

Regression for Recovery Rates with both Continuous and Discrete Characteristics

Raffaella Calabrese

Abstract I propose to consider the recovery rate as a mixed random variable, obtained as the mixture of a Bernoulli random variable and a beta random variable with support $(0,1)$. The recovery rate is considered as a dependent variable in a regression model. I suggest to apply the logistic regression model for the extreme values of the recovery rates. For the recovery rates belonging to the interval $(0,1)$, I propose to model jointly the conditional expectation and the conditional dispersion by using two link functions. This model accommodates skewness and heteroscedastic errors. The estimation procedure is the maximum likelihood method. Finally, the proposed estimation methodology is applied to the survey on loan recovery process of Italian banks, conducted in 2000 by the Bank of Italy.

Key words: recovery rate, mixed random variable, joint beta regression model, statistical analysis of financial data

1 Introduction

Although estimates of recovery rates are key variables for banking practice, such pivotal topic is relatively unexplored by the literature. Within this research field, the main aim of this work is to propose a model in order to model jointly the conditional mean and the conditional variance of the recovery rate, given some explanatory variables. At first, I propose to consider the recovery rate as a mixed random variable to represent the high concentration of data at total recovery and total loss. In particular, I assume that the recovery rate is a mixture of a Bernoulli and a beta random variables.

In order to analyze the influences of some explanatory variables on the con-

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

tinuous part, I propose the joint beta regression methodology that leads to model jointly the conditional expectation and the conditional dispersion by using two link functions. The parameters of the model are estimated by the maximum likelihood method. This model accommodates skewness, multimodality and heteroscedastic errors. For the discrete part of the recovery rate I apply a logistic regression. By this proposal I can examine the different influences of the same covariates on the continuous and the discrete parts. Finally, 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. In order to constrain this variable within the interval $[0,1]$, the recovery rates are computed by the methodology proposed by Calabrese and Zenga (2008).

The present paper is organized as follows. Section 2 explains the proposal of considering the recovery rate as a mixed random variable given by the mixture of a beta and a Bernoulli random variables. Under this assumption, I propose a regression model for the recovery rate in the following section. Successively, section 4 presents the Bank of Italy's database, to which I apply the suggested methodology and I present the main empirical results.

2 The Recovery Rate as a Mixed Random Variable

Analogously to many models in the literature (i.e. Gupton et al., 1997), in this work I consider the recovery rate as a random variable. Moreover, some approaches (Gupton et al., 1997) 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), Caselli et al. (2008), Dermine and Neto de Carvalho (2006), Grunert and Weber(2009), 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, I propose to consider the recovery rate R as a mixed random variable, given by the mixture of a Bernoulli random variable B and a beta random variable Y

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

where F_Y denotes the distribution function of the beta random variable Y and $P\{R = j\}$ is the probability that the recovery rate R is equal to j with $j = 0, 1$. This parametric model provides a good fit to data (Calabrese and Zenga, 2009).

An important issue is to propose an estimation methodology for regression model whose dependent variable is a mixed random variable.

3 A Regression Model for the Recovery Rate

In order to understand the determinants of the conditional 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 $g(\mu) = \mathbf{x}^T \alpha$. 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 $(0,1)$ on to the whole real line.

When the conditional distribution of Y is a beta distribution, the range of the parameter μ is the interval $(0,1)$. By applying a methodology similar to GLM models, Ferrari and Cribari-Neto (2004) propose the beta regression model, where the conditional distribution of the dependent variable is a beta distribution. The probability density function of a beta distribution with parameters $p > 0$ and $q > 0$ is

$$f(y; p, q) = \frac{y^{p-1}(1-y)^{q-1}}{B(p, q)} = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}$$

where $y \in (0, 1)$, $B(\cdot, \cdot)$ denotes the beta function and $\Gamma(\cdot)$ the Gamma function. They propose a reparameterization that translates p and q into a location parameter μ and a dispersion parameter ϕ

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

where $\phi = p+q$. Similarly to GLM models, Ferrari and Cribari-Neto (2004)'s approach models only the conditional mean μ and they consider the dispersion parameter $\phi = p+q$ as a nuisance parameter.

On the contrary, I model jointly the conditional mean μ and the dispersion parameter ϕ , defined as

$$\mu = \frac{p}{p+q} \quad \phi = p+q. \quad (3)$$

for the beta random variable Y by applying the same reparameterization proposed by Ferrari and Cribari-Neto (2004).

