Journal of the Operational Research Society (2010) 61, 411–420

www.palgrave-journals.com/jors/

Modelling credit risk of portfolio of consumer loans

M Malik and LC Thomas*

University of Southampton, Southampton, UK

One of the issues that the Basel Accord highlighted was that, though techniques for estimating the probability of default and hence the credit risk of loans to individual consumers are well established, there were no models for the credit risk of portfolios of such loans. Motivated by the reduced form models for credit risk in corporate lending, we seek to exploit the obvious parallels between behavioural scores and the ratings ascribed to corporate bonds to build consumer-lending equivalents. We incorporate both consumer-specific ratings and macroeconomic factors in the framework of Cox Proportional Hazard models. Our results show that default intensities of consumers are significantly influenced by macro factors. Such models then can be used as the basis for simulation approaches to estimate the credit risk of portfolios of consumer loans.

Journal of the Operational Research Society (2010) **61,** 411–420. doi:10.1057/jors.2009.123
Published online 28 October 2009

Keywords: finance; credit risk; survival analysis; credit scoring

Introduction

The main objective of credit scoring is to develop a scoring system that can correctly rank the customers in terms of their relative default risk so that the customers above some cut-off score are less riskier than those who are below. Credit scoring models can broadly be classified into two types, application scoring and behavioural scoring. The objective of both is to classify whether a customer will default (bad) or not default (good) in a given time period, which leads to estimates of probability of default (PD) of the customer in that period. Application scores are used to predict customers' default risk, say 12 months into the future, at the time an application is made for the loan. Past customers are classified as good or bad based on whether they defaulted, which usually means 90+ days delinquent, during the first 12 months of the starting of the loan. The information available at the time of application is in the form of application variables and credit bureau records. These are used to estimate the probability of being good/bad in the given time period. Behavioural scoring is similar in principal to application scoring except that in behavioural scores we observe the recent payment and purchase behaviour of customers who have been granted loans (say from the last 12 months). This information and the information available from application scoring is used to predict the PD in the next 12 months or some other fixed time horizon. As the name suggests in behavioural scoring the individual's behaviour with a particular lender and on a specific product is considered in addition to the information the lender has through credit bureaus.

**Correspondence: L Thomas, School of Management, University of Southampton, Southampton, Hampshire SO17 1BJ, UK.*
E-mail: L.Thomas@soton.ac.uk

Behavioural scoring is used to make lending decisions on current customers, like increasing/decreasing credit limits, offering new financial products or offering new interest rates. Lenders update their behavioural scores monthly by using the most recent information on their customers.

With the advent of the Basel II banking regulation (BCBS, 2004) it is not just enough to correctly rank customers according to their default risk but also one needs to measure accurately the PD, as the predicted PDs are used to calculate the minimum capital needed to set aside for the portfolio of consumer loans. Moreover, PD has to be predicted not just at an individual level but also for segments of the loan portfolio. The limitation of the above approach of developing scorecards is that it uses a snapshot of customers who joined say during certain months in calendar time (for application score) or who are on books during certain months in calendar time (for behavioural scores). This does not allow for changes in economic conditions and the quality of loans over time. Motivated by the reduced form models for portfolio credit risk in corporate lending (Lando, 1994; Jarrow *et al*, 1997) we seek to exploit the obvious parallels between behavioural scores and the ratings ascribed to corporate bonds to build consumer-lending equivalents. Similar recent studies conducted in corporate credit risk include (Duffie *et al*, 2007) who studied multi-period corporate default prediction with time-varying covariates. They modelled the time-series dynamics of the macroeconomic and firm-specific covariates and combined these with a short-horizon default model to estimate the likelihood of default over several future periods. (Campbell *et al*, 2008) in their recent study did not model the time-series evolution of the predictor variables but instead estimated separate logit models for firm default probabilities at short and long horizons. (Figlewski *et al*, 2007) fitted

Cox intensity models for credit events, including defaults or major upgrades and downgrades in credit rating. Their models incorporated both firm-specific factors related to a firm's credit rating history and a broad range of macroeconomic variables. Their results showed that, in addition to being strongly influenced by ratings related factors, intensities of occurrence of credit events are significantly affected by macroeconomic factors. (Shumway, 2001) used a discrete duration model with time-dependent covariates and demonstrated that hazard models are statistically superior to static models that do not take into account the fact that a firm is exposed to the risk of a credit event over multiple periods. However there has been no work on building duration models for the credit risk of portfolios of consumer loans.

In this paper, we incorporate consumer-specific ratings (behavioural score) and macroeconomic factors in the framework of Cox Proportional Hazard to build a model for customer's default probability in the next 12 months. This uses current information available on an individual along with the values of macroeconomic factors for 1 year ahead. The results of our analysis show that default intensities of consumers are significantly influenced by macroeconomic factors and the time of origination of the loan. The information contained in behavioural score, which is developed on the history of loans that started during a certain period in calendar time, does not capture in full the driving force behind the dynamics of default behaviour. Finally, we will demonstrate how our model of individual consumers default risk can be used to simulate the distribution of defaults in a portfolio of consumer loans.

The development of our model will not affect the rules governing Basel II, but could be expected to have an impact on the way banks segment and stress test their portfolios under its regulations. Such models will also be of considerable use in the segmentation and pricing of portfolios of consumer loans for securitization purposes—an area where the theories of corporate and consumer risk management should, but as yet do not, meet. The failure of the credit rating agencies' models to correctly price residential mortgage backed securities during the credit crunch of 2007/2008 shows how important it is to be able to develop robust models for portfolios of consumer loans.

