Rapporto n. 200

Predicting Bank Loan Recovery Rates in a Mixed Continuous-Discrete model

Raffaella Calabrese

Novembre 2010

Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali
Università degli Studi di Milano Bicocca
Via Bicocca degli Arcimboldi 8 - 20126 Milano - Italia
Tel +39/02/64483102/3 - Fax +39/2/64483105
Segreteria di redazione: Andrea Bertolini

Predicting Bank Loan Recovery Rates in a Mixed Continuous-Discrete model

Raffaella Calabrese

I propose to consider the recovery rate as a mixed random variable, obtained as the mixture of a Bernoulli and a beta random variables. I estimate the mixture weights and the Bernoulli parameter by two logistic regression models. For the recovery rates belonging to the interval (0,1), I model, jointly, the mean and the dispersion by using two link functions, so I propose the joint beta regression model that accommodates skewness and heteroscedastic errors. The estimation procedure is the maximum likelihood method. Finally, the methodological proposal is applied to a comprehensive survey on loan recovery process of Italian banks. Macroeconomic variables are relevant to explain the recovery rate and allow to estimate it in downturn conditions, as Basel II requires.

Key words: downturn recovery rate, mixed random variable, joint beta regression model, logistic regression model

1 Introduction

While the prediction of the probability of default has been the subject of many analyses during the past few decades, the prediction of recovery rates is relatively unexplored by the literature. The recovery rate is defined as the payback quota of the loan. The Basel II Accord (Basel Committee on Banking Supervision (BCBS), 2004a, paragraph 286-317) prefers to consider the "Loss Given Default"(LGD) which denotes the loss quota in the case of the borrower's default and it is defined as one minus the recovery rate. In this framework, banks adopting the advanced Internal-Rating-Based (IRB) approach are allowed to use their own estimates of LGDs. Basel II requires that the internal estimates reflect economic downturn conditions wher-

Raffaella Calabrese
University of Milano-Bicocca, Via Bicocca degli Arcimboldi, 6, Milano 20123, e-mail: raffaella.calabrese1@unimib.it

ever necessary to capture risk accurately (BCBS, 2004a, paragraph 468). Therefore, for each exposure, the LGD must not be lower than the average long-term loss rate, weighted for all observed defaults for the type of facility in question. Moreover, banks must account for the possibility that the LGD may exceed the weighted average value when credit losses are higher than average, thus modeling the so-called “downturn LGD”. In the assessment of capital adequacy the downturn LGD is also useful to stress testing processes (BCBS, 2004a, paragraph 434).

Within this research field, the main aim of this work is to propose a regression model for the recovery rate in order to model jointly its mean and variance, given some explanatory variables. At first, to represent the high concentration of data at total recovery and total loss I consider the assumption introduced by Calabrese and Zenga (2010): the recovery rate is a mixed random variable, given by the mixture of a Bernoulli random variable and a continuous random variable with support (0,1). In particular, in this work I assume that the continuous part is modeled by a beta random variable, consistently with the distribution function estimate obtained by Calabrese and Zenga (2010).

In order to analyze the influences of some explanatory variables on the continuous part, I propose the joint beta regression that models jointly the expectation and the dispersion by using two link functions and two covariate sets that could be different. The model parameters are estimated by the maximum likelihood method. The joint beta regression model accommodates skewness and heteroscedastic errors. To estimate the Bernoulli parameter and the mixture weights I propose to apply two logistic regression models.

The main advantage of my proposal is that it allows to examine the different influences of the same covariates on the extreme values and the recovery rates belonging to the interval (0,1). By this characteristic I can analyse the assumption that special conditions make a debtor pay back the full amount of debt or to pay back nothing, rather than just a portion. This topic is pivotal in many works on the recovery risk, e.g. Bellotti and Crook (2009), Grunert and Weber (2008), Schuermann (2005). Furthermore, the regression model here proposed supplies accurate estimations for the extreme values of the recovery rates, which are really important in recovery risk analysis. Finally, another positive aspect of my proposal is that it allows to estimate both the mean and the variance of the recovery rate knowing the covariate values.

At last, I apply the proposed approach to a comprehensive database (Banca d’Italia, 2001) of recovery rates on Italian bank loans. This survey is really important since very few analyses on recovery rates of bank loans focus on continental Europe. Moreover, I introduce some macroeconomic variables that let to obtain an estimate in downturn conditions and to stress testing. Analogously to some results in the literature (e.g. Acharya et al., 2007; Altman et al, 2005; Bellotti and Crook, 2009; Caselli et al., 2008; Figlewski et al., 2007), these variables are significant to estimate the recovery rates.

On the Bank of Italy’s data I compare the predictive accuracy of my proposal with those of the fractional response model, proposed by Papke and Wooldridge (1996). This model is applied by Bastos (2010), Chalupka and Kopecsni (2009), Dermine and Neto de Carvalho (2006), Grippa et al. (2005) with different link functions (the logit, the log-log and the complementary log-log functions). The model here proposed shows the highest out-of-time predictive accuracy in terms of the mean absolute error and the mean square error for different forecasting periods of time.

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 regression model is described. In section 4 the first subsection describes the dataset of the Bank of Italy and the second shows the estimation results by applying the proposed model to these data. Then, subsection 4.3 presents the fractional response model, proposed by Papke and Wooldridge (1996). In the following subsection the out-of-time predictive accuracies of the fractional response models with different link functions are compared with that of my proposal on the Bank of Italy's data. Finally, the last section is devoted to conclusions. In appendix I report the score functions and the Fisher information matrix of the parameters of the joint beta regression model.

2 Literature review

In order to be compliant with the IRB approach of Basel II, banks must estimate the probability of default, the LGD and the Exposure At Default (EAD). Most empirical research focuses on modeling and estimating default probabilities, while only recently the recovery analysis is attracting attention. Several studies consider recovery rates on corporate bonds (e.g. Bruche and González-Aguado, 2008; Renault and Scaillet, 2004; Schuermann, 2003), while few 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 is available for empirical analyses. Noticeably, recovery rates on corporate bonds and on bank loans are significantly different. In particular, Carty and Lieberman (1996), Schuermann (2003) show that the average recovery rate on bank loans is higher than the one on bonds. On the contrary, the results about the variability are discordant: Araten et al. (2004) assert that LGDs on bank loans have greater variability than recovery rates on bonds, instead Schuermann (2003) finds the contrary.

Most of these empirical studies concern the U.S. banking system (Araten et al., 2004; Asarnow and Edwards, 1995; Bruche and González-Aguado, 2008; Carty and Lieberman, 1996; Emery et al., 2004; Friedman and Sandow, 2003; Gupton and Stein, 2002; Renault and Scaillet, 2004). More recent works consider the European market (Bastos, 2009; Bellotti and Crook, 2009; Calabrese and Zenga, 2010; Caselli et al., 2008; Dermine and Neto de Carvalho, 2006; Grunert and Weber, 2009).

In recovery risk analysis a pivotal topic is the forecasting of recovery rates. Gupton and Stein (2002) assume that the recovery rate is beta distributed, so they transform the LGDs of 1,800 U.S. defaulted loans, bonds and preferred stock from Beta to Normal space. Finally, on the transformed market prices of these instruments after default they apply a linear regression model. The model validation is performed out-of-time, as later-explained this means that the model is fit using data from one time period and tested on a subsequent period. On the one hand, also Bruche and González-Aguado (2008) assume that the recovery rate is beta distributed. On the other hand, they extend the static beta distribution assumption of CreditMetrics (Gupton et al., 1997) and KMV Portfolio Manager by modeling the beta parameters as functions of systematic risk. In particular, Bruche and González-Aguado (2008) evaluate the out-of-time predictive accuracy of the model by the log-likelihood ratio, the Akaike's Information Criterion and the Bayesian information criterion on 2,000 defaulted bonds of US firms from 1974 to 2005.

