Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

MASTERS THESIS

Competing Risks of Default and Prepayment of Mortgage Market

Author:
Oluwatobiloba John Olajubu

Supervisor:
Prof. Simon Wilson

A thesis submitted in fulfilment of the requirements for the degree of Master of Science

in the

School of Computer Science and Statistics, Statistics

April 2, 2020

Declaration

I hereby declare that this project is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at `http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

Signed: ______________________ Date: ______________________

Abstract

Using a large data set on the Single family home loans from The Federal Home Loan Mortgage Corporation (FHLMC), sponsored by the US government, this research studies the economic factors affecting the competing risks of prepayment and default. Since the subprime crisis, understanding the competing risks of prepayment and default on the single-family mortgage market have become increasingly important. Refinancing incentives on the mortgage market has been triggered by the current low interest rate environment. The paper presents a unified model of the competing risks of default and default mortgage termination, considering the two hazards as independent competing risks, which are jointly estimated. It also accounts for the unobserved heterogeneity among borrowers, and estimates the unobserved heterogeneity simultaneously with the parameters and baseline hazards associated with default and prepayment functions. This thesis also examines the different models that can be used to predict current and future prepayment rates. In mortgage research applications the time to occurrence of one event of interest which may be caused by another competing event is investigated. The most widely used prepayment and default models are competing risk models and option theoretic models. However, this thesis places more emphasis on competing risk models. As we know, survival analysis examines and models the time to an event (usually failure of a system and possibly subject to right censoring) and one of the most common tools 'Cox proportional-hazards regression model' is used to study the dependency of survival times on the predictor variables. Important risks determinants include the specific characteristics of the borrower, specific characteristics of the loan and macro-economic variables. Variable selection procedures are used to identify the most significant risk drivers. Tests on parameter stability are conducted to check forecasting accuracy for models. A problem, however, is heterogeneity between mortgages-suggesting there is future potential research to be done on this topic.

Keywords: Cox Proportional-hazards model . Competing risks . Mortgages . Default. Prepayment.

Acknowledgements

I would first like to thank my thesis advisor Professor Simon Wilson of the School of Computer Science and Statistics at University of Dublin, Trinity College Dublin. He consistently allowed this paper to be my own work, but steered me in the right direction whenever he thought I needed it. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Masters research study.

Finally, I must express my very profound gratitude to my mum and to my Wife Christianah for providing me with unfailing support and continuous encouragement throughout my year of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Abbreviations

FG Fine and Gray

CIF Cumulative Incidence Function

PH Proportional hazards

csh cause-specific hazards

BIC Bayesian information criterion

AIC Akaike information criterion

FHLMC Federal Home Loan Mortgage Corporation

MI Mortgage insurance percentage

CLTV Combined loan-to-value

DTI Debt-to-income

UPB Unpaid principal balance

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Competing risks analysis is a topic in survival analysis, where individuals can fail from one of several competing causes. Investigators are often interested in the distribution of time to failure for a main event in the presence of other competing risks.

One of the key challenges of competing risk survival analysis lies in the estimation of the survival function $S(t)$ or equivalently the cumulative incidence function $F(t)$. The method of estimating cumulative incidence, calculated as 1 - $S(t)$ from the Kaplan-Meier (KM) curve, is known to be inappropriate and may lead to the overestimation of cumulative incidence in the presence of competing events, due to the fact that the Kaplan-Meier curve assumed the individuals who experienced the competing events as censored.

A mortgage borrower is technically delinquent after a monthly mortgage payment due date has been missed. However, most lenders give the borrower a substantial period of time (usually 90 days but varying depending on the lender) to bring the loan to its current status by making up all the missed payments plus the associated late fees. If the borrower is still delinquent after a certain period of time, the lender shall initiate a foreclosure procedure. Indeed, some of these loans may be prepaid, reinstated or modified (term extension or other changes to reduce the monthly payment) or may have other alternative outcomes and these outcomes can be seen as competing risks. As Ambrose and Goetzmann (1998) have noted, however, defaults are usually defined in the literature as being delinquent in mortgage payment for ninety days.

There is a rich literature on mortgage risk modeling; Quercia and

Stegman (1992) and Leece (2012) are a literature review. A large category of models is based on the ruthless default assumption that a rational lender will maximize his or her assets by defaulting on mortgage if the value of the mortgage exceeds the house value and pre-pay if the house's market value exceeds that of the book. These models take an option theoretic approach and presume, for instance (Kau, 1990), the mortgage value and the prepayment and default options are calculated by the stochastic behaviour of variables such as property prices and the interest rates. As a result, other factors, such as characteristics of the borrower, transaction costs and many more, are assumed to have no effect on the value of the mortgage and the property underlined under the option theoretic approach. The ruthless assumption of default is not widely accepted in literature, and several evidence against the validity of the assumption has been presented in many works. Furthermore, the implementation of this model requires the availability of performance level data on individual loans over time, which is usually difficult to obtain (Soyer and Xu, 2010a).

Soyer and Xu (2010b) explained a review work by Deng and Gabriel (2006) that evaluates hazards rates under the default and prepayment, viewed as competing risks, and the distribution of the mortgage duration time, which is obtained by the minimum of the random variables. These latest advances boost the significance of default risk to mortgage lenders, be they are portfolio lenders or mortgage backed securities (MBS) customers. Whilst my upmost focus is on time to event (default or prepayment), I understand that assessment of default time is a key value for the financial institutions who offer loan modification and loss mitigation programs which are available in US as well as Europe, as in Bhattacharya et al. (2019). In this thesis, we considered modeling duration of single-home mortgages. In doing so, we model default and prepayment probabilities simultaneously using competing risks proportional hazards models. Also, we adopted a survival viewpoint in the analysis and develop predictive inferences using the Cox proportional hazard model. In the model, we included both individual and aggregate level covariates. Our interest is modeling time to default and prepayment (failure rates) using competing risks model. In particular the use of the model to predict when a mortgage might default (so we are modelling the time to default from the start of repayment). One question is whether such a model is a good one to explain a particular mortgage data set?

Also we examined the effect of macroeconomic variables on default and prepayment rates, as well as the characteristics of borrower, property and loan. Here we adopted an important model in reliability and risk analysis, Cai and Zeng (2014) where the failure rate is defined in terms of the explanatory variables and a regression on the failure times as a function of those variables is done. There are several independent causes of an event occurring, and so it actually occurs when the first of these 'competing' causes occurs. Thus, the time to each cause is modelled with a failure rate. It is important to examine the following questions when the probability of prepayment and default are simultaneously high : what economic factors affect competitive risks of prepayment and default and to what extent do they influence the risks?

1.1 Research proposal

1.1.1 Problem statement

A survival analysis can be complicated because of the issues related with censoring. The most common form (usually in the case of survival analysis) is right censoring where the study ends or the subject withdraws before the endpoint event has occurred (Sączewska-Piotrowska, 2015a). The tools applied force the ideal method of collecting statistical data. In the study of competing risk data, Prentice et al. (1981) identified three different issues:

- The estimation of the relationship between covariates and the rate of occurrence of cause-specific failure.
- The analysis of the interrelation between types of failure under a specific set of conditions.
- The estimation of failure rates for certain types of failure given the removal of some or all other failure types.

1.1.2 Objectives of the research

The goal of this study is to determine the potential of applying survival time analysis techniques in competing risks model estimation for default and prepayment of mortgage data. Survival analysis is a branch of statistics that models the time until an event happens. the event can be, for instance, competing risk of mortgage default and prepayment. The time until such an event happens is called survival time. The individual exposed to this risk is the "subject" of the event. One of the strengths of Survival analysis is that it can handle the censoring of observations. Observations are called censored when the information about their survival time is incomplete; the most commonly encountered form is right censoring.

1.1.3 Research questions

In order to achieve the objective of this study, the following main research question is formulated: (1) what is the potential of applying survival analysis methods for estimating "time to default and prepayment (failure rate)" and which metrics should be used to evaluate the performance of these methods? (2) the competing risks model used to predict time to default from the start of prepayment, whether such a model is a good one to explain a particular mortgage data set?

1.2 Organisation of thesis

The rest of the thesis is organized as follows:

In Chapter 2, a literature review and outline of the problem at hand is presented, with a description of the most common competing risks models, as well as an overview of central concepts that are important in any time-to-event analysis, including censoring, survival functions, the hazard function, and cumulative hazards. Bound up with this, it also highlights key issues that need to be considered when faced with a competing risk scenario. In addition, a review of the extent to which competing risks are recognised is given in the broad related literature. The latter identifies research gaps that motivate an empirical analysis of my mortgage default and prepayment case study in the next chapter.

In Chapter 3, the dataset, explanatory variables and dependent variables used in this paper are explicitly discussed.

Chapter 4 outlines the methodologies we used. We first introduce the idea of Competing Risk model in a rigorous sense and then follow this up outlining Cox proportional-hazards (PH) model, its essential use to obtaining Survival estimates that compliments multivariable regression models for time-to-event data, both for prediction and illustration of covariate effects. The chapter focus on proportional cause-specific hazards Cox regression for each of the two competing risks default and prepayment.Furthermore, we presents the Fine and Gray (FG) approach, also a Cox model, but for the sub-distribution hazard attached to the cummulative incidence, to competing risks which provides a clear understanding of the competing risk process by considering both the hazard and the cumulative incidence of each competing risk.

Chapter 5 is our results and observation section. This chapter shows summary of results. Visualization is performed using tables and charts. However, due to a large number of analyzes and outcomes, it was not feasible to demonstrate all the outcomes in these parts and, furthermore, to maintain the clarity of this paper. As a result, on the basis of personal preference, most of the relevant results are displayed.

Chapter 6 concludes the thesis with main insights gained, limitations and as well as an outlining of potential future work.

Chapter 2

Literature Review

2.1 Introduction

This chapter gives an overview of the competing risk methodology within the framework of survival analysis, giving the background to the empirical analysis in the latter chapter. Within this chapter I aim to outline the pre-existing literature in competing risks model and as well highlights key issues that need to be considered when faced with competing risks. The concepts will be illustrated with reference to literature on mortgage market, in order to effectively link the ideas. This chapter introduces different statistical methods to account for competing risks and will address some of the statistical challenges arising from mortgage studies. This chapter will outline my works contribution to academia, and how it advances the current literature.

2.2 Competing risks in the survival analysis

Survival analysis is commonly used in time-to-event data where an individual is followed up to see whether, and when, they experience an event of interest, and major methodological advances have recently been made in the regression analysis of such data.

In the literature of Bhattacharya et al. (2019), the authors introduced a bayes model, i.e, a bayesian approach to modelling the time to mortgage prepayment or default as a function of mortgage covariates. They proposed a competing risks model that allows one to take account of the fact that an observation of a mortgage prepay is also a censored observation of a default, and vice versa; hence observation of one does contain information about the other that should be used in inference (Bhattacharya et al., 2019). There have been some difficulties with the inference in their contributions, specifically for defaults that were only a small percentage of the data. In addition, the extension of the authors' research considered heterogeneity between mortgages using random effects model, which would allow one to explore the variability between mortgages and identify clusters of mortgages with similar properties. Some difficulties with inference were encountered, resulting to slowdown in the computation of time.

There are many methods available for modeling competing risk data. Typical approaches used for the analysis include Kaplan-Meier survival estimator and Cox proportional hazard regression modelling. However, there are often situations where borrowers are at risk of two or more mutually exclusive events (in our case study, default and prepayment) affecting each other's risk, and this requires a different approach. There is a competing risk scenario in such situations. The risks are said to be "competing" with each other to be the first event, for instance, observing whether a borrower is likely to default or prepay. The default and prepayment are "competing risks", and the first of them to occur is observed.

One of the most widely studied topics in the literature on mortgage financing is the termination of loans by default of the borrower. In early studies using option models in real estate finance, researchers analyze either prepayment or default decisions, but not both. The pricing of mortgage contracts is complicated, mainly because of the options available to the borrower for default or prepayment. These options are distinct but not independent. Thus, the economic value of the default option can not be precisely calculated without, at the same time, considering the financial incentive for prepayment with risk preferences and other distinctive differences across borrowers varying.

The contingency compensation models established by Jonathan (2018), John C. et al. (1985) and others provide a consistent guide to borrower behavior and this model has been applied to the mortgage market by a number of studies. Quigley and Order (1990) approximate a risk-based prepayment function and check how borrowers exercise their prepayment rights in a manner consistent with the ideal approach defined in the contingency claims models. The dataset used is from the Federal Home Loan Mortgage Corporation (Freddie Mac) and includes 6375 30-year fixed-rate single-family mortgages issued during the period 1976-1980. The results show that the value of the call (prepayment) option being "in money" has a very positive effect on the risk of prepayment, and that the initial equity has an insignificant and negative impact on the risk of prepayment. However, after several testing on prepayment behaviour, the authors draw a conclusion on prepayment option not be exercised Implicitly. Furthermore, the authors stated default is not considered in the choice model, thus, a more general model should consider that loans are about to be prepaid as well as defaulted.

2.3 Kaplan-Meier Survival analysis

The previous section highlighted the need to recognize competing risks when they exist, because a different approach to analysis is required. This section explains why the Kaplan-Meier approach in standard survival analysis is not appropriate when competing risks are present.

Non-parametric methods are not related to an underlying distribution. The mostly used non-parametric method is the Kaplan-Meier estimator, also known as the product limit estimator. There is a significant difference between competing risks and standard survival analysis. With standard survival analysis, there is a one-to-one correspondence between the hazard and survival functions. This implies that when the hazard increases, the probability (1-survival) of the event occurring will increase. Similarly, a decrease in the hazard will lead to a decrease in the probability of the event occurring.

In standard survival analysis, the survival time of subjects who do not experience the outcome of interest (default or prepayment) during the observation period is censored at the end of follow-up. In those cases, we do not know whether and when such a borrower will experience the event, we only know that he or she has not done so by the end of the observation period. Conventional methods for survival analysis that ignore competing events, such as the Kaplan – Meier method and the standard Cox proportional hazard regression, may not be appropriate in the presence of competing risks, and alternative methods specifically designed to analyze competing risk data should then be applied.

