Faculty of Economics and Business
The University of Sydney

OME WORKING PAPER SERIES

Survival Analysis for Credit Scoring: Incidence and Latency

John Watkins
Faculty of Economics and Business
The University of Sydney

Andrey Vasnev
Faculty of Economics and Business
The University of Sydney

Richard Gerlach
Faculty of Economics and Business
The University of Sydney

Abstract

Duration analysis is an analytical tool for time-to-event data that has been borrowed from medicine and engineering to be applied by econometricians to investigate typical economic and finance problems. In applications to credit data, time to the pre-determined maturity events have been treated as censored observations for the events with stochastic latency. A methodology, motivated by the cure rate model framework, is developed in this paper to appropriately analyse a set of mutually exclusive terminal events where at least one event may have a predetermined latency. The methodology is applied to a set of personal loan data provided by one of Australia's largest financial services institutions. This is the first framework to simultaneously model prepayment, write off and maturity events for loans. Furthermore, in the class of cure rate models it is the first fully parametric multinomial model and the first to accommodate for an event with pre-determined latency. The simulation study found this model performed better than the two most common applications of survival analysis to credit data. In addition, the result of the application to personal loans data reveals particular explanatory variables can act in different directions upon incidence and latency of an event and variables exist that may be statistically significant in explaining only incidence or latency.

November 2009

OME Working Paper No: 03/2009
http://www.econ.usyd.edu.au/ome/research/working_papers

Survival Analysis for Credit Scoring: Incidence and Latency

John Watkins, Andrey Vasnev and Richard Gerlach

November 29, 2009

Duration analysis is an analytical tool for time-to-event data that has been borrowed from medicine and engineering to be applied by econometricians to investigate typical economic and finance problems. In applications to credit data, time to the pre-determined maturity events have been treated as censored observations for the events with stochastic latency. A methodology, motivated by the cure rate model framework, is developed in this paper to appropriately analyse a set of mutually exclusive terminal events where at least one event may have a predetermined latency. The methodology is applied to a set of personal loan data provided by one of Australia's largest financial services institutions. This is the first framework to simultaneously model prepayment, write off and maturity events for loans. Furthermore, in the class of cure rate models it is the first fully parametric multinomial model and the first to accommodate for an event with pre-determined latency. The simulation study found this model performed better than the two most common applications of survival analysis to credit data. In addition, the result of the application to personal loans data reveals particular explanatory variables can act in different directions upon incidence and latency of an event and variables exist that may be statistically significant in explaining only incidence or latency.

1 Introduction

Credit scoring and risk assessment of retail credit is dominated by logistic and probit regression techniques. These models are most commonly employed to establish the incidence of default over a twelve month time horizon [4, Altman & Saunders (1998) pp. 1723] [23, Crook, Edelman & Thomas (2007) pp. 1448]. Bucay & Rosen (2000) employ pseudo logistic and probit regression techniques to analyse revolving credit. More recently, researchers have investigated the use of survival analysis as a model to assess credit risk. Applications such as Andreeva (2006) and Stepanova & Thomas (2002) apply the time-to-event analysis technique to the credit risks of prepayment and default separately.

The papers of Banasik, Crook & Thomas (1999), Stepanova & Thomas (2002), Andreeva (2006) and Bellotti & Crook (2007) each examine prepayment and default individually. The models treat all other failure times as censored observations for the event of interest. These risks are examined simultaneously in the papers of Deng, Quigley & Van Order (2000), Pavlov (2001) and Ciochetti, Deng, Gao & Yao (2002). Deng et al (2000) emphasise the importance of the jointness of the decision to default or prepay on mortgages. The event of loan maturity is incorrectly treated as a censored observation in all previous research. The framework common to these latter three papers used to simultaneously analyse the time to prepayment and default is developed in the papers of Han & Hausman (1990), Sueyoshi (1992) and McCall (1996) which has been coined HHSM. The work in the HHSM series of papers develops a proportional hazards survival analysis framework for the examination of labour market problems. The factors affecting the time to transition to non-terminal employment states are assessed using this framework.

In credit data, the events of prepayment, maturity and write off are terminal. Although the simultaneous estimation of the prepayment and write off risks is important, the treatment of the maturity events has not been adequate. A class of mixture models, known as Cure Rate models in the medical literature, provide motivation for the model developed in this paper to address this issue. This class of survival analysis model mixes a binary distribution, most commonly logistic, with a typical distribution used for the analysis of failure time data. The methodology was pioneered as early as the 1950's in the papers of Boag (1949) and Berkson & Gage (1952) for the analysis of the fraction of patients cured after experiencing cancer therapies. The methodology has continued to be used in the papers of Farewell (1982), Sy & Taylor (2000), Peng & Dear (2000) and Cancho, Bolfarine & Ortega (2008). The use of such models in analysis of failure time data typical in medical research is motivated by a biological possibility of cure and often evidenced by heavy censoring and Kaplan-Meier (KM) non-parametric survival function estimates which plateau to values strictly greater than zero [46, Sy & Taylor (2000) p. 22]. These papers add additional complexities to the methodology and extend this class of models to the non-parameteric sphere of analysis.

The paper of Hoggart & Griffin (2001) uses the cure rate methodology to analyse the problem of customer attrition in the banking industry. The authors adopt the Bayesian cure rate methodology developed by Chen, Ibrahim, & Sinha (1999). In this framework there are N independent and identically distributed (iid) risks which may cause the event under study to occur. The risks follow a Poisson distribution with constant mean and a Bayesian partition method is used to assess the binary cure rate model. The research in the paper of Cancho, Bolfarine & Ortega (2008) uses the same framework in analysis of a clinical study on cancer patients. Tsodikov, Ibrahim & Yakovlev (2003) extend this framework of Chen et al (1999) to a multinomial non-parametric Bayesian cure rate methodology. In addition, the authors argue that extending the cure rate model to a multinomial parametric methodology would be theoretically and computationally cumbersome. However, the work in this paper reveals this is not the case, at least in the case of credit data.

The methodolody developed in this paper contributes to the current literature in three ways:

i.) the seminal work of Deng, Quigley & Van Order (2000) is extended to the simultaneous estimation of prepayment, write off and maturity events;

ii.) the methodology is the first fully parametric multinomial model in the class of cure rate models, and;

iii.) the model is the first in its class to allow for the simultaneous modelling of a set of mutually exclusive events, where one of the event's duration times may be non-stochastic or pre-determined.

The model is applied to a unique data set of over one million personal loan observations provided by one of Australia's largest financial services organisations. The extent to which the Option Theoretic and the Permanent Income Hypothesis influence the debtor's loan termination decision in the Australian market are explored. The application of this methodology simultaneously estimates parameters for both incidence and latency of credit events, allowing the flexibility in framework to account for a variable that does not operate in the same direction on an event's incidence and latency.

Through an empirical study we find that there are results where variables influence the incidence and latency of an event in opposite directions. Furthermore, these results are logically consistent with the expected behaviours of debtors. In addition, through the simulation study we find that the methodology developed in this paper is superior at estimating the true parameter values and does not suffer from biases caused by treating maturity observations as censored prepayment and default events.

The rest of the paper is divided into the following sections: Section 2 examines survival analysis methodologies and cure rate models in empirical applications; Section 3 develops the model; Section 4 presents the results of the empirical and simulation study; and, Section 5 concludes.

2 Motivation, Development and Model

Loan terminations can be grouped into three broad categories of prepayment, maturity and write off. Deng et al (2000) argue prepayment and default is a consequence of debtors exercising "in-the-money" call and put options, respectively, on their debt facility. Their empirical study examines data from the US market and found that unobserved heterogeneity, including the degree of debtor's financial savvy, are important determinants of loan termination. The authors also observe many non-optimal option exercises and attempt to account for these events by including control variables such as national divorce rate figures.

Exercising in-the-money put and call options of write off and prepayment, respectively, are also reasons for debtors to terminate their personal loans in the Australian market. However, the Australian market has significant structural barriers to option exercise, not present in the US market. These features include full recourse loans (liability is not limited to the mortgaged asset), heavy penalties against future borrowing in the event of write off and early repayment adjustments on fixed rate products in addition to early exit fees. Despite these financial penalties prepayment and write off events are still observed. In addition, interest rates were increasing over the period of data collection, meaning there was no optimal point to exercise prepayment options in order to refinance. It is believed that consumption optimisation of the debtor is the main reason for observing prepayment events, whilst write off events are due to severe shocks to debtor income. The permanent income hypothesis, pioneered by Milton Friedman, provides motivation to the observation of prepayment and write off events through shocks to expected income paths and changes to subjective intertemporal discount factors. The papers of Carroll (2001) and Browning & Crossley (2001) provide a summary of research realted to the permanent income hypothesis of Milton Friedman.

In recent years focus has been placed on retail credit risk assessment techniques with the advent of the Basel II Capital Accord, which is a set of guiding principles for Austhorised Deposit-taking Instituations (ADIs) stipulating minimum standards and requirements for in house risk assessment methodologies. Specific “loss characteristics” such as probability of default (PD), exposure at default (EAD) and loss given default (LGD) are defined with specific measurement techniques. The criteria of the Basel II Capital Accord are met most commonly by methodologies such as logistic and probit regression. Survival analysis techniques also meet the requirements for measuring PD, however, as succintly explained in Bansik et al (1999) they "answer not only if, but when" these events will happen. This facet of survival analysis ensures it is useful for profit scoring, measuring EADs and matching the term of a banks funding with that of their asset protfolio.

The fundamental quantity under assessment is time to event data and from a risk assessment perspective, the event of interest may be default or write off, where the failure time would be measured from loan origination to loan closure. The set of observable failure times exists in the set of non-negative reals, such that $T = \{t : t \in R^{+}\}$. Each observed failure time, t_i, is believed to be a random variable with a probability density function (pdf), $f(t)$. The cumulative density function (cdf), $F(t)$, is used to define the Survival Function, such that $S(t) = 1 - F(t)$. The focus of many applications is to estimate the distribution for the failure time variable, however, non-parametric estimation techniques are also frequently used.

In the papers of Banasik et al (1999), Stepanova & Thomas (2002), Andreeva (2006) and Bellotti & Crook (2007), the authors apply duration analysis to credit data, treating the events of prepayment and default as indendepent. The observed maturity events have been treated as censored prepayment and default event times. Under this independent competing risks assumption the prepayment and default observations are analysed separately, treating all other observed failure times as censored default or prepayment times, respectively. The likelihood function across $i = 1, ..., N$ observations is:

$$L(\theta) = \prod_{i=1}^{N} f(t)^{1-\delta_i} \, S(t)^{\delta_i} \tag{1}$$

where δ_i takes the value of 1 or 0 if censored or uncensored, respectively.

Deng, Quigley & Van Order (2000) extend the framework to simultaneously model the events of prepayment and default. This methodology was originally developed in the series of seminal papers HHSM. The HHSM authors developed this framework specifically for the time-to-event data typical in labour market economic problems, which is characterised by transitional events and stochastic time processes for every event. Deng et al (2000) augment the methodology for the terminal event times of interest in credit market problems. The framework developed in the Deng et al (2000) paper was applied in the research of Pavlov (2001) and Ciochetti, Deng, Gao & Yao (2002). The data is split into the mutually exclusive sets of prepayment, default, censoring and unkown event types. The set of censored events contains all maturity observations. The log-likelihood function ($\mathcal{L}(\theta)$) maximised across the observations $i = 1, ..., N$ can be written most simply as:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \{\delta_{Pi} \ln [F_P(t_i)] + \delta_{Di} \ln [F_D(t_i)] + \delta_{Ui} \ln [F_U(t_i)] + \delta_{Ci} \ln [F_C(t_i)]\} \quad (2)$$

where $F_j(t_i)$ for $j = P, D, U, C$ are the probabilities of mortgage termination due to (P)repayment, (D)efault, (U)nkown reason and (C)ensoring, respectively. The δ_{ji} for $j = P, D, U, C$ are indicator variables taking value of unity when the i^{th} individual experiences event j.

