International Journal of Management, Accounting and Economics
Vol. 5, No. 1, January, 2018
ISSN 2383-2126 (Online)

www.ijmae.com

# Modelling LGD Using Survival Analysis

Rusul Alsarray[1]
Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Iran

## Abstract

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.In this paper, We compare linear regression and survival analysis models for modelling recovery rates and recovery amounts, in order to predict the LGD for unsecured consumer loans or credit cards.

**Keywords:** Recovery rate, Linear regression, Survival analysis, Loss Given Default forecasts

Cite this article: Alsarray, R. (2018). Modelling LGD Using Survival Analysis. *International Journal of Management, Accounting and Economics, 5*(1), 28-42.

## Introduction

With the introduction of Basel II, banks are allowed to use internally developed models for calculating its regulatory capital. This is called the Internal Ratings-Based (IRB) approach. Otherwise, the bank should calculate its regulatory capital based on the standardized approach which results in higher capital requirements. The purpose of the development of rating models is to identify and combine those factors that differentiate between facilities the best in terms of riskiness. The terms used to describe risk of facilities are probability of default (PD), loss given default (LGD) and exposure at default (EAD).Within Rabobank International (RI) models have been developed by the Modelling and Research department.

In order to determine PD, credit scoring systems were built. They try to answer the question how likely an applicant for credit is to default within a certain period. Many models are available of which currently the most commonly used is the logistic regression

[1] *Corresponding author's email: sweetrussul86@gmail.com*

(LR) approach, see e.g. Stepanova & Thomas (2002). In recent years, survival analysis has been introduced into credit scoring. This collection of statistical methods tries to model the time to default and has advantages compared to other credit scoring systems.

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 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 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 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.

In this paper, we use linear regression and survival analysis models to build predictive models for the recovery rate, and hence LGD. The comparison will be made based on a case study involving data from an in-house collection process for personal loans. This consisted of collection data on 27,000 personal loans over the period from 1989 to 2004.

## Methodology

### *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 *CFt* 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).

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*· *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.

### *Linear regression model*

Linear regression is the most obvious predictive model to use for recovery rate (RR) modelling, and is also widely used for prediction in other financial areas. Formally, a linear regression model fits a response variable y to a function of regressor variables $x_1, x_2$, ..., xm and parameters. The general linear regression model has the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon, \qquad (2.1)$$

where, in this case,

$y$ is the recovery rate or recovery amount;

$\beta_0, \beta_1, \dots, \beta_m$ are unknown parameters;

$x_1, x_2, \dots, x_m$ are independent variables which describe various characteristics of the loan and the borrower; and ε is a random error term.

In linear regression, one assumes that the mean of each error component (random variable ε) is zero, and each error component follows an approximate normal distribution. However, the distribution of the recovery rate tends to be a bathtub shape, so the error component of the linear regression model for predicting the recovery rate does not satisfy these assumptions.

### *Survival analysis models*

#### Survival analysis concepts

In survival analysis, one is normally dealing with the time that an event occurs; however, in some cases the event has not yet occurred, and so the data are censored. In our recovery rate approach, the target variable is how much has been recovered before the collection process stops, at which point collection is still underway in some cases, and thus the recovery rate is censored. The debts which had been written off are uncensored events, while the debts which are still being paid are censored events, because we do not know how much more money will be paid or could be paid. If the whole loan is paid off, we could treat this as a censored observation, as the recovery rate(RR) is greater than 1 in some cases. If one assumes that the recovery rate will never exceed 1, then such observations are not censored. Since we redefine the cases where RR >1 to RR = 1, we will consider all recovery rates of 1 to be censored.

