

Extending logistic approach to risk modelling through semiparametric mixing

Marco Alfò⁽¹⁾, Stefano Caiazza⁽²⁾, Giovanni Trovato⁽²⁾

⁽¹⁾Dipartimento di Statistica, Probabilità e Statistiche Applicate,
Università degli Studi "La Sapienza" di Roma,

⁽²⁾Dipartimento di Economia e Istituzioni, Università di Roma "Tor Vergata"

Abstract

The New Proposal of Basel Committee on banking regulation issued in January 2001 allows banks to use Internal Rating Systems to classify firms. Within this context, the main problem is to find a model that fits data as better as possible, providing at the same time good prediction and explicative capabilities. In this paper, our aim is to compare two kind of classification models applied to credit worthiness using weighted classification error as performance function: the standard logistic model and a mixed logistic model, adopting respectively a parametric and a semiparametric approach. As it is well known, the main problem of the former is related to the assumption of i.i.d. hypothesis, while it often turns out necessary to consider the possible presence of unobservable heterogeneity, that characterizes microeconomic data. To better consider this phenomenon we defined and applied a random effect logistic model, avoiding parametric assumptions upon the random effect distribution. This leads to a likelihood which is defined as the integral of the kernel density with respect to the mixing density which has no analytical solution. This problem can be obviated by approximating the integral with a finite sum of kernel densities, each one characterized by a different set of model parameters. This discrete nature helps us in detecting non-overlapping clusters characterized by homogeneous values of insolvency risk, and in classifying firms to one of these clusters by means of estimated posterior probabilities of component membership.

Keywords: bankruptcy risk, logistic model, finite mixtures, nonparametric maximum likelihood.

1 Introduction

Random effects models are frequently used to analyze complex data structures in the presence of significant sources of heterogeneity among individuals. Such models have been introduced in a wide variety of empirical applications, ranging from overdispersed to clustered observations. One of the possible applications of these models is in the credit risk framework. It has recently known a great interest due to the relevant impact of unsound credits on banks balances and to the proposal to modify the minimum regulatory capital by Basel Committee (2001).

The new Basel Proposal (Basel Committee on Banking Supervision, 2001) and its latest revision in April 2003 (2003) is based on the so called three-pillar approach to capital adequacy:

- i) the minimum capital requirements;
- ii) the review of the supervisory process of internal bank assessments of capital;
- iii) the market disclosure involving the quality of information provided to the market.

One of the most important innovations of the first pillar is the chance for banks to develop an internal rating system. The output of this system should "(...) play an essential role in credit approval, risk management, internal capital allocations, and corporate governance functions of banks (...)" (Basel Committee, 2003, pagg. 203, note 460). The procedure to define an internal rating system can be basically divided into three steps (Moody's, 2000):

- i) choice of the classification model which assigns a posterior probability (or a score) to each borrower to belong to group of sound or unsound borrowers;
- ii) starting from posterior probabilities definition of a "splitting rule" to assign each borrower to one of the several discrete classes in the rating system;
- iii) evaluation of the Probability of Default (PD) for each class, which is one of the input variable to work out capital requirements.

The first point deals with the choice and the evaluation of the model and, in many cases, with selection of relevant input variables (specification of the model) and the choice of the cut-off point, the value of posterior probability used to classify observations into classes of sound and unsound debtors. In the framework of credit risk analysis, the logistic approach represents a widely used tool to screen among firms on the basis of their predicted bankruptcy probability. However, in many cases, especially when the bankruptcy risk is extremely low or extremely high (i.e. when we register few or too many events), the logistic approach could fail and must be extended to take into account of potential heterogeneity sources among firms. This can be done by defining a mixed effect logistic model, where the random structure is assumed to account for potential departures from the i.i.d. hypothesis. Following this idea, we present a semiparametric approach which provides a useful tool to deal with heterogeneous binary responses extending the standard logistic approach to logistic approach with random effects, whose distribution is left unspecified and estimated by nonparametric maximum likelihood (in the following NPML).

