Survival Analysis in LGD Modeling[#]

Jiří WITZANY[*] *– Michal RYCHNOVSKÝ*[**] *–*
Pavel CHARAMZA[***]

Introduction

Loss Given Default (*LGD*) is one of the key parameters needed in order to estimate expected and unexpected credit losses necessary for credit pricing as well as for calculation of the regulatory Basel II requirement (BCBS, 2006). While the credit rating and probability of default (*PD*) techniques have been well developed in recent decades, *LGD* has attracted little attention before 2000s. One of the first papers on the subject (Schuermann, 2004) provides an overview of what has been known about *LGD* at that time. Since the first Basel II consultative papers being published there has been an increasing amount of research on *LGD* estimation techniques (see e.g. Altman – Resti – Sironi, 2004; Frye, 2003; Gupton, 2005; Huang – Oosterlee, 2008; etc.).

One of the issues financial institutions estimating *PD* and *LGD* face is lack of data. Besides the problem of short time series the most recent development is usually represented only by partial, i.e. censored data on defaults and recoveries. If default is defined as a legal bankruptcy or 90 days past due observed in the standard 12 month horizon then it is difficult to use data on loans granted during the last 12 months to predict *PD* for new applications. The problem is even more serious for *LGD* where financial institutions have started to collect data on recoveries from defaulted

[#] The article has been supported by the Czech Science Foundation grant no. 402/09/0732 *Market Risk and Financial Derivatives* and by the institutional support project VŠE IP10040.

[*] RNDr. Jiří Witzany, Ph.D. – associate professor; Department of Banking and Insurance, Faculty of Finance and Accounting, University of Economics, Prague, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic; <witzanyj@vse.cz>.

[**] Ing., Mgr. Michal Rychnovský, MSc. – doctoral student; Department of Probability and Statistics, Faculty of Informatics and Statistics, University of Economics, Prague, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic; <michal.rychnovsky@gmail.com>.

[***] Pavel Charamza – credit portfolio manager, Home Credit Asia N.V., Moravské náměstí 249/8 Brno-město, 602 00 Brno; <pavel.charamza@homecredit.asia>.

receivables in systematic manner relatively recently and moreover the recovery process usually takes up to three or even more years. Hence even if a bank observed recoveries on loans that defaulted in the past five years many or majority of *LGD* observations may be incomplete. It may be then difficult or impossible to estimate the *LGD* satisfying the regulatory requirements (BCBS, 2005) as well as the point-in-time *LGD* important for actual credit pricing that should reflect the most recent trends.

It is natural to apply the statistical technique of survival time analysis to model the probability of default. The technique allows to utilize censored default data as well as to model consistently probabilities of default in different time horizons. There is a relatively extensive literature on the subject (see e. g. Narain, 1992; Andreeva, 2006; Chava – Stefanescu – Turnbull, 2008) and the technique is used by some banks and practitioners. On the other hand with the exception of Rychnovsky (2009) there is no literature to the authors' knowledge on possible applications of the survival time modeling techniques to *LGD* modeling. This can be explained by the fact that the *LGD* estimation techniques are generally less developed and the interpretation of recovery data as time survival data is less straightforward than in the case of defaults.

The goal of this paper is to study possible applications of survival time analysis techniques, in particular the proportional Cox model and its modifications to *LGD* estimations. The methods are applied to real banking data and compared with more classical techniques like the linear and logistic regression. The definitions and methodological approach are outlined in Section 2, the empirical results are given and discussed in Section 3, and concluding remarks are made in Section 4.

1 Methodology

1.1 Recovery Rates and Loss given Default

First we need to specify the notions of realized (ex post) and expected (ex ante) Recovery rate (*RR*) and the complementary Loss Given Default (*LGD*). Realized *RR* can be observed only on defaulted receivables while the expected recovery rate is estimated for non defaulted receivables based on available information. The *RR* and *LGD* are expressed as percentages out of the exposure outstanding at default (*EAD*) and $LGD = 1 - RR$ is simply the complementary loss rate based on the recovery rate that is

usually less than 1. For market instruments like bonds or other debt securities we may define the market *RR* as the market value out of the principal (plus coupon accrued at default) of the security shortly (e.g. one month) after the default. Applicability of the definition assumes existence of an efficient and sufficiently liquid market for defaulted debt. For other receivables we have to observe the net recovery cash flows CF_t from the receivable generated by a work-out process. The work-out process may be internal or external where a collection company is paid a fee for collecting the payment on behalf of the receivable owner. The process may also combine an ordinary collection and sale of the receivable to a third party. In any case the work-out process involves significant costs that must be deducted from the gross recoveries. The net cash flows must be finally discounted with a discount rate r appropriately reflecting the risk (BCBS, 2005).

