BRITISH COLUMBIA | Financial Institutions Commission

March 6, 2015
File No.: 450-04
Ref. No.: 0877

To: BC Credit Unions

Re: Research Report on Residential Mortgage Default Probability

This letter announces a new research report by the Financial Institutions Commission (FICOM) on models and methods for analyzing the probability of default of residential mortgages.

FICOM has conducted a review of academic literature to identify some of the theories and models for estimating the probability of default for credit union personal real estate secured loans. Please find the report summarizing the results at http://www.fic.gov.bc.ca/index.aspx?p=fid/publications.

We anticipate that the information presented will be of assistance to credit unions in developing more effective credit risk models that can potentially be used for stress testing and other risk management activities.

For any questions or concern regarding the report, please contact me at mehrdad.rastan@ficombc.ca.

Yours truly,

[redacted]

Mehrdad Rastan
Executive Director, Risk Surveillance & Analytics

ML/ac

- Superintendent of Financial Institutions
- Superintendent of Pensions
- Superintendent of Real Estate
- Registrar of Mortgage Brokers

2800-555 West Hastings
Vancouver, BC V6B 4N6
Telephone: 604 660-3555
Facsimile: 604 660-3365
www.fic.gov.bc.ca

Financial Institutions Commission

Research Paper

October 2014

Residential Mortgage Probability of Default Models and Methods

by

Mingxin Li

Risk Surveillance and Analytics

Financial Institutions Commission

About this report

Mingxin Li is a PhD candidate in the Beedie School of Business at Simon Fraser University. The research was completed under the supervision of the Financial Institutions Commission staff. The views expressed in this paper are those of the author. No responsibility for them should be attributed to the Financial Institutions Commission.

Acknowledgements

I would like to take this opportunity to acknowledge the following individuals:

I would like to thank Dr. Evan Gatev and Dr. Christina Atanasova for reviewing the paper and providing many helpful comments.

I would like to thank Mr. Mehrdad Rastan, Mr. Gilbert Yuen, Mr. Peter Lee, and Mr. Jack Ni for the intensive discussion and valuable feedback during the development of this project.

I would also like to thank Ms. Angel Chen for proofreading and editing this research paper.

Table of Contents

Executive Summary **3**

I. Introduction **4**

II. Evaluating Mortgage Default Risk in the Early Days **5**

III. Models for Default Risk of an Individual Loan **6**

Model 1: linear regression analysis on default risk 6

Model 2: logistic model 9

Model 3: survival analysis 13

Model 4: optimization model 17

IV. Models for Default Probability of a Loan Portfolio **20**

Model 5: linear regression analysis on default rates 20

Model 6: linear regression analysis on log odds 24

V. Default Determinants Implied from Economic Theories **25**

Explaining default in the early days 26

Competing theories of default behavior 27

Option-based theory of default behavior 29

Macroeconomic factors 32

VI. Issue of Model Stability **35**

VII. Conclusion **36**

Appendix 1.a: Overview of Models **38**

Appendix 1.b: Loan-level Model versus Portfolio-level Model **39**

Appendix 2: Determinants of Residential Mortgage Default Risk **40**

References **42**

Executive Summary

Stress testing is the investigation of an entity's performance under abnormal circumstances. Financial institutions should conduct stress tests to gauge the resilience of their balance sheets to substantial macroeconomic shocks. One way to measure the performance of a financial institution is by assessing the institution's loan portfolio loss under stressed scenarios. The first step in assessing loan loss is to estimate the probability of default (PD). Understanding PD is necessary for the purpose of stress testing and risk management. Financial institutions may also find it beneficial as insights from default modeling can be incorporated to guide improvements on good underwriting practice and competitive mortgage pricing.

This paper serves as a rigorous background research on PD. We draw upon academic literature on residential mortgage default and research papers on stress testing published by other regulatory bodies, and pull together six models (five statistical models and one economic model) that can be used to generate quantitative assessments of PD. We also comb through the development of economic theories aimed at explaining default behaviors. The economic theories provide the basis for selecting default determinants, which in turn are used as inputs in statistical models to predict PD. This paper sheds light on the questions of what drives default and how to model the probability of default for residential mortgages and mortgage portfolios.

Our goal is to present available methods for the purpose of modeling PD, rather than to recommend specific models or default determinants for financial institutions to use. FICOM and the credit unions, in choosing a model, should assess the suitability of the model giving consideration to specific business requirements. Further research into the model may be required for seamless execution.

I. Introduction

Although default rates on residential mortgages in BC have been relatively low in the past, credit unions should still be concerned about mortgage default for several reasons. First, residential mortgages make up a large portion of the asset portfolios of BC credit unions. According to data, almost 68 per cent of BC credit unions' total loans are personal real estate backed assets. Secondly, home mortgages represent a large bulk of outstanding household debt. As of the second quarter of 2014, mortgages account for 47 per cent of total consumer debt in BC.[1] Default is costly to everyone involved. Costs to the lender and the insuring institution incur when net cash recouped from foreclosure is less than the remaining balance of the defaulted mortgage. In the extreme case, systemic defaults may impair the soundness of lending institutions. Default is also costly to the borrower. Examples include the loss of a home, a lower credit rating, an impaired ability to acquire financing, and even mental distress. In addition, default risk is of particular concern given the continuously climbing housing price in the Greater Vancouver area. When the US last experienced a housing price run-up, what followed was a disastrous crash, the effects of which still persist today. Acknowledging the differences between the real estate and mortgage markets of BC and those of the US, we do not attempt to make predictions of the housing market in BC; rather, we emphasize the importance of understanding the risk of mortgage default, as real estate backed loans play a key role in our financial system. Understanding mortgage default risk will not only provide guidance for designing stress testing scenarios but also help improve underwriting practices and enhance pricing of mortgage products.

The goal of this paper is to provide an overview of alternative methods that can be applied to answer the question – How should lending institutions assess the default probability on a pool of mortgage loans? Firstly, section II briefly discusses how default risk was assessed in the early days and why that is insufficient in understanding default risk today. Then section III and IV describe six models that can be used to estimate default probability given certain factors. Appendix 1 offers an overview. The models are introduced in the order as they were first applied in studies of residential mortgage default. Adoptions of later models are often spurred by some inadequacy of earlier ones in answering the question of interest or are inspired by new developments in statistical methods and computer programming capabilities. Model 1, 2, 3 and 4 are for individual loans; Model 5 and 6 are for loan portfolios. Model 1 uses a linear probability

[1] Information of BC credit unions asset mix and total household debt distribution are from FICOM DTI Q2 2014 report.

function to model default risk; it is simple and robust in discriminating loans based on a predicted default risk index; however, this model does not provide a number for the default probability. Model 2 overcomes this shortfall and uses a logistic function to model default probability. Model 3 applies a time-to-event method to model the length of time before a mortgage terminates. Model 4 departs from these regression-type models; instead, for every possible outcome for house prices and interest rates over a period of time, it simulates a borrower's decision over three choices: continuing with the current mortgage, defaulting, or prepaying the current mortgage. Model 5 and 6 view a mortgage portfolio as a whole and analyze the default rate of the portfolio.

Section III and IV do not discuss (except for Model 4) the factors that one would input into the models. These factors are macroeconomic measures, loan-, and borrower-specific characteristics that potentially drive default behavior. They are sometimes referred to as default determinants. These models have flexibility in terms of the factors they accept as inputs. It is up to the users to choose the factors. Section V discusses these factors as suggested by economic theories. Appendix 2 presents a summary of default determinants. Finally, section VI discusses the issue of model stability, and section VII concludes the paper. Models and methods discussed hereafter draw upon studies done in the past by researchers. A list of references is provided at the end for further investigation.

II. Evaluating Mortgage Default Risk in the Early Days

Prior to the 1980's, the evaluation of mortgage default risk was largely established on rules of thumb and risk ratings based on experience ([34]). Mortgage applications were scored or rated on a grid given borrower-, loan-, and property-related criteria. Four ratios were employed back then and are still in use today. They are the loan to value ratio, the monthly mortgage payment to gross income ratio, the total debt obligation to gross income ratio, and the house value to gross income ratio. These ratio analysis and risk ratings specify some indicators of default risk; however, they are insufficient mainly in two ways. Firstly, they look at the likelihood of default during the life time of a mortgage, but do not deal with the timing of default. As shown by researchers, marginal probabilities of default display a rising-then-falling pattern over time.[2] Secondly, the risk ratings do not provide quantitative assessments of the likelihood of default.

[2] Von Furstenberg ([34]) is the first to reveal this pattern. For his loan sample, default rates peak around 3 to 4 years after origination and subsequently fall and become negligible after half the term of a mortgage has passed.

The shortcoming is twofold. A rating or score of, say, 1 out of 10 may indicate that the mortgage is likely to default, but it does not tell how likely it is to default (i.e., whether there is a 90 per cent or 60 per cent probability of default). Also, these risk ratings do not estimate the degree of impact each criterion has on the likelihood of default. In turn, a differential in rating indicates that one mortgage is more or less likely to default than another, lacking insights on how much more or less the likelihood is.

III. Models for default risk of an individual loan

This section outlines four default risk models, where one considers individual mortgages as the subject of study. Model 1, 2, and 3 are statistical models that predict default risk by estimating relationships between default risk and default determinants. Model 4 is an economic model based on optimization, which estimates default risk by describing a borrower's behavior under certain economic forces. A description is provided for each model, followed by the model implementation with data structure examples; the model is then compared to earlier ones to show the advantages and disadvantages.

Model 1: Linear regression analysis on default risk

Description

Regression analysis looks for the relationships between default risk and an array of variables that may have impacts on default behavior. Default risk is treated as a dependent variable, which can be explained by some independent or explanatory variables. The relationship between default risk and its explanatory factors is assumed to be linear. A common formulation[3] is

$$default\ risk = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k + \varepsilon \quad (1)$$

where $X_1, X_2, \ldots X_k$ are explanatory variables, factors or predictors that may help determine default risk; α is a constant; $\beta_1, \beta_2, \ldots, \beta_k$ are coefficients that capture the impact that each factor may have on default risk; and ε is an error term, which is assumed to be independent and is sometimes in addition assumed to be normally distributed. *Default risk* here is not measured by the probability of default, as a loan is either in default or not in default. One does not observe a "probability" for a single loan; rather, the status of the loan is observed. Loan status is used as a proxy for *default risk*. If a mortgage is in good standing, then the *default risk* measure takes a

[3] See Quercia and Stegman ([29]) for a list of studies.

value of zero; if a mortgage is in default (either in delinquency or foreclosure), then the *default risk* measure takes a value of one. Explanatory variables, $X's$, are any factors that may affect the default risk of a mortgage. These factors can be macroeconomic, loan specific, borrower-, lender-, or property- related. They are derived from economic reasoning as well as empirical observations. In the simplest specification, the default risk is assumed to have a linear relationship with the factors. Factors may be transformed before entering the regression equation. We discuss the selection of explanatory variables later.