The idea of employing survival analysis for building credit-scoring models was first introduced by (Narain, 1992) and then developed further by (Thomas *et al*, 1999; Stepanova and Thomas, 2002). Thomas *et al* (1999) compared performance of exponential, Weibull and Cox's nonparametric models with logistic regression and found that survival-analysis methods are competitive with, and sometimes superior to, the traditional logistic-regression approach.

In the next section, we shall briefly discuss the notion of hazard rate and survival probability, and the theory associated with the Cox Proportional Hazard rate model. We then develop a hazard rate model to predict future hazard rates of customers conditional on the information on customers available today. These predicted hazards are then combined to predict the PD for 12 months ahead. Finally, we discuss the results of the simulations to construct the default distribution of portfolio of loans and draw some conclusions.

Preliminaries

Let T be a nonnegative continuous random variable representing the time to default of an individual from a homogeneous population. In survival analysis, the probability distribution of T is described in the following three most popular ways: the survivor function denoted $S(t)$, which gives the probability of surviving beyond time t, the probability density function denoted $f(t)$ and the hazard function (hazard rate) denoted $h(t)$, which is the risk of defaulting at time t, given the person has survived till time t. Mathematically,

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leqslant T < t + \delta t | T \geqslant t)}{\delta t} \right\}$$

$$P(T \leqslant t) = 1 - S(t) = 1 - e^{-\int_0^t h(x)\,dx}$$

The hazard function fully specifies the distribution of T and so determine both the density and the survivor function.

For a heterogeneous population, let $\mathbf{x}(t)$ denote the time covariate vector at time t and $\mathbf{X}(t) = \{\mathbf{x}(u) : 0 \leqslant u < t\}$ specify the path or history of the covariate process up to time t. The components of $\mathbf{x}(t)$ may include fixed covariates measured at time 0 as well as measurements of risk factors on an individual or on the environment that varies over time. Given the covariate path up to time t the conditional hazard function is defined as

$$h(t; \mathbf{X}(t)) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leqslant T < t + \delta t | \mathbf{X}(t), T \geqslant t)}{\delta t} \right\}.$$

In the presence of time dependent covariates the relative risk model (Cox, 1972) describes the hazard rate process as

$$h(t; \mathbf{X}(t)) = h_0(t) \exp[\boldsymbol{\beta}^T \mathbf{Z}(t)]$$

where $\mathbf{Z}(t)$ are functions of $\mathbf{X}(t)$ and t and have left-continuous sample paths, and $h_0(t)$ is an unknown function (baseline) giving the hazard for the standard set of conditions, when $\mathbf{X} = 0$. From the above, it is easy to observe that the Cox model assumes, at any time t, the ratio of the hazard rates of two different individuals does not involve the baseline hazard. In particular, the ratio of hazard rates stay constant over time if the covariates are all time independent. The Cox regression model is therefore often referred to as the proportional hazards model. In our analysis we employ time-dependent covariates, and hence the ratio of hazard rates of two individuals will change over time.

Let $t_1 < t_2 < \cdots < t_n$ be n distinct default times where the individual who defaults at time t_i has characteristics $Z_i(t_i)$ at default. Let R_i be the set of individuals who are at risk of default just before time t_i, $1 \leqslant i \leqslant n$. Then the likelihood function of the observed data is:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\exp[\boldsymbol{\beta}^T \mathbf{Z}_i(t_i)]}{\sum_{l \in R_i} \exp[\boldsymbol{\beta}^T \mathbf{Z}_l(t_i)]}$$

In our data, an individual can leave the risk set either due to default or because of censoring, where the account was closed or the information about the loan reached the end of the sample period. In estimating $h(t; \mathbf{X}(t))$, any event other than default is treated as right-censored. An individual that is censored at time t will contribute to the risk set R_i if $t_i \leqslant t$, and will be excluded if $t < t_i$.

The ith term in the partial likelihood function above is the conditional probability that an individual with covariates $\mathbf{Z}_i(t_i)$ experiences a default event at t_i given the individuals in R_i and that exactly one individual default at t_i.

Cox's original model assumes that the hazard functions are continuous. However, credit performance data are normally recorded only monthly so that multiple defaults at one time can be observed. These are tied default times, and the likelihood function must be modified because it is now unclear which individuals to include in the risk set at each failure time t_1, t_2, t_3, etc. There are two alternative definitions of proportional hazards models for such discrete time situations. Kalbfleisch and Prentice (1980) suggested considering it as an interval censored version of the continuous time version whereas Cox (1972) suggested a linear log odds model. This paper uses the first of these approaches, which means one can continue to use the continuous time notation.

So let d_i denote the number of defaults at time t_i and D_i be the set of individuals defaulted. Let S denote any subset of d_i individuals drawn from the risk set R_i and let S_i be the collection of all these subsets. The generalized version of the likelihood function is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\exp\left[\boldsymbol{\beta}^T \sum_{j \in D_i} \mathbf{Z}_j(t_i)\right]}{\sum_{S \in S_i} \exp\left[\boldsymbol{\beta}^T \sum_{j \in S} \mathbf{Z}_j(t_i)\right]}$$

The summation in the denominator is difficult to calculate, so easier approximations were proposed by (Breslow, 1974) and (Efron, 1977).