Since the recovery process takes so long, survival analysis has an advantage over the regression approaches, in that one can also use the data for the cases which are still in the recovery process, rather than having to wait until they have either been paid off completely or been written off. Thus, in the regression approach one is using data on cases which are at least five years since default, on average. Suppose that T is the random variable of the percentage of the debt recovered (defined as RR in this case), which has a probability density function f . If an observed outcome, t of T , always lies in the interval [0,+∞), then T is a survival random variable. The cumulative density function F for this random variable is

$$F(t) = P(T \leq t) = \int_0^t f(u)du, \qquad (2.2)$$

and the survival function is defined as:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(u)du. \qquad (2.3)$$

Likewise, given S, one can calculate the probability density function f (u),

$$f(u) = -\frac{d}{du}S(u). \qquad (2.4)$$

The hazard function h(t) is an important concept in survival analysis, because it models the imminent risk. Here, the hazard function is defined as the instantaneous rate of no further payment of the debt, given that t% of the debt has been repaid:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t | T \geq t)}{\Delta t}. \qquad (2.5)$$

The hazard function can be expressed in terms of the survival function,

$$h(t) = \frac{f(t)}{S(t)}, \qquad t > 0. \qquad (2.6)$$

Rearranging, we can also express the survival function in terms of the hazard,

$$S(t) = e^{-\int_0^t h(u)du}. \qquad (2.7)$$

Finally, the cumulative hazard function, which relates to the hazard function h(t), namely

$$H(t) = \int_0^t h(u)du = -\ln S(t), \qquad (2.8)$$

is used widely.

It should be noted that f , F , S, h and H are related, and only one of the functions is required for us to be able to calculate the other four. There are two types of survival analysis models which connect the characteristics of the loan to the amount recovered: accelerated failure time models and Cox proportional hazards regression.

### Accelerated failure time models

In an accelerated failure time model, the explanatory variables act multiplicatively on the survival function, and either speed up or slow down the rate of 'failure'. If g is a positive function of x and S0 is the baseline survival function, then an accelerated failure model can be expressed as

$$S_x(t) = S_0\big(t.g(x)\big), \qquad (2.9)$$

where the failure rate is speeded up when g(x) < 1. By differentiating Eq. (2.9), the associated hazard function is

$$h_x(t) = h_0[tg(x)]g(x). \qquad (2.10)$$

For survival data, accelerated failure models are generally expressed as log-linear models, which occurs when $(x) = e^{\beta^T x}$ . In that case, one can show that the random variable T satisfies

$$log_e T_x = \mu_0 + \beta^T x + \sigma Z, \qquad (2.11)$$

where Z is a random variable with a zero mean and unit variance. The parameters β are then estimated via maximum likelihood methods. As a parametric model, Z is often specified as the Extreme Value distribution, which corresponds to T having an Exponential, Weibull, Log-logistic or other distribution. When building an accelerated failure model, the type of distribution of the dependent variable has to be specified. Using accelerated failure time ideas to model recovery rates leads to problems, in that they do not allow the target variable to have a zero value, nor can there be any value $t^*$ so that $S(t^*) = 1$ in all cases. Thus, in order to use this approach one must allow RR > 1, not redefine such recovery rates to be 1; one also needs to use a logistic regression model to first classify which loans will have a zero recovery rate, and use the accelerated failure approach on those which are predicted to have a positive recovery rate.

Cox proportional hazards regression

Cox (1972) proposed the following model:

$$h(t; x) = e^{(\beta^T x)} h_0(t), \qquad (2.12)$$

where β is a vector of unknown parameters, x is a vector of covariates and h0(t) is called the baseline hazard function. The advantages of this model are that we do not need to know the parametric form of $h_0(t)$ in order to estimate β, and also the distribution type of the dependent variable does not need to be specified. Cox (1972) showed that one can estimate β by using only the rank of the failure times to maximise the likelihood function.

## Case study

### *Data*

The project data relates to defaulted personal loans from a UK bank. The debts occurred between 1987 and 1999, and the repayment pattern was recorded until the end of 2003. Data on a total of 27,278 debts were recorded in the data set, of which 20.1% were paid off before the end of 2003, 14% were still being paid, and 65.9% were written off beforehand. The debt amount ranged from £500 to £16,000; 78% of debts were less than or equal to £5000, and only 3.6% of them were greater than £8000. Loans for multiples of a thousand pounds were most frequent, especially 1000, 2000, 3000 and 5000. Twenty one characteristics of the loan and the borrower were available in the data set, including the ratio of the loan to income, employment status, age, time with the bank, and purpose and term of the loan. The recovery amount is calculated as:

default amount − last outstanding balance (for non-write- off loans)

OR default amount−write off amount (for write-off loans).

