Recovery Rates in Consumer Lending: Empirical Evidence and Model Comparison

Samuel Prívara[a,*], Marek Kolman[b], Jiří Witzany[b]

[a]*Department of Control Engineering, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic*
[b]*Department of Banking and Insurance, Faculty of Finance and Accounting, University of Economics, Czech Republic*

Abstract

The bank regulation embodied in the Basel II Accord has opened-up a new era in estimating recovery rates or complementary loss given default in retail lending credit evaluation process. In this paper we investigate the properties of survival analysis models applied for recovery rates in order to predict loss given default for retail lending. We compare the results to standard techniques such as linear and logistic regressions and discuss the pros and cons of the respective methods. The study is performed on a real dataset of a major Czech bank.

JEL classification: G21, G28, C14

Keywords: recovery rates, loss given default, retail lending, survival analysis

1. Introduction

1.1. Motivation and literature review

Basel II Accord requires banks to calculate the capital requirement by the Standardized approach[1] or by the Internal Ratings Based (IRB) approach[2]. The possibility as well as the need of computing bank capital requirements invoked the demand for advanced modelling approaches and better models of credit risk.

A very good overview on modelling of the credit risk models was provided by (Altman, 2006). The paper presents a thorough guide to historical development of a large number of different approaches to credit risk modelling as such, or its components, Probability of Default (PD), Loss Given Default (LGD) and Exposure at Default (EAD), respectively.

The models here were divided into two basic groups, (i) those treating credit pricing, and (ii) portfolio credit Value at Risk (VAR) models. Apart from the review, the major contribution of the paper lies in an analysis of the relationship between PD and recovery

[*] Corresponding author at Department of Control Engineering, Faculty of Electrical Engineering, Czech Technical University in Prague, Technická 2, 166 27 Praha 6, Czech Republic, tel.: +420 776 697 672.

Email address: samuel.privara@fel.cvut.cz (Samuel Prívara)

[1] In the standardized approach, each asset class is assigned a prescribed risk weight.

[2] The IRB approach enables banks to compute the components of the credit risk to evaluate the capital requirement using their own estimates.

rates (RR or LGD, see the discussion below) as these were proven to be correlated (Acharya et al., 2007; Miu and Ozdemir, 2006).

Modelling of PD has been thoroughly studied for some time, however, LGD has not received that kind of attention, yet. This was caused, as pointed out by (Zhang and Thomas, 2012), due to (i) censored character of the data, which are not easily handled by classical regression techniques, and (ii) different reasons, why the debtors defaulted, which in turn leads to a different repayment patterns.

Because of the above mentioned reasons, we shall pay a special attention to LGD in the following. LGD is the incurred loss when an obligor defaults on a loan, given as the fraction of EAD unpaid after some period of time. The other definition by Commision (2006) puts LGD as the ratio of the loss on an exposure due to the default of a counterparty to the amount outstanding at default. Very often, a complementary variable recovery rate ($RR = 1 - LGD$) can be, using net cash flows (CF) at time t_i (i.e. payments from the debtor net recovery process costs) defined as

$$RR = \frac{1}{EAD} \sum_{t=t_1}^{T_{max}} \frac{CF_t}{(1+r)^t}, \tag{1}$$

with r being an appropriate risk-adjusted interest rate used for discounting.[3] It is usual to assume (but not necessary) that $LGD \in [0, 1]$, where 0 means that the balance is fully recovered ($RR = 1$) and 1 means the total loss of EAD ($RR = 0$).

There is quite a large number of LGD and RR modelling applications. The paper by (Schuermann, 2004) was one of the first reactions on the Basel II Accord and its influence on LGD modelling. The paper tries to answer to questions such as (i) what does LGD mean and what is its role in IRB, (ii) how is it defined and measured, (iii) what drives the differences in LGD, and (iv) what approaches can be taken to model or estimate LGD. Even though the paper provides a very nice introduction to the problematics, it does not treat the problem of creating the model for predictions of LGD whatsoever.

A study on the complementary measure, RR was provided by (Qi and Zhao, 2011) where the determinants of creditor recoveries from defaulted debt instruments were investigated. Instead of the traditional techniques such as Ordinary Least Squares (OLS) regression, a Fractional Response Regression (FRR) and Non-Parametric Regression Trees (NPRT) were applied to a portfolio of corporate loans. While the former method (especially useful for computation of RR as it is intended for continuous variables bounded in the interval [0; 1]) was able to provide quite consistent results on both training and validation data samples, the latter was performing worse on the validation data set, which was however rectified by applying a model complexity control.

LossCalc model by the Moody's KMV (Gupton et al., 2002) is a model for predicting LGD for US bonds, loans and preferred stock. The model produces estimates of LGD for defaults occurring immediately and for defaults occurring in one year, thus the two point-in-time estimates are used to predict LGD over the holding periods. For prediction of LGD, LossCalc incorporates information on instrument, firm, industry, and economy. As far as the

[3] The common banking practice is to use the actual facility interest rate at the time of default. However, a more consistent approach is to define the discount rate as the risk-free rate with maturity corresponding to average recovery process duration plus a risk marging reflecting uncertainty of the recovery rates.

used technique is concerned, LossCals transforms LGD (or RR) to a normal variable using a Beta distribution, then a regression is run on the mentioned variables, and finally, the inverse transform is applied.

The sector of credit cards including the macroeconomic variables was treated by (Bellotti and Crook, 2012). The authors exercised several models such as those based on account level data, including Tobit, a decision tree model, a Beta and fractional logit transformation. However, the best results were achieved by the OLS.

The modelling of LGD for commercial companies was detailed by (Grunert and Weber, 2009). The study is focused on German companies where the analyses on the distribution of RRs and the impact of the quota of collateral, the creditworthiness of the borrower, the size of the company and the intensity of the client relationship on the RR were examined. The paper, however, treats mainly the influence of the corresponding covariates on the response variable. The problem of constructing the predictor is again missing.

