

Survival models and credit scoring: some evidence from Italian Banking System.

Francesca Giambona* and Vincenzo Lo Iacono**

The purpose of this contribution is to introduce a survival analysis to study the risk of default and to propose the empirical evidence by the Italian Banking System. Banks currently use credit scoring to determine the probability of default of the borrowers. The timing when borrowers become insolvent is very interesting to predict the probability of default over the lifetime of a loan and it is a relevant information to perform profit scoring. Results obtained through a discrete time hazard model adapted to the life of a loan provided clear evidence that time when the default occurs is an important element to predict the probability of default in time. The hazard model, estimated for a population of loans, involves different probability of default if the explanatory variables and the time when the default occurs are jointly considered.

Field of Research: credit scoring, hazard models to estimate probability of default.

1. Introduction

During the second half of the Nineties, banks have been developing credit risk models aimed to measure the potential loss, with a predetermined confidence level, that a portfolio of credit exposures might suffer within a specified time horizon (generally one year) (BIS, 2004). It is very important for banks to predict the probability of default for a homogeneous group of loans: losses on any single loan will not cause a bank to become insolvent. Credit scoring is a suitable objective model to evaluate the risk of default. This is a multivariate statistical model that examines the different borrower's characteristics attributing a different weight to explanatory variables on risk of default reaching a probability of default (PD) for each borrower.

Then, the purpose of credit scoring is to quantify the predicted values of probability of default, in order to identify and to quantify the borrower's characteristics, that determine which variables involve a higher default rate. In this contribution a model is introduced to quantify PD proposing a credit scoring model that introduces also the time when the default occurs. The purpose of this paper is to introduce and to show an approach in evaluating and monitoring PD. A survival credit scoring model is introduced by a discrete time hazard model that introduce the time when the default occurs. An application to dataset made by Italian Central Bank shows the usefulness of this approach.

*Francesca Giambona, Department of National Accounting and Social Processes Analysis, University of Palermo, email: francesca.giambona@unipa.it

**Vincenzo Lo Iacono, Department of Quantitative Methods for Human Sciences, University of Palermo, email: lojacono@unipa.it

2. Literature Review

During the past 10-15 years, marked progress has been achieved in measuring credit risk. Most approaches involve the estimation of three parameters: the probability of default on individual loans or pools of transactions (PD); the estimation of the losses-given-default (LGD) and the correlation between defaults (Crouhy *et al.*, 2000; Duffie and Singleton, 2003). In the new capital accord agreed on June 2004 (BIS, 2004), financial institutions are invited, in the internal rating-based (IRB) approach, to estimate the one-year probability of default and the expected LGD. There are three broad categories of traditional models used to estimate PD: i) expert systems, including artificial neural networks; ii) rating systems; and iii) credit scoring models.

More details of expert and rating systems are discussed in Allen *et al.* (2004). The most common model to measure PD in credit risk measurement methodology is credit scoring analysis. Credit scoring models are commonly structured along the lines of Altman's (1968) Z-score model using historical loan and borrower data to identify which borrower characteristics are able to distinguish between defaulted and non defaulted loans. During this development stage, decisions about the estimation of the model had to be made and all potentially relevant borrower characteristics to be selected and their impact of default established. Now, a credit scoring model can be applied to new loan applications for which the PD is not known. Based on the estimated of credit scoring, a credit score can be calculated for each new loan where a higher score indicates better expected performance of the borrower and thus a lower PD.

Other general introductions to credit scoring are by Mays (1998), Hand and Henley (1997), Mester (1997), Viganò (1993), and Lewis (1990). A *credit-scoring* model is a formula that puts weight on different characteristics of a borrower, lender, and loan. There are four methodological forms of multivariate credit scoring models: (1) the linear probability model; (2) the logit model; (3) the probit model; (4) the multiple discriminant analysis model; and (5) decision trees. Based on findings by Boyle, Crook, Hamilton, and Thomas (1992), Desai, Crook, and Overstreet (1996, 1997), Henley (1995), Srinivasan and Kim (1987), and Yobas, Crook, and Ross (2000) as reported by Thomas (2000), in this paper the logistic regression method is adopted. Credit scoring models are relatively easy to be implemented and do not suffer from the subjectivity and inconsistency of expert systems.

