



**UCD GEARY INSTITUTE  
DISCUSSION PAPER SERIES**

**Regression Model for Proportions with  
Probability Masses at Zero and One**

**Raffaella Calabrese**

Geary Dynamic Lab, Geary Institute  
University College Dublin

Geary WP2012/09  
March 2012

UCD Geary Institute Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Any opinions expressed here are those of the author(s) and not those of UCD Geary Institute. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

# Regression Model for Proportions with Probability Masses at Zero and One

Raffaella Calabrese

**Abstract** In many settings, the variable of interest is a proportion with high concentration of data at the boundaries. This paper proposes a regression model for a fractional variable with nontrivial probability masses at the extremes. In particular, the dependent variable is assumed to be a mixed random variable, obtained as the mixture of a Bernoulli and a beta random variables. The extreme values of zero and one are modelled by a logistic regression model. The values belonging to the interval  $(0,1)$  are assumed beta distributed and their mean and dispersion are jointly modelled by using two link functions. The regression model here proposed accommodates skewness and heteroscedastic errors. Finally, an application to loan recovery process of Italian banks is also provided.

**Key-words:**

- proportions, mixed random variable, beta regression, skewness, heteroscedasticity.

**AMS Subject Classification:**

- 62J02, 62M10, 62FXX.

## 1 Introduction

Fractional response variables arise in many settings. Examples include the proportion of crude oil converted to gasoline after distillation and fractional [10], the 401(k) plan participation rate [20]. In some cases, the observations at one or both boundaries occur with a large frequency. For instance, the proportion of a firm's total capital accounted for by its debt capital [8] represents a fractional variable with mass point at zero. Furthermore, there are many applications with high concentration of data at both the boundaries, e.g. the ability of the patient to perform activities

---

Raffaella Calabrese  
Geary Dynamics Lab, University College Dublin, e-mail: raffaella.calabrese@ucd.ie

measured by Barthel index [16], the rate of voluntary option exercise [7], the proportion of asset allocations and equities [1], the loan recovery rates [6]. Since the last set of applications is relatively unexplored, the aim of this paper is to propose a regression model for proportions with mass points at the boundaries.

The main regression models for fractional response variable are briefly analysed. At first, the linear model is not appropriate to examine how a set of explanatory variables  $\mathbf{x}$  influences a fractional response since it does not guarantee that the predicted values of the dependent variable are restricted to the unit interval. In order to analyse the determinants  $\mathbf{x}$  of the conditional mean  $\mu$  of the fractional response variable, a widely applied approach is the generalized linear model [18] that considers a strictly monotonic and twice differentiable *link function*  $g(\cdot)$  that maps the interval  $(0,1)$  onto the whole real line and such that  $g(\mu) = \mathbf{x}'\lambda$ .

By denoting  $G(\cdot)$  the inverse of the link function, four functions are commonly used

- the logit function  $G(\mathbf{x}'\lambda) = \frac{1}{1 + \exp(-\mathbf{x}'\lambda)}$
- the probit function  $G(\mathbf{x}'\lambda) = \Phi(\mathbf{x}'\lambda)$
- the log-log function  $G(\mathbf{x}'\lambda) = \exp[-\exp(-\mathbf{x}'\lambda)]$
- the complementary log-log function  $G(\mathbf{x}'\lambda) = 1 - \exp[-\exp(-\mathbf{x}'\lambda)]$ .

A widely applied approach that handles data observed on the closed interval  $[0,1]$  (e.g. Bastos [4], Carpenter et al. [7], Grippa et al. [12], Dermine and Neto de Carvalho [9]) is the model proposed by Papke and Wooldridge [20]. The estimation procedure of the fractional response model of Papke and Wooldridge is a quasi-likelihood method that consists of the maximization of the Bernoulli log-likelihood function

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

for  $i = 1, 2, \dots, n$ , where  $n$  is the number of observations. Because equation (1) is a member of the linear exponential family, the quasi-maximum likelihood estimator of  $\lambda$ , obtained from the maximization problem

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

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

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

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

The censored normal regression model (i.e. the Tobit model) is sometimes used to address data observed on the closed interval  $[0,1]$  (e.g. Agnew et al. [1]). There are also some problems with this approach. As Maddala [17] observes, a Tobit model is appropriate to describe censored data in the interval  $[0,1]$  but its application to data defined only in that interval is not easy to justify: observations at the boundaries

of a fractional variable are a natural consequence of individual choices and not of any type of censoring. Furthermore, the Tobit model is very stringent in terms of assumptions, requiring normality and homoscedasticity of the dependent variable, prior to censoring.