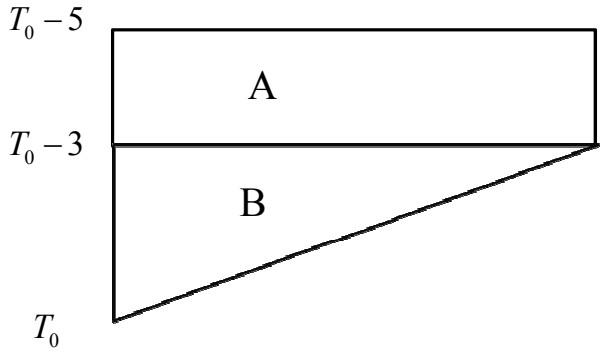
$$RR = \frac{1}{EAD} \sum_{i=1}^{n} \frac{CF_{t_i}}{(1+r)^{t_i}}., \tag{1}$$

The work-out recovery rate should in a sense mimic the market recovery rates. The relationship between the two ex ante notions is an analogy between the fundamental value and the market value of a stock. Hence the discount rate can be based on a measure of the *RR* systematic risk and a general price of risk (see Witzany, 2009). Since the market recovery rate is never negative and can be hardly larger than 1 we normally assume that *RR* as well as $LGD = 1 - RR$ lie in the interval [0,1]. The calculation of the work-out recovery rates according to (1) may however in some cases lead to negative values due to high costs and low or no recoveries, and on the other hand to values larger than 1 in the case of large and successfully collected late fees.

Having collected and calculated the realized recovery rates the next task is to estimate *LGD* for non defaulted accounts. In case of new loan applications banks need to estimate not only the probability of default (i) in the 12 month or longer horizon but also the *LGD* in the same horizon. The loan interest rate margin should cover the expected loss $PD \cdot LGD$ besides the cost of funds, administrative costs, minimum profit, etc. The ex ante *LGD* must be also calculated by banks applying the Advanced Internal Rating Based Approach (AIRB) in order to calculate the capital requirement for every non-defaulted receivable as defined by the Basel

(2006) regulation. Looking on the recovery cash flow data the typical situation may be illustrated by Figure 1.

Fig. 1: Ex post recovery data



The recovery cash flow finishes at time t_n from (1) if the past due receivable is fully collected, or the uncollected receivable is written-off abandoning further collection or due to a sale of receivables, or when the recovery time exceeds certain maximum time, e.g. 3 years. Hence if T_0 denotes the current time then the ultimate recovery rate information is systematically available only for receivables that defaulted between the time $T_0 - 5$ years and $T_0 - 3$ years, i.e. in the part A of Figure 1. Between $T_0 - 3$ and T_0, i.e. in the part B, the recorded recovery rate history will be for many receivables only partial. For example for receivables that defaulted 6 month ago, i.e. at $T_0 - 0.5$ only for a minority the collection process could have been finished due to a full repayment, sale of receivable, or a write-off caused by some legal reasons. For majority of the defaulted receivables there is only partial recovery history information and the ultimate result of the recovery process is not known. Consequently the decision to use, for the sake of ex ante estimations, the completed recoveries from the part B but discard the incomplete recoveries may cause a significant bias and an estimation error. So applying methodologies based on ultimate recoveries we should limit ourselves just to data from the part A. Such a dataset may be clearly insufficient in terms of number of observations and more importantly we are losing the information on recent developments that might be important in particular in times of a financial turmoil like the recent one.

Regarding the basic *LGD* estimation techniques we distinguish the pool level and account level estimations. The pool level estimations are designed for pools of receivables that are assumed to be homogenous in

terms of expected *LGD*, typically defined by product, collateral level, and other properties. For example we may observe realized recovery rates for unsecured consumer loans collected through a standardized internal process and estimate the expected *LGD* for non-defaulted unsecured consumer loans as one minus an average ultimate recovery rate observed in the part A of Figure 1. A more advanced approach is to try estimating expected *LGD* based on a set of explanatory variables, i.e. on specific properties of every non-defaulted receivable based on historical recovery rates and the observed values of the explanatory variables. We will go in this direction and compare classical linear and logistic regressions utilizing only the ultimate recoveries (part A, Figure 1) and the survival analysis techniques that can also consistently exploit the complete and incomplete recoveries in the part B.

1.2 Goodness of Fit Measures

Before we start analyzing various regression methods that could be applied to estimation of ex ante *LGD* let us specify our target in terms of appropriate goodness of fit measures. The goal is to find, based on available historical data, a function $\hat{L}(a) = F(\mathbf{x}(a))$ that gives predictions of the Loss Given Defaults based on given explanatory variables $\mathbf{x}(a)$ for any non-defaulted receivable a in the product class for which the function has been developed. The performance of the function should be measured only on receivables that default within the 12 month horizon from the estimation time. So if we develop the function at time T_0 on the data shown on Figure 1, optimally we need to calculate all the predictions based on covariates as of T_0, then wait 12 month to record the set D of all defaults in the observed class of receivables, and moreover wait up to 3 more years to obtain the realized $LGD(a), a \in D$. Given all the data we may finally calculate e.g. the *EAD* weighted R-squared as a standard goodness of fit measure:

$$R^2 = 1 - \frac{\sum_{a \in D} EAD(a) \cdot \left(LGD(a) - \hat{L}(a) \right)^2}{\sum_{a \in D} EAD(a) \cdot \left(LGD(a) - \mu \right)^2}, \qquad (2)$$

The indicator $R^2 = R^2(D, \mu)$ depends on the set of defaulted accounts used and on the mean μ. The *EAD* weighted mean of $LGD(a), a \in D$ would

be a standard choice but the logic of the measure is to compare the performance of an advanced prediction function with a basic *LGD* mean estimate that could be produced at the time T_0. However at that time we may calculate only the mean of ultimate *LGD*s in the rectangle part A of the historical data, hence further on we shall use $\mu = \sum_{a \in A} EAD(a) \cdot LGD(a) / \sum_{a \in A} EAD(a)$.