Implementation

One can observe the performance status of a sample of loans and conduct regression analysis. There are two ways to do it: 1) a cross-sectional dataset is obtained if a sample is observed at one point in time; 2) a panel dataset is obtained if a sample is observed at multiple points in time.

If data is prepared as a snapshot of a loan profile at one point in time, the regression is cross-sectional. Figure1 gives an example of cross-sectional loan data.

Figure 1. Cross-section data on individual mortgages: data structure example

Loan ID	Loan Status	X_1: loan-to-value	X_2: term of mortgage	X_3: borrower occupation
1	0	80%	20	3
2	0	85%	25	4
3	1	90%	25	2
……	.....	……	……	……

Fitting the model with data yields estimates of the coefficients, $\beta's$, in equation (1). The coefficients estimate the impact of each factor on default risk, by how much default risk changes when a factor changes by a particular amount. Alternatively, the estimation may suggest that a factor does not have a significant impact on default risk. Using estimated coefficients and given values of explanatory variables, we can compute the predicted default risk for a particular mortgage from equation (1).

If data is prepared such that there are multiple mortgages in the sample and each mortgage is observed at multiple points in time, one would have a panel dataset. Estimation of the model then follows panel regression techniques. An example of panel loan data is shown in Figure 2.

Figure 2. Panel data on individual mortgages: data structure example

Loan ID	Date	Loan Status	X_1: loan-to-value	X_2: term of mortgage	X_3: borrower occupation	X_4: GDP growth
1	2005	0	80%	20	3	1.5%
1	2006	1	85%	20	3	1.2%
2	2005	0	85%	25	4	1.5%
2	2006	0	84%	25	4	1.2%
2	2007	0	80%	25	4	1.3%
2	2008	0	83%	25	4	1.0%
3	2005	1	90%	25	2	1.5%
……	……	.....	……	……	……	……

Advantage and disadvantage

The linear regression model is easy to implement and the interpretation of the output is straightforward. Equation (1) can have good discriminating power and can be used to rank mortgages by estimated default risk; lower output values indicate lower default risk and high output values indicate higher default risk. However, the model has several problems in general. When default risk is measured by loan status, it only assumes a value of either zero or one. From equation (1), one can see that with a dichotomous dependent variable, the error term ε is dichotomous as well. This is inconsistent with the model assumption on normally distributed errors. Predictions from a linear probability function may be difficult to interpret. In order to have a probability interpretation, the output of the estimated equation should be a number between zero and one, even when particular values are assigned to the explanatory variables. For example, when designing stress scenarios, one may set the house price index at a stressed level to estimate the resulting default probability. If the output of equation (1) is negative or above one for some set of factors, then one cannot interpret the estimated default risk as a probability of default. So the output of the model may be viewed as a default risk index rather than a default probability of a mortgage. The model does not answer the questions of interest – What is the probability of default given values of the explanatory variables?

Model 2: Logistic model

Description

The performance status of a mortgage loan is often described as current, 30-, 60-, 90-day delinquent, foreclosed, refinanced, et cetera. In statistical analysis, this information is qualitative data, and is represented using categorical indicators.[4] A logistic model is particularly suitable for empirical studies with qualitative data. Consider the loan status, a binary variable which takes a value of either zero (for mortgages that are performing) or one (for non-performing mortgages). A logistic model formulates the probability of a loan being non-performing as a logistic function of some combination of explanatory variables[5]:

$$P(loan\ status = 1) = \frac{1}{1+e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots \beta_k X_k)}} \quad (2)$$

where $P(loan\ status = 1)$ is the probability of a mortgage being non-performing.[6] Equation (2) can be seen as a transformation of equation (1), a positive monotone transformation that maps the linear probability predictor into a unit interval. Such a transformation will retain the linear structure of the model while ensuring the estimated output stays between zero and one.

Implementation

Suppose that the one-year default probability is desired and one draws a loan sample in 2010. All loans that are outstanding at the beginning of 2010 enter the sample, and one observes the loan status at the end of 2010. An example of loan data looks like Figure 1.

The model is estimated using likelihood techniques, and goodness-of-fit tests can be conducted to assess whether or not the model fits the data on hand. Logit coefficients, $\beta's$, estimate the impact of a unit change in factors on the natural logarithm of odds. Odds have the intuitive meaning of $\frac{\pi}{1-\pi}$, where π is the probability of a mortgage being non-performing. For example, the odds of a loan being in default are the probability of default versus the probability of non-

[4] For example, if mortgages in a portfolio are either current or non-current, one may use a value of zero for mortgages that are current and one for mortgages that are non-current. If mortgages in a portfolio are current, delinquent, or foreclosed, one may use a value of one for mortgages that are current, two for delinquency, and three for foreclosure.

[5] See Quercia and Stegman ([29]) for a list of studies.

[6] McFadden ([25]) shows that the logistic function is an appropriate representation of consumer choice behavior under reasonable assumptions. In this application, it is the borrower's choice of continuing servicing current mortgage, becoming delinquent, defaulting, or prepaying.

default. A simple conversion gives the impact of factors on the default probability π. Using estimated coefficients and given values of explanatory variables, one can compute the predicted probability of default for a particular mortgage from equation (2). The predicted probability can also be used to classify mortgages. For example, one may choose a cut-off value, say 0.5, and classify mortgages with predicted probability above 0.5 into a group of predicted default loans and mortgages with predicted probability below 0.5 into a group of predicted performing loans.

If one has a separate record for each time period in which each mortgage is observed, the data structure is similar to Figure 2, and panel regression techniques apply.

Sometimes, one may have a finer categorization of mortgages, more than just "performing" and "non-performing". Consider a loan sample consisting of three groups of mortgages based on their performance status. Loan status equals to 1 for mortgages that are performing, 2 for mortgages that default, and 3 for mortgages that are prepaid. In this case, one would use a multinomial logistic model. Figure 3 is an example of multinomial data with loan sample observed at one point in time.

Figure 3. Cross-section data on individual loans: data structure example

Loan ID	Loan Status	X_1: loan-to-value	X_2: term of mortgage	X_3: borrower occupation
1	3	90%	20	3
2	2	85%	25	4
3	1	80%	25	2
……	.....	……	……	……

Estimation of a multinomial logistic model takes one group as the base group and identifies coefficients for the rest of the groups. For example, if one uses performing mortgages (group 1) as the base group, then the model estimates two sets of coefficients, one for each of default mortgages (group 2) and prepaid mortgages (group 3). The natural logarithm of odds of a mortgage falling in group i versus the base group is

$$\ln\left(\frac{P(loan\ status=i)}{P(loan\ status=1)}\right) = \beta_0^i + \beta_1^i X_1 + \cdots \beta_k^i X_k \qquad (3)$$

where group $i = 2\ or\ 3$, and $\beta^{i\prime}s$ are coefficients quantifying impact of factors on a mortgage falling in group i versus the base group. For example, β_1^2 of 0.5 is interpreted such that a one unit increase in X_1 results in a 0.5 increase in the natural logarithm of odds that the loan falls into

group 2 versus group 1; or the odds of falling into group 2 versus group 1 increase by $e^{0.5}$ as a result of one unit increase in X_1. Coefficients for the base group, $\beta^{1\prime}s$, are set to zero for the purpose of estimating the model. From equation (3), one can derive the probability of a mortgage falling in group i to be

$$P(loan\ status = i) = \frac{e^{(\beta_0^i + \beta_1^i X_1 + \cdots \beta_k^i X_k)}}{\sum_i e^{(\beta_0^i + \beta_1^i X_1 + \cdots \beta_k^i X_k)}} \quad (4)$$

where group $i = 1, 2, or\ 3$. For a particular mortgage with a given set of explanatory variables, equation (4) computes the predicted probability of the mortgage falling in group i. Any of the three groups can be used as the base group. Coefficient estimates are different depending on the choice of the base group; however, predicted probabilities will be the same regardless of the choice. For classification, a mortgage would be assigned to the group with the largest predicted probability. For example, fitting the model with data one can estimate β^2 and β^3; β^1 for the base group is set to 0. Then for a mortgage with given values of $X^{\prime}s$, from equation (4) we predict 70 per cent probability for it falling in the performing group, 20 per cent probability for it falling in the default group, and 10 per cent for it falling in the prepayment group. And one would classify this mortgage into a group of predicted performing loans.

Figure 4 is an example of a panel dataset where a particular mortgage has multiple records from multiple points in time. Panel regression techniques apply.

Figure 4. Panel data on individual loans: data structure example

Loan ID	Date	Loan Status	X_1: loan-to-value	X_2: term of mortgage	X_3: borrower occupation	X_4: GDP growth
1	2005	1	90%	20	3	1.5%
1	2006	3	85%	20	3	1.2%
2	2005	1	85%	25	4	1.5%
2	2006	1	84%	25	4	1.2%
2	2007	2	86%	25	4	1.3%
3	2005	1	80%	25	2	1.5%
3	2006	1	76%	25	2	1.2%
3	2007	1	75%	25	2	1.3%
3	2008	1	77%	25	2	1.0%
……	……	.....	……	……	……	……

Advantage and disadvantage

If one is only concerned with the significance of the relationship between loan status and explanatory factors, both the linear regression model and the logistic model may yield similar results. If the goal is to estimate the probability of an event, such as mortgage default, then the logistic model is better. It overcomes the problems of a linear regression model in analysing categorical data, and fits the observed loan status better than a linear regression model. Coefficient estimates under the logistic model are efficient and well behaved even when the sample size is relatively small.[7] Equation (2) and (4) can used to predict probabilities of default for mortgages. The outputs from the model fall within a sensible range between zero and one. The output of default probability predictions is on a loan-by-loan basis.[8]

