ON THE USING OF ECONOMETRIC MODELS IN THE CREDIT SCORING SYSTEMS

V.I. Malugin, N.V. Hryn
Belorussian State University
Minsk, BELARUS
e-mail: malugin@bsu.by

Abstract

The paper is devoted to the problem of analysis and forecasting of the solvency of banks borrowers on the base of econometric models, which takes into account the influence of the exogenous economic factors on the balance coefficients of borrowers. The credit score algorithms based on multivariate linear regression model and the "plug-in" decision rule for classification of the sample from the mixture of the multivariate regression observations distributions are suggested and examined.

1 Introduction

Credit scoring system are based on automatic classification algorithms, which are used for classification of commercial banks borrowers on the given number of classes of solvency in according with the accessible information. The results of credit scoring are used for decision-making relative to possibility and conditions of granting of credits. Usually used in credit scoring systems algorithms allow to classify borrowers on the basis of the information, concerning only borrowers. For an evaluation of the influence of macroeconomic environment on various categories of borrowers as well as on the credit market as a whole, it is reasonably to use econometric models.

The relationships between the individual financial indicators (e.g. balance sheet coefficients) of borrowers, considered as endogenous variables, and general for all borrowers economic indicators (exogenous variables) may be described by means of different type of multivariate econometric models. Within the frame of this research the multivariate linear regression models are used to describe the stacked vector of features including endogenous and exogenous variables. The problems of classification of banks borrowers on two classes of solvency are considered. The algorithms of discriminant and cluster analysis of multivariate regression observations realizing "plug-in" Bayesian decision rule for classified and non-classified learning samples are suggested and examined.

2 Credit scoring algorithms based on multivariate regression models

Let the potential borrower of bank ω is characterized by a set of individual balance indicators of which it is formed N-dimensional *a vector of factors* $\mathbf{x} = (x_1, ..., x_N)^T =$

$\mathbf{x}(\omega) \in \Re^N$. The purpose credit scoring is reference of the potential borrower of bank ω to one of $L \geq 2$ classes $\{\Omega_i\}$ $(i = 1, ...L)$, which differ with reliability degree, on the basis of the analysis of a vector of factors $\mathbf{x}(\omega), \omega \in \Omega_0 \cup \Omega_1$.

Let's assume further $L = 2$ and Ω_0 – a class of the reliable borrowers possessing high degree of solvency; Ω_1 – a class of unreliable borrowers possessing low degree of solvency. Concerning the borrower ω from a class Ω_0 $(\omega \in \Omega_0)$ full performance of credit obligations is expected. Delivery of the credit to the borrower ω from a class Ω_1 $(\omega \in \Omega_1)$ can be refused, as concerning it default in full credit obligations is expected.

True number of a class $d^0 = d^0(\omega) \in S = \{0, 1\}$, to which the borrower ω belongs, is a discrete random variable with distribution of probabilities:

$$P\{d^0(\omega) = i\} = \pi_i > 0, \ i \in S; \ \pi_0 + \pi_1 = 1,$$

where $\pi_0, \ \pi_1 = 1 - \pi_0$ – *a priori probabilities of classes*. Let's assume that at an estimation (forecasting) of a condition of the borrower at the moment of the termination of term of the credit contract along with a vector of controllable factors $x = (x_1, ..., x_N)^T \in \Re^N$ the vector of exogenous variables $\mathbf{z} = (z_1, ..., z_M)^T \in \mathbf{Z} \subset \Re^M$ is used, describing influence on a condition of borrowers from outside the general external economic factors. We will assume that existing between vectors $x \in \Re^N$ and $z \in \mathbf{Z}$ statistical dependence is described by econometric multivariate linear regression model:

$$\mathbf{x} = B_i \mathbf{z} + \mathbf{u}, \ i \in S = \{0, 1\}, \tag{1}$$

where for a class of borrowers Ω_i: $B_i = (b_{ije})$ – $(N \times M)$-matrix of regression coefficients, $Z \subset \Re^M$ – the limited closed area, such that $(B_2 - B_1)z \neq 0$ with $z \in \mathbf{Z}$; $\mathbf{u} \in \Re^N$ – Gaussian random vector with zero average value and regular covariance $(N \times N)-$ matrix $\Sigma = (\sigma_{jk})$.