The indicator R-squared is a conventional econometric measure that has many technical advantages. Nevertheless it does not exactly fit the practical perspective of the *LGD* estimation users, i.e. banks and the regulators. The banks and the regulators will rather measure the absolute difference of realized losses and of the predictions (in currency units). The banks will not be happy if the predictions overshoot the real losses since the high predictions cause unnecessary capital requirements or too conservative prices. The central bank will not accept systematically low predictions reducing the capital requirement that should serve as a buffer against unexpected losses. Hence we propose to rather look on the modified R based on the absolute sum of differences:

$$\tilde{R} = 1 - \frac{\sum_{a \in D} EAD(a) \cdot \left| LGD(a) - \hat{L}(a) \right|}{\sum_{a \in D} EAD(a) \cdot \left| LGD(a) - \mu \right|}, \qquad (3)$$

Finally we have to consider feasible data sets at which the goodness of fit measures could be evaluated. To get the full out-of-sample measures as described above we would need at least 9 years of data, 5 years for the estimations and 4 years for the out-of-sample calculations. Since we have a shorter period of data we will have to use an in-sample or a mix between in-sample and out-sample approach. The first possibility is to evaluate the goodness of fit measures on the set A of receivables with ultimate recoveries. The measures would however clearly give an advantage to regression functions developed only on A not taking into account the data from the part B (Figure 1). Hence to get a fair goodness of fit measure we will assume that we know the ultimate recoveries of all the accounts in the part B. This can be achieved waiting sometime after T_0 until all the partial recoveries are completed, or retrospectively by using only a part of the historical data for the regression and remaining part to obtain the completed recoveries. Let B be the set of all receivables

in the part B of our development dataset and let $C = A \cup B$. The key goodness of fit measures we shall use will be $R^2(C,\mu)$ and $\tilde{R}(C,\mu)$ with the *EAD* weighted *LGD* mean μ calculated on the set A.

1.3 Linear and Logistic Regression

The simplest way to model *LGD* is to use the *OLS* regression $LGD(a) = \mathbf{x}(a)'\boldsymbol{\beta} + \varepsilon$, i.e. to search for the function L in the form $L(a) = \mathbf{x}(a)'\boldsymbol{\beta}$, $\mathbf{x}(a)$ containing the constant covariate 1, minimizing the sum of squared errors with the *EAD* weights on the given sample, i.e. looking for the coefficients $\boldsymbol{\beta}$ minimizing the expression $\sum_{a\in A} EAD(a)\cdot\big(LGD(a) - L(a)\big)^2$. The solution that can be expressed analytically by definition maximizes $R^2(A,\mu)$ but not necessarily $R^2(C,\mu)$ or $\tilde{R}(C,\mu)$.

The second possibility we will explore is the logistic regression based on the idea dividing the observed and future *LGD*s on "low" and "high" values. Let $l \in (0,1)$ be a threshold and define an LGD value to be "low" if $LGD < l$. Hence for $a \in A$ we have the indicator function $low(a) \in \{0,1\}$ and for a non defaulted receivables we want to find the logistic function

$$\pi(a) = \frac{\exp(\mathbf{x}(a)'\boldsymbol{\beta})}{1+\exp(\mathbf{x}(a)'\boldsymbol{\beta})},$$

estimating the probability that the loss will be "low" if the account defaults. To estimate the ex ante *LGD* we combine appropriately the *EAD* weighted mean of low observed *LGD*s and high observed *LGD*s, i.e.

$$L(a) = \pi(a)\cdot\mu_{low} + (1-\pi(a))\cdot\mu_{high},$$

$$\mu_{low} = \frac{\sum_{a\in A, low(a)} EAD(a)\cdot LGD(a)}{\sum_{a\in A, low(a)} EAD(a)}, \qquad (4)$$

$$\mu_{high} = \frac{\sum_{a \in A, \neg low(a)} EAD(a) \cdot LGD(a)}{\sum_{a \in A, \neg low(a)} EAD(a)}.$$

The vector of parameters $\boldsymbol{\beta}$ is obtained by maximizing the likelihood

$$L = \prod_{a \in A} \pi(a)^{low(a) \cdot EAD(a)} \left(1 - \pi(a)\right)^{(1 - low(a)) \cdot EAD(a)}.$$

The solution can be found numerically e.g. solving $\sum_{a \in A} EAD(a) \cdot \left(low(a) - \pi(a)\right)\mathbf{x}(a) = 0$ with the Newton-Raphson algorithm. The performance of the resulting function (4) may be tested for different values of the threshold l, e.g. 0.1, 0.2, …, 0.9.

1.4 Survival Analysis *LGD* Modeling

The survival analysis is appropriate in situations where we observe a population of objects that stay in certain state (survive) for some time until an exit (death or failure) happens. Typically some observations are censored, i.e. the objects are known to have survived until certain time but no more information is available. The goal is to study the time until failure and the probability of survival or failure in a given time period. In the case of defaulted receivables the idea is to consider the currency units or certain elementary amounts as the individuals that are in the collection process until they exit by a repayment.