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. They test several linear regression models with different explanatory variables and they evaluate the out-of-time predictive accuracy. 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. Unlike the previous work, they evaluate the goodness of fit by an adjusted R^2. Analogously to this work, Grunert and Weber (2009) attach a great importance to very high or very low recovery rates, so they investigate whether some factors influence the fact that banks receive the EAD almost completely or only minimally by two logistic regression models.

A model widely applied to forecast the recovery rate is the the fractional response model proposed by Papke and Wooldridge (1996) and explained in subsection 4.3. Dermine and Neto de Carvalho (2006) apply the fractional response model with the log-log link function on 373 non-performing loans granted to SMEs over the period 1995 to 2000. On the contrary, Grippa et al. (2005) choose the logit link function for the fractional response model that is applied on more than 22,000 recovery rates gathered by the same survey of the Bank of Italy analysed in this work. I specify that they apply a different expression to compute the recovery rate from the one used in this work and proposed by Calabrese and Zenga (2008, 2010).

Chalupka and Kopecsni (2009) compare the fractional response models with different link functions (logit, log-log and complementary log-log functions) and they obtain that the log-log link function performs better to LGDs of Czech firm loans defaulted in the period 1989-2007. Bellotti and Crook (2007) evaluate the performance of different regression approaches on over 55,000 credit loans in default over the period 1999 to 2005 in UK and they obtain that the fractional logit regression shows the best out-of-sample predictive accuracy in terms of mean absolute error.

Although the need to estimate the downturn LGD is clearly framed (BCBS, 2004b, 2005), Basel II does not provide a specific approach that banks must use in calculating this variable. In particular, the paragraph 468 (BCBS, 2004a) states that banks have to consider macroeconomic downturn conditions when predicting recovery rates. The BCBS (2005) states that banks should use the growth of GDP and the rate of unemployment as factors for the recovery rate prediction.

Different conclusions are obtained on this topic. 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 and al. (2007). On the contrary, the results agree on the relevance of the unemployment rate to explain the LGD (Acharya et al., 2007; Bellotti and Crook, 2009; 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, 2009; Figlewski et al., 2007), the stock market return (Acharya et al., 2007; Figlewski et al., 2007), the investment growth (Bruche and González-Aguado, 2008; Caselli et al., 2008) and the inflation (Figlewski et al., 2007).

A pivotal topic in the recovery risk analysis is the relationship between the default probability and the recovery rate (Altman et al., 2005). Bruche and González-Aguado (2008) and Altman et al. (2005) agree that recovery rates and default rates are strongly correlated.

3 Modeling approach

Analogously to many models in the literature (e.g. Gupton et al., 1997), in this work I consider the recovery rate as a random variable. Moreover, some approaches (Gupton et al., 1997; Gupton and Stein, 2002) assume that the recovery rate is a beta random variable.

A pivotal characteristic of the recovery rate distribution is the high concentration of data at total recovery and total loss, as showed by Asarnow and Edwards (1995), Calabrese and Zenga (2008, 2010), 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.

In order to supply accurate estimations for the extreme values, Calabrese and Zenga (2010) propose to consider the recovery rate R as a mixed random variable, given by the mixture of a Bernoulli random variable and a continuous random variable Y with support (0,1)

$$F_R(r) = \begin{cases} P\{R=0\} & \text{r=0;} \\ P\{R=0\}+[1-P\{R=0\}-P\{R=1\}]F_Y(r) & \text{r}\in\text{(0,1)} \\ 1 & \text{r=1.} \end{cases} \tag{1}$$

where F_Y denotes the cumulative distribution function of the random variable Y and $P\{R=j\}$ is the probability that the recovery rate R is equal to j with $j=0,1$. Since the beta probability density function is flexible, I assume that Y is a beta random variable. By a nonparametric density estimation Calabrese and Zenga (2010) show that this parametric model provides a good fit to data.

Hence, an important issue is to propose an estimation methodology for regression model whose dependent variable is a mixed random variable. In order to understand

the determinants of the mean μ of the dependent variable, the generalized linear models (McCullagh and Nelder, 1989) apply a strictly monotonic and twice differentiable *link function* g$(\cdot)$ such that $g(\mu) = x'\alpha$. There are several possible choices for the link function $g(\cdot)$. When $0 < \mu < 1$, a link function should satisfy the condition that maps the interval (0,1) onto the whole real line.
The recovery rates $r_1, r_2, ..., r_n$ are assumed to be the observed values of independent random variables $R_1, R_2, ..., R_n$ such that R_i has cumulative distribution function given by (1). In order to propose the regression model for the recovery rates $R_1, R_2, ..., R_n$ I prefer to use the following parametrization

$$a_i = \frac{P\{R_i = 1\}}{P\{R_i = 0\} + P\{R_i = 1\}} \qquad b_i = P\{R_i = 0\} + P\{R_i = 1\}. \tag{2}$$

The beta density function of the random variable Y_i with parameters $p_i > 0$ and $q_i > 0$ is given by

$$f(y_i; p_i, q_i) = \frac{y_i^{p_i - 1}(1 - y_i)^{q_i - 1}}{B(p_i, q_i)} = \frac{\Gamma(p_i + q_i)}{\Gamma(p_i)\Gamma(q_i)} y_i^{p_i - 1}(1 - y_i)^{q_i - 1} \quad 0 < y_i < 1 \tag{3}$$

where $B(\cdot,\cdot)$ denotes the beta function and $\Gamma(\cdot)$ the Gamma function.
By using (1) and (2) and by considering the n-vectors $\mathbf{a}' = [a_1, a_2, ..., a_n]$, $\mathbf{b}' = [b_1, b_2, ..., b_n]$, $\mathbf{p}' = [p_1, p_2, ..., p_n]$ and $\mathbf{q}' = [q_1, q_2, ..., q_n]$, the log-likelihood function based on a sample of n independent random variables $R_1, R_2, ..., R_n$ is

$$\begin{aligned} l(\mathbf{a}, \mathbf{b}, \mathbf{p}, \mathbf{q}; \mathbf{r}) = & \sum_{r_i = 0} \ln(1 - a_i) + \sum_{r_i = 1} \ln a_i + \sum_{r_i = 0} \ln b_i + \sum_{r_i = 1} \ln b_i + \\ & + \sum_{0 < r_i < 1} \ln(1 - b_i) + \sum_{0 < r_i < 1} \ln f(p_i, q_i, r_i) \end{aligned} \tag{4}$$

where $f(p_i, q_i, r_i)$ is the beta probability density function defined in (3).
I consider the Bernoulli random variable

$$Z_i = \begin{cases} 1, \ \{R_i = 0\} \bigcup \{R_i = 1\}; \\ 0, \ \text{otherwise.} \end{cases} \tag{5}$$

The log-likelihood function (4) so becomes

$$\begin{aligned} l(\mathbf{a}, \mathbf{b}, \mathbf{p}, \mathbf{q}; \mathbf{r}) = & \sum_{r_i = 0 \bigvee r_i = 1} [r_i \ln a_i + (1 - r_i) \ln(1 - a_i)] + \\ & + \sum_{z_i = 0 \bigvee z_i = 1} [z_i \ln b_i + (1 - z_i) \ln(1 - b_i)] + \sum_{0 < r_i < 1} \ln f(p_i, q_i, r_i). \end{aligned} \tag{6}$$