One caveat is that the logistic function may not fit a particular dataset. If the probability of default is not monotonic in relation to an explanatory variable, then logistic regression would not fit the data. For example, von Furstenberg ([36]) reveals a single-peaked pattern for the term structure of mortgage default. Term structure is the relationship between the default rate and the mortgage age. On average, default rates increase and peak a few years after origination and subsequently decrease until they become negligible. Here default probability is not a monotonic function of the mortgage age. To accommodate this, one can include both the mortgage age and the squared mortgage age as explanatory variables. Alternatively, one can use multiple age categories and dummy variables to account for the non-linear relationship between age-of-mortgage and default probability.[9]

Another consideration before using a multinomial logistic model is that the model relies on an assumption of independence of irrelevant alternatives, which says that odds of one group relative to another are unaffected by the presence or absence of the third group. In our example, this is to assume that 1) the possibility of refinancing is irrelevant to how likely a mortgage would be in default rather than performing; 2) the possibility of default is irrelevant to how likely a mortgage would be refinanced rather than performing; and 3) the possibility of performing is irrelevant to

[7] Another advantage of logistic model arises with the consideration of sampling schemes: whether it is prospective sampling or retrospective sampling. Logistic model specifies functional form for odds of one outcome relative to another, instead of for probabilities directly. Odds are identical regardless of the sampling scheme.
[8] The prediction resulting from the model is something like: loan #1 has 2% probability of default; loan #2 has 1% probability of default, and so on. It is not like: default rate of the portfolio is 3%. The latter is dealt with in Model 6.
[9] An APRA working paper by Coleman et al. ([8]) uses multiple age categories (dummy variables) to account for the non-linear relationship between age-of-mortgage and default probability.

how likely a mortgage would be refinanced rather than in default. Whether or not this assumption holds for our application is debatable, and which is a limitation. However, even if the assumption were violated, multinomial logistic model would still be more effective than other models that do not rely on this assumption.

Model 3: survival analysis

Description

Survival analysis is a modelling technique for time-to-event data or duration data. Consider the life course of a mortgage. At each point in time, the mortgage may enter one of a number of mutually exclusive states, such as performing, default, and prepayment. With the passage of time, the mortgage moves between these states (or it remains static). It is likely that the mortgage will start in the performing state and later stay in the performing state or move into either default or prepayment. Survival analysis is a tool to study the length of time the mortgage spends within the performing state, in other words, how long the mortgage survives before it defaults or prepays. We seek the relationship between mortgage status and the passage of time along with other explanatory variables. A common formulation for survival analysis[10] is

$$h(t) = h_0(t)e^{(\beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k)} \tag{5}$$

where $h(t)$ is the hazard rate, or the conditional probability that a mortgage survives until time t but fails during the next time interval;[11] time t represents age of mortgage; $h_0(t)$ is the baseline hazard, which captures the shape of the hazard function and summarizes how the probability of mortgage termination (either default or prepay) changes over time; $X_1, X_2, \ldots, X_k$ are explanatory variables that also influence risk of mortgage termination; and $\beta_1, \beta_2, \ldots, \beta_k$ are coefficients that measure the impacts of the explanatory variables on the hazard rate.

Implementation

Before conducting survival analysis, one first organizes data into a loan-period format. One now needs an event indicator (loan status) and time variables that can be used to imply duration of

[10] Refer to Survival Analysis by Stephen P. Jenkins.

[11] This interpretation is appropriate for discrete time hazard rate. In continuous time, $h(t)\Delta t$ has similar interpretation. In this context, survival of a mortgage means that it stays in the performing state; failure of a mortgage means that it exits the performing state and moves into either default or prepayment.

time before a mortgage moves out of the performing state. Figure 5 is an example of the data structure where loan sample is observed at one point in time.

Figure 5. Duration data on individual loans: data structure example

Loan ID	Origination date	Event date	Event type	X_1: initial loan-to-value	X_2: term of mortgage	X_3: borrower occupation
1	2005Q1	2009Q3	2	90%	20	3
2	2003Q3	2009Q4	1	85%	25	4
3	2001Q2	2010Q1	0	80%	25	2
……	……	……	……	……	……	……

In this example, the sample period ends in 2010 Q1. At this time loan #3 is still in the performing state; the "Event" indicator for loan #3 is 0, and "Event date" is the same as the end of the sample period. Loan #1 prepays in 2009 Q3 with "Event" indicator equal to 2; loan #2 defaults in 2009 Q4 with "Event" indicator equal to 1.

Now suppose that one draws a sample period from 2009 Q2 to 2010 Q1, and observes the loan sample every quarter. Figure 6 gives an example of the data structure with time varying explanatory variables.

Figure 6. Duration data on individual loans: data structure example

Loan ID	Origination date	Event date	Event type	X_1: current loan-to-value	X_2: term of mortgage	X_3: borrower occupation	X_4: GDP growth
1	2005Q1	2009Q2	0	80%	20	3	1.5%
1	2005Q1	2009Q3	2	85%	20	3	1.2%
2	2003Q3	2009Q2	0	85%	25	4	1.5%
2	2003Q3	2009Q3	0	83%	25	4	1.2%
2	2003Q3	2009Q4	1	84%	25	4	1.3%
3	2001Q2	2009Q2	0	61%	25	2	1.5%
3	2001Q2	2009Q3	0	63%	25	2	1.2%
3	2001Q2	2009Q4	0	62%	25	2	1.3%
3	2001Q2	2010Q1	0	60%	25	2	1.1%
……	……	……	……	……	……	……	……

The data structure example presented above represents a competing nature of two types of mortgage termination risk – risk of default and risk of prepayment. A lender of an outstanding

mortgage faces these two types of termination risk. The mortgage terminates once one of the two risks is realized. The two risks are jointly present; however, their realizations are mutually exclusive. A mortgage that prepays will not default, whereas a mortgage that defaults will not prepay. Thus default and prepayment are "competing" risks. For this reason, a complete model of mortgage termination risk should simultaneously consider both default risk and prepayment risk.[12] The most current method for mortgage termination under this framework is a competing risks hazard model, which is illustrated by Deng, Quigley and Van Order ([12]). The competing risk hazard model is a unified model that analyzes the joint choices of default and prepayment, and estimates the influence of factors on default decision as well as prepayment decision.

The estimation of the model uses likelihood techniques, which try to find the values of coefficients, $\beta's$, that maximize the probability of observing the data on hand. Assumptions on the functional form of the baseline hazard, $h_0(t)$, are not required to estimate coefficients. Cox-regression fits the model to data and estimates equation (5) by maximizing the partial likelihood function derived from the equation; the baseline hazard is common to all mortgages and its contributions cancel out in the partial likelihood expression, so when the estimation process maximizes partial likelihood, the baseline hazard does not make a difference and only the coefficients, $\beta's$, are estimated.[13] Using estimated coefficients and empirical baseline hazard, one can compute conditional default probabilities from equation (5) for a particular mortgage with given values of explanatory variables.

The likelihood estimation techniques allow one to control for unobserved variables. Equation (5) implies that two mortgages with same age, t, and same explanatory characteristics, $X's$, have identical default risks. For simplicity, consider only two factors influencing default, loan-to-value ratio (LTV) and borrower occupation. Suppose two mortgages, both one year old with LTV of 70 per cent, and both borrowers in the same occupational category. Equation (5) would predict an identical probability of default for these two mortgages. The problem is that the two borrowers are most likely to differ in ways that are not captured by occupation; they may differ in consumer behavior, habit, ability to pull external financial resources, et cetera. These

[12] Using the option theory, we can consider the prepayment option as a call option and the default option as a put option. The borrower of an outstanding mortgage holds both options. Once one of the two options is exercised, the mortgage is terminated and the other option is foregone. This is to say, when the borrower makes the decision to exercise one option, he/she would bear in mind the value of the other option. Hence a model of default risk should also address the presence of prepayment risk.

[13] Refer to Survival Analysis by Stephen P. Jenkins and Buis ([4]) for more explanation.

differences are unobserved or unmeasured, but they do play a role in the borrowers' default decisions. To deal with this issue, one can assume that mortgages in the sample belong to some number of groups; mortgages in each group are similar in terms of the unobserved characteristics. Mortgages are not pre-assigned into groups. The likelihood estimation process will generate coefficient estimates taking into account the presence of unobserved characteristics. One can start with two groups and subsequently increase the number of groups until the model performance no longer improves. Incorporating unobserved characteristics enhances the estimation results. This paper does not discuss details of the estimation process. Buis ([4]) explains how the likelihood estimation process incorporates unobserved characteristics.[14]

Advantage and disadvantage

The survival analysis method is well accepted in studies of default probability because it matches the life course of a mortgage and its termination process. The estimated output provides forecasts of default probabilities as a function of time (the mortgage age) and other default determinants. It models both probability of default and time dependence of the probability. This is an advantage over the logistic model. In a logistic model, the predicted probability has a fixed time horizon; to have prediction for a different time horizon, one needs to revise the loan sample and repeat the estimation process. Survival analysis can estimate default risk for any time horizon. Also, survival analysis handles censored data, which is an issue not addressed by the logistic model.[15] Another advantage of survival analysis is its versatility, because assumptions on the functional form of the baseline hazard are not required to estimate the model. To predict default probability for a particular mortgage over time, the model uses the empirical baseline hazard along with estimated coefficients and given values of explanatory variables.

[14] Logistic and multinomial logistic models with panel dataset may also be estimated with fixed effects which control for time-invariant, borrower-specific unobserved variables.

[15] A time-to-event (survival time) is censored if we only know the observation either entered or exited within the sample period, and the total length of survival time is not known exactly. For example, a mortgage that is still outstanding at the end of the sample period is censored because we only know that it has not defaulted yet but we do not know whether it will mature without default or not. Another example, a mortgage exits the sample during sample period because lender sold the mortgage, thus performance status of this mortgage is not observed. For the former example, logistic model implicitly ignores the issue; for the latter example, logistic model abandons the observation due to missing data. Survival analysis incorporates censored data in its estimation process.

Model 4: optimization model

Description

An optimization model of default attempts to capture the core structure of economic dynamics surrounding the default process. This type of model assumes that a borrower makes mortgage payment decisions with objectives to maximize wealth and utility or minimize housing-related costs. At one point in time, a vector of choices available to the borrower normally includes:

1) to make the scheduled mortgage payment and continue with the current mortgage,
2) to prepay the current mortgage, or
3) to default on the current mortgage.[16]

There are various wealth effects associated with each of the choices. A borrower compares these wealth effects and chooses to default if it meets his/her objective better than the other alternatives.