The key survival analysis concepts (Greene, 2003, Kalbfleisch, Prentice, 2002, Collet, 2003) are the survival function and the hazard rate. Let T be the random variable representing the time of exit of an object, $f(t), t \geq 0$ its continuous probability density function, and $F(t)$ the cumulative distribution function. Then $F(t)$ is the probability of exit (failure) of an individual until the time t while the survival function $S(t) = 1 - F(t)$ gives the probability of survival until t. The hazard rate is defined as $\lambda(t) = \frac{f(t)}{S(t)}$. It gives the rate at which objects that have survived

until the time t and exit right at t; specifically $\lambda(t)\delta t$ is approximately the probability of exit in the time interval $(t, t+\delta t]$ provided the object is still alive at t. It is also useful to define the cumulative hazard function $\Lambda(t) = \int_0^t \lambda(s)ds$ as it can be seen that $S(t) = e^{-\Lambda(t)}$. If the concepts are applied to recovery data as indicated above then $F(t)$ corresponds to the expected recovery rate at time t, while $S(t)$ to the expected loss rate if the recovery process was terminated at t. The hazard rate $\lambda(t)$ corresponds to the incremental recovery rate or to the speed of recovery measured on the unrecovered amount at time t after default.

The models are specified through the hazard function given in a parametric or semi-parametric form. The parameters are moreover allowed to depend on explanatory variables characterizing the objects under observation. For example the parametric Weibull model is specified by

$$\lambda(t) = \lambda p(\lambda t)^{p-1},\ S(t) = e^{-(\lambda t)^p}\ , \tag{5}$$

while the Loglogistic model has the form

$$\lambda(t) = \lambda p(\lambda t)^{p-1} / [1+(\lambda t)^p],\ S(t) = \frac{1}{1+(\lambda t)^p}.\ , \tag{6}$$

The coefficient $\lambda = e^{-\mathbf{x}'\boldsymbol{\beta}}$ in both cases depends on the vector of covariates $\mathbf{x}$ (without the constant 1). The coefficients $\boldsymbol{\theta} = (\boldsymbol{\beta}, p)$ are estimated using a maximum likelihood method maximizing in general

$$\ln L(\boldsymbol{\theta}) = \sum_{\substack{\text{uncensored} \\ \text{observations}}} \ln \lambda(t \mid \boldsymbol{\theta}) + \sum_{\text{all observations}} \ln S(t \mid \boldsymbol{\theta})\,, \tag{7}$$

The two parameterizations can be formulated as accelerated failure time models where $\ln T = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$ and ε has a specific distribution.

The parametric models are attractive for their simplicity but may impose too much restriction on the structure of data. Fewer restrictions are imposed by the Cox (1972) proportional hazard model we shall focus on. The proposed hazard function has a semi-parametric form

$$\lambda(t, \mathbf{x}) = \lambda_0(t) \exp(\mathbf{x}'\boldsymbol{\beta}),$$

where $\lambda_0(t)$ is called baseline hazard function independent on the explanatory variables $\mathbf{x}$.

The baseline hazard is a step function estimated on a discrete set of points where exits or censoring take place. The corresponding survival function is in the form

$$S(t, \mathbf{x}) = \exp\left(-\int_0^t \lambda_0(s) \exp(\mathbf{x}'\boldsymbol{\beta}) \right) = S_0(t)^{\exp(\mathbf{x}'\boldsymbol{\beta})},$$
$$\text{where } S_0(t) = \exp\left(-\int_0^t \lambda_0(s) \right) \tag{8}$$

The coefficients $\boldsymbol{\beta}$ are estimated using the partial likelihood: if an object i with covariates $\mathbf{x}_i$ exits at time t_i, if we assume that there is only one exit at that time, and if A_i is the set of objects alive at t_i then the partial likelihood is

$$L_i(\boldsymbol{\beta}) = \frac{\lambda(t_i, \mathbf{x}_i)}{\sum_{j \in A_i} \lambda(t_j, \mathbf{x}_j)} = \frac{\exp(-\mathbf{x}_i'\boldsymbol{\beta})}{\sum_{j \in A_i} \exp(-\mathbf{x}_j'\boldsymbol{\beta})} \tag{9}$$

The coefficients $\boldsymbol{\beta}$ are then obtained maximizing $\ln L = \sum_{i=1}^{K} \ln L_i$ numerically using the Newton-Raphson algorithm. In general, in particular in the case of recovery process modelling, we need to handle ties, i.e. multiple exits at the same time. The partial likelihood function (9) can be generalized in a straightforward manner for the case of d_i ties (frequency weights) at the same time t_i. However due to computational complexity the exact partial likelihood function is usually approximated by an estimate due to Breslow (1974) or due to Efron (see Kalbfleisch, Prentice, 2002). Given $\boldsymbol{\beta}$ the baseline hazard function values are estimated separately for each of the unit time intervals where it is assumed to be constant maximizing the likelihood function

$$L_t = \prod_{i=1}^{n} \left[\lambda_0(t)\exp(\mathbf{x}_i'\boldsymbol{\beta})\right]^{dN_i(t)} \exp\left(-\lambda_0(t)\exp(\mathbf{x}_i'\boldsymbol{\beta})Y_i(t)\right),$$

where $dN_i(t)$ is an indicator of the fact that subject i died in the time interval $(t-1,t]$, and $Y_i(t)$ is an indicator of the fact that subject i is at the time $t-1$ still alive. The maximum likelihood estimator of the baseline hazard function is then in the Breslow-Crowley form

$$\hat{\lambda}_0(t) = \frac{\sum_{i=1}^{n} dN_i(t)}{\sum_{i=1}^{n} \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right) Y_i(t)}. \tag{10}$$

If there are no explanatory variables, i.e. $\exp(\mathbf{x}_i'\boldsymbol{\beta}) = 1$, then the estimator gives the estimate of the Kaplan-Meier hazard rate function and the corresponding Kaplan-Meier survival function.