Modelling methodology

In this paper, we propose to build a hazard rate model for predicting the PD of a customer in the next 12 months, given all the current information available on a customer along with the values of macroeconomic factors for 1 year ahead. The results of our analysis show that default intensities of consumers are significantly influenced by macro factors and by the (calendar) time when the loan started. It will show that the information contained in macroeconomic variables and the origination quality of loans are major drivers of the dynamics of default probability.

Data description

The original data set contains records of customers of a major UK bank who were on the books as of January 2001 together with all those who joined between January 2001 and December 2005. The data set consists of customers' monthly behavioural score along with the information on their time to default or time to the information being censored. There is no information on the time to default of those individuals who have joined and defaulted before January 2001. For model fitting purpose we consider approximately. 50 000 records of customers' behavioural scores for each month from January 2001–December 2004, along with information on age of loan and default status. We tested our results using customers' performance during 2005 from a hold out sample. In our analysis, individuals who closed the account during the above time period or whose history is truncated at the end of observation period, that is, December 2004, are considered censored. Anyone, who became 90 days delinquent, was charged off or became bankrupt between January 2001–December 2004 is considered as having defaulted. The time to default/censored is measured in months from the start of the loan.

Macroeconomic variables

Traditionally behavioural score models are built on customers performance with the bank over a 12 month or some other fixed-time performance period, using characteristics like average account balance, number of times 30+ days delinquent, etc. In that sense behavioural score can be considered as capturing the idiosyncratic risk of customers. However, it is observed in corporate credit risk models (Lando, 1994; Duffie *et al*, 2007), that though idiosyncratic risk is a major risk factor, during economic slowdowns systemic risk factors emerge and have had a substantial effect on the default risk in a portfolio of loans. In the context of the UK economy, we analyse the following four macroeconomic variables that have been found to affect the default behaviour of consumers in the consumer finance literature (Tang *et al*, 2007) and represents the general economic and investment climate. The variables considered are:

(a) Percentage change in consumer price index over 12 months: reflects the inflation felt by customers where high levels may cause a rise in customer default rate.
(b) Monthly average Sterling inter-bank lending rate: higher values might correspond to general tightness in the economy and increased difficulty in raising cash to make debt service payments.
(c) Annual return on log of FTSE 100: gives the yield from stock market and also reflects the fact that a buoyant stock market may encourage purchasing of financial products.

(d) Percentage growth in GDP for a quarter compared with the comparable quarter in the previous year.

A positive sign of the coefficient of a macroeconomic variable in the hazard equation is associated with the increase in risk of default and vice versa. Unemployment rate was considered in our model but it always entered with a wrong sign and also affected the sign and significance of other macroeconomic covariates. This could be because there is no significant variation in UK's unemployment rate during 2001–2004, so spurious effects are occurring. Therefore we decided to drop unemployment rate from our models. There is a general perception in economics (Figlewski *et al*, 2007) that change in economic conditions does not have an instantaneous effect on default rate. To allow for this effect we considered lagged values of macroeconomic covariates in the form of weighted average over a 6-months period with an exponentially declining weight of 0.88. This choice is motivated by a recent study made by (Figlewski *et al*, 2007).

Hazard rate models

In our analysis, we modelled the hazard rate of customers by incorporating information on:

(a) Age of loan
(b) Behavioural scores
(c) Macroeconomic variables
(d) Time loan taken out (vintage)

If t is time since loan started, the hazard of default at time t for a person i from $Vintage_i$ under current economic conditions $EcoVar_i(t)$ and having behavioural score $BehScr_i(t)$ is

$$h_i(t) = h_0(t)e^{aBehScr_i(t)+bEcoVar_i(t)+cVintage_i}$$

In the above equation, $h_0(t)$ is the baseline hazard that represents the risk due to the age of loan. On could think of the idiosyncratic risk, systemic risk and the risk due to the vintage quality of loans as being represented by $BehScr(t)$, $EcoVar(t)$ and *Vintage*, respectively. Vintages is described for each quarter from 2001 to 2004 as a set of binary variables with a value 1 if an individual starts a loan in a particular vintage and 0 otherwise. The partial likelihood method, first proposed by Cox (1972), estimates the parameters a, b and c in the hazard function without knowledge of the baseline hazard $h_0(t)$. Since in our analysis, we need to predict the PD, say, in next 12 months from now, we use the Nelson–Aalen form of the estimator (Kalbfleisch and Prentice, 1980), to recover the baseline hazard function $h_0(t)$. This is given by

$$h_0(t) = \frac{d_t}{\sum_{l \in R_t} \exp \boldsymbol{\beta}^T \mathbf{Z}_l(t)}$$

where d_t represents the people defaulted at time t and R_t denotes the risk set at time t.

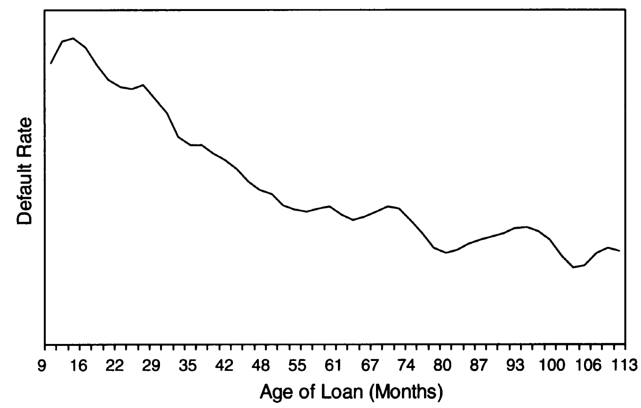

Figure 1 Default rate *versus* age of loan.

