UCD GEARY INSTITUTE
DISCUSSION PAPER SERIES

Estimating bank loans loss given default by generalized additive models

Raffaella Calabrese
University of Milano-Bicocca

Geary WP2012/24
October 2012

Estimating bank loans loss given default by generalized additive models

Raffaella Calabrese

With the implementation of the Basel II accord, the development of accurate loss given default models is becoming increasingly important. The main objective of this paper is to propose a new model to estimate Loss Given Default (LGD) for bank loans by applying generalized additive models. Our proposal allows to represent the high concentration of LGDs at the boundaries. The model is useful in uncovering nonlinear covariate effects and in estimating the mean and the variance of LGDs. The suggested model is applied to a comprehensive survey on loan recovery process of Italian banks. To model LGD in downturn conditions, we include macroeconomic variables in the model. Out-of-time validation shows that our model outperforms popular models like Tobit, decision tree and linear regression models for different time horizons.

Key words: downturn LGD, generalized additive model, Basel II

1 Introduction

In Basel II framework, banks adopting the advanced Internal-Rating-Based (IRB) approach are allowed to use their own estimates of Loss Given Default (LGD) , which denotes the loss quota in the case of the borrower's default. LGD is defined as one minus the recovery rate, which denotes the payback quota of the loan. Basel II requires that the internal estimates reflect economic downturn conditions (Basel Committee on Banking Supervision (BCBS), 2004a, paragraph 468). For each exposure, LGD must not be lower than the average long-term loss rate, weighted for all observed defaults for the type of facility in question. Therefore, LGD may exceed the weighted average value when credit losses are higher than average, thus

Raffaella Calabrese
University of Milano-Bicocca
e-mail: raffaella.calabrese1@unimib.it

the downturn LGD is obtained. In the assessment of capital adequacy the downturn LGD is also useful for stress testing purposes (BCBS, 2004a, paragraph 434).

In this paper we focus on modelling and forecasting LGD for Italian bank loans. We propose a regression model for LGD prediction by using generalized additive models (GAM). To our knowledge there is no study applying GAM to LGD prediction. GAM is a generalization of the linear regression model. It replaces the usual linear function of some covariates with a sum of unspecified smooth functions, helping us dicover potential nonlinear shapes of covariate effects.

Data on LGDs typically have a high concentration at total loss and total recovery. For this reason, we consider nontrivial probability masses at the endpoints of LGD (Calabrese and Zenga, 2010; Calabrese, 2012). We suggest to model the total recovery and the total loss by two logistic additive models. The LGDs bounded to the (0,1) interval are modeled by a beta random variable. The beta distribution is well suited to the modeling of LGDs, as it has support [0,1] and, in spite of requiring only two parameters, is quite flexible. Bruche and Gonzalez-Aguado (2008), Gupton et al. (1997), Gupton and Stein (2002) similarly assume that LGD is beta distributed.

To predict LGD on the (0,1) interval, we propose the joint beta additive regression that models jointly the expectation and the dispersion by using two GAMs. The joint beta additive model accommodates skewness and heteroscedastic errors.

Particular conditions are likely to make the debtor either pay full or not pay at all, which can be different from the factors that determine the partial payback. This topic is analysed in many works on recovery risk, e.g. Bellotti and Crook (2012), Calabrese (2012), Grunert and Weber (2008), Schuermann (2005). The main advantage of our proposal is that it allows to model the different influences of the same explanatory variables on the total recovery, the total loss and the partial loss. Also the model suggested by Calabrese (2012) shows this important characteristic and for this reason both the models supply accurate estimates for the endpoints of the LGDs.

The previous proposal (Calabrese, 2012) involves the strong assumption about the linear relationship between the explanatory variable and the predictor. Such assumption may force the fitted relationship away from its natural path at critical points. To overcome this drawback, a pivotal improvement of the model here suggested is the capability of estimating the relationship between the explanatory variables and the predictor, which can not be linear. Analogously to Calabrese (2012), another positive aspect of our proposal is that it allows to estimate both the mean and the variance of the LGD.

We compare our proposal with the joint beta regression model (Calabrese, 2012), the Tobit model (Bellotti and Crook, 2012), the linear regression (Caselli et al., 2008, Grunert and Weber, 2009) and the decision tree model (Bellotti and Crook, 2012)[1] on a comprehensive database of LGDs on Italian bank loans (Banca d'Italia, 2001). This survey is interesting since few analyses of LGDs of bank loans focus on continental Europe. Moreover, we include macroeconomic variables that enable us to obtain estimates of LGDs in downturn conditions. Analogously to some results

[1] We do not consider the fractional response model proposed by Papke and Wooldridge (1996) since it shows less accurate estimates of recovery rates in Calabrese's (2012) analysis.

in the literature (e.g. Acharya et al., 2007; Altman et al, 2005; Bellotti and Crook, 2012; Caselli et al., 2008; Figlewski et al., 2007), these variables are significant in predicting LGD.

Our proposal shows the highest out-of-time predictive accuracy in terms of the mean absolute error and the mean square error for different forecasting periods and for different sample percentages of the extreme values of the LGDs. Finally, it shows the best performance also when the number of years of data included when building the model changes.

The present paper is organized as follows. The next section is a brief literature review. Section 3 presents the regression approach here proposed in which the joint beta additive model is described. In Section 4, the first subsection describes the dataset of the Bank of Italy and the second shows the covariate effects identified through the proposed model on these data. In the following subsection, tree decision, Tobit, linear regression models are compared with those of our proposal on the Bank of Italy's data. Finally, the last section contains some concluding remarks.

