Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD

Abstract

Estimating Recovery Rate and Recovery Amount has become important in consumer credit because of the new Basel Accord regulation and because of the increase in number of defaulters due to the recession. We compare linear regression and survival analysis models for modelling Recovery rates and Recovery amounts, so as to predict Loss Given Default (LGD) for unsecured consumer loans or credit cards. We also look at the advantages and disadvantages of using single distribution models or mixture distribution models for estimating these quantities.

Key words: Recovery Rate, Linear regression, Survival analysis, Mixture distribution, Loss Given Default forecasts

1. Introduction

The New Basel Accord allows a bank to calculate credit risk capital requirements according to one of two approaches. The first, the standardized approach requires a percentage of the risk weighted assets to be set aside where the percentage is given in the regulations. The second, the internal ratings based (IRB) approach, allows a bank to use internal estimates of components of credit risk to calculate credit risk capital. Institutions using IRB need to develop methods to estimate the following components for each segment of their loan portfolio:

– PD (probability of default in the next 12 months);

– LGD (loss given default);

– EAD (expected exposure at default).

Modelling PD, the probability of default has been the objective of credit scoring systems for fifty years but modelling LGD is not something that had really been addressed in consumer credit until the advent of the Basel regulations. Modelling LGD appears to be more difficult than modelling PD, because of two reasons. Firstly, much of the data may be censored (debts still being paid) because of the long time scale of recovery. Linear regression does not deal that well with censored data and even the Buckley-James approach (Buckley and James 1979) does not cope well with this form of censoring. Second, debtors have different reasons for defaulting and these lead to different repayment patterns. For example, some people do not want to repay; some people can not repay because of permanent changes in their situation, while for others the reason for non repayment is temporary. One distribution may find it hard to model the outcomes of these different reasons. Survival analysis though can handle censored data, and segmenting the whole default population is helpful to modelling LGD for defaulters with different reasons for defaulting.

Most LGD modelling research has concentrated on corporate lending where LGD (or its opposite Recovery Rate RR, where RR=1-LGD), was needed as part of the bond pricing formulae. Even in this case, until a decade ago LGD was assumed to be a deterministic value obtained from a historical analysis of bond losses or from bank work out experience (Altman et al 1977). Only when it was recognised that LGD was part of the pricing formula and that one could use the price of non defaulted risky

bonds to estimate the market's view of LGD were models of LGD developed. If defaults are rare in a particular bond class then it is likely the LGD got from the bond price is essentially a subjective judgment by the market. The market also trades defaulted bonds and so one can get directly the market values of defaulted bonds (Altman and Eberhart 1994). These values of LGD obtained from defaulted bonds or implied in the price of non-defaulted bonds were used to build regression models that related LGD to relevant factors, such as the seniority of the debt, country of issue, size of issue, size of firm, industrial sector of firm but most of all to economic conditions which determined where the economy was in relation to the business cycle. The most widely used model is the Moody's KMV model, LossCalc (Gupton 2005). It transforms the target variable into a normal distribution by a Beta transformation, regresses the transformed target variable on a few characteristics, and then transforms back the predicted values to get the LGD prediction. Another popular model, Recovery Ratings, was created by Standard & Poor's Ratings Services (Chew and Kerr 2005); it classifies the loans into 6 classes which cover different recovery ranges. Descriptions of the models are given in several books and reviews (Altman, Resti, Sironi 2005, De Servigny and Oliver 2004, Engelmann and Rauhmeier 2006, Schuermann 2005).

Such modelling is not appropriate for consumer credit LGD models since there is no continuous pricing of the debt as is the case on the bond market. The Basel Accord (BCBS 2004 paragraph 465) suggests using implied historic LGD as one approach in determining LGD for retail portfolios. This involves identifying the realised losses (RL) per unit amount loaned in a segment of the portfolio and then if one can estimate the default probability PD for that segment, one can calculate LGD since

RL=LGD.PD. One difficulty with this approach is that it is accounting losses that are often recorded and not the actual economic losses. Also since LGD must be estimated at the segment level of the portfolio, if not at the individual loan level there is often insufficient data in some segments to obtain robust estimates.

The alternative method suggested in the Basel Accord is to model the collections or work out process. Such data was used by Dermine and Neto de Carvalho (Dermine and Neto de Carvalho 2006) for bank loans to small and medium sized firms in Portugal. They used a regression approach, albeit a log-log form of the regression to estimate LGD.

The idea of using the collection process to model LGD was suggested for mortgages by Lucas (2006). The collection process was split into whether the property was repossessed and the loss if there was repossession. So a scorecard was built to estimate the probability of repossession where Loan to Value was key and then a model used to estimate the percentage of the estimated sale value of the house that is actually realised at sale time. For mortgage loans, a one-stage model, was build by Qi and Yang (2009). They modelled LGD directly, and found LTV (Loan to Value) was the key variable in the model and achieved an adjusted R square of 0.610, but only a value of 0.15 without including LTV.

For unsecured consumer credit, the only approach is to model the collections process, and now there is no security to be repossessed. The difficulty in such modelling is that the Loss Given Default, or the equivalent Recovery Rate, depends both on the ability and the willingness of the borrower to repay, and on decisions by the lender on how

vigorously to pursue the debt. This is identified at a macro level by Matuszyk et al (2010), who use a decision tree to model whether the lender will collect in house, use an agent on a percentage commission or sell off the debts, - each action putting different limits on the possible LGD. If one concentrates only on one mode of recovery in house collection for example, it is still very difficult to get good estimates. Matuszyk et al (2010) look at various versions of regression, while Bellotti and Crook (2009) add economic variables to the regression. Somers and Whittaker (2007) suggest using quantile regression, but in all cases the results in terms of R-square are poor - between 0.05 and 0.2. Querci (2005) investigated geographic location, loan type, workout process length and borrower characteristics for data from an Italian bank, but concludes none of them is able to explain LGD though borrower characteristics are the most effective.

In this paper, we use linear regression and survival analysis models to build predictive models for recovery rate, and hence LGD. Both single distribution and mixture distribution models are built to allow a comparison between them. This analysis will give an indication of how important it is to use models – survival analysis based ones- which cope with censored debts and also whether mixed distribution models give better predictions than single distribution model.