When modelling LGD, a bi-modality is often observed alongside with boundedness of the distribution. The former is quite intuitive as the debtor (when default occurs) can either recover and repay the debt only with minor losses for creditor due to administrative costs linked to the recovery process, or he does not recover and the losses are huge creating a characteristic U-shaped distribution. The bi-modality was treated by introducing the mixture of two beta distributions by (Hlawatsch and Ostrowski, 2011) where Expectation Maximization (EM) algorithm (Dempster et al., 1977; Moon, 1996) was used to estimate the parameters of the distribution mix.

The field of consumer loans was targeted by (Zhang and Thomas, 2012). The paper addresses the problem of LGD modelling by two different classes of methods, namely single distribution models and mixed distribution models. The former class includes well know approaches such as linear or logistic regressions, or survival analysis modelling, Cox regression among others. The latter class is aimed at improving the specific features of LGD modelling, such as U-shaped distribution by introducing so called FM models where the effort is put to identify the specific group of data, to ungroup them and to analyse them separately. The popular techniques here are, e.g., classification and regression trees, group classification with sequel linear or logistic regression, or survival analysis applied to the individual groups of data. Surprisingly, the best results were recorded by the linear analysis (even though some of the observations were censored).

1.2. Contribution of the paper

LGD is one of the pivotal parameters for computation of the expected and unexpected credit losses needed for credit pricing and to comply with the regulatory requirements. The credit rating and PD have been studied for a long time and the estimation techniques are well developed, the situation with LGD is different. One of the main issues when estimating LGD is lack of data. Not only the short time series, but also only partially available, i.e. censored data, represent serious difficulties for standard estimation techniques such as OLS or logistic regression.

The situation can be solved by applying the statistical technique of survival time analysis which allows to utilize censored data. Survival analysis was relatively extensively used for estimation of PD (see e.g. Narain (1992); Andreeva (2005)), however, its application to estimate LGD is rather scarce even though some exception apply, e.g. Witzany et al. (2010); Bonini and Caivano (2012); Zhang and Thomas (2012).

In this paper we provide a comparison of the standard estimation techniques (OLS, logistic regression) and some modifications of survival analysis methods applied to a real data from the environment of a major Czech bank.

We also aim to compare our methods and results with Zhang and Thomas (2012) where survival analysis-like approaches were applied. In that study, the authors also use Cox proportional hazard model but in a substantially different way than we do. In contrast with this paper, the authors assume the RR being the time of exit T. Thus, the more is collected (in RR sense), the longer the survival period T which implies that better debts are attributed with higher T, and conversely, worse quality debts are attributed with lower T. That approach also leads to treatment of the debt as a one item which might not be as sensitive as the precise unit-by-unit Cox analysis we conduct. The perception of time also differs in Zhang and Thomas (2012) and our paper. The Cox approach the authors suggest in fact dampens the time factor since the time is regarded as a RR. In our paper, the treatment of time is more explicit. Because the time is a key component of survival analysis the achieved results can be (and also are) substantially different. Last but not least, output of the Cox approach in the aforementioned study results into a *distribution* of RR for each debt and so an optimal quantile must be picked to return unbiased estimates of RR. In our paper a direct estimate of RR or a probability that RR for a given debt is a member of low/high RR group, is returned.

1.3. Structure of the paper

The paper is further structured as follows. The methodology used for estimation RR/LGD is discussed in detail in Section 2. The performance of the respective methods is evaluated on the real data from a major Czech bank in Section 3. The last section drafts possible improvements and concludes the paper.

2. Methods and models

2.1. Methodology

The market RR for the instruments that are traded on the market with satisfactory liquidity, can be computed as the market value out of the principal (plus coupon accrued at default) of the security achieved after a defined period after default, see, e.g., Witzany et al. (2010). For other instruments, the computation of RR can be performed according to Eq.(1), i.e. it is based on a work-out process. Sometimes, the recovery rate can be negative due to workout process costs and negligible or no recovered amounts from the debtor. The other case, $RR > 1$ may be caused by large late fees that are paid in full by the debtor. To mimic the market RR which is rarely outside $[0, 1]$ we constrain ourselves to this interval. Note that this adjustment simplifies the survival analysis that, in fact ,needs to work with non-negative values.

In order to achieve nonnegative recovery rates and cash flows, the CF from Eq.(1) will be adjusted as follows:

$$CF_i = max(0, repayment_i - cost_i). \tag{2}$$

At the same time we disregard cash flows CF_i, or their parts, that would make the recovery rate larger than 0.

Up to this point we have not differentiated between the observed RRs and the ultimate realized RR. To systematically build-up a model which predicts LGD (RR) for non-defaulted

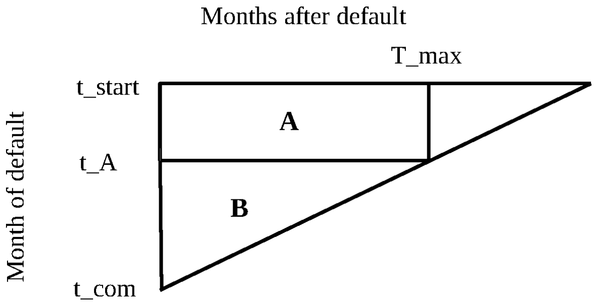


Figure 1: Concept of the data availability

accounts based on observed RRs and certain specific characteristics, we have to realize, that the full information set is not available for all cases under investigation at a given moment as the work-out process might not have been completed. The recovery CF is typically measured only over some defined maximum period as within this period, the debt is usually either collected, sold or written-off. Further on, we denote the maximum number of months for which recoveries are observed as T_{max}.

We will illustrate the whole concept in Fig. 1 where individual entries represent observed monthly recoveries of defaulted loans on individual or aggregate level. The horizontal axis depicts the number of months after default, whilst the vertical axis depicts the month of default with t_{com} denoting the observation time when the computations are done. T_{max} is shown on the horizontal axis and t_{start} on the vertical axis represents the month when the observations started and the first default occurred. Analogically, the time span between t_{start} and $t_A = t_{com} - T_{max} + 1$ includes all the defaults that begin within this period and ended (written-off, sold or repaid) until T_{max}. This rectangular area (denoted as A) covers all the cases, where the full information (in the horizon of length T_{max}) is available, i.e. the full realized recovery rates can be computed. The area denoted as B contains the cases where only the partial information is available (with respect to the moment of computation t_{com} the work-out process is still not finished), thus only partial realized recovery rates can be computed. The real life situation is that only the minor part of the collected cases belong to the area A, i.e. have the complete set of information. For majority of cases, only partial information is available hence the ultimate RR can not be computed. Further on, A will also denote the set of observed loan accounts where the recovery process could be observed for at least T_{max} months, B will denote the set of accounts where the observation period is shorter than T_{max}.