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

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

where $y_i \in (0, 1)$, $1 > \mu_i > 0$ and $\phi_i > 0$, with mean and variance given by the equations (2). To joint model the conditional mean μ_i and the parameter ϕ_i with $i = 1, 2, \dots, n$, I propose to use two different link function $g(\cdot)$ and $h(\cdot)$ such that

$$g(\mu) = \mathbf{x}_i^T \alpha \quad h(\phi) = -\mathbf{w}_i^T \beta, \quad (4)$$

with $i = 1, 2, \dots, n$, where α and β are vectors of respectively k and m unknown regression parameters, $\mathbf{x}_i$ and $\mathbf{w}_i$ are two vectors of observations on respectively k and m covariates ($k + m < n$), which are assumed fixed and known. In order to define the link function $h(\cdot)$ I use the negative sign in the second equation (4) since I prefer to model the dispersion of the dependent variable Y .

Furthermore, I point out that the conditional mean μ and the parameter ϕ depend on two different vectors of covariates, respectively $\mathbf{x}$ 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 conditional mean μ_i . Since $0 < \mu_i < 1$ and $\phi_i > 0$ with $i = 1, 2, \dots, n$, I suppose that the link function $g(\cdot)$ is the logit function and the link function $h(\cdot)$ is the log function, it follows that

$$\mu_i = \frac{1}{1 + e^{-\mathbf{x}_i^T \alpha}} \quad \phi_i = e^{-\mathbf{w}_i^T \beta}, \quad (5)$$

with $i = 1, 2, \dots, n$. I underline that the variance of Y_i is a function of μ_i and ϕ_i , as given by the equation (2) and, as a consequence, of the covariate values $\mathbf{x}_i$ and $\mathbf{w}_i$, so such model accommodates the heteroscedastic errors. Moreover, since the beta distribution is flexible, the skewness and multimodality are also accomodated.

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

$$\begin{aligned} l(\alpha, \beta) = \sum_{i=1}^n & \left[\ln \Gamma(e^{-\mathbf{w}_i^T \beta}) - \ln \Gamma\left(\frac{e^{\mathbf{x}_i^T \alpha - \mathbf{w}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \alpha}}\right) - \ln \Gamma\left(\frac{e^{-\mathbf{w}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \alpha}}\right) + \right. \\ & \left. + \left(\frac{e^{\mathbf{x}_i^T \alpha - \mathbf{w}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \alpha}} - 1\right) \ln(y_i) + \left(\frac{e^{-\mathbf{w}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \alpha}} - 1\right) \ln(1 - y_i) \right]. \end{aligned}$$

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 $\psi(z) = \frac{\partial \Gamma(z)}{\Gamma(z)}$. The score function is obtained by differentiating the log-likelihood function with respect to the unknown parameters α and β

$$\frac{\partial l(\alpha, \beta)}{\partial \alpha_j} = \sum_{i=1}^n x_{ij} \frac{e^{\mathbf{x}_i^T \alpha - \mathbf{w}_i^T \beta}}{[1 + e^{\mathbf{x}_i^T \alpha}]^2} \left[\phi\left(\frac{e^{-\mathbf{w}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \alpha}}\right) - \phi\left(\frac{e^{\mathbf{x}_i^T \alpha - \mathbf{w}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \alpha}}\right) + \log \frac{y_i}{1 - y_i} \right]$$

$$\begin{aligned} \frac{\partial l(\alpha, \beta)}{\partial \beta_h} = \sum_{i=1}^n -w_{ih} \frac{e^{-\mathbf{w}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \alpha}} \left[\left(1 + e^{\mathbf{x}_i^T \alpha} \right) \phi \left(e^{-\mathbf{w}_i^T \beta} \right) - e^{\mathbf{x}_i^T \alpha} \phi \left(\frac{e^{\mathbf{x}_i^T \alpha - \mathbf{w}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \alpha}} \right) + \right. \\ \left. - \phi \left(\frac{e^{-\mathbf{w}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \alpha}} \right) + \log(1 - y_i) + e^{\mathbf{x}_i^T \alpha} \log(y_i) \right], \end{aligned} \quad (6)$$

with $j = 1, 2, \dots, k$; $h = 1, 2, \dots, m$; $i = 1, 2, \dots, n$, where y_i is a realization of the recovery rate.