The problem of credit scoring consists in reference of the borrower of commercial bank $\omega \in \Omega$ to one of classes $\{\Omega_i\}_{i \in S}$ on set of its indicators $\mathbf{x} = \mathbf{x}(\omega)$, that is in estimation of unknown number of a class $d^0 = d^0(\omega) \in S$ for the borrower ω on known value of its indicators $\mathbf{x} = \mathbf{x}(\omega) \in \Re^N$, where value of a vector of factors $\mathbf{x}_{it} \in \Re^N$ is defined on the basis of (1) for a preset value of a vector of factors $\mathbf{z}_{it} \in \mathrm{Z}$ $\quad (t = 1, ..., T)$.

In the case of known parameters of model (1) a minimum of probability of an error at observation classification $(\mathbf{x}, \mathbf{z})$ is reached for a bayesian decision rule of a kind:

$$d_z = d_z(\mathbf{x}) = \begin{cases} 0, & G_z(\mathbf{x}) < 0, \\ 1, & G_z(\mathbf{x}) \geq 0 \end{cases}$$

with linear discriminant function

$$G_z(\mathbf{x}) = \mathbf{z}^T (B_1 - B_0)^T \Sigma^{-1} \mathbf{x} - \frac{1}{2} \mathbf{z}^T (B_1 + B_0)^T \Sigma^{-1} (B_1 - B_0) \mathbf{z} - \ln \frac{\pi_0}{1 - \pi_0}. \tag{2}$$

True values of parameters $\{\pi_i, B_i, \Sigma\}_{i \in S}$ $\quad (i \in S)$ of model (1) in practice are not known. We will assume that in this case there is *a classified training sample* of values of indicators $X = X_0 \cup X_1$ $(X_i = \{\mathbf{x}_{i1}, ..., \mathbf{x}_{in_i}\},\ i \in S)$ volume $n = n_0 + n_1$ from classes $\Omega_0 \cup \Omega_1$, corresponding to sequence of values of factors $Z = Z_0 \cup Z_1$, $(Z_i = \{\mathbf{z}_{i1}, ..., \mathbf{z}_{in_i}\},\ i \in S)$ where value of a vector of factors $\mathbf{x}_{it} \in \Re^N$ is defined on the basis of (1) for a preset value of a vector of factors $\mathbf{z}_{it} \in \mathrm{Z}\quad (t = 1, ..., T)$. That sample is used for statistical estimation of probability characteristics $\{\pi_i,\ B_i,\ \Sigma\}_{i\in S}\quad (i \in S)$ and construction of a solving rule of classification which then is applied to classification of new borrowers by observation corresponding to them $(\mathbf{x}_\tau,\ \mathbf{z}_\tau),\ \tau = T+1, ..., T+n,\ n \geq\ 1$.

In this case can be constructed a "plug-in" bayesian decision rule by substitution of estimations on a maximum likelihood method $\{\widehat{\pi}_i, \widehat{B}_i\}$, $\widehat{\Sigma}$ the unknown parameters, calculated on training sample of regression observations X, Z volume n.

Results of experimental researches [1] allow to draw a conclusion, that if at an estimation of reliability of borrowers to not consider statistical dependence between indicators of borrowers and external economic conditions or to consider not adequately accuracy of accepted decisions can be not satisfactory. The correct account of statistical dependence between analyzed parameters by using of adequate econometric models can essentially increase accuracy of accepted decisions.

The case when there is not classified training sample of values of factors $X = \{\mathbf{x}_1, ..., \mathbf{x}_T\}$ volume T from classes $\Omega_0 \cup \Omega_1$, which corresponds sequence of values of factors $Z = \{\mathbf{z}_1, ..., \mathbf{z}_T\}$. It can be considered as random sample of a mixture of distributions of regression observations, the density of distribution of probabilities for which looks like

$$p_\pi(\mathbf{x};\mathbf{z},\theta) = \pi_0\varphi_N(\mathbf{x}|\, B_0\mathbf{z},\Sigma) + \pi_1\varphi_N(\mathbf{x}|\, B_1\mathbf{z}, \Sigma),\ \mathbf{x} \in \Re^N,\ \mathbf{z} \in \mathrm{Z}, \qquad (3)$$

where $\theta \in \Theta \subset \Re^m (m = 2MN + N(N+1)/2 + 1)$ – the compound vector of parameters, formed of independent elements of matrixes $\{B_\alpha\}$, Σ and a priori probability π_0 $(\pi_1 = 1 - \pi_0)$; $\varphi_N(\mathbf{x}|\, B_\alpha\mathbf{z}, \Sigma)\quad -$ density of distribution N-dimensional normal distribution with a mean $B_\alpha\mathbf{z}$ and a regular covariance matrix Σ. Thus following tasks of the analysis of a mixture (3) on not classified training sample $\{X, Z\}$ take place:

1) statistical estimation of a vector of parameters $\theta \in \Theta \subset \Re^m$ (calculation of maximum likelihood estimators (ML-estimators) $\left\{\widehat{\pi}_\alpha\right\}, \left\{\widehat{B}_\alpha\right\}, \widehat{\Sigma}$);

2) classifications of training sample $\{X, Z\}$, that is estimation of a vector of classification of sample $\mathbf{d} = (d_t) \in S^T$.