Capozza, Kazarian, and Thomson ([7]) provide an example of utilizing a dynamic optimization model for estimating residential mortgage default behavior. Consider a time line that is divided into monthly intervals.[17] At each time interval, a borrower makes a decision around mortgage payment and chooses the least costly action. The borrower assesses whether it is less costly to default, to refinance, or to continue with the current mortgage. At each interval, a borrower's choice can be written as:

$$P_t(H_t, r_t) = \min[P_t^d(H_t, r_t), P_t^r(H_t, r_t), P_t^w(H_t, r_t)] \quad (6)$$

where P_t is the borrower's housing cost at time t; P_t^d is the housing cost if the borrower chooses to default; P_t^r is the housing cost if the borrower chooses to refinance; P_t^w is the housing cost if the borrower chooses to continue with the current mortgage; H_t is the property value at time t, modeled as a stochastic process; and r_t is the interest rate at time t, modeled as another stochastic process. The stochastic processes are functions that specify how house prices and interest rates will evolve over time. With an initial value, one can use the processes to simulate possible house prices and interest rates over time.

[16] The specification of borrower's choices follows Capozza, Kazarian, and Thomson ([7]). Souissi ([33]) has a similar setup. Others may have finer or coarser differentiation among choices. For example, another model may distinguish prepayment by sale of property from prepayment by refinancing.

[17] Each time interval is one time-step; length of the time interval can be one month to represent monthly mortgage payment.

Implementation

How is the default probability estimated from the model? First, one generates possible outcomes for house prices and interest rates over the term of the mortgage. This is done by assuming stochastic processes for house prices and interest rates, respectively. At each time interval, this provides the distributions of the house price and the interest rate at that time. Then one explicitly expresses the functions of P_t^d, P_t^r, and P_t^w. At a time interval t, the cost of default, P_t^d, is the property value at that time plus the transaction costs of default[18]; the cost of refinancing, P_t^r, is the periodic mortgage payment plus the outstanding mortgage balance along with the deadweight cost of refinancing[19]; the cost of continuing with the current mortgage, P_t^w, is the periodic mortgage payment plus the expected mortgage cost in the future, $E(P_{t+1})$. The expected future mortgage cost does not affect the cost of default or refinancing, because the mortgage is terminated after default or refinancing. Equation (6) is recursive – the borrower repeatedly makes such decision at every time interval until the mortgage matures or terminates, whichever comes first; also, the borrower's choice today is influenced by the borrower's expectation of housing costs tomorrow – P_t^w is a function of P_{t+1}, which is the same as equation (6) except with subscript $t+1$ instead of t. In addition, the function of P_t^w can be modified to include a trigger event. A trigger event is a random event, such as divorce or unemployment, which can happen to an average borrower and that "triggers" mortgage termination. The "trigger event" is modeled by Capozza, Kazarian, and Thomson ([7]) as a probability that such a random event happens. With a 10-year mortgage, equation (6) represents a system of 120 equations, one for every monthly time interval.[20] Finally, from distributions of H_t and r_t at each time interval the model generates a distribution of mortgage payment choices, from which one calculates probability of default for that time interval. The solution techniques and calculation of default probabilities are described in Capozza, Kazarian, and Thomson ([7]).[21]

[18] Assume that the mortgagor losses the property if he/she defaults. Transaction costs of default can incorporate a wide range of monetary and non-monetary things, including moving expenses, legal fees, a negative impact on the borrower's credit quality, mental stigma, and so on.

[19] Assume that the new mortgage starts after the current time interval. Deadweight cost is a transaction cost for refinancing; it may be a fixed amount plus a variable amount as a percentage of outstanding mortgage balance.

[20] The equation for the final time interval is the boundary condition; it has a slightly different form and is simpler as housing cost during the last month before mortgage matures is cost of default (property value plus transaction costs of default) or periodic mortgage payment, whichever is less.

[21] One way to understand the solution technique is to draw reference from binomial option pricing model.

The model can be used in at least three ways. First, it can be used to generate an estimation of the default probability for a given loan over a chosen time horizon. Secondly, the model can be used to generate probabilities of default assuming differentiated values for a particular parameter (for example, higher house price volatility versus lower house price volatility) to assess the impact of that parameter on the default probability. Lastly, the model can be used to simulate the performances of hypothetical mortgages; this simulated sample can then be used to test the robustness of a statistical model of default.

Advantage and disadvantage

An optimization model of default differs from previous models. Model 1, 2, and 3 are statistical models that reduce the economic structure behind the mortgage default process; they use statistical properties inherent in loan data to draw inferences. An optimization model, on the other hand, tries to tell a story on what happens when a borrower chooses to default, and capture the dynamics using equations. The advantage is that the optimization model does not make the assumption that the default probability is a given function of the explanatory variables. Probability estimation in the model is due to different economic forces that drive the borrower's behavior. However, the model requires extensive programming and is more difficult to implement than previous models. Outcomes of the model rely on assumptions made to construct the model. Here, assumptions are made on how house prices and interest rates evolve over time. Bad assumptions lead to poor predictions of default probability. Also due to its reliance on certain economic structures, the model is less flexible. For example, the specification described above mainly incorporates impacts of house prices and interest rates. If one wants to add in the impact of GDP growth, it is not easy to do.

Another advantage of this model is that its ability to estimate default probability relies more on the economic structure and less on the historical data. It can be useful when one has a poor collection of loan data. The model estimates the default probability for a "typical" mortgage; "typical" is characterized by a set of initial values assigned as model inputs, which include parameters in house price and interest rate processes, and parameters in housing cost functions. Depending on the purpose of the estimation, one can vary the set of initial values and generate the default probability for a particular loan, the median loan in a portfolio, or a stressed scenario.

The model described by equation (6) may be criticized for not considering the borrower's ability to continue making periodic mortgage payments. The borrower may be forced to default because of insufficient cash to meet mortgage obligations. One justification is that the ability-to-pay is accounted for by the inclusion of the "trigger event". Also, the ability-to-pay, perhaps more precisely inability-to-pay, can be implicitly accounted for by the transaction costs of default. For example, for a borrower who is financially distressed, default on mortgage relieves the borrower of an unaffordable financial burden, which can be reflected in the costs of default. One possible modification that directly deals with the ability-to-pay is to introduce another stochastic process for income (net of non-housing expenditures) or non-housing wealth, revising the borrower's decision by incorporating the liquidity constraints.

Another group of optimization models falls under a utility-maximization framework that is often used in economics to model household consumption. This type of model defines household utility as a function of non-durable consumptions over time, housing consumptions over time and/or terminal wealth (financial wealth and housing wealth). At one point in time, a borrower's housing and mortgage decisions are outcomes resulting from maximizing expected lifetime utility.[22] These models are highly structured; they make assumptions that balance model tractability and its accuracy in describing consumer behaviors as well as housing and mortgage market practices.

IV. Models for default probability of a loan portfolio

Models in section III treat an individual mortgage as the subject of study. In this section, one would view a portfolio of mortgages as one subject.

Model 5: Linear regression analysis of default rate

Description

Regression analysis looks for the relationship between default risk and an array of explanatory variables. Default risk is treated as a dependent variable, which can be explained by some independent or explanatory variables. For a portfolio of loans, the default rate is calculated as the

[22] Two current working papers, Garriga and Schlagenhauf ([18]) and Campbell and Cocco ([5]), are examples of structural models for household mortgage decisions. The former study emphasizes the multiplier effect of leverage in increasing default risk; leverage position of a mortgagor is measured by the loan-to-value ratio. The latter offers a dynamic model incorporating income, house price, inflation, and interest rate risks. Both studies provide ample implications for an empirical model of default risk.

number of loans in default over the total number of loans in the portfolio.[23] The *default rate* in turn serves as a measure of default risk for the loan portfolio. The regression model is formulated as:

$$default\ rate = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k + \varepsilon \quad (7)$$

where $X_1, X_2, \ldots X_k$ are explanatory variables, factors or predictors that may help determine default risk; α is a constant; $\beta_1, \beta_2, \ldots, \beta_k$ are coefficients that capture the impact that each factor may have on default risk; and ε is an error term.

Implementation

When viewing a loan portfolio as one subject, the question arises as to how explanatory variables in equation (7) are measured. For example, one may use the loan-to-value (LTV) ratio as one predictor of default risk; each mortgage in the portfolio has a LTV. Then the question is how to measure the LTV for the portfolio. Broadly speaking, there are two ways to construct the sample.

1. A particular lending institution may consider its entire mortgage portfolio as one subject under examination. The average or median measures for the explanatory variables (e.g., LTV) may be used in the analysis. Periodically, one observes the default rate and explanatory variables for the entire portfolio over time. An example of data looks like Figure 7.

Figure 7. Time series data on a loan portfolio: data structure example

Date	Default rate	X_1: average loan-to-value	X_2: average term of mortgage	X_3: GDP growth
2005	1%	52%	21	2%
2006	2%	53%	22	3%
2007	3%	55%	22	1.5%
……	……	……	……	……

When using average portfolio measures, one should bear in mind the variations in the sizes of loans in the portfolio. Instead of using a simple average LTV, one may use

[23] Alternatively, one may calculate the default rate as loan value in default over total value of loan portfolio. An IMF working paper by Hardy and Schmieder ([20]) suggests that credit loss rate (dollar loan loss from profit and loss account over total dollar of loan stock) account for both probability of default and loss given default.

weighted average by selected weighting factors. For example, one may calculate the average LTV weighted by individual loan sizes relative to portfolio size. Forming a sample this way is simple and straightforward. As a result, the institution gets a time series dataset on its entire mortgage portfolio as a whole. A disadvantage of this approach is that certain explanatory factors (e.g., LTV) are smoothed out over time due to averaging and the impacts of these variables are not well estimated.

2. Another way to construct the sample is to group the entire mortgage portfolio into smaller sub-portfolios and view each sub-portfolio as a subject under examination. Mortgages in one sub-portfolio share some common combination of characteristics. The characteristics are criteria used to group loans; they also enter the regression equation as explanatory variables. For example, one may use two criteria to group mortgages: 1) loan term of 20 or 25 years; and 2) initial LTV of above 90 per cent, between 80 and 90 per cent, or below 80 per cent. As shown in Figure 8, using this grid one sorts all mortgages in the loan portfolio into 6 layered groups, or cohorts, each of which is a sub-portfolio under examination.