Current applications of survival analysis to credit data treat the terminal pre-determined maturity event observations as censored prepayment and default events. As subsequently shown, this treatment can lead to bias in the parameter estimates. The class of models known as Cure Rate Models offers motivation for the solution developed in this paper. The methodology was developed in response to the possibility of cure given the biology of the disease under study, as evidenced by non-parametric survival function estimates that plateau to non-sero values and heavy right censoring, as discussed in Sy & Taylor (2000). The Cure Rate Models were pioneered in the work of Boag (1949) and Berkson & Gage (1952). These models are a class of mixture models, where most frequently a binary distribution is mixed with a typical failure time data distribution with support on $\mathbb{R}^+$.

The cure rate models are applied to time-to-event data where there are individuals susceptible and insusceptible to the risks under study. In addition, it is not known ab initio to which group an individual belongs. Tsodikov, Ibrahim & Yakovlev (2003) define the surviving proportion as the non-zero asymptoic value, p, of the survivor function, $\overline{S}(t)$, as t tends to infinity and T is the survival time with cdf $S(t) = 1 - \overline{S}(t)$.

$$p = \lim_{t \to \infty} \overline{S}(t) = \exp\left\{-\int_0^\infty \lambda(u)\, du\right\} \quad (3)$$

where $\lambda(u) = f(u)/S(u)$ is the hazard function.

This framework leads to what has largely been labelled as the two-component (binary) mixture model and Tsodikov et al (2003) show it can be generalised as:

$$\overline{S}(t) = E\left\{\left[\overline{S}(t|M=1)\right]^M\right\} = (1-p) + p\overline{S}(t|M=1) \quad (4)$$

where M is a binary variable taking values 0 and 1 with probability $(1-p)$ and p, respectively. The surviving fraction is $(1-p)$ and the incidence of susceptible individuals is p with latency described by the conditional survival function, $\overline{S}(t|M=1)$.

Hoggart & Griffin (2001) apply the cure rate methodology to the empirical study of time to customer attrition from banks. The method used in this paper assumes there are N iid poisson risks with mean θ, resulting in the probability that an individual is insusceptible to attrition being parameterised as $\exp\{-\theta\}$. This method is also applied to clinical data on patients suffering from cancer in Cancho, Bolfarine & Ortega (2008). Farewell (1982) parameterises the incidence proportion using logistic regression and the latency distribution using the Weibull density function. Sy & Taylor (2000) and Peng & Dear (2000) develop semi-parametric techniques for the binary cure rate model. Tsodikov et al (2003) develop non-parametric and semi-parametric Bayesian multinomial methods for cure rate models. In the following section a fully-parametric model incorporating cure rate techniques is developed.

3 Modle for Simultaneous Estimation of Prepayment, Maturity and Default

There are three terminal and mutually exclusive events of maturity, write off and prepayment. Let the set of labels for these observable permanent events be respectively:

$$\mathbf{M} = \{0, 1, 2\}$$

The observed time to each event for an account is represented by $\widetilde{T}_{ij}$, where $j = 0, 1,$ or 2 to indicate the event type and $i = 1, ..., N$ indicates the i^{th} individual. This variable is calculated as the time from loan origination to the time the account experiences an event in set $\mathbf{M}$. The support for these variables is outlined below. First, let $\overline{a}$ ="days to maturity" such that Pr $\left[\widetilde{T}_{i0} = \overline{a}\right] = 1$. Then we can define:

$$\widetilde{T}_{i0} = \overline{a}, \quad \widetilde{T}_{i1} \in [0, \infty) \quad \text{and} \quad \widetilde{T}_{i2} \in [0, \overline{a}) \tag{5}$$

Define $\widetilde{\mathbf{q}}$ as the vector of labels for the N individuals under observation. The i^{th} element of $\widetilde{\mathbf{q}}$, $\widetilde{q}_i$ for $i = 1, \ldots, N$, takes the label from set $\mathbf{M}$ corresponding to the observed terminal event.

Three binary indicator variables are defined to signal when a failure time for the events in set $\mathbf{M}$ is observed for individual i. Let:

$$y_{ij} = \left\{ \begin{array}{ll} 1 & \text{if } \widetilde{q}_i = j \\ 0 & \text{otherwise} \end{array} \right\}, \quad \text{for} \quad j \in \mathbf{M} \tag{6}$$

The density of $\widetilde{\mathbf{q}}$ over the observed failures follows a multinomial distribution which can be characterised as $\prod_{i=1}^{N} \prod_{j=0}^{2} p_{ij}^{y_{ij}}$.

The probability of incidence for each event is: $\Pr(y_{ij} = 1) = p_{ij} = F_j(\mathbf{x}_i, \boldsymbol{\beta})$; where $\mathbf{x}_i$ and $\boldsymbol{\beta}$ are $(k \times 1)$ column vectors of individual specific regressors and corresponding coefficients, respectively. The function, F_j, must satisfy the following conditions: $p_{ij} \in [0, 1]$ and $\sum_{l=0}^{2} p_{il} = 1$. These restrictions ensure that the p_{ij} satisfy the properties of probabilities for a set of mutually exclusive events.

The functional form of F_j will be chosen to be the alternative-invariant form of the Multinomial Logit (MNL). The MNL is characterised as:

$$F_j\left(\mathbf{x}_i, \boldsymbol{\beta}\right) = \frac{\exp\left(\mathbf{x}_i^T \boldsymbol{\beta}_j\right)}{\sum_{l=0}^{2} \exp\left(\mathbf{x}_i^T \boldsymbol{\beta}_l\right)} \tag{7}$$

In addition, the identification restriction translates to setting the parameters for one alternative (maturity) in the MNL to the null vector. The incidence of maturity becomes the base category for comparison in relative risk assessments.

The observed failures, $\widetilde{T}_{ij}$, are conditionally iid across i with pdf $f_j\left(t \mid \mathbf{x}_i, \phi_j, y_{ij} = 1\right)$ for $j = 0, 1, 2$, where ϕ_j is the set of parameters for distribution f_j, and $\mathbf{x}_i$ and y_{ij} have been defined above. In the case that $j = 0$, we find that $\Pr\left[\widetilde{T}_{i0} = \overline{a}\right] = 1$. In the case that $j = 1$ or 2, a density with support over the positive reals is used to define the latency to these events.

The methodology also deals with censored observations. Let C_i be the time to censoring for the i^{th} individual where $i = 1, \ldots, N$. The censoring time is measured from loan origination to the time data collection ceased. Each individual in the sample will have a censoring time, however, only a subset of individuals will have censoring times without an observed failure time. The time to an event or censoring is defined as:

$$T_i = \widetilde{T}_{ij} \wedge C_i \tag{8}$$

Let $\mathbf{T} = (T_1 \cdots T_N)'$ be the vector of failure and censoring times for all individuals. Correspondingly, a binary indicator variable is defined to signal if an event has an observed event time or is still active. Let the indicator be:

$$\delta_i = \begin{cases} 1 & \text{if } \widetilde{T}_{ij} \leq C_i \\ 0 & \text{if } \widetilde{T}_{ij} > C_i \end{cases} \tag{9}$$

Under this "right" censoring mechanism, the C_i for $i = 1, \ldots, N$, are iid random variables across i with pdf and cdf of v_i and V_i, respectively. Conditional on the observed regressors for individual i, the data pairs (T_i, δ_i) are independent. The censoring mechanism in this data set is consistent with definitions of non-informative censoring mechanisms detailed in Kalbfleisch & Prentice (2002).

The uncensored events of the data set are observed with probability:

$$\begin{aligned} &\Pr\left[T_i \in [t, t+dt), \delta_i = 1 \mid \mathbf{x}_i, \phi_j\right] \\ &= \Pr\left[C_i \geq t + dt\right] \Pr\left[y_{ij} = 1\right] \Pr\left[\widetilde{T}_{ij} \in [t, t+dt) \mid \mathbf{x}_i, \phi_j, y_{ij} = 1\right] \\ &\simeq \left[1 - V_i(t)\right] p_{ij} f_j\left(t \mid \mathbf{x}_i, \phi_j, y_{ij} = 1\right) dt \quad \text{for } j = 1, 2 \end{aligned} \tag{10}$$

In the case where $j = 0$, we obtain for the maturity event:

$$\Pr\left[y_{i0} = 1\right] \Pr\left[T_i \in [t, t+dt), \delta_i = 1 \mid \mathbf{x}_i, \phi_j, y_{i0} = 1\right] \simeq V_i(t)\, p_{i0} \tag{11}$$

Whilst the censored event time observations are observed with probability:

$$\Pr\left[T_i \in [t, t+dt), \delta_i = 0 \mid \mathbf{x}_i, \phi_j\right]$$

$$
\begin{aligned}
&= \Pr\left[C_i \in [t, t+dt)\right] \sum\nolimits_{j=0}^{2} \Pr\left[y_{ij}=1\right] \Pr\left[\widetilde{T}_{ij} \geq t \mid \mathbf{x}_i, \phi_j, y_{ij}=1\right] \\
&\simeq v_i(t) \left\{ p_0 + \sum\nolimits_{j=1}^{2} p_{ij} S_j\left(t \mid \mathbf{x}_i, \phi_j, y_{ij}=1\right) \right\} dt \quad \text{for } j=1,2
\end{aligned} \tag{12}
$$

Note that for $j = 0$:

$$
\begin{aligned}
S_0\left(t \mid \mathbf{x}_i, \phi_0, y_{i0}=1\right) &= \Pr\left[\widetilde{T}_{i0} \geq t = \overline{a} \mid \mathbf{x}_i, \phi_0, y_{i0}=1\right] \\
&= 1 - \Pr\left[\widetilde{T}_{i0} < t = \overline{a} \mid \mathbf{x}_i, \phi_0, y_{i0}=1\right] = 1
\end{aligned} \tag{13}
$$

since conditional on $y_{i0} = 1$, t is the maturity date.

If the censoring mechanism is noninformative then the terms relating to the pdf and cdf of the censoring variables can be dropped as constants of proportionality. The resulting likelihood for the set of parameters $\boldsymbol{\theta} = \left(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \phi_1', \phi_2'\right)'$ with independent and noninformative right censoring times is:

$$
L\left(\boldsymbol{\theta} \mid \mathbf{X}, \widetilde{\mathbf{q}}, \mathbf{T}, \boldsymbol{\delta}\right) \propto \prod_{i=1}^{n} \left\{ (p_0)^{y_{i0}} \prod_{j=1}^{2} \left[p_j f_j(t)\right]^{y_{ij}} \right\}^{\delta_i} \left\{ p_0 + \sum_{j=1}^{2} p_j S_j(t) \right\}^{1-\delta_i} \tag{14}
$$

where $\mathbf{X} = (\mathbf{x}_1 \cdots \mathbf{x}_N)'$; $\boldsymbol{\delta}$ has typical element δ_i as defined in equation 9. Where $j = 1, 2$, $f_j(t)$ is $f_j\left(t \mid \mathbf{x}_i, \phi_j, y_{ij} = 1\right)$ and $S_j(t)$ is the survival function $S_j\left(t \mid \mathbf{x}_i, \phi_j, y_{ij} = 1\right)$ where the conditional statements have been dropped for notational ease. In addition, $f_0(t)$ and $S_0(t)$ take values of unity as in 13.

There are three distributions applied to the latency events in this paper. The distirbutions are the Gamma, Weibull and Log-Normal which are respectively represented by the labels: G_j, W_j, and N_j, for $j = 1, 2$ corresponding to the events of write off and prepayment, respectively. The distributions are characterised by their pdfs and survival functions outlined below.