It is important to note that in (6) the addend of the first sum represents the log-likelihood function of the Bernoulli random variable I_i

$$I_i = \begin{cases} 1, \ R_i = 1; \\ 0, \ R_i = 0 \end{cases}$$

where $P\{I_i = 1\} = a_i$. Analogously, the addend of the second sum represents the log likelihood function of the Bernoulli random variable Z_i, defined in (5), where $P\{Z_i = 1\} = b_i$. For that reason, in order to estimate the vectors **a** and **b**, I propose to consider two logistic regression models given by

$$a_i = \frac{\exp(\mathbf{x}_i'\alpha)}{1+\exp(\mathbf{x}_i'\alpha)} \qquad b_i = \frac{\exp(\mathbf{x}_i'\beta)}{1+\exp(\mathbf{x}_i'\beta)} \tag{7}$$

where $\alpha' = [\alpha_0, \alpha_1, ..., \alpha_p]$ and $\beta' = [\beta_0, \beta_1, ..., \beta_p]$ are the unknown parameter vectors and $\mathbf{x}_i' = [1, x_{i1}, ..., x_{ip}]$ is the covariate vector. This approach is similar to the one proposed by Bellotti and Crook (2009) that apply two logistic regression and the ordinary least square methodologies in a decision tree model.

By considering the equations (7), the log-likelihood function (6) can be expressed also as a function of α and β

$$\begin{aligned} l(\alpha,\beta,\mathbf{p},\mathbf{q};\mathbf{r}) &= \sum_{r_i=0 \bigvee r_i=1} [r_i\mathbf{x}_i'\alpha - \ln(1+\exp(\mathbf{x}_i'\alpha))] + \\ &+ \sum_{z_i=0 \bigvee z_i=1} [z_i\mathbf{x}_i'\beta - \ln(1+\exp(\mathbf{x}_i'\beta))] + \\ &+ \sum_{0<r_i<1} \ln f_Y(p_i,q_i,r_i). \end{aligned} \tag{8}$$

The maximum likelihood method is applied to estimate the unknown parameters α and β. The derivative of the log likelihood function (8) with respect to α and β are respectively

$$\begin{aligned} \frac{\partial}{\partial\alpha_j} l(\alpha,\beta,\mathbf{p},\mathbf{q};\mathbf{r}) &= \sum_{r_i=0 \bigvee r_i=1} \left[x_{ij}\left(r_i - \frac{\exp(\mathbf{x}_i'\alpha)}{1+\exp(\mathbf{x}_i'\alpha)} \right) \right] \\ \frac{\partial}{\partial\beta_j} l(\alpha,\beta,\mathbf{p},\mathbf{q};\mathbf{r}) &= \sum_{z_i=0 \bigvee z_i=1} \left[x_{ij}\left(z_i - \frac{\exp(\mathbf{x}_i'\beta)}{1+\exp(\mathbf{x}_i'\beta)} \right) \right] \end{aligned} \tag{9}$$

with $j = 0,1,...,p$ and $x_{i0} = 1\ \forall i$. By making the score functions for α and β (9) equal to zero, the maximum likelihood estimates of α and β, respectively, are obtained by using a nonlinear optimization algorithm, such as a Newton algorithm or a quasi-Newton algorithm (McLachlan and Krishnan, 1997).

By applying the model (1) and by knowing the covariates **x**, in order to estimate the mean and the variance of the recovery rates I need to estimate the two vectors **p** and **q** that represent the parameters of the beta probability density function $f(p_i,q_i,r_i)$ used in the log-likelihood function (8). With this aim I propose the joint beta regression model in the following subsection.

3.1 Joint beta regression model

This subsection focuses just on the continuous part of the recovery rate in the model (1), given by the beta random variable Y. An interesting model where the dependent variable is beta distributed is the beta regression approach, proposed by Ferrari and Cribari-Neto (2004). In this model they prefer to use a reparameterization that translates p and q into a location parameter μ and a dispersion parameter ϕ

$$E(Y) = \frac{p}{p+q} = \mu \qquad var(Y) = \frac{pq}{(p+q)^2(p+q+1)} = \frac{\mu(1-\mu)}{\phi+1}, \tag{10}$$

where $\phi = p+q$. Similarly to GLM models, Ferrari and Cribari-Neto (2004)'s approach models only the mean μ and they consider the dispersion parameter $\phi = p+q$ as a nuisance parameter. On the contrary, by applying the same reparameterization proposed by Ferrari and Cribari-Neto (2004) I model, jointly, the mean μ and the dispersion parameter ϕ of the response beta random variable Y, following a similar approach to that used by Nelder and Lee (1991).

In particular, let $Y_1, Y_2, ..., Y_n$ be independent beta random variables, where each Y_i, with $i = 1,2,...,n$, follows the density

$$f(y_i;\mu_i,\phi_i) = \frac{y_i^{\mu_i\phi-1}(1-y)^{\phi_i-\mu_i\phi_i-1}}{B(\mu_i\phi_i,\phi_i-\mu_i\phi_i)} = \frac{\Gamma(\phi_i)}{\Gamma(\mu_i\phi_i)\Gamma(\phi_i-\mu_i\phi_i)} y^{\mu_i\phi_i-1}(1-y_i)^{\phi_i-\mu_i\phi_i-1}$$

where $y_i \in (0,1)$, $1 > \mu_i > 0$ and $\phi_i > 0$, with mean and variance given by the equations (10). To joint model the mean μ_i and the parameter ϕ_i, since $0 < \mu_i < 1$ and $\phi_i > 0$ with $i = 1,2,...,n$, I suppose that the link function $g(\cdot)$ is the logit function and the link function $h(\cdot)$ is the log function

$$g(\mu_i) = log\frac{\mu_i}{1-\mu_i} = \mathbf{v}_i'\eta \qquad h(\phi_i) = log(\phi_i) = -\mathbf{w}_i'\theta, \tag{11}$$

it follows that

$$\mu_i = \frac{1}{1+e^{-\mathbf{v}_i'\eta}} \qquad \phi_i = e^{-\mathbf{w}_i'\theta}, \tag{12}$$

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

Furthermore, I point out that the mean μ and the parameter ϕ depend on two different vectors of covariates, respectively $\mathbf{v}$ and $\mathbf{w}$. By such characteristic, the model here proposed can consider some variables in the vector $\mathbf{w}_i$ that are relevant just for the parameter ϕ_i and not for the mean μ_i. I underline that the variance of Y_i is a function of μ_i and ϕ_i, as given by the equation (10) and, as a consequence, of the covariate values $\mathbf{v}_i$ and $\mathbf{w}_i$, so such model accommodates the heteroscedastic errors. Moreover, since the beta distribution is flexible, the skewness and the multimodality are also accommodated.

For the interpretation of the parameter vector η, I suppose that the value of the

j-th regressor (with $j = 1, 2, .., k$) is increased by c unit and all other independent variables remain unchanged. Let μ^* denote the mean of y under the new covariate values, whereas μ denotes the mean of Y under the original covariate values. It is easy to show that

$$e^{c\eta_j} = \frac{\mu^*/(1-\mu^*)}{\mu/(1-\mu)} \tag{13}$$

$e^{c\eta_j}$ equals the odds ratio $\frac{\mu^*/(1-\mu^*)}{\mu/(1-\mu)}$ (with $j = 1, 2, .., k$). Hence, if η_j is positive, from the first equation in (12), the mean μ^* is higher than μ and from the equation (13) also the odds ratio $\frac{\mu^*/(1-\mu^*)}{\mu/(1-\mu)}$ is higher than one.