When the dependent variable  $Y$  is observed on the interval  $(0,1)$  and is beta distributed, by applying a methodology similar to GLM models, Ferrari and Cribari-Neto [10] propose the beta regression model. 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. Ferrari and Cribari-Neto propose a reparameterization that translates  $p$  and  $q$  into a location parameter  $\mu$  and a dispersion parameter  $\phi$

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

where  $\phi = p + q$ . First, Ferrari and Cribari-Neto model only the conditional mean  $\mu$  and they consider the precision parameter  $\phi = p + q$  as a nuisance parameter. Second, the response variable is restricted to the interval  $(0,1)$ , neglecting the information of the boundaries, analogously to GLM models.

By applying the same regression model to both the values belonging to the interval  $(0,1)$  and to the extreme values of zero and one, it is assumed that the two sets of data show the same characteristics and follow the same model. A violation of this assumption is the cause of sample selection bias, which Heckman [14] demonstrates is another type of specification error. For instance, in the application to recovery rates considered in this paper, many authors (e.g. Friedman and Sandow [11], Grunert and Weber [13]) 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.

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 dispersion of a proportion with mass points at zero and one, given some explanatory variables. At first, the dependent variable is considered as a mixed random variable to represent the high concentration of data at the boundaries. In particular, the dependent variable is assumed to be a mixture of a Bernoulli random variable and a beta random variable. In order to analyze the influences of some explanatory variables on the continuous part, a regression model is proposed 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 dependent variable a logistic regression is applied. Finally, the proposed approach is applied to a comprehensive database of recovery rates on Italian bank loans.

The present paper is organized as follows. In section 2 the regression for a propor-

tion with mass points at zero and one is proposed. Successively, section 3 presents the Bank of Italy's database, to which the suggested methodology is applied and the main empirical results are presented. Finally, the last section is devoted to conclusions.

## 2 The Model and the Estimation

In order to supply accurate estimations for the extreme values of  $Y$ , the dependent variable  $Y$  is considered as a mixed random variable, given by the mixture of a Bernoulli random variable and a beta random variable  $B$

$$F_Y(y) = \begin{cases} P\{Y = 0\} & y=0; \\ P\{Y = 0\} + [1 - P\{Y = 0\} - P\{Y = 1\}]F_B(y) & y \in (0,1) \\ 1 & y=1 \end{cases} \quad (3)$$

where  $F_B$  denotes the distribution function of the beta random variable  $B$  and  $P\{Y = j\}$  is the probability that the dependent variable  $Y$  is equal to  $j$  with  $j = 0, 1$ . The attention is focused on the beta random variable  $B$  considered in the model (3). As above-mentioned, the beta regression approach, proposed by Ferrari and Cribari-Neto [10], assumes that the dependent variable is beta distributed. By applying the same reparameterization proposed by Ferrari and Cribari-Neto, this work models jointly the conditional mean  $\mu$  and the precision parameter  $\phi$ , defined in equations (2).

In particular, let  $B_1, B_2, \dots, B_d$  be independent random variables, where each  $B_i$ , with  $i = 1, 2, \dots, d$ , follows the density

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

where  $b_i \in (0, 1)$ ,  $1 > \mu_i > 0$  and  $\phi_i > 0$ , with mean and variance given by the equations (2). To joint model the conditional mean  $\mu_i$  and the parameter  $\phi_i$  with  $i = 1, 2, \dots, p$ , two different link function  $g(\cdot)$  and  $h(\cdot)$  are used such that

$$g(\mu_i) = \mathbf{x}'_i \alpha \quad h(\phi_i) = -\mathbf{w}'_i \beta, \quad (4)$$

with  $i = 1, 2, \dots, d$ , where  $\alpha$  and  $\beta$  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 ( $k + m < d$ ), which are assumed fixed and known. In order to define the link function  $h(\cdot)$ , in the second equation of (4), the negative sign is considered in order to model the dispersion of the dependent variable  $B$ .