But credit scoring does not consider the time when the default occurs. The mortality approach in risk credit management, introduced by Altman (1989), was used in different applications (Altman and Suggitt, 2000). Studies on corporate bonds reported information on the probability of default over time for different bond ratings, on recovery rates based on market prices at the time of default, on estimates of rating transition matrices, and on the degree of correlation between default frequencies and recovery rates (Allen *et al.*, 2004; Acharya *et al.*, 2003). In Italy the mortality approach was suggested by the Italian Central Bank (Banca d'Italia, 2001; ABI, 2001) by introducing the mortality rate to evaluate and monitoring Italian Banking System.

3. Methodology and Research Design

Researchers use the survival analysis in a variety of contexts that share a common characteristic: interest centers on describing whether or when events occur. It is necessary to use the survival analysis if we are interested in whether and when an event occurs (Allison, 1984). In this context the event occurrence represents a borrower's transition from one state, loan "in bonis" that is not in default, to another state, the default. To introduce survival approach to loans we assume that:

Assumption 1: a generation (or cohort) of loans is formed by loans granted by Italian banks in a same year;

Assumption 2: the death of the loan occurs with the default (the definition of default given by the Italian Central Bank is adopted in this contribution);

Assumption 3: the death of a loan in an uncertain event both "when" and "if";

Assumption 4: loan survival is the difference between two time: the time when a loan has been granted and the time when a loan becomes default;

Assumption 5: a loan is censored when, in the period of study (named follow-up), it is not in default or it goes out of the study to verify an event different of default. Thus, a loan is censored when: 1) a loan is *in bonis* so it is survived in all time, 2) the loan has been repaid.

An appropriate statistical tool used in survival analysis is a discrete-time hazard model introduced by Cox (1972), and later discussed and reviewed (Singer and Willet, 2003; Fahrmeir and Tutz, 1994; Allison, 1984). The fundamental quantity in survival analysis is the hazard rate: at a given time point t , the hazard is the probability of experiencing the event of interest at time t conditional on being still at risk and on the value of the covariates.

$$h_i(t | \mathbf{x}_{it}) = P(T_i = t | T_i \geq t, \mathbf{x}_{it}) \quad (1)$$

where the vector \mathbf{x}_{it} includes all the covariates of subject i at time t . The covariates can be time-invariant or time-varying. Time-varying covariates are extremely useful in building a proper model for the hazard, but they are rarely available in practice because of the difficulty of an accurate measurement, especially in retrospective surveys. Since the hazard function is bounded between 0-1, a linear model for the hazard itself is not suitable, but one can apply a linear model to an appropriate transformation of the hazard:

$$g(h_i(t | \mathbf{x}_{it})) = \alpha_t + \beta' \mathbf{x}_{it} \quad (2)$$

where the transform $g(\cdot)$, called *link function*, maps the (0,1) interval onto the real line. On the right-hand side, β is the vector of regression coefficients and $(\alpha_1, \alpha_2, \dots, \alpha_p)$ are time-specific intercepts representing the baseline hazard, i.e. the hazard for the hypothetical subject with all the covariates set to zero. The number of

time-specific intercepts is P , the maximum number of time points (intervals) in the data.