The paper is organized as follow: in section 2 we will briefly discuss main results of the literature; in section 3 we will discuss the proposed approach and define a suitable EM algorithm for parameter estimation; in section 4 the data-base and obtained results are discussed. Last section is devoted to conclusions and future research agenda.

2 Economic background

Even if there is a sceptical view about classification (scoring) models due to some well-known shortcomings (Allen, 2002), the use of these models has significantly increased

since Altman (1968) proposed to use Linear Discriminant Analysis (LDA) to predict up to three years in advance bankruptcy in a sample of 33 sound and 33 unsound manufacturing firms. Several proposals have tried to improve Altman's results using different parametric, semiparametric and nonparametric models, stressing on both explicative and predictive capabilities. The latter correspond to the capacity to correctly predict future status of the debtors, evaluating results on an *out of sample* set, i.e. a sub sample of debtors not included in the estimation sample. For this purpose, nonparametric models present excellent results due to their robustness in picking up nonlinear relationship in the data. Conversely, they present a higher possibility to overfit and a lower explicative capability with respect to parametric approaches.

A common parametric approach used in literature for classification purposes is the logit model. As reported in Barniv and McDonald (1999) 178 articles in accounting and finance journals between 1989 and 1996 used this model. Platt and Platt (1990) applied logit analysis to predict bankruptcy with interesting result in terms of classification performance. A larger data set has been used by Laitinen (1999) who predicted insolvency for 3200 Finland firms using 15 variables selected with partial support of automatic selection procedures on a set composed by 35 variables. In the out of sample set (400 firms) sensitivity (correctly classifying a firm as unsound) was 93.75% with specificity (correctly classifying a sound) equal to 96.35%. Moody's (2001) applied the logistic approach to estimate the default probabilities of 4655 European public firms, covering 26 countries, comparing European and USA firms. With a low number of defaulted firms (81), results in *out of sample* were quit good for USA firms but were not reported for European firms. Logit approach has been compared to hazard model to forecast insolvency of property-liability insurers (Lee and Urrutia, 1996) and to analyze the probability to default for banks. Kolari et al. (2002) applied logit analysis to predict large US commercial banks failure between 1989 and 2002. The model showed good prediction performance in *out of sample* set one year prior failure but was not significantly better than chance two years prior to failure. Parametric approach, such as logit and LDA, are very common not only in the academic literature but also in studies conducted by Central Banks in Austria, France, Germany, Italy, United Kingdom (see Ooghe, et al., 1999). Cannari and Signorini (1995) used the logit model with 14 independent variables selected by stepwise forward procedure, working on the population of 800 cooperative credit banks from 1984 to 1992. The model reached a free-of-error classification rate of 70% for sound and 80% for unsound banks. Laviola and Trapanese (1997) used a sample of 3270 firms from 1991 to 1995 in commerce, industrial and building sectors: 202 were sound and 1271 were unsound firms. Starting from a covariates set of 35 variables, they selected with standard stepwise forward procedure 7 ratios for industrial sector, 3 for building sector and 5 for commerce. In the *in sample* set logit reached a correct classification rate of 95% for sound firms and 85% for unsound firms, with an average classification rate of 91%. In the *out of sample* set, logit got a classification power of 60%. More recently Fabi, Laviola and Marullo Reedtz (2002) built a rating system based on logit model to evaluate the impact of Basel Proposal on a sample of 180.000 Italian firms at 1998.

However when classification results, and in particular prediction results, of logit approach are compared to other models, especially nonparametric models, logit usually underperforms. These findings are corroborated, among others, by Caiazza (2004) who

compared LDA, logit, classification trees, neural networks and fuzzy algorithms, by Borra and Caiazza (2002) who compared logit and classification trees using generalized additive models (bagging and boosting), and by Galindo and Tamayo (2000) who compared classification trees, neural network and K-nearest neighbour. One possible explanation of these poor results is that whenever the bankruptcy risk is extremely low or extremely high (i.e. when we register few or too many events), the logistic approach could fail and must be extended to take into account of potential heterogeneity sources among firms. This can be done by defining a mixed effect logistic model, where the random structure is assumed to account for potential departures from the i.i.d. hypothesis.

