Modelling LGD for unsecured retail loans using Bayesian methods

Katarzyna Bijak and Lyn C. Thomas, University of Southampton

Abstract

Loss Given Default (LGD) is the loss borne by the bank when a customer defaults on a loan. LGD for unsecured retail loans is often found difficult to model. In the frequentist (non-Bayesian) two-step approach, two separate regression models are estimated independently, which can be considered potentially problematic when trying to combine them to make predictions about LGD. The result is a point estimate of LGD for each loan. Alternatively, LGD can be modelled using Bayesian methods. In the Bayesian framework, one can build a single, hierarchical model instead of two separate ones, which makes this a more coherent approach. In this paper, Bayesian methods as well as the frequentist approach are applied to the data on personal loans provided by a large UK bank. As expected, the posterior means of parameters which have been produced using Bayesian methods are very similar to the frequentist estimates. The most important advantage of the Bayesian model is that it generates an individual predictive distribution of LGD for each loan. Potential applications of such distributions include the downturn LGD and the stressed LGD under Basel II.

Keywords: Loss Given Default, downturn LGD, Bayesian, regression models

Introduction

Loss Given Default (LGD) is the loss borne by the bank when a customer defaults on a loan. The Directive 2006/48/EC defines LGD as "the ratio of the loss on an exposure due to the default of a counterparty to the amount outstanding at default" (European Union, 2006, Article 4(27)), where 'loss' means "economic loss, including material discount effects, and material direct and indirect costs associated with collecting on the instrument" (European Union, 2006, Article 4(26)). According to the European Banking Authority guidelines, "the data used to calculate the realised LGD of an exposure should include all relevant information" (European Banking Authority, n.d., section 3.3.2.2). Among the relevant information, the guidelines mention: outstanding amount of the exposure at default (including principal as well as interests and fees), recoveries (e.g. proceeds from the sale of collateral or the loan) and work-out costs (including the costs of both in-house and outsourced collection).

Under the Basel II Advanced Internal Ratings-Based (AIRB) approach, banks are required to use their own estimates of LGD, PD (Probability of Default) and EAD (Exposure at Default). One of the requirements is that "credit institutions shall use LGD estimates that are appropriate for an economic downturn if those are more conservative than the long-run average" (European Union, 2006, Annex VII, Part 4, point 74). This is referred to as the 'downturn LGD'. The estimation of the downturn LGD can be challenging, since there is no Basel formula for it but only a principles-based approach was suggested (Basel Committee on Banking Supervision, 2005). Under the AIRB approach, banks are also expected to stress test the risk parameters, including LGD. More on LGD in Basel II can be found in books by Thomas (2009) and van Gestel and Baesens (2009).

LGD for corporate loans has been assessed for a much longer time than for retail loans, first with a fixed value based on historical data, and then using more complicated models (Thomas, 2009). Various approaches to modelling corporate LGD were presented e.g. by Altman *et al* (2005). Since the sale of collateral can have a large impact on LGD, there are separate models for secured and unsecured loans. In particular, mortgage LGD can be modelled either directly or as a combination of repossession and haircut models, where a 'haircut' is the ratio of the sale price to the estimated value of a property. Examples include models by Somers and Whittaker (2007), Qi and Yang (2009), Leow *et al* (2009 and 2010), Zhang *et al* (2010) and Tong *et al* (2011).

This paper is on modelling LGD for unsecured retail loans. Because of the LGD distribution shape, it is often difficult to fit a model to the data. Therefore, multi-stage models were proposed, such as the two-step approach presented by Matuszyk *et al* (2010). In this frequentist (non-Bayesian) approach, two separate models are estimated independently, which can be considered potentially problematic when trying to combine them to predict LGD. The first model (logistic regression) separates positive values from zeroes, whereas the second model (e.g. linear regression) allows for the estimation of the positive values. The result is a point estimate of LGD for each loan. In order to apply this approach, one has either to set a cut-off for the first model or to calculate a product of the estimated value and probability that this value is greater than zero. One can also draw a number from a Bernoulli distribution with the estimated probability, whether to assign the value or zero, which is equivalent to using a random cut-off.

Alternatively, LGD can be modelled using Bayesian methods. The Bayesian framework offers a more coherent approach, since there is a single, hierarchical model instead of two separate ones. The result is an individual predictive distribution of LGD for each loan, rather than just a single number. Having a distribution, one can use its characteristics such as quantiles. The predictive distributions can be used, for example, in the LGD stress testing process or to approximate the downturn LGD. In this paper, Bayesian methods as well as the frequentist approach are applied to the data on personal loans that were provided by a large UK bank. The data are such that the empirical distribution of LGD has a high peak at zero, which justifies the use of multi-stage approaches. With regard to Bayesian methods, they are argued to be an appropriate choice here, because they allow for an integrated estimation of hierarchical models.

The paper is structured as follows. The next section is on the research background that covers various techniques of LGD modelling as well as a short introduction to Bayesian statistics. In the third section, the frequentist and Bayesian approaches to LGD modelling are described. In the fourth section, the data and the empirical results are presented. The fifth section is a discussion on the possible uses of the results, whereas the last section includes conclusions.

Background

LGD modelling for unsecured retail loans

LGD usually takes values from the interval [0,1] and some models cannot cope with values outside this interval. However, LGD can exceed one, if a bank hardly manages to recover any of the loan and adds in its collection costs. LGD can also be negative, if the principal, interests, fees and penalties which have been paid sum up to more than the outstanding amount plus work-out costs. The LGD distribution often has a high peak at zero, since there are many customers who default but finally pay in full. This peak can be partly due to ‘cures’, i.e. defaulters who get back on track before the bank takes any action against them. There is sometimes another peak at one when many customers pay nothing. In consequence, LGD is generally found difficult to model.

LGD is typically modelled for recovery periods that are longer than typical outcome periods in PD models. Under the IRB approach, the observation period for retail LGD must cover at least five years. LGD models for unsecured retail loans can be classified as either one-stage or multi-stage approaches. As far as the former are concerned, a number of regression models were suggested: Ordinary Least Squares (OLS) regression (e.g. Querci, 2005, Bellotti and Crook, 2008 and 2009, Loterman *et al*, 2009), Least Absolute Value (LAV) regression (Bellotti and Crook, 2008 and 2009), robust and ridge regression (Loterman *et al*, 2009), beta regression (Loterman *et al*, 2009, Arsova *et al*, 2011) and fractional regression (Arsova *et al*, 2011). Other one-stage models include tobit (Bellotti and Crook, 2008) and two-tailed tobit (Bellotti and Crook, 2009). Moreover, Zhang and Thomas (2012) used survival analysis, whereas Loterman *et al* (2009) applied such techniques as Classification and Regression Trees (CART), neural networks (NN), Multivariate Adaptive Regression Splines (MARS) and Least Squares Support Vector Machines (LSSVM).