Gamma Distribution:

$$
f_j(t) = \frac{\exp\left(-\mathbf{x}_i^T \boldsymbol{\beta}_{Lj} \gamma_{Lj}\right) t^{\gamma_{Lj}-1} \exp\left\{-\exp\left[\ln(t_i) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj}\right]\right\}}{\Gamma\left(\gamma_{Lj}\right)} \tag{15}
$$

$$
S_j(t) = 1 - I\left(\exp\left[\ln(t_i) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj}\right], \gamma_{Lj}\right) \tag{16}
$$

where $I(\alpha, \beta)$ is the incomplete gamma function. Details on the gamma and incomplete gamma functions are provided in the appendix to this paper.

Weibull Distribution:

$$
f_j(t) = \gamma_{Lj} \exp\left(-\mathbf{x}_i^T \boldsymbol{\beta}_{Lj} \gamma_{Lj}\right) t^{\gamma_{Lj}-1} \exp\left\{-\exp\left[\gamma_{Lj}\left(\ln(t_i) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj}\right)\right]\right\} \tag{17}
$$

$$
S_j(t) = \exp\left\{-\exp\left[\gamma_{Lj}\left(\ln(t_i) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj}\right)\right]\right\} \tag{18}
$$

Log-Normal Distribution:

$$f_j(t) = \left(\sqrt{2\pi}\gamma_{Lj}t\right)^{-1} \exp\left\{\left[\frac{\ln(t_i) - \mathbf{x}_i^T\boldsymbol{\beta}_{Lj}}{\sqrt{2}\gamma_{Lj}}\right]^2\right\} \quad (19)$$

$$S_j(t) = 1 - \Phi\left(\gamma_{Lj}^{-1}\left[\ln(t_i) - \mathbf{x}_i^T\boldsymbol{\beta}_{Lj}\right]\right) \quad (20)$$

These three distributions are used to model time to the events of write off and prepayment. There are nine combinations of these distributions, the simplex method was used to find the minimum of the negative log-likelihood in each case. The results from the simplex method of optimisation have been detailed in the following section. The score and hessian functions are detailed in the appendix accompanying this paper.

4 Empricial Application

4.1 Data and Summaries

The data set of over one million observations contains information on personal loans which originated between 01 March 2001 and 31 March 2008 provided by one of Australian's largest financial services institutions. The loans can be contracted for terms of whole years ranging from 1 to 7 years. In addition to application and performance data at the account level, sufficient information on opening and closing dates of accounts and the reason for their terminations was provided to conduct the research within this paper. The list of personal loan application data provided for this research is outlined in table 4.1.

Table 4.1: List of Application Data

Number of Applicants per Loan	Time with Current Employer
Total Assets	Time with Previous Employer
Total Liabilities	Current State
Other Bank Home Loan	Time at Current Address
Other Bank Liabilities	Time at Previous Address
House Value	Guarantor
Other Value	Number of Installments
Accommodation Status	Total Loan Amount
Gender	Interest Rate at Application
Age at application	Repayment Amount

Table 4.2 details the proportion of maturity, write off, prepayment and censoring observations across each contracted loan term. There is a decreasing trend in the proportion of maturity events within each loan term stratum, whilst the proportions of of write off and prepayment events both increase. There is insufficient data to adequately identify the maturity events in the 72 and 84 month personal loan strata. These accounts have been

excluded from detailed analysis for this reason and focus placed on the 12 to 60 month personal loans. Censored events are a dominant component of the available information illustrated in table 4.2, particularly for the longer term loans.

Table 4.2: Count of Accounts Experiencing Defined Permanent Events and Censoring with Contracted Term as Stratum

TERM	Full Term	Write Off	Prepayment	Censored	TOTAL
12	43.69%	1.63%	40.87%	13.79%	1.68%
24	20.04%	2.44%	63.37%	14.16%	8.20%
36	8.56%	3.10%	69.08%	19.27%	12.40%
48	3.41%	4.23%	70.55%	21.79%	10.10%
60	1.38%	5.28%	62.98%	30.37%	24.10%
72	0.54%	5.96%	63.99%	29.49%	4.03%
84	0.01%	6.70%	52.89%	40.38%	39.50%
TOTAL	4.14%	5.18%	60.20%	30.50%	100.00%

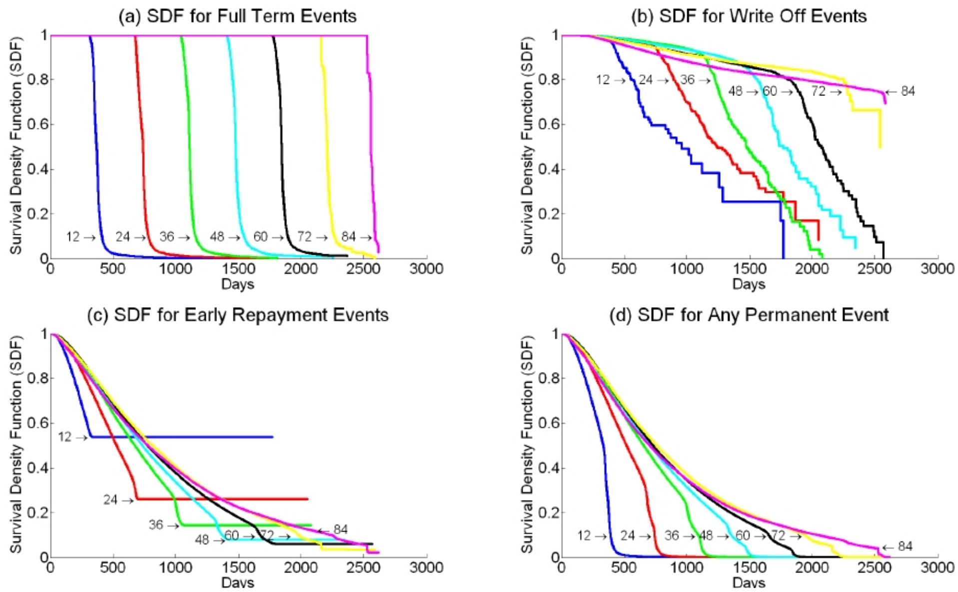
The variables pertaining to age, gender, time at current address and time with current employer provide a picture of the demographic of personal loan holders. The mode of the age distribution is around 18 to 22 years across the entire data set. The large frequency of young adults taking out personal loans is consistent with the permanent income hypothesis. Moreover, people in early adulthood have not reached their full income potential, however, are able to form expectations of their future income path. Based on this expectation and subjective intertemporal consumption discount factor, they may need to borrow in order to finance their optimised consumption path.

The Kaplan-Meier (KM) survival curve estimates in panel (a) of figure 1 are for the time to the predetermined maturity date. As expected, whilst treating all other observations as censored, the survival curve at the yearly marks drops away from unity toward zero for each term. In contrast, once the survival curves in panel (c) reach the contracted term date they plateau at a level equivalent to the proportion of uncensored accounts accounts that repaid as contracted.

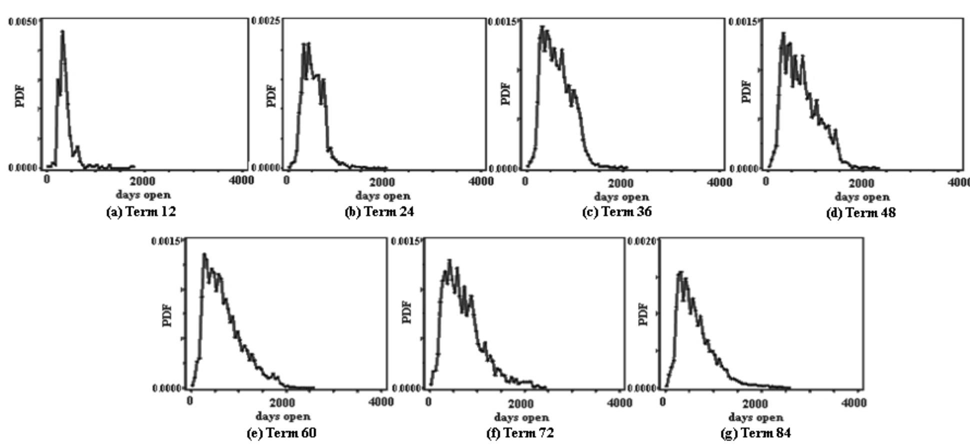
Hazard and pdf plots for the write off variable were created by focusing exclusively on write off events. The hazard of write off is increasing initially, then decreases and becomes more volatile with fewer observations at larger T, then after the maturity date the hazard significantly increases. In addition, this characteristic is further intensified by the fact that Australian banks have a policy of holding delinquent accounts up to 180 days past due.

The life-table estimates of the pdfs for each term are displayed in figure 2. Without the other observed events and treating only active accounts as censored, the time to write off pdf estimates seem more reasonable than the KM survival curve estimates where all other events were also treated as censored. The pdfs for all terms have a mode above the one year mark then begin to decline. The rate of decline decreases as the term increases. This feature is consistent with the statement in Banasik et al (1999) for a rule of thumb that states “if they

1. Kaplan Meier survival function estimates on the full data set with 12 to 84 month loan term stratum labels.



2. Life-Table pdf estimates for write off on the full sample



[personal loans] go bad, they go bad early". This general observation is not apparent in the KM survival curve estimates where all other observations have been treated as censored.

4.2 Results of Model Fitting

A simulation study was conducted to assess the performance of the methodology developed in this paper in evaluating the true parameter values and compared to that of previous applications. The results of the simulation study have been included in the appendix accompanying this paper. The simulation scenarios generated independent event times for write off and prepayment. Write off events could occur at any time. The results found that the model developed in this paper was far superior to the survival analysis methodologies which examine prepayment and write off events separately, treating maturity events as censored observations rather than terminal event times. In addition, the following observations can be made from the simulation study:

- censoring enduces more errors in the parameter estimates for all methodologies;
- the error and variance in the estimation results for the methodology of separte treatment of latency events increases as the proportion of maturity observations increases;
- the variance of the parameter estimates is around four times smaller in the methodology developed in this paper than those of the separate treatment method, and;
- as the relative frequency of an event with stochastic latency increases, the variance of the parameter estimates for the same latency decreases in both models.

Table 4.3: Variables used in empirical application

LVR	$\ln\left(\text{Loan Amount}/\text{Total Assets}\right)$	HV	$\ln\left(\text{House Value in 1000's}\right)$
TL	$\ln\left(\text{Total Liabilities in 1000's}\right)$	Lamt	Total Loan Amount
TA	$\ln\left(\text{Total Assets in 1000's}\right)$	TCA	Time at Current Address
TCE	Time with Current Employer	TPA	Time at Previous Address
TPE	Time with Previous Employer	GEN	1 if Female, 0 otherwise
T_DL	1 if Period 57 prior to system change, else 0	NTA	Net Assets = $e^{TA} - e^{TL}$
P_DL	1 if Period 58 after system change, else 0	ATL	$\ln\left(e^{TA} / e^{TL}\right)$
PCR	1 if Period 36 to 56 of low credit quality, else 0	AddYrs	TCA + TPA
Guar	1 if guarantor on loan, 0 otherwise	EmpYrs	TCE + TPE
Int	Interest Rate at Application	Age	Age in years

The results of the Maximum Likelihood Estimation (MLE) were obntained using the Nelder-Mead method of simplexes. The chosen distributions are the Gamma, Log-Normal and Weibull, which will be denoted by "G_j", "N_j", and "W_j", respectively, where the subscript j associates the distribution with event $j = 1, 2$. For example, the application of the Gamma and Weibull distributions to write off and prepayment latencies, respectively, will be denoted by "G_1W_2" in the presented results. The set of parameters for the k

regressors can be charactised as $\boldsymbol{\theta} = \left(\boldsymbol{\beta}'_{I1}, \boldsymbol{\beta}'_{I2}, \phi'_1, \phi'_2\right)'$, where $\phi_j = \left(\boldsymbol{\beta}'_{Lj}, \gamma_j\right)'$ and each $\boldsymbol{\beta}$ vector is $k \times 1$ and the γ's are scalar, bringing the total number of parameters in the model to $(k \times 4) + 2$. Note that a subscript "I" on a parameter denotes "incidence", whilst a sub script "L" denotes "latency".