The recovery rate

$$\frac{\text{Recovery amount}}{\text{Default amount}}$$

is more useful, as it relates to the percentage of the debt which is recovered. The average recovery rate in this data set is 0.42 (not including debts still being paid). Some debts could have a negative recovery rate, if the defaulted amounts generated interest and fees in the months after default, but the debtors did not pay anything, so that the outstanding balance kept increasing. Whether fees and interest are allowed to be added after default or not is determined by banking rules and the lender's accounting conventions. The vast majority of UK lenders do not add fees, and thus the amount owed is frozen at default, and the recovery rate is the amount repaid as a percentage of this. We use this convention in this paper, and thus recovery rates only increase with time. This also means that we redefine all negative recovery rates to be zero.  If fees and interest are included, it is possible for the recovered amount to exceed the amount at default. In this case, should one allow RR > 1 or redefine it to be 1?

We choose the latter course of action, which is consistent with fees being a cost in the recovery process, not part of the debt which is repaid. This is what mortgage and car finance companies do, in that the fees are taken out of the money received for selling the repossessed property before addressing the question of whether or not the remainder is enough to cover the defaulted balance of the loan. There is less uniformity for credit card and personal loan recoveries, but a collections department will not normally charge fees or add interest to the defaulted balance during the recovery process. With these conventions, the distribution of the recovery rate is a bathtub shape. 30.3% of debts have a 0 recovery rate, and 23.9% debts have a 100% recovery rate, while the others are distributed relatively evenly between 0 and 1. (This distribution excludes the debts which are still being paid.)The whole data set is split randomly into 2 parts: the training sample contains 70% of observations and is used for building models, while the test sample contains 30% of observations and is used for testing and comparing models. The modelling details are presented in the following sections. The results from linear regression and survival analysis models are compared.

### *Linear regression*

Two multiple linear regression models are built, one with the recovery rate as the target variable and one with the recovery amount as the target variable. In the former case, the predicted recovery rate can be multiplied by the default amount, and therefore the recovery amount can be predicted indirectly; while in the latter case, a predicted recovery rate can be obtained by dividing the predicted recovery amount by the default amount.

The stepwise selection method was used for all regression models. A coarse classification was used on the categorical variables, so that attributes with similar average target variable values are put in the same class. The two continuous variables ‘default amount’ and ‘ratio of default amount to total loan’ were transformed into ordinal variables as well, and their functions (square root, logarithm, and reciprocal) and original form were also included in the model building in order to find the best fit for the recovery rate. The results are reported using a number of measures; R2, the coefficient of determination, is a common measure of the goodness-of-fit for regression models, in that it measures how much of the square of the differences between the recovery rates of individual debtors and the mean recovery rate is explained by the RR model. Although R2 values of up to 0.8 are common in time series analysis, R2 values of around 0.1–0.2 are not unusual in real problems involving individual people. If one is only interested in how well the model ranks the debtors, the Spearman coefficient is more appropriate. On the other hand, if one is concerned about the error between the actual RR and the predicted RR for each individual, then the mean absolute error (MAE) or mean squared error (MSE) would be the important measure (the MAE and MSE values for the recovery amount will be much greater than those for the recovery rate, as the latter is always bounded between 0 and 1).

Consistent with the findings of previous studies (Bellotti & Crook, 2009; Dermine & de Carvalho, 2006; Matuszyk et al., 2010), the $R^2$ values for these models are small (see Table 1, which gives the results on the training samples); however, they are statistically significant. The Spearman rank correlation reflects the accuracy of the ranking of the predicted values. From the results, we can see that modelling the recovery rate directly is better than modelling it indirectly by first estimating the recovery amount. Surprisingly, better recovery amount results are also obtained by predicting the recovery rate first and then calculating the recovery amount, rather than estimating the amount directly. The details of the recovery rate models, the results of which are given in Table 1, are provided in Zhang and Thomas (2012).