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

$$\begin{aligned} -E \left(\frac{\partial^2 l(\alpha, \beta)}{\partial \alpha_j \partial \alpha_q} \right) &= \sum_{i=1}^n \frac{x_{ij} x_{iq} e^{\mathbf{x}_i^T \alpha - 2\mathbf{w}_i^T \beta} [1 - e^{\mathbf{x}_i^T \alpha}]}{[1 + e^{\mathbf{x}_i^T \alpha}]^3} \left[\phi' \left(\frac{e^{\mathbf{x}_i^T \alpha - \mathbf{w}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \alpha}} \right) + \phi' \left(\frac{e^{-\mathbf{w}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \alpha}} \right) \right] \\ -E \left(\frac{\partial^2 l(\alpha, \beta)}{\partial \beta_h \partial \beta_u} \right) &= \sum_{i=1}^n w_{ih} w_{ui} e^{-\mathbf{w}_i^T \beta} \left[\left(\frac{e^{\mathbf{x}_i^T \alpha}}{1 + e^{\mathbf{x}_i^T \alpha}} \right)^2 \phi' \left(\frac{e^{\mathbf{x}_i^T \alpha - \mathbf{w}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \alpha}} \right) - \phi' \left(e^{-\mathbf{w}_i^T \beta} \right) + \right. \\ &\quad \left. + \frac{1}{1 + e^{\mathbf{x}_i^T \alpha}} \phi' \left(\frac{e^{-\mathbf{w}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \alpha}} \right) \right] \\ -E \left(\frac{\partial^2 l(\alpha, \beta)}{\partial \alpha_j \partial \beta_h} \right) &= \sum_{i=1}^n \frac{w_{ih} x_{ij} e^{\mathbf{x}_i^T \alpha - 2\mathbf{w}_i^T \beta}}{[1 + e^{\mathbf{x}_i^T \alpha}]^2} \left[\phi' \left(\frac{e^{-\mathbf{w}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \alpha}} \right) - e^{\mathbf{x}_i^T \alpha} \phi' \left(\frac{e^{\mathbf{x}_i^T \alpha - \mathbf{w}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \alpha}} \right) \right] \end{aligned}$$

with $j, q = 1, 2, \dots, k$; $h, u = 1, 2, \dots, m$; $i = 1, 2, \dots, n$. From the Fisher's information matrix I note that the parameter vectors α and β are not orthogonal, so their maximum likelihood estimators are dependent and can not be computed separately.

The maximum likelihood estimators of α and β are obtained by making the score function (6) 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 α the ordinary least squares estimate of this parameter vector obtained from a linear regression of the transformed response

$$\psi_i = \frac{\mu_i(1 - \mu_i)}{\text{var}(Y_i)} - 1.$$

By applying the delta method I derive the following approximation

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

so I obtain that

$$\text{var}(Y_i) \approx \text{var}[\text{logit}(Y_i)]\mu_i^2(1-\mu_i)^2.$$

Hence, I use the approximation

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

with $\hat{\mu}_i = \frac{e^{\mathbf{x}_i^T \hat{\alpha}}}{1 + e^{\mathbf{x}_i^T \hat{\alpha}}}$, where $\hat{\alpha}$ and $\hat{e}$ are, respectively, the ordinary least squares estimate and residual from the linear regression of the transformed response. As initial point estimate for β I use the ordinary least squares estimate obtained from a linear regression of the transformed value $-\ln(\hat{\psi}_i)$ on $\mathbf{w}_i^T$.

I define this approach *joint beta regression model* since the conditional distribution of the dependent variable is assumed to be a beta distribution, analogously to the beta regression model proposed by Ferrari and Cribari-Neto (2004), but unlike I model *jointly* the conditional expectation and the conditional dispersion.

Under the assumption (1) the conditional expectation and the conditional variance of the recovery rate R are

$$\begin{aligned} E(R/\mathbf{x}, \mathbf{w}) &= E(B/\mathbf{x}, \mathbf{w})P\{(R=0) \cup (R=1)\} + E(Y/\mathbf{x})P\{0 < R < 1\} \\ \text{var}(R/\mathbf{x}, \mathbf{w}) &= \text{var}(B/\mathbf{x}, \mathbf{w})P\{(R=0) \cup (R=1)\} + \text{var}(Y/\mathbf{x}, \mathbf{w})P\{0 < R < 1\} + \\ &\quad + [E(B/\mathbf{x}, \mathbf{w}) - E(R/\mathbf{x}, \mathbf{w})]^2 P\{(R=0) \cup (R=1)\} + \\ &\quad + [E(Y/\mathbf{x}) - E(R/\mathbf{x}, \mathbf{w})]^2 P\{0 < R < 1\} \end{aligned}$$

I propose to apply the logistic regression model (Hosmer and Lemeshow, 2000) for the discrete part and the joint beta regression model for the continuous part of the recovery rate. Finally, I estimate the mixture weights by the corresponding relative frequencies.

4 Empirical Evidence of recovery rates

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. The database comprises 149,378 defaulted borrowers. 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 on the 31/12/1998 and written off by the end of 1999.

I apply the regression model proposed in this work to the Bank of Italy's

database. Considering the recovery rate as a mixed random variable, I estimate the probabilities in (1) by the respective relative frequencies. In addition, I model separately the discrete and the continuous parts of the recovery rate R .

On the one hand, I apply the logistic regression model for the $n_2 = 45,867$ extreme values of the recovery rates. On the other hand, for $n_1 = 103,511$ data exhibit recovery rates belonging to the interval $(0,1)$ I apply the regression model proposed in the third section in order to analyze the determinants of the recovery rates. Such methodology leads to analyze the different influences of the covariates on the discrete and the continuous parts of the recovery rate.