A controversial issue with competing risks is whether survival is an appropriate quantity to estimate. Some have stated that the ideal estimate is net or marginal survival e.g Dignam and Kocherginsky (2008). The

Kaplan-Meier approach is based on the assumption of non-informative censoring. The censoring mechanism should not provide any information that would influence the distribution of event times. For instance, at the time of censoring, borrowers who are censored should not be any different from those borrowers who are uncensored.

In standard survival analysis, as there is only one possible event, mortgagors either experience the event or are censored at the time they were last known not to have the event. For each specific event, the analysis is carried out under a competing risk scenario by treating them as an event of interest.

Competing events are censored since they are not an event of interest. Using Kaplan-Meier method introduces bias estimates that increases the probability of the event of interest. The explanation is that the uncensored individuals left in the dataset are not indicative of all those who did not experience the event. For example, a censored case where the loan did not experience default at the period of data gathering. Early repayment and mature cases are subject to censoring. This type of censoring is used in models where default is the only event of interest. There are several contributions in the literature that demonstrate the bias in the Kaplan-Meier approach by contrasting the Kaplan-Meier curve with the corresponding cumulative incidence curve, taking into account competing risks.

However, for standard survival analysis, the probability of a specific event occurring at any given time is of interest in the presence of competing risks. The cumulative incidence that takes into account competing risks is an appropriate estimate for this. The calculation of this involves the hazard of a specific event, as well as the hazard of each competing risk. This is explained in detail in the next sub-section.

2.3.1 Cumulative incidence for competing risks

As noted earlier, there is a need to take into account of any competing risks when estimating the cumulative incidence of an event of interest. Therefore, it acknowledged the fact that the cumulative incidence of a specific event should take into account both the hazards of each competing event and the event of interest. There are many contributions describing how to estimate the cumulative incidence function, see Dignam et al. (2012); Andersen et al. (2012); Bakoyannis and Touloumi (2012).

The formal equation to calculate the cumulative incidence denoted as

$CIF_k(t)$, that is, the probability of the occurrence of the event due to k before or at the time t is defined (Dignam et al. (2012) as:

$$CIF_k(t) = Pr(\text{ failure time } T \leq t, \text{cause} = k) = \int_0^t S(u)\lambda_k(u)\mathrm{d}u$$

where $S(t)$ is the survival free from any of the events up to time t and $\lambda_k(t)$ is the cause-specific hazard for the event of interest.

The key driver behind the cumulative incidence in the presence of competing risks is the cause-specific hazard. As described by Dignam et al. (2012), this represents the likelihood of failure due to cause k at a time when no failure of any kind has yet occurred. This is formally defined (Putter et al. (2007)) as:

$$\lambda_k(t) = \lim_{\Delta t \to 0} \frac{\text{Prob}(t \leq T < t + \Delta t, D = k | T \geq t)}{\Delta t}$$

where T is the failure time and D is the cause of failure. It is the instantaneous rate of failure of cause k in the small time interval [t, t + Δ t] given that no failure of any kind occurred prior to time t.

The $S(t)$ element of the cumulative incidence $CIF_k(t)$ also incorporate the cause-specific hazard of each event, that is, the event of interest and any competing events. Putter et al. (2007) outlined $S(t)$ as:

$$S(t) = \exp(-\sum_{k=-1}^{K} \Lambda_k(t)) \qquad \text{where } \Lambda_k(t) = \int_0^t \lambda_k(s)\mathrm{d}s$$

The general survival function S(t) can be estimated with the widely applied Kaplan-Meier estimator. Bakoyannis and Touloumi (2012) outlined a non-parametric plug-in estimators for estimation purposes as:

$$\widehat{F} = \sum_{m:t_m \leq t} \widehat{\lambda}_k(t_m)\widehat{S}(t_m - 1), \qquad \text{where } \widehat{\lambda}_k(t_m) = \frac{d_k m}{n_m}, \widehat{S}(t_m - 1) = \prod_{i=1}^{m-1}(1 - \frac{d_i}{n_i})$$

t_m is the mth-ordered failure time,

n_m is the number of subjects at risk at the time t_m,

d_m is the total number of failures (from any cause) and

$d_k m$ is a number of events that have occurred due to the cause k.

In addition, as Pfirrmann et al. (2011) points out explicitly, the cumulative incidence estimator above can be expressed as:

$$\widehat{F_k(t_m)} = F_k(t_m - 1) + \widehat{p}_k(t_m),$$

where $\widehat{p}_k(t_m) = \widehat{\lambda}_k(t_m) \times \widehat{S}(t_m - 1)$ is the unconditional probability of observing event k at the time t_m.

Many of the authors that outline the calculation for cumulative incidence further offer a step-by-step demonstration of that calculation. Each of these findings provides a useful perspective on how the cumulative incidence of each event becomes influenced by the cause-specific hazard: Dignam et al. (2012), Pfirrmann et al. (2011), Bakoyannis and Touloumi (2012), Andersen et al. (2012) and many others.

A literature by Satagopan et al. (2004) provides illustration on non-parametric estimation of cumulative incidence of the event of interest taking into account the informative nature of censoring due to competing risks. In their discussion, they provide a two-steps illustration of the cumulative incidence. The authors state: " the probability failure for the event of interest is one minus the probability of survival up to a given time". However, as already mentioned in Section 2.3, survival is not an appropriate term to be used for a specific event in a context of competing risks and focus should instead be on cumulative incidence.

Interest is often expressed in comparing the cumulative incidence curves between different groups. Robert J. (1988) developed a K-test for this purpose similar to the log-rank text for the Kaplan-Meier survival curve comparison.

2.4 Competing risks modelling approaches

Under the framework of the competing risks, the previous Section 2.3 specified a non-parametric cumulative incidence estimator with no covariates considered. This section explains in detail the approaches to modeling competing risks that can accommodate and account for the effect of covariates on different hazard functions.

The section continues by summarising the two most commonly used approaches under the classical latent failure times competing risks framework, namely the FG sub-distribution proportional hazard model and the cause-specific hazard Cox approach. The cause-specific approach has a long history of application, and proportional hazards models are typically assumed (Prentice et al., 1978). The cause-specific hazard is a major driving force in analysis that takes into account competing risks. In contrast, the model of Fine and Gray (1999) targets the subdistribution function directly and has become a very popular choice in practice, the approach is often seen as default choice of modelling competing risks events, but in fact the cause-specific hazard and the subdistribution hazard (cumulative incidence) complement each other, thus, they should be modelled for each competing risks (Bakoyannis and Touloumi, 2012; Andersen et al., 2012). The calculation of the cumulative incidence for a specific event involves the cause-specific hazard of each event. Latouche et al. (2013) argue that a proper competing-risks analysis should report results on all the cause-specific hazards and cumulative incidence functions. In this thesis we consider both approaches for modelling competing risk to aid interpretation and the perspectives they offer.

2.4.1 Classical latent failure times competing risks

The term ' competing risk problem' has come to encompass the study of any failure process in which there is more than one distinct cause or failure types: see R. L. et al. (1978). Latent failure times approach to competing risks assumes that each individual has a potential failure time for each type of failure. In fact, the failure that occurs first is observed, with the other failure times considered to be latent.

The formation of competing risks problem (Gail, 1975) introduces a latent failure times Y_1 , . . . , Y_m for each of m failure types and that only $T =$ min $(Y_1, \ldots, Y_m)$ is observed. Some distinct problems arise in the analysis of failure times in competing risks which are; estimation of failure rates of some causes given the removal of some or all other causes, inference on the effect of covariates on type of failure and lastly, is the interrelations among failure types under specified conditions (R. L. et al., 1978).

A fundamental concept in the approach to latent failure times is that of the joint survival function

$$Q(Y_1 , \ldots , y_m ;\mathbf{z}) = P(Y_1 \geq y_1, \ldots, Y_m \geq y_m),$$

where z is a covariate vector restricted to be time independent in this section. All the problems stated above are then formulated in the terms of Q. There is also a related concept which is the marginal "survival" distribution

$$Q(Y_j \; ; \mathbf{z}) = Q(0, ..., 0, Y_j, \; . \; . \; ., \; 0 \; : \mathbf{z}).$$

Nevertheless, both the marginal "survival" distribution and the joint survivor function are affected by the identification problem. They can only be identified from the data observed with additional assumptions. One such strong assumption is the independence of the various latent failure times: see (R. L. et al., 1978), (Bakoyannis and Touloumi, 2012). These assumptions are not testable. Thus, other approaches based on observable quantities have largely superseded latent failure times approach. Two of these approaches are described in the next two parts of this section.

2.4.2 FG subdistribution hazard modelling approach to competing risks

In the presence of competing risks, the one-to -one correspondence between hazard and survival that exists with the standard survival analysis does not necessarily hold (as mentioned earlier in previous section 2.3). As a result, the effect of a covariate on a cause-specific hazard for a particular cause may differ from its corresponding effect on the probability of an event occurring.

Fine and Gray (1999) have developed an alternative approach that maintains a one-to - one link. This method overcomes problems with the interpretation of a cause-specific hazard approach. Its main advantage in setting competing risks is that, unlike the cause-specific hazard, it corresponds one - to - one to the cumulative incidence of the event of interest [He et al. (2016)].

Based on Gray (1988) subdistributional hazard technique, Fine and Gray (1999) proposed a proportional subdistribution hazards model. The hazard of the subdistribution takes the usual form of a hazard function, based on the improper τ :

$$
\begin{aligned}
\lambda_k(t;\mathbf{Z}) &= \lim_{\Delta_t \to 0} \frac{1}{\Delta_t} Pr(t < \tau \leq t + \Delta | \tau > t) \\
&= \lim_{\Delta_t \to 0} \frac{1}{\Delta_t} Pr(t < T \leq t + \Delta, \delta = k | T > t \text{ or } [T \leq t, \delta \neq k], \mathbf{Z}) \\
&= \frac{\partial (F_k(t;\mathbf{Z}) / \partial t}{1 - F_k(t;\mathbf{Z})} \\
&= \frac{-\partial \log[1 - F_k(t;\mathbf{Z})]}{\partial t}
\end{aligned}
$$

where T is the failure time, $F_k(t;\mathbf{Z}) = Pr(t \leq T, \delta = k|\mathbf{Z})$,

$\mathbf{Z}$ is the p $\times$ 1 bounded time independent covariate vector and

δ ε (1, ..., k) is the cause of failure.

The risk set associated with this subdistribution hazard is somewhat counter-intuitive. The borrower experiencing a competing event remains at risk even if the occurrence of the event results in a zero or otherwise altered likelihood of the event under consideration. Analogous to cause-specific proportional hazards and standard survival analysis, subdistribution hazard can be modelled in a subdistribution hazards approach (Fine and Gray (1999)) as:

$$\lambda_k(t;\mathbf{Z}) = \lambda_0(t) \exp(\mathbf{Z}^T \beta)$$

where $\lambda_0(t)$ is some common baseline subdistribution hazard function. However, $\lambda_k(t;\mathbf{Z})$ can be extended to time-varying covariates by replacing $\exp(\mathbf{Z}^T \beta)$ for $\exp(\mathbf{Z}(\mathbf{u})^T \beta)$. The subdistribution hazard for event k is defined as the probability for a subject to fail from cause k in an infinitesimal small time interval, given the subject experienced no event until time t or experienced an event other than k before time t.

There is a direct relationship between the CIF and subdistribution hazard function (He et al. (2016)):

$$F_k(t;\mathbf{Z}) = 1 - \exp\left\{ -\left(\int_0^t \lambda_0(u) \mathrm{d}u \right) \exp^{(\mathbf{Z}(\mathbf{u})^T \beta)} \right\}.$$

The approach used to model the hazard of subdistribution for a particular cause depends on the reason for any censoring that may occur. The competing events are censored at a time just after last event of the type and also infinitely retained in the risk set with a time for the event under consideration. This perception is the same for subdistribution hazards (SH) λ_k and acting as improper random variable T written as:

$$T^* = I(\delta = k) \times T + 1 - \{I(\delta = k)\} \times \infty.$$

The cumulative incidence function $F_k(t;\mathbf{Z})$ stated above is also an improper probability distribution in the sense that $F_k(\infty;\mathbf{Z}) < 1$.

It has been shown that, in financial research studies, censoring time can depend on some covariates. To make asymptotically unbiased inferences, the censoring distribution is modeled and the survival probability for each individual is also estimated. While censoring is purely practical, the analysis can be carried out using Cox's proportional hazard approach. This is explicitly discussed in the next section.

2.4.3 Cox's proportional hazards (PH) modelling approach to competing risks

Another method which is commonly used in survival analysis is the Cox Proportional Hazards model (Cox, 1972). The Cox Proportional Hazards model can be used to model the impact of explanatory variables on the survival function and enables greater flexibility than other types of models. This paper focus on Cox PH modelling approach in the presence of competing risks. The approach is called cause-specific hazards Cox regression. It involves the fitting, for an individual with covariate values $\mathbf{Z} = (Z_1, ..., Z_p)$, a separate Cox hazard regression model for each of the competing events (causes).

FG proposed a Cox type regression model, i.e the cause-specific hazard Cox model of cause k ($k = 1, 2$), conditioned for an individual with covariate vector $\mathbf{Z} = (Z_1, Z_2, ..., Z_p)$, is defined as:

$$\begin{aligned} \lambda_k(t|\mathbf{Z}) &= \lambda_{k0} \exp\,[\sum_{k=1}^{p} \beta_k \times Z_k] \\ &= \lambda_{k0} \exp\left\{\beta_k^{\mathbf{T}} \mathbf{Z}\right\}, \end{aligned}$$

where λ_{k0} is the baseline function, i.e, cause-specific hazard of cause k,

$\beta_k^{\mathbf{T}} = (\beta_1, \beta_2, ..., \beta_p)$ denotes the covariate effects on cause k, i.e, the vector we try to estimate.