The comparison will be made based on a case study involving data from an in house collections process for personal loans. This consisted of collections data on 27K personal loans over the period from 1989 to 2004. In section two we briefly review the theory of linear regression and survival analysis models. In section three we explain the idea of mixture distribution models as they are applied in this problem. In

section four we build and compare single distribution models using linear regression and survival analysis based models, while in section five we create mixture distribution models, so that comparisons can be made. In section 6 we summarise the conclusion obtained.

2 Single distribution models

2.1 Linear regression model

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

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m + \varepsilon \tag{2.1}$$

Where in this case

y is the recovery rate or recovery amount

$\beta_0, \beta_1, ... \beta_m$ are unknown parameters

$x_1, x_{2,}, ..., x_m$ are independent variables which describe characteristics of the loan or the borrower

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 recovery rate tends to be bathtub shape, so the error component of linear regression model for predicting recovery rate does not satisfy these assumptions.

2.2 Survival analysis models

Survival analysis concepts

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

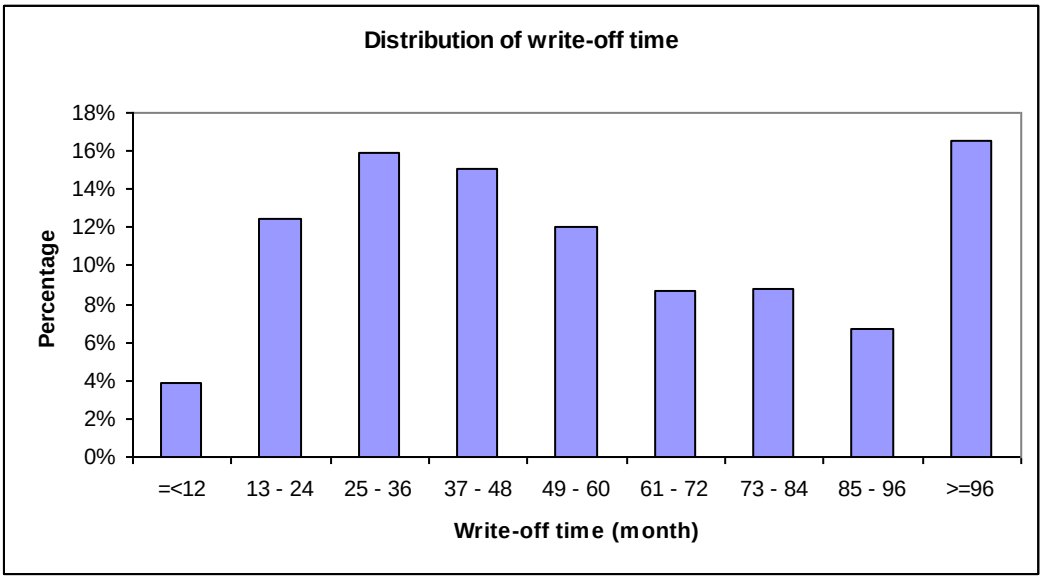

Figure 1: Distribution of writeoff/payoff times

Since the recovery process takes so long survival analysis has an advantage over the regression approaches, in that one can use the data for the cases in the recovery

process, and not have to wait until they have either paid off completely or been written off. Figure 1 shows the distribution of time between default and being written off or paid off in full for the data set of the case study described in section 4. It shows the mean write-off/pay off time is 58 month, with a standard deviation of 34 months, and a longest time of 173 months. So in the regression approach one is using data on cases which on average are at least five years since default.

Suppose T is the random variable of the percentage of the debt recovered (defined as RR in this case) which has probability density function f. If an observed outcome, t of T, always lies in the interval $[0, +\infty)$, then T is a survival random variable. The cumulative density function F for this random variable is

$$F(t) = P(T \le t) = \int_0^t f(u)du \tag{2.2}$$

The survival function is defined as:

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

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

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

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

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

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

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

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

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

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

$$H(t) = \int_0^t h(u)du = -\ln S(t) \tag{2.8}$$

is widely used.

It should be noted that f, F, S, h and H are related, and only one of the function is needed 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. They either speed up or slow down the rate of 'failure'. If g is a positive function of x and S_0 is the baseline survival function then an accelerated failure model can be expressed as

$$S_x(t) = S_0(t \cdot g(x)) \tag{2.9}$$

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

$$h_x(t) = h_0[tg(x)]g(x) \tag{2.10}$$

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

$$\log_e T_x = \mu_0 + \beta^T x + \sigma Z \tag{2.11}$$

where Z is a random variable with zero mean and unit variance. The parameters, β, are then estimated through 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 types of distribution. When building an accelerated failure models, 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 a value t* so that S(t*)=1 for all cases. Thus to use this approach one must allow RR>1 and not redefine such recovery rates to be 1; one also needs to use a logistic regression model to first classify which loans will have zero recovery rate, and use the accelerated failure approach on those which are predicted to have positive recovery rate.

Cox proportional hazards regression

Cox (1972) proposed the following model

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

Where β is a vector of unknown parameters, x is a vector of covariates and $h_0(t)$ is called the baseline hazard function.

The advantage of this model is that we do not need to know the parametric form of $h_0(t)$ to estimate β, and also the distribution type of 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.

3 Mixture distribution models

Models may be improved by segmenting population and building different models for each segment, because some subgroups maybe have different features and distributions. For example, small and large loans have different recovery rates, long established customers have higher recovery rate than relatively new customers (the latter may have high fraudulent elements which lead to low RR), and recovery rate of house owners is higher than that of tenants (because the former has more assets which may be realisable). Segmenting on recovery rate is a way of splitting who will not pay or permanently cannot pay from those who temporarily cannot pay. One could develop more sophisticated segments but using the RR values is an obvious first approach to a mixture model.

The development of finite mixture (FM) models dates back to the nineteenth century. In recent decades, as result of advances in computing, FM models proved to offer powerful tools for the analysis of a wide range of research questions, especially in social science and management (Dias, 2004). A natural interpretation of FM models is that observations collected from a sample of subjects arise from two or more unobserved/unknown subpopulations. The purpose is to unmix the sample and to identify the underlying subpopulations or groups. Therefore, the FM model can be seen as a model-based clustering or segmentation technique (McLachlan and Basford, 1998; Wedel and Kamakura, 2000).

In order to investigate different features and distributions in subgroups, we model the recovery rate by segmenting first. A classification tree model is built to generate segments with different features. Then, linear regression and survival models are built for each segment, so that mixture distribution models can be created.