Table 4.4: 36 month term data set parameter estimates

54 parameters; BIC $= 1,682,327$

	Incidence		Latency	
	Write Off	Prepayment	Write Off	Prepayment
Constant	-0.7830^a (0.0084)	2.5279^a (0.0024)	5.0659^a (0.0004)	6.3545^a (0.0005)
LVR	0.0260^b (0.0122)	-0.0707^a (0.0027)	-0.0328^a (0.0006)	0.0270^a (0.0006)
TL	-0.0215^a (0.0008)	0.0128^a (0.0002)	-0.0005^a (0.0000)	-0.0042^a (0.0000)
HV	-0.0545^a (0.0004)	0.0010^a (0.0001)	0.0038^a (0.0000)	0.0006^a (0.0000)
EmpYrs	-0.0031 (0.0035)	-0.0005 (0.0009)	0.0003^c (0.0002)	0.0001 (0.0002)
AddYrs	-0.0031 (6.8344)	-0.0002 (2.4803)	0.0001 (0.3554)	0.0001 (0.4047)
Age	-0.0082 (0.1669)	-0.0194 (0.0377)	-0.0023 (0.0067)	0.0049 (0.0084)
Int	-0.1050 (0.0712)	-1.2405^a (0.0188)	0.1744^a (0.0033)	0.0915^a (0.0037)
Guar	0.0002 (0.2204)	0.1353^b (0.0687)	-0.0182^c (0.0104)	0.0250^b (0.0115)
GEN	-0.0895 (0.1578)	0.1281^a (0.0346)	-0.0064 (0.0069)	0.0362^a (0.0071)
T_DL	-0.0689 (0.0815)	-0.0886^a (0.0261)	-0.0270^a (0.0040)	-0.0835^a (0.0047)
P_DL	-0.0746 (0.2930)	-0.0659 (0.0422)	-0.1134^b (0.0486)	-0.0759^a (0.0187)
PCR	0.0234^a (0.0085)	0.0123^a (0.0011)	-0.0372^a (0.0018)	-0.0397^a (0.0062)
γ_{L_j}			1.4784^c (0.8163)	2.0482^a (0.0252)

a, b, & c indicate the parameter estimates are signficant at the1%, 5% and 10% levels, respectively

The results of the MLE are displayed in table 4.4 to 4.9 in this section. An intensive study exclusively focusing on personal loans of 36 month term was performed and the results from the model with the lowest Bayesian Information Criteria (BIC) are displayed in Table 4.4. Tables 4.5 to 4.9 display the results for personal loan terms from 12 to 60 of the models with the lowest BIC and less than 40 parameters. These parameter estimates illustrate that regressors can act in opposite directions upon the incidence and latency of an event. This is evident in the results for the Loan to Value Ratio (LVR) variable. Table 4.5 shows that increases in LVR lead to increases in the incidence of write off. This is consistent with a priori expectation. In addition, the negative LVR coefficient estimates in table 4.7 show that conditional on experiencing write off, those accounts with a higher LVR progressed more slowly to this event. This results is also seen in table 4.4. In the case of LVR, this is consistent with the actions banks take to mitigate reputational risks surrounding high LVR lending practices.

Table 4.5: Write off incidence parameter estimates $\{\beta_{I1}\}$

. .	12 N_1W_2	24 G_1W_2	36 G_1W_2	48 G_1W_2	60 G_1W_2
Constant	-2.7773^a	-1.3675^a	-0.3377^a	0.6039^a	1.2310^a
	(0.0180)	(0.0124)	(0.0083)	(0.0080)	(0.0052)
LVR	0.2108^a	0.1739^a	0.1291^a	0.1542^a	0.1643^a
	(0.0023)	(0.0013)	(0.0008)	(0.0007)	(0.0004)
TL	-0.0671^a	-0.0418^a	-0.0332^a	-0.0175^a	-0.0232^a
	(0.0009)	(0.0005)	(0.0004)	(0.0004)	(0.0002)
EmpYrs	-0.0048	-0.0043	-0.0029	-0.0041	-0.0031
	(0.0079)	(0.0048)	(0.0034)	(0.0033)	(0.0021)
AddYrs	-0.0027	-0.0032	-0.0032	-0.0032	-0.0027
	(1.9394)	(0.2976)	(0.1916)	(0.1847)	(0.1435)
Age	-0.0002	-0.0028	-0.0098	-0.0058	-0.0052
	(0.1564)	(0.1056)	(0.0704)	(0.0675)	(0.0443)
Guar	-1.4361^a	-1.1311^a	-1.1199^a	-1.1441^a	-1.4184^a
	(0.0465)	(0.0233)	(0.0263)	(0.0458)	(0.0452)
GEN	0.0466^a	-0.1988^a	-0.1012^a	0.0008	-0.0926^a
	(0.0111)	(0.0060)	(0.0068)	(0.0119)	(0.0117)

a, b, & c indicate the parameter estimates are signficant at the1%, 5% and 10% levels, respectively

The LVR variable influences the incidence and latency of prepayment in opposite directions. A larger LVR relative to total assets is often correlated with a lower income debtor and the results are consistent with this, parameter estimates in tables 4.4 and 4.6 show a higher LVR leads to a lower incidence of prepayment. However, table 4.8 illustrates that given the debtor prepays, a debtor with a higher LVR will progress faster to prepayment. This results is still consistent with the previously mentioned correlation with income, since those who can afford to repay early are likely to be those with a high LVR but have high disposable income to repay their personal loan. In addition, this may be a refinancing decision where the debtor borrows at cheaper rates and repays their more expensive debt.

The Total Liabilities (TL) variable also works in opposite directions upon the incidence and latency of an event, see tables 4.5 to 4.8. Moreover, the individuals with very large liabilities can afford them through large servicing capacities and are less likely to experience write off. However, conditional on experiencing write off, those with higher TL progressed faster to this event, see table 4.7 for terms greater than 24 months. In terms of 12 and 24 months the TL variable work in the same direction upon write off incidence and latency, however, the coefficient estimates are economically insignificant. This result is consistent with table 4.4. In addition, TL works in the opposite direction on prepayment incidence and latency, see table 4.4 and tables 4.6 and 4.8. This may be a result of debtors refinancing this more expensive consumer credit facility to a cheaper alternative. However, more data would be required to appropriately determine if this is the case.

Table 4.6: Prepayment incidence parameter estimates $\{\beta_{I2}\}$

. .	12 N_1W_2	24 G_1W_2	36 G_1W_2	48 G_1W_2	60 G_1W_2
Constant	0.3850^a	1.6696^a	2.3822^a	3.2075^a	3.6645^a
	(0.0038)	(0.0021)	(0.0023)	(0.0041)	(0.0038)
LVR	-0.0544^a	-0.0733^a	-0.0621^a	-0.0469^a	0.0239^a
	(0.0003)	(0.0001)	(0.0002)	(0.0002)	(0.0002)
TL	0.0017^a	0.0124^a	0.0110^a	0.0046^a	0.0310^a
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
EmpYrs	0.0001	−0.0001	0.0000	−0.0002	−0.0003
	(0.0017)	(0.0008)	(0.0009)	(0.0015)	(0.0015)
AddYrs	−0.0004	−0.0003	−0.0002	−0.0003	−0.0000
	(0.1166)	(0.0374)	(0.0376)	(0.0583)	(0.0596)
Age	−0.0180	−0.0214	−0.0199	−0.0224	−0.0244
	(0.0314)	(0.0167)	(0.0188)	(0.0319)	(0.0306)
Guar	−0.2471	-0.1028^a	-0.0887^a	-0.1864^a	-0.0525^a
	(0.2552)	(0.0043)	(0.0049)	(0.0070)	(0.0055)
GEN	-0.0675^a	0.0564^a	0.1094^a	0.1615^a	0.1078^a
	(0.0021)	(0.0007)	(0.0010)	(0.0014)	(0.0014)

a, b, & c indicate the parameter estimates are signficant at the1%, 5% and 10% levels, respectively

The inclusion of a guarantor on a personal loan is statiscally and economically significant in decreasing the incidence of write off. This is consistent with a priori expectations, however, given an individual does write off their loan facility, the existence of a guarantor increases the rate of progression to this event as evidenced by the positive and highly significant parameter estimates in table 4.7. Prepayment incidence is less likely when a guarantor is supporting the loan facility, which is consistent with the lower income group whom require guarantor support to obtain access to credit, see table 4.6 results. Given a debtor prepays their loan facility, the existence of a guarantor is not a statistically significant determinant of the latency of prepayment in table 4.8. In table 4.4, however, the coefficient on the guarantor latency variable is positive and significant, suggesting that given a debtor prepays, the presence of a guarantor increases the speed to prepayment. A more detailed study examining the types of guarantor support would be required to decisively conclude whether the effect of a guarantor upon the latency of an event is significant.

Table 4.7: Write off latency parameter estimates $\{\boldsymbol{\beta}_{L1}, \gamma_{L1}\}$

.

	12 N_1W_2	24 G_1W_2	36 G_1W_2	48 G_1W_2	60 G_1W_2
Constant	5.7465^a	4.4881^a	5.0458^a	5.3545^a	5.5759^a
	(0.0013)	(0.0003)	(0.0004)	(0.0005)	(0.0005)
LVR	0.0159^a	-0.0007^a	-0.0099^a	-0.0151^a	-0.0115^a
	(0.0002)	(0.0000)	(0.0000)	(0.0001)	(0.0001)
TL	-0.0001^c	-0.0011^a	0.0029^a	0.0035^a	0.0081^a
	(0.0001)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
EmpYrs	0.0004	0.0006^a	0.0003^c	0.0005^b	0.0005^a
	(0.0006)	(0.0001)	(0.0002)	(0.0002)	(0.0002)
AddYrs	-0.0002	0.0001	0.0002	0.0003	0.0001
	(0.2644)	(0.0081)	(0.0091)	(0.0137)	(0.0153)
Age	0.0036	-0.0008	-0.0015	-0.0004	0.0000
	(0.0114)	(0.0029)	(0.0033)	(0.0047)	(0.0037)
Guar	0.1307^a	0.0996^a	0.1779^a	0.2550^a	0.3292^a
	(0.0145)	(0.0060)	(0.0054)	(0.0068)	(0.0053)
GEN	0.1034^a	0.0489^a	0.0117^a	0.0290^a	0.0642^a
	(0.0036)	(0.0016)	(0.0014)	(0.0018)	(0.0014)
$\ln(\gamma_{L1})$	-0.9542^a	1.7555^a	1.4760^a	1.2425^a	1.0702^a
	(0.3478)	(0.1549)	(0.1040)	(0.0990)	(0.0652)

a, b, & c indicate the parameter estimates are signficant at the1%, 5% and 10% levels, respectively

Table 4.8: Prepayment latency parameter estimates $\{\boldsymbol{\beta}_{L2}, \gamma_{L2}\}$

.