To apply the survival analysis approach to recovery data we assume to have a set of defaulted receivables $a \in C$ and observed (discounted net) recovery cash flows $RCF(a,t) = \frac{CF(a,t)}{(1+r)^t}$ (see (1)) taking nonnegative integer values. The recovery time t is measured in month (or some other units) and takes only values in $\{1,2,...,K\}$, i.e. the maximum length of the recovery process is K month. The observed recovered amounts end at a time $t_{end}(a) \le K$. If $t_{end}(a) < K$ then the recovery process has been either successfully finished, or abandoned with a write-off, or the process has not been completed, but we have no more observations. Defaulted receivables with complete recovery history are marked by the indicator $fin(a) \in \{0,1\}$. If $t_{end}(a) = K$ then the recovery process is always considered to be complete, i.e. $fin(a) = 1$. Moreover for each receivable there is an initial exposure at default $EAD(a)$ again being a positive integer and a vector of explanatory variables $\mathbf{x}(a)$ (personal and/or behavior information). We assume that the cumulative recovery cash flow $CRCF(a,t) = \sum_{s=1}^{t} RCF(a,s)$ never exceeds the exposure at default. In particular the observed ultimate recovery rate

$RR(a) = RR(a, t_{end}) = CRCF(a, t_{end}) / EAD(a)$ (corresponding to (1)) will be always in the interval $[0,1]$. Finally, the survival time data set must contain not only the information on amounts that have been recovered but also the information on amounts that were not recovered. We will construct it as follows:

1. For every $a \in C, t \le t_{end}(a)$ with $RCF(a,t) > 0$ include an observation of $RCF(a,t)$ objects with covariates $\mathbf{x}(a)$ exiting at time t, i.e. censor = 0 (for exit) and frequency weight $d = RCF(a,t)$. This means that the amount of $RCF(a,t)$ was recovered at the time t.

2. For every $a \in C$ such that the recovery process is incomplete ($fin(a) = 0$) and $CRCF(a, t_{end}(a)) < EAD(a)$ include an observation of $d = EAD(a) - CRCF(a, t_{end}(a))$ objects with covariates $\mathbf{x}(a)$ censored at the time $t_{end}(a)$. This means that the amount of d has not been recovered until the time t, i.e. survived the time t with no future information (censoring).

3. For every $a \in C$ such that the recovery process is complete ($fin(a) = 1$) and $CRCF(a, t_{end}(a)) < EAD(a)$ include an observation of $d = EAD(a) - CRCF(a, t_{end}(a))$ objects with covariates $\mathbf{x}(a)$ censored at the time $t = K$. In this case we know that there were no recoveries until the last observation time and we have no more future information.

Having applied one of the parametric or semi-parametric survival models described above we get a survival function $S(t, \mathbf{x})$ and our final ex ante *LGD* estimate for a receivable a will be the survival probability

$$\hat{L}(a) = S(K, \mathbf{x}(a)),$$

i.e. the probability (given by the covariates of a) of a currency unit not being recovered until the maximal recovery time.

1.5 Pseudo Survival Models for *LGD*

The main advantage of the proposed application of survival models to *LGD* estimations is a consistent utilization of all available recovery data

including partial recoveries. On the other hand it appears that the maximum likelihood estimation approach used by the standard survival analysis model is a weak point with respect to the targeted goodness of fit measures, i.e R-squared and the modified R. Moreover the likelihood estimation (7) or (9) takes into account the sequence of all partial recoveries while the R-squared and modified R indicators measure performance of the predictions only with respect the ultimate recovery rates.

Our proposed modification is to use an appropriate survival model functional form $S(t,\mathbf{x}\,|\,\theta)$ and to fit the parameters θ not using MLE but simply minimizing an appropriate sum of squared errors. Similarly we could minimize a sum of absolute differences but the minimization would be generally numerically less efficient due to many singularities of the function to be minimized. Taking into account only the ultimate or last available recovery rates the *EAD* weighted sum of squared errors is

$$SSE(\theta)=\sum_{a\in C,\,fin(a)=1} w(a)\cdot EAD(a)\cdot\big(S(K,x(a)\,|\,\theta)-RR(a)\big)^2+ \qquad (11)$$
$$+\sum_{a\in C,\,fin(a)=0} w(a)\cdot EAD(a)\cdot\big(S(t_{end}(a),x(a)\,|\,\theta)-RR(a,t_{end}(a))\big)^2$$