2 Literature review

The Basel II (BCBS, 2004) allows banks the opportunity to estimate LGD using their own models via the IRB approach Several studies consider LGDs on corporate bonds (e.g. Bruche and González-Aguado, 2008; Renault and Scaillet, 2004; Schuermann, 2003), while some authors deal with bank loans (e.g. Araten et al., 2004; Asarnow and Edwards, 1995; Calabrese and Zenga, 2010; Caselli et al., 2008; Chalupka and Kopecsni, 2009; Dermine and Neto de Carvalho, 2006; Emery et al., 2004; Grippa et al. 2005; Grunert and Weber, 2009). Since loans are private instruments, few data are available. Noticeably, LGDs on corporate bonds and on bank loans are significantly different (Carty and Lieberman, 1996; Schuermann, 2003).

In recovery risk analysis a pivotal topic is the forecasting of LGDs. In order to predict LGDs, some authors apply a linear regression model. Caselli et al. (2008) examine 11,649 distressed loans to households and small and medium size companies from 1990 to 2004. LGD is estimated from cash-flows recovered after the default event. A similar methodology is applied by Grunert and Weber (2009) on 120 recovery rates of German defaulted companies in the years from 1992 to 2003. Grunert and Weber (2009) attach great importance to very high or very low recovery rates, so they investigate whether some factors influence banks receiving the EAD almost completely or only minimally by using two logistic regression models.

Gupton and Stein (2002) transform the LGDs of 1,800 U.S. defaulted loans, bonds and preferred stock from beta to normal space. They apply a linear regression model to the transformed market prices of the assets soon after the default event. Bruche and González-Aguado (2008) also assume that the recovery rate is beta distributed, but they extend the static beta distribution assumption by modeling the beta parameters as functions of systematic risk. Such model is applied to 2,000 defaulted bonds of US firms from 1974 to 2005. For the bimodal nature of LGD, Bellotti and

Crook (2012) apply Tobit and a decision tree model to 55,000 credit card accounts defaulted in UK over the period 199 to 2005.

To estimate the downturn LGD (BCBS, 2004b, 2005), Basel II (BCBS, 2004a, paragraph 468) suggests that banks have to consider macroeconomic downturn conditions when predicting recovery rates. In particular, the BCBS (2005) states that banks should use the growth of GDP and the rate of unemployment as factors for the recovery rate prediction.

The forecasting of LGD for retail credit using macroeconomic variables is relatively enexplored by literature. The growth rate of GDP is significant in calculating the loss rate for Altman et al. (2005) on US bonds and for Figlewski et al. (2007) also on US bonds. The same variable is not significant for Bruche and González-Aguado (2008) and Acharya et al. (2007). The results agree on the relevance of the unemployment rate to explain the LGD (Acharya et al., 2007; Bellotti and Crook, 2012; Bruche and González-Aguado, 2008, Caselli et al., 2008).

Other macroeconomic covariates chosen in the literature to predict the recovery rates are the interest rate (Bellotti and Crook, 2012; Figlewski et al., 2007), stock market return (Acharya et al., 2007; Figlewski et al., 2007), investment growth (Bruche and González-Aguado, 2008; Caselli et al., 2008) and inflation (Figlewski et al., 2007). The growing empirical evidence shows a negative correlation between default and LGDs, as shown by Bruche and González-Aguado (2008) and Altman et al. (2005).

3 Prediction models

A pivotal characteristic of the LGD distribution is the high concentration of data at total recovery and total loss, as shown by Asarnow and Edwards (1995), Bellotti and Crook (2012), Calabrese and Zenga (2008, 2010), Calabrese (2012), Caselli et al. (2008), Dermine and Neto de Carvalho (2006), Grunert and Weber (2009), Renault and Scaillet (2004), Schuermann (2003). Hence, the estimates of total loss and total recovery are crucially important for banks.

This section analyses the main models proposed to supply accurate estimates for the endpoints of LGDs.

3.1 Tobit model

Since LGD has a truncated distribution, Bellotti and Crook (2012) suggest to apply the two-tailed Tobit model (Maddala, 1987, Chapter 6). To model boundary cases, the latent variable $z = \mathbf{x}'\beta + \varepsilon$ is considered, where $y = min(1, \max(0, z))$. To apply the maximum likelihood method, the conditional distribution of the residual ε is assumed to be normal $N(0, \sigma^2)$ with variance σ^2, so the following log-likelihood function is obtained

$$
\begin{aligned}
l(\beta,\sigma) \; &= \sum_{0<lgd_i<1} log\left[\sigma^{-1}\phi\left(\frac{lgd_i - \mathbf{x}_i'\beta}{\sigma}\right)\right] + \sum_{lgd_i=0} log\left[1-\Phi\left(\frac{\mathbf{x}_i'\beta}{\sigma}\right)\right] \\
&+ \sum_{lgd_i=1} log\left[\Phi\left(\frac{\mathbf{x}_i'\beta - 1}{\sigma}\right)\right] \qquad (1)
\end{aligned}
$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability and cumulative density functions for the standard normal distribution, respectively.

3.2 Decision tree model

Bellotti and Crook (2012) suggest a decision tree model using two logistic regression models for total loss and total recovery, i.e. $LGD = 1$ and $LGD = 0$ respectively, as binary classification problems. Then, an Ordinary Least Squares (OLS) regression is used to model LGD for the group $0 < LGD < 1$.
To describe this model, we use the following parametrization

$$
a = P\{LGD = 1\} \qquad\qquad b = P\{LGD = 0\}. \qquad (2)
$$

and the cumulative distribution function of LGD proposed by Calabrese and Zenga (2010)

$$
F_{LGD}(lgd) = \begin{cases} a & lgd = 0; \\ a + [1-a-b]F_Y(lgd) & lgd \in (0,1) \\ 1 & lgd = 1, \end{cases} \qquad (3)
$$

where Y is a continuous random variable with support $(0,1)$. From the expression (3), the expected value of LGD is given by

$$
E(LGD) = \hat{b}(\hat{a}+\hat{b})[(1-\hat{a}-\hat{b})\hat{y}] \qquad (4)
$$

where $\hat{a}$ and $\hat{b}$ are estimated by two logistic regression models and $\hat{y}$ is computed from a OLS model inder the assumption that the loss is fractional. Bellotti and Crook use the expression (4) to forecast LGD.