This is semi-parametric Cox regression approach (Cox, 1972) with a flexible unspecified cause-specific baseline hazard rate $\lambda_{k,0}$. As in common Cox regression, the cause-specific hazard rates are assumed to be proportional translating to covariate effects that are constant over time. This assumption can be checked by graphical methods using Schoenfeld residuals (Schoenfield, 1982). Several works also described extensions of the classical model presented by Cox (1972), modifications of the model as inclusion of time dependent covariates or time dependent effects for contructing a valid Cox regression model, for example (Therneau and Grambsch, 2000). Instead of the semi-parametric Cox regression model parametric models assuming e.g. exponentially, Weibull or gamma distributed event times can be performed (see (Klein and Moeschberger, 2003).

There is absolutely no 'error' term in this model as the randomness is implicit to the survival process. The baseline hazard function is assumed to be non-negative, time-constant and independent on the covariates. The baseline hazard function is the hazard function obtained when all covariates are set to zero. The cause-specific hazard is the instantaneous rate of cause k faced by those who have not encountered cause k or any of the competing events. A simple way to apply this cause-specific hazard approach is to fit a different Cox model for each cause k, censoring any competing events at their time of occurrence.

He et al. (2016) suggested that, in practice, the standard model control procedure can be used to examine the Cox model assumption for censoring distribution and to use the standard model building technique to identify the risk factors associated with censoring time. Let z_i be the fixed observed value for the pth individual's covariates, the predicted censoring survival probability model for the cumulative incidence curve in the form

$$\begin{aligned} G_k^{cox}(t:\mathbf{z}_p) &= P(C > t|\mathbf{Z}_p = \mathbf{z}_p) \\ &= \exp\left\{-\Lambda_{k0}(t)\ \exp(\beta_k^{\mathbf{T}}\mathbf{z}_p)\right\} \end{aligned}$$

is estimated as

$$\widehat{G_k^{cox}}(t:\mathbf{z}_p) = \exp\left\{-\widehat{\Lambda}_{k0}(t)\ \exp(\widehat{\beta_k}^{\mathbf{T}}\mathbf{z}_p)\right\}, \text{ He et al. (2016)}$$

where $\widehat{\beta}$ is the maximum partial likelihood estimate for β_k and $\widehat{\Lambda}_{k0}(t)$ is the standard Breslow estimator for the cumulative baseline censoring hazard

$\Lambda_{k0}(t) = \int_0^t \lambda_0(u)\mathrm{d}u$, i.e, the baseline survival function and also dependent on time only. Note that any administrative censoring events at time τ is not considered to be an event in the regression estimation. In this framework, it is easy to decide if covariates have a significant effect on the cumulative incidence curve for a specific cause of failure.

Nonetheless, it might be essential to also include time-dependent covariates. The Cox Proportional Hazards model can therefore be extended which implies that the covariates for a given subject may change over time. This is not a major focus in this thesis. The Extended Cox model, which includes time-dependent and time-independent explanatory variables, is defined as follows:

$$\lambda(t|\mathbf{Z}(t) = \lambda_0(t)\exp\left[\sum_{k=1}^{p_1}\beta_k Z_k + \sum_{j=1}^{p_2}\delta_j Z_j(t)\right],$$

where $Z_k = (z_1, z_2, ..., z_{p1})$ is a vector of the time-independent covariates and $Z_j = (z_1, z_2, ..., x_{p2})$ is a vector of the time-dependent covariates. The vector $\delta = (\delta_1, \delta_2, ..., \delta_{p2})$ represents the overall effect on the survival time of the time-dependent covariates and is not time-dependent. Here, an assumption of this model is that the covariate vector $Z_j(t)$ can change over time but the hazard model provides only one value of the coefficients for each time-dependent covariate in the model.

Cox regression is run on an augmented data set (Lunn and McNeil, 1995). Another preferred way to implement cause-specific hazard modeling is to use the "data argumentation" technique, which offers more flexibility and helps to solve the problems of over-fitting. Data-Augmentation approach requires setting up data in a long format, each individual having as many rows as the number of causes. An individual may encounter only one cause in a competing risk scenario, any other causes being censored at the time the cause occurred. It allows the relationship between the covariate and the cause to be applied, allowing a precise measure of the difference in the influence of the covariate between the causes. It also allows the researcher to check the effect of an interaction using hypothesis testing, which lets the modeler determine whether a specific effect of covariate can be appropriate across all causes. Furthermore, it allows the effect of covariate on one cause to be proportional to the effect on another cause, offering a different option for over-fitting. The contributions by Putter et al. (2007) and Lunn and McNeil (1995) explicitly provides more information on the implementation of cause-specific hazard \ data augmentation technique. There are other approaches to modeling competing risks, although they are not the primary focus of the paper. This will be discussed in the next section.

2.4.4 Other methods to modeling competing risks

Many literature discusses alternative approaches and extensions to the modeling of cause-specific hazards and cumulative incidences. Some of these contributions include the mixed modeling (Lau et al., 2008) and the vertical modelling (Nicolaie et al. (2010)). This thesis applies the two key approaches to the competing risks regression describe in the last two sections above.

Certain interesting methods explored by these authors include the extension of standard case where all event times are observed precisely or right-censored, to interval-censored and probably truncated data: see (Robins and Finkelstein, 2000) and (Hudgens et al., 2014). Further contributions describe approaches that may be used when a proportional hazard assumption is not met include modeling of time-varying cause-specific hazard (csh) ratios, those involving flexible modeling of competing risks (Scheike and Mei-Jie, 2008), and stratified competing risk regression (Zhou et al., 2011).

2.5 The effect of a covariate on the cumulative incidence and on the cause-specific hazards

There is a significant difference between competing risks and standard survival analysis. With standard survival analysis, there is a one-to-one correspondence between the hazard and survival functions. This section explains some key consequences of this one fundamental difference that affects the analysis and interpretation of competing risks.

One of the consequences, as previously discussed in section 2.3, is that Kaplan-Meier approach in standard survival analysis introduces bias estimates that increases the probability of the event of interest, thus, inappropriate approach to use when competing risks are present. The effect of the covariate on the cause-specific hazard is not generally the same as its effect on the cumulative incidence. It is difficult to identify covariates that predict the cumulative incidence likelihood through modeling the cause-specific hazards.

This is another main consequence. Zhou et al. (2011) contributions in literature explains that while estimates of the cumulative incidence curve are relatively easy to obtain, one issue is that it is difficult to summarize the impact of covariates on the cumulative incidence curve in a simple manner.

Furthermore, substantial covariate effects for the cause-specific hazards will have a significant impact on the cumulative incidence curve, but it is probable that the effect on the cause-specific hazard for cause 1 will be counteracted by an adverse effect on overall survival, and even though the covariate may be highly significant for all cause-specific hazards, it may not have an impact on the cumulative incidence (Zhou et al. (2011)). There are many papers in the literature that demonstrate this important aspect of competing risks, for example, Zhou et al. (2011), Putter et al. (2007), and Lambrecht et al. (2003). One factor that is not always emphasized in the literature is that the baseline hazard of the event of interest, and any competing risk, may also affect the occurrence of the cumulative incidence of an event. Nevertheless, the authors Putter et al. (2007), contributions in the literature show simulations of varying baseline hazards to demonstrate the effect this may have on the cumulative incidence.

2.6 Variable selection techniques with competing risks

It is important for any financial institution to examine the predicting power of different risk drivers to the model estimates. When developing the models one has to make a decision about which of these risk drivers to include in the model. There are several variable selection techniques used in statistics. With the competitive risk analysis, variable selection is less straightforward. In this thesis, we adopt the two main methods for selecting the variables: sequential methods and all-subset methods.

The first method is based on a sequence of tests and determines whether a variable should be included or removed from the current model. Adding or removing depends on the criteria for inclusion and normally includes a level of significance of $\alpha= 0.05$. Sequential methods are the most common methods for selecting variables, as well as the approach used in this analysis. Sequential methods may use forward selection, backward selection, or a combination of both. The second method is based on the fitting of all possible models and the selection of the best model based on the accepted criteria see (Sauerbrei et al., 2007).

The model is tested by using a chosen model fit criterion and repeated until no improvements in the model are exhibited. A combination of both forward selection and backward selection procedures is used test at each step which variable to be included or excluded (Sauerbrei et al., 2007). The likelihood-based approaches such as likelihood-ratio tests and information criteria, e.g Bayesian information criterion (BIC) and Akaike information criterion (AIC) are used in standard survival analysis because they are intuitive and easy to use. However, BIC will under-fit. AIC works well if there are three competing models to choose from. AIC is just a (smarter) restatement of $\mathbf{P}$-values so doesn't solve the fundamental problem if the number of competing models exceeds a few. There are some problems associated with stepwise variable selection.

The next two sections discusses the drawbacks of variable selection under the approaches to competing risk modeling with focus on both cause-specific hazards (csh) Cox approach and FG subdistribution proportional hazard modelling approach.

2.6.1 Variable selection and cause-specific hazards Cox approach

As previously mention in this thesis that there is a one-to-one correspondence between the hazard and survival functions. Section 2.5 explicitly outlined the consequence, stating that the effect of the covariate on the cause-specific hazard is not generally the same as its effect on the cumulative incidence. Therefore, in the context of competing risks, the likelihood-ratio test results and information criteria focused solely on a single cause-specific hazard should be treated with caution. Understanding one cause-specific hazard in isolation is similar to the naive Kaplan-Meier or Cox predictions. The problem with such an approach is that it does not have a direct interpretation of the survival probabilities, that are obviously reduced by the other risk: a modeller is basically disregarding the information about the events due to it. The results are liable to mis-interpretation and misleading conclusions.

However, a modeller could wondered whether to consider overall survival in modeling competing events. There is another approach previously discussed in section 2.4.2, based on Fine and Gray (1999) subdistribution proportional hazard, i.e. the cumulative incidence function. This is discussed in the next section.

2.6.2 Variable selection and FG subdistribution proportional hazard approach

FG subdistribution proportional hazard approach as previously discussed in section 2.4.2. There are many pitfalls and high probability of getting a misleading result in the field of model selection. The drawback here is that, in the risk set for a given event, individuals who may not have the event itself (because the competing event has already occurred) are still included, thus, their estimates are not based on probability in the proper sense. Nevertheless, their approach has a clear interpretation in terms of survival probabilities: it allows to calculate the likelihood of each event.

Kuk and Varadhan (2013) developed a technique called information criteria-based stepwise regression procedure used with FG model. This version of their technique is based on the well recognised AIC and BIC type of criteria and also the new criteria that the authors propose to be called the BICcr. This approach addresses the problem driven by the importance of reliable models for predicting risk in financial research and the lack of covariate selection techniques with Fine and Gray (1999) model.

2.7 Summary of the review

Several contributions, as described in sub-section 2.4.2 and sub-section 2.4.3, that involved competing risks analysis found that either Fine and Gray's proportional subdistribution hazard model or cause-specific hazard Cox modelling was adopted. However, the presentation of the analysis of the two competing risks was very limited between the contributions using the Fine and Gray's model. However, earlier in this chapter, there was emphasis made on the cause-specific hazard as a key driving force in the calculation of cumulative incidence. In addition, it was emphasized that, in order to fully understand the impact on the outcomes of concern, the cause-specific hazard and the cumulative incidence for each of the competing risks should be examined.

Chapter 3

Data Description and Summary Statistics

3.1 Mortgage data description

The Federal Home Loan Mortgage Corporation(FHLMC),known as Freddie Mac is a public company that is sponsored by the United States government. It was formed in 1970 to expand the secondary market for mortgages in the US. It has provided a data-set about single family loan-level credit performance on a portion of fully amortizing fixed-rate mortgages that the company purchased or guaranteed. The data-set contains information about approximately 21.5 million fixed-rate mortgages that originated between January 1, 1999, and December 31, 2014. The data-set can be down loaded from the Freddie Mac website and is organised as two files for each quarter:

1. the origination data file that contains data concerning the setup of the loan;

2. the monthly performance data file that contains the monthly performance of each loan e.g. amount repaid, the outstanding principal, whether it is in default, etc.

For the purposes of this thesis, we focus on FHLMC loans for which the time to mortgage default and prepayment were initiated. There is also a smaller sample data set containing a simple random sample of 30734 selected loans per year and a proportionate number of loans from subsequent years (the actual definition is 30734 selected loans from each full vintage year and a proportionate number of loans from each partial vintage year of the full family loan-level data set). The sample data set also contains an annual origination and performance file. The raw data has been processed in order to transform it into a format that can be analyzed by this model. Each loan was tracked through the data in order to categorize it as defaulted, prepaid or active. Because loans in the data set originated in year 1999 and are valid for thirty years, no loans were categorized as mature and could be ignored in this category. For the purposes of this thesis, we focus on the loan for which the foreclosure proceedings have been initiated. For purposes other than default and prepayment, we exclude loans for which the foreclosure process ends; service was moved to a variety of service providers and the values for explanatory variables are missing. Our final sample size is 30734 according to the outlined exclusion criteria, 29243 of which are prepaid, 510 default and 992 still active. In addition, we measure the duration of the loan as the number of months from the start of the foreclosure until the loan is either defaulted or prepaid. We denote as 'right censored' a loan that is still in the foreclosure proceedings at the end of the observation period and measure its lifetime as the number of months from the beginning of the foreclosure to 31 December 2003.

3.1.1 Classification of Loan

We noticed that some loans are discontinued in the data set without clear data, hence, excluded from our analysis. Four fields in the data were used to classify each loan as default, prepaid or active, and to describe the time observed:

- **zerobalance** defines whether or not the balance of a particular loan has been reduced to 0 and has the following codes:
  01 Prepaid or Matured (Voluntary pay off);
  03 Foreclosure Alternative Group (Short Sale, Third Party Sale, Charge Off or Note Sale);
  06 Repurchase prior to Disposition;
  09 REO Deposit;
  empty Not Applicable
- **reportingdate** is the month in which the observation is made
- **monthremain** is the number of months until the loan's legal expiry
- **deliquency** provides a value corresponding to the number of days the borrower has not paid the loan, according to the due date of the last payment of the installment, or if the loan is acquired by REO, code as

follows:

0 Current or less than 30 days in the past due;
1 from 30 to 59 days delinquent;
2 from 60 to 89 days delinquent;
3 from 90 to 119 days delinquent;
R REO acquisition;
empty Unavailable;

Also, we have the loan status defined as;

- **Default** if there is a month in which zero balance = 03, 06 or 09 exists. The default time t_d in this case is the time from the origination of the loan to the reporting date where this first occurs.