Figure 8. Layered groups: example

	LTV > 90%	**80% < LTV < 90%**	**LTV < 80%**
Term: 20 years	Loan portfolio 1	Loan portfolio 3	Loan portfolio 5
Term: 25 years	Loan portfolio 2	Loan portfolio 4	Loan portfolio 6

The grouping technique transforms loan-by-loan data into a cohort-by-cohort sample. Each cohort (sub-portfolio) is then treated as one subject, and is observed in each period. The more criteria one adds to the grid and the finer one defines the grid, the more sub-portfolios one has in the sample. Suppose one adds a third criterion, borrower's income – above median or below median, this increases the number of sub-portfolios from 6 to 12. Figure 9 is an example of data structure using this approach.

Figure 9. Panel data on loan portfolios: data structure example

Loan portfolio ID	Date	Default rate	X_1: LTV>90%	X_2: 80%<LTV<90%	X_3: LTV<80%	X_4: term of mortgage	X_5: GDP growth
1	2009	1.5%	1	0	0	20	1.5%
1	2010	2.2%	1	0	0	20	1.2%
2	2009	2.3%	1	0	0	25	1.5%
2	2010	2.8%	1	0	0	25	1.2%
3	2009	1.0%	0	1	0	20	1.5%
3	2010	1.2%	0	1	0	20	1.2%
4	2009	4%	0	1	0	25	1.5%
4	2010	6%	0	1	0	25	1.2%
……	……	.....	……	……	……	……	……

Instead of using dummy variables for LTV, we can also use average or weighted average LTV as described earlier. The data structure then looks like Figure 10.

Figure 10. Panel data on loan portfolios: data structure example

Loan portfolio ID	Date	Default rate	X_1: average LTV	X_2: term of mortgage	X_3: GDP growth
1	2009	1.5%	91.0%	20	1.5%
1	2010	2.2%	91.0%	20	1.2%
2	2009	2.3%	90.5%	25	1.5%
2	2010	2.8%	90.5%	25	1.2%
3	2009	1.0%	87.3%	20	1.5%
3	2010	1.2%	87.3%	20	1.2%
4	2009	4%	85.0%	25	1.5%
4	2010	6%	85.0%	25	1.2%
……	……	.....	……	……	……

Advantage and disadvantage

Using a loan-portfolio dataset, one can use the default rate as a measure of default risk. This cannot be achieved in an individual-loan dataset. Also, the construction of cohorts helps alleviate collinearity between the explanatory variables in ungrouped data. Similarly to the regression analysis for individual loans, the model can be used to rank loan portfolios by predicted default rates. However, the predicted default rates from this model may fall outside of the range between

0 and 1, which cannot be understood as percentages of loans in a portfolio that is estimated to default. In order to ensure that the predicted default rates fall between 0 and 1, one can use a censored regression (a two-limit Tobit model with a upper limit of one and a lower limit of zero) to estimate equation (7).[24]

Model 6: Linear regression analysis of log odds

Description

When the dependent variable is a probability, a linear relationship between the dependent and explanatory variables is generally inappropriate. Instead, one can specify the natural logarithm of odds ratio as a linear function of the explanatory variables:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k + \varepsilon \qquad (8)$$

where π is the probability of default for the mortgage portfolio under examination. π is not the probability that the entire portfolio defaults. It should be interpreted as the average default probability of mortgages in the portfolio, or as the number of mortgages that default as a fraction of the total number of mortgages in the portfolio.[25]

Implementation

The two ways of forming loan portfolios and constructing samples we discussed for Model 5 also apply here. To implement this model, we need one more step in the data preparation. For each loan portfolio-date observation, we calculate log odds from the portfolio default rate. This transforms Figure 7 into Figure 11.

[24] Webb ([38]) is an example of estimating default risk with a Tobit model. Another interesting point of Webb's study is the use of a "potential delinquency" measure to proxy for default risk. A loan is potentially delinquent if the borrower is forced to choose between delinquency on mortgage and a reduction in non-mortgage expenses. Possible situations that force the borrower to make such a choice share one common condition - an increase in the mortgage-payment-to-income ratio. So a potentially delinquent loan is one whose mortgage-payment-to-income ratio increases. This idea may be appealing when and where actual delinquency data is not available or there is a lack of default experience historically.

[25] This model draws upon a Bank of Canada working paper by Misina, Tessier, and Dey ([27]). An APRA working paper by Coleman et al. ([8]) also uses a similar formulation to assess relationship between default rate and default determinants on multiple layered portfolios separately.

Figure 11. Log odds of loan portfolio: data structure example

Date	Log odds	X_1: average loan-to-value	X_2: average term of mortgage	X_3: GDP growth
2005	-4.60	52%	21	2%
2006	-3.89	53%	22	3%
2007	-3.48	55%	22	1.5%
……	.....	……	……	……

The estimates of $\beta's$ are used to predict log odds for given values of the explanatory variables from equation (8). The predicted probability of default is then calculated as:

$$\pi = \frac{1}{1+e^{-\log\ odds}} \tag{9}$$

Advantage and disadvantage

This model is a common approach to modeling default rates of loan portfolios. It can generate predictions of default rates for loan portfolios; the predicted default rates are expected to be between zero and one. However, the time horizons of the predictions are restricted by the way that the dataset is constructed. For example, suppose that the sample consists of annual observations, like the one outlined in Figure 11. The predicted default rates are at a one-year horizon. Aggregating predictions for multiple periods going forward yields predicted default rates for longer time horizons. The achievable time horizons are limited to multiples of the sample observation interval, which is one year in this example. Also, aggregating default rates over multiple periods may not be straightforward; one possibility is to assume that the total number of loans stays constant.

V. Default determinants implied from economic theories

Section III and IV outlined statistical methods (Model 1, 2, and 3) and economic methods (Model 4) that are used in empirical studies of mortgage default probability. An appropriate interpretation of results from these models requires careful understanding of the model application in context. The functional forms of statistical models rely little on any economic theory rationalizing default behavior. They treat default risk as functions of the explanatory variables. So far, we have not touched on what these explanatory variables are. This is the focus

of the current section, which intends to stimulate thoughts that lead to the designing of a model best suitable for specific business requirements.

Explanatory variables are the determinants of default. Selections of these variables are drawn from the economic theories explaining the processes and motivations of default. This section discusses such economic theories, and the resultant explanatory variables to consider in modeling the probability of default for residential mortgages. We first present some variables from earlier studies; then we discuss the equity- and cash-flow- theories of default behavior; next, the option-based theory of default behavior is considered; finally, the importance of macroeconomic factors is recognized, especially for portfolio-level studies.

The theories offer alternative explanations to the default process. They do not mean to contradict each other; rather, they emphasize different angles of reasoning. The variables suggested by one theory may not be exclusive of those by another. Later theories often introduce new variables that are not considered by earlier ones.

Appendix 2 summarizes these variables and their effects on residential mortgage default. The discussions in this section and Appendix 2 should not be taken as an exhaustive list of default determinants.

Explaining default in the early days

Why do some mortgages end up defaulting, while others continue performing? Since lenders assess mortgage applications before extending loans, a natural question is whether or not the criteria used in the loan approval process are efficient in weeding out potentially high risk loans. Since lenders also collect information about borrowers through the mortgage application process, one may also ask how certain borrower characteristics are associated with mortgage status. In earlier studies, researchers often use information at loan origination to explain mortgage status realized later in time. These include loan specific measures (e.g., loan-to-value ratio, term to maturity, rate type, and loan purpose), borrower characteristics (e.g., income level, payment-to-income ratio, debt-to-income ratio, occupation, marital status, and number of dependents), and property information (e.g., property type, property condition, and location). Initial loan-to-value

ratio is one of the dominating factors explaining default. Its positive relationship with default risk[26] is first validated by Herzog and Earley ([21]) and confirmed by many studies that follow.

A disadvantage of using only information at loan origination is obvious. Loan specific measures can change over time. For example, as the principal being paid down, the loan-to-value ratio decreases; when house price falls, the loan-to-value ratio increases; the term-to-maturity of a mortgage reduces as the mortgage ages over time. A borrower's default decision is not made at the time of mortgage origination, but rather at a time later down the line when the mortgage is no longer affordable, either because the borrower's ability-to-pay deteriorates or a decline in property value deems the mortgage too expensive. Hence, not only information at loan origination matters in understanding default risk, but also does contemporaneous mortgage profile.[27] A model for default risk should seek motivations from the process involved in the mortgagor's default decision.

Competing theories of default behavior

Theory

There are two competing theories of residential mortgage default behavior: the equity theory of default, and the cash-flow or ability-to-pay theory of default ([22]). Consider a borrower making mortgage decisions at one point in time. The equity theory of default holds that a borrower chooses between continuing servicing the mortgage and defaulting on the mortgage to maximize the equity in mortgaged property. The borrower's equity is either (a) without default, the borrower's equity equals to the property value at that time minus the outstanding mortgage balance; or (b) with default, the borrower's equity is zero. The borrower decides to default if the value of (b) is greater than that of (a) (i.e., when the borrower's equity is negative). Based on the equity theory, the probability of default is then equal to the probability that the property value is less than the outstanding mortgage balance. Alternatively, the ability-to-pay theory of default suggests that a borrower refrains from defaulting as long as the borrower is able to meet the

[26] The initial loan-to-value ratio is positively correlated with default. Higher initial loan-to-value ratio is associated with higher default probability.

[27] The probability of default considering only information at mortgage origination is sometimes referred to as unconditional probability; while the probability taking account for current information (and information evolved since origination) is referred to as conditional probability. A Factor that has a large effect on the unconditional probability of default may not have a significant impact on the conditional probability of default. For example, Capozza, Kazarian and Thomson ([7]) suggest that age-of-mortgage greatly affects the unconditional probability of default, but has little effect on the conditional probability of default once current loan-to-value ratio is used.

periodic mortgage payments. It means that the borrower will continue servicing the mortgage as long as his/her income, net of necessary expenditure, is sufficient to make mortgage payments. According to the ability-to-pay theory, the probability of default is equal to the probability that the borrower's income net of expenditure falls below the periodic mortgage payment amount. Jackson and Kaserman ([22]) formally test these two alternative motivations of mortgage default and find that the equity theory dominates the ability-to-pay theory.