3.3 Joint beta regression model

Since the beta density function is flexible, Calabrese (2012) proposes to consider LGD as a mixture of a dummy random variable and a beta random variable. Analogously Bellotti and Crook (2012), two logistic regression models are used to estimate the parameters a and b defined in the equations (2). To estimate the LGD on the interval (0,1), the parametrization suggested by Ferrari and Cribari-Neto (2004) is considered

$$E(Y) = \mu \qquad var(Y) = \frac{\mu(1-\mu)}{\phi+1}, \tag{5}$$

This means that the random variable Y has the following density function

$$f(y;\mu,\phi) = \frac{y^{\mu\phi-1}(1-y)^{\phi-\mu\phi-1}}{B(\mu\phi,\phi-\mu\phi)} \quad 0<y<1 \tag{6}$$

with $0<\mu<1$ and $\phi>0$ and where $B(\cdot,\cdot)$ denotes the beta function .
Generalized Linear Models (McCullagh and Nelder, 1989) model only the mean μ and they consider the precision parameter $\phi = p+q$ as a nuisance parameter. Calabrese (2012) models, jointly, the mean μ and the precision parameter ϕ of the response beta random variable Y. By choosing the logit and the log functions as link functions, it follows that

$$\mu_i = \frac{1}{1+e^{-\mathbf{v}_i'\eta}} \qquad \phi_i = e^{-\mathbf{w}_i'\theta},$$

with $i = 1,2,...,m$, where η and θ are vectors of respectively k and l unknown regression parameters, $\mathbf{v}_i$ and $\mathbf{w}_i$ are the two vectors of observations on respectively k and l covariates ($k+l<m$), which are assumed fixed and known.

3.4 Generalized additive model

Generalized Additive Models (GAMs) are semi-parametric regression type models. They are parametric, in that they require a parametric distribution assumption for the dependent variable, and ”semi” in the sense that the modeling of the parameters of the distribution, as functions of explanatory variables, involve using non-parametric smoothing functions.
Hastie and Tibshirani (1986) propose GAM. These models assume that the mean of the dependent variable depends on an additive predictor through a nonlinear link function. GAM extends the GLM by replacing the linear form $\mathbf{x}'\beta$ with the additive form $\sum_{i=1}^{k} f_i(x_i)$. The smooth function $f_i(\cdot)$ are estimated on data through iterative smoothing operations, i.e. backfitting algorithm (Hastie and Tibshirani, 1990, 90-91).
The main adavantage of GAMs is that they provide a flexible method for identifying nonlinear covariate effects. This means that GAMs can be used to understand the effect of covariates and suggest parametric transformations of the explanatory variables.

4 A new regression model for LGD

Calabrese's (2012) assumption that LGD is a mixture of a dummy random variable and a beta random variable is considered. In order to estimate the vectors **a** and **b**, we propose to apply two logistic additive regression models given by

$$a = \frac{1}{1-\exp[\alpha+\sum_{j=1}^{p} f_j(x_j)]}, \qquad b = \frac{1}{1-\exp[\beta+\sum_{j=1}^{p} g_j(x_j)]}, \tag{7}$$

where $f_j(\cdot)$ and $g_j(\cdot)$ with $j = 1,2,...,p$ are arbitrary smooth functions, α and β are the unknown parametrs and $\mathbf{x}' = [x_1,...,x_p]$ is the covariate vector.

4.1 Joint beta additive regression model

By considering the assumption (3), LGD on the interval (0,1) is represented by the beta random variable Y, with probability density function (6). A relevant disadvantage of the joint beta regression model is the strong asumption about the linear relationship between the explanatory variables and the predictor. We show in the application of this work that some covariates do not satisfied this assumption. To overcome this problem, we suggest to replace teh linear predictor with an additive one.

Following Calabrese (2012), the mean μ and the parameter ϕ, defined in (5), are modeled by using two link functions given by the logit and the log functions, respectively. It follows that

$$\mu = \frac{1}{1+e^{-\eta+\sum_{i=1}^{k} h_i(v_i)}} \qquad \phi = e^{-\theta+\sum_{j=1}^{l} t_j(w_j)}, \tag{8}$$

where η and θ are vectors of respectively $k+1$ and $l+1$ unknown parameters, $h_i(\cdot)$ and $t_j(\cdot)$ are arbitrary smooth functions, **v** and **w** are the two vectors of observations on respectively k and l covariates.

Furthermore, we point out that the mean μ and the parameter ϕ_i depend on two different vectors of covariates, respectively **v** and **w**. Such a characteristic means that some covariates could be determinant only for one of the two parameters ϕ and μ. Furthermore, the variance of Y is a function of μ and ϕ, as given by the second equation in (5) and such a model thus accommodates heteroscedastic errors. Moreover, since the beta distribution is flexible, the skewness is also accommodated. In order to estimate the parameters η and θ and the functions $h_1(\cdot),h_2(\cdot),...,h_k(\cdot)$ and $t_1(\cdot),t_2(\cdot),...,t_l(\cdot)$, a penalized maximum likelihood method performed. The log-likelihood function is such that