Figure 1 shows the movement of default rate with age of loan for 10 years of duration in our data set. It is evident from the graph that propensity of default is high during the early stage of a loan. This goes well with the market assumption that default rates are high during the first 12–24 months of the starting of the loan, which is the reason why many banks use 12–24 months of performance window in their scorecards to classify good/bad customers. However, we also notice from the graph that the default rate does not drop immediately after the first year. Instead, it gradually comes down for up to 8 years, which is the reason why age of loan should not be ignored as one of the factors while developing credit risk models.

The second and most important factor in our model, behavioural score, represents the risk due to the individual's performance with the bank. Behavioural score is related to the PD by a log odds relationship. In our data set behavioural score predicts the PD in the next 6 months. Banks use behaviour score to rank their customers and decide on their policies accordingly. One of the problem in using behavioural score as the sole representation of individual's PD is that the relationship of behavioural score to PD may change over time. It means that the behavioural score of two individuals in different months or the behavioural score of the same individual in different months may not represent the same PD. This is because there is a change in the score to log odds relationship in a credit scorecard over time and is a major motivation behind introducing macroeconomic and vintage variables in our hazard equation.

The point is that the behavioural score, $s_t(x)$, developed at time t_0, which is being used at time t to measure the risk of a borrower with characteristics x, is the sum of two terms where one relates to the population odds at the development time and the other is the information odds of that individual at the current time t, see Thomas (2009).

$$s_t(x) = s_{pop}(t_0) + s_{inf}(x, t)$$

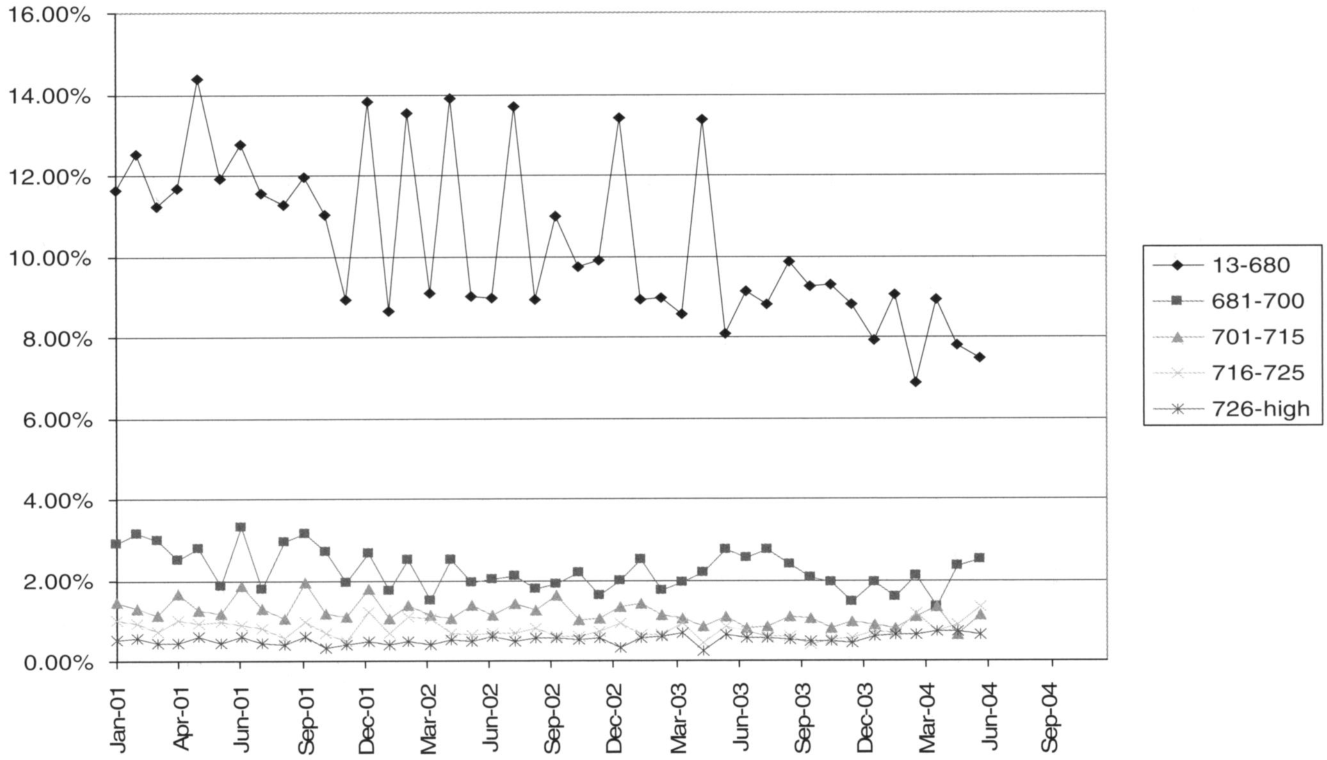

Figure 2 Six-month default rates for different scorecard bands.