This concept is pivotal to understand when choosing the suitable modelling technique tools such as OLS/GLS that can be applied only on the cases from the area A, while survival analysis techniques can use even the partial information from the area B.

2.2. *Linear and logistic regression models*

Linear and logistic regression techniques are well described in literature, see e.g. Menard (2001); Draper et al. (1966) and belong to popular choices for financial applications.

A linear regression model can be expressed as

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon, \tag{3}$$

where y represents LGD or RR, β_i are unknown parameters to be searched for, x_i are independent variables, characterizing the specific debt case and ε is the error term. The key assumption is that the error of the estimation follows the white zero mean Gaussian noise. Then the estimated output $\hat{y}$ is obtained as a result of the optimization task minimizing the sum of weighted squared errors $\sum w(y - \hat{y})^2$ tuning parameters β_i.

The logistic regression belongs to the class of Generalized Linear Model (GLM) regression which is a special class of nonlinear regressions using linear methods. Note that the outcome variable now acquires only binary values $\{0; 1\}$, "dead" or "alive", or for our specific case "low" and "high" recovery rates as will be explained later. Using the notation from the linear regression we have

$$\begin{aligned} y = f(\pi) &= \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon, \\ f(\pi) &= \ln\left(\frac{\pi}{1-\pi}\right), \end{aligned} \tag{4}$$

where y is linear predictor that is equal to the link function $f(\pi)$ and π is a mean of the corresponding distribution (binomial for our case).

The whole procedure is performed in three steps

- **Transformation of RR.** Because the measured RR are from the interval $[0, 1]$ and for the logistic regression only the binary data are needed, we will introduce a threshold function f_T which transforms the continuous RR to binary RR as follows

  $$f_T(RR) = \begin{cases} 0, & \text{if } RR < RR_T \\ 1, & \text{if } RR \geq RR_T \end{cases} \tag{5}$$

  with RR_T being a chosen threshold value.

- **Logistic regression.** The parameters of the model defined by Eq.(4) are computed e.g. by Newton-Raphson algorithm Ostrowski (1960) or by standard solvers for generalized linear models.

- **Inverse transformation of RR.** The final estimate of recovery rate is computed as

  $$\widehat{RR} = RR_H\hat{\pi} + RR_L(1 - \hat{\pi}), \tag{6}$$

  where $\hat{\pi}$ is the estimated inverse link function from Eq.(4) and

  $$RR_H = \frac{\sum\limits_i EAD_i RR_i}{\sum\limits_i EAD_i}, i : RR_i \geq RR_T, \tag{7}$$

  with RR_L is defined analogously.

2.3. Survival analysis models

Survival analysis methods (SAM) have two properties which make them superior from the theoretical point of view to the classical regression methods. As mentioned before, the classical regression methods need the ultimate information set. Since the debts where the work-out process has not been finished yet can not be included into the regression we are losing a part of the information. SAM, on the other hand, enables us to treat the partial information as censored and utilize thus all information available at the point of observation. As pointed out by Zhang and Thomas (2012), the other reason is the critical assumption on normality of the distribution of residuals. As mentioned several times, this assumption does not hold in case of modelling recovery rates, therefore, there are fundamental problems with applying the classical regression. In contrary, SAM enables us to create model using an abundant number of distributions – Li (1999); Hosmer et al. (2011); Bellotti and Crook (2008), even empirical Cox (1972); Draper et al. (1966).

The basic principle is as follows. Given T is a random variable as an instant of exit[4] of an object, its statistical properties can be expressed by the probability density function $f(t)$ and the cumulative distribution function $F(t)$ with the standard statistical meaning. Then the survival function $S(t) = 1 - F(t)$ represents the cumulative probability of the object being still alive until time t. Hazard function, or hazard rate is defined as

$$h(t) = \frac{f(t)}{S(t)} \tag{8}$$

and represents the rate at which the objects are exiting exactly at t given survival until t. Then the cumulative hazard function, using an equivalent concept as in case of $F(t)$, is defined as

$$H(t) = \int_0^t h(s)ds = -\ln(S(t)). \tag{9}$$

So far, the general remarks regarding SAM were given. Among the most popular implementation belong accelerated failure time models, or Cox proportional hazard models to which we put special attention in the next section.

2.3.1. Cox proportional hazard regression

Cox proportional hazard models implicitly enable bigger flexibility than other types of models. Cox (1972) introduced the following parametrization of the hazard function Eq.(8)

$$h(t, x) = h_0(t)e^{x^T\beta}, \tag{10}$$

with β being the vector of unknown parameters and x being the covariates (explanatory variables). Note that h_0 (often called as baseline hazard) is independent of parameters and is a function of time only. Following Eq.(9) we can finally parametrize survival function as

$$S(t, x) = e^{-H(t,\beta)} = e^{-\int_0^t h(s,\beta)ds} \tag{11}$$

or equivalently

$$S(t, x) = (S_0(t))^{e^{x^T\beta}}, \tag{12}$$

where $S_0(t) = e^{-\int_0^t h_0(s)ds}$ is the baseline survival function dependent only on time. The vector of parameters β is estimated by the maximum likelihood method using standard statistical packages.

[4] Exit is very often referred to as death. The opposite term is survival, i.e. all the object observed up to some time which did not exit are still alive, i.e. survived.

2.4. Censoring

A key question when dealing with Cox regression analysis is how to handle the incomplete information. We have already indicated in Section 2.1 how the data can be viewed upon. Now, we explain the censoring within the Cox regression framework as there are several possibilities what and how to censor.