	12 N_1W_2	24 G_1W_2	36 G_1W_2	48 G_1W_2	60 G_1W_2
Constant	5.4251^a	6.0404^a	6.3416^a	6.4742^a	6.5554^a
	(0.0012)	(0.0005)	(0.0005)	(0.0006)	(0.0005)
LVR	0.0102^a	0.0159^a	0.0258^a	0.0341^a	0.0577^a
	(0.0001)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
TL	0.0005^a	-0.0015^a	-0.0041^a	-0.0063^a	-0.0069^a
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
EmpYrs	-0.0001	0.0001	0.0002	0.0002	0.0002
	(0.0005)	(0.0002)	(0.0002)	(0.0003)	(0.0002)
AddYrs	0.0000	0.0000	0.0001	0.0001	0.0001
	(0.0501)	(0.0102)	(0.0085)	(0.0100)	(0.0082)
Age	0.0001	0.0032	0.0050	0.0066	0.0081^b
	(0.0099)	(0.0042)	(0.0037)	(0.0045)	(0.0035)
Guar	0.0560	0.0240^b	0.0293	0.0333	0.0650
	(0.0983)	(0.0111)	(0.0183)	(0.0211)	(0.0119)
GEN	-0.0034	0.0242^a	0.0357^a	0.0361^a	0.0493^a
	(0.0271)	(0.0087)	(0.0060)	(0.0058)	(0.0035)
γ_{L2}	2.7903^a	2.4056^a	2.0529^a	1.8044^a	1.6060^a
	(0.0383)	(0.0272)	(0.0218)	(0.0218)	(0.0168)

a, b, & c indicate the parameter estimates are signficant at the1%, 5% and 10% levels, respectively

Across the 24, 36 and 60 month term data sets, the results indicate that women were less likely to experience write off on their personal loan facility, see table 4.4 and 4.5. In table 4.4, gender was not a significant determinant of the speed to write off, however, in table 4.7 gender was found to statistically significantly increase the speed to this event. In addition, women were more likely to repay early, and would do so at a faster rate than men as seen in tables 4.4, 4.6 and 4.8. The only exception to this being in the 12 month personal loans data set representing 1.68% of the total sample. The gender variable results may be due to women being relatively more risk averse than men.

The variables of EmpYrs, AddYrs, and Age were statistically insignificant across all latencies and incidences. This is consistent with a priori expectations that given the inclusion of financial variables, debtor age, housing and employment stability (measured in years) are not as important factors as a debtors capacity to service their loan facility. The only exception being statistical significance showing that increases in the EmpYrs variable increase the speed to write off conditional on experiencing that event. This result appears counterintuitive, but is consistent across all studies and data sets. A possible explanation could be a skilling problem, where an individual who has been in the one job for many years loses it due to new technological innovation and adoption, leaving the individual with a redundant skill set. Additional data on final employment status would be useful in determining if the previous explanation is the primary influence of the observed result.

Table 4.9: Bayesian Information Criteria for distribution pairs

	12	24	36	48	60
G_1G_2	130,550.09	999,829.49	1,691,494.65	1,457,727.29	3,233,059.25
G_1N_2	132,062.25	1,011,948.31	1,709,943.73	1,470,431.32	3,252,580.28
G_1W_2	129,316.02	Ŧ992,198.73	Ŧ1,683,291.80	Ŧ1,454,077.57	Ŧ3,230,943.09
N_1G_2	130,518.20	999,958.15	1,691,834.76	1,458,044.18	3,233,504.08
N_1N_2	132,029.61	1,012,086.80	⊠1,710,278.91	⊠1,470,751.17	⊠3,253,084.87
N_1W_2	Ŧ129,285.01	992,336.37	1,683,622.32	1,454,381.49	3,231,329.38
W_1G_2	130,645.22	999,976.88	1,691,509.26	1,457,749.99	3,233,488.50
W_1N_2	⊠132,155.98	⊠1,012,089.73	1,709,957.94	1,470,447.48	3,252,984.85
W_1W_2	129,411.40	992,348.72	1,683,307.10	1,454,103.83	3,231,386.54

Note: Ŧ: minimum; and; ⊠:maximum for each term data set.

The BIC for each of the nine distribution combinations across the five data sets is displayed in Table 4.9. The BIC indicate that the Weibull distribution for prepayment consistently lead to the best fit, whilst the log-normal distribution for prepayment consistently lead to the worst model fit. The Gamma and Weibull distributional mix for write off and prepayment, respectively, resulted in the best model fit in the largest four of the five data sets.

4.3 Diagnostics

Profile log-likelihoods were examined across the parameter pairs. The surfaces reveal clear maxima with clear features of uniformity, symmetry and concavity. The profile log-likelihoods suggest that there is a clear global maximum for all distributional pairs. An example of the profile log-likelihoods is illustrated in figure 3.

Anderson-Darling and Kolmogorov-Smirnov tests were applied to write off and prepayment latency residuals across the nine distribution assumptions. The tests unanimously found sufficient evidence to reject the null hypothesis that the appropriately standardised residuals came from the standard normal distribution and the results have been included in the appendix accompanying this paper. This suggests a misspecification of distributions. However, this may be mitigated fractionally given the increasing sensitivity of these tests as the data set grows in size.

A graphical test for linearity can be performed to assess the appropriateness of the Weibull distributional assumptions. Figure 4 plots the log negative log of the KM survival function against log time and should have an intercept term of $-\widehat{\boldsymbol{\beta}}_{Lj}\widehat{\gamma}_{Lj}$ and a slope coefficient of $\widehat{\gamma}_{Lj}$. The plots for the time to prepayment events in panel (b) of figure 4 appear to be the most linear of the two plots. The plot of the write off variables in panel (a) of figure 4 do not illustrate a linear relationship. Overall, there is only a very weak linear relationship for this variable and it casts doubt over the appropriateness of the Weibull distributional assumption for the write off latency.

A regression was performed such that:

$$y_i = \widehat{\eta}_a + \widehat{\eta}_b \ln(t_i) + u_i \tag{21}$$

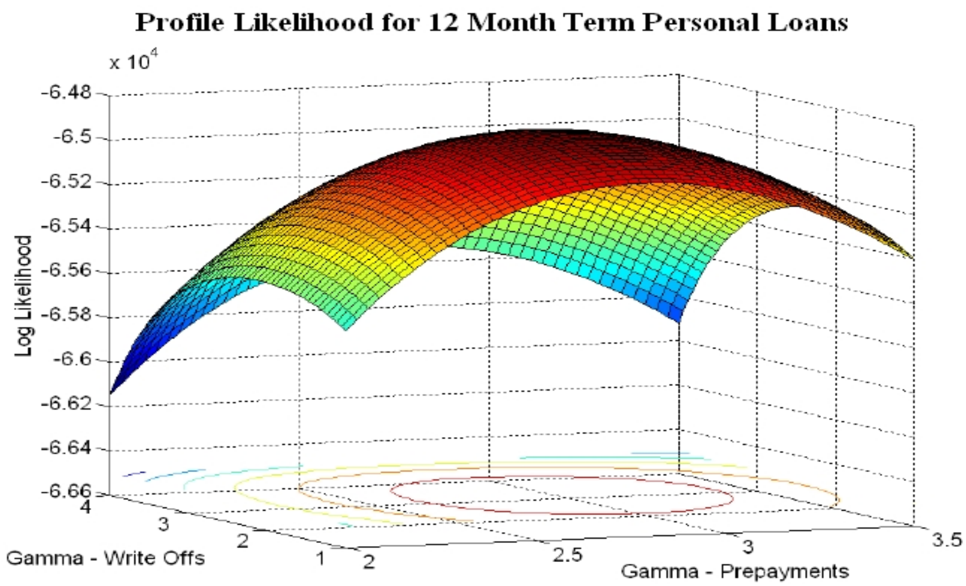
where $y_i = \ln\left\{-\ln\left[\widehat{S}_{KM}(t_i)\right]\right\}$ and was regressed on the natural logarithm of the observed failure times (t_i) separately for each event. Should the time to the particular event be distributed Weibull then the estimates should hold the following relationship with the optimised parameters of the parametric survival estimation:

$$\widehat{\eta}_a = -\widehat{\gamma}_{Lj}\widehat{\boldsymbol{\beta}}_{Lj} \quad \text{and} \quad \widehat{\eta}_b = \widehat{\gamma}_{Lj} \tag{22}$$

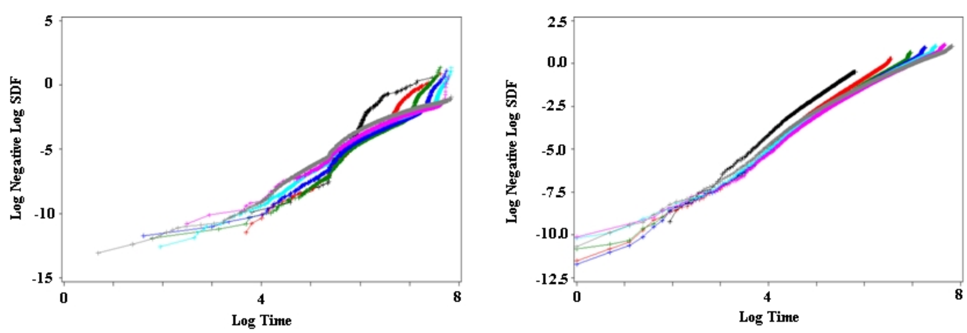
A Wald test was performed to test the null hypothesis $H_o : \widehat{\eta} - \eta = 0$ against the alternative that it is not equal to zero, across all distribution pairs and the full set of results has been included in the appendix accompanying this paper. In all except two instances, the null hypotheses were rejected at the one percent significance level. The two exceptions were both for the prepayment event in the 48 Month Term data set for the Weibull Weibull and Log-Normal Weibull distribution combintations.

Upon examination of the histogram plots in figure 5 it is apparent that the Weibull Weibull distribution assumption appears to match the EVM pdf plot in figure 5 most closely of the nine histograms. The Gamma Weibull (that is Gamma write offs and Weibull prepayments) and the Log-Normal Weibull depart the most from the pdf plot of the EVM distribution, characterised by an approximate ten fold decrease in the domain

3. Profile log-likelihood over the gamma latency parameters

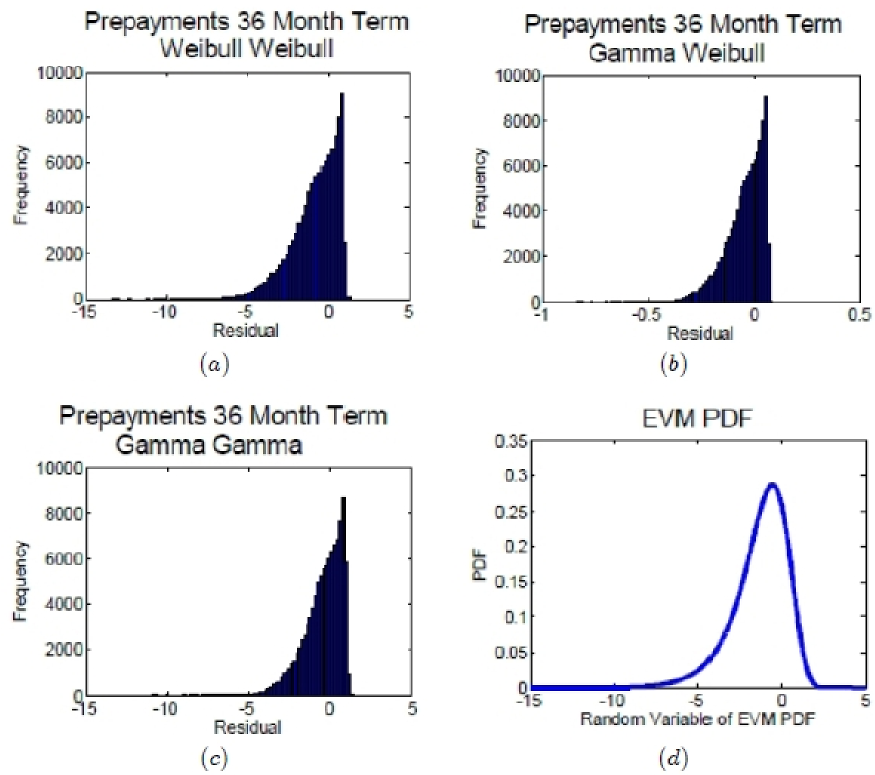


4. Plots for log negative log of the Kaplan-Meier survival function versus log time



(a) Write Off (b) Prepayment

5. Time to prepayment residual histograms and EVM pdf plot



of the residuals. These plots also correspond to the estimates of Model I and II with the lowest BIC across the nine distribution combinations. In addition, there may be correlation between the write off and prepayment events that needs to be addressed.