Furthermore, the conditional mean  $\mu$  and the parameter  $\phi$  depend on two different vectors of covariates,  $\mathbf{x}$  and  $\mathbf{w}$  respectively. By such characteristic, the model here proposed can consider some variables in the vector  $\mathbf{w}_i$  that are relevant just for the parameter  $\phi_i$  and not for the conditional mean  $\mu_i$ . Since  $0 < \mu_i < 1$  and  $\phi_i > 0$  with

$i = 1, 2, \dots, d$ , it is supposed that the link function  $g(\cdot)$  is the logit function and the link function  $h(\cdot)$  is the log function, it follows that

$$\mu_i = \frac{1}{1 + e^{-x_i' \alpha}} \quad \phi_i = e^{-w_i' \beta}, \quad (5)$$

with  $i = 1, 2, \dots, d$ . The variance of  $B_i$  is a function of  $\mu_i$  and  $\phi_i$ , 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 heteroscedastic errors. Moreover, since the beta distribution is flexible, the skewness and multimodality are also accommodated.

In order to estimate the two vectors  $\alpha$  and  $\beta$  of parameters, the maximum likelihood method is performed. The log-likelihood function is

$$\begin{aligned} l(\alpha, \beta) = \sum_{i=1}^d \left[ \ln \Gamma(e^{-w_i' \beta}) - \ln \Gamma\left(\frac{e^{x_i' \alpha - w_i' \beta}}{1 + e^{x_i' \alpha}}\right) - \ln \Gamma\left(\frac{e^{-w_i' \beta}}{1 + e^{x_i' \alpha}}\right) \right. \\ \left. + \left(\frac{e^{x_i' \alpha - w_i' \beta}}{1 + e^{x_i' \alpha}} - 1\right) \ln(b_i) + \left(\frac{e^{-w_i' \beta}}{1 + e^{x_i' \alpha}} - 1\right) \ln(1 - b_i) \right]. \end{aligned}$$

Since the beta distribution is a two parameter full exponential family and the log-likelihood function satisfies a given condition (Barndorff-Nielsen [3], pp. 151), the maximum likelihood estimators exist and are unique.

The score functions 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  $\Gamma(\cdot)$

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

For  $m = 0$  this function is called digamma function  $\psi(z) = \frac{\partial \Gamma(z)}{\Gamma(z)}$ . The score functions are obtained (see Appendix) by differentiating the log-likelihood function with respect to the unknown parameters  $\alpha$  and  $\beta$

$$\begin{aligned} \frac{\partial l(\alpha, \beta)}{\partial \alpha_j} &= \sum_{i=1}^d x_{ij} \frac{e^{-x_i' \alpha - w_i' \beta}}{[1 + e^{x_i' \alpha}]^2} \left[ -\psi\left(\frac{e^{-w_i' \beta}}{1 + e^{x_i' \alpha}}\right) + \psi\left(\frac{e^{x_i' \alpha - w_i' \beta}}{1 + e^{x_i' \alpha}}\right) + \log \frac{b_i}{1 - b_i} \right] \\ \frac{\partial l(\alpha, \beta)}{\partial \beta_h} &= \sum_{i=1}^d -w_{ih} \frac{e^{-w_i' \beta}}{1 + e^{-x_i' \alpha}} \left[ (1 + e^{x_i' \alpha}) \psi(e^{-w_i' \beta}) - e^{x_i' \alpha} \psi\left(\frac{e^{x_i' \alpha - w_i' \beta}}{1 + e^{x_i' \alpha}}\right) \right. \\ &\quad \left. - \psi\left(\frac{e^{-w_i' \beta}}{1 + e^{x_i' \alpha}}\right) + \log(b_i) + e^{x_i' \alpha} \log(1 - b_i) \right], \quad (6) \end{aligned}$$

with  $j = 1, 2, \dots, k$ ;  $h = 1, 2, \dots, m$ ;  $i = 1, 2, \dots, d$ , where  $b_i$  is a realization of the dependent variable belonging to the interval  $(0, 1)$ .