Analogously, for the interpretation of the parameter vector θ, I suppose that the value of the h-th regressor (with $h = 1, 2, .., m$) is increased by c unit and all other independent variables remain unchanged. I denote the parameter ϕ under the original covariate values and ϕ^* under the new covariate values. It is easy to show that

$$e^{-c\theta_h} = \frac{\phi^*}{\phi} \tag{14}$$

with $h = 1, 2, .., m$. If θ_h is positive, from the second equation in (12) the parameter ϕ^* results lower than ϕ. Moreover, if the mean μ^* does not depend on the h-th regressor, from the second equation in (10) the variance of the dependent variable Y increases when the value of the h-th regressor is increased by c unit. If also the mean depends on the h-th regressor, the variance of Y increases when it is satisfied the following condition

$$\frac{\mu^*(1-\mu^*)}{\mu(1-\mu)} > \frac{\phi^*+1}{\phi+1}. \tag{15}$$

Some sufficient conditions for the inequality (15) to hold are

- $\eta_h > 0, \theta_h > 0$ and $\theta_h > \eta_h$; (16)
- $\eta_h > 0, \theta_h < 0$ and $\theta_h < -\eta_h$. (17)

A sufficient condition so that the variance of Y decreases is

- $\eta_h < 0$ and $\theta_h > 0$, (18)

when μ and ϕ depend on the same h-th covariate.

In order to estimate the two vectors η and θ of parameters, the maximum likelihood method is performed. The log-likelihood function is

$$\begin{aligned} l(\eta,\theta) = \sum_{i=1}^{n} \Bigg[& ln\Gamma(e^{-\mathbf{w}_i'\theta}) - ln\Gamma\left(\frac{e^{\mathbf{v}_i'\eta-\mathbf{w}_i'\theta}}{1+e^{\mathbf{v}_i'\eta}}\right) - ln\Gamma\left(\frac{e^{-\mathbf{w}_i'\theta}}{1+e^{\mathbf{v}_i'\eta}}\right) + \\ & + \left(\frac{e^{\mathbf{v}_i'\eta-\mathbf{w}_i'\theta}}{1+e^{\mathbf{v}_i'\eta}} - 1\right) ln(y_i) + \left(\frac{e^{-\mathbf{w}_i'\theta}}{1+e^{\mathbf{v}_i'\eta}} - 1\right) ln(1-y_i) \Bigg]. \end{aligned} \tag{19}$$

Since the beta distribution is a two parameter full exponential family and the log-likelihood function satisfies a given condition (Barndorff-Nielsen, 1978, pp. 151), the maximum likelihood estimators exist and are unique.
The score function and the Hessian can be obtained explicitly in terms of the polygamma function, where the polygamma function of order m is defined as the $(m+1)^{th}$ derivative of the logarithm of the gamma function $\Gamma(\cdot)$

$$\frac{\partial^m \psi(z)}{\partial^m z} = \frac{\partial^{m+1} ln\Gamma(z)}{\partial^{m+1} z}.$$

For $m = 0$ this function is called digamma function $\psi(z) = \dfrac{\partial log\Gamma(z)}{\partial z} = \dfrac{\partial \Gamma(z)}{\Gamma(z)}$ for $z > 0$. The score functions are obtained by differentiating the log-likelihood function with respect to the unknown parameters η and θ, respectively,

$$\frac{\partial l(\eta,\theta)}{\partial \eta_j} = \sum_{i=1}^{n} \; x_{ij} \frac{e^{\mathbf{v}_i'\eta - \mathbf{w}_i'\theta}}{\left[1+e^{\mathbf{v}_i'\eta}\right]^2} \left[\phi\left(\frac{e^{-\mathbf{w}_i'\theta}}{1+e^{\mathbf{v}_i'\eta}}\right) - \phi\left(\frac{e^{\mathbf{v}_i'\eta - \mathbf{w}_i'\theta}}{1+e^{\mathbf{v}_i'\eta}}\right) + log\frac{y_i}{1-y_i}\right]$$

$$\begin{aligned}\frac{\partial l(\eta,\theta)}{\partial \theta_h} = \sum_{i=1}^{n} \; -w_{ih} \frac{e^{-\mathbf{w}_i'\theta}}{1+e^{\mathbf{v}_i^T\eta}} \Bigg[& \left(1+e^{\mathbf{v}_i'\eta}\right)\phi\left(e^{-\mathbf{w}_i'\theta}\right) - e^{\mathbf{v}_i'\eta}\phi\left(\frac{e^{\mathbf{v}_i'\eta-\mathbf{w}_i'\theta}}{1+e^{\mathbf{v}_i'\eta}}\right) + \\ & -\phi\left(\frac{e^{-\mathbf{w}_i'\theta}}{1+e^{\mathbf{v}_i'\eta}}\right) + log(1-y_i) + e^{\mathbf{v}_i'\eta} log(y_i)\Bigg],\end{aligned} \tag{20}$$

with $j = 1,2,...,k$; $h = 1,2,...,m$, where y_i is a realization of the recovery rate with $0 < y_i < 1$ and $i = 1,2,...,n$.
The asymptotic standard errors of the maximum likelihood estimators of the parameters in the models are given by the Fisher information matrix whose elements are

$$-E\left(\frac{\partial^2 l(\eta,\theta)}{\partial \eta_j \partial \eta_q}\right) = \sum_{i=1}^{n} \; \frac{v_{ij}v_{iq}e^{\mathbf{v}_i^T\eta - 2\mathbf{w}_i^T\theta}\left[1-e^{\mathbf{v}_i^T\eta}\right]}{\left[1+e^{\mathbf{v}_i^T\eta}\right]^3} \left[\phi'\left(\frac{e^{\mathbf{v}_i^T\eta-\mathbf{w}_i^T\theta}}{1+e^{\mathbf{v}_i^T\eta}}\right) + \phi'\left(\frac{e^{-\mathbf{w}_i^T\theta}}{1+e^{\mathbf{v}_i^T\eta}}\right)\right]$$

$$\begin{aligned}-E\left(\frac{\partial^2 l(\eta,\theta)}{\partial \theta_h \partial \theta_u}\right) = \sum_{i=1}^{n} \; w_{ih}w_{ui}e^{-\mathbf{w}_i^T\theta} \Bigg[& \left(\frac{e^{\mathbf{v}_i^T\eta}}{1+e^{\mathbf{v}_i^T\eta}}\right)^2 \phi'\left(\frac{e^{\mathbf{v}_i^T\eta-\mathbf{w}_i^T\theta}}{1+e^{\mathbf{v}_i^T\eta}}\right) - \phi'\left(e^{-\mathbf{w}_i^T\theta}\right) + \\ & + \frac{1}{1+e^{\mathbf{v}_i^T\eta}}\phi'\left(\frac{e^{-\mathbf{w}_i^T\theta}}{1+e^{\mathbf{v}_i^T\eta}}\right)\Bigg]\end{aligned}$$

$$-E\left(\frac{\partial^2 l(\eta,\theta)}{\partial \eta_j \partial \theta_h}\right) = \sum_{i=1}^{n} \; \frac{w_{ih}v_{ij}e^{\mathbf{v}_i^T\eta-2\mathbf{w}_i^T\theta}}{\left[1+e^{\mathbf{v}_i^T\eta}\right]^2} \left[\phi'\left(\frac{e^{-\mathbf{w}_i^T\theta}}{1+e^{\mathbf{v}_i^T\eta}}\right) - e^{\mathbf{v}_i^T\eta}\phi'\left(\frac{e^{\mathbf{v}_i^T\eta-\mathbf{w}_i^T\theta}}{1+e^{\mathbf{v}_i^T\eta}}\right)\right] \tag{21}$$

