1 Credit Risk Management through Robust Generalized Linear Models

Luigi Grossi[1] and Tiziano Bellini[2]

[1] Dipartimento di Economia,
Università di Parma, Italy
luigi.grossi@unipr.it
[2] Ufficio Risk Management,
Banca Monte Parma, Italy
tiziano.bellini@monteparma.it

Abstract: In this work, a robust methodology is developed for the classification of a sample of small and medium firms on the basis of their default probability. The importance of this classification procedure is emphasized by the New Basel Capital Accord (Basel II) for the capital adequacy of internationally active banks. The Basel accord introduces the possibility to adopt models of internal rating for the estimation of the default probability of customers' banks. The reference framework of this paper is the class of generalized linear models which allows to classify units avoiding strict assumptions such those required by the linear discriminant analysis. Another advantage of generalized linear models is the possibility to explore different links between the expected value of the dependent variable and the linear predictor. Parameters are estimated using balance ratios and data coming from Centrale dei Rischi for a set of firms which are customers of a medium sized bank of Northern Italy. Finally, we perform a robust analysis of the model estimates through the forward search in order to monitor the influence of outliers on the final classification.

Keywords: Insolvency prediction, forward search, rating system, robust estimation.

1.1 Introduction

A financial institution deciding whether to supply credit assesses if the potential borrower will be able to redeem the credit. According to this goal, financial institutions are engaged in developing rating systems which graduate customers on the basis of their future ability to refund the money supplied and may be applied to classify potential new customers. The main issue of a credit rating system is to identify criteria which separate "good" creditors from "bad" creditors. This issue, apart from its theoretical attractiveness, is gaining importance in financial institutions considering the role of rating systems, not only in day by day lending activity, but also in determining

the adequacy of regulatory capital under the Basel Capital Accord (Bank of International Settlements, 2004). Using financial and non-financial risk factors of a sample of more than 600 firms extracted from a financial institution database merged with "Centrale dei Rischi" database, we adopt generalized linear models in order to classify healthy and potentially insolvent firms in classes according with their default probability. One of the most relevant innovation of this paper is the introduction of a robust analysis for distress prediction methods using the forward search methodology (Atkinson and Riani, 2002). The main contribution of the forward search in the framework of rating system is the possibility to improve the classification rule avoiding the influence of outlying firms.

1.2 Brief description of the method: generalized linear models and the forward search

In the literature about insolvency prediction, linear discriminant analysis models and multiple logistic regression models have been widely used to discriminate between failed and non-failed firms on the basis of financial ratios. Altmans popular Z-Score and Ohlsons O-Score are, for example, respectively based on linear discriminant analysis and on logistic regression. Neural network models have become a popular alternative with the ability to incorporate a very large number of features in an adaptive nonlinear model. For a survey of business failure classification models see, for example, Hand and Henley (1997) and, more recently, Giudici (2003). In the present paper, we adopt generalized linear models (GLM) which are a family of models including logistic regression as a special case. The choice of GLM can be justified in various ways: the first relevant reason is that it is a good compromise between linear discriminant analysis, which can be applied only under strict conditions (such as equality of covariance matrices) and nonlinear methods (such as neural networks and genetic algorithms) which are nonparametric models and generally show good forecasting performances, but are black-boxes hard to interpret. Another reason to prefer GLM with respect to discriminant analysis is that GLM methodology allows to use categorical variables. With respect to previous papers based on logistic regression, two aspects of the present paper are original:

1. the application of a robust analysis based on the forward search,
2. the introduction of links different from logit which could lead, for some data sets, to better forecasting performances.

When handling insolvency data, it is natural to label one of the categories as success (healthy) and the other as failure (default). Generally speaking, let Y be the binary response variable which can assume two values according to a particular event which can happen (success) or not (unsuccess) defined as follows:

$$y_i = \begin{cases} 0 & \text{unsuccess} \\ 1 & \text{success} \end{cases} \tag{1.1}$$

Let $E(Y_i) = \mu_i$ and x_i is a vector containing the values of the explanatory variables for the $i-th$ unit and β the corresponding parameter vector, the linear predictor η_i and the mean μ_i are related by the link function

$$g(\mu_i) = \eta_i = x_i'\beta. \tag{1.2}$$

Binary data, such as defaulting and healthy firms, require a link such that the mean lies between zero and one. The most widely used links for binary data that satisfy this properties are logit, log-log and complementary log-log (cloglog from now on) which are reported as follows:

$$\begin{aligned} g(\mu_i) &= \log\left(\tfrac{\mu_i}{1-\mu_i}\right), && \text{logit} \\ g(\mu_i) &= \log(\log(\mu_i)), && \text{log-log} \\ g(\mu_i) &= \log(-\log(1-\mu_i)), && \text{clog-log} \end{aligned}$$