Table 1. Linear regression models (results from the training sample).

| | $R^2$ | Spearman | MAE | MSE |
|---|---|---|---|---|
| Recovery rate from recovery rate model | 0.1066 | 0.3183 | 0.3663 | 0.1650 |
| Recovery rate from recovery amount model | 0.0354 | 0.2384 | 0.4046 | 0.2352 |
| Recovery amount from recovery amount model | 0.1968 | 0.2882 | 1239.2 | 2774405.4 |
| Recovery amount from recovery rate model | 0.2369 | 0.3307 | 1179.6 | 2637470.7 |

The most significant variable is the ratio of the default amount to the total loan, which is negatively related to the recovery rate. This gives some indication of how much of the loan was still owed when default occurred, and if a substantial proportion of the loan was repaid before default then the recovery rate is likely to be high. The second most significant variable is ‘second applicant status’, where loans with a second applicant have a higher recovery rate than loans without a second applicant. Other significant variables, using the t-value as a measure, include employment status, residential status, and default amount. The coefficient of the reciprocal of the default amount looks very large but is only multiplying small values, and thus the overall impact, although significant, is not the largest effect. The years of default were also allowed as variables, since they represent the best that one could hope to do by using economic variables to represent the temporal changes in the credit environment. The fact that they were not very significant means that it was felt that adding in economic variables would have only a minor impact in these models. In the recovery amount model, the variables which entered the model are very similar to those in the recovery rate model. Because predicting the recovery amount from the recovery amount model directly is worse than predicting it indirectly via the recovery rate model, the coefficient details of the recovery amount model are not given in this paper.

### *Survival analysis*

There are two reasons why survival analysis may be a useful approach for recovery rate and LGD modelling. Firstly, debts which are still being repaid cannot be included in the standard linear regression approach. Survival analysis models can treat such repayments as censored, and easily include them in the model building. Secondly, the recovery rate is not normally distributed, and therefore modelling it using a linear regression violates the assumptions of linear regression models. However, survival analysis models can handle this problem: different distributions can be set in accelerated models, and the Cox model's approach allows any empirical distribution. Survival analysis models for modelling both the recovery rate and the recovery amount can be built. The variable of interest is the percentage recovered when the debt is written off, so written-off debts are treated as uncensored, while debts which have been paid off or were still being paid are treated as censored. All of the independent variables which are used in building the linear regression model are used here as well, and once again they are coarse classified and dummy variables are used to represent the various classes so created. Continuous variables were firstly split into 10–15 bins to become 10–15 dummy variables, and these were used in a proportional hazard model without any other characteristics. Observing the coefficients from the model output, bins with similar coefficients were combined. The same method was used for nominal variables. Two continuous variables, 'default amount' and 'ratio of default amount to total loan', were included in the models, both in their original form and as coarse classified versions. Because accelerated failure time models cannot handle zeros existing in the target variable, observations with a recovery rate of zero should be removed from the training sample before building the accelerated failure time models. This is something that could also be done for the proportional hazards model, so that one is estimating the spike at RR = 0 separately from the rest of the distribution. This leads to a new task: a classification model is needed to classify recovery zeros and non-zeros (a recovery rate greater than 0). Therefore, a logistic regression model based on the training sample is built before building the accelerated failure time models. In the logistic regression model, the variables 'month until default' and 'loan term' are very significant, even though they were not very important in the linear regression models previously. The other variables selected in the model are similar to those in the previous regression models. The Gini coefficient is 0.32, and the logistic regression model predicted 57.8% of zeros as non-zeros and 21.5% of non-zeros as zeros. Cox regression models allow zeros to exist in the target variable, so two variants of the Cox model were built: one where those with RR = 0 were first separated out by building a logistic regression model, and a one stage model where all of the data were used to build the Cox model. For the accelerated failure life models, the type of distribution of the survival time needs to be chosen. After some simple distribution tests, the Weibull, Log-logistic and Gamma distributions were chosen for the recovery rate models; and the Weibull and Log-logistic distributions were chosen for the recovery amount models.