with $j,q = 1,2,...,k$; $h,u = 1,2,...,m$; $i = 1,2,...,n$. From the Fisher's information matrix I note that the parameter vectors η and θ are not orthogonal, so their maximum likelihood estimators are dependent and can not be computed separately.
The maximum likelihood estimators of η and θ are obtained by making the score functions (20) equal to zero and do not have closed-form. Hence, they need to be obtained by numerically maximizing the log-likelihood function using a nonlinear optimization algorithm, such as a Newton algorithm or a quasi-Newton algorithm (McLachlan and Krishnan, 1997). The optimization algorithms require the specification of initial values to be used in iterative scheme.
My suggestion is to use as an initial point estimate for η the ordinary least squares estimate of this parameter vector obtained from a linear regression of the transformed response

$$\phi_i = \frac{\mu_i(1-\mu_i)}{var(Y_i)} - 1,$$

with $i =, 1,2,...,n$. By applying the delta method I derive the following approximation

$$var[logit(Y_i)] \approx var\left[logit(\mu_i) + (Y_i - \mu_i)\frac{\partial}{\partial \mu_i} logit(\mu_i)\right],$$

so I obtain that

$$var(Y_i) \approx var[logit(Y_i)]\mu_i^2(1-\mu_i)^2,$$

with $i =, 1,2,...,n$. Hence, I use the approximation

$$\hat{\phi}_i \approx \left|\frac{1}{\hat{e}_i{}^2\hat{\mu}_i(1-\hat{\mu}_i)} - 1\right|$$

with $\hat{\mu}_i = \frac{e^{\mathbf{v}_i'\hat{\eta}}}{1+e^{\mathbf{v}_i'\hat{\eta}}}$, where $\hat{\eta}$ and $\hat{e}_i$ are, respectively, the ordinary least squares estimate and residual from the linear regression of the transformed response. As initial point estimate for θ I use the ordinary least squares estimate obtained from a linear regression of the transformed value $-ln(\hat{\phi}_i)$ on $\mathbf{w}_i'$, with $i =, 1,2,...,n$.
I define this approach *joint beta regression model* since the distribution of the dependent variable is assumed to be a beta distribution, analogously to Ferrari and Cribari-Neto (2004), but, unlike the beta regression model, I model *jointly* the expectation and the dispersion of the dependent variable.

3.2 Mean and variance estimates of recovery rates

By considering the model (1) here proposed, the mean and the variance of the recovery rate R are respectively

$$\begin{aligned} E(R_i) &= E(I_i)P\{(R_i=0)\cup(R_i=1)\} + E(Y_i)P\{0<R_i<1\} \qquad (22)\\ var(R_i) &= var(I_i)P\{(R_i=0)\cup(R_i=1)\} + var(Y_i)P\{0<R_i<1\} + \end{aligned}$$

$$+[E(I_i)-E(R_i)]^2P\{(R_i=0)\cup(R_i=1)\}+[E(Y_i)-E(R_i)]^2P\{0<R_i<1\}$$

with $i=1,2,...,n$, where I denotes a Bernoulli random variable and Y a beta random variable. I underline that my model allows to estimate both the mean and the variance of the recovery rate.

By using the estimates of α and β that make the score functions (9) equal to zero, and by obtaining μ and ϕ from the joint beta regression model, the mean and the variance estimates of the recovery rates are given by

$$\begin{aligned}\hat{E}(R_i) &= \hat{a}_i\hat{b}_i+\hat{\mu}_i(1-\hat{b}_i) \\ v\hat{a}r(R_i) &= \hat{a}_i\hat{b}_i(1-\hat{a}_i)+(1-\hat{b}_i)\left[\frac{\hat{\mu}_i(1-\hat{\mu}_i)}{\hat{\phi}_i+1}+\hat{b}_i(\hat{\mu}_i-\hat{a}_i)^2\right]\end{aligned} \quad (23)$$

with $i=1,2,...,n$.

4 Data application

4.1 The Bank of Italy's survey

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. I consider 144,996 data for which all the covariate values are known. I 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 the 31/07/1975 to the 31/12/1998 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 given by the inclusion of transitory non-performing debts.

Since this survey considers loans privately held, it is difficult to assign them a market price, so necessarily the Bank of Italy applies the ultimate recovery approach (Friedman and Sandow, 2003) to compute the recovery rate. Hence, the exposure represents the outstanding debt at the time of default and the Bank of Italy considers the recovery as the actual recovery amount. I highlight that in this analysis I apply the expression proposed by Calabrese and Zenga (2008, 2010) to compute the recovery rate that constrains this variable within the interval [0,1].

This survey collects two main kinds of information: aggregate information about recovery procedures of banks and individual characteristics of loans whose recovery process is concluded by 1999. With regard to the former subject[1], private agreements are the most used recovery procedure and they show the shortest length of recovery procedure with a mean of 4.5 years. About this topic, Caselli et al. (2008) highlight that the Italian bankruptcy discipline, in force when the survey is conducted, do not bring about a quick liquidation of defaulted firms' assets. The length of recovery proceedings influences the costs of recovery procedures, which are then reflected in a higher LGD. Hence, different insolvency laws could cause national differences in the recovery process, so great importance is attached to our analysis of the Italian banking system.

However, the average recovery rate is 0.3846, the median value is 0.3333 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.

4.2 Estimation results

I apply the regression model proposed in this work to the Bank of Italy's database. I consider the recovery rate as a mixed random variable whose cumulative distribution function is given by (1). This application is interesting since it concerns loans, on which the availability of data is very difficult, in the Italian recovery process, which could be different from other countries.

The recovery rate is considered as a dependent variable in a regression model. In particular, I apply the model here proposed to estimate the recovery rate. In order to model jointly the mean and the variance of the recovery rate R, given by the equations (23), I need to estimate the parameters a, b, μ, ϕ. As above explained in the section 3, in order to estimate the parameter a I apply a logistic regression model, represented by the first equation in (7), just on the extreme values of the recovery rates. Moreover, in order to estimate the parameter b I apply a second regression model, represented by the second equation in (7), whose dependent variable is given by the dummy variable Z defined in (5). Finally, to estimate the parameters μ and ϕ I apply the joint beta regression model proposed in the subsection 3.1.

Since the aim of this regression model is to estimate the recovery rate at the time of default, all the covariates are known in that moment and not during or in the end of the recovery process. In a previous work (Calabrese and Zenga, 2008) the presence of collateral or personal guarantee and the exposure at default result significant in estimating the recovery rate. For that reason I consider a dummy variable that represents the presence of collateral or personal guarantee (CG) and the logarithm of the exposure at default (lnEAD) as explanatory variables. In order to investigate the influence of the geographic areas on the recovery rate, I introduce four dummy variables that represent five Italian macro areas (SI=South Italy, CI=Central Italy,

[1] For more details see Banca d'Italia (2001) and Grippa et al. (2005).

NEI=North East Italy, NWI=North West Italy) in the regression model.

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 the recovery rate, I introduce a dummy variable 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 analogously to many empirical studies (Acharya et al., 2007; Altman et al., 2005; Bellotti and Crook, 2009; Bruche and González-Aguado, 2008; Caselli et al., 2008; Figlewski et al., 2007), the chosen macroeconomic variables are the interest on delayed payment (IR), the unemployment rate (UR), the growth of GDP (GDP) and the default rate (DR), 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.