The Fisher information matrix (see Appendix) is computed in order to obtain the

asymptotic standard errors of the maximum likelihood estimators of the parameters. Since the Fisher's information matrix is not a diagonal matrix, the parameter vectors  $\alpha$  and  $\beta$  are not orthogonal, so their maximum likelihood estimators are dependent and can not be computed separately.

The maximum likelihood estimators of  $\alpha$  and  $\beta$  are obtained by making the score functions (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 [19]. The optimization algorithms require the specification of initial values to be used in iterative scheme.

It is suggested to use as an initial point estimate for  $\alpha$  the ordinary least squares estimate of this parameter vector obtained from a linear regression of the transformed response

$$\phi_i = \frac{\mu_i(1 - \mu_i)}{\text{var}(B_i)} - 1.$$

By applying the delta method the following approximation is derived

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

so it is obtained that

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

Hence, the following approximation is used

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

with  $\hat{\mu}_i = \frac{e^{\mathbf{x}_i' \hat{\alpha}}}{1 + e^{\mathbf{x}_i' \hat{\alpha}}}$ , where  $\hat{\alpha}$  and  $\hat{e}$  are, respectively, the ordinary least squares estimate and residual obtained from the linear regression of the transformed responses  $\text{logit}(y_i)$  on  $\mathbf{x}_i$ . As initial point estimate for  $\beta$ , the ordinary least squares estimate obtained from a linear regression of the transformed value  $-\ln(\hat{\phi}_i)$  on  $\mathbf{w}_i$  is used. After estimating the conditional mean and variance of the beta random variable  $B$ , a great advantage of the regression model proposed in this paper is that, under the assumption (3), it allows to estimate both the conditional mean and variance of the dependent variable  $Y$  given by

$$E(Y/\mathbf{x}, \mathbf{w}) = E(I/\mathbf{x}, \mathbf{w})P\{(Y = 0) \cup (Y = 1)\} + E(B/\mathbf{x})P\{0 < Y < 1\}$$

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

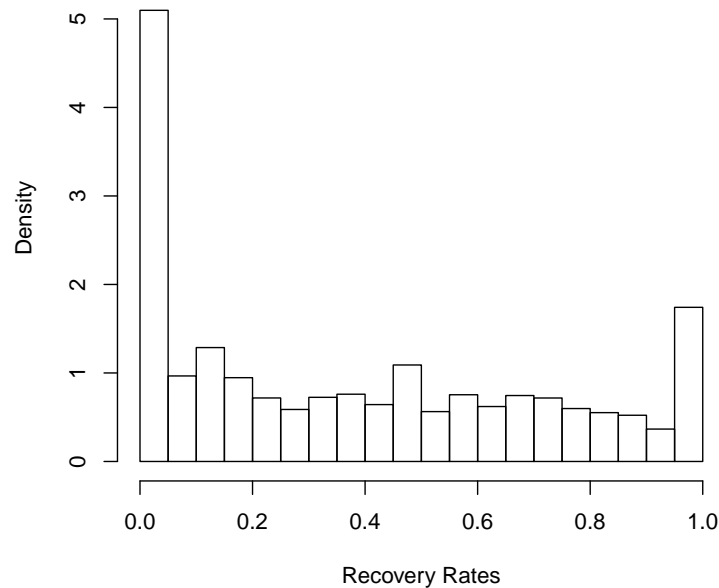
$$+[E(I/\mathbf{x}) - E(B/\mathbf{x}, \mathbf{w})]^2 P\{0 < Y < 1\}$$

where  $I$  is a Bernoulli random variable whose conditional mean is estimated by the logistic regression model (see Hosmer and Lemeshow [15]). Finally, the mixture weights are estimated by the corresponding relative frequencies.

### 3 An application to 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. 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. The expression proposed by Calabrese and Zenga [5, 6] is applied to compute the recovery rate and it shows the advantage that the recovery rate is constrained within the interval  $[0,1]$ . The distribution of the Bank of Italy's recovery rates is shown in Figure 1.