Therefore, using the time indicators (α_i) as well explanatory variables, each intercept parameters represents the value of logit hazard (the log odds of event occurrence) in that particular time period for individuals in the baseline group; each slope parameter assesses the effect of a one unit difference in that predictor on event occurrence, statistically controlling for the effects of all other predictors in the model. When the link $g(\cdot)$ is the *logit* function the corresponding model is called *proportional hazard odds model*:

$$\log\left(\frac{h_i(t | it)}{1 - h_i(t | it)}\right) = \alpha_t + \beta' X_{it} \quad (3)$$

or, in terms of the hazard function:

$$h_i(t | it) = \frac{1}{1 + \exp(-\alpha_t - \beta' X_{it})} \quad (4)$$

The interpretation of the regression coefficients requires some care, since β is the change in the *logit* of the hazard following a unit increase in the k -th covariate. As in logistic regression it is rare to interpret the estimated parameters. Commonly, the odds ratio (OR) is defined as is the odds of an event occurring in one group to the odds of it occurring in another group. An odds ratio of 1 indicates that the condition or event under study is equal in both groups. The OR is defined as:

$$OR = \exp(\beta_k) \quad (5)$$

Estimation can be carried out using standard software for binary response models. In fact, the likelihood of a discrete-time survival model on the original dataset is the same as the likelihood of a binary response model on the *person-period dataset*. To obtain the person-period dataset, each original record is replicated as many times as the observed time and the new response variable is the indicator of the event of interest.

4. Discussion of Findings

In this section the determinants of hazard rates for Italian loans will be empirically estimated. A discussion of the choice of explanatory variables and the econometric specification is followed by the empirical results. The databank used is by the Italian Central Bank considering 11 cohorts of loans (from 1985 to 1995) for a 10 year period. Italian Central Bank records default at the end of each year. The number of loans analysed grouped 1.302.186 borrowers, that between 1985 and 1995 obtained a loan by an Italian bank. For these borrowers, some characteristics as possible explanatory variables or risk factors for the default have been selected. The dependent variable is the default. The loans were tracked for 10 years in order to study the survival function in the first ten years from their origination.

Explanatory variables include the amount of the loan, the institutional sectors of the borrowers, the geographical area of borrowers and the year when the loan was granted, named below as “cohort loan”. Time indicator variables are included to define the periods in which the hazard function is higher or lower. A loan is censored when, in the period of study, it is not in default or it goes out of the study to verify an event different than default. Mainly a loan is censored when: 1) it is *in bonis*, so it survives; 2) it has been repaid.

Preliminarily, the matrix of data was transformed into a person-period dataset. A discrete-time hazard model was fitted introducing ten intercepts parameters for time indicators (value of logit hazard in that particular time period for borrowers in baseline group) and a slope parameter (effect of a one unit difference in that predictor on event occurrence). A logit regression was used to regress the event indicator (the default) on the time indicators and on the selected explanatory variables in the person-period dataset. Explanatory variables introduced in the model are dichotomous (productive sector, amount class of the loan and loan cohort) and polytomous (territorial area), in order to study the effects of these in the probability of default (which determine the drop out of the cohort).

The loans were shared in homogeneous groups. A score is attributable not only in relationship to the “risk factors” that cause higher value of PD, but also considering the years in which the loan could enter in default. This model is to be able to attribute, to a new loan, a diversified score for year (*survival score*) that means the predictors variables of default and the year in which the default occurs. For each loan it is possible to classify the borrowers by:

- geographical area of borrower: Northern regions (baseline), Central and Southern regions;
- productive economic sector: producer families (baseline) and enterprises;
- generation or cohort of belonging: loans granted between 1985 and 1992 (baseline) and loans granted between 1993 and 1995;
- amount of the loan: less than 250.000 euro (baseline), more than 250.000 euro.

Baseline was selected with reference to the loan that had a lower value of the probability of default; in such way it was possible to define two bound profiles, minus and plus hazardous, inside which were included all possible combinations of risk factors in the selected period (in this case ten years). We specified a proportional odds model, to understand the relation between default and explanatory variables that include also time indicators of default (see table 1).