Following this idea, we present a semiparametric approach which provides a useful tool to deal with heterogeneous binary responses extending the standard logistic approach to include individual (firm-specific) random effects, whose distribution is left unspecified and estimated by nonparametric maximum likelihood (in the following NPML). The solution proposed in this paper is characterized by the flexibility and high predictability of a nonparametric model, which however maintains an explicative capability similar to that of standard parametric logit. We found an impressive indication of the goodness of this approach, both in terms of explicative capability (all variables enter in the model with the right signs) and in terms of predictive capability, as shown by out of sample classification errors.

3 The Modelling approach

Assume that we are analyzing a sample of n firms and that we know, for each firm $i=1, \dots, n$, the values assumed by a binary response variable Y_i with a vector of p covariates $\mathbf{x}_i=(x_{i1}, x_{i2}, \dots, x_{ip})^\top$, where superscript \top denotes vector transpose.

Binary data are often analyzed through the logistic model; in this framework, the response variable is assumed to be Bernoulli distributed:

$$f(y_i | \theta_i) = p_i^{y_i} (1 - p_i)^{1-y_i} = \exp[y_i \theta_i - \log(1 + \exp(\theta_i))]$$

and the canonical parameter $\theta_i = \text{logit}(p_i)$ has an identity link to the linear predictor:

$$\theta_i = \alpha + \beta^\top \mathbf{x}_i \quad (1)$$

where $p_i = \Pr\{Y_i=1\}$, α represents the overall intercept and β is a vector of p regression parameters.

In general, failure of the model to fit the data could be due to various reasons: individual responses may be correlated or the model could be misspecified. A simple way of expressing these problems, which unifies these possibilities, is through omitted variables: we assume that some (say q) fundamental covariates were not considered in the model specification. If these additional variables were surveyed, they could be added in a new formulation of the model leading to updated values for parameter estimates. In this case the linear predictor (1) can be rewritten as:

$$\theta_i = \alpha + \beta^\top \mathbf{x}_i + \lambda^\top \mathbf{v}_i \quad (2)$$

where $\mathbf{v}_i=(v_{i1},v_{i2},\dots,v_{iq})^\top$ is the vector of q additional covariates and $\lambda=(\lambda_1,\lambda_2,\dots,\lambda_q)^\top$ the corresponding parameter vector. Unfortunately, the reality of many social and economic phenomena is different: variables that should be added to improve model fit are often unknown. Hence, we have no available information on the \mathbf{v}_i 's.

To solve this problem, we assume that we can model the influence of omitted covariates through the addition of a set of unobserved variables $u_i, i=1,\dots,n$, each representing the conjoint effect of the components of \mathbf{v}_i . Following this approach, the linear predictor in (1) becomes:

$$\theta_i = \alpha + \beta^\top \mathbf{x}_i + u_i \quad (3)$$

which will be referred to as mixed effect logistic model.

The values $u_i, i=1,\dots,n$, represent firm-specific features varying over the data set in an unknown way; they are usually considered, according to Kiefer and Wolfowitz (1956), as drawn from n i.i.d. random variables U_i with a common, unknown, density function $g(\cdot)$. Treating the U_i s as nuisance parameters and integrating them out, the likelihood function can be expressed as:

$$L(\cdot)= \prod_{i=1}^n \left\{ \int_{\mathcal{V}} f(y_i | \theta_i) dG(U_i) \right\} \quad (4)$$

For the computation of the maximum likelihood estimator, the results of Lindsay (1983a, 1983b) are of great interest. He showed that, for fixed β , the likelihood is maximized with respect to $G(\cdot)$ by at least one discrete distribution $\hat{G}_n(\cdot)$ with at most n support points. As a prerequisite for a well behaved estimation of $G(\cdot)$ we should assume that the mixture is identifiable, i.e. that two sets of parameters which do not agree after permutation cannot yield the same mixture distribution. Teicher (1963) pointed out that no mixing distribution can be identified in mixtures of Bernoulli distributions. However, in the regression setting something changes: Follmann and Lambert (1991) gave simple bounds for the number of support points ensuring identifiability of $G(\cdot)$ in random effects models for binary responses if covariates values can be adequately grouped.