In order to apply the log odds relationship between score $\tilde{s}_t(x)$ and PD in the next 6 months from t, $p_t(x)$, namely

$$\tilde{s}_t(x) = \log\left(\frac{1 - p_t(x)}{p_t(x)}\right)$$

then the score that should be used is $\tilde{s}_t(x) = s_{pop}(t) + s_{inf}(x, t)$. So the score needs regular recalibration. Including economic variables is one way of estimating the difference $s_{pop}(t) - s_{pop}(t_0)$ where if the scorecard has not been updated t_0 is the time when the sample on which the scorecard was built was observed. Otherwise it is the time of the last recalibration.

The financial organization reported that the behavioural score had not been recalibrated during the period 2000–2005. Partial confirmation of this can be seen in Figure 2 that shows the 6-month default rates for different scorecard bands. In almost all cases, there is a gradual drop in the default rate over the 4 years, which suggests no major recalibration, as that would result in jumps returning the default rates to their initial values.

To understand the relevance of vintage variables and macroeconomic variables in our model we consider the following example. Suppose two individuals A and B start their loan during different periods in calendar time say the first quarter of 2001 and the first quarter of 2003, respectively. Suppose they both have the same score x after 6 months into the loan. It could be possible that the score x may not represent the same default probability for these two customers.

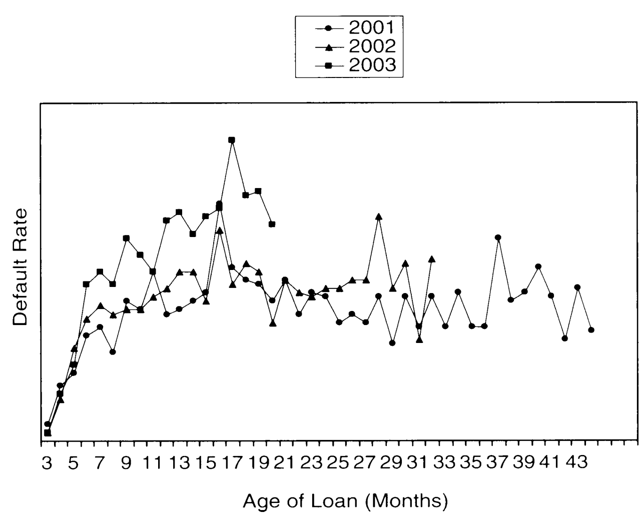

Figure 3 Default rate by year of origination.

Figure 3 shows the default rates of customers based on the year they joined. It seems that customers who joined during 2003 are at higher risk (have higher default rates) than those who joined during 2001–2002. This gives us the intuition to incorporate year of origination of loan to test the dependency of hazard rate on the time the customers joined. To capture this effect we can define binary variables corresponding to whether the loan was started in a specific quarter or not. If these variables come out to be nonzero and significant in the model then it is a clear indication that the hazard rate of two individual's n months into the loan with the same score in month n would depend not only on the score but also on the calendar time during which the individuals started the loan.

As outlined earlier, the economic variables in the hazard function allow for the variations over time of the score to PD relationship. A log odd score gives this relationship at a fixed time—when the score was developed—and the economic variables allow for the dynamics of the relationship so as to get Point in Time (PIT) probability of default (PD) estimates. If one had take a through the cycle (TTC) sample to build the credit scorecard, which is rarely done, then one could argue that the economic variables allow one to translate the score to TTC PD relationship into a score to PIT PD relationship.

Forecasting multi-period hazard rate and default probability

To predict the PD of customers for 1 year ahead in calendar time we first predict their hazard rate in time as the loan started and then transform it into a calendar time for 12 months ahead. This is achieved by developing 12 hazard rate models each with the lagged behavioural score variable in the hazard equation. In particular, for customer i who survived up to t months on books, the hazard rate for k months ahead is given by

$$h_i(t+k) = h_0(t+k)e^{a_k BehScr_i(t) + b_k EcoVar_i(t+k) + c_k Vintage_i}$$

Each hazard model has economic variables taken as a weighted average over the last 6 months. Take, for example, a customer with behavioural scores in December 2003 and who started his loan in January 2003. To predict his hazard rate in February 2004, based on the score in December 2003, we can select parameter estimates from the hazard rate model with a behavioural score of lag 2 and use weighted average of the values of economic variables for the 6 months from August 2003–January 2004. (If one really were in December 2003 then the January 2004 value in this weighted average would be a forecast of the future economic situations). We can then predict the hazard rate for 14 months into the loan, which gives us the predicted hazard rate for a customer in February 2004. We coarse classify the behavioural scores into five bins in order of decreasing risk with people having low behavioural scores, that is, who are at higher risk of default, in the first bin. We also define one special value bin that represents people who do not currently have a behavioural score which means they have only recently become active or have been inactive for the past year. We then define new time-varying covariates BehScore_i ($i = 1, \ldots, 5$) that are binary $(0, 1)$ variables corresponding to the behavioural score bins defined above in terms of worst to best ranking.