4 Case Study – Single distribution model

4.1 Data

The data in the project is data on 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. In total 27278 debts were recorded in the data set, of which, 20.1% debts were paid off before the end of 2003, 14% debts were still being paid, and 65.9% debts were written off beforehand. The range of the debt amount was from £500 to £16,000; 78% of debts are less than or equal to £5,000 and only 3.6% of them are greater than £8,000. Loans for multiples of thousands of pound are most frequent, especially 1000, 2000, 3000 and 5000. Twenty one characteristics about the loan and the borrower were available in the data set such as the ratio of the loan to income, employment status, age, time with bank, and purpose and term of 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 distribution of recovery amount is given in Figure 2, ignoring debts that are still being repaid but this graph could be misleading as it does not describe the original debt.

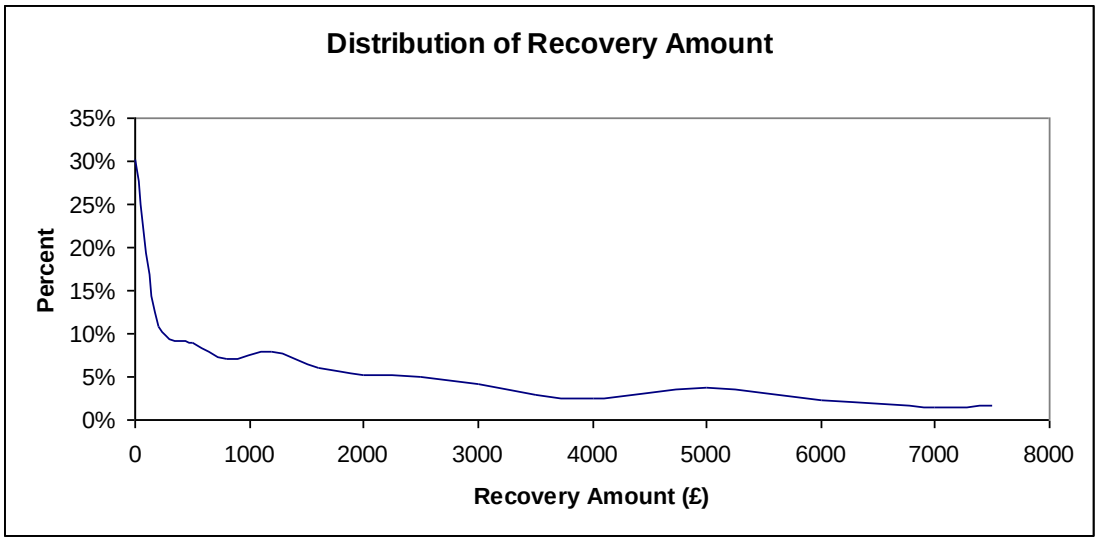

Figure 2: Distribution of Recovery Amount in the data set

The recovery rate $\frac{\text{Recovery Amount}}{\text{Default Amount}}$

is more useful as it describes what percentage of the debt is recovered. The average recovery rate in this data set is 0.42 (not including debts still being paid). Some debts could have negative recovery rate, if the defaulted amounts generate interest and fees in the months after default, but the debtors did not pay anything, so the outstanding balance keeps increasing. Whether fees and interest are allowed to be added after default is determined by banking rules and the lender's accounting conventions. The vast majority of UK lenders do not add fees and so 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 so recovery rates only increase with time. It also means 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 and 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 whether the remainder is enough to cover the defaulted balance of the loan. For credit card and personal loan recoveries there is less uniformity but normally a collections department will not charge fees or add interest to the defaulted balance during the recovery process.

With these conventions, the distribution of recovery rate is a bathtub shape, see Figure 3. 30.3% debts have 0 recovery rate, and 23.9% debts have 100% recovery rate, others are relatively evenly distributed between 0 and 1. (This distribution excludes the debts still being paid.)

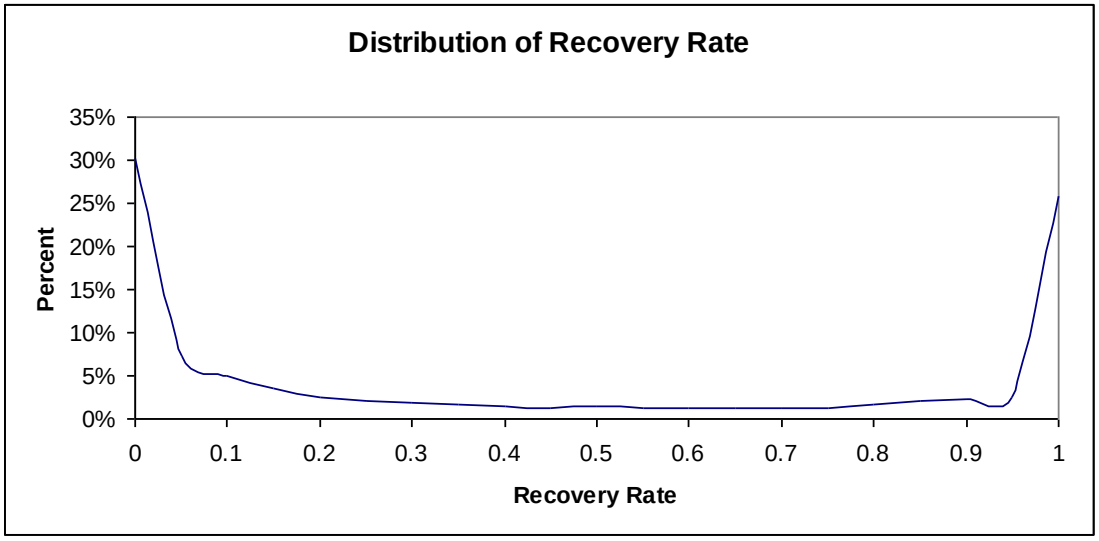

Figure 3: Distribution of recovery rate in the data set

The whole data is randomly split into 2 parts; the training sample contains 70% of observations for building models, and the test sample contains 30% of observations for testing and comparing models.

In the following sections, the modelling details are presented. The results from linear regression and survival analysis models are compared, as are the results from single distribution models and mixture distribution models.