- **Prepayment** if **zerobalance** = 01 and **repurchase** = "N" exists for a month, then the prepaid time t_p in this case is the time from the origination of the loan to the reporting date where this first occurs.

- **Active** if the loan can not be classified as Prepayment or Default and the recent loan reporting date is later than 01/01/2014 AND zero balance is void at the recent date AND delinquency is not equivalent to R at the latest date. The active time is the time from the origination of the loan to the reporting date.

3.1.2 Covariates

In the data set, the following covariates (fixed term) are available: number of borrowers, number of units, combined loan-to-value (CLTV), debt-to-income (DTI), Mortgage insurance percentage (MI), credit score, unpaid principal balance (UPB), original interest rate, first home buyer, occupancy status, property type, property state and current interest rate.

The covariate Original interest rate refers to the rate at the start of the loan whereas Current interest rate includes monthly interest rates since the start time of the loan. The latter is the only variable whose values changes in time. The covariates property state has been re-categorised into judicial or non-judicial stated (renamed Foreclosure state), where in a judicial state, the lender needs to go through court system for the foreclosure process. The categorical variables converted into indicator variables are first home buyer, property type, property state and occupancy status. Some of the categorical variables were of low frequency. For instance, there are 6 categories in variable property type, of which approximately 81 percent were single-family home and some categories leasehold as low as 0.0003 percent. It was decided to group categories with extremely low frequencies for all the categorical variables. Furthermore strong correlation have been found between the covariates MI and CLTV, and between original and current interest rates which led us to drop the latter in both cases. All quantitative variables was standardized and since the covariate Current interest rate has been dropped, we do not have to operate with a covariate based on time.

3.2 Analysis of the Data

After developing our modeling structure, we are now turning to the implementation of our proposed models to the Freddie Mac Single Family mortgage loan data set. This study investigated the effect of competing risks of default and prepayments on mortgage market. The data includes information on times (in years), events (0 = active, 1 = Prepaid and 2 = Default) and nine covariates (non categorical) used as predictors are credit score, unpaid principal balance (UPB), combined loan-to-value (CLTV), debt-to-income (DTI),Mortgage insurance percentage (MI), number of units, original interest rate, current interest rate and loan age.

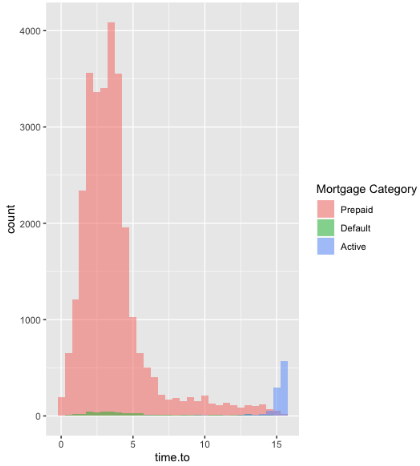
In the analysis, we consider modelling the duration of single-home mortgages. In doing so, I model default and prepayment Probabilities simultaneously using competing risks proportional hazards models. The time to default of the mortgage is the endpoint of interest, this variable is measured in months.

3.2.1 Explanatory data analysis

The categorical predictors used hare include the First Time Home-buyer Indicator (Yes, No, or Not applicable), Occupancy Owner and Property Type. Active are right censored for both risks while Default and Prepaid are the competing events. The data set consisted of 30734 mortgages originating in the year 1999, 29234 (95.1 percent) were Prepaid, 510 (1.66 percent) are Defaults and 992 (3.23percent) are Active.

The data set is widely unbalanced with over 95 percent of the mortgages being Prepaid and the smallest percentage of the mortgage to the Default category. This enormous imbalance is obvious in the histogram of time to default of Default, Prepaid and Active times for each category shown in figure 3.1. The distribution shows a high spark around 4 and drastically

Figure 3.1: Histogram of time to Default, Prepaid and Active times for each category



decreasing afterward. As we have already mentioned, covariates implementation that is default or prepayment predicting factors, has an important effect on the realization of the mortgage duration hazard rate. We have chosen nine default variables to build our model, namely which are stated in Section 2.4 above.

Chapter 4

Competing risk analysis with the Mortgage Study

The previous chapter described the background to the competing risk approach outlined the methodological issues that need to be addressed, and reviewed the literature on the extent of its use. In particular, the review focused on mortgage default and prepayment time.

The chapter discusses empirical competing risk analyses with a mortgage case study, presents different aspects of competing risk analysis to demonstrate the insights that can be gained from using standard survival analysis. The chapter begins by presenting an overview of the Single Family Loans case study. The chapter focus on proportional cause-specific hazards Cox regression for each of the two competing risks default and prepayment. Furthermore, we presents the Fine and Gray (FG) approach, also a Cox model, but for the subdistribution hazard attached to the cummulative incidence, to competing risks which provides a clear understanding of the competing risk process by considering both the hazard and the cumulative incidence of each competing risk. In the sequel, we reserve the term Cox model for Cox regression on the cause-specific hazard (CSH), and we use Fine and Gray (FG) regression for proportional hazards modeling of the subdistribution hazard (SH). The cumulative hazard and cumulative incidence plots are also included. Finally, a review of the analysis provided in this chapter and the additional insight gained on standard survival analysis.

4.1 Overview of Single-Family home loans study

This chapter has been reviewed in detail with the scale of data in the Freddie Mac data set, a case study used to demonstrate the potential in hypothetical modelling duration of single-home mortgages of the methods described in this thesis. In brief, the study included 30736 subjects with mortgage loans for default (n = 510), prepaid (n = 29234) or active (n = 992), and a median follow-up duration of 3.33 months. Active loans are right-censored observations of both default and prepayment times. The dataset contains a cohort of 30736 mortgages. By the end of follow-up, there are 510 survivors and 30226 non-survivors. The time is the month after a loan cures or terminates, and status is vital status (0= censored, 1= default and 2= prepaid), where time to mortgage default is an event of interest and prepayment is a competing event. The competing hazard model implies that prepay times are also right-censored observations of default time, and vice versa. We model default and prepayments likelihoods simultaneously using cause-specific Cox regression approach and FG subdistribution hazard approach, thus, allowing comparison with the works of Bhattacharya et al. (2019). This is our main interest in this chapter.

4.2 Cause-specific hazard:Cox proportional (PH) Regression model

The Cox proportional (PH) hazard approach was used to estimate the relationship between the hazard (default and prepaid) rates and covariates effects without having to make assumptions about the shape of the baseline hazard function. The Cox PH model is widely applicable and can handle censored and/or truncated observations. In this section, we explore regression analysis which was used to identify the risk factors. To determine the covariates to be included in the final model, the univariate Cox proportional hazard regression analysis is applied first to identify the impact of individual variable on time to event before proceeding with more complicated model selection. Covariates are identified as significant at 5% level of significance in the univariate analysis.

Table 4.1 presents the univariate Cox PH analysis, showing the effects of the covariates on the cause-specific hazard of each of the events at the univariable level. Each covariates are assessed through separate univariate Cox regressions. According to the analysis, nearly all the covariates, with the exception of no of units ($p = 0.15$ for default and $p = 0.36$ for prepaid), are statistically significant and selected as significant factors for risk of default and prepaid from loan origination to the month the observation is made.

Furthermore, we then carried out multivariate Cox PH analysis (by using stepwise selection process) including all the potential risk factors that had a p-value of less than or equal 0.05 in the univariate Cox PH analysis. In order to select the subgroup of covariates in our model, the approach of stepwise discussed in section 2.6 was applied. Table 4.2 displayed the results of the stepwise proportional hazard regression. The first section in Table 4.2 adds all the covariates into the model, and to further optimise our Cox model, the covariates with the highest p-value and over threshold of significance are removed from the model one by one. No. of brorrowers, property types, credit scores, CLTV and original interest rate have significant impact on the prediction of default rate. Furthermore, we see in Table 4.3 nealry all the covariates, with exception of mortgage insurance % and original DTI, turn out to be significant.

The final model for each events was obtained as in Table 4.2 and Table 4.3, showing that most of the covariates are significant with their p-values less than 5%. The standard appropriate checks on violations of proportional hazards were carried out, although they are not shown for reasons of brevity. However, none of the interactions among the covariates are significant since an investigation of Schoenfeld residuals revealed no evidence against the assumption of proportionality for both types of event. Hence, the final Cox regression model for the failure type (default) was fitted to estimate the effect of the five covariates no of borrowers, property type, credit score, CLTV and original interest rate on the cause-specific hazards using the R function coxph from the library survival. Also, the final Cox PH model was fitted for failure type (prepaid) which estimates the effects of the eight covariates no of borrowers, property type, credit score, CLTV, UPB, original interest rate, occupancy status and foreclosure state on cause-specific hazards.

To estimate the cumulative incidence function for all the covariates, the influence of the covariates on all types of failures have been assessed using the approach as previously explained in section 2.4.3. For each type of failure mortgagers experiencing the competing event were considered as censored observations. No of borrowers, property type, credit score, CLTV and original interest rates had a significant effect on the cause-specific hazards for both default and prepaid, in addition to significant effect on the cause-specfic hazards for prepaid are UPB, occupancy status and foreclosure states (results are shown in Table 4.4 and Table 4.5). Combined loan-to-value (CLTV) had a greater influence on the cause-specific hazard for default with a cause-specific hazard ratio (HR) of 1.94(95% C.I 1.701 to 2.218) compared to 1.04(95% C.I 1.024 to 1.049) for prepaid events. The analysis for prepaid revealed an increased risk for mortgagors from the original interest rate with cause-specfic hazard ratio (HR) of 1.45(95% C.I 1.435 to 1.469) which is similar to the cause-specific hazard ratio of 1.49(95% C.I 1.372 to 1.608) for default risk. No of borrowers group with cause-specific hazard ratio of 0.55(95% C.I 0.462 to 0.658) had an increased risk on cause-specific hazard for default but the effect was much lower compared with cause-specific hazard for prepaid with cause-specific hazard ratio of 1.12(95% C.I 1.091 to 1.146). In terms of credit scores, the results indicate that loans with higher credit scores are less likely to default and more likely to prepaid.

Table 4.1: Univariable cause-specific hazard competing risks Cox modelling results.

Covariates	COMPETING RISK REGRESSION Univariable				
	cause-specific hazard (default)		cause-specific hazard (prepaid)		
	hazard ratio for default HR (95% C.I)	P-value	hazard ratio for prepaid HR (95% C.I)	P-value	Baseline
Mortgage Insurance %	$0.702(0.27-0.4)$		$0.198(0.86-0.9)$		<0
	$0.702(1.4-3)$	$2.5e-35$	$0.198(1.1-1.3)$	$2.1e-33$	$0-1$
No of borrowers	$0.548(0.46-0.65)$	$2.3e-11$	$1.17(1.1-1.2)$	$8.8e-38$	1
Number of Units	$0.87(0.56-9)$	0.15	$-0.0626(0.7-1.3)$	0.36	
Property Type	$1.55(1.2-2)$	0.0014	$0.899(0.87-0.93)$	$6.7e-13$	*Yes*
Credit score	$0.468(0.43-0.51)$	0.000	$1.04(1-1.1)$	$3.1e-11$	
Original DTI	$1.24(1.1-1.4)$	$1.2e-06$	$1.06(1-1.1)$	$1.5e-20$	
CLTV	$2.41(2.1-2.8)$	$1.6e-37$	$1.08(1.1-1.1)$	$6.9e-37$	
UPB	$0.914(0.83-1)$	0.068	$1.26(1.2-1.3)$	0	
Original interest rate	$1.76(1.6-1.9)$	$4.4e-48$	$1.32(1.3-1.3)$	0	
Occupancy status	$1.62(0.99-2.7)$	0.057	$1.06(1-1.1)$	0.026	*Yes*
Foreclosure state	$1.02(0.86-1.2)$	0.81	$0.937(0.92-0.96)$	0.000	*Yes*

The final multivariate Cox proportional hazard (PH) models, for both default and prepaid, are then given by (see section 2.4.3) :

$$\begin{aligned}\text{Mod 1 (Default)} : \lambda_k(t) = \lambda_0 \exp(&\text{- 0.59558 No of borrowers}\\ &+ \text{0.40042 Property Type - 0.63589 Credit Score}\\ &+ \text{0.39566 Original interest rate}\\ &+ \text{0.71321 CLTV})\end{aligned} \tag{1}$$

$$\begin{aligned}\text{Mod 2 (Prepaid)} : \lambda_k(t) = \lambda_0 \exp(&\text{1.118 No of borrowers}\\ &+ \text{0.926 Property Type + 1.128 Credit Score}\\ &+ \text{1.337 UPB + 1.452 Original interest rate}\\ &+ \text{1.223 Occupancy status + 1.036 CLTV}\\ &+ \text{0.958 Foreclosure state})\end{aligned} \tag{2}$$

4.3 Cause-specific hazard model vs Fine-Gray subdistribution model

As stated in the previous section, an understanding of a competing risk scenario can be gained by investigating the hazards and cumulative incidence functions for each of the competing risks. In the works of Latouche et al. (2013), the authors argued that a complete understanding of the event

Table 4.2: Multivariate Cox PH regression analysis for default mortgages.