Since NPML estimate of a mixing distribution is a discrete distribution on a finite number of locations, say K , the likelihood function can be expressed as:

$$L(\delta)= \prod_{i=1}^n \sum_{k=1}^K f(y_i | \theta_{ik}) \pi_k \quad (5)$$

where $\theta_{ik}=\alpha + \beta^\top \mathbf{x}_i + u_k$ and δ is the corresponding ‘‘global’’ vector of model parameters. The parameter u_k represents the deviation of the k -th component intercept from the overall intercept α . We can, therefore, follow the path described by Aitkin (1999) in the framework of clustered data. Writing:

$$f_{ik} = f(y_i | \theta_{ik}) \quad (6)$$

then, we have:

$$\frac{\partial \log[L(\delta)]}{\partial \delta} = \sum_{i=1}^n \sum_{k=1}^K w_{ik} \frac{\partial \log(f_{ik})}{\partial \delta} \quad (7)$$

where:

$$w_{ik} = \frac{f_{ik} \pi_k}{\sum_{k=1}^K f_{ik} \pi_k} \quad (8)$$

represents the posterior probability that i -th firm comes from the k -th component of the mixture. Differentiating the log-likelihood under the constraint $\sum_{k=1}^K \pi_k = 1$, we obtain:

$$\hat{\pi}_k = \frac{\sum_{i=1}^n w_{ik}}{n} \quad (8)$$

a standard result for ML in finite mixture models.

Equating to zero gives likelihood equations which are simple weighted sums of those for an ordinary logistic model with weights w_{ik} . Alternatively solving these equations for a given set of weights and updating weights from the current parameter estimates is an EM algorithm (Dempster, Laird, Rubin, 1977).

4 Data and Results

Our purpose is to obtain a reliable forecast for the probability of being unsound at future times given actual firms status and conditionally on some covariates recorded at time t , as well as to define a modelling approach characterized by good explicative capabilities. The analyzed data are collected by Bureau van Dijk S.A. on 230,000 Italian small and large firms. Selecting for juridical status and default information we derived a sub-sample of 81,912 manufacturing firms for years 1999-2002. Tables 1 and 2 shows firms included in the analyzed sample by localization and size.

Table 1 about here

Table 2 about here

For our scope, the event of default (unsound firms) is recorded when firms are bankrupted or are going to be liquidated; in table 3 we report corresponding frequencies. The variable status assumes value 0 for all sound firms and value 1 for the unsound ones.

Table 3 about here

Previous table shows the distribution of the default event through years: as can be easily noted, there is a peak in 1999, which could be due to the beginning of the worldwide economic slowdown. The total amount of defaulted firms is 1.32% of the analyzed

sample. The geographical distribution of sound and unsound firms is reported in table 4, while table 5 reports sound and unsound firms distribution with respect to firms size.

Table 4 about here

Table 5 about here

To check the explicative capability of logit and mixed effect logit, we focused on statistical significance for covariates and values of maximized log-likelihood. In evaluating the predictive capability we have split observations in two sets, respectively the *in sample* and the *out of sample* set. We used the former to estimate parameters for each fitted model, and the latter to test the corresponding prediction behaviours.

The *in sample* set is made up by all firms which got unsoundness in 1999 or 2000, while the *out of sample* set is made up by all firms who got failed between 2001 and 2002. With respect to sound firms, we have decided to randomly draw firms for the *in sample* and *out of sample* sets, in order to leave unchanged the percentage of sound firms across the two sets.

Thus, the estimated probability of default in the *in sample* set has to be interpreted as the average probability of default during 1999 and 2000. At the same time, the estimated probability for the out of sample set has to be interpreted as average default probability of default during 2001 and 2002. The underlying hypothesis is that the effect of covariates on default probability remains constant over the 2 analyzed periods (1999-2000 and 2001-2002). Considering the *in sample* set previously defined, we fitted three models: first a simple logit model with a minimal covariates set just to give a formal standard of comparison. After, we had considered a wider set of covariates adopting a logit model with stepwise forward procedure based on LR statistic. This could be considered as the real benchmark for the proposed mixed effect approach, which is the last model we have employed.