Table 1 presents parameter estimates from the two sets of models Model A and Model B for each month up to 12 months ahead. Model A predicts hazard rate using just behavioural score as a time covariate. Model B uses time-varying macroeconomic variables and time of origination of the loan, which is captured by vintage variables, in addition to the behavioural scores. We employ 'stepwise selection' procedure to keep only variables that contribute significantly to the explanatory power of the model. The likelihood ratios and the associated p-values show that incorporating macroeconomic and vintage-level information provides a significant improvement across all periods. However, the improvement in fit of Model B over Model A decreases with the increase in prediction horizon. In Model B, the coefficients of behavioural score bins, leaving aside the special value bin, go from high to low for all periods, which agrees with the risk associated with each bin. The prediction power of behavioural score decreases with the increase in prediction horizon, which is evident from the decrease in the difference of coefficient values between behavioural score bins with each increasing month. However, as was observed for corporate defaults by (Figlewski *et al*, 2007), there has been little change in the value of the estimated coefficients for behavioural score bins for each month as we move from Model A to Model B. This means that the information contained in macroeconomic and vintage variables though significant is incremental to that which is captured by the behavioural score alone. This could be possibly because economic conditions remained approximately static during 2001–2004. Three macroeconomic variables namely Interest rate, GDP growth and CPI (rate of inflation) proved significant at various horizons. The signs of the macroeconomic variables remain consistent throughout the 12 models that forecast forward default rates 1, 2, 3 up to 12 months ahead. Increases in interest rates and increases in the price index lead to increases in default whereas increases in GDP lead to falls in the default rate.

As to vintages, the results show that the default rate of the same behavioural score bands increases the more recent the vintage even allowing for month on books effects. For some reason, be it lender policy, changes in the application population or variations in marketing and targeting policies, the relationship between score and default rate is changing. Note that the sample size decreases as the covariate time lag increases because the minimum period of observation is increasing. It appears though that the effect of vintages gradually diminishes as we increase the forecast horizon.

Suppose $h(1), h(2), h(3), \ldots, h(12)$ represent the monthly default hazard rates of an individual, in the next month, 2 months into the future, 3 months into the future, etc. From the definition of hazard rate and the law of conditional probability, the probability of default in the next 12 months (PD12) can be estimated by $\text{PD12} = 1 - \prod_{n=1}^{12}(1 - h(n))$.

We tested our results on a hold out sample for 2005. We considered people who joined between January 2002–December 2004 and have neither defaulted nor closed their accounts as of December 2004. We then estimate their PD12 for 2005 using the actual economic values that occurred

Table 1 Coefficients of Models A and B

	1-month predictive model Parameter estimates		*2-month predictive model Parameter estimates*		*3-month predictive model Parameter estimates*		*4-month predictive model Parameter estimates*		*5-month predictive model Parameter estimates*		*6-month predictive model parameter estimates*	
	Model A	*Model B*	*Model A*	*Model B*	*Model A*	*Model B*	*Model A*	*Model B*	*Model A*	*Model B*	*Model A*	*Model B*
BehScore bins												
BehScorer1	−2.24671	−2.3323	−2.6242	−2.70791	−3.09697	−3.18273	−2.75823	−2.84724	−2.48733	−2.56478	−2.61022	−2.67749
BehScorer2	−8.54738	−8.66029	−6.98778	−7.09739	−5.62776	−5.73832	−4.9138	−5.02491	−4.08235	−4.18073	−3.90613	−3.99306
BehScorer3	−9.60406	−9.71268	−8.08176	−8.19148	−6.08841	−6.20220	−5.57769	−5.69440	−4.84958	−4.95517	−4.50302	−4.60184
BehScorer4	−10.03671	−10.14294	−9.68832	−9.79372	−6.96915	−7.07702	−5.72644	−5.83734	−5.03777	−5.13704	−4.83499	−4.92699
BehScorer5	−11.5715	−11.66931	−11.9021	−11.99937	−7.60843	−7.71063	−6.20157	−6.30834	−5.42494	−5.52223	−5.22294	−5.31600
SpecialValue	−6.01966	−6.11543	−6.47088	−6.56873	−5.80204	−5.90250	−5.1372	−5.24298	−4.58765	−4.68248	−4.48938	−4.57994
Macroeconomic factors												
Interest Rate		0.26229		0.25069		0.24416		0.22181		0.22013		0.23729
GDP		−0.42539		−0.38256		−0.31511		−0.24414		−0.27600		−0.34237
CPI												
Vintages												
VintQrt1		−1.21870		−1.15786		−0.99363		−0.79594		−0.74695		−0.79301
VintQrt2		−1.13194		−1.07397		−0.96448		−0.80352		−0.74484		−0.79143
VintQrt3		−1.01896		−0.92263		−0.77844		−0.65098		−0.64334		−0.70042
VintQrt4		−0.75628		−0.68562		−0.59465		−0.50605		−0.46615		−0.49181
VintQrt5		−0.81972		−0.77152		−0.71633		−0.65304		−0.60409		−0.66492
VintQrt6		−0.73978		−0.69347		−0.62936		−0.62144		−0.60719		−0.64168
VintQrt7		−0.51075		−0.44722		−0.39576		−0.35188		−0.34361		−0.36271
VintQrt8		−0.38299		−0.31696		−0.26262		−0.24582		−0.21872		−0.23177
VintQrt9		−0.26098		−0.22958		−0.18531		−0.21149		−0.19159		−0.19938
VintQrt10		−0.13124										
Vintage11												
Vintage12												
Vintage13										0.19690		
Vintage14												
VintQrt15		0.77901		1.00105								
Likelihood ratio test: Model B *versus* Model A		218.5007		213.7275		175.9484		139.9434		122.2438		109.3105
p-value		<0.0001		<0.0001		<0.0001		<0.0001		<0.0001		<0.0001

Table 1 (continued)