- **Cox regression with censoring of RR**. The key measure to observe is RR defined in a specific time span. Let $a \in A$ be a default case from Fig. 1. Then $CF(a, t_i)$ will denote the corresponding discounted cash flow at time instant t_i, with $\{t_i : i = 1 \ldots, n\}$ being the moments of payments since the time of default. The transformation of the original data into a Cox-regression-ready dataset is depicted by the simplified algorithm 1.

Algorithm 1: Default cases from dataset A

```
Input: $\forall a \in A$, $t(a) = [t_1, \ldots, t_n]$ is the vector of times of payments that are $\leq T_{max}$,
       and $rr(a) = \sum_{i=1}^{n} CF(a, t_i)/EAD(a)$
Output: Censored and weighted data prepared for Cox regression
1  if $rr(a) < 1$ then
2      %censoring at $t_n$
3      $censor(a) = [0, \ldots, 0, 1]$;
4      %how much remained unpaid
5      $CF2pay = EAD - \sum_{i=1}^{n} CF(a, t_i)$;
6      $f_{weight} = [CF_1, ...CF_n, CF2pay]$;
7      %response variable y
8      $y = [t_1, \ldots, t_n, T_{max}]$;
9  else
10     %$rr(a) = 1$
11     $censor(a) = [0, \ldots, 0]$;
12     $f_{weight} = [CF_1, ..., CF_n]$;
13     $y = [t_1, \ldots, t_n]$;
14 end
```

Here, the pivotal idea to realize is that each payment is perceived as a single frequency weighted case with the weights corresponding to the cash amount recovered at that time instant. Then each payment can be marked by a yes/no censor label. A payment case can be censored due to two reasons: i) the sum of the payments received up to the time instant of the case under consideration is less than debt amount, i.e. the debt was not yet payed-off (or at the time instant under consideration, the defaulted receivable was written-off or sold), i.e. there remains some part of it to be paid and this is the censored amount, or ii) the work-out process lasted longer than is its maximal allowed length T_{max} and the amount collected until that moment is less than the EAD. Note, e.g., on line 6 of algorithm 1 that each payment within the defined time interval represents the frequency weighting. The last payment $CF2pay$ corresponds to the unpaid amount $EAD(a) - \sum_{i=1}^{n} CF(a, t_i)$ censored at the very last moment of the defined time interval. The algorithm for debt cases from B proceeds in a very same fashion as 1 but T_{max} is replaced with $t_{com} - D_{beg}$, where D_{beg} denotes the time of default. In general, we can define the observation time horizon as $T_{obs} = \min(T_{max}, t_{com} - D_{beg})$ and use it instead of T_{max} in both cases.

- **High-low Cox regression (HLCR) with censoring of RR**. The preparation of data for HLCR is in principle different from the one described by algorithm 1. The debt cases are now perceived as single entities, i.e. the respective payments can not be treated separately, i.e. can not be censored individually as only the account (debt case) as such is treated in this way, and the weight is always the EAD.

  The key principle is the following. Choose[5] a threshold recovery rate RR_T, for example 90%. Then each account that has $RR < RR_T$ is not finished (has survived and $censor = 1$,) while all the others ($RR \geq RR_T$) are "dead" for the subsequent HLCR ($censor = 0$.) The exit or censoring time is set equal to the last observed month (at most T_{obs}) or to the month when the minimum recovery RR_T was achieved.

- **The Cox implementation in Zhang and Thomas (2012)** was already outlined in Section 1.2. Since the model's conception of time is different, censoring also differs from the above mentioned approaches. The authors basically classify accounts into either finished (i.e. written off) or unfinished, where the collection process is still underway. For better explanation we use death/survival analogy. Observations of the accounts which are written-off are uncensored, which in the analogy means the patient has been seen to die. The time of exit is set equal to the realized recovery rate. On the other hand, observations for the accounts not written-off are censored because either i) the collection process has not been finished yet, or ii) it is the case when RR = 1 (and thus the patient has completely recovered). The time of censoring is set equal to the realized recovery rate. Of course, we must also take into account the information availability in the sense that we must respect T_{obs} in any case. This means that an account written off at some time $T > T_{obs}$ will be treated as censored.

3. Case study

3.1. Data

The data used for investigation of LGD/RR modelling approaches origin from a major Czech bank and relate to defaulted unsecured retail loans. We have been provided with $10,000$ randomly chosen accounts with events measured over 10-year period starting in 2000. Respective defaults occurred between the end of 2001 and the end of 2003, i.e. all the defaults started within 24 months. Monthly net account level repayments were alreadz discounted by the bank based on the facility rates. Because the observation horizon is relatively short, we do not take into account any macroeconomic background. The repayment story is not homogeneous as some defaults (618) where the recovery process ended already after one month being in default, while others had much worse repayment discipline (the longest with the recovery process accounting for 126 months). The histogram of recovery lengths of defaulted loans is depicted in Fig. 2 with an average being 20 months.

[5] The choice of the threshold recovery rate RR_T results from optimization with respect to R_2. In our case a brute force was used, when we examined all the recovery rates between 0 and 1 with step 0.05

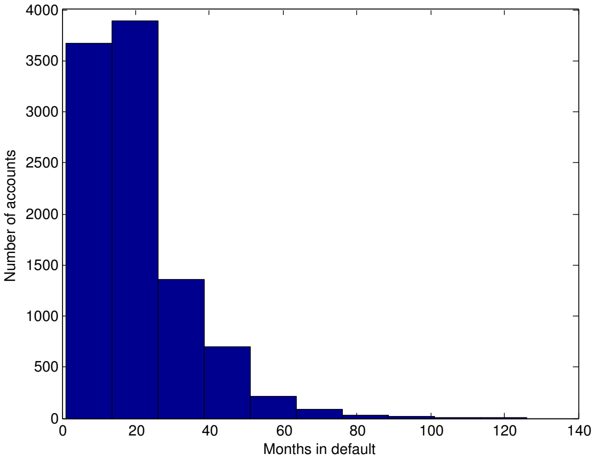


Figure 2: Histogram of recovery lengths of defaulted loans in months

There are 8 demographic variables at hand, namely the borrower's sex, age, marital status, education, employer, kind of employment and housing status. All 8 variables are categorical. Apart from the demographic data, several other characteristics are available, namely EAD, month where the account defaulted, end of the recovery month, default end status and the loan limit. The average EAD is around $37,800$ CZK, median $26,800$ CZK and the maximal exposure almost 1.5 million CZK. Note that 90% of the loans have exposure less than $86,000$ CZK. The other characteristics, such as ratio of EAD and limit of the loan were computed. As the provided payments as well as the costs related to repayment process were already discounted, the realized CF could be computed as a simple sum with an exception to non-positive flows that were adjusted according to Eq.(2).