3) classification of new observations $(\mathbf{x}_\tau,\ \mathbf{z}_\tau),\ \tau = T+1, ..., T+n,\ n \geq 1$.

For the decision of tasks 1, 2 the algorithm of splitting of a mixture of distributions of regression observations (3) from a class *EM*-algorithms is offered, allowing simultaneously to calculate ML-estimators of parameters $\{\widehat{\pi}_\alpha\}$, $\{\widehat{B}_\alpha\}$, $\widehat{\Sigma}$ and to carry out classification of training sample $\{X, Z\}$. In algorithm representations for ML-estimators of parameters of a mixture of distributions of regression observations (3), depending on a posteriori probabilities $p_{\alpha t}(\mathbf{x}_t; \mathbf{z}_t)$ references regression observation$(\mathbf{x}_t, \mathbf{z}_t)$ to a class Ω_α are used. A posteriori probabilities for preset values of parameters of a mixture $\{\pi_\alpha\}, \{B_\alpha\}, \Sigma$ are determined under the formula

$$p_{\alpha t}(\mathbf{x}_t, \mathbf{z}_t) = \frac{\exp\left(g'_\alpha \mathbf{x}_t + h_\alpha\right)}{\sum\limits_{\alpha \in S} \exp\left(g'_\alpha \mathbf{x}_t + h_\alpha\right)},\ \mathbf{x}_t \in \Re^N,$$

$$g_\alpha = \Sigma^{-1} B_\alpha \mathbf{z}_t,\ h_\alpha = \frac{1}{2}\mathbf{z}'_t B'_\alpha \Sigma^{-1} B_\alpha \mathbf{z}_t + \ln\left(\pi_\alpha\right),\ \mathbf{z}_t \in \mathrm{Z},\ t = 1, ..., T,\ \alpha \in S.$$

For preset values $\{p_{\alpha t}\}$ $(t = 1, ..., T,\ \alpha \in S)$ ML-estimators $\left\{\widehat{\pi}_\alpha\right\}, \left\{\widehat{B}_\alpha\right\}, \widehat{\Sigma}$ parameters of a mixture (3) suppose representations

$$\widehat{B}_\alpha = \sum_{t=1}^{T} p_{\alpha,t} \mathbf{x}_t \mathbf{z}'_t \left(\sum_{t=1}^{T} p_{\alpha,t} \mathbf{z}_t \mathbf{z}'_t \right)^{-1},$$

$$\widehat{\Sigma} = \sum_{\alpha=0}^{1} \sum_{t=1}^{T} p_{\alpha,t} \left(\mathbf{x}_t - \widehat{B}_\alpha \mathbf{z}_T \right) \left(\mathbf{x}_t - \widehat{B}_\alpha \mathbf{z}_T \right)',$$

$$\widehat{\pi}_0 = \frac{1}{T} \sum_{t=1}^{T} p_{0,t} \left(\widehat{\pi}_1 = 1 - \widehat{\pi}_0 \right).$$

Application of the offered algorithms leads to an establishment of statistical dependencies which adequacy can be researched by using of a standard set of the tests applied to check of adequacy of regression models, including the analysis of the statistical importance of factors of regress, the analysis of the rests, etc.

One of key problems of the discriminant analysis tasks is presence of representative training sample of supervision. Carried out researches [2] testify to impossibility correct classification in case of unsatisfactorily prepared training sample. Use of the such iterative algorithms based on methods of the cluster analysis of multivariate regression observations [3], can essentially increase quality of training sample. In this case initial classification can be set by expert opinion.

References

[1] Malugin V.I., Hryn N.V. (2008). The analysis and forecasting of credit risk on the basis of econometric models *Economy, modelling, forecasting: the collection of scientific papers of Scientific and research economic institute of Ministry of Economy of the Republic of Belarus* . Vol. **2**, pp. 260–277.

[2] Malugin V.I., Korchagin O.I., Hryn N.V. (2008). Examining efficiency of algorithms of classifying of bank borrowers (on the basis of balance ratios) *The Bank bulletin.* Vol. **7**, pp. 26–33.

[3] Malugin V.I. (2008). Statistical analysis of the mixtures of the regression observations. *Informatics.* Vol. **4 (20)**, pp. 79–88.