4.2 Single distribution models

Linear regression

Two multiple linear regression models are built, one is for recovery rate as the target variable and one is for recovery amount as the target variable. In the former case, the predicted recovery rate could be multiplied by the default amount, and so the recovery amount could be predicted indirectly; 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. Coarse classification was used on 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 also their functions (square root, logarithm, and reciprocal) and their original form were 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, R^2, the coefficient of determination is a common measure of goodness of fit for regression models , in that it measures how much of the square of the differences between the recovery rate of individual debtors and the mean recovery rate is explained by the RR model. Although R^2 of up to 0.8 are common in time series analysis, in real problems

involving individual people, R^2 around 0.1 to 0.2 are not unusual. If one is only interested in how well the model is ranking the debtors, the Spearman coefficient is more appropriate. If one is concerned about the error between the actual RR and the predicted RR for each individual then Mean absolute error (MAE) or Mean square error (MSE) would be the measure of importance. (MAE and MSE values for Recovery Amount will be much greater than those for Recovery Rate as the latter is always bounded between 0 and 1).

The R-squares for these models are small, (see Table 1, which gives the results on the training samples). This is consistent with previous authors (Bellotti and Crook 2009, Dermine and Neto de Carvalho 2006, Matuszyk et al 2010), but they are statistically significant. The Spearman rank correlation reflects how accurate was the ranking of the predicted values. From the results, we can see modelling recovery rate directly is better than indirect modelling by first estimating the recovery amount. Surprisingly, better recovery amount results are also obtained by predicting recovery rate first and then calculating recovery amount rather than estimating the amount directly.

	R-square	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

Table 1: Linear regression models (results are from training sample)

The details of these recovery rate models whose results are given in Table 1 are given in Appendix 1. The most significant variable is 'the ratio of default amount to total loan', which has a negative relation with recovery rate. This gives some indication of how much of the loan was still owed before default occurs, and if a substantial portion of the loan was repaid before default then the Recovery Rate is also likely to be high. The second most significant variable is 'second applicant status', where loans with a second applicant have higher recovery rate than loans without a second applicant. Other significant variables, using t value as a measure, include: employment status, residential status, and default amount. The coefficient of the reciprocal of default amount looks very large but is only multiplying small values; so the overall impact although significant is not the largest effect. The years of default were also allowed as variables since they represent the best one could hope to do if one used economic variables to represent the temporal changes in the credit environment. The fact they were not that significant means it was felt that adding in economic variables would have a minor impact in these models.

In the recovery amount model, the variables which entered the model are very similar to recovery rate model. Because predicting recovery amount directly from the recovery amount model is worse than that predicting it indirectly via the recovery rate model, the coefficient details of recovery amount model are not given in this paper.

Survival analysis

There are two reasons why survival analysis may be a useful approach to Recovery Rate and LGD modelling. Firstly, debts still being repaid cannot be included in the standard linear regression approach. Survival analysis models can treat such

repayments as censored, and include them easily in the model building. Secondly, the recovery rate is not normally distributed, so modelling it using linear regression violates the assumptions of linear regression. Survival analysis models can handle this problem; different distributions can be set in accelerated models and Cox model's approach allows any empirical distribution.

Survival analysis models can be built for modelling both recovery rate and recovery amount. The event of interest is the percentage recovered when the debt is written off, so written-off debts are treated as uncensored; debts which were paid off or were still being paid are treated as censored. All the independent variables which are used in the linear regression model building are used here as well, and they are coarse classified again and dummy variables used to represent the various classes created. Continuous variables were firstly split into 10 to 15 bins to become 10 to 15 dummy variables, and these 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 can not handle 0's existing in target variable, observations with recovery rate 0 should be removed off from the training sample before building the accelerated failure time models. This is also something that could be done for 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 0's and non-0's (recovery rate greater than 0).

Therefore, a logistic regression model is built based on the training sample before building the accelerated failure time models. In the logistic regression model, the variables 'month until default' and 'loan term' are very significant, though they were not so important in the linear regression models before. The other variables selected in the model are similar to those in the previous regression models. The Gini coefficient is 0.32 and 57.8% 0's were predicted as non-0's and 21.5% non-0's were predicted as 0's by logistic regression model. Cox regression models allow 0's to exist in the target variable; so two variants of the Cox model were built – one where one first separated out those with RR=0 by building a logistic regression model, and a one stage model where all the data was used to build the Cox model.

For the accelerated failure life models, the type of distribution of survival time needs to be chosen. After some simple distribution tests, Weibull, Log-logistic and Gamma distributions were chosen for the recovery rate models; and Weibull and Log-logistic distributions were chosen for the recovery amount models.

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 2: Survival analysis models results for recovery rate (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

Table 3: Survival analysis models results for recovery amount (training sample)

Unlike linear regression, survival analysis models generate a predicted distribution of the 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. In all the survival models, the mean and median values are not good predictors, because they are too big and generate large MAE and MSE 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 are found with quantile levels lower than median, and the results from the training sample models are listed in Table 2 and Table 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 et al 2005). The model details of Cox-with 0 recoveries are found in Appendix 2., while the baseline hazard function for the model excluding the RR=0 values is given in Figure 4