As GLM estimates are strongly influenced by outliers, we apply a forward search analysis. The forward search is a general method which has been introduced originally in linear regression models and subsequently extended to other fields of statistics such as multivariate techniques (Atkinson *et al.* (2004)), structural time series models (Riani, (2004)) and financial time series models (Grossi and Laurini (2004)). The main steps of this procedure are:

1. identification of a basic subset free from outliers. In the case of linear regression models least median of squares estimators are used in order to select the units belonging to the basic subset;
2. ordering of observations according to their degree of accordance to the underlying model using, in the case of linear regression, squared residuals computed on a subset of m observations. The subset size is increased from m to $m+1$ by selecting the least outlying observations from the previous graduation.
3. monitoring of statistics, such as parameter estimates, t-values, and so on along each step of the search.

The output is a complete monitoring of estimates which do not suffer from masking and smearing effects typical of classical backward methods for outlier detection. The forward search for generalized linear models is similar to that for linear regression except that we replace squared least squares residuals with squared deviance residuals. Another point which deserves to be stressed with respect to linear regression models is that, as we are analyzing binary data, we have to avoid including during the search only observations of one kind. This can be done through a balanced search in order to maintain a balance of both kinds of firms (bad and good), that is the ratio of bad and good in the various subsets of the procedure is maintained as close as possible to

the ratios in the complete set of n observations (Riani and Atkinson (2001)). For a detailed explanation of the steps of the forward search in generalized linear models, see Atkinson and Riani (2000).

1.3 Data, model application and main results

1.3.1 Brief description of the data

The sample is composed of 653 firms from a database of a medium sized bank. Sample units have been selected randomly stratifying by industry and by size and excluding both small (turnover under 1 million) and large companies (turnover over 10 million). The remaining firms represent the 10 most important business areas defined by the Bank of Italy. The sample represents roughly 40% of the universe of firms with the former requirements. Because the model we use is based on maximum likelihood and does not work with missing values, we omit from the analysis units which contain a missing value in at least one of the explanatory variables. The number of firms on which we conduct the analysis becomes 523. We merge the December 2004 release of the financial institution database with the December 2004 release of "Centrale dei Rischi" database. The first database contains qualitative variables such as juridical form, economic sector, etc. joined with information on balance sheet for years 2001, 2002, 2003. Into the second we find the monthly history of firms financial exposure with respect to the whole financial system over the period January 2004-November 2004. The total sample of firms has been divided into two groups: healthy and insolvent firms. The first one, excluding firms with missing values, is formed by 456 firms and the second by 67. We define insolvent the firms both defaulted (bankrupted or not) and likely to become defaulting in the near future according with the financial institution qualitative credit rating. Because of the large number of variables found to be sensible indicators of corporate economic capabilities, we decide to use some popular indicators concerning liquidity, profitability, leverage and solvency obtained from balance sheet. We concentrate on year 2003 data. To introduce a dynamic view of the economic performance, we calculate the variation of balance ratios in year 2003 with respect to the previous year 2002. In addition, we use some "Centrale dei Rischi" indicators in order to give evidence of the relationship between the amount of credit the overall banking system offers to the firm (Accorded), the amount of credit used by firms (Usage), the credit usage over the limit fixed by the bank (OverUsage) and warranties supplied (Warranties). In particular, we resume the monthly variables computing the arithmetic mean in the period January 2004-November 2004 and focus the analysis on the following ratios: Usage/Accorded, OverUsage/Accorded, Warranties/Accorded. The final design matrix is formed by 38 variables: juridical form, economic sector, 15 balance ratios, 15 variations of balance ratios, 6 ratios derived from Centrale

dei Rischi database. The complete list of variables is not reported for lack of space, but it is available by the authors.

1.3.2 Construction of explanatory variables

Considering the high number of variables, a first challenge of our analysis is to verify whether some variables are more powerful in discriminating healthy from insolvent firms. This selection has been made by means of a forward variable selection method based on the likelihood ratio under the hypothesis of the logit link. That procedure leads to the choice of five variables:

- spread in 2003;
- banking debts over turnover in 2003;
- average credit delay in 2003;
- variation of the duration of monetary cycle in 2003 with respect to 2002;
- usage/accorded with respect to the whole banking system in 2004.