Unlike linear regression, survival analysis models generate a predicted distribution of recovery values for each debt, rather than a precise value. Thus, to give a precise value, the quantile or mean of the distribution needs to be chosen. For all of the survival models, the mean and median values are not good predictors, because they are too big and generate large MAE and MSE values compared with predictions from some other quantiles. The optimal predicting quantile points are chosen based on minimising the MAE and/or MSE. The lowest MAE and MSE values are found using quantile levels which are lower than the median, and the results from the training sample models are listed in Tables 2 and 3. The optimal quantiles are obtained empirically, but it would be interesting to see whether there is any theoretical justification for them which would be useful in using quantile regression in LGD modelling (Whittaker, Whitehead, & Somers, 2005). The model details of Cox with 0 recoveries are found in Zhang and Thomas (2012).

Using a quantile value has some advantages in this case, and quantile regression has previously been applied in credit scoring research. Whittaker et al. (2005) use quantile regression to analyse collection actions, and Somers and Whittaker (2007) use quantile regression for modelling distributions of profit and loss. Benoit and Van den Poel (2009) apply quantile regression to the analysis of customer life value. Using quantile values to make predictions avoids the influence of outliers. When using survival analysis in particular, the mean value of a distribution is affected by the number of censored observations in the data set, so the use of a quantile value is a good idea when making predictions. If the Spearman rank correlation test is the criterion by which the model is judged, we can see from the above tables of results (Tables 2 and 3) that the accelerated failure time model with a log-logistic distribution is the best of several survival analysis models. We can also see that the optimal quantile point is almost the same, regardless of the distribution in accelerated failure time models. In addition, the number of censored observations in the training sample does influence the optimal quantile point. If some of the censored observations are deleted from the training sample, the optimal quantile points move towards the median.

Table 2. Survival analysis model results for the recovery rate(training sample).

| Recovery rate | Optimal quantile(%) | Spearman | MAE | MSE |
|---|---|---|---|---|
| Accelerated(Weibull) | 34 | 0.24731 | 0.3552 | 0.1996 |
| Accelerated(log-logistic) | 34 | 0.25454 | 0.3532 | 0.2015 |
| Accelerated(gamma) | 36 | 0.16303 | 0.3597 | 0.1968 |
| Cox with 0 recoveries | 46 | 0.24773 | 0.3631 | 0.2092 |
| Cox without 0 recoveries | 30 | 0.24584 | 0.3604 | 0.2100 |

Table 3. Survival analysis model results for the recovery amount(training sample).

| Recovery amount | Optimal quantile(%) | Spearman | MAE | MSE |
|---|---|---|---|---|
| Accelerated(Weibull) | 34 | 0.30768 | 1129.7 | 3096952 |
| Accelerated(log-logistic) | 34 | 0.31582 | 1117.0 | 3113782 |
| Cox with 0 recoveries | 46 | 0.29001 | 1174.5 | 3145133 |
| Cox without 0 recoveries | 30 | 0.30747 | 1140.25 | 3112821 |

### *Model comparison*

The models are compared based on the results using the test sample. For debts which are still being paid, the final recovery amount and recovery rate are not known, and they cannot be measured properly; thus, these observations are removed from the test sample.