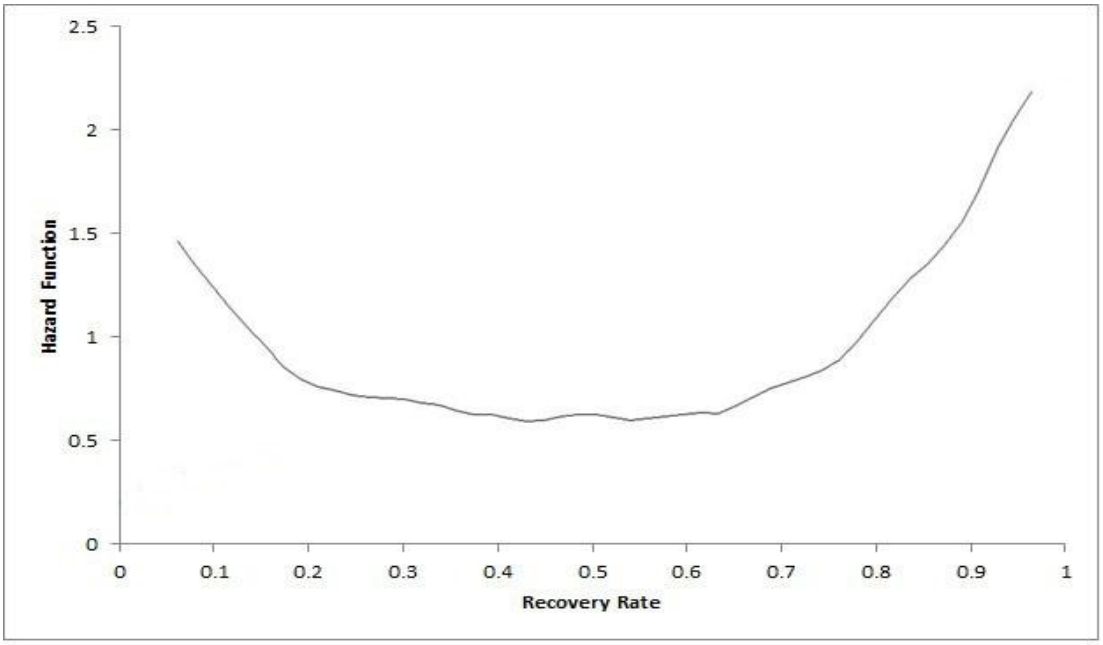

Figure 4: Baseline hazard function obtained from Cox model excluding RR=0

Using a quantile value has some advantages in this case and quantile regression has 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 analyse customer life value. Using quantile values to make prediction avoids outlier influences. In particular when using survival analysis, the mean value of a distribution is affected by the amount of censored observations in the data set, so use a quantile value is a good idea when making predictions using it.

If the Spearman rank correlation test is the criterion to judge the model, we can see, from the above results tables (table2 and table3), the accelerated failure time model with log-logistic distribution is the best one among several survival analysis models. We can also see the optimal quantile point is almost the same regardless of the distribution in accelerated failure time models. Also the number of censored observations in the training sample does influence what is the optimal quantile point.

If some of the censored observations are deleted from the training sample, the optimal quantile points move towards the median.

Model comparison

The comparison of the models is based on the results using the test sample. For debts still being paid, the final recovery amount and recovery rate are not known, and they can't be measured properly, thus these observations are removed from the test sample.

Recovery Rate	R-square	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 0's	0.0673	0.27261	0.3546	0.2006
(6) Cox – excluding 0's	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 0's*	0.0425	0.22646	0.3693	0.2216
(11) Cox – excluding 0's*	0.0504	0.23269	0.3624	0.2108

*: results from recovery amount models

Table 4: Comparison of recovery rate from single distribution models (test sample)

This is unfortunate since it means one is comparing the methods only using debts which have been completely written off or paid off. Yet one of the advantages of survival analysis is that it can deal with loans which are still paying. The results from all the single distribution models when applied to the test sample are listed in Tables 4 and 5.

From the recovery rate Table 4, if R-square and Spearman ranking test are the criterion to judge a model, we can see (1) Linear Regression is the best one, and (5)

Cox-including 0's is the second best model. In the training sample, accelerated failure time model with log-logistic distribution outperforms the Cox models, but for the test sample, the Cox model including 0's is more robust than the accelerated failure models. In terms of MSE, linear regression always achieves the lowest MSE as one would expect as it is minimising that criterion. All the survival models have similar results. For MAE, the results are very consistent, except the linear regression models

Recovery Amount	R-square	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 0's	0.1572	0.31788	1138.9	2887499
(5) Cox – excluding 0's	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 0's*	0.1628	0.34619	1101.9	2906821
(11) Cox – excluding 0's*	0.1377	0.31246	1107.4	3028183

*: results from recovery rate models

Table 5: Comparison of recovery amount from single distribution models (test sample)

are poor. Modelling recovery rate directly (rows 1 to 6 in Table 4) gives better results than modelling it indirectly via recovery amount, whose results are in rows 7 to 11 of Table 4. Almost all the R-square and Spearman test from recovery amount models are lower than these from recovery rate models.

From the recovery amount results in Table 5, we see that modelling recovery amount directly (rows 1 to 5) is not as good as estimating recovery rate first (rows 6 to 11).

The (6) Linear Regression* model achieves the highest R-square while (10) Cox-including 0's* model achieves the highest Spearman ranking coefficient. Both of them are recovery rate models and the predicted recovery amount is calculated by multiplying predicted recovery rate by the default amount. Regression models and Cox-including 0's models outweigh the accelerated failure time models. In the test sample, Cox-including 0's model beats the other survival models. The reason is that the logistic regression model which is used before the other models to classify 0 recoveries and non-0 recoveries generates more errors in the test sample, but Cox-including 0's model is not affected by this model.

5 Mixture distribution models

Mixture distribution models have the potential to improve prediction accuracy and they have been investigated by other researchers for modelling RR. Matuszyk et al (2010) suggested to separate LGD=0 and LGD>=0 for unsecured personal loans, and then modelling LGD by using different models in each segment. Bellotti and Crook (2009) suggested to separate RR=0, 0<RR<1, and RR=1 for credit cards, and then for the group 0<RR<1, use Ordinary Least Squares regression or Least Absolute Value regression to model RR and achieved R-square 0.077. One possible reason for modelling RR by mixture distribution is people's different views about repayment. Some debtors want to pay back, but they have financial troubles and can't pay back; but some debtors deliberately do not want to pay.

For these reasons, we build a mixture model where the segments aim to have different recovery rate ranges. There are other ways of segmenting – age and size of loan, percentage of loan already paid off - which may also separate out the won't pays from

the can't permanently pays and can't temporarily pays, but using Recovery Rate to segment has the advantage of building on the work of others and of the inherent view that RR=0 must contain the won't pays. The default years were not considered as variables to segment on because they did not appear significant in the single distributions, but it might be worth exploring this further in due course. We describe two approaches to achieving appropriate segments.

Method 1

The recovery rate is treated as a continuous variable and also the target variable, and a classification tree model is built to split the whole population into a few subgroups, in order to maximise the difference of average recovery rate between the subgroups.