**Fig. 1** The distribution of the Bank of Italy's recovery rates.



The regression model proposed in this work is applied to the Bank of Italy's database. Considering the recovery rate as a mixed random variable, the probabilities in (3) are estimated by the corresponding relative frequencies. In addition, the discrete and the continuous parts of the recovery rate  $Y$  are modeled separately. On the one hand, the logistic regression model is applied on  $n_2 = 45,867$  extreme values of the recovery rates. On the other hand,  $n_1 = 103,511$  data exhibit recovery rates belonging to the interval  $(0,1)$  on which the regression model proposed in this paper is applied. Such methodology allows to analyze the different influences of the covariates on the boundaries and the continuous part of the recovery rate.

As above-mentioned, some authors (e.g. Friedman and Sandow [11]) hypothesize that the extreme values of the recovery rates show different characteristics from the ones belonging to the interval  $(0,1)$ . In order to verify this statement, the same set of covariates for  $\mathbf{x}$  and  $\mathbf{w}$  are chosen given by six determinants of the the recovery risk: recovery amount, logarithm of the Exposure At Default (EAD), time in default, interest on delayed payment, legal costs and amount of collateral or personal guarantee.

Table 1 reports the parameter estimates and the p-values in round brackets obtained by the application of the methodological proposal of this work to 149,378 data of the Bank of Italy<sup>1</sup>.

	Logistic Regression	Regression for Continuous Part	
		$\alpha$	$\beta$
Constant	-14.099(0.000)	-0.082(0.000)	-1.127(0.000)
Recovery amount	26.005(0.000)	0.011(0.000)	0.058(0.000)
Log 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)

**Table 1** Parameter estimates on 149,378 data of the Bank of Italy.

Choosing a level of significance of 0.05, for both the discrete and the continuous parts of the recovery rate  $Y$ , interest on delayed payment and legal costs are not significant. Furthermore, Table 1 shows that the 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 interesting to analyse the influence of EAD on the recovery rates since some empirical studies lead to different conclusions on this topic: Asarnow and Edwards [2] find no significant influence of the loan size on LGDs, instead Dermine and Neto de Carvalho [9], Grippa et al. [12] hit upon that the recovery rates decrease when the loan size increases. From Table 1, the logarithm of EAD has a different influence on the means of the extreme values and of the beta random variable.

<sup>1</sup> These results are obtained by using the package "LogicReg" and the procedure "optim" with the method "Nelder-Mead" of R-program.

In order to analyse the performance of the regression model proposed in this work, the predictive accuracy of this approach is compared with the one of the fractional response model proposed by Papke and Wooldridge [20] and mentioned in section 1. Since some authors (e.g. Bastos [4], Grippa et al. [12], Dermine and Neto de Carvalho [9]) apply this fractional response model by considering the logit and the log-log link functions, defined in equations (1) and (1), respectively. Furthermore, in this article the complementary log-log link function, defined in equation (1), is considered.

The predictive accuracy of the models is assessed using two performance measures: the Root Mean Square Error (RMSE) and the Mean Absolute Error are defined as

$$RMSE = \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right\}^{1/2} \quad MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

where  $y_i$  and  $\hat{y}_i$  are the actual and the predicted recovery rates on loan  $i$ , respectively. Models with lower RMSE and MAE estimate actual recoveries more accurately. Since the models may overfit the data, resulting in over-optimistic estimates of the predictive accuracy, the RMSE and the MAE must be assessed on a sample which is different from that used in estimating the model parameters. In order to develop models with a large fraction of the available data and evaluate the predictive accuracy with the complete dataset, a 1000-fold cross-validation [21] is implemented. In this approach, the sample is randomly partitioned into 1000 subsamples of approximately equal size. Of the 1000 subsamples, a single subsample is retained for testing the model and the remaining 999 subsamples are used for estimating the model parameters. The cross-validation is repeated 1000 times with each of the 1000 subsamples used once as test data. The mean of MAE and RMSE from the 1000 folds are reported in Table 2. Moreover, Table 2 shows in round brackets the relative frequency of times that the error of the fractional response model is higher than the one of the continuous-discrete model on the 1000 folds.