Implication

From the perspectives discussed above, here are some considerations one may have when analysing default probability:

1. Improved ability to forecast property value is important to the successful forecasting of mortgage default. Since all mortgages start with positive equity and mortgage balances decrease over time, negative equity can only occur as a result of declines in the property value. So a model of default should take into consideration the property value over time.
2. A contemporaneous loan-to-value[28] ratio, instead of the ratio at loan origination, may be included. A contemporaneous LTV is calculated using the remaining mortgage balance and an updated market value of the mortgaged property. The remaining mortgage balance can be computed from the loan amortization schedule. Choosing the updated market value of the mortgaged property is an empirical issue, as the value is not directly observed. One may estimate this value using an appropriate house price index, or approximate it from sales data of similar properties. How to construct a contemporaneous LTV is subject to practical considerations.
3. Other housing market variables, such as house price volatilities, may also be considered.
4. One should not look at the equity effect and the cash-flow effect as mutually exclusive. They can both play a role in explaining default risk. Elul et al. ([16]) show that borrower's liquidity constraints have a significant impact on default behavior as well as on negative equity.[29]

[28] Loan-to-value (LTV) ratio is the ratio of mortgage balance over the value of the mortgaged property; it is a measure of the borrower's equity position in the mortgaged property. LTV greater than one means the borrower's equity is negative.

[29] Elul et al. ([16]) use credit bureau information on individual borrowers to measure their liquidity constraints. They find that both negative equity and liquidity constraint have comparable effects on mortgage default. The results are

Option-based theory of default behavior

Theory

Under a contingent-claim framework, a mortgage contract can be viewed as an ordinary debt instrument with various options embedded in it. In particular, the mortgage termination risk is characterized by two options. A borrower's ability to default can be viewed as a put option where the underlying asset is the mortgaged property whose price fluctuates over time; the strike price is the outstanding mortgage balance; and the term to maturity is the remaining life of the mortgage contract. When the borrower defaults (i.e., when the borrower exercises the put option), the borrower is essentially "selling" the mortgaged property to the lender for an amount equal to the outstanding mortgage balance. The borrower would only consider doing so when the property value is lower than the outstanding mortgage balance (i.e., when the borrower has negative equity), because in such a case the borrower would be "selling" the property for a price higher than its worth (i.e., profiting from exercising an in-the-money put option).

Another option available to the borrower is the prepayment option. The borrower's ability to prepay can be viewed as a call option where the underlying asset is the outstanding mortgage, the market value of which vary over time because changing mortgage rates alter the present value of the remaining mortgage payments; the strike price is the book value of the current mortgage, which is the outstanding mortgage balance; and the term to maturity is the remaining life of the mortgage contract. When the borrower prepays (i.e., when the borrower exercises the call option), the borrower is essentially "buying" the mortgage out at a price equal to its book value. Also, the borrower would only consider doing so when current mortgage rate is lower than the original rate, because in such a case the borrower would be "buying" the mortgage at a price below its market value (i.e., profiting from exercising an in-the-money call option).

How does the option theory-based approach enhance our understanding of mortgage default risk?

1. Setting mortgage termination in an option-based framework is especially fruitful for pricing mortgage contracts, although this is not the concern of this paper. The Black-Scholes-Merton formulas allow us to calculate option prices from only a few variables,

particularly relevant as they study mortgages originated in 2005 and 2006 for a sample period through to 2009; these mortgages are likely to have negative equity given the housing crisis in the U.S. during that period.

including loan-to-value ratio, outstanding mortgage balance, interest rate, house price volatility, and the remaining life of the mortgage.

2. In terms of default risk, using this framework and the theory of stochastic calculus, one can solve for the probability of default over time. Kau, Keenan, and Kim ([23]) provide an example of this type of exercise.
3. More importantly, viewing a borrower's right to default as a put option allows us to assess how certain factors affect default risk. For example, an option is more likely to be in-the-money and by a larger amount when the underlying asset price is volatile. When house prices are volatile, the default option is more likely to be in-the-money and default is more likely to be "profitable", thus the probability of default is greater. Also, an option is more likely to be exercised the more it is in-the-money. The current loan-to-value ratio is a good proxy for the degree of moneyness[30]; higher loan-to-value ratio means that the default option is deeper in-the-money. Thus the option theory predicts a higher probability of default for a higher current loan-to-value ratio.[31]
4. Finally, testing the option theory-based framework extends our knowledge on default behavior. If default indeed occurs as predicted by this framework, then a model that includes only variables suggested by option theory should be sufficient in explaining default risk. Otherwise, we should consider additional explanation outside of the option-based framework. Empirical evidence ([17]) uncovers that default is not "ruthless" as suggested by a frictionless option-based model – not all mortgages with negative equity default, while others default even with positive equity. How can we explain the empirical observations that are not captured by the option theory?

Elul ([14]) outlines several reasons why some borrowers do not appear to default as soon as their equities become negative and others actually do default with positive equity. First, default option is an "American option", where borrower can exercise at any time until its maturity.[32] The option theory has shown that it may not be optimal to exercise an American put option as soon as it is

[30] Under the option theory, moneyness is used to describe the relationships between the strike price and the current value of the underlying asset. A call option is in the money when the strike price specified by the call option is below the current value of the underlying asset. A put option is in the money when the strike price specified by the put option is above the current value of the underlying asset.

[31] The current loan-to-value ratio is calculated as the outstanding mortgage balance over the current property value. The outstanding mortgage balance can be calculated using the amortization schedule. The current property value may not be observed easily. Deng, Quigley, and Van Order ([12]) use house price index to estimate for each mortgage at a point in time the probability that the property value is below the outstanding mortgage balance. This estimated probability is used as proxy that the default option is in-the-money.

[32] This is in contrast to a European option, which can be exercised only on maturity date.

in-the-money; option holder may be better off waiting until later to make the exercise decision. In the case of residential mortgages, borrowers may not go through an option evaluation process when making default choices; however, they do assess the probability that the house price may go up or the mortgage rate may go down in the future. The default decision from a borrower's consideration of future house prices is consistent with that implied by the option theory for American put options.

Secondly, a borrower has not only the default option but also a prepayment option. When the borrower exercises the default option, he/she foregoes the prepayment option. Exercise decisions for these two options should be made jointly rather than independently.

Another reason for a borrower's reluctance to default is attributed to the transaction costs of default. Moving expense is an example of transaction costs. The negative impact of default on the borrower's reputation and credit quality is another form of transaction costs; borrowers with default history may have difficulty obtaining loans in the future or may only be able to borrow at a higher interest rate. Psychological stigma is also a transaction cost of default; while some borrowers may feel morally wrong to default, being forced to move out of their homes may cause mental distress to others.

Finally, default may be triggered by personal crisis and events other than negative equity. Examples of "trigger event" include relocation, change of employment status, illness, and change of marital status, among others. It is now widely accepted that the "trigger event" plays an important role in mortgage default; however, it is difficult to incorporate it explicitly into a theoretical model, because such events are subjective to too much individual dissimilarity. The significance of trigger events offers support to the cash-flow or ability-to-pay theory that borrowers default when they are liquidity constrained, meaning that they do not have the financial resources to cover mortgage payments and they cannot borrow freely even when the cash shortfall is only temporary.

Implication

Based on the option-based theory of default and the testing of this theory, here are some directions one may take to model default risk:

1. Measures for the worthiness of exercising the default and prepayment options should be included in a default model. These measures may be calculated from variables including the current property value, the outstanding mortgage amount, the contract interest rate, and the current mortgage rate. How these variables enter a statistical model is subject to practical considerations. For example, some use a current loan to value ratio, with the current property value calculated from scaling the original purchase price by an appropriate house price index; others use more complicated measures that calculate the probability of negative equity. Deng, Quigley, and Van Order ([12]) offer an example of the latter. To incorporate mortgage rates, some use a percentage difference between the current rate and the contract rate, while others use a percentage difference between the present values of payment streams calculated with the current mortgage rate and with the contract mortgage rate.
2. Default determinants are not limited to those variables suggested by the option theory. A borrower's ability-to-pay is of great concern for default risk of individual loans. Some use local unemployment rates or low down-payments as a proxy for a borrower's cash-flow constraint. These measures are typically imperfect, such that the former is taken at a geographical region level, which hardly reflects the situation of a particular individual; and the latter is taken at loan origination, which may indicate a borrower's overall financial resource but the quality of this indication is arguable. Elul et al. ([16]) provide a more direct way to measure a borrower's current liquidity position, using information on the borrower's credit line utilization rate from credit bureau files. A borrower draws down his/her credit line and uses a larger fraction of available credit when facing income flow problems. So an increased utilization rate is likely to be associated with increased default risk. Depending on the information and data available, one may construct other measures of liquidity constraints.

Macroeconomic factors

Theory

So far, the discussion has focused on default behavior from an individual borrower's perspective. The recent financial crisis has highlighted the importance of understanding macroeconomic conditions in managing credit risk. A few studies have looked at the interaction between the

macroeconomic environment and default risk both aggregated and at loan-level.[33] In a top-down approach of investigating default risk for the system as a whole or for a large mortgage portfolio, macroeconomic variables are at the centre of attention; these variables represent systemic variations in the economic environment. Dissimilarities in loan specific and borrower specific factors are non-systemic given sufficient diversification; they are less essential in explaining aggregate default rates over time. The macroeconomic conditions, without doubt, affect aggregated default behavior. For example, interest rate and level of personal disposable income affect borrowers' ability-to-pay; real house price affects borrowers' equity positions in mortgaged properties; and tightening of credit supply imposes liquidity constraints on borrowers. It is reasonable to argue that macroeconomic measures should be included in a mortgage default model. How are they incorporated then?

In modeling default probability as a function of macroeconomic variables, we need to address the issues of dynamic interactions between default probabilities and macroeconomic factors, as well as between two macroeconomic factors. There are two channels through which a macroeconomic factor may impact default: the direct impact of changes in the macroeconomic factor, and the indirect impact through its influences on other macroeconomic factors. Also, the impact of a macroeconomic shock may kick in after time lags or persist for more than one period of time. For example, an increase in the unemployment rate in the first quarter may not be associated with an increase in default rates in the first quarter, but there may be a resultant rise in defaults in the second and third quarter, and the effect may eventually diminish by the fourth quarter. In this hypothetical example, the unemployment rate has an impact on the default rate up to two time lags. Using vector autoregressive (VAR) analysis on a series of default rates and macroeconomic measures over time will offer insights into the dynamics among these variables. A VAR model can be written as:

$$\boldsymbol{X}_t = \varphi_1 \boldsymbol{X}_{t-1} + \varphi_2 \boldsymbol{X}_{t-2} + \cdots + \varphi_p \boldsymbol{X}_{t-p} + u_t \tag{10}$$

where $\boldsymbol{X}_t$ is a vector that contains time t values of the default rate and the macroeconomic variables of interest; the equation represents (a) the relationship between one variable in the

[33] Ali and Daly ([1]), Crook and Banasik ([9]), and Misina, tessier, and Dey ([27]) examine the relationship between aggregate default rates and macroeconomic measures; Bellotti and Crook ([3]) and de Silva Correa and Marins ([11]) are examples of analyzing loan-level default risk with macroeconomic factors as explanatory variables. An IMF working paper by Hardy and Schmieder ([20]) also emphasizes the importance of "overall conjunctural conditions prevailing in the economy".

current time period and itself in the past p time periods, and (b) the relationship between the variable and other variables in the current and past time periods. Estimating the model tells us how default rates respond to shocks in macroeconomic environment over time, which is an interesting result in its own right. The statistical significances of coefficients tell us what macroeconomic variables and their lags to include in a default risk model. Consider the hypothetical example of unemployment rate again. Suppose that estimating equation (10) suggests that the unemployment rate has an impact on default up to two time lags. We want to include unemployment rate as an explanatory factor for default probability, say, using a model like the one specified by equation (8). The VAR analysis tells us that for current period's default risk we should consider unemployment not only in the current period, but also in the previous two periods. In stress testing, estimated coefficients, $\varphi's$, can also be used to simulate future realizations of macroeconomic variables for a given set of initial values.