	7-month predictive model Parameter estimates		*8-month predictive model Parameter estimates*		*9-month predictive model Parameter estimates*		*10-month predictive model Parameter estimates*		*11-month predictive model Parameter estimates*		*12-month predictive model Parameter estimates*	
	Model A	*Model B*	*Model A*	*Model B*	*Model A*	*Model B*	*Model A*	*Model B*	*Model A*	*Model B*	*Model A*	*Model B*
BehScore bins												
BehScorer1	−2.3721	−2.43000	−2.45326	−2.49524	−2.4375	−2.48732	−2.42873	−2.48829	−2.33096	−2.38991	−2.26133	−2.32921
BehScorer2	−3.60232	−3.67689	−3.54707	−3.60116	−3.44737	−3.50893	−3.33606	−3.40538	−3.25984	−3.32759	−3.12304	−3.19961
BehScorer3	−4.13394	−4.21920	−3.94328	−4.00594	−3.94487	−4.0159	−3.78145	−3.85942	−3.59847	−3.67288	−3.51835	−3.60081
BehScorer4	−4.36663	−4.44559	−4.35139	−4.40785	−4.27954	−4.34375	−4.05897	−4.13097	−4.05529	−4.12304	−3.9288	−4.00519
BehScorer5	−4.85296	−4.93306	−4.81405	−4.87324	−4.71469	−4.78202	−4.61785	−4.69305	−4.46676	−4.54106	−4.32225	−4.40504
Special Value	−4.1005	−4.17712	−4.06151	−4.11792	−3.99174	−4.05697	−3.89062	−3.96378	−3.78074	−3.85201	−3.66479	−3.74210
Macroeconomic factors												
Interest Rate		0.24696		0.27168		0.26628		0.33990		0.24378		0.37177
GDP		−0.28462						−0.10516		−0.29266		−0.14673
CPI				0.39202		0.4977		0.63936				
Vintages												
VintQrt1		−0.52755										
VintQrt2		−0.54694										
VintQrt3		−0.50450										
VintQrt4		−0.27505		0.10792								0.13935
VintQrt5		−0.46768		−0.14904		−0.16601		−0.16383				−0.19260
VintQrt6		−0.45223		−0.18930		−0.21061		−0.17845				−0.17424
VintQrt7		−0.18336								0.32766		
VintQrt8										0.33385		
VintQrt9										0.35307		
VintQrti10										0.51792		
Vintage11										0.45821		
Vintage12												
Vintage13		0.51457		0.66400								
Vintage14												
VintQrti15												
Likelihood ratio test: Model B *versus* Model A		98.2929		60.3232		51.8505		52.2771		59.3396		54.3668
p-value		<0.0001		<0.0001		<0.0001		<0.001		<0.0001		<0.0001

Table 2 Results of out of time testing

Model	*Test sample size*	*Actual no. of defaults in 2005*	*Expected no. of defaults in 2005*
Model A	14 091	959	565
Model B	14 091	959	1022

that year. The results of our analysis are shown in Figure 3 and Table 2 below.

One way of validating the models is by their discriminating ability. Given a model, one can order each of the loans in the sample according to their predicted default value PD12 for 2005. This gives a ranking of how likely each loan is to default and this can be compared with which ones actual did then default. The results are shown in the Receiver Operating Characteristic (ROC) curve in Figure 3, where each point on the curve comes from a predicted PD value. Its horizontal distance is the proportion of the actual defaults with predicted PDs above this value; the vertical distance is the proportion of the actual non-defaults with predicted PDs above the value. The closer the curve is to the ideal point (1,0), where all the actual defaults have higher predicted PD values than the non-defaults, the better the discrimination. Figure 3 shows that there is very little difference in discriminating power between Models A and B, though Model B is slightly superior. This is not surprising because the economic variables are the same for all loans in the sample and so do not affect the rankings. Even the vintage variables only change the ranking a little.

Table 2, on the other hand, describes how well the models predict the total default level of the portfolio. The predicted PD of each loan summed over all the loans in the sample gives the model's prediction of the expected number of defaults in the portfolio. Table 2 shows that Model B gives a much closer prediction (1022) to the actual number of defaults (959) than Model A. Thus the economic variables have a considerable effect on improving the prediction of the portfolio level default rate and do seem to be one way of modelling the dynamics of the changes in the default rate from year to year.

Portfolio default distribution

In a recent paper Duffie *et al* (2007) studied the time series dynamics of the covariate processes in the context of multi-period corporate default prediction. Motivated by this we model interest rate, GDP and CPI as an independent autoregressive time series of order 1 and simulate the default distribution of loans from a hold out sample for the year 2005. We assume that given the value of economic variables for a fixed time period in calendar time, the defaults are conditionally independent. Table 3 shows the parameter estimates for each of the macroeconomic variable in our model that uses the following specification

$$y_t - \mu = \rho(y_{t-1} - \mu) + \sigma e_t,$$

Table 3 Estimates from an AR(1) model for macroeconomic covariates

Variable	*Mean* (μ)	*Coefficient* (ρ)	*Standard Error* (σ)
Interest rate	5.656389	0.98704	0.209164
GDP	2.925	0.93775	0.212757
CPI	1.465741	0.87823	0.237134

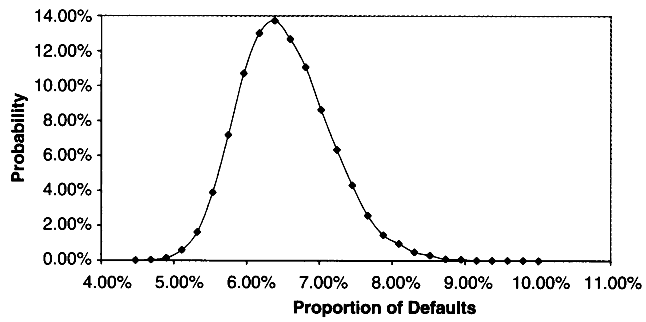

Figure 4 Simulated distribution of proportion of defaults in a loan portfolio.

where μ is the estimated mean of the series y_t and e_t are independent standard normal errors.