Covariates	The stepwise proportional hazard regression (Multivariable)					
	Cox regression (Event : default)		Elimination of variable with high p- value by Stepwise		The final Cox PH model	
	hazard ratio for default HR (95% C.I)	P-value	hazard ratio for default HR (95% C.I)	P-value	hazard ratio for default HR (95% C.I)	P-value
Mortgage Insurance %	$1.15(0.824-1.607)$	0.409				
	$1.34(0.895-2.052)$	0.151				
No of borrowers (No=1)	$0.57(0.468-0.685)$	$5.15e-09$	$0.55(0.462-0.658)$	$3.93e-11$	$0.55(0.462-0.658)$	$3.93e-11$
Property Type (Yes)	$1.51(1.133-2.013)$	0.004	$1.55(0.46-0.658)$	0.004	$1.55(0.46-0.658)$	0.004
Credit score	$0.52(0.478-0.572)$	0.000	$0.53(0.487-0.576)$	0.0000	$0.53(0.487-0.576)$	0.0000
Original DTI	$1.09(0.989-1.211)$	0.078				
CLTV	$2.07(1.614-2.647)$	$8.74e-09$	$1.94(1.701-2.218)$	0.0000	$1.94(1.701-2.218)$	0.0000
UPB	$0.99(0.893-1.118)$	0.968				
Original interest rate	$1.50(1.378-1.633)$	0.000	$1.49(1.372-1.608)$	0.000	$1.49(1.372-1.608)$	0.000
Occupancy status (Yes)	$1.09(0.636-1.885)$	0.745				
Foreclosure state (Yes)	$0.90(0.747-1.084)$	0.268				

Table 4.3: Multivariate Cox PH regression analysis for prepaid mortgages.

Covariates	The stepwise proportional hazard regression (Multivariable)					
	Cox regression (Event : prepaid)		Elimination of variable with high p- value by Stepwise		The final Cox PH model	
	hazard ratio for prepaid HR (95% C.I)	P-value	hazard ratio for prepaid HR (95% C.I)	P-value	hazard ratio for prepaid HR (95% C.I)	P-value
Mortgage Insurance %	$0.97(0.933-1.001)$	0.057				
	$0.93(0.864-1.011)$	0.093				
No of borrowers (No=1)	$1.12(1.093-1.151)$	0.000	$1.12(1.091-1.146)$	0.0000	$1.12(1.091-1.146)$	0.0000
Property Type (Yes)	$0.93(0.901-0.958)$	$2.72e-06$	$0.92(0.899-0.954)$	$2.77e-07$	$0.92(0.899-0.954)$	$2.77e-07$
Credit score	$1.12(1.114-1.143)$	0.000	$1.12(1.115-1.142)$	0.0000	$1.12(1.115-1.142)$	0.0000
Original DTI	$1.01(0.999-1.025)$					
CLTV	$1.03(1.009-1.044)$	0.003	$1.04(1.024-1.049)$	$4.86e-09$	$1.04(1.024-1.049)$	$4.86e-09$
UPB	$1.33(1.316-1.350)$	0.000	$1.34(1.321-1.353)$	0.000	$1.34(1.321-1.353)$	0.000
Original interest rate	$0.69(1.419-1.455)$	0.000	$1.45(1.435-1.469)$	0.000	$1.45(1.435-1.469)$	0.000
Occupancy status (Yes)	$1.20(1.132-1.269)$	$6.36e-10$	$1.22(1.157-1.292)$	$1.17e-12$	$1.22(1.157-1.292)$	$1.17e-12$
Foreclosure state (Yes)	$0.96(0.935-0.982)$	0.0007	$0.96(0.937-0.982)$	0.0005	$0.96(0.937-0.982)$	0.001

Table 4.4: Analyses of cause-specific and subdistribution hazards models for characterizing levels of default mortgages.

Covariates	COMPETING RISK REGRESSION Multivariable				
	cause-specific hazard (default)		sub-distribution hazard (default)		
	hazard ratio for default		hazard ratio for default		
	HR (95% C.I)	P-value	HR (95% C.I)	P-value	Baseline
No of borrowers	$0.55(0.462-0.658)$	$3.93e-11$	$0.42(0.354-0.504)$	$0.0e+00$	1
Property Type	$1.55(0.46-0.658)$	0.004	$1.84(1.403-2.414)$	$1.1e-05$	*Yes*
Credit score	$0.53(0.487-0.576)$	0.000	$0.47(0.436-0.509)$	$0.0e+00$	
CLTV	$1.94(1.701-2.218)$	0.000	$1.74(1.526-1.984)$	$2.2e-16$	
Original interest rate	$1.49(1.372-1.608)$	0.000	$1.15(1.056-1.260)$	$1.6e-03$	

dynamics requires that both hazards and cumulative incidence be analyzed side-by-side, and that this is generally the most rigorous scientific approach to analyzing competing risks data. Thus, suggesting that the effects of covariates should be assessed in the framework of both cumulative hazard plots and cumulative incidence plots. The objective of this section is to demonstrate the additional insight that such a comprehensive analysis can provide to help interpret in the presence of competing risks.

Although modern statistical packages allow easy analysis of cause-specific hazard models and sub-distribution hazard models on a per-form basis, it is worth considering some analytical strategies on the basis of two main objectives which are; association and prediction.

In the regression analysis of competing risk data, the effects of covariates on the cause-specific hazard function or cumulative risk function can be investigated using either the cause-specific hazard model or the Fine-Gray (subdistribution hazard) model (Zhang et al., 2018).

A proportional hazards regression model for both the cause-specific hazards (CSH) and the sub-distribution (SH) is employed for the univariate and multivariate analyses. In the case-specific hazard (CSH) model, the hazard ratio represents the instantaneous relative risk of an event of interest (default or prepayment, as the case may be) in the presence of the covariate, for instance, the hazard ratio (HR) corresponding to the conditions described by the two various levels of the explanatory variable, all other covariates being the same. However, this hazard ratio (HR) cannot be directly translated to the cumulative incidence function.

As indicated in previous chapter 2, the cummulative incidence of the event of interest has a one-to-one correspondence with the estimated Fine and Gray (FG) subdistribution. Thus, the coefficients quantify the direct effects of covariates on the cummulative incidence. Nevertheless, the approximate hazard ratio of sub-distribution from the Fine-Gray model has no clear interpretation as the survival times of mortgagers who have not encountered the event of interest are transformed into censored times that are theoretically extended to infinity. In addition, the probabilistic relationship between the cumulative incidences of various types of events and the marginal survival function is lost (Zhang et al., 2018).

Thus, for the above reason, this section explicitly reports simultaneously these two models in the analysis of competing risks. Firstly, we fitted the subdistribution hazard regression (Fine and Gray) models using an interface R function FGR(), where the event of interest (event 1 = default or event 2 = prepaid) is specified. All covariates of interest, found statistically significant in Table 4.2 and Table 4.3, are included for both cases.

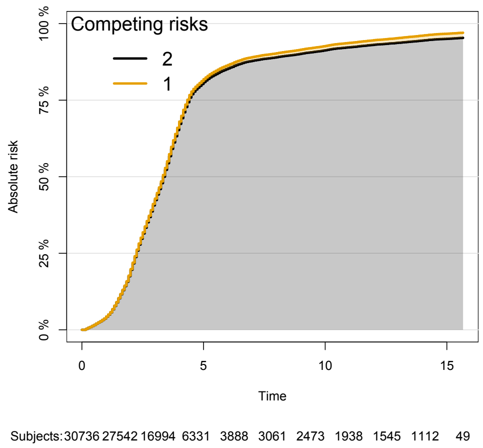
It can be seen in Table 4.2 that Number of borrowers who are obligated to repay the mortgage note secured by the mortgaged property, property type, credit score of borrower's creditworthiness, original combined loan-to-value ratio (CLTV) and original interest rate were each found to be associated with a higher hazard of default. It can also be seen from Table 4.3 that there was a higher hazard of the composite outcome associated with number of borrowers who are obligated to repay the mortgage note secured by the mortgaged property, property type, credit score of borrower's creditworthiness, original combined loan-to-value ratio (CLTV), original interest rate, the current actual unpaid principal balance (UPB),occupancy status and foreclosure state. Also, a full model is fitted using the Fine and Gray sub-distribution model including all the eight covariates, using an interface R function FGR().

Results of the four regression models investigating the influence of the covariates on both the cause-specific hazard and the subdistribution hazard are shown in Table 4.4 and Table 4.5 for default and prepaid rate respectively. Furthermore, the effects on the subdistribution hazard can be translated directly to effects on the cumulative incidence function. For default rate a subdistribution hazard ratio (HR) comparing the combined loan-to-value (CLTV) group of 1.74(95% C.I 1.526 to 1.984) was observed, indicating a much higher incidence of default for mortgagors. The effect of CLTV was slightly weaker for prepaid with subdistribution hazard ratio of 1.02(95% C.I 1.010 to 1.035). Effects on no. of borrowers, property type, credit score and original interest rate are higher on subdistribution hazard of prepaid.

Table 4.5: Analyses of cause-specific and subdistribution hazards models for characterizing levels of prepaid mortgages.

Covariates	COMPETING RISK REGRESSION Multivariable				
	cause-specific hazard (prepaid)		sub-distribution hazard (prepaid)		
	hazard ratio for prepaid HR (95% C.I)	P-value	hazard ratio for prepaid HR (95% C.I)	P-value	Baseline
No of borrowers	$1.12(1.091-1.146)$	0.000	$1.14(1.114-1.175)$	$0.0e+00$	1
Property Type	$0.93(0.899-0.954)$	$2.77e-07$	$0.92(0.890-0.947)$	$8.4e-08$	*Yes*
Credit score	$1.13(1.115-1.142)$	0.000	$1.15(1.136-1.167)$	$0.0e+00$	
CLTV	$1.04(1.024-1.049)$	$4.86e-09$	$1.02(1.010-1.035)$	$3.6e-04$	
UPB	$1.34(1.321-1.353)$	0.000	$1.31(1.289-1.327)$	$0.0e+00$	
Original interest rate	$1.45(1.435-1.469)$	0.000	$1.40(1.374-1.422)$	$0.0e+00$	
Occupancy status	$1.22(1.157-1.292)$	$1.17e-12$	$1.21(1.138-1.283)$	$7.6e-10$	*Yes*
Foreclosure state	$0.96(0.937-0.982)$	0.001	$1.03(0.944-0.992)$	$1.1e-02$	*Yes*

Figure 4.1: Non-parametric estimates of the cumulative incidence functions for mortgage and other causes. This gives the probabilities of getting cause 1 and cause 2 in the competing risks model, and represent the relevant probabilities in the model. This what we see in the mortgage loans when 30736 mortgagors start at time 0, then after 15.666 months we can give the percentage that have associated with cause 1 (default) or cause 2 (prepaid) or are still active (censored =0) under risk.



Overall, for each event, the two approaches showed similar results in terms of the covariate effects being in the same direction. The rest of this section will discuss the effect of covariates on the hazard of each event and how this relates into the corresponding effect on the subdistribution / accumulative incidence of each event.

In figure 4.1, the plot revealed that 98% of the mortgagors experienced cause 1 (default), and around 92% experienced cause 2 (prepaid) over the the length of the study. The probabilities were from the underlying cause specific hazards. The plots in figures 4.2 and 4.3 shows that predictions are somewhat similar to the non-parametric estimates for both causes default and prepaid. There are variations observed, but it is difficult to evaluate how

Figure 4.2: Cumulative incidence functions for the 2 levels of No of borrowers based on cause-specific hazard for both default (left) and prepaid (right).

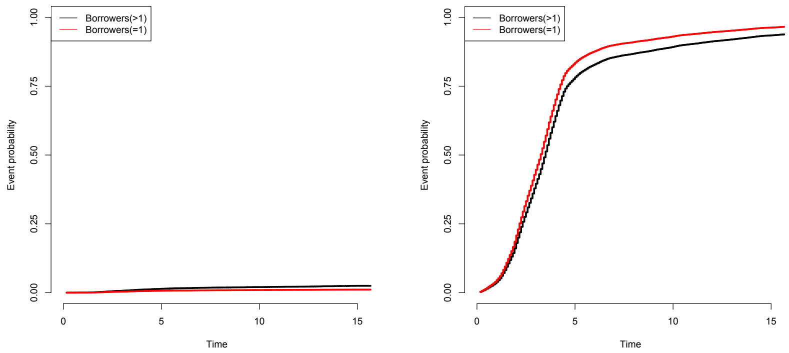


Figure 4.3: Cumulative incidence functions for the 2 levels of property type based on cause-specific hazard for both default (left) and prepaid (right).

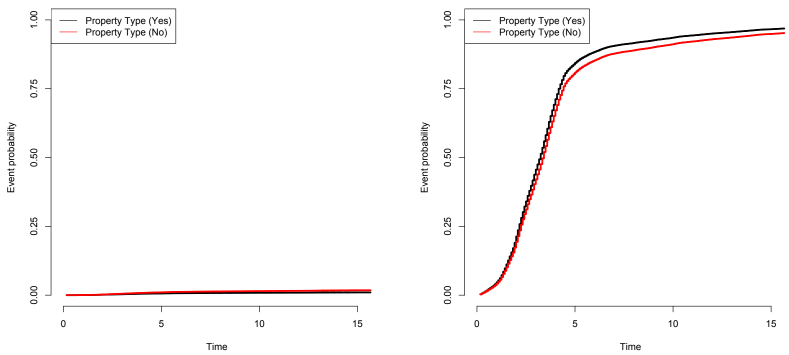


Figure 4.4: Cumulative incidence functions for the 2 levels of No of borrowers based on Fine and Gray's subdistribution hazard for both default (left) and prepaid (right).