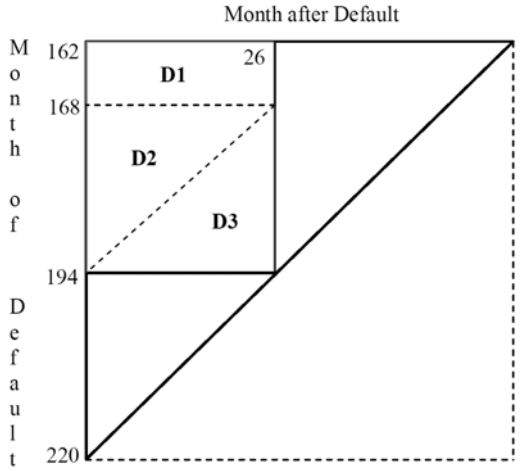
The purpose of the weights $w(a)$ is to differentiate completed recovery observations and partial observations. Note that in the Cox regression an account contributes to the likelihood function with a number of terms (9) corresponding to the number of monthly observations. Partial recoveries based on short observations should have lower weights than the results based on a full or almost completed recovery process. Consequently we propose to set $w(a)=1$ for completed observations and $w(a)=\frac{t_{end}(a)}{K}$ for incomplete recoveries. The estimation procedure can be directly realized in the case of the parametric Weibull (5) or Loglogistic model (6) where $\boldsymbol{\theta}=(\boldsymbol{\beta},p)$. To apply the idea to the Cox model we must specify the baseline survival function in (8). We will use simply the Kaplan-Meier estimate $S_0(t)$, and the vector of coefficients to be estimated will be just $\boldsymbol{\theta}=\boldsymbol{\beta}$ in this case also including the constant coefficient changing the overall level of the baseline survival, hence

$$S(t, \mathbf{x} \mid \boldsymbol{\beta}) = S_0(t)^{exp(\boldsymbol{\beta}'\mathbf{x})} \tag{12}$$

2 Empirical Results

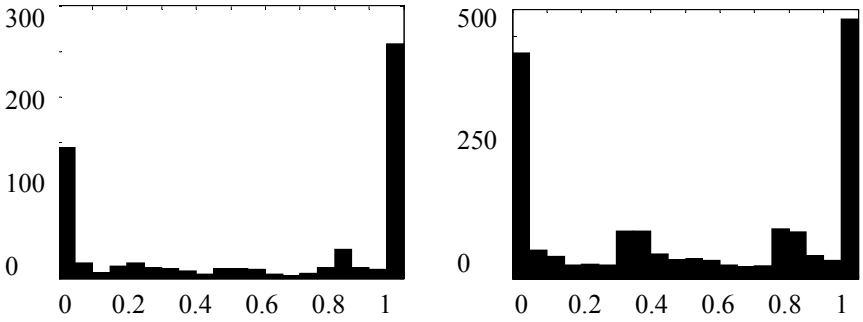
We have used an *LGD* data set of 4 000 defaulted unsecured retail loans obtained from a large Czech retail bank. The loans defaulted in a recent period (preceding the year 2008 but not exactly specified due to confidentiality reasons) of 57 months in a numbering used by the bank starting with the month $m_1 = 162$ and ending with the month $m_2 = 218$. The last month when we have observed recoveries is $m_3 = 220$, thus the recovery process has not been completed for many accounts. The data set contains account level information on net discounted monthly recovery cash flows as well as some basic application and behavior explanatory variables. Ultimate recovery rates are achieved by a sale of receivable, write-off, or full recovery, with majority of cases (87%) being resolved in 27 months. To test the survival methods in the context outlined in Section 2 we need data of the type shown in Figure 1 and at the same time to have the information on ultimate recoveries for all accounts in the data set. In order to achieve that we not only need to move retrospectively back, e.g. to restrict ourselves only to accounts that defaulted between the months 162 and 194, but also to shorten the maximum recovery time to a shorter period, e.g. setting $K = 27$. Figure 2 shows the structure of the original and modified data set.

Fig. 2: The original and modified data sets



Hence the development of various ex ante *LGD* functions will be done as of the month 194 on the data sets D1 and D2, but the goodness of fit measures will be calculated on ultimate recovery rates, i.e. also on the data set D3 available from the perspective of the month 220. Since for the purpose of survival analysis method testing we admit only nonnegative cash flows and recovery rates in [0,1] we had to omit negative cash flows and adjust the exposure at default to the cumulative recovery rate in case it exceeded the original *EAD*. The resulting distributions of the ultimate recovery rates on the data set D1 and on D3 shown on Figure 3 are highly bimodal due to the fact that original data contained an unusually high number of recoveries below 0 and over 1. Note that the recovery rate distribution on D3 (unknown at the development month 194) differs quite significantly from the distribution on D1.

Fig. 3: Histograms of ultimate recovery rates on the data sets D1 (left) and on D3 (right)



A descriptive statistics of the datasets D1, D2, and D3 focusing on the ultimate or last available recovery rates in the case of D2 is shown in Table 1.

The number of observations is obviously still more than sufficient to calibrate the model. There are 8 available explanatory variables including time in books, exposure, and other application or behavior properties not disclosed by the bank. One categorical variable with 10 possible values has been decoded into 9 dummy variables; hence the total number of the regression variables not including the intercept coefficient is 16.

Tab. 1: Descriptive statistics of the ultimate or last available recovery rates on D1, D2, and D3

	Ultimate RR on D1	Last RR on D2	Ultimate RR on D3
Num	605	1739	1739
Max	1	1	1
Min	0	0	0
Mean	0.5951	0.3508	0.5253
Median	0.8174	0.1136	0.5260
Range	1	1	1
Std	0.4270	0.4133	0.4010

The last available (partial) recovery rates on D2 cannot be used for the linear and logistic regressions development. The results of the regressions developed on ultimate recoveries on D1 in terms of the R-squared and modified R goodness of fit indicators measured on D1, D3, and D1+D3 are presented in Table 2 and 3.