Spread in 2003 is a profitability indicator that is obtained deducting from ROA (operational income/total assets) the cost of debt (financial interests/debts). This ratio stresses that firms with higher operational profitability, cleaned from financial interest costs, are usually more likely to be solvent than less profitable firms. The relationship between banking debts at the end of the year 2003 and year turnover gives evidence of firm's financial structure. From an economical point of view, traditionally, firms with a weak financial structure are usually more likely to become insolvent than those with stronger financial structure. The delay accorded to customers to pay their debt is usually expression of the trading power of the firm into the market. When the firm has some marketing difficulties the average delay accorded usually increases. The monetary cycle duration is obtained by difference between the delay in paying suppliers (in days) and the sum of:

- delay accorded to customers (in days);
- stock period (in days).

Our analysis emphasizes the role played by the evolution of the duration of monetary cycle from year 2002 till year 2003. The growth of monetary cycle duration shows financial difficulties that will probably lead to insolvency. The ratio usage/accorded stresses the relationship between credit accorded by the whole banking system to the firm and its use of the credit. When a firm reaches high level of credit usage (compared to accorded credit), it is likely that its financial capabilities are not really strong and it will probably fall into insolvency.

1.3.3 Application of the forward search to GLM

Using the selected indicators we estimate three generalized linear models for binary data applying the links cited in section 1.2.

Fig. 1.1. *Goodness of link test along the forward search: logit, cloglog and loglog links.*

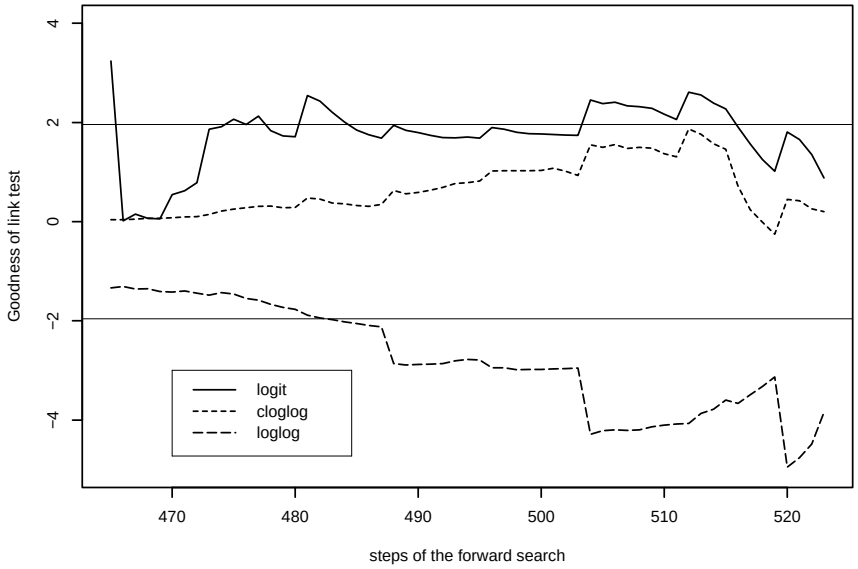


Figure 1.1 reports the goodness of link test during the last 60 steps of the forward search. According to the forward search approach, the first values of the lines represented on the figure are computed on the basis of the first 464 observations included in the subset ordered considering the degree of accordance to the model omitting the remaining observations. At each step of the search further observations are included in the subset by respecting the rule of minimising the squared deviance residuals. In the last step, which in the figure corresponds to the extreme right values of the lines, the test is computed using all observations. Horizontal lines indicate 5% asymptotic confidence region. As can be noted, the cloglog link (dotted lines) always lies inside the region during the search, the logit link lies across the upper limit, whereas the loglog link lies outside the region in the majority of the steps and must be surely rejected. Thus, the best link looks to be the cloglog, even if the logit is very close to the acceptance region. It is worth stressing that the difference of the three links is not so clear observing the values based on the entire sample (last value of the lines).

Logit and cloglog models have been both estimated and lead to very similar results (see Table 1.1): the total classification error is about 7% with both models when all units are included in the estimation procedure. Neverthe-

Table 1.1. *Classification error of logit and cloglog links at the end of the forward search.*