$$\begin{aligned} l(\eta,h_i(\cdot),\theta,t_j(\cdot)) = \sum_{i=1}^{m} \Bigg[& ln\Gamma(e^{-\theta+\sum_{j=1}^{l} t_j(w_j)}) - ln\Gamma\left(\frac{e^{\eta+\sum_{i=1}^{k} h_i(v_i)-\theta+\sum_{j=1}^{l} t_j(w_j)}}{1+e^{\eta+\sum_{i=1}^{k} h_i(v_i)}}\right)+ \\ & +\left(\frac{e^{-\theta+\sum_{j=1}^{l} t_j(w_j)}}{1+e^{\eta+\sum_{i=1}^{k} h_i(v_i)}}-1\right) ln(1-y_i) - ln\Gamma\left(\frac{e^{-\theta+\sum_{j=1}^{l} t_j(w_j)}}{1+e^{\eta+\sum_{i=1}^{k} h_i(v_i)}}\right)+ \\ & +\left(\frac{e^{\eta+\sum_{i=1}^{k} h_i(v_i)-\theta+\sum_{j=1}^{l} t_j(w_j)}}{1+e^{\eta+\sum_{i=1}^{k} h_i(v_i)}}-1\right) ln(y_i)\Bigg]. \end{aligned} \tag{9}$$

The smoothing functions $h_1(\cdot),h_2(\cdot),...,h_k(\cdot)$ and $t_1(\cdot),t_2(\cdot),...,t_l(\cdot)$ are estimated by applying the cubic splines (Hastie and Tibshirani, 1990, chapter 2). Under the assumption that all the functions $t_i(\cdot)$ and $h_j(\cdot)$ are twice continuously diffentiable, we maximize a penalized log-likelihood, given by (9) subject to penalty terms of the form $\sum_{j=1}^{l} \lambda_j \int_{-\infty}^{\infty}[t_j''(q)]^2 dq$ and $\sum_{i=1}^{k} \varpi_i \int_{-\infty}^{\infty}[h_i''(q)]^2 dq$ (Hastie and Tibshirani, 1990, 149-151).

We define this approach the *Joint Beta Additive Model* (JBAM) since the distribution of the dependent variable is assumed to be a beta distribution, analogously to Ferrari and Cribari-Neto (2004) and Calabrese (2012). Futhermore, similarly to Calabrese (2012), we model jointly the expectation and the dispersion of the dependent variable but, unlike Calabrese (2012), generalized additive models are used. The main advantage of JBAMs is the capability of identifying nonlinear covariate effects on the transformed (by the link functions) mean and variance of LGDs. In this way, parametric transformation can be suggested to estimate both the mean and the variance of LGDs.

5 Data

The Bank of Italy conducts a comprehensive survey on the loan recovery process of Italian banks in the years 2000-2001. Its purpose is to gather information on the main characteristics of the Italian recovery process and procedures, by collecting information about recovered amounts, recovery costs and timing.

By means of a questionnaire, about 250 banks are surveyed. Since they cover nearly 90% of total domestic assets of 1999, the sample is representative of the Italian recovery process. We consider 134,937 instances for which all the covariate values are known. We highlight that the data concern individual loans which are privately held and not listed on the market. In particular, loans are towards Italian resident defaulted borrowers from 1985 to 1999 and written off by the end of 1999.

The definition of default the Bank of Italy chooses in its survey (Banca d'Italia, 2004 p.II10) is tighter than the one Basel Committee on Banking Supervision (BCBS) (2004a, paragraph 452) proposes. The difference is the inclusion of transitory nonperforming debts.

To constrain LGD within the interval [0,1], we compute it as one minus the expression proposed by Calabrese and Zenga (2010) to compute the recovery rate.

Figure 1 around here

The distribution of the Bank of Italy's LGDs is shown in Figure 1. The mode of LGD distribution is the extreme value 1, with 23.43% of the observations. LGD equal to 0 also shows a high percentage (7.88%). Hence, we obtain a bimodal distribution analogously to that found in Araten et al. (2004), Asarnow and Edwards (1995), Caselli et al. (2008), Friedman and Sandow (2003).
The average LGD is 0.6154, the median value is 0.6667 and the standard deviation is 0.3395. These values show a less efficient Italian recovery process than the one represented by Caselli et al. (2008) for the period 1990-2004. In fact, in that analysis the average LGD is 0.54, the median is 0.56 and the standard deviation is 0.43.

5.1 Covariate effects

Calabrese and Zenga (2008) show that the exposure at default and the presence of collateral or personal guarantee are significant in estimating the recovery rate. For that reason we consider a dummy variable that represents the presence of collateral or a personal guarantee (CG) and the logarithm of the exposure at default (ln EAD) as explanatory variables.
Caselli et al. (2008) show that the LGDs for loans to households and to small and medium enterprises are statistically different. In order to understand if this characteristic is a determinant of LGD, we introduce a dummy variable[2] that is equal to one when the borrower belongs to a consumer family (CF).
Since internal estimates for the LGD must reflect economic downturn conditions (BCBS, 2004 paragraph 468), as explained in Section 2, macroeconomic variables are introduced in the regression model in order to represent the state of the economic cycle.
Compliant to Basel II (BCBS, 2005) and in line with many empirical studies (Acharya et al., 2007; Altman et al., 2005; Bellotti and Crook, 2012; Bruche and González-Aguado, 2008; Caselli et al., 2008; Figlewski et al., 2007), the chosen macroeconomic variables are the interest on delayed payment (Interest rate), the unemployment rate (Unemployment rate), the growth of GDP (GDP growth rate) and the default rate (default rate), all evaluated at the time of default. The source of the first and the third variables is the Statistical Bulletin of the Bank of Italy, for the others the International Monetary Fund.
In this analysis we consider 144,966 loans that defaulted between January 1985 and December 1999 and which are written off within December 1999. We underline that the size of the sample here considered is significantly higher than the sample size considered in most of empirical studies in the literature. For example, Bellotti and Crook (2009) examine over 55,000 credit card accounts in default and Caselli et al. (2008) consider 11,649 bank loans. We specify that the sample size for Grippa et al.

[2] For a dummy variable, it is evident tht only the parameter and not the spline function can be estimated.