Tab. 2: The goodness of fit measures for the *LGD* linear regression

D1			D3			D1+D3		
Num	R^2	Mod R	Num	R^2	Mod R	Num	R^2	Mod R
605	15.18%	**11.96%**	1739	6.72%	**8.63%**	2344	8.9%	**9.47%**

Tab. 3: The goodness of fit measures for the *LGD* logistic regression with different cut-offs

	D1			D3			D1+D3		
Cutoff	Num	R^2	ModR	Num	R^2	ModR	Num	R^2	ModR
0.1	605	12.73%	**8.67%**	1739	13.11%	**9.78%**	2344	13.02%	**9.50%**
0.2	605	14.32%	**10.28%**	1739	7.99%	**8.09%**	2344	9.62%	**8.65%**
0.3	605	12.14%	**9.20%**	1739	6.99%	**7.46%**	2344	8.32%	**7.90%**
0.4	605	15.63%	**11.59%**	1739	6.91%	**8.32%**	2344	9.15%	**9.15%**
0.5	605	15.08%	**11.15%**	1739	6.68%	**8.13%**	2344	8.84%	**8.89%**
0.6	605	15.06%	**11.28%**	1739	6.47%	**8.01%**	2344	8.68%	**8.84%**
0.7	605	14.79%	**10.99%**	1739	5.41%	**7.00%**	2344	7.82%	**8.01%**
0.8	605	13.93%	**10.70%**	1739	3.81%	**6.26%**	2344	6.41%	**7.38%**
0.9	605	12.85%	**9.30%**	1739	4.08%	**5.33%**	2344	6.33%	**6.33%**

Our key goodness of fit indicator, i.e. the modified R on D1+D3, does not show a superior performance with values below 10%. The low R indicates a weak explanatory power of the covariates which is nevertheless normal in the case of *LGD* predictions according to the authors' experience. The linear and logistic regressions with the recovery rate cut-off threshold at 10% show the best performance. Looking also on the R-squared one would prefer the logistic regression predictions. It is interesting to note that while the linear regression fits well the data set D1 and poorly the data set D3, the logistic regression predictions appear to be more balanced.

Next we have performed the Cox regression based on maximum likelihood estimation of the coefficients with the same covariates but extending the data set D1 with partial recoveries in D2. The goodness of fit measures in Table 4 indicate that the predictions fit much better the ultimate recovery rates given by D3 due to the partial recovery history information. The overall performance on D1+D3 is significantly superior to the linear and logistic regression.

Tab. 4: The goodness of fit measures for the Cox regression

D1			**D3**			**D1+D3**		
Num	R^2	Mod R	Num	R^2	Mod R	Num	R^2	Mod R
605	7.26%	**6.91%**	1739	14.53%	**12.99%**	2344	12.66%	**11.45%**

The Cox survival function and a particular shape of the baseline hazard function in Figure 4 and 5 indicate that the parametric hazard functions might be difficult to fit to the given data. The Weibull and Loglogistic models we have tested provide indeed weaker results compared to the Cox regression.

Fig. 4: The baseline hazard function given by the Cox regression

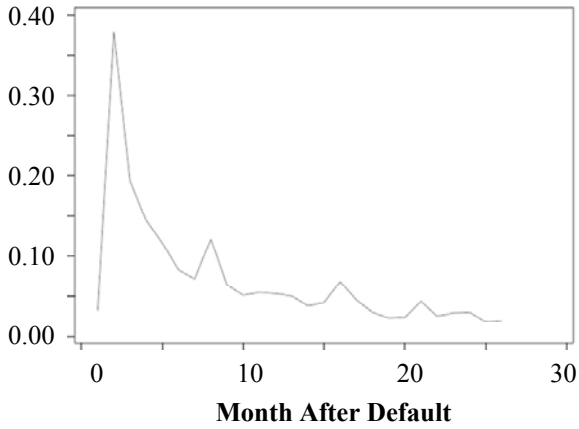
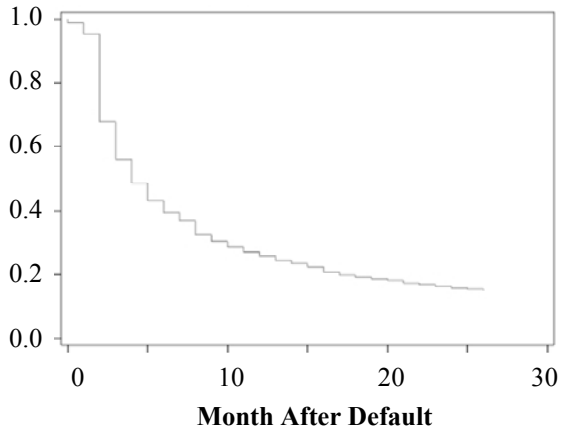


Fig. 5: The survival function for the first account given by the Cox regression



Finally we have estimated the Kaplan-Meier survival function $S_0(t)$ and found the coefficients of the Cox-like function (12) minimizing the sum of squared errors (11) on D1+D2. As we expected the predictions yield significantly better performance with modified R almost 13% and R-squared on D1+D3 exceeding 15%. The parametric functions again did not show a better result.