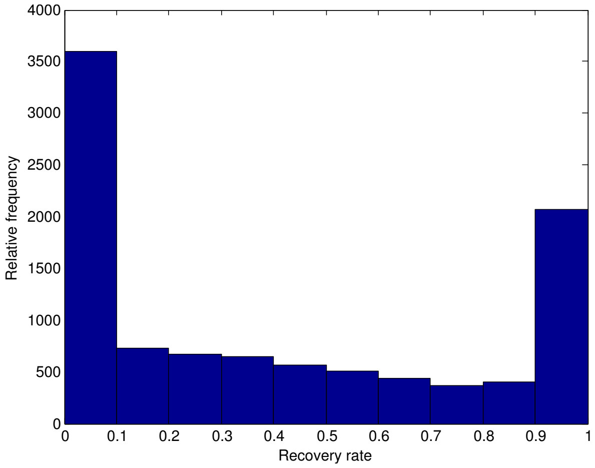


Figure 3: Distribution of the recovery rates.

Distribution of the recovery rates is given in Fig. 3 with the average of 0.4. Note that the data were adjusted to fit within the interval $[0, 1]$ as we apply the RR floor 0 and cap 1. Let us now continue with properties of the analyzed data. 28% of the cases have a 0% recovery rate, while 15% of the cases fully repaid the debt. The distributions of the recovery rates for datasets A and B (see Section 1.2 for definition od these sets), respectively, expressed in absolute values are depicted in Fig. 4. The U-shape distribution is in-line with many other research papers Witzany et al. (2010); Zhang and Thomas (2012).

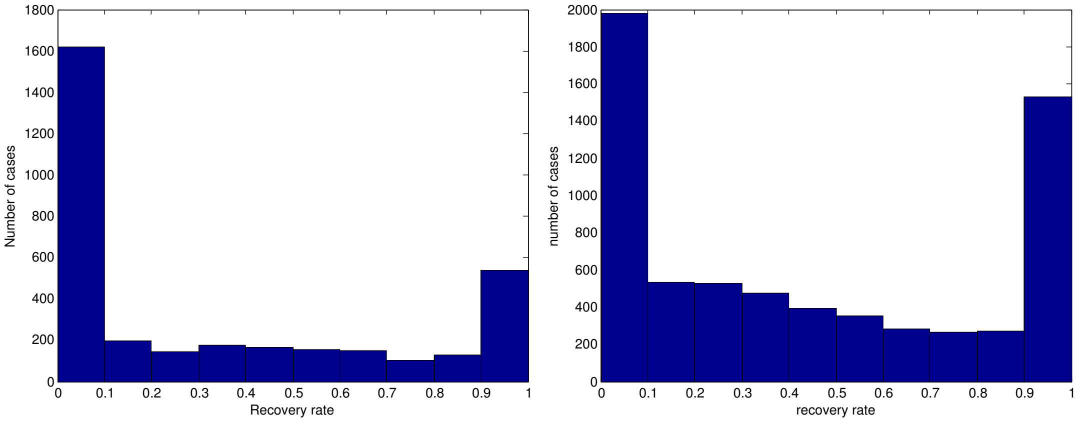
3.2. Modelling approaches

Over 70% of accounts will finish the work-out process within 24 months. Moreover, as stated above, all the debt cases have defaulted within 24 months. To comply with Fig. 1, T_{max} in algorithms from Section 2.4 was chosen to be 24 as well.

A model for linear and logistic regressions was built with recovery rate as an exogenous variable and all the demographic variables from above with addition of EAD and a ratio $EAD/limit$ as explanatory variables (covariates). As for linear regression, two methods were tested, namely i) classic linear regression, and ii) step-wise selection procedure with linear terms in the model. Logistic regression is performed on two basic models, namely i) logistic regression with linear terms only, and ii) logistic regression with linear terms and products of the covariates.

Cox regression with censoring of RR essentially divides the exposure (EAD) into individual monetary units (crowns) and models their survival in time, i.e. models course of repayment of each individual unit with respect to the month in default. Those units that survived (no exit, i.e. no repayment) until some time (T_{max}) were marked as censored. The modelling outcomes provide the model parameters β and the baseline survival function $S_0(t)$ according to Eq.(12). $S_0(t)$ is dependent on nothing but time, while the account specific survival function $S(t, a)$ of Eq.(11) depends on both time and parameters β related to a specific account. The baseline survival function for classic and HLCR regressions are depicted in Fig. 6(a) and Fig. 6(b), respectively.

The survival functions for two accounts with different characteristics (hence different β) are presented in Figs. (7(a) – 7(b)). In the first case (see Fig. 7(a)) $S(T_{max}, 34) = 0.6004$, while in the second case $S(T_{max}, 340) = 0.2907$. The interpretation is as follows. The first account has repaid almost 40% of the debted amount within the observation period (default start, T_{max}). The rest of the debt "survived", i.e. was not repaid. For the second case, the situation is different. This account has significantly different characteristics which results in