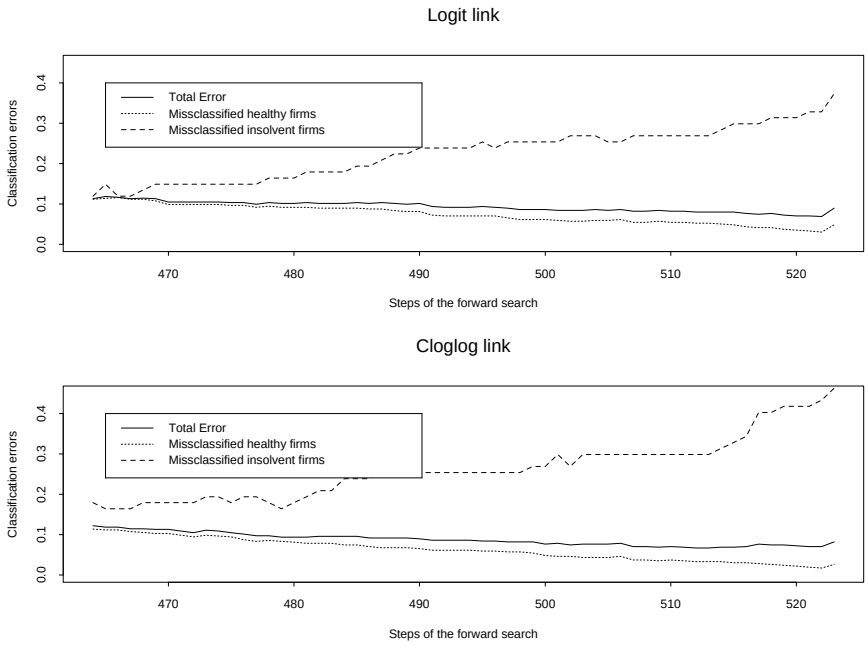
LOGIT				
			Predicted	
		Good	Sufference	Total
Observed	Good	0.97	0.03	1
	Sufference	0.33	0.67	1
	Total	0.89	0.11	1

CLOGLOG				
			Predicted	
		Good	Sufference	Total
Observed	Good	0.98	0.02	1
	Sufference	0.43	0.57	1
	Total	0.91	0.09	1

less, the logit link returns a classification error of insolvent firms (insolvent firms classified as healthy) substantially lower than that given by the cloglog link: 33% vs 43% given all observations. Thus, the cloglog better classifies healthy firms (classification error of healthy firms: 1.7% vs 3.1%) and worse the insolvent. From the financial institution point of view, it is more serious to misclassify an insolvent firm as healthy than the opposite and the logit link should be preferred. The analysis of classification errors given till now does not consider the presence of outliers. To this purpose the robust analysis of the forward search could be very useful.

Figure 1.2 shows trajectories of classification errors during the forward search for logit and cloglog links. It is very interesting to note that adding observations to the initial subset causes a slow decrease of total classification error and error in classifying healthy firms (lower lines). On the opposite the proportion of misclassified insolvent firms increases very quickly: in the case of the logit link this error goes from about 10% when the subset is composed by 460 observations to 33% corresponding to estimates on the whole sample. This particular behavior can be explained considering that observations which are included in the last steps are healthy firms which are financially similar to insolvent firms and wrongly influence the classification rule. For example, the last observation included by the forward search is a healthy firm with debts to banks three times greater than total sales: this ratio is clearly greater than the maximum values among insolvent firms. Finally, observing Figure 2, we gain that the cloglog link is more influenced by outliers with respect to logit: the trajectory of misclassified insolvent firms presents a neat jump around step 515 and, around step 500, classification error of insolvent firms with the two links is equivalent.

Fig. 1.2. *Classification errors during the forward search: logit (upper panel) and cloglog (lower panel) link.*



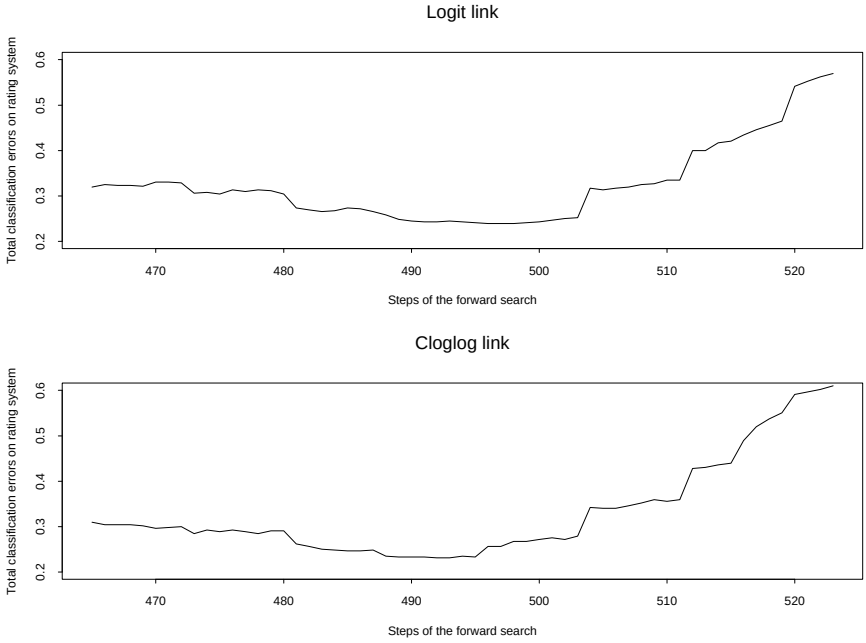
1.3.4 The forward search on rating system

The Basel Agreement states that banks can use internal rating systems to determine the regulatory capital according to the ability of customers to refund their debts (see for example Altman and Saunders (2001)). In this section we evaluate the ability of the model estimated in the previous section to forecast the right rating class of borrowers. This goal can be reached comparing the average probability of default (PD) observed in each category of customer according to the internal qualitative rating system with the posterior probability of default predicted by the model. The qualitative classification of customers given by the bank identifies four classes of customers called A, B, C and DEF, going from the best to the worst customers, the last class being composed by defaulted firms. The rating class for each firm has been predicted by the model according to the empirical PD observed in each class, that is: ($PD < 0.0023$): class A, ($0.0023 < PD < 0.02$): class B, ($0.02 < PD < 0.28$): class C, ($PD > 0.28$): DEF.

Figure 1.3 reports the forward trajectories of the total error obtained forecasting the qualitative rating system of the banks through the logit (upper panel) and the cloglog link (lower panel). As can be noted, the trajectories remain roughly stable until step 500 and begin to increase sharply after that

step, that is when the most outlying observations are included in the subset. Thus, the best forecasting performance of the model is reached estimating the parameters on a subset of 500 observation out of the total sample size. Note that the forecasting error of the model is computed at each step of the procedure by excluding the observations less in accordance with the underlying model, which can be considered as out of sample. Therefore, the forward search gives in-sample forecasts for observations included in the main subset and out-of-sample forecasts for the remaining units.

Fig. 1.3. *Total classification errors using the categories of the internal rating system along the last 60 steps the forward search: logit (upper panel) and cloglog (lower panel) link.*



1.4 Final remarks and extensions for further research

Starting from the statements of the Basel Agreement, we have analyzed a method to classify bank customers according to their future ability to refund money. In the literature about rating system the problem of influential observations has not been deeply analyzed. Notwithstanding outliers can strongly bias model estimates and the forecasting performance of the procedure. In

this paper a robust analysis of generalized linear models for the classification of firms has been presented. The robustification of the models has been made through the forward search, which is an iterative procedure allowing to monitor the influence of single o group of observations on the model estimates. Among all possible link functions which can be used in the GLM framework, the application of a goodness of link test suggested by Atkinson and Riani (2000) has restricted the attention to logit and cloglog links. The forward search analysis has shown that the presence of outlying firms can dramatically influence the classification errors, particularly that which leads to misclassify insolvent firms. In this paper five variables have been selected on the basis of the likelihood ratio under the hypothesis of a logit link and without considering the influence of outliers. Greater attention deserve in the next future the selection of variables which should be integrated in the forward search. Another point which should be deepened in future research regards the development of calibratory tools to evaluate how significant is the influence of observations on parameters. This will be done by means of simulated envelopes.

References

ALTMAN, E.I. and SAUNDERS, A. (2001): An analysis and critique of the BIS proposal on capital adequacy and ratings. *Journal of Banking and Finance, 25, 25–46.*

ATKINSON, A.C. and RIANI, M. (2000): *Robust Diagnostic Regression Analysis.* Springer Verlag, New York.

ATKINSON, A.C., RIANI, M. and CERIOLI, A. (2004): *Exploring Multivariate Data with the Forward Search.* Springer Verlag, New York.

BANK for INTERNATIONAL SETTLEMENTS (2004): International Convergence of Capital Measurement and Capital Standards: a Revised Framework. *Report, Basel, http://www.bis.org/publ/bcbs107.pdf.*

GIUDICI, P. (2003): *Applied Data Mining.* Wiley, West Sussex, U.K.

GROSSI, L. and LAURINI, F. (2004): Effects of extremal observations on the test for heteroscedastic components in economic time series. *Applied Stochastic Models in Business and Industry, 20, 115–130.*

HAND, D. and HENLEY, W. (1997): Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society, Series A, 160, 523–541.*

RIANI, M. (2004): Extension of the forward search to time series. *Studies in non Linear Dynamics and Econometrics, 8, Article 1.*

RIANI, M. and ATKINSON, A.C. (2001): A unified approach to outliers, influence, and transformations in discriminant analysis. *Journal of Computational and Graphical Statistics, 10, 513–544.*