Table 1: Discrete time hazard model, probability of default (baseline omitted¹)

Default	Coef.	Std. Err.	P>z	[95% Conf.	Interval]
<i>Time Indicator variables (α_t)</i>					
1.00	-5.051	0.010	0.000	-5.071	-5.031
2.00	-5.058	0.010	0.000	-5.078	-5.038
3.00	-5.187	0.010	0.000	-5.207	-5.166
4.00	-5.302	0.011	0.000	-5.324	-5.281
5.00	-5.460	0.011	0.000	-5.483	-5.438
6.00	-5.644	0.012	0.000	-5.668	-5.621
7.00	-5.837	0.013	0.000	-5.862	-5.812
8.00	-6.006	0.014	0.000	-6.033	-5.979
9.00	-6.238	0.015	0.000	-6.268	-6.209
10.00	-6.467	0.017	0.000	-6.499	-6.434
<i>Explanatory Variables (risk factors, β)</i>					
Central regions	0.528	0.007	0.000	0.514	0.542
Southern regions	0.808	0.007	0.000	0.795	0.822
Amount of the Loan > 125.000 euro	0.585	0.007	0.000	0.571	0.599
Enterprises	0.272	0.006	0.000	0.260	0.284
Cohort 1985-1992	0.483	0.007	0.000	0.469	0.497

¹Baseline: Northern Italian regions, amount of the loan < 125.000 euro, cohort 1993-1995, producer families.

The odds ratios associated with specified explanatory variables are shown in table 2.

Table 2: Odds Ratios (OR).

Risk Factors	Odds Ratios
Central Regions (/ Northern regions)	1,6953
Southern regions (/ Northern regions)	2,2444
Amount of the Loan > 125.000 euro (/ < 125.000 euro)	1,7950
Cohort 1985-1992 (/Cohort 1993-1995)	1,6214
Enterprises (/Producer families)	1,3130

Time indicator variables show that the probability of default for the baseline decreases over time and the explanatory variables are risk factors for default. Factors as: a loan granted in central/southern regions, or to a enterprise, or in the year between 1993-1995, or for an amount > 125.000 euro, increase the probability of default by a factor correspondent to the odds ratio. Observing the results of Table 1, it is seen that the loan cohort is an explanatory variable statistically significant: loans granted between 1985 and 1992 have an hazard of default higher than those granted between 1993 and 1995. The estimate confirms that the actions taken by Italian banks in order to decrease the default have produced some expected results. Differences can be also found with respect to the institutional sector of the borrower; the hazard is higher for firms than for a producer family (OR=1,31).

Territorial coefficients show higher values of hazard for borrowers in central or southern regions, especially for the latest. The OR are respectively 1,70 and 2,24; the probability that a borrower in Southern regions defaulted is about 2 times higher than a borrower in Northern regions while for a borrower in central regions this probability is 1.7 times higher. Results confirmed the dualistic Italian credit market: the defaults are more evident in the South than in North Italy where economic conditions also favour the credit market.

The briefly mentioned aspects, that underline the fragility of Southern regions economy, have a feedback in credit market where the defaults are still too high, as referred by the literature (Cusimano and Vassallo, 2007; Cusimano, 2006; Cannari and Panetta, 2006; Bongini and Ferri, 2005; Mattesini and Messori, 2004). This evidence suggests differentiated bank policies related to borrowers characteristics (already achieved in credit scoring models) and also by year of “life” of loan. Using the formula (5) we can try hazard rate (thus PD) in each period.

For example in the first year the hazard rate for baseline loan (thus a loan granted between 1993 and 1995, to a producer family, in North and for an amount < 125.000 euro) is about 0,052; in the same year for a loan granted between 1985 and 1992, to an enterprise, in South and for an amount of loan > 125.000 euro is about 0,006. This evidence suggests that a different score must be attribute to loans considering time and risk factors of default.

5. Conclusion

The probability of default is a crucial problem for banks. In the last years international agreements, as Basel Accord and the following Basel 2, have incentivised the banks to adopt objectives systems of evaluating and monitoring risk of default in order to predict PD for new loans based on borrower’s characteristics. The literature confirms that credit scoring is the model utilised by the banks. In this paper a revised version of credit scoring is presented and a first application to Italian bank system reported. The time when the default occurs was introduced in a credit scoring model by using a survival approach through a discrete time hazard model.