As far as the multi-stage approach is concerned, there are two and sometimes three stages, in which separate models are estimated. The first model usually discriminates positives from zeroes (and negatives, if any). In the two-stage approach, the second model allows for the estimation of the positive values. In the three-stage approach, the second model separates ones-or-greater from the rest, whereas the third model is built for the estimation of the remaining values, i.e. those from the interval (0,1).

In the first two stages, logistic regression and decision trees can serve as the discrimination models (e.g. Bellotti and Crook, 2008 and 2009, Matuszyk *et al*, 2010, Zhang and Thomas, 2012). One can also combine two discrimination tasks into one using ordinal logistic regression (Arsova *et al*, 2011). In the last stage, the following models were tried out: OLS and LAV (Bellotti and Crook, 2008), robust, ridge and beta regression, CART, NN, MARS and LSSVM (Loterman *et al*, 2009) as well as survival analysis (Zhang and Thomas, 2012). Another multi-stage approach was presented by Loterman *et al* (2009): one can estimate a linear regression in the first stage and correct it using a non-linear model in the second stage. The nonlinear model is applied to estimate the error of the linear regression.

Linear regression is usually better than survival analysis (Zhang and Thomas, 2012), tobit models and simple decision trees (Bellotti and Crook, 2008), but it tends to be outperformed by nonlinear models such as NN and MARS (Loterman *et al*, 2009). However, such findings may depend on the performance measures used. For example, in one research, OLS was better than LAV for MSE, while for MAE the opposite was true (Bellotti and Crook, 2008).

Apart from Mean Square Error (MSE) and Mean Absolute Error (MAE), the following performance measures are used for LGD models: Root Mean Square Error (RMSE), coefficient of determination (R-squared), Pearson's, Spearman's and Kendall's correlation coefficients as well as area over the Regression Error Characteristic curve (AOC) and area under the Receiver Operating Characteristic curve (AUC) (Loterman *et al*, 2009). The correlation coefficients measure correlation between the observed and predicted LGD. The AOC estimates the expected error. The AUC requires a binary variable such as the observed LGD classified into below-the-mean and over-the-mean. Thus, the AUC measures how well the model separates lower and higher values of LGD. However, Somers' D would be more suitable for this purpose, since it does not need any arbitrary classification of the dependent variable. Regardless of the measure used, most LGD models perform rather weakly.

In order to improve model performance and/or produce a more normal-shaped distribution, transformations of the original LGD are introduced. In particular, Beta transformation is often applied (e.g. Gupton and Stein, 2005, Loterman *et al*, 2009, Matuszyk *et al*, 2010). Other possible transformations include: log, fractional logit and probit (Bellotti and Crook, 2008) as well as the Box-Cox transformation (Loterman *et al,* 2009, Matuszyk *et al*, 2010). However, transformations do not necessarily lead to a better model performance (Loterman *et al*, 2009).

Ideally, an LGD model should be characterised by good performance (low errors and high correlation coefficients), stability and intuitive covariates. The covariates can be classified into five groups: socio-demographic variables (e.g. customer's age), customer's financial situation (e.g. income), account details (e.g. loan amount), payment history (e.g. outstanding balance) and macroeconomic variables. A similar, yet not identical, classification was suggested by Bellotti and Crook (2008). Using macroeconomic variables is one way to assess the downturn LGD (Caselli *et al*, 2008, Bellotti and Crook, 2009).

Bayesian statistics

So far, LGD modelling has been based on frequentist (classical) statistics, in which inference is made using sample data as the only source of information. Bayesian statistics, in turn, allows for the incorporation of other sources of information (e.g. expert knowledge). This extra knowledge is called the ‘prior information’, and is described with the prior probability distributions of the model parameters. The prior distributions are then updated using data, which yields the posterior distributions of the parameters, conditional on the observations. Providing a full distributional profile of the parameters is one of the advantages of Bayesian statistics. Other advantages include a coherent description of uncertainty in the model and direct interpretation of confidence (‘credible’) intervals. Bayesian statistics also enables an integrated estimation of complex and multilevel models (Lynch, 2007).

Since data and the prior information can to some extent compensate for each other, Bayesian methods can be successfully applied even if there is little data or no additional knowledge. The relationship between the prior and posterior distributions of the parameters can be described using Bayes’ theorem (e.g. Bernardo and Smith, 2003, Congdon, 2004). In order to generate samples from the posterior distributions, stochastic simulation methods are usually employed with Markov chain Monte Carlo (MCMC) being the most popular ones (e.g. Lynch, 2007, Ntzoufras, 2009). For more details on Bayesian statistics, it is recommended to refer to the literature cited above.

Bayesian methods have been successfully applied in credit scoring for at least 10 years. Since Bayesian statistics can effectively deal with data scarcity, it is found a useful tool for low default portfolios, LDPs (Dwyer, 2007, Kiefer, 2009, Fernandes and Rocha, 2011). It can also be employed in the stress testing process (Park *et al*, 2010). Bayesian statistics allows for the incorporation of expert knowledge or some extra information into a model, e.g. for risk-based pricing (Konstantinos *et al*, 2003). It also offers a way to update an old scorecard with new data that are not sufficient to build a new model (Ziemba, 2005). If there are no data on performance of the rejected applicants, one can use a Bayesian reject inference technique (Chen and Åstebro, 2003). Finally, Bayesian methods can be applied as an alternative to the frequentist ones, e.g. to estimate PD (Miguéis *et al*, 2012) or to find the best scorecard (Giudici, 2001).

Methodology

Frequentist approach

In this paper, Bayesian methods are compared with the frequentist approach. The latter is similar to the two-step approach presented by Matuszyk *et al* (2010). Let y_i denote LGD of the *i*th loan (i = 1, …, N). The first of the two models separates positives from zeroes and negatives. It takes the form of a logistic regression:

$$P(y_i > 0) = \frac{1}{1 + e^{-\boldsymbol{\beta}_1 \boldsymbol{x}_i}}$$

where $\boldsymbol{\beta}_1$ are the parameters and $\boldsymbol{x}_i$ are the covariates. The second model allows for the estimation of the positive values. It is a linear regression with parameters $\boldsymbol{\beta}_2$ and covariates $\boldsymbol{z}_i$:

$$E(y_i | y_i > 0) = \boldsymbol{\beta}_2 \boldsymbol{z}_i$$

The logistic regression predicts, whether there will be a (positive) loss or not. Here, its result will be referred to as the 'probability of loss'. The linear model yields the estimated LGD, provided that there is a loss. In this application, the estimation has been performed using SAS. The models have been developed on the training sample and tested on the validation sample. Based on the findings of Loterman *et al* (2009), no transformations have been applied to the original LGD. The covariates of both regressions have been chosen using the stepwise selection (they have been selected because of their statistically significant relationship with the dependent variable and not because of their role in the recovery process).