(2005)'s multivariate analysis is over 22,000 loans. Although Grippa et al. (2005) consider the same survey of the Bank of Italy (Banca d'Italia, 2001), their sample size is much lower than the one here analysed since they consider only the loans for which all data are available.

Analogously to Calabrese (2012), the model here propsoed allows to analyse the different influences of the same covariates on subsets of LDGs. Calabrese (2012) analyse the covariate effects on two groups: the endpoints and the value belonging to the interval (0,1). On the contrary, in this paper the covariate effects are analysed on three subsets: total, null partial losses. Some authors (e.g. Bellotti and Crook, 2012; Friedman and Sandow, 2003; Grunert and Weber, 2009; Schuermann, 2003) hypothesize that the extreme values of the recovery rates show different characteristics from the ones belonging to the interval (0,1), but they can not verify this statement with an appropriate methodology. In order to achieve this aim, the covariate sets **x**, **v** and **w**, considered in the equations (7) and (8), coincide.

Some authors in the literature, e.g. Bellotti and Crook (2012) and Dermine and de Carvalho (2005), maintain that a good macroeconomic model of LGD should have training data across the entire business cycle. In fact, the Bank of Italy's data concern a long recovery period of 14 years, covering the expansion in the 80s and the recession in the early 90s in Italy.

The covariate effects here analysed are estimated on 134,937 defaults between 1985 and 1998. To measure the severity of multicollinearity we compute the Variance Inflation Factor (VIF) (Greene, 2000, p.257-258) for each macroeconomic variable in an ordinary least square regression model. Since VIF values are all lower than 5, the level of multicollinearity is tolerable.

Table 1 reports[3] the parameter estimates of the intercept and the dummy variables CG (the presence of guarantee) and CF (consumer family).

Table 1 about here

The presence of collateral or personal guarantee strongly affects the mean of recovery rates, as shown by Chalupka and Kopecsni (2009), Friedman and Sandow (2003), Grippa et al. (2005), and the signs of the estimates coincide with expectations. The dummy variable for the consumer family (CF) shows that the probability of null loss (or total loss) is higher (or lower) for consumer family. The cause of this interesting result could be the larger resort to collateral and personal guarantee for consumer families.

The results on the influence of the EAD on the recovery rates are interesting since empirical studies lead to contrasting conclusions on this topic: Asarnow and Edwards (1995), Carty and Lieberman (1996) find no significant influence of the loan size on LGDs, while Dermine and Neto de Carvalho (2006), Grippa et al. (2005) find that the recovery rates decrease when the loan size increases. Figure 1 shows

[3] We obtain these results by using the package R "GAMLSS" (Stasinopoulos and Rigby, 2007) for the JBAM and the logistic additive model.

that the effect of the logarithm of the EAD on the mean of internal LGDs is clearly unlinear. We obtain the same result for the growth rate of GDP and the default rate. Analogously to Bellotti and Crook (2012), the time with a bank has a strong influence on the mean of internal LGDs. as Figure 1 shows.

The effect of the interest on delayed payment coincides with expectations, this could be due to a short-term influence of this macroeconomic variable on LGDs. Since this influence could be of long-term for the unemployment rate, in Figure 1 we obtain a incoherent result for the unemployment rate. Similar results are obtained by Bellotti and Crook (2012) and Calabrese (2012). The correlation between LGD and PD is a pivotal topic in credit risk analysis (Altman et al., 2005a). The contrasting results obtained in the literature (Altman et al., 2005a) could be due to the wrong assumption of a liner relathionsip between the default rate and the LGD.

For semplicity, we report in Figure 1 the covariate effects only on the logit of the mean of the internal LGDs. By applying the model here proposed, the same analysis could be developed for the probability of total or null losses and for the dispersion parameter ϕ defined in (8).

5.2 Model comparisons

In this subsection we compare the predictive accuracy of some popular models (the Tobit, the decision tree and the linear regression models) and the model here proposed. The predictive accuracy of each model is assessed using the Mean Square Error (MSE) and the Mean Absolute Error (MAE). Models with lower MSE and MAE forecast actual LGDs more accurately. Since the developed models may overfit the data, resulting in over-optimistic estimates of the predictive accuracy, the MSE and the MAE must be assessed on a sample which is different from that used in estimating the model parameters. Since this work focuses on predictive accuracy, the models are fitted on data referring to a period of time and the predictive accuracy is measured on a subsequent period. The accuracy so evaluated is known as out-of-time predictive accuracy.

We compare the out-of-time predictive accuracies obtained by the above-mentioned models[4] on the Bank of Italy's data for different time horizons. The time horizon is important to consider since a pivotal decision for a bank is to decide when the recovery process should be written off. On the one hand, a bank continue its recovery process with the aim of increasing the recovered amounts. Perhaps the best solution to choose the right time for written off is to continuously use models with different forecasting periods of time by comparing their results.

Since the Bank of Italy's data cover the period from 1985 to 1999, the predictive accuracy within one year is evaluated on defaults that occurred in 1999 and the models are fitted on loans defaulted from 1985 to 1998. For the forecast within two years the models are fitted on loans defaulted from 1985 to 1997 and the out-of-time sam-

[4] In order to estimate the parameters of the linear regression model the least squares method is applied.

ple is given by the defaults from 1998 to 1999. Finally, for the forecast within three years the models are developed using defaults from 1985 to 1996 and the accuracy is measured on defaults that occurred from 1997 to 1999.
By the results reported in Table 2 our proposal shows the lower MAE and MSE than the correspondent errors of all the analysed models for each forecasting horizon.