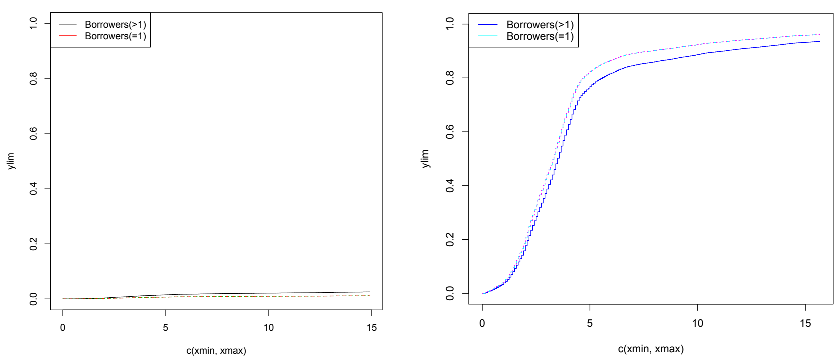
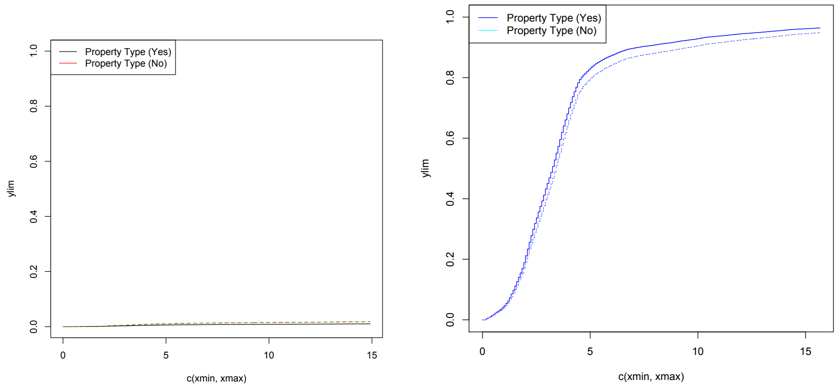


Figure 4.5: Cumulative incidence functions for the 2 levels of property type based on Fine and Gray's subdistribution hazard for both default (left) and prepaid (right).



important these aspects are. The plots in figure 4.2 also revealed that number of borrowers (2 or more) who are obligated to repay the mortgage note secured by the mortgaged property leads to more default events, and vice versa for prepaid events. Similarly, figure 4.3 showed that the property type (Yes) condominium, leasehold, planned unit development (PUD), cooperative share, manufactured home, or Single Family home leads to more prepaid events, and vise versa. This reflects the competing risk aspect of the mortgage dataset.

The cumulative hazard plots are based on the Cox cause-specific modeling, and the cumulative incidence curves are based on the Fine and Gray subdistribution hazard model. The latter were calculated using the default weighted equation technique in cmprsk from R. Figures 4.5 and 4.4 displayed the estimate cumulative incidence for property type and number of borrowers via the Fine and Gray model, with obvious view that predictions are somewhat similar to the non-parametric estimates for both causes default and prepaid. It provides useful predictions and tests for specific levels of the covariates, but the regression coefficient is hard to interpret explicitly at the sub-distribution hazard level. A vital point in relation to the results presented in Tables 4.4 and 4.5 are that the effects of covariates are shown as hazard ratios (subdistribution) and as such are relative effects.

Chapter 5

Model Evaluation

5.1 Model checking

We have fitted models which is ideally the first part of regression analysis as it is based on certain assumptions. In this section, we adopt Regression Diagnostic approaches, which are used to access the adequacy of a fitted model after a model has been constructed and to investigate whether or not observations have a significant undue influence on the analysis.

5.1.1 The PH Assumption Checking

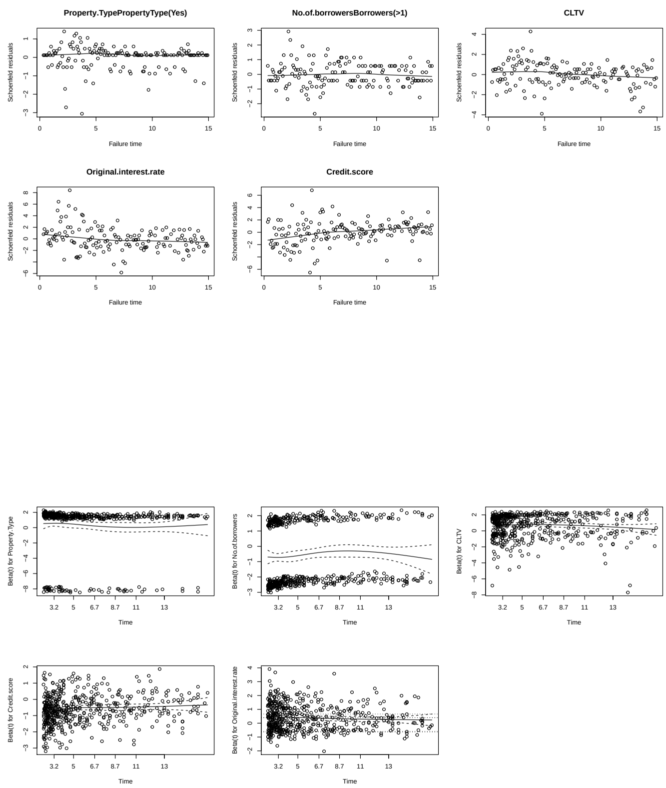
There are several methods for verifying that a model satisfies the assumption of proportionality (Graphical method, Scaled Schoenfeld residuals, Adding time dependent covariate) (Collett, 1994). The final models (see section 4.2) are based on a major assumption that the hazards between groups are proportional. To test the assumption of proportionality, the scaled Schoenfeld residuals and log-log survival plot have been used.

Testing for PH assumptions using Scaled Schoenfeld residuals

From our understanding, the main assumption of the Cox proportional hazards model is proportional hazards, which mean that the hazard ratio (HR) is constant over time. Scaled Schoenfeld residuals is based on the principle that, for a given regressor, the assumption restricts $\beta(t_k) = \beta$ for all t_k, thus, implies that a plot of $\beta(t_k)$ against time will have a slope of zero (Medhat, 2015).

We plot the Schoenfeld residuals against failure time for each covariate

Figure 5.1: Schoenfeld residuals against failure time for each covariate for Subdistribution hazard model (top) and cause-specific hazard model (bottom) for default. Dotted horizontal lines indicate the estimated coefficient values of these covariates.



for subdistribution hazard model and cause-specific hazard (see Figure 5.1). If the proportional assumption is true, the residual should have a constant mean across time. A scatterplot smoother is added for each covariate to check for the assumption. In Figure 5.1, it appears that the covariates, credit scores and original interest rate, for cause-specific hazard model has non-constant residuals across time, indicating a potential violation to the proportional assumption. To further formally check the assumption for these covariates, a global test is considered (see Table A1.2). The P value in Table A1.2 indicates the significance probability of the Schoenfeld residual test for the credit score and original interest rate used, and such values indicate a violation of the proportional hazard assumption, which means that there is an interaction between these covariates and time.

The covariates violating the proportional hazard assumption in the cause-specific hazard (CSH) model (see eqn 1) should be adequately adjusted. To fit the estimated cause-specific hazard CPH model with

Table 5.1: Stratified Cox regression results for the No-interaction and Interaction Models.

Covariates	No interaction model		Interraction model			
	Event : default		Strata 1 : Credit score		Strata 2 : Original interest rate	
	hazard ratio HR (Coef)	P-value	hazard ratio HR (Coef)	P-value	hazard ratio HR (Coef)	P-value
Property Type (Yes)	1.58(0.443)	0.00136	0.00(−14.99)	0.9939	1.36(0.310)	0.0139
No of borrowers (>1)	0.60(−0.507)	0.000	0.00(−15.639)	0.9923	0.97(−0.021)	0.7983
CLTV	2.06(0.724)	0.000	0.98(−0.019)	0.9733	0.89(−0.114)	0.0586

Log likelihood for no interaction model = -2789.1
Log likelihood for interaction model = -2781.9

covariates violating the proportional hazard assumption, one method we applied is a stratified CPH model. This approach results in one combined result from the outcomes of each stratum containing a categorical variable classified on the basis of a certain criteria. In contrast to the Mantel – Haenszel method, which is based on the sample size of each stratum, the stratification in the CPH model sets a different baseline hazard corresponding to each stratum, and a statistical estimate is then used to achieve prevalent coefficients for the remaining covariates, with the exception of stratified covariates (Collett, 1994). This covariate credit score and original interest rates are made using a time-independent variable and time, or a function of time. The method contrasts two models, namely the CPH model, which assumes that the proportional hazard assumption has not been violated, and the CPH model, which uses a combination of the explanatory variable and time (or time function) in the estimated CPH model, however, for comparison purposes, a likelihood ratio test or Wald statistics are used. This approach has some advantages, including a simultaneous comparison with multiple covariates and various time functions, although the results may change depending on the covariates and types of functions selected (Collett, 1994). Thus, providing a hazard ratio of the controlled effects of covariates that violate the proportional hazard assumption(In and Lee, 2019).

Table 5.1 displays the application of stratified Cox regression with no interaction and interaction model, to determine which model was more statistically suitable for the likelihood ratio test, which compares log-likelihood statistics for the interaction model and the no-interaction model. It is obvious that there are different baseline hazard functions for each stratum in the no-interaction model, but the coefficients are the same. However different baseline hazard functions and different coefficients are obtained for each stratum in the interaction model. A Likelihood ratio statistics value of 14.2 was calculated (see Medhat (2015)), which is not significant at the 0.05 level of significance. Thus, in-spite the numerical difference between the coefficients of the covariates in the Interaction models, there is no statistical difference. Hence, for our mortgage data the stratified Cox model with no-interaction model is acceptable at 0.05 level of significance and the covariates property type, number of borrowers, and CLTV are statistically significant and as such selected as significant factors for the risk of default following the origination of the loan at 0.05 level of significance.

5.2 Model validation

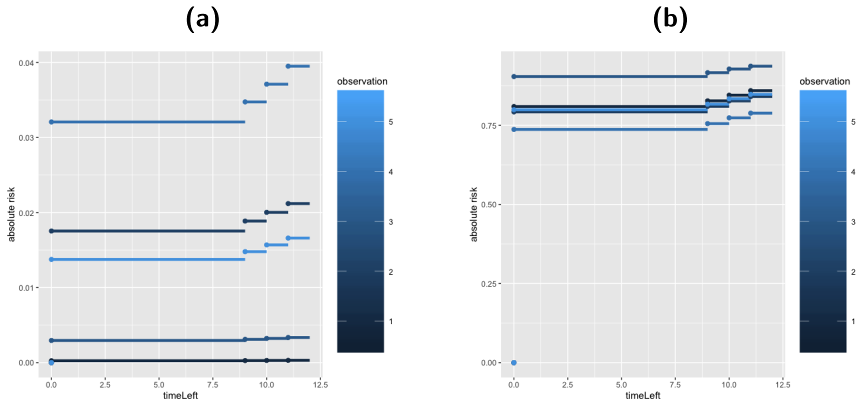
In this chapter, we illustrate methods for validating competing risks regression models. For time-to-event data, model validation is performed at each analysis time point. The prediction accuracy of competing risks models can be assessed by calibration or discrimination. We used the Area under the ROC Curve (AUC) and calibration plots as measures of discrimination and calibration, for our final models (illustration on time to default as our primary focus) respectively. We performed a one-time splitting method that randomly split the original mortgage data into training (80%) and testing (20%) datasets. Although, this method reduces the sample size of both training and test datasets which may produce a different results by different splitting processes. Thus, the cross-validation method is more appealing (Zhang et al., 2018).

5.2.1 Model Prediction

In this section, we used our fitted models to obtain predictions about future cumulative incidence functions (CIF) for mortgage default and prepaid rates respectively. For the Fine and Gray model, predicting CIFs for the event you are modeling is a straightforward task, because the subdistribution hazard is being modeled directly and the CIF is only one transformation away. In contrast, predicting CIFs on the basis of the cause-specific Cox models involves more tedious calculations (Guo, 2018).

The R code used to obtain predictions of the cumulative incidence functions (CIF) is almost identical between the cause-specific survival analysis and the Fine and Gray analysis. As a requirement, it is essential to designate a cause of interest for the cause-specific survival analysis enable prediction of cumulative incidence for different causes of interest simultaneously (see Appendix A1.1). Figure 5.2 shows the cumulative incidence estimates (i.e., absolute risk) for both default and prepaid for each of the five individuals. The plot was produced using the predict() function in R, which returns covariates that have been used for prediction and report the absolute risk. It

Figure 5.2: Cumulative Incidence function for mortgagors #1 to #5 over months 9 to 12 after loan origination for both default (a) and prepaid (b) using the cause-specific hazard regression



appears that loans #4 has the highest mortgage default rate over the entire follow-up period.

Meanwhile, the absolute risk of each individaual loan can be obtained by using the predictRisk() function in R but does not return covariates that have been used for prediction, for instance, for loan #60, the cumulative incidence estimates by 14 months for default from mortgage loan are 0.0299 under the cause-specific proportional hazards model and 0.0268 under the Fine and Gray subdistributio model (see Appendix A1.1). The difference between estimates of the cumulative incidence function from the cause-specific hazard models and Fine-Gray model is due to different proportionality assumptions — proportional subdistribution hazards for the Fine-Gray model and proportional cause-specific hazards for both Cox regression models (Guo, 2018).

We further predict new observations with given combinations of covariates using our fitted Fine-Gray model. For instance, a new dataset containing three mortgagors was created with covariates of property type, CLTV and no of borrowers are defined for them. The data frame was then transformed to matrix and factor variables be transformed to dummy variables. The generic function plot() in R was applied to draw CIF for each observation (see figure 5.3).

5.2.2 Calibration with cross validation method

This section discusses the importance of calibration with cross validation method for improving prediction accuracy. We used the calibration curve was used to compare the predicted probability with observed probability at a certain time using the plotCalibration() function in R. Figure 5.5 shows the calibration curve for our final model. The validation was performed on the

Figure 5.3: Parametric estimates of cumulative incidence functions (CIF) for three mortgagors with given covariate values.

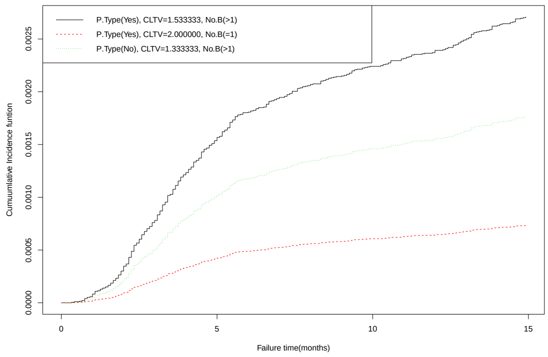


Figure 5.4: The AUC (left) for both models. Note that the cause-specific hazard model has different values to that of the Fine-Gray model. AUC, area under operating characteristic curve. Brier scores (right) for all models, with confidence interval.