This is unfortunate, since it means that one is comparing the methods using only debts which have been completely either written off or paid off, even though one of the advantages of survival analysis is that it can deal with loans which are still being paid. The results from the single distribution models when applied to the test sample are listed in Tables 4 and 5. From the recovery rate (Table 4), if the $R^2$ value and the Spearman ranking test are the criteria for judging a model, we can see that (1) Linear regression is the best one, and (5) Cox—including zeros is the second best. In the training sample, the accelerated failure time model with a log-logistic distribution outperforms the Cox models, but for the test sample, the Cox model including zeros is more robust than the accelerated failure models. In terms of the MSE, linear regression always achieves the lowest MSE, as one would expect, given that it is minimising that criterion. All of the survival models have similar results. For the MAE, the results are very consistent, except that the linear regression models are poor. Modelling the recovery rate directly (rows 1–6 in Table 4) gives better results than modelling it indirectly via the recovery amount (rows 7–11 of Table 4). Almost all of the R2 and Spearman test results from the recovery amount models are lower than those from the recovery rate models. From the recovery amount results in Table 5, we can see that modelling the recovery amount directly (rows 1–5) is not as good as estimating the recovery rate first (rows 6–11). The (6) Linear regression∗ model achieves the highest R2, while the (10) Cox—including zeros∗ model achieves the highest Spearman ranking coefficient. Both of these are recovery rate models, and the predicted recovery amount is calculated by multiplying the predicted recovery rate by the default amount. The regression models and Cox including zeros models outperform the accelerated failure time models. In the test sample, the Cox including zeros model beats the other survival models. This is because the logistic regression model which is used before the other models to classify zero and non-zero recoveries generates more errors in the test sample, but this model does not affect the Cox—including zeros model.

Table 4. Comparison of the recovery rates from different single distribution models(test sample).

| Recovery rate | $R^2$ | Spearman | MAE | MSE |
|---|---|---|---|---|
| (1)Linear regression | 0.0904 | 0.29593 | 0.3682 | 0.1675 |
| (2)A-Weibull | 0.0598 | 0.25306 | 0.3586 | 0.2042 |
| (3) A-log-logistic | 0.0638 | 0.25990 | 0.3560 | 0.2060 |
| (4) A-gamma | 0.0527 | 0.23496 | 0.3635 | 0.2015 |
| (5)Cox-including zeros | 0.0673 | 0.27261 | 0.3546 | 0.2006 |
| (6) Cox-excluding zeros | 0.0609 | 0.25506 | 0.3564 | 0.2072 |
| (7) Linear regression* | 0.0292 | 0.22837 | 0.4077 | 0.2432 |
| (8) A-Weibull* | 0.0544 | 0.24410 | 0.3606 | 0.2070 |
| (9) A-log-logistic* | 0.0591 | 0.25315 | 0.3575 | 0.2077 |
| (10)Cox- including zeros* | 0.0425 | 0.22646 | 0.3693 | 0.2216 |
| (11)Cox-excluding zeros* | 0.0504 | 0.23269 | 0.3624 | 0.2108 |
| *Results from the recovery amount models. | | | | |

Table 5. Comparison of recovery amounts from different single distribution models(test sample).

| Recovery amount | $R^2$ | Spearman | MAE | MSE |
|---|---|---|---|---|
| (1) Linear regression | 0.1807 | 0.28930 | 1212.1 | 2634270 |
| (2) A-Weibull | 0.1341 | 0.30594 | 1123.5 | 3026908 |
| (3) A-log-logistic | 0.1318 | 0.31178 | 1111.7 | 3047317 |
| (4) Cox-including zeros | 0.1572 | 0.31788 | 1138.9 | 2887499 |
| (5) Cox-excluding zeros | 0.1400 | 0.30437 | 1125.3 | 3017661 |
| (6) Linear regression* | 0.2068 | 0.32522 | 1162.4 | 2549591 |
| (7) A-Weibull* | 0.1424 | 0.31149 | 1116.1 | 2982477 |
| (8) A-log-logistic* | 0.1396 | 0.31697 | 1105.9 | 3014320 |
| (9) A-gamma* | 0.1413 | 0.30139 | 1141.5 | 2972807 |
| (10) Cox-including zeros* | 0.1628 | 0.34619 | 1101.9 | 2906821 |
| (11) Cox-excluding zeros* | 0.1377 | 0.31246 | 1107.4 | 3028183 |
| *Results from the recovery rate models. | | | | |

## Conclusions

Estimating the recovery rate and recovery amount has recently become much more important, both because of the new Basel Accord regulation and because of the increase in the number of defaulters due to the recession.