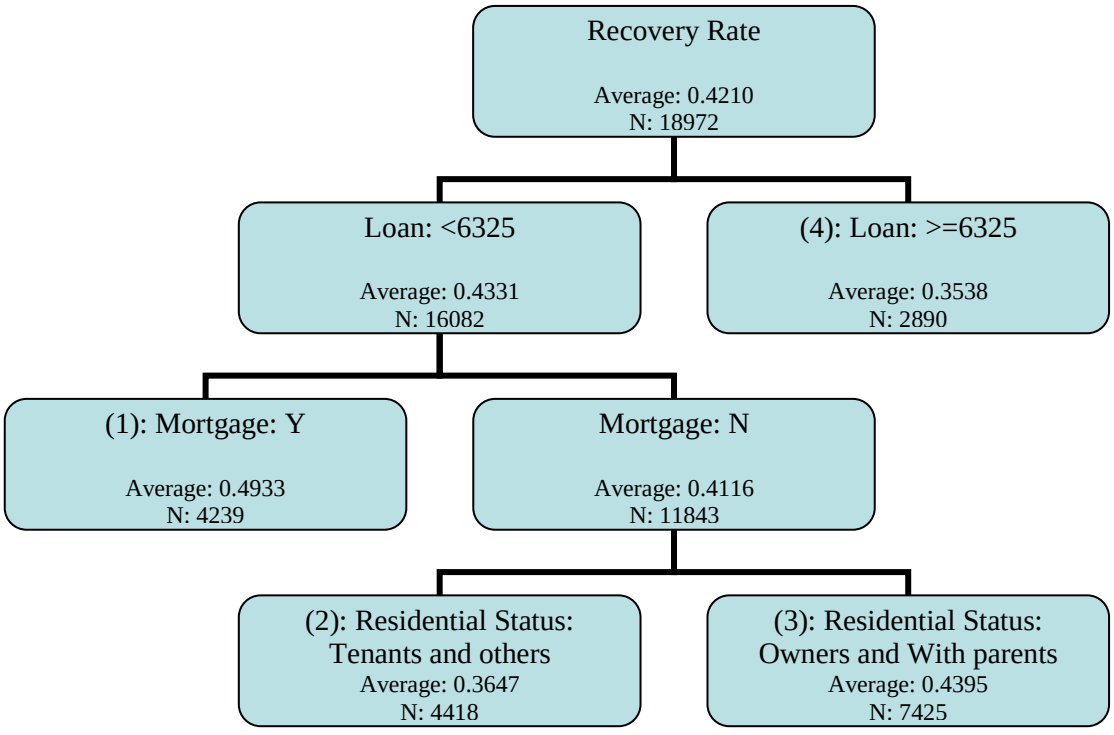

Figure 5: Classification tree for recovery rate as continuous variable

As is seen from the tree in Figure 5, the whole population was eventually split into 4 segments. Generally, large amount loans have lower recovery rate than small amount loans; if the debtors have a mortgage with this bank, then their loans have higher

recovery rate than those without a mortgage with the bank; house owners or living with parents have higher recovery rates than people of tenants or those with 'other' residential status.

Linear regression model and survival models are built on each of the segments. The previous research shows that better predicted recovery amount results are obtained from predicting recovery rate first and then multiplying by the default amount, so only recovery rate models are built here. The models are built based on training samples and tested on test samples.

Recovery Rate	R-square	Spearman	MAE	MSE
Regression	0.0840	0.28544	0.3693	0.1688
Accelerated	0.0660	0.26625	0.3549	0.2055
Cox-including 0's	0.0752	0.28581	0.3518	0.1967
Cox-excluding 0's	0.0636	0.26236	0.3549	0.2067

Table 6: Recovery rate from mixture distribution models of method 1 (test sample)

In all four segments, linear regression is always the best modelling technique, as it has the highest R-square and Spearman coefficient; so after piecing together the 4 segments, linear regression model still has the highest R-square. Among the accelerated failure time models, the best fit in the first three segments are achieved with the log-logistic distribution models, and the best fit in the last segment is with Weibull distribution model. So the test results for the accelerated failure time models are made up of three log-logistic distribution models and one Weibull distribution model. In the Cox-regression modelling, the Cox model including 0's (without logistic regression to predict 0 or non-0 recoveries) performs better than Cox model excluding 0's (with logistic regression first) in all four subgroups. This means it is not better to predict 0 recoveries by logistic regression first. The results of the four

approaches are given for the recovery rate in Table 6 and for the recovery amount in Table 7.

Recovery Amount	R-square	Spearman	MAE	MSE
Regression	0.1942	0.31824	1166.7	2593870
Accelerated	0.1346	0.31820	1102.3	3030185
Cox-including 0's	0.1574	0.35314	1100.5	2976283
Cox-excluding 0's	0.1357	0.31564	1105.8	3068188

Table 7: Recovery amount from mixture distribution models of method 1 (test sample)

In terms of R-square, among mixture distribution models, the linear regression models are the best; but in terms of Spearman ranking test, the Cox model-including 0's outperforms the linear regression model, especially for predicting recovery amount.

Compared with the analysis from single distribution models, the results from mixture distribution models are disappointing and are somewhat worse than the results from the single distribution models. In terms of R-square, the best mixture distribution model is linear regression, but its R-square is still lower than that from the single distribution linear regression model. In terms of Spearman ranking coefficient, the best mixture distribution model is the Cox model-including 0's. The Spearman ranking coefficient for the recovery rate is a little bit lower than 0.29593 which is the best one in the single distribution models; the Spearman ranking coefficient for the recovery amount is higher than 0.34619 which is the highest in the single distribution models. Thus, it seems mixture distribution models only improve the Spearman rank coefficient in the case of recovery amount predictions.

Method 2

Another way to separate the whole population is to split the target variable into three groups: the first group RR<0.05 (almost no recoveries), the second group 0.05<RR<0.95 (partial recoveries), and the third group RR>0.95 (full recoveries). These splits correspond to essentially no, partial or full recovery.