5 Conclusion

Credit risk assessment has been dominated by logistic and probit regression techniques. Research into the application of duration analysis to credit data has become increasingly abundant in recent years. Typical applications examine the credit events of default and prepayment individually. There have been applications treating the aforementioned events as dependent competing risks and have simultaneously estimated their parameters. However, all applications have failed to adequately treat credit maturity events which will lead to biases in parameter estimation.

This paper has developed the first integrated methodology for the analysis of a set of mutually exclusive events, where the duration time to an event may be non-stochastic or pre-determined. It has been motivated by the Cure Rate methodologies in the medical literature, augmenting these binary models to a fully parametric multinomial mixture model framework, best applied to credit data. Incidence and latency of each event in the system are estimated simultaneously.

The results from the model estimation with Australian retail credit data provide the first evidence of regressors acting in opposite directions upon the incidence and latency of an event. In particular, as the Loan to Value Ratio (LVR) at application of the personal loans rises, the incidence of write off increases whilst the incidence of prepayment decreases and the conditional latencies of write off and prepayment are progressed to more slowly and faster, respectively. Similarly, for the Total Liabilities (TL) at application, positive and negative coefficients were estimated for the prepayment incidence and latency effects, respectively. This suggests that the higher the TL at application the more likely a loan facility is to progress to prepayment, but the slower this will occur, conditional on experiencing prepayment.

The same set of results were also the first to provide evidence of regressors in credit data which are significant in explaining the conditional latency and insignificant in explaining the incidence of the same event. The regressor for the number of months an applicant has been working at their last two jobs (Emp Yrs) is not significant in explaining the incidence of write off whilst it is significant in explaining the conditional latency of write off. However, it is only of marginal economic significance given the low magnitude of the positive coefficient estimate.

The results within this paper were unattainable using previous methodologies. This aspect of the model allows for a deeper and more rigorous examination of credit data. In addition, the results of the simulation study indicate that the methodology developed in this paper was superior in predicting the parameter values compared to the previous two general frameworks. The model corrects the biases that existed in previous studies die to the treatment of maturity observations. There are far reaching applications of this model ranging from profit scoring to portfolio funding optimisation. Future research can extend this framework to explicitly examine the dependence structure between prepayment and write off through either a copula or bivariate framework.

6 Bibliography

[1] Alexander, C., 2005, *The Present and Future of Financial Risk Management*, Journal of Financial Econometrics, 3(1), 3-25.

[2] Abramowitz, M, & Stegun, I., 1965, *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*, New York, Dover.

[3] Allen, L., DeLong, G., & Saunders, A., 2004, *Issues in the credit risk modeling of retail markets*, Journal of Banking and Finance, 28(4), 727-752.

[4] Altman, E., & Saunders, A., 1998, *Credit Risk Measurement: Developments over the last 20 years*, Journal of Banking & Finance, 21(12), 1721-1742.

[5] Anderson, P., Borgan, Ø., Gill, R., & Keiding, N., 1993, *Statistical Models Based on Counting Processes*, New York, Springer-Verlag.

[6] Anderson, T., & Darling, D., 1954, *A Test of Goodness of Fit*, Journal of the American Statistical Association, 49(268), pp. 765-769.

[7] Andreeva, G., 2006, *European Generic Scoring Models using Survival Analysis*, The Journal of the Operational Research Society, 57(10), 1180-1187.

[8] Banasik, J., Crook, J., & Thomas, L., 1999, *Not if but when will borrowers default*, The Journal of the Operational Research Society, 50(12), 1185-1190.

[9] Bellotti, T., & Crook, J., 2007, *Credit Scoring with Macro Variables using Survival Analysis*, Working Paper, University of Edinburgh.

[10] Berenson, M., Levine, D., and Krehbiel, T., *Basic Business Statistics - Concept and Applications*, Eighth Edition, New Jersey: Prentice Hall, 2002.

[11] Berkson, J., & Gage, R., 1952, *Survival Curve for Cancer Patients Following Treatment*, Journal of the American Statistical Association, 47(259), pp. 501-515.

[12] Black, F., & Scholes, M., 1973, *The Pricing of Options and Corporate Liabilities*, The Journal of Political Economy, 81(3), pp. 637-654.

[13] Bluhm, C., Overbeck, L., & Wagner, C., 2003, *An Introduction to Credit Risk Modelling*, Florida, Chapman & Hall/CRC.

[14] Boag, J. W., 1949, *Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy*, Journal of the Royal Statistical Society, Series B (Methodological), 11(1), pp. 15-53.

[15] Browning, M., & Crossley, T., 2001, *The Life-Cycle Model of Consumption and Saving*, The Journal of Economic Perspectives, 15(3), pp. 3-22.

[16] Bucay, N., & Rosen, D., 2000, *Applying Portfolio Credit Risk Models to Retail Portfolios*, ALGO Research Quarterly, 3(1), 45-74.

[17] Cameron, A., & Trivedi, P., 2005, *Microeconometrics: Methods and Applications*, New York, Cambridge University Press.

[18] Cancho, V., Bolfarine, H., & Ortega, E., 2008, *The Log-exponentiated-Weibull Regression Models with Cure Rate*, Working Paper, University of São Paulo.

[19] Carroll, C., 2001, *A Theory of the Consumption Function, with and without Liquidity Constraints*, The Journal of Economic Perspectives, 15(3), pp. 23-45.

[20] Chen, M. H., Ibrahim, J. G., and Sinha, D., 1999, *A New Bayesian Model for Survival Data with a Surviving Fraction*, Journal of the American Statistical Association, 94, 909-919

[21] Ciochetti, D., Deng, Y., Gao, B., & Yao, R., 2002, *The Termination of Commercial Mortgage Contracts through Prepayment and Default: A Proportional Hazards Approach with Competing Risks*, Real Estate Economics, 30(4), pp. 595-633.

[22] Cox, D., 1975, *Partial Likelihood*, Biometrika, 62(2), 269-276.

[23] Crook, J., Edelman, D., & Thomas, L., 2007, *Recent Developments in Consumer Credit*

Risk Assessment, European Journal of Operational Research, 183(3), pp. 1447-1465.

[24] Dalla Valle, A., 2007, *A Test for the Hypothesis of Kew-Normality in a Population*, Journal of Computation and Simulation, 77(1), pp. 63-77.

[25] Deng, Y., Quigley, J., & Van Order, R., 2000, *Mortgage Terminations, Heterogeneity, and the Exercise of Mortgage Options*, Econometrica, 68(2), 275-307.

[26] Drzik, J., 2005, *New Directions in Risk Management*, Journal of Financial Econometrics, 3(1), 26-36.

[27] Efron, B., 1977, *The efficiency of Cox's Likelihood Function for Censored Data*, Journal of the American Statistical Association, 72(359), 557-565.

[28] Farewell, V. T., 1982, *The Use of Micture Models for the Analysis of Survival Data with Long-Term Survivors*, Biometrics, 38(4), pp. 10414-1046.

[29] Greene, W., 2003, Econometric Analysis, 5th Edition, New Jersey, Prentice Hall

[30] Han, A., & Hausman, J., 1990, *Flexible Parametric Estimation of Duration and Competing Risk Models*, Journal of Applied Econometrics, 5(1), pp. 1-28.

[31] Hoggart, C., & Griffin, J. E., 2001, *A Bayesian Partition Model for Customer Attrition*, In: George, E. I. (ed.), *Bayesian Method with Applications to Science, Policy, and Official Statistics*, Selected Papers from the ISBA 2000, pp. 223-232

[32] Hőrmann, W., Leydold, J., & Derflinger, G., 2004, *Automatic Nonuniform Random Number Generation*, New York, Springer

[33] Hougaard, P., 1986, *Survival Models for heterogeneous populations derived from stable distributions*, Biometrika, 73(2), 387-396.

[34] Ibrahim, J., Chen, M.-H., & Sinha, D., 2001, *Bayesian Survival Analysis*, New Yok, Springer-Verlag.

[35] Kalbfleisch, J., & Prentice, R., 2002, *The Statistical Analysis of Failure Time Data*, 2nd Edition, New Jersey, John Wiley & Sons.

[36] Kau, J. B., Keenan, D. C., Muller, W. J., and Epperson, J. F., 1992, *A Generalized Valuation Model for Fixed-Rate Residential Mortgages*, Journal of Money, Credit and Banking, 24(3), pp. 279-299

[37] Klein, J., & Meoschberger, M., 1997, *Survival Analysis: Techniques for Censored and Truncated Data*, New York, Springer-Verlag.

[38] McCall, B., 1996, *Unemployment Insurance Rules, Joblessness, and Part-Time Work*, Econometrica, 67(3), pp. 647-682

[39] Merton, R., 1974, *On the Pricing of Corporate Debt: The Risk Structure of Interest Rates*, The Journal of Finance, 29(2), pp. 449-470

[40] Narain, B., 1992, *Survival Analysis and the Credit Granting Decision*, In: Thomas, L., Crook, J., & Edelman (eds.), *Credit Scoring and Credit Control*, Oxford, Oxford University Press, 109-121.

[41] Pavlov, A., 2001, *Competing Risks of Mortgage Termination: Who Refinances, Who Moves and Who Defaults*, Journal of Real Estate Economics and Finance, 23(2), pp. 185-211

[42] Peng, Y., & Dear, K., 2000, *A Nonparametric Micture Model for Cure Rate Estimation*,

Biometrics, 56(1), pp. 227-236

[43] Stepanova, M., & Thomas, L., 2001, *PHAB scores - proportional hazards analysis behavioural scores*, The Journal of the Operational Research Society, 41(9), 1007-1016.

[44] Stepanova, M., & Thomas, L., 2002, *Survival analysis methods for personal loan data*, Operations Research Quarterly, 50(2), 277–289.

[45] Sueyoshi, G., 1992, *Semiparametric Proportional Hazards Estimation of Competing Risks Models with Time-Varying Covariates*, Journal of Econometrics, 51(1-2), pp. 25-58.

[46] Sy, J., & Taylor, J., 2000, *Estimation in a Cox Proportional Hazards Cure Model,* Biometrics, 56(1), pp 227-236

[47] Thomas, L., 2000, *A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers*, International Journal of Forecasting, 16(2), 149-172.

[48] Tsodikov, A. D., Ibrahim, J. G., & Yakovlev, A. Y. (2003). *Estimating Cure Rates from Survival Data: An Alternative to Two-Component Mixture Models*, Journal of the American Statistical Association, 98(464), pp. 1063-1078

[49] Van den Berg, G., 2001, *Duration Models - Specification, Identification, and Multiple Durations*, in Handbook of Econometrics, J. J. Heckman and E. Leamer (eds.), Volume 5, pp. 3381-3460, Amsterdam, North-Holland.

Appendix A.

7 Gamma Function

The Gamma Function and Incomplete Gamma Function are defined below, respectively:

$$\Gamma\left(\gamma_{Lj}\right) = \int_0^\infty e^{-t} t^{\gamma_{Lj}-1} dt \tag{A-1}$$

$$I\left(c, \gamma_{Lj}\right) = \int_0^c e^{-t} t^{\gamma_{Lj}-1} dt \Bigg/ \int_0^\infty e^{-t} t^{\gamma_{Lj}-1} dt \tag{A-2}$$

8 Score Function

The incidence components for this model are given a Multinomial Logit (MNL) functional form. Equation **??** details the expressions for p_{i0}, p_{i1} and p_{i2}, respectively. Now let the expression

$$\left[1 + \sum_{j=1}^{2} e^{\mathbf{x}_i^T \boldsymbol{\beta}_{Ij}} S_{\mathfrak{t}_j}\left(t\right)\right] = \Xi_{\mathfrak{t}_1 \mathfrak{t}_2} \tag{A-3}$$

for notational convenience. In addtion, let the indicator function be used such that $\mathbf{1}(\mathfrak{t}_j = W_j)$ takes the value of one when the statement inside is true, and zero otherwise. It is used analogously for all other distribution labels.

