pp. 98-100, 2016.

[7] J. Li, Application of Lasso-Logistic and Group Lasso-Logistic Model in Birth Defect Research , Shanxi Medical University, 2016.

[8] T.T. Zhang and Y.C. Jing, “Adaptive Lasso-Logistic Regression Analysis of Personal Credit Scores,” Mathematics in Practice and Theory, vol. 46(18), pp. 92-99, 2016.

[9] W.N. Fang, G.J. Zhang, and H.Y. Zhang, “Personal Credit Risk Early Warning Method Based on Lasso-logistic Model,” Quantitative Economics & Technology Research, vol. 31(02), pp. 125-136, 2014.

[10] H.Y. Mo, J. Wang, and H.L. Niu, “Exponent Back Propagation Neural Network Forecasting for Financial Cross-Correlation Relationship,” Expert Systems With Applications, vol. 53(01), pp. 106-116, 2016.

[11] G. Fu, P. Zeng, and G.L. Chen, “Empirical Study on Enterprise Financial Crisis Early Warning under Economic New Normal,” Journal of Finance and Economics, vol. 09, pp. 88-99, 2016.

[12] Z.M. Qin, Research on the Selection of Financial Early Warning Variables for Listed Companies in China, Dongbei University of Finance and Economics, 2012.

[13] H. Yan, K. Gang, and P. Yi, “Nonlinear Manifold Learning for Early Warnings in Financial Markets,” European Journal of Operational Research, vol. 258(02), pp. 692-702, 2016.

[14] M. Tian and J.Y. Gong, “Strong convergence of a modified proximal algorithm for solving the lasso,” Journal of Inequalities and Applications, vol. (01), pp. 1-15, 2015.

[15] Y. Song, J.M. Zhu, and W. Li, “Research on Enterprise Financial Early Warning Based on Big Data,” Journal of Central University of Finance and Economics, vol. (06), pp. 55-64, 2015.

[16] K.M. Wang and M.G. Ji, “Research on Financial Early Warning of Losing Companies Based on Financial and Non-Financial Indicators—Taking Company ST as an Example,” Journal of Finance and Economics, vol. (07), pp. 63-72, 2006.

[17] K. Tharmaratnam, M. Sperrin, T. Jaki, S. Reppe, and A. Frigessi, “Tilting the Lasso by Knowledge-based Post-processing,” BMC Bioinformatics, vol. 17 (01), pp. 344-344, 2016.