This paper compares Several models of predicting the recovery rate for unsecured consumer loans. Linear regression and survival analysis are the two main techniques used in this research, where survival analysis can cope with censored data better than linear regression. For survival analysis models, we investigated the use of proportional hazard models and accelerated failure time models, although the latter have certain problems that need to be addressed: they do not allow zeros to exist in the target variable and the recovery rate cannot be bounded above. This can be overcome by not defining RR > 1 to be censored at 1 and by first using a logistic regression model to classify which loans have zero and non-zero recovery rates. Cox's proportional hazard regression models can deal with zeros in the target variable and with the requirement that RR ≤ 1 for all loans, so that approach was tried, both with logistic regression used first to split off the zero recoveries and without using logistic regression first. In all cases, the approaches were used to model both the recovery rate and the recovery amount, and for all of the models it proved to be better to model the recovery rate and then use this estimate to calculate the recovery amount, rather than modelling the recovery amount directly.

In our comparison, it has been shown that linear regression is better than survival analysis models in most situations. For recovery rate modelling, linear regression achieves a higher $R^2$ value and Spearman rank coefficient than the survival analysis models. The Cox model without the logistic regression first is the best of the survival analysis models. This is surprising, given the flexibility of distribution that the Cox approach allows. Of course, one would expect the minimum MSE to be obtained by the linear regression on the training sample, because that is what the linear regression tries to do. However, the superiority of the linear regression also holds for the other measures, on both the training and test sets. One reason for this may be the need to separate the zero recovery rate cases in the accelerated failure time approach. This is obviously difficult to do, and the errors from this first stage result in a poorer model at the second stage.

Another reason for the survival analysis approach not doing so well is that in performing these comparisons we used test sets where the recovery rate was known for all of the debtors. That is, they had all been either paid off or written off. Thus, there was no opportunity to test the model's predictions on those who were still paying, which is of course the type of data that are used by the survival analysis models, though not the regression based models. Finally, in the survival analysis approach there is the question of whether loans with RR = 1 are really censored or not. Assuming that they are not censored would lead to lower estimates of RR, which might be more appropriate for the conservative philosophy of the Basel Accord.

These results are based on the case study data set, which, though quite large, is from only one UK lender. The results require further validation from either the use of other data sets or some theoretical underpinning for them to be considered valid for all types of unsecured consumer credit LGD modelling.

## References

Altman, E. I., Resti, A., & Sironi, A. (2005). Loss given default; a review ofthe literature in recovery risk. In E. I. Altman, A. Resti, & A. Sironi (Eds.), Recovery risk (pp. 41–59). *London: Risk Books*.

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

Basel Committee on Banking Supervision (BCBS) (2004, updated 2005). International convergence of capital measurement and capital standards: a revised framework. Basel: Bank of International Settlement.

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

Bellotti, T., & Crook, J. (2009). Calculating LGD for credit cards. In Conference on risk management in the personal financial services sector. http://www3.imperial.ac.uk/mathsinstitute/programmes/research/bankfin/qfrmc/events/past/jan09conference.

Benoit, D. F., & Van den Poel, D. (2009). Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: an application in financial services. *Expert Systems with Applications*, 36, 10475–10484.

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

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, Series B, 34, 187–220.

Dermine, J., & de Carvalho, C. N. (2006). Bank loan losses given default: a case study. *Journal of Banking and Finance*, 30, 1219–1243.

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

Gupton, G. (2005). Estimating recovery risk by means of a quantitative model: lossCalc. In E. I. Altman, A. Resti, & A. Sironi (Eds.), Recovery risk (pp. 61–86). *London: Risk Books*.

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

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

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.

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

Schuermann, T. (2005). What do we know about loss given default? In E. I. Altman, A. Resti, & A. Sironi (Eds.), Recovery risk (pp. 3–24). *London: Risk Books*.

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

Whittaker, J., Whitehead, C., & Somers, M. (2005). The neglog transformation and quantile regression for the analysis of a large credit scoring database. *Journal of the Royal Statistical Society*, Series C, 54, 863–878.