(a) Dataset A

(b) Dataset B

Figure 4: Distribution of the recovery rates

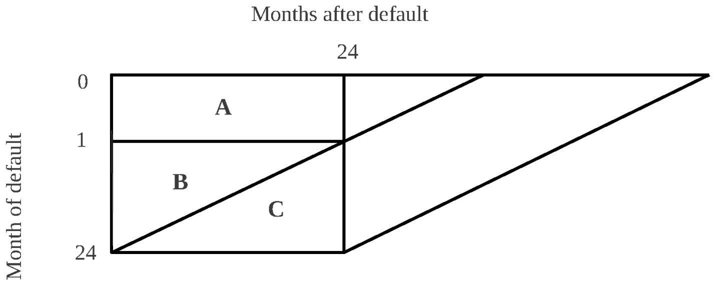
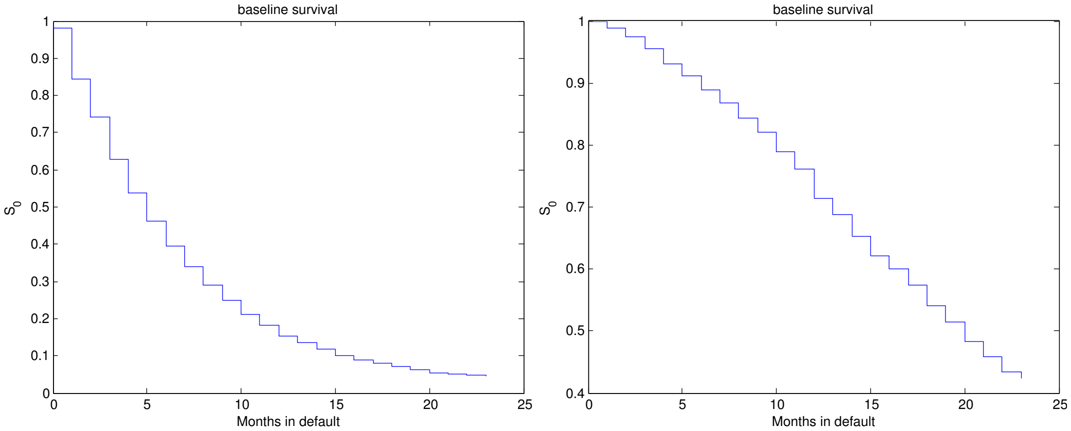


Figure 5: Full and partial information dataset

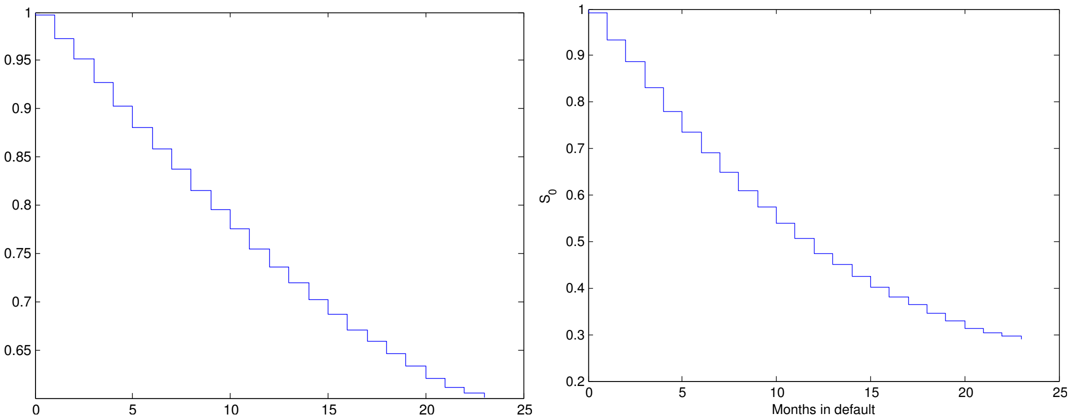


(a) Cox baseline survival

(b) HLCR baseline survival

Figure 6: Baseline surival functions

a different repayment pattern and more than 70% of the debt being repaid ($RR(T_{max} - 1) = 1 - S(T_{max}, 340)$).



(a) Survival function for account 34

(b) Survival function for account 340

Figure 7: Cox regression: Survival functions

The input data for HLCR has been censored as described in Section 2.4, i.e. the high-low RR principle was used. The outcome of the modelling procedure is again (as in the previous case) baseline survival function S_0 and the survival function S depending on both time and parameters. The interpretation is however different for this case. Because the input data was in high-low form, the resulting survival function S must be transformed back in a similar way as described for logictic regression inverse transformation of RR, i.e. we need to compute RR_H according to Eq.(7) (and RR_L analogously) and then the estimate of recovery rate can be formulated as

$$\hat{RR}_{HLCR} = S(T_{max})RR_L + (1 - S(T_{max}))RR_H, \tag{13}$$

with $S(t)$ being the survival function estimated for HLCR.

3.3. Results

The quality of the models can be described by the R^2 coefficient (coefficient of determination) defined as

$$R^2(x) = 1 - \frac{\sum_i (x_i - \hat{x}_i)^2}{\sum_i (x_i - \overline{x})^2}, \tag{14}$$

with $\hat{x}$ denoting the estimate of x and $\overline{x}$ the mean of x. The coefficient is closely related to the "fit factor" (see e.g. MATLAB (2012)) or normalized root mean square error (NRMSE) Lee (2011)

$$NRMSE = \sqrt{\frac{\sum_i (x_i - \hat{x}_i)^2}{\sum_i (x_i - \overline{x})^2}}, \tag{15}$$

with the same meaning of the variables as in Eq.(14). In most of the technical as well as financial application the R^2 of the model exceeds 80 or 90%, however, in modelling RR/LGD the numbers are significantly lower and usually around 3 – 12%, (see e.g. Zhang and Thomas (2012); Bellotti and Crook (2009, 2008); Matuszyk et al. (2009)).

A practical implementation of the Cox algorithm does not divide the exposure into individual monetary units as described in the previous section, however uses the so called frequency weighting, which follows the respective payments of the specific account in time and in fact aggregates individual monetary units to a larger sums - respective payments being repaid in a discrete time instants. Hence Eq.(14) has to be slightly adjusted to reflect the weighting as follows

$$R^2(rr) = 1 - \frac{\sum_i w_i (rr_i - \hat{rr}_i)^2}{\sum_i w_i \left(rr_i - \sum_i w_i rr_i \right)^2}, \tag{16}$$

with $w_i = EAD_i / \sum_i EAD_i$ being the weight, rr_i observed recovery rate (actual recovery rate in the horizon T_{max}) and $\hat{rr}_i$ modeled recovery rate, i denotes i-th account. Recall that logistic regression uses weighting as well, see Eq.(7), which is then used for computation of R^2 according to Eq.(16).