<i>Error</i>	<i>Models</i>			
	<i>Continuous-discrete</i>	<i>Fractional</i>		
		<i>log-log</i>	<i>logistic</i>	<i>complementary log-log</i>
MAE	0.2271	0.3124(1)	0.3385(1)	0.3016(1)
RMSE	0.2607	0.3539(0.991)	0.3883(1)	0.3421(0.995)

**Table 2** Out-of-sample accuracy measures of different models in 1000-fold cross validation.

By the results reported in Table 2, the model proposed in this paper exhibits the means of both the MAE and the RMSE lower than the means of the respective errors of all the three fractional response models. Furthermore, on each fold the MAE of the continuous-discrete model is lower than that of all the other models. Finally, among the fractional response models the complementary log-log link function exhibits the best accuracy in terms of both the MAE and the RMSE. This result is due

to the left-skewness of the recovery rate distribution, as a nonparametric distribution estimation [6] of the Bank of Italy's data shows.

## 4 Conclusion remarks

In this work a regression model for a proportion with high concentration of data at the boundaries is proposed. At first, the dependent variable is assumed to be a mixed random variable, obtained as the mixture of a Bernoulli random variable and a beta random variable. A logistic regression model is applied to estimate the parameter of the Bernoulli random variable. For the continuous part of the dependent variable, a regression is proposed that accommodates skewness and heteroscedastic errors. The main advantage of this proposal is that it allows to analyse the different influences of the same covariates on the boundaries and on the continuous part of the dependent variable. Moreover, the model proposed in this article allows to estimate both the mean and the variance of the dependent variable, knowing the covariates. Afterwards, the regression model here proposed is applied to the Bank of Italy's data. An interesting result is the different influence of the Exposure At Default on the means of the extreme values and of the values belonging to the interval  $(0,1)$ . Finally, the model here proposed shows lower out-of-sample accuracy than the one of the fractional response model for different link functions.

## References

1. Agnew, J., Balduzzi, P., Sundén, A.: Portfolio Choice and Trading in a Large 401 (k) Plan. *American Economic Review*. **93** (1) 11–23 (1995)
2. Asarnow, E., Edwards, D.: Measuring loss on default bank loans: A 24-year study. *Journal of Commercial Lending*. **77**, 11–23 (1995)
3. Barndorff-Nielsen, O.: *Information and exponential families in statistical theory*. Wiley, New York (1978)
4. Bastos J. A.: Forecasting bank loans loss-given-default. *Journal of Banking and Finance* **34**(10), 2510–2517 (2010)
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* **34** (5) 903–911 (2010)
7. Carpenter, J. F., Stanton, R., Wallace, N.: Estimation of Employee Stock Option Exercise Rates and Firm Cost. *Finance Working Paper* (2009) Available via DIALOG. <http://archive.nyu.edu/handle/2451/29546>
8. Cook, D. O., Kieschnick, R., McCullough, B. D.: Regression analysis of proportions in finance with self selection. *Journal of Empirical Finance* **15**, 860–867 (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. Heckman, J.: Sample selection bias as a specification error. *Econometrica* **47**, 153–161 (1979)
15. Hosmer, D. W., Lemeshow, S. *Applied logistic regression*. Wiley, New York (2000)
16. Lesaffre, E., Scheys, I., Fröhlich, J., Bluhmki, E.: Calculation of Power and Sample Size with Bounded Outcome Scores. *Statistics in Medicine* **12**, 1063–1078 (1993)
17. Maddala, G. S.: A perspective on the use of limited-dependent and qualitative variables models in accounting reserach. *Accounting Review* **66** (4), 788-807 (1991)
18. McCullagh, P., Nelder, J.A.: *Generalized Linear Models*. Chapman & Hall/CRC, London (1989)
19. McLachlan, G. J., Krishnan, T.: *The EM Algorithm and Extentions*. Wiley, New York (1997)
20. 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) doi:10.2307/2288403
21. Picard, C., Cook, D.: Cross-Validation of Regression Models. *Journal of the American Statistical Association* **79** (387), 575583
22. Ramalho, E. A., Ramalho, J.: Alternative Estimating and Testing Emprirical Strategies for Fractional Regression Models. *Journal of Economic Surveys*. (2010) doi: 10.1111/j.1467-6419.2009.00602