Table 2 around here

Since data concern defaults written off by the end of 1999, the percentage of null recoveries in the out-of-time sample decreases as the forecasting horizon increases. In the out-of-time sample of 1 year the percentage of total loss is 61.91%, for 2 years is 38.11% and for 3 years is 32.48%. Thie means that our proposal is preferable for different sample percentages of total and null losses.
We suspect that several actors in the market use only the most recent data when building LGD prediction models. This is justified by the fact that the most recent data best reflect the characteristics of the data on which it will be used. But then the assumption is made that these characteristics change from year to year, and if this is true the developed model will not be interesting anyway since it will only be applicable on contemporary data. To understand the efficiency of such methodology, we compare models estimated on different periods of time. The 12-years model is built on defaults occurred from 1987 to 1998, while the 6-years and the 3-years models are built on data from 1993 to 1998 and from 1996 to 1998, respectively. The forecasting horizons of the three models are one-year (data from 1999). The fact that the 12-years model outperforms the other models on the 1999 data is interesting. This means that even if the characteristics of the data changes, the introduction of macroeconomic variable allows to include these changes in a LGD model, as Bellotti and Crook (2012) show. Hence, we can deduce that a financial institution should build LGD prediction model by using all the available data.

6 Conclusions remarks

Following the incresing use of IRB approach, LGD modeling has become one of the leading topics in modern finance. In this paper, we propose a new regression model for LGD. The main advantage of our proposal is that it allows to uncover nonlinear covariate effects. Another positive aspect is that our proposal allows to estimate both the mean and the variance of LGD. In particular, banks and financial institutions would be able to improve their credit recovery measurement on the basis of our proposal.

References

1. Acharya Viral V., Bharath Sreedhar T., Srinivasan A.: Does industry-wide distress affect defaulted firms? Evidence from creditor recoveries. Journal of Financial Economics **85**(3), 787–821 (2007)
2. Altman E. I., Resti A., Sironi A.: The PD/LGD Link: Implications for Credit Risk Modeling. In: Altman E. I. , Resti A. and Sironi A. (eds.) The Next Challenge in Credit Risk Management, pp. 253-266. Riskbooks, London (2005)
3. Altman E. I., Brady B., Resti A., Sironi A.: The link between default and recovery rates. Theory, empirical evidence and inplications. Journal of Business **78**, 2203–2228 (2005b)
4. Araten M., Jacobs Jr. M., Varshney P.: Measuring LGD on commercial loans: An 18-year internal study. Journal of Risk Management Association **4**, 96–103 (2004)
5. Asarnow, E., Edwards, D.: Measuring loss on default bank loans: A 24-year study. Journal of Commercial Lending. **77**, 11–23 (1995)
6. Banca d'Italia: Principali risultati della rilevazione sull'attivitá di recupero dei crediti. Bollettino di Vigilanza. 12, December (2001)
7. Basel Committee on Banking Supervision: International convergence of capital measurement and capital standards: A revised framework. Bank for International Settlements. Basel, June (2004a)
8. Basel Committee on Banking Supervision: Background note on LGD quantification. Bank for International Settlements. Basel, December (2004b)
9. Basel Committee on Banking Supervision Guidance on paragraph 468 of the framework document. Bank for International Settlements. Basel, July (2005)
10. Bastos J. A.: Forecasting bank loans loss-given-default. Journal of Banking and Finance **34**(10), 2510–2517 (2010)
11. Bellotti T. and Crook J.: Loss given default models incorporating macroeconomic variables for credit cards. **28**, 171–182 (2012)
12. Bruche, M., Gonzalez-Aguado C.: Recovery Rates, Default Probabilities and the Credit Cycle. CEMFI, working paper (2008)
13. Calabrese, R., Zenga, M.: Measuring loan recovery rate: Methodology and empirical evidence. Statistica & Applicazioni. **6**, 193–214 (2008)
14. Calabrese, R., Zenga, M.: Bank loan recovery rates: Measuring and nonparametric density estimation. Journal of Banking and Finance **34**(5), 903–911 (2010)
15. Calabrese, R.: Predicting bank loan recovery rates with mixed continuous-discrete model. Applied Stochastic Models in Business and Industry. To appear
16. Carty, L., Lieberman, D.: Defaulted bank loan recoveries. Moody's special comment. November (1996)
17. Caselli, S., Gatti, S. Querci, F.: The sensitivity of the loss given default rate to systematic risk: New empirical evidence on bank loans. Journal of Financial Services Research. **34**, 1–34 (2008)
18. Chalupka R. and Kopecsni, J.; Modeling Bank Loan LGD of Corporate and SME Segments: A Case Study. Czech Journal of Economics and Finance. **59**, 360–382 (2009)
19. Dermine, J., Neto de Carvalho, C.: Bank loan losses-given-default: A case study. Journal of Banking and Finance. **30**, 1219–1243 (2006)
20. Emery, K., Cantor, R., Arner, R.: Recovery Rates on North American Syndicated Bank Loans, 1989-2003. Moody's special comment. March (2004)
21. Ferrari, S., Cribari-Neto, F.: Beta regression for modeling rates and proportions. Journal of Applied Statistics. **31**, 799–815 (2004)
22. Figlewski S., Frydman H., Liang W.: Modeling the Effect of Macroeconomic Factors pn Corporate Default and Credit Rating Transitions. NYU Stern Finance Working Paper, November (2007)
23. Friedman, C., Sandow, S.: Ultimate recoveries. Risk. **16**, 69–73 (2003)
24. Frye J.: Depressing Recoveries. Risk **13** 11, 108–111 (2000)
25. Greene W. H.: Econometric Analysis. Prentice Hall, New York (2000)