[6] Mean Square Error

Table 1: Numerical results of the modelling approaches

Model	R^2[%]	MSE[6]	NRMSE
Linear	8.15	0.137	0.958
Linear step-wise	8.89	0.136	0.955
Logit linear	10.46	0.109	0.946
Logit products	10.45	0.110	0.946
Classic Cox	15.21	0.103	0.903
HLCR	14.48	0.105	0.925
Cox (Zhang, Thomas)	3.23	0.154	0.982
Cox (Zhang, Thomas)*	4.52	0.150	0.975

(*) denotes the model with a grouping technique

The models used within the linear and logistic regression framework were built-up using data from dataset A, see Fig. 5, while survival models were able to utilize data from dataset B as well. Since the data from dataset B carry only a partial information the former group of methods is incapable of incorporating this information. The validation is performed on the full dataset without any limitation.

The results for both the classical regression methods (linear and logistic regressions) and the SAM are summarized in Tabs. (1–5), respectively.

Tab. 1 provides the R^2, MSE and NRMSE results for all the six tested methods (plus the method applied by Zhang and Thomas (2012)). It can be seen, that linear and logistic regression recorded results which are in-line with those reported by literature, see e.g. Zhang and Thomas (2012); Bellotti and Crook (2009, 2008); Matuszyk et al. (2009). Both regressions used categorical variables (denoted with subscript number). Step-wise selection of the variables recorded slightly better performance than the linear regression with all variables counted in. Tab. 2 and Tab. 3 report the estimates of the model coefficient with the corresponding p-values. Note that both Tab. 2 and Tab. 3 presents only those coefficient with $p < 0.05$. The former table stands for step-wise variable selection, the latter results from classical linear regression with all the covariates included. The estimate of coefficients for logistic regression are recorded in Tab. 4. The threshold RR_T was selected by brute force as $RR_T = 0.3$[7].

By using Cox regression formulated in Section 2.3.1 we obtained superior results comparing to all other methods. These results easily surpassed those obtained by Zhang and Thomas (see Table 4. and Table 8., respectively, presented in Zhang and Thomas (2012)). The coefficient estimates as well their p-values for both classic and HLCR regressions are recorded by Tab. 5. Note that all the variables are considered significant on 5% significance level for HLCR. Moreover, the optimal threshold for high-low determination of RR_T was selected by the same procedure as in case of logistic regression, here as $RR_T = 0.3$.

[7] The choice of the threshold recovery rate RR_T results from optimization with respect to R^2. In our case a brute force was used in the sense that we have examined all the recovery rates between 0 and 1 with the step 0.05

Table 2: Linear regression results: step-wise selection

Name	β	p
Intercept	0.5437	0
EAD/LIM	-0.1968	0
EAD	0	0.0001
Limit	0	0.0074
Sex_1	-0.0343	0.0373

Table 3: Linear regression explanatory variables

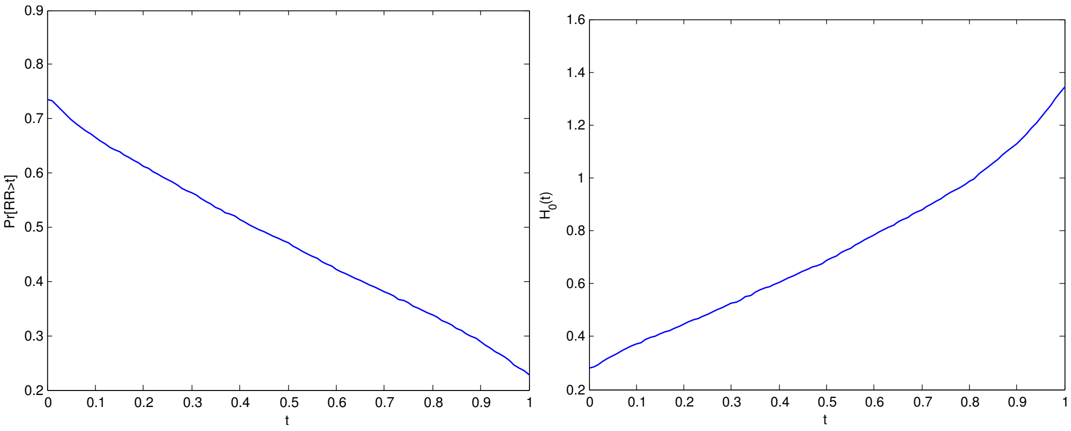
Name	β	p
Intercept	0.4739	0.014
EAD/LIM	-0.1899	0
EAD	0	0.000
Limit	0	0.021
Education_5	0.0579	0.017
Education_6	0.0878	0.048
House_3	-0.0502	0.084
House_6	-0.1718	0.045
Employment_7	-0.2934	0.023

Table 4: Logistic regression explanatory variables

Name	β	p
EAD/LIM	-0.4988	0.0088
EAD	0	0
Limit	0	0.0001
Education_5	0.2576	0.0399
Education_6	0.668	0.0032
Empl.since_9	2.3313	0.0461
Employer_8	-0.3138	0.0363
Housing_3	-0.3272	0.0311
Employment_4	-1.5044	0.0331
Employment_7	-2.1959	0.0034
Employment_8	-2.1331	0.0089
Employment_{10}	-1.3943	0.0435

Table 5: Cox regression explanatory variables

Name	$\beta_{classic}$	$p_{classic}$	β_{binary}	p_{binary}
EAD/LIM	-2.15E+0	0.00	-3.20E+0	0.00
EAD	-1.93E-6	0.00	2.84E-5	0.00
Limit	9.23E-7	0.00	1.32E-7	0.00
Sex	-1.38E-1	0.00	-2.50E-2	0.00
Age	-2.89E-2	0.00	6.54E-2	0.00
Marriage	-3.95E-2	0.00	-7.75E-2	0.00
Education	3.19E-2	0.00	5.37E-2	0.00
Empl. since	2.49E-2	0.00	1.48E-2	0.00
Employer	-1.87E-2	0.00	-3.96E-2	0.00
House	-2.75E-3	0.00	4.78E-2	0.00
Employment	-1.17E-2	0.00	5.84E-2	0.00