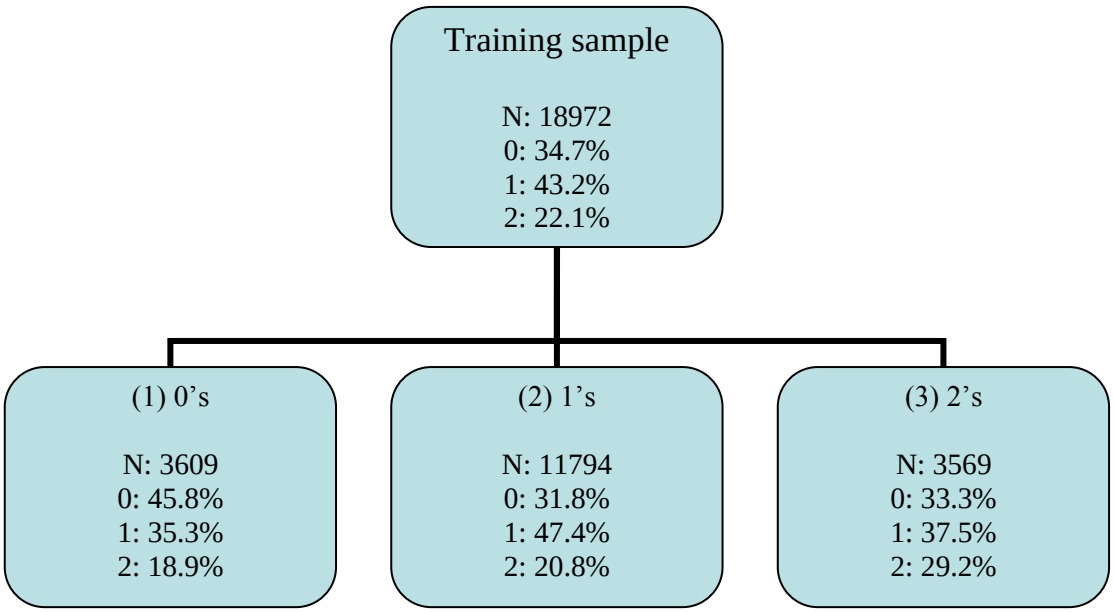

Figure 6: Classification tree for recovery rate as ordinal variable

Recovery rate can be treated as an ordinal variable, with three classes - recovery rate less than 0.05 is set to 0, recovery rate between 0.05 and 0.95 is set 1, and recovery rate greater than 0.95 is set 2. A classification tree with the three classes as the target variable was tried, but the results were disappointing because each end node had similar distribution over the three classes. As an alternative a classification tree was first built to separate 0's and non-0's, so the whole data is split into two groups. Then a second classification tree was built for the non-0's group, in order to separate them into 1's and 2's. So again the population was split into 3 subgroups and this gave slightly better results. The population in the first segment (most zero repayments)

have the following attributes: no mortgage and loan term less than or equal to 12 months, OR no mortgage, time at address less than 78 months and have a current account. The population in the third segment (highest full repayment rate) have attributes: loan less than £4320 and insurance accepted. The rest of the population are allocated to the second segment as is shown in Figure 6.

This classification is very coarse. Group (1) aims at debts with recovery rate less than 0.05, but only 45.8% debts actually belong to this group; group (2) is for the debts with recovery rate between 0.05 and 0.95, but only 47.4% debts are in this range; group (3) is for the debts with recovery rate greater than 0.95, but, only 29.2% debts in this group have recovery rate greater than 0.95.

In the previous analysis, the linear regression model and Cox-including 0's model are the two best models, so here only the linear regression model and the Cox-including 0's regression model are built for each of the three segments. The results from the combined test sample are compared with the results from previous research in Tables 8 and 9.

Recovery Rate	R-square	Spearman	MAE	MSE
Regression	0.0734	0.26453	0.3695	0.1688
Cox including 0's	0.0570	0.25869	0.3588	0.2051

Table 8: Recovery rate from mixture distribution models of method 2 (test sample)

Recovery Amount	R-square	Spearman	MAE	MSE
Regression	0.2054	0.31356	1169.4	2564149
Cox including 0's	0.1669	0.33888	1125.7	2930725

Table 9: Recovery amount from mixture distribution models of method 2 (test sample)

From Tables 8 and 9, we can see that, for recovery rate, the linear regression model is still better than the Cox regression model in terms of R-square and Spearman coefficient; for recovery amount, the R-square of the linear regression model is higher than that of the Cox regression model, but the Spearman coefficient of linear regression is lower than that of the Cox model. Compared with the results from single distribution models, these mixture models do not improve the R-square or the Spearman ranking coefficient.

5 Conclusions

Estimating Recovery Rate and Recovery Amount has 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 makes a comparison between single distribution and mixture distribution models of predicting 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 though the latter have certain problems that need to be addressed-they do not allow 0's 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 which have non zero recovery rates. Cox's proportional hazard regression models can deal with 0's in the target variable and can deal 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 case one used the approaches to model both recovery rate and recovery amount, and for all the models it turns out it is better to model recovery rate and then use the estimate to calculate the recovery amount rather than modelling the recovery amount directly.

.

In the comparison of the single distribution models, the research result shows that linear regression is better than survival analysis models in most situations. For recovery rate modelling, linear regression achieves higher R-square and Spearman rank coefficient than survival analysis models. The Cox model without logistic regression first is the best model among all the survival analysis models. This is surprising given the flexibility of distribution that the Cox approach allows. Of course one would expect MSE to be minimised using linear regression on the training sample because that is what linear regression tries to do. However, the superiority of linear regression holds for the other measures both on the training and the test set. One reason may be the need to split off the zero recovery rate cases in the accelerated failure time approach. This is obviously difficult to do and the errors from this first stage results in a poorer model in the second stage. This could also be the reason that the mixture models do not give a real improvement. Finding suitable segments is difficult and the resultant subgroups are not as homogeneous as one would wish.

Another reason for the survival analysis approach not doing so well is that to make comparisons we used test sets where the recovery rate was known for all the debtors. That is they all had either paid off or been written off. So there was no opportunity to test the models predictions on those who were still paying, which is of course the type of data that is used by the survival analysis models but not by 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 they are not censored would lead to model lower estimate of RR, which might be more appropriate for the conservative philosophy of the Basel Accord.

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

References:

Altman E., Eberhart A., (1994), Do Seniority Provisions protect bondholders' investments, J. Portfolio Management, Summer, pp67-75

Altman E., Haldeman R., Narayanan P., (1977), ZETA Analysis: A new model to identify bankruptcy risk of corporations, Journal of banking and Finance 1, pp29-54

Altman E. I., Resti A., Sironi A. (2005), Loss Given Default; a review of the literature in Recovery Risk, *Recovery Risk*, ed by Altman E.I., Resti A, Sironi A. Risk Books, London, pp 41-59

Basel Committee on Banking Supervision (BCBS), (2004, updated 2005), International Convergence of Capital Measurement and Capital standards: a revised framework, Bank of International Settlement, Basel.