Since the macroeconomic variables are available from 1985, in this analysis I consider 144,966 loans that defaulted between January 1985 and December 1999 and whose recovery process is written off within December 1999. I underline that the size of the sample here considered is significantly much 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. I specify that the sample size for Grippa et al. (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.

I point out that the model here proposed allows to analyze the different influences of the same covariates on the discrete and the continuous parts of the recovery rate. Some authors (e.g. Bellotti and Crook, 2009; 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 $\mathbf{x}$, $\mathbf{v}$ and $\mathbf{w}$, considered in the equations (7) and (12), coincide.

Since bank aims at forecasting the recovery rate, in order to avoid the overfitting, data are divided in two groups that refer to different periods of time. The model is fitted on the data concerning a given period and the predictive accuracy is evaluated in the aftermath. Hence, the parameters of the regression model here proposed are estimated on 134,937 defaults occurred from 1985 to 1998. In subsection 4.4 the predictive accuracy is measured on the out-of-time sample of 10,059 loans that defaulted in 1999. Moreover, in order to analyse the accuracy for different forecasting periods of time, in the same subsection I consider also two out-of-time samples given by the defaults occurred from 1998 to 1999 and from 1997 to 1999. The model is fitted on loans defaulted respectively from 1985 to 1997 and from 1985 to 1996.

Some authors in the literature, e.g. Bellotti and Crook (2009) 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 time of 14 years, including Eighties expansion and the early Nineties recession in Italy.

High correlation among macroeconomic variables is a drawback since it leads to multicollinearity in the regression model and therefore the parameter estimators could be biased. In order to measure the severity of multicollinearity I 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 lower than 5, the level of multicollinearity is tolerable.

The following table[2] reports the parameter estimates and the p-values in round brackets obtained by the application of the methodological proposal of this work to the Bank of Italy's data. The p-values lower than 0.0001 are omitted in Table I. Choosing a level of significance of 0.1%, all the variables are significant except for the unemployment and the GDP growth rates in the logistic regression model for α estimate, the two dummy variables for the Centre and the South of Italy in the logistic regression for β estimate model and the logarithm of the EAD for the expectation of the joint beta regression model.

	Logistic Regression	Logistic Regression	Joint Beta Regression	
	α	β	η	θ
Constant	-1.5308(0.0004)	-1.5126	-0.9955	-3.1461
IR	-0.0713	-0.1111	-0.0466	-0.0251
UR	0.1087(0.0061)	0.3727	0.1229	0.2066
GDP	-0.0331(0.1137)	-0.1693	-0.0138	-0.0661
DR	0.0975	-0.1052	-0.0206	-0.0364
lnEAD	-0.0897	-0.6459	-0.0023(0.0045)	0.0916
CG	1.1140	-0.2333	0.5084	-0.1132
CF	0.4357	-0.5883	-0.0682	-0.0564
CI	-0.6407	-0.0804(0.00722)	0.0982	-0.0052
SI	-0.1912	0.0877(0.0060)	0.0248(0.0005)	0.0674
NEI	-0.7677	0.4448	-0.0753	0.1688
NWI	-0.7440	0.6032	-0.1272	0.1221

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

The results on the influence of the EAD on the recovery rates are very 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, instead Dermine and Neto de Carvalho (2006), Grippa et al. (2005) hit upon that the recovery rates decrease when the loan size increases. From this work the logarithm of the EAD has an inverse relationship with the mean recovery rate for both the discrete and the continuous parts, as Table 1 shows. Since

[2] I obtain these results by using the package "LogicReg" for the logistic regression model and as optimization procedure "optim" with the method "Nelder-Mead" of R-program.

the condition (18) is satisfied, the EAD has an inverse relationship with the variance of the continuous part of the recovery rate, coherently with Calabrese and Zenga (2008).

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 estimate coincide with the expectations. For the discrete part, the dummy variable for the consumer family (CF) shows that the mean recovery rate given that the borrower belongs to a consumer family is higher than the one for nonconsumers. This ordering is reverse for the continuous part of the recovery rates. This interesting result shows that for a bank the full recovery is easier if the borrower belongs to a consumer family. The cause of this characteristic could be the larger resort to collateral and personal guarantee for consumer families. Analogously to Grippa et al. (2005), the geographic areas are relevant determinants for the recovery rates.

I highlight that the default rate has a different influence on the means of the discrete and the continuous parts. This difference could be due to different discrimatory powers achieved by bank for the two groups: the extreme values of recovery rates could be characterized by higher discrimatory power than the one achieved for the continuous part.

Moreover, Table 1 shows that the macroeconomic variables have the same influence on the means of the discrete and the continuous parts. In particular, the sign of the estimate for the interest rate on delayed payment coincides with the expectations. Instead, this agreement fails for the unemployment and the GDP growth rates. The cause of these results could be that the interest on delayed payment has a short-term influence on the recovery rates, instead this influence could be of long-term for the unemployment and the GDP growth rates. A similar consideration is involved in Bellotti and Crook (2009)'s work.

4.3 The comparison with the fractional response model

As above-mentioned, in order to guarantee that the predicted recovery rates lie in the unit interval (0,1), in GLMs the link function $g(\cdot)$ maps the interval (0,1) on the real axis. A wide choice of link functions $g(\cdot)$ is available. Three functions commonly used are

- the logit function $$E\left\{log\left(\frac{R}{1-R}\right)|\mathbf{x}\right\} = \mathbf{x}'\lambda \qquad (24)$$
- the log-log function $$E\left[-log\left(-log(R)\right)|\mathbf{x}\right] = \mathbf{x}'\lambda \qquad (25)$$
- the complementary log-log function $$E\left[-log\left(-log(1-R)\right)|\mathbf{x}\right] = \mathbf{x}'\lambda\,. \qquad (26)$$

The main drawback of these link functions is that the equations (24), (25) and (26) cannot be true if R takes on the values 0 or 1 with positive probability. Since the extreme values 0 and 1 have a pivotal role in recovery risk analysis, some authors, i.e.

Bastos (2009), Chalupka and Kopecsni (2009), Grippa et al. (2005), Dermine and Neto de Carvalho (2006), apply the fractional response model proposed by Papke and Wooldridge (1996) in order to overcome this problem. The estimation procedure of the fractional response model is a quasi-likelihood method that consists of the maximization of the Bernoulli log-likelihood function

$$l_i(\hat{\lambda}) = r_i log[G(\mathbf{x}_i'\hat{\lambda})] + (1-r_i)log[1-G(\mathbf{x}_i'\hat{\lambda})], \tag{27}$$

for $i = 1,2,...,n$, where $0 < G(\cdot) < 1$ is the inverse of the link function $g(\cdot)$. Because equation (27) is a member of the linear exponential family, the quasi-maximum likelihood estimator of λ, obtained from the maximization problem

$$\max_{\lambda} \sum_{i=1}^{n} l_i(\lambda),$$

is consistent for λ and $\sqrt{n}$-asymptotically normal, regardless of the distribution of R_i conditional on $\mathbf{x}_i$, provided that

$$E(R_i|\mathbf{x}_i) = G(\mathbf{x}_i'\lambda),$$

with $i = 1,2,...,n$. In particular, Grippa et al. (2005) apply the fractional response model with a logit link function, so its inverse results

$$G(\mathbf{x}_i'\lambda) = \frac{\exp(\mathbf{x}_i'\lambda)}{1+\exp(\mathbf{x}_i'\lambda)}.$$

On the contrary, Dermine and Neto de Carvalho (2006) choose the log-log link function, so they obtain

$$G(\mathbf{x}_i'\lambda) = \exp(-\exp(-\mathbf{x}_i'\lambda)).$$

In Bastos (2009)'s analysis the logit and the log-log link functions do not exhibit substantial differences in forecasting performance.