Tab. 5: The goodness of fit measures for the pseudo Cox regression

D1			D3			D1+D3		
Num	R^2	Mod R	Num	R^2	Mod R	Num	R^2	Mod R
605	10.21%	**8.81%**	1739	17.66%	**14.37%**	2344	15.68%	**12.92%**

Conclusions

We have described and tested four regression methods, linear regression, logistic regression, survival, and pseudo survival, to estimate future recovery rates and *LGD*s. The recovery data have been limited to only non negative cash flows and the recovery rates not exceeding one. Without those assumptions the survival methods can be hardly expected to be applicable. This prerequisite could be however achieved separating the gross recovery amounts from the costs and scaling the data appropriately, e.g. using a discount rate corresponding to the penalizing interest rates and the late fees. The general experience from banking practice is that standard regression *LGD* predictions perform quite poorly with R^2 below or around 10%. In spite of that banks do apply the regression analysis at least to sort exposures into appropriate *LGD* pools. Thus any improvement in the account level *LGD* prediction methods is desirable. The results confirmed that the survival methods utilizing partial recovery observations provide significantly better ex ante predictions with R^2 exceeding 15%. We have identified the Cox proportional model compared to the parametric as more flexible and appropriate to fit empirical recovery data with different patterns. Our proposed modification of the survival methods, in particular the pseudo Cox model, based on minimization of squared differences on last known recovery rates outperformed all the other methods.

References

[1] Altman, E. – Resti, A. – Sironi, A. (2004): *Default Recovery Rates in Credit Risk Modelling: A Review of the Literature and Empirical Evidence*. Economic Notes by Banca dei Paschi di Siena SpA, 2004, vol. 33, no. 2, pp. 183-208.

[2] Andreeva, G. (2006): *European Generic Scoring Models Using Survival Analysis*. Journal of the Operational Research Society, 2006, vol. 57, no. 10, pp. 1180-1187.

[3] BCBS (2005): *Basel Committee on Banking Supervision, Guidance on Paragraph 468 of the Framework Document*. Basel, Basel Committee on Banking Supervision, 2005.

[4] BCBS (2006*): International Convergence of Capital Measurement and Capital Standards. A Revised Framework – Comprehensive Version.* Basel, Basel Committee on Banking Supervision, 2006.

[5] Breslow, N. A. (1974): *Covariance Analysis of Censored Survival Data*. Biometrics, 1974, vol. 30, no. 1, pp. 89-99.

[6] Chava, S. – Stefanescu, C. – Turnbull, S. (2008): *Modeling the Loss Distribution*. [on-line], London, London Business School, c2008, [cit. 25th May, 2012], <http://faculty.london.edu/cstefanescu/Chava_Stefanescu_Turnbull.pdf >.

[7] Collet, D. (2003): *Modelling Survival Data in Medical Research*. London, Chapman & Hall / CRC, 2003.

[8] Cox, D. R. (1972): *Regression Models and Life-Tables*. Journal of the Royal Statistical Society, 1972, vol. 34, no. 2, pp. 187-220.

[9] Frye, J. (2003): *A False Sense of Security*, Risk, vol. 16, no. 8, pp. 63-67.

[10] Greene, W. H. (2003): *Econometric Analysis*, Englewod Cliffs, Prentice Hall, 2003.

[11] Gupton, G. M. (2005): *Advancing Loss Given Default Prediction Models: How the Quiet Have Quickened*, Economic Notes by Banca dei Paschi di Siena SpA, 2005, vol. 34, no. 2, pp. 185-230

[12] Huang, X. – Oosterlee, C. W. (2008): *Generalized Beta Regression Models for Random Loss-Given-Default*. Delft, Delft University of Technology Report 08-10, 2008.

[13] Kalbfleisch, J. D. – Prentice, R. L. (2002): *The Statistical Analysis of Failure Time Data*. Hoboken, New York, Wiley, 2002.

[14] Narain, B. (1992): *Survival Analysis and the Credit Granting Decision.* In: Thomas, L. C. – Crook, J. N. – Edelman, D. B. (eds):

Credit Scoring and Credit Control. Oxford, Oxford University Press, 1992, pp. 109-122.

[15] Rychnovsky, M. (2009): *Mathematical Models of LGD*, *Diploma Thesis*. Praha, Charles University, Faculty of Mathematics and Physics, April 2009.

[16] Schuermann, T. (2004): *What Do We Know About Loss Given Default*. Credit Risk Models and Management, London, Risk Books, 2004.

[17] Witzany, J. (2009): *Unexpected Recovery Risk and LGD Discount Rate Determination*. European Financial and Accounting Journal, 2009, vol. 4, no. 1, pp. 61-84.

European Financial and Accounting Journal, 2012, vol. 7, no. 1, pp. 6-27.

Survival Analysis in LGD Modeling

Jiří WITZANY – Michal RYCHNOVSKÝ – Pavel CHARAMZA

ABSTRACT

The paper proposes an application of the survival time analysis methodology to estimations of the Loss Given Default (*LGD*) parameter. The main advantage of the survival analysis approach compared to classical regression methods is that it allows exploiting partial recovery data. The model is also modified in order to improve performance of the appropriate goodness of fit measures. The empirical testing shows that the Cox proportional model applied to *LGD* modeling performs better than the linear and logistic regressions. In addition a significant improvement is achieved with the modified "pseudo" Cox *LGD* model.

Key words: Credit risk; Recovery rate; Loss given default; Correlation; Regulatory capital.

JEL classification: G21, G28, C14.