The simple logit model considered only basic covariates, such as *area* (geographical localization in 4 classes), *size* (n. of employees in 4 classes), *roe₁₉₉₉* and *roi₁₉₉₉*.

We considered these last ratios to be consistent with previous literature, where they usually enter as covariates. In particular *roe* represents a useful indicator to explain unsoundness whenever this state is represented by bankruptcy. The correlation between this indicators, at 1999, for the whole sample is lower than 0.65. The corresponding linear predictor (with canonical link adopted) is the following:

$$\text{logit}(p_{ij}) = \beta_0 + \sum_{l=2}^4 \beta_{1l} \text{area}_l + \sum_{m=2}^4 \beta_{2m} \text{size}_m + \beta_3 \text{roe}_{1999} + \beta_4 \text{roi}_{1999}$$

where $p_{ij} = \Pr(Y_{ij}=1|\mathbf{x}_{ij}) = \Pr(\text{status}=1|\mathbf{x}_{ij})$, i represents individual firms within sectors, while j indexes economic sectors.

Table 6 reports estimates for the logit model while table 7 reports the corresponding classification matrix for *out of sample* observations. The intercept term, *roe* and *roi* are all statistically significant at the 5% level. Adopting this model specification, it is worth noting that both size and area variables, with the only exception of center Italy modality (*area*=3), seem to play no role in explaining the probability of default (PD), i.e.

geographical localization and firm size have no effect on PD. The estimated effects corresponding to *roe* and *roi* are negative: therefore, the lower are return on equity and return on investment, the greater is the probability of default. This result is coherent with standard working hypothesis, discussed in the relevant literature. As can be noted from Table 7, the predictive capability evaluated in the *out of sample* set is quite unsatisfactory. Since sample is overwhelming unbalanced between sound and unsound firms (default event is in fact a rare event), logit model is not able to discriminate between the two classes; sensitivity and specificity are equal, respectively, to 48.22% and 71.55%.

Table 6 about here

Table 7 about here

To improve the simple logit model we decided to adopt a more complex model considering as potential predictors all covariates available at 1999. Due to the wide dimension of the resulting covariates set we proceeded employing a stepwise forward selection procedure based on LR statistic to select only a relevant covariates subset. The results are reported in table 8, with respect to parameter estimates, and in table 9 with respect to out of sample prediction.

Table 8 about here

Table 9 about here

As we could have expected, more balance sheet indicators enter in model specification, a lot of them being highly significant. In particular, stepwise selected four variables relative to reserves, even if just two of them are significant at 5% level. The negative sign of these variables correctly identify that the lower are reserves, the higher is PD. “Financial costs” variable is highly significant and indicates that higher financial costs imply higher PD as indicated by economic theory on asymmetric information. Variable “primary products” is highly significant and seems to indicate that bankrupted firms have a lower amount of primary products; *roi* appears highly significant but with the wrong sign with respect to the expected one.

Other ratios enter with negative sign, coherently with our working hypotheses. It is interesting to point out that both *roe* and *roi* were selected by stepwise procedure, strengthening our *a priori* choice. Dummies for size and area enter in model specification only for category 3, i.e. for medium sized and center-located firms. Their estimated effects are, however, positive and controversial: a medium sized firm seems to be more prone to default than very small and small firms. At the same time, firms located in center Italy are more likely to experience a default than firms in other regions and this result is somewhat striking with empirical evidence. The goodness of fit is higher than in previous specification (with just two covariates) as shown by the higher value of maximized log likelihood. Even looking at the predictive capability (table 9) in the *out of sample* set, we can see changes in the classification matrix, even if the results are not completely satisfactory; sensitivity and specificity are, respectively, equal to 49.24% and 73.54%. This means that, should we have decided to use previous models, we would have had huge losses in the next two years after estimation. Therefore, we

decided to estimate the default probability using a mixed effects logit model, where the linear predictor has been defined as follows:

$$\text{logit}(p_{ij}) = \beta_0 + \sum_{l=2}^4 \beta_{1l} \text{area}_l + \sum_{m=2}^4 \beta_{2m} \text{size}_m + \beta_3 \text{roe}_{1999} + \beta_4 \text{roi}_{1999} + u_j$$

where $p_{ij} = \Pr(Y_{ij}=1 | \mathbf{x}_{ij}, u_j) = \Pr(\text{status}=1 | \mathbf{x}_i, u_j)$. In this context, the random effects u_j represent economic sector-specific random variation from the overall intercept, whose distribution is left unspecified as discussed in section 3. This means that we associated a varying effect to intercept terms in each sector, to model heterogeneity present at sector level. Due to change in model fit and based on penalized likelihood criteria, we estimated a two-component mixed effects logit model. The first component has a prior probability equal to 0.9472 reflecting sound firms: in this component intercept value is negative reflecting a lower propensity to default event in this group. The second component, which has a prior probability of 0.0528, is mainly associated to unsound firms and reveals a positive component-specific intercept term which leads to a near-zero constant term. Estimated fixed effects for *roe* and *roi* are both statistically significant at the 5% level, as well as the effect corresponding to medium sized firms, thus providing empirical evidence coherent results obtained by the logit model. Table 10 and table 11 report estimation results for mixed effect logit model and the corresponding classification matrix for *out of sample* observations.

Table 10 about here

Table 11 about here

The maximized log-likelihood value is greater than those obtained by the simple logit model and the stepwise logit model, and is associated to a better out of sample prediction, as can be noted from table 11, which reveals that the mixed effects logit outperforms. The prediction capability is quite satisfactory: in this context, sensitivity and specificity are, respectively, equal to 59.39% and 83.09%. A possible explanation of this result should be found in the higher robustness of the mixed approach to model misspecification and in its capability to catch up potential heterogeneity sources among observations as well as to account for potential departures from standard parametric hypotheses.

5 Conclusions

In this paper, we have discussed an extension of standard logit method to classify firms with respect to default probability. As shown, the standard logit model is usually biased when the default event is particularly rare; therefore we proposed and discussed the application of a semiparametric logit model assuming heterogeneity is present at the economic sector level. The adopted random effects are considered as random variables with a common but unknown distribution: our proposal is to proceed to nonparametric estimation of this distribution using finite mixtures as detailed in Aitkin (1999). The performance of the mixed effect logit has been compared with both simple logit and

more complex stepwise forward logit model, stressing that the proposed approach has better performances on the out of sample set in terms of classification errors.

The adopted approach is not however, the right answer to all questions: first, the analyzed dataset is a particular one: more research is needed to verify the behavior of proposed approach in a wide variety of empirical situations, maybe through a Monte Carlo study. Secondly, identifiability and convergence properties of the mixed effect model must be carefully checked to obtain adequate results. Third, large sample sizes are necessary to consistently estimate the mixing distribution and, therefore, to produce satisfactory out of sample prediction. Nevertheless, the performance is so higher that it could be worth of more complex analysis.

References

- Aitkin, M.A., (1999) A general maximum likelihood analysis of variance components in generalized linear models, *Biometrics*, **55**, 117-128.
- Allen, L., (2002) Credit risk modeling of middle market, presented at Conference on Credit Risk Modeling and Decisioning, Wharton FIC, University of Pennsylvania, <http://fic.wharton.upenn.edu/fic/creditrisk.html>.
- Altman, E. I., (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *The Journal of Finance*, **23**, 589-609.
- Barniv, R., McDonald J. B., (1999) Review of Categorical Models for Classification issues in accounting and finance, *Review of Quantitative Finance and Accounting*, **13**, 39-62.
- Basel Committee on Banking Supervision, (2001) The new Basel capital accord, January.
- Basel Committee on Banking Supervision, (2003) The new Basel capital accord, April.
- Borra, S., Caiazza, S., (2002) Comparative performance of credit scoring models using aggregated predictors, *Data Mining III*, WIT Press, 747-756.
- Caiazza, S., (2004), *The comparative performance of credit scoring models: an empirical approach*, in Monetary Integration, Markets and Regulation, Research in Banking and Finance, 4, Elsevier Science Ltd, forthcoming.
- Cannari, L., Signorini, F., (1995) L'analisi discriminante per la previsione delle insolvenze delle micro-banche, *Temi di Discussione, Banca d'Italia*, 258.
- Dempster, A.P., Laird, N.M., Rubin, D.B., (1977) Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, **B**, 39, 1-38.
- Fabi, F., Laviola, S., Marullo, R. P., (2002) The impact of Basel II on the Italian corporate sector and on Italian bank, presented at the Bocconi Conference on Risk and Stability in the Financial System: What Role for Regulators, Management and Market Discipline, Milan, 13-14 June.
- Follmann, D.A., Lambert, D., (1991) Identifiability of finite mixtures of logistic regression models. *Journal of Statistical Planning and Inference*, **27**, 375-381.
- Galindo, J., Tamayo, P., (2000) Credit risk assesment using statistical and machine learning: basic methodology and risk modelling applications, *Computational Economics*, **15**, 107-143.