There are two problems inherent in this approach. Firstly, the two models are estimated independently, although the use of the second model is conditional on the outcome of the first one. In this situation, their independent estimation can be considered potentially problematic when trying to combine them to predict LGD: the approach is incoherent in terms of handling uncertainty. Since there is no joint probability framework, uncertainty is not propagated from the first to the second model and then into the output. Thus, a part of uncertainty about the LGD estimates is ignored. In particular, this may lead to confidence intervals that are too narrow and give a false impression of accuracy.

Secondly, it is not clear how to use the frequentist approach, once the models have been built, i.e. which value should be taken as the predicted LGD for a given loan. One option is to set a cut-off for the first model. Then zero is taken, if the probability of loss is less than the cut-off, and the estimated LGD is taken otherwise (‘cut-off approach’). This raises another question, though, which is how to set the cut-off. Alternatively, it is possible to randomly decide, whether there will be a loss or not. One can draw a number from a Bernoulli distribution with parameter equal to the probability of loss. If the result is zero, zero is taken, and if the result is one, the estimated LGD is taken. Equivalently, one can draw a cut-off from a standard uniform distribution for each loan separately (‘random cut-off approach’).

Yet another option is to calculate the predicted LGD as a product of the probability of loss and the estimated LGD (‘probability times value approach’). This product can be viewed as a mean of the discrete distribution, in which a random variable takes a value of the estimated LGD with the probability of loss, and zero with the complement probability. Regardless of the approach, the result is a point estimate of LGD for each loan. Instead, one can use the above-mentioned simple distribution with only two possible values.

Bayesian approach

In this research, Bayesian methods have been chosen, since they allow for an integrated estimation of hierarchical models. In consequence, the Bayesian approach is free from the problems that are discussed in the previous section. In this approach, there is a single, hierarchical model instead of two separate ones. The structure of the model, which resembles the random cut-off approach, is illustrated in Figure 1. Implementing the same hierarchical structure, including the same covariates, in both the Bayesian and frequentist approaches makes these approaches directly comparable.

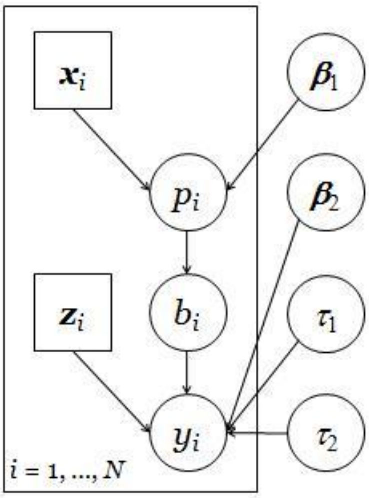

Figure 1 Bayesian hierarchical model

For each loan from the training sample, the probability of loss p_i is calculated using the logistic regression formula with parameters $\boldsymbol{\beta}_1$ and variables $\boldsymbol{x}_i$. Subsequently, a number b_i is drawn from a Bernoulli distribution with parameter p_i. If b_i equals zero, then y_i follows a normal distribution with zero mean and precision τ_1. If b_i equals one, then y_i follows a normal distribution with mean computed using the linear regression formula with parameters $\boldsymbol{\beta}_2$ and variables $\boldsymbol{z}_i$, and precision τ_2. Then the observed value of y_i is used to update the parameters $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, τ_1 and τ_2. This is the only place where it is fed into the model. The upper part of the model is not provided with additional information, whether there was a loss or not.

For each loan from the validation sample, the same operations are performed as described above, except for disclosing the observed value of y_i and updating the parameters. As a result, for each loan there is an individual predictive distribution of LGD that is a mixture of the two normal distributions mentioned above: $N(0, \tau_1^{-1})$ and $N(\boldsymbol{\beta}_2\boldsymbol{z}_i, \tau_2^{-1})$. The resulting predictive distributions are bimodal. The adopted approach is similar to the (non-Bayesian) model suggested by Hlawatsch and Ostrowski (2011) who employed a mixture of two Beta distributions to account for the bimodality of LGD for corporate loans. For comparison purposes, the probability times value approach is also applied in this paper, which produces the predictive distributions of LGD calculated as LGD* = $p_i\boldsymbol{\beta}_2\boldsymbol{z}_i$.

As far as the prior distributions are concerned, weakly informative priors are adopted for all model parameters. More informative priors are not necessary, since there is a large training sample. For each element of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, the prior is a normal distribution with zero mean and small precision (large variance). The parameter τ_1 is assumed to follow a Gamma distribution with shape parameter 10 and inverse scale parameter 0.00001. Thus, τ_1 has a very large expected value (10^6) and an even larger variance (10^{11}). In the model, τ_1 serves as precision of the normal distribution with zero mean, so the larger the τ_1, the smaller the variance of this distribution. This is designed to model the peak of the LGD distribution at zero.

The parameter τ_2 is assumed to follow a Gamma distribution with parameters 0.01 and 0.01. Hence, the expected value of τ_2 is one and its variance equals 100, which gives relatively small precision (large variance) of the normal distribution with mean based on the linear regression formula. This aims to model the rest of the LGD distribution. The initial values of all model parameters are set to be equal to the expected values of their prior distributions.

The model has been developed using OpenBUGS. The first 10000 iterations have been discarded as the burn-in period, and the next 100000 iterations have provided the MCMC output. Since relatively high autocorrelations up to lag four have been observed, a sampling lag (thinning interval) $L = 5$ has been used to obtain an independent sample.

Empirical results

Data

The methods presented above have been applied to the data on personal loans that were granted by a large UK bank between 1987 and 1998 and defaulted between 1988 and 1999 (see Table 1). The data cover the recovery periods until 2004, when some loans were still being paid. There have been ca. 50000 records in the dataset. After the removal of outliers and missing values of LGD, ca. 48000 records have remained. Subsequently, the training and validation samples of 10000 loans each have been randomly selected from the dataset. Since the period covered by the data is long enough to include the whole economic cycle, “out of time” validation does not seem necessary here.

Characteristics	***Values***
Original dataset size	49943
Dataset size w/o outliers and missing values	47853
Training sample size	10000
Validation sample size	10000
Loan open dates	1987-1998
Default dates	1988-1999
Recovery periods	Until 2004
Loan amounts at opening (in £)	500-16000
Loan terms (in months)	12-60
LGD	–0.04-1.23

Table 1 Data characteristics

The empirical distribution of LGD is demonstrated in Figure 2. Since ca. 30% of the loans were paid in full, it has a high peak at zero. There is no information on which customers were ‘cures’. Less than 10% of the loans were not repaid at all. There are many cases of LGD greater than one and few cases of LGD less than zero. They have been kept unchanged, since the models which are used in this application can cope with such values. The mean and median are equal to 0.5 and 0.59, respectively. The standard deviation equals 0.39.

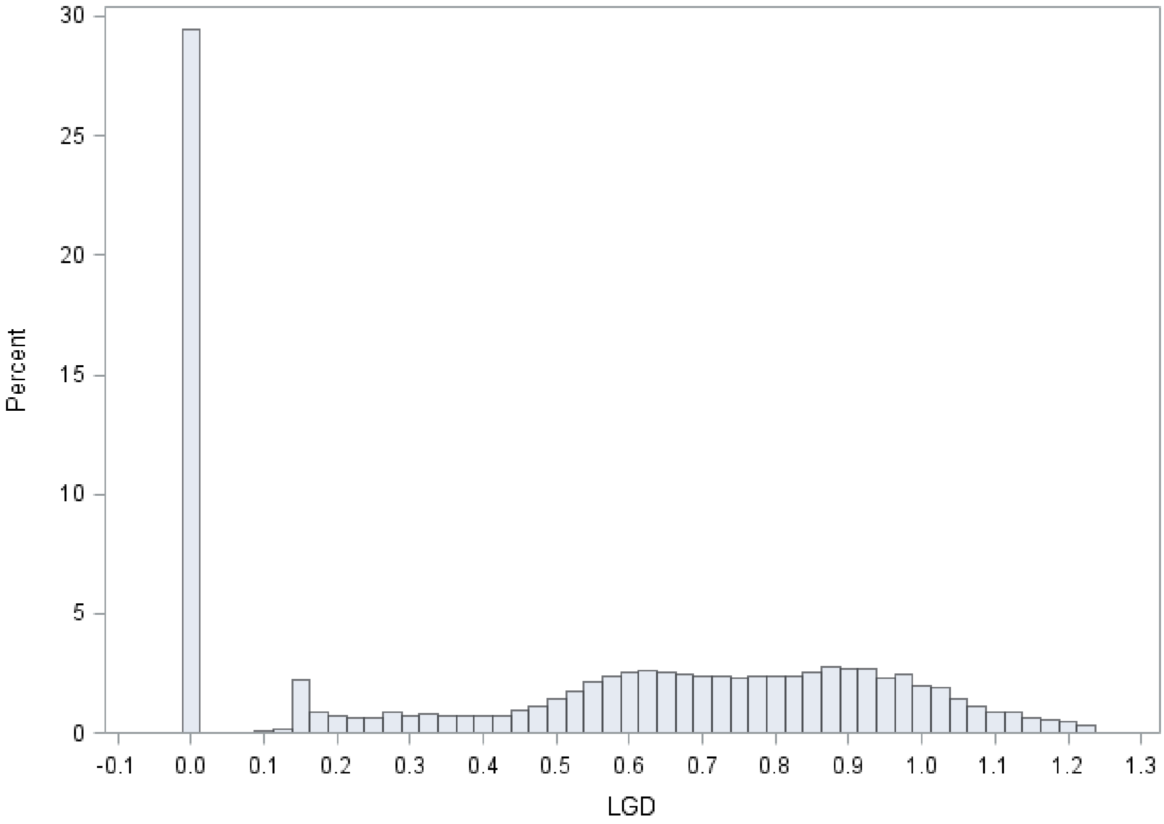

Figure 2 Empirical distribution of LGD

In the dataset, there are variables from four out of five groups mentioned in the section on LGD modelling. Macroeconomic variables have not been used in this application. Socio-demographic variables have been collected at application. Some account details reflect the situation at opening and some at default. The payment histories cover the period until default. Thus, the life of the loan means the time from opening to default, whereas the last 12 months mean the last year before default etc. The variables have been standardised.

Model convergence and performance

In the frequentist approach, the quality of each of the two models has been assessed separately before measuring the performance of the entire LGD model. The logistic regression discriminatory power has been measured with the Gini coefficient and the KS statistic, whereas the linear regression goodness of fit has been assessed using the R-squared. In the training sample, the Gini coefficient and the KS statistic equal 0.42 and 0.31, respectively. Almost the same values of these measures have been obtained on the validation sample, which means that the discriminatory power of the first model is good and stable. The R-squared of the linear regression is equal to 0.16 on both the training and validation samples. Thus, the goodness of fit of the second model is rather poor but stable. This is in line with the findings of Matuszyk *et al* (2010).

In the Bayesian approach, the monitoring of the MCMC algorithm convergence has been based on autocorrelations, quantiles and (‘trace’) plots of the generated values as well as the Monte Carlo (MC) errors that measure variability of the parameter estimates due to the simulation (Ntzoufras, 2009). The autocorrelations are low due to the use of a sampling lag. In the successive iterations, the quantiles and generated values of each parameter have been remaining within their zones with no visible tendencies, which demonstrates that the algorithm has converged. The MC errors are relatively low, since they do not exceed 1.6% of the posterior standard deviations of the parameters (see Table 2). This shows that the posterior means of the parameters have been estimated with high precision.

Parameter	***Frequentist***	***Bayesian***			
	Estimate (std. error)	***Posterior mean***	***Posterior std. dev.***	***MC error***	***MC %***
$\boldsymbol{\beta}_1$					
Intercept	1.084 (0.026)	1.087	0.026	1.19E-04	0.45
Age of exposure (months)	–0.545 (0.061)	–0.545	0.062	8.93E-04	1.45
Amount of loan at opening	0.338 (0.025)	0.339	0.025	9.93E-05	0.39
Total number of advances/ arrears within the whole life of the loan	–1.478 (0.062)	–1.481	0.062	5.25E-04	0.84
Number of months with arrears >0 within the life of the loan	0.073 (0.078)	0.076	0.078	1.23E-03	1.57
Number of months with arrears >1 within the last 12 months	–0.529 (0.040)	–0.531	0.040	3.08E-04	0.76
$\boldsymbol{\beta}_2$					
Intercept	0.719 (0.003)	0.718	0.003	9.14E-06	0.32
Joint applicant present	–0.012 (0.003)	–0.012	0.003	8.53E-06	0.29
Total number of advances/ arrears within the whole life of the loan	–0.143 (0.016)	–0.146	0.015	1.89E-04	1.23
Term of loan (months)	–0.037 (0.003)	–0.037	0.003	1.01E-05	0.32
Worst arrears within the life of the loan	0.178 (0.016)	0.180	0.016	1.91E-04	1.22
Number of months with arrears >2 within the last 12 months	–0.053 (0.004)	–0.053	0.004	1.36E-05	0.31
τ_1	-	$1.46 \cdot 10^8$	$3.83 \cdot 10^6$	12600	0.33
τ_2	-	17.580	0.294	9.37E-04	0.32

Table 2 Estimation results

In the frequentist and Bayesian approaches, the LGD model performance has been measured and compared using MSE and MAE as well as Pearson’s, Spearman’s and Kendall’s correlation coefficients. As mentioned earlier, it is not clear how to use the frequentist LGD model. Therefore, its performance has been assessed using three approaches (cut-off, random cut-off and probability times value). In the cut-off approach, the performance measures have been calculated for a number of cut-offs. Figure 3 shows that the results strongly depend on the cut-off level.

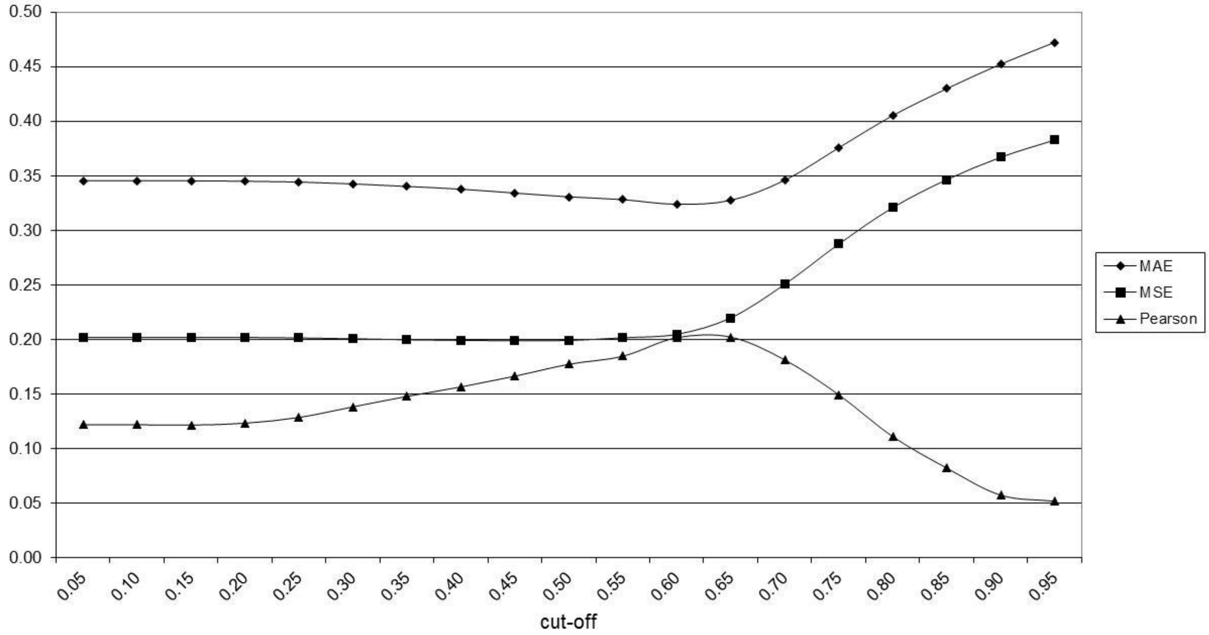

Figure 3 Performance of the frequentist LGD model (cut-off approach, validation sample)

The random cut-off approach has been implemented in the Bayesian framework. Thus, there are the posterior distributions of the performance measures applied. The posterior means and standard deviations are presented in Table 3. The results of the frequentist random cut-off approach vary from one use to another, since there is random drawing involved. Therefore, the bootstrap has been performed in order to produce the distributions of MSE, MAE and the correlation coefficients. In the bootstrap algorithm, the frequentist random cut-off approach has been applied to 10000 generated samples. The bootstrap estimates of means and standard deviations of the performance measures are almost identical as those produced in the Bayesian approach (the differences are only in the fourth decimal place). The model performance is stable. Similar values of the errors and the correlation coefficients were obtained on some datasets by Loterman *et al* (2009).

Performance measure	***Training sample***		***Validation sample***	
	Mean	***Std. dev.***	***Mean***	***Std. dev.***
MSE	0.244	0.003	0.245	0.003
MAE	0.364	0.003	0.365	0.003
Pearson’s correlation	0.081	0.010	0.085	0.010
Spearman’s correlation	0.107	0.010	0.115	0.010
Kendall’s correlation	0.084	0.007	0.090	0.007

Table 3 Model performance measures (random cut-off approach)

In addition, the probability times value approach has been applied. It has also been implemented in the Bayesian framework. The values of the performance measures which have been calculated in the frequentist probability times value approach are almost exactly the same as the corresponding posterior means presented in Table 4. The posterior standard deviations of the performance measures are not shown in this paper since they are very small. The results are stable and slightly better than those yielded in the random cut-off approach. The individual predictive distributions of LGD* are unimodal and extremely concentrated.

Performance measure	***Training sample***	***Validation sample***
MSE	0.142	0.143
MAE	0.328	0.329
Pearson’s correlation	0.256	0.268
Spearman’s correlation	0.241	0.255
Kendall’s correlation	0.169	0.179

Table 4 Model performance measures (probability times value approach)

Parameter estimates

As expected, the posterior means of the parameters which have been produced in the Bayesian approach are very similar to the estimates obtained in the frequentist approach, and so are the posterior standard deviations and the standard errors (see Table 2). The similarity of the posterior means and the corresponding frequentist estimates was also observed e.g. by Fernandes and Rocha (2011). These similarities are likely to result from the large sample sizes. They may also be related to using non-informative (as in Fernandes and Rocha, 2011) or weakly informative priors (as in this application): when informative priors are not used, data remain the only source of information for inference, as in frequentist statistics.

In this paper, the following interpretation of the posterior means (or the frequentist estimates) of the parameters $\boldsymbol{\beta}_1$ is suggested. The newer the exposure and the larger the loan amount, the higher is the probability that there will be a loss. However, the larger the number of arrears within the loan life and the larger the number of months with arrears >1 within the last year, the lower is the probability that there will be a loss. Matuszyk *et al* (2010) explained similarly surprising findings using the metaphor of ‘falling off a cliff’. The customers who tend to be in arrears (‘to keep their heads above water’) are more likely to succeed than those who have no delinquencies prior to default (‘going underwater’). The explanation is that the latter default because of some sudden changes in their lives (‘falling off a cliff’) which may affect their ability to pay forever.

The posterior means (or the frequentist estimates) of the parameters $\boldsymbol{\beta}_2$ can be interpreted as follows. The longer the term of a loan, the lower is the LGD. The presence of a joint applicant has a negative impact on LGD. Moreover, the larger the number of arrears within the loan life and the larger the number of months with arrears >2 within the last year, the lower is the LGD. The posterior means of τ_1 and τ_2 are larger than their prior means. Thus, the variances of the normal distributions are smaller than initially assumed. This is especially true of the distribution that is designed to model the peak at zero.

Predictive distributions of LGD

In the Bayesian approach, there is an individual predictive distribution of LGD for each loan, rather than just a point estimate as in the frequentist approach. Examples of such distributions for three selected loans from the validation sample are shown in Figures 4a, 4b and 4c. Each of them is a mixture of two normal distributions that are mixed in various proportions. Thus, the predictive distributions are bimodal. In fact, they have much narrower peaks at zero, but a smoothing method (kernel density estimation with a Gaussian kernel) has been used here for visualisation purposes. The dashed lines mark the observed values of LGD.

Having the predictive distributions, one can use their characteristics such as means and quantiles. If the predictive mean of LGD is treated as a point estimate for each loan, then the performance measures take the same values as presented in Table 4. Using the predictive median or other quantiles instead of the mean does not considerably improve the model performance. For the median, only MAE is slightly lower than for the mean, with values of 0.316 and 0.319 on the training and validation samples, respectively.

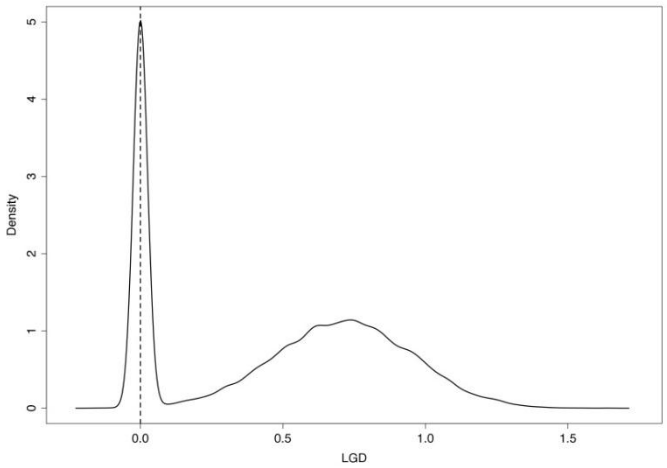

Figure 4a Predictive distribution of LGD for the loan (1)

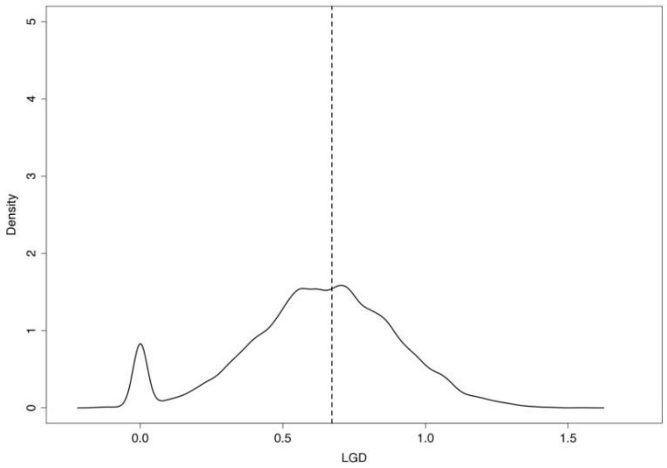

Figure 4b Predictive distribution of LGD for the loan (2)

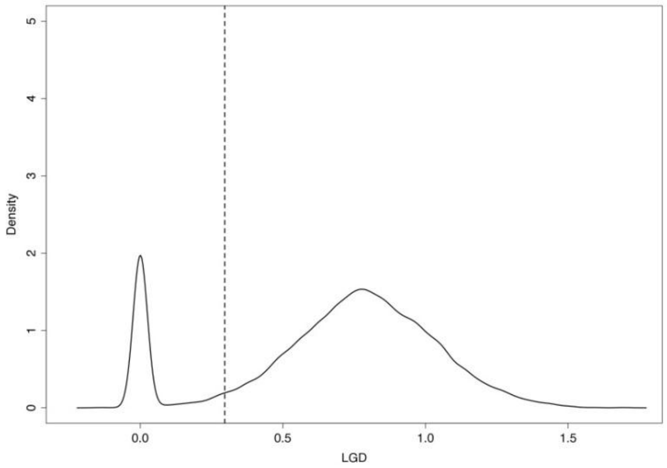

Figure 4c Predictive distribution of LGD for the loan (3)

Discussion

The individual predictive distributions provide much more information and offer more possibilities than the point estimates of LGD.

Kim (2006) proposed using the theoretical distributions to produce various LGD estimates, including the downturn LGD, for corporate exposures, in a non-Bayesian framework. In the Bayesian framework, one could approximate the downturn LGD with a certain quantile of the predictive distribution for each loan. The posterior distributions of the parameters reflect all reasonable sources of uncertainty in a Bayesian model (Gelman *et al*, 2004); what is usually not reflected is the model uncertainty. Thus, all reasonable sources of uncertainty are handled and – explicitly or implicitly – incorporated in the model, including uncertainty arising from inability to capture each and every influence on the dependent variable in the model (e.g. uncertainty related to such omitted factors as the changing macroeconomic conditions or systematic risk). Kim (2006) defined the economic downturn as "the state that the systematic risk factor takes on value at the 99.9% quantile". From the equivariance of quantiles under monotonic transformations (e.g. Hao and Naiman, 2007), it follows that if LGD is assumed to be a monotonic function of the systematic risk factor, then the selected quantile of the LGD distribution will correspond to the quantile of the same order of the underlying systematic risk factor distribution. Hence, e.g. the 0.999th quantiles will reflect both the downturn conditions and the downturn LGD. According to Kim (2006), the choice of the quantile depends on the user's perception of the severity of downturns and the 0.999th quantile can be used for extremely severe downturns. In the validation sample, choosing the 0.9th and 0.95th quantiles results in the average predicted downturn LGD of 0.97 and 1.06, respectively (while the average observed LGD of these loans was equal to 0.5 in the changing economic conditions of over a decade). Choosing the 0.75th quantile leads to the average predicted downturn LGD of 0.8, which means that such a quantile may reflect moderate downturn conditions.

In the presented example, the data cover the whole economic cycle. Had the data been collected over a shorter period of time, it could be argued that the predictive distributions would change in the downturn conditions. In particular, one could then expect them to have heavier tails. Therefore, a conservative approach would be to choose a higher quantile than if the data had covered a longer time period (e.g. the 0.99th instead of the 0.95th). Alternatively, one could correct y_i by incorporating a variable that represents the state of the economy in the

Bayesian model. After deriving the distribution of this variable from long series of historical data, such a model could be used to assess the downturn LGD.

In addition, selected quantiles of the predictive distributions can be used as the stressed LGD. One can also apply the methodology proposed by Park *et al* (2010), who stressed the coefficients instead of the corresponding financial variables in the PD model where PD was a symmetric function of the variables and their coefficients. They used the 75th percentiles of the posterior distributions of the coefficients as reflecting a stress situation. Within the approach suggested in this paper, one can stress the model parameters instead of such variables as the number of months with arrears >2 within the last 12 months. Then the appropriate quantiles of the posterior distributions of these parameters can be used to generate the stressed LGD.

Moreover, the predictive distributions of LGD can be a useful tool in the collection process. For example, a bank may wish to identify and try to recover only those loans that are likely to be paid at least partially, if not in full. Based on the predictive distributions, the bank can select the loans, for which 90% credible intervals do not include one: $P(LGD < 1) \geq 0.9$. In this application, such loans make up ca. 60% of the validation sample (in fact, 96% of them were paid at least partially). Another bank may be able to try to recover e.g. only 25% of the defaulted loans. The bank can order the loans by $P(LGD < 1)$ and take actions against the one-fourth with the highest probabilities. Yet another bank refrains from punitive actions once half of the debt has been recovered. Thus, that bank may wish to know which loans are likely to be paid in more than 50%, e.g. $P(LGD < 0.5) \geq 0.9$. Generally, the predictive distributions can be used to diversify collection strategies in order to improve the work-out process. Understandably, changing the collection process will generate the need to update the LGD model. In order to test effectiveness of the new model based strategies, a champion/challenger approach can be used.

Furthermore, the predictive distributions of LGD can help set a cut-off for the score used to accept and reject applicants. This should be based on a sample of similar loans that have already been granted. The loans need to be ranked according to the scores at application. Having the estimates of PD, LGD and EAD, one can compute the expected loss for each loan from the sample (this 12-month estimate would need to be adjusted for the loan lifetime expected loss to take a long term perspective). One can also calculate the expected profit made with the complement probability (1 – PD). Then the probability-weighted sum of the expected profit and loss can be computed for each loan. As a result, there can be an estimate of profit/loss on the entire portfolio for each level of the cut-off. The above calculations can involve the LGD quantile which reflects possible worsening of the economic situation (in particular, the downturn LGD can be used along with the downturn PD). Then a cut-off can be chosen that corresponds to the break-even point, i.e. neither profit nor loss on the portfolio. With such a cut-off, normally there should be a profit, but even in adverse economic conditions, loss is unlikely.

Finally, the individual predictive distributions, and credible intervals in particular, offer the benchmarks which can help confirm that the selected LGD estimates are sufficiently conservative.

Conclusions

In this paper, Bayesian methods have been compared and contrasted with the frequentist two-step approach to modelling LGD for unsecured retail loans. Two ways of combining the two steps (random cut-off and probability times value) have been implemented in the Bayesian framework. Then both approaches have been applied to the data on personal loans granted by a large UK bank.

As expected, the posterior means of the parameters which have been produced in the Bayesian framework are very similar to the frequentist estimates. The posterior means and standard deviations of the model performance measures are also almost identical as the corresponding bootstrap estimates that have been generated in the frequentist random cut-off approach. In comparison with the random cut-off approach, the probability times value approach has yielded slightly better posterior means of the performance measures.

In spite of the similar performance, the Bayesian model is free from the drawbacks of the frequentist approach. It is more coherent and allows for a much better description of uncertainty. The most important advantage of the Bayesian model is that it generates an individual predictive distribution of LGD for each loan, whereas the frequentist approach only produces a point estimate. The predictive distributions provide a lot of information (including benchmarks for LGD estimates) and can be used, among other purposes, for stress testing and approximating the downturn LGD.

Obviously, it is possible to generate some distributions of LGD within the frequentist framework. One way is taking into account the standard error of the predicted LGD from the second model (linear regression). This allows for the determination of confidence intervals after the adoption of the normality assumption (e.g. Maddala, 2001). If the error term is assumed to follow a normal distribution, then the predicted LGD follows a normal distribution, too. That approach has serious drawbacks. It assumes normality of the error term and – in consequence – also of the LGD distribution, whereas empirical LGD distributions are known for being far from normal-shaped. Furthermore, it ignores uncertainty from the first model (logistic regression), which may lead to confidence intervals being too narrow.

Another way to generate LGD distributions is using bootstrap methods. If the sample is large, the results may be numerically similar. However, if the sample is small, the Bayesian approach offers the advantage of utilising the prior information, which can be useful e.g. in case of LDPs. It is also worth remembering that Bayesian methods yield distributions of the model parameters, whereas the bootstrap only produces distributions of their estimators (Rubin, 1981). As a result, Bayesian credible intervals have much more natural and straightforward interpretation than bootstrap-based confidence intervals (Jaynes, 1976). Differences between the two approaches are both technical and philosophical, and the choice is up to the potential user.

Yet another way to obtain LGD distributions is using survival analysis (Zhang and Thomas, 2012). In survival analysis, the time until an event occurs is usually modelled. Zhang and Thomas (2012) applied the Cox proportional hazards model, but instead of the time, they estimated how much is recovered until the end of the collection process (or – in case of censored observations – the end of the period covered by data). As a result, they obtained a probability of being in the collection process for each value of the Recovery Rate (RR = 1 – LGD), which gives the RR distribution for each loan. However, the distributions derived from the Cox proportional hazards model have a major drawback. Since hazard function lines of different loans never cross one another, the ranking of loans is the same for each quantile of the distributions. The Bayesian approach which has been proposed in this paper is free from such limitations and thus much more flexible.

Further modifications of this approach could include using more informative priors, which might be beneficial in case of smaller samples than in this application. Moreover, one could apply more complex Bayesian graphical models and/or Bayesian model selection to find the best covariates of the logistic and linear regressions. In the Bayesian framework, one could also use more sophisticated models than the regressions and employ some transformations of LGD, as it could be done in the frequentist framework.

References

Altman, E., Resti, A. and Sironi, A. (2005) Recovery Risk. London: Risk Books.

Arsova, A., Haralampieva, M. and Tsvetanova, T. (2011) Comparison of regression models for LGD estimation. Credit Scoring and Credit Control XII, Edinburgh.

Basel Committee on Banking Supervision (2005) Guidance on Paragraph 468 of the Framework Document. Basel: Bank for International Settlements.

Bellotti, T. and Crook, J. (2008) Modelling and estimating Loss Given Default for credit cards. *Credit Research Centre Working Paper*, WP 08/1.

Bellotti, T. and Crook, J. (2009) Loss Given Default models for UK retail credit cards. *Credit Research Centre Working Paper*, WP 09/1.

Bernardo, J.M. and Smith, A.F.M. (2003) *Bayesian Theory*. Chichester: Wiley.

Caselli, S., Gatti, S. and Querci, F. (2008) The Sensitivity of the Loss Given Default Rate to Systematic Risk: New Empirical Evidence on Bank Loans. *Journal of Financial Services Research*, 34(1), pp. 1-34.

Chen, G. and Åstebro, T. (2003) Bound and Collapse Bayesian Reject Inference When Data are Missing not at Random. In: Åstebro, T., Beling, P., Hand, D., Oliver, B. and Thomas, L.B. (eds) *Mathematical Approaches to Credit Risk Management: Conference Proceedings*. Banff, Alberta: Banff International Research Station for Mathematical Innovation and Discovery.

Congdon, P. (2004) *Applied Bayesian Modelling*. Chichester: Wiley.

Dwyer, D.W. (2007) The distribution of defaults and Bayesian model validation. *The Journal of Risk Model Validation*, 1(1), pp. 23-53.

European Banking Authority (n.d.) *Electronic Guidebook*. Available at: http://www.eba.europa.eu/Publications/Compendium-of-guidelines.aspx (Accessed: 9/10/11).

European Union (2006) Directive 2006/48/EC of the European Parliament and of the Council of 14 June 2006 relating to the taking up and pursuit of the business of credit institutions (recast). *Official Journal of the European Union* L 177 of 30 June 2006.

Fernandes, G. and Rocha, C.A. (2011) Low default modelling: a comparison of techniques based on a real Brazilian corporate portfolio. Credit Scoring and Credit Control XII, Edinburgh.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004) *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Giudici, P. (2001) Bayesian data mining with application to benchmarking and credit scoring. *Applied Stochastic Models in Business and Industry*, 17(1), pp. 69-81.

Gupton, G.M. and Stein, R.M. (2005) LossCalc v2: dynamic prediction of LGD, modelling methodology. Moody's KMV.

Hao, L. and Naiman, D.Q. (2007) *Quantile Regression*, Thousand Oaks, CA: Sage Publications.

Hlawatsch, S. and Ostrowski, S. (2011) Simulation and estimation of loss given default. *The Journal of Credit Risk*, 7(3), pp. 39-73.

Jaynes, E.T. (1976) Confidence Intervals vs Bayesian Intervals. In: Harper, W.L. and Hooker, C.A. (eds) *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*. Dordrecht: D. Reidel, pp. 175-257.

Kiefer, N.M. (2009) Default estimation for low-default portfolios. *Journal of Empirical Finance*, 16(1), pp. 164-173.

Kim, M.-J. (2006) Downturn LGD, Best Estimate of Expected Loss, and Potential LGD under Basel II. *Journal of Economic Research*, 11(2), pp. 203-223.

Konstantinos, S., Dimitrios, V. and Georgios, A. (2003) Risk-Based Pricing (RBP) Using Bayesian Statistics: How to Market RBP in the Context of New Credit Card Customers. Credit Scoring and Credit Control VIII, Edinburgh.

Leow, M., Mues, C. and Thomas, L. (2009) LGD Modelling for Mortgage Loans. Credit Scoring and Credit Control XI, Edinburgh.

Leow, M., Mues, C. and Thomas, L. (2010) Competing Risks Survival Model for Residential Mortgage Loans. European Conference on Operational Research EURO XXIV, Lisbon.

Loterman, G., Brown, I., Martens, D., Mues, C. and Baesens, B. (2009) Benchmarking State-Of-The-Art Regression Algorithms For Loss Given Default Modelling. Credit Scoring and Credit Control XI, Edinburgh.

Lynch, S.M. (2007) *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. New York: Springer.

Maddala, G.S. (2001) *Introduction to Econometrics*. Third Edition. Chichester: Wiley.

Matuszyk, A., Mues, C. and Thomas, L.C. (2010) Modelling LGD for unsecured personal loans: decision tree approach. *Journal of the Operational Research Society,* 61(3), pp. 393-398.

Miguéis, V.L., Benoit, D.F. and van den Poel, D. (2012) Enhanced decision support in credit scoring using Bayesian binary quantile regression. *Journal of the Operational Research Society*, doi:10.1057/jors.2012.116.

Ntzoufras, I. (2009) *Bayesian Modeling Using WinBUGS*. Hoboken, NJ: Wiley.

Park, Y., Sirakaya, S. and Kim, T.Y. (2010) A Dynamic Hierarchical Bayesian Model for the Probability of Default. *Center for Statistics and the Social Science Working Paper,* 98.

Qi, M. and Yang, X. (2009) Loss given default of high loan-to-value residential mortgages. *Journal of Banking & Finance,* 33(5), pp. 788-799.

Querci, F. (2005) Loss Given Default on a medium-sized Italian bank's loans: an empirical exercise. European Financial Management Association Annual Meetings, Milan.

Rubin, D.B. (1981) The Bayesian bootstrap. *The Annals of Statistics*, 9(1), pp. 130-134.

Somers, M. and Whittaker, J. (2007) Quantile regression for modelling distributions of profit and loss. *European Journal of Operational Research*, 183(3), pp. 1477-1487.

Thomas, L.C. (2009) *Consumer Credit Models: Pricing, Profit, and Portfolios*. New York: Oxford University Press.

Tong, E., Mues, C. and Thomas, L. (2011) A zero-adjusted gamma model for estimating loss given default on residential mortgage loans. Credit Scoring and Credit Control XII, Edinburgh.

Van Gestel, T. and Baesens, B. (2009) *Credit Risk Management. Basic concepts: financial risk components, rating analysis, models, economic and regulatory capital*. New York: Oxford University Press.

Zhang, J. and Thomas, L.C. (2012) Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. *International Journal of Forecasting,* 28(1), pp. 204-215.

Zhang, Y., Ji, L. and Liu, F. (2010) Local Housing Market Cycle and Loss Given Default: Evidence from Sub-Prime Residential Mortgages. *IMF Working Paper,* WP/10/167.

Ziemba, A. (2005) Bayesian updating of generic scoring models. Credit Scoring and Credit Control IX, Edinburgh.