Some authors (i.e. Friedman and Sandow, 2003; 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 reach this goal I choose the same set of covariates for $\mathbf{x}$ and $\mathbf{w}$ given by six variables relevant to the recovery risk: the capitalized recovery amount, the logarithm of the capitalized EAD, the time in default, the interest on delayed payment, the legal costs and the amount of collateral or personal guarantee.

The following table¹ 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.

	Logistic Regression	Joint Beta Regression	
		α	β
Constant	-14.099(0.000)	-0.082(0.000)	-1.127(0.000)
Capitalized recovery amount	26.005(0.000)	0.011(0.000)	0.058(0.000)
Log capitalized EAD	-0.022(0.009)	0.001(0.016)	-0.036(0.068)
Time in default	0.003(0.009)	-0.460(0.089)	-0.277(0.046)
Interest on delayed payment	-1.602(0.191)	-0.169(0.230)	-0.006 (0.187)
Legal costs	-1.605(0.182)	0.009(0.128)	-0.006(0.243)
Collateral or personal guarantee	-0.057(0.010)	1.035(0.035)	-3.897(0.123)

Choosing a level of significance of 0.05, for both the discrete and the continuous parts of the recovery rate R interest on delayed payment and legal costs are not significant. The capitalized recovery amount has a strong influence on the extreme values of the recovery rates. For the continuous part an expected result is that as the capitalized recovery amount increases the dispersion also increases.

It is really interesting the results on the influence of EAD on the recovery rates since some empirical studies lead to different 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.

¹ I obtain these results by using the package "LogicReg" and the procedure "optim" with the method "Nelder-Mead" of R-program.

(2005) hit upon that the recovery rates decrease when the loan size increases. From the results of this work the logarithm of capitalized EAD has an inverse relationship with the dependent variable for the discrete part and has a direct relationship with the conditional expectation for the continuous part. This result is coherent with the outcomes of a descriptive analysis showed by Calabrese and Zenga (2008) on the same data.

I also highlight the different influences of the time in default on the extreme values (direct) and on the conditional expectation of the recovery rates belonging to the interval $(0,1)$ (inverse). Finally, the logarithm of capitalized EAD and the amount of collateral or personal guarantee are not significant for the dispersion parameter ϕ .

References

1. Asarnow, E., Edwards, D.: Measuring loss on default bank loans: A 24-year study. *Journal of Commercial Lending*. **77**, 11–23 (1995)
2. Banca d'Italia: Principali risultati della rilevazione sull'attività di recupero dei crediti. *Bollettino di Vigilanza*. **12**, December (2001)
3. Barndorff-Nielsen, O.: Information and exponential families in statistical theory. Wiley, New York (1978)
4. Basel Committee on Banking Supervision International convergence of capital measurement and capital standards: A revised framework. Bank for International Settlements. Basel, June (2004)
5. Calabrese, R., Zenga, M.: Measuring loan recovery rate: Methodology and empirical evidence. *Statistica & Applicazioni*. **6**, 193–214 (2008)
6. Calabrese, R., Zenga, M.: Bank loan recovery rates: Measuring and nonparametric density estimation. *Journal of Banking and Finance* (2009) Available online <http://dx.doi.org/10.1016/j.jbankfin.2009.10.001>
7. Carty, L., Lieberman, D.: Defaulted bank loan recoveries. Moody's special comment. November (1996)
8. 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)
9. Dermine, J., Neto de Carvalho, C.: Bank loan losses-given-default: A case study. *Journal of Banking and Finance*. **30**, 1219–1243 (2006)
10. Ferrari, S., Cribari-Neto, F.: Beta regression for modeling rates and proportions. *Journal of Applied Statistics*. **31**, 799–815 (2004)
11. Friedman, C., Sandow, S.: Ultimate recoveries. *Risk*. **16**, 69–73 (2003)
12. 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)
13. Grunert, J., Weber, M.: Recovery rate of commercial lending: Empirical evidence for German companies. *Journal of Banking and Finance*. **33**, 505–513 (2009)
14. Gupton, G. M., Finger, C. C., Bhatia, M.: CreditMetrics. Technical document, J. P. Morgan (1997)
15. Hosmer, D. W., Lemeshow, S. Applied logistic regression. Wiley, New York (2000)
16. McCullagh, P., Nelder, J.A.: Generalized Linear Models. Chapman & Hall/CRC, London (1989)
17. McLachlan, G. J., Krishnan, T.: The EM Algorithm and Extensions. Wiley, New York (1997)
18. Schuermann, T.: What Do We Know About Loss Given Default? Recovery Risk. Working Paper, Federal Reserve Bank of New York (2003)