Discrete time proportional hazard model showed that PD is not constant over time and the explanatory variables are risk factors for default. Considering, the time and the risk factors jointly, a different PD can be modelled. Time is a variable not considered in credit scoring and we put evidence that a survival model (we used a time discrete hazard model, but it is possible to adapt other survival models, e.g. Cox’s model for continuous time variables) is a suitable tool to estimate PD following a longitudinal approach.

References

- ABI. 2001, “I tassi di default per l’analisi del rischio di credito”, *Bancaria*, no. 6, pp. 75-76.
- Acharya, V.V., Bharath, S.T., Srinivasan, A. 2003, Understanding the recovery rates on defaulted securities, Mimeo.

- Allen, L., De Long, G. and Saunders, A. 2004, "Issues in the credit risk modeling of retail markets", *Journal of Banking and Finance*, vol. 28, no. 4, pp. 727–752.
- Allison, P. D. 1984, Event history analysis: Regression for longitudinal data, Sage University paper series on quantitative applications in the social sciences, Beverly Hills, Sage.
- Altman, E. I. 1968, "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy", *Journal of Finance*, no.18, pp. 589-609.
- Altman, E. I. 1989, "Measuring Corporate Bond Mortality and Performance", *Journal of Finance*, no. 44, pp. 909-922.
- Altman, E. I. and Suggitt, H. J. 2000, "Default rates in the syndicated bank loan market: A mortality analysis", *Journal of Banking and Finance*, no. 24, pp. 229-253.
- Banca d'Italia. 2001, Bollettino statistico, Roma.
- BIS. 2004, International Convergence of Capital Measurement and Capital Standards, Basilea.
- Bongini, P. and Ferri, G. 2005, Il sistema bancario meridionale, Laterza, Bari.
- Cannari, L. and Panetta, F. 2006, Il sistema finanziario e il mezzogiorno, Cacucci, Bari.
- Cox, D.R. 1972, "Regression Models and Life Tables", *Journal of the Royal Statistical Society, Serie B*, no. 34, pp. 187-220.
- Crouhy, M., Galai, D., Mark, R. 2000, "A comparative analysis of current credit risk models", *Journal of Banking and Finance* n. 24, pp. 57–117.
- Cusimano, G. 2006, Sul sistema bancario italiano tra il 1999 ed il 2005, *Minerva Bancaria*, no. 6, pp. 9-37.
- Cusimano, G. and Vassallo, E. 2007, Caratterizzazioni territoriali nella distinzione dimensionale delle banche italiane, *Minerva Bancaria*, no. 5-6, pp. 10-22.
- Duffie, D. and Singleton, K. J. 1998, "Simulating correlation defaults", *Bank of England Conference on Credit Risk Modeling and Regulatory Implications*, London, 21–22 September.
- Fahrmeir, L. and Tutz, G. 1994, Multivariate Statistical Modelling Based on Generalised Linear Models, Springer, New York.
- Hand, D.J. and Henley W.E. 1997, "Statistical Classification Methods in Consumer Credit Scoring: A Review", *Journal of the Royal Statistical Association Series A*, vol. 160, Part 3, pp. 523-541.

- Lewis, E. M. 1990, An Introduction to Credit Scoring, Athena Press, San Rafael.
- Mattesini, F. and Messori, M. 2004, L'evoluzione del sistema bancario meridionale: problemi aperti e possibili soluzioni, Il Mulino, Bologna.
- Mays, E. 1998, Credit Risk Modeling: Design and Application, New York: Glenlake
- Mester, L. J. 1997, "What's the Point of Credit Scoring?", *Business Review*, no. 5, pp. 3-16.
- Singer, J. R., and Willett J. B. 1993, "It's about time: using discrete-time survival analysis to study duration and the timing of events", *Journal of Educational Statistics*, no. 18, pp. 155-195.
- Thomas, L. C. 2000, "A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers", *International Journal of Forecasting*, no. 16, pp. 149-172.
- Viganò, L. 1993, "A Credit Scoring Model for Development Banks: An African Case Study", *Savings and Development*, vol. 17, no. 4, pp. 441-482