Using this terminology we can define the score functions for Model II, they are outlined below for $j = 1, 2$:

$$\begin{aligned} \frac{\partial \mathcal{L}\left(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{q}, \mathbf{T}, \boldsymbol{\delta}\right)}{\partial \boldsymbol{\beta}_{Ij}} &= \sum_{i=1}^{N} -\delta_i \left[y_{i0} p_{ij} - y_{ij} + y_{i1} p_{ij} + y_{i2} p_{ij}\right] \mathbf{x}_i - \left(1 - \delta_i\right) p_{ij} \mathbf{x}_i \\ &+ \left(1 - \delta_i\right) e^{\mathbf{x}_i^T \boldsymbol{\beta}_{Ij}} S_{\mathfrak{t}_j}\left(t \mid \mathbf{x}_i, \phi_{\mathfrak{t}_j}, y_{ij} = 1\right) \left(\Xi_{\mathfrak{t}_1 \mathfrak{t}_2}\right)^{-1} \mathbf{x}_i \end{aligned} \tag{A-4}$$

$$
\begin{aligned}
\frac{\partial \mathcal{L}\left(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{q}, \mathbf{T}, \boldsymbol{\delta}\right)}{\partial \boldsymbol{\beta}_{Lj}} &= \sum_{i=1}^{N} \delta_i y_{ij} \left\{\gamma_{Lj}\left(\exp\left[\gamma_{Lj}\left(\ln\left(t_i\right) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj}\right)\right] - 1\right)\right\}^{\mathbf{1}(\mathfrak{t}_j = W_j)} \\
&\quad \times \left[-\gamma_{Lj} \exp\left(\ln\left(t_i\right) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj}\right)\right]^{\mathbf{1}(\mathfrak{t}_j = G_j)} \\
&\quad \times \left\{\gamma_{Lj}^{-2} \exp\left[2\left(\ln\left(t_i\right) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj}\right)\right]\right\}^{\mathbf{1}(\mathfrak{t}_j = \text{Ł}_j)} \mathbf{x}_i \\
&\quad + \left(1 - \delta_i\right) e^{\mathbf{x}_i^T \boldsymbol{\beta}_{Ij}} \left\{\gamma_{Lj} e^{\gamma_{Lj}\left[\ln\left(t_i\right) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj}\right]} S_{W_j}\left(t \mid \mathbf{x}_i, \phi_{W_j}, y_{ij} = 1\right)\right\}^{\mathbf{1}(\mathfrak{t}_j = W_j)} \\
&\quad \times \left\{-\gamma_{Lj} e^{\ln\left(t_i\right) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj}} e^{-\exp\left[\ln\left(t_i\right) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj}\right]} \left[\Gamma\left(\gamma_{Lj}\right)\right]^{-1}\right\}^{\mathbf{1}(\mathfrak{t}_j = G_j)} \qquad \text{(A-5)} \\
&\quad \times \left\{\left(\sqrt{2\pi}\gamma_{Lj}\right)^{-1} \exp\left[\frac{\ln\left(t_i\right) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj}}{\sqrt{2}\gamma_{Lj}}\right]\right\}^{\mathbf{1}(\mathfrak{t}_j = \text{Ł}_j)} \left(\Xi_{\mathfrak{t}_1 \mathfrak{t}_2}\right)^{-1} \mathbf{x}_i
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \mathcal{L}\left(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{q}, \mathbf{T}, \boldsymbol{\delta}\right)}{\partial \gamma_{Lj}} &= \sum_{i=1}^{N} \delta_i y_{ij} \left\{\gamma_{Lj}^{-1} + \left[\ln\left(t_i\right) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj}\right]\left[1 - e^{\gamma_{Lj}\left[\ln\left(t_i\right) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj}\right]}\right]\right\}^{\mathbf{1}(\mathfrak{t}_j = W_j)} \\
&\quad \times \left[\ln\left(t_i\right) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj} - \psi\left(\gamma_{Lj}\right)\right]^{\mathbf{1}(\mathfrak{t}_j = G_j)} \\
&\quad \times \left\{-\gamma_{Lj}^{-1} + \gamma_{Lj}^{-3} \exp\left[2\left(\ln\left(t_i\right) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj}\right)\right]\right\}^{\mathbf{1}(\mathfrak{t}_j = \text{Ł}_j)} \\
&\quad + \left(1 - \delta_i\right) e^{\mathbf{x}_i^T \boldsymbol{\beta}_{Ij}} \left\{\left[\ln\left(t_i\right) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj}\right] e^{\gamma_{Lj}\left[\ln\left(t_i\right) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj}\right]} S_{W_j}\left(t\right)\right\}^{\mathbf{1}(\mathfrak{t}_j = W_j)} \\
&\quad \times \left\{\Gamma\left(\gamma_{Lj}\right)\left[{}_2\widetilde{F}_2\left(\begin{array}{cc} \gamma_{Lj} & \gamma_{Lj} \\ \gamma_{Lj} + 1 & \gamma_{Lj} + 1 \end{array} \middle| -e^{\ln\left(t_i\right) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj}}\right)\right.\right. \\
&\quad \left.\times e^{\left[\ln\left(t_i\right) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj}\right]\left(\gamma_{Lj} - 1\right)} + I\left(0, \gamma_{Lj}\right) - I\left(e^{\ln\left(t_i\right) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj}}, \gamma_{Lj}\right)\right] \\
&\quad \left. \times \left[\ln\left(t_i\right) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj} - \psi\left(\gamma_{Lj}\right)\right]\right\}^{\mathbf{1}(\mathfrak{t}_j = G_j)} \qquad \text{(A-6)} \\
&\quad \times \left\{\left(\sqrt{2\pi}\gamma_{Lj}\right)^{-1} \exp\left[-\frac{1}{2}\left(\frac{\ln\left(t_i\right) - \mathbf{x}_i^T \boldsymbol{\beta}_{Lj}}{\gamma_{Lj}}\right)^2\right]\right\}^{\mathbf{1}(\mathfrak{t}_j = \text{Ł}_j)} \left(\Xi_{\mathfrak{t}_1 \mathfrak{t}_2}\right)^{-1}
\end{aligned}
$$

The digamma function is the derivative of the log Gamma function with respect to its only argument. The expression is outlined below:

$$
\begin{aligned}
\psi\left(\gamma_{Lj}\right) &= digamma = \frac{\partial\left(\ln\left[\Gamma\left(\gamma_{Lj}\right)\right]\right)}{\partial \gamma_{Lj}} = \frac{\partial \Gamma\left(\gamma_{Lj}\right) / \partial \gamma_{Lj}}{\Gamma\left(\gamma_{Lj}\right)} \\
&= -C + \sum_{n=1}^{\infty} \frac{\gamma_{Lj}}{n\left(\gamma_{Lj} + n\right)} \qquad \text{(A-7)}
\end{aligned}
$$

Where C is Euler's Constant and is defined as:

$$C = \lim_{n\to\infty}\left[1+\frac{1}{2}+...+\frac{1}{n}-\ln(n)\right] \approx 0.57712566490153 \qquad \text{(A-8)}$$

In addition, the Regularised Hypergeometic Function $\left({}_2\widetilde{F}_2\right)$ is used in equation A-6 whenever the Gamma distribution is applied to any of the latencies. The Regularised Hypergeometic Function is characterised as:

$${}_2\widetilde{F}_2\left(\begin{array}{cc} \gamma_{Lj} & \gamma_{Lj} \\ \gamma_{Lj}+1 & \gamma_{Lj}+1 \end{array}\middle| -e^{\ln(t_i)-\mathbf{x}_i^T\boldsymbol{\beta}_{Lj}}\right) = \frac{{}_2F_2\left(\begin{array}{cc} \gamma_{Lj} & \gamma_{Lj} \\ \gamma_{Lj}+1 & \gamma_{Lj}+1 \end{array}\middle| -e^{\ln(t_i)-\mathbf{x}_i^T\boldsymbol{\beta}_{Lj}}\right)}{\Gamma\left(\gamma_{Lj}+1\right)\Gamma\left(\gamma_{Lj}+1\right)}$$

where ${}_2F_2$ is the Generalised Hypergeometric Function which characterised as:

$${}_pF_q\left(\begin{array}{ccc} a_1 & \cdots & a_p \\ b_1 & \cdots & b_q \end{array}\middle| x\right) = \sum_{k=0}^{\infty}\frac{(a_1)_k\cdots(a_p)_k}{(b_1)_k\cdots(b_q)_k}\frac{x^k}{k!}$$

where the Pochammer Notation, $(a_1)_k$, represents:

$$(a)_k \equiv \frac{\Gamma(a+k)}{\Gamma(a)}$$

This information is available on the Wolfram MathWorld web pages.

The Hessian Matrix can be obtained as (i) $\frac{\partial \ln L(\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}_0\partial\boldsymbol{\theta}_0^T}$ or (ii)$\left(\frac{\partial \ln L(\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}_0}\right)^{\otimes 2}$, where $\mathbf{a}^{\otimes 2}=\mathbf{a}\mathbf{a}'$. The Hessian for the Weibull Weibull case of this model was calculated using the first method, whilst the second method was used for all other distribution combinations.

9 Simulation

The first three scenarios generate from the Log-Normal and Weibull distributions for write off and prepayment events, respectively (Ł$_1$W$_2$), differing in parameter values and fixed incidence proportions. The next two scenarios labelled 04 and 05, generate the latencies for each event from two independent but identical Weibull distributions (W$_1$W$_2$). These first five scenarios are generated with only an intercept term and thus occur in fixed proportions. The fixed proportions are created using a uniform (0,1) random variable labelled U_I. A different and independently generated uniform (0,1) vector labelled U_δ is used to simulate random noninformative right censoring. Models I and III are used to estimate the parameters on sets of 1,000 observations generated 20,000 times.

Table 5.1: Parameter values used to generate simulation scenarios 01 to 05

Scenario	p_0	p_1	p_2	$\ln(\sigma_1)$	β_1	γ_2	β_2
Sim01 Ł$_1$W$_2$	0.05	0.20	0.75	ln(0.6)	4.50	7.00	5.50
Sim02 Ł$_1$W$_2$	0.05	0.75	0.20	ln(0.6)	4.50	7.00	5.50
Sim03 Ł$_1$W$_2$	0.40	0.20	0.40	ln(0.6)	4.50	7.00	5.50
Sim04 W$_1$W$_2$	0.60	0.10	0.30	ln(0.6)	4.50	7.00	5.50

Once an incidence of maturity, write off, or prepayment has been randomly assigned to each of the 1,000 generated data points, a time vector, $\mathbf{T}$, can be constructed. Each element of T will correspond to the elements of the randomly and independently generated ($Ł_1 W_2$) and ($W_1 W_2$) event times. This vector of event times and the corresponding event and censoring indicator variables are then used in an optimisation routine to estimate the parameters for Model i: developed in this paper; Model ii: same as model i except treats maturity events as censored; Model iii: simulataneous estimation of prepayment and default without separation of incidence and latency; Model iv: examines prepayment individually treating all other events as censored observations; and Model v: examines write off individually treating all other events as censored observations.