(a) Distribution of RR for account 34

(b) Cumulative baseline hazard

Figure 8: Cox by Zhang and Thomas metohodology

Last but not least we have also conducted Zhang and Thomas-like analysis. Using their Cox-technique, one obtains a set of coefficients for Cox proportional hazard model and (cumulative) baseline hazard function. This baseline hazard function is related to the observation horizon T_{obs} and the t in $h(t, x)$ must be seen as recovery rate rather than time. The output of Cox model with this architecture returns probability distribution of RR, for every account. One such an example is given in Figure 8(a). Having the distribution we have to pick a quantile value t_α of this distribution where α represents probability and t_α is treated as estimate of RR. We are thus looking for some α (ceteris paribus) that maximizes $R^2 := R^2(\alpha)$. Such α must be searched for on an iterative basis. The model was tested in two different manners. The first one treats recovery (represented by time of exit) as a continuous variable. One can also observe, e.g. from Figure 8, that both the distribution function and cumulative baseline hazard function are nearly continuous lines. This is because exit times (represented by RR) were densely dispersed over [0,1]. The second one uses a grouping technique where recoveries (i.e. exit times) are placed into groups spaced by 0.05. Results of both variants are depicted in Table 1. According to the results shown, the second approach performed slightly better than the first one. The optimal α values were found to be 0.38 and 0.41 for the above specified models, respectively (similarly to Zhang and Thomas (2012)).

4. Conclusions

Due to an ongoing recession carrying higher than usual defaults as well as Basel Accord regulation, the modelling of the recovery rates has become an important part of the banking practice. Several regression methods for estimation of RR/LGD have been examined. Linear and logistic regression techniques, Cox regression using censoring of RR and high-low Cox regression were implemented and tested on a dataset of 10 000 non-secured consumer loans from a large Czech bank.

The usual experience from banking practice shows R^2 around or under 10% for estimation of RR. When using Cox regression (both the approach utilizing censoring of RR and HLCR), we have recorded superior results in comparison to the classical linear and logistic regression techniques. The Cox regression was even able to exceed 15% in R^2 for ex-ante predictions.

In addition to comparison of Cox regressions and standard tools (OLS, logistic regression) we have compared our implementation of Cox regression to the one presented in Zhang and Thomas (2012). We deem that the survival analysis treatment in Zhang and Thomas (2012) was outperformed by the methods we have proposed.

5. Acknowledgements

This research has been supported by the grant of the Grant Agency of Czech Republic (GACR) no. P402/12/G097 “Dynamical Models in Economics”.

References

Acharya, V., Bharath, S., Srinivasan, A., 2007. Does industry-wide distress affect defaulted firms? Evidence from creditor recoveries. Journal of Financial Economics 85 (3), 787–821.

Altman, E., 2006. Default recovery rates and LGD in credit risk modeling and practice: an updated review of the literature and empirical evidence. New York University, Stern School of Business.

Andreeva, G., 2005. European generic scoring models using survival analysis. Journal of the Operational research Society 57 (10), 1180–1187.

Bellotti, T., Crook, J., 2008. Credit scoring with macroeconomic variables using survival analysis. Journal of the Operational Research Society 60 (12), 1699–1707.

Bellotti, T., Crook, J., 2009. Calculating LGD for credit cards. In: QFRMC Conference on Risk Management in the Personal Financial Services Sector.

Bellotti, T., Crook, J., 2012. Loss given default models incorporating macroeconomic variables for credit cards. International Journal of Forecasting 28 (1), 171–182.

Bonini, S., Caivano, G., 2012. Beyond basel2: Modeling loss given default through survival analysis. Mathematical and Statistical Methods for Actuarial Sciences and Finance, 43–52.

Commision, E., 2006. Directive 2006/48/ec. Official Journal of the European Union 4 (27).

Cox, D., 1972. Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological), 187–220.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 1–38.

Draper, N., Smith, H., Pownell, E., 1966. Applied regression analysis. Vol. 3. Wiley New York.

Grunert, J., Weber, M., 2009. Recovery rates of commercial lending: Empirical evidence for german companies. Journal of Banking & Finance 33 (3), 505–513.

Gupton, G., Stein, R., Salaam, A., Bren, D., 2002. Losscalc: Model for predicting loss given default (LGD). Moody's KMV, New York.

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

Hosmer, D., Lemeshow, S., May, S., 2011. Applied survival analysis: regression modeling of time to event data. Vol. 618. Wiley-Interscience.

Lee, J., 2011. Advanced Electrical and Electronics Engineering. Vol. 2. Springer.

Li, D., 1999. On default correlation: a copula function approach. Available at SSRN 187289.

MATLAB, 2012. version 8.0.0.783 (R2012b). The MathWorks Inc., Natick, Massachusetts.

Matuszyk, A., Mues, C., Thomas, L., 2009. Modelling LGD for unsecured personal loans: decision tree approach. Journal of the Operational Research Society 61 (3), 393–398.

Menard, S., 2001. Applied logistic regression analysis. Vol. 106. Sage Publications, Incorporated.

Miu, P., Ozdemir, B., 2006. Basel requirements of downturn loss given default: modeling and estimating probability of default and loss given default correlations. Journal of Credit Risk 2 (2), 43–68.

Moon, T., 1996. The expectation-maximization algorithm. Signal Processing Magazine, IEEE 13 (6), 47–60.

Narain, B., 1992. Survival analysis and the credit granting decision. Credit Scoring and Credit Control. Oxford University Press: Oxford, 109–121.

Ostrowski, A., 1960. Solution of equations and systems of equations. New York and London.

Qi, M., Zhao, X., 2011. Debt structure, market value of firm, and recovery rate.

Schuermann, T., 2004. What do we know about loss given default?

Witzany, J., Rychnovsky, M., Charamza, P., 2010. Survival analysis in LGD modeling. Available at SSRN 1574452.

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