Financial Ratio Selection for Distress Classification

Roberto Kawakami Harrop Galvão
Instituto Tecnológico de Aeronáutica, Brazil

Victor M. Becerra
University of Reading, UK

Magda Abou-Seada
Middlesex University, UK

INTRODUCTION

Prediction of corporate financial distress is a subject that has attracted the interest of many researchers in finance. The development of prediction models for financial distress started with the seminal work by Altman (1968), who used discriminant analysis. Such a technique is aimed at classifying a firm as bankrupt or nonbankrupt on the basis of the joint information conveyed by several financial ratios.

The assessment of financial distress is usually based on ratios of financial quantities, rather than absolute values, because the use of ratios deflates statistics by size, thus allowing a uniform treatment of different firms. Moreover, such a procedure may be useful to reflect a synergy or antagonism between the constituents of the ratio.

BACKGROUND

The classification of companies on the basis of financial distress can be performed by using linear discriminant models (also called *Z*-score models) of the following form (Duda, Hart, & Stork, 2001):

$$Z(\mathbf{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{x} \qquad (1)$$

where $\mathbf{x} = [x_1\, x_2 \ldots x_n]^{\mathrm{T}}$ is a vector of n financial ratios, $\boldsymbol{\mu}_1 \in \Re^n$ and $\boldsymbol{\mu}_2 \in \Re^n$ are the sample mean vectors of each group (continuing and failed companies), and $\mathbf{S}_{n\times n}$ is the common sample covariance matrix. Equation 1 can also be written as

$$Z = w_1x_1 + w_2x_2 + \ldots + w_nx_n = \mathbf{w}^{\mathrm{T}}\mathbf{x} \qquad (2)$$

where $\mathbf{w} = [w_1\, w_2 \ldots w_n]^{\mathrm{T}}$ is a vector of coefficients obtained as

$$\mathbf{w} = \mathbf{S}^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \qquad (3)$$

The optimal cut-off value for classification z_c can be calculated as

$$z_c = 0.5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\mathbf{S}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \qquad (4)$$

A given vector $\mathbf{x}$ should be assigned to Population 1 if $Z(\mathbf{x}) > z_c$, and to Population 2 otherwise.

The generalization (or prediction) performance of the *Z*-score model, that is, its ability to classify objects not used in the modeling phase, can be assessed by using an independent validation set or cross-validation methods (Duda et al., 2001). The simplest cross-validation technique, termed "leave-one-out," consists of separating one of the m modelling objects and obtaining a *Z*-score model with the remaining $m-1$ objects. This model is used to classify the object that was left out. The procedure is repeated for each object in the modeling set in order to obtain a total number of cross-validation errors.

Resampling techniques (Good, 1999) such as the Bootstrap method (Davison & Hinkley, 1997) can also be used to assess the sensitivity of the analysis to the choice of the training objects.

The Financial Ratio Selection Problem

The selection of appropriate ratios from the available financial information is an important and nontrivial stage in building distress classification models. The best choice of ratios will normally depend on the types of companies under analysis and also on the economic context. Although the analyst's market insight plays an important role at this point, the use of data-driven selection techniques can be of value, because the relevance of certain ratios may only become apparent when their joint contribution is considered in a multivariate context. Moreover, some combinations of ratios may not satisfy the statistical assumptions required in the modeling process, such as normal distribution and identical covariances in the groups being classified, in the case of standard linear discriminant analysis (Duda et al., 2001). Finally, collinearity between ratios may cause the model to have poor prediction ability (Naes & Mevik, 2001).

Techniques proposed for ratio selection include normality tests (Taffler, 1982), and clustering followed by stepwise discriminant analysis (Alici, 1996).

Most of the works cited in the preceding paragraph begin with a set of ratios chosen from either popularity in the literature, theoretical arguments, or suggestions by financial analysts. However, this article shows that it is possible to select ratios on the basis of data taken directly from the financial statements.

For this purpose, we compare two selection methods proposed by Galvão, Becerra, and Abou-Seada (2004). A case study involving 60 failed and continuing British firms in the period from 1997 to 2000 is employed for illustration.

MAIN THRUST

It is not always advantageous to include all available variables in the building of a classification model (Duda et al., 2001). Such an issue has been studied in depth in the context of spectrometry (Andrade, Gomez-Carracedo, Fernandez, Elbergali, Kubista, & Prada, 2003), in which the variables are related to the wavelengths monitored by an optical instrumentation framework. This concept also applies to the *Z*-score modeling process described in the preceding section. In fact, numerical ill-conditioning tends to increase with $(m - n)^{-1}$, where m is the size of the modeling sample, and n is the number of variables (Tabachnick & Fidell, 2001). If $n > m$, matrix $\mathbf{S}$ becomes singular, thus preventing the use of Equation 1. In this sense, it may be more appropriate to select a subset of the available variables for inclusion in the classification model.

The selection procedures to be compared in this article search for a compromise between maximizing the amount of discriminating information available for the model and minimizing collinearity between the classification variables, which is a known cause of generalization problems (Naes & Mevik, 2001). These goals are usually conflicting, because the larger the number of variables, the more information is available, but also the more difficult it is to avoid collinearity.

Algorithm A (Preselection Followed by Exhaustive Search)

If N variables are initially available for selection, they can be combined in $2^N - 1$ different subsets (each subset with a number of variables between 1 and N). Thus, the computational workload can be substantially reduced if some variables are preliminarily excluded.

In this algorithm, such a preselection is carried out according to a multivariate relevance index $W(x)$ that measures the contribution of each variable x to the classification output when a *Z*-score model is employed. This index is obtained by using all variables to build a model as in Equation 1 and by multiplying the absolute value of each model weight by the sample standard deviation (including both groups) of the respective variable.

An appropriate threshold value for the relevance index $W(x)$ can be determined by augmenting the modeling data with artificial uninformative variables (noise) and then obtaining a *Z*-score model. Those variables whose relevance is not considerably larger than the average relevance of the artificial variables are then eliminated (Centner, Massart, Noord, Jong, Vandeginste, & Sterna, 1996).

After the preselection phase, all combinations of the remaining variables are tested. Subsets with the same number of variables are compared on the basis of the number of classification errors on the modelling set for a *Z*-score model and the condition number of the matrix of modeling data. The *condition number* (the ratio between the largest and smallest singular value of the matrix) should be small to avoid collinearity problems (Navarro-Villoslada, Perez-Arribas, Leon-Gonzalez, & Polodiez, 1995). After the best subset has been determined for each given number of variables, a cross-validation procedure is employed to find the optimum number of variables.

Algorithm B (Genetic Selection)

The drawback of the preselection procedure employed in Algorithm A is that some variables that display a small relevance index when all variables are considered together could be useful in smaller subsets. An alternative to such a preselection consists of employing a genetic algorithm (GA), which tests subsets of variables in an efficient way instead of performing an exhaustive search (Coley, 1999; Lestander, Leardi, & Geladi, 2003).

The GA represents subsets of variables as individuals competing for survival in a population. The genetic