Finally, Chalupka and Kopecsni (2009) consider also the complementary log-log link function, whose inverse is

$$G(\mathbf{x}_i^T\lambda) = 1-\exp(-\exp(-\mathbf{x}_i^T\lambda)).$$

By comparing the three above-mentioned link functions, Chalupka and Kopecsni (2009) show that the log-log model performs better.

The main difference among these three link functions is that the logit link function is symmetric, the log-log is right-skewed and the complementary log-log is left-skewed. This means that the best link function in terms of forecasting performance depends on the distribution skewness of recovery rates.

4.4 Out-of-time predictive accuracy

In this subsection I compare the predictive accuracy of the regression model here proposed with the one of the fractional response model. In the latter model I consider the logit, the log-log and the complementary log-log link functions, analysed in the previous subsection. The predictive accuracy of the models is assessed using two performance measures. The Mean Square Error (MSE) is defined as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(r_i - \hat{r}_i)^2$$

where r_i and $\hat{r}_i$ are the actual and the predicted recovery rates on loan i, respectively. The Mean Absolute Error (MAE) is defined as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|r_i - \hat{r}_i|.$$

Models with lower MSE and MAE forecast actual recoveries 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 the 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.
At this point, I compare the out-of-time predictive accuracy of my proposal with the fractional response model on the Bank of Italy's data for different forecasting period of time. 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 sample 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.
Table 2 reports the MAE and the MSE for each model and for each forecasting period of time. Since for a bank the overestimation of the recovery rate is more risky than the underestimation, I compute the parts of the MSE and the MAE due to the negative errors $r_i - \hat{r}_i < 0$. I indicate them by MSE^- and MAE^-, respectively, and also their values are reported in Table 2. Therefore, the percentages of the errors MSE^- and MAE^- are worked out on the respective errors MSE and MAE and their values are reported in Table 2 between round brackets.
By the results reported in Table 2 my proposal exhibits both the MAE and the MSE lower than the respective errors of all the three fractional response models for each forecasting period of time. By the results reported in Table 2 my proposal exhibits both the MAE and the MSE lower than the respective errors of all the three fractional response models for each forecasting period of time.

Forecasting period of time	*Error*	*Models*			
		Continuous-discrete	*Fractional*		
			log-log	*logistic*	*complementary log-log*
within one year	MAE	0.3328	0.4077	0.4040	0.4005
	MSE	0.1370	0.1802	0.1768	0.1745
	MAE^-	0.2131 (64.03%)	0.3162 (77.56%)	0.3100 (76.73%)	0.3043 (75.98%)
	MSE^-	0.0641 (46.79%)	0.1349 (74.86%)	0.1295 (73.25%)	0.1253 (71.81%)
within two years	MAE	0.3478	0.3675	0.3664	0.3682
	MSE	0.1500	0.1570	0.1564	0.1583
	MAE^-	0.1681 (48.33%)	0.2133 (58.04%)	0.2117 (57.78%)	0.2220 (60.29%)
	MSE^-	0.0544 (36.27%)	0.0843 (53.69%)	0.0831 (53.13%)	0.0911 (57.55%)
within three years	MAE	0.3349	0.3477	0.3424	0.3460
	MSE	0.1435	0.1455	0.1431	0.1447
	MAE^-	0.1448 (43.24%)	0.1853 (53.29%)	0.1711 (49.97%)	0.1776 (51.33%)
	MSE^-	0.0441 (30.73%)	0.0681 (46.80%)	0.0599 (41.86%)	0.0641 (44.30%)

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

Furthermore, I can observe that the fractional response models with different link functions (logit, log-log and complementary log-log) do not exhibit substantial differences in forecasting performance. A similar result is obtained by Bastos (2010) only for fractional logit and log-log models. On the contrary, Chalupka and Kopecsni (2009) show that fractional log-log model performs better than fractional logit and complementary log-log models.

By focusing my attention just on the negative errors $r_i - \hat{r}_i < 0$, Table 2 shows that both the MAE^- and the MSE^- of my proposal are significantly lower than the respective errors of all the three fractional response models for each forecasting period of time. This result is mainly due to the overestimation of the null recovery rates by the fractional response models. Analogously to Araten et al. (2004), Asarnow and Edwards (1995), Caselli et al. (2008), Friedman and Sandow (2003), the percentage of null recovery rates is relevant (22.88%).

It is interesting to note that for a given model, as the forecasting period of time increases, the importance of the underestimation errors decreases. Since data concern defaults by the end of 1998 and written off by the end of 1999, the percentage of null recoveries in the out-of-time sample decreases as the forecasting period of time increases. Consequently, the underestimation errors decreases. From this characteristic I can deduce that my proposal is preferable for different sample percentages of the extreme values of the recovery rates.

5 Conclusions remarks

In this work I aim at proposing a regression model for the recovery rate. At first, to represent the high concentration of data at total recovery and total loss I assume that the recovery rate is a mixed random variable, given by the mixture of a Bernoulli and a beta random variables. To estimate the Bernoulli parameter and the mixture weights I propose to apply two logistic regression models. For the continuous part of the recovery rate I propose the joint beta regression that accommodates skewness and heteroscedastic errors.

The main advantage of my proposal is that it allows to analyse the different influences of the same covariates on the extreme values and the recovery rates belonging to the interval (0,1). Another positive aspect is that my proposal allows to estimate both the mean and the variance of the recovery rate, knowing the covariates.

Afterwards, I apply the regression model here proposed to the Bank of Italy's data. Compliant to Basel II (BCBS, 2005) I introduce some macroeconomic variables that are significant in predicting recovery rates. An interesting result is the different influence of the default rate on the means of the discrete and the continuous parts. Since the extreme values of the recovery rates have a pivotal role in the recovery risk analysis, the fractional response model, proposed by Papke and Wooldridge (1996), is widely used in the literature.

The comparison of the out-of-time predictive accuracy of my proposal and the one of the fractional response model shows that the first is preferable for different forecasting periods of time and for different sample percentages of the extreme values of the recovery rates.

References

1. Acharya Viral V., Bharath Sreedhar T., Srinivasan A.: Does industry-wide distress affect defaulted firms? Evodence from creditor recoveries. Journal of Financial Economics. **85**(3), 787–821 (2007)
2. 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 (2005)
3. 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)
4. Asarnow, E., Edwards, D.: Measuring loss on default bank loans: A 24-year study. Journal of Commercial Lending. **77**, 11–23 (1995)
5. Banca d'Italia: Principali risultati della rilevazione sull'attivitá di recupero dei crediti. Bollettino di Vigilanza. 12, December (2001)
6. Banca d'Italia: Bolletino Statistico, March (2000a)
7. Banca d'Italia: Sintesi delle note sull'andamento dell'economia delle regioni italiane nel 1999, (2000b)
8. Barndorff-Nielsen, O.: Information and exponential families in statistical theory. Wiley, New York (1978)
9. Basel Committee on Banking Supervision: International convergence of capital measurement and capital standards: A revised framework. Bank for International Settlements. Basel, June (2004a)
10. Basel Committee on Banking Supervision: Background note on LGD quantification. Bank for International Settlements. Basel, December (2004b)
11. Basel Committee on Banking Supervision Guidance on paragraph 468 of the framework document. Bank for International Settlements. Basel, July (2005)
12. Bastos J. A.: Forecasting bank loans loss-given-default. Journal of Banking and Finance **34**(10), 2510–2517 (2010)