The simulation steps to estimate the default distribution of portfolio of loans are explained below.

(1) Generate a sample path of the three macroeconomic variables above for 12 months from January 2005 taking initial value Y_0 in AR(1) equation from December 2004 and independently drawing 12 random variables corresponding to error terms from standard normal distribution.
(2) Calculate the hazard rate and hence PD12 using Model B for each customer in the hold out sample for 2005 using the sample path of macroeconomic values generated in Step 1 above.
(3) Generate random numbers from a uniform distribution between 0 and 1 corresponding to each observation.
(4) Define an individual as default (1) if PD12 $>$ random value and non-default (0) if PD12 $<$ random value.
(5) Calculate the total number of defaults in Step 4.
(6) Repeat steps 1, 2, 3 and 4 with random seed values and draw the distribution of proportion of defaults.

Figure 4 shows the smoothed default distribution of proportion of defaults for 2005 in a hold out sample of approximately 14 000 customers. The peak of the distribution is at a default rate of approximately 6.5%. The distribution is though somewhat asymmetric with higher probabilities of high default rates. The simulation suggest that there is a 0.1% probability of having default rates above 10%.

One can use more sophisticated models of the time series dynamics of macroeconomic variables and still apply the above simulation procedure using the hazard rate model we have developed. Our objective is to show such simulations can be undertaken as there are many choices of macroeconomic time series models, see (Duffie *et al*, 2007).

Conclusions

In this paper, we build a proportional hazard rate model for customers' default probability for up to 12 months ahead using recent information on customers' behavioural scores along with the values of general macroeconomic factors. We do not predict customers' future behavioural scores as we believe the current behavioural score contains the best estimate of future default risk based only on customer-specific information. Instead, we can use existing models for predicting the time series dynamics of macroeconomic factors that are the drivers of the dynamics of default behaviour. We estimated 12 hazard rate models for each month up to 12 months ahead by lagging behavioural score covariate. Our major conclusion is that inclusion of macroeconomic factors and time of origination improves the default predictions significantly. However, it does not have a major impact on the discriminating ability of the model beyond that given by the behavioural score. Such models lend themselves to building credit risk estimates of portfolios of consumer loans in which the correlations between defaults of different loans are given by changes in the macroeconomic conditions.

Acknowledgements—This work was supported by an EPSRC grant that pump primed the Quantitative Financial Risk Management Centre. We are grateful to the referees for their useful suggestions on an earlier version of this paper.

References

Basel Committee on Banking Supervision (BCBS) (2004). *International Convergence of Capital Measurement and Capital of Capital Standards: A revised framework*. Bank of International Settlements: Basel.

Breslow NE (1974). Covariance analysis of censored survival data. *Biometrics* **30**: 89–99.

Campbell J, Hilscher J and Szilagyi J (2008). In search of distress risk. *J Financ* **63**: 2899–2939.

Cox DR (1972). Regression models and life-tables (with discussion). *J Roy Stat Soc* **B 74**: 187–220.

Duffie D, Saita L and Wang K (2007). Multi-period corporate default prediction with stochastic covariates. *J Financ Econ* **83**: 635–665.

Efron B (1977). The efficiency of Cox's likelihood function for censored data. *J Am Stat Ass* **72**: 557–565.

Figlewski S, Frydman H, Liang W (2007). *Modelling the effect of macroeconomic factors on corporate default and credit rating transitions*. Working Paper no. FIN-06-007. NYU Stern School of Business.

Jarrow R, Lando D and Turnbull S (1997). Pricing derivatives on financial securities subject to credit risk. *J Financ* **50**: 53–86.

Kalbfleisch JD and Prentice RL (1980). *The Statistical Analysis of Failure Time Data*. Wiley: New York.

Lando D (1994). *Three essays on contingent claims pricing*. PhD thesis, Cornell University, Ithaca, NY.

Narain B (1992). Survival analysis and the credit granting decision. In: Thomas LC, Crook JN and Edelman DB (eds). *Credit Scoring and Credit Control*. OUP: Oxford, pp 109–121.

Shumway T (2001). Forecasting bankruptcy more accurately: A simple hazard model. *J Bus* **74**(1): 101–124.

Stepanova M and Thomas LC (2002). Survival analysis methods for personal loan data. *Oper Res* **50**: 277–289.

Tang L, Thomas LC, Thomas S and Bozzetto J-F (2007). It's the economy stupid: Modelling financial product purchases. *Int J Bank Market* **25**: 22–38.

Thomas LC (2009). *Consumer Credit Models: Pricing, Profit and Portfolios*. Oxford University Press: Oxford.

Thomas LC, Banasik J and Crook JN (1999). Not if but when will borrowers default. *J Opl Res Soc* **50**: 1185–1190.

Received December 2007; accepted August 2009 after two revisions