Bellotti T., Crook J., (2009), Calculating LGD for Credit Cards, presentation in Conference on Risk Management in the Personal Financial Services Sector, London, 22-23 January 2009
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 (2009) 10475-10484

Buckley, J. J., James, I. R. (1979), Linear regression with censored data, *Biometrika* 66, 429-436.

Chew W.H., Kerr S.S., (2005), Recovery Ratings: Fundamental Approach to Estimation Recovery Risk, *Recovery Risk*, ed by Altman E.I., Resti A, Sironi A. Risk books, London, p87-97

Cox D.R., (1972), Regression Models and Life Tables (with discussion), *Journal of the Royal Statistical Society*, B34, 187-220

De Servigny A., Oliver R., (2004), Measuring and managing Credit Risk, McGraw Hill, Boston

Dermine J., Neto de Carvalho C., (2006), Bank loan losses given default: A case study, Journal of banking and Finance 30, 1219-1243.

Dias Jose G. (2004), *Finite Mixture Models*, Rijksuniversiteit Groningen

Engelmann B., Rauhmeier R., (2006), The Basel II Risk Parameters, Springer, Heidelberg

Gupton G. (2005), Estimation Recovery Risk by means of a Quantitative Model: LossCalc, *Recovery Risk*, ed by Altman E.I., Resti A, Sironi A. Risk books, London, p61-86

Lucas A.: Basel II Problem Solving; QFRMC Workshop and conference on Basel II & Credit Risk Modelling in Consumer Lending, Southampton 2006;

McLachlan G. J., Basford K. E. (1998), *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.

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

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

Querci F., (2005), Loss Given Default on a medium-sized Italian bank's loans: an empirical exercise, The European Financial Management Association, Genoa, Genoa University. http://www.efmaefm.org/efma2005/papers/206-querci_paper.pdf

Schuermann T., (2005), What Do We Know About Loss Given Default? *Recovery Risk*, ed by Altman E.I., Resti A, Sironi A. Risk books, London, p3-24

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

Wedel M., Kamakura W. A., (2000), *Market Segmentation. Conceptual and Methodological Foundations* (2nd ed.), International Series in Quantitative Marketing, Boston: Kluwer Academic Publishers.

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

Appendix 1

Coefficients of variables in single distribution linear regression model for RR

Variable	Parameter Estimate	Standard Error	P-value
Intercept	0.682	0.029	**
Employment status 1	0.098	0.013	**
Employment status 2	0.144	0.015	**
Mortgage	0.047	0.009	**
Visa card	-0.036	0.010	*
Insurance indicator 2	-0.053	0.009	**
No. of dependant 2	0.027	0.010	*
Personal loan account	0.024	0.008	*
Residential status 1	-0.037	0.011	*
Residential status 3	-0.041	0.017	
Residential status 4	-0.113	0.013	**
Saving account	0.014	0.007	
Loan term1	-0.063	0.019	*
Loan term2	-0.027	0.010	*
Loan term4	0.042	0.011	*
Second applicant status 1	-0.107	0.014	**
Second applicant status 2	-0.051	0.017	*
Second applicant status 3	-0.127	0.009	**
Loan purpose 1	-0.069	0.016	**
Loan purpose 2	-0.040	0.009	**
Loan purpose 3	-0.051	0.012	**
Loan purpose 4	-0.044	0.010	**
Time at address 2	0.033	0.011	*
Time at address 3	0.037	0.010	*
Time at address 4	0.051	0.013	**
Time at address 5	0.066	0.015	**
Time at address 6	0.074	0.015	**
Time at address 7	0.090	0.014	**
Time with the bank 1	-0.030	0.015	
Time with the bank 5	0.032	0.010	*
Time in occupation 1	0.029	0.013	
Time in occupation 2	0.039	0.013	*
Time in occupation 3	0.044	0.015	*
Time in occupation 4	0.047	0.015	*
Time in occupation 5	0.090	0.016	**
Monthly expenditure	0.036	0.016	
Monthly income 1	0.066	0.013	**
Monthly income 2	0.060	0.013	**
Affordability 3	0.057	0.016	*
Default year 90	0.031	0.010	*
Default year 96	0.029	0.011	*
SQR default amount	-0.003	0.000	**
REC default amount	-58.398	8.933	**
Default rate	-0.012	0.001	**

**: $p<0.0001$; *: $p<0.01$

Appendix 2

Coefficients of variables in single distribution Cox regression model (including 0 recoveries) for recovery rate

Variable	Parameter Estimate	Standard Error	P-value
Mortgage	-0.142	0.024	**
Visa card	0.106	0.027	**
Personal loan account	-0.087	0.021	**
Employment status 1	-0.079	0.040	
Employment status 2	0.064	0.033	
Employment status 3	0.328	0.045	**
Insurance indicator 2	0.099	0.030	*
Insurance indicator 3	0.115	0.032	*
Marital status	0.090	0.031	*
No. of dependant	-0.064	0.021	*
Residential status 1	0.092	0.029	*
Residential status 3	0.265	0.029	**
Second applicant status 1	-0.225	0.025	**
Second applicant status 2	-0.145	0.046	*
Loan purpose 1	0.146	0.022	**
Loan purpose 2	0.130	0.026	**
Age of applicant	-0.051	0.024	
Time at address	-0.163	0.023	**
Time in occupation	-0.147	0.024	**
Time with the bank 1	-0.060	0.023	
Time with the bank 2	-0.115	0.030	**
Time with the bank 3	-0.215	0.031	**
Affordability	0.170	0.031	**
Default rate 1	0.090	0.027	*
Default rate 2	0.183	0.028	**
Default rate 3	0.324	0.039	**
Default rate 4	0.340	0.050	**
Default rate 5	0.439	0.052	**
Default amount 1	0.112	0.044	
Default amount 3	-0.068	0.027	
Default amount 4	0.059	0.027	
Default amount 5	0.183	0.040	**
Default amount 6	0.210	0.044	**
Month until default 1	0.120	0.039	*
Month until default 2	0.067	0.027	
Default year 91	0.101	0.027	*
Default year 92	0.082	0.038	
Default year 93	0.116	0.045	
Default year 95	-0.105	0.050	
Default year 96	-0.203	0.044	**
Default year 97	-0.190	0.046	**
Default year 98	-0.216	0.046	**
Default year 99	-0.165	0.064	*

**: $p<0.0001$; *: $p<0.01$