13. Bellotti T. and Crook J.: Loss Given Default models for UK retail credit cards. Credit Research Centre, working paper (2009)
14. Bruche, M., González-Aguado C.: Recovery Rates, Default Probabilities and the Credit Cycle. CEMFI, working paper (2008)
15. Calabrese, R., Zenga, M.: Measuring loan recovery rate: Methodology and empirical evidence. Statistica & Applicazioni. **6**, 193–214 (2008)
16. Calabrese, R., Zenga, M.: Bank loan recovery rates: Measuring and nonparametric density estimation. Journal of Banking and Finance **34**(5), 903–911 (2010)
17. Carty, L., Lieberman, D.: Defaulted bank loan recoveries. Moody's special comment. November (1996)
18. 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)
19. 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)
20. Dermine, J., Neto de Carvalho, C.: Bank loan losses-given-default: A case study. Journal of Banking and Finance. **30**, 1219–1243 (2006)
21. Emery, K., Cantor, R., Arner, R.: Recovery Rates on North American Syndicated Bank Loans, 1989-2003. Moody's special comment. March (2004)
22. Ferrari, S., Cribari-Neto, F.: Beta regression for modeling rates and proportions. Journal of Applied Statistics. **31**, 799–815 (2004)
23. 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)
24. Friedman, C., Sandow, S.: Ultimate recoveries. Risk. **16**, 69–73 (2003)
25. Frye J.: Depressing Recoveries. Risk **13** 11, 108–111 (2000)
26. Greene W. H.: Econometric Analysis. Prentice Hall, New York (2000)
27. 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)
28. Grunert, J., Weber, M.: Recovery rate of commercial lending: Empirical evidence for German companies. Journal of Banking and Finance. **33**, 505–513 (2009)
29. Gupton, G. M., Finger, C. C., Bhatia, M.: CreditMetrics. Technical document, J. P. Morgan (1997)
30. Gupton, G. M., Stein, R. M.: LosscalcTM: Model for predicting Loss Given Default (LGD), Moody's Investors Service (2002)
31. Hagmann, M., Renault O., Scaillet, O.: Estimation of Recovery Rate Densities: Nonparametric 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)
32. Hosmer, D. W., Lemeshow, S. Applied logistic regression. Wiley, New York (2000)
33. McCullagh, P., Nelder, J.A.: Generalized Linear Models. Chapman & Hall/CRC, London (1989)
34. McLachlan, G. J., Krishnan, T.: The EM Algorithm and Extentions. Wiley, New York (1997)
35. 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)
36. 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)
37. Schuermann, T.: What Do We Know About Loss Given Default? Recovery Risk. Working Paper, Federal Reserve Bank of New York (2003)

6 Appendix

In this appendix I obtain the score functions and the Fisher information matrix for η, θ. The notation used here is defined in the subsection 3.1. At first, in order to compute the score functions I consider the following equations

$$\frac{\partial l_i(\eta,\theta)}{\partial \eta_j} = \frac{\partial l_i(\mu_i,\phi_i)}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_j} \qquad \frac{\partial l_i(\eta,\theta)}{\partial \theta_h} = \frac{\partial l(\mu_i,\phi_i)}{\partial \phi_i}\frac{\partial \phi_i}{\partial \theta_h} \tag{28}$$

with $j = 1,2,...,k$; $h = 1,2,...,m$. From equations (12) and (19) I obtain that

$$\begin{aligned}
\frac{\partial l_i(\mu_i,\phi_i)}{\partial \mu_i} &= -\phi_i\psi(\mu_i\phi_i) + \phi_i\psi(\phi_i - \mu_i\phi_i) + \phi_i log\frac{y_i}{1-y_i}\\
\frac{\partial l_i(\mu_i,\phi_i)}{\partial \phi_i} &= \psi(\phi_i) - \mu_i\psi(\mu_i\phi_i) - (1-\mu_i)\psi(\phi_i - \mu_i\phi_i) + \mu_i log(y_i) + (1-\mu_i)log(1-y_i)\\
\frac{\partial \mu_i}{\partial \eta_j} &= \frac{x_j e^{\mathbf{X}_i'\eta}}{\left[1+e^{\mathbf{X}_i'\eta}\right]^2} \qquad\qquad \frac{\partial \phi_i}{\partial \theta_h} = -w_h e^{-\mathbf{W}_i'\theta}
\end{aligned}$$

with $j = 1,2,...,k$; $h = 1,2,...,m$; $i = 1,2,...,n$. Substituting the former results and the expressions (12) in equations (28) the score functions (20) are obtained.
The second order partial derivatives of the log-likelihood function with respect to parameters (μ_i,ϕ_i) are

$$\begin{aligned}
\frac{\partial^2 l_i(\mu_i,\phi_i)}{\partial^2 \mu_i} &= -\phi_i^2[\psi'(\mu_i\phi_i) - \psi'(\phi_i - \mu_i\phi_i)]\\
\frac{\partial^2 l_i(\mu_i,\phi_i)}{\partial^2 \phi_i} &= \phi'(\psi_i) - \mu_i^2\phi'(\mu_i\psi_i) - (1-\mu_i)\phi'(\psi_i - \mu_i\psi_i)
\end{aligned} \tag{29}$$

and the second order partial derivatives of the parameters μ_i and ϕ_i with respect to the regression parameters η and θ are

$$\begin{aligned}
\frac{\partial^2 \mu_i}{\partial \eta_j \partial \eta_q} &= \frac{x_j x_h e^{\mathbf{X}_i'\eta\left[1-e^{\mathbf{X}_i'\eta}\right]}}{\left[1-e^{\mathbf{X}_i'\eta}\right]^3}\\
\frac{\partial^2 \phi_i}{\partial \theta_h \partial \theta_u} &= w_h w_u e^{-\mathbf{W}_i'\theta}
\end{aligned} \tag{30}$$

with $j,q = 1,2,...,k$; $h,u = 1,2,...,m$; $i = 1,2,...,n$. The Fisher information is the negative of the expectation of the second derivatives of the log-likelihood with respect to the regression parameters η and θ

$$-E\left(\frac{\partial^2 l_i(\eta,\theta)}{\partial \eta_j \partial \eta_q}\right) = \frac{\partial^2 l_i(\eta,\theta)}{\partial^2 \mu_i}\frac{\partial^2 \mu_i}{\partial \eta_j \partial \eta_q}$$

$$
\begin{aligned}
-E\left(\frac{\partial^2 l_i(\eta,\theta)}{\partial\theta_h\partial\theta_u}\right) &= \frac{\partial^2 l_i(\eta,\theta)}{\partial^2\phi_i}\frac{\partial^2\phi_i}{\partial\theta_h\partial\theta_u} \\
-E\left(\frac{\partial^2 l_i(\eta,\theta)}{\partial\eta_j\partial\theta_h}\right) &= \frac{\partial\left[\frac{\partial l_i(\eta,\theta)}{\partial\eta_j}\right]}{\partial\theta_h}
\end{aligned}
\qquad (31)
$$

with $j,q = 1,2,...,k$; $h,u = 1,2,...,m$; $i = 1,2,...,n$. In the first two equations of (31) I substitute the results (29) and (30), so I obtain the first two equations of (21). From the last equation of (31), I compute the derivative of the first result in (20) with respect to θ_h, so I obtain the last equation of (21).