- Kiefer, J., Wolfowitz, J., (1956) Consistency of the maximum likelihood estimation of a mixing distribution, *Journal of the American Statistical Association*, **73**, 805-811.
- Kolari, J., Glennon D., Shin H., Caputo M., (2002) Predicting large US commercial bank failure, *Journal of Economics and Business*, **54**, 631-387.
- Laitinen, E. K., (1999) Predicting a corporate credit analyst's risk estimate by logistic and linear models, *International Review of Financial Analysis*, **8**(2), 97-121.
- Laviola, S., Trapanese, M., (1997) Previsione delle insolvenze delle imprese qualità del credito bancario: un'analisi statistica, *Temi di Discussione, Banca d'Italia*, 318.
- Lee, S. H., Urrutia J. L., (1996) Analysis and prediction of insolvency in the property-liability insurance industry: a comparison of logit and hazard models, *The Journal of Risk and Insurance*, 63 (1), 121-130.
- Lindsay, B.G., (1983a) The geometry of mixture likelihoods: a general theory. *Annals of Statistics*, **11**, 86-94.
- Lindsay, B.G., (1983b) The geometry of mixture likelihoods, part II: the exponential family. *Annals of Statistics*, **11**, 783-792.
- Moody's, (2000) RiskCalc™ for private companies: Moody's default model, www.moody.com
- Moody's, (2001) RiskCalc™ Public - Europe, www.moody.com
- Ooghe H., Claus H., Sierens N., Camerlynck J., (1999) International comparison of failure prediction models from different countries: an empirical analysis, *Working Paper*, 99/79, Univesity of Ghen.
- Platt, H.D., Platt, M.D., (1990) Development of a class of stable predictive variables: the case of bankruptcy prediction, *Journal of Business Finance and Accounting*, **17**, 31-51.
- Teicher, H., (1963) Identifiability of finite mixtures, *Annals of Mathematical Statistics*, **34**, 244-248.

Tables

<i>Size</i>	Freq.	Percent	Cum.
[0, 15)	21'660	26.44	26.44
[15, 50)	27'798	33.94	60.38
[50, 250)	9'745	11.9	72.28
[250, +)	22'709	27.72	100
Total	81'912	100	

Table 1: Distribution of analyzed firms by firm size.

<i>Area</i>	Freq.	Percent	Cum.
NorthEast	12'999	15.87	15.87
NorthWest	40'679	49.67	65.54
Centre	17'215	21.02	86.56
South	11'012	13.44	100
Total	81'905	100	

Table 2: Distribution of analyzed firms by firm localization.

<i>Year</i>	Freq.	Percent
1999	508	47.65%
2000	370	34.71%
2001	131	12.29%
2002	57	5.35%
Total	1'066	

Table 3: Unsound firms by year of default.

<i>Area</i>	<i>Status</i>		Total
	0	1	
NorthEast	12'846	153	12'999
NorthWest	40'221	458	40'679
Centre	16'939	276	17'215
South	10'833	179	11'012
Total	80'839	1'066	81'905

Table 4: Status of analyzed firms by firm size.