26. Grippa, P., Iannotti, S., Leandri, F.: Recovery rates in the banking: Stylised facts emerging from Italian experience. In: Altman E. I. , Resti A. and Sironi A. (eds.) The Next Challenge in Credit Risk Management, pp. 121-141. Riskbooks, London (2005)
27. Grunert, J., Weber, M.: Recovery rate of commercial lending: Empirical evidence for German companies. Journal of Banking and Finance. **33**, 505–513 (2009)
28. Gupton, G. M., Finger, C. C., Bhatia, M.: CreditMetrics. Technical document, J. P. Morgan (1997)
29. Gupton, G. M., Stein, R. M.: LosscalcTM: Model for predicting Loss Given Default (LGD), Moody's Investors Service (2002)
30. Hagmann, M., Renault O., Scaillet, O.: Estimation of Recovery Rate Densities: Non-parametric and Semi-parametric Approaches versus Industry Practice. In: Altman E. I. , Resti A., Sironi A. (eds.) The Next Challenge in Credit Risk Management, pp. 323-346. Riskbooks, London (2005)
31. Hastie, T., Tibshirani, R.: Generalized additive models. Statistical Science, 1, 297-318.
32. Hosmer, D. W., Lemeshow, S. Applied logistic regression. Wiley, New York (2000)
33. Maddala, G. S.: Limited-Dependent and Qualitative Variables in Econometrics. Econometric society monographs, Cambridge (1987)
34. Papke, L. E., Wooldridge, J. M.: Econometric Methods for Fractional Response Variables With an Application to 401(K) Plan Participation Rates. Journal of Applied Econometrics **11**, 619–632 (1996)
35. Renault, O., Scaillet, O.: On the Way to Recovery: A Nonparametric Bias Free Estimation of Recovery Rate Densities. Journal of Banking and Finance **28**, 2915–2931 (2004)
36. Stasinopoulos, D. M., Rigby, R. A.: Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. Journal of Statistical Software **23** (7) (2007)
37. Renault, O., Scaillet, O.: On the Way to Recovery: A Nonparametric Bias Free Estimation of Recovery Rate Densities. Journal of Banking and Finance **28**, 2915–2931 (2004)
38. Schuermann, T.: What Do We Know About Loss Given Default? Recovery Risk. Working Paper, Federal Reserve Bank of New York (2003)

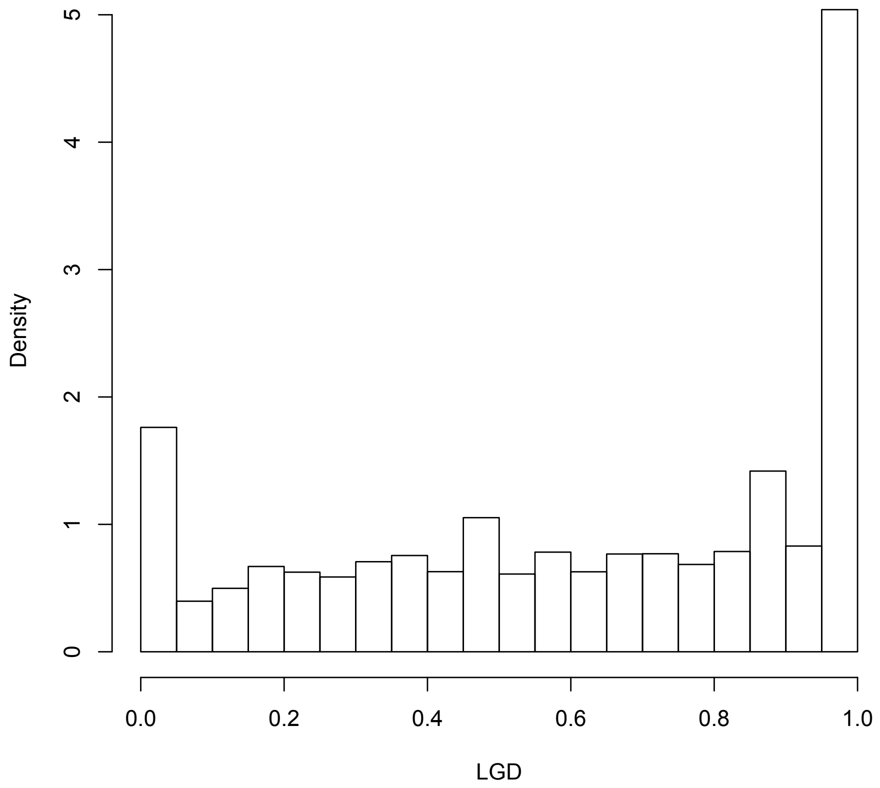
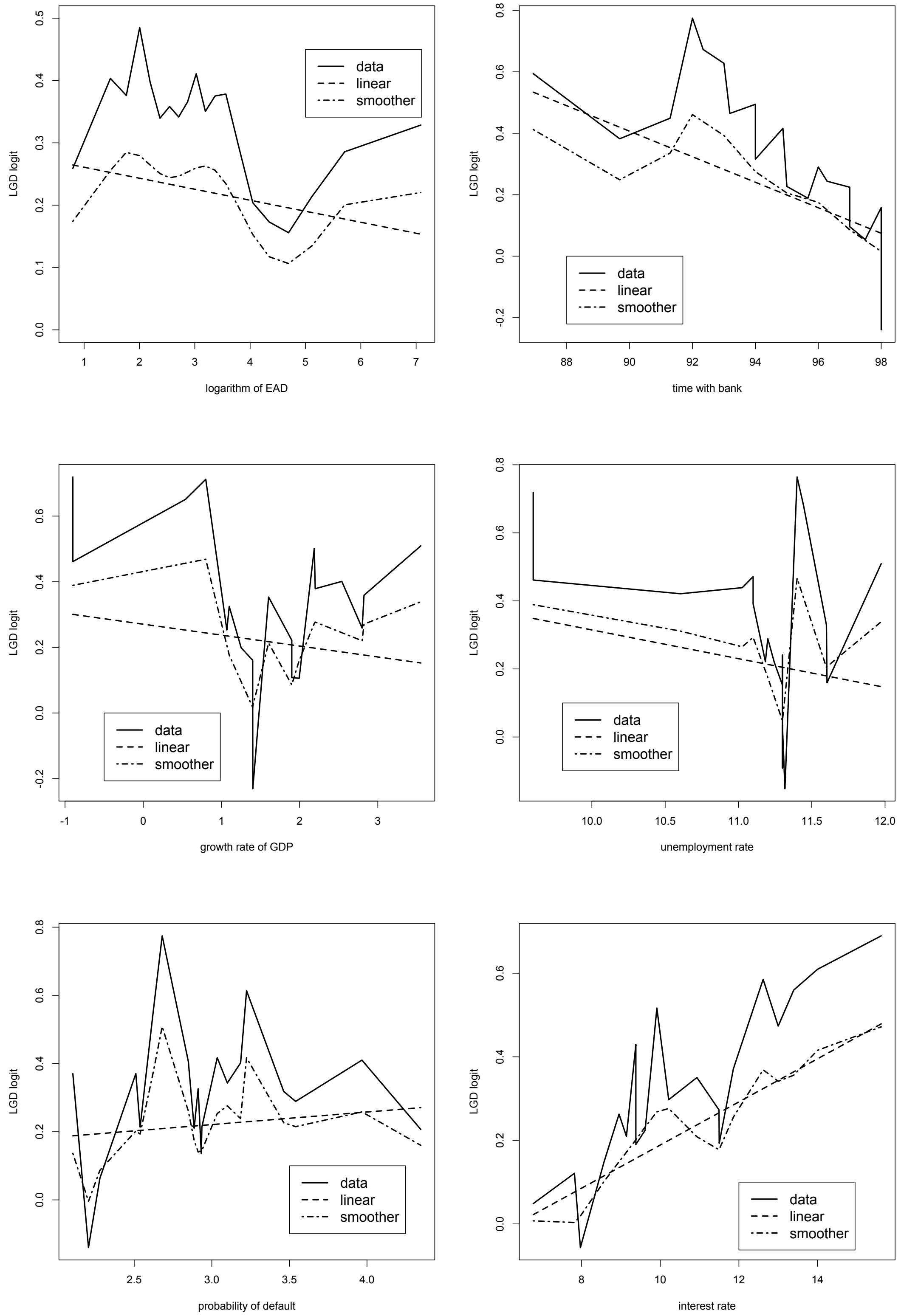