Implication

Macroeconomic measures are potential default determinants, especially at portfolio level. Here are some important considerations:

1. Macroeconomic measures should be included in a default model, especially when one examines default rates of loan portfolios. Examples of such measures include GDP and GDP growth, interest rates, unemployment rate, housing market index, debt-to-GDP ratio, population growth, and so on.
2. Other measures, such as industrial productions, immigration policy and foreign investments, may also be relevant given the natures of the loan portfolios and the economic region under examination. For example, natural resource is one of the key industry sectors in BC; so the performance of this industry cuts into housing and labor markets in this province, which in turn may have some influence on the default rates of the mortgage portfolios of lending institutions whose asset books may be heavily concentrated in this area.
3. When using macroeconomic measures in default models, one should consider time lags. Lagged values of these measures may be appropriate.

4. Industry lending practices do have consequences. The financial crisis episode around 2007 tells the story that private securitized loans turn out to be riskier.[34] An institution would know best whether or not and to what degree securitization affects its own lending practice. If securitization is indeed a factor, it would also be used as a default predictor.

VI. Issue of model stability

One reason to estimate a statistical model is to predict default probabilities or default rates going forward. Estimations of a statistical model utilize historical data on mortgages and their default experience to quantify the relationship between default risk and default determinants. When estimations based on historical data are used to forecast default risk in the future, it is implicitly assumed that the relationship between default risk and its determinants in the future resembles the relationship between the two in the past. In occasions when this assumption is unsound, the statistical model breaks down – it generates inaccurate predictions of defaults. This phenomenon was widespread during the recent financial crisis; models that use historical data to estimate coefficients and predict defaults performed poorly in the period from 2007 onwards.[35] Rajan, Seru, and Vig ([30]) offer one reason for the failure of these statistical models – they rely on "hard information" about borrowers to predict defaults and ignore changes in lenders' incentives to collect "soft information" during the loan approval process.[36] Two borrowers who present the same "hard information" may differ in terms of "soft information"; and "soft information" is potentially important in driving default choices. The exponential growth in securitization during the 2000's imposes severe moral hazard problems in mortgage markets. The originator of a loan has less incentive to collect "soft information" about the borrower because securitization distances the originator from investors who actually bear the default risk. This suggests that there is a structural break in mortgage markets. Changes in lenders' incentive caused by securitization lead to changes in the nature of approved loans; borrowers who would have been denied credit in a low-securitization regime now are able to obtain loans in a high-securitization one – the average quality of mortgage borrowers worsens even though "hard information", such as credit score, appears the same. As a result, the relationships between explanatory variables representing

[34] See Elul ([15]) for a discussion on why securitized loans might be risker. Also see Keys et al. ([24]) for empirical evidences of adverse selection in securitized loan markets.
[35] An et al. ([2]) assess the model instability problem during the recent crisis. Rajan, Seru and Vig ([30]) note that in 2007 Standard & Poor's adjusted its default model to increase predicted defaults on no-documentation loans by approximately 60 percent.
[36] "Hard information" includes measures such as the loan-to-value ratio and the borrower's FICO credit score. "Soft information" is unverifiable to a third party; the borrower's income risk is an example.

"hard information" and credit quality of mortgage (default probability) change. Consequently, a statistical model using loan samples from a low-securitization era under-predicts default rates in a high-securitization era.

What this says is that statistical models themselves are fine as long as they are used in accordance with model assumptions; the problem is how the models are applied and how the outcomes are interpreted. Results from a model are shaped by assumptions made by the model, and a model breaks down when those assumptions are violated. This tells us that instead of using a model blindly, one should be aware of changes in industry standards and lending practices to identify whether or not such a systemic break as described above is prevalent. One needs to scrutinize and assess appropriateness before applying any model.

Some possible remedies are proposed[37] to ease the potential model instability issue. First, using a larger sample size over a longer history may improve the coefficient estimations. Secondly, use a rolling window of sample period, instead of a static sample period, to incorporate as much recent information as possible.[38] This may reduce prediction errors but not completely eliminate them. Thirdly, incorporate forward-looking macroeconomic variables like forecast of house price index. However, bad forecasts of these variables further deteriorate the predictive accuracy of a statistical model. Fourthly, bring market signals in the statistical model. One example of market signals is market price of loans. Using market signals attempts to capture information from market participants who may have an information advantage.[39] Finally, a structural model that accommodates changes in relevant sectors of the economy may generate fruitful results.

VII. Conclusion

This paper discusses six models that can be used to assess the default risk of residential mortgages. Model 1, 2, 3, and 4 look at default probabilities of individual loans; Model 5 and 6 turn to default rates of loan portfolios. Appendix 1 offers a quick summary and comparison of these models. The paper also combs through the development of economic theories that attempt to explain a borrower's default behavior. Based on these theories we gather a list of factors that

[37] They are suggested by Rajan, Seru, and Vig ([30]) and An, Deng, Rosenblatt, and Yao ([2]).

[38] A rolling window specification involves: to predict default probabilities on loans issued in year t, use a loan sample over year 1 through t-1 to fit a statistical model and estimate the coefficients. For example, to predict default probabilities of loans issued in 2009, use loans issued up to 2008 to fit a model; to predict default probabilities of loans issued in 2013, use loans issued up to 2012 to fit a model. Refer to Rajan, Seru, and Vig ([31]) for details.

[39] Market price reflects the perceived risk of an asset and investors' expectation of the future given that investors are rational and capital markets are efficient.

potentially drive default. The factors and their impacts on default are assembled in Appendix 2. Finally, we emphasize the importance of understanding industry phenomenon while applying any model. We want to point out that there are limitations to this paper. One may use it as a well-informed starting point. Should a model be selected, further investigation is strongly recommended to fully comprehend all practical aspects.

Appendix 1.a: Overview of models

	Subject of Study	Description	Pro	Con
Model 1: Linear regression analysis on default risk	Individual mortgages	Estimate default risk as a linear combination of various factors	• Simple to apply • Easy to interpret results • Can rank or classify mortgages	• Default probability predictions may fall outside of a sensible range
Model 2: Logistic model	Individual mortgages	Estimate default probability as a logistic function of a combination of various factors	• Appropriate for qualitative response data • Predict probability of default • Efficient yet easy to implement and interpret	• Default probability predictions are for a fixed time horizon
Model 3: Survival analysis	Individual mortgages	Use time-to-event methodology to estimate the relationship between default probability and the passage of time along with other factors	• Match life course of mortgages • Default probability predictions are for flexible time horizons • Adjust for censored data • Avoid certain bad model assumptions using semi-parametric estimation process	• Data preparation can be complicated • Require some programming, especially for post estimation analysis
Model 4: Optimization model	Individual mortgages	Estimate default probability by simulating a borrower's decisions on mortgage payment for possible outcomes of house prices and interest rates over time	• Model default behavior from economic forces • Default probability predictions rely less on historical loan data	• Require extensive programming • Wrong assumptions on model inputs lead to bad predictions • Less flexible for additional economic forces
Model 5: Linear regression analysis on default rate	Mortgage portfolios	Estimate default rate as a linear combination of various factors	• Predict default rates • Simple to apply • Easy to interpret results	• Default rate predictions may fall outside of a sensible range
Model 6: Linear regression analysis on log odds	Mortgage portfolios	Estimate log odds as a linear combination of various factors	• Predict default rates • Efficient yet easy to implement and interpret	• Default rate predictions are for a fixed time horizon

Appendix 1.b: Loan level model versus portfolio level model

This appendix discusses when to use a loan level model and when to use a portfolio level model.

1. The answer depends on the purpose of the study. If one wants to assess default risk for individual loans, especially for credit approval or loan pricing, one would use a loan level model. Instead, if an institution's loan loss provision is of concern, default rate of the entire loan portfolio matters more than default probability of each individual loan separately. In such case, a portfolio level model is sufficient.
2. The answer also depends on data availability. For loan level study, mortgage and borrower data are particularly important. Examples of loan specific data include loan interest rate, loan size, age of loan, loan to value ratio, rate type, loan term, and loan purpose. Examples of borrower specific data include income level, income risk, age of borrower, non-housing wealth, credit quality, marital status, dependents, and so on. Also it is better to have current information rather than that at loan origination. At portfolio level, individual variations are of less relevance. Macroeconomic conditions, such as interest rate, GDP growth, unemployment, house price volatility, et cetera, would be of greater necessity. These data are available from various data sources.
3. Another difference lies in the need for aggregating either input or output. For a loan level study, one does not aggregate input data, and the estimation output is loan by loan. After that, one can estimate loan loss for the portfolio as a whole by aggregating estimated loan loss for each individual loan. For a portfolio level study, one aggregates input date before estimation. For example, weighted average of LTV for the portfolio from individual loan LTV is calculated and the portfolio weighted average LTV then enters the estimation process as an input.

The table below summarizes above discussion.

	Loan level models	**Portfolio level models**
Purpose of study	Assess default probability of individual loans	Assess default rate of loan portfolio as a whole
Data	More reliance on borrower and loan data	More reliance on macroeconomic data
Implementation	Input data is not aggregated Output is default probability predictions loan by loan Output can be aggregated for a portfolio loan loss estimation	Input data is aggregated Output is a default rate prediction for the portfolio Do not have loan by loan estimations

Appendix 2: Determinants of residential mortgage default risk

The table summarizes default determinants for residential mortgages.