## 5 Appendix

In this appendix we obtain the score functions and the Fisher information matrix for  $\alpha$ ,  $\beta$ . The notation used here is defined in the section 2. In order to compute the score functions we consider the following equations

$$\frac{\partial l_i(\alpha, \beta)}{\partial \alpha_j} = \frac{\partial l_i(\mu_i, \phi_i)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \alpha_j} \quad \frac{\partial l_i(\alpha, \beta)}{\partial \beta_h} = \frac{\partial l(\mu_i, \phi_i)}{\partial \phi_i} \frac{\partial \phi_i}{\partial \beta_h} \quad (7)$$

with  $j = 1, 2, \dots, k$ ;  $h = 1, 2, \dots, m$ ;  $i = 1, 2, \dots, d$ .

The score function and the Hessian can be obtained explicitly in terms of the polygamma function, where the polygamma function of order  $r$  is defined as the  $(r + 1)^{th}$  derivative of the logarithm of the gamma function  $\Gamma(\cdot)$

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

For  $r = 0$  this function is called digamma function  $\psi(z) = \frac{\partial \log \Gamma(z)}{\partial z} = \frac{\partial \Gamma(z)}{\partial z} \frac{1}{\Gamma(z)}$

for  $z > 0$ .

From equations (5) and (6) we obtain that

$$\frac{\partial l_i(\mu_i, \phi_i)}{\partial \mu_i} = -\phi_i \psi(\mu_i \phi_i) + \phi_i \psi(\phi_i - \mu_i \phi_i) + \phi_i \log \frac{y_i}{1 - y_i}$$

$$\begin{aligned} \frac{\partial l_i(\mu_i, \phi_i)}{\partial \phi_i} &= \psi(\phi_i) - \mu_i \psi(\mu_i \phi_i) - (1 - \mu_i) \psi(\phi_i - \mu_i \phi_i) + \mu_i \log(y_i) + (1 - \mu_i) \log(1 - y_i) \\ \frac{\partial \mu_i}{\partial \alpha_j} &= - \frac{x_{ij} e^{-\mathbf{x}'_i \alpha}}{[1 + e^{-\mathbf{x}'_i \alpha}]^2} & \frac{\partial \phi_i}{\partial \beta_h} &= -w_{ih} e^{-\mathbf{w}'_i \beta} \end{aligned}$$

with  $j = 1, 2, \dots, k$ ;  $h = 1, 2, \dots, m$ ;  $i = 1, 2, \dots, d$ . Substituting the former results and the expressions (5) in equations (7) the score functions (6) are obtained.

The second order partial derivatives of the log-likelihood function with respect to parameters  $(\mu_i, \phi_i)$  are

$$\begin{aligned} \frac{\partial^2 l_i(\mu_i, \phi_i)}{\partial^2 \mu_i} &= -\phi_i^2 [\psi'(\mu_i \phi_i) - \psi'(\phi_i - \mu_i \phi_i)] \\ \frac{\partial^2 l_i(\mu_i, \phi_i)}{\partial^2 \phi_i} &= \psi'(\phi_i) - \mu_i^2 \psi'(\mu_i \phi_i) - (1 - \mu_i)^2 \psi'(\phi_i - \mu_i \phi_i). \end{aligned} \quad (8)$$

The second order partial derivatives of the parameters  $\mu_i$  and  $\phi_i$  with respect to the regression parameters  $\alpha$  and  $\beta$  are

$$\begin{aligned} \frac{\partial^2 \mu_i}{\partial \alpha_j \partial \alpha_q} &= \frac{x_{ij} x_{iq} e^{-\mathbf{x}'_i \alpha}}{[1 - e^{-\mathbf{x}'_i \alpha}]^2} \\ \frac{\partial^2 \phi_i}{\partial \beta_h \partial \beta_u} &= w_{ih} w_{iu} e^{-\mathbf{w}'_i \beta} \end{aligned} \quad (9)$$