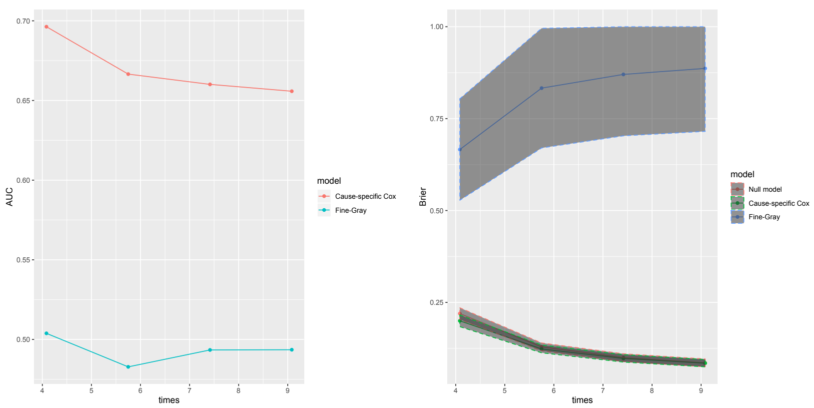
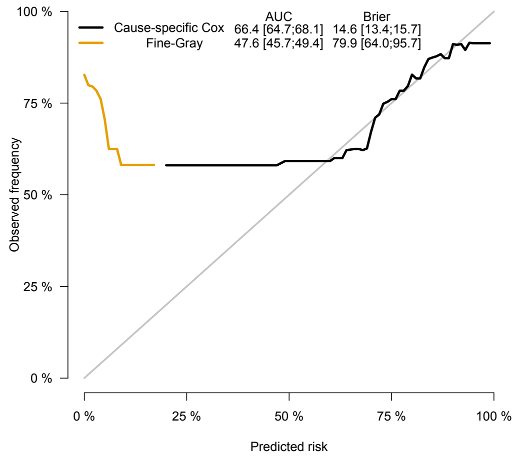


Figure 5.5: Calibration curves for the cause-specific proportional hazards model and Fine-Gray model. The validation is performed on the test dataset. AUC and Brier score are expressed as the point estimates and 95% confidence intervals.



test dataset. The cause-specific Cox model seemed to be ideal as the pairs of observed and predicted probabilities lie on the 45-degree angle line, implying that the two probabilities match each other well. The closer of a calibration curve of a model to the diagonal, the better of the model (Zhang et al., 2018). Furthermore, the area under operating characteristic curve (AUC) and Brier score are displayed at the top of Figure 5.5. The Brier score measures discrimination and calibration at the same time while the area under operating characteristic curve (AUC) assesses the discrimination of the model, with AUC > 0.8 indicating that the discriminatory accuracy of that model is good (Schoop et al., 2011). The cause-specific Cox model with the smallest Brier score and largest AUC indicates a better model than the Fine-Gray model.

A widely used cross-validation method is k-fold cross-validation, that has also been applied to competing risk regression models (Giuliana et al., 2013; Zhang et al., 2018). This approach uses all the data for both model training and validation, thus, taking into account k equal sampled sized subsamples partitioned randomly from the original sample. Of the k subsamples, the validation data for testing a model is retained as a single subsample, and the remaining $k - 1$ subsamples are used as training data. The process of cross-validation is then repeated k-times. The validation data for every k subsample is used exactly once. The results of the k can then be averaged in order to produce a single estimate. The ideal of using the cross validation was to check our model for improving prediction accuracy.

Figure 5.6: Calibration curves for the cause-specific proportional hazards model obtained by using cross validation. The validation is performed on the whole mortgage dataset. AUC and Brier score are expressed as the point estimates and 95% confidence intervals.

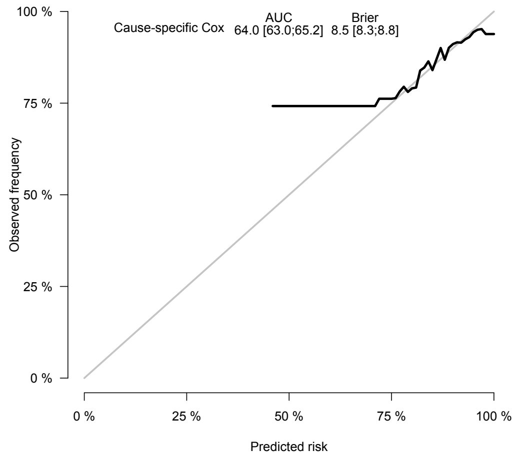


Figure 5.7: nomogram for predicting cumulative risk at 7 and 12 months with cause-specific hazard model. The mortgagor No 35 is illustrated in the nomogram by mapping its values to the covariate scales. The probabilities of mortgage-caused default by months 7 and 12 are estimated to be 0.013 and 0.0258, respectively.

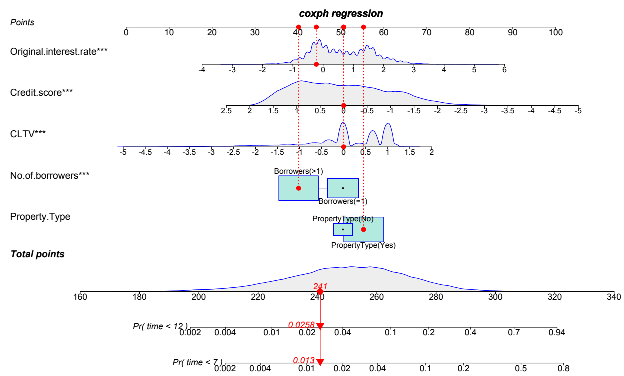


Figure 5.6 shows the calibration plots for cause-specific model at a time (in months) 11, using the whole dataset to fit the model. Furthermore, we accessed the dataset using area under the receiver operating characteristic curve (AU-ROC) plot. A graphical display (see Figure 5.4) shows the prediction accuracy of the model over entire follow up time period. The discrimination measures how well the model can distinguish between mortgagors who default and those who prepaid.

Chapter 6

Main insights, limitations and potential areas of future research

6.1 Introduction

The overall aim of this thesis was to explore the potential of competing risks analysis in a financial context. The thesis described the background to the methods, which included emphasising the importance of recognising a competing risks scenario when it exists and the extra insight that can be gained from using the methods over standard survival analysis. The research was centred around the mortgage case studies and the specific outcomes default and prepaid. In this thesis, we studied the dynamics of mortgage default and prepayment behavior over the Single family loan from The Federal Home Loan Mortgage Corporation (FHLMC) sponsored by the US government. A data description and summary statistics of the dataset was presented. We found out that the dataset was enormously unbalanced with over 95% of the mortgages being prepaid.

The background chapter included reviews of the extent of use of competing risks (section 2.2). The chapter outlined the approaches, including the literature reviews on their implementation common to the case study, illustrated misunderstandings and misunderstandings relevant to some of the methodological issues. The aim of this was to ease the confusion that might arise from some of the mixed signals in the literature. Section 6.2 of this final chapter summarises the main insights from the empirical chapters of this thesis, demonstrating where appropriate some of the issues and considerations emphasised in the background chapter.

6.2 Main insights

6.2.1 Insights gained from the competing risks analysis over standard survival analysis

Chapter 4 met the goal of demonstrating the additional insight acquired from the competing risk analysis of the mortgage case study, with particular focus on cause-specific hazard (CSH) and Fine-Gray (FG) subdistribution hazard modeling approaches involving its use.

Section 4.2 of this thesis introduced a Cox proportional hazard (PH) competing risks model for the time to mortgage prepayment or default as a function of mortgage covariates. The model considers the two hazards to be dependent on competing risks and jointly estimates them, allows one to take into account the fact that the observation of a mortgage default is also a censored observation of prepayment, and vice versa; consequently, the observation of mortgage default contains information about the mortgage prepayment that should be used in inference. The model accounts for the unobserved heterogeneity between mortgagors, estimates the unobserved heterogeneity at the same time as the baseline hazards and parameters associated with prepayment and default functions. Each of the remaining sections presented various different aspects of competing risks analysis to illustrate the insights that can be gained over standard survival analysis.

Furthermore, in section 4.2, we presented a univariate Cox proportional hazard analysis showing the effects of the covariates on the cause-specific hazard of each events prepaid and default. These covariates were assessed through seperate univariate Cox regression and all covariates with significant effects were selected, as risk factors, for both events. Multivariable analysis was then carried out, taking into account the covariates together in the model, allowing the influence of each covariate to be adjusted by other relevant variables. A stepwise selection was carried out until the final model was obtained for both mortgage default and prepaid, respectively.

In section 4.3 of this thesis, it was shown that a greater understanding of the competing risk scenario can be achieved by considering both the hazard and the cumulative incidence of each competing risk. This showed in particular the impact of covariates on the hazard ratios (subdistribution) of each competing risk, along with cumulative hazard and cumulative incidence plots. In the works of Latouche et al. (2013), the authors suggested showing both cumulative hazard plots and cumulative incidence plots to evaluate the impacts of covariates. In addition, it emphasized how the results observed in the competitive risk scenario have been interpreted in this way. It showed that the impact of a covariate on the cumulative incidence.of an outcome may not necessarily translate into the same impact on the the cause-specific hazard. We showed, in some instances, that exploring both the hazards and the cumulative incidences of each of the two competing risks may help to interpret certain outcomes, which would not be possible if only the hazard or cumulative incidence were taken into account in isolation.

6.2.2 Insights gained from validating models

The Single Family mortgages case study used in this thesis is part of the Federal Home Loan Mortgage Corporation(FHLMC), known as Freddie Mac. FHLMC Freddie Mac is a project that brings together individual mortgagor level data from various studies and trials at the direction of its regulator, the Federal Housing Finance Agency (FHFA) as part of a larger effort to increase transparency and help investors build more accurate credit performance models for validation. Thus, a valuable resource in which to validate the competing risks models for mortgage default rates and mortgage prepaid rates developed in Chapter 4 of this thesis. We used the calibration with cross validation method and the Area under the ROC Curve (AUC) to aid model checking for improving prediction accuracy.

6.3 Limitations

The study was to demonstrate competing risk modeling in areas in which their potential was not conceived in a financial context. This section highlights such constraints and suggests solutions that could overcome them in the future, if possible.

6.3.1 Variable selection with competing risks modelling

When developing predictive models, with combination of risk factors of a specific endpoint, careful consideration should be given to the predictive variable selected for inclusion in the model. The models play such an important role in predicting the outcomes of individual mortgagors with these risk factors.

Other models deemed to be parsimonious in nature are quite easy to use to accomplish a desired level level of prediction with as few predictor variables (risk factors) as possible, thus, more likely to be generalised to different population. The recognized strategies to reduce the number of covariates include choosing only those that are causative, eliminating obvious collinearity and ignoring those with data quality issues even though there are cases of dataset with high level of missing data. Moreover, even following such techniques to ignore covariates of apparent limited value, the number of variables may still exceed the standard thumb rule of Peduzzi et al. (1995) of 10 events per variable. Thus, variable selection methods based on modeling are often necessary. We have defined in this section certain challenge(s) and limitations of the selection of the variable in the sense of competing risk modelling.

Variable selection with cause-specific hazard modelling and Fine and Gray's proportional subdistribution hazard modelling

The information criterion-based test developed by Kuk and Varadhan (2013), that appears to be the first article to address variable selection using Fine-Gray's sub-distribution model, have been criticised for stability problems and their automation process. Thus, unfit for prognostic modelling analysts may not consider all variables as appropriate to ensure that they are financially relevant. However, the contributions of Kuk and Varadhan (2013) detailed the issues associated with this procedure which are deemed useful and also a starting point to build our models on.

In recent literature, there are approaches developed to enhance variable selection for competing risks models. The penalised regression approach with Fine and Gray's sub-distribution model developed by Fu et al. (2017) provides an alternative stepwise techniques to overcome difficulties associated with over-fitting. This approach allows the elimination of covariates and the most common example of this may be lasso-based regression. For the reasons outlined above, this approach adds value to the variable selection literature for competing risk models.

Section 2.5 of this study has previously mentioned that a prediction with a cause-specific hazard approach to competing risk modeling is feasible. Nonetheless, this approach makes variable selection much more challenging because the phenomenon that the effect of a covariate on the cause-specific hazard does not necessarily mean the same effect on its cumulative incidence. Likewise, the cumulative incidence may also be affected by the cause-specific hazards of any competing events. Therefore, it is not desirable to concentrate on the cause-specific hazard for one event isolated when selecting covariates that must be included in the modeling. Rather, each of the events could be viewed simultaneously, and the effects of a given covariate on each event taken together as a group when selecting that covariate.

In the works of Möst et al. (2016), the authors pointed out, that a distinction can be made between variable selection and parameter selection in multiple response models such as competing risks, that is, the effect of one covariates can be represented by several parameters. In addition, the authors (Möst et al., 2016) further emphasize the recommendation when selecting variables instead of selecting individual parameters for parsimonious modeling cases using a cause-specific hazard approach in order to use them for prediction purposes. Although, the works by Möst et al. (2016) is based on a discrete-time rather than a continuous time, competing risk framework neither does their works shows methods that perform efficient variable selection with explicit modeling of the heterogeneity in the population.

Many literature still considered prediction using cause-specific hazard modelling, when variable selection is a more challenging process than it is with Fine-Gray subdistribution modelling. As regards why we chose the cause-specific Cox regression approach, our reasons for it are: it allows comparison with the work in Bhattacharya et al. (2019) that we are interested in; the model parameters in the model have an interpretation (in terms of how the odds of each risk changes per unit change in the value of each explanatory variables); it allowed relatively easy implementation through existing R packages that could handle the scale of data in the Freddie Mac data set.

In the case where proportional hazard (PH) assumptions for covariates are met for cause-specific hazards approach but not with Fine-Gray subdistribution approach, the former may be used for prediction. In the instance where both approaches simultaneously meet the proportional hazard (PH) assumptions for covariates, then it is up to the analyst to decide which is favourable. Thus, the lack of a recognized appropriate method of variable selection in preparation for calculating predictions is a notable limitation in the cause-specific hazard approach to competing risks. The lack of an appropriate variable selection method may be possible to overcome by adapting variable selection techniques using lasso-based penalisation for multinomial logit models developed by the authors Turlach et al. (2005) and Tutz et al. (2015). Although their methods is not designed for the framework for the survival study, their contribution can, however, provide some insight into a competing risk solution.

