International Journal of Forecasting 28 (2012) 204–215

Contents lists available at SciVerse ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD

Jie Zhang *, Lyn C. Thomas

School of Management, University of Southampton, Building 2 Highfield Campus, Southampton SO17 1BJ, United Kingdom

ARTICLE INFO

Keywords:
Recovery rate
Linear regression
Survival analysis
Mixture distribution
Loss Given Default forecasts

ABSTRACT

Estimating the recovery rate and recovery amount has become important in consumer credit due to the new Basel Accord regulation and the increase in the number of defaulters as a result of the recession. We compare linear regression and survival analysis models for modelling recovery rates and recovery amounts, in order to predict the loss given default (LGD) for unsecured consumer loans or credit cards. We also look at the advantages and disadvantages of using single and mixture distribution models for estimating these quantities.

1. Introduction

The New Basel Accord allows banks to calculate their credit risk capital requirements according to one of two approaches. The first, namely the standardized approach, requires a percentage of the risk weighted assets to be set aside, where the percentage is given in the regulations. The second, known as the internal ratings based (IRB) approach, allows banks to use internal estimates of the components of credit risk to calculate their credit risk capital. Institutions using IRB need to develop methods of estimating the following components for each segment of their loan portfolio:

- PD (probability of default in the next 12 months);
- LGD (loss given default); and
- EAD (expected exposure at default).

Modelling the probability of default (PD) has been the objective of credit scoring systems for fifty years, but modelling LGD is not something that was really addressed in consumer credit prior to the advent of the Basel regulations. Modelling LGD appears to be more difficult than modelling PD, for two reasons. Firstly, much of the data may be censored (debts still being paid) because of the long time scale of recovery. Linear regression does not deal very well with censored data, and even the Buckley-James approach (Buckley & James, 1979) does not cope well with this form of censoring. Second, debtors have different reasons for defaulting, which lead to different repayment patterns. For example, some people do not want to repay, and some people cannot repay because of permanent changes in their situation; while for others the reason for non-repayment may be temporary. One distribution may find it hard to model the outcomes of these different reasons. However, survival analysis can handle censored data, and segmenting the whole default population is helpful in modelling LGD for defaulters with different reasons for defaulting.

Most LGD modelling research has concentrated on corporate lending, where LGD (or its opposite, the recovery rate (RR), where RR $= 1 - $ LGD) was needed as part of the bond pricing formulae. Even in this case, LGD was assumed until a decade ago to be a deterministic value obtained from a historical analysis of bond losses or from bank experience (Altman, Haldeman, & Narayanan, 1977). Only when it was recognised that LGD was part of the pricing formula and that one could use the price of non-defaulted risky bonds to estimate the market's view of LGD were models of LGD developed. If defaults are rare in a particular bond class, then it is likely that the LGD obtained

* Corresponding author. Tel.: +44 23 8059 7718; fax: +44 23 8059 3844.
E-mail addresses: jz3g08@soton.ac.uk (J. Zhang), L.Thomas@soton.ac.uk (L.C. Thomas).

0169-2070/$ – see front matter
doi:10.1016/j.ijforecast.2010.06.002

from the bond price is essentially a subjective judgment by the market. The market also trades defaulted bonds, and thus one can obtain the market values of defaulted bonds directly (Altman & Eberhart, 1994). These values of the LGD, whether obtained from defaulted bonds or implied in the price of non-defaulted bonds, were used to build regression models that related LGD to relevant factors such as the seniority of the debt, country of issue, size of issue, size of the firm and industrial sector of the firm, but most of all to the economic conditions which determined where the economy was in relation to the business cycle. The most widely used model is Moody's KMV model, LossCalc (Gupton, 2005). It transforms the target variable into a normal distribution using a Beta transformation, regresses the transformed target variable on a few characteristics, and then transforms the predicted values back, to get the LGD prediction. Another popular model, Recovery Ratings, was created by Standard and Poor's Ratings Services (Chew & Kerr, 2005); it divides the loans into 6 classes which cover different recovery ranges. Descriptions of the models are given in several books and reviews (Altman, Resti, & Sironi, 2005; De Servigny & Oliver, 2004; Engelmann & Rauhmeier, 2006; Schuermann, 2005).