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

$$\begin{aligned} -E \left( \frac{\partial^2 l_i(\alpha, \beta)}{\partial \alpha_j \partial \alpha_q} \right) &= -E \left( \frac{\partial^2 l_i(\alpha, \beta)}{\partial \mu_i \partial \alpha_q} \frac{\partial \mu_i}{\partial \alpha_j} + \frac{\partial l_i(\alpha, \beta)}{\partial \mu_i} \frac{\partial^2 \mu_i}{\partial \alpha_j \partial \alpha_q} \right) \\ -E \left( \frac{\partial^2 l_i(\alpha, \beta)}{\partial \beta_h \partial \beta_u} \right) &= -E \left( \frac{\partial^2 l_i(\alpha, \beta)}{\partial \phi_i \partial \beta_u} \frac{\partial \phi_i}{\partial \beta_h} + \frac{\partial l_i(\alpha, \beta)}{\partial \phi_i} \frac{\partial^2 \phi_i}{\partial \beta_h \partial \beta_u} \right) \\ -E \left( \frac{\partial^2 l_i(\alpha, \beta)}{\partial \alpha_j \partial \beta_h} \right) &= -E \left( \frac{\partial \left[ \frac{\partial l_i(\alpha, \beta)}{\partial \alpha_j} \right]}{\partial \beta_h} \right) \end{aligned} \quad (10)$$

with  $j, q = 1, 2, \dots, k$ ;  $h, u = 1, 2, \dots, m$ ;  $i = 1, 2, \dots, d$ . By substituting the results (8) and (9) in the first two equations of (10) and by computing the derivative of the first result in (6) with respect to  $\beta_h$  (according to the last equation of (10), we obtain the Fisher information matrix whose elements are

$$-E \left( \frac{\partial^2 l(\alpha, \beta)}{\partial \alpha_j \partial \alpha_q} \right) = \sum_{i=1}^m \frac{x_{ij} x_{iq} e^{\mathbf{x}'_i \alpha - 2\mathbf{w}'_i \beta} [1 - e^{\mathbf{x}'_i \alpha}]}{[1 + e^{\mathbf{x}'_i \alpha}]^3} \left[ \psi' \left( \frac{e^{\mathbf{x}'_i \alpha - \mathbf{w}'_i \beta}}{1 + e^{\mathbf{x}'_i \alpha}} \right) + \psi' \left( \frac{e^{-\mathbf{w}'_i \beta}}{1 + e^{\mathbf{x}'_i \alpha}} \right) \right]$$

$$\begin{aligned}
-E\left(\frac{\partial^2 l(\alpha, \beta)}{\partial \beta_h \partial \beta_u}\right) &= \sum_{i=1}^m w_{ih} w_{ui} e^{-w'_i \beta} \left[ \left( \frac{e^{x'_i \alpha}}{1 + e^{x'_i \alpha}} \right)^2 \psi' \left( \frac{e^{x'_i \alpha - w'_i \beta}}{1 + e^{x'_i \alpha}} \right) - \psi' \left( e^{-w'_i \beta} \right) + \right. \\
&\quad \left. + \frac{1}{1 + e^{x'_i \alpha}} \psi' \left( \frac{e^{-x'_i \beta}}{1 + e^{x'_i \alpha}} \right) \right] \\
-E\left(\frac{\partial^2 l(\alpha, \beta)}{\partial \alpha_j \partial \beta_h}\right) &= \sum_{i=1}^m \frac{w_{ih} x_{ij} e^{x'_i \alpha - 2w'_i \beta}}{[1 + e^{x'_i \alpha}]^2} \left[ \psi' \left( \frac{e^{-w'_i \beta}}{1 + e^{x'_i \alpha}} \right) - e^{x'_i \alpha} \psi' \left( \frac{e^{x'_i \alpha - w'_i \beta}}{1 + e^{x'_i \alpha}} \right) \right] \quad (11)
\end{aligned}$$

with  $j, q = 1, 2, \dots, k$ ;  $h, u = 1, 2, \dots, m$ ;  $i = 1, 2, \dots, d$ .