6.4 Areas of future research

This section provides an overview of possible directions for future research. The natural extension of this work is to consider the heterogeneity between mortgages, so that the variability between mortgages could be explored and clusters of mortgages with similar properties identified. A heterogeneous model is a way to make this possible. By this I mean that each mortgage has its own covariate coefficient parameters in the Cox model (so $\theta_{D,i}$ and $\theta_{P,i}$ for mortgage i) and then there is a population-level model $p(\theta_{D,i}|\psi_D)$ and $p(\theta_{P,i}|\psi_P)$. This is also known as a hierarchical model in Bayesian modelling.

Generalise to time-dependent covariates e.g. current interest rate, house price inflation. The difficulty with this is prediction: if we want to predict when a particular mortgage might prepay or default then we have to model or predict how their covariates might change in the future.

Bibliography

Pricing commercial mortgages and their mortgage-backed securities. *The Journal of Real Estate Finance and Economics*, 1990. ISSN 08955638. doi: 10.1007/BF00178857.

Brent W. Ambrose and William N. Goetzmann. Risks and Incentives in Underserved Mortgage Markets. *Journal of Housing Economics*, 1998. ISSN 10511377. doi: 10.1006/jhec.1998.0232.

Per Kragh Andersen, Ronald B Geskus, Theo de Witte, and Hein Putter. Competing risks in epidemiology: possibilities and pitfalls. *International Journal of Epidemiology*, 41(3):861–870, 01 2012.

Giorgos Bakoyannis and Giota Touloumi. Practical methods for competing risks data: a review. *Statistical Methods In Medical Research*, 21(3):257 – 272, 2012. ISSN 1477-0334. URL `http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cmedm&AN=21216803`.

Arnab Bhattacharya, Simon P. Wilson, and Refik Soyer. A Bayesian approach to modeling mortgage default and prepayment. *European Journal of Operational Research*, 274(3):1112–1124, 2019. ISSN 03772217. doi: 10.1016/j.ejor.2018.10.047.

Jianwen Cai and Donglin Zeng. Cox Proportional Hazard Model. In *Wiley StatsRef: Statistics Reference Online*. 2014. doi: 10.1002/9781118445112.stat06880.

D. Collett. Modelling survival data in medical research. *Chapman Hall, London*, pages XVII + 347, 1994. URL `https://doi.org/10.1002/sim.4780140936`.

D R Cox. JSTOR: Journal of the Royal Statistical Society. Series B (Methodological), Vol. 34, No. 2 (1972), pp. 187-220. *Journal of the Royal Statistical Society Series B (. . .* , 1972.

Yongheng Deng and Stuart A. Gabriel. Risk-Based Pricing and the Enhancement of Mortgage Credit Availability among Underserved and Higher Credit-Risk Populations. *Journal of Money, Credit, and Banking*, 2006. doi: 10.1353/mcb.2006.0079.

James J. Dignam and Maria N. Kocherginsky. Choice and interpretation of statistical tests used when competing risks are present. *Journal of Clinical Oncology*, 26(24):4027–4034, 2008. doi: 10.1200/JCO.2007.12.9866. URL https://doi.org/10.1200/JCO.2007.12.9866. PMID: 18711194.

James J. Dignam, Qiang Zhang, and Masha Kocherginsky. The use and interpretation of competing risks regression models. 18(8):2301–2308, 2012.

Jason P. Fine and Robert J. Gray. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, 1999. ISSN 1537274X. doi: 10.1080/01621459.1999.10474144.

Zhixuan Fu, Chirag R. Parikh, and Bingqing Zhou. Penalized variable selection in competing risks regression. *Lifetime Data Analysis: An International Journal Devoted to Statistical Methods and Applications for Time-to-Event Data*, 23(3):353–376, 2017. URL https://EconPapers.repec.org/RePEc:spr:lifeda:v:23:y:2017:i:3:d:10.1007_s10985-016-9362-3.

M Gail. A review and critique of some models used in competing risk analysis. *Biometrics*, 31(1):209 – 222, 1975. ISSN 0006-341X. URL http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cmedm&AN=1164533.

Cortese Giuliana, Gerds Thomas A., and Andersen Per K. *Comparing predictions among competing risks models with time-dependent covariates.* New York, NY: John Wiley Sons, 2013., 2013. ISBN 0277-6715. URL http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ir01221a&AN=eld.81055.vdc.100024541913.0x000001.

Robert J. Gray. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, 16(3): 1141–1154, 1988. ISSN 00905364. URL http://www.jstor.org/stable/2241622.

Changbin Guo. Cause-specific analysis of competing risks using the phreg procedure. 2018.

Peng He, Frank Eriksson, Thomas H. Scheike, and Mei - Jie Zhang. A proportional hazards regression model for the subdistribution with covariates-adjusted censoring weight for competing risks data. *Scandinavian Journal of Statistics*, 43(1):103 – 122, 2016. ISSN 03036898. URL http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=112997835.

Michael G. Hudgens, Chenxi Li, and Jason P. Fine. Parametric likelihood inference for interval censored competing risks data. *Biometrics*, 70(1):1 – 9, 2014. ISSN 0006341X. URL http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=94857292.

Junyong In and Dong Kyu Lee. Survival analysis: part ii - applied clinical data analysis. *Korean journal of anesthesiology*, 72(5):441—457, October 2019. ISSN 2005-6419. doi: 10.4097/kja.19183. URL https://europepmc.org/articles/PMC6781220.

Cox John C., Jr. Jonathan E. Ingersoll, and Ross Stephen A. A theory of the term structure of interest rates. *Econometrica*, 53(2):385, 1985. ISSN 00129682. URL http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edsjsr&AN=edsjsr.10.2307.1911242.

Jonathan. Black–scholes option-pricing model. *A Dictionary of Finance and Banking*, 2018. ISSN 978-0-19-183143-0. URL http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edsoro&AN=oro.9780198789741.013.4235.

John P. Klein and 1940-2019 Moeschberger, Melvin L. *Survival analysis: techniques for censored and truncated data 2nd ed;.* Statistics for biology and health. Springer, 2003. ISBN 038795399X. URL http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=brd&AN=69279610.

Deborah Kuk and Ravi Varadhan. Model selection in competing risks regression. *STATISTICS IN MEDICINE*, 32(18):3077 – 3088, 2013. ISSN 02776715. URL http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edswsc&AN=000324991400003.

Bart Lambrecht, W. R. M. Perraudin, and S. (Stephen) Satchell. Mortgage Default and Possession Under Recourse: A Competing Hazards Approach. *Journal of Money, Credit, and Banking*, 2003. doi: 10.1353/mcb.2003.0021.

Aurelien Latouche, Arthur Allignol, Jan Beyersmann, Myriam Labopin, and Jason P Fine. A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. *Journal of Clinical Epidemiology*, page epub ahead of print, February 2013. doi: 10.1016/j.jclinepi.2012.09.017. URL https://hal.archives-ouvertes.fr/hal-00794995.

Bryan Lau, Stephen R. Cole, Richard D. Moore, and Stephen J. Gange. Evaluating competing adverse and beneficial outcomes using a mixture model. *Statistics in Medicine*, 27(21):4313, 2008. ISSN 02776715. URL http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=63566086.

D. Leece. Mortgage choice. Classical economics. In *International Encyclopedia of Housing and Home*. 2012. ISBN 9780080471716. doi: 10.1016/B978-0-08-047163-1.00142-9.

Mary Lunn and Don McNeil. Applying cox regression to competing risks. *Biometrics*, 51(2):524–532, 1995. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2532940.

Sally Hossam Eldin Ahmed Zakria Medhat. Modeling survival data by using cox regression model. *American Journal of Theoretical and Applied Statistics*, 4(6):504–512, 2015. URL http://www.sciencepublishinggroup.com/journal/paperinfo.aspx?journalid=146&doi=10.11648/j.ajtas.20150406.21.

S.and Pößnecker Möst, W., and G. Tutz. Variable selection for discrete competing risks models. *Quality and Quantity*, 50:1589–1610, 2016. doi: 10.1007/s11135-015-0222-0.

M. A. Nicolaie, Hans C. van Houwelingen, and H. Putter. Vertical modeling: A pattern mixture approach for competing risks modeling. *STATISTICS IN MEDICINE*, 29(11):1190 – 1205, 2010. ISSN 02776715. URL http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edswsc&AN=000277960900004.

Peter Peduzzi, John Concato, Alvan R. Feinstein, and Theodore Holford. Importance of events per independent variable in proportional hazards regression analysis. ii. accuracy and precision of regression estimates. *Journal of clinical epidemiology*, 48 12:1503–10, 1995.

M. Pfirrmann, A. Hochhaus, M. Lauseker, S. Saussele, R. Hehlmann, and J. Hasford. Recommendations to meet statistical challenges arising from endpoints beyond overall survival in clinical trials on chronic myeloid leukemia. *LEUKEMIA*, 25(9):1433 – 1438, 2011. ISSN 08876924. URL http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edswsc&AN=000294665400007.

R. L. Prentice, J. D. Kalbfleisch, A. V. Peterson, N. Flournoy, V. T. Farewell, and N. E. Breslow. The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4):541–554, 1978. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2530374.

R. L. Prentice, B. J. Williams, and A. V. Peterson. On the regression analysis of multivariate failure time data. *Biometrika*, 68(2):373–379, 1981. ISSN 00063444. URL http://www.jstor.org/stable/2335582.

H. Putter, M. Fiocco, and R. B. Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389, 2007. ISSN 02776715. URL http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=63565395.

Roberto G Quercia and Michael a Stegman. Residential Mortgage Default : A Review of the Literature *. *Journal of Housing Research*, 1992.

John M. Quigley and Robert Van Order. Efficiency in the Mortgage Market: The Borrower's Perspective. *Real Estate Economics*, 18(3):237–252, September 1990. doi: 10.1111/1540-6229.00520. URL https://ideas.repec.org/a/bla/reesec/v18y1990i3p237-252.html.

Prentice R. L., Kalbfleisch J. D., Jr. A. V. Peterson, Flournoy N., Farewell V. T., and Breslow N. E. The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4):541, 1978. ISSN 0006341X. URL http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edsjsr&AN=edsjsr.10.2307.2530374.

Gray Robert J. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, 16(3):1141,

1988. ISSN 00905364. URL
http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edsjsr&AN=edsjsr.2241622.

J. M. Robins and D. M. Finkelstein. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics*, 2000. ISSN 0006341X. doi: 10.1111/j.0006-341X.2000.00779.x.

J. M. Satagopan, L. Ben-Porat, M. Berwick, M. Robson, D. Kutler, and A. D. Auerbach. A note on competing risks in survival data analysis. *British Journal of Cancer*, 91(7):1229, 2004. ISSN 00070920. URL
http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=14542104.

Willi Sauerbrei, Patrick Royston, and Harald Binder. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Statistics in Medicine*, 26(30):5512, 2007. ISSN 02776715. URL
http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=63565680.

Thomas H. Scheike and Zhang Mei-Jie. Flexible competing risks regression modeling and goodness-of-fit. *Lifetime Data Analysis*, 14(4):464 – 483, 2008. ISSN 13807870. URL
http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=34995430.

DAVID Schoenfield. Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239, 1982. ISSN 00063444. URL
http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=57617078.

Rotraut Schoop, Jan Beyersmann, Martin Schumacher, and Harald Binder. Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical journal. Biometrische Zeitschrift*, 53 1:88–112, 2011.

Refik Soyer and Feng Xu. Assessment of mortgage default risk via Bayesian reliability models. *Applied Stochastic Models in Business and Industry*, 2010a. ISSN 15241904. doi: 10.1002/asmb.849.

Refik Soyer and Feng Xu. Assessment of mortgage default risk via Bayesian reliability models. *Applied Stochastic Models in Business and Industry*, 26(3):308–330, 5 2010b. ISSN 15241904. doi: 10.1002/asmb.849.

Terry M. Therneau and Patricia M. Grambsch. *Modeling survival data: Extending the Cox Model.* Statistics for Biology and Health series., 2000. URL http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eoh&AN=0553067.

Berwin A. Turlach, William N. Venables, and Stephen J. Wright. Simultaneous variable selection. *Technometrics*, 47(3):349–363, 2005. ISSN 00401706. URL http://www.jstor.org/stable/25471027.

Gerhard Tutz, Wolfgang Pößnecker, and Lorenz Uhlmann. Variable selection in general multinomial logit models. *Comput. Stat. Data Anal.*, 82: 207–222, 2015.

Zhongheng Zhang, Giuliana Cortese, Christophe Combescure, Roger Marshall, Minjung Lee, Hyun Ja Lim, Bernhard Haller, and written on behalf of AME Big-Data Clinical Trial Collaborative Group. Overview of model validation for survival regression model with competing risks using melanoma study data. *Annals of Translational Medicine*, 6(16), 2018. ISSN 2305-5847. URL http://atm.amegroups.com/article/view/21020.

Bingqing Zhou, Aurelien Latouche, Vanderson Rocha, and Jason Fine. Competing risks regression for stratified data. *Biometrics*, 67(2):661 – 670, 2011. ISSN 1541-0420. URL http://elib.tcd.ie/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cmedm&AN=21155744.

Appendix A1

Appendix

A1.1 R Code :

```
> predictRisk(csc.mortgage, newdata = dftest[60,],
    times = 14, cause = 1)
    [1,] 0.02993827
> predictRisk(fgr1.mortgage, newdata = dftest[60,],
    times = 14)
    [1,] 0.02675687
```

A1.2 Tables :

Table A1.1: Test PH assumption by using Scaled Schoenfeld residuals for Model 1

	Chisq	df	P-value
Property.Type	1.52	1	0.21834
No.of.borrowers	1.03	1	0.31114
CLTV	1.72	1	0.18966
Credit.score	13.93	1	0.00019
Original.interest.rate	6.09	1	0.01358
GLOBAL	21.46	5	0.00066