LGD Parameter Scoring using Beta Regression Model

Jan Polívka*

Abstract
Basel II gave birth to broader world of modelling opportunities in credit risk. One of them - established due to standardization of risk parameters led to development of models for LGD scoring. We present a new parametric model to score LGD here. The model theoretical properties are laid out and the model is compared on real-life data to the benchmark parametric model used by industry. Despite model great flexibility our study shows that its benefits are marginal. The main reason is that it is difficult to find covariates which would, with sufficient predictive power, explain changes in LGD distribution.

Keywords
Basel II, LGD parameter, Scoring, Maximum likelihood.

1 Introduction

Banks help clients satisfy their financial needs. Bank’s retail clients acquire consumer finance products to settle down their day to day shopping and expenses or to buy durable goods. There are clients with different ability to repay acquired loans. Bank sets up its policy of how much of credit risk it wishes to bear.

Once a decision about risk policy was taken on, bank has in its balance sheet a portion of portfolio which is in arrears. Bank seeks optimal approach to work with such portfolio of clients in order to rescue them from default.

When a client becomes past due he or she is asked to cooperate with bank to pay off past due items. Bank tries to find optimal decision strategy how to work with such a client. It is not main bank business to work with clients which are past due. Hence fast solution is sought. The key factor is willingness or ability to cooperate. Statistical tools to predict success or failure of retail client to pay off his or her debt are then of importance. There is a unique number which can describe to what extent has been bank successful: \( \text{LGD} \) or \( \text{RR} := 1 - \text{LGD} \) parameter. Bank then decides on a prospect of recovery of an exposure or it is more economical to sell the exposure to third party: decision might be based on \( RR \) value within fixed time frame after default or combination of \( RR \) and other negative information about a client.

*Mgr. Ing. Jan Polívka, Ph. D., Česká Spořitelna, Erste Bank Group e-mail: jenda.polivka@tiscali.cz

The paper expresses opinion and experience of the author.
Two groups of models have been applied in this area: parametric ones and non-parametric ones. Parametric families of models comprise logistic regression [5], transformed dependent models [6] and models for censored data [8] and [1]. Non-parametric models include neural networks [3], vector support machines [11] and decision and regression trees [10] and generalized additive models. Whereas parametric models are easier to explain to a regulator due to their lucid structure non-parametric models are able to better cope with non-linearity in data. When transformation of input variables (covariates) is allowed then even a logistic regression is able to handle non-linearity in data. Non-parametric models are usually used more for situations where structure of model is not published. They became more popular as they started to be a part of common industry used statistical packages.

Further segmentation is based on fact whether a modeled response is categorical or continuous. While non-linearity in dependency between response and explanatory variables is not a main issue, there is a substantial effort put to modeling of real life features of distributions observed. Figure 1 shows one of them.

Figure 1: Distribution of RR modeled using one beta density
Distribution is bi-modal or multi-modal with heavy tails around zero and one. There are observations above one and below zero sometimes. This special type of response distribution is arising under workout recovery rate(RR) measurement methodology and is underpinned by the following facts:

- Definition of default. There is a group of exposures which are marked as defaulted although they are not in arrears. This happens when bank adopts a client based
approach to definition of default. An exposure might become defaulted because one of its holders become defaulted in a role of a co-debtor or in a role of a guarantor on defaulted exposure of some third party.

- **Write-off (charge-off) policies.** An exposure is written-off as its expected recovery cost to revenues ratio is well above one.

- **An exposure is sold in auction sale.** This effective and fast workout solution leads to clustering of RRs around selling price. Selling of exposures with periods of varying length after default leads to new mode in RR distribution even when selling price is fixed.

- **Discount rate choice and sanction penalty.** Discount rate (contractual or effective interest rate) and sanction penalty are applied by exposures having often different periods of length. Consequently they lead to RRs above 100% or losses higher than 100%.

All above mentioned facts inspired researchers to seek parametric model which would encompass most of features mentioned above. Recent paper [7] nicely surveys beta regression models applied to portfolio credit risk models. However we are motivated by RR scoring of retail consumer or mortgage products rather than by working with simulated or time-series data over longer time span. We came to one instance of beta regression model by different motivation. Also factors determining RRs are different in our case.

## 2 Model properties

Figure 1 showed us real-life distribution of individual \( RR \) observations. The distribution is mixture of several distributions. We can hardly ever find one parametric distribution which would fit general framework of multi-modality. However beta distribution is one of flexible parametric families advocated by practitioners [9].