Such modelling is not appropriate for consumer credit LGD models, since there is no continuous pricing of the debt as there is on the bond market. The Basel Accord (Basel Committee on Banking Supervision, 2004, paragraph 465) suggests using the implied historic LGD as one approach for determining the LGD for retail portfolios. This involves identifying the realised losses (RL) per unit amount loaned in a segment of the portfolio and estimating the default probability PD for that segment, from which one can calculate LGD, since RL = LGD.PD. One difficulty with this approach is that it is often accounting losses that are recorded rather than the actual economic losses. Also, since LGD must be estimated at the segment level of the portfolio, if not at the individual loan level, in some segments there is often insufficient data segments to obtain robust estimates.

The alternative method suggested in the Basel Accord is to model the collection or work out process. Such data were used by Dermine and de Carvalho (2006) for bank loans to small and medium sized firms in Portugal. They used a regression approach, albeit a log–log form of the regression, to estimate LGD.

The idea of using the collection process to model LGD for mortgages was suggested by Lucas (2006). The collection process was split into whether the property was repossessed or not, and the loss if there was repossession. Thus, a scorecard was designed to estimate the probability of repossession, where Loan to Value was key, and then a model was used to estimate the percentage of the estimated sale value of the house that is actually realised at sale time. For mortgage loans, a one-stage model was built by Qi and Yang (2009). They modelled LGD directly, and found that LTV (Loan to Value) was the key variable in the model; they achieved an adjusted R^2 value of 0.610, but this dropped to 0.15 if LTV was excluded.

For unsecured consumer credit, the only available approach is to model the collection process, but now there is no security to be repossessed. The difficulty in such modelling is that the loss given default, or the equivalent recovery rate, depends both on the ability and willingness of the borrower to repay, and on decisions made by the lender as to how vigorously they will pursue the debt. This is identified at a macro level by Matuszyk, Mues, and Thomas (2010), who use a decision tree to model whether the lender will collect in house, use an agent on a percentage commission, or sell off the debts, with different actions putting different limits on the possible LGD. Even if one concentrates on one mode of recovery only (for example, in house collection), it is still very difficult to get good estimates. Matuszyk et al. (2010) look at various versions of regression, while Bellotti and Crook (2009) add economic variables to the regression. Somers and Whittaker (2007) suggest using quantile regression, but the results in terms of R^2 are poor in all cases—between 0.05 and 0.2. Querci (2005) investigated data from an Italian bank on geographic location, loan type, workout process length and borrower characteristics, but concluded that none of them was able to explain LGD, though borrower characteristics were the most effective.

In this paper, we use linear regression and survival analysis models to build predictive models for the recovery rate, and hence LGD. Both single distribution and mixture distribution models are built, and we compare the two approaches. This analysis will give an indication of how important it is to use models—survival analysis based ones—which can cope with censored debts, and will also investigate whether mixed distribution models give better predictions than single distribution models.

The comparison will be made based on a case study involving data from an in-house collection process for personal loans. This consisted of collection data on 27,000 personal loans over the period from 1989 to 2004. In Section 2 we briefly review the theory of linear regression and survival analysis models. In Section 3 we explain the idea of mixture distribution models as they are applied in this problem. In Section 4 we build and compare single distribution models using linear regression and survival analysis based models, while in Section 5 we create mixture distribution models, to enable us to compare them. In Section 6 we summarise the conclusions reached.

2. Single distribution models

2.1. Linear regression model

Linear regression is the most obvious predictive model to use for recovery rate (RR) modelling, and is also widely used for prediction in other financial areas. Formally, a linear regression model fits a response variable y to a function of regressor variables $x_1, x_2, \ldots, x_m$ and parameters. The general linear regression model has the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon, \tag{2.1}$$

where, in this case,

y is the recovery rate or recovery amount;
$\beta_0, \beta_1, \ldots, \beta_m$ are unknown parameters;