Fig. 1 The distribution of the Bank of Italy's LGDs.


LGD logit
0.0
0.1
0.2
0.3
0.4
0.5
data
linear
smoother
1
2
3
4
5
6
7
logarithm of EAD
LGD logit
-0.2
0.0
0.2
0.4
0.6
0.8
data
linear
smoother
88
90
92
94
96
98
time with bank
LGD logit
-0.2
0.0
0.2
0.4
0.6
data
linear
smoother
-1
0
1
2
3
growth rate of GDP
LGD logit
0.0
0.2
0.4
0.6
0.8
data
linear
smoother
10.0
10.5
11.0
11.5
12.0
unemployment rate
LGD logit
0.0
0.2
0.4
0.6
0.8
data
linear
smoother
2.5
3.0
3.5
4.0
probability of default
LGD logit
0.0
0.2
0.4
0.6
data
linear
smoother
8
10
12
14
interest rate

JBAM for μ				
Variable	*Estimate*	*Std. Error*	*t value*	$Pr(>\|t\|)$
Intercept	2.557	0.202	12.683	7.824e-37
CF	0.001	0.007	0.067	9.465e-01
CG	-0.485	0.015	-31.858	1.396e-221
JBAM for σ				
Variable	*Estimate*	*Std. Error*	*t value*	$Pr(>\|t\|)$
Intercept	-1.125	0.187	-6.000	1.978e-09
CF	-0.068	0.007	-9.723	2.473e-22
CG	-0.036	0.014	-2.588	9.653e-03
Logistic additive regression for zeros				
Variable	*Estimate*	*Std. Error*	*t value*	$Pr(>\|t\|)$
Intercept	-14.994	0.811	-18.489	3.139e-76
CF	0.100	0.024	4.114	3.886e-05
CG	0.887	0.047	18.900	1.431e-79
Logistic additive regression for ones				
Variable	*Estimate*	*Std. Error*	*t value*	$Pr(>\|t\|)$
Intercept	-2.939	1.566	-1.876	6.059e-02
CF	-0.663	0.020	-33.874	1.795e-250
CG	-1.051	0.059	-17.763	1.637e-70

Table 1 Parameter estimates on 134,937 defaults occurred in Italy from 1985 to 1998.

FORECASTING HORIZON	*ERRORS*	*MODELS*			
		JBAM	*JBM*	*Tobit*	*Linear*
1 year	MAE	0.2529	0.3543	0.3378	0.3014
	MSE	0.1278	0.1563	0.1516	0.1439
2 years	MAE	0.3308	0.3681	0.3484	0.4661
	MSE	0.1634	0.1898	0.1400	0.3213
3 years	MAE	0.3404	0.3928	0.3693	0.5042
	MSE	0.1594	0.2101	0.1448	0.3647

Table 2 Forecasting accuracy measures of regression models over different forecasting horizons on the out-of-time sample.

ESTIMATION PERIOD	*ERRORS*	*MODELS*			
		JBAM	*JBM*	*Tobit*	*Linear*
12 years	MAE	0.2601	0.3596	0.3712	0.3088
	MSE	0.1293	0.1566	0.1521	0.1457
6 years	MAE	0.2782	0.3610	0.3794	0.3103
	MSE	0.1386	0.1685	0.1552	0.1523
3 years	MAE	0.2897	0.4151	0.4076	0.3167
	MSE	0.1472	0.1753	0.1598	0.1581

Table 3 Forecasting accuracy measures of regression models over different horizons of the sample.