Table 5.2: All models parameter estimates with 10% random censoring Sim01

		$\widehat{p}_1$	$\widehat{p}_2$	$\ln(\widehat{\sigma}_1)$	$\widehat{\beta}_1$	$\widehat{\gamma}_2$	$\widehat{\beta}_2$
Sim01	Population	0.20	0.75	ln(0.6)	4.50	7.00	5.50
Model i	Mean	0.1846	0.7480	-0.4914	4.5220	7.0952	5.5099
	Std	0.0125	0.0143	0.0563	0.0467	0.2129	0.0057
	Prct Err.	-7.68%	-0.26%	-3.80%	0.49%	1.36%	0.18%
Model ii	Mean	0.2576		0.0462	5.0747	7.0921	5.5096
	Std	0.0148		0.0557	0.0846	0.2129	0.0057
	Prct Err.	28.82%		-109.04%	12.77%	1.32%	0.18%
Model iii	Mean			0.0802	5.1720	7.0842	5.5059
	Std			0.0532	0.0859	0.2127	0.0057
	Prct Err.			-115.70%	14.93%	1.20%	0.11%
Model iv	Mean					4.3454	5.5711
	Std					0.2784	0.0102
	Prct Err.					-37.92%	1.29%
Model v	Mean			0.4791	6.7995		
	Std			0.0403	0.1105		
	Prct Err.			-193.78%	51.10%		

Table 5.3: All models parameter estimates with 10% random censoring Sim02

		$\widehat{p}_1$	$\widehat{p}_2$	$\ln(\widehat{\sigma}_1)$	$\widehat{\beta}_1$	$\widehat{\gamma}_2$	$\widehat{\beta}_2$
Sim02	Population	0.75	0.20	ln(0.6)	4.50	7.00	5.50
Model i	Mean	0.7199	0.2133	-0.4930	4.5340	7.1623	5.5067
	Std	0.0150	0.0139	0.0284	0.0238	0.4183	0.0111
	Prct Err.	-4.01%	6.66%	-3.48%	0.75%	2.32%	0.12%
Model ii	Mean	0.7893		-0.2830	4.6852	7.1526	5.5066
	Std	0.0138		0.0311	0.0298	0.4185	0.0111
	Prct Err.	5.24%		-44.61%	4.12%	2.18%	0.12%
Model iii	Mean			-0.2961	4.6575	7.1541	5.5145
	Std			0.0349	0.0317	0.5273	0.0255
	Prct Err.			-42.04%	3.50%	2.20%	0.26%
Model iv	Mean					3.4853	5.8049
	Std					0.2636	0.0363
	Prct Err.					-50.21%	5.54%
Model v	Mean			-0.1351	4.8827		
	Std			0.0250	0.0302		
	Prct Err.			-73.54%	8.51%		

Table 5.4: All models parameter estimates with 10% random censoring Sim03

		$\widehat{p}_1$	$\widehat{p}_2$	$\ln(\widehat{\sigma}_1)$	$\widehat{\beta}_1$	$\widehat{\gamma}_2$	$\widehat{\beta}_2$
Sim03	Population	0.20	0.40	ln(0.6)	4.50	7.00	5.50
Model i	Mean	0.1829	0.3789	-0.5062	4.5105	7.0732	5.5028
	Std	0.0124	0.0158	0.0547	0.0451	0.2924	0.0079
	Prct Err.	-8.55%	-5.28%	-0.91%	0.23%	1.05%	0.05%
Model ii	Mean	0.2673		-0.2494	5.0045	3.4154	5.9650
	Std	0.1800		0.4460	0.9691	1.7145	0.2275
	Prct Err.	33.67%		-51.18%	11.21%	-51.21%	8.45%
Model iii	Mean			0.6531	6.9373	7.0832	5.5040
	Std			0.0396	0.1428	0.3064	0.0108
	Prct Err.			-227.86%	54.16%	1.19%	0.07%
Model iv	Mean					2.6689	6.0124
	Std					0.0619	0.0231
	Prct Err.					-61.87%	9.32%
Model v	Mean			0.6483	7.2475		
	Std			0.0366	0.1306		
	Prct Err.			-226.91%	61.06%		

Table 5.5: All models parameter estimates with 10% random censoring Sim04

		$\widehat{p}_1$	$\widehat{p}_2$	$\ln(\widehat{\sigma}_1)$	$\widehat{\beta}_1$	$\widehat{\gamma}_2$	$\widehat{\beta}_2$
Sim04	Population	0.10	0.30	ln(0.6)	4.50	7.00	5.50
Model i	Mean	0.0908	0.2778	-0.5142	4.5063	7.0630	5.5015
	Std	0.0091	0.0144	0.0775	0.0646	0.3383	0.0091
	Prct Err.	-9.21%	-7.39%	0.66%	0.14%	0.90%	0.03%
Model ii	Mean	0.0870		-0.4724	4.5313	2.3566	6.3422
	Std	0.0089		0.0963	0.0753	0.0574	0.0329
	Prct Err.	-13.00%		-7.52%	0.70%	-66.33%	15.31%
Model iii	Mean			0.8487	8.2127	3.4140	5.8806
	Std			0.0628	0.3786	1.6457	0.1644
	Prct Err.			-266.15%	82.50%	-51.23%	6.92%
Model iv	Mean					2.4096	6.2594
	Std					0.0562	0.0296
	Prct Err.					-65.58%	13.81%
Model v	Mean			0.8850	8.8388		
	Std			0.0481	0.2418		
	Prct Err.			-273.24%	96.42%		

The results for Models I & III are displayed above in tables 7.3 to 7.4 without random censoring results. The tables display the average of the $20,000$ parameter estimates along with the standard deviation for these estimates. The percent error (labelled "Prct Err.") is calculated as $\left(\widehat{\theta}-\theta\right)\Big/\theta$ and provides an indication of how well each of the models performs in finite samples.

10 Diagnostics Test Results

The following section details results for the Kolmogorov-Smirnov, Anderson-Darling and Wald Tests summarised in section 4.3 of this paper. The Aderson Darling and Kilmogorov Smirnov statistics are tests of distribution assumptions with the null hypothesis being that the variables follow the standard normal distribution. In each case the tests are applied to normalised residuals to determine if they follow the distribution assumptions specified in this paper. Results have been presented below for the Gamma-Weibull distribution assumptions.

Table 6.1: Gamma-Weibull Models - Anderson Darling (AD) and Kolmogorov-Smirnov (KS) Distributional Test Results

	Term				
	12	**24**	**36**	**48**	**60**
	Write Off (Gamma)				
Sample Size	334	2446	4680	5229	15876
AD	6.42	3.09	13.03	11.72	17.67
AD Adj	6.43	3.09	13.04	11.72	17.67
AD P-Value	0.0000	0.0000	0.0000	0.0000	0.0000
KS	0.96	0.92	0.85	0.76	0.68
KS P-Value	0.0000	0.0000	0.0000	0.0000	0.0000
Critical Value	0.0665	0.0247	0.0179	0.0169	0.0097
	Early Repayment (Weibull)				
Sample Size	8398	63479	104296	87385	185408
AD	239.04	1643.16	1997.01	1057.99	1204.29
AD Adj	239.06	1643.18	1997.02	1058.00	1204.29
AD P-Value	0.0000	0.0000	0.0000	0.0000	0.0000
KS	0.75	0.73	0.70	0.67	0.64
KS P-Value	0.0000	0.0000	0.0000	0.0000	0.0000
Critical Value	0.0133	0.0049	0.0038	0.0041	0.0028

Note: "AD Adj" is the AD statistic with an adjustment for small sample sizes

The Wald test was conducted to assess if the parameter estimates of a regression of the log negative log of the empircal survival function (dependent variable) on the log of time (explanatory variable) were different from the optimised parameters of the model in this paper. The results to test the hypotheses $\widehat{\eta}_1 + \widehat{\gamma}_{Lj}\widehat{\boldsymbol{\beta}}_{Lj} = 0$ and $\widehat{\eta}_2 - \widehat{\gamma}_{Lj} = 0$ are presented in the tables 6.2 to 6.6 below. There were only two occasions when there was insufficient evidence to reject the null hypothesis. Also see equations 21 and 22 for regression specifications.

Table 6.2: Wald test results for 12 month term relevant models

	Write Off			Prepayment		
12 Term	W_1W_2	W_1G_2	$W_1Ł_2$	W_1W_2	G_1W_2	$Ł_1W_2$
$-\widehat{\gamma}_{Lj}\widehat{\boldsymbol{\beta}}_{Lj}$	-12.852	-12.852	-12.851	-15.078	-15.077	-15.104
$\widehat{\eta}_1$	-28.768	-28.768	-28.768	-12.932	-12.932	-12.932
P-Value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\widehat{\gamma}_{Lj}$	2.1105	2.1106	2.1105	2.7883	2.7883	2.7931
$\widehat{\eta}_2$	4.2315	4.2315	4.2315	2.1771	2.1771	2.1771
P-Value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 6.3: Wald test results for 24 month term relevant models

24 Term	Write Off			Prepayment		
	W_1W_2	W_1G_2	$W_1Ł_2$	W_1W_2	G_1W_2	$Ł_1W_2$
$-\widehat{\gamma}_{Lj}\widehat{\boldsymbol{\beta}}_{Lj}$	-16.137	-16.145	-16.150	-14.705	-14.705	-14.705
$\widehat{\eta}_1$	-24.368	-24.368	-24.368	-12.773	-12.773	-12.773
P-Value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\widehat{\gamma}_{Lj}$	2.5228	2.5240	2.5248	2.3963	2.3962	2.3962
$\widehat{\eta}_2$	3.3045	3.3045	3.3045	1.9885	1.9885	1.9885
P-Value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 6.4: Wald test results for 36 month term relevant models

36 Term	Write Off			Prepayment		
	W_1W_2	W_1G_2	$W_1Ł_2$	W_1W_2	G_1W_2	$Ł_1W_2$
$-\widehat{\gamma}_{Lj}\widehat{\boldsymbol{\beta}}_{Lj}$	-15.547	-15.566	-15.586	-13.242	-13.242	-13.243
$\widehat{\eta}_1$	-21.476	-21.476	-21.476	-12.235	-12.235	-12.235
P-Value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\widehat{\gamma}_{Lj}$	2.3426	2.3456	2.3491	2.0385	2.0385	2.0386
$\widehat{\eta}_2$	2.7725	2.7725	2.7725	1.8308	1.8308	1.8308
P-Value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 6.5: Wald test results for 48 month term relevant models

48 Term	Write Off			Prepayment		
	W_1W_2	W_1G_2	$W_1Ł_2$	W_1W_2	G_1W_2	$Ł_1W_2$
$-\widehat{\gamma}_{Lj}\widehat{\boldsymbol{\beta}}_{Lj}$	-14.060	-14.092	-14.134	-11.947	-13.242	-11.950
$\widehat{\eta}_1$	-18.572	-18.572	-18.572	-11.911	-11.911	-11.911
P-Value	0.0000	0.0000	0.0000	0.1437	0.0000	0.1131
$\widehat{\gamma}_{Lj}$	2.0786	2.0837	2.0911	1.7893	2.0385	1.7897
$\widehat{\eta}_2$	2.3214	2.3214	2.3214	1.7518	1.7518	1.7518
P-Value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 6.6: Wald test results for 60 month term relevant models

60 Term	Write Off			Prepayment		
	W_1W_2	W_1G_2	$W_1Ł_2$	W_1W_2	G_1W_2	$Ł_1W_2$
$-\widehat{\gamma}_{Lj}\widehat{\boldsymbol{\beta}}_{Lj}$	-12.776	-12.819	-12.908	-10.783	-10.786	-10.792
$\widehat{\eta}_1$	-17.070	-17.070	-17.070	-11.217	-11.217	-11.217
P-Value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\widehat{\gamma}_{Lj}$	1.8747	1.8815	1.8970	1.5911	1.5914	1.5925
$\widehat{\eta}_2$	2.1172	2.1172	2.1172	1.6247	1.6247	1.6247
P-Value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000