<i>Size</i>	<i>Status</i>		Total
	0	1	
[0, 15)	21'361	299	21'660
[15, 50)	27'415	383	27'798
[50, 250)	9'632	113	9'745
[250, +)	22'438	271	22'709
Total	80'846	1'066	81'912

Table 5: Status of analyzed firms by firm localization.

		<i>Coef.</i>	<i>Std. Err.</i>	<i>z</i>	<i>P> z </i>
<i>Area</i>	<i>NorthWest</i>	-0.013	0.188	-0.07	0.95
	<i>Center</i>	0.428	0.176	2.42	0.02
	<i>South</i>	0.190	0.197	0.96	0.34
<i>Size</i>	[15, 50)	-0.132	0.104	-1.28	0.20
	[50, 250)	-0.080	0.175	-0.46	0.65
	[250, +)	-0.251	0.218	-1.15	0.25
<i>roi</i>		-0.050	0.015	-3.33	0.00
<i>roe</i>		-0.008	0.002	-4.38	0.00
<i>cons</i>		-4.468	0.184	-24.26	0.00

log likelihood -2456.128

Table 6: Status equation; parameter estimates for logit model. In sample observations.

<i>Status (observed)</i>	<i>Status (predicted)</i>		Total
	0	1	
0	10495	4174	14669
1	102	95	197
Total	10597	4269	14866

Table 7: Status equation; logit model, out of sample classification matrix.

		<i>Coef.</i>	<i>Std. Err.</i>	<i>z</i>	<i>P> z </i>
<i>Legal reserve</i>		-0.009	0.003	-3.72	0
<i>roi</i>		-0.077	0.023	-3.33	0.001
<i>roe</i>		-0.006	0.002	-3.73	0
<i>Statutory reserve</i>		-0.003	0.001	-2.54	0.011
<i>Revaluation reserve</i>		-0.004	0.003	-1.47	0.142
<i>Area</i>	Centre	0.371	0.098	3.8	0
<i>Size</i>	[50, 250)	0.584	0.144	4.06	0
<i>ros</i>		0.026	0.013	2.01	0.044
<i>Other reserves</i>		0.000	0.000	-0.68	0.496
<i>Financial costs</i>		0.000	0.000	4.27	0
<i>Depreciation / devaluation</i>		0.000	0.000	-0.97	0.333
<i>Primary products</i>		0.000	0.000	-1.89	0.059
<i>cons</i>		-4.230	0.096	-44.24	0
log likelihood		-2277.724			

Table 8: Status equation; parameter estimates for stepwise forward LR logit model. In sample observations.

<i>Status (observed)</i>	<i>Status (predicted)</i>		Total
	0	1	
0	10788	3881	14669
1	100	97	197
Total	10888	3978	14866

Table 9: Status equation; stepwise forward LR logit model, out of sample classification matrix.

		<i>Coef.</i>	<i>Std. Err.</i>	<i>z</i>	<i>P> z </i>
<i>Area</i>	<i>NorthWest</i>	0.014	0.148	0.10	0.92
	<i>Center</i>	0.355	0.157	2.26	0.02
	<i>South</i>	0.155	0.182	0.85	0.40
<i>Size</i>	[15, 50)	-0.095	0.109	-0.87	0.39
	[50, 250)	-0.027	0.147	-0.19	0.85
	[250, +)	-0.280	0.197	-1.42	0.15
<i>Roi</i>		-0.049	0.008	-6.03	0.00
<i>Roe</i>		-0.008	0.002	-4.92	0.00
<i>Cons</i>	Comp. 1	-4.799	0.162	-29.63	0.00
	Comp. 2	-0.585	0.032	-18.27	0.00
σ^2		0.889			
log likelihood		-2247.051			

Table 10: Status equation; parameter estimates for mixed effect logit model. In sample observations.

<i>Status (observed)</i>	<i>Status (predicted)</i>		Total
	0	1	
0	12188	2481	14669
1	80	117	197
Total	12268	2598	14866

Table 11: Status equation; mixed effects logit model, out of sample classification matrix.