	Factor	Effect on Default[40]	Evidence
Macro-economic	Unemployment rate	(+)	[3] [6] [11] [13] [28]
	House price volatility	(+)	[28]
	Debt-to-GDP ratio	(+)	[1]
	Personal loan interest rate	(+)	[9]
	House price	(–) default increases when housing prices decline, and the increase is more severe for Graduated Payment Mortgages due to high leverage; the effect also comes with time lag	[9] [18] [19]
	GDP, GDP growth	(–)	[1] [3] [11] [20]
	Population	(–) population is used as a proxy for mortgage market size	[11]
	Personal disposable income	(–)	[9]
	Stock market index	(–) weak significance	[3]
	Interest rate	(*) lower interest rate implied by easing monetary policy leads to more lending to riskier borrowers and increased default; response of default to interest rate change comes with time lag; following an increase in interest, default rate initially falls but subsequently rises	[1] [11] [13]
	Industrial production	(\)	[1]
	Consumer confidence	(\)	[3]
Loan	Loan-to-value ratio	(+) LTV's both at origination and at current time have a positive effect on default; the positive effect kicks in after LTV exceeds certain threshold	[5] [6] [10] [21] [22] [28] [32] [35] [36]
	Loan rate volatility	(+)	[28]
	Loan size	(–)	[13] [32]
	Loan interest rate	(–) increased mortgage rate from loan origination is associated with lower default probability, this negative effect is amplified by high LTV	[6] [7] [10]
	Loan purpose	(*) loans used for cash-out refinancing default more; so do secondary mortgages	[5] [21] [32]
	Age of mortgage	(*) defaults display a rise-then-fall pattern as mortgages age	[6] [8] [10] [36]
	Mortgage term	(*) some find longer term mortgages default more; others find the opposite or the insignificance of mortgage term	[21] [22] [32]
	Loan interest rate	(*) some find higher default is related to higher loan rate; others find that loan rate is insignificant	[9] [13] [22]
	Mortgage rate type	(*) with low initial rates, fixed rate mortgages default less than variable rate mortgages; with high initial rates, it is the reverse; also, when the borrower's income is correlated to interest rate, fixed rate mortgages default more than variable rate ones	[5] [21]

[40] (+): positive relationship between explanatory variable and default; higher probability of default is associated with higher value of the variable.
(–): negative relationship between explanatory variable and default; higher probability of default is associated with lower value of the variable.
(*): significant relationship between explanatory variable and default, but the pattern may be non-monotonic, non-linear or inconsistent among studies, or the relationship does not have a positive or negative interpretation.
(\): insignificant relationship between explanatory variable and default.

Borrower	Income risk	(+) the borrower's income risk increases default and the effect is stronger if the initial loan rate is high	[5] [35]
	Leverage and indebtedness	(+) borrowers who are highly indebted or who utilize a larger fraction of their available credit lines are subject to higher default risk	[16] [26]
	Income	(–) higher income relates to lower default, but the effect is weaker or reversed for Graduated Payment Mortgages	[18] [37]
	Non-housing wealth	(–)	[35]
	Credit quality	(–)	[11] [28]
	Payment-to-income ratio or debt service ratio	(*) some find a positive effect comes in once income is below a certain threshold; some find a negative or insignificant effect; one explanation is that borrowers with high payment-to-income ratio can only obtain loans if they are indeed of low default risk due to rigorous underwriting practices	[5] [10] [21] [28] [32] [35]
	Occupation	(*) this may be category of occupation, duration in current job, or other related measures	[10] [21] [35]
	Dependents	(*) some find that the number of dependents matters; others find it does not	[21] [35]
	Age of borrower	(*) a fall-rise-fall feature across borrower age cohorts; borrower age is at the time when loan status is observed instead of at loan origination	[10] [18] [21]
	Marital status	(\)	[21] [35]
Property	Property condition	(*)	[6] [32] [37]
	Region	(*)	[21]
	Neighborhood	(*)	[26] [32] [35] [37]

References

[1] Ali, Asghar, and Kevin Daly, 2010, Macroeconomic determinants of credit risk: recent evidence from a cross country study, International Review of Financial Analysis 19: 165-171.

[2] An, Xudong, Yongheng Deng, Eric Rosenblatt, and Vincent W. Yao, 2012, Model stability and the subprime mortgage crisis, The Journal of Real Estate Finance and Economics 45-3: 545-568.

[3] Bellotti, T., and J. Crook, 2009, Credit scoring with macroeconomic variables using survival analysis, The Journal of the Operational Research Society 60-12: 1699-1707.

[4] Buis, Maarten L., 2006, An introduction to survival analysis, working paper from author's website http://maartenbuis.nl/wp/survival.html.

[5] Campbell, John Y., and Joao F. Cocco, 2014, A model of mortgage default, working paper.

[6] Campbell, Tim S., and J. Kimball Dietrich, 1983, The determinants of default on insured conventional residential mortgage loans, The Journal of Finance 38-5: 1569-1581.

[7] Capozza, Dennis R., Dick Kazarian, and Thomas A. Thomson, 1996, The conditional probability of mortgage default, Real Estate Economics 26-3: 359-389.

[8] Coleman, Anthony, Neil Esho, Ilanko Sellathurai, and Niruba Thavabalan, 2005, Stress testing housing loan portfolios: a regulatory case study, Australian Prudential Regulation Authority working paper.

[9] Crook, Jonathan, and John Banasik, 2012, Forecasting and explaining aggregate consumer credit delinquency behaviour, International Journal of Forecasting 28: 145-160.

[10] Cunningham, Donald F., and Charles A. Capone, Jr., 1990, The relative termination experience of adjustable to fixed-rate mortgages, The Journal of Finance 45-5: 1687-1703.

[11] Da Silva Correa, Arnildo, Jaqueline Terra Moura Marins, Myrian Beatriz Eiras das Neves, and Antonio Carlos Magalhaes da Silva, 2011, Credit default and business cycles: an empirical investigation of Brazilian retain loans, Banco Central Do Brasil working paper series 260.

[12] Deng, Yongheng, John M. Quigley, and Robert van Order, 2000, Mortgage terminations, heterogeneity and the exercise of mortgage options, Econometrica 68-2: 275-307.

[13] Divino, Jose Angelo, Edna Souza Lima, and Jaime Orrillo, 2013, Interest rates and default in unsecured loan markets, Quantitative Finance 13-12: 1925-1934.

[14] Elul, Ronel, 2006, Residential mortgage default, Business Review Q3 2006: 21-30.

[15] Elul, Ronel, 2010, What have we learned about mortgage default? Business Review Q4 2010: 12-19.

[16] Elul, Ronel, Nicholas S. Souleles, Souphala Chomsisengphet, Dennis Glennon, and Robert Hunt, 2010, What "triggers" mortgage default?, The American Economic Review 100-2, 490-494.

[17] Foster, Chester, and Robert van Order, 1985, FHA terminations: a prelude to rational mortgage pricing, AREUEA Journal 13-3: 273-291.

[18] Garriga, Carlos, and Don E. Schlagenhauf, 2010, Home equity, foreclosures, and bail-out programs during the subprime crisis, working paper.

[19] Hamilton, Dan, Rani Isaac, and Kirk Lesh, 2010, Using aggregate time series variables to forecast notices of default, Business Economics 45-1: 8-15.

[20] Hardy, Daniel C., and Christian Schmieder, 2013, Rules of thumb for bank solvency stress testing, IMF working paper.

[21] Herzog, John P., and James S. Earley, 1970, The major determinants of differential mortgage quality, NBER http://www.nber.org/books/herz70-1.

[22] Jackson, Jerry R., and David L. Kaserman, 1980, Default risk on home mortgage loans: a test of competing hypotheses, The Journal of Risk and Insurance 47-4, 678-690.

[23] Kau, James B., Donald C. Keenan, and Taewon Kim, 1994, Default probabilities for mortgages, Journal of Urban Economics 35: 278-296.

[24] Keys, Benjamin J., Tanmoy Mukherjee, Amit Seru, and Vikrant Vig, 2010, Did securitization lead to lax screening? Evidence from subprime loans, The Quarterly Journal of Economics 125-1: 307-362.

[25] McFadden, Daniel, 1973, Conditional logit analysis of qualitative choice behavior, Frontiers in Econometrics, New York: Academic Press.

[26] Mian, Atif, and Amir Sufi, 2009, The consequences of mortgage credit expansion: evidence from the U.S. mortgage default crisis, The Quarterly Journal of Economics 124-4: 1449-1496.

[27] Misina, Miroslav, David Tessier, and Shubhasis Dey, 2006, Stress testing the corporate loans portfolio of the Canadian banking sector, Bank of Canada working paper 2006-47.

[28] Quercia, Roberto G., Anthony Pennington-Cross, and Chao Yue Tian, 2011, Mortgage default risk and local unemployment, working paper.

[29] Quercia, Roberto G., Michael A. Stegman, 1992, Residential mortgage default: a review of the literature, Journal of Housing Research 3-2: 341-379.

[30] Rajan, Uday, Amit Seru, and Vikrant Vig, 2010, Statistical default models and incentives, The American Economic Review 100-2: 506-510.

[31] Rajan, Uday, Amit Seru, and Vikrant Vig, 2012, The failure of models that predict failure: distance, incentives and defaults, working paper.

[32] Sandor, Richard L., and Howard B. Sosin, 1975, The determinants of mortgage risk premiums: a case study of the portfolio of a savings and loan association, The Journal of Business 48-1: 27-38.

[33] Souissi, Moez, 2007, An approach to stress testing the Canadian mortgage portfolio, Bank of Canada Financial System Review – December 2007.

[34] Vandell, Kerry D., 1993, Handing over the keys: a perspective on mortgage default research, Journal of the American Real Estate and Urban Economics Association 21-3: 211-246.

[35] Vandell, Kerry D., and Thomas Thibodeau, 1985, Estimation of mortgage defaults using disaggregate loan history data, Real Estate Economics 13-3: 292-316.

[36] Von Furstenberg, George M., 1969, Default risk on FHA-insured home mortgages as a function of the terms of financing: a quantitative analysis, The Journal of Finance 24-3: 459-477.

[37] Von Furstenberg, George M., and R. Jeffery Green, 1974, Home mortgage delinquencies: a cohort analysis, The Journal of Finance 29-5: 1545-1548.

[38] Webb, Bruce G., 1982, Borrower risk under alternative mortgage instruments, The Journal of Finance 37-1: 169-183.