Figure 1 shows that it might not fit well in our case. However, conditionally there is a hope that there are covariates that could predict parameters of underlying distributions that are base for \( RR \) distribution. We expect that the resulting mixture distribution would have much better fit than model that assumes that one beta distribution fits data well [9]. This situation is illustrated on figure 2. The new curves are fitted using beta distribution family. The left most one and the right most one illustrate conditional densities of beta distributions whose parameters depend on covariates explaining propensity towards lower or higher recovery rates. Mixture of distributions would fit data well in case that there are covariates that identify groups of observations with different beta distribution parameters. We formulate such model in next few paragraphs.
Bank observes recovery rate $Y_1, \ldots, Y_n$ for pool of exposures. Recovery rate distribution is specified by

$$Y_i \sim B(a(x_i), b(x_i)), \quad EY_i(x_i) = \mu(x_i) = \alpha \frac{a(x_i)}{a(x_i) + b(x_i)}$$

Here $\sim$ in (1) stands for ”has distribution” and $B$ denotes cumulative distribution function of a beta distribution. Expected value of $RR$ is modeled through the same link function as in logistic regression. For practical application it is convenient to assume that $\varphi = (\alpha, \beta, \gamma, \delta)^T \in \Psi$ is open subset of $\mathbb{R}^+ \times \mathbb{R}^K \times \mathbb{R}^+ \times \mathbb{R}^L$ and $\overline{\Psi}$ is bounded. Our model closely resembles to a class of family of generalized linear models with two dimensional link function. This class of models is usually estimated by iterative weight least squares method which is an iterative approximation to maximum likelihood estimation and under certain conditions leads to the same estimator properties. We decided to use directly maximum likelihood to avoid assumptions of iterative weighted least squares and obstacles with choice of the starting point. Here we find that direct estimation by maximum likelihood (maximization of log likelihood of data) is acceptable in terms of speed and is well theoretically justified (see below).

Likelihood equation and log-likelihood for $Y_i, i = 1, \ldots, n$ independent not equally dis-
distributed observations is
\[ L(\varphi) = \prod_{i=1}^{n} f_i(x_i, \varphi) \]  
(2)

\[ \frac{\partial \prod_{i=1}^{n} f_i(x_i, \varphi)}{\partial \varphi} = 0 \]  
(3)

with \( f_i(x_i, \varphi) \) density of \( B(a(x_i), b(x_i)) \). We come to following consistency and asymptotic efficiency result:

2.1 Theorem. (i) Given likelihood function (2), model specification (1) and
\[ \varphi = (\alpha, \beta, \gamma, \delta)^T \in \Psi, \]
\( \Psi \) open subset with \( \overline{\Psi} \) bounded and if \( \hat{\varphi}_n \) represents solution to likelihood equation (3) for \( \varphi \) and if \( \varphi_0 \) represent true value for \( \varphi \), then as \( n \to \infty \), \( \hat{\varphi}_n \) is consistent estimator of \( \varphi_0 \).

(ii) Matrix \( \left( \frac{\partial^2 \ln f}{\partial \varphi_r \partial \varphi_s} \right) |_{\varphi=\hat{\varphi}_n} \) is negative definite with probability approaching unity where \( \hat{\varphi}_n \) is consistent estimator of \( \varphi_0 \) and a solution to (3) as \( n \to \infty \).

(iii) Of all possible solutions to (3), one and only one tends in probability to true parameter vector \( \varphi_0 \).

(iv) If \( \hat{\varphi}_n \) is vector of maximum likelihood estimators and \( \varphi_0 \) is vector of true parameter values, then \( \sqrt{n}(\hat{\varphi}_n - \varphi_0) \) has asymptotically as \( n \to \infty \) multivariate normal distribution with zero and mean covariance matrix \( J(\varphi_0)^{-1} \):
\[ J(\varphi_0) := \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} E\left( \frac{\partial \ln f_i(x_i, \varphi)}{\partial \varphi_r} \frac{\partial \ln f_i(x_i, \varphi)}{\partial \varphi_s} \right) \bigg|_{\varphi=\varphi_0} \]
\[ = -\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} E\left( \frac{\partial^2 \ln f_i(x_i, \varphi)}{\partial \varphi_r \partial \varphi_s} \right) \bigg|_{\varphi=\varphi_0}. \]

Proof. Utilizes properties proved in paper [4] and is forthcoming in next paper.

3 Model comparison

We compare two models applied to RR scoring: standard industry used model [9] which is parametric transformed dependent model against our new model (1). Transformed dependent model [9] assumes
\[ N^{-1}B(Y_i, u, v) = x_i^T \eta + \epsilon_i, \epsilon_i \sim \text{i.i.d} \]
(4)

\( \epsilon_i \) have normal distribution. Parameters \( u, v, \eta \) are common for all observations and \( N, B \) denote standard normal and cumulative beta distribution functions. This is our benchmark model.

The comparison of models is performed on two exposure pools from retail consumer finance products. One of them is revolving facility product. Population samples and models comparison setting is described in table below:

176
Table 1: Numerical study set up

The choice of covariates is crucial for models predictive power. Here, we chose the same covariates for both models to have equal conditions for model comparison. Covariates were re-scaled and non-linearly transformed. Data samples covered observations over several years were changes in workout conditions and legislative settings took place. We estimated all parameters by maximum likelihood estimation procedure except for parameter $\alpha$, $\alpha := 1$ in our beta regression model. Observations of recovery rate above 100% and below 0% were randomized to observations close to 100% and slightly above 0% respectively.

As evaluation metric was chosen Sommer’s D [2] which is a robust rank-based statistics. It gives the difference between the two probabilities for randomly chosen two pairs of (score, recovery rate): the probability that the larger of the two score values is associated with the larger of the two recovery rates minus the probability that the larger of the two score values is associated with the smaller of recovery rates. All probabilities are conditional on probability that scores are different as we corrected for ties in score variable. Sommer’s D around 50% might be regarded as excellent value for recovery rate scoring. Each validation sample was scored using model developed on training sample - we carried out out of sample predictive power validation. Validation and training sample had equal sizes.

Our findings are summarized below:

Table 2: Product 1 Sommer’s D (%), out of sample validation

<table>
<thead>
<tr>
<th>Model</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM</td>
<td>44.39</td>
<td>36.47</td>
<td>43.04</td>
<td>36.35</td>
</tr>
<tr>
<td>TB</td>
<td>44.12</td>
<td>35.32</td>
<td>43.87</td>
<td>43.34</td>
</tr>
</tbody>
</table>

*Bootstraped confidence interval at 5% confidence level

Table 3: Product 2 Sommer’s D (%), out of sample validation

<table>
<thead>
<tr>
<th>Model</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM</td>
<td>18.84</td>
<td>18.74</td>
<td>19.29</td>
<td>20.11</td>
</tr>
<tr>
<td>TB</td>
<td>18.05</td>
<td>19.44</td>
<td>18.97</td>
<td>20.08</td>
</tr>
</tbody>
</table>

Beta regressions does not significantly improve a predictive power in our setting. Main reason is that a random component of beta distribution not explained by covariates is so high that drawbacks of model (4) are of lower importance. Potential bias in parameter estimates of the model (4) is of a lower order in comparison to an unexplained noise component. We can recommend beta regression only if factors determining recovery rate are strongly identified. In this case, the model potential could be exploited.
4 Conclusions

We have shown an application of beta regression models to recovery rate scoring on real-life data. Beta regression model in comparison to transformed dependent regression model proved to have no added value in terms of predictive power. In our case, this was due to relatively low ability of explanatory covariates to identify separate groups of observations having different beta distributions.

References


Acknowledgement The paper is based on research activities sponsored through GAČR (Czech Science Foundation – Grantová Agentura České Republiky) project under No. 402/08/1237. The support is greatly acknowledged.
Summary

Scoring prarametru LGD pomoci beta regresního modelu

Koncept Basel II rozšířil příležitosti k modelování úvěrového rizika (credit risk) i množství jejich aplikací. Jedním ze směrů je vzhledem ke standardizaci rizikových parametrů rozvoj modelů LGD (ztráta daná selháním) scoringu. V článku je pro tuto oblast prezentován nový parametrický model. Nejprve jsou položeny teoretické základy modelu, dále dochází k jeho aplikaci na reálných datech z bankovního prostředí, včetně srovnání. Jako benchmark je použit model běžně užívaný v odvětví. Přes nepodstatně větší flexibilitu navrženého modelu naše studie ukazuje, že přínosy jsou omezené, zejména vzhledem k obtížím při hledání proměnných, které by s dostatečnou vypovídací schopností vysvětlovaly změny v rozdělení pravděpodobnosti LGD.