Data Mining for Internationalization

Luciana Dalla Valle
University of Milan, Italy

INTRODUCTION

The term "internationalization" refers to the process of international expansion of firms realized through different mechanisms such as export, strategic alliances and foreign direct investments. The process of internationalization has recently received increasing attention mainly because it is at the very heart of the globalization phenomenon. Through internationalization firms strive to improve their profitability, coming across new opportunities but also facing new risks.

Research in this field mainly focuses on the determinants of a firms' performance, in order to identify the best entry mode for a foreign market, the most promising locations and the international factors that explain an international firms' performance. In this way, scholars try to identify the best combination of firms' resources and location in order to maximize profit and control for risks (for a review of the studies on the impact of internationalization on performance see Contractor et al., 2003).

The opportunity to use large databases on firms' international expansion has raised the interesting question concerning the main data mining tools that can be applied in order to define the best possible internationalization strategies.

The aim of this paper is to discuss the most important statistical techniques that have been implemented to show the relationship among firm performance and its determinants.

These methods belong to the family of multivariate statistical methods and can be grouped into *Regression Models* and *Causal Models*. The former are more common and easy to interpret, but they can only describe direct relationships among variables; the latter have been used less frequently, but their complexity allows us to identify important causal structures, that otherwise would be hidden.

BACKGROUND

We now describe the most basic approaches used for internationalization. Our aim is to give an overview of the statistical models that are most frequently applied in International Finance papers, for their easy implementation and the straightforwardness of their result interpretation. In this paragraph we also introduce the notation that will be used in the following sections.

In the class of *Regression Models,* the most common technique used to study internationalization is the *Multiple Linear Regression*. It is used to model the relationship between a continuous response and two or more linear predictors.

Suppose we wish to consider a database, with N observations, representing the enterprises, the response variable (firm performance) and the covariates (firm characteristics). If we name the dependent variable Y_i, with $i = 1, ..., N$, this is a linear function of H predictors $x_1, ..., x_H$, taking values $x_{i1}, ..., x_{iH}$, for the i–th unit. Therefore, we can express our model in the following way:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_H x_{iH} + \varepsilon_i, \qquad (1)$$

where $\beta_1, ..., \beta_i$ are the regression coefficients and ε_i is the error term, that we assume to be normally distributed with mean zero and variance σ^2 (Kutner et al., 2004).

Hitt, Hoskisson and Kim (1997), for example, applied the technique shown above to understand the influence of firm performance on international diversification, allowing not only for linear, but also for curvilinear and interaction effects for the covariates.

When our dependent variable is discrete, the most appropriate regression model is the *Logistic Regression* (Giudici, 2003).

The simplest case of Logistic Regression is when the response variable Y_i is binary, so that it assumes only the two values *0* and *1*, with probability p_i and $1\text{-}p_i$, respectively.

We can describe the Logistic Regression model for the *i*-th unit of interest by the logit of the probability p_i, linear function of the predictors:

$$\log it(p_i) = \log \frac{p_i}{1-p_i} = \eta_i = \underline{x}_i' \underline{\beta}, \qquad (2)$$

where $\underline{x}_i$ is the vector of covariates and $\underline{\beta}$ is the vector of regression coefficients.

For more details about Logit models the reader can consult, for example, Mardia, Kent and Bibby (1979) or Cramer (2003).

For an interesting application, the reader can consult the study by Beamish and Ruihua (2002), concerning the declining profitability from foreign direct investments by multinational enterprises in China. In their paper the authors used subsidiaries' performance as the response variable, with the value "1" indicating "profitable" and value "0" denoting "break-even" or "loss" firms. Beamish and Ruihua show that the most important determinants of profitability are subsidiaries-specific features, rather than macro-level factors.

Concerning *Causal models*, *Path Analysis* is one of the first techniques introduced in the internationalization field.

Path Analysis is an extension of the regression model, where causal links are allowed between the variables. Therefore, variables can be at the same time both dependent and independent (see Dillon and Goldstein, 1984 or Loehlin, 1998).

This approach could be useful in an analysis of enterprise internationalization, since, through a flexible covariate modeling, it shows even indirect links between the performance determinants.

Path Analysis distinguishes variables into two different types: *exogenous* and *endogenous*. The former are always independent and they do not have explicit causes; the latter can be dependent as well as independent and, since they are stochastic, they are characterized by a disturbance term as an uncertainty indicator.

The association between a couple of endogenous variables caused by an exogenous one is defined as *spurious correlation*. This particular correlation is depicted by an arrow in a circle-and-arrow figure, called a *path diagram* (see figure 1).

The object of Path Analysis is to find the best causal model through a comparison between the regression weights predicted by some models and the observed correlation matrix for the variables. Those regression weights are called *path coefficients* and they show the direct effect of an independent variable on a dependent variable in the path model.

For more details about Path Analysis, see, for example, Bollen (1989) or Deshpande and Zaltman (1982).

MAIN FOCUS

In this section we analyze the statistical methodologies most recently used for internationalization, with particular attention to the applications.

Regression Models

Logistic Regression is very useful when we deal with a binary response variable, but when we consider a

Figure 1: The path diagram. A, B and C are correlated exogenous variables; D